The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification

Pirazh Khorramshahi[*1], Neehar Peri[*1], Jun-cheng Chen[2], and
Rama Chellappa[1]

[1] Center for Automation Research, UMIACS, and the Department of Electrical and
Computer Engineering, University of Maryland, College Park
[2] Research Center for Information Technology Innovation, Academia Sinica

Abstract. In recent years, the research community has approached the problem of vehicle re-identification (re-id) with attention-based models, specifically focusing on regions of a vehicle containing discriminative information. These re-id methods rely on expensive key-point labels, part annotations, and additional attributes including vehicle make, model, and color. Given the large number of vehicle re-id datasets with various levels of annotations, strongly-supervised methods are unable to scale across different domains. In this paper, we present Self-supervised Attention for Vehicle Re-identification (SAVER), a novel approach to effectively learn vehicle-specific discriminative features. Through extensive experimentation, we show that SAVER improves upon the state-of-the-art on challenging VeRi, VehicleID, Vehicle-1M and VERI-Wild datasets.

Keywords: Vehicle Re-Identification, Self-Supervised Learning, Variational Auto-Encoder, Deep Representation Learning

1 Introduction

Re-identification (re-id), the task of identifying all images of a specific object ID in a gallery, has been recently revolutionized with the advancement of Deep Convolutional Neural Networks (DCNNs). This revolution is most notable in the area of person re-id. Lou *et al.* [28] recently developed a strong baseline method that supersedes state-of-the-art person re-id methods by a large margin, using an empirically derived "Bag of Tricks" to improve the discriminative capacity of DCNNs. This has created a unique opportunity for the research community to develop innovative yet simple methods to push the boundaries of object re-id.

Specifically, vehicle re-id has great potential in intelligent transportation applications. However, the task of vehicle re-id is particularly challenging since vehicles with different identities can be of the same make, model and color. Moreover, the appearance of a vehicle varies significantly across different viewpoints. Therefore, recent DCNN-based re-id methods focus attention on discriminative regions to improve robustness to orientation and occlusion. To this end, many

---

[*] The first two authors equally contributed to this work.
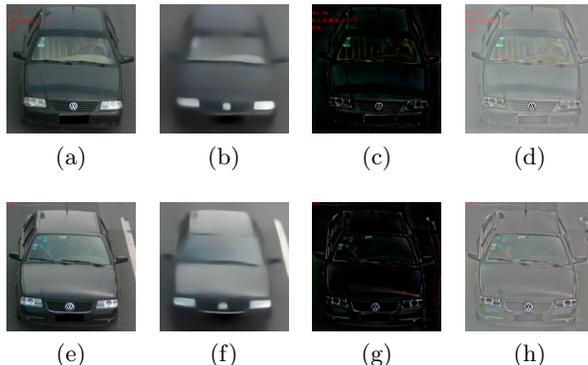
Fig. 1. Vehicle image decomposition into coarse reconstruction and residual images, left-most column (a,e): vehicle image, second column (b,f): coarse reconstruction, third column (c,g): residual, right-most column (d,h): normalized residual (for the sake of visualization). Despite having the same coarse reconstruction, both vehicles have different residuals highlighting key areas, *e.g.*, the windshield stickers, bumper design.

high performing re-id approaches rely on additional annotations for local regions that have been shown to carry identity-dependent information, *i.e.* key-points [41, 16, 17] and parts bounding boxes [11, 46] in addition to the ID of the objects of interest. These extra annotations help DCNNs jointly learn improved global and local representations and significantly boost performance [16, 48] at the cost of increased complexity. Despite providing considerable benefit, gathering costly annotations such as key-point and part locations cannot be scaled to the growing size of vehicle re-id datasets. As manufacturers change the design of their vehicles, the research community has the burdensome task of annotating new vehicle models. In an effort to re-design the vehicle re-id pipeline without the need for expensive annotations, we propose SAVER to automatically highlight salient regions in a vehicle image. These vehicle-specific salient regions carry critical details that are essential for distinguishing two visually similar vehicles. Specifically, we design a Variational Auto-Encoder (VAE) [19] to generate a vehicle image template that is free from manufacturer logos, windshield stickers, wheel patterns, and grill, bumper and head/tail light designs. By obtaining this coarse reconstruction and its pixel-wise difference from the original image, we construct **residual** image. This residual contains crucial details required for re-id, and acts as a pseudo-saliency or pseudo-attention map highlighting discriminative regions in an image. Fig. 1 shows how the residual map highlights valuable fine-grained details needed for re-identification between two visually similar vehicles.

The rest of the paper is organized as follows. In section 2, we briefly review recent works in vehicle re-id. The detailed architecture of each step in the proposed approach is discussed in section 3. Through extensive experimentation in section 4, we show the effectiveness of our approach on multiple challenging vehicle re-id benchmarks [43, 22, 9, 27, 24], obtaining state-of-the-art results. Finally, in section 5 we validate our design choices.

2   Related Works

Learning robust and discriminative vehicle representations that adapt to large viewpoint variations across multiple cameras, illumination and occlusion is essential for re-id. Due to a large volume of literature, we briefly review recent works on vehicle re-identification.

With recent breakthroughs due to deep learning, we can easily learn discriminative embeddings for vehicles by feeding images from large-scale vehicle datasets, such as VehicleID, VeRi, VERI-Wild, Vehicle-1M, PKU VD1&VD2 [43], CompCars [44], and CityFlow [40], to train a DCNN that is later used as the feature extractor for re-id. However, for vehicles of the same make, model, and color, this global deep representation usually fails to discriminate between two similar-looking vehicles. To address this issue, several auxiliary features and strategies are proposed to enhance the learned global appearance representation. Cui *et al.* [4] fuse features from various DCNNs trained with different objectives. Suprem *et al.* [36] propose the use of an ensemble of re-id models for vehicle identity and attributes for robust matching. [41, 23, 46, 11, 16] propose learning enhanced representation by fusing global features with auxiliary local representations learned from prominent vehicle parts and regions, *e.g.,* headlights, mirrors. Furthermore, Peng *et al.* [31] leverage an image-to-image translation model to reduce cross-camera bias for vehicle images from different cameras before learning auxiliary local representation. Zhou *et al.* [50] learn vehicle representation via viewpoint-aware attention. Similarly, [48, 32] leverage attention guided by vehicle attribute classification, *e.g.,* color and vehicle type, to learn attribute-based auxiliary features to enhance the global representation. Metric learning is another popular approach to make representations more discriminative. [47, 2, 3, 21] propose various triplet losses to carefully select hard triplets across different viewpoints and vehicles to learn an improved appearance-robust representation.

Alternatively, to augment training data for more robust training, [45] adopts a graphic engine and [42, 39] use generative adversarial networks (GANs) to synthesize vehicle images with diverse orientations, appearance variations, and other attributes. [25, 26, 34, 38, 14, 29, 15] propose methods for improving the matching performance by also making use of spatio-temporal and multi-modal information, such as visual features, license plates, inter-camera vehicle trajectories, camera locations, and time stamps.

In contrast with prior methods, SAVER benefits from self-supervised attention generation and does not assume any access to extra annotations, attributes, spatio-temporal and multi-modal information.

3   Self-Supervised Attention for Vehicle Re-identification

Our proposed pipeline is composed of two modules, namely, **Self-Supervised Residual Generation, and Deep Feature Extraction**. Fig. 2 presents the proposed end-to-end pipeline. The self-supervised reconstruction network is responsible for creating the overall shape and structure of a vehicle image while
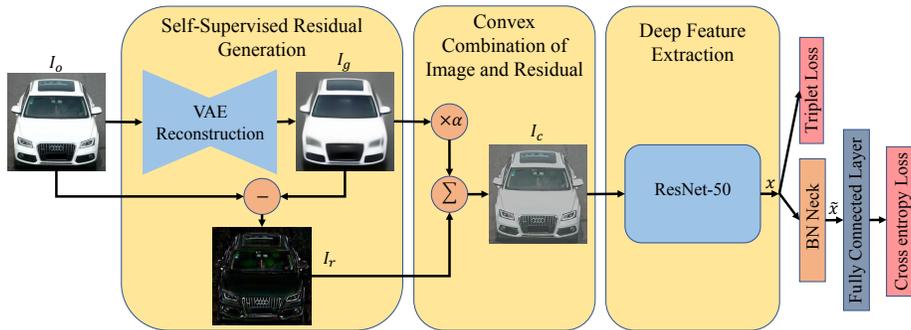
Fig. 2. Proposed SAVER pipeline. The input image is passed through the VAE-based reconstruction module to remove vehicle-specific details. Next, the reconstruction is subtracted from the input image to form the residual image containing vehicle-specific details. Later, the convex combination (with trainable parameter α) of the input and residual is calculated and passed through the re-id backbone for deep feature extraction. The entire pipeline is trained via triplet and cross entropy losses, separated via a batch normalization layer (BN Neck) proposed in [28].

obfuscating discriminative details. This enables us to highlight salient regions and remove background distractors by subtracting the reconstruction from the input image. Next, we feed the convex combination (with trainable parameter α) of the residual and original input images to ResNet-50 [12] model to generate robust discriminative features. To train our deep feature extraction module, we use techniques proposed in "Bag of Tricks" [28] and adapt them for vehicle re-identification, offering a strong baseline.

3.1   Self-Supervised Residual Generation

In order to generate the crude shape and structure of a vehicle while removing small-scale discriminative information, we leverage prior work in image segmentation [1] and generation [19]. Specifically, we construct a novel VAE architecture that down-samples the input image of spatial size $H \times W$ through max-pooling into a latent space of spatial size $\frac{H}{16} \times \frac{W}{16}$. Afterwards, we apply the re-parameterization trick introduced in [19] to the latent features via their mean and covariance. Next, we up-sample the latent feature map as proposed by [30] to prevent checkerboard artifacts. This step generates the reconstructed image of size $H \times W$. Fig. 3 illustrates the proposed self-supervised reconstruction network.

Formally, we pre-train our reconstruction model using the mean squared error (MSE) and Kullback-Leibler (KL) divergence such that

$$\mathcal{L}_{reconstruction} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{KL} \tag{1}$$
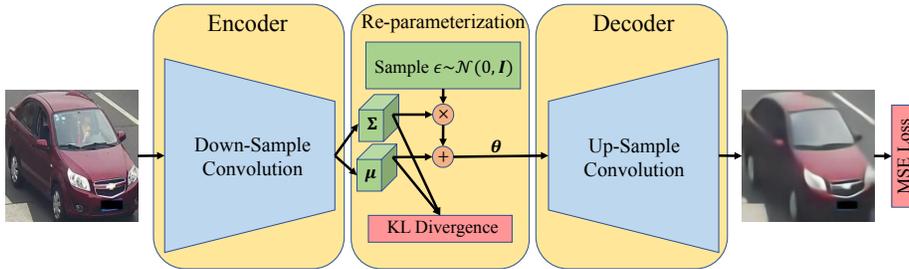
Fig. 3. Self-Supervised image reconstruction required for subsequent residual generation. The input image goes through the convolutional encoder and is mapped to 3-dimensional latent variable. Using the VAE re-parameterization trick, a sample from the standard multivariate Gaussian ϵ is drawn and scaled via mean μ and co-variance Σ of the latent variable. Lastly, θ is up-sampled with a convolutional decoder to generate the input image template with most fine grained details removed.

where

$$\mathcal{L}_{MSE} = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} |I_o(j,k) - I_g(j,k)|^2 \tag{2}$$

and

$$\mathcal{L}_{KL} = \frac{1}{2 \times (\frac{H}{16} \times \frac{W}{16})} \sum_{m=1}^{M} \left[\mu_m^2 + \sigma_m^2 - \log(\sigma_m^2) - 1 \right] \tag{3}$$

In Eq. 1, λ sets the balance between the MSE and KL objective functions. Also, I_o and I_g in Eq. 2 refer to the original and generated images respectively. Finally, in Eq. 3, M is the dimensionality of the latent features $\theta \in \mathbb{R}^M$ with mean $\mu = [\mu_1, \ldots, \mu_M]$ and covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_M^2)$, that are re-parameterized via sampling from standard multivariate Gaussian $\epsilon \sim \mathcal{N}(\mathbf{0}, I_M)$, *i.e.* $\theta = \mu + \Sigma^{\frac{1}{2}} \epsilon$.

We pre-train this model on the large-scale Vehicle Universe dataset, introduced in section 4.2.1, prior to training our end-to-end pipeline, as described in section 4. This pre-training allows the reconstruction model to generalize to vehicle images with a larger variety of make, model, color, orientation, and image quality. Hence, it captures domain invariant features that can later be fined-tuned for a particular dataset. Additionally, pre-training improves the rate of convergence for end-to-end pipeline training. It is important to note that unlike traditional VAE implementations, we use three-dimensional latent feature maps, *i.e.*, channel, height and width dimensions, rather than one-dimensional latent vectors with only channel dimension, for improving the reconstruction quality and preserve more spatial information. Moreover, we scale $\mathcal{L}_{KL}$ when calculating Eq. 1 to improve the reconstruction quality. We further explore the effect of the KL divergence scaling factor λ in section 5. Once the self-supervised image reconstruction network generates the coarse image template I_g, we subtract it from original input to obtain the residual image, *i.e.* $I_r = I_o - I_g$.

3.2   Deep Feature Extraction

Since vehicle images reside on a high-dimensional manifold, we employ a DCNN to project the images onto a lower-dimensional vector space while preserving features that can effectively characterize a unique vehicle identity. To this end, we use a single-branch ResNet-50. To train this model, we use techniques proposed in "Bag of Tricks" [28], which are shown to help a DCNN traverse the optimization landscape using gradient-based optimization methods more effectively. In particular, we observe that the following techniques significantly contribute to the performance of the vehicle re-id baseline model:

1 - **Learning Rate Warm-Up**: [6] has suggested increasing the learning rate linearly in initial epochs of training to obtain improved weight initialization. This significantly contributes to the enhanced performance of our baseline.
2 - **Random Erasing Augmentation (REA)**: To better handle the issue of occlusion, [13] introduced REA with the goal of encouraging a network to learn more robust representations.
3 - **Label Smoothing**: In order to alleviate the issue of over-fitting to the training data, [37] proposed smoothing the ground-truth labels.
4 - **Batch Normalization (BN) Neck**: To effectively apply both classification and triplet losses to the extracted features, a BN layer is proposed by [28]. This also significantly improves vehicle re-id performance.

The ResNet-50 feature extractor model is trained to optimize for triplet and cross entropy classification losses which are calculated as follows:

$$\mathcal{L}_{triplet} = \frac{1}{B} \sum_{i=1}^{B} \sum_{a \in b_i} \left[\gamma + \max_{p \in \mathcal{P}(a)} d(x_a, x_p) - \min_{n \in \mathcal{N}(a)} d(x_a, x_n) \right]_{+} \qquad (4)$$

and

$$\mathcal{L}_{classification} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(W_{c(\hat{x}_i)}^T \tilde{x}_i + b_{c(\hat{x}_i)})}{\sum_{j=1}^{C} \exp(W_j^T \tilde{x}_i + b_j)} \qquad (5)$$

In Eq. 4, B, b_i, a, γ, $\mathcal{P}(a)$ and $\mathcal{N}(a)$ are the total number of batches, i^{th} batch, anchor sample, distance margin threshold, positive and negative sample sets corresponding to a given anchor respectively. Moreover, x_a, x_p, x_n represent the ResNet-50 extracted features associated with anchor, positive and negative samples. In addition, function $d(.,.)$ calculates the Euclidean distance of the two extracted features. Note that in Eq. 4, we used the batch hard triplet loss [13] to overcome the computational complexity of calculating the distances to all unique triplets of data points. Here we construct batches so that they have exactly K instances of each ID used in a particular batch, *i.e.* B is a multiple of K. In Eq. 5, $\tilde{x}_i$ and $c(\hat{x}_i)$ refer to the extracted feature for the ith image in the training set after passing through the BN Neck layer and its corresponding ground-truth class label respectively. Furthermore, W_j, b_j are the weight vector and bias associated with class j in the final classification layer. N and C represent the total number of samples and classes in the training process respectively.

3.3   End-To-End Training

After pre-training the self-supervised residual generation module, we jointly train the VAE and deep feature extractor. We compute the convex combination of input images and their respective residuals using a learnable parameter α, *i.e.* $I_c = \alpha \times I_o + (1 - \alpha) \times I_r$, allowing the feature extractor network to weight the importance of each input source. Moreover, the end-to-end training helps the entire pipeline adapt the residual generation such that it is suited for the re-id task. In summary, the loss function for end-to-end training is the following:

$$\mathcal{L}_{total} = \mathcal{L}_{triplet} + \mathcal{L}_{classification} + \eta\mathcal{L}_{reconstruction} \tag{6}$$

In Eq. 6, the scaling factor η is empirically set to 100.

4   Experiments

In this section, we first present the different datasets on which we evaluate the proposed approach and describe how vehicle re-identification systems are evaluated in general. Next, we present implementation details for the proposed self-supervised residual generation, deep feature extraction and end-to-end training steps. Finally, we report experimental results of the proposed approach.

4.1   Vehicle Re-Identification Datasets

We evaluate SAVER on six popular vehicle re-id benchmarks, including VeRi, VehicleID, VERI-Wild, Vehicle-1M and PKU VD1&VD2. Table 1 presents the statistics of these datasets in terms of the number of unique identities, images and cameras. Additionally, we highlight four additional datasets of unconstrained vehicle images, including CityFlow, CompCars, BoxCars116K [35], and Stanford-Cars [20], used in the pre-training of our self-supervised reconstruction network.

Table 1. Vehicle re-id datasets statistics. ID, IM, Cam refer to number of unique identities, images and cameras respectively. Note that the evaluation set of VehicleID, VERI-Wild, Vehicle-1M, VD1 & VD2 are partitioned into small (S), medium (M) and large (L) splits.

	Split Set	VeRi	VehicleID			VERI-Wild			Vehicle-1M			VD1			VD2		
							Vehicle Re-id Datasets										
Train	ID	576	13164			30671			50000			70591			39619		
	IM	37746	113346			277797			844571			422326			342608		
	Cam	20	-			173			-			-			-		
			S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
Gallery	ID	200	800	1600	2400	3000	5000	10000	1000	2000	3000	18000	131275	141757	12000	70755	79764
	IM	11579	800	1600	2400	38861	64389	128517	1000	2000	3000	104887	602032	1095649	103550	455910	805260
	Cam	19	-	-	-	146	153	161	-	-	-	-	-	-	-	-	-
Query	ID	200	800	1600	2400	3000	5000	10000	1000	2000	3000	2000	2000	2000	2000	2000	2000
	IM	1678	5693	11777	17377	3000	5000	10000	15123	30539	45069	2000	2000	2000	2000	2000	2000
	Cam	19	-	-	-	105	113	126	-	-	-	-	-	-	-	-	-

Re-id systems are commonly evaluated using the Cumulative Match Curve (CMC) and Mean Average Precision (mAP). A fixed gallery set is ranked with respect to the similarity score, e.g., L_2 distance, of its images and a given query image. CMC@K measures the probability of having a vehicle with the same ID as the query within the top K elements of the ranked gallery. It is a common practice to report CMC@1 and CMC@5. Similarly, mAP measures the average precision over all images in a query set.

4.2   Implementation Details

Here we discuss the implementation of both the self-supervised residual generation and deep feature extraction modules. In general, we resize all the images to $(256, 256)$ and normalize them by a mean and standard deviation of 0.5 across RGB channels before passing them through the respective networks. Moreover, similar to [17], we pre-process all images across all the experiments with the Detectron object detector [7] to minimize background noise.

4.2.1   Self-Supervised Residual Generation To pre-train the self-supervised residual generation module, we construct the large-scale Vehicle Universe dataset. We specifically consider vehicles from a variety of distributions to improve the robustness of our model. We utilize data from several sources, including Comp-Cars, StanfordCars, BoxCars116K, CityFlow, PKU VD1&VD2, Vehicle-1M, VehicleID, VeRi and VeRi-Wild. In total, Vehicle Universe has 3706670, 1103404 and 11146 images in the train, test and validation sets respectively.

4.2.2   Deep Feature Extraction As mentioned in section 3.2, we use ResNet-50 for feature extraction. In all of our experiments learning rate starts from $3.5e-5$ and is linearly increased with the slope of $3.1e-5$ in the first ten epochs. Afterwards, it is decayed by a factor of ten every 30th epoch. In total, the end-to-end pipeline is trained for 150 epochs via Adam [18] optimizer. Furthermore, we use an initial value of $\alpha = 0.5$ for convex combination and $\gamma = 0.3$ for the triplet loss in Eq. 4.

4.3   Experimental Evaluation

In this section, we present evaluation results of the global appearance model (baseline) and global appearance model augmented with self-supervised attention (SAVER) on different re-id benchmarks discussed in section 4.1.

4.3.1   VeRi Table 2 reports the evaluation results on VeRi, a popular dataset for vehicle re-id. SAVER improves upon the strong baseline model. Most notably, SAVER gives 1.4% improvement on the mAP metric. We note that α in the convex combination of the input and residual saturates at 0.96, which means the model relies on 96% percent of the original image and 4% of the residual to construct more robust features.

(a)   (b)   (c)   (d)   (e)
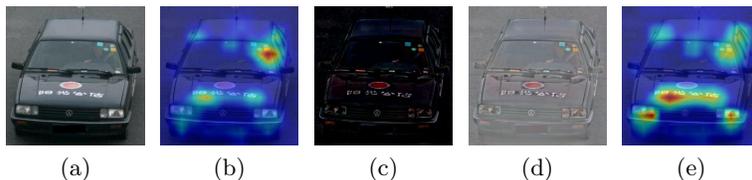
Fig. 4. Grad-CAM visualization of baseline and SAVER; (a) original image, (b) Grad-CAM visualization corresponding to the baseline model, *e.i.*, $\alpha = 1$, (c),(d) are residual and normalized residual maps (for the sake of visualization) obtain via our proposed self-supervised model respectively. (e) is the Grad-CAM visualization of proposed model, *e.i.*, $\alpha = 0.97$ in VehicleID dataset.

4.3.2   VehicleID Table 3 presents results of baseline and SAVER on test sets of varying sizes. Performance improvement of +1.0% in CMC@1 over the baseline model can be observed for all the test splits. To better demonstrate the discriminating capability of the proposed model, we visualize the attention map of both baseline and the proposed SAVER models on an image of this dataset using Gradient Class Activation Mapping (Grad-CAM) [33]. In Figure 4, it is clear that SAVER is able to effectively construct attention on regions containing discriminative information such as headlights, hood and windshield stickers.

4.3.3   VERI-Wild Evaluation results on the VERI-Wild dataset are presented in Table 4. Notably, our proposed residual generation model is improved upon the baseline by +2.0% and +1.0% for mAP and CMC@1 metrics on all evaluation splits respectively. The final alpha value $\alpha = 0.94$ suggests that the residual information contributes more in extracting robust features in this dataset.

4.3.4   Vehicle-1M Table 5 reports the results of baseline and the proposed methods. Similar to VehicleID dataset, Vehicle-1M does not include fixed evaluation sets, therefore we randomly construct the evaluation splits and keep them fixed throughout the experiments. With the value of $\alpha = 0.98$ the proposed self-supervised residual generation module improves upon the baseline model in all metrics across all evaluation sets.

4.3.5   PKU VD1&2 Table 6 highlights the evaluation results on both PKU VD datasets. Similar to most re-id datasets, VD1&2 have S/M/L evaluation sets. However, due to the extreme size of these data splits, as shown in Table 1, we are only able to report numbers on the small evaluation set. The performance of SAVER is comparable to our baseline model. Moreover, the final value of $\alpha = 0.99$ indicates that baseline models is already very strong, and has almost no room for improvement. We can conclude that our performance on these data sets are saturated. Qualitatively, in Figure 5 we show two failure cases of SAVER on these datasets. Note that how extremely similar these images are and it is nearly impossible to differentiate them based on only visual information.

Table 2. Performance Comparison on VeRi

Model	mAP(%)	CMC@1(%)	CMC@5(%)	
Baseline	78.2	95.5	97.9	
SAVER	**79.6**	**96.4**	**98.6**	$\alpha = 0.96$

Table 3. Performance Comparison on VehicleID

Model	CMC@1(%)			CMC@5(%)			
	S	M	L	S	M	L	
Baseline	78.4	76.0	74.1	92.5	89.1	86.4	
SAVER	**79.9**	**77.6**	**75.3**	**95.2**	**91.1**	**88.3**	$\alpha = 0.97$

Table 4. Performance Comparison on VERI-Wild

Model	mAP(%)			CMC@1(%)			CMC@5(%)			
	S	M	L	S	M	L	S	M	L	
Baseline	78.5	72.8	65.0	92.9	91.3	88.1	97.3	96.8	95.0	
SAVER	**80.9**	**75.3**	**67.7**	**94.5**	**92.7**	**89.5**	**98.1**	**97.4**	**95.8**	$\alpha = 0.94$

Table 5. Performance Comparison on Vehicle-1M

Model	CMC@1(%)			CMC@5(%)			
	S	M	L	S	M	L	
Baseline	93.6	94.9	91.7	97.9	99.1	98.0	
SAVER	**95.5**	**95.3**	**93.1**	**98.0**	**99.4**	**98.6**	$\alpha = 0.98$

Table 6. Performance Comparison on VD1&VD2

Dataset	Model	mAP(%)	CMC@1(%)	CMC@5(%)	
VD1	Baseline	96.4	96.2	98.9	
	SAVER	**96.7**	**96.5**	**99.1**	$\alpha = 0.99$
VD2	Baseline	**96.8**	**97.9**	99.0	
	SAVER	96.7	97.8	99.0	$\alpha = 0.99$

(a) Query   (b) Top 1   (c) Top 2   (d) Top 3

(e) Query   (f) Top 1   (g) Top 2   (h) Top 3
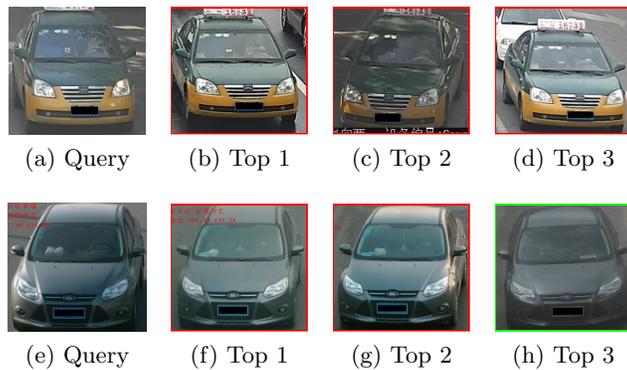
Fig. 5. Examples of SAVER failure on VD1 (sub-figures (a-d)) and VD2 (sub-figures (e-h)). The overall appearance of the query and top ranked images of the gallery are nearly identical. Visual cues such as windshield sticker placement are almost indistinguishable.

4.3.6   State-of-the-Art Comparison In this section, we present the latest state-of-the-art vehicle re-id methods and highlight the performance of the proposed SAVER model. Table 7 reports the state-of-the-art on re-id benchmarks. It can be seen that our proposed model, despite its simplicity, surpasses the most recent state-of-the-art vehicle re-id works without relying on any extra annotations or attributes. For the case of VeRi and VERI-Wild datasets, we also try the method of re-ranking suggested in [49] and achieved considerable mAP scores of 82.0 and 84.4 respectively.

Table 7. Comparison with recent state-of-the-arts methods

Method	Dataset								
	VeRi			VehicleID					
				S		M		L	
	mAP(%)	CMC(%)		CMC(%)		CMC(%)		CMC(%)	
		@1	@5	@1	@5	@1	@5	@1	@5
AAVER [16]	66.35	90.17	94.34	74.69	93.82	68.62	89.95	63.54	85.64
CCA [31]	68.05	91.71	96.90	75.51	91.14	73.60	86.46	70.08	83.20
BS [21]	67.55	90.23	96.42	78.80	**96.17**	73.41	**92.57**	69.33	**89.45**
AGNet [48]	71.59	95.61	96.56	71.15	83.78	69.23	81.41	65.74	78.28
VehicleX [45]	73.26	94.99	97.97	79.81	93.17	76.74	90.34	73.88	88.18
PRND[11]	74.3	94.3	**98.7**	78.4	92.3	75.0	88.3	74.2	86.4
Ours	79.6	**96.4**	98.6	**79.9**	95.2	**77.6**	91.1	**75.3**	88.3
Ours + Re-ranking	**82.0**	**96.9**	97.7						

Method	Dataset																					
	VERI-Wild									Vehicle-1M						VD1			VD2			
	S			M			L			S		M		L			CMC			CMC		
	mAP	CMC		mAP	CMC		mAP	CMC		CMC		CMC		CMC		mAP			mAP			
		@1	@5		@1	@5		@1	@5	@1	@5	@1	@5	@1	@5		@1	@5		@1	@5	
BS[21]	70.54	84.17	95.30	62.83	78.22	93.06	51.63	69.99	88.45	-	-	-	-	-	-	87.48	-	-	84.55	-	-	
AAVER[16]	62.23	75.80	92.70	53.66	68.24	88.88	41.68	58.69	81.59	-	-	-	-	-	-	-	-	-	-	-	-	
TAMR [10]	-	-	-	-	-	-	-	-	-	95.95	99.24	94.27	98.86	92.91	98.30	-	-	-	-	-	-	
Ours	80.9	94.5	98.1	75.3	92.7	97.4	67.7	89.5	95.8	95.5	98.0	95.3	99.4	93.1	98.6	96.7	96.5	99.1	96.7	97.8	99.0	
Ours + Re-ranking	84.4	95.3	97.6																			

5   Ablation Studies

In this section, we design a set of experiments to study the impact of different neural network architectures on the quality of reconstructed images, and also understand the impact of key hyper-parameters. In addition, we are interested in understanding how we can maximally exploit the reconstructed images in deep feature extraction. The experimental results of the reconstruction network are evaluated on the Vehicle Universe dataset, and experiments regarding the deep feature extraction module are evaluated on VeRi and VehicleID datasets.

5.1   Residual Generation Techniques

5.1.1   Effect of Different Reconstruction Architectures Here, we study the reconstruction quality of Auto-Encoder (AE) [1], VAE [19], and GAN [8]

(a) Original   (b) AE   (c) VAE   (d) GAN   (e) BF

(f) Original   (g) AE   (h) VAE   (i) GAN   (j) BF
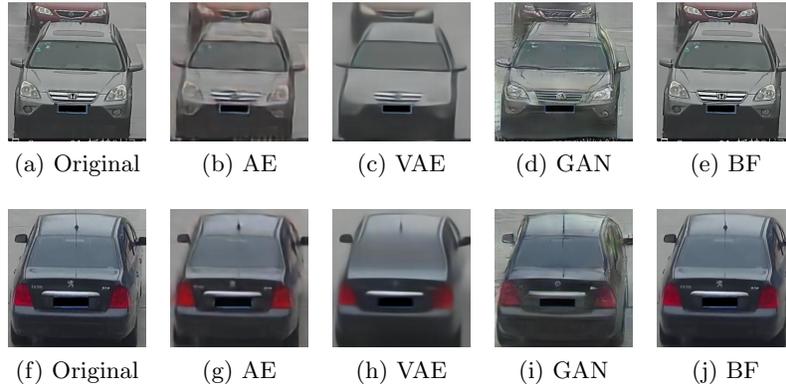
Fig. 6. Different image reconstruction methods.

methods. Moreover, we study the use of Bilateral Filtering (BF) as a baseline for texture smoothing, subsequent residual generation and vehicle re-id. Figure 6 qualitatively illustrates the reconstruction of each method for a given vehicle identity. We notice that both AE and GAN models attempt to recreate fine-grained details, but often introduce additional distortions. Specifically, the GAN model generates new textures, modifies the logo and distorts the overall shape of the vehicle. As a result, GANs produce sharper images with various artifacts that diminish the quality of the residual image required by the re-id network. Also note that although bilateral filtering attempts to smooth images, it is unable to remove the critical details needed in residuals and vehicle re-id. The VAE is able to reconstruct the image by removing minute details and smoothing out textures. As a result, the VAE is able to generate the detailed residual maps needed for our proposed re-id method. Table 8 presents evaluation metrics on VeRi-766 and VehicleID for each of the generative models and bilateral filtering.

Table 8. Performance comparison of different image reconstruction methods

Method	Dataset								
	VeRi			VehicleID					
	mAP(%)	CMC(%)		S		M		L	
				CMC(%)		CMC(%)		CMC(%)	
		@1	@5	@1	@5	@1	@5	@1	@5
AE	79.0	96.0	98.2	79.0	93.9	76.8	90.5	74.9	87.9
VAE	**79.6**	**96.4**	**98.6**	**79.9**	**95.2**	**77.6**	**91.1**	**75.3**	**88.3**
GAN	78.3	95.6	98.1	78.5	93.0	75.6	89.1	73.4	85.7
BF	78.5	95.5	97.6	78.7	76.6	74.5	94.2	90.2	87.4

5.1.2 Effect of Scaling Kullbeck-Leibler Divergence Coefficient λ in Eq. 1 In this experiment, we are particularly interested in the scaling parameter

λ used in training the VAE model. Figure 7 demonstrates how larger values of λ result in a more blurry reconstruction. Intuitively, this parameter offers a natural level for balancing the reconstruction quality of fine-grained discriminative features. As λ approaches 0, our VAE model approximates the reconstruction quality of a traditional Auto-Encoder. Empirically, we found that $\lambda = 1e - 3$ produces higher quality vehicle templates, while removing discriminative information across all datasets.

(a) Original   (b) $\lambda = 1e - 1$ (c) $\lambda = 1e - 2$ (d) $\lambda = 1e - 3$
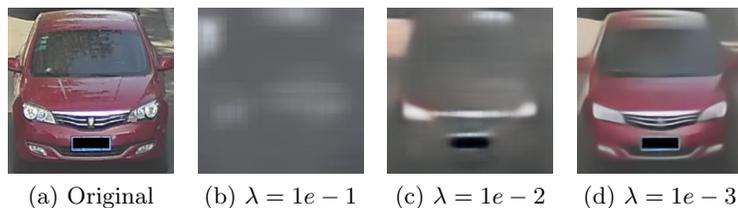
Fig. 7. Effect of scaling KL loss in image generation

5.2   Incorporating Residual Information

To effectively exploit complimentary information provided by the residuals, we design a set of four additional experiments on the VeRi and VehicleID datasets as follows:

A. We only feed the VAE reconstruction I_g as input to the re-id network. The purpose of this experiment is to understand how much critical information can be inferred from the VAE reconstruction.
B. We only feed the residual image I_r into the re-id pipeline. In this experiment we are interested to find out how much identity-dependent information can be extracted from only the residual image.
C. We use the residual maps to excite the actual image of the vehicle through point-wise matrix multiplication.
D. We concatenate the residual image with the original input image. Therefore, in this experiment we feed a six-channel tensor to the feature extraction module.

Table 9 presents the results of experiments A to D and highlights their performance against the baseline and SAVER models. In experiment A, the deep feature extractor is trained using the reconstructed image from the VAE. Intuitively, this method provides the lowest performance since all discriminating details are obfuscated. Interestingly, experiment B, training a deep feature extractor using only residual images, is able to perform nearly as well as our standard baseline. This reaffirms the idea that local information is essential for vehicle re-id. Experiment C performs considerably worse than the baseline model, indicating that point-wise multiplication with the sparse residual removes key information.

Lastly, experiment D performs lower than our baseline. This can be attributed to the ImageNet [5] weight initialization, which is not well suited for six-channel images.

Table 9. Evaluation of different designs of employing residuals

Experiment	VeRi			VehicleID					
				S		M		L	
	mAP(%)	CMC(%)		CMC(%)		CMC(%)		CMC(%)	
		@1	@5	@1	@5	@1	@5	@1	@5
A	67.5	91.4	96.4	64.2	80.6	62.9	76.3	59.4	73.5
B	77.5	94.5	98.2	77.9	92.7	74.7	89.0	73.4	86.2
C	71.4	91.9	96.4	76.3	92.6	73.3	86.8	70.7	83.5
D	75.7	94.8	98.3	78.9	93.1	75.3	89.2	73.3	86.1
Baseline	78.2	95.5	97.9	78.4	92.5	76.0	89.1	74.1	86.4
SAVER	**79.6**	**96.4**	**98.6**	**79.9**	**95.2**	**77.6**	**91.1**	**75.3**	**88.3**

6   Conclusion

In this paper we have shown the benefits of using simple, highly-scalable network architectures and training procedures to generate robust deep features for the task of vehicle re-identification. Our model highlights the importance of attending to discriminative regions without additional annotations, and outperforms existing state-of-the-art methods on benchmark datasets including VeRi, VehicleID, Vehicle-1M, and VeRi-Wild.

Acknowledgement

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)

2. Bai, Y., Lou, Y., Gao, F., Wang, S., Wu, Y., Duan, L.Y.: Group-sensitive triplet embedding for vehicle reidentification. IEEE Transactions on Multimedia **20**(9), 2385–2399 (2018)

3. Chu, R., Sun, Y., Li, Y., Liu, Z., Zhang, C., Wei, Y.: Vehicle re-identification with viewpoint-aware metric learning. In: IEEE International Conference on Computer Vision (ICCV). pp. 8282–8291 (2019)

4. Cui, C., Sang, N., Gao, C., Zou, L.: Vehicle re-identification by fusing multiple deep neural networks. In: International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2017)

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

6. Fan, X., Jiang, W., Luo, H., Fei, M.: Spherereid: Deep hypersphere manifold embedding for person re-identification. Journal of Visual Communication and Image Representation **60**, 51–58 (2019)

7. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron (2018)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)

9. Guo, H., Zhao, C., Liu, Z., Wang, J., Lu, H.: Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 6853–6860. AAAI Press (2018)

10. Guo, H., Zhu, K., Tang, M., Wang, J.: Two-level attention network with multi-grain ranking loss for vehicle re-identification. IEEE Transactions on Image Processing **28**(9), 4328–4338 (2019)

11. He, B., Li, J., Zhao, Y., Tian, Y.: Part-regularized near-duplicate vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3997–4005 (2019)

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. pp. 770–778. IEEE Computer Society (2016)

13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)

14. Hsu, H.M., Huang, T.W., Wang, G., Cai, J., Lei, Z., Hwang, J.N.: Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), AI City Challenge Workshop (2019)

15. Huang, T.W., Cai, J., Yang, H., Hsu, H.M., Hwang, J.N.: Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 434–442 (2019)

16. Khorramshahi, P., Kumar, A., Peri, N., Rambhatla, S.S., Chen, J.C., Chellappa, R.: A dual-path model with adaptive attention for vehicle re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

17. Khorramshahi, P., Peri, N., Kumar, A., Shah, A., Chellappa, R.: Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations (2014)

20. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13) (2013)

21. Kumar, R., Weill, E., Aghdasi, F., Sriram, P.: Vehicle re-identification: an efficient baseline using triplet embedding. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2019)

22. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2167–2175 (2016)

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

24. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo, ICME. pp. 1–6. IEEE Computer Society (2016)

25. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European conference on computer vision (ECCV). pp. 869–884. Springer (2016)

26. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia **20**(3), 645–658 (2017)

27. Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.: Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 3235–3243. Computer Vision Foundation / IEEE (2019)

28. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)

29. Lv, K., Deng, W., Hou, Y., Du, H., Sheng, H., Jiao, J., Zheng, L.: Vehicle reidentification with the location and time stamp. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)

30. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016)

31. Peng, J., Jiang, G., Chen, D., Zhao, T., Wang, H., Fu, X.: Eliminating cross-camera bias for vehicle re-identification. arXiv preprint arXiv:1912.10193 (2019)

32. Qian, J., Jiang, W., Luo, H., Yu, H.: Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. arXiv preprint arXiv:1910.05549 (2019)

33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

34. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: IEEE International Conference on Computer Vision (ICCV). pp. 1900–1909 (2017)

35. Sochor, J., Špaňhel, J., Herout, A.: Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. IEEE Transactions on Intelligent Transportation Systems pp. 1–12 (2018)
36. Suprem, A., Lima, R.A., Padilha, B., Ferreira, J.E., Pu, C.: Robust, extensible, and fast: Teamed classifiers for vehicle tracking in multi-camera networks. arXiv preprint arXiv:1912.04423 (2019)
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
38. Tan, X., Wang, Z., Jiang, M., Yang, X., Wang, J., Gao, Y., Su, X., Ye, X., Yuan, Y., He, D., et al.: Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 275–284 (2019)
39. Tang, Z., Naphade, M., Birchfield, S., Tremblay, J., Hodge, W., Kumar, R., Wang, S., Yang, X.: Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: IEEE International Conference on Computer Vision (ICCV). pp. 211–220 (2019)
40. Tang, Z., Naphade, M., Liu, M., Yang, X., Birchfield, S., Wang, S., Kumar, R., Anastasiu, D.C., Hwang, J.: Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 8797–8806. Computer Vision Foundation / IEEE (2019)
41. Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., Wang, X.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 379–387 (2017)
42. Wu, F., Yan, S., Smith, J.S., Zhang, B.: Joint semi-supervised learning and re-ranking for vehicle re-identification. In: International Conference on Pattern Recognition (ICPR). pp. 278–283. IEEE (2018)
43. Yan, K., Tian, Y., Wang, Y., Zeng, W., Huang, T.: Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 562–570 (2017)
44. Yang, L., Luo, P., Loy, C.C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 3973–3981. IEEE Computer Society (2015)
45. Yao, Y., Zheng, L., Yang, X., Naphade, M., Gedeon, T.: Simulating content consistent vehicle datasets with attribute descent. arXiv preprint arXiv:1912.08855 (2019)
46. Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., Shen, C.: Part-guided attention learning for vehicle re-identification. arXiv preprint arXiv:1909.06023 (2019)
47. Zhang, Y., Liu, D., Zha, Z.J.: Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1386–1391. IEEE (2017)
48. Zheng, A., Lin, X., Li, C., He, R., Tang, J.: Attributes guided feature learning for vehicle re-identification (2019)
49. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
50. Zhou, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)