Two-Stage Progressive Learning for Vehicle Re-Identification in Variable Illumination Conditions

Zhihe Wu, Zhi Jin, Xiying Li

Abstract: Vehicle matching in variable illumination environments can be challenging due to the heavy dependence of vehicle appearance on lighting conditions. To address this issue, we propose a two-stage progressive learning (TSPL) framework. In the first stage, illumination-aware metric learning is enforced using a two-branch network via two illumination-specific feature spaces, used to explicitly model differences in lighting. In the second stage, discriminative feature learning is introduced to extract distinguishing features from a given vehicle. This process consists of a local feature extraction attention module, a local constraint, and a balanced sampling strategy. During the metric learning phase, the model expresses the union of local features, extracted from the attention module, with illumination-specific global features to form joint vehicle features. As part of the study, we construct a large-scale dataset, termed VERI-DAN (vehicle re-identification across day and night), to address the current lack of vehicle datasets exhibiting variable lighting conditions. This set is composed of 200,004 images from 16,654 vehicles, collected in various natural illumination environments. Validation experiments conducted with the VERI-DAN and Vehicle-1M datasets demonstrated that our proposed methodology effectively improved vehicle re-identification Rank-1 accuracy.

Keywords: vehicle re-identification; dataset; illumination aware; detail aware; discriminative feature learning

1. Introduction

Vehicle re-identification (re-ID) aims to match a target vehicle across multiple non-overlapping surveillance cameras with varying viewpoints, illumination, and resolution. The proliferation of surveillance cameras in urban areas has led to a significant increase in the demand for vision-based re-ID techniques, which could facilitate the management of smart cities [1–3]. While the development of deep learning and existing annotated datasets have greatly facilitated vehicle re-ID research, vision-based vehicle re-ID still suffers from low resolution, blurred motion, and extreme weather conditions, such as fog, rain, and snow. Meanwhile, variations in viewpoint, illumination, and background can pose significant challenges for vehicle re-ID. Specifically, illumination presents a major challenge, as vehicle appearance can depend heavily on subtle changes in light intensity. Existing models have been trained primarily using datasets that exhibit limited lighting variability, while also ignoring the obstruction of notable visual cues caused by dramatic fluctuations in illumination. Furthermore, the conventional approach of collecting images in a single-feature space also underestimates the challenges posed by inconsistent lighting. As shown in Figure 1, two vehicles of the same model in similar lighting conditions (Figure 1a) may appear to be more similar than the same vehicle observed under different lighting (Figure 1b). In addition, lower lighting levels may obstruct visual cues, such as logos, stickers, and body features, as seen in Figure 1c. One possible solution to this problem is to...
enhance images featuring poor illumination. Although low-light enhancement techniques
have been successful in improving visual quality for classification tasks [4–8], extreme
variations in illumination still pose significant challenges. As part of this study, experiments
were conducted using images enhanced by a state-of-the-art (SOTA) method [6]. The results
are summarized in Table 1, which indicates that re-ID accuracy experience a significant drop-off when images undergo enhancement. We suggest that some distinguishing details, which
are essential for differentiating similar vehicles, may have been lost during enhancement.
In addition, images containing unnatural noise may have further reduced re-ID accuracy.

Figure 1. Challenges posed by variations in illumination. (a) Two different vehicles may appear to be
similar under consistent lighting, especially when they are of the same model. (b) In contrast, images
of the same vehicle may appear to be quite different if the illumination varies drastically. (c) Dramatic
changes in lighting may also alter distinctive visual cues, such as inspection marks on windshields,
headlights, and car logos.

Table 1. Re-ID performance with augmented images. Origin denotes the weight offered by the author.
Retrain denotes a weight retrained on VERI-DAN. The best results are given in bold.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Augment</th>
<th>Query</th>
<th>Gallery</th>
<th>Origin</th>
<th>mAP</th>
<th>Retrain</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank-1</td>
<td></td>
<td>Rank-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>88.9</td>
<td>94.4</td>
<td>56.5</td>
<td>88.9</td>
<td>94.4</td>
<td>56.5</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>75.8</td>
<td>87.0</td>
<td>48.7</td>
<td>75.5</td>
<td>85.2</td>
<td>41.4</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>72.8</td>
<td>83.6</td>
<td>47.6</td>
<td>64.8</td>
<td>79.0</td>
<td>34.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>84.0</td>
<td>91.0</td>
<td>52.1</td>
<td>81.1</td>
<td>90.2</td>
<td>47.2</td>
</tr>
</tbody>
</table>

Inspired by the manual process of recognizing vehicles, we propose to address
this problem by using a two-stage progressive learning (TSPL) strategy, consisting of
an illumination-aware metric-learning stage and a detail-aware discriminative feature-
learning stage. In the first phase, we used coarse-grained labels to describe the illumination
of each image (i.e., “daytime” and “nighttime”). Samples with the same illumination label
were then assigned as S-IL image pairs, and samples with different illumination labels were
assigned as D-IL image pairs, as illustrated in Figure 1. Two separate deep metrics were
then learned for the S-IL and D-IL images in the two illumination-specific feature spaces.
Specifically, we measured the similarity of images with the same illumination label in the
S-IL feature space and images with different illumination labels in the D-IL feature space.
The within-space and cross-space constraints were then enforced to explicitly learn robust visual representations against variations in lighting. In the second stage, we designed a detail-aware discriminative feature learning process, which was learned with the guidance of a local constraint, to extract distinguishing features among similar types of vehicles (as shown in Figure 2). Specifically, a local feature extraction module was introduced to generate local features, and a triplet loss, optimized by triplets of the same model, was designed to enforce the local constraint. Experiments confirmed that both stages were critical for improving re-ID accuracy (see Section 5.4.1).

Figure 2. Visual markers that could be used to distinguish similar vehicles under varying illumination, especially when the negative sample includes the same model with a positive sample under similar illumination. Personalized decorations on the windshield were of particular importance, as shown in the blue boxes. (a) Anchor. (b) Positive. (c) Negative.

As part of the study, we constructed a comprehensive, large-scale dataset termed VERI-DAN (vehicle re-identification across day and night) to address the lack of re-ID datasets exhibiting significant changes in illumination across samples. VERI-DAN included 200,004 images from 16,654 vehicles, each of which was photographed at least 4 times in variable lighting conditions using real-world cameras. The primary contributions of this study can be summarized as follows:

- A large-scale dataset termed VERI-DAN is developed to facilitate vehicle re-ID in various illumination conditions. VERI-DAN is the first dataset of its type to represent changes in lighting on this scale. As such, it simulated a relatively challenging scenario (matching D-IL pairs), which is both common and useful in real-world scenes.

- A two-stage progressive learning strategy is proposed for vehicle re-ID with variable illumination. In Stage I, we introduced an illumination-aware network, which significantly improved the learning of robust visual representations in extreme lighting conditions. In Stage II, we developed a discriminative local feature learning process, which facilitated the ability to distinguish among vehicles with a similar appearance.

- We assessed the effectiveness of this approach using two datasets involving obvious lighting changes (VERI-DAN and Vehicle-1M). Despite only 12% of the training set being usable in Vehicle-1M (due to insufficient lighting variability), the proposed technique achieved SOTA performance for the original testing set.

2. Related Works

2.1. Vehicle Re-ID Methods

Vehicle re-ID, a variation of person re-ID [9], has received increasing attention in recent years, as its viability continues to improve. Zapletal and Herout [10] were the first to collect a large-scale dataset for this purpose, conducting a vision-based study by utilizing color and oriented gradient histograms. Liu et al. [11] combined traditional hand-crafted features with CNN-based deep features, thereby demonstrating that deep features were more discriminative. Liu et al. [12] proposed “PROVID” and made progress with the use of license plate information. Other studies [13,14] have shown that spatial
and temporal information from vehicle images have contributed to improving vehicle re-ID performance. For example, PROVID [12] re-ranks vehicles using spatio-temporal properties based on a simple from-near-distant principle. Wang et al. [15] achieved spatio-temporal regularization for vehicle re-ID by considering the delay time between cameras. These techniques, however, are limited in their application because they require complex spatio-temporal labels.

Several studies have described vehicle re-ID as a metric learning problem and have introduced a series of metric losses to obtain better vehicle representations. Specifically, triplet loss has achieved great success in person re-ID [16–18] and has been adopted in vehicle re-ID. Zhang et al. [12] combined classification loss with triplet loss, providing further benefit. Yan et al. [19] proposed a multi-grain ranking loss to discriminate vehicles with a similar appearance. Studies have also shown that attributes, such as color, brand, and wheel pattern, can further improve re-ID efficacy [1,20–22]. Other strategies [20,21,23–27] have exploited the indirect attributes of a vehicle, such as camera perspective information and background information, making considerable improvements. These techniques, however, have overlooked valuable information present in the image beyond the vehicle itself, such as lighting conditions. In this study, we have suggested that the significance of an image extends beyond its perceptible features. By incorporating less apparent elements, such as illumination, the proposed deep learning model can produce more robust representations in a variety of lighting conditions. Table 2 summarizes related work on vehicle re-ID.

Table 2. Summary of related works on vehicle re-ID.

<table>
<thead>
<tr>
<th>Classification Basis</th>
<th>Method Type</th>
<th>References</th>
<th>Advantage</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature representation approaches</td>
<td>Hand-crafted</td>
<td>Refs. [10,11]</td>
<td>Highly interpretable</td>
<td>High time complexity; low recognition accuracy</td>
</tr>
<tr>
<td></td>
<td>Spatio-temporal information based</td>
<td>Refs. [12–14]</td>
<td>Works well for hard samples</td>
<td>Need extra complex spatio-temporal labels</td>
</tr>
<tr>
<td></td>
<td>Metrics learning based</td>
<td>Refs. [12,19]</td>
<td>High recognition accuracy</td>
<td>High cost</td>
</tr>
<tr>
<td></td>
<td>Multidimensional information based</td>
<td>Refs. [1,20–27]</td>
<td>Sensitivity to the special appearance of vehicles</td>
<td>Vulnerable to variations in viewpoints and illuminations</td>
</tr>
</tbody>
</table>

2.2. Vehicle Re-ID Datasets

Re-ID algorithms that have been applied to public datasets, such as VehicleID [28], VeRi-776 [12], VERI-Wild [29], and Vehicle-1M [30], have conventionally underestimated the importance of illumination. These conditions are limited in existing datasets, as the lighting in each image is typically consistent. For example, all of the samples in VeRi-776 were collected between 4:00 p.m. and 5:00 p.m. We inspected vehicle images in these public datasets and counted the number of samples collected during both the daytime and nighttime. This preliminary evaluation suggested that 90% of samples in VehicleID, 94% of samples in VERI-Wild, and 70% of samples in Vehicle-1M exhibited little variation in background luminance. In this paper, we propose a large-scale vehicle re-ID dataset termed VERI-DAN, which provides a more challenging classification task because each vehicle appears several times in different lighting conditions. The set contains 482 refined vehicle models (e.g., MG3-2016) with highly similar features. As such, this set was more suitable for evaluating the performance of vehicle re-ID methods in challenging scenarios.
3. Methodology

The network architecture for the proposed algorithm is illustrated in Figure 3. This framework constituted a two-stage progressive deep learning process involving illumination conditions and vehicle attributes. During the first stage (Section 3.2), we established an illumination-aware network (IANet) consisting of two branches with identical structures and applied it to both S-IL and D-IL image pairs. This two-branch network leveraged the coarse-grained illumination labels to supervise the learning of illumination-specific features. The Stage I model (IANet) then enabled the retrieval of samples under different lighting conditions. During the second stage (Section 3.3), we introduced a guided local feature extraction process to generate local features. This process included an illumination-aware local feature extraction module (IAM) and a detail-aware local feature extraction module (DAM). This attention mechanism facilitated the learning of distinguishing features among different vehicles with similar appearances, under the supervision of fine-grained model labels. The Stage II model was specifically designed to extract discriminative features from local areas to distinguish among similar types of vehicles. We adopted triplet loss as a learning baseline metric, as discussed in Section 3.1.

Figure 3. The TSPL architecture. The “branch_simi” and “branch_diff” modules accept the output of “shared conv” and transform N images into 2N features (i.e., global_simi and global_diff). TSPL then appends an attention process to the “shared conv” to generate 3N local features (i.e., local_simi, local_diff, and local_features). In the illumination-specific feature space, TSPL expresses the fusion of local features extracted from the IAM (with global features extracted from the branch) as joint vehicle features. TSPL also generates an \( N \times N \) distance matrix using the local features and two \( N \times N \) distance matrices from the fusion features (i.e., matrix3, matrix1, and matrix2). In matrix3, TSPL only uses distances from same-model pairs to calculate the loss function \( L_t \). In matrix1 (matrix2), TSPL only uses the green cells (brown cells) to compute the loss function \( L_S \) (\( L_D \)). In addition, TSPL incorporates matrix1 and matrix2 into the fusion matrix to calculate \( L_{CROSS} \).
3.1. Metric Learning Baseline

We adopted triplet loss to construct a metric learning baseline. Given an image pair \( P = (x_i, x_j) \), distances were calculated using \( D(P) = D(x_i, x_j) = \| f(x_i) - f(x_j) \|_2 \), where \( x_i \) and \( x_j \) represent images from the dataset \( X \) and \( D \) denotes the Euclidean distance between features. The function \( f \) then mapped raw images to their respective features. An example is provided given three samples: \( x, x^+, \) and \( x^- \), where \( x \) and \( x^+ \) belong to the same class (i.e., the same vehicle ID), while \( x \) and \( x^- \) belong to different classes. A positive pair \( P^+ = (x, x^+) \) and a negative pair \( P^- = (x, x^-) \) can then be formed, for which the triplet loss is defined as follows:

\[
L_{Tri}(x, x^+, x^-) = \max\{D(P^+) - D(P^-) + \alpha, 0\},
\]

where \( \alpha \) is a margin enforced between positive and negative pairs. Equation (1) aims to minimize the distance between samples with the same ID while maximizing the distance between samples with different IDs.

3.2. Illumination-Aware Metric Learning

Inspired by previous work [25], we propose an illumination-aware network that learns two separate deep metrics for S-IL and D-IL samples. A coarse-grained classification was included to divide the images into two distinct illumination types. We then employed IANet to learn illumination-specific metrics via the explicitly modeling of lighting conditions. Since it is difficult to manually annotate real-world environments with fine-grained labels, coarse-grained labels were assigned to the images (i.e., daytime and nighttime). Images including annotated timestamps ranging from 06:00 to 18:00 were labeled daytime, and those spanning from 18:00 to 06:00 were labeled nighttime. Datasets lacking a timestamp were categorized using an illumination predictor trained on VERI-DAN samples. Images with the same illumination label were denoted as S-IL pairs, and those with different labels were denoted as D-IL pairs. This convention produced four types of image pairs: \( P_s^+ \) (S-IL positive), \( P_d^+ \) (D-IL positive), \( P_s^- \) (S-IL negative), and \( P_d^- \) (D-IL negative).

Images were mapped into two distinct illumination-specific feature spaces using two convolutional branches with identical structures, which did not share any parameters. Each branch layer could be viewed as a function for illumination-specific feature extraction (i.e., \( f_s \) and \( f_d \)). For each image in a mini-batch, IANet generated two distinct features using \( branch_{simi} \) and \( branch_{diff} \), as illustrated in Figure 3. Pair-wise distances in the S-IL feature space were then calculated from \( D_s(P) = \| f_s(x_i) - f_s(x_j) \|_2 \), and distances in the D-IL feature space were determined by \( D_d(P) = \| f_d(x_i) - f_d(x_j) \|_2 \). Equation (1) was decomposed into two types of constraints: within-space and cross-space constraints. The within-space constraints expect \( D_s(P^+_s) \) to be smaller than \( D_d(P^-_s) \) in the S-IL feature space, and \( D_d(P^+_d) \) to be smaller than \( D_d(P^-_d) \) in the D-IL feature space. The cross-space constraints expect \( D_d(P^+_s) \) to be smaller than \( D_s(P^-_s) \), and \( D_d(P^+_s) \) to be smaller than \( D_d(P^-_s) \).

Within-space constraints: Two triplet loss terms were introduced, one in each of the S-IL and D-IL feature spaces, to ensure that positive samples were closer to each other than negative samples. Triplet loss in the S-IL feature space was defined as follows:

\[
L_s = \max\{D_s(P^+_s) - D_s(P^-_s) + \alpha, 0\},
\]

and in the D-IL feature space as follows:

\[
L_d = \max\{D_d(P^+_d) - D_d(P^-_d) + \alpha, 0\}.
\]

Within-space constraints were then implemented through a summation of \( L_s \) and \( L_d \) as follows:

\[
L_{within} = L_s + L_d.
\]
Within each illumination-specific feature domain, the correlating loss function operated solely on illumination-specific samples. In other words, we used only S-IL pairs to calculate $L_s$, while $L_d$ was optimized solely by D-IL pairs.

**Cross-space constraints:** Focusing solely on single-feature spaces runs the risk of underestimating the complex issue of illumination variability, which in turn could limit re-ID accuracy. As such, we further proposed cross-space constraints between $(P_s^-, P_d^+)$ and $(P_s^+, P_d^-)$, which were implemented using the following triplet loss function:

$$L_{\text{cross}} = \max\{D_s(P_d^+) - D_s(P_s^-) + \alpha, 0\} + \max\{D_s(P_s^+) - D_d(P_d^-) + \alpha, 0\}. \quad (5)$$

**Loss functions in the Stage I model:** The total triplet loss enforced in the first stage can then be expressed as follows:

$$L = L_{\text{within}} + L_{\text{cross}}. \quad (6)$$

### 3.3. Detail-Aware Discriminative Feature Learning

We observed that vehicles with similar appearances often exhibited differences in localized regions, such as windshield decorations, as depicted in Figure 2. Thus, we suggested that re-ID accuracy could be improved by enhancing an algorithm’s capacity to capture these distinctive local details in the second stage. The neural network encoded images in a progressive process \cite{31,32}, beginning with fine-grained details and gradually expanding to local and global information. Thus, mid-level features from the middle layers of the network facilitated the extraction of local area features for the vehicle. As such, based on this approach, we proposed a detail-aware discriminative feature learning process for vehicle re-ID. This process incorporated a local feature extraction module with attention mechanisms included to extract local features. Local constraints were then introduced to guide the generation of these features and devise an illumination-balanced sampling strategy to optimize local constraints.

**Attention-guided local feature extraction module (AG):** Different vehicle parts play varying roles in distinguishing among vehicles that are similar in appearance. Specifically, areas such as a car logo or windshield, for which marked dissimilarities exist between individual vehicles, are more important than common features, such as doors and hoods. To this end, we introduced an attention mechanism to learn from these distinctive areas. This attention-guided process consisted of a detail-aware local feature extraction module (DAM) and an illumination-aware local feature extraction module (IAM), as illustrated in Figure 4.

![Figure 4](image)

**Figure 4.** The attention-guided local feature extraction module (AG). (a) The detail-aware local feature extraction module (DAM), (b) The illumination-aware local feature extraction module (IAM).

The DAM generated detail-aware discriminative local features to enforce local constraints, as shown in Figure 4a. Mid-level features are denoted in the figure by $F$, with dimensions $H \times W \times C$, where $H$, $W$, and $C$ represent the height, width, and number of...
channels in the feature layer, respectively. The attention local feature map $F'$ was then generated with the following equation:

$$F' = A \otimes F = \sigma(g(F) \otimes F),$$

where $g(\cdot)$ is a convolution kernel, $\sigma$ is the sigmoid function, and $\otimes$ denotes element-wise multiplication between two tensors. Global maximum pooling was then applied to $F'$ to produce the final local feature vector (i.e., local features in Figure 3). Each channel in $F'$ represents a specific vehicle region, and spatial points within the vector indicate the significance of each region. In this way, the incorporated attention map was able to guide the network’s focus toward the significant areas of each vehicle.

As demonstrated in Figure 4b, IAM generated two different types of local features that were discriminative in S-IL and D-IL feature spaces, respectively. The appearance of certain distinguishing areas, such as headlights, differed significantly as the illumination changed. In other words, specific visual cues may become more or less significant in different feature spaces. To this end, we further introduced squeeze and excitation modules [33] to identify illumination-specific local features for the S-IL and D-IL space. The corresponding feature map $F''$ is obtained as follows:

$$F'' = A' \otimes F = F_{SE}(A) \otimes F = \sigma(F_{SE}(g(F)) \otimes F),$$

where $F_{SE}$ denotes the squeeze-and-excitation block. Consequently, we obtained two different local features from IAM (i.e., local simi and local diff), as shown in Figure 3. We then employed the union of illumination-specific local features with global features (extracted from the branch network) to enforce within-space and cross-space constraints. This process was distinguished from the formulation defined in Section 3.2 by using $L_s$, $L_d$, and $L_{cross}$ to denote corresponding loss terms calculated from the fusion features.

**Detail-aware local constraints:** In real-world scenarios, differences between vehicles of the same model were concentrated primarily in regions such as inspection marks and personalized decorations. As such, training a triplet loss function using hard negatives from the same vehicle model served as a guiding mechanism that directed network attention toward relevant discriminative local regions. In the following notation, $x_m$ denotes a vehicle belonging to model $m$. A typical triplet in the local constraints is denoted by $(x_m, x^+_m, x^-_m)$, where $x_m$ and $x^+_m$ exhibit the same ID, while $x^-_m$ has a different ID but shares a model type with $x_m$ and $x^+_m$. Following these definitions, $P^+_m$ (same-model positive pair) and $P^-_m$ (same-model negative pair) are denoted as $(x_m, x^+_m)$ and $(x_m, x^-_m)$, respectively. Formally, local constraints were enforced through triplet loss as follows:

$$L_I = L_I(x_m, x^+_m, x^-_m) = \max\{D(P^+_m) - D(P^-_m) + \beta, 0\},$$

where $\beta$ is a margin enforced between positive and negative pairs. All negative samples in the proposed local constraints shared a model type with the anchor, which was conducive to guiding the generation of discriminative local features. Note that $L_I$ is an advanced version of $L_{cross}$, since different vehicles of the same model were prone to generate the hardest negatives in S-IL and D-IL feature spaces. Thus, we removed $L_{cross}$ from the final model after introducing local constraints.

**Loss functions in the Stage II model:** The total triplet loss function in the Stage II model can be expressed as follows:

$$L = L_s + L_d + L_I.$$  

**Illumination-balanced sampling strategy:** Maintaining a balance between S-IL and D-IL pairs is necessary in a mini-batch to train an illumination-aware network against variations in lighting. However, in most cases, the number of daytime images in each mini-batch is much larger than that of the nighttime images. As a result, the network may tend to learn from images captured in the daytime and may not be able to identify a robust
correlation among samples with different illumination. To address this issue, we designed a function to ensure that each vehicle provided an equal number of images for both types of lighting. Specifically, the algorithm selected N daytime images and N nighttime images for each vehicle ID in a minibatch. If a vehicle ID exhibited fewer than N daytime images, the algorithm duplicated these samples to produce N images. The effectiveness of this balanced sampling strategy will be illustrated in Section 5.4.2.

3.4. Training and Inference

TSPL expressed the fusion of local features with global features extracted from “branch_conv” as joint vehicle features, as demonstrated in Figure 3. During training, “branch_conv” output dual global features for each image in different feature spaces (i.e., global_simi in S-IL and global_diff in D-IL). In contrast, IAM output dual illumination-specific local features to form joint representation with global features (i.e., local_simi and local_diff). DAM then output detail-aware local features to optimize the local constraints. Given N input images, TSPL generated two illumination-specific distance matrices containing $N \times N$ distance value elements (i.e., matrix1 and matrix2). Only $D_s(P_s)$ and $D_d(P_d)$ (denoted by the green and brown cells in Figure 3) contributed to $L_s$ and $L_D$ loss, respectively. TSPL also generated a local distance matrix containing $N \times N$ distance value elements calculated from local features. Distances from the same model (denoted by the colored cells in matrix3) were then used to calculate the triplet loss $L_t$ and to enforce detail-aware local constraints.

Note that $L_{CROSS}$ was not incorporated into the final model but was a component of the ablation study. The generation of $L_{CROSS}$ was illustrated by the red dashed line box shown in Figure 3. In the fusion matrix, the green cells were related to S-IL pair distance values in matrix1, while the brown cells were related to D-IL distance values in matrix2. During the testing phase, a specific procedure was followed based on the illumination conditions of query and gallery images. If these images were identified as a S-IL pair, their distance was calculated by $D_s(P_s)$ using a S-IL branch; otherwise, $D_d(P_d)$ was employed through a D-IL branch. Distances were also calculated from local features and the union of these results provided joint distances between the query and gallery images.

4. The VERI-DAN Dataset

Existing re-ID datasets either exhibit limited illumination variability or lack annotations to quantify luminance. Therefore, we carefully constructed the VERI-DAN dataset to provide a variety of lighting conditions for each vehicle. The set included 200,004 total images from 16,654 vehicles, collected by 120 cameras in a large urban district in natural environments. Statistics for this dataset are provided in Figure 5 and sample images are shown in Figure 6a. Table 3 presents a comparison of VehicleID [28], VeRI-776 [12], VERI-Wild [29], Vehicle-1M [30], and VERI-DAN. The distinctive properties of VERI-DAN can be summarized as follows:

Balanced illumination conditions: VERI-DAN was generated from $120 \times 24 \times 8 = 23,040$ h of video footage collected in various illumination conditions. Specifically, every vehicle appeared multiple times in both daytime and nighttime settings, as shown in Figure 6b.

Refined model information: We meticulously annotated each image with one of 482 refined vehicle classes, denoting the make, model, color, and year (e.g., “Audi-A6-2013”). VERI-DAN included many similar vehicles of the same model, which facilitated the training of a network to differentiate hard negative samples.

Spatio-temporal geographic coordinate information (S-T): Details such as camera ID, timestamp, and geodetic coordinates were provided to facilitate research based on camera networks [34–36].
Figure 5. VERI-DAN dataset statistics. (a) The number of identities collected by camera (i.e., 1–120). (b) The number of IDs captured by hour. (c) The distribution of captured images by time of day.

Figure 6. An overview of the VERI-DAN dataset, including (a) sample images and (b) examples of changes in illumination.

Table 3. A comparison of re-ID datasets. S-T, spatio-temporal geographic coordinate information; ID-DL, ID with diverse lighting.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>VehicleID</th>
<th>Vehicle-1M</th>
<th>VERI-Wild</th>
<th>VERI-DAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>221,763</td>
<td>936,051</td>
<td>416,314</td>
<td>200,004</td>
</tr>
<tr>
<td>Identities</td>
<td>26,267</td>
<td>55,527</td>
<td>40,671</td>
<td>16,654</td>
</tr>
<tr>
<td>Models</td>
<td>250</td>
<td>400</td>
<td>153</td>
<td>482</td>
</tr>
<tr>
<td>S-T</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ID-DL (%)</td>
<td>10.3%</td>
<td>29.9%</td>
<td>6.3%</td>
<td>100%</td>
</tr>
</tbody>
</table>

5. Experiments

5.1. Datasets

We evaluated the proposed method using the VERI-DAN and Vehicle-1M datasets, both of which included vehicles exhibiting significant illumination changes. Following a common practice [28], we divided VERI-DAN into a training set (141,470 images from 13,454 vehicles) and a test set, which contained the remaining 58,534 images from 4800 vehicles. We further divided the test set into three subsets, denoted Small, Medium, and Large, as shown in Table 4.
5.2. Evaluation Protocols

During evaluation, we followed the protocol proposed by Liu et al. [12,28], in which mean average precision (mAP) and cumulative matching characteristics (CMC) were used as performance metrics. CMC estimates the probability of finding a correct match in the top K returned results, while MAP is a comprehensive index that considers both the precision and recall of the results. The final CMC and mAP values were averaged over 10 iterations.

5.3. Implementation Details

We adopted the InceptionV3 [37] network as the backbone model. All layers preceding the Inception (7a) module were implemented as "share conv", and layers ranging from Inception (7a) to the global average pooling layer were appended as "branch conv". Since mid-level features facilitated the extraction of discriminative local vehicle features, we added an attention module to the Inception (5d) layers to generate local feature maps of dimensions $35 \times 35 \times 288$. The input images were then resized to $299 \times 299$ without augmentation, using processes such as color jitter and horizontal flip. The model was trained for 120 epochs using the Adam optimizer with a momentum of 0.9 and a weight decay of 0.05. The learning rate was initialized to 0.001, and decreased by a factor of 0.1 every 20 epochs. The margins $\alpha$ and $\beta$ were both set to 1.0. Each mini-batch contained 128 images (32 IDS, each with 4 images) on VERI-DAN as well as on Vehicle-1M. We adopted a batch hard-mining strategy to reduce the triplet loss.

The illumination predictor was trained using cross-entropy based on InceptionV3. We coarsely categorized all images into two illumination classes: daytime and nighttime. The daytime—daytime and nighttime—nighttime samples were then defined as the S-IL pairs, and the daytime—nighttime samples were defined as the D-IL pairs. In addition to triplet loss, we incorporated cross-entropy loss into the model to learn differences between the individual vehicle models, drawing inspiration from several existing re-ID methods [1,25,38]. Specifically, we appended the model classifier into the feature-embedding layer. The classifier was then implemented with a fully connected layer and a softmax layer. The output of the softmax was supervised by the model labels applied to the training images, and optimized by the cross-entropy loss step.

5.4. Ablation Study

We conducted ablation studies on the two large-scale datasets to validate the effectiveness of the proposed strategies both quantitatively and qualitatively. We provided a detailed analysis of the impacts arising from constraints in Section 5.4.1 and the sampling strategy in Section 5.4.2, respectively.

5.4.1. Constraint Influence

We conducted a series of comparison experiments to validate the effectiveness of the included illumination-aware metric learning and detail-aware discriminative feature learning. Specifically, we performed comprehensive ablation studies on combinations of $L_1$ with $L_{within}$ and $L_{cross}$. This was carried out to verify the benefits of fusing global and local features, as demonstrated in Table 5. Note that $TSPL_{-3}$ and $TSPL_3$ used the same model to generate features, although $TSPL_{-3}$ did not consider local features when calculating distances. It is evident from the table that our proposed strategy significantly improved vehicle re-ID performance over the baseline method.
Table 5. Evaluation results for the small VERI-DAN and Vehicle-1M test sets (%). The best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>$L_{\text{within}}$</th>
<th>$L_{\text{cross}}$</th>
<th>$L_{\text{local}}$</th>
<th>VERI-DAN</th>
<th>Vehicle-1M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank-1</td>
<td>Rank-5</td>
</tr>
<tr>
<td>(a) Baseline</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>96.26</td>
<td>99.23</td>
</tr>
<tr>
<td>(b) IANet</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>98.83</td>
<td>99.70</td>
</tr>
<tr>
<td>(c) TSPL1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>99.55</td>
<td>99.93</td>
</tr>
<tr>
<td>(d) TSPL2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>99.40</td>
<td>99.94</td>
</tr>
<tr>
<td>(e) TSPL3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>99.52</td>
<td>99.95</td>
</tr>
<tr>
<td>(f) TSPL3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>99.54</td>
<td>99.95</td>
</tr>
</tbody>
</table>

IANet produced significant improvements despite relatively coarse-grained illumination classification. Comparing IANet with the single-branch baseline demonstrated Rank-1 accuracy improvements of +6.74% for Vehicle-1M and +2.57% for VERI-DAN. Consistent performance improvements across both datasets confirmed the effectiveness of a multi-branch network following the proposed progressive strategy.

Local constraints provide significant benefits for vehicle re-ID. Every variation of IANet, with the introduction of local constraints (i.e., TSPL1, TSPL2, and TSPL3) produced considerable improvements over IANet. Specifically, compared with IANet, local constraints yielded Rank-1 accuracy improvements of +3.69% for Vehicle-1M and +0.72% for VERI-DAN. In addition, combining either $L_{\text{within}}$ or $L_{\text{cross}}$ with $L_{\text{local}}$ resulted in better performance than the combination of within-space and cross-space constraints. Combining cross-space constraints and local constraints resulted in Rank-1 accuracy increases of +2.91% for Vehicle-1M and +0.57% for VERI-DAN. The joint optimization of within-space and local constraints produced similar results, that is, +4.37% for Vehicle-1M and +0.71% for VERI-DAN. TSPL3 achieved the best performance, while TSPL3 outperformed IANet by ~1.8% (Rank-1 accuracy), even when local features were not involved in distance matrix calculations during testing. This result suggested that local constraints, which focus on extracting and leveraging local fine-grained visual cues, are also well suited for processing the variations and challenges introduced by illumination.

Within-space constraints are critical for re-ID. We observed a performance degradation when comparing TSPL2 with TSPL1 (−0.78% Rank-1 accuracy for Vehicle-1M and −0.15% Rank-1 accuracy for VERI-DAN.) In contrast, either TSPL3 or TSPL3 achieved better performance than TSPL2 and TSPL1. This was reasonable because without within-space constraints, TSPL was not able to learn from two relatively common scenarios: both the positive pairs and the negative pairs are observed under S-IL (or D-IL). We thus inferred that within-space constraints played a critical role in enhancing the retrieval capacity.

5.4.2. Sampling Strategy Influence

We developed a special test set to validate the importance of maintaining a balance between S-IL and D-IL pairs during training. All query and gallery samples were D-IL pairs, as summarized in Table 6. For each individual in a mini-batch, a ratio of 1:3 indicated that one image was captured during the daytime, while the other three images were taken during the nighttime. As demonstrated in Table 6, both the baseline and IANet achieved the best performance when this ratio was 2:2. This outcome suggested that maintaining an illumination-balanced sampling strategy was beneficial for query retrieval from the D-IL gallery.
Table 6. The effect of sampling strategy (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>Daytime: Nighttime</th>
<th>Test Size = 800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank-1</td>
<td>Rank-5</td>
</tr>
<tr>
<td>Baseline</td>
<td>Random</td>
<td>89.45</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>88.8</td>
</tr>
<tr>
<td></td>
<td>2:2</td>
<td><strong>92.05</strong></td>
</tr>
<tr>
<td></td>
<td>3:1</td>
<td>89.60</td>
</tr>
<tr>
<td>IANet</td>
<td>Random</td>
<td>96.30</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>96.37</td>
</tr>
<tr>
<td></td>
<td>2:2</td>
<td><strong>96.82</strong></td>
</tr>
<tr>
<td></td>
<td>3:1</td>
<td>96.15</td>
</tr>
</tbody>
</table>

5.5. Comparison with SOTA Methods

We compared the proposed method with a variety of SOTA vehicle re-ID methods on VERI-DAN and Vehicle-1M. We used InceptionV3 [37] as our baseline model, which was pretrained on ImageNet [39]. C2F-Rank [30] designed a multi-grain ranking loss to efficiently learn feature embedding with a coarse-to-fine structure. GSTN [40] automatically located vehicles and performed division for regional features to produce robust part-based features for re-ID. DFR-ST [13] involved appearance and spatio-temporal information to build robust features in the embedding space. DSN [41] utilized a cross-region attention to enhance spatial awareness of local features. The comparison results on the two datasets are detailed in Sections 5.5.1 and 5.5.2, respectively.

5.5.1. Evaluation with VERI-DAN

We verified the effectiveness of the proposed methodology in the presence of significant illumination changes by conducting comprehensive validation experiments using the VERI-DAN dataset, while also drawing comparisons among the InceptionV3 [37], IANet, and TSPL. As shown in Table 7, TSPL achieved significant improvements in Rank-1 accuracy and mAP over the baseline. The superiority of TSPL was evident through visual inspection as well, as illustrated in Figure 7. When compared with the baseline, we observed that TSPL could identify more correct matches among retrieved ranking lists in which the correct matches had lower rank values.

Figure 7. A visual comparison of baseline (top) and TSPL (bottom) performance. Images with blue contours show query vehicles, and images with green and red contours indicate correct and incorrect predictions, respectively.
Table 7. Performance with the VERI-DAN dataset (%). The best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>InceptionV3 [37]</td>
<td>96.26</td>
<td>93.59</td>
<td>93.27</td>
</tr>
<tr>
<td>IANet</td>
<td>98.83</td>
<td>97.65</td>
<td>98.42</td>
</tr>
<tr>
<td>TSPL</td>
<td>99.54</td>
<td>98.91</td>
<td>99.51</td>
</tr>
</tbody>
</table>

5.5.2. Evaluation with Vehicle-1M

We selected a group of 6448 (out of 50,000) vehicles in Vehicle-1M, whose images involved a variety of illumination conditions, as the training set. TSPL was compared with a variety of SOTA methods developed in recent years, including C2F-Rank [30], GSTN [40], DFR-ST [13], and DSN [41], as shown in Table 8. Despite the limited utilization of training samples, TSPL achieved remarkably competitive performance without any modified training strategies or additional detection modules. Specifically, compared with the second-place method, TSPL achieved +1.83%, +1.46%, and +1.66% Rank-1 improvements for the small, medium, and large test sets, respectively. This outcome demonstrated the superiority of the proposed two-stage progressive learning framework. Re-ranking and data augmentation may further improve performance.

Table 8. A comparison with SOTA algorithms applied to Vehicle-1M (%). The best results are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Training Set</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDS Images</td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>C2F-Rank (2018)</td>
<td>GoogLeNet</td>
<td>50,000</td>
<td>844,571</td>
<td>67.1</td>
<td>87.1</td>
</tr>
<tr>
<td>GSTN (2019)</td>
<td>ResNet18</td>
<td>50,000</td>
<td>844,571</td>
<td>95.14</td>
<td>96.29</td>
</tr>
<tr>
<td>DFR-ST (2022)</td>
<td>ResNet50</td>
<td>50,000</td>
<td>844,571</td>
<td>93.04</td>
<td>96.70</td>
</tr>
<tr>
<td>DSN (2023)</td>
<td>ResNet50</td>
<td>50,000</td>
<td>844,571</td>
<td>92.9</td>
<td>93.7</td>
</tr>
<tr>
<td>TSPL (Ours)</td>
<td>InceptionV3</td>
<td>6448</td>
<td>525,808</td>
<td>96.97</td>
<td>95.40</td>
</tr>
</tbody>
</table>

6. Conclusions

To address the challenging problem posed by dramatic changes in illumination, we proposed a novel two stage-progressive learning (TSPL) strategy for vehicle re-identification. This technique consisted of illumination-aware metric learning and detail-aware discriminative feature learning. Unlike existing methods that only learn a single metric for all types of lighting conditions, Stage I of TSPL aimed to learn proper vehicle representations under different illumination conditions using a two-branch network called IANet, which learned two separate metrics for images with similar and differing illumination conditions. By enforcing corresponding constraints (i.e., within-space constraints and cross-space constraints), IANet improved re-ID accuracy when retrieving D-IL images. Stage II in TSPL enabled the network to learn discriminative local features through an attention guided local feature extraction module (AG), which was optimized by local constraints. The proposed attention module not only facilitated the distinguishing of vehicles with similar appearances but also increased the associated robustness against variations in illumination. Additionally, the large-scale VERI-DAN was developed as part of the study, to provide images with significant changes in lighting. VERI-DAN is expected to facilitate the development of new re-ID methods by suppressing the distractions introduced by variable illumination. The implementation of each proposed metric learning strategy consistently improved re-ID performance with both VERI-DAN and Vehicle-1M, which further verified the effectiveness of TSPL. Despite the limited number of training samples in Vehicle-1M, TSPL achieved SOTA Rank-1 accuracy for the original test set, thereby demonstrating the superiority of this approach.
Author Contributions: Conceptualization, Z.W., Z.J. and X.L.; methodology, Z.W., Z.J. and X.L.; software, Z.W.; validation, Z.W., Z.J. and X.L.; formal analysis, Z.W.; investigation, Z.W.; resources, X.L.; data curation, Z.W.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W., Z.J. and X.L.; visualization, Z.W.; supervision, Z.J. and X.L.; project administration, Z.J. and X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant No. U21B2090) and Natural Science Foundation of Guangdong Province, China (grant No. 2022A1515010361).

Data Availability Statement: The data presented in this study are available in the article. VERI-DAN is available at http://www.openits.cn/openData4/825.jhtml (accessed on 8 November 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

\[ P = (x_i, x_j) \]  
An image pair consists of image \( x_i \) and \( x_j \)

\[ D(x_i, x_j) \]  
Euclidean distance between \( x_i \) and \( x_j \)

\( x, x^+, \) and \( x^- \)  
x and \( x^+ \) share the same vehicle ID, while \( x \) and \( x^- \) have different vehicle ID

S-IL  
Similar-illumination

D-IL  
Different-illumination

\( P^+_s \)  
S-IL positive

\( P^+_d \)  
D-IL positive

\( P^-_s \)  
S-IL negative

\( P^-_d \)  
D-IL negative

\( P^+_m \)  
same-model positive pair

\( P^-_m \)  
same-model negative pair

AG  
Attention-guided local feature extraction module

CMC  
Cumulative matching characteristics

DAM  
Detail-aware local feature extraction module

IAM  
Illumination-aware local feature extraction module

IANet  
Illumination-aware network

mAP  
Mean average precision

re-ID  
Re-identification

SOTA  
State-of-the-art

S-T  
Spatio-temporal geographic coordinate information

TSPL  
Two-stage progressive learning

VERI-DAN  
Vehicle re-identification across day and night

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.