Self-supervised Geometric Features Discovery via Interpretable Attention for Vehicle Re-Identification and Beyond

Ming Li*   Xinming Huang   Ziming Zhang

Worcester Polytechnic Institute

100 Institute Rd, Worcester, MA, USA

ming.li@u.nus.edu, {xhuang, zzhang15}@wpi.edu

Abstract

To learn distinguishable patterns, most of recent works in vehicle re-identification (ReID) struggled to redevelop official benchmarks to provide various supervisions, which requires prohibitive human labors. In this paper, we seek to achieve the similar goal but do not involve more human efforts. To this end, we introduce a novel framework, which successfully encodes both geometric local features and global representations to distinguish vehicle instances, optimized only by the supervision from official ID labels. Specifically, given our insight that objects in ReID share similar geometric characteristics, we propose to borrow self-supervised representation learning to facilitate geometric features discovery. To condense these features, we introduce an interpretable attention module, with the core of local maxima aggregation instead of fully automatic learning, whose mechanism is completely understandable and whose response map is physically reasonable. To the best of our knowledge, we are the first that perform self-supervised learning to discover geometric features. We conduct comprehensive experiments on three most popular datasets for vehicle ReID, i.e., VeRi-776, CityFlow-ReID, and VehicleID. We report our state-of-the-art (SOTA) performances and promising visualization results. We also show the excellent scalability of our approach on other ReID related tasks, i.e., person ReID and multi-target multi-camera (MTMC) vehicle tracking.

1. Introduction

Vehicle ReID is a fundamental but challenging problem in video surveillance due to subtle discrepancy among vehicles from identical make and large variation across viewpoints of the same instance. The success of recent works suggests that the key to solving this problem is to incor-

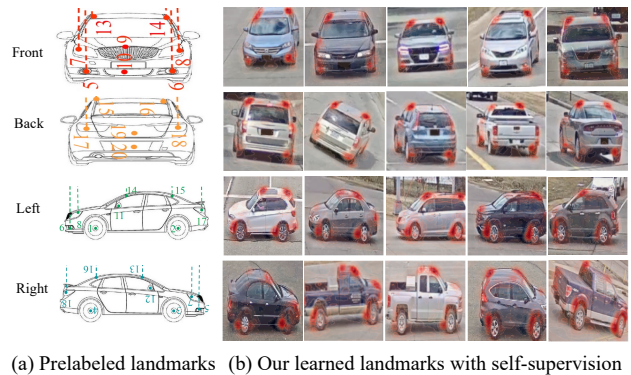(a) Prelabeled landmarks (b) Our learned landmarks with self-supervision

Figure 1: **(Top)** In the literature, a labor-intensive fine-grained labels annotation is often required to capture local discriminative features, such as (a) labelling landmarks in [55] to learn orientation invariant features. In contrast, we manage to discover such geometric features (denoted as red in (b)) in a self-supervised way. **(Bottom)** As for its generalization ability, our approach can also consistently locate critical parts of a deformable human body, *e.g.,* head, upper arms, and knees, with no using corresponding ground truths. Best viewed in color.

porate explicit mechanisms to discover and concentrate on informative vehicle parts (*e.g.,* wheels, manufacturer logos) for discriminative feature extraction in addition to capturing robust features from a holistic image. They all sought, however, to edit original data to provide supplementary su-

pervisions, *e.g.,* view segmentation [39], keypoints [24, 55], vehicle orientation [24, 55, 8, 7] or key parts [17], for training their deep classifiers. Although these methods perform satisfactorily, their annotation processes inevitably involve intensive human efforts, which significantly limits the applicability of such approaches. For example, when deployed in a new scenario, [17] demands the informative parts have to be manually localized to optimize their YOLO detector. Afterwards, they are able to embed local patterns from detected regions-of-interest to assist ReID. So it is desirable to develop approaches which are capable of concentrating on informative details of a vehicle body but do not require corresponding ground truths.

On the other hand, while their power is demonstrated by various computer vision tasks [12, 2, 24, 4], existing attention mechanisms, like channel attention [21], spatial attention [58], and self-attention [53], are all pretty sophisticated and obscure. That is to say, their architectures are difficult to explain and attention maps are learned all by themselves. In self-attention [53], for example, high-dimensional embeddings Q, K, V are first projected from an input by convolutional or linear operations and then entry-wise correlation (attention) is obtained through matrix multiplication between Q, K. V is weighted by the resulting correlation matrix as the attentional output. Although the workflow seems to make sense, the underlying principle why it works is still a black-box like other deep networks. Additionally, their learned attentions usually spread over a holistic object without specific concerns. Otherwise, an interpretable attention module, whose design should be easy to understand, can reveal what is critical for recognition and help to guide further improvement.

In light of the above observations, we propose a novel framework that can successfully learn discriminative geometric features, under the assistance of self-supervised learning and a simple but interpretable attention, in addition to global representations for vehicle ReID. In specific, self-supervised learning is performed to optimize an encoder network, which is shared to condense low-level vehicle representations, under the supervision of automatically generated ground truths. The encoded vehicle representations are fed into the introduced interpretable attention mechanism to acquire an attention map. By weighting it on another low-level vehicle representations, we obtain the regions-of-interest emphasized features for vehicle ReID.

In summary, our key contributions in this work are:

- We are the first to successfully learn informative geometric features for vehicle ReID without supervisions from fine-grained annotations.
- An interpretable attention module, whose design is easy to explain and whose concentrations are physically important locations, is introduced to highlight the automatic regions-of-interest.

- We report the SOTA performances of our proposed approach on widely used vehicle ReID benchmarks, *i.e.,* VeRi-776 [33], CityFlow-ReID [50], and VehicleID [31], compared with all existing works including those involving more supervisions from manual annotations. We also visualize the reliable and consistent geometric features learned by our framework.
- The excellent scalability of the proposal is demonstrated by our directly transferring experiments on person ReID and MTMC vehicle tracking.

2. Related works

Vehicle ReID. Most of existing works in this field struggled to explore extra supervisions in addition to identity labels to guide ReID. These works can be grouped into three mainstreams as follows: (1) exploiting vehicle attributes (*e.g.,* color and model) [16, 62, 34, 33, 35, 75] or temporal information in data [55, 44] to regularize representation learning; (2) editing official datasets to provide more fine-grained annotations, like critical part locations [17], view segmentation [39], keypoints or vehicle body orientation [24, 55, 8, 7], to supervise local feature discovery; (3) assembling multiple datasets together [69] or synthesizing more vehicle images [36, 49, 59] to train more powerful networks. Additionally, there are a couple of works aiming to enhance representation learning from the perspective of metric learning [3, 8, 1, 66]. In contrast, our work manages to capture discriminative local patterns without corresponding supervision. Furthermore, unlike recent well-performing works which relied on another auxiliary pre-trained network to indicate informative parts [39, 17, 7, 8], our framework is elegant and end-to-end trainable.

Visual attention. Various attention architectures have been proposed in computer vision community, *e.g.,* self-attention [53, 12], channel-wise attention [21], and spatial-wise attention [58], which also spread into ReID field [4, 2, 74, 24]. For instance, [2] and [74] proposed using attention gains and multi-level foreground consistency to regularize ReID feature extraction, respectively. All these attention networks are pretty complicated and computational costly, especially hard to explain, which limits their generalization and future improvement. The attentive branch in [4], for example, incorporated Channel Attention Module (CAM) and Position Attention Module (PAM) in parallel. The reason why the latter employed stacked convolutional layers to perform Q, K, V projection but the former just utilized identity layer (copy) instead is unknown. In this case, we have no idea to improve it further, *e.g.,* making the positional attention focus on more distinguishable parts rather than a large general area of human body in [4]. Differently, our attention is composed of only a couple of learnable operations and each step is reasonable and easy to explain.
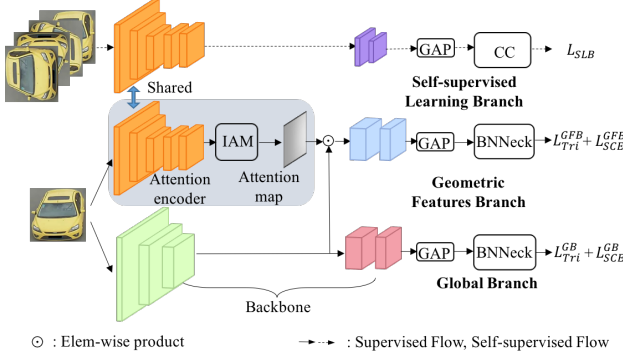
Figure 2: Overview of our framework: Self-supervised Geometric Features Discovery via Interpretable Attention, which consists of Global Branch (GB), Self-supervised Learning Branch (SLB), and Geometric Features Branch (GFB). Some key components are Interpretable Attention Module (IAM), Batch Normalization Neck (BNNeck) [37], Cosine Classifier (CC) [13], Global Average Pooling (GAP), hard mining Triplet loss (Tri) [20], and Smoothed Cross Entropy loss (SCE) [48].

Self-supervised learning. The success of self-supervised learning hinges on devising an appropriate pretext task to supervise model optimization. In this literature, a variety of visual tasks have been constructed, for instance, image completion [22], colorization [65], patch position prediction [9, 29], patch order prediction [26, 29] and rotation recognition [64, 11]. Besides, recently contrastive learning of multi-view images [5, 18] has demonstrated its efficacy. Furthermore, these pretext tasks can also be borrowed as an auxiliary task to strengthen a targeted one [13, 6]. Ours is significantly different from these works. We conduct self-supervised learning to facilitate geometric features discovery, which has not been explored by other works yet.

We are aware that recently Khorramshahi *et al.* [25] proposed their Self-supervised Attention for Vehicle Re-identification (SAVER) framework to pay attention to details of a vehicle body. Although our title shares terms "self-supervised" and "attention" with theirs, our approach is completely different from theirs. In principle, SAVER took the residual of removing the reconstructed image by Variational Auto-Encoder (VAE) [28] from an input as discriminative details for feature extraction, which they called "self-supervised attention". However, we actually propose a novel approach of borrowing self-supervised learning to regularize our interpretable attention learning. Besides, our proposal obviously differs from SAVER on these aspects, at least:

- Ours can robustly and consistently locate geometric features with physical interpretation across vehicle instances and viewpoints.
- Our deep framework is more concise with no need of

extra offline pretraining (like VAE in SAVER) or customized image pre-processing, *i.e.,* removing background noise from all images using object detector.
- Our results are significantly better. For instance, even on the most challenging testing scenario of VehicleID, *i.e.,* Large gallery size, our approach outperforms SAVER by 5.5% and 5.4% on Top-1 and Top-5 accuracy, respectively.

3. Self-supervised geometric features discovery via interpretable attention

As illustrated in Figure 2, in order to learn self-supervised geometric features as well as global representations simultaneously, our framework is composed of *Global Branch* (GB), *Self-supervised Learning Branch* (SLB) and *Geometric Features Branch* (GFB). Each branch has its own function and also interacts with each other. Generally, GB is employed to encode robust global codes from an input image. SLB performs the auxiliary self-supervised representation learning. By sharing its encoder with SLB, GFB is able to discover discriminative features from automatically discovered geometric locations without corresponding supervision. In remaining subsections, we elaborate each main component in turn.

3.1. Problem setup

Given a query image, vehicle ReID is to obtain a ranking list of all gallery images according to the similarity between query and each gallery image. The similarity score is typically calculated from deep embeddings, *i.e.,* $cos(f(x_q; \theta), \; f(x_g; \theta))$. Here $f(\cdot; \theta)$ represents a deep network with learnable parameters θ; x_q, x_g are query and gallery image respectively; $cos(\cdot)$ denotes cosine similarity computation. $f(\cdot; \theta)$ is optimized on a 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.

3.2. Self-supervised learning for highlighting geometric features

Self-supervised learning is equivalent to optimizing a deep network under the supervision of machine generated pseudo labels. Among them, image rotation degree prediction, *i.e.,* rotating image by a random angle and training a classifier to predict it, has demonstrated its capacity in many tasks [14, 64, 11, 29]. Vehicle ReID can be regarded as an instance-level classification problem, *i.e.,* all images contain the same species but many instances. Thus salient object in each image has similar geometry properties, *e.g.,* shape, outline, and skeleton. We argue that training a network to predict the rotation degree of a randomly rotated vehicle image encourages it to focus on these reliable and shared geometric properties (it is the same for per-
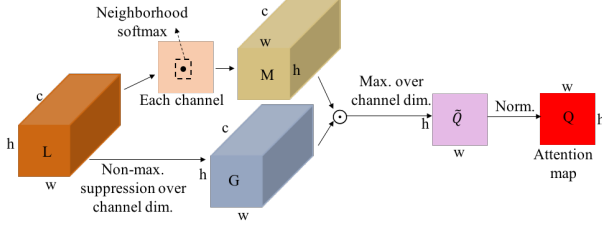
Figure 3: Interpretable Attention Module (IAM).

son ReID), which can help to easily recognize the rotation of an object. This geometric information has been proven crucial and discriminative for distinguishing a vehicle instance [55, 24] although it was represented by manually annotated keypoints as shown in Figure 1 (a).

Concretely, we first rotate an image x_i from D by $0°$, $90°$, $180°$ or $270°$ (assigning class 0, 1, 2 or 3 respectively) to generate a new dataset $D_{SL} = \{x_{i,r}, y_r\}_{r=1}^4$, $i = 1, ..., N$. Subsequently, the image $x_{i,r}$ is fed into an shared encoder $f_{ae}(\cdot; \theta_{ae})$ (namely attention encoder in Figure 2) to extract low-level semantics, $f_{ae}(x_{i,r}; \theta_{ae})$. To predict rotation class, high-level representations need to be further condensed from $f_{ae}(x_{i,r}; \theta_{ae})$. We append another deep module $f_{se}(\cdot; \theta_{se})$ to achieve this. Thus a high-dimensional embedding vector is obtained:

$$F_{SL}(x_{i,r}) = GAP[f_{se}(f_{ae}(x_{i,r}; \theta_{ae}); \theta_{se})], \quad (1)$$

where $GAP[\cdot]$ denotes Global Average Pooling operation. To generate more compact clusters in embedded space, the Cosine Classifier (CC) [13] is employed to assign the rotation class. The learnable parameters of CC is $W_{CC} = [w_1, \ldots, w_j, \ldots, w_b]$, $w_j \in \mathbb{R}^d$, where d is the dimension of vector F_{SL} and b is the number of classes (*i.e.,* $b = 4$). The probabilities of assigning the input image into each class can be represented as $P(x_{i,r}) = [p_1, \ldots, p_j, \ldots, p_b]$, where each element is

$$p_j = Softmax\left[\gamma\cos\left(F_{SL}(x_{i,r}), w_j\right)\right]. \quad (2)$$

$Softmax[\cdot]$ and γ represent respective normalized exponential function and a learnable scalar. Finally, the objective function of self-supervised learning is:

$$\mathcal{L}_{SLB} = \mathbb{E}_{D_{SL}}[CE(P(x_{i,r}), y_r)], \quad (3)$$

where $CE(\cdot)$ is Cross Entropy loss function. Obviously, the optimization of $\mathcal{L}_{SLB}$ enforces the deep classifier, especially the subnetwork $f_{ae}(\cdot; \theta_{ae})$, to capture geometric features from the input image.

3.3. Discriminative features discovery via interpretable attention

Through performing self-supervised learning (Section 3.2), low-level geometric features have been extracted by

the shared encoder $f_{ae}(\cdot; \theta_{ae})$. We argue that the best way to discovering discriminative local patterns is to aggregate spatial locations of high response and concentrate on corresponding features of these points. Definitely, the former step is pretty important and a spatial attention may be a choice of achieving this. However, existing works in attention learning usually have two well-known drawbacks: (1) the unexplainable workflows, *i.e.,* the architectures were usually heuristically devised and their loads of parameters were completely learned by themselves; (2) the scattered concerning areas, *i.e.,* the high response regions were too large to indicate discriminative patterns. Alternatively, we introduce an interpretable attention module whose deriving process can be reasonably explained and does not contain any learnable parameters. Furthermore, visualizations demonstrate that our attention can successfully focus on more accurate regions-of-interest that have physical meanings.

We illustrate the Interpretable Attention Module (IAM) in Figure 3, where $L \in \mathbb{R}^{c \times h \times w}$ is a 3D tensor extracted by $f_{ae}(\cdot; \theta_{ae})$ from an input image x_i and c, h, w denote the channel, height, and width dimension, respectively. To discover local points of interest on spatial dimensions, a $Softmax[\cdot]$ over neighborhood of each point is first conducted along each channel in L, *i.e.,*

$$M(k, u, v) = \frac{\exp\left(L(k, u, v)\right)}{\sum_{(m,n)\in\mathcal{N}(u,v)} \exp(L(k, m, n))}, \quad (4)$$

where $\mathcal{N}(u, v)$ denotes the squared neighborhood set with side length $\mathcal{K}$ around location (u, v) in the k-th channel. In parallel, a Non-Maximum Suppression (NMS) computing across all channels is performed from L to highlight important feature channels, *i.e.,*

$$G(k, u, v) = \frac{L(k, u, v)}{\max_{t=1,\ldots,c} L(t, u, v)}. \quad (5)$$

To take the local spatial maxima and channel-wise maxima above into account altogether, $\tilde{Q}$ is obtained by the element-wise product of M and G followed by the maximization over channel dimension, *i.e.,* $\tilde{Q}(u, v) = \max_{t=1,\ldots,c}\{M(t, u, v) \cdot G(t, u, v)\}$. Our final attention Q is obtained by the spatial normalization of $\tilde{Q}$, which considers all local maxima together and aggregates global points of interest:

$$Q(u, v) = \frac{\tilde{Q}(u, v)}{\sum_{(m,n)} \tilde{Q}(m, n)}. \quad (6)$$

Q represents the spatial emphasis of the activation tensor L, namely, crucial points of the input image x_i. So it is reasonable to weight another global representations extracted from x_i with Q as discriminative geometric features as in Figure 2. Our attention is partially inspired by the soft landmark detection in [10] but significantly different from theirs.

3.4. Overall optimization objectives

To optimize the whole framework, we combine CE loss from SLB, hard mining Triplet loss (Tri) [20] and Smoothed Cross Entropy loss (SCE) [48] from GFB and GB together as our final objectives. Tri and SCE loss are optimized referring to the combination mechanism of Batch Normalization Neck (BNNeck) [37]. Our overall objectives are

$$\mathcal{L}_{overall} = \lambda_{Tri}^{GB}\mathcal{L}_{Tri}^{GB} + \lambda_{SCE}^{GB}\mathcal{L}_{SCE}^{GB} + \lambda_{Tri}^{GFB}\mathcal{L}_{Tri}^{GFB} \quad (7)$$
$$+ \lambda_{SCE}^{GFB}\mathcal{L}_{SCE}^{GFB} + \lambda_{SLB}\mathcal{L}_{SLB}.$$

To avoid heavy tuning of hyperparameters, we simply set the importance coefficients λ_{Tri}^{GB}, λ_{SCE}^{GB}, λ_{Tri}^{GFB}, λ_{SCE}^{GFB} to 0.5 in all experiments. Only λ_{SLB} is fine-tuned in ablation study and set to 1.0 in final experiments.

During inference, SLB is abandoned. Two feature vectors from GB and GFB are concatenated as representations of an input image.

3.5. Network architecture

We give the architecture configurations in Figure 2 and each color represents a subnetwork. Referring to the literature, we choose ResNet50 [19], with $stride = 2$ in $conv5\_x$ replaced with $stride = 1$, as the backbone of GB. It is divided into two subnetworks, *i.e.,* the first ($conv1$, $conv2\_x$, $conv3\_x$) and the second ($conv4\_x$, $conv5\_x$), denoted by green and red respectively. The shared encoder between SLB and GFB is implemented by ResNet18 (orange) whose $stride$ in $conv4\_x$, $conv5\_x$ is set as 1. In SLB, another subnetwork (purple), consisting of two basic ResNet blocks [19] with $stride = 2$, is appended to the encoder to further condense features. In GFB, each image is first downsampled by 8 times by passing through the attention encoder and then the obtained tensor is processed by IAM to get the attention map. By element-wise multiplication, it is broadcast to every channel of the features from the first subnetwork of GB backbone, followed by another subnetwork (blue) composed of $conv4\_x'$, $conv5\_x'$.

4. Experiments

Datasets. We conduct experiments on three vehicle ReID benchmarks. *VeRi-776* [33] contains 49,357 images of 776 vehicles and 37,778 images of 576 identities compose its training set. *CityFlow-ReID* [50] is a challenging dataset where images are captured by 40 cameras under diverse environments. 36,935 images from 333 identities form the training set. *VehicleID* [31] is a large-scale benchmark containing 221,763 images of 26,267 vehicles. Its gallery set only contains one randomly selected image for each identity and thus we report our results as the mean over 10 trials. There are three numbers of gallery images widely used for

Method	Venue	ES	tmAP	imAP	Top-1	Top-5
OIFE [55]	ICCV17	Y	48.0	-	65.9	87.7
OIFE+ST [55]	ICCV17	Y	51.42	-	68.3	89.7
NuFACT [34]	TMM17	Y	53.42	-	81.56	95.11
VAMI [75]	CVPR18	Y	50.13	-	77.03	90.82
AAVER [24]	ICCV19	Y	58.52	-	88.68	94.10
RS [49]	ICCV19	Y	-	63.76	90.70	94.40
R+MT+K [49]	ICCV19	Y	-	65.44	90.94	96.72
VANet [8]	ICCV19	Y	66.34	-	89.78	95.99
PART [17]	CVPR19	Y	74.3	-	94.3	**98.7**
SAN [40]	MST20	Y	72.5	-	93.3	97.1
CFVMNet [47]	MM20	Y	-	77.06	95.3	98.4
PVEN [39]	CVPR20	Y	-	79.5	95.6	98.4
SPAN [7]	ECCV20	Y	68.9	-	94.0	97.6
DMML [3]	ICCV19	N	-	70.1	91.2	96.3
UMTS [23]	AAAI20	N	-	75.9	95.8	-
SAVER [25]	ECCV20	N	79.6	-	96.4	98.6
Ours	-	N	**86.2**	**81.0**	**96.7**	98.6

Table 1: Results comparison on VeRi-776.

testing, *i.e.,* 800 (Small), 1600 (Medium), and 2400 (Large).

Implementation. We choose PyTorch to implement our framework and Adam optimizer [27] with default betas ($\beta_1 = 0.9$, $\beta_2 = 0.999$), weight decay 5e-4 to optimize it. During training, random cropping, horizontally flipping, and erasing are performed to augment data samples. None of them is adopted to process testing images. All images are resized to 256×256 and experiments are conducted on one NVIDIA GEFORCE RTX 2080Ti GPU. The batch size on VeRi-776 and CityFlow-ReID is 28 and that on VehicleID is 40, with 4 images from each instance. On VeRi-776 and CityFlow-ReID, the initial learning rate is 1e-4 and the margin of triplet loss is set as 0.5 empirically. The number of training epochs is 80 and the learning rate is decreased by a factor of 0.1 at 20th, 40th, and 60th epoch. On VehicleID, the margin is 0.7 and the number of learning epochs is 120. The learning rate is increased linearly from 0 to 1e-4 during the first 10 epochs, decreased with cosine scheduler to 1e-7 at 100th epoch, and to 0 at the last epoch.

Evaluation protocols. Unlike some previous methods, we do not use any post-processing techniques like k-reciprocal re-ranking [71] to refine our results. We evaluate our approach by four widely used metrics in ReID literature, *i.e.,* image-to-track retrieval mean Average Precision (tmAP) (if tracks are available in one dataset), image-to-image retrieval mAP (imAP), Top-1, and Top-5 accuracy. Particularly, we report both tmAP and imAP on VeRi-776 for comprehensive evaluation. These scores are shown as percentages and the best are marked in bold. In Table 1, 2, and 3, ES (Y/N) indicates whether Extra Supervision besides ID labels is employed to train a corresponding method.

4.1. Performance comparison with SOTA works

VeRi-776. We compare our approach with SOTA ones in Table 1. We can see most works utilized extra supervisions

Method	Venue	ES	imAP	Top-1	Top-5
FVS [52]	CVPRW18	Y	5.08	20.82	24.52
RS [49]	ICCV19	Y	25.66	50.37	61.48
R+MT+K [49]	ICCV19	Y	30.57	54.56	66.54
SPAN [7]	ECCV20	Y	**42.0**	59.5	61.9
Xent [72]	arXiv19	N	18.62	39.92	52.66
Htri [72]	arXiv19	N	24.04	45.75	61.24
Cent [72]	arXiv19	N	9.49	27.92	39.77
Xent+Htri [72]	arXiv19	N	25.06	51.69	62.84
BA [30]	IJCNN19	N	25.61	49.62	65.02
BS [30]	IJCNN19	N	25.57	49.05	63.12
Ours	-	N	37.14	**60.08**	**67.21**

Table 2: Results comparison on CityFlow-ReID.

Method	Venue	ES	Small		Medium		Large	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
GoogLeNet [63]	CVPR15	Y	47.90	67.43	43.45	63.53	38.24	59.51
MD+CCL [31]	CVPR16	Y	49.0	73.5	42.8	66.8	38.2	61.6
OIFE [55]	ICCV17	Y	-	-	-	-	67.0	82.9
NuFACT [34]	TMM17	Y	48.90	69.51	43.64	65.34	38.63	60.72
VAMI [75]	CVPR18	Y	63.1	83.3	52.9	75.1	47.3	70.3
AAVER [24]	ICCV19	Y	72.47	93.22	66.85	89.39	60.23	84.85
VANet [8]	ICCV19	Y	**88.12**	97.29	83.17	95.14	80.35	92.97
PART [17]	CVPR19	Y	78.4	92.3	75.0	88.3	74.2	86.4
SAN [40]	MST20	Y	79.7	94.3	78.4	91.3	75.6	88.3
CFVMNet [47]	MM20	Y	81.4	94.1	77.3	90.4	74.7	88.7
PVEN [39]	CVPR20	Y	84.7	97.0	80.6	94.5	77.8	92.0
UMTS [23]	AAAI20	N	80.9	-	78.8	-	76.1	-
SAVER [25]	ECCV20	N	79.9	95.2	77.6	91.1	75.3	88.3
Ours	-	N	86.8	**97.4**	**83.5**	**95.6**	**80.8**	**93.7**

Table 3: Results comparison on VehicleID.

Figure 4: Discovered consistent geometric features from various viewpoints of the same vehicle (each row).

to achieve their performances. For instance, VANet annotated 5,000 images from each dataset to train a viewpoint predictor and learned distinct metrics for similar and dissimilar viewpoint pairs. PART defined three types of vehicle parts, *i.e.,* lights, windows, and brands, to train a YOLO [42]. When training the ReID model, they extracted local features from detected regions by YOLO as supplementary information for global representations. PVEN provided view segmentations of 3,165 images to train a U-Net segmentor, whose output mask was used for view-aware feature alignment when optimizing their model. Although no enhancement from extra labels was utilized in SAVER, Detectron [15] was needed to pre-process all images to remove background noise. In contrast, our approach does not involve any extra annotations to assist local feature learning. Although our training batch size 28 is much smaller than other methods (*e.g.,* 256 in SAN), our method can still outperform other competitors significantly on tmAP and imAP. Regarding Top-5 accuracy, ours is only 0.1% lower than the best that used a larger image size 512×512. That was demonstrated to promote their performances considerably [17]. When comparing under the same condition, ours are much better than theirs on all indicators.

CityFlow-ReID. The results are reported in Table 2. This dataset is quite challenging because images are taken from five scenarios, covering a diverse set of location types and traffic conditions. Results of metric learning methods (Xent, Htri, Cent, Xent+Htri) and batch-based sampling ones (BA, BS) are acquired without using extra annotations. To assist ReID, real and synthetic images were exploited by RS, while R+MT+K employed keypoints, vehicle type, and color class to perform multi-task learning. SPAN adopted vehicle orientation information to guide visible feature extraction and computed a co-occurrence part-attentive distance for each image pair. As we see, except for SPAN, our approach surpasses others by large margins on all three metrics, *e.g.,* ∼7.0% imAP, ∼6.0% Top-1, and ∼1.0% Top-5 accuracy compared with R+MT+K.

VehicleID. We list the results for comparison in Table 3.

Note that VANet and PVEN required much larger batch size 128 and 256, respectively. Even so, our approach beats all competitors in almost every test setting. In particular, compared with SAVER, ours achieves much better performances on all gallery sizes, *i.e.,* 6.9% Top-1, 2.2% Top-5 on Small, 5.9% Top-1, 4.5% Top-5 on Medium, 5.5% Top-1, 5.4% Top-5 on Large higher, although it involved some specific pre-processing steps.

4.2. Visualizations of discovered geometric features through self-supervision

We cover an input image with its attention map from GFB to visualize critical vehicle parts learned by our framework. Even though our geometric features are discovered without using accurate supervision like others, qualitative visualizations demonstrate the superiority of our method.

Comparison with defined landmarks by other works. Previous works manually annotated a specific number of landmarks on a vehicle body [24, 55] to assist their discriminative feature learning. These landmarks visible from each viewpoint (front, back, left or right) are illustrated in Figure 1 (a) which is borrowed from [55] and re-organized vertically. To compare with these human annotations thoroughly, we visualize our learned geometric features from

each viewpoint accordingly in Figure 1 (b). We can easily observe that our approach focuses on many similar locations to predefined ground truths on each viewpoint, *e.g., left (right)-front corner of vehicle top, left (right) fog lamp* on front view, *right-front corner of vehicle top, right-front (back) wheel, and right headlight* on right view, which demonstrates that our framework can successfully discover critical and informative vehicle parts for ReID without the supervision of ground truths.

Consistency across viewpoints and scenarios. To validate the consistency of learned geometric features across viewpoints and scenarios, we select a couple of images, belonging to an identical vehicle instance but taken from various viewpoints and by different cameras, for visualization in Figure 4. Each row represents an vehicle instance. Although viewpoint, object scale, and background of each image vary largely, identical vehicle parts, *e.g.,* fog lamps, vehicle tops, and wheels, are discovered for the same instance. This validates the stability and reliability of our approach in handling viewpoint and scenario changes, which are the key points of solving ReID problems.

Generalization to human part discovery. To demonstrate the generalization ability of our framework to person ReID, we conduct experiments on two popular benchmarks. Please refer to Section 4.4 for more experiment details and here we just analyze the visualization results shown in Figure 1 Bottom. As a deformable object, discovering geometric features from a human body is much more challenging. For saving space, we just select three images for each person. Obviously, identical human parts, *e.g.,* head, upper arms, and knees, are discovered by our approach even though human pose, viewpoint, and background change so much among images. These parts are critical to estimate human pose which has been demonstrated to play an important role in person ReID [61, 32, 45].

Discussion. As mentioned in Section 3.2, person or vehicle ReID is an instance classification problem, *i.e.,* all images in a task are taken from the same category but different individuals. So salient objects in these images have a lot in common, *e.g.,* geometric shapes (for vehicles), compositions, and skeletons. It is reasonable that conducting self-supervised learning encourages a deep network to discover these geometric features because they are reliable and repeatable clues to completing the self-supervised pretext task successfully. Visualizations in this section demonstrate this claim sufficiently. In view of their high similarity to ReID, we will expand our approach to other fine-grained classification tasks [54, 38] in future work.

4.3. Ablation study

To evaluate the effect of each proposal of our framework, we conduct extensive experiments on VeRi-776 and

Method	VeRi-776		CityFlow-ReID	
	tmAP	imAP	imAP	Top-1
GB w/o attention	84.0	78.3	32.04	56.27
GB+ResNet18 w/o attention	85.2	79.5	34.63	57.98
GB+GFB ($\mathcal{K} = 7$)	<u>85.9</u>	<u>80.7</u>	<u>36.63</u>	<u>59.98</u>
GB+GFB ($\mathcal{K} = 11$)	85.8	80.6	35.94	59.70
GB+GFB ($\mathcal{K} = 15$)	85.5	80.2	36.32	58.56
GB+GFB+SLB ($\lambda_{SLB} = 0.1$)	85.8	80.5	36.61	59.13
GB+GFB+SLB ($\lambda_{SLB} = 1.0$)	**86.2**	**81.0**	**37.14**	**60.08**
GB+GFB+SLB ($\lambda_{SLB} = 2.0$)	86.1	80.9	36.54	59.60

Table 4: Results of ablation study. We underline the corresponding results of selected values for $\mathcal{K}$ and λ_{SLB}. The performance improvement upon each component is consistent across datasets.

Figure 5: Learned attention maps (top) without and (bottom) with self-supervised learning from the same image.

CityFlow-ReID. Here we report tmAP and imAP, imAP and Top-1 on them respectively because these metrics are more important on each dataset. Results are in Table 4.

Effect of simply incorporating another branch upon baseline. Our framework employs a ResNet50 as the backbone of GB and a ResNet18 as the shared encoder between SLB and GFB. Although more branches and larger networks were usually utilized to perform ReID in previous works, we still conduct experiments to show that our performance gains upon the baseline (GB w/o attention) come from our proposals rather than an extra branch. To this end, we implement a new framework termed as "GB+ResNet18 w/o attention", consisting of two independent branches with respective ResNet50 and ResNet18 as backbone. Compared with our final results, we can see performances are marginally beneficial from adding a ResNet18 based branch. However, this also suggests that our framework can be stronger if involving more branches like others.

Pure IAM still bringing much improvement. IAM based GFB is the bridge of incorporating self-supervised learning (SLB) into our whole framework. To demonstrate the effectiveness of IAM even without the regularization from SLB, we conduct experiments "GB+GFB" with different $\mathcal{K}$ while fixing other hyperparameters. Results in the second part of Table 4 show that IAM are pretty robust *w.r.t.* values of $\mathcal{K}$. We set $\mathcal{K} = 7$ by default in subsequent experiments. Besides, comparing "GB+GFB" with "GB+ResNet18 w/o attention", it is seen that our IAM based GFB is much

Method	Venue	Market-1501			DukeMTMC-reID		
		imAP	Top1	Top5	imAP	Top1	Top5
DG-Net [70]	CVPR19	86.0	94.8	-	74.8	86.6	-
Bag [37]	TMM19	85.9	94.5	-	**76.4**	86.4	-
PCB [46]	ECCV18	81.6	93.8	97.5	69.2	83.3	90.5
RGA [67]	CVPR20	**88.4**	**96.1**	-	-	-	-
OSNet [73]	ICCV19	84.9	94.8	-	73.5	**88.6**	-
Ours	-	86.1	94.3	**98.3**	75.7	85.7	**93.6**

Table 5: Results comparison on person ReID benchmarks.

more powerful than a ResNet18 branch. For example, on the challenging CityFlow-ReID, the former and the latter bring improvement about 4.6% *vs.* 2.6% on imAP and 3.7% *vs.* 1.7% on Top-1 accuracy upon the baseline. This is attributed to the interpretable attention IAM. Although it is not able to discover specific parts without the help of self-supervised learning (referring to Figure 5), it can focus on a holistic vehicle body from cluttered background for extracting more effective representations.

Physically meaningful attention discovery through self-supervised learning. To enforce attention to emphasize crucial vehicle parts (*i.e.,* physically meaningful locations), we conduct experiments with the full framework "GB+GFB+SLB" with different λ_{SLB} while keeping other hyperparameters identical. The third part results in Table 4 tell that our framework are robust *w.r.t.* values of λ_{SLB} and we select $\lambda_{SLB} = 1.0$ as our decision. It is observed that self-supervised learning improves performances consistently on all metrics compared with "GB+GFB". Especially, from the attention maps comparison in Figure 5, we can see self-supervised learning helps to shift attentions distracted by background vehicles to the main concern. And our framework overcomes the interference from diverse background distractors, *e.g.,* traffic lights and road signs, and discovers meaningful vehicle parts successfully.

4.4. Generalizing to other ReID related tasks

In this section, we demonstrate the potential application of our approach in person ReID and multi-target multi-camera (MTMC) vehicle tracking.

Person ReID. Instead of identifying individual vehicles, this task aims to associate the same person in images taken from different cameras. We conduct experiments on Market-1501 [68] and DukeMTMC-reID [43], two most widely used benchmarks for person ReID. The training details keep identical to those on VehicleID. We compare our performances with recent works in Table 5. As we see, though our approach is not intentionally proposed and tuned for person ReID, its performances are still very promising. We believe it will perform much better if the hyperparameters are fine-tuned accordingly.

MTMC vehicle tracking. As a complicated video surveil-

Rank	1	2	3	4	5	6
Team ID	Ours	92	141	11	163	63
IDF1 Score	**0.4930**	0.4616	0.4552	0.4400	0.4369	0.3677

Table 6: MTMC vehicle tracking results comparison on AI City Challenge 2020.

lance task, MTMC vehicle tracking is commonly composed of four steps, *i.e.,* vehicle detection, multi-target single-camera (MTSC) tracking, vehicle re-identification, and tracklet synchronization. Among them, vehicle ReID is the crucial stage for a satisfactory tracking result. It is much more challenging than operating on well-calibrated ReID benchmarks because of large object-scale variation and heavy blur caused by distance changes between a camera and vehicles. To verify the generalization ability of our approach under cross-dataset testing, we perform experiments on the data provided by City-Scale Multi-Camera Vehicle Tracking of AI City 2020 Challenge [51] using our trained model on VeRi-776 without any fine-tuning. Considering that ReID is only our concern, we simply adopt an efficient MTMC tracking pipeline, similar to [41], to achieve this. Specifically, we first employ Mask R-CNN from Detectron2 [60] to detect vehicles from each video frame. Then we utilize Deep SORT [57] with the association strategy from [56] to perform MTSC vehicle tracking. Finally, our trained model is directly applied to capture ReID representations from cropped vehicle images, followed by tracklet synchronization with identical rules to [41]. Refer to [41] for more details due to page limitation. Our approach achieves 0.4930 regarding the official evaluation metric IDF1 score [50], which is much higher than 0.4585 from [41], although they trained their ReID model on the officially provided dataset. Besides, we compare our result with other submissions in Table 6 and ours outperforms others significantly.

5. Conclusion

In this paper, based on our observation that salient objects in ReID images share similar properties, we propose a novel framework to learn geometric features, without supervision from fine-grained annotations, for vehicle ReID through performing a self-supervised task. To this end, an interpretable attention module is also introduced to discover physically reasonable features. Comprehensive experiments demonstrate the effectiveness and generalization ability of our approach qualitatively and quantitatively. In the future, we plan to generalize it to addressing fine-grained classification problems.

Acknowledgement

References

[1] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399, 2018. 2

[2] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, October 2019. 2

[3] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. Deep meta metric learning. In *ICCV*, October 2019. 2, 5

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, October 2019. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *CVPR*, pages 12154–12163, 2019. 3

[7] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. *arXiv preprint arXiv:2008.11423*, 2020. 2, 5, 6

[8] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *ICCV*, pages 8282–8291, 2019. 2, 5, 6

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 3

[10] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, June 2019. 4

[11] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *CVPR*, pages 10364–10374, 2019. 3

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, June 2019. 2

[13] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, pages 8059–8068, 2019. 3, 4

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[15] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron, 2018. 6

[16] Haiyun Guo, Kuan Zhu, Ming Tang, and Jinqiao Wang. Two-level attention network with multi-grain ranking loss for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(9):4328–4338, 2019. 2

[17] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *CVPR*, June 2019. 2, 5, 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3, 5

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[23] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11165–11172, 2020. 5, 6

[24] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *ICCV*, pages 6132–6141, 2019. 2, 4, 5, 6

[25] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. *arXiv preprint arXiv:2004.06271*, 2020. 3, 5, 6

[26] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 3

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[29] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, June 2019. 3

[30] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019. 6

[31] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 2, 5, 6

[32] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), June 2018. 7

[33] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*, pages 869–884. Springer, 2016. 2, 5

[34] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017. 2, 5, 6

[35] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2

[36] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(8):3794–3807, 2019. 2

[37] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019. 3, 5, 8

[38] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 7

[39] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 2, 5, 6

[40] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. *Measurement Science and Technology*, 2020. 5, 6

[41] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 588–589, 2020. 8

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 6

[43] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 8

[44] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *ICCV*, pages 1900–1909, 2017. 2

[45] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3960–3969, 2017. 7

[46] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 8

[47] Ziruo Sun, Xiushan Nie, Xiaoming Xi, and Yilong 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*, pages 3523–3531, 2020. 5, 6

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 3, 5

[49] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *ICCV*, pages 211–220, 2019. 2, 5, 6

[50] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR*, pages 8797–8806, 2019. 2, 5, 8

[51] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR*, June 2019. 8

[52] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In *CVPR Workshops*, pages 108–115, 2018. 6

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 7

[55] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *ICCV*, pages 379–387, 2017. 1, 2, 4, 5, 6

[56] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019. 8

[57] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 8

[58] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module.

In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[59] Fangyu Wu, Shiyang Yan, Jeremy S Smith, and Bailing Zhang. Joint semi-supervised learning and re-ranking for vehicle re-identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 278–283. IEEE, 2018. 2

[60] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 8

[61] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[62] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *ICCV*, pages 562–570, 2017. 2

[63] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015. 6

[64] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, October 2019. 3

[65] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, pages 1058–1067, 2017. 3

[66] Yiheng Zhang, Dong Liu, and Zheng-Jun Zha. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1386–1391. IEEE, 2017. 2

[67] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020. 8

[68] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 8

[69] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *arXiv preprint arXiv:2004.06305*, 2020. 2

[70] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[71] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 5

[72] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019. 6

[73] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 8

[74] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *ICCV*, pages 8040–8049, 2019. 2

[75] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *CVPR*, June 2018. 2, 5, 6