sensors

MDPI

Article

Progressively Hybrid Transformer for Multi-Modal Vehicle Re-Identification

Wenjie Pan [1], Linhan Huang [1], Jianbao Liang [1], Lan Hong [1,*] and Jianqing Zhu [1,2,*]

[1] College of Engineering, Huaqiao University, Quanzhou 362021, China
[2] Xiamen Yealink Network Technology Company Limited, No. 666, Hu'an Road, High-Tech Park, Huli District, Xiamen 361015, China
* Correspondence: hhhlam@hqu.edu.cn (L.H.); jqzhu@hqu.edu.cn (J.Z.)

Abstract: Multi-modal (i.e., visible, near-infrared, and thermal-infrared) vehicle re-identification has good potential to search vehicles of interest in low illumination. However, due to the fact that different modalities have varying imaging characteristics, a proper multi-modal complementary information fusion is crucial to multi-modal vehicle re-identification. For that, this paper proposes a progressively hybrid transformer (PHT). The PHT method consists of two aspects: random hybrid augmentation (RHA) and a feature hybrid mechanism (FHM). Regarding RHA, an image random cropper and a local region hybrider are designed. The image random cropper simultaneously crops multi-modal images of random positions, random numbers, random sizes, and random aspect ratios to generate local regions. The local region hybrider fuses the cropped regions to let regions of each modal bring local structural characteristics of all modalities, mitigating modal differences at the beginning of feature learning. Regarding the FHM, a modal-specific controller and a modal information embedding are designed to effectively fuse multi-modal information at the feature level. Experimental results show the proposed method wins the state-of-the-art method by a larger 2.7% mAP on RGBNT100 and a larger 6.6% mAP on RGBN300, demonstrating that the proposed method can learn multi-modal complementary information effectively.

Keywords: multi-modal image; transformer; vehicle re-identification

Citation: Pan, W.; Huang, L.; Liang, J.; Hong, L.; Zhu, J. Progressively Hybrid Transformer for Multi-Modal Vehicle Re-Identification. *Sensors* **2023**, *23*, 4206. https://doi.org/10.3390/s23094206

Academic Editor: Omprakash Kaiwartya

Received: 9 February 2023
Revised: 5 April 2023
Accepted: 17 April 2023
Published: 23 April 2023

1. Introduction

The aim of vehicle re-identification (ReID) [1–3] is to retrieve a specific vehicle image from a large-scale vehicle gallery captured by non-overlapping cameras, which receives a lot of attention from the artificial intelligence research field due to its significant role in intelligent transportation systems for building smart cities. Most existing vehicle ReID methods [4–16] are only based on single-modal visible images, i.e., RGB images, which would suffer from weak performance because of the poor imaging quality under low light environments.

To overcome low illumination conditions, Li et al. [17] firstly proposed using three-modal (i.e., visible, near-infrared, and thermal-infrared) images for vehicle ReID, and constructed a vehicle ReID benchmark that shows that three-modal vehicle ReID greatly improves accuracy in low illumination conditions. Although a non-visible spectrum could show good night imaging results to play good complements to visible images, different spectra have different imaging characteristics, which could be a challenge even to a strong global feature modeling model [16]. As shown in Figure 1, the contrast between the foreground (i.e., vehicles) and background in near-infrared images is lower than that in visible images. Visible images have a stronger ability to reflect texture detail information of vehicles than near-infrared images in the daytime. Thermal-infrared images contain more noise than visible and near-infrared images. As a result, although non-visible images have great potential to boost vehicle ReID performance in low illumination environments, there is an open question in multi-modal ReID in practice: how to effectively fuse the complementary information from multi-modal data?

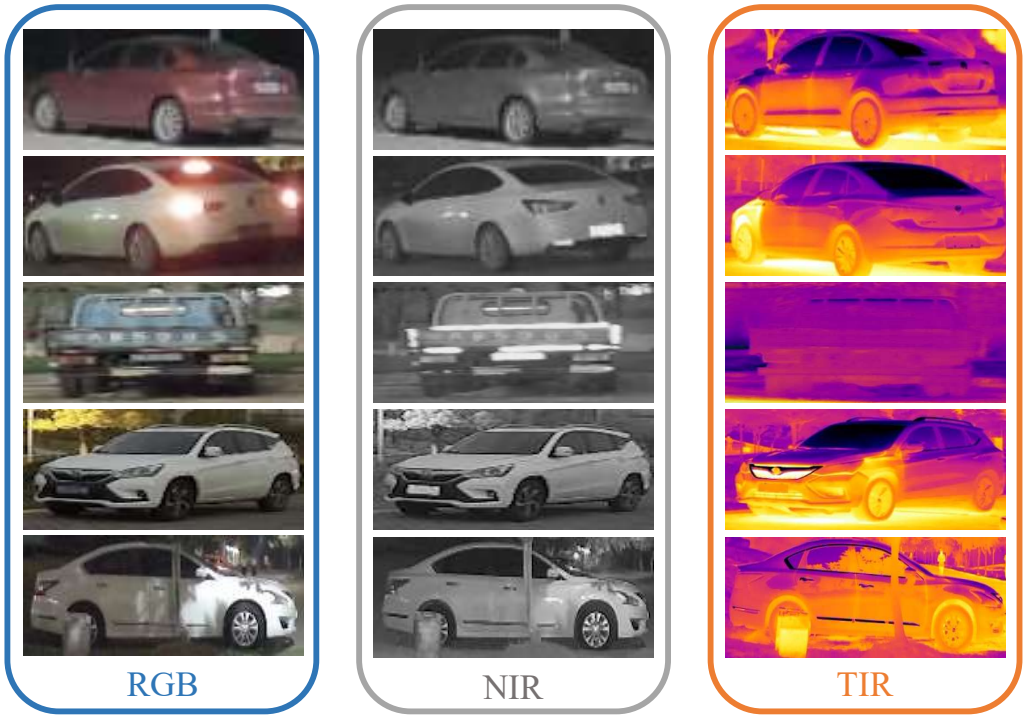

Figure 1. Multi-modal vehicle image examples. Here, RGB, NIR, and TIR are abbreviations for visible, near-infrared, and thermal-infrared, respectively.

Existing multi-modal vehicle Re-ID [17–20] most focus on learning modal robust features. For example, Wang et al. [20] designed a cross-modal interacting module and a relation-based embedding module to exchange useful information from multi-modal features so as to enhance features' richness. Both cross-modal interacting and relation-based embedding modules are convolutional neural network (CNN) branches. Zheng et al. [19] proposed a cross-directional consistency network to mitigate cross-modal discrepancies and adjust individual feature distributions for learning modal robust features. Li et al. [17] proposed a heterogeneity collaboration aware multi-stream convolutional neural network to constrain scores of different instances of the same identity to be coherent. Guo et al. [21] proposed a generative and attentive fusion network to fuse and align features of the original data. Although they have acquired great progress for multi-modal vehicle ReID, there is still room for designing an effective multi-modal fusion manner to improve multi-modal vehicle ReID. Specifically, there are two reasons for emphasizing multi-modal fusion. First, current multi-modal vehicle ReID works [17–22] are based on CNNs that use local kernels having a limited receptive field, which is inadequate in fusing global features of multi-modal data. Hence, this paper designs a multi-modal hybrid transformer to use the transformer's long-distance dependency learning ability to realize a global feature fusion of multi-modal data. Second, current multi-modal vehicle ReID methods only pay attention to the feature level fusion, and the image level fusion is underestimated. Therefore, this paper proposes a random hybrid augmentation to fuse multi-modal complementary information at the image level. Consequently, combing the multi-modal hybrid transformer and the random hybrid augmentation, a progressively hybrid transformer is constructed in this paper, which fuses multi-modal complementary information at both image and feature levels.

The contributions of this paper are summarized as follows:

- This paper proposes a multi-modal hybrid transformer, which applies the feature hybrid mechanism (FHM) to fuse multi-modal information at the feature level by the modal-specific controller and modal information embedding.

- This paper designs a random hybrid augmentation (RHA) to fuse multi-modal information at the image level, which upgrades the multi-modal hybrid transformer into a progressively hybrid transformer (PHT) that fuses multi-modal information at both image and feature levels.
- Experimental results on RGBNT100 and RGBN300 demonstrate that the proposed PHT outperforms state-of-the-art methods.

This paper is an extended version of the preliminary work [23]. Compared with the preliminary work [23], this paper is improved in two aspects. (1) This paper designs a new data augmentation method (i.e., random hybrid augmentation) to form a more comprehensive multi-modal information fusion which outperforms [23] a larger 0.9% mAP on RGBNT100 and a larger 0.3% mAP on RGBN300. (2) This paper implements more experiments to analyze the proposed method. The rest of this paper is organized as follows. Section 2 contains recent works related to the proposed method. Section 3 describes the proposed method in detail. Section 4 presents experimental results and analysis to show the proposed method's advantage. Section 5 concludes this paper.

2. Related Works

2.1. Visible Re-Identification

Most of the existing vehicle re-identification methods are based on visible images and they have acquired great progress [6,8,9,24–28]. Several representative works are reviewed as follows. Zhu et al. [5] extracted the final similarity by using orientation and camera similarity as auxiliaries to alleviate the difficulty of similar appearances. Cai et al. [29] proposed a multi-level feature extracting approach to learn global features from whole vehicle images and learn local discriminative features from different local region channels. Meng et al. [7] proposed a part perspective transformation module to map the different vehicle parts into a unified perspective to deal with viewpoint variations. Zhou et al. [8] proposed a viewpoint-aware attentive multi-view inference model cooperating with visual information to handle viewpoint variations. Li et al. [27] proposed an efficient transformer to learn multi-view part-wise correlations to deal with complex viewpoint variations. Zeng et al. [30] proposed an illumination identity disentanglement (IID) network to dispel different scales of illumination away while maintaining each identity's discriminant information. Zhang et al. [31] proposed using an illumination teacher model trained by the differences between the illumination-adjusted and original images to separate the ReID features from lighting features to enhance ReID performance. Although low illumination promotes vehicle ReID, extremely unsatisfactory illumination conditions are still killers of vehicle ReID.

2.2. Deep Architecture

Thanks to the rapid development of deep learning, many excellent deep networks have emerged in computer vision research fields, which could be divided into two categories: (1) convolutional-based networks [32–40] and (2) vision transformer-based networks [41–49].

The first convolutional neural network (CNN) is proposed by LeCun [32], which shows an impressive performance for document recognition. Krizhevsky et al. [33] proposed the famous AlexNet via stacking more convolutional layers followed by max-pooling layers and fully connected layers, acquiring good results on the large-scale image classification benchmark [50]. Simonyan et al. [34] emphasized using more small convolutional kernels to construct a deeper VGG network. Szegedy et al. [35] first designed GoogLeNet with an inception structure utilizing sparse structure to achieve deep and wide networks. Ioffe et al. [36] designed a batch normalization layer playing in a convolution layer and an activation function to reduce internal covariate shifts to improve the training convergence of GoogLeNet. Furthermore, Szegedy et al. [37] explored factorizing convolutions with large kernels to avoid representational bottlenecks of inception structures of GoogLeNet. In addition to inception series, residual networks [39,51,52] are another popular family.

He et al. [39] firstly designed residual layers to effectively alleviate the problem of gradient vanishing, allowing for training ultra-deep networks, namely, residual networks (ResNet). Hu et al. [52] designed a squeeze-and-excitation (SE) block to learn channel-wise information to upgrade the ResNet to the SE-ResNet. Xie et al. [51] proposed ResNeXt by combining the residual layer and the inception structure. Szegedy et al. [38] also combined the inception structure and the residual layer to improve their networks.

More recently, vision transformer [49], known for its ability to learn global features from its self-attention mechanism, has done an impressive job in computer vision tasks. Wu et al. [53] proposed a pyramid pooling method to acquire a stronger multi-head self-attention that could more properly deal with multi-scale information. Zhang et al. [24] introduced a transformer-based feature calibration to integrate low-level detail information as a global prior for high-level semantic information. Chen et al. [54] proposed a structure-aware positional transformer network to utilize the structural and positional information and learn semantic-aware features. Especially, for the visible modal person/vehicle ReID task, He et al. [16] first proposed a pure transformer-based object ReID framework, which achieves state-of-the-art performance on most person/vehicle re-identification benchmarks.

2.3. Data Augmentation

Zhong et al. [55] proposed a data augmentation method to randomly select a rectangle region in an image and erase its pixel with a random value, which reduces the risk of over-fitting and makes a deep network robust to occlusions. The random patch method [56] firstly creates a patch pool of random image patches and then pastes a random patch from the patch pool onto an input image at a random position. Because [55,56] could heavily occlude images, Chen et al. [57] believed these two methods would harm the models' ability to mine salient local information, so they proposed soft random erasing, in which an erased area is not completely replaced with random pixels but also retains a proportion of the original pixels. Li et al. [58] combined different regions of different identities to generate virtual regional perceptual data pairs. Qjagh et al. [59] proposed a data preprocessing strategy to generate the missing data by average, maximum, and weighted average. Lin et al. [60] proposed an illuminate-aware data-augmentation method that estimates the illuminate distribution from the training data and generates synthesis images under different illumination. Huang et al. [61] designed an adversarial learning-based occlusion image generation method to enhance the person ReID model's generalization ability.

Considering these data augmentation methods perform well by introducing useful complementary information and the complementary information between different modalities is essential for multi-modal vehicle ReID, a random hybrid augmentation (RHA) method is designed to improve the previous work [23] in the fusion of the image level. Compared with the previous work, [23], which only fuses multi-modal information at the feature level, this paper fuses multi-modal information at both image and feature levels. Specifically, in addition to the multi-modal information fusion at the feature level, this paper fuses multi-modal information at the image level by exchanging information between different modalities at image regions with random positions, random numbers, random sizes, and random aspect ratios.

3. Methodology

Figure 2 shows the overall framework of the proposed progressively hybrid transformer (PHT), including (1) random hybrid augmentation (RHA) and (2) a feature hybrid mechanism (FHM)-based multi-modal hybrid transformer. RHA brings local structural characteristics of all modalities, mitigating modal differences at the beginning of feature learning. The FHM assigns the distribution of modal-specific layers to improve multi-modal feature fusion.

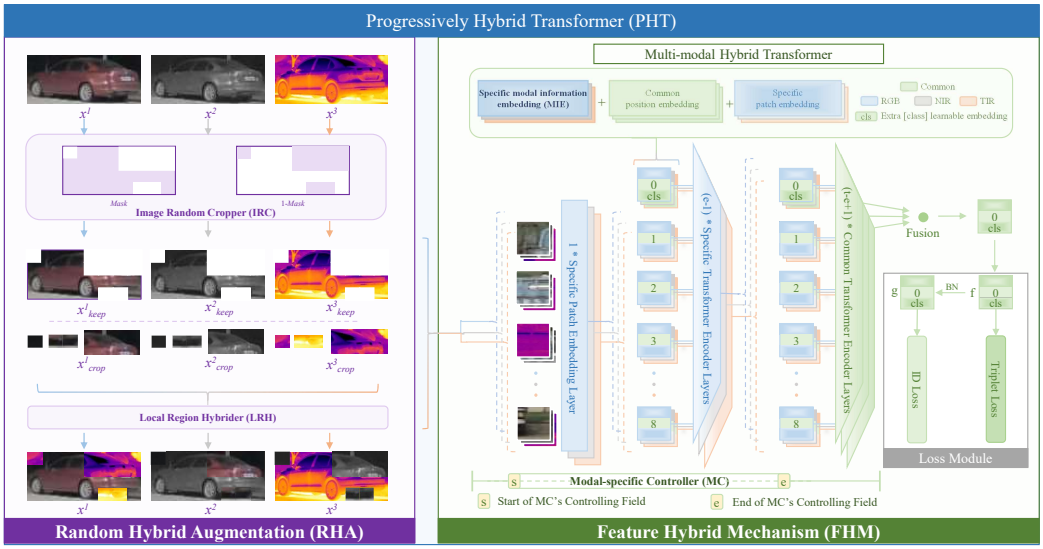

Figure 2. The overall framework of the proposed progressively hybrid transformer.

3.1. Random Hybrid Augmentation

As shown in Figure 2, the RHA has two processors: (1) a image random cropper (IRC) and (2) a local region hybrider (LRH). The IRC extracts multi-modal-specific information by simultaneously cropping multi-modal images of random positions, random numbers, and random sizes. The LRH captures multi-modal complementary information by fusing the cropped regions to let regions of each modal take local structural characteristics of multi-modalities.

Given a group of n-modal images $\{x^i \in \mathbb{R}^{H\times W}, i = 1, 2, \ldots, n\}$, where H and W denote the height and width of each modal image. For the convenience of description, the IRC is parameterized by n_{region} and p_{region}, which, respectively, denotes the max number of cropped regions and the max proportion of the cropped edge and the image original edge. As shown in Figure 2, the IRC's workflow is described as follows.

(1) Initializing a $H \times W$ sized *Mask* whose elements are equal to 1.
(2) Random zero setting $l \in [0, n_{region}]$ local regions of *Mask*, that is,

$$Mask(m,n) = \begin{cases} 0 & m,n \in \cup_{j=1}^{l} R_j, \\ 1 & \text{otherwise}, \end{cases} \tag{1}$$

where $m \in [1, H]$ and $n \in [1, W]$ are y-coordinate and x-coordinate, respectively; R_j is the j-th zero setting region that has a random aspect ratio and a random area. Please note that each zero setting region's max height and width are $H \times p_{region}$ and $W \times p_{region}$.

(3) Cropping each modal image as follows.

$$\begin{aligned} x^i_{crop} &= x^i \otimes (1 - Mask), \\ x^i_{keep} &= x^i \otimes Mask, \\ i &= 1, 2, ..., n, \end{aligned} \tag{2}$$

where $\otimes$ is element-wise multiplication operation; x^i_{crop} is the cropped part of the i-th modal image, and x^i_{keep} is the rest part that keeps unchanging.

Based on Equation (2), the LRH calculation is formulated as follows:

$$x^i = x^i_{keep} + Hybrid(x^1_{crop}, x^2_{crop}, ..., x^n_{crop}), \tag{3}$$

where *Hybrid* is the fusion function. In this paper, five types of fusion functions are designed. (1) The average method, which simply averages all modal cropped regions. (2) The self-excluding average, which first excludes cropped regions of its own modality and then averages cropped regions of all remaining modalities. Similarly, two Hadamard product versions are also designed, i.e., (3) the Hadamard product and (4) the self-excluding Hadamard product. (5) Randomly swapping, in which $\{x_{crop}^1, x_{crop}^2, ..., x_{crop}^n\}$ are stochastically scheduled and then each element is used to replace the cropped regions of a modality. Based on Equations (2) and (3), the RHA module could bring local structural characteristics of all modalities, reducing modal differences at the beginning of feature learning.

3.2. Feature Hybrid Mechanism-Based Multi-Modal Hybrid Transformer

As shown in Figure 2, this paper designs a multi-modal hybrid transformer, which is a multi-branch transformer simultaneously extracting features from multi-modal images. Each branch is a vision transformer proposed by [16,42], which consists of a patch embedding layer and a list of encoders. The patch embedding layer is responsible for mapping the image patch into a vector. The encoder is a combating of layer normalization and multi-head self-attention with residual connections to complex features of vectors generated by the patch embedding layer. Features from each branch are fused to form multi-modal features and are fed into the loss function for training. In this paper, three feature fusion methods are applied, i.e., (1) average, (2) Hadamard product, and (3) concatenation.

The multi-modal hybrid transformer only fuses multi-modal information at one and only one depth position. Hence, the feature hybrid mechanism (FHM) is proposed to improve the multi-modal hybrid transformer. The FHM has two modules: (1) modal-specific controller (MC), and (2) modal information embedding (MIE). The MC module is designed for allocating the modal-specific parts of vision transformer branches. The MIE module is designed to attach modal information to patch embeddings. The details of the MC and MIE are described as follows.

3.2.1. Modal-Specific Controller

The MC module assigns the sharing attribute of three structures, i.e., (1) position embedding, (2) patch embedding layers, and (3) encoders. For the position embedding, the MC module default set the position embedding to be modal-common, considering that spatial position information is more likely to be modal independent.

For patch embedding layers and encoders, the MC module can flexibly assign common or specific attributes with a modal-specific controlling field and the number of modal-specific layers. The modal-specific controlling field is denoted as $v = [s, e)$, where s and e are natural numbers, and the number of modal-specific layers is written as k, where $k \leq e - s$. Given a transformer model of one patch embedding layer and t encoders, the MC workflow is formulated in Equation (4).

$$MC(k, s, e, i) = \begin{cases} \text{modal-specific}, & i \in [s, e) \cap [s, s + k), \\ \text{modal-common}, & \text{otherwise}, \end{cases} \tag{4}$$

where $i \in [0, t + 1)$ represents the transformer component index, and the patch embedding layer index is $i = 0$. Through Equation (4) of the MC module, the first s layers are modal common, the next k layers are modal specific, and the last $t + 1 - k$ layers are modal common.

Figure 2 shows the case that has $s = 0, e = t, k = e - s$. For example, as a transformer model has 12 encoders, in the medium modal-specific configuration of $k = 9, v = [1, 10)$, the patch embedding layer is modal common, the first 9 encoders layers are modal specific, and the rest of the three encoders layers are modal common.

3.2.2. Modal Information Embedding

Different from the position embedding, P is set as the modal-common default, the modal information embedding MIE is always set as modal specific to freely encode modal in-

formation to alleviate the feature deviations towards modal variations. Inspired by [41], the modal information embedding is formulated in Equation (5), as follows:

$$Z = [x_{cls}; E(x_p^1); E(x_p^2); \ldots; E(x_p^N)] + P + MIE, \tag{5}$$

where Z denotes the output of patch-embedding layers (i.e., $E(\cdot)$); x_{cls} is a learnable token embedding; x_p is a image patch, and N is the number of patches; P is a learnable position embedding; MIE is a learnable modal information embedding.

3.3. Progressively Hybrid Transformer

Combining the proposed RHA and FHM designed in previous subsections, the multi-modal hybrid transformer would be upgraded into a progressively hybrid transformer (PHT) because both image and feature level information is progressively fused. As shown in Figure 2, the PHT's loss module consists of a triplet loss and a classification loss. The triplet loss is the hard-miming triplet loss function [62] formulated in Equation (6), as follows:

$$L_{tri} = log[1 + exp(\|f_a - f_{hp}\|^2 - \|f_a - f_{hn}\|^2)], \tag{6}$$

where f_a is the multi-modal fusion feature of an anchor sample, f_{hp} is the multi-modal fusion feature of a hard positive sample that is the farthest away from the anchor sample and has the same class as the anchor sample, and f_{hn} is the multi-modal fusion feature of a hard negative sample that is close to the anchor sample and has a different class from the anchor sample. The classification loss is the commonly used cross-entropy loss function [16] formulated in Equation (7), as follows:

$$L_{cls} = -\delta(y == c) log(p(y|g)), \tag{7}$$

where δ is an indicator function that is equal to 1 if the equation in the formula is true, otherwise 0, g is the batch normalized multi-modal fusion feature of a sample, and y and c are the sample's prediction and truth class labels, respectively.

4. Experiments and Analysis

To show the proposed method's advantage, this paper compares the PHT method with state-of-the-art methods on two challenging multi-modal vehicle datasets, namely, RGBNT100 [17] and RGBN300 [17]. The RGBNT100 is a three-modal dataset, including visible, near-infrared, and thermal images of 100 subjects, and the RGBN300 is a two-modal dataset, containing visible and near-infrared images of 300 subjects. Following [17], on both RGBNT100 and RGBN300 datasets, half of the dataset is used for training and the other half is for testing. The cumulative matching characteristic (CMC) curve [63] and the mean average precision (mAP) [64] are applied as the performance metric. R1, R5, and R10 denote rank-1, rank-5, and rank-10 identification rates on a CMC curve, respectively.

4.1. Implementation Details

The software tools are Pytorch 1.7 [65], CUDA 11.1, and python 3.8. The hardware device is one GeForce RTX 3090 GPU. All images of each modality are resized to 192 × 192 sized images. The random horizontal flipping, padding, random cropping, and random erasing [55] are applied for data augmentation, as performed in [16]. Each mini-batch contains 16 subjects, and if on the RGBNT100 dataset, each subject has 4 visible images, 4 near-infrared images, and 4 thermal images, otherwise, on the RGBN300 dataset, each subject has 4 visible images and 4 near-infrared images. The ImageNet pre-trained vision transformer (ViT) is applied as the backbone as performed in [16]. Following [16], the momentum and weight decay of the stochastic gradient descent (SGD) optimizer [33] are set to 0.9 and 0.0001, respectively, the learning rate is initialized as 0.008 with cosine learning rate decay, and the patch size and stride size are both set to 16 × 16. As RGBNT100 and RGBN300 are three-modal and double-modal datasets, the PHT's backbone is corre-

spondingly made to have three ViT branches and two ViT branches on the RGBNT100 and RGBN300. As each ViT branch has 1 patch embedding layer and 12 transformer encoder layers, the controlled field of the modal-specific controller (MC) is limited to $v = [s, e) | 0 \leq s \leq e \leq 13$.

4.2. Comparison with State-of-the-Art

The performance comparison between the proposed PHT and state-of-the-art methods is shown in Table 1. Those state-of-the-art methods could be divided into two categories: (1) CNN-based methods, namely, HAMNet [17], GAFNet [21], CCNet [19], and DANet [22]; (2) the transformer-based method, namely, TransReID [16]. Several interesting observations are as follows.

Table 1. The performance comparison between the proposed PHT and other state-of-the-arts methods on both RGBNT100 and RGBN300.

Methods	RGBNT100				RGBN300			
	mAP (%)	**R1 (%)**	**R5 (%)**	**R10 (%)**	**mAP (%)**	**R1 (%)**	**R5 (%)**	**R10 (%)**
HAMNet [17]	65.4	85.5	87.9	88.8	61.9	84.0	86.0	87.0
TransReID [16]	60.1	82.2	83.7	84.7	67.1	86.5	88.0	88.7
GAFNet [21]	74.4	93.4	94.5	95.0	72.7	91.9	93.6	94.2
CCNet [19]	77.2	96.3	97.2	97.7	N/A	N/A	N/A	N/A
DANet [22]	N/A	N/A	N/A	N/A	71.0	89.9	90.9	91.5
PHT (Proposed)	79.9	92.7	93.2	93.7	79.3	93.7	94.8	95.3

First, the transformer-based method TransReID [16] is inferior to those CNN-based methods. For example, the mAP of TransReID [16] is 5.3% smaller than the earliest CNN-based method called HAMNet [17]. This observation illustrates that without an appropriate multi-modal information fusion, even using a strong transformer, there is no accuracy performance advantage.

Second, the proposed method (i.e., PHT) greatly improves TransReID [16] and outperforms those CNN-based methods. On RGBNT100, the PHT's mAP is 1.8% larger than that of the strongest CNN-based method, i.e., CCNet [19], although R1, R5, and R10 of the PHT are inferior to those of CCNet [19]. According to [64], mAP is a more comprehensive performance indicator than R1, R5, and R10, who are isolated points on a CMC curve. Therefore, the PHT is better overall than CCNet [19]. Similarly, on RGBN300, the PHT gains good performance, which defeats the strongest one (i.e., GAFNet [21]) by a 6.6% larger mAP. These results suggest that the full fusion working at both image and feature levels is a great help for a transformer model to improve multi-modal vehicle ReID.

4.3. Analysis of Feature Hybrid Mechanism

4.3.1. Influence of Modal-Specific Controller

To investigate the influence of using modal-specific layers at different positions, five types of modal-specific controller (MC) configurations are formed based on Equation (4), as shown in Table 2. These configurations of the MC are conducted on RGBNT100. Furthermore, position embedding is set to be modal-common and disabled RHA to avoid their influence. The experimental results are shown in Figure 3.

From Figure 3 one can see that three partial modal-specific (i.e., shallow modal-specific, medium modal-specific, and deep modal-specific) configurations outperform fully modal-specific and fully modal-common configurations. Especially, when the deep modal-specific configuration has the number of modal-specific layers $k = 5$ and controlled field $v = [8, 13)$, the best performance (79.0% mAP) is achieved. Furthermore, among three partial modal-specific configurations, the deep modal-specific configuration outperforms shallow modal-specific and medium modal-specific configurations. The strength of the deep modal-specific configuration setting shallow layers of a transformer to be modal-

common is that the fusion computation works on a deep location requiring complementary features of different modalities so that modal-common layers should be configured at shallow positions while modal-specific layers should be configured at deep positions near to the fusion computation for fusing multi-modal complementary information better.

Table 2. Five types of modal-specific controller (MC) configurations.

Type	v		k	Patch Embedding Layer	Transformer Encoder Layers
	s	e			
Fully modal common	0	0	0	Common	Common: all Layers
Fully modal specific	0	$t+1$	$t+1$	Specific	Specific: all Layers
Shallow modal specific	0	t	$1 \leq k \leq t$	Specific	Specific: the first $k-1$ layers
Medium modal specific	1	$t+1$	$1 \leq k \leq t$	Common	Specific: the first k layers
Deep modal specific	1	$t+1$	$1 \leq k \leq t$	Common	Specific: the last k layers

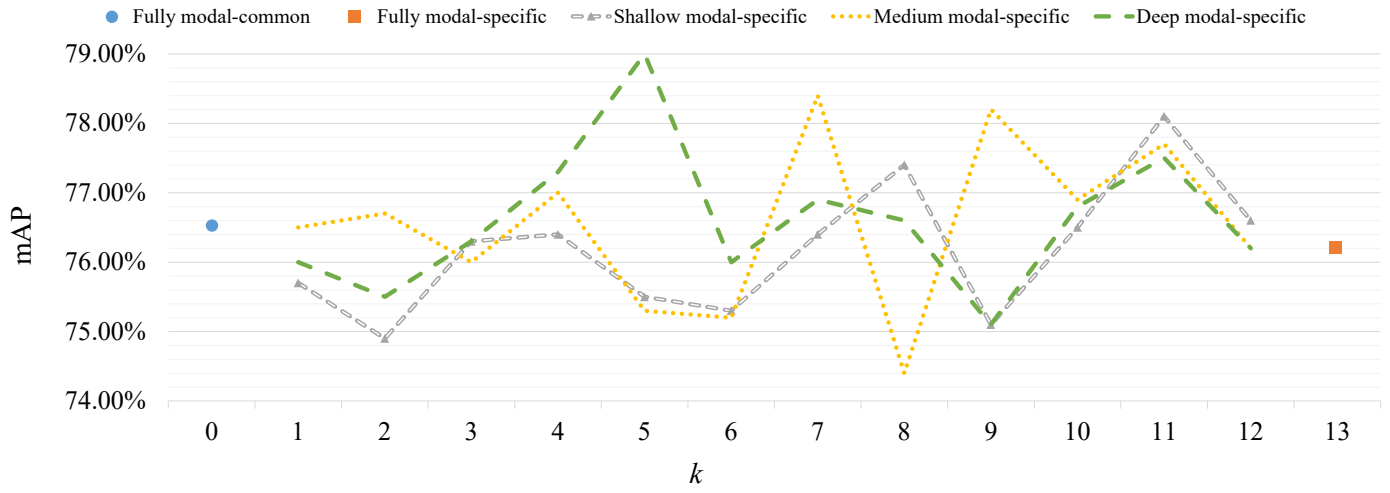

Figure 3. The comparison of modal-specific controller configurations on RGBNT100.

4.3.2. Role of Modal Information Embedding

Based on the observation on the analysis of the modal-specific controller (MC) in Section 4.3.1, each type's best MC configuration is chosen and RHA is discarded, and then the role of modal information embedding (MIE) is analyzed, as follows.

From Figure 4, one can see that PHT with MIE outperforms the PHT without MIE by a 1.9% larger mAP on RGBNT100 and a larger 0.5% mAP on RGBN300, respectively, under the modal-specific configuration of $v = [8, 13)$. Unfortunately, using MIE brings a negative impact on RGBNT100 and RGBN300 under the fully modal-specific configuration of $v = [0, 13)$. This is because the fully modal-specific configuration has no modal-common layers, hindering MIE from learning modal invariant characteristics. Consequently, MIE is useful to alleviate feature deviations towards modal variations and is helpful to enhance multi-modal complementary information fusing but requires a proper MC configuration.

4.3.3. Impact of Position Embedding

Similar to the experiment settings in the previous model information embedding (MIE) analysis, each type's best MC configuration is chosen and RHA is discarded, and then the performance resulting from modal-specific and modal-common position embedding on RGBNT100 and RGBN300 is compared.

From Table 3, one can find that most modal-common position embedding cases are stronger than modal-specific position embedding. For example, on RGBNT100, regarding the $v = [1, 10)$ case, the mAP of modal-common position embedding is 1.5% larger than that of the modal-specific position embedding. Similarly, for the $v = [8, 13)$ case, the modal-common position embedding outperforms the modal-specific position embedding by a

1.4% mAP improvement. These results mean that the modal-common position embedding is more robust than the modal-specific position embedding. The reason for this situation is deduced to the modal-common position embedding requiring fewer parameters than the modal-specific position embedding so that it is easier to be well trained.

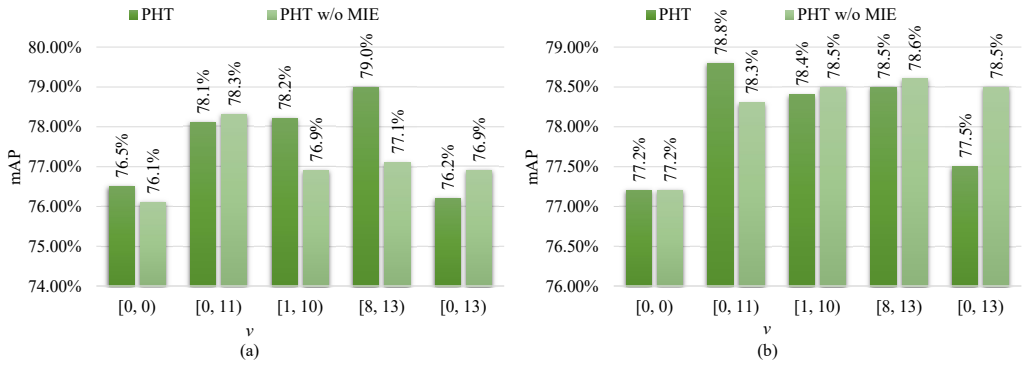

Figure 4. The ablation study of modal information embedding (MIE) on (**a**) RGBNT100 and (**b**) RGBN300 datasets. Here, k is configured to $k = e - s$.

Table 3. The comparison of the modal-specific position embedding and the modal-common position embedding on RGBNT100 and RGBN300.

v	k	Type	RGBNT100				RGBN300			
			mAP (%)	R1 (%)	R5 (%)	R10 (%)	mAP (%)	R1 (%)	R5 (%)	R10 (%)
[0, 0)	0	Common	76.5	91.5	93.1	93.6	77.2	91.2	92.5	93.1
		Specific	76.1	91.5	92.9	93.4	77.8	92.8	93.6	93.8
[0, 11)	11	Common	78.1	91.9	92.7	93.2	78.8	93.5	94.5	95.2
		Specific	77.7	92.1	92.9	93.7	79.0	93.7	94.7	95.1
[1, 10)	9	Common	78.2	93.4	94.2	94.8	78.4	93.4	94.4	94.8
		Specific	76.7	91.7	93.1	93.9	78.4	93.2	94.2	94.8
[8, 13)	5	Common	79.0	93.4	94.4	95.3	78.5	92.3	93.1	93.7
		Specific	77.6	90.6	91.6	92.1	78.4	92.8	93.7	94.2
[0, 13)	13	Common	76.2	92.7	93.6	94.3	77.5	92.4	93.3	94.0
		Specific	76.9	92.8	94.2	94.6	77.2	92.5	93.2	93.7

4.3.4. Effect of Feature Fusion

According to Figure 3, the best configuration (i.e., $k = 5$ and $v = [8, 13)$ in deep modal specific) are selected to compare the average, Hadamard product [66], and concatenating fusion methods. Here, the modal-common position embedding is applied and RHA is still disabled.

From Table 4, one can observe that the average fusion method gains the best result, that is, 79.0% mAP, 93.4% R1, 94.4% R5, and 95.3% R10 on RGBNT100, and 78.5% mAP, 92.3% R1, 93.1% R5, and 93.7% R10 on RGBN300. The preponderance of the average fusion method suggests that the low-pass effect of average fusion could filter out multi-modal heterogeneity of multi-modal data, so as to improve performance more significantly.

4.4. Analysis of Random Hybrid Augmentation

4.4.1. Comparison with the Preliminary Work

To straightforwardly show the role of random hybrid augmentation (RHA), this paper compares the proposed PHT to the preliminary work [23], namely, H-ViT, which does not utilize RHA. As shown in Figure 5, the PHT in this paper consistently outperforms H-ViT [23] on both RGBNT100 and RGBN300. This comparison illustrates that the fusion

at the image level of RHA supplements the fusion at the feature level, further boosting multi-modal vehicle ReID. More detailed analyses of RHA are constructed as follows.

Table 4. Results of different fusion methods on RGBNT100 and RGBN300.

Fusion	RGBNT100				RGBN300			
	mAP (%)	R1 (%)	R5 (%)	R10 (%)	mAP (%)	R1 (%)	R5 (%)	R10 (%)
Average	79.0	93.4	94.4	95.3	78.5	92.3	93.1	93.7
Hadamard Product	45.2	63.0	65.9	67.6	72.0	89.1	90.5	91.2
Concatenating	74.9	92.4	93.5	94.1	75.6	91.2	92.3	92.9

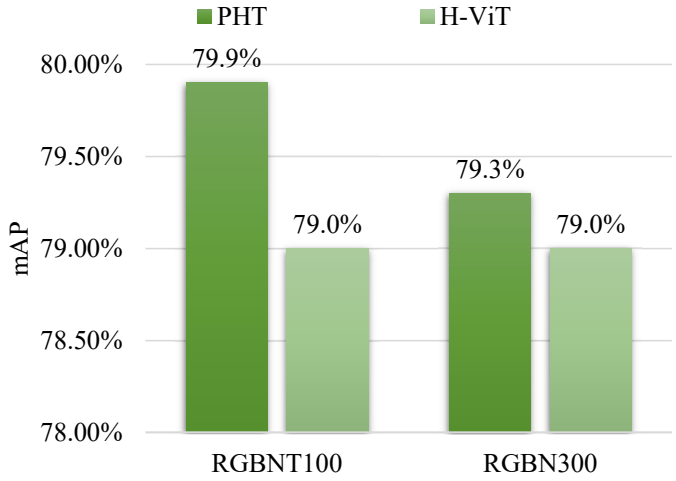

Figure 5. The comparison between PHT and the previous work H-ViT on RGBNT100 and RGBN300 datasets.

4.4.2. Role of Image Random Cropper

According to Figure 3, the best MC configuration (i.e., the deep modal-specific configuration of $v = [8, 13)$) is fixed and two key parameters of the image random cropper (IRC), i.e., n_{region} and p_{region}, are changed to validate the role of IRC. The results are shown in Figure 6a,b.

As shown in Figure 6a, one can see that the best n_{region} value is 3 which brings 0.9% mAP performance improvements but most of the rest of the values cause performance degradation. This paper believes this small n_{region} could not bring data augmentation while a too dominant n_{region} could damage the original image information. Based on a similar reason, as shown in Figure 6b, the p_{region} has a similar performance fluctuation trend, that is, performance improvements followed by performance degradation. Therefore, good RHA should have proper n_{region} and p_{region} settings for better multi-modal complementary information learning, as performed in existing data augmentation works [55,56].

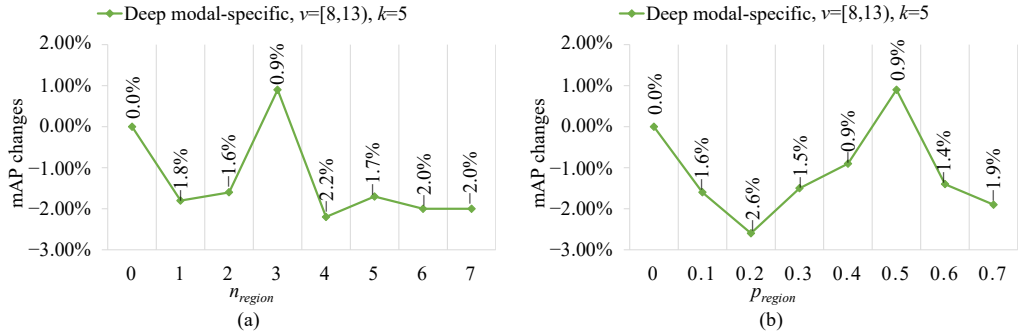

Figure 6. The role of parameters (**a**) n_{region} and (**b**) p_{region} in image random cropper on RGBNT100.

4.4.3. Impact of Local Region Hybrider

Based on observations in the previous subsection (i.e., Section 4.4.2) of IRC analysis, the impact of the local region hybrider is further analyzed by using different hybrid methods, including (1) average, (2) self-excluding average, (3) Hadamard product, (4) self-excluding Hadamard product, and (5) randomly swapping. The results are shown in Table 5.

From Table 5, it can be found that average reaches the best performance, i.e., 79.9% mAP, which defeats self-excluding average, Hadamard product, self-excluding Hadamard product, and randomly swapping by a 1.2%, 1.1%, 3.1%, and 1.9% mAP, respectively. This result is in line with the average preponderance of the average fusion method in the feature hybrid mechanism (i.e., Section 4.3.4), which demonstrates that the low-pass effect of average fusion could filter out multi-modal heterogeneity of multi-modal data again to improve performance more significantly.

Table 5. Results of different local region hybrider on RGBNT100.

Local Region Hybrider	mAP (%)	R1 (%)	R5 (%)	R10 (%)
Average	79.9	92.7	93.2	93.7
Self-excluding Average	78.7	91.8	92.6	93.1
Hadamard Product	78.8	91.7	92.9	93.6
Self-excluding Hadamard Product	76.8	91.1	92.1	92.5
Randomly Swapping	78.0	91.0	92.1	92.7

4.5. Discussion

Based on the comparison with state-of-the-art methods in Section 4.2, the performance strength of the PHT is demonstrated. Specifically, the proposed PHT method is superior to the transformer-based method TransReID [16] by 19.8% mAP on RGBNT100 [17] and 12.2% mAP on RGBN300 [17]. Compared to two strong CNN-based methods, namely, GAFNet [21] and CCNet [19], the proposed PHT method outperforms GAFNet [21] by 2.7% mAP on RGBNT100 [17] and CCNet [19] by 6.6% mAP on RGBN300 [17]. Furthermore, based on ablation experiments in Sections 4.3 and 4.4, the performance advantage of the PHT is demonstrated. Especially, compared to the preliminary work H-ViT [23], the proposed PHT mAP is 0.9% larger on RGBNT100 [17]. The victory of the proposed PHT in this paper demonstrates that image level information fusion is beneficial to feature level information fusion. The victory is actually expected because the fusion at the image level could be seen as a data augmentation, which is naturally conducive to the subsequent feature learning.

5. Conclusions

To comprehensively fuse multi-modal complementary information for multi-modal vehicle ReID, this paper proposes a progressively hybrid transformer (PHT). The PHT is constructed with two aspects: random hybrid augmentation (RHA) and a feature hybrid mechanism (FHM). At the image level, the RHA emphasizes structural characteristics of all modalities by fusing random regions of multi-modal images. At the feature level, the FHM allows for a multi-modal feature interaction by encoding modal information and fusing different modal features in different positions. The experiments show that (1) the proposed PHT surpasses the state-of-the-art methods on both RGBNT100 and RGBN300 datasets; (2) the multi-modal hybrid transformer built on the FHM is more advantageous than the single-branch transformer; (3) the fusion at the image level of RHA supplements the fusion at the feature level to further boost multi-modal vehicle ReID. Although the PHT is effective for multi-modal vehicle ReID, there is still a limitation of the PHT because it requires a manual setting of fusion configurations (e.g., fusion locations and fusion manners). In the future, a network architecture search approach will be explored to automatically determine fusion locations and manners to realize an adaptive fusion for multi-modal vehicle ReID.

Author Contributions: Conceptualization, W.P. and J.Z.; writing—original draft preparation, W.P. and J.L.; writing—editing, W.P., J.L. and L.H. (Linhan Huang); writing—review, L.H. (Lan Hong) and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the grant 2022J06023, in part by the National Natural Science Foundation of China under the grant 61976098, and in part by Collaborative Innovation Platform Project of Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone under the grant 2021FX03. The authors would like to thank Yun Liao for his help in this work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

ReID	Re-identification
PHT	Progressively hybrid transformer
RHA	Random hybrid augmentation
FHM	Feature hybrid mechanism
NIR	Near-infrared
TIR	Thermal-infrared
CNN	Convolutional neural networks
IRC	Image random cropper
LRH	Local region hybrider
MC	Modal-specific controller
MIE	Modal information embedding
CMC	Cumulative matching characteristic
mAP	Mean average precision
R1	Rank 1 identification rate
R5	Rank 5 identification rate
R10	Rank 10 identification rate
ViT	Vision transformer
SGD	Stochastic gradient descent

References

1. Avola, D.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Pannone, D.; Piciarelli, C. Bodyprint—A meta-feature based LSTM hashing model for person re-identification. *Sensors* **2020**, *20*, 5365. [CrossRef] [PubMed]
2. Paolanti, M.; Romeo, L.; Liciotti, D.; Pietrini, R.; Cenci, A.; Frontoni, E.; Zingaretti, P. Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. *Sensors* **2018**, *18*, 3471. [CrossRef] [PubMed]
3. Uddin, M.K.; Bhuiyan, A.; Bappee, F.K.; Islam, M.M.; Hasan, M. Person Re-Identification with RGB–D and RGB–IR Sensors: A Comprehensive Survey. *Sensors* **2023**, *23*, 1504. [CrossRef]
4. Deng, J.; Hao, Y.; Khokhar, M.S.; Kumar, R.; Cai, J.; Kumar, J.; Aftab, M.U. Trends in vehicle re-identification past, present, and future: A comprehensive review. *Mathematics* **2021**, *9*, 3162.
5. Zhu, X.; Luo, Z.; Fu, P.; Ji, X. Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020 ; pp. 602–603.
6. Wang, Z.; Tang, L.; Liu, X.; Yao, Z.; Yi, S.; Shao, J.; Yan, J.; Wang, S.; Li, H.; Wang, X. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017 ; pp. 379–387.
7. Meng, D.; Li, L.; Wang, S.; Gao, X.; Zha, Z.J.; Huang, Q. Fine-grained feature alignment with part perspective transformation for vehicle reid. In Proceedings of the ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 619–627.
8. Zhou, Y.; Shao, L. Aware attentive multi-view inference for vehicle re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6489–6498.

9. Zhu, J.; Zeng, H.; Huang, J.; Liao, S.; Lei, Z.; Cai, C.; Zheng, L. Vehicle re-identification using quadruple directional deep learning features. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 410–420. [CrossRef]
10. Khan, S.D.; Ullah, H. A survey of advances in vision-based vehicle re-identification. *Comput. Vis. Image Underst.* **2019**, *182*, 50–63. [CrossRef]
11. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef]
12. Yang, Q.; Wang, P.; Fang, Z.; Lu, Q. Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification. *Sensors* **2020**, *20*, 4431. [CrossRef]
13. Chen, Y.; Yang, T.; Li, C.; Zhang, Y. A Binarized segmented ResNet based on edge computing for re-identification. *Sensors* **2020**, *20*, 6902. [CrossRef] [PubMed]
14. Si, R.; Zhao, J.; Tang, Y.; Yang, S. Relation-based deep attention network with hybrid memory for one-shot person re-identification. *Sensors* **2021**, *21*, 5113. [CrossRef] [PubMed]
15. Lorenzo-Navarro, J.; Castrillón-Santana, M.; Hernández-Sosa, D. On the use of simple geometric descriptors provided by RGB-D sensors for re-identification. *Sensors* **2013**, *13*, 8222–8238. [CrossRef]
16. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15013–15022.
17. Li, H.; Li, C.; Zhu, X.; Zheng, A.; Luo, B. Multi-spectral vehicle re-identification: A challenge. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11345–11353.
18. Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; Tang, J. Robust multi-modality person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 3529–3537.
19. Zheng, A.; Zhu, X.; Li, C.; Tang, J.; Ma, J. Multi-spectral Vehicle Re-identification with Cross-directional Consistency Network and a High-quality Benchmark. *arXiv* **2022**, arXiv:2208.00632.
20. Wang, Z.; Li, C.; Zheng, A.; He, R.; Tang, J. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virginia, VA, USA, 17–19 November 2022; Volume 36, pp. 2633–2641.
21. Guo, J.; Zhang, X.; Liu, Z.; Wang, Y. Generative and Attentive Fusion for Multi-spectral Vehicle Re-Identification. In Proceedings of the International Conference on Intelligent Computing and Signal Processing, Beijing, China, 21–24 October 2022; pp. 1565–1572.
22. Kamenou, E.; Rincon, J.; Miller, P.; Devlin-Hill, P. Closing the Domain Gap for Cross-modal Visible-Infrared Vehicle Re-identification. In Proceedings of the International Conference on Pattern Recognition, Montréal, QC, Canada, 21–25 August 2022; pp. 2728–2734.
23. Pan, W.; Wu, H.; Zhu, J.; Zeng, H.; Zhu, X. H-ViT: Hybrid Vision Transformer for Multi-modal Vehicle Re-identification. In Proceedings of the CAAI International Conference on Artificial Intelligence, Beijing, China, 27–28 August 2022; pp. 255–267.
24. Zhang, G.; Zhang, P.; Qi, J.; Lu, H. Hat: Hierarchical aggregation transformers for person re-identification. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 516–525.
25. Khorramshahi, P.; Kumar, A.; Peri, N.; Rambhatla, S.S.; Chen, J.C.; Chellappa, R. A dual-path model with adaptive attention for vehicle re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October – 2 November 2019; pp. 6132–6141.
26. Guo, H.; Zhu, K.; Tang, M.; Wang, J. Two-level attention network with multi-grain ranking loss for vehicle re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4328–4338. [CrossRef]
27. Li, M.; Liu, J.; Zheng, C.; Huang, X.; Zhang, Z. Exploiting Multi-view Part-wise Correlation via an Efficient Transformer for Vehicle Re-Identification. *IEEE Trans. Multimed.* **2021**, *25*, 919–929. [CrossRef]
28. Gu, X.; Chang, H.; Ma, B.; Bai, S.; Shan, S.; Chen, X. Clothes-changing person re-identification with rgb modality only. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 1060–1069.
29. Cai, J.; Deng, J.; Aftab, M.U.; Khokhar, M.S.; Kumar, R. Efficient and deep vehicle re-identification using multi-level feature extraction. *Appl. Sci.* **2019**, *9*, 1291.
30. Zeng, Z.; Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.Y.; Satoh, S. Illumination-adaptive person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 3064–3074. [CrossRef]
31. Zhang, Z.; Da Xu, R.Y.; Jiang, S.; Li, Y.; Huang, C.; Deng, C. Illumination adaptive person reid based on teacher-student model and adversarial training. In Proceedings of the 2020 IEEE International Conference on Image Processing, Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2321–2325.
32. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *IEEE* **1998**, *86*, 2278–2324. [CrossRef]
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.

35. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
36. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Nord, France, 6–11 July 2015; pp. 448–456.
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
38. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Liu, Q.; Chen, D.; Chu, Q.; Yuan, L.; Liu, B.; Zhang, L.; Yu, N. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing* **2022**, *483*, 333–347. [CrossRef]
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017 ; Volume 30.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
43. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual Only, 18–24 July 2021; pp. 10347–10357.
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
45. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 6881–6890.
46. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
47. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
48. Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; Gavves, E. Nformer: Robust person re-identification with neighbor transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–25 June 2022; pp. 7297–7307.
49. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
50. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
51. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
52. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
53. Wu, Y.H.; Liu, Y.; Zhan, X.; Cheng, M.M. P2T: Pyramid pooling transformer for scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, 1–12. [CrossRef]
54. Chen, C.; Ye, M.; Qi, M.; Wu, J.; Jiang, J.; Lin, C.W. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Trans. Image Process.* **2022**, *31*, 2352–2364. [CrossRef] [PubMed]
55. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.
56. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5056–5069. [CrossRef]
57. Chen, M.; Wang, Z.; Zheng, F. Benchmarks for corruption invariant person re-identification. *arXiv* **2021**, arXiv:2111.00880.
58. Li, Q.; Yu, Z.; Wang, Y.; Zheng, H. TumorGAN: A multi-modal data augmentation framework for brain tumor segmentation. *Sensors* **2020**, *20*, 4203. [CrossRef] [PubMed]
59. Ojagh, S.; Cauteruccio, F.; Terracina, G.; Liang, S.H. Enhanced air quality prediction by edge-based spatiotemporal data preprocessing. *Comput. Electr. Eng.* **2021**, *96*, 107572. [CrossRef]
60. Lin, Z.; Liu, C.; Qi, W.; Chan, S.C. A Color/Illuminance Aware Data Augmentation and Style Adaptation Approach to Person Re-Identification. *IEEE Access* **2021**, *9*, 115826–115838. [CrossRef]

61. Huang, H.; Li, D.; Zhang, Z.; Chen, X.; Huang, K. Adversarially occluded samples for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5098–5107.
62. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
63. Gray, D.; Brennan, S.; Tao, H. Evaluating appearance models for recognition, reacquisition, and tracking. In Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Arusha, Tanzanian, 14 October 2007; Volume 3, pp. 1–7.
64. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1116–1124.
65. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
66. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–18 June 2020; pp. 10076–10085.