Navigating Diverse Salient Features for Vehicle Re-identification

Wen Qian, Zhiqun He, Chen Chen, Silong Peng

***Abstract*—Mining sufficient discriminative information is vital for effective feature representation in vehicle re-identification. Traditional methods mainly focus on the most salient features and neglect whether the explored discriminative information is sufficient. This paper tackles the above limitation by proposing a novel Salience-Navigated Vehicle Re-identification Network (SVRN) which explores diverse salient features at multi-scale. For mining sufficient salient features, we design SVRN from two aspects: 1) network architecture: we propose a novel salience-navigated vehicle re-identification network, which mines diverse features under a cascaded suppress-and-explore mode. 2) feature space: cross-space constraint enables the diversity from feature space, which restrains the cross-space features by vehicle and image identifications (IDs). Extensive experiments demonstrate our method's effectiveness, and the overall results surpass all previous state-of-the-arts in three widely-used Vehicle ReID benchmarks (VeRi-776, VehicleID, and VERI-WILD), i.e., we achieve an 84.5% mAP on VeRi-776 benchmark that outperforms the second-best method by a large margin (3.5% mAP).**

***Index Terms*—Vehicle Re-identification; Sufficient Salient Feature; Suppress-and-explore Mode; Grid-based Salient Navigation; Cross-space Constraints.**

I. Introduction

VEHICLE Re-identification (ReID) aims to retrieve a specific vehicle-of-interest in images captured by disjoint cameras, which is broadly applied in cross-camera tracking, surveillance systems and intelligent transportation [1]–[7]. The task is challenging for two reasons: 1) the inter-class difference: two images that share the same vehicle ID may have different appearances due to variations in illumination and viewpoints; 2) the intra-class similarity: two images from different vehicles may have similar appearances since they share almost the same color and type.

The inter-class difference and the intra-class similarity pose evident challenges for existing Vehicle ReID methods which can be alleviated by a robust fine-grained feature representation. However, the previous Vehicle ReID methods [1], [2], [4] tend to generate a global feature representation, which only focuses on the most salient features and neglects the diversity of features. Recently, we notice some methods [8]–[14] that focus on the exploration of diverse and fine-grained features, and summarize them as (Fig. 1): the methods based on prior knowledge (predefined methods), e.g., the part-based information and the attribute-based information; the methods based on adaptive online feature learning (self-learning methods), e.g., the attention-based information.

The attributes (colors or types of vehicles) and the local specifications (windows or logos) all need to be predefined by human annotations, and we term the relevant attribute-based methods [8], [9], [15]–[17] and part-based methods [10], [11], [18]–[22] as **predefined methods**. Although these methods have achieved remarkable performance on public benchmarks with the assistance of predefined information, there still exists some problems: 1) the predefined methods tend to focus on fixed regions and neglect the potential salient information outside the predefined range that may be important for distinguishing the difficult samples. 2) the predefined methods rely heavily on detection or segmentation networks [20], [23], [24] for obtaining the predefined information, which is time-consuming and computational.

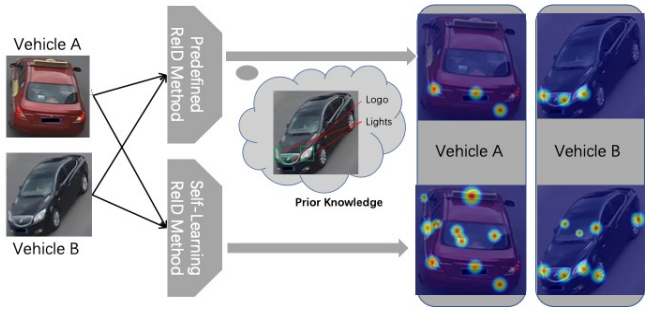

Fig. 1. A comparison between the predefined methods and self-learning methods for Vehicle ReID: given vehicle A and B to both predefined method and self-learning method, and the self-learning method mines more flexible and diverse salient regions than the predefined methods.

As mentioned before, the fixed focus regions in predefined methods restrict the performance of Vehicle ReID, and the extra modules, e.g., detection or segmentation networks, are time costing. So researchers propose some methods that explore the discriminative fine-grained features adaptively [12]–[14], [25], [26] and term them as **self-learning methods**. For example, the attention-based methods [12]–[14], [25], [26] use the attention mechanism to mine the discriminative information adaptively; the residual information [13] that comes from the coarse reconstruct images and the raw images can also be employed as a kind of attention. We conclude the advantages of self-learning methods as 1) more flexibly fine-grained features can be explored 2) no human annotation is needed. However, the aforementioned attention-based methods are supervised by classification or metric learning tasks indirectly, and thus the feature representation tends to overfit the most salient regions with the convergence of the network.

Based on the above observations, we find that both previous predefined and self-learning methods focus on how to mine for differentiated clues, but they neglect that whether the mined discriminative information is sufficient for Vehicle

ReID. Vehicle image information usually contains redundant parts, and only part of the information can make ReID model convergence better. However, redundant information is significant for the ReID model since it can help to distinguish some difficult samples.

For mining sufficient discriminative and fine-grained information adaptively in Vehicle Re-identification, we propose a **salience-navigated vehicle re-identification network (SVRN)**. SVRN employs a cascaded suppress-and-explore mode at multi-stages, which explores the most salient features and then suppresses them for exploring the second salient features at the next stage as shown in Fig. 2. In addition, SVRN uses a grid-based salience ranking layer that directly guides the selection of the most salient regions under the supervision of vehicle identification, making it more suitable for the ReID task than previous methods. Finally, considering that the features from different stages tend to distribute on various subspaces with natural gaps, we propose a cross-space constraint to keep the diversity of them from the feature space. Cross-space constraint contains several sub-components: a vehicle-based constraint that aims to push cross-space features into a unified distribution, an image-based constraint that aims to restrict the features from the same image closer, and a diverse constraint that aims to enhance the diversity of the unified features.

To summarize, our main contributions are in three folds:

- We propose a novel salience-navigated vehicle re-identification network (SVRN), which employs a grid-based critic module for navigating sufficient diverse salient features adaptively under the supervision of the ReID task straightforwardly.
- We propose a novel cross-space constraint that keeps the diversity of features from feature space, and it is consists of a vehicle-based constraint (VCC), an image-based constraint (ICC), and a diverse constraint (DCC).
- Experiments on three Vehicle benchmarks achieve superior performance over previous SOTA methods, which verify the effectiveness of SVRN and the necessity of sufficient diverse salient information for Vehicle ReID.

II. Related Work

We introduce the predefined and self-learning methods that focus on alleviating the inter-class difference and the intra-class similarity challenges. Predefined methods aim to solve this problem with the assistance of local specifications or attributes information [1], [2], [4], [6], [27]; while self-learning methods are usually attention-based which can mine diverse salient information adaptively.

Predefined Methods Recently, people employ some predefined information to improve the performance of Vehicle ReID, which can be classified as part-based information [8], [9], [17], attribute-based information [11], [18], [20], and many other information such as viewpoint information [24], [28], [29] and multimedia information [30].

In part-based methods, there exist two kinds of local specification: the unsupervised coarse specification [8], [9], [15]–[17] and supervised local specification [24], [31]. In unsupervised coarse specification methods, researchers divide the feature map into stripes to capture fine-grained information, which has been proved effective in PCB [15]: Wang et al. [32] use a two-branch architecture to decompose the vehicle feature from stripes at different scales; Qian et al. [25] design a stripe-based branch to decompose the vehicle features and integrate them with the global one for performance boosting. In supervised local specification methods, people employ the detection network or segmentation network for obtaining the specified local parts and use them to enhance the Vehicle ReID models, e.g., Wang et al. [16] propose to use a detection branch for obtaining the predefined local parts and then integrate it with the global ReID modules. Another group of part-based methods employs the key points to emphasize the effectiveness of localized features [12], [33], [34].

Considering that attributes such as types, appearance, and colors are vital for Vehicle ReID, researchers propose the attribute-based Vehicle ReID methods [19], [21], [22], [31]. Li et al. [35] propose to train the attribute classification and ReID tasks jointly for improving the ReID performance. Zeng et al. [36] propose a novel deep network architecture, which works under the guidance of meaningful attributes includes camera views, vehicle types, and colors. Wang et al. [37] propose a novel Attribute-Guided Network (AGNet), which could learn global representation with the abundant attribute features in an end-to-end manner.

Recently, many other approaches are proposed such as viewpoint-based methods [24], [28], [29] or multimedia-based methods [30] are proposed. Meng et al . [24] propose PVEN for capturing the stable discriminative information of vehicles under different views, and then these features are integrated for better feature alignment. Chen et al. [10] propose a dedicated Semantics-guided Part Attention Network (SPAN), which can robustly predict part attention masks for different views of vehicles given only image-level semantic labels during training. People further find that knowledge transfer can also improve the performance of ReID models [38], high discrimination can be ensured by distilling identity-relevant features from the removed information.

Self-learning Methods As mentioned above, the predefined methods achieve encouraging improvement when compared with traditional ReID methods [39], but they are suffered from the fixed information from the predefined regions and can hardly mine diverse feature representations outside these regions dynamically. So researchers propose the self-learning methods for ReID [12]–[14], [25], [26] which aims to mine diverse salient and fine-grained features adaptively. The attention mechanism is employed in the self-learning Vehicle ReID methods to decide the salient degree of different regions. For example, Pirazh et al. [12] present a novel dual-path adaptive attention model for vehicle re-identification, which learns to capture localized discriminative features by focusing attention on the most informative key-points; Apart from using heat maps as the attention, Pirazh et al. [13] also propose to use the residual images from the coarse reconstruction images and the raw images as another kind of attention in Vehicle ReID.

Finally, we conclude that both previous predefined methods and the self-learning methods are focus on how to mine salient

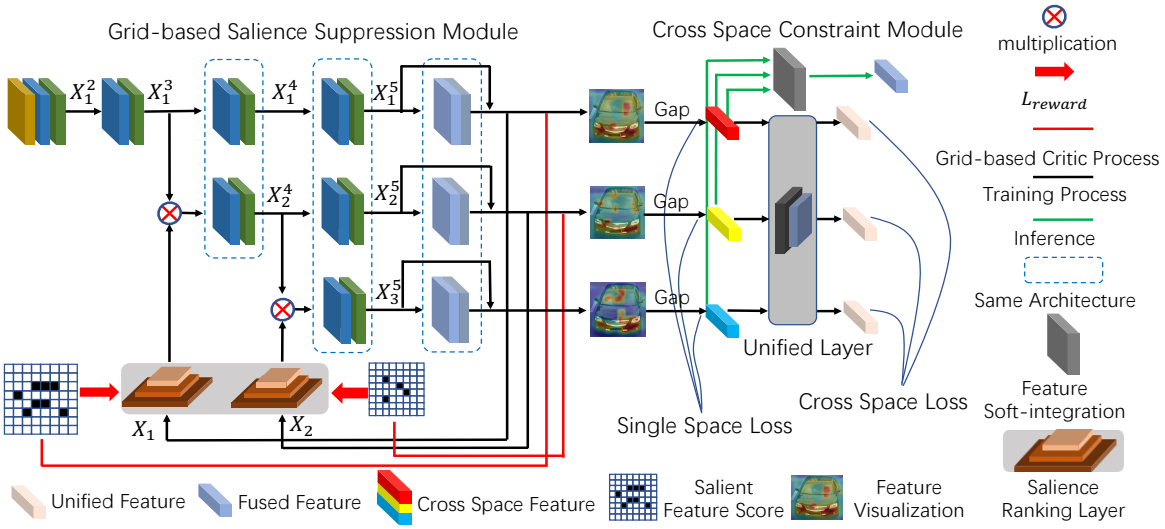

Fig. 2. An overview of SVRN consists of: a) a grid-based salience boosting module which mines diverse salient regions under a suppress-and-explore mode; b) a cross-space constraint that consists of vehicle-based constraint, image-based constraint, and diverse constraint. The salient feature map in the final stage does not require a grid-based evaluation process again.

and fine-grained features, but they tend to neglect whether the explored information is sufficient for the Vehicle ReID task. The SCSN [26] tries to mine more potential features under a suppress-and-explore mode which is similar to our work, so we compare it with our SVRN here: 1) SCSN proposes to explore salient features with indirect supervision, and SVRN employs a grid-based critic module for salient region navigation directly under the supervision of Vehicle ID. The direct supervision in our SVRN makes it more suitable for the ReID task. 2) the CNN-based networks can recovery the masked region from its neighbor parts thanks to the reception field. Instead of just the suppress-and-explore mode in SCSN, we further propose a cross-space constraint to ensure the diversity of features in feature space.

III. Methods

We propose a salience-navigated vehicle re-identification network (SVRN) to explore enough saliency features adaptively. As shown in Fig. 2, SVRN consists of two main subcomponents: a grid-based salience boosting module and a cross-space constraint. The grid-based salience boosting module is illustrated in section III-A, which mines diverse salient features (such as drivers, tabletops, etc.) in a suppress-and-explore cascaded manner at multi-stages with the assistance of a grid-based salience ranking layer. We introduce the cross-space constraint in section III-B, which aims to keep the diversity of salient features from the feature space. Finally, we introduce the total loss function and inference process in section III-C.

A. Grid-based Salience Boosting Module

The grid-based salience boosting module is consists of a pyramid suppression network and a grid-based salience ranking layer. The pyramid suppression network mines non-overlapping salient features from multi-stages, which provides sufficient discriminative information for Vehicle ReID. The salience ranking layer predicts the salience of different regions under the supervision of ReID tasks, which can give a more accurate salience score than previous attention-based methods [12], [13], [26].

Pyramid Suppression Network As aforementioned, the pyramid suppression network employs a multi-stage architecture for exploring diverse salient feature representations, and we take a three-stage pyramid suppression network as an example in this section (as shown in Fig. 2). For convenience, in stage t, we refer to X_t^n/X_t^{n+1} as the input/output feature map of layer n, X_t as the input of salience ranking layer, and S_t as the predicted salience ranking score.

Given a training set $P = [p_1, p_2, ..., p_N]$, where p_i is a vehicle image, N is the number of images. Pyramid suppression network aims to learn a set of feature embedding functions $\phi(\theta, I) = \{\phi_t(\theta_t; I_i)|t = 1, 2, 3\}$, where $\phi_t(\theta_t; I_i)$ denotes the embedding function of stage t, I_i represents the input sample, and the parameters of branch t are collectively denoted as θ_t. In this way, we obtain three feature representations $[F_0^i, F_1^i, F_2^i]$ from each image p_i, and the suppress-and-explore process is employed to keep the diversity of features.

Following [23], [25], we adopt Resnet-50 as the backbone, and the pyramid suppression network is organized in the manner of a pyramid by duplicating blocks in ResNet-50. For a given input p_i, the output feature of backbone in stage t is the basic feature map X_t^5. Then a spatial information mining module that consists of the channel-wise pooling operation and several convolutional layers is further employed to get the spatial feature X_t. After getting the spatial feature X_t, we feed it as the input of the salience ranking layer. Moreover, we feed the spatial feature X_t into a block for classifying the samples into corresponding classes during the training process, which consists of a reduction layer, a global average pooling layer, a batch normalization layer, and a fully connected layer (classified layer) Finally, the pyramid suppression network

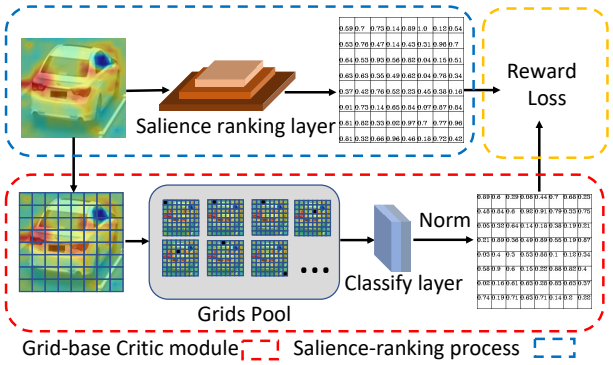

Fig. 3. The overview of grid-based salience ranking layer.

outputs a vehicle representation F_t in stage t, and these representations focus on different salient regions.

$$F_t = \phi_t(\theta_t; I_t), \tag{1}$$

where I_0 is p_i, I_1 is the output feature of stage1, and I_2 is the output feature of stage2.

Grid-based Salience Ranking Layer Instead of previous attention-based methods [12]–[14], [26] which generate attention with indirect supervision, we propose the grid-based salience ranking layer to generate the salient ranking score (attention) under the supervision of the Vehicle ID. A generation process of the salience ranking score S_t is shown in Fig. 2 and Fig. 3, which navigates the most salient regions. We decompose the training process of the grid-based salience ranking layer into three parts: the prediction of salience ranking score, the generation of a grid-based salient map, and the computation of reward loss.

Firstly, given the feature $X_t \in R^{2048\times16\times16}$ from the pyramid suppression network, the salience ranking layer uses an average pooling layer to downsample the feature X_t into $R^{2048\times8\times8}$. Then, a 1×1 convolution block is employed to reduce the channel numbers and explore the salience score of corresponding regions. Consequently, we term the output of the salience ranking layer as salience ranking score $S_t \in R^{8\times8}$, and the above process can be seen in the blue part of Fig. 3. Each score $S_t(i,j), i,j \in [1,8]$ represents the salience degree of relative regions, which can be reflected to a $R^{32\times32}$ region in the original image.

Secondly, for integrating the training process of the salience ranking layer and Vehicle ReID task better, we propose a grid-based critic module as shown in the red part of Fig. 3. The grid-based critic module generates the grid-based salient score map R_{score}, which can be used as the label for the training of the salience ranking layer. After obtaining the same input X_t as the salience ranking layer, we divide it into 8×8 feature grids ($R^{32\times2048\times2\times2}$ for each one). For the feature map X_t, we mask one grid at position (i,j) each time, and the setting of masked feature maps is termed as grids pool as shown in Fig. 3. Each masked feature map $Xm_t(i,j)$ in grids pool will be fed into the classification layer of stage t, and we can get a cross-entropy loss (ReID supervision) $C_s(i,j)$ for each masked feature map:

$$C_s(i,j) = \sum_{k=1}^{N} -q_k(i,j)log(p_k(i,j)) \begin{cases} q_k(i,j) = 0 & y \neq k \\ q_k(i,j) = 1 & y = k, \end{cases} \tag{2}$$

where (i,j) are the coordinates of the masked grid, y is the ground-truth Vehicle ID of the masked feature map, p_k is the predicted score of class k, and N represents the number of classes. We use the value of cross-entropy loss $C_s(i,j)$ to represent the salience degree of the masked region. A large $C_s(i,j)$ indicates that the masked region at (i,j) contains information that is important for Vehicle ReID, and if we mask it will lead to evident performance drops. The computation of $C_s(i,j)$ will be repeated 8×8 times parallelly, which doesn't need gradient back-propagation. The final ranking score R_{score} can be represent as:

$$R_{score} = (\frac{C_s - \min C_s}{\max C_s - \min C_s}). \tag{3}$$

Finally, we use the ranking score R_{score} to supervise the training process of salience ranking score S_t, and formulate the salience ranking loss as:

$$L_{reward} = |R_{score} - S_t|^2. \tag{4}$$

The above formula indicates that we train the salience ranking layer under the supervision of the cross-entropy loss, which makes the salience ranking score more suitable for the Vehicle ReID task.

The Procedure of Suppress-and-explore. The combination of salience ranking layer and pyramid suppression network motivates the suppress-and-explore mode in SVRN, which enables SVRN to learn sufficient diverse salient information at multi-scales. After getting the salience ranking score S_t and the hyper-parameter α_t which denotes the suppression percentage in stage t, we generate a salience ranking mask M_t to navigate the most salient regions as shown in Algorithm 1:

Algorithm 1 Salience ranking algorithm

Input1: Salience ranking score S_t;
Input2: Suppression percentage α_t.
Output: Salience ranking mask M_t
Main(args):

Reshape S_t into one-dimension;
Sort the salience score map S_t in a descending order, and get a value list V_t and a rank list R_t;
Get $n = \alpha_t \times len(S_t)$;
Get Suppression threshold α_v^t by finding the top n-th value in V_t;
if $S_t(i,j) > a_v^t$ **then**
    $M_t(i,j) = 0$
else
    $M_t(i,j) = 1$
end if
return $M_t(i,j)$

After getting the salience ranking mask M_t, we obtain the input X_{t+1} of the next stage $t+1$ based on M_t:

$$X_{t+1} = O_t \times Up(M_t), \tag{5}$$

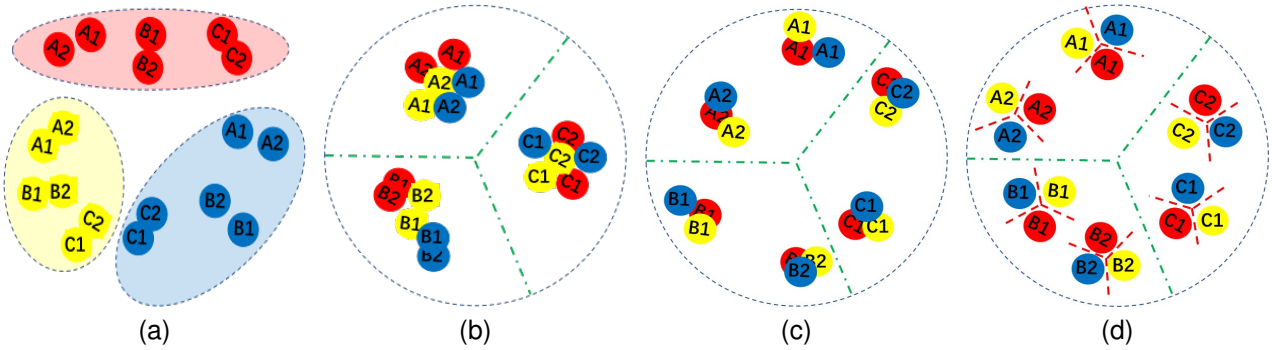

Fig. 4. An illustration of cross-space constraint: a) the distribution of original cross-space features; b) the feature distribution after the vehicle-based constraint; c) the feature distribution after the image-based constraint; d) the feature distribution after the diverse constraint; where A, B, C represent different vehicle IDs; A1, A2 represent different samples of vehicle ID A; the color red, yellow, and blue represent cross-space features from multi-stages.

where O_t is the output feature of stage t and $Up()$ is the up-sampling operation. We need to map M_1 with X_1^3 by up-sampling since they have different sizes. The pyramid-like suppress-and-explore mode masks the most salient regions at the current stage and explores non-overlapping salient regions at the next stage on a smaller scale. Once a region has been chosen, it won't be chosen again during the following stages. Hence, SVRN can mine sufficient diverse salient information from bottom to up.

B. Cross-space Constraint

We argue that the masked region in the aforementioned grid-based salience boosting module can be easily recovered from its neighbor parts, and thus it is harmful to the diversity of features. To make up for the shortcoming of the grid-based salience boosting module, we further propose the cross-space constraint to (CSC) to enhance the diversity of features from feature space. The cross-space constraint is consists of a vehicle-based constraint, an image-based constraint, and a diverse constraint.

Vehicle-based Cross-space Constraint Given an vehicle image set $P = p_1, p_2, ..., p_N$, SVRN outputs three groups of cross-space features F_1^i, F_2^i, and F_3^i, where $i \in [1, N]$ is the index of different images. It's sub-optimal to fuse these cross-space features straightforwardly by coarse adding/concatenating since they are from different stages of SVRN and tend to have gaps in distributions. So vehicle-based constraint solves this by projecting the cross-space features into a unified distribution, and a unified layer $g(\theta_g, F_t^i)$ is employed to project these cross-space features:

$$G_t^i = g(\theta_g, F_t^i) \tag{6}$$

where θ_g is parameters of the unified layer which is composed of several convolution layers, and G_t^i denotes the unified feature of image i in stage t.

We propose a novel vehicle-based cross-space constraint (VCC) for training the unified layer, and VCC aims to push features from the same vehicle closer and features from different vehicles farther apart. The above process is an expansion of triplet loss on cross-space features, and the formula is:

$$L_{vcc} = \max(D(V^+) - D(V^-) + \beta_V, 0), \tag{7}$$

where D is a distance function and β_V is a margin enforced between positive and negative pairs. The positive vehicle pair V^+ and negative vehicle pair V^- in vehicle-based cross-space features G_t^i can be represent as:

$$\begin{aligned} D(V^+) &= \max_{n,m,i,j} (D(G_n^j, G_m^i)) \, s.t. \, ID(G_n^j) = ID(G_m^i), \\ D(V^-) &= \min_{n,m,i,j} (D(G_n^j, G_m^i)) \, s.t. \, ID(G_n^j) \neq ID(G_m^i), \end{aligned} \tag{8}$$

where i, j denote the image identifications (IDs), n, m denote the stages of SVRN, $ID(G_t^i)$ denotes the vehicle identifications (vehicle IDs) of feature G_t^i, G_n^j denotes the anchor feature, and G_m^i denotes any feature except the anchor one. The V^+ aims to find the farthest feature to the anchor one among features which have the same vehicle ID of the anchor, while V^- aims to find the closest feature to the anchor one among those features which have different vehicle ID of the anchor. The vehicle-based cross-space constraint asks the farthest distance between the anchor feature and the feature from the same vehicle ID should smaller than the closest distance between the anchor feature and the feature from the different vehicle ID, and the process is shown in Fig. 4a and Fig. 4b.

Image-based Cross-space Constraint We propose an image-based cross-space constraint (ICC) that enables features of the same image to have similar distributions. We formulate the image-based cross-space constraint as :

$$L_{icc} = \max(D(I^+) - D(I^-) + \beta_I, 0), \tag{9}$$

where β_I is used to ensure the margin between I^+ and I^-. We further formulate the representation of the positive image pair I^+ and negative image pair I^- of a given input p_i as:

$$\begin{aligned} D(I^+) &= \max_{n,m} (D(G_n^j, G_m^i)) \quad s.t. \quad n \neq m \quad i = j \\ D(I^-) &= \min_{n,m,i,j} (D(G_n^j, G_m^i)) \quad s.t. \quad G_n^j \neq G_m^i \end{aligned} \tag{10}$$

where keeps the same settings as the vehicle-based cross-space constraint. The I^+ aims to find the feature farthest to the anchor among those features which have the same image ID of the anchor feature, and I^- aims to find the feature closest to the anchor among all features which have different image ID of the anchor feature. As can be seen in Fig. 4d, the cross-space

features from the same image show a better cluster distribution after employing the image-based cross-space constraint.

Diverse Constraint The features that come from different vehicles and images can be distinguished after the operations of the vehicle-based and image-based cross-space constraints, but some of the cross-space features are still hard to distinguish (as shown in Fig. 4d). So we further propose the diverse constraint, which aims that the cross-space features from different stages of SVRN can have non-overlapping interest regions. For a given image p_i, $G^i = [G_0^i, G_1^i, G_2^i] \in R^{3\times L}$ represents the output cross-space features from SVRN, where L represents the dimension of each feature. We compute the gram of G^i, and ask it to be close to an identity matrix under Frobenius norm:

$$L_{dcc} = ||G^i G^{iT} - I||_F. \tag{11}$$

The joint optimization of cross-space constraint loss is:

$$L_{cross} = L_{vcc} + L_{icc} + \frac{1}{N}\sum_{i=1}^{N} L_{dcc}. \tag{12}$$

C. The Total Loss Function and Inference Process

The total loss function We train the SVRN in an end-to-end manner that takes the grid-based salience boosting module and the cross-space constraint into a joint consideration. Finally, we integrates the ReID loss (Triplet loss L_{tri} and Cross-entropy loss L_{id}), the cross-constraint loss (L_{cross}), and the reward loss (L_{reward}) of the grid-based salience boosting module into a total loss function of SVRN:

$$\begin{aligned} L_{id} &= \sum_{t=1}^{3}\sum -q_i log(p_i) \begin{cases} q_i = 0 & y \neq i \\ q_i = 1 & y = i \end{cases} \\ L_{tri} &= \sum_{t=1}^{3} \max(D(Z^+) - D(Z^-) + \beta_T, 0) \\ L_{total} &= L_{tri} + L_{id} + L_{cross} + L_{reward}, \end{aligned} \tag{13}$$

where t denotes stages of SVRN, y denotes the label for vehicles, Z^+/Z^- denotes the positive/negative feature pair at each stage, β_T denotes the threshold value for distinguishing the positive/negative feature pair Z^+/Z^-.

The inference process For a given figure, we get three cross-space features (the red, yellow, and blue block after the global average pooling (GAP) operation as shown in Fig. 2) from different stages which focus on non-overlapping salient regions. Then these features can be operated by a feature soft-integration module or just adding them for getting the final representation. We choose to add the cross-space features straightforwardly since we aim to provide a fair comparison with other methods and emphasize more on the main contribution of our paper.

IV. Experiments

We detail the implementation and evaluation of SVRN in this section. The datasets and evaluation metrics are introduced in section IV-A; the implementation details are introduced in section IV-B; the comparisons with State-of-the-Arts are introduced in section IV-C; the ablation study and evaluation based on SVRN are introduced in section IV-D, finally, we do further analysis on the visualization and time costing as introduced in section IV-E1.

A. Datasets and Evaluation Metrics

Dataset We evaluate our method on three popular Vehicle ReID benchmarks: VeRi776 [2], VehicleID [40], and VERI-WILD [6]. VeRi776 [2] is a classic Vehicle ReID benchmark, which contains 776 identities collected by 20 cameras in a real-world environment. VehicleID [40] is a large-scale dataset collected by multiple cameras during the daytime on the open road, which contains 26,267 vehicles and 221,763 images in total. VERI-WILD [6] is another large-scale dataset, and it consists of 40,671 vehicles and 416,314 images.

Evaluation Metrics We follow the same official evaluation protocols in [40]–[42], and employ the Mean Average Precision (mAP) and the cumulative matching characteristics at Rank1 (CMC@1) to evaluate the performance of SVRN. Moreover, it should be noticed that the VehicleID [40] benchmark pays more attention to CMC@1.

B. Implementation Details

Experimental Setting All experiments are conducted in PyTorch with 8 NVIDIA Titan Xp GPU. We resize the images to $256 \times 256 \times 3$ and use the random erasing and flipping operations for augmentation. We employ Adam as the optimizer with the weight decay factor of 1e-4 and initialize the learning rate to 1e-4 that decreased by a factor of 0.1 after the 40^{th} and 70^{th} epoch.

Network The three-stage grid-based salience boosting module is modified from ResNet-50 [51]. The first stage in SVRN that starts from block 1 in ResNet-50 suppresses the top 20% salient regions of the output feature map for the second stage, the second stage starts from block 3 in ResNet-50 with the top 10% salient regions suppressed, and the third stage starts from block 4 in ResNet-50. The two hyper-parameter β_V and β_I in cross-space constraint are set as 0.3 and 0.15, while the β_T for triplet loss is also set as 0.3. Each stage in SVRN has a classifier layer for classifying the vehicles in training space by ID loss, and this layer is further applied to calculate in C_{score} for the salience ranking layer.

C. Comparisons with State-of-the-Art Methods

We compare SVRN with a wide range of state-of-the-arts Vehicle ReID methods, including (1) part-based approaches: PGAN [43], PRN [23], PVEN [24], and GLAMOR [44]; (2) attribute-based approaches: AGNet-ASL [37], DJDL [35], XG-6-sub-multi [31], and SAN [25]; (3) attention-based approaches: AAVER [12] and SEVER [13]; (4) other interesting approaches: GSTE [45], VAMI [46], DCDLearn [47], GB+GFB+SLB [49], CAL [48] and TransReID [50].

We show the comparison results in Table I, and get the following conclusions: 1) SVRN achieves state-of-the-art performance on all benchmarks, consistently outperforming the best competitor by up to 3.5% mAP on VeRi-776, 0.4% CMC@1 on VehicleID, and 2.1% mAP on VERI-WILD. 2)

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS. IT INCLUDES MAP AND CMC@1 ON VERI-776; CMC@1 ON THREE TEST SETS OF SMALL, MEDIUM, AND LARGE ON VEHICLEID; MAP ON THREE TEST SETS OF SMALL, MEDIUM, AND LARGE ON VERI-WILD. FOR THE THREE TEST SETS ON VEHICLEID AND VERI-WILD, THEY ARE REPRESENTED BY S, M, AND L RESPECTIVELY. FINALLY, "/" INDICATES MISSING PARTS OF THE EXPERIMENTS.

	Method	VeRi-776		VehicleID			VERI-WILD		
		mAP	CMC@1	CMC@1 (S)	CMC@1 (M)	CMC@1 (L)	mAP (S)	mAP (M)	mAP (L)
Part-based	PGAN [43]	79.3	96.5	77.8	/	/	74.1	/	/
	PRN [23]	74.3	94.3	78.4	75.0	74.2	/	/	/
	PVEN [24]	79.5	95.6	84.7	80.6	77.8	82.5	77.0	69.7
	GLAMOR [44]	80.3	96.5	78.6	/	/	77.2	/	/
Attribute-based	AGNet-ASL [37]	71.59	95.61	71.15	69.23	65.74	/	/	/
	DJDL [35]	/	/	78.6	74.7	72.0	/	/	/
	XG-6-sub-multi [31]	/	/	76.1	73.1	71.2	/	/	/
	SAN [25]	72.5	93.3	79.7	78.4	75.6	/	/	/
Attention-based	AAVER [12]	61.2	89.0	74.7	68.6	63.5	/	/	/
	SEVER [13]	79.6	96.4	79.9	77.6	75.3	83.4	78.7	71.3
Others	GSTE [45]	59.4	/	87.1	82.1	79.8	/	/	/
	VAMI [46]	61.3	89.5	63.1	52.9	47.3	/	/	/
	DCDLearn [47]	70.4	92.8	82.9	78.7	75.9	/	/	/
	CAL [48]	74.3	95.4	82.5	/	/	/	/	/
	GB+GFB+SLB [49]	81.0	96.7	86.8	/	/	/	/	/
	TransReID [50]	80.6	96.9	82.8	/	/	/	/	/
Ours	Baseline	80.8	96.7	82.9	80.1	78.3	83.5	79.0	74.6
	SVRN	**84.5**	**97.2**	**87.5**	**84.6**	**81.8**	**85.5**	**81.5**	**76.3**

TABLE II
ABLATION STUDY OF COMPONENTS IN SVRN, WHERE GSB AND CSC REPRESENT GRID-BASED SALIENCE BOOSTING MODULE AND CROSS-SPACE CONSTRAINT, RESPECTIVELY.

Method	GSB	CSC	mAP	CMC@1
Baseline	×	×	80.7	95.6
Scheme a	✓	×	83.0	96.5
Scheme b	✓	✓	84.5	97.4

TABLE III
EXPERIMENT RESULTS OF DIFFERENT STAGES OF SVRN ON THREE POPULAR BENCHMARKS.

Method	VeRi776		VehicleID	VeRi-WILD
	mAP	CMC@1	CMC@1	mAP
Baseline	80.7	95.6	82.9	84.0
SVRN+2 stages	83.1	96.6	84.4	84.3
SVRN+3 stages	84.5	97.2	87.5	85.5
SVRN+4 stages	79.6	95.6	82.0	82.0

Compared to the part-based and attribute-based methods [23]–[25], [31], SVRN achieves significant improvement, e.g., up to 4.2% mAP on VeRi-776, 2.8% mAP on VehicleID, and 3.0% mAP on VERI-WILD, which validates that the flexible salient feature extraction ability in SVRN can lead to an obvious performance improvement. 3) Compared to the attention-based methods [12], [13], SVRN achieves significant improvement, e.g. up to 4.9% mAP on VeRi-776, 7.6% mAP on VehicleID, and 2.1% mAP on VERI-WILD. The above experiments validate that SVRN performs better than previous attention-based methods, and we ascribe it to the direct supervision from the Vehicle ID in the grid-based ReID specified supervision.

D. Ablation Study and Evaluation

Ablation of Components in SVRN SVRN is consists of a grid-based salience boosting module and a cross-space constraint. We do ablation studies on the performance of these components on the VeRi-776 benchmark, and Table II reports the ablation results. 1) With the employment of grid-based salience boosting module, the model achieves significant improvement than baseline, e.g., up to 2.3% mAP on VeRi-776 benchmark. The suppress-and-explore mode cooperates well with a salient ranking mechanism to mine sufficient non-overlapping salient features beneficial to Vehicle ReID. 2) The proposed cross-space constraint brings a significant improvement, e.g., 1.5% mAP improvement on VeRi-776, which validates that the cross-space constraint is more suitable to the cross-space features than the single-space constraint.

Ablation of Stages in SVRN SVRN uses a multi-stage design for mining sufficient diverse salient information, and we do ablation studies on the influence of the number of stages as shown in Table III. 1) The three-stage SVRN achieves the best performance on all benchmarks, which outperforms the second-best one by 1.4%, 3.1%, and 1.2% mAP on VeRi-776, VehicleID, and VERI-WILD. 2) The two-stage SVRN mines insufficient salient features, so there still exists potential salient regions which can provide discriminative information for Vehicle ReID. 3) In the four-stage SVRN, we observe a significant performance degradation and conclude this phenomenon into several reasons: a) the former stages have explored a lot of information, leaving little ones for the last stage. So it is hard to converge and results in negative migration during back-propagation since the stages have a shared feature extraction module. b) When we integrate the cross-space features during inference, the fourth-stage features with less useful information and greater randomness will lead to a decline in performance;

Ablation of the Suppression Percentage in SVRN The suppress-and-explore mode suppresses the most salient regions

TABLE IV
ABLATION STUDY OF THE SUPPRESSED PERCENTAGES OF EACH STAGE IN SVRN.

Method	Stage 1	Stage 2	CMC@1	mAP
Baseline			95.6	80.7
Scheme a	30%	30%	96.2	82.5
Scheme b	20%	20%	96.7	83.7
Scheme c	10%	10%	96.5	83.1
Scheme d	5%	5%	96.9	82.1
Scheme e	20%	10%	97.4	84.5

TABLE V
ABLATION STUDY OF SUB-COMPONENTS IN CROSS-SPACE CONSTRAINT: CSC REPRESENTS CROSS-SPACE CONSTRAINT WHICH CONSISTS OF VEHICLE-BASED CONSTRAINT (VCC), IMAGE-BASED CONSTRAINT (ICC) AND DIVERSE CONSTRAINT (DCC).

Method	VCC	ICC	DC	mAP	CMC@1
Baseline	×	×	×	80.7	95.6
SVRN w/o CSC	×	×	×	83.0	96.5
SVRN+VCC	✓	×	×	83.8	96.8
SVRN+ICC	×	✓	×	82.1	96.7
SVRN+DCC	×	×	✓	83.5	96.9
SVRN+CSC	✓	✓	✓	84.5	97.4

TABLE VI
ABLATION STUDY OF HYPER-PARAMETERS β_V AND β_I IN CROSS-SPACE CONSTRAINT.

Method	β_V	β_I	mAP	CMC@1
SVRN+VCC	0.15	/	83.3	96.5
SVRN+VCC	0.30	/	83.8	96.8
SVRN+VCC	0.45	/	82.7	96.1
SVRN+CSC	0.30	0.30	83.2	96.9
SVRN+CSC	0.30	0.15	84.5	97.4
SVRN+CSC	0.30	0.10	83.8	97.2

in the current stage and transmits the masked feature to the next one for exploring sufficient non-overlapping salient features. We experiment on the percentage of suppressed regions in each stage in Table IV and get the following conclusions: 1) when keeping the same suppression percentage in stages, we compare different θ_t which ranges from 30%, 20%, 10%, and 5%, and we find that suppressing the top 20% salient features achieves the best performance. A higher θ_t will lead to little useful information left for the later stages, while a smaller θ_t can't mine sufficient salient ones. 2) we also experiment with an attenuated factor on the suppressed percentage for broadcasting more information to the later stages. Scheme e in Table IV denotes the attenuated suppression method that outperforms Scheme b by 0.8% mAP, which shows that the attenuated suppress factor can leave more information for the last stage and achieve better performance.

Ablation of Sub-components in Cross-space Constraint The cross-space constraint is consists of three components: the vehicle-based constraint (VCC), the image-based constraint (ICC) and the diverse constraint (DCC), and we do ablation study on them as shown in Table V (VeRi-776). We find that SVRN+VCC and SVRN+DCC outperform SVRN w/o CSC by 0.8%, 0.5% in mAP, and it demonstrates that the single

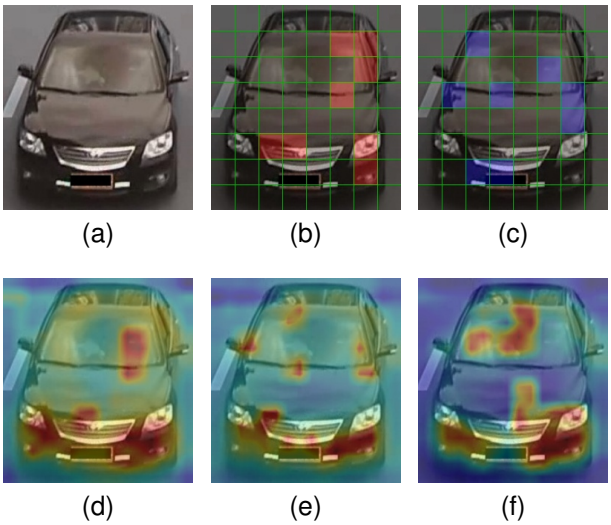

Fig. 5. The visualization of salience-navigated vehicle re-identification network (three stages): a) the original input image; b) the most salient regions in stage 1; c) the most salient regions in stage 2; d) the feature visualization in stage 1; e) the feature visualization in stage 2; f) the feature visualization in stage 3.

VCC or DCC is beneficial to the performance of Vehicle ReID. In addition, CSC integrated these sub-constraints and produced significant improvements, such as a 1.5Nevertheless, we notice that SVRN+ICC results in a 0.9% decline in mAP and conclude that the cross-space features need to be unified before conducting image-based constraints.

Ablation of Hyper-parameters in Cross-space Constraint We further do experiments to study the influence on different choices of hyper-parameters β_V and β_I as shown in Table VI. Firstly, we experiment on the choice of β_V in VCC, e.g., SVRN+VCC achieves an mAP of 83.3%, 83.8%, and 82.7% with β_V of 0.15, 0.30 and 0.45 respectively. The experiments show that the model achieves the best result when $\beta_V = 0.3$. We find a similar phenomenon in the choice process of β_T in regular triplet loss since that VCC is an extension of triplet loss from single-space features into cross-space features. Secondly, SVRN+CSC reaches a mAP of 83.2%, 84.5%, and 83.8% with β_I of 0.3, 0.15 and 0.1 respectively. We conclude that β_I should be smaller than β_V since the image IDs should have a closer distribution than vehicle IDs.

E. Further Analysis

1) Visualization of Salient Information in SVRN: To better show how the suppress-and-explore mechanism works in SVRN, we visualize the intermediate features from different stages as shown in Fig. 5. Given an image (as shown in Fig. 5a), we first explore the most salient regions in the first stage, e.g., Fig. 5b represents the top 10Then we suppress the marked regions in Fig. 5b and aim to explore more diverse salient information in the next stage. We show a similar visualization in the second stage, e.g., Fig. 5c represents the most salient regions explored in the second stage, Fig. 5e visualize the feature map of the second stage. Finally, Fig. 5f visualizes the feature map in the third stage. We can find

TABLE VII
THE TRAINING TIME AND INFERENCE TIME OF BASELINE AND SVRN.

Method	Training Time (ms/frame)	Inference Time (ms/frame)
Baseline	40	23
SVRN	65	37

that the stages focus on diverse salient information which is complementary (as shown in Fig. 5d, Fig. 5e, and Fig. 5f), and these visualizations show the effectiveness of our methods.

2) Time Analysis of SVRN: As can be seen in Fig. 2, SVRN employs some extra modules, e.g., the grid-based salience suppression module and the cross-space constraint, which are time-costing. We compare the training and inference time of baseline and SVRN as shown in Table VII. The baseline costs 40ms/step during training which has only one stage and no extra component. Although the training process of the grid-based salience suppression module needs to calculate a classification loss value $C_s(i,j)$ for each masked feature, SVRN computes the $C_s(i,j)$ parallelly and needs no back-propagation. So we find that the SVRN with the grid-based salience suppression module and the cross-space constraint cost 65ms/step during training, which cost more time from the cascaded architecture but is still acceptable. A similar phenomenon is shown in the inference time, with baseline and SVRN taking 23 ms and 37 ms per frame, respectively.

V. CONCLUSION

In this paper, we propose a salience-navigated vehicle re-identification network (SVRN) to explore sufficient diverse salient features from both network structure and feature space. Firstly, SVRN adopts a pyramid suppression network, which can mine more discriminative features adaptively for the following reasons: 1) the multi-scale suppress-and-explore mode that imitates the human's attention mechanism from bottom-to-up; 2) the direct supervision from the ReID loss in the grid-based critic module. Secondly, we further proposed a cross-space constraint to ensure the diversity of features in feature space, which consists of vehicle-based constraint, image-based constraint, and diverse constraint. Extensive experiments achieve superior performance on three main benchmarks, e.g., VeRi-776, VehicleID, and VERI-WILD, which demonstrate the effectiveness of SVRN.

REFERENCES

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

[2] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[3] X. Liu, L. Li, S. Wang, Z. Zha, D. Meng, and Q. Huang, "Adaptive reconstruction network for weakly supervised referring expression grounding," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2611–2620. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00270

[4] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[5] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multim.*, vol. 20, no. 3, pp. 645–658, 2018. [Online]. Available: https://doi.org/10.1109/TMM.2017.2751966

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

[7] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "Skeletonnet: A hybrid network with a skeleton-embedding process for multi-view image representation learning," *IEEE Trans. Multim.*, vol. 21, no. 11, pp. 2916–2929, 2019. [Online]. Available: https://doi.org/10.1109/TMM.2019.2912735

[8] H. Chen, B. Lagadec, and F. Bremond, "Partition and reunion: A two-branch neural network for vehicle re-identification." in *CVPR Workshops*, 2019, pp. 184–192.

[9] Y. Chen, L. Jing, E. Vahdani, L. Zhang, M. He, and Y. Tian, "Multi-camera vehicle tracking and re-identification on ai city challenge 2019." in *CVPR Workshops*, vol. 2, 2019.

[10] T. Chen, C. Liu, C. Wu, and S. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12347. Springer, 2020, pp. 330–346. [Online]. Available: https://doi.org/10.1007/978-3-030-58536-5_20

[11] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *European Conference on Computer Vision*. Springer, 2014, pp. 134–146.

[12] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6132–6141.

[13] P. Khorramshahi, N. Peri, J. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12359. Springer, 2020, pp. 369–386. [Online]. Available: https://doi.org/10.1007/978-3-030-58568-6_22

[14] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2285–2294.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[17] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 410–420, 2019.

[18] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[19] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[20] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3739–3747.

[21] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognition*, vol. 75, pp. 77–89, 2018.

[22] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, and J. Lai, "Adversarial attribute-image person re-identification," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 2018.

[23] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3997–4005.

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

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

[26] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3300–3310.

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

[28] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8282–8291.

[29] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 379–387.

[30] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, "Learning longterm representations for person re-identification using radio signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 699–10 709.

[31] Y. Zhao, C. Shen, H. Wang, and S. Chen, "Structural analysis of attributes for vehicle re-identification and retrieval," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 2, pp. 723–734, 2019.

[32] H. Wang, J. Peng, G. Jiang, F. Xu, and X. Fu, "Discriminative feature and dictionary learning with part-aware model for vehicle re-identification," *Neurocomputing*, vol. 438, pp. 55–62, 2021.

[33] A. Kanaci, M. Li, S. Gong, and G. Rajamanoharan, "Multi-task mutual learning for vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 62–70.

[34] P. Khorramshahi, N. Peri, A. Kumar, A. Shah, and R. Chellappa, "Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding." in *CVPR Workshops*, 2019, pp. 239–246.

[35] Y. Li, Y. Li, H. Yan, and J. Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 395–399.

[36] A. Zheng, X. Lin, C. Li, R. He, and J. Tang, "Attributes guided feature learning for vehicle re-identification," *arXiv preprint arXiv:1905.08997*, 2019.

[37] H. Wang, J. Peng, D. Chen, G. Jiang, T. Zhao, and X. Fu, "Attribute-guided feature learning network for vehicle re-identification," *arXiv preprint arXiv:2001.03872*, 2020.

[38] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3143–3152.

[39] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

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

[41] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2909–2918.

[42] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.

[43] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle re-identification," *CoRR*, vol. abs/1909.06023, 2019. [Online]. Available: http://arxiv.org/abs/1909.06023

[44] A. Suprem and C. Pu, "Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention," *arXiv preprint arXiv:2002.02256*, 2020.

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

[46] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6489–6498.

[47] R. Zhu, J. Fang, H. Xu, H. Yu, and J. Xue, "Dcdlearn: Multi-order deep cross-distance learning for vehicle re-identification," *arXiv preprint arXiv:2003.11315*, 2020.

[48] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1025–1034.

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

[50] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90