Abstract: Vehicle re-identification (vehicle ReID) is designed to recognize all instances of a specific vehicle across various camera viewpoints, facing significant challenges such as high similarity among different vehicles from the same viewpoint and substantial variance for the same vehicle across different viewpoints. In this paper, we introduce the RAND network, which is equipped with relation attention mechanisms, nuance, and disparity masks to tackle these issues effectively. The disparity mask specifically targets the automatic suppression of irrelevant foreground and background noise, while the nuance mask reveals less obvious, sub-discriminative regions to enhance the overall feature robustness. Additionally, our relation attention module, which incorporates an advanced transformer architecture, significantly reduces intra-class distances, thereby improving the accuracy of vehicle identification across diverse viewpoints. The performance of our approach has been thoroughly evaluated on widely recognized datasets such as VeRi-776 and VehicleID, where it demonstrates superior effectiveness and competes robustly with other leading methods.

Keywords: vehicle re-identification; relation attention; nuance mask; disparity mask; representation learning

1. Introduction

Vehicle re-identification (ReID) has increasingly become a focal task due to its significant applications in intelligent transportation systems [1–5]. The aim of vehicle ReID lies in recognizing all instances of a specific vehicle across various camera viewpoints. Substantial progress has been made in this domain, where a common approach involves the construction of feature representations for each vehicle image, followed by the calculation of similarity rankings between these images based on the computed feature distances. Unlike person ReID, the primary challenge in vehicle ReID stems from the rigid structure of vehicles, which results in excessive similarity among different vehicles from the same viewpoint and considerable variance for the same vehicle across different viewpoints.

Recent research has yielded notable achievements. Liu et al. [6] introduced a progressive framework leveraging multi-modality information, while Shen et al. [7] developed a Hybrid Pyramidal Graph Network for feature extraction at various granularities, enhancing vehicle detail retention. Similarly, Zheng et al. [8] utilized public datasets to craft a two-stage approach for robust visual representation, emphasizing holistic vehicle appearance. Despite these successes, a challenge remains in capturing diversified discriminative parts through global structural information alone. The importance of fine-grained, local area features has led to the emergence of attention-based methods. Chen et al. [9] introduced a
high-order attention module to leverage complex information within the attention mechanism, and Zhang et al. [10] proposed stacking relations of feature information to effectively capture both global structure and local details. These innovations highlight the balance between global scope and localized attention to detail.

However, these methodologies, despite their effectiveness, do not fully resolve all issues. As shown in Figure 1, approaches emphasizing local features—such as stickers on the front window or the view of the interior through it [11]—aid in differentiating between similar vehicle models but may become ineffective due to occlusions or changes in viewpoint because these features do not always appear in images from all viewpoints. Attempts [12,13] have been made to counter the loss of critical features by increasing the weight of viewpoint-independent features or concatenating an extra feature vector representing the vehicle direction. However, these solutions either require additional hard-to-obtain information or their discriminative capability for learned features is often insufficient to differentiate similar vehicle models. For instance, DFnet [12] uses directional labels to learn specific orientation-related features and attr-net [13] focuses on body color or vehicle model attributes as viewpoint-independent features. Most of these methods use IBNnet [14] as the backbone network.

Addressing these challenges necessitates a framework capable of mitigating background and foreground occlusions; discerning multiple discriminative regions; and learning adaptive, viewpoint-related structural information without extra labeling. In response, we propose a vehicle ReID network endowed with relation attention mechanisms and nuanced and disparity masks. The proposed network is based on the following phenomenon: “Simple negative classes”, which differ markedly in vehicle model and color from the target and tend to activate at unrelated aspects such as background or foreground occlusions during training. In contrast, “hard negative classes”, which share similar vehicle models, focus on car emblems, headlights, or tire regions. Through the intersection of class activation maps from both simple and hard negative identities, we discern vehicle background and sub-discriminative regions, encapsulated in the disparity mask and nuance mask, respectively. Furthermore, our relation attention module, leveraging an enhanced transformer architecture, facilitates self-attention across vehicle component information, distilled into patches from each $c \times 1 \times 1$ feature map unit, fostering a network understanding of the inter-relations among vehicle components. This methodology adeptly minimizes intra-class distances across varying viewpoints of the same ID, requiring no additional data beyond...
vehicle IDs for network training. Our approach achieved a SOTA performance on VERI-776 and VehicleID datasets.

The primary contributions of this paper are summarized as follows:

1. We propose RANDnet, a network designed to mitigate background and foreground occlusions, discern multiple discriminative regions, and adaptively learn structural information related to vehicle viewpoints without the need for additional labeling, thereby significantly enhancing the accuracy of vehicle re-identification.

2. The introduction of the disparity mask for the automatic suppression of foreground noise or the background through simple negative class response intersections, and the nuance mask, which leverages hard negative class response intersections to reveal sub-discriminative vehicle image regions, facilitating the acquisition of robust global features.

3. We present a relation attention module incorporating an advanced transformer architecture, which enables the network to assimilate a relationship vector among various vehicle components, significantly narrowing the intra-class distances across divergent viewpoints.

4. Extensive experiments and visualization results demonstrate that our approach achieves the SOTA in the task of vehicle re-identification.

2. Related work
2.1. Vehicle Re-Identification

To achieve discriminative feature representation in vehicle re-identification, some previous methods utilize orientation information. As vehicles are shot by cameras from various angles, images of an identical car from distinct orientations would vary largely, and sometimes it is difficult to identify similar vehicles of the same orientation. Wang et al. [11] present a novel framework composed of the orientation invariant feature embedding and the spatial–temporal regularization. In addition, to obtain vehicle key points, which can be used to distinguish similar cars via subtle differences, Wang et al. also adopt a key point regressor. Chu et al. [15] tackle the vehicle re-identification problem by learning viewpoint-aware deep metrics through a two-branch network. With adversarial training architecture, Zhou Yi and Shao Ling [16] utilize a viewpoint-aware attention model to realize multi-view feature inference from single-view input. Zhou et al. [17] focus on the uncertainty of different vehicle orientations and utilize LSTM to model transformations across continuous orientation variations of the same vehicle. Zhu et al. [18] propose train re-identification models for vehicles, orientations, and cameras, respectively. Then, the final similarity between the testing images is penalized by the orientation and camera similarity.

Besides orientations, some previous works address vehicle re-identification problems by exploiting local details and regions [19–21]. Meng et al. [22] design a parser to segment four views of a vehicle and propose a parsing-based view-aware embedding network to generate fine-grained representations. To address the near-duplicate problem in vehicle re-identification, He et al. [19] propose a part-regularized approach that can enhance local features. Liu et al. [23] introduce a multi-branch model that learns global and regional features simultaneously. In [23], ratio weights of regional features are adaptively predicted for the fusion process. Furthermore, to optimize the distance within and across vehicle image groups, a Group–Group loss is also proposed. Shen et al. [24] propose a two-stage framework which incorporates important visual–spatial–temporal path information for regularization. Khorramshahi et al. [25] present self-supervised attention for vehicle re-identification to extract vehicle-specific discriminative features. However, these methods may only identify too small discriminable regions and thus reduce the generalization of the network. We visualized the feature map of IBNnet [14], a commonly used backbone network of these methods. As shown in the left column of Figure 2, the network pays too much attention to the position of the car lights. In this paper, we also utilize discriminative regions to improve the re-identification accuracy. Specifically, a region mask is generated to highlight vehicle details. Note that no additional manual annotation is involved in this process.
Figure 2. The class activation maps generated by IBNNet. For the target vehicle, only a single most discriminative region is activated in the positive category (left). In contrast, for negative categories (right), simple negative category (above) activations are predominantly located in the background, whereas complex negative category (below) activations focus on other details of the vehicle. The areas with the highest activation are highlighted in red boxes.

2.2. Attention Mechanism

Vehicle re-identification (ReID) tasks inherently require the ability to distinguish subtle differences between highly similar vehicles, a challenge that particularly benefits from the use of attention mechanisms. These mechanisms enhance the model’s ability to focus selectively on the most informative parts of an image, crucial for identifying vehicles across different camera captures.

Attention mechanisms were first introduced in the field of natural language processing, as evidenced by several foundational studies [26–29]. Bahdanau et al. [26] extended the encoder–decoder model to include a mechanism that learns to align and translate jointly, which enables the model to focus on relevant parts of a source sentence dynamically rather than relying on rigid segmentation. Luong et al. [30] introduced both global and local attention mechanisms to neural machine translation, where global attention assesses all source words and local attention targets a subset at any given time. Vaswani et al. [31] further innovated in this space by proposing the transformer, which utilizes multi-headed self-attention to replace recurrent layers, thus basing the entire sequence transduction model on attention mechanisms. Shaw et al. [32] expanded on self-attention by incorporating relative positions and distances between sequence elements.

The utility of attention mechanisms in enhancing the performance of deep neural networks has led to their adoption in various other domains such as image classification, semantic segmentation [33], and object recognition [34]. Hu et al. [35] introduced the Squeeze-and-Excitation block to adaptively adjust channel-wise feature responses by explicitly modeling interdependencies, focusing on the relationships between channels. Wang et al. [36] developed the Residual Attention Network, stacking attention modules to produce features that are attention-aware. Woo et al. [37] proposed the Convolutional Block Attention Module, which refines features using attention across both channel and spatial dimensions. Park et al. [38] enhanced the representational power of networks with a Bottleneck Attention Module. Bello et al. [39] introduced a two-dimensional relative self-attention mechanism as an alternative computational primitive to traditional convolutions. Li et al. [40] devised a Harmonious Attention Convolutional Neural Network that integrates attention selection with feature representation learning, incorporating a cross-attention mechanism. Liu et al. [41] designed a Multi-Task Attention Network to handle multiple learning tasks simultaneously, with each task benefiting from customized attention modules that facilitate the learning of both shared and task-specific features.
3. Vehicle Re-Identification With RANDnet

In this section, we initially formulate the vehicle re-identification problem and define the intersecting features, which respectively represent the significant differential characteristics and the subtle distinguishing features among vehicles. Then, we introduce the proposed nuance–disparity masks and discuss the rational attention module in great detail. Finally, we propose a novel network called RANDnet to immigrate the modules.

3.1. Problem Statement and Intersecting Features

Vehicle re-identification involves analyzing a given vehicle image and retrieving all corresponding images of the same identity from a gallery. These gallery images are captured by multiple non-overlapping cameras positioned at various viewpoints, emphasizing the need for robust analysis to accurately match vehicle identities across diverse perspectives. The vehicle re-identification problem can be formulated as follows: Within the image set represented as \( \langle x_n, y_i \rangle^{N,I} \), we aim to establish the correspondence between the vehicle image \( x_n \) and its corresponding ID \( y_i \). Here, \( N \) and \( I \), respectively, denote the total number of vehicle images and IDs. The common method is to first obtain a feature map \( F \in \mathbb{R}^{C \times H \times W} \) of the query image, then yield the corresponding feature vector by undergoing a global average pooling layer. The similarity of an image pair is measured by calculating the distance of corresponding feature vectors, where the pair with the highest similarity is considered to have the same ID.

Through the visualization of Figure 2, an intriguing phenomenon was observed: aside from unique features, incorrect IDs can elicit a response in specific parts of the query image. We refer to these features as intersecting features between different IDs.

In an image pair of entirely different vehicle models, intersecting features might include common background areas, such as roads, trees, road signs, or fences. In a pair with similar car models, intersecting features generally encompass aspects like similar colors, headlights, tires, or emblems. For the pairs with more similarity but not in the same ID, intersecting features often appear in more detailed adornments, like sunroofs. It can be posited that the intersecting features of simple and hard identities do not overlap. In essence, the characteristics of a vehicle can be considered a conglomeration of these intersecting features and the unique features for the ground truth identity. To address these features distinctly, we propose the RANDnet framework with nuance–disparity masks and a relation attention module.

3.2. Nuance–Disparity Masks

In this section, we introduce the construction principles and methodologies for the nuance–disparity masks, which are automatically generated by the network. These masks are designed to adaptively mine a broader spectrum of discriminative regions within vehicle images while concurrently suppressing extraneous information. The nuance mask is specifically engineered to augment the network’s capability to recognize a more comprehensive set of target features, whereas the disparity mask focuses on mitigating background interference. This enables a refined focus on essential details critical for effective vehicle identification. The N&D mask generation process is shown in Figure 3.

IAF. In order to better describe the composition of the disparity and nuance masks, we first define the identity-associated features (IAFs) and identity intersection features (IIFs).

In the re-identification network, a global average pooling (GAP) layer and a fully connected (FC) layer are involved to obtain a vehicle ID embedding. Let the feature vector after GAP operation be \( f = [f_1, f_2, \ldots, f_i, \ldots, f_c] \), where \( c \) denotes the number of channels. \( f_k \) is defined as

\[
f_k = \frac{1}{h \cdot w} \sum_{i=1}^{h} \sum_{j=1}^{w} F_{i,j,k}, \tag{1}
\]
where $F$ denotes an $h \times w \times c$ feature map embedded by the backbone network. $h$ and $w$ denote the height and width of the feature map, respectively. Then, the result of the GAP layer is converted into a 1D tensor via the FC layer.

Figure 3. N&D masks generation process. We assume there are 4 IDs in the training set. A nuance or disparity mask is obtained by summing 3 IIFs belonging to other categories. The activation score in attention map reflects the importance of corresponding spatial units for recognition.

For a given vehicle image, its identity is labeled as $gt$. Then, the IAF is denoted as $M_{gt}^{i,j} \in \mathbb{R}^{h \times w}$:

$$M_{gt}^{i,j} = \sum_{k=1}^{c} W_{gt,k} \cdot F_{i,j,k},$$

where each spatial unit $(i,j)$ of the feature map $F$ is utilized, together with the FC layer parameter $W$. Note that, for each class, the average value of the IAF is equal to the corresponding dimension of the softmax input, for category $gt$,

$$\frac{1}{h \cdot w} \sum_{j=1}^{h} \sum_{k=1}^{w} M_{gt}^{i,j} = \arg \max \text{softmax}(Z),$$

where $Z$ is the input of softmax, namely $Z = W \times f$.

For an IAF, a larger $M_{gt}^{i,j}$ means that the spatial position $(i,j)$ contains more significant clues that lead to a higher predicted probability of label $gt$.

**IIF.** Identity intersection features are proposed based on the following analysis. The IAF adeptly captures the most discriminative regions of a vehicle specific to a particular ID. However, it is not good enough to determine the discrimination area directly through the IAF. The region indicated by the highly active position in $M_{gt}^{i,j}$ may not be sufficient, and some other important information might be neglected. This is especially prominent when there are some exclusive characteristics in a particular vehicle image, e.g., special cargo or self-made signs. Thus, utilizing solely IAFs can diminish the model’s generalization capability. Furthermore, when the IAF vanishes from the feature maps due to likely occurrences of viewpoint changes or foreground occlusions, the network exhibits a deficiency in localizing sub-discriminative regions within these feature maps. This significantly exacerbates the increase in intra-class variance.

For distant classes with significant differences in car types, the IIF of the input vehicle images predominantly occurs in the background, indicating that both positive and negative samples share a similar backdrop. In contrast, in the case of ambiguous classes where car types are subtly distinct, the IIF tends to concentrate on specific vehicle components such as emblems, headlights, and tires, signifying that the positive and negative samples possess similar vehicular parts.

Consequently, by suppressing the IIF in distant classes, we can effectively mitigate the interference of background noise on model performance. Conversely, enhancing the IIF in ambiguous classes facilitates the model’s focus on regions with sub-discriminative features, thereby augmenting the robustness of feature representation. Stemming from this concept, we introduce the disparity and nuance masks, designed to adaptively discern and
accentuate the IIF at various levels within input images, promoting more nuanced and differentiated feature learning.

The detailed procedure for obtaining D&N masks is outlined in Algorithm 1, wherein the parameters \( \theta_{\text{disp}} \) and \( \theta_{\text{nuc}} \) dictate the involvement of hard or easy categories in defining sub-discriminative or background regions, respectively. In addition to these parameters, another parameter set is introduced to modulate the extent of area covered by the mask within the image. The area of discriminative regions can be governed by a quantile threshold. Consequently, \( \sigma_{\text{disp}} \) and \( \sigma_{\text{nuc}} \) are conceptualized to denote the proportion of discriminative regions relative to the entire image. An in-depth analysis of these parameters is provided in Section 4.3.

### Algorithm 1: Nuance–Disparity Mask

1: **Input:** \( F \in \mathbb{R}^{BS \times C \times H \times W} \), ID count thresholds for Disparity & Nuance masks \( \theta_{\text{disp}} \) and \( \theta_{\text{nuc}} \), area thresholds \( \sigma_{\text{disp}} \) and \( \sigma_{\text{nuc}} \)
2: **Output:** Nuance Mask \( M_{\text{nuc}} \), Disparity Mask \( M_{\text{disp}} \)
3: \( f = \text{BN}(\text{GAP}(F)) \)
4: \( S = \text{softmax}(f) \)
5: \( \text{sort}(S) \)
6: \( M_{\text{disp}} = \text{zeros}(1, h, w), M_{\text{nuc}} = \text{zeros}(1, h, w) \)
7: **for** \( b = 1 \) to \( BS \) **do:**
8: \( m_{\text{disp}} = \text{zeros}(1, h, w), m_{\text{nuc}} = \text{zeros}(1, h, w) \)
9: **for** \( s = S_0 \) to \( s = S_{\theta_{\text{nuc}}} \) **do:**
10: \( IIF_{ij}^{s} = \sum_{k=1}^{C} W_{s,k} \cdot F_{ij,k} \)
11: \( m_{\text{nuc}} = m_{\text{nuc}} + IIF_{ij}^{s} \)
12: **end for**
13: \( m_{\text{nuc}} = \text{reshape}(m_{\text{nuc}}, 1, 1 \times h \times w) \)
14: \( \text{Idx} = \text{top}(m_{\text{disp}} \times w \times \sigma_{\text{nuc}}) \)
15: \( m_{\text{nuc}} = 0, m_{\text{nuc}}^{\text{idx}} = 1 \)
16: **for** \( s = S_{\theta_{\text{disp}}} \) to \( s = S_{-1} \) **do:**
17: \( IIF_{ij}^{s} = \sum_{k=1}^{C} W_{s,k} \cdot F_{ij,k} \)
18: \( m_{\text{disp}} = m_{\text{disp}} + IIF_{ij}^{s} \)
19: **end for**
20: \( m_{\text{disp}} = \text{reshape}(m_{\text{disp}}, 1, 1 \times h \times w) \)
21: \( \text{Idx} = \text{top}(m_{\text{disp}} \times w \times \sigma_{\text{disp}}) \)
22: \( m_{\text{disp}} = 1, m_{\text{disp}}^{\text{idx}} = 0 \)
23: \( \text{concat}(M_{\text{disp}}, m_{\text{disp}}) \) along \( b \)-axis
24: \( \text{concat}(M_{\text{nuc}}, m_{\text{nuc}}) \) along \( b \)-axis
25: **end for**

### 3.3. Relation Attention Module

Introducing the D&N masks, we adeptly extracted feature representations from both the most and sub-discriminative regions of vehicles, thereby amalgamating these elements to forge a holistic representation of the vehicle’s entirety. However, an intrinsic challenge arises due to the immutable rigid geometric structure of vehicles, which manifests as pronounced disparities in images of identical vehicle IDs when viewed from divergent perspectives. This phenomenon necessitates the identification of consistent elements that persist across varied photographic angles, focusing particularly on the inter-relationships among the vehicle’s components. Consequently, our goal is to unearth invariant features across different photographic angles of the same vehicle, specifically focusing on the inter-relations among vehicle components. To address this, we introduce the ra-module, a framework devised to capture these consistent relationships.

The transformer architecture, renowned for its exceptional capability in extracting relational information within sequences, has found extensive applications in the field of Computer Vision. Classification networks such as \cite{42} incorporate a class token to serve as
the feature vector input for classifiers. Detection and segmentation networks utilize the decoder module to restore the resolution of feature maps altered by the encoder, facilitating detailed pixel-level outcomes. Diverging from these approaches, the re-identification task necessitates the derivation of a feature vector that is not only succinct but also exhibits heightened generalizability to unseen classes.

First of all, an LA layer is applied to the feature map of the network’s final layer to extract more condensed regional information. Contrary to the traditional transformer approach, which partitions an image into fixed-size patches, our method divides the feature map into $1 \times 1 \times c$ units. Subsequently, an LA composed of parameter-sharing $1 \times 1$ convolution and batch normalization (BN) is employed to extract information from these segments:

$$\text{Units} = \{F_{ij} | F_{ij} \in \mathbb{R}^{1 \times 1 \times c}, \forall i \in \{1, \ldots, h\}, \forall j \in \{1, \ldots, w\}\}$$  \hspace{1cm} (4)

where $F_{ij}$ represents the unit at position $(i, j)$ in the original feature map $F$. Each $F_{ij}$ is a vector of length $c$, encapsulating the channel values at that specific spatial location within $F$.

Then, the $Q$, $K$, and $V$ vectors are generated by

$$Q = \text{Proj} (\text{Concat}^0 (\text{BN} (\text{conv} (\text{Unit})))) W^Q$$  \hspace{1cm} (5)

$$K = \text{Proj} (\text{Concat}^0 (\text{BN} (\text{conv} (\text{Unit})))) W^K$$  \hspace{1cm} (6)

$$V = \text{Proj} (\text{Concat}^0 (\text{BN} (\text{conv} (\text{Unit})))) W^V$$  \hspace{1cm} (7)

where $W^Q \in \mathbb{R}^{d_{RA} \times d_k}$, $W^K \in \mathbb{R}^{d_{RA} \times d_k}$, and $W^V \in \mathbb{R}^{d_{RA} \times d_V}$. The notation $\text{Concat}^0$ denotes the concatenation of matrices along the $0$th dimension. Then, self-attention and multi-head attention are calculated as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$  \hspace{1cm} (8)

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O$$  \hspace{1cm} (9)

where $W^O \in \mathbb{R}^{hd_v \times d_{LA}}$. The length of the spliced vector is $hd_v$, and the length of each vector in $\text{MultiHead}(Q, K, V)$ is $d_{LA}$. Therefore, the shape of features remains unchanged after the self-attention process. The construction details of RA module are shown in Figure 4. In the framework, we add the relation module after the backbone.

As shown in Figure 5, the relation attention module consists of $L$ self-attention blocks. Each block has multi-head and layer normalization operations. The calculation process can be represented by

$$F_0 = F$$  \hspace{1cm} (10)

$$F'_l = \text{MultiHead}(\text{LayerNorm}(F_{l-1})) + F_{l-1}$$  \hspace{1cm} (11)

$$F_l = \text{MLP}(\text{LayerNorm}(F'_l)) + F'_l$$  \hspace{1cm} (12)
Figure 5. The network architecture of RANDnet incorporates disparity mask and nuance mask. (a) Details of N&D masks. During the testing phase, the representation vector of a vehicle is composed of the three vectors. (b) Details of RA Module.

3.4. Framework

In this section, we describe the architecture of our proposed relation attention with the nuance–disparity masks network (RANDnet). We adopt IBNNet [14] as our backbone model. Two pivotal components, the N&D masks and the RA (Region Attention) module, are integrated as branch networks into the backbone architecture, as shown in Figure 5. The feature vectors extracted from these two branches are concatenated to generate the final image representation.

We implement a weight matrix by N&D masks on the feature map before the GAP operation. In order to obtain a concrete importance score for the feature weight, a parameter $\text{impact}$ is introduced.

Moreover, an $h \times w$ tensor is defined as the weight matrix. The calculation procedure can be expressed as

$$F'_{\text{last}} = F_{\text{last}} \odot (\text{impact} \times \text{Masks} + 1)$$

where operation $\odot$ denotes the Hadamard product. $F'_{\text{last}}$ represents the refined 3D feature tensor, which is then fed into the GAP layer. Let $F_{\text{last}}$ denote the final feature map produced by the backbone network. The hyper-parameter $\text{impact}$ is employed to modulate the extent to which the N&D masks influence this feature map.

In the RANDnet framework, the backbone network processes the input vehicle image to produce a 3D feature map, which is subsequently relayed to each branch. Within each branch, an FC layer is employed to compute the cross-entropy loss, treating the re-identification task as a classification challenge. The cross-entropy loss is formulated as follows:

$$\text{loss}_{\text{CE}} = -\sum_{i=1}^{\text{BS}} y_i^T \log(\hat{y}_i),$$

where $\text{BS}$ denotes the batch size utilized during training. Here, $y_i$ represents the one-hot encoded vector corresponding to the ground truth label of the image $I_i$, and $\hat{y}_i$ signifies the predicted probability distribution across all categories for the given branch.

In addition to the cross-entropy loss, triplet loss is employed to enhance the distribution of features. The function $\Theta(\cdot)$ symbolizes the deep learning model, serving as a transformation from raw images to feature vectors. The triplet loss is defined as

$$\text{loss}_{\text{Tri}} = \max(d(\Theta(I_a), \Theta(I_p)) - d(\Theta(I_a), \Theta(I_n)) + \text{margin}, 0),$$

where $d(\cdot, \cdot)$ denotes the distance function.
where \( I_a \) is the anchor image, \( I_p \) is a positive instance sharing the same label as \( I_a \), and \( I_n \) is a negative instance with a differing label. These instances are selected from a training batch following a strategy aimed at mining hard examples. The function \( d(\cdot) \) calculates the \( L_2 \) distance, and the inclusion of a margin promotes discriminative learning between positive and negative pairs.

The cumulative loss is expressed as

\[
\text{loss} = \text{loss}_{CE} + \alpha \cdot \text{loss}_{Tri},
\]

(16)

where \( \alpha \) is a hyper-parameter that adjusts the relative influence of the two losses. The losses from branch-1 and branch-2 are denoted as \( \text{loss}_1 \) and \( \text{loss}_2 \), respectively. Thus, the overall loss for RANDnet is calculated by

\[
\text{loss}_{RANDnet} = \text{loss}_1 + \beta \cdot \text{loss}_2,
\]

(17)

introducing \( \beta \) as a mechanism to modulate the gradient flow proportionally between the branches during backpropagation, thereby optimizing the learning dynamics of the network.

4. Experiments

In this section, we initially delineate the evaluation datasets, metrics, and experimental setup. Subsequently, we analyze the introduced parameters and conduct an ablation study. Finally, a comparative analysis with state-of-the-art methodologies is presented.

4.1. Datasets and Metrics

For evaluation purposes, experiments are carried out on three datasets, namely VeRi-776 [43] and VehicleID [44], which are pivotal in vehicle re-identification research.

The VeRi-776 dataset [43] comprises 51,035 images representing 776 vehicles captured by 20 cameras. Specifically, the training set consists of 37,778 images of 576 vehicles, while the testing set includes 13,257 images of 200 vehicles. The testing set is further divided, where 1678 images form the query set and the remaining serve as the gallery. Each probe image in the query set corresponds to multiple images with the same identity in the gallery set.

The VehicleID dataset [44] contains 221,763 images of 26,267 vehicles, with 110,178 images of 13,134 vehicles designated for training and the remainder for testing. The test set is segmented into three subsets of varying sizes: small, medium, and large, containing 800, 1600, and 2400 vehicles, respectively. Each query image is matched with only one image in the gallery. Unlike VeRi-776, the VehicleID dataset predominantly features two vehicle orientations: front and back.

To quantitatively assess the efficacy of our proposed method, we employ two standard metrics: mean Average Precision (mAP) and Cumulative Match Curve (CMC). While both indicators effectively reflect the model performance, each has its distinct emphasis in a re-identification task. The mAP metric is particularly well-suited for an \( n \) vs. \( n \) scenario, where each probe image corresponds to multiple matching images in the gallery, as exemplified by the VeRi-776 dataset. Conversely, the CMC metric is more applicable to an \( n \) vs. one scenario, characterized by the presence of only one matching image in the gallery, typical of datasets like VehicleID.

4.2. Experiments Setup

(1) Backbone Configuration.

In our study, we employ IBN-Net-50 [14] as the backbone architecture, a variation of ResNet-50. IBN-Net-50, while retaining a similar structure to ResNet-50, integrates instance normalization with batch normalization in certain layers, enhancing feature discriminability with minimal computational overhead. Moreover, we adjust the stride of the final pooling layer to one to yield larger feature maps. Our model is implemented using the Pytorch deep learning framework.
(2) Data Augmentation.
We incorporate random erasing and horizontal flipping as part of our data augmentation strategy during the training phase. Adhering to the optimal parameters specified in [45], tailored for image classification tasks, we set $s_l = 0.02$, $s_h = 0.4$, and $r_1 = 0.3$. The area of the erasure rectangle is randomly chosen from the range $(s_l, s_h)$, and $r_1$ governs the aspect ratio of this rectangle. Random erasing is applied with a probability of 0.5. Similarly, horizontal flipping is executed with a 0.5 probability.

(3) Training and Evaluation.
Input images are resized to $256 \times 256$ in both the training and testing phase. The batch size for training is configured to 64. The distribution of images per category within each batch is adjusted based on the specific characteristics of the dataset being used. For instance, in the VeRi-776 dataset, the average number of images per vehicle identity is approximately 75.6, resulting in the selection of 16 random images for each vehicle identity per batch. In contrast, for the VehicleID dataset, each category is represented by four images within a batch.

The training process spans over 120 epochs to ensure the underlying re-identification network stabilizes and learns effectively. Disparity masks remain dormant until the 60th epoch, activating only in the latter half of the training to refine the model’s focus on intricate details. Nuance masks, designed to enhance the classification accuracy by emphasizing distinct features, are triggered after the 80th epoch. Simultaneously, the RA module begins to leverage the regional features highlighted by these masks, enriching the model’s understanding and representation of vehicle identities.

Stochastic Gradient Descent (SGD) is employed as the optimization method, starting with an initial learning rate of 0.01. The first 10 epochs incorporate a warm-up strategy, where the learning rate linearly increases from 0.0001 to 0.01, preparing the model for more intense training phases. Post the 30th epoch, the learning rate is methodically reduced to $7.7 \times 10^{-5}$ following a cosine annealing schedule, optimizing the training process by adjusting the learning rate in a smooth, cyclical pattern.

A batch normalization layer follows the GAP layer in each branch of the network, ensuring the consistent normalization of the features. For evaluation purposes, the output features of the batch normalization layer are utilized as the branch output features, which contribute to the stability and performance of the model during the testing phase.

4.3. Parameter Analysis
In this section, we delve into the analysis of key parameters integral to our method, focusing on two aspects: the visualization of discriminative region identification and the accuracy in re-identification tasks. All experiments referenced here are conducted on the VeRi-776 dataset.

(1) N&D Masks.
To accurately and reliably identify discriminative region masks, we advocate the use of multiple spatial attention maps. We experimented with various values of $\theta_{\text{disp}}$, visualizing the discriminative regions as shown in Figure 6. Empirical results are presented for $\theta_{\text{disp}} \in \{1, 3, 5, 10, 50, 100, 200\}$ and with $\sigma_{\text{disp}} = 0.5$. As depicted, an increase in $\theta_{\text{disp}}$ leads to more precise and focused recognition of discriminative regions. Specifically, regions identified with $\theta_{\text{disp}} = 200$ are densely concentrated on distinct vehicle areas, such as headlights and roofs, compared to $\theta_{\text{disp}} = 1$. The variations between $\theta_{\text{disp}} = 50$, $\theta_{\text{disp}} = 100$, and $\theta_{\text{disp}} = 200$ are subtly discernible. Consequently, we set $\theta_{\text{disp}}$ to 200 for subsequent experiments.

The variations between $\theta_{\text{disp}} = 50$, $\theta_{\text{disp}} = 100$, and $\theta_{\text{disp}} = 200$ are subtly discernible. Regarding the parameter $\sigma_{\text{disp}}$, which controls the area of the rigid body in the image, we consider the vehicle’s proportion within the image. Extreme proportions, i.e., too small or too large, are deemed impractical and are therefore excluded. Empirical tests with $\sigma_{\text{disp}} \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and $\theta_{\text{disp}} = 200$ were conducted. The visual results, displayed in Figure 7, clearly show that as $\sigma_{\text{disp}}$ increases, the mask progressively encompasses the
entire vehicle body. The impact on re-identification is illustrated in Figure 8a, where the X-axis and Y-axis represent the proportion $\sigma_{\text{disp}}$ and the mAP evaluated on VeRi-776 [43], respectively. Initially, an increase in $\sigma_{\text{disp}}$ correlates with a rise in the mAP peaking at $\sigma_{\text{disp}} = 0.7$. However, further escalation in $\sigma_{\text{disp}}$ leads to a decline in accuracy. This is attributed to the fact that a smaller $\sigma_{\text{disp}}$ challenges the N mask’s ability to learn all details, while a larger $\sigma_{\text{disp}}$ risks including irrelevant background elements in the D mask.

A similar experiment was conducted to identify the optimal values for $\theta_{\text{nuc}}$ and $\sigma_{\text{nuc}}$. The candidate values for $\theta_{\text{nuc}}$ were set to $\{2, 3, 5, 10, 20\}$, selected based on the frequency of each vehicle type within the dataset. Experimental results indicate that the best mAP performance for nuance masks is achieved when $\theta_{\text{nuc}}$ is set to five. Correspondingly, $\sigma_{\text{nuc}}$, the parameter for nuance masks, is determined to be 0.3 based on these outcomes.

![Figure 6. Parameter analysis of $\theta_{\text{disp}}$ in the setting $\sigma_{\text{disp}} = 0.5$. The left column is original image, and right columns show the effect of visualized results with increasing $\theta_{\text{disp}}$. When more IIFs are involved, recognized discriminative regions become more accurate and compact.](image)

![Figure 7. Parameter analysis of $\sigma_{\text{disp}}$ in the setting $\theta_{\text{disp}} = 200$. It can be seen that, with the increase in $\sigma_{\text{disp}}$, discriminative regions cover more and more vehicle details.](image)

The parameter impact signifies the attention weight assigned to the vehicle body during training. We executed experiments with impact $\in \{0.1, 0.2, 0.3\}$, keeping upper parameters remaining. The experimental outcomes are illustrated in Figure 8a. Observations reveal that N&D masks attain the highest mAP when impact is set to 0.2.
(a) Parameter analysis of impact and $\sigma_{\text{disp}}$. The value of impact is represented by different colors, and X-axis denotes the proportion $\sigma_{\text{disp}}$. Experiments are all evaluated on the VeRi dataset. The optimal values are $\text{impact} = 0.2$ and $r = 0.7$, respectively.

(b) The distance statistics between vehicles with the same ID but different orientations were analyzed both before and after the introduction of the RA module. Post the incorporation of the RA module, the distribution of distances is situated at a lower level.

**Figure 8.** Numerical analysis of RANDnet components.

Drawing from this analysis, we identified the optimal values for the key parameters. All subsequent experiments are conducted under these settings.

(2) RA module.

The hyper-parameters associated with the relation attention module encompass the loss function weight $\beta$, the number of self-attention modules $L$, and the feature vector length $d_{\text{RA}}$. To ascertain suitable values for these parameters, we devised a framework akin to that illustrated in Figure 5, featuring two branches: one retains the baseline structure while the other incorporates the relation attention module. This framework serves as the basis for our analysis of the parameters in the relation attention module.

The re-identification accuracy exhibits sensitivity to the weight of the loss function. The model’s framework, composed of dual branches, shares the same backbone network. The loss function weight essentially represents the gradient proportion of the two branches relative to the backbone network. Disproportionate gradients from the branches can disrupt the training of the backbone network if the weight setting is imbalanced. We tested five distinct values of $\beta$, as tabulated in Table 1: 0.01, 0.05, 0.1, 0.2, and 1.0. With $\beta = 0$, the branch with the relation attention module ceases gradient contribution, reducing the overall framework to the baseline approach. Optimal performance is observed at $\beta = 0.1$, with the mAP and CMC@1 reaching 81.6% and 96.3%, respectively.

In addition, we conducted experiments on the Veri776 dataset to investigate the impact of the weighting of triplet loss relative to the overall loss. The experimental results are depicted in Figure 9. Based on the obtained curve results, we determined the weighting of triplet loss to be one, indicating an equal weight allocation between triplet loss and cross-entropy classification loss, with a ratio of 1:1.

**Figure 9.** The weighting of triplet loss represents the ratio between triplet loss and cross-entropy loss. When the weight equals 1.0, the CMC results are relatively optimized.
Results from the parameter analysis concerning the number of self-attention modules $L$ and feature length $d_{RA}$ are presented in Figure 10a,b. The chosen values for $L$ in our experiments are 1, 2, 3, and 4, while for $d_{RA}$, they are 128, 256, and 512. These experiments reveal that the highest accuracy is achieved with $L = 1$, and the impact of varying $d_{RA}$ values appears to be inconclusive.

### Table 1. Parameter analysis of $\beta$.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>mAP</th>
<th>CMC@1</th>
<th>CMC@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.816</td>
<td>0.957</td>
<td>0.983</td>
</tr>
<tr>
<td>0.01</td>
<td>0.817</td>
<td>0.954</td>
<td>0.985</td>
</tr>
<tr>
<td>0.05</td>
<td>0.830</td>
<td>0.964</td>
<td>0.985</td>
</tr>
<tr>
<td>0.10</td>
<td><strong>0.842</strong></td>
<td>0.981</td>
<td><strong>0.995</strong></td>
</tr>
<tr>
<td>0.20</td>
<td>0.824</td>
<td><strong>0.982</strong></td>
<td>0.986</td>
</tr>
<tr>
<td>1.00</td>
<td>0.818</td>
<td>0.973</td>
<td>0.985</td>
</tr>
</tbody>
</table>

Bold numbers indicate the best results.

Figure 10. Experimental results of parameter analysis on $L$ and $d_{LA}$.

Consequently, the parameter configuration for subsequent experiments is set as follows: $\beta = 0.1$, $L = 1$, and $d_{RA} = 256$.

### 4.4. Ablation Study

In this section, we present a series of ablation studies designed to assess the performance enhancements brought forth by our method. Utilizing the VeRi-776 dataset, the results of these studies are documented in Table 2.

(1) Effectiveness of the N&D masks.

The results showcased in Table 2 reveal that our N&D masks register an mAP of 83.6%, thereby achieving a notable improvement of 1.7% over the established baseline. Unlike the baseline model, which tends to be misled by irrelevant background objects such as guardrails and green belts, the model incorporating the N mask directs its focus more accurately onto the vehicle body, effectively ignoring these distractions. Furthermore, the N mask fosters a more refined attention distribution, whereas the D mask expands the focus areas to prevent overly concentrated responses, demonstrating the different benefits of each mask.

### Table 2. Ablation study.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>CMC@1</th>
<th>CMC@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ibn-net)</td>
<td>0.819</td>
<td>0.970</td>
<td>0.990</td>
</tr>
<tr>
<td>+ Disparity mask</td>
<td>0.834</td>
<td>0.976</td>
<td>0.991</td>
</tr>
<tr>
<td>+ Buance mask</td>
<td>0.818</td>
<td>0.963</td>
<td>0.988</td>
</tr>
<tr>
<td>+ N&amp;D masks</td>
<td>0.836</td>
<td>0.977</td>
<td>0.992</td>
</tr>
<tr>
<td>+ RA module</td>
<td>0.837</td>
<td>0.979</td>
<td>0.992</td>
</tr>
</tbody>
</table>
We conducted a visual analysis of the sorting results to demonstrate the effectiveness of our method. In Figure 11, we present several examples where the re-identification results were poor in the baseline. Compared to the baseline, the use of the N&D mask allows the model to suppress interference from background and foreground noise, as illustrated in the comparison of the results in the first two rows of the figure. Additionally, by enhancing focus on less discriminative regions, the model can adaptively mine key components related to the ID, such as the sticker on the windshield of the minivan and the roof of the cargo truck, as shown in the third and fourth rows.

Figure 11. Visualization of ranking list on vehicle ReID task. The image on the left is the query image and the red/blue boxes indicate right/wrong results.

We further conducted a visual analysis of the feature maps, as shown in Figure 12. Through the implementation of the D mask, the network's focus is centralized on regions with the highest ID discriminability, rather than on the background. However, this approach results in an excessively small area of attention. By applying the N mask, the network expands its focus to include sub-discriminative regions, thereby enhancing the overall effectiveness of the identification process.

Figure 12. The visualization of the class activation areas learned by the network is displayed, where the second column shows the regions learned by the baseline, the third column presents those learned after applying the D Mask, and the fourth column illustrates the regions discerned post-application of the N&D Mask. This representation provides a comparative insight into the model’s learning progression and the effectiveness of the employed masking techniques.
(2) Effectiveness of the Relation Attention Module.

Table 2 indicates that incorporating the relation attention module elevates the model’s mAP to 81.6% and CMC@1 to 96.3%, outstripping the baseline by margins of 2.3% and 0.6%, respectively. These enhancements underscore the module’s capacity to facilitate the model’s understanding of the interconnections between different vehicle components, thereby significantly boosting the re-identification precision.

Additionally, to gauge the model’s resilience to variations in vehicle pose, we curated a validation dataset comprising 517 manually selected vehicle image pairs from VeRi-776, each pair showcasing different angles of the same vehicle. By calculating the cosine distance between the embeddings of each image pair, we were able to evaluate the model’s pose adaptability. The results, plotted in Figure 8b, with the X-axis representing the cosine distance and the Y-axis denoting the count of image pairs, demonstrate that models augmented with positional embedding consistently exhibit narrower cosine distances when compared to the baseline. This indicates a substantial enhancement in the model’s ability to adapt to different vehicle poses.

To conclude, our findings confirm that each introduced module not only enriches the model’s feature representation but also significantly fortifies its capability to extract more resilient features, thereby contributing to the overall efficacy of the model.

4.5. Comparison with State-of-the-Art

We compare our method with recent state-of-the-art methods, including RAM [20], VAMI [16], AAVER [46], PRN citepart-regularized, SAVER [25], PGAN [47], PVEN [22], and FDA-Net [48].

Experiments are conducted on three large-scale public datasets. On VeRi-776 [43], mAP, CMC@1, and CMC@5 are adopted to evaluate the performance quantitatively, and on VehicleID [44], we apply CMC@1 and CMC@5. Table 3 presents the comparison on VeRi-776 [43] and VehicleID [44].

Table 3. Results comparison on VeRi-776 and VehicleID.

<table>
<thead>
<tr>
<th>Method</th>
<th>VeRi-776</th>
<th>VehicleID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td>CMC@1</td>
</tr>
<tr>
<td>VAMI [16]</td>
<td>0.501</td>
<td>0.770</td>
</tr>
<tr>
<td>AAVER [46]</td>
<td>0.612</td>
<td>0.890</td>
</tr>
<tr>
<td>EALN [49]</td>
<td>0.574</td>
<td>0.844</td>
</tr>
<tr>
<td>RAM [20]</td>
<td>0.615</td>
<td>0.886</td>
</tr>
<tr>
<td>PRN [19]</td>
<td>0.743</td>
<td>0.943</td>
</tr>
<tr>
<td>SAVER [25]</td>
<td>0.796</td>
<td>0.964</td>
</tr>
<tr>
<td>PGAN [47]</td>
<td>0.793</td>
<td>0.965</td>
</tr>
<tr>
<td>PVEN [22]</td>
<td>0.795</td>
<td>0.956</td>
</tr>
<tr>
<td>TransReID [50]</td>
<td>0.823</td>
<td>0.971</td>
</tr>
<tr>
<td>DFNet [12]</td>
<td>0.810</td>
<td>0.971</td>
</tr>
<tr>
<td>SGFD [51]</td>
<td>0.810</td>
<td>0.967</td>
</tr>
<tr>
<td>VAL [52]</td>
<td>0.814</td>
<td>0.967</td>
</tr>
<tr>
<td>PAMNet [53]</td>
<td>0.816</td>
<td>0.965</td>
</tr>
<tr>
<td>Fast [54]</td>
<td>0.819</td>
<td>0.970</td>
</tr>
<tr>
<td>HRCN [55]</td>
<td>0.831</td>
<td>0.973</td>
</tr>
<tr>
<td>VNet [8]</td>
<td>0.834</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Baseline(ibrn-50) | 0.819 | 0.970 | 0.990 | 0.866 | 0.979 | 0.829 | 0.960 | 0.806 | 0.939 |
| Ours     | 0.835    | 0.977     | 0.990     | 0.885  | 0.987  | 0.833  | 0.965  | 0.810  | 0.945  |

Bold numbers indicate the best results.

Results on VeRi-776. Table 3 presents the comparison on VeRi-776 [43]. Our approach, RANDnet, achieves a mean Average Precision (mAP) of 83.5% and a Cumulative Matching Characteristic at Rank 1 (CMC@1) of 97.7%. These results demonstrate that our method
significantly surpasses other state-of-the-art techniques by a considerable margin. Notably, there is a substantial improvement in the mAP, whereas the gains in CMC@1 and CMC@5 are relatively modest. This indicates a strong overall performance, particularly in accurately matching the top-ranked vehicle identities. We attribute this phenomenon to the fact that incorporating additional secondary discriminative features assists the network in elevating the accuracy of re-identifying targets belonging to similar vehicle models. However, the impact on differentiation between two categories that are extremely similar is relatively limited.

**Results on VehicleID.** The evaluation results are shown in Table 3. Our method achieves 88.5%, 83.3%, and 81.0% in CMC@1 on three test sets. Meanwhile, RANDnet achieves 98.7%, 96.5%, and 94.5% in CMC@5. It is clear that our method outperforms other recent works in almost all metrics. Our method demonstrates a substantial improvement in the Rank 1 metric on this dataset. We attribute this enhancement to the fact that the VehicleID dataset comprises only two directional viewpoints—frontal and rear views. This circumstance allows our relation attention (RA) module to effectively mitigate feature distance disparities between different viewpoints.

5. Conclusions

Our study introduces a pioneering vehicle ReID network that integrates relation attention mechanisms with nuance and disparity masks to effectively address the challenges of background and foreground occlusions, as well as the identification of multiple discriminative regions without the need for extra labeling. By capitalizing on the differentiation between simple and hard negative classes during training, our model is able to identify irrelevant background features and focus on key discriminative areas, such as emblems, headlights, or tires. The application of our relation attention module, which utilizes an advanced transformer architecture for self-attention on vehicle component information, enables our network to significantly reduce intra-class distances for the same vehicle ID across different viewpoints. This approach does not require any data beyond vehicle IDs for training, simplifying the process while ensuring robustness. Our proposed method reaches the state-of-the-art (SOTA) performance on the commonly used vehicle re-identification datasets VERI-776 and VehicleID, achieving results of 0.835 mAP and 0.885 Top1 accuracy, respectively, showcasing its potential to significantly advance the field of vehicle re-identification.

**Author Contributions:** Methodology, Y.H.; writing—original draft preparation, Y.H.; writing—review and editing, H.S. and W.K.; visualization, Y.H.; supervision, H.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is partially supported by the National Key RD Program of China (No.2022YFB3306500), the National Natural Science Foundation of China (No.62394332), and the Open Fund of the State Key Laboratory of Software Development Environment (No.SKLSDE-2023ZX-11) and the Haiyou Plan Fund.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The VeRi-776 dataset used in this paper was provided by Xinchen Liu and accessed under the necessary permissions. The dataset’s homepage can be found at [https://vehicleraid.github.io/VeRi/](https://vehicleraid.github.io/VeRi/). The VehicleID dataset utilized in this study was provided by Hongye Liu, also accessed under the appropriate permissions, with its homepage located at [https://www.pkuml.org/resources/pku-vehicleid.html](https://www.pkuml.org/resources/pku-vehicleid.html). Interested researchers are required to contact the data providers directly and may need to sign a usage agreement.

**Acknowledgments:** Thank you for the support from HAWKEYE Group.

**Conflicts of Interest:** Author Hao Sheng was employed by the company Sinenux. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
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.