Exploiting Multi-view Part-wise Correlation via an Efficient Transformer for Vehicle Re-Identification

Ming Li, Jun Liu *Member, IEEE*, Ce Zheng, Xinming Huang *Member, IEEE*, Ziming Zhang* *Member, IEEE*,

***Abstract*—Image-based vehicle re-identification (ReID) has witnessed much progress in recent years thanks to the advances of deep neural networks. However, most of existing works struggled to extract robust and discriminative features from a single image at each feedforward to represent one vehicle instance. We argue that images taken from distinct viewpoints, *e.g.,* front and back, have significantly different appearances and patterns for recognition. In order to "memorize" each vehicle, existing models often have to capture consistent "ID codes" from totally different view images, which causes learning difficulties. Additionally, we claim that correspondences among views, *i.e.,* various vehicle parts observed from the identical image and the same part observed from different viewpoints, contribute to instance-level feature learning as well. Motivated by these observations, we propose to extract comprehensive instance-specific representations of the same vehicle from multiple views through modelling part-wise correlations. To this end, we present an efficient transformer-based framework to exploit both inner- and inter-view correlations for vehicle ReID. In specific, we first adopt a deep encoder to condense a series of patch embeddings from each view image. Then our efficient transformer, consisting of a distillation token and a noise token in addition to a regular class token, is constructed to enforce all patch embeddings to interact with each other regardless of whether they are taken from identical or different views. For inference, one testing image together with its augmented counterparts (pseudo views) are regarded as multi-view inputs and fed into our framework to capture its representations. We conduct extensive experiments on widely used vehicle ReID benchmarks, and our approach achieves the state-of-the-art performance, showing the effectiveness of our method.**

***Index Terms*—Vehicle Re-identification, Transformer, Multi-view Learning, Correlation Exploiting.**

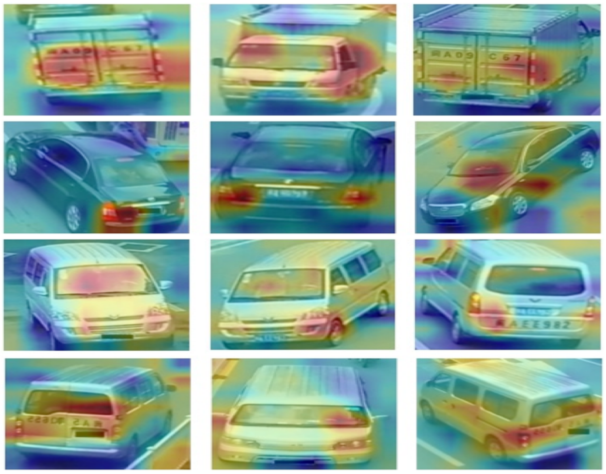


Fig. 1: Part-wise correlation learning results across input views of our approach. Our framework aims to learn instance representations for vehicle ReID from multiple views simultaneously, by modelling patch-wise inter-dependencies. Images shown in each row are triplet inputs and taken from the same individual. They are covered with extracted feature maps for visualizing the high response regions, and red regions indicate regions-of-interest (ROIs) for identifying corresponding vehicles. Our framework successfully concentrates on various informative vehicle parts from multi-view images. Furthermore, different locations on one view (*i.e.,* inner-view correlation) and the same parts from multiple views (*i.e.,* inter-view correlation) are exploited by our approach.

I. Introduction

Vehicle re-identification (ReID) aims to associate vehicle images captured from a camera network spreading across a variety of traffic scenarios, *e.g.,* living areas and highways, in a large city [1, 2, 3, 4]. This task is pretty challenging mainly due to the following factors. First, there are often only subtle valid discrepancies in the same view, which can be used to distinguish vehicles with the same model [5, 6]. Second, differences between various viewpoint images of the same instance are so large that ReID models have to focus on significantly different patterns, *e.g.,* brands on front or back views but wheels on side images [7, 8], at each feeding.

Given these challenges, most researchers have attempted to redevelop official datasets and provide more supervisions in addition to vehicle identities, to direct extraction of robust and discriminative features. For example, in existing works, view segmentation [6], vehicle body keypoints [8, 9], orientations [8, 9, 10] or informative parts [7] were manually annotated and then another auxiliary network was optimized to perform the corresponding recognition task for guiding ReID representation learning. While performing well, they inevitably involved intensive human efforts, which constrains the applicability of these approaches. Besides, some studies [11] suggested adopting a self-supervised manner to encourage discriminative local patterns discovery for vehicle ReID. Other

M. Li is with Institute of Data Science, National University of Singapore, Singapore. E-mail: ming.li@u.nus.edu. This work was done when he visited the VISLab at WPI.

J. Liu is with Information Systems Technology and Design, Singapore University of Technology and Design, Singapore. E-mail: jun_liu@sutd.edu.sg.

C. Zheng is with Department of Computer Science, University of Central Florida, USA. E-mail: cezheng@knights.ucf.edu.

X. Huang is with Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, USA. E-mail: xhuang@wpi.edu.

*Corresponding author: Z. Zhang is with Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, USA. E-mail: zzhang15@wpi.edu.

researchers investigated this challenging task through presenting new metric learning methods [10, 12, 13, 14]. However, all these works narrowed their methods to capturing ReID representations from a single image.

As stated before, vehicle images taken from different viewpoints vary significantly, so do patterns for recognition. As a result, existing models have to embed instance-related features from only one randomly selected view image at each feeding. That is, ReID models are supposed to encode as similar representations as possible for one vehicle instance but from different perspective images during the course of optimization process. To adapt to recognition from various viewpoints, model parameters often oscillate a lot, causing instability during training and thus hurting final performances. Some works [15, 16] in the ReID literature have observed this phenomenon, which however has not been investigated as our perspective. On the other hand, only seeing a single image captured from one viewpoint is not what people actually do when distinguishing a vehicle in practice. They usually search around a vehicle body to discover distinguishable clues for recognition from different angles. These multi-view information is then integrated for recognizing the specific instance. Inspired by these, we propose to learn vehicle representations from multiple view images of each instance at the same time.

It is known that vehicle parts usually contain informative and even discriminative details for identifying an individual. Many works proposed their own strategies of encouraging deep models to concentrate on these parts from a single image [7, 8, 9, 11]. In our work, we expect that our framework is able to discover informative parts from multi-view inputs. However, from Figure 1 we can see there are several discriminative parts in one view and some important vehicle parts are visible from different viewpoints. In this case, encoding high-level features from each view separately and then fusing them together simply is surely an intuitive but relatively coarse method, which ignores the correlations among inner- and inter-view parts and therefore hurts the quality of extracted instance representations. Inspired by recent advances in transformers [17, 18, 19], we propose to employ a transformer-based framework to automatically model global part-wise correlations existing in multiple views. Systematically, due to lack of prior knowledge on image properties, *e.g.,* localization, rotation and shifting invariance in its principal design, the optimization of a popular visual transformer often needs a massive dataset [19, 20]. It is unavailable for most vehicle ReID benchmarks. To address this issue, we propose a customized transformer, which does not require any pretraining on ImageNet-scale dataset [21] and thus can be trained easily and efficiently on vehicle ReID benchmarks.

Note that [22] also attempted to extract representations from multiple vehicle images, while our work is significantly different from theirs. They randomly selected a couple of images belonging to the same vehicle and treated them as a frame sequence. Then they used Long Short Term Memory (LSTM) network to extract and aggregate image-level features for vehicle ReID. However, these images are taken from random viewpoints rather than coherent angles, whose confusing correspondences can not be reasonably represented by LSTM in principle. This does not matter for our transformer-based framework. Besides, they encoded representations from each view independently and simply combined image-level representations together for vehicle ReID, which took neither local features nor their correspondences (across views) into account. In our work, nevertheless, an efficient transformer is presented to learn part-level correlations across views.

To be more specific, we first employ a convnet to extract structural patch representations from each view image. These embeddings are incorporated with a class token, a distillation token, and a noise token to comprise the input of stacked transformer layers. The distillation token is proposed to distill knowledge from the convnet to the transformer for facilitating its learning. And the noise token is designed to encourage classification entropy maximization for preventing the overfitting when training transformer on a ReID dataset. It is worth noting that our efficient transformer is learned from scratch and does not rely on the pretraining on a large-scale dataset, which suggests that ours is significantly different from and much more challenging than another concurrent work on transformer-based ReID [23]. It is on the basis of fine-tuning a pretrained visual transformer on vehicle ReID benchmarks and also aims to condense vehicle representations from a sing view like other works. To summarize, our main contributions are as follows:

- We are the first to present a transformer driven framework to capture comprehensive instance codes from multiple view images for vehicle ReID.
- The proposed transformer is able to exploit part-wise correlations across views successfully.
- Our efficient transformer, containing a distillation token and a noise token, can be trained from scratch on vehicle ReID datasets with no need of sophisticated pretraining on ImageNet.
- Our approach outperforms recent state-of-the-art (SOTA) works on popular vehicle ReID benchmarks.

II. Related Works

Vehicle ReID. The majority of recent works in this field struggled to extract robust global and discriminative local representations under guidance of extra supervisions as well as ID labels [6, 7, 8, 10, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Typically, recognizing vehicle attributes, *e.g.,* color and model [28, 29, 30, 31, 32, 33], was incorporated into multi-task optimization scheme to regularize representation learning. [7] located informative parts of vehicles, such as the front or back window and the license plate, by human labor and trained a separate YOLO network to predict the positions of these parts as regions-of-interest (ROIs) of the ReID model. Similarly, ground truths of view segmentation were provided in [6] and a U-Net semantic parser was optimized to segment vehicle bodies for further view-aware feature alignment. Instead, several studies sought self-supervised approaches for enforcing ReID models to focus on distinguishable vehicle details [11]. Additionally, metric learning is also an integral section of ReID literature [10, 12, 13, 14]. While, our work is distinct from all these investigations. We aim to learn instance-specific

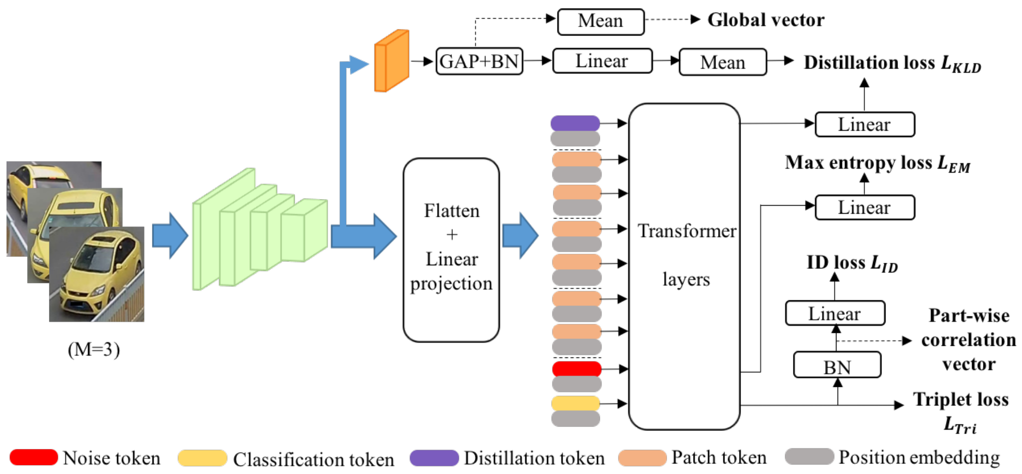


Fig. 2: Overview of our framework, which aims to learn instance-distinguishable representations from M views by modelling part-wise dependencies. In this figure, we take $M = 3$ as an example for illustration. The framework is composed of two modules: the convnet encoder and the transformer. BN and GAP are short for Batch Normalization and Global Average Pooling, respectively. The dashed lines indicate data flows that only run during the inference phase, where a testing image and its augmented versions serve as pseudo multi-view inputs.

representations from multiple views by modeling global correlations among vehicle parts. Significantly, our proposal does not require extra supervisions except from vehicle identities.

Visual Transformer. Transformer based on Multi-Head Self-Attention (MHSA) originates from [17] and shows its power on modelling long-range dependencies in Natural Language Processing (NLP) [18]. By dividing an image into a series of patches and feeding patch embeddings into a stack of MHSA layers, ViT [19] introduced transformer into visual recognition successfully. However, due to the lack of prior knowledge on intrinsic image properties in its architecture design, the performance of ViT on ImageNet depended on the heavy pretraining on a tremendous private dataset, which was unfeasible for most researchers. DeiT [20] proposed to optimize a transformer on ImageNet efficiently through knowledge distillation from a convnet teacher to the student transformer. Similarly, to facilitate transformer learning, T2T-ViT [34] proposed to take local rigid information of images into account. They presented a tokens-to-token transformer which merged structured neighboring tokens into one token gradually. Motivated by the success of residual network structure in deep learning, Bottleneck Transformer [35] simply replaced the 3×3 convoluational layers of the last group in ResNet [36] with MHSAs to capture global dependencies among high-level features. Except from image classification, visual transformers also demonstrate desirable potentials in other computer vision areas, *e.g.,* semantic segmentation [37, 38], object detection [38, 39], point cloud computing [40], and medical image analysis [41].

In contrast, our transformer is devised for a fine-grained and open-set classification problem, namely vehicle re-identification. It is much more difficult, which is because of not only disjoint classes between the training and testing set but also the quite smaller scale of training set (usually containing approximate ten thousands samples). Concurrently, [23] also attempts to employ a visual transformer to address ReID problem. Nevertheless, their work relies on fine-tuning an ImageNet-pretrained transformer on an object ReID benchmark and does not pay attention to optimizing it directly on a small ReID dataset. As known to all, the latter is much more challenging. Our transformer differs significantly from all existing equivalents in terms of network width, depth, and tokens. More importantly, ours can be optimized on a ReID dataset efficiently and easily, with random parameter initialization.

III. Transformer Driven Inter- and Inner-View Part-wise Correlation Learning

As shown in Figure 2, our framework consists of two phases: condensing a sequence of patch embeddings from multiple view images by the convnet encoder and afterwards aggregating instance representations through patch-wise interacting in the transformer. In the remaining of this section, we introduce each key component in turn.

A. Preliminaries

Given a query image, vehicle ReID is to obtain a ranking list of all gallery images according to similarity scores between query and each gallery image. The similarity score is usually calculated from deep embeddings, *i.e.,* $cos(f(x_q;\boldsymbol{\theta}),\ f(x_g;\boldsymbol{\theta}))$. Here $f(\cdot;\boldsymbol{\theta})$ represents a deep feature network with the learnable parameter set $\boldsymbol{\theta}$; x_q, x_g are a query and gallery image respectively; $cos(\cdot)$ denotes the cosine similarity function. $f(\cdot;\boldsymbol{\theta})$ is optimized on the training set $D = \{x_i,\ y_i\}_{i=1}^{N}$, where x_i, y_i are a vehicle image and its identity label and N is the number of training samples. It is obvious that the capability of the deep model is crucial for high-performing ReID.

B. Convolutional Patch Embedding

As discussed before, the optimization of a pure transformer network often necessitates a large number of training samples since there is no prior knowledge of image inherent properties embedded in its architecture design [19, 20]. However, the sizes of most vehicle ReID benchmarks are much smaller than those of ImageNet-alike large-scale datasets. To enable successful training of our framework, we choose to first extract structured patch representations from each view image by a convnet. Here, we employ ResNet50 [36] to achieve this considering that it is widely used in ReID. Our final goal is to perform correlation learning among patch embeddings rather than high-level representations so that we take the outputs of the third convoluational group to derive subsequent patch tokens of our transformer.

As shown in Figure 2, we denote the number of input views as M and take $M = 3$ as an example for subsequent formulations. In specific, the input images are represented by $\{x_i,\ x_j,\ x_k\}$ and they are taken from D with $y_i = y_j = y_k$, *i.e.,* they belong to the same vehicle. The convnet encoder is divided into two subnetworks (green and orange), represented by $f(\cdot;\boldsymbol{\theta}_{b1})$ and $f(\cdot;\boldsymbol{\theta}_{b2})$ respectively. By passing input images through $f(\cdot;\boldsymbol{\theta}_{b1})$, we obtain a series of 3D tensors $\{\ f(x_i;\boldsymbol{\theta}_{b1}),\ f(x_j;\boldsymbol{\theta}_{b1}),\ f(x_k;\boldsymbol{\theta}_{b1})\}$ with each holding the dimension of $\mathbb{R}^{c\times h\times w}$. Then the spatial dimensions of each tensor are flattened, yielding $\mathbb{R}^{c\times hw}$ as the final shape. By transposing their first and second dimension and performing a linear projection, we get a set of 2D tensors with shape $\mathbb{R}^{hw\times t}$, where t is the embedding dimension of our transformer tokens. All tokens are incorporated in the order of classification, noise, view patches, and distillation, as illustrated in Figure 2. They are concatenated together, achieving a new tensor with the size of $\mathbb{R}^{(3hw+3)\times t}$. Adding the position embeddings with the same size to it, we eventually get the input tensor $I \in \mathbb{R}^{(3hw+3)\times t}$ of our transformer.

C. Distilling Knowledge for Regularizing Transformer Learning

As explained before, the optimization of a transformer network from scratch on ReID benchmarks is pretty difficult since there are not enough data-ground truth pairs. It is well recognized that the learning of a convnet on small datasets is relatively easier, attributed to the superiority of its architecture design, *e.g.,* shared convolutional kernels. In order to train our transformer effectively and efficiently, we propose to employ the knowledge from the convnet encoder to regularize our transformer learning. The convnet encoder, in this case, serves two purposes in our framework: encoding patch representations and facilitating transformer optimization. And it is pretrained in advance on a ReID dataset.

M tensors $\{\ f(x_i;\boldsymbol{\theta}_{b1}),\ f(x_j;\boldsymbol{\theta}_{b1}),\ f(x_k;\boldsymbol{\theta}_{b1})\}$ with the dimension $\mathbb{R}^{c\times h\times w}$ are extracted from $\{x_i,\ x_j,\ x_k\}$ by the subnetwork $f(\cdot;\boldsymbol{\theta}_{b1})$, as stated in Section III-B. We then acquire the tensor $f(\ f(x_i;\boldsymbol{\theta}_{b1});\boldsymbol{\theta}_{b2})$ from the image x_i by processing $f(x_i;\boldsymbol{\theta}_{b1})$ using the second subnetwork $f(\cdot;\boldsymbol{\theta}_{b2})$. The other two tensors embedded from x_j and x_k are in the similar formations and we only take the features of x_i as an example to simplify the description. After that, Global Average Pooling (GAP), Batch Normalization (BN), and the linear classification are performed to further process these tensors. Then the generated logits for x_i, x_j and x_k are averaged to get the final ones $[z_1,\ldots,z_C]$, where C is the number of vehicle instances in D. The logits are scaled by a coefficient and converted to the classification distribution P by a softmax operation, with the hth entry calculated as:

$$p_h = \frac{\exp\left(z_h/\tau\right)}{\sum_r \exp\left(z_r/\tau\right)}, \tag{1}$$

where τ is the so-called temperature coefficient for adjusting the distribution sharpness. We set $\tau = 16$ in our final experiments.

Meanwhile, the transformer is responsible for learning long-range dependencies among tokens of the input I, *i.e.,* inter- and inner-view correlation learning. After the computation of a stack of MHSAs in the transformer, its output tensor O keeps the same dimensions with I, *i.e.,* $O \in \mathbb{R}^{(3hw+3)\times t}$. Then a linear classifier is also utilized to change the feature dimension of the distillation token and derive the logits $[w_1,\ldots,w_C]$. Similarly, a softmax operation is adopted to transform the logits after scaling and the resultant classification distribution is denoted as Q, whose hth element is:

$$q_h = \frac{\exp\left(w_h/\tau\right)}{\sum_r \exp\left(w_r/\tau\right)}. \tag{2}$$

Subsequently, Kullback-Leibler Divergence is employed to evaluate the distance of these two distributions, *i.e.,*

$$L_{KLD}(P\|Q) = \sum_r P(r)\log\frac{P(r)}{Q(r)}. \tag{3}$$

By minimizing L_{KLD}, we enforce the distillation token of our transformer to capture as similar representations as those of the convnet. Meanwhile, it transmits effective information to other tokens including the classification token through interactions of self-attention computing automatically.

D. Maximizing Classification Entropy

When training a transformer network on a small dataset, over-fitting can often be observed. Data augmentation is an important technique for extending the training set and avoiding it to some extent. However, we find that this strategy is not enough to train a generalizable transformer-based ReID model. We propose a simple yet effective method to tackle this problem. That is, in addition to the classification and distillation token, another token can be employed to predict a uniform classification distribution. In other words, we expect that the input multiple views would be assigned to any class at random with the same probability. We term this token as the noise token because this strategy is equivalent to adding noise into the framework optimization, which has been shown to be helpful for network training [42, 43, 44, 45, 46].

Concretely, the noise token is split from O according to its embedded position in I. It is projected into class prediction logits by a linear layer. And processed by a softmax operation, the class distribution is denoted as $G = [g_1,\ldots,g_C]$. So the entropy maximization objective function can be formulated as:

$$L_{EM} = \sum_{r} G(r) \log G(r). \tag{4}$$

The optimization of L_{EM} encourages the uniform classification of input samples, which is demonstrated to be effective for improving inference performances of our framework.

E. ReID Related Losses

In the literature, hard mining triplet loss (Tri) [47] and smoothed cross entropy loss (ID) [48] are widely used to optimize ReID models. They are typically integrated by the Batch Normalization neck (BNNeck) proposed in [15]. We borrow the same mechanism to derive the triplet loss L_{Tri} and ID loss L_{ID} from the classification token of our transformer, as shown in Figure 2.

F. Overall Objectives

To optimize our framework, we combine all loss functions by linear summation as our final objectives:

$$L_{Overall} = \lambda_1 L_{KLD} + \lambda_2 L_{EM} + \lambda_3 L_{Tri} + \lambda_4 L_{ID}, \tag{5}$$

where λ_1, λ_2, λ_3, λ_4 are balancing coefficients and their values are simply chosen according to initial scales of corresponding loss terms without heavy tuning. And their values are set as 1e4, 1e-2, 1.0, and 1.0, respectively.

G. Network Architecture

Generally, our framework is composed of the convnet encoder and the transformer. We employ the ResNet50, with $stride = 2$ in $conv5\_x$ replaced with $stride = 1$, as the patch embedding encoder. Referring to Figure 2, the green subnetwork consists of sequential $conv1$, $conv2\_x$, $conv3\_x$, and $conv4\_x$. The last convolutional group $conv5\_x$ of the ResNet50 composes the orange subnetwork, followed by a GAP layer and a BNNeck structure. In order to reduce the number of patch tokens for saving computation memory, the output tensor of the first subnetwork is actually processed by a max pooling operation with $stride = 2$ and $kernel\ size = 3$ before being flattened. All classifiers and the linear projector are acted by fully connected layers with different input and output channels accordingly. The design of our transformer layers refers to that of DeiT [20]. However, to model the more complex part correlations among multi-view images, we increase the dimension of the token embeddings to 1,920 and the number of the attention heads in MHSAs to 96. The feature dimension expanding ratio in multilayer perceptron (MLP) is set as two instead of four to reduce memory cost. Furthermore, unlike widely used transformers, ours contains much less MHSA-based layers, *i.e.,* four layers for VeRi-776 [31], six layers for VehicleID [49] and VERI-Wild [50].

TABLE I: Results comparison on VeRi-776. Our approach outperforms other methods, including those involving more human efforts, on two most important metrics, *e.g.,* tmAP and imAP.

Method	Venue	ES	tmAP	imAP	Top-1	Top-5
Siamese+Path [51]	ICCV17	Y	58.27	-	83.49	90.04
OIFE [8]	ICCV17	Y	48.0	-	65.9	87.7
OIFE+ST [8]	ICCV17	Y	51.42	-	68.3	89.7
NuFACT [30]	TMM17	Y	53.42	-	81.56	95.11
VAMI [33]	CVPR18	Y	50.13	-	77.03	90.82
AAVER [9]	ICCV19	Y	58.52	-	88.68	94.10
RS [26]	ICCV19	Y	-	63.76	90.70	94.40
R+MT+K [26]	ICCV19	Y	-	65.44	90.94	96.72
VANet [10]	ICCV19	Y	66.34	-	89.78	95.99
PART [7]	CVPR19	Y	74.3	-	94.3	**98.7**
SAN [52]	MST20	Y	72.5	-	93.3	97.1
CFVMNet [53]	MM20	Y	-	77.06	95.3	98.4
PVEN [6]	CVPR20	Y	-	79.5	95.6	98.4
SPAN [24]	ECCV20	Y	68.9	-	94.0	97.6
DMML [12]	ICCV19	N	-	70.1	91.2	96.3
UMTS [54]	AAAI20	N	-	75.9	95.8	-
SAVER [11]	ECCV20	N	79.6	-	**96.4**	98.6
Ours	-	N	**86.0**	**80.9**	96.2	98.4

IV. Experiments

Datasets. 1) VeRi-776 is constructed by [31] and contains 49,357 multi-view images of 776 vehicles, which are captured by 20 non-overlapping cameras. And 37,778 images of 576 identities and 11,579 images of the remaining 200 instances have been chosen as training and testing set, respectively. The query set contains 1,678 images selected from the testing set. When testing, the gallery set contains all the images in the testing set except those which share the same identity and camera ID as the probe. 2) VehicleID [49] is a widely used large-scale vehicle ReID dataset which contains 221,763 images of 26,267 vehicles taken from front or rear view. 113,346 images of 13,164 vehicles are selected for training and others are reserved for evaluation. The gallery set only contains one image for each testing identity. There are three gallery sizes widely used for evaluation, *i.e.,* 800 (Small), 1,600 (Medium), and 2,400 (Large). Obviously, VehicleID is a much more challenging benchmark because it involves plenty of vehicle instances. So training ReID models on it usually requires more epochs and more complicated learning rate scheduler. 3) VERI-Wild [50] is another large-scale vehicle ReID benchmark, composed of 416,314 images of 40,671 vehicles taken by 174 cameras. Its images are captured from various viewpoints, under different weathers and lighting conditions. The testing set contains 138,517 images of 10,000 identities in total, divided into three subsets with 3,000 (Small), 5,000 (Medium), and 10,000 (Large) vehicles respectively.

Implementation. We choose PyTorch to implement our framework and Adam optimizer [55] with default betas ($\beta_1 = 0.9$, $\beta_2 = 0.999$), weight decay 5e-4 to optimize it. The initial learning rate in all experiments is 1e-4. During training, random cropping, horizontally flipping, and erasing are performed to augment data samples. All images are resized to 256×256. When training the patch embedding encoder on VeRi-776, VehicleID, and VERI-Wild, the batch size is respective 28, 40, and 100, with 4 images from each instance. The margin of triplet loss is set as 0.5 empirically. On VeRi-776, the total

TABLE II: Results comparison on VehicleID. It is observed that our approach achieves SOTA performances on five out of six indicators.

Method	Venue	ES	Small		Medium		Large	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
GoogLeNet [57]	CVPR15	Y	47.90	67.43	43.45	63.53	38.24	59.51
MD+CCL [49]	CVPR16	Y	49.0	73.5	42.8	66.8	38.2	61.6
OIFE [8]	ICCV17	Y	-	-	-	-	67.0	82.9
NuFACT [30]	TMM17	Y	48.90	69.51	43.64	65.34	38.63	60.72
VAMI [33]	CVPR18	Y	63.1	83.3	52.9	75.1	47.3	70.3
AAVER [9]	ICCV19	Y	72.47	93.22	66.85	89.39	60.23	84.85
VANet [10]	ICCV19	Y	**88.12**	97.29	83.17	95.14	80.35	92.97
PART [7]	CVPR19	Y	78.4	92.3	75.0	88.3	74.2	86.4
SAN [52]	MST20	Y	79.7	94.3	78.4	91.3	75.6	88.3
CFVMNet [53]	MM20	Y	81.4	94.1	77.3	90.4	74.7	88.7
PVEN [6]	CVPR20	Y	84.7	97.0	80.6	94.5	77.8	92.0
UMTS [54]	AAAI20	N	80.9	-	78.8	-	76.1	-
SAVER [11]	ECCV20	N	79.9	95.2	77.6	91.1	75.3	88.3
Ours	-	N	86.8	**97.3**	**83.4**	**95.8**	**80.6**	**93.1**

training epochs is 80 and the learning rate is decreased by a factor of 0.1 after running every 20 epochs. On VehicleID and VERI-Wild, the total training epochs is 120 and the learning rate is adjusted using a warmup cosine annealing schedule (WCA), *i.e.,* it is increased linearly from 0 to 1e-4 during the first 10 epochs, decreased with a cosine annealing scheduler.

When training our transformer, we freeze the learned parameters of the patch encoder. On VeRi-776, VehicleID, and VERI-Wild, M is set as 3, 2, and 2 respectively. The training batch size is multiplied by M, namely with $4M$ images from each instance. And randomly selected M images from the same instance form each input group of our framework. The number of total training epochs is 50 and the WCA is also utilized to adjust the learning rate. Our framework attempts to extract rich patters from multiple views. So it is reasonable to push the hard-negative "samples" farther away from the hard-positive ones *w.r.t.* the anchors by setting a larger margin 0.7 in triplet loss.

During inference, we combine the original image, its horizontally flipped version (and centrally cropped version on VeRi-776) together as pseudo multi-view inputs of our framework. The global vector and part-wise correlation vector in Figure 2 are concatenated along the feature dimension as the final vehicle representations.

Evaluation Protocols. We do not adopt any post-processing strategies, *e.g.,* k-reciprocal re-ranking [56], to further improve our performances. Our framework is evaluated by four popular metrics in the ReID literature, *i.e.,* image-to-track retrieval mean Average Precision (tmAP) (only for VeRi-776 with tracks information), image-to-image retrieval mAP (imAP), Top-1, and Top-5 accuracy. Unlike other works providing either tmAP or imAP on VeRi-776, we report both for comprehensive comparison. These metric scores are shown as percentages and the best ones in each table are marked by bold. For fair comparison, in Table I, II, and III, we use the abbreviation ES (Y/N) to indicate whether any Extra Supervision in addition to vehicle identities is adopted to enhance a corresponding method.

A. Performance Comparison with SOTA Methods

VeRi-776. In Table I, we compare our results with those of recent works. A few of them focused on performing vehicle ReID only under the guidance of identity labels like ours, and the majority struggled to exploit more supervisions to implement their approaches. In [10], for example, vehicle body orientations of respective 5,000 images from VeRi-776 and VehicleID were manually labelled to train a viewpoint classifier. Distinct metrics for similar and dissimilar viewpoint images were learned for ReID based on this viewpoint predictor. Similarly, in [7], front and back lights, front and back windows, and brands were thought as informative and discriminative vehicle parts. Thus the locations of these parts were annotated and a YOLO detector [58] was optimized to detect them in an offline manner. When performing ReID, local features were condensed from these regions as supplementary of global representations. Besides, a U-Net based semantic parser was trained in [6] by using the supervision from vehicle view segmentation of 3,165 images from VeRi-776. The parser was responsible for predicting a view mask which facilitated the proposed view-aware feature alignment when optimizing their ReID model. AAVER [9] relied on predicting the defined keypoints and orientation to guide their discriminative feature learning. Also, SAN [52] adopted various attribute annotations to supervise their attribute-aware branch. Although no enhancement from additional supervisions was borrowed in SAVER [11] explicitly, they required that the Detectron object detector [59] had to be utilized to pre-process all vehicle images for removing background noise, which indirectly involved the information from object detection benchmarks.

In contrast, our framework does not need any extra supervisions at all. Only with vehicle identity annotations, we aim to conduct correlation learning among discriminative parts of multi-views for the extraction of integrated instance representations. Besides, our training batch size is much smaller than those of other methods, *e.g.,* 256 in SAN [52], because of GPU resource limitation. As we know, a larger image batch in forward feeding contains more similar negative samples and dissimilar positive samples to each anchor, definitely enforcing a ReID model to be more distinguishable, when optimizing it using triplet loss. Even under these hostile conditions, our approach can still beat all other competitors on tmAP and imAP. In terms of Top-5 accuracy, our result is just 0.3% worse than that of PART [7] whose training image size 512×512 is much larger than ours. It was demonstrated in [7] that increasing image resolution was able to improve the performances of their proposal considerably. If compared with their results under our image resolution, ours are significantly better on all metrics.

VehicleID. Most of the SOTA methods on this dataset are identical to those on VeRi-776, except from MD+CCL [49] and GoogLeNet [57]. Thus we do not elaborate them in details. Their results are listed in Table II for comparison. Please be aware that PVEN [6] required much larger batch size 256 to achieve their performances in addition to involving extra supervisions. Despite these challenges, our approach

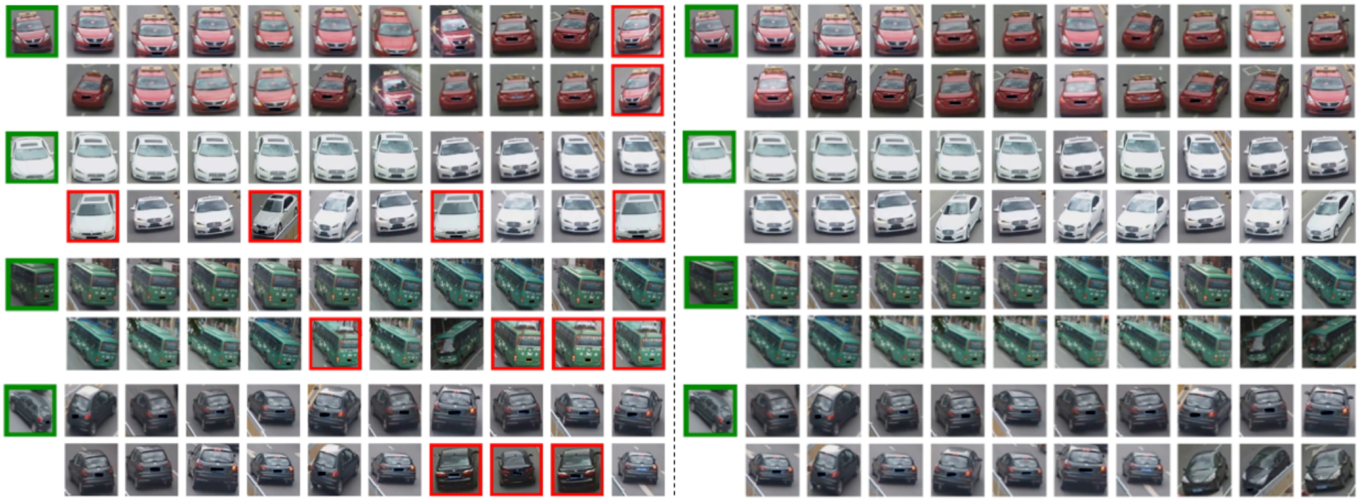


Fig. 3: Visualization comparison of ranking results: the Baseline method (left) and our approach (right). Two query images (indicated by green boxes) on each row are the same, followed by the respective top 20 candidates from the gallery set. They are sorted according to their similarities with the query in descending order. The images in red boxes are negative candidates. Obviously our framework is much more powerful than the baseline in capturing discriminative details from vehicle images and therefore eliminating pretty challenging distractors.

outperforms all these proposals on almost every metric of three gallery sizes. Especially on difficult testing scenarios, *i.e.,* Medium and Large gallery, our framework is the best one. When compared with SAVER [11], our approach performs significantly better on all evaluation indicators, *i.e.,* 6.9% Top-1, 2.1% Top-5 on Small gallery, 5.8% Top-1, 4.7% Top-5 on Medium gallery, and 5.3% Top-1, 4.8% Top-5 on Large gallery higher, although the former required customized pre-processing strategies.

VERI-Wild. The results comparison on this benchmark is reported in Table III. Note that PVEN [6] does not adopt the same testing setting with others, *i.e.,* the candidate gallery images with identical vehicle ID and camera ID to the query are not removed for evaluation. This operation is thought to enhance the performances largely. For a fair comparison, we also report our results under the same setting. Our approach can easily beat other methods on all metrics although some of them employ auxiliary information. On the most challenging evaluation scenario (*i.e.,* Large testing set), for example, our result on imAP is 2.4% better than the second best [60].

B. Ablation Study

We conduct detailed ablation study experiments to validate the efficacy of our proposals and report all results in Table IV, V, and VI.

Effectiveness of Each Proposed Component. In Table IV, we release our experiment results of adding each component into the baseline method (Baseline) in turn. The Baseline refers to a ReID model consisting of a ResNet50 as its backbone and a BNNeck mechanism, which is commonly treated as the baseline in recent ReID literature [15]. We report tmAP, imAP, and Top-1 accuracy on VeRi-776 and Top-1, Top-5 accuracy on VehicleID for thorough comparison. The Transformer indicates simply using a common transformer (without our distillation and noise token) to exploit part-wise correlation for ReID. Except indicated, we only adopt the part-wise correlation vector derived from the transformer (referring to Figure 2) as representations of each testing image to better evaluate our proposals. From the second section of Table IV, margin improvements for all metrics are beneficial from directly using Transformer, which, compared with our final improvements, demonstrates the significance of our proposed components. Then through adding the distillation token, we conduct experiments with distilling knowledge from the convnet encoder to our transformer (+ Distill). The improvements upon Transformer are pretty impressive on most metrics. To prevent the over-fitting phenomenon when training the transformer-based framework on a ReID dataset, we further involve our noise token to enforce the classification entropy maximization (+ EM). It can be observed that all indicators are further promoted. By concatenating the part-wise correlation vector from our transformer and the global vector from the convnet encoder together, we acquire the final version of our framework (Ours). We can also conclude that the convolutional features, in addition to the part-wise correlation learning based transformer representations, further contribute to our final performances.

Improvements of Multi-view Testing. We report both Single-view and Multi-view testing results in Table V. When extracting the representation vector for a testing image, we treat its original and augmented versions as the multi-view inputs. To demonstrate that our framework is capable of capturing more distinguishable features from the "pseudo" multiple views (Multi-view), we also perform another testing procedure, *i.e.,* encoding representations from an original image and its copies (Single-view). Considerable performance gains on two datasets are seen from the Multi-view testing.

Robustness in Terms of Hyperparameter Adjustment.

TABLE III: Performance comparisons on VERI-Wild. $*$ These testings do not remove gallery images with the same identity and camera as the query. Apparently, our method is significantly better than existing works.

Method	Venue	ES	Small			Medium			Large		
			imAP	Top-1	Top-5	imAP	Top-1	Top-5	imAP	Top-1	Top-5
AAVER [9]	ICCV19	Y	62.2	75.8	92.7	53.7	68.2	88.9	41.7	58.7	81.6
PGAN [60]	TITS20	Y	83.6	95.1	-	78.3	92.8	-	70.6	89.2	-
DFNet [61]	TPAMI21	Y	83.1	94.8	98.1	77.3	93.2	97.5	69.9	89.4	96.0
FDA [50]	CVPR19	N	35.1	64.0	82.8	29.8	57.8	78.3	22.8	49.4	70.5
BW [62]	IJCNN19	N	70.5	84.2	95.3	62.8	78.2	93.1	51.6	70.0	88.5
SAVER [11]	ECCV20	N	80.9	93.8	97.9	75.3	92.7	97.5	67.7	89.5	95.8
Ours	-	N	**85.0**	**95.3**	**98.6**	**79.8**	**93.3**	**98.1**	**73.0**	**90.2**	**96.2**
PVEN [6]*	CVPR20	Y	82.5	96.7	99.2	77.0	95.4	98.8	69.7	93.4	97.8
Ours*	-	N	**87.1**	**97.2**	**99.4**	**82.5**	**96.0**	**99.1**	**76.2**	**94.3**	**98.2**

TABLE IV: Ablation study on VeRi-776 and VehicleID. The Baseline refers to a ReID model employing a ResNet50 as its backbone. The Transformer here represents directly using a regular transformer to perform part-wise correlation learning. The Distill and EM indicate adding the distillation token for knowledge distillation and the noise token encouraging classification entropy maximization into the Transformer. Gradually improved performances demonstrate their efficacy when incorporating more proposals into the baseline network.

Method	VeRi-776			Small		Medium		Large	
	tmAP	imAP	Top-1	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Baseline	83.2	77.2	95.5	80.1	93.3	77.2	90.6	75.0	87.1
Transformer	83.8	77.5	95.9	80.2	93.3	77.4	90.7	75.4	87.9
+ Distill	85.3	79.6	95.6	86.1	96.8	82.1	94.9	79.6	92.6
+ EM	85.7	80.0	95.9	86.5	97.0	82.6	95.2	80.0	92.7
Ours	**86.0**	**80.9**	**96.2**	**86.8**	**97.3**	**83.4**	**95.8**	**80.6**	**93.1**

TABLE V: Effect of multi-view testing on VeRi-776 and VehicleID. Single-view: feeding a testing image and its copies into our framework for feature extraction, Multi-view: encoding representations from an original image and its "pseudo" multiple views. We can see that performances on all metrics benefit from "pseudo" multi-view inputs.

Method	VeRi-776			Small		Medium		Large	
	tmAP	imAP	Top-1	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Single-view	85.8	80.4	95.8	86.7	96.9	82.7	94.9	79.8	93.0
Multi-view	**86.0**	**80.9**	**96.2**	**86.8**	**97.3**	**83.4**	**95.8**	**80.6**	**93.1**

To demonstrate the robustness of our approach, we conduct experiments of adjusting some critical hyperparameters, *i.e.,* the temperature coefficient τ, the margin of Triplet loss m, and the number of input views M, on VeRi-776. Grid search is used to tune these hyperparameters, and corresponding results are listed in Table VI. Note that our final selections for these hyperparameters are $\tau = 16.0$, $M = 3$, and $m = 0.7$. We adjust only one of them in each experiment and keep others identical to their final values. Each section of Table VI gives results of tuning one hyperparameter. Very small performance changes are seen from this Table when adjusting these hyperparameters heavily. Namely, our approach is sufficiently robust to the adjustment of key hyperparameters.

TABLE VI: Results of hyperparameter adjustment on VeRi-776, *w.r.t.* the temperature coefficient τ, the margin of Triplet loss m, and the number of input views M. The stable performances demonstrate the excellent robustness of our framework.

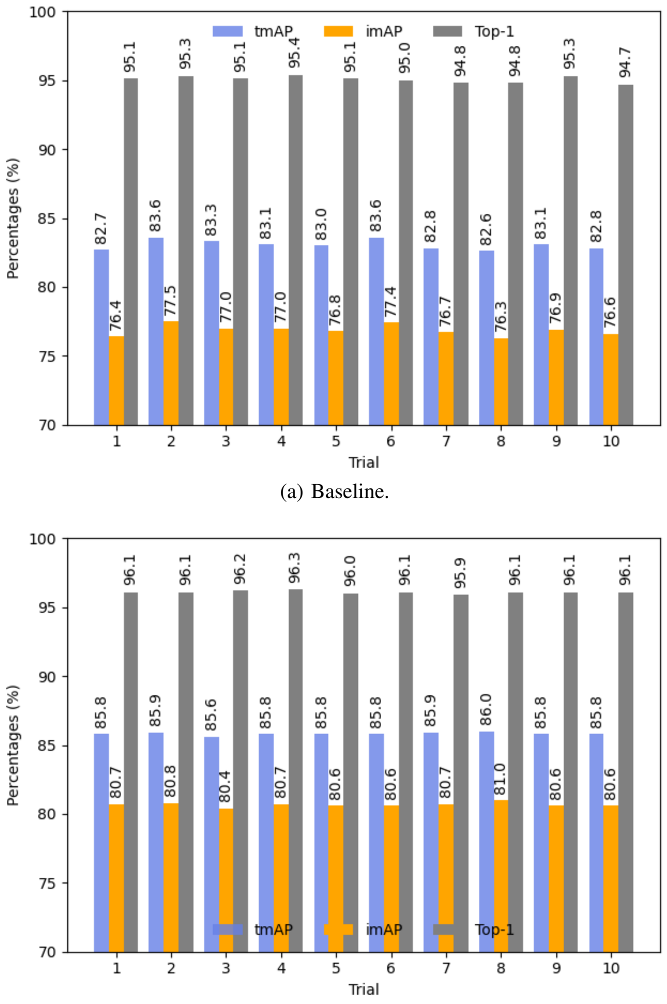
Hyperparameters	tmAP	imAP	Top-1
$\tau = 10.0$	85.9	80.9	96.1
$\tau = 50.0$	85.7	80.5	96.0
$\tau = 100.0$	85.7	80.5	96.1
$M = 2$	85.8	80.5	95.9
$M = 4$	84.8	79.3	95.5
$m = 0.5$	85.3	80.1	96.2
$m = 0.6$	85.6	80.4	96.2
$m = 0.8$	85.9	80.6	96.2
Ours	**86.0**	**80.9**	**96.2**

C. Visualizations

To verify the superiority of our framework qualitatively, we show its correlation learning visualization, retrieval rankings, and performance oscillation of random trails in this section.

Part-wise Correlation Discovery across Views. To visualize the results of our inter- and inner-view correlation learning, we cover the learned feature maps on the input samples to show where the framework pays more attention. Concretely, we first obtain the patch tokens yielded by the last transformer layer during feedforward. Then the feature dimensions of these embeddings are squeezed by summation. Through reshaping and upsampling operations, we recover their spatial resolutions and place them on top of the input multi-view images after colorization, as illustrated in Figure 1. Every triplet inputs from the same vehicle are displayed in one row.

Obviously, our framework is capable of discovering discriminative parts from each view image successfully. Surprisingly, these automatically detected vehicle parts, *e.g.,* lights and



(a) Baseline.

(b) Ours.

Fig. 4: Performance oscillation of random ten trials of the Baseline (a) and our approach (b) on VeRi-776. We conclude that learning comprehensive vehicle representations from multiple views through transformer-based part-wise correlation modelling can apparently stabilize the training process and finally contribute to consistent performances at random running.

windows, are highly consistent with those manually annotated by other works [7]. It releases one of the important reasons why our approach performs better than other competitors. Furthermore, it can be seen that our framework performs inter- and inner-view correlation learning among vehicle parts successfully. On one hand, various vehicle parts activate high responses in one view image, *e.g.,* the annual inspection sign (the right-upper corner of the front window) and the left window in the middle image of the first vehicle, the left and right back lights in the middle image of the second individual. These parts contain instance-specific information for condensing ID representations through inner-view correlation learning. On the other hand, the identical vehicle parts, if visible from multiple viewpoints, are focused on. For example, almost identical locations on the back of the last vehicle are discovered from the left and right image. The back lights of the second vehicle are detected from the left and middle image. Similarly, the texts on the back of the first vehicle are highlighted from the left and right image. All these suggest that perfect part-wise interdependencies among views are modelled implicitly by our approach.

Qualitative Comparison of Ranking Results. To validate the power of our approach in distinguishing vehicles qualitatively, we compare our ranking results with those of the Baseline method when retrieving the same query image. For each one, we predict a sorted list of all gallery images according to their similarities with the query in embedded space. The closer the image is to the query one, the higher it will be ranked. Here we just show the top 20 candidates for each query due to the space limitation. The rankings of the Baseline (left) and ours (right) are shown in Figure 3. Query images (indicated by green boxes) of the same row are identical. Candidates circled by red boxes are false items, *i.e.,* they belong to different vehicles with the corresponding query. We can see that there are many negative distractors which have extremely similar appearances but actually distinct IDs with the query in the gallery set. They can not be distinguished by the Baseline correctly, even humans. Thus they are placed at very high ranking positions by mistake in the left column. In contrast, our framework is able to identify true positive candidates successfully from those negative distractors by modelling intrinsic relationships among discriminative vehicle parts.

Consistent Performances of Random Trials. As mentioned before, many researchers noticed the problem of unstable ReID performances at different running regarding their methods, while none of them released their results of multiple random trials. To demonstrate the great stability and reliability of our framework, we provide the results of the Baseline and ours for ten random trials on VeRi-776 in Figure 4a and 4b, respectively. We provide tmAP, imAP, and Top-1 accuracy of two approaches and choose bar charts to visualize these numbers for better comparing their fluctuations. These trials are conducted by setting different randomness seeds for the whole ReID system. We can see that our framework achieves almost identical performances at each trial and thus the results are easily reproduced by random running. This demonstrate that our approach of capturing comprehensive vehicle representations from multiple views by modelling part-wise correlations is superior on stabilizing ReID feature learning.

D. Cross-dataset Transferring

To verify the generalization ability of our framework, we perform a cross-dataset testing using our learned model on the source dataset without any fine-tuning on the target one. Moreover, we compare the results with those of the fully supervised method proposed in VERI-Wild [63]. From Table VII, our results for imAP and Top1 accuracy are generally better than VERI-Wild but worse for Top5. At least, both methods are comparable. Note that VehicleID is a relatively smaller dataset and the results of VERI-WILD are obtained under fully supervised training. This comparison does demonstrate the excellent scalability of our method.

TABLE VII: Our domain transferring results from VehicleID to VERI-Wild. DT: directly transferring, FST: Fully Supervised Training. Overall, our approach can beat the fully supervised method proposed in VERI-Wild even without seeing the training data.

Scale	Ours (DT)			VERI-Wild (FST)		
	imAP	Top-1	Top-5	imAP	Top-1	Top-5
Small	**35.5**	63.6	79.3	35.1	**64.0**	**82.8**
Medium	**31.6**	**58.5**	74.8	29.8	57.8	**78.3**
Large	**26.1**	**51.5**	68.2	22.8	49.4	**70.5**
Mean	**31.1**	**57.9**	74.1	29.2	57.1	**77.2**

V. Conclusion

In this paper, we propose a novel transformer-based ReID framework which is capable of extracting vehicle representations from multiple views through exploiting part-wise correlation. To cope with the difficulty of optimizing a transformer directly on a small ReID dataset, we present an efficient one, composed of a distillation token and a noise token for encouraging the knowledge distillation from the convnet to MHSA layers and classification entropy maximization, respectively. We perform comprehensive experiments on three widely used vehicle ReID benchmarks. Quantitative performances and qualitative visualizations demonstrate the superiority of our framework.

Acknowledgement

This work was supported partially by NSF (CCF-2006738).

References

[1] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *CVPR*, 2019, pp. 8797–8806.

[2] X. Lin, R. Li, X. Zheng, P. Peng, F. Huang, Y. Wu, and R. Ji, "Aggregating global and local visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, 2020.

[3] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, 2020.

[4] M. Li, X. Huang, and Z. Zhang, "Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 194–204.

[5] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7103–7112.

[7] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *CVPR*, June 2019.

[8] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *ICCV*, 2017, pp. 379–387.

[9] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *ICCV*, 2019, pp. 6132–6141.

[10] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *ICCV*, 2019, pp. 8282–8291.

[11] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," *arXiv preprint arXiv:2004.06271*, 2020.

[12] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *ICCV*, October 2019.

[13] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2385–2399, 2018.

[14] Y. Zhang, D. Liu, and Z.-J. Zha, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 1386–1391.

[15] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, 2019.

[16] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Advances in Neural Information Processing Systems*, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[18] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *arXiv preprint arXiv:2012.07436*, 2020.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image

database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] S. A. S. Alfasly, Y. Hu, T. Liang, X. Jin, Q. Zhao, and B. Liu, "Variational representation learning for vehicle re-identification," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3118–3122.

[23] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *arXiv preprint arXiv:2102.04378*, 2021.

[24] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," *arXiv preprint arXiv:2008.11423*, 2020.

[25] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, 2019.

[26] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *ICCV*, 2019, pp. 211–220.

[27] F. Wu, S. Yan, J. S. Smith, and B. Zhang, "Joint semi-supervised learning and re-ranking for vehicle re-identification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 278–283.

[28] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4328–4338, 2019.

[29] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *ICCV*, 2017, pp. 562–570.

[30] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2017.

[31] ——, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European conference on computer vision*. Springer, 2016, pp. 869–884.

[32] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: a region-aware deep model for vehicle re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[33] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *CVPR*, June 2018.

[34] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[35] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," *arXiv preprint arXiv:2101.11605*, 2021.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[37] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *arXiv preprint arXiv:2012.15840*, 2020.

[38] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[39] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," *arXiv preprint arXiv:2011.10881*, 2020.

[40] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *arXiv preprint arXiv:2012.09688*, 2020.

[41] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *arXiv preprint arXiv:2102.08005*, 2021.

[42] Y. Solomon, A. Wagner, and P. Bendich, "A fast and robust method for global topological functional optimization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 109–117.

[43] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," in *International conference on machine learning*. PMLR, 2016, pp. 3059–3068.

[44] Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba, "Interplay between optimization and generalization of stochastic gradient descent with covariance noise," *arXiv preprint arXiv:1902.08234*, 2019.

[45] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, "Adding gradient noise improves learning for very deep networks," *arXiv preprint arXiv:1511.06807*, 2015.

[46] L. Xiao, Z. Zhang, and Y. Peng, "Noise optimization for artificial neural networks," *arXiv preprint arXiv:2102.04450*, 2021.

[47] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[49] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *CVPR*, 2016, pp. 2167–2175.

[50] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *CVPR*, June 2019.

[51] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *ICCV*, 2017, pp. 1900–1909.

[52] J. Qian, W. Jiang, H. Luo, and H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," *Measurement Science and Technology*, 2020.

[53] Z. Sun, X. Nie, X. Xi, and Y. Yin, "Cfvmnet: A multi-

branch network for vehicle re-identification based on common field of view," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3523–3531.

[54] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 165–11 172.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 1318–1327.

[57] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *CVPR*, 2015, pp. 3973–3981.

[58] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[59] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," 2018.

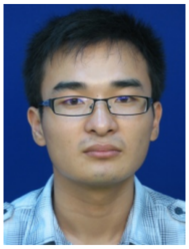
[60] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle instance retrieval," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[61] Y. Bai, J. Liu, Y. Lou, C. Wang, and L. Duan, "Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

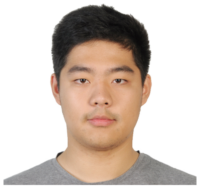
[62] R. Kuma, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: an efficient baseline using triplet embedding," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.

[63] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

Ming Li received his B.S. degree from Xidian University in 2015 and M.S. degree from Peking University in 2018. He worked as a Research Scholar in The University of North Carolina at Chapel Hill from Aug. 2018 to Oct. 2019. He was a Visiting Scholar at Worcester Polytechnic Institute from Nov. 2019 to Apr. 2021. Now he is pursuing his Ph.D. degree at National University of Singapore. His research interests lie in vehicle re-identification, self-supervised visual learning, and domain generalization.



Jun Liu is an Assistant Professor at Singapore University of Technology and Design. He obtained the Ph.D. degree from Nanyang Technological University, Singapore. His research interests include video analysis and deep learning. He is an Associate Editor of Elsevier Neurocomputing, Springer Signal, Image and Video Processing, IET Computer Vision, and SPIE Journal of Electronic Imaging. He is an Area Chair of ICLR 2022 and WACV 2022.



Ce Zheng received his B.S. degree from Wuhan University of Science and Technology in 2016 and M.S. degree from Tufts University in 2019. He was a Research Assistant at Worcester Polytechnic Institute from Oct 2019 to May 2020. Now he is pursuing his Ph.D. degree at University of Central Florida. His research interests focus on human pose estimation, 3D human mesh recovery and image/video processing.

Xinming Huang received the Ph.D. degree in electrical engineering from Virginia Tech in 2001. Since 2006, he has been a faculty in the Department of Electrical and Computer Engineering at Worcester Polytechnic Institute (WPI), where he is currently a chair professor. Previously he was a Member of Technical Staffs with the Bell Labs of Lucent Technologies. His main research interests are in the areas of circuits and systems, with emphasis on autonomous vehicles, deep learning, IoT and wireless communications.

Ziming Zhang is an assistant professor at Worcester Polytechnic Institute (WPI). Before joining WPI he was a research scientist at Mitsubishi Electric Research Laboratories (MERL) in 2017-2019. Prior to that, he was a research assistant professor at Boston University in 2016-2017. Dr. Zhang received his PhD in 2013 from Oxford Brookes University, UK, under the supervision of Prof. Philip H. S. Torr. His research areas include object recognition and detection, zero-shot learning, deep learning, optimization, large-scale information retrieval, visual surveillance, and medical imaging analysis. His works have appeared in TPAMI, IJCV, CVPR, ICCV, ECCV, ACM Multimedia, ICDM, ICLR and NIPS. He won the R&D 100 Award 2018.