This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UCM-VeID V2: A Richer Dataset and A Pre-training Method for UAV Cross-Modality Vehicle Re-Identification

Xingyue Liu Jiahao Qi Chen Chen KangCheng Bin Ping Zhong*
National University of Defense Technology, China
{liuxingyue18, qijiahao1996, chenchen21c, binkc21, zhongping}@nudt.edu.cn

Abstract

Cross-Modality Re-Identification (VT-ReID) aims to achieve around-the-clock target matching, benefiting from the strengths of both RGB and infrared (IR) modalities. However, the field is hindered by limited datasets, particularly for vehicle VT-ReID, and by challenges such as modality bias training (MBT), stemming from biased pre-training on ImageNet. To tackle the above issues, this paper introduces an dataset benchmark, named UCM-VeID V2, for vehicle VT-ReID, and proposes a new self-supervised pre-training method, Cross-Modality Patch-Mixed Self-supervised Learning (PMSL). UCM-VeID V2 dataset features a significant increase in data volume, along with enhancements in multiple aspects. PMSL addresses MBT by learning modality-invariant features through Patch-Mixed Image Reconstruction (PMIR) and Modality Discrimination Adversarial Learning (MDAL), and enhances discriminability with Modality-Augmented Contrasting Cluster (MACC). Comprehensive experiments are carried out to validate the proposed method.

1. Introduction

Cross-modality vehicle re-identification (ReID), also referred to as vehicle visible-infrared (VI) ReID [22], involves retrieving a matching vehicle image from a gallery set captured in the visible (RGB) or infrared (IR) spectrum, given a query vehicle image in the alternate modality. This process, where images are obtained from cameras of differing modalities, has garnered significant attention in applications such as video surveillance and public security [1, 28], primarily due to its advantages in enabling continuous, intelligent monitoring across day and night conditions.

Despite its importance, research on vehicle VT-ReID remains relatively underdeveloped. A primary challenge lies in the lack of essential resources, including a comprehensive large-scale dataset and widely accessible benchmark results. Existing remote sensing vehicle ReID datasets, such as VRAI [31], UAV-VeID [29], VRU [24], and VeRi-UAV [19], primarily focus on the single RGB modality, lacking the infrared data necessary for cross-modality analysis. For this, [22] has introduced the first UAV RGB-IR vehicle ReID dataset, UCM-VeID, to support cross-modality vehicle ReID research. However, this dataset is constrained by its single data acquisition platform and limited number of

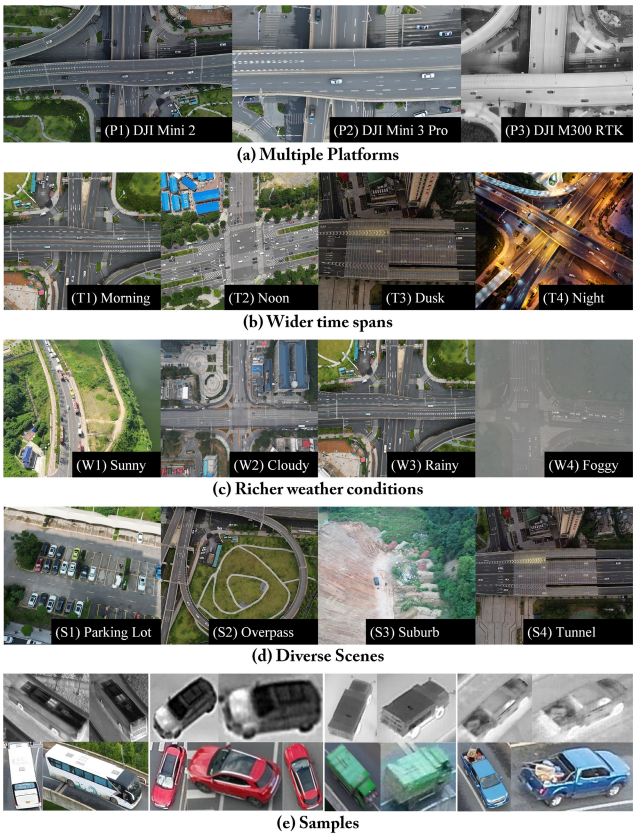


Figure 1. Increments and samples of UCM-VeID V2. (a)(b)(c)(d) are varies platforms, time periods, weather conditions and locations for data acquisition, respectively. (e) shows the 4 IDs with different modalities, view-points and scales samples.

*Corresponding author

identities, reducing its suitability for generalizing to diverse real-world scenarios. Consequently, developing large-scale and more diverse vehicle VT-ReID datasets is essential to advance research and better meet the requirements of practical applications.

Another key challenge in vehicle VT-ReID development is the need for further research on VT-ReID methods, particularly to address the modality bias training (MBT) problem, which remains under-explored. As introduced by [16], MBT refers to the modality imbalance that arises when VT-ReID models are initialized with pre-trained ImageNet weights [6]. The modality imbalance stems from training on RGB-exclusive ImageNet data, which skews feature learning towards the RGB modality and reduces the IR discrimination capability. Moreover, the MBT issue is amplified in large-scale multi-modal datasets: as shown in Fig. 2, models initialized with ImageNet pre-trained weights exhibit a performance drop on larger datasets, particularly in terms of discriminative ability. The potential underlying cause of the above phenomenon is that ImageNet pre-trained weights perform well when handling infrared data with lower modality diversity. However, they lack sufficient generalization ability when faced with larger datasets that exhibit higher modality diversity, which exacerbates modality bias and degrades model performance.

To address these challenges, this paper introduces a more comprehensive dataset for vehicle VT-ReID and a novel pre-training paradigm called Cross-modality Patch-mixed Self-supervised Learning (CPSL) to mitigate the MBT problem. The proposed dataset, UCM-VeID V2, includes 113,929 RGB images and 51,162 IR images of 3,407 vehicles captured by five cameras. Compared to UCM-VeID [22], it is significantly expanded across multiple dimensions, including additional cameras and acquisition platforms, extended capture times, a broader range of weather conditions, and more diverse shooting scenarios, as illustrated in Fig. 1. For the MBT challenge, inspired by [13], CPSL aligns the RGB and IR domains by incorporating a Patch-Mixed Image Reconstruction (PMIR) self-supervised task. To further eliminate the modality bias, CPSL designs a Modality Discrimination Adversarial Learning (MDAL) strategy to exact modality invariant features through an adversarial training manner. In addition, considering discrimination insufficiency, we introduce the Modality-Augmented Contrasting Learning (MACL) task, where patch-mixed images are used as augmented positive examples for contrastive learning.

In summary, our main contributions are as follows:

- We introduce the UCM-VeID V2 dataset, which significantly enhances the original UCM-VeID with greater scale and diversity across multiple dimensions.
- We propose a novel self-supervised paradigm to pre-train an encoder, replacing the ImageNet-trained ResNet50, to address the MBT issue.

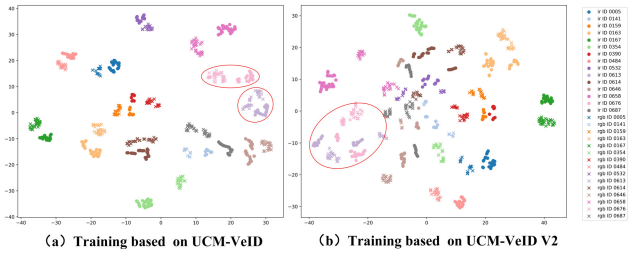


Figure 2. The feature distribution of pre-trained embeddings is visualized. Panels (a) and (b) show the embeddings in 2D space for models trained on the UCM-VeID and UCM-VeID V2 datasets, respectively, with dots and crosses representing different identities in RGB and IR modalities. The experiment uses the AGW baseline [37]. As indicated by the red circles in (a), the same-ID samples are more compact. In contrast, the same-ID samples are more spread out in (b), which ultimately degrades Re-ID performance.

2. Related Works

2.1. Vehicle Re-ID Datasets

Vehicle Re-ID research has flourished over the past few years, significantly benefited by the availability of vehicle Re-ID datasets. The most widely-employed datasets are primarily captured via fixed surveillance cameras, such as VehicleID [20], VeRI-776 [21], VERI-Wild [23] and Vehicle-1M [10]. Compared to fixed surveillance, UAV-based cameras offer superior mobility, flexibility, and convenience, leading to the development of key UAV-based vehicle Re-ID datasets such as VRAI [31], UAV-VeID [29], VRU [24], and VeRi-UAV [19]. Furthermore, Liu *et al.* [22] have advanced the first RGB-IR vehicle Re-ID dataset, named UCM-VeID, to support the research of cross-modality vehicle Re-ID. However, the UCM-VeID dataset exhibits limitations in both size and diversity, highlighting an urgent need for the further expansion and enrichment of the dataset, ensuring it remains a potent benchmark for advancing the frontiers of vehicle Re-ID research in the face of evolving challenges and complexities.

2.2. Modality Bias Training

The MBT problem is essentially a problem of unbalanced training. Imbalanced model learning has received extensive attention in multi-modal learning [15, 32], and great progresses have been made in VT-ReID [16, 30]. To our best knowledge, there are two distinct approaches to tackle the MBT issue. The first approach is to control the training process, such as learning rate adjusting [15], gradient modulation [27], and ID-exclusive sampling strategy [16]. The second approach is to use modal balanced weights rather than ImageNet pre-trained ResNet50 avoiding MBT problems from the beginning. Wan et al. [30] devised a self-

Dataset	Venue	Identities	Images	Modalities	Cameras	Platform	Camera View	Target State	Color	Type	Orientation
VRAI	ICCV2019	13022	137613	RGB	2	UAV	Flexible	Move	9	7	—
UAV-VeID	IJCV2021	4601	41917	RGB	2	UAV	Flexible	Move	—	—	4
UCM-VeID	TMM2024	753	29928	RGB/IR	2	UAV	Flexible	Move&Static	12	9	8
UCM-VeID V2	ours	3407	165091	RGB/IR	5	UAV	Flexible	Move&Static	12	9	8

Table 1. Comparisons among other vehicle Re-ID datasets and UCM-VeID V2

supervised pre-training method to obtain a set of new unbiased weights, which achieves comparable performance to that of ImageNet-trained ResNet50. The second approach addresses the MBT problem from the outset by using pre-trained weights that are modally balanced, instead of relying on subsequent adjustments. In contrast to the first approach, the second method more fundamentally resolves the modality bias, offering stronger generalization and more stable performance. Consequently, we follow the second approach to carry out the MBT problem.

2.3. Self-Supervised Learning

Self-supervised learning, as a subset of unsupervised learning, constructs pretext tasks to extract valuable insights from large-scale unlabeled data, thereby learning representations beneficial for downstream tasks. Common methods involve contextual pretext tasks to learn discriminative features, such as Word2Vec [25], Jigsaw [26], Rotation [8], Colorization [38], and MAE [14]. Another prominent approach, contrastive learning, builds discriminative representations by comparing instances within the same feature space, as seen in methods like MoCo [12] and BYOL [9], both of which perform well on downstream tasks. Nevertheless, above methods are not specifically tailored for the Re-ID task, particularly for VT-ReID task. In this paper, we leverage the consistency of different modalities to construct a patch-mixed image reconstruction self-supervised task aimed at extracting shared features, thereby mitigating the modality discrepancy.

3. Benchmark: UCM-VeID V2

To further enhance the UAV cross-modal vehicle re-identification dataset, we have significantly expanded the first edition, UCM-VeID [22], by an order of magnitude in terms of UAV platforms, identities and so on. This section details the expanded dataset called UCM-VeID V2.

3.1. Dataset Description

(1) **Dataset Partition**. UCM-VeID V2 comprises a total of 3407 identities with 165091 samples. The dataset is split into a training set and a test set in an $8:2$ ratio, resulting in a training set with 2,693 identities and 131,451 samples, while the test set contains 674 identities and 33,640 samples. Furthermore, the query set and gallery set are derived from the test set. Specifically, one sample is drawn from each ID within the test set to constitute the query set, while the remaining samples collectively form the gallery set. It is worth noting that the extraction process is random, and crucially, every sample in the query set has a corresponding sample with the same ID in the gallery set.

(2) **Dataset Statistics**. To maximize the utility of the UCM-VeID V2 dataset, we perform a detailed statistical analysis of its key attributes, providing insights into the data patterns relevant to Re-ID tasks.

- **ID distribution of 5 cameras**. Data is collected using 3 drone platforms equipped with 5 different cameras. $D300_z$ camera captures fewer IDs compared to the other four cameras, as Fig. 3(a) shown. The $D300_z$ camera captures high-resolution images, primarily focused on nighttime data, thereby enhancing the dataset's diversity.
- **Distribution of ID samples**. Samples are collected from different camera perspectives, ensuring diversity and ample representation for each ID. As shown in Fig. 3(b), most IDs have 20-40 samples, providing comprehensive coverage of target's appearance across different conditions.
- **Distribution of Resolution**. As shown in Fig. 3(c), resolutions vary significantly across cameras. For example, $D300_ir$ samples cluster around 10,000-15,000 pixels, while $DM3P$ samples are predominantly 45,000-50,000 pixels, highlighting a resolution disparity that presents a core challenge in UAV-based VT-ReID.
- **Energy Distribution of IR Samples**. The dataset spans all seasons and covers a 24-hour period, capturing both dynamic and static vehicles, leading to diverse infrared energy levels. Fig. 3(d) reveals that most infrared energy values are between 80-100, corresponding to moderate daylight. Extreme conditions, like intense heat or cold, are less common but contribute to the dataset's breadth.

(3) **Dataset Attributes**. Attributes annotation is an integral part of building reliable, high-performance data sets. In this dataset, we annotate the below attributes.

- **Vehicle Type**. Vehicles are classified into categories such as car, SUV, minibus, MPV, bus, mini truck, truck, and pickup. As shown in Fig. 3(e), cars and SUVs make up the majority, while trucks and pickups are less frequent. This distribution reflects the urban focus of our data collection, where cars and SUVs are more prevalent.
- **Vehicle color**. We categorize vehicle colors into thirteen

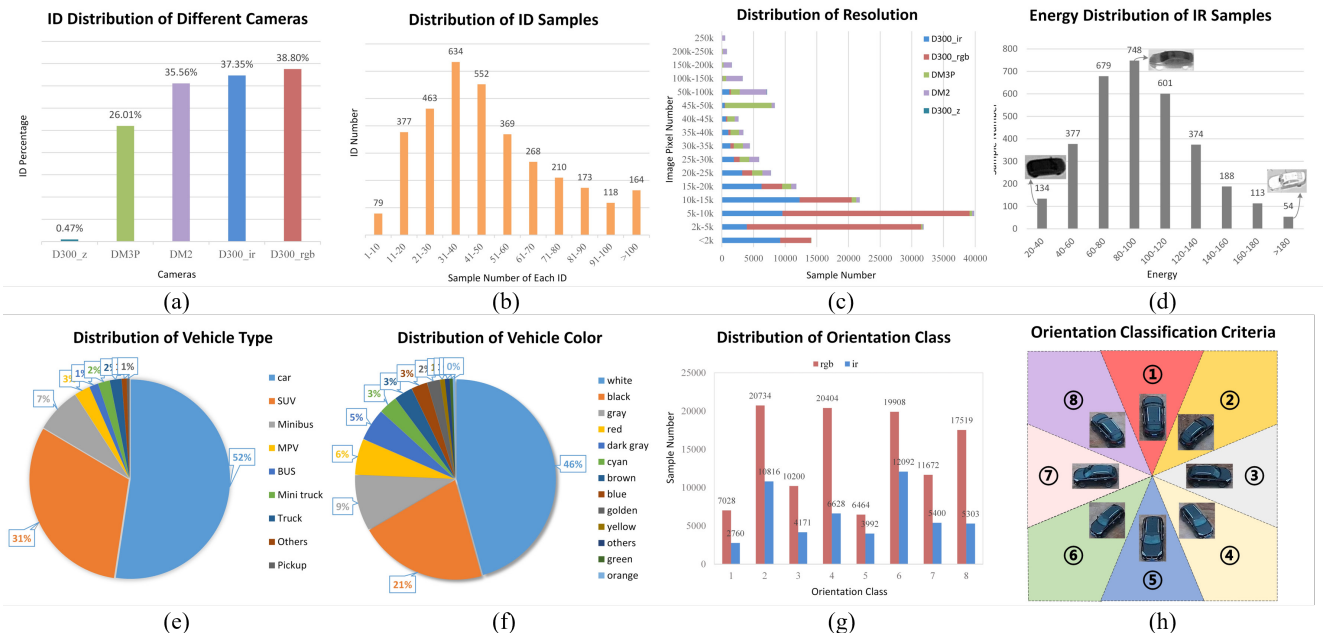


Figure 3. The statistics and annotation attributes of UCM-VeID V2 dataset. Notation that, the 'k' in (c) means the unit of 1000 pixels.

classes, shown in Fig. 3(f), ordered by proportion as follows: white, black, gray, red, dark gray, cyan, brown, blue, golden, yellow, others, green, and orange. Despite some imbalance, this distribution mirrors real-world vehicle color trends.

- **Vehicle Orientation**. Based on [22] and Fig. 3(h), we define eight orientation categories. Fig. 3(g) shows that orientations 2, 4, 6, and 8 are most common, a result of the UAV's flexible viewing angles. This orientation imbalance highlights the dataset's realism and the challenges in UAV-based Re-ID tasks.

3.2. Comparison with Other Datasets

In contrast to other vehicle Re-ID datasets, as detailed in Tab. 1, our dataset has several unique characteristics:

1. **Multi modal information**. Compared with RGB-based datasets, such as VRAI [31], UAV-VeID [29], UCM-VeID [22] and UCM-VeID V2 integrate both RGB and IR modalities, facilitating cross-modality vehicle Re-ID tasks, which holds promise as a viable solution to achieve round-the-clock vehicle Re-Identification.
2. **Flexible viewpoint, scale and orientations**. The UCM-VeID V2 dataset, captured via a drone-mounted camera, offers varied perspectives, scales, and orientations that differ from fixed-camera datasets like RGBN300 [18] and MSVR310 [40]. This diversity substantially enriches the dataset, making it a valuable resource for robust vehicle Re-ID research.
3. **Overall enhancement**. Compared to UCM-VeID, the UCM-VeID V2 dataset offers a substantial increase in data volume and significant improvements in several dimensions, including additional cameras and platforms, extended capture times, a broader variety of weather conditions, and a wider range of shooting scenarios, as shown in Fig. 1. These enhancements make it a richer and more comprehensive dataset, supporting diverse and robust data-driven research and applications.

4. Proposed pre-training method

4.1. Overall Model Structure.

The proposed pre-training VT-ReID method is illustrated in Fig. 4. Firstly, style learning and image generation techniques are used to create pairs of RGB and IR images. Then, we segment these images into patches and randomly swap patches between different modalities to create two new patch-mixed images. Finally, the patch-mixed images are fed into the encoder to complete Patch-Mixed Image Reconstruction (PMIR), Modality Discrimination Adversarial Learning (MDAL) and Modality-Augmented Contrasting Learning (MACL) tasks, where PMIR and MDAL are adopted to learn modality-invariant features, and MACL is utilized to enhance ID discriminability.

4.2. Patch-Mixed Unified Embedding

Cross-modality data generation. Given the limited size of Re-ID dataset, we first train a CycleGAN [41] to generate cross-modality versions of ImageNet. Specifically, we use CycleGAN to learn style mapping between RGB and IR images from the UCM-VeID V2 dataset with a gener-

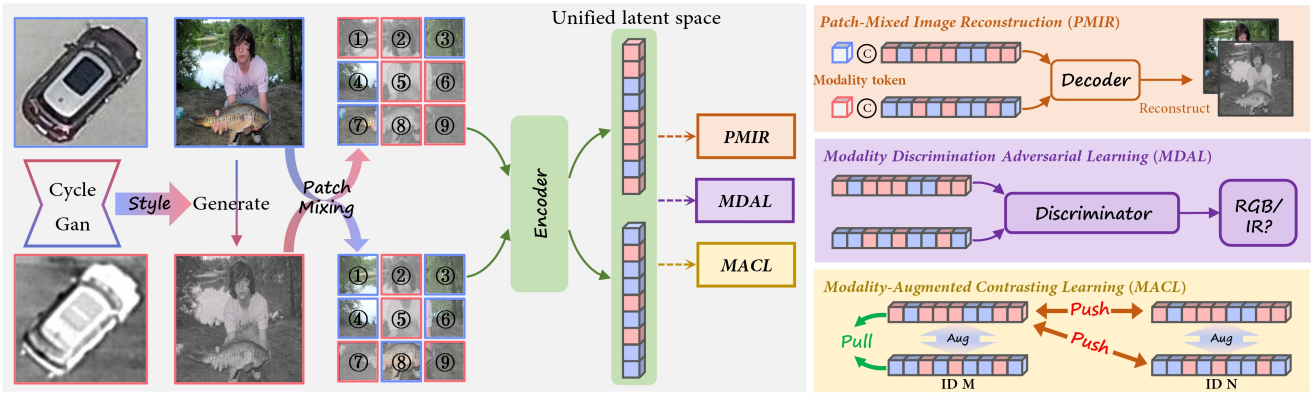


Figure 4. Pre-training paradigm, which comprises preprocessing and three self-supervised tasks. Preprocessing involves obtaining paired RGB and IR images through style learning and image generation, followed by patches segmentation and swapping, PMIR and MDAL then focus on learning modality-invariant features, while MACL aims to enhance the ID discriminability.

ator G. To avoid data leakage, we use ImageNet dataset for pre-training. For an input RGB image $\boldsymbol{x}_i^{\text{RGB}}$ in the ImageNet, we generate the corresponding IR image as $\boldsymbol{x}_i^{\text{IR}} = G(\boldsymbol{x}_i^{\text{RGB}})$. The resulting cross-modality ImageNet dataset is then $\{(\boldsymbol{x}_i^{\text{RGB}}, \boldsymbol{x}_i^{\text{IR}})\}_{i=1}^{N}$, where N is the total number.

Random patch mixture. Our method aims to teach the backbone to fully leverage the cross-modality information for modality bias elimination. Therefore, for the data input aspect, we mixture the RGB and IR image pair in a patch swapping manner. Mathematically, given a RGB-IR image pair $(\boldsymbol{x}_i^{\text{RGB}}, \boldsymbol{x}_i^{\text{IR}})$, we first establish a patch mask $\boldsymbol{M}$ with the same size as $\boldsymbol{x}_i^{\text{RGB}}$ ($\boldsymbol{x}_i^{\text{IR}}$) that consists of $K \times K$ regular non-overlapping patches. With a predefined mask ratio η, $\lfloor \eta \times K \times K \rfloor$ patches in $\boldsymbol{M}$ are randomly selected following a uniform distribution and set as all-zeros matrices. In this paper, we predefined a mask ratio $\eta = 0.75$. A higher η leads to more patch exchanges, increases the difficulty of the reconstruction task, and enables the model to learn more shared features. Then, the process of the proposed random path mixture can be formulated as:

$$\begin{aligned} \hat{\boldsymbol{x}}_i^{\text{RGB}} &= \boldsymbol{x}_i^{\text{RGB}} \odot \boldsymbol{M} + \boldsymbol{x}_i^{\text{IR}} \odot (1 - \boldsymbol{M}) \\ \hat{\boldsymbol{x}}_i^{\text{IR}} &= \boldsymbol{x}_i^{\text{IR}} \odot \boldsymbol{M} + \boldsymbol{x}_i^{\text{RGB}} \odot (1 - \boldsymbol{M}), \end{aligned} \tag{1}$$

Then, based on the Eq. (1), the cross-modality patch-mixed unified embeddings are finally achieved as:

$$\boldsymbol{f}_i^{\text{RGB}}, \boldsymbol{f}_i^{\text{IR}} = \boldsymbol{F}(\hat{\boldsymbol{x}}_i^{\text{RGB}}), \boldsymbol{F}(\hat{\boldsymbol{x}}_i^{\text{IR}}), \tag{2}$$

where $\boldsymbol{F}(\cdot)$ is the feature extractor and ResNet50 [11] is employed in this paper. $\boldsymbol{f}_i^{\text{RGB}}$ and $\boldsymbol{f}_i^{\text{IR}}$ are the latent embeddings of mixed images.

4.3. Self-supervised Multi-task Learning.

With the proposed unified embedding module, we can acquire the latent features containing cross-modality information. However, these features can not be directly employed to the Re-ID task due to modality-bias information and insufficient discriminability [30, 34, 37]. To address above these issues, we devise three self-supervised tasks.

Patch-Mixed Image Reconstruction. In this task, we utilize the mixed features, $\boldsymbol{f}_i^{\text{RGB}}$ and $\boldsymbol{f}_i^{\text{IR}}$, to reconstruct the input images, $\boldsymbol{x}_i^{\text{RGB}}$ and $\boldsymbol{x}_i^{\text{IR}}$, respectively. This approach is motivated by the inherent advantage of image reconstruction as a naturally unsupervised learning task, eliminating the need for additional labels and making it suitable for training large models. Moreover, for RGB image reconstruction, $\boldsymbol{f}_i^{\text{RGB}}$ contains minimal RGB information but abundant IR information, compelling the network to focus on modality-shared features. The limited RGB information acts as style information that supports the reconstruction of the RGB image. Similarly, for IR image reconstruction, $\boldsymbol{x}_i^{\text{IR}}$ using $\boldsymbol{f}_i^{\text{IR}}$, where the feature contains mostly RGB information with only a small amount of IR data, thus also ensuring the model learns to emphasize cross-modality features effectively. According to above analysis, the process of patch-mixed image reconstruction can be formulated as:

$$\begin{aligned} \bar{\boldsymbol{x}}_i^{\text{RGB}} &= \boldsymbol{H}_{\text{RGB}}(\boldsymbol{f}_i^{\text{RGB}}) \\ \bar{\boldsymbol{x}}_i^{\text{IR}} &= \boldsymbol{H}_{\text{IR}}(\boldsymbol{f}_i^{\text{IR}}), \end{aligned} \tag{3}$$

where $\bar{\boldsymbol{x}}_i^{\text{RGB}}$ and $\bar{\boldsymbol{x}}_i^{\text{IR}}$ are the reconstructed RGB and IR images. $\boldsymbol{H}_{\text{RGB}}(\cdot)$ and $\boldsymbol{H}_{\text{IR}}(\cdot)$ are networks for RGB and IR image reconstruction respectively. Finally, a pixel-wise image reconstruction loss with l_2-norm is employed to guide the reconstruction process as:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \left(\|\bar{\boldsymbol{x}}_i^{\text{RGB}} - \boldsymbol{x}_i^{\text{RGB}}\|_2^2 + \|\bar{\boldsymbol{x}}_i^{\text{IR}} - \boldsymbol{x}_i^{\text{IR}}\|_2^2 \right). \tag{4}$$

Modality Discrimination Adversarial Learning. In this task, we introduce adversarial training to encourage

the learned features to be modality-invariant. Specifically, we employ a discriminator $\boldsymbol{D}(\cdot)$ to distinguish between the mixed features $\boldsymbol{f}_i^{\text{RGB}}$ and $\boldsymbol{f}_i^{\text{IR}}$. Initially, we fix the feature extractor $\boldsymbol{F}$ and train the discriminator $\boldsymbol{D}$ to correctly classify the mixed features based on their corresponding modality labels. This step ensures that $\boldsymbol{D}$ can accurately distinguish between features from different modalities. The objective for the discriminator $\boldsymbol{D}$ can be formulated as:

$$\mathcal{L}_{\text{D}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \boldsymbol{D}(\boldsymbol{f}_i) + (1-y_i)\log(1-\boldsymbol{D}(\boldsymbol{f}_i))\right], \quad (5)$$

where y_i is the modality label, $\boldsymbol{f}_i$ represents either $\boldsymbol{f}_i^{\text{RGB}}$ or $\boldsymbol{f}_i^{\text{IR}}$, and K is the batch size. In the second step, we fix the discriminator $\boldsymbol{D}$ and update the feature extractor $\boldsymbol{F}$ to generate mixed features that the discriminator cannot distinguish. This step forces the feature extractor to produce modality-invariant features, effectively minimizing the modality bias in the learned features. The objective for the feature extractor $\boldsymbol{F}$ during this phase is to maximize the classification error of $\boldsymbol{D}$ as:

$$\mathcal{L}_{\text{F}} = -\frac{1}{N}\sum_{i=1}^{N}\left[(1-y_i) \log \boldsymbol{D}(\boldsymbol{f}_i) + y_i\log(1-\boldsymbol{D}(\boldsymbol{f}_i))\right]. \quad (6)$$

The adversarial training process iterates between these two steps until the model converges, achieving a balance where the discriminator $\boldsymbol{D}$ cannot reliably distinguish the modality of the features generated by $\boldsymbol{F}$.

Modality-Augmented Contrasting Cluster. In this task, we focus on enhancing the discriminative power of the learned features by employing a contrastive learning approach. Unlike traditional methods that rely on data augmentation to generate negative samples, our approach does not introduce additional data augmentation but instead utilizes the inherent modality-specific features for contrastive learning. We treat $\boldsymbol{f}_i^{\text{RGB}}$ and $\boldsymbol{f}_i^{\text{IR}}$ as positive pairs within the contrastive learning framework. The contrastive loss function aims to maximize the similarity between the mixed features of different modalities, which can be formulated as:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\boldsymbol{f}_i^{\text{RGB}}, \boldsymbol{f}_i^{\text{IR}}))}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{f}_i^{\text{RGB}}, \boldsymbol{f}_j^{\text{IR}}))}, \quad (7)$$

where $\text{sim}(\cdot,\cdot)$ denotes a similarity measure (such as cosine similarity) between features. The loss function encourages the feature extractor to learn representations that pull together positive pairs ($\boldsymbol{f}_i^{\text{RGB}}$ and $\boldsymbol{f}_i^{\text{IR}}$) while pushing apart other features. By iteratively optimizing this contrastive loss, the feature extractor $\boldsymbol{F}$ is guided to produce more discriminative features, thereby improving the overall performance of the cross-modal image re-identification task.

In summary, the overall loss function for the entire training process can be expressed as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{rec}} + (\mathcal{L}_{\text{D}} + \mathcal{L}_{\text{F}}) + \mathcal{L}_{\text{contrast}}. \quad (8)$$

5. Experiments

5.1. Implementation and Evaluation Protocols

Pre-training Stage. For pre-training, we implement CPSL using PyTorch on an NVIDIA RTX A100. Optimization is conducted with an SGD optimizer (momentum of 0.9, weight decay of 0.0005), an initial learning rate of 0.1 with linear warm-up, and a cosine decay schedule without restarts. Data augmentation includes random gray-scaling, horizontal flipping, Gaussian filtering, and interpolation sampling. Images are resized to 256×256, and the mask size N is set to 4.

Fine-Tuning Stage. For fine-tuning, we implement VT-ReID models using PyTorch on an NVIDIA GeForce RTX 3090 GPU. Images are resized to 180×256. Optimization uses an SGD optimizer (momentum of 0.9, weight decay of 0.0005) with an initial learning rate of 0.1 and linear warm-up over 80 epochs. The model is initialized with CPSL checkpoints while other hyper-parameters and data augmentations remain unchanged, as this setup is not specifically tuned for CPSL. Further optimization may improve results, which we leave for future work.

Evaluation Protocols. Following [34], we use the cumulative matching characteristic (CMC) curve and mean average precision (mAP) for evaluation. CMC assesses rank-k accuracy, with rank-1, rank-10, and rank-20 scores reported in our experiments. mAP reflects the overall retrieval performance across all query samples.

5.2. Comparison with pre-trained weights

To validate the effectiveness of our proposed pre-training method, we employ three different backbone initialization strategies during fine-tuning: random initialization (Random Init.), initialization with ImageNet pre-trained weights (ImageNet-1K), and initialization using the proposed pre-training weights (CPSL). To enhance experimental rigor, we perform supervised fine-tuning with four state-of-the-art VT-ReID methods: AGW [37], DDAG [36], MMD-ReID [17], and DEEN [39].

From Tab. 2, we can observe that the VT-ReID models trained from scratch generally under-performance, with all methods failing to converge. The primary cause of the gradient explosion is attributed to the triplet loss, suggesting that randomly initialized models struggle to establish semantically meaningful Euclidean feature spaces.

Our method CPSL achieves superior results across all testing and query modes. In contrast to ImageNet-1K pre-training, we do not rely on additional training data or incorporate any supervised training processes. The superior performance is primarily attributed to the design of our self-supervised tasks, which account for modality bias and eliminate cross-modality differences.

It can also be observed that CPSL consistently improves

Method	Venue	Pre-Train	Single-shot						Multi-shot					
			IR2RGB			RGB2IR			IR2RGB			RGB2IR		
			rank1	rank10	mAP	rank1	rank10	mAP	rank1	rank10	mAP	rank1	rank10	mAP
AGW [37]	TPAMI 2021	Random Init.	4.35	18.75	9.38	4.89	19.7	9.29	7.22	27.26	4.82	6.67	26.17	5.53
		ImageNet-1K	15.75	45.98	25.74	16.17	47.58	26.54	20.1	58.12	16.21	20.45	58.16	18.3
		CPSL	**23.47**	**53.06**	**33.69**	**25.36**	**58.54**	**36.57**	**31.91**	**67.54**	**24.13**	**32.64**	**68.07**	**24.01**
DDAG [36]	ECCV 2020	Random Init.	5.22	22.81	11.48	5.06	22.69	11.22	8.55	32.72	5.84	7.06	30.05	6.25
		ImageNet-1K	20.46	48.78	30.12	19.46	49.15	29.21	27.58	64.75	20.28	26.69	65.54	22.87
		CPSL	**22.23**	**50.86**	**31.99**	**23.45**	**51.33**	**32.94**	**31.96**	**67.94**	**21.7**	**31.16**	**66.58**	**24.51**
MMD-ReID [17]	BMVC 2021	Random Init.	4.34	15.27	9.86	4.92	18.53	9.93	5.99	19.37	10.78	5.69	20.34	10.42
		ImageNet-1K	18.93	48.64	28.99	19.27	49.14	29.08	22.04	55.87	19.64	22.96	59.27	31.41
		CPSL	**21.98**	**51.85**	**32.39**	**22.52**	**53.38**	**31.99**	**28.34**	**62.09**	**22.1**	**29.04**	**62.85**	**21.8**
DEEN [39]	CVPR 2023	Random Init.	7.29	31.74	15.24	6.11	27.06	13.29	7.88	37.44	7.48	7.39	32.81	7.34
		ImageNet-1K	18.26	53.01	30.32	20.46	56.23	32.09	23.83	64.14	22. 06	22.65	66.61	21.43
		CPSL	**22.57**	**59.85**	**31.09**	**23.7**	**59.56**	**33.35**	**25.62**	**67.52**	**23.44**	**24.53**	**68.63**	**23.97**

* The **bold entries** represent the best performance in table.

Table 2. Comparisons between the proposed CPSL and some state-of-the-art methods on UCM-VeID V2 datasets. The rank-k (%) and mAP (%) are demonstrated.

performance across various VT-ReID methods, which indicates that the weights generated by CPSL exhibit strong robustness and adaptability. This suggests that the CPSL method is effective in enhancing the generalization capabilities of the models, allowing them to perform well under different conditions and with diverse Re-ID techniques.

5.3. Comparisons With Other SSL Methods

We compare CPSL with multiple ResNet50 pre-trained models on the UCM-VeID V2. The methods involved can be categorized into four groups: (1) method based on VT-ReID fine-tuning (MMGL [30]), (2) method based pretext task design (Jigsaw [26], Rotation [8]), (3) method based on self-supervised learning (SSL) pre-training (InstDisc [35], SimCLR [3], SwAV [2], BYOL [9], SimSiam [4], MoCo V2 [5], DenseCL [33], Relative-Loc[7]).

As shown in Tab. 3, MMGL performs suboptimally. This is partly because MMGL is tailored for pedestrian targets, using a top-down human segmentation approach that struggles with vehicles' variable orientations. Additionally, the released MMGL weights were fine-tuned on pedestrian VT-ReID datasets, embedding pedestrian-specific features that do not generalize well to vehicle targets.

SSL methods share a similar goal with our approach by aiming to endow the encoder with discriminative capabilities. However, methods such as SwAV [2] and BYOL [9] perform poorly, as indicated by the stagnant triplet loss during fine-tuning, suggesting these methods fail to establish a meaningful Euclidean space from the outset. Although some SSL methods outperform those based on pretext task design, they still underperform compared to CPSL. This is because they focus solely on target discriminability through contrastive learning without addressing modality differences. In contrast, our PMIR and MDAL modules effectively mitigate cross-modality differences, enabling shared feature learning.

Method	Rank1	Rank10	Rank20	mAP
method based on VT-ReID fine-tuning:				
MMGL [30]	21.82	56.56	68.03	16.14
method based pretext task design:				
Jigsaw [26]	21.24	56.77	67.25	15.66
RotationNet [8]	26.62	63.03	72.77	19.32
method based on SSL pre-training:				
InstDisc [35]	24.09	62.07	73.16	20.04
SimCLR [3]	12.43	48.57	62.15	11.17
SwAV [2]	6.57	28.89	40.22	5.17
BYOL [9]	6.89	25.51	36.87	5.17
SimSiam [4]	27.24	65.74	73.13	22.36
MoCo V2 [5]	28.89	67.46	74.98	22.44
DenseCL [33]	27.89	64.99	75.12	22.41
Relative-Loc[7]	18.91	55.79	68.06	15.53
ours	**32.64**	**68.07**	**76.32**	**24.01**

Table 3. Comparison with other SSL methods. The rank-k (%) and mAP (%) are demonstrated.

5.4. Ablation Study

To assess the effectiveness of each component in our method, we conducted ablation studies on the proposed pre-training method, removing specific modules while keeping all other experimental settings constant. Our method includes three key modules: PMIR (R), MDAL (D), and

Method	Single-shot			Multi-shot		
	rank1	rank10	mAP	rank1	rank10	mAP
Random Init.	4.89	19.7	9.29	6.67	26.17	5.53
ImageNet-1K	16.17	47.58	26.54	20.45	58.16	18.3
R	7.64	26.93	12.21	9.33	42.72	10.59
R+D	14.63	49.57	22.46	18.88	54.95	18.32
R+C	22.49	55.48	30.34	28.97	65.16	21.96
R+D+C	**25.36**	**58.54**	**36.57**	**32.64**	**68.07**	**24.01**

Table 4. The quantitative results of ablation studies. The rank-k (%) and mAP (%) are demonstrated.

MACL (C). Here, R and D target modality alignment, while C enhances discriminative representation. The ablation study results are shown in Tab. 4.

Effectiveness of PMIR. PMIR aims to create a shared latent representation space for RGB and IR modalities via reconstruction tasks. As shown in Tab. 4, compare to the results of Random Init., RANK1 and mAP improve by 2.57% and 2.92% under the single-shot query mode, indicating PMIR's effectiveness in reducing modality discrepancies.

Effectiveness of MDAL. MDAL uses modality adversarial learning to further align modalities. Tab. 4 shows a RANK1 and mAP increase of 6.99% and 10.25% under single-shot query mode, demonstrating that MDAL helps guide the encoder toward shared feature learning.

Effectiveness of MACL. MACL employs modality contrastive learning to enhance feature clustering and discriminative power. As evidenced in Tab. 4, RANK1 and mAP improve by 7.86% and 7.88% in single-shot query mode, confirming MACL's success in learning discriminative features for VT-ReID.

P_n	2×2	3×3	4×4	5×5
rank1	24.66	27.32	**32.64**	29.05
mAP	16.38	20.91	**24.01**	22.44

Table 5. Different partition stripe numbers P_n $(N \times N)$. The rank-k (%) and mAP (%) are demonstrated.

5.5. Parameter Analysis

The partition stripe number P_n directly determines the size of the image blocks to be swapped and is an important hyperparameter of our method. To verify the impact of this hyperparameter, we conducted experiments with different values for the partition stripe number, with the relevant results presented in Tab. 5. From the table, it can be observed that as the partition stripe number increases, the model's performance initially improves before declining. This is because, as the partition stripe number increases, the semantic content within each image block transitions from being more discrete to becoming sparser. This indicates that in practical applications, appropriate adjustment of this parameter is necessary to achieve optimal performance.

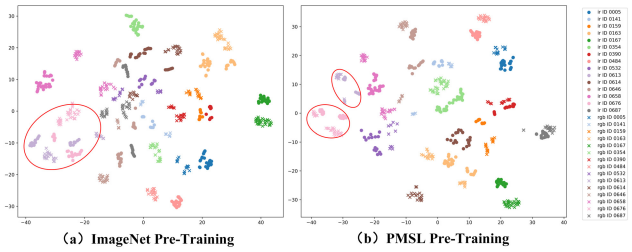


Figure 5. Visualization of the pre-learned features. Different colors represent distinct identities, while circle and cross symbols denote RGB and IR modalities, respectively. The experiment is conducted on UCM-VeID V2 dataset with the AGW baseline [37].

5.6. Visualization of the feature distribution

To better understand the proposed pre-training paradigm, we visualize the t-SNE distributions of features pre-learned using ImageNet-1K and the CPSL checkpoints on the UCM-VeID V2 dataset, utilizing AGW [37]. As shown in Fig. 5, we randomly select 15 identities to visualize all corresponding images.

Fig. 5 demonstrates that, the modal gap between PMSL features is smaller than that between ImageNet features (highlighted by the red circle). Specifically, PMSL features show less variation across modalities, indicating better consistency and robustness when processing RGB and IR images. In contrast, ImageNet-pretrained features exhibit significant modal variation, which can negatively impact performance in VT-ReID tasks. Thus, PMSL not only eliminates cross-modality discrepancy but also improves the model's discriminative ability for downstream tasks, providing stronger support for cross-modal vehicle re-identification.

6. Conclusion

In this paper, we introduce a novel dataset that expands existing ones in terms of scale, acquisition platforms, collection scenarios, time, and climate conditions. To address the MBT problem, we break it into two sub-problems: modality discrepancy and insufficient discriminability. We propose an end-to-end cross-modal pre-training framework to address these issues. To reduce modality discrepancies, the framework incorporates self-supervised tasks based on image block swapping and modality adversarial learning. For enhanced discriminability, we design an unsupervised clustering-guided multi-modal contrastive learning approach. Extensive experiments on the new dataset validate the effectiveness of our method.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62201586).

References

[1] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Trans. Multim.*, 20(9):2385–2399, 2018. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 9912–9924, 2020. 7

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 7

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 15750–15758, 2021. 7

[5] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009. 2

[7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1422–1430. IEEE Computer Society, 2015. 7

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proc. Int. Conf. Learn. Represent.*, 2018. 3, 7

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 21271–21284, 2020. 3, 7

[10] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proc. AAAI Conf. Artif. Intell.*, pages 6853–6860, 2018. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 5

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9726–9735, 2020. 3

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 15979–15988, 2022. 2

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 15979–15988, 2022. 3

[15] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proc. Int. Conf. Mach. Learn.*, pages 9226–9259, 2022. 2

[16] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Alleviating modality bias training for infrared-visible person re-identification. *IEEE Trans. Multim.*, 24:1570–1582, 2022. 2

[17] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. MMD-ReID: A simple but effective solution for visible-thermal person reid. In *Proc. BMVC*, pages 12026–12035, 2021. 6, 7

[18] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo. Multi-spectral vehicle re-identification: A challenge. In *Proc. AAAI Conf. Artif. Intell.*, pages 11345–11353, 2020. 4

[19] Chunsheng Liu, Ye Song, Faliang Chang, Shuang Li, Ruimin Ke, and Yinhai Wang. Posture calibration based cross-view and hard-sensitive metric learning for UAV-based vehicle re-identification. *IEEE Trans. Inf. Forensics Secur.*, 23:19246–19257, 2022. 1, 2

[20] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2167–2175, 2016. 2

[21] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proc. Eur. Conf. Comput. Vis.*, pages 869–884, 2016. 2

[22] Xingyue Liu, Jiahao Qi, Chen Chen, Kangcheng Bin, and Ping Zhong. Beyond sharing weights in decoupling feature learning network for UAV rgb-infrared vehicle re-identification. *IEEE Trans. Multim.*, 2024. 1, 2, 3, 4

[23] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3235–3243, 2019. 2

[24] Mingming Lu, Yongchuan Xu, and Haifeng Li. Vehicle re-identification based on UAV viewpoint: Dataset and method. *Remote. Sens.*, 14(18):4603, 2022. 1, 2

[25] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. Int. Conf. Learn. Represent.*, 2013. 3

[26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. Eur. Conf. Comput. Vis.*, pages 69–84, 2016. 3, 7

[27] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8228–8237, 2022. 2

[28] Shangzhi Teng, Shiliang Zhang, Qingming Huang, and Nicu Sebe. Multi-view spatial attention embedding for vehicle re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, 31 (2):816–827, 2021. 1

[29] Shangzhi Teng, Shiliang Zhang, Qingming Huang, and Nicu Sebe. Viewpoint and scale consistency reinforcement for UAV vehicle re-identification. *Int. J. Comput. Vis.*, 129(3): 719–735, 2021. 1, 2, 4

[30] Lin Wan, Qianyan Jing, Zongyuan Sun, Chuang Zhang, Zhihang Li, and Yehansen Chen. Self-supervised modality-aware multiple granularity pre-training for rgb-infrared person re-identification. *IEEE Trans. Inf. Forensics Secur.*, 18: 3044–3057, 2023. 2, 5, 7

[31] Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, and Yanning Zhang. Vehicle re-identification in aerial imagery: Dataset and approach. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 460–469, 2019. 1, 2, 4

[32] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 12692–12702, 2020. 2

[33] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3024–3033. Computer Vision Foundation / IEEE, 2021. 7

[34] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 5390–5399, 2017. 5, 6

[35] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3733–3742, 2018. 7

[36] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. *Proc. Eur. Conf. Comput. Vision*, 12362:229–247, 2020. 6, 7

[37] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44:2872–2893, 2020. 2, 5, 6, 7, 8

[38] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proc. Eur. Conf. Comput. Vis.*, pages 649–666, 2016. 3

[39] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2153–2162, 2023. 6, 7

[40] Aihua Zheng, Xianpeng Zhu, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Inf. Fusion*, 100:101901, 2023. 4

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2242–2251, 2017. 4