# Multi-Spectral Vehicle Re-Identification: A Challenge

**Hongchao Li,**[1] **Chenglong Li,**[1] **Xianpeng Zhu,**[1] **Aihua Zheng,**[1,*] **Bin Luo**[1,2]
[1]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education,
School of Computer Science and Technology, Anhui University, Heifei, China
[2]Key Lab of Industrial Image Processing & Analysis of Anhui Province, Hefei, China
{lhc950304, ahzheng214, lcl1314, xpzhu6325}@foxmail.com, luobin@ahu.edu.cn

## Abstract

Vehicle re-identification (Re-ID) is a crucial task in smart city and intelligent transportation, aiming to match vehicle images across non-overlapping surveillance camera views. Currently, most works focus on RGB-based vehicle Re-ID, which limits its capability of real-life applications in adverse environments such as dark environments and bad weathers. IR (Infrared) spectrum imaging offers complementary information to relieve the illumination issue in computer vision tasks. Furthermore, vehicle Re-ID suffers a big challenge of the diverse appearance with different views, such as trucks. In this work, we address the RGB and IR vehicle Re-ID problem and contribute a multi-spectral vehicle Re-ID benchmark named RGBN300, including RGB and NIR (Near Infrared) vehicle images of 300 identities from 8 camera views, giving in total 50125 RGB images and 50125 NIR images respectively. In addition, we have acquired additional TIR (Thermal Infrared) data for 100 vehicles from RGBN300 to form another dataset for three-spectral vehicle Re-ID. Furthermore, we propose a Heterogeneity-collaboration Aware Multi-stream convolutional Network (HAMNet) towards automatically fusing different spectrum features in an end-to-end learning framework. Comprehensive experiments on prevalent networks show that our HAMNet can effectively integrate multi-spectral data for robust vehicle Re-ID in day and night. Our work provides a benchmark dataset for RGB-NIR and RGB-NIR-TIR multi-spectral vehicle Re-ID and a baseline network for both research and industrial communities. The dataset and baseline codes are available at: https://github.com/ttaalle/multi-modal-vehicle-Re-ID.

## Introduction

Vehicle re-identification (Re-ID) is to identify vehicle images from the gallery that shares the same identity as the given probe. It is an active and challenging computer vision task and has drawn much attention due to its wide applications in video surveillance, social security, smart city, and intelligent transportation, to name a few. Despite recent breakthroughs in vehicle Re-ID, it still faces huge challenging especially in adverse illumination conditions, such as strong or poor lighting, shadow or black night. The main reason is the RGB camera can only capture visible light $(0.38-0.78\mu m)$ reflected by subjects, which significantly affects the imaging quality of the visible spectrum, as shown in Fig. 1 (Middle). Therefore, identifying the vehicles with various illumination environments in day and night is an imminent question in vehicle Re-ID.




Figure 1: Examples of images captured by RGB and NIR cameras from six camera views. Middle: the RGB images of six camera views. Bottom: the corresponding NIR images.

The NIR (near infrared) camera can capture near infrared light $(0.78-3\mu m)$ reflected by subjects and is not affected in dark environments and bad weathers, which are more crucial for social security. As shown in Fig. 1 (Bottom), NIR information is able to handle the imaging limitations of visible ones, herein can handle the conventional RGB vehicle Re-ID in adverse illumination conditions and expand its application from daytime to nighttime. Recently, some works propose RGB-Infrared cross-modality Re-ID to overcome the limitation of RGB imagination in dark environments. However, infrared images contain no color information. The large heterogeneous issue between RGB and NIR modalities brings a big challenge for cross-modal matching. Furthermore, existing vehicle Re-ID methods and datasets only focus on the single visible spectrum, which limits its capability of real-life applications in adverse environments such as dark environments and bad weathers.

Table 1: Publicly available benchmark datasets for person/vehicle re-identification (Re-ID). In the column of modality, '+' and '/' denote multi-modality and cross-modality respectively.

| | Benchmark | cameras | ID | images | modality | Multiview |
|---|---|---|---|---|---|---|
| person | VIPER | 2 | 632 | 1264 | RGB | no |
| | iLIDS | 2 | 119 | 476 | RGB | no |
| | CUHK01 | 2 | 972 | 1942 | RGB | no |
| | Market | 6 | 1501 | 32668 | RGB | yes |
| | DukeMTMC-ReID | 8 | 1404 | 36411 | RGB | yes |
| | PAVIS | - | 79 | 788 | RGB+D | no |
| | BIWI | - | 50 | 39280 | RGB+D | no |
| | RegDB | - | 412 | 8240 | RGB+T | no |
| | SYSU-MM01 | 6 | 491 | 303357 | RGB/N | yes |
| vehicle | VeRi-776 | 20 | 776 | 49357 | RGB | yes |
| | VehicleID | - | 26267 | 221763 | RGB | no |
| | Vehicle-1M | - | 55527 | 936051 | RGB | no |
| | CityFlow-ReID | 40 | 666 | 229680 | RGB | yes |
| | RGBN300 | 8 | 300 | 100250 | RGB+N | yes |

In this work, we contribute a comprehensive RGB-NIR multi-spectral vehicle Re-ID dataset called RGBN300. RGBN300 contains 50125 image pairs of 300 vehicles from 8 camera views with RGB and NIR spectra. To the best of our knowledge, this dataset provides the first time for the study of multi-spectral vehicle Re-ID. Compared with other existing commonly used Re-ID datasets (Gray, Brennan, and Tao 2007; Zheng, Gong, and Xiang 2009; Dong et al. 2011; Ristani et al. 2016; Barbosa et al. 2012; Munaro et al. 2014; Nguyen et al. 2017; Wu et al. 2017; Liu et al. 2016b; 2016a; Guo et al. 2018; Tang et al. 2019) as shown in Table 1, our dataset has the following major advantages. 1) It contains spatially aligned RGB-NIR vehicle image pairs. 2) It includes a large number of video frames, which enables users to perform large-scale performance evaluations. 3) It contains 2-8 views per vehicle, which supports vehicle matching from different views. In addition, we have acquired additional TIR (Thermal Infrared) data for 100 vehicles from RGBN300 to form another dataset for three-spectral vehicle Re-ID.

Although a number of works investigate the TIR (thermal infrared) or depth information as the complementary modality in person Re-ID (Barbosa et al. 2012; Munaro et al. 2014; Nguyen et al. 2017; Wu, Zheng, and Lai 2017), Most of the existing works directly utilized the TIR or depth information as auxiliary information or connective feature for person Re-ID. We argue that it is essential to explore the ability of heterogeneous data to collaborate to identify the same ID and effective ways to aware of the contribution of different spectra to the same class. To provide a powerful baseline algorithm, we propose a Heterogeneity-collaboration Aware Multi-stream convolutional Network (HAMNet) towards automatically fusing different spectrum features in an end-to-end learning framework. First, we naturally build a multi-stream convolutional network and use two independent ID losses to constrain two heterogeneous data with the same identity. Second, we propose to constrain multi-spectral heterogeneous data with a similar score distribution and combine it with the spectra independent ID losses to form a heterogeneity-collaboration loss. In other words, we enforce the similarity between score distribution of heterogeneous data to be consistent, in addition to be with the same class/identity. Inspired by Class Activation Map (CAM) (Zhou et al. 2016), which can indicate the locations of informative parts and the richness of features, we further propose to measure the importance of each spectrum for classification by the CAMs of different spectra. Finally, we combine class activation maps from different spectra according to the importance of CAM and constrain it with a class-aware ID loss.

Comprehensive experiments on prevalent networks show that our HAMNet can effectively integrate multi-spectral data for robust vehicle Re-ID in day and night. With our new benchmark dataset, we propose a novel multi-stream convolutional network to adaptively incorporate the information from multi-spectrum images for vehicle Re-ID.

The contribution of this paper can be summarized as follows.

- We are the first time to contribute a standard benchmark dataset RGBN300 to support the study of multi-spectral vehicle Re-ID. We also construct another benchmark datasets with RGB, near infrared and thermal infrared images for related researches and applications. These benchmark datasets will be open to the public for free academic usage.
- We propose a Heterogeneity-collaboration Aware Multi-stream convolutional Network (HAMNet) towards automatically fusing spectrum-specific features in the network for multi-spectral matching, which provide a powerful baseline algorithm for the future study.
- Comprehensive experiments on our challenging benchmark datasets RGBN300 and RGBNT100 validate the superior performance of our model for multi-spectral vehicle Re-ID.

## Related Work

We briefly review the related works in the following three folds, i.e., vehicle Re-ID, cross-modality, and multi-modality person Re-ID.

### Vehicle Re-ID

With the development of person Re-ID, the vehicle re-identification task has gained more and more attention in recent years. Liu et al. (2016b) proposed a dataset VeRi-776 and built a coarse-to-fine progressive search framework by adding license plates and spatio-temporal labels. Liu et al. (2016a) released a benchmark dataset, called VehicleID and presented a deep relative distance learning method to learn a Euclidean space where distance can be directly used to measure the similarity of vehicle images. Wang et al. (2017) proposed an orientation invariant feature embedding module and a spatial-temporal regularization module for vehicle Re-ID. Shen et al. (2017) proposed a two-stage framework that incorporates complex spatio-temporal information for effectively regularizing the vehicle Re-ID results. Yan et al. (2017) modeled the relationships of vehicle images as a multi-grain list and proposed two ranking methods for vehicle Re-ID. Lou et al. (2019) resorted to adversarial learning to generate cross views examples, while He

et al. (2019) proposed a part-regularized approach to enhance the discriminative capability of global features for vehicle Re-ID. Recently, Guo et al. (2018) proposed a larger dataset Vehicle-1M, including 936,051 images of 55,527 vehicles with only the head and rear of the vehicles. Tang et al. (2019) proposed a large city-scale benchmark for vehicle Re-ID, called CityFlow-ReID, containing 666 vehicles from 40 cameras. However, existing vehicle Re-ID datasets and methods only devote to the single RGB modality, while neglecting the crucial mission of Re-ID task in adverse illumination conditions, which are even important in social security and video surveillance.

### RGB-IR Cross-Modality Person Re-ID

One pioneer cross-modality person Re-ID problem is the text-to-image person retrieval (Ye et al. 2015; Li et al. 2017a; 2017b), which aimed to search a person with a natural language description. To relieve the illumination limitation in the dark area or nighttime, the RGB-IR cross-modality person Re-ID emerges. Wu et al. (2017) contributed for the first time a standard benchmark RGB-NIR cross-modality Re-ID dataset SYSU-MM01 and proposed a deep zero-padding network for modality shareable feature representations learning. Ye et al. (2018) introduced a two-stage framework to learn sequentially discriminate features and distance metrics for RGB-NIR cross-modality person Re-ID. Recently, Dai et al. (2018) presented a cross-modality generative adversarial network to jointly discriminate the modalities and identities. Ye et al. (2019) designed a dual-path network with bi-directional dual constrained top-ranking loss to learn discriminative feature representations for RGB-NIR cross-modality Re-ID. Nguyen et al. (Nguyen et al. 2017) proposed a novel RGB-TIR dataset RegDB with RGB-TIR image pairs, and it is widely used for cross-modality person Re-ID but only captured by one RGB-T camera. However, the heterogeneous issue across different modalities brings an additional challenge for the Re-ID task.

### RGB-D Multi-Modality Person Re-ID

To integrate additional modality into Re-ID, some works proposed the indoor RGB-Depth (RGB-D) multi-modality datasets (Barbosa et al. 2012; Munaro et al. 2014). John et al. (John, Englebienne, and Krose 2013) combined RGB-Height histogram and gait feature of depth information. Pala et al. (Pala et al. 2016) improved the accuracy of appearance descriptors by fusing them with anthropometric measures extracted from depth data. Wu et al. (Wu, Zheng, and Lai 2017) proposed a locally rotation invariant depth shape descriptor for depth data and empirically fused the traditional RGB features to identify a person. However, it is hard to capture the depth information outdoor which significantly limits the application in real-life surveillance. Furthermore, most of the existing works directly utilize the multi-modality information as auxiliary information or connective feature for Re-ID while ignoring the common expression ability of the multi-modality information.

## RGBN300 Benchmark

To overcome the illumination limitation in the conventional RGB-based vehicle Re-ID, we are the first time to contribute a new dataset named RGBN300 for multi-spectral vehicle Re-ID. We will elaborate on the acquisition details, followed by the dataset constructions and the challenges in our dataset in this section.

### Imaging Platform

The dataset is captured in a campus by eight RGB-NIR camera pairs shooting at eight views of vehicles, as shown in Fig. 2. Our image acquisition platform is based on two HIKVSION cameras, each consisting of one near infrared camera and one visible camera with the same imaging parameters. We manually align the images of two modalities by moving the visible image to totally contain the entire near infrared one. Then, we manually crop the RGB-NIR image pair with a small part of the background with different views.

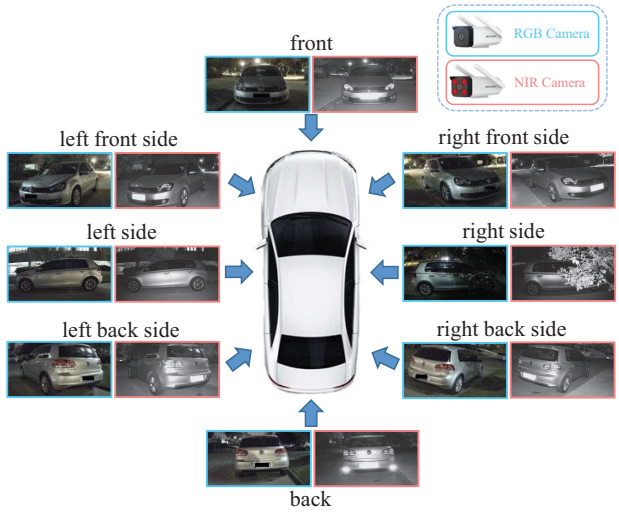
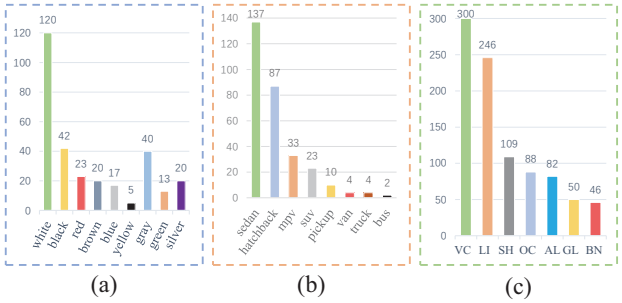



Figure 2: Illustration of eight camera views and corresponding samples in our dataset.

### Dataset Construction

RGBN300 contains 50125 image pairs of 300 vehicle identities in both RGB and near infrared modalities. The number of image pairs of each vehicle varies from 50 to 200. We provide at least 2 and at most 8, coming up with averagely 6.7 camera views per vehicle. We randomly select 150 vehicles with 25200 image pairs as the training set, while the rest 150 vehicles with 24925 image pairs as the testing set (gallery). We further randomly selected 4985 image pairs from the gallery as the query (probe). Fig. 3 (a) and (b) demonstrates the distribution of the attributes of 8 types and 9 colors in RGBN300, which covers the most types and colors of modern vehicles.

**Challenges.** In addition to the common challenges, such as view changes (VC) as Fig. 4 (a, b) and occlusion (OC) as Fig. 4 (c), our dataset contains more challenges including




(a) (b) (c)

Figure 3: The distribution of (a) nine colors, (b) eight types, and (c) seven challenges in RGBN300.

abnormal lighting (AL) and glaring (GL) as shown in Fig. 4 (d) and (f), which significantly affect the color appearance on the vehicle images in RGB spectrum. The poor illumination caused by shadow (SH) (Fig. 4 (e) ), low illumination (LI) (Fig. 4 (g)) and even black night (BN) (Fig. 4 (h)) also bring severe challenges for vehicle Re-ID. The distribution of challenges in our dataset is shown in Fig. 3 (c).

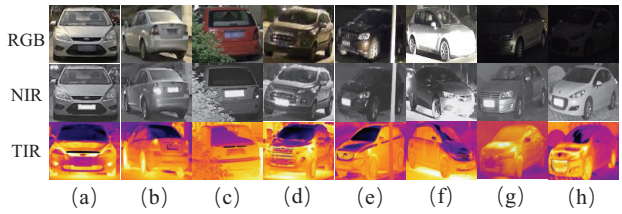



Figure 4: Challenges of our dataset. (a) normal condition. (b) view change (VC). (c) occlusion (OC). (d) abnormal lighting (AL). (e) shadow (SH). (f) glare (GL). (g) low illumination (LI). (h) black night (BN).

**Extension.** To fully verify the robustness of the proposed multi-spectral vehicle Re-ID method in the next section, we have captured additional 17250 TIR (thermal infrared) images of 100 vehicles from RGBN300. Fig. 4 (bottom) shows the TIR image in each challenge scenario. Supplemental TIR data and corresponding RGB-NIR image pairs constitute the dataset with 17250 image triples named RGBNT100 in this paper. In RGBNT100, the training set contains 50 vehicles with 8675 image triples, while the other 50 vehicles with 8575 image pairs for testing/gallery, from which 1715 image triples are randomly selected as the query/probe. From Fig. 4 (bottom) we can see, TIR images reflect the thermal temperature information of the object which is not affected by illumination changes. In particular, both RGB and NIR cameras are with the resolution of $1920 * 1080$, while the TIR camera is $640 * 480$. The cameras from three modalities are with the same frame rate of 25 fps.

## Baseline Approach

We follow a widely used ID-discriminative embedding (IDE) model (Zheng, Zheng, and Yang 2018) as standard baseline, and utilize five state-of-the-art networks as the backbones, including ResNet50 (He et al. 2016), SeResNet50 (Hu, Shen, and Sun 2018), Densenet121 (Huang et al. 2017), InceptionResNetV3 (Szegedy et al. 2017) and MobileNetV2 (Sandler et al. 2018), all of which are pre-trained on ImageNet (Deng et al. 2009) and with the same hyper-parameter settings. The sizes of input images are fixed to $128 \times 256$ for ResNet50, Densenet121 and MobileNetV2, $224 \times 224$ for SeResNet50, and $299 \times 299$ for Inception-ResNetV3. We design a multi-stream convolutional network (MSN) for multi-spectral vehicle Re-ID, where multiple CNN models are trained with the sum of ID losses from each stream (spectrum). Then we sum up the features from each stream/spectrum in the test procedure. The generated feature dimension is between 512 and 2048 for final matching.

# Heterogeneity-collaboration Aware Multi-Stream Convolutional Network

To maintain the similarity between heterogeneous spectra about aligned images and measure the importance of each spectrum for consistent vehicle identity, we propose a Heterogeneity-collaboration Aware Multi-stream convolutional Network (HAMNet) for multi-spectral vehicle Re-ID.

## Network Architecture

As shown in Fig. 5, HAMNet consists of multi-stream equivalent backbones to extract multi-spectral features. To guarantee the predicted ID for each stream to be consistent with the ground truth, we firstly design the multi-stream identification loss $L_{msid}$ by the summation of independent identification loss $L_{id}^{j}$ from the corresponding backbone (stream). Considering the consistency of intrinsic geometric structures among different spectral images, we secondly propose a novel loss function called heterogeneous score coherence loss ($L_{hsc}$) to maintain the score coherence of the multi-spectral information of the same identity. $L_{hsc}$ is summed with $L_{msid}$ called heterogeneity-collaboration ($L_{hc}$) loss, to enforce the similarity consistency among heterogeneous data, in addition to the ID consistency. Meanwhile, inspired by Class Activation Map(CAM) (Zhou et al. 2016), which can indicate the locations of informative parts and features. We use the information parts indicated by CAM to learn the class-aware weights associated with the multi-spectrum. We adaptively fuse class activation maps of multi-spectrum based on the weights, and constrain the final class activation map with a class-aware identification loss ($L_{caid}$). Finally, our HAMNet model is trained by the summation of $L_{hc}$ and $L_{caid}$.

## Loss Functions

**Heterogeneity-Collaboration loss ($L_{hc}$).** Given the vehicle image $I^j$ from the $j$-th spectrum, we can obtain the class activation maps $M^j \in R^{h \times w \times C}$. The global average pooling (GAP) is used to transfer $M^j$ into the class scores $S^j \in R^C$.

$$S^j = GAP(M^j) \tag{1}$$

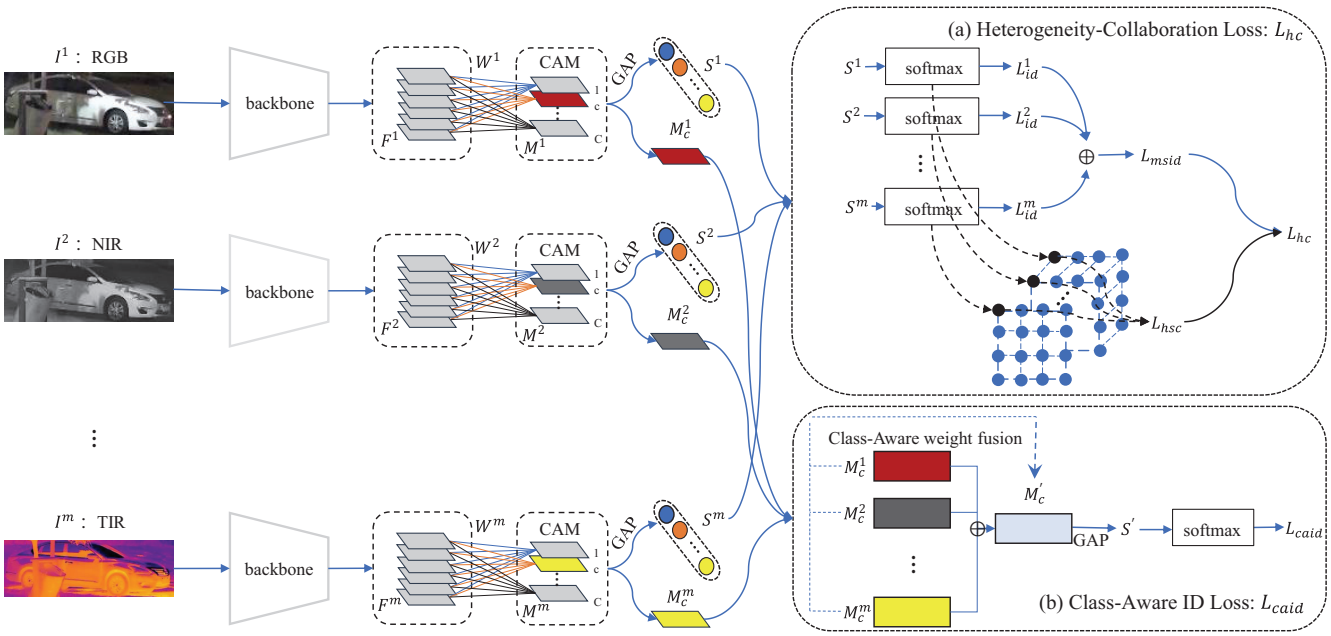



Figure 5: Pipeline of Heterogeneity-Collaboration Aware Multi-Stream Convolutional Network (HAMNet). Given the aligned images $\{I^1, I^2, \ldots, I^m\}$ with label $c$, we first extract the corresponding tensors $\{F^1, F^2, \ldots, F^m\}$ via $m$-stream equivalent backbones. Then, we weight $F^j \in R^{h\times w\times d}, j = \{1, \ldots, m\}$ by $W^j \in R^{d\times C}$ to obtain the class activation maps (CAMs) $M^j \in R^{h\times w\times C}$. The class scores $S^j \in R^{1\times C}$ is obtained by the Global Average Pooling (GAP). In (a), we enforce the multi-stream identification loss $L_{msid}$ by the sum of identification losses $L_{id}^j$s to guarantee the predicted ID from each stream $j$ to be consistent with the ground truth. Meanwhile, we enforce the heterogeneous score coherence ($L_{hsc}$) loss by multiple class scores $\{S^1, S^2, \ldots, S^m\}$ to maintain the score similarity among $m$ streams. The heterogeneity-collaboration ($L_{hc}$) loss is then constructed by $L_{msid}$ and $L_{hsc}$. In (b), we use $M_c^j \in R^{h\times w}$ measures the class-aware weights from each stream $j$ so that the information in different streams can fuse to one class activation map $M_c'$ adaptively. $M_c'$ is constrained by the class-aware identification loss ($L_{caid}$), which enforces the same class information of multi-spectral data. Finally, our HAMNet model is trained by the sum of $L_{hc}$ and $L_{caid}$.

$S^j$ is further normalized by a softmax function into a probability distribution $\hat{p}^j \in R^C$:

$$\hat{p}^j = softmax(S^j) \tag{2}$$

The identification loss corresponding to image $I^j$ is calculated as the cross entropy between the predicted probability $\hat{p}^j$ and the ground-truth class $c$:

$$L_{id}^j = \sum_{i=1}^{C} -p_i^j \log(\hat{p}_i^j), i \in \{1, 2, \ldots, C\} \tag{3}$$

where $\hat{p}^j$ is the predicted probability, $p_i^j$ is the target probability. $p_i^j = 0$ for all $i$ except $p_c^j = 1$.

Our baseline multi-stream convolutional network (MSN) is trained by the multi-stream identification loss $L_{msid}$, which is defined as:

$$L_{msid} = \sum_{j=1}^{m} L_{id}^j \tag{4}$$

Although MSN takes into account the different spectral losses and joint training, these losses are essentially independent of each other. In multi-spectral Re-ID, we argue that the heterogeneous feature maps derived from the multi-spectral images of the same vehicle tend to have similar geometric distribution in their source domains, as shown in Fig. 5. Herein we propose to enforce this constraint on multi-spectral images to obtain the consistent probability distribution. Specifically, we propose a heterogeneous score coherence loss ($L_{hsc}$) to consider the score consistency among multi-spectral features, which is defined as:

$$L_{hsc} = 1 - \sum_{i=1}^{C}(\prod_{j=1}^{m} \hat{p}_i^j) \tag{5}$$

where $\prod$ denotes the element multiplication. $\hat{p}_i^j$ denotes the $i$-th class predicted probability of $j$-th spectrum.

We combine the heterogeneous score coherence loss $L_{hsc}$ and the multi-stream identification loss $L_{msid}$ to form a heterogeneity-collaboration loss ($L_{hc}$) to maintain the data heterogeneity and score coherence of different spectra pointing to the same identity as:

$$L_{hc} = L_{msid} + \alpha L_{hsc} \tag{6}$$

where $\alpha$ is a hyper-parameter setting as 0.001.

**Class-Aware ID loss ($L_{caid}$).** One can directly concatenate the multi-spectral features for multi-spectral vehicle Re-ID. However, it may not be reliable enough because the information provided by different spectra is various in challenge scenarios. Inspired by Class Activation Mapping (CAM), which can highlight the class-specific discriminative regions. We propose to measure the importance of different spectra by CAM, and further obtain one more robust class activation map by different class-aware wights in different spectra. The fused class activation map is finally constrained by the class-aware identification loss $L_{caid}$.

As shown in Fig. 5, after obtaining the CAM, i.e., $M^j \in R^{h\times w\times C}$, for each spectrum, we first normalize it into $0 \sim 1$ by the sigmoid function:

$$A_c^j = \frac{1}{1+\exp(-M_c^j)} \tag{7}$$

The most interesting regions of the heterogeneous activation maps can be expressed as:

$$B_c = \max(A_c^1, A_c^2, \ldots, A_c^m) \tag{8}$$

The response value of each class activation map can be obtained by following operation:

$$\eta_c^j = \sum A_c^j \odot B_c \tag{9}$$

where $\odot$ denotes the element multiplication. The class-aware weight of each spectrum for classification can be measured as:

$$\alpha_c^j = \frac{\eta_c^j}{\sum_{j=1}^m \eta_c^j} \tag{10}$$

The fused class activation map can be expressed as:

$$M_c' = \sum_{j=1}^{m} \alpha_c^j M_c^j \tag{11}$$

After global average pooling (GAP) over the fused class activation map $M_c'$, we obtain the corresponding class score $S_c'$, which is further normalized by a softmax function into a probability value $y_c'$. The class-aware ID loss $L_{caid}$ is employed to guarantee the predicted ID for fused activation map to be consistent with the ground truth:

$$L_{caid} = -\log(y_c') \tag{12}$$

The final objective function for the HAMNet model is defined as the sum of the heterogeneity-collaboration loss $L_{hc}$ and the class-aware ID loss $L_{caid}$.

$$L_{total} = L_{hc} + L_{caid}. \tag{13}$$

# Experimental Results

To evaluate the effectiveness of the proposed HAMNet on multi-spectral vehicle Re-ID, we integrate HAMNet into the five state-of-the-art networks, including ResNet50 (He et al. 2016), SeResNet50 (Hu, Shen, and Sun 2018), Densenet121 (Huang et al. 2017), InceptionResNetV3 (Szegedy et al. 2017) and MobileNetV2 (Sandler et al. 2018).

## Experimental Settings

**Evaluation Metrics.** Following (Zheng, Zheng, and Yang 2018), we use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) for evaluation. CMC scores reflect the retrieval precision, where Rank-1 (R-1), Rank-5 (R-5), Rank-10 (R-10) scores are reported in our experiments. mAP measures the mean of all queries of average precision (the area under the Precision Recall curve) which reflects the recall.

**Implementation Details.** We adopt the network that pretrained on ImageNet (Deng et al. 2009) as the backbone. The common backbone in the different streams does not share parameters. The classifier weights are randomly initialized. For data augmentation, standard random cropping and horizontal flipping are used during training. The Adam (Kingma and Ba 2014) optimizer is used with the batch size of 16. We use warmup (Fan et al. 2019) to bootstrap the network, which spent 10 epochs linearly increasing the learning rate from $3.5\times10^{-5}$ to $3.5\times10^{-4}$. The learning rate is decayed to $3.5\times10^{-5}$ and $3.5\times10^{-6}$ at 40-th epoch and 70-th epoch respectively. Our model is trained in total 120 epochs.

## Evaluation on RGBN300 Dataset

Table 2 reports the evaluation results of five backbones on RGBN300 dataset. One-stream denotes the Re-ID on the corresponding spectrum, e.g., single-spectral Re-ID. MSN denotes the multi-stream network for multi-spectral Re-ID without the heterogeneous score coherence loss and the class-aware ID loss, which is the baseline approach of our HAMNet.

From which we can see, i) The accuracy on the RGB spectrum clearly outperforms the NIR spectrum as expected on all the five backbones, since RGB images contain richer appearance information than NIR ones. ii) By integrating NIR to RGB, the MSN significantly improves the performance which verifies the effectiveness of the multi-spectral information in vehicle Re-ID. iii) By introducing the heterogeneous score coherence loss and the class-aware ID loss, our HAMNet consistently beats both MSN on RGB-NIR multi-spectral vehicle Re-ID and the single spectrum ones, which validates the contrition of our method.

## Evaluation on RGBNT100 Dataset

To further verify the robustness of multi-spectral methods, we additionally evaluate our HAMNet on the extended RGB-NIR-TIR dataset RGBNT100. The evaluation results on RGBNT100 are shown in Table 3. From which we can see, i) By integrating NIR or TIR to RGB, the multi-spectral results of MSN on either RGB-NIR or RGB-TIR outperforms the single-spectrum ones on RGB, NIR or TIR spectrum, which consistently verifies the effectiveness of multi-spectrum information for vehicle Re-ID. ii) By integrating TIR to RGB-NIR, MSN on RGB-NIR-TIR significantly improves the performance, which further confirms the contribution of multi-spectrum information. iii) Our HAMNet further boosts the performance on RGB-NIR-TIR comparing to MSN, which validates the effectiveness of the proposed method.

Table 2: State-of-the-art networks for vehicle Re-ID on RGBN300 (in %).

| Modality | | ResNet50 | | | | SeResNet50 | | | | Densenet121 | | | | InceptionResNetV3 | | | | MobileNetV2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 |
| RGB | one-stream | 49.5 | 72.6 | 76.4 | 78.6 | 48.2 | 72.9 | 76.2 | 78.1 | 36.1 | 61.4 | 65.1 | 67.3 | 41.1 | 67.6 | 70.7 | 72.5 | 44.9 | 68.7 | 73.4 | 76.0 |
| NIR | one-stream | 42.1 | 61.9 | 65.4 | 67.5 | 41.2 | 65.1 | 67.3 | 68.5 | 28.8 | 50.5 | 53.2 | 55.2 | 34.4 | 57.5 | 61.0 | 64.0 | 39.1 | 62.7 | 67.1 | 69.1 |
| RGB-N | MSN | 56.9 | 77.2 | 79.9 | 81.4 | 56.9 | 80.5 | 82.1 | 83.2 | 46.1 | 74.0 | 76.2 | 77.6 | 52.8 | 77.1 | 78.8 | 80.2 | 54.3 | 78.9 | 81.2 | 82.7 |
| RGB-N | HAMNet | **61.9** | **84.0** | **86.0** | **87.0** | **61.8** | **84.3** | **86.6** | **87.9** | **46.7** | **75.1** | **76.9** | **78.2** | **56.6** | **82.0** | **84.2** | **85.5** | **56.9** | **80.0** | **82.1** | **83.5** |

Table 3: State-of-the-art networks for vehicle Re-ID on RGBNT100 (in %).

| Modality | | ResNet50 | | | | SeResNet50 | | | | Densenet121 | | | | InceptionResNetV3 | | | | MobileNetV2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 |
| RGB | one-stream | 41.0 | 58.5 | 63.6 | 66.9 | 39.1 | 57.9 | 60.5 | 62.2 | 35.2 | 54.4 | 58.5 | 61.3 | 37.5 | 57.1 | 62.2 | 64.9 | 37.3 | 58.1 | 63.3 | 66.4 |
| NIR | one-stream | 37.1 | 52.8 | 56.4 | 59.1 | 37.3 | 53.2 | 56.4 | 58.4 | 27.2 | 43.8 | 46.0 | 47.8 | 31.1 | 51.6 | 56.2 | 58.3 | 36.9 | 56.9 | 62.4 | 65.0 |
| TIR | one-stream | 35.7 | 61.8 | 66.5 | 69.9 | 39.1 | 67.1 | 72.6 | 74.9 | 27.5 | 50.8 | 56.3 | 59.6 | 40.5 | 70.8 | 75.3 | 78.2 | 39.9 | 68.3 | 73.1 | 75.6 |
| RGB-N | MSN | 43.1 | 65.4 | 70.6 | 73.6 | 43.7 | 64.6 | 68.2 | 70.4 | 40.9 | 61.5 | 63.2 | 64.9 | 45.9 | 71.8 | 75.6 | 77.6 | 48.2 | 70.3 | 73.4 | 75.5 |
| RGB-T | MSN | 56.2 | 80.7 | 83.3 | 86.2 | 57.3 | 81.4 | 83.2 | 85.4 | 45.9 | 73.8 | 76.9 | 78.4 | 49.3 | 78.0 | 82.4 | 85.5 | 52.9 | 79.1 | 82.8 | 84.7 |
| RGB-N-T | MSN | 60.5 | 82.6 | 85.7 | 87.1 | 60.0 | 82.6 | 85.2 | 87.3 | 51.0 | 76.8 | 78.9 | 80.1 | 53.9 | 81.2 | 84.1 | 86.1 | 57.1 | 82.3 | 85.7 | 87.8 |
| RGB-N-T | HAMNet | **64.1** | **84.7** | **88.0** | **89.4** | **65.4** | **85.5** | **87.9** | **88.8** | **54.3** | **80.9** | **83.1** | **84.8** | **56.3** | **85.0** | **86.0** | **87.6** | **58.2** | **83.9** | **86.2** | **88.5** |

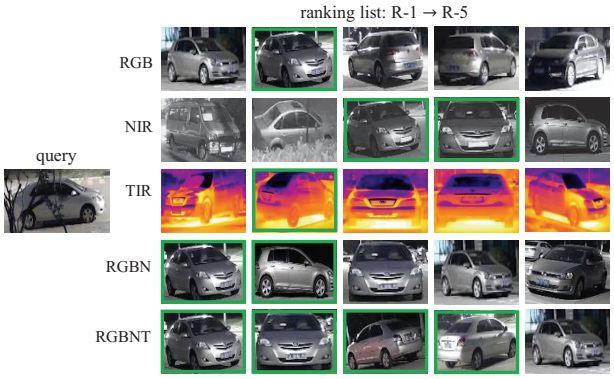



Figure 6: Top 5 ranking results on RGBNT100. The images with green bounding boxes and the rest ones indicate the correct and wrong matchings respectively.

## Comparing to State-of-the-art Methods

As the first work on multi-spectral vehicle Re-ID, we extend the state-of-the-art RGB single spectrum Re-ID methods to multi-spectral cases for comparison. Specifically, we first train the methods on the three spectra respectively and then directly sum up the learned deep features from each spectrum for testing. Table 4 reports the results comparing to three state-of-the-art Re-ID methods, including PCB (Sun et al. 2018), MGN (Wang et al. 2018) and ABD (Chen et al. 2019)) on RGBN300 and RGBNT100.Note that few of vehicle Re-ID methods have released their codes, we herein extend the recently advanced methods on person Re-ID task, which shares a common objective as vehicle Re-ID. From Table 4 we can see that the compared methods achieve good performance at Rank-1 due to their powerful capabilities of feature extraction. However, the mAPs are overshadowed comparing to our HAMNet, which implies that, HAMNet can better capture the complementary information provided among multiple heterogeneity spectral information even on the simple backbone (ResNet-50).

Table 4: Comparing to State-of-the-art Re-ID methods (in %). (Backbone: ResNet-50).

| Methods | RGBN300 | | RGBNT100 | |
|---|---|---|---|---|
| | mAP | Rank-1 | mAP | Rank-1 |
| PCB (ECCV2018) | 57.7 | 82.0 | 57.2 | 83.5 |
| MGN (MM2018) | 60.5 | 83.7 | 58.1 | 83.1 |
| ABD (ICCV2019) | 58.9 | 83.1 | 60.4 | 85.1 |
| OURS | 61.9 | 84.0 | 64.1 | 84.7 |

Table 5: Ablation study on RGBN300 and RGBNT100 (in %). (Backbone: ResNet-50).

| Models | RGBN300 | | RGBNT100 | |
|---|---|---|---|---|
| | mAP | Rank-1 | mAP | Rank-1 |
| baseline (MSN) | 56.9 | 77.2 | 60.5 | 82.6 |
| + $L_{hsc}$ | 59.1 | 80.1 | 61.4 | 83.4 |
| + $L_{caid}$ | 59.4 | 79.2 | 61.4 | 83.1 |
| + $L_{hsc}$ + $L_{caid}$ | 61.9 | 84.0 | 64.1 | 84.7 |

## Ablation Study

To verify the contribution of the components in our model, we implement the ablation study of several variants of our method on RGBN300 and RGBNT100 datasets, as reported in Table 5. Note that, both heterogeneous score coherence loss $L_{hsc}$ and class-aware ID loss $L_{caid}$ surpass the baseline MSN, which demonstrates the contribution of each loss. By simultaneously enforcing both losses, our HAMNet can further boost the performance.

## Summary

To our best knowledge, this is the first work to identify the RGB-IR multi-spectral vehicle Re-ID problem. We have contributed two new multi-spectral Re-ID datasets, together with a novel multi-spectral Re-ID method. Comparing with RGB-RGB vehicle Re-ID, additional infrared im-

ages can help identify vehicles in adverse illumination conditions. Extensive experiments demonstrate the promising performance of the proposed method. In addition, based on evaluation results, we highlight some critical observations for RGB-IR multi-spectral vehicle Re-ID. First, the ability of heterogeneous data to identify the same ID is worth considering. Second, adaptive fusion is effective. Third, powerful feature representations are essential for high-performance achievement. Improving above mentioned components will further advance the state-of-the-art of RGB-IR multi-spectral vehicle Re-ID.

## Acknowledgement

This research is supported in part by the National Natural Science Foundation of China (Nos. 61976002, 61976003, 61860206004 and 61671018), the National Laboratory of Pattern Recognition (NLPR) (201900046), and the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033).

## References

Barbosa, I. B.; Cristani, M.; Del Bue, A.; Bazzani, L.; and Murino, V. 2012. Re-identification with rgb-d sensors. In *ECCV Workshops*, 433–442.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *Proc.of ICCV*, 8351–8361.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *Proc.of IJCAI*, 677–683.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc.of CVPR*, 248–255.

Dong, S. C.; Cristani, M.; Stoppa, M.; Bazzani, L.; and Murino, V. 2011. Custom pictorial structures for re-identification. In *Proc.of BMVC*, volume 1, 6.

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

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshops*, volume 3, 1–7.

Guo, H.; Zhao, C.; Liu, Z.; Wang, J.; and Lu, H. 2018. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proc.of AAAI*, 6853–6860.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc.of CVPR*, 770–778.

He, B.; Li, J.; Zhao, Y.; and Tian, Y. 2019. Part-regularized near-duplicate vehicle re-identification. In *Proc.of CVPR*, 3997–4005.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proc.of CVPR*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proc.of CVPR*, 4700–4708.

John, V.; Englebienne, G.; and Krose, B. 2013. Person re-identification using height-based gait in colour depth camera. In *Proc.of ICIP*, 3345–3349.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Proc.of ICCV*, 1890–1899.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017b. Person search with natural language description. In *Proc.of CVPR*, 1970–1979.

Liu, H.; Tian, Y.; Yang, Y.; Pang, L.; and Huang, T. 2016a. Deep relative distance learning: Tell the difference between similar vehicles. In *Proc.of CVPR*, 2167–2175.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016b. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proc.of ECCV*, 869–884.

Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; and Duan, L.-Y. 2019. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing* 28(8):3794–3807.

Munaro, M.; Fossati, A.; Basso, A.; Menegatti, E.; and Van Gool, L. 2014. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*. 161–181.

Nguyen, D.; Hong, H.; Kim, K.; and Park, K. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17(3):605.

Pala, F.; Satta, R.; Fumera, G.; and Roli, F. 2016. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology* 26(4):788–799.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc.of ECCV*, 17–35.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc.of CVPR*, 4510–4520.

Shen, Y.; Xiao, T.; Li, H.; Yi, S.; and Wang, X. 2017. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proc.of ICCV*, 1900–1909.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc.of ECCV*, 480–496.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc.of AAAI*, 4278–4284.

Tang, Z.; Naphade, M.; Liu, M.-Y.; Yang, X.; Birchfield, S.; Wang, S.; Kumar, R.; Anastasiu, D.; and Hwang, J.-N. 2019. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proc.of CVPR*, 8797–8806.

Wang, Z.; Tang, L.; Liu, X.; Yao, Z.; Yi, S.; Shao, J.; Yan, J.; Wang, S.; Li, H.; and Wang, X. 2017. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proc.of ICCV*, 379–387.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, 274–282.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *Proc.of ICCV*, 5380–5389.

Wu, A.; Zheng, W. S.; and Lai, J. H. 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*. 26(6):2588–2603.

Yan, K.; Tian, Y.; Wang, Y.; Zeng, W.; and Huang, T. 2017. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *Proc.of ICCV*, 562–570.

Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; Chen, J.; and Liu, J. 2015. Specific person retrieval via incomplete text description. In *Proc.of ICMR*, 547–550.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. C. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Proc.of AAAI*, 7501–7508.

Ye, M.; Lan, X.; Wang, Z.; and Yuen, P. C. 2019. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security* 1–13.

Zheng, W. S.; Gong, S. G.; and Xiang, T. 2009. Associating groups of people. In *Proc.of BMVC*, 1–11.

Zheng, Z.; Zheng, L.; and Yang, Y. 2018. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14(1):13.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proc.of CVPR*, 2921–2929.