Robust and Scalable Vehicle Re-Identification via Self-Supervision

Pirazh Khorramshahi, Vineet Shenoy and Rama Chellappa
Artificial Intelligence for Engineering and Medicine (AIEM) Lab, Johns Hopkins University
{pkhorra1,vshenoy4,rchella4}@jhu.edu

Abstract

Many state-of-the-art solutions for vehicle re-identification (re-id) mostly focus on improving the accuracy on existing re-id benchmarks using additional annotated data. To balance the demands of accuracy, availability of annotated data, and computational efficiency, we propose a simple yet effective hybrid solution empowered by self-supervised learning which is free of intricate and computationally-demanding add-on attention modules often seen in state-of-the-art approaches. Through extensive experimentation, we show our approach, termed Self-Supervised and Boosted VEhicle Re-Identification (SSBVER), is on par with state-of-the-art alternatives in terms of accuracy without introducing any additional overhead during deployment. Additionally, we show that our approach, generalizes to different backbone architectures which accommodates various resource constraints and consistently results in a significant accuracy boost. Our code is available at https://github.com/Pirazh/SSBVER.

1. Introduction

The problem of vehicle re-id is a retrieval task in which a query vehicle image is presented and matches are retrieved from a large gallery set. The gallery is composed of vehicle images captured at different times of day, from traffic cameras mounted at different locations and under varying weather conditions. Vehicle re-id task becomes quite challenging as a given vehicle's appearance can drastically vary under different viewpoints, camera and lighting conditions. On the other hand, many vehicles can appear similar due to relatively small variations in vehicle manufacturers, models, trims, years and colors. To address this task and associated challenges, discriminative visual representation learning via Deep Neural Networks (DNNs) has become the de facto approach. Note that vehicle re-id is objectively different than vehicle classification task where the goal is to identify a vehicle's model rather than its instance. Therefore, Vehicle re-id requires more fine-grained features, particularly within local regions, to highlight differences in similar looking vehicles. As a result, several research works have been undertaken to integrate attention mechanisms into the DNNs' pipeline in both implicit [17, 41, 44] and explicit ways [11, 18]. While these approaches are successful in improving the state-of-the-art, they often require rich data annotations and demand heavy computation that raise scalability issues. The burden of deploying such models in real-time applications such as city-scale multi-camera tracking quickly becomes evident as hundreds of traffic cameras should be processed simultaneously under limited computational resources. Consequently, it is paramount to design a vehicle re-id module that effectively learns discriminative representations without relying on the existence of additional annotations beyond ID labels, *e.g.* vehicle's manufacturer, model, color, key-points or parts' location.

Recently, there have been great strides in the area of Self-Supervised Learning (SSL) particularly for the task of image classification to learn robust embeddings without using human-generated labels. As a result, the performance gap between self-supervised and fully-supervised learning has become smaller. In addition, SSL methods outperform mainstream supervised pretraining when transferred to down-stream tasks such as object detection and demonstrate better data efficiency [3, 5, 9, 12]. This has motivated us to explore the viability of recent self-supervised learning techniques in the context of vehicle re-id. A great number of recent works in SSL classify [8] or discriminate [5, 12] each image as a separate class known as *Instance Classification* and *Instance Discrimination* respectively via contrastive learning. While these approaches yield robust representations for the image classification, they cannot be extended to object re-id where there are multiple images corresponding to the same ID which should not be discriminated against one another. To address this issue, supervised contrastive learning [21], a generalization of Triplet loss [43], has been proposed so that similar images are considered as positives during training. This is identical to the current practice in object re-id which employs triplet loss as standard. In contrast, the recent SSL method [4] casts the SSL as self-distillation and establishes the connection between

Knowledge Distillation and SSL in the absence of labels without performing any discriminatory task among images. As we discuss in section 3, this creates the opportunity to re-formulate a novel training framework to enrich the learning of a re-id model with the self-supervisory signal and encourages a simple baseline to build local to global correspondences, without any explicit attention mechanism. Our contributions can be summarized as the following:

1- Introduction of a novel training framework for vehicle re-identification that is a hybrid of supervised and self-supervised objective functions.

2- Presenting a simple, efficient and accurate design with no intricate attention modules that requires no extra labels for training and no extra overhead for inference.

4- Achieving state-of-the-art results for VeRiWild dataset in terms of Cumulative Match Curve metrics.

3- Advocating for the joint optimization of accuracy and efficiency when designing vehicle re-id systems to account for deployment constraints.

The rest of the paper is organized as follows. Section 2 reviews recent works in vehicle re-id. Proposed method and its detailed architecture is discussed in section 3. Through extensive experimentation in section 4, we show the effectiveness of our approach with different backbone architectures on multiple challenging vehicle re-id benchmarks, and obtain state-of-the-art in terms of accuracy-efficiency trade-off. We also highlight the benefits of our model in a multi-camera vehicle tracking scenario. In section 5 we further analyze our method and validate our design choices. Section 6 concludes the paper.

2. Related Work

Vehicle re-id has recently attracted a significant amount of attention thanks to its critical role in the development of smart transportation technologies. Here we review a number of selected works that has been published recently.

Learning discriminative features for vehicles demands curated datasets of vehicles' images of diverse makes, models, colors with high number of identities. Several datasets have been introduced over the past several years which contributed to the current landscape of vehicle re-id. Among these are VeRi [25], VehicleID [24], VeRiWild [30], Vehicle1M [10] and CityFlow Re-id [40]. Each of these has different attributes and variations in terms of scale and resolution; however, only VeRi, VeRiWild, and CityFlow Re-id capture vehicles from diverse views that is more representative of unconstrained vehicle re-id. Since vehicle re-id is concerned with subtle cues and small-scale details on vehicle images, [42] annotated images in VeRi dataset with view point and key-point information such as the location of logo, and head and tail lights. This helps to devise supervised attention models to adaptively extract local features based on vehicle's orientation [18]. Similarly, [11] annotated images in VehicleID dataset with parts' bounding box information to detect and extract fine-grained features. While having extra annotations help to learn where to look for discriminative information, it is not scalable due to the increasing number of new vehicle models year in year out. To address this issue, a variational auto-encoder model was developed in [19, 20, 36] to generate coarse vehicle images and obtain self-supervised saliency maps which highlight identity-dependant information to either adjust images directly or excite intermediate feature maps of an underlying DNN, inspired by the idea of Curriculum learning [2]. Similarly, [23] proposed a self-supervised model based on the pretext task of image rotation to learn geometric features. A jigsaw patch module for vision transformers was introduced in [15] which forces local features to be globally distinguishable. As orientation is one of the factors to negatively bias the learned embeddings, [32] and [1] propose to learn view-aware aligned features and to disentangle orientation from visual features respectively. To extract region-specific features, [46] introduced a heterogeneous relational graph-based model to encode the relation of the different local regions into a unified representation. These methods are mainly designed to enhance the re-id accuracy; consequently, efficiency and large-scale deployment has not been considered as discussed in 4.6 section. Therefore, we present SSBVER, a hybrid learning approach that employs the power of self-supervision to boost vehicle re-id performance while preserving the computational complexity and inference time of a baseline model.

3. Method

In this section we discuss the details of the proposed Self-Supervised Boosted Vehicle Re-identification pipeline shown in Fig. 1.

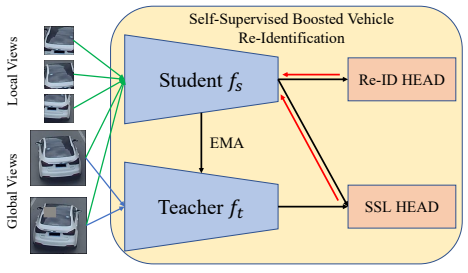


Figure 1. Self-supervised and Boosted Vehicle Re-identification.

3.1. Backbone Feature Extractors

Inspired by recent SSL methods, our approach benefits from a student and teacher pairing, where both have iden-

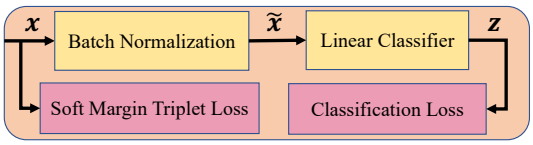


Figure 2. Re-Id Head: Extracted visual features x from the student model f_s are passed through a bottleneck layer implemented by 1-dimensional batch normalization to obtain feature $\tilde{\mathbf{x}}$. Subsequently, classification logits $\mathbf{z}$ are obtained from a linear classifier. Soft-margin Triplet and Cross Entropy loss functions constrain $\mathbf{x}$ and $\mathbf{z}$ respectively.

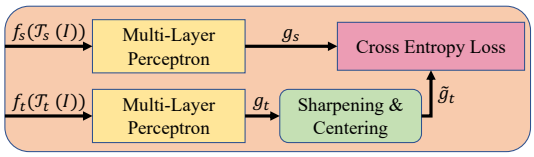


Figure 3. Self-Supervised Learning Head: Image I is randomly augmented using $\mathcal{T}_s$ and $\mathcal{T}_t$ transformations and passed to student and teacher networks followed by multi-layer perceptrons to obtain g_s and g_t prediction vectors. After applying sharpening and centering to the teacher's output to avoid collapse, the cross entropy of student and teacher predictions is minimized.

tical architectures. The choice of architecture is arbitrary and can be selected based on the application and resource constraints. In this work, we adopt multiple candidate architectures including ResNet [13], ResNet_IBN [35], Vision Transformer (ViT) [7], SWIN Transformer [26], and ConvNext [27] to study the generalization capability of SSBVER. The teacher model is considered a momentum encoder as it is a low-pass version of the student via taking the exponential moving average over the course of training iterations with the momentum parameter λ, *i.e.* ${\theta_t}^i = \lambda {\theta_t}^{i-1} + (1-\lambda){\theta_s}^i$ where θ_t, θ_s and i are teacher and student model parameters and the current training iteration respectively. Note that both models are initialized from the same set of ImageNet pre-trained weights, *i.e.* $\theta_t^0 = \theta_s^0$.

3.2. Re-Identification Head

SSBVER uses the re-id head to constrain the extracted features $\mathbf{x}$ by the student model so that those corresponding to the same identity are embedded close together while the ones belonging to different identities kept apart. This goal is realized by employing Triplet loss in conjunction with Cross Entropy loss and results in a strong baseline model as demonstrated in prior works [14, 19, 31, 36]. Fig. 2 outlines the inner workings of the re-id head. The soft-margin triplet loss with batch-hard sampling method is computed via the following formulation:

$$\mathcal{L}_t = \sum_{a \in b_i} \log(1 + \exp(\max_{p \in \mathcal{P}(a)} ||\mathbf{x}_a - \mathbf{x}_p||_2 - \min_{n \in \mathcal{N}(a)} ||\mathbf{x}_a - \mathbf{x}_n||_2)) \quad (1)$$

In Eq. 1, b_i denotes the i^{th} training batch. In addition, a, $\mathcal{P}(a)$ and $\mathcal{N}(a)$ are an anchor sample and its corresponding positive and negative sets defined within b_i. Once the representation vector $x \in \mathbb{R}^d$ is computed, it is passed to a batch normalization layer to obtain $\tilde{\mathbf{x}}$. Authors in [31] showed that employing this bottleneck layer helps the consistency of Triplet and Cross Entropy loss functions in the context of re-id. Afterwards, the linear classifier computes the class logit vector $\mathbf{z} \in \mathbb{R}^k$ (k is the number of training IDs) through linear operation $\mathbf{z} = W\tilde{\mathbf{x}} + B$. $W \in \mathbb{R}^{k \times d}$ and $B \in \mathbb{R}^k$ are the weight matrix and bias of the classifier correspondingly. The classification loss is computed as follows:

$$\mathcal{L}_c = -\sum_{j=1}^{k} y_j^i \log \hat{y}_j^i, \quad \hat{y}_j^i = \frac{e^{z_j^i}}{\left(\sum_{m=1}^{k} e^{z_m^i}\right)} \quad (2)$$

$\hat{y}_j^i$ is the prediction probability that i^{th} sample belongs to the class j. In addition, we employ label-smoothing as a regularization method following the work of [39]. Therefore, rather than considering the ground-truth vector as a one-hot encoding vector, it is computed as:

$$y_j^i = \begin{cases} 1 - \frac{k-1}{k}\epsilon & j = k(i) \\ \frac{\epsilon}{k} & \text{otherwise} \end{cases} \quad (3)$$

where $\epsilon \in [0, 1]$ and $k(i)$ are a hyper-parameter and the class label of the i^{th} sample. While optimizing Triplet and Cross Entropy classification loss functions on the extracted representations results in a strong and efficient baseline re-id model, incorporation of an attention mechanism to focus on local regions of vehicle images and extract subtle cues can further improve the performance. However this is achieved at the expense of increased complexity and computation time that can be prohibitive when applied to large-scale and real-time scenarios. To overcome this shortcoming with the goal of minimizing inference complexity while enjoying enhanced accuracy, we incorporate an SSL objective in the training phase to encourage local to global correspondence and to mimic an attention mechanism. This is discussed in the following section.

3.3. Self-Supervised Learning Head

To improve the performance of re-id without the incorporation of any additional annotation on vehicles' parts and attributes, we propose to apply self-supervised optimization objective based on self-training and knowledge distillation. [4] presents a self-supervised learning paradigm with multi-crop strategy [12] in which semantically rich representations can be learned and demonstrates competitive performance when transferred to down-stream tasks. Compared to

contrastive learning-based methods, knowledge distillation-based approach does not solve the instance discrimination task and therefore does not rely on any negative sampling; this is a fitting choice for re-id as each identity is represented by multiple images that should not be discriminated against each other.

After sampling an image I, we create two sets of views on the fly, namely $V_g(I)$ and $V_l(I)$, where $V_g(I) = \{I_{g_1}, I_{g_2}\}$ contains two different global views and $V_l(I) = \{I_{l_1}, \ldots, I_{l_L}\}$ has L local views of image I. Images in $V_g(I)$ are generated by $\mathcal{T}_t$ with randomly cropping a region with random area ratio a_g, padding zeros on the edges, flipping horizontally, jittering colors and erasing a random patch to simulate occlusion [16]. To generate images in $V_l(I)$, $\mathcal{T}_s$ crops a random portion of image I with random area ratio of a_l, randomly flips in horizontal direction and jitters color. The teacher only processes images in $V_g(I)$ while the student model is fed by images in both sets *i.e.* $V_g(I) \cup V_l(I)$. As shown in Fig. 3, when $\mathbf{x}_s = f_s(\mathcal{T}_s(I))$ and $\mathbf{x}_t = f_t(\mathcal{T}_t(I))$ are obtained, they are mapped to another space using multi-layer perceptrons (MLP) with four hidden layers, and Gaussian Error Linear Units (GELU) to yield E-dimensional vectors g_s and g_t. A common problem that is associated with SSL-based approaches with a pair of networks, is the issue known as collapse where both encoders learn to output trivial embeddings irrespective of the input images to minimize the respective loss function. There has been a number of techniques to prevent collapse including contrastive learning with negative pairs [5], stop-gradient [6], clustering [3], momentum encoder [9], and redundancy reduction of the learnt representation's dimensions [45]. Knowledge distillation objective is optimized by minimizing the cross entropy loss between student and teacher networks' output so that student can match teacher's prediction. While it uses momentum encoder and stop gradient techniques to battle collapse, collapse can still occur in the form of either outputting uniform predictions or having a single dimension dominating others. To counteract, centering and sharpening of the teacher's outputs are proposed [4]. In centering, an exponential moving average $\mathbf{c}$ of teacher's predictions is recorded and subtracted from its predictions to prevent the domination of a single dimension. On the other hand, a relatively small temperature is applied to the teacher network to battle the uniformity of the outputs. Sharpening and centering operations attempt to establish a balance in which collapse does not occur. This cross entropy loss is calculated with the following formulation:

$$\mathcal{L}_s = -\sum_{I \in V_g(I)} \sum_{I' \in V_g(I) \cup V_l(I),\, I' \neq I} \sum_{i=1}^{E} p_t^i(I) \log(p_s^i(I')) \quad (4)$$

Where $p_s^i(I') = \frac{\exp(g_s^i(f_s(I'))/\tau_s)}{\sum_j \exp(g_s^j(f_s(I'))/\tau_s)}$ and $p_t^i(I) = \frac{\exp((g_t^i(f_t(I)) - c^i)/\tau_t)}{\sum_j \exp((g_t^j(f_t(I)) - c^j)/\tau_t)}$. Also τ_s, τ_t are student and teacher's temperatures. c^i is the i^{th} element of the vector $\mathbf{c}$ that is the exponential moving average of g_t. Note that the cross entropy is only calculated when student and teacher process different augmentations of an image, *i.e.* $I' \neq I$.

3.4. End-to-End Training

We first establish a baseline model setup in which only $\mathcal{L}_c$ and $\mathcal{L}_t$ are used to train the student. The teacher which is the exponential moving average of the student model over the training iterations is used for evaluation. Afterwards, the setup of the SSBVER outlined in Fig. 1 is used for model training and the total loss function is calculated as follows:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_c + \lambda_t \mathcal{L}_t + \lambda_s \mathcal{L}_s \quad (5)$$

Coefficients λ_c, λ_t, and λ_s are the weights corresponding to each of the loss terms and are empirically set. We emphasize that gradients of the loss functions in Eqs. 4, 2, 1 are computed with respect to only student's parameters θ_s.

4. Experimental Results

To evaluate SSBVER method and understand how much it can benefit the re-identification without introducing any additional annotations and overhead during test time, we use the three widely used VeRi, VehicleID and VeRiWild datasets. Additionally, we use ResNet, ResNet_IBN, ViT, SWIN and Convnext backbone models to study the extent to which SSBVER generalizes to different architectures. Here we discuss vehicle re-id datasets, evaluation metrics, implementation details, and present our experimental results.

4.1. Datasets

VeRi is the first multi-view vehicle re-id dataset. It is regarded as a large-scale dataset; however, compared to the size of more recent datasets it is relatively small. The training and testing sets contain $37,778$ and $13,257$ images of 576 and 200 vehicle identities respectively.

VehicleID is a comparatively larger benchmark as it contains $113,346$ ($108,417$) images of $13,164$ ($13,103$) unique vehicles in the training (testing) set. In contrast to VeRi, images in Vehicle ID are mainly captured from either front or rear of vehicles which impacts the dataset's representativeness of the real-world scenarios. For evaluation, multiple splits of different sizes are created from the original test set and referred to as small, medium and large containing 800, 1600, and 2400 unique identities.

VeRiWild with $416,314$ images of $40,671$ individual identities is the largest multi-view vehicle re-id dataset in the wild that is captured via 174 traffic cameras and have variations in lighting and weather conditions. The train set contains $277,797$ images of $30,671$ identities and test set is

split into three small, medium and large sets of 3000, 5000, and $10,000$ unique identities.

4.2. Evaluation Metrics

Mean Average Precision (mAP) and **Cumulative Match Curve (CMC)** are widely adopted in the re-id community to measure the success of re-id systems. Upon receiving a query image, visual representations are computed for query and the entire gallery. Afterwards, a distance measure, *e.g.* Euclidean or Cosine, is used to compute the similarity scores and rank the gallery. mAP shows how well the gallery is ranked with respect to the query image. All the corresponding true matches to the query identity participate in the calculation of mAP. CMC @ k yields the probability that there exists at least one correct match to the query image in the top k items in the ranked gallery.

4.3. Implementation Details

In the baseline experiments we record the exponential moving average with momentum parameter $\lambda = 0.9995$ of the feature extractor' parameters to replicate the teacher in SSBVER, which is used for evaluation. Total training epochs is set to 120. Label smoothing parameter is set to $\epsilon = 0.2$. To create global and local views $a_g \in [0.8, 1]$ and $a_l \in [0.1, 0.4]$ are randomly selected. For ResNet and ResNet_IBN architectures, we use the Adam [22] optimizer, learning rate of $\eta = 0.0005$ with Gamma decay factor $\gamma = 0.1$ at $40^{th}, 70^{th}, 100^{th}$ epochs and weight decay of 0.001. For ViT, SWIN and Convnext we use the *base* model variant, AdamW [29] optimizer and cosine learning rate decay scheduling [28] with $\eta_{max} = 0.0001$ and $\eta_{min} = 1.6e-5$. In addition, weight decay is set to 0.0001. Finally, linear learning rate warm-up (with rate 0.099) is adopted for the first ten epochs. In the SSL head, the student's temperature is set to $\tau_s = 0.1$ while the teacher's temperature τ_t is increased linearly from 0.0005 to 0.001 in the first ten epochs and remains fixed for the rest of training.

4.4. Evaluation Results

4.4.1 VeRi Dataset

Table 1 reports experimental results on VeRi dataset. It is seen SSBVER performs better than baseline in almost every evaluation metric and architecture with the exception of ViT for the CMC@1. We should note that CMC@1 is more sensitive compared to other metrics as it only considers the first item in the ranked gallery which is either a hit or miss. This can be attributed to the fact that ViT has a high capacity to learn in data-abundant regime as noted in [7] which is not the case for VeRi. In addition, we highlight that the performance of ResNet50 and ResNet50_IBN models are significantly improved by adopting self-supervision compared to architectures with larger number of parameters that can easily overfit the data and suffer from high variance. We would like to highlight that the performance gained here is cost-free for inference in that SSLBVER preserves the complexity of the baseline model.

Table 1. Comparison of SSBVER model against baseline on VeRi dataset. Note that bold black figures denote the higher performance for each architecture while bold red figures are the highest among all models and architectures.

Architecture	Model	Evaluation Metric		
		mAP (%)	CMC@1 (%)	CMC@5 (%)
ResNet50	Baseline	78.03	95.89	97.85
	SSBVER	**80.94**	**97.02**	**98.45**
ResNet50_IBN	Baseline	79.88	96.13	97.97
	SSBVER	**82.11**	**97.08**	**98.45**
ViT_Base	Baseline	77.72	**96.48**	98.39
	SSBVER	**77.74**	95.95	**98.51**
SWIN_Base	Baseline	78.40	95.65	97.85
	SSBVER	**79.35**	**95.74**	**97.91**
ConvNext_Base	Baseline	78.73	96.13	98.03
	SSBVER	**79.01**	**96.36**	**98.27**

Table 2. Performance comparison between SSBVER and baseline models on VehicleID dataset.

Architecture	Model	Evaluation Metric								
		mAP (%)			CMC@1 (%)			CMC@5 (%)		
		S	M	L	S	M	L	S	M	L
ResNet50	Baseline	88.77	86.05	82.91	82.75	80.26	76.79	97.00	94.06	90.92
	SSBVER	**90.73**	**86.57**	**83.82**	**85.61**	**80.34**	**77.26**	**97.73**	**94.92**	**92.59**
ResNet50_IBN	Baseline	89.19	84.95	82.73	83.44	78.81	76.79	96.82	93.13	90.53
	SSBVER	**90.88**	**87.36**	**84.83**	**85.61**	**81.62**	**78.91**	**97.72**	**94.92**	**92.60**
ViT_Base	Baseline	88.70	84.88	82.65	82.50	78.53	76.33	97.22	**93.61**	90.75
	SSBVER	**89.09**	**85.23**	**83.13**	**82.93**	**79.05**	**76.64**	**97.33**	93.56	**91.78**
SWIN_Base	Baseline	89.77	86.74	84.35	83.84	80.86	77.90	97.61	94.91	92.17
	SSBVER	**90.58**	**86.98**	**84.68**	**85.19**	**81.02**	**78.62**	**97.96**	**95.08**	**93.27**
ConvNext_Base	Baseline	88.95	85.54	83.14	82.84	**79.46**	76.69	97.03	93.69	**91.51**
	SSBVER	**89.10**	**85.81**	**83.24**	**83.42**	79.38	**77.13**	**97.17**	**94.17**	91.44

4.4.2 VehicleID Dataset

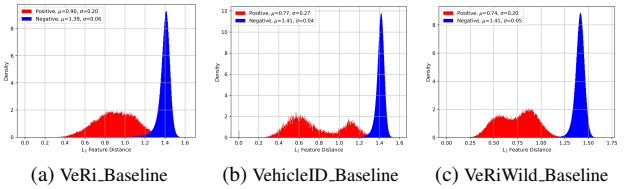
Test set of VehicleID has three splits: small, medium and large which are enumerated by S, M, and L in Table 2. Images in VehicleID are only captured from either front or rear and the extent to which a network can exploit small-scale information in overlapping views is limited. Similar to VeRi, self-supervised objective contributes to performance improvement across all evaluation metrics and architectures. Due to relatively larger size compared to VeRi, the performance of bigger models, namely ViT, SWIN and ConvNext is more pronounced. The superior performance of SWIN compared to ViT shows the benefit of hierarchical design and multi-resolution feature maps in a transformer-based model as it can extract information at various scales.

4.4.3 VeRiWild Dataset

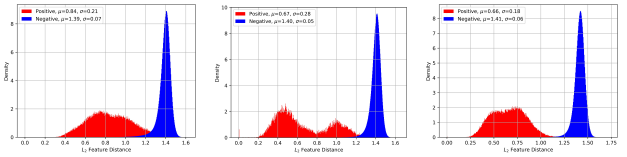
The test set is split into three small, medium and large sets consisting of 41861, 69389, and 138517 images respectively. The performance metrics are reported in Table 3. For this multi-view dataset the benefit of self-supervision is quite evident as every evaluation metric across all test splits and architectures is improved by a significant margin. This

Table 3. Performance comparison between SSBVER and baseline models on VeRiWild dataset.

Architecture	Model	Evaluation Metric								
		mAP (%)			CMC@1 (%)			CMC@5 (%)		
		S	M	L	S	M	L	S	M	L
ResNet50	Baseline	78.20	72.43	64.43	93.14	90.62	86.93	97.82	96.89	94.71
	SSBVER	**80.41**	**74.77**	**67.02**	**93.88**	**91.44**	**88.26**	**98.03**	**96.93**	**94.98**
ResNet50_IBN	Baseline	81.46	75.74	67.70	93.24	90.76	86.41	97.82	96.51	94.20
	SSBVER	**82.64**	**77.49**	**70.09**	**95.11**	**93.37**	**90.14**	**98.53**	**97.45**	**95.67**
ViT_Base	Baseline	81.76	76.13	67.71	93.44	91.56	86.77	98.59	97.57	95.55
	SSBVER	**83.81**	**78.25**	**70.55**	**94.98**	**92.71**	**89.65**	**98.69**	**97.83**	**95.98**
SWIN_Base	Baseline	84.94	79.64	71.93	94.58	92.05	87.89	98.80	97.55	95.97
	SSBVER	**86.05**	**81.28**	**74.07**	**95.62**	**93.75**	**90.27**	**99.10**	**98.23**	**96.76**
ConvNext_Base	Baseline	83.44	78.12	69.93	93.74	91.32	86.69	98.33	97.61	95.59
	SSBVER	**84.34**	**79.08**	**71.29**	**94.21**	**92.29**	**88.14**	**98.76**	**97.75**	**96.10**



(a) VeRi_Baseline (b) VehicleID_Baseline (c) VeRiWild_Baseline



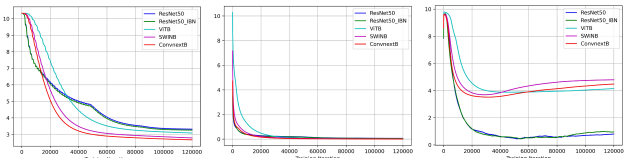
(d) VeRi_SSBVER (e) VehicleID_SSBVER (f) VeRiWild_SSBVER

Figure 4. Distribution of the distances in feature space. Embeddings are extracted from ResNet50_IBN architecture.

shows distillation-based SSL objective effectively regulates model training to exploit more fine-grained features that are favorable for the vehicle re-id task. Moreover, because of the large number of training samples, ViT, SWIN, and Convnext achieve substantially higher performance compared to ResNet-based models. **SSBVER model with SWIN backbone achieves the highest performance on VeRi-Wild among all published research works.**

4.4.4 Intra-class Compactness & Inter-class Separation

We are interested to know how the incorporation of self-supervision impacts the L_2 distance between extracted features. To this end, we plot the distribution of Euclidean distances of positive and negative image pairs in the feature space. Qualitatively, this measures the intra-class compactness and inter-class separation. Fig. 4 shows this comparison between the baseline and self-supervised boosted models across the test set of different re-id benchmarks. It is seen that SSL objective reduces the mean μ of positive pair distance distribution by 6.6%, 12.9%, and 10.8% for VeRi, VehicleID, and VeRiWild respectively. However, the means of negative pair distance distributions are roughly unchanged. This analysis shows that SSBVER helps the intra-class compactness since the student model is constrained to match the predictions of the teacher model for



(a) Classification Loss (b) Triplet Loss (c) Self-supervised Loss

Figure 5. Progression of objective functions involved in the SSBVER pipeline over the course of training for VeRiWild.

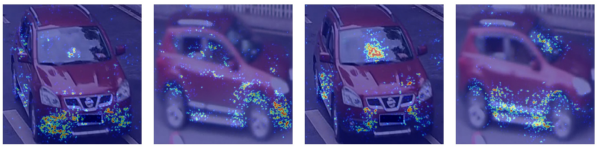
different views of a same object. Interestingly, for the case of VehicleID, there are two prominent peaks in the distribution of positive pair distances. As images are either captured from the rear or front of vehicles, for positive pairs, the distance is small when both images are from the same views and is larger when they are from opposing views. Therefore, two peaks stand out in the corresponding distribution.

4.4.5 Convergence Analysis

We examine the convergence of SSBVER model. We plot the $\mathcal{L}_c$, $\mathcal{L}_t$, and $\mathcal{L}_s$ during the course of training on VeRi-Wild in Fig. 5. Since we use label smoothing for classification objective, its does not converge to zero. However, Convnext, SWIN, and ViT achieve lower classification objective compared to ResNet-based models due to their higher capacity to fit the data. Triplet loss converges to zero for all models; although, for ViT over the initial training iterations the maximum L_2 distance between the features of positive pairs is significantly larger than the minimum distance between the features of negatives resulting in a higher objective value. This can be attributed to the fact that unlike ResNet, SWIN, and Convnext, ViT does not have a hierarchical design and instead has a global receptive field from the first layer. Self-supervised objective evolves differently for ViT, SWIN, and ConvNext compared to ResNet-based architectures. While collapse is avoided and the $\mathcal{L}_s$ converges for both groups, it converges to a much lower value for ResNet-based models. This difference can be potentially justified based on the fact that ViT, SWIN, and ConvNext all use patchification strategy in their initial layer compared to the down-sampling in ResNet-based models. Down-sampling can provide a better chance to match teacher's prediction and achieve a lower objective value.

4.4.6 SSL: An Implicit Attention Mechanism?

We visualize the regions in a query and gallery image pair that are most sensitive to the similarity score obtained for the pair. More precisely, we compute the input saliency maps for query m_q and gallery m_g via computing the gradient of similarity score for extracted feature vectors with



(a) m_q_Baseline (b) m_g_Baseline (c) m_q_SSBVER (d) m_g_SSBVER

Figure 6. Saliency maps of baseline and SSBVER models for a pair of query-gallery images. Similarity scores generated from baseline and SSBVER are 0.96 and 0.98. Adopted architecture is ResNet50_IBN and images are chosen from VeRi dataset. More results are provided in the supplementary material.

Table 4. Performance comparison between SSBVER and baseline models with ResNet_IBN backbone architecture on CityFlow multi-camera vehicle tracking task.

Model	IDF1	IDP	IDR
Baseline	0.6903	0.6852	0.6956
SSBVER	**0.7288**	**0.7208**	**0.7369**

respect to the input images:

$$m_q = \nabla_{I_q} f_t(I_q).f_t(I_g), \quad m_g = \nabla_{I_g} f_t(I_q).f_t(I_g) \tag{6}$$

where . is the dot product and I_q, I_g represent query and gallery images respectively. Here we did not use the Grad-CAM [38] method as it computes gradient maps in much lower resolutions which results in blob-like salient regions when up-sampled to the original image size which might not be descriptive enough. In Fig. 6 input saliency maps are depicted for baseline and SSBVER models for a pair of test images selected from VeRi dataset. It is seen that baseline model mainly focused on vehicle's front bumper and slightly attended its side as it is a shared portion between the two images. However, SSBVER asserts more attention on discriminative cues such as the white napkin box on the dashboard and the car's side skirt. Also similarity scores obtained for the pair via baseline and SSBVER models are 0.96 and 0.98 respectively. This examples qualitatively explains the intra-class compactness shown in Fig. 4. While the baseline is a strong re-id model, incorporation of self-supervision helps to learn more discriminative and locally-distinguishable information without employing any explicit and computation-demanding attention mechanism.

4.5. A Use Case: Multi-Camera Vehicle Tracking

We study the impact of SSBVER on multi-camera vehicle tracking task where re-id features and spatio-temporal cues are used to associate vehicle identities across multiple cameras. For this we used NVIDIA AI City CityFlow [33,34,40] dataset which contains 3.5 hours of traffic videos collected from 46 cameras spanning 16 intersections. The evaluation metric for this benchmark is $IDF1 = (2{*}IDP{*}IDR)/(IDP + IDR)$ [37] that is the harmonic mean of identification precision (IDP) and recall (IDR). We adopted a multi-camera tracking system and only replaced the baseline feature extractor with SSBVER. Table 4 highlights that SSBVER improves the overall performance without imposing any additional costs.

4.6. Comparison with the state-of-the-art

We compare SSBVER with ResNet50_IBN architecture against recent works on vehicle re-id in terms of evaluation and efficiency metrics. The reason we chose ResNet50_IBN for SSBVER compared to ResNet50, ViT, SWIN, and ConvNext is that it maintains a comparatively high level of accuracy on all benchmarks while keeping the inference speed and resource utilization low. From Table 5 HRCN appears to be the superior model in terms of evaluation metrics; however, it takes 10.84 ms to compute 3584-dimensional embeddings which is more than twice the time required by SSBVER to obtain embeddings of size 2048. In addition, SSBVER performs better in terms of CMC for VeRiWild which is the largest multi-view benchmark. Therefore, SSBVER is a simple and light weight approach that does not rely on additional annotations, and has a performance that is comparable to HRCN and higher than other computationally expensive alternatives such as TransReID and PVEN. In Fig. 7 we plotted accuracy versus efficiency in terms of inference time, number of model's parameters, memory usage and embedding size. SSBVER achieves highest accuracy-efficiency trade-off among state-of-the-art models. We would like to emphasize that computational efficiency is one of the key contributions of our work which is often overlooked in recent works as we had to measure them by re-implementing or adopting the corresponding works. Lastly, we point out that reported numbers for Time should be considered for relative comparison. These can be further reduced depending on the hardware and adopting inference time optimizations libraries.

5. Ablation Studies

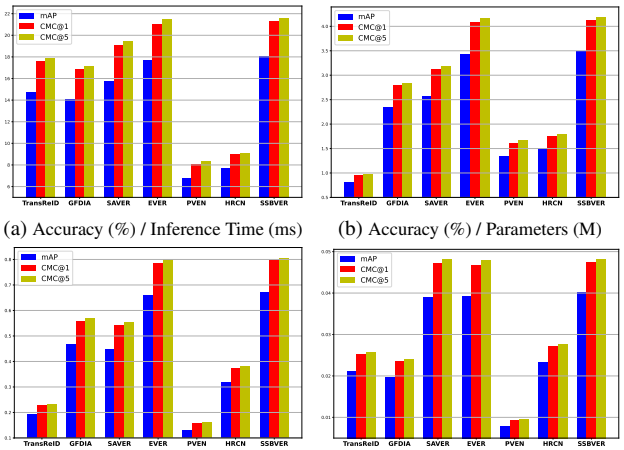
First we replace the cross entropy loss in Eq. 4 between the outputs of student and teacher branches with the L_2 norm of their difference which is often referred to as Root Mean Squared Error (RMSE) loss. Therefore, the self-supervised loss $\mathcal{L}_s$ is calculated by:

$$\mathcal{L}_s = \sum_{I \in V_g(I)} \sum_{I' \in V_g(I) \cup V_i(I),\, I' \neq I} ||g_s(f_s(I')) - g_t(f_t(I))||_2 \tag{7}$$

Table 6 presents the result of this comparison for VeRi dataset and ResNet50_IBN architecture. It is seen that minimizing the cross entropy between the predictions of student and teacher models performs better compared to directly minimizing their RMSE. This observation is consistent with the findings of authors of [4] where they attempt to learn

Table 5. Comparison with recent state-of-the-arts methods. Note that * denotes the number is not reported in the original paper and is computed by implementing the corresponding work or adopting the official repository upon availability. NVIDIA RTX 2080 GPU card for time measurements.

Method	Evaluation Metrics									Efficiency Metrics			
	VeRi			VehicleID (L)			VeRiWild (S)						
	mAP	CMC		mAP	CMC		mAP	CMC		Params (M)	Dims	Time (ms/image)	Memory (MB)
		@1	@5		@1	@5		@1	@5				
TransReID [15]	81.4	96.8	98.4	84.9	78.7	93.2	81.2*	92.3*	98.0*	101*	3840*	5.51*	423*
GFDIA [23]	81.0	96.7	98.6	-	80.0	93.7	-	-	-	34.7*	4096*	5.74*	173*
SAVER [19]	79.6	96.4	98.6	82.9	75.3	88.3	80.9	94.5	98.1	31*	2048*	5.06*	178*
EVER [36]	80.4	95.8	97.9	84.3	78.4	92.3	80.7	93.7	97.8	23.5*	2048*	4.55*	122*
PVEN [32]	79.5	95.6	98.4	-	77.8	92.0	79.8	94.0	98.0	59.2*	10240*	11.79*	603*
HRCN [46]	**83.1**	**97.3**	**98.9**	**85.9***	**79.5***	**94.8***	**85.2**	94.0	98.3*	55.4*	3584*	10.84*	260*
SSBVER	82.1	97.1	98.4	84.8	78.9	92.6	82.6	**95.1**	**98.5**	**23.5**	**2048**	**4.55**	**122**



(a) Accuracy (%) / Inference Time (ms) (b) Accuracy (%) / Parameters (M)

(c) Accuracy (%) / Memory Usage (MB) (d) Accuracy (%) / Embedding Size

Figure 7. Accuracy versus efficiency comparison of SSBVER and current state-of-the-art models for VeRi Dataset.

Table 6. Comparison of different objective choices for $\mathcal{L}_s$. VeRi dataset and ResNet50_IBN model are used.

Objective Function	Evaluation Metrics		
	mAP(%)	CMC@1(%)	CMC@5(%)
RMSE	80.80	96.54	98.39
Cross Entropy	**82.11**	**97.08**	**98.45**

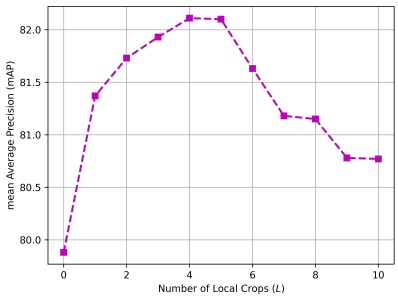


Figure 8. Impact of the number of local crops L on mAP. ResNet50_IBN architecture and VeRi dataset are used.

purely self-supervised features from the scratch. Additionally, we would like to understand how the performance of the SSBVER varies with respect to the number of local crops L that student model observes for each sample during training. Fig. 8 demonstrates the re-id performance in terms of mAP as a function of L for the case of VeRi dataset and ResNet50_IBN architecture. In Fig. 8 there is a significant jump from baseline ($L = 0$) to apply self-supervision with only $L = 1$ local crop. The maximum mAP occurs for $L = 4$ crops. The reduction in performance for higher L can be attributed to observing more regions of a target image by student network and putting less effort to match the output of teacher. This leads to learning less discriminative representations. Nevertheless, we should note that performance is considerably higher than the baseline model.

6. Conclusions

This work presents a novel hybrid training framework that engages self-supervision through self-training and knowledge distillation. This yields performance improvements consistently on public benchmarks irrespective to the choice of DNN architecture. In contrast to alternatives, our approach only requires a forward pass of a single DNN without extra computational overhead. As vehicle re-id technology becomes more mature, its large-scale deployment seems to be reachable more than ever. As a result, more emphasis should be directed towards efficiency metrics such as throughput, and the memory footprint which are often overlooked in the community. The importance of such metrics becomes evident in real-time and at scale applications where the amount of data to be processed and managed is overwhelming. SSBVER obtains performance on par to state-of-the-art in spite of being computationally far less demanding. Therefore, we advocate for efficiency metrics and hope this work motivates further research to develop efficient and lightweight frameworks suited for large-scale applications.

7. Acknowledgment

This work is partially supported by the ONR MURI grant N00014-20-1-2787.

References

[1] Yan Bai, Yihang Lou, Yongxing Dai, Jun Liu, Ziqian Chen, Ling-Yu Duan, and ISTD Pillar. Disentangled feature learning network for vehicle re-identification. In *IJCAI*, pages 474–480, 2020. 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, pages 9912–9924, 2020. 1, 4

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 4, 7

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 4

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 1

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, pages 21271–21284, 2020. 1, 4

[10] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 1, 2018. 2

[11] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[12] 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. 1, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 3

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 2, 8

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6141, 2019. 1, 2

[19] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*, pages 369–386. Springer, 2020. 2, 3, 8

[20] Pirazh Khorramshahi, Sai Saketh Rambhatla, and Rama Chellappa. Towards accurate visual and natural language-based vehicle retrieval systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2021. 2

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, pages 18661–18673, 2020. 1

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

[23] Ming Li, Xinming Huang, and Ziming 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*, pages 194–204, 2021. 2, 8

[24] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2167–2175, 2016. 2

[25] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos.

In *2016 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2016. 2

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 3

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 2

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[32] 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, 8

[33] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa. The 6th ai city challenge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3346–3355. IEEE Computer Society, June 2022. 7

[34] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E. Lopez, Anuj Sharma, Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. The 5th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021. 7

[35] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 3

[36] Neehar Peri, Pirazh Khorramshahi, Sai Saketh Rambhatla, Vineet Shenoy, Saumya Rawat, Jun-Cheng Chen, and Rama Chellappa. Towards real-time systems for vehicle re-identification, multi-camera tracking, and anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 3, 8

[37] 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. 7

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[40] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8797–8806, June 2019. 2, 7

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 1

[42] 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 *Proceedings of the IEEE international conference on computer vision*, pages 379–387, 2017. 2

[43] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, (2), 2009. 1

[44] 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. 1

[45] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 4

[46] Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 205–214, 2021. 2, 8