Closing the domain gap for cross-modal visible-infrared vehicle re-identification

Kamenou, E., Martinez del Rincon, J., Miller, P., & Devlin-Hill, P. (2022). Closing the domain gap for cross-modal visible-infrared vehicle re-identification. In *Proceedings of the 26th International Conference on Pattern Recognition* (pp. 2728-2734). (International Conference on Pattern Recognition (ICPR)). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ICPR56361.2022.9956381

Published in:
Proceedings of the 26th International Conference on Pattern Recognition

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Closing the Domain Gap for Cross-modal Visible-Infrared Vehicle Re-identification

Eleni Kamenou*, Jesus Martinez del Rincon*, Paul Miller* and Patricia Devlin-Hill†
*Centre of Secure Information Technologies (CSIT)
Queen's University, Belfast, United Kingdom
†Thales UK

***Abstract*—Traditional vehicle re-identification (ReID) approaches, based on visible spectrum data achieve high performance, but have limited capability of real-life applications, as they perform poorly under occluded visibility conditions, such as night-time and bad weather. In such cases, the use of infrared spectrum thermal imagery offers complementary and persistent information when the visual data contribution is inadequate. It is therefore highly beneficial to create a vehicle ReID framework that can exploit both modalities, if available, and is able to apply cross-modality matching, when ReID is required across single modal sensors. This is an extremely challenging task because the nature of the two data modalities induces high discrepancy between the two domains. In this paper we propose a robust end-to-end 2-stream vehicle ReID system that aims to solve the multi-modal and cross-modal ReID problem together by minimising the domain shift between infrared and visible distributions. Our framework consists of a shared network part, following the 2 independent streams, to extract shareable features, along with a domain alignment technique to narrow the gap between the two domains and inter-modality learning to address the cross-domain matching problem. The proposed system achieves state-of-the-art results on RGBN300 dataset, when both modalities are available at inference time. Moreover, our work is the first to explore the cross-modal settings for vehicle ReID and attempts to reduce the performance drop of the cross-modal scenario, when the query and the gallery images come from different modalities. We first measure the baseline cross-modal performance, and then prove that the proposed method improves up to 11% in mAP and 16% in Rank-1 score against the baseline.**

I. Introduction

Vehicle re-identification (ReID) aims at keeping the identities of the vehicles tracked while moving across multiple sensors with non-overlapping fields of view and has various applications for surveillance and security purposes. In computer vision, ReID is modelled as a retrieval task where given a query image of a vehicle we search over numerous gallery images to match it with images that show the same vehicle captured by different cameras. To this day it is still considered an open research problem in computer vision; due to numerous challenges related to the data properties. Different camera angles, scene occlusions, inter-class similarity and intra-class variability impose large differences between a vehicle's appearances in different camera views. In general, the difficulty of the problem is exacerbated by the differences across the sensors being used.

Another prominent issue that prevents the applicability of the so far proposed methods is the failure to apply ReID in adverse dark environments or at night time. This is because visual spectrum data provides limited information in such conditions. To this end, thermal spectrum imaging is able to overcome the aforementioned challenges.

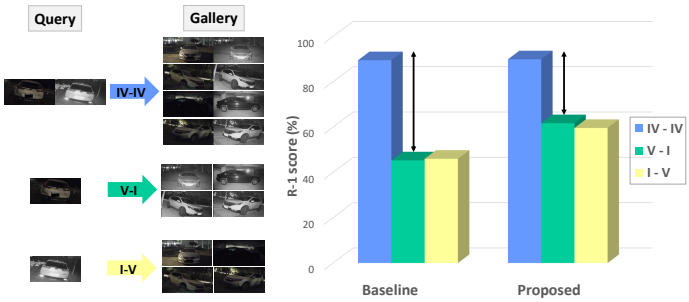


Figure 1: Illustration of the performance on RGBN300 dataset, in multi-modal and cross-modal ReID settings. "IV-IV" refers to both visible and infrared images being available for each sample in query and gallery sets. "V-I" and "I-V" denote the ReID performance when having visible query and infrared gallery set, and vice-versa, respectively. It is clear that the ReID problem across modalities is much more difficult because the 2 modalities are intrinsically distinct. The black arrows point out the performance drop of cross-modal ReID of baseline compared to our proposed approach. Our system achieves clear improvement in the cross-modal ReID scenarios while keeping a very high performance in the multi-modal ReID case (see Table I).

While visible images captured by RGB sensors measure the reflected energy and provide information similar to what the human eye would process, infrared spectrum imagery depends on the emitted energy from objects and potential absorption/emission from the background. In particular NIR (near infrared) sensors can capture near infrared light ($0.78 - 3\mu m$ wavelength) reflected by subjects, which is not affected by low illumination and occlusions due to bad weathers. It is therefore very important to be able to apply ReID from both visible and infrared imagery in situations when the first does not provide semantic information.

Realistic scenarios would also require the ability to identify a vehicle across the two modalities. The problem of cross-modal visible-infrared matching arises which is currently a highly under-studied field. Successfully matching vehicle images that belong to heterogeneous modalities is extremely challenging due to the large differences in visualising the two

Manuscript 852 submitted to 26th International Conference on Pattern Recognition. Received January 24, 2022.

light spectrums. In the cases where the query and gallery images belong to different modalities (cross-modal ReID), the performance is lower compared to the scenario where both modalities are available for each sample (multi-modal ReID), as Figure 1 shows.

In this paper, we aim to build a system that is able to identify vehicles when both modalities are available, but also in the more challenging case of cross-modal ReID. To do this, we propose a 2-stream deep neural architecture for multi-spectral vehicle ReID. The main contributions of this work are summarised below:

- We introduce an end-to-end architecture that uses both data modalities at training phase and is capable of applying ReID even in cases when one of the modalities is missing at testing phase. We provide baseline results that achieve state-of-the-art performance on RGBN300 [1] dataset exceeding the current performance in the literature by a large margin.
- We propose a novel domain alignment method to reduce the domain gap between the feature spaces of the 2 modalities projections. This method offers $+16\%$ higher Rank-1 score in the cross-modal settings against our baseline.
- Moreover, given the aligned domains, cross-domain metric learning is employed to expose the model to cross-modal matching and jointly project the embeddings from both modalities grouped into identity clusters. This reaches $+11\%$ higher mAP score in the cross-modal settings against our baseline, while maintaining high performance in the multi-modal ReID case.
- Finally, this work provides a strong baseline for future works in domain generalisation, domain alignment and cross-modal retrieval between visible and infrared domains for similarity learning tasks. Note that, our work is the first to explore the cross-modal visible-infrared scenario for vehicle ReID.

II. Related Work

Although there have been numerous works [2]–[7] addressing the vehicle ReID problem from visible spectrum imagery, the use of infrared vehicle data has not been reviewed yet. In fact, [1] is the first work to introduce the multi-spectral vehicle ReID problem along with the first dataset for this task, proving that the use of heterogeneous data is beneficial and allows higher ReID performance against using only one modality, traditional RGB data. Specifically, to encourage the consistency among different spectrum images, Li et al. [1] propose a score and similarity coherence loss function along with adaptive fusion applied to the activation maps of different spectrums.

As multi-modal ReID research evolves, the cross-modal matching problem arises. High discrepancy between the two domains renders cross-modal ReID a significantly challenging research problem that remains unstudied. To the best of our knowledge, there are no current works on cross-modal vehicle ReID, we herein present some of the few recent methods on cross-modal ReID on person data, which shares a common objective as vehicle ReID. [8] first introduced the cross-modal ReID problem between RGB and infrared modalities on multi-spectral person data and analysed different network structures.

Several works [9]–[12] employ generative adversarial network (GAN) architectures to address the cross-modal matching. [9] proposed a cross-modality generative adversarial network to handle the lack of insufficient discriminative information. Similarly, [11] intends to decrease the heterogeneousness of feature representations between the modalities with a modality discriminator as an adversary which guides the feature extractor to enhance the homogeneousness of features. A one-stream network is used as feature extractor to generate modality-invariant feature maps that are then projected into an embedding space as discriminative feature vectors. The discriminator aims to recognize the modalities of feature maps, and the feature extractor generates modality-invariant features to deceive the modality discriminator. [12] proposes a method of dual-level discrepancy reduction learning to map images of different modalities into a unified space using generative adversarial network. To this end, cascaded sub-networks are used to obtain discriminative feature representations and decompose the mixed modality.

Cross-modal ReID frameworks mainly follow two branches architectures for the two modalities. [13] proposed an end-to-end dual-path learning framework using a two-stream backbone structure while applying the bi-directional dual-constrained top-ranking loss to train the network. The parameters of two feature extractors are different in order to extract the modality-specific information. Then the authors use the embedding FC layer to learn multi-modal shareable features. [14] proposes a two-stage framework of feature learning and metric learning. A 2-stream convolutional architecture is proposed to learn the multi-modal shareable feature representations of the two heterogeneous modalities, and a Hierarchical Cross-modality Metric Learning (HCML) approach is adopted to transform two heterogeneous modalities into a consistent space. Finally, [15] explores the potential of combining modality-shared information with modality-specific characteristics. proposed a dual-path cross-modality feature learning framework, including a dual-path spatial-structure preserving common space network and a contrastive correlation network, which preserves intrinsic spatial structures and attends to the difference of input cross-modality image pairs.

III. Proposed Method

Figure 2 depicts the training pipeline of the proposed framework. As can be seen, it is based on a 2-stream convolutional architecture, utilising the two modalities simultaneously. Each stream is independent and it consists of a backbone model followed by Adaptive Average Pooling (AAP) and Batch Normalization (B-Norm) layers. Following the independent streams part, there is a shared network consisting of a pair of Fully Connected (FC) layers, where the first, FC_e, generates the embeddings, and the second, FC_c, acts as classification layer. The integration of shared-weight layers encourages the

Manuscript 852 submitted to 26th International Conference
on Pattern Recognition. Received January 24, 2022.

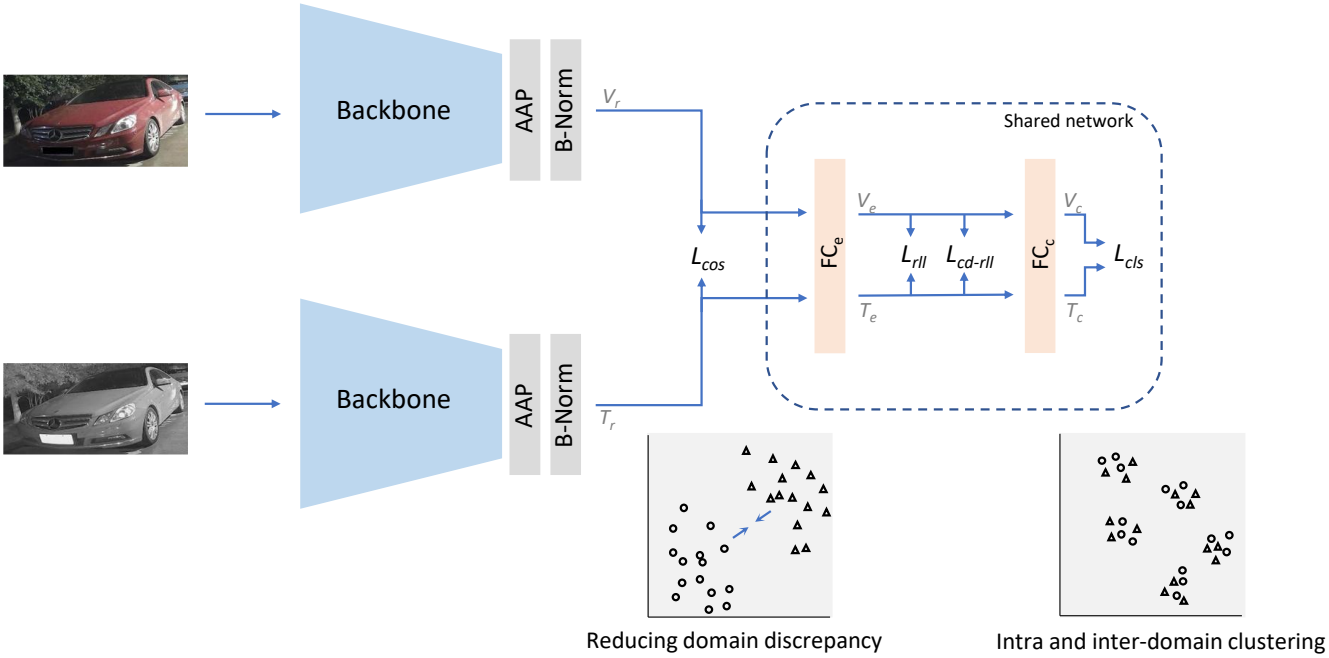


Figure 2: Pipeline of the proposed learning framework. Given the aligned images, we first extract the corresponding heterogeneous feature distributions V_r T_r via the 2-stream equivalent backbones. L_{cos} applies domain alignment by bringing the 2 distrubutions together. Then, these features are inserted to the FC_e embedding layer to obtain the embeddings V_e and T_e and within domain and cross domain metric learning are applied to group together the identity clusters regardless of the modality. Finally classification loss is applied to the classification vectors V_c and T_c. The model is trained by the summation of all the loss components in a single back-propagation.

generalization between the heterogeneous feature maps generated by the 2 independent branches. To facilitate this process, before passing the features to the shared network we apply domain alignment loss. As for similarity learning, within-domain and across-domain metric learning is employed to jointly project the embeddings from both modalities grouped into identity clusters. In this way, we create a common embedding space that enables us to apply ReID regardless of the modality.

The following subsection, outlines all the loss function components that are combined together to train our framework.

A. Loss Functions

In a batch of size B, there exist B pairs of aligned images of the visible and the infrared spectrum. Passing the input images through the network, we get feature representations at different layers of the architecture. V^r, V^e V^c correspond to the outputs of B-Norm, FC_e and FC_c layers of the network respectively, for visible modality samples. Accordingly, T^r, T^e T^c are the corresponding feature vectors regarding the infrared samples.

1) Domain Alignment: To reduce the discrepancy between visible and infrared domains, cosine similarity loss is applied on the pairs of feature representations belonging to the 2 modalities before passing them to the shared network. Let $V_r = \{v \quad | v_i \in R^{D_r}, \quad i = 1, ..., B\}$ and $T_r = \{t | t_i \in R^{D_r}, \quad i = 1, ..., B\}$ be the sets of features generated by the visible and infrared branches, respectively. L_{cos} is given at the equation below:

$$L_{cos}(V_r, T_r) = \frac{\sum_{i=1}^{B} 1 - \frac{v^i \cdot t^i}{\|v^i\| \|t^i\|}}{B} \tag{1}$$

2) Intra-modality Learning: For similarity learning within each modality, we employ both metric learning, specifically the Ranked-list Loss (RLL) function [16], applied to V_e and T_e embedding sets separately, and classification loss applied to V_c and T_c classification vectors.

Ranked-List Loss: In a batch of samples $X = \{x_c^i \quad | c = 1, ..., C, \quad i = 1, ..., K\}$ consisting of C identities and K samples per identity, every sample acts as anchor sample, having $K-1$ positive pair samples and $(C-1) \times K$ negative pair samples in total. RLL aims to ensure that the separation between negative samples is greater than a distance a, and the separation between the positives is less than $a - m$. Let the set of positive and negative pair samples that produce non-zero loss for an anchor sample x_c^i, are $P_{c,i} = x_c^j \mid j \neq i, \quad d_{ij} > (a-m)$ and $N_{c,i} = x_k^j \mid k \neq c, \quad d_{ij} < a$ respectively, where $d_{ij} = \| x^i - x^j \|_2$ denotes the euclidean distance between two samples. The positive and negative loss equations are:

$$L_p(x_c^i) = \frac{1}{\mid P_{c,i} \mid} \sum_{x_c^j \in P_{c,i}} d_{ij} - (a - m) \tag{2}$$

$$L_n(x_c^i) = \frac{1}{\mid N_{c,i} \mid} \sum_{x_k^j \in N_{c,i}} a - d_{ij} \tag{3}$$

Finally the RLL loss, L_r, is computed by summing L_p and L_n:

$$L_r(X) = \frac{\sum_{c=1}^{C}\sum_{i=1}^{K} L_p(x_c^i) + L_n(x_c^i)}{CK} \tag{4}$$

We apply the above equations separately to the V_e and T_e sets of embeddings and sum them up.

$$L_{rll} = L_r(V_e) + L_r(T_e) \tag{5}$$

Classification Loss: As classification loss function, we employ cross-entropy [17] with label smoothing regularization [18], denoted as $L_{ID}(.)$ applied to V_c and T_c features coming from the classification layer FC_c.

$$L_{cls} = L_{ID}(V_c) + L_{ID}(T_c) \tag{6}$$

3) Inter-Modality Learning: To enhance cross-modal training, we also expose the model to similarity learning between visible and infrared sample projections by applying the RLL in a cross-modal manner. In this case, the positive and negative pair sets $P_{c,i}$ and $N_{c,i}$ come from a different modality than the anchor x_c^i. Cross-domain RLL loss, L_{cd-rll}, is formally presented below.

Let $V_e = V = \{v^1, ..., v^B\}$ and $T_e = T = \{t^1, ..., t^B\}$ be the embedding sets for the visible and the infrared domains respectively for a batch of size $B = CK$. At this point, we drop the subscript e for simplicity. In the first case, visible samples act as anchors and the distance d_{ij} is now computed only between cross-modal samples: $d_{ij} = \| v^i - t^j \|_2$. Subsequently, for an anchor v_c^i, the positive and negative pair set are $P_{c,i}^v = \{t_c^j \mid j \neq i, \quad d_{ij} > (a - m)\}$ and $N_{c,i}^v = \{t_k^j \mid k \neq c, \quad d_{ij} < a\}$.

$$L_p(v_c^i) = \frac{1}{\mid P_{c,i}^v \mid} \sum_{t_c^j \in P_{c,i}^v} d_{ij} - (a - m) \tag{7}$$

$$L_n(v_c^i) = \frac{1}{\mid N_{c,i}^v \mid} \sum_{t_k^j \in N_{c,i}^v} a - d_{ij} \tag{8}$$

Accordingly, when the anchor t_c^i belongs to the infrared modality, the distance is $d_{ij} = \| t^i - v^j \|_2$, and the pair sets are $P_{c,i}^t = \{v_c^j \mid j \neq i, \quad d_{ij} > (a - m)\}$ and $N_{c,i}^t = \{v_k^j \mid k \neq c, \quad d_{ij} < a\}$.

$$L_p(t_c^i) = \frac{1}{\mid P_{c,i}^t \mid} \sum_{v_c^j \in P_{c,i}^t} d_{ij} - (a - m) \tag{9}$$

$$L_n(t_c^i) = \frac{1}{\mid N_{c,i}^t \mid} \sum_{v_k^j \in N_{c,i}^t} a - d_{ij} \tag{10}$$

The final cross-domain RLL loss is computed below:

$$L_{cd-rll} = \frac{1}{CK} \sum_{c=1}^{C} \sum_{i=1}^{K} L_p(v_c^i) + L_n(v_c^i) + L_p(t_c^i) + L_n(t_c^i) \tag{11}$$

Finally, all the aforementioned components are combined for the final loss:

$$L = (1 - \lambda)(L_{cls} + L_{rll}) + \lambda L_{cos} + w L_{cd-rll} \tag{12}$$

IV. Experiments

In this section, we present the experimental set up and analyze the results of our experimental process.

A. Dataset

RGBN300 [1] is currently the only dataset designed for vehicle ReID that includes cropped vehicle images from both visible and infrared imagery. It contains 50125 aligned image pairs of 300 vehicle identities in both RGB and near infrared modalities captured by eight pairs of RGB-NIR cameras. The number of image pairs of each vehicle varies from 50 to 200. 150 vehicles of 25200 image pairs constitute the training set, while the remaining 150 vehicles are used for testing and are split into 24925 gallery image pairs and 4985 query image pairs.

B. Implementation Details

As backbone, we adopt the ResNet50 architecture, pretrained on ImageNet [19]. The parameters of the 2 streams backbones are not shared. FC_e and FC_c layers' weights are shared and randomly initialised. Also, the dimensionality of V_r and T_r vectors is $D_r = 2048$ and the embedding size is $D_e = 512$. Following the implementation details in [1], for data augmentation, standard random cropping and horizontal flipping are applied during training, the Adam [20] optimizer is used with the batch size of 16, consisting of $C = 8$ identity classes and $K = 4$ images per class, and the learning rate scheduler follows the implementation details in [1]. Weighting factors λ and w are set to 0.5 and 3.0 after experimental analysis (see Tables III and V). Our model successfully converges after 6-8 epochs.

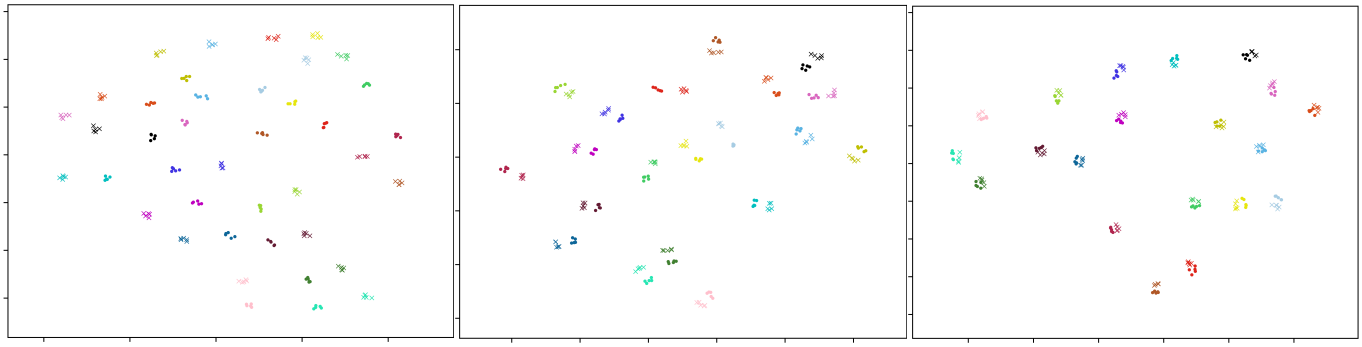
At inference time, V_e and T_e, extracted from FC_e are used as testing feature vectors for the two modalities, and for the multi-modal settings, the feature vector is the aggregate of the two: $V_e + T_e$.

C. Comparison with Baseline

In this section, we extensively examine the impact of each component of the proposed method. First, we built an initial framework with just necessary loss functions and basic architecture, and we gradually add the proposed shared-weights part, domain alignment and cross-domain metric learning to evaluate their effectiveness. Our baseline architecture follows the training configuration proposed in [1] for the 2-stream case. According to this approach, embeddings are extracted directly from the two convolutional streams without using any shared embedding layer, and these features are used for the loss computation. The loss function for training under our baseline settings includes Intra-modality metric learning, $L_{rll} = L_r(V_r) + L_r(T_r)$, on the embeddings together with classification loss L_{cls} applied after FC_c layer, as reported in Equations 5 and 6.

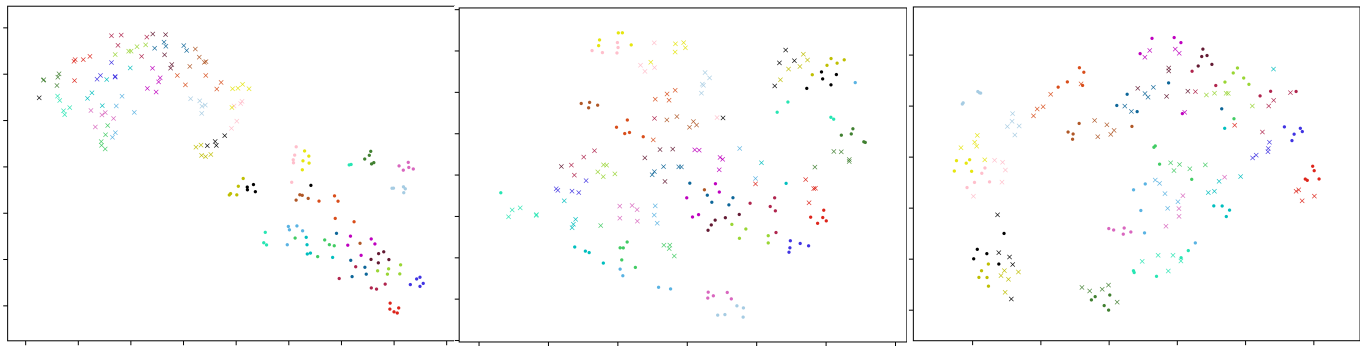
	IV - IV				V - I				I - V			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Baseline	70.2	89.5	**90.9**	**91.9**	35.0	45.3	49.9	52.5	34.9	47.5	51.1	53.2
Baseline + FC_e	69.5	88.6	89.9	90.9	35.3	47.1	50.9	52.9	34.5	44.4	48.7	50.9
Baseline + L_{cos}	67.1	86.4	88.1	89.3	41.5	56.6	59.8	61.6	41.6	55.0	57.7	59.7
Baseline + FC_e + L_{cos}	69.7	89.3	90.3	90.9	44.4	**61.7**	**64.3**	**65.9**	43.8	**59.6**	**62.6**	**64.6**
Baseline + FC_e + L_{cos} + L_{cd-rll}	**71.0**	**89.9**	**90.9**	91.5	**46.0**	59.6	63.6	65.9	**45.1**	57.7	61.8	63.8

Table I: Our method's performance against Baseline on RGBN300.



(a) Baseline (b) Baseline + FC_e (c) Baseline + FC_e + L_{cos}

Figure 3: The feature space projections of 20 training classes (5 samples per class for each modality) with circle and cross shapes corresponding to visual and infrared samples, respectively. The impact of FC_e and L_{cos} is apparent as the domain gap is strongly alleviated.



(a) Baseline (b) Baseline + FC_e (c) Baseline + FC_e + L_{cos}

Figure 4: The feature space projections of 20 testing classes (5 samples per class for each modality) with circle and cross shapes corresponding to visual and infrared samples, respectively. The integration of FC_e creates a feature space where the two distributions come closer compared to the baseline. This effect is even stronger in (c) after adding L_{cos} domain alignment.

Observing the baseline results at the first row of Table I, there is a big accuracy drop in the cross-modal testing settings compared to the IV-IV settings. To eliminate this performance gap and create a system that is capable of effectively applying ReID across the two spectrums, we applied certain modifications to our baseline. Regarding the architecture, an extra embedding layer, FC_e, with shared weights, is added to create a unified feature space for the two domains. Before passing the features to this layer, domain alignment is applied on the two distributions. Moreover, as well as employing metric learning within each domain to group together the embeddings that correspond to the same vehicle, we also expose the model to cross-domain similarity learning by applying metric learning in a cross-modal manner.

According to the results on Table I, extracting embeddings from the shared FC_e layer is only effective when it is combined with prior domain alignment on the features, using L_{cos}. On the other hand, domain alignment applied on features extracted from the streams without shared embedding layer, offers a 6% mAP and up to 11% Rank-1 improvement in cross-modal cases against baseline, at the cost of 3% mAP accuracy drop in multi-modal settings. Combining the two components boosts the cross-modal accuracy and also prevents accuracy drop in multi-modal case. Finally, the Inter-Modality learning that is introduced by L_{cd-rll} improves the multi-modal ReID accuracy while having a positive effect on the mAP scores of the cross-modal columns but not a clear improvement on the rank scores. mAP score is a more robust metric in ReID [21], therefore we could argue that L_{cd-rll} provides a moderate improvement also in cross-modal settings.

The above results are also supported by the visualised sample distributions in Fig. 3 and 4, where the embedding projections of visible (circle-shaped) and infrared (cross-shaped) modalities, for 20 identity classes of training and testing data are illustrated. The impact of employing a unified embedding space in bringing the 2 domains closer is apparent in both Fig. 3(b) and 4(b) as the samples from the two distribution move closer. Furthermore, this effect becomes even stronger after adding domain alignment on the features before the embedding layer, as demonstrated in Fig. 3(c) and 4(c).

D. Comparison with State-of-the-Art

Table II summarizes our method's performance against the state-of-the-art on the RGBN300 dataset. As there are not enough works on multi-spectral vehicle ReID, we provide results of visible spectrum ReID methods [22]–[24] extended to multi-spectral settings for comparison, as reported in [1]. Our work is closely comparable to HAMNet [1], that attempts to automatically fuse spectrum-specific features for multi-spectral matching. In addition to the reported result in [1], we also enhance HAMNet pipeline with metric learning to remove the influence of this factor in the comparison. Through our experimental process we found that adding a metric learning loss function, like Triplet (L_{trpl}) or RLL (L_{rll}) to HAMNet, boosts the accuracy and sets fairer comparison settings. Specifically, HAMNet+L_{rll} achieves comparable to our method's results in multi-modal settings, but significantly lower performance in the cross-modal cases. In particular, regarding the most challenging case of cross-modal matching our method outperforms HAMNet (even combined with metric learning) by up to 13% in mAP and 16% in Rank-1 score.

	IV - IV		V - I		I - V	
	mAP	R-1	mAP	R-1	mAP	R-1
PCB [22]	57.7	82.0	-	-	-	-
MGN [23]	60.5	83.7				
ABD-Net [24]	58.9	83.1				
HAMNet [1]	61.9	84.0	-	-	-	-
HAMNet + L_{trpl}	68.3	89.2	32.5	45.9	36.5	49.1
HAMNet + L_{rll}	**71.4**	89.7	36.1	47.0	32.8	41.7
Ours	71.0	**89.9**	**46.0**	**59.6**	**45.1**	**57.7**

Table II: Comparison with State-of-the-Art on RGBN300.

E. Experimental Analysis on Parameter Tuning

In this section, we extensively examine the impact of parameters and weighting factors used in the proposed approach on the performance.

1) Impact of Domain Alignment L_{cos}: To study the effect of domain alignment, we tune the value of λ weighting factor in Equation 12, while keeping out the L_{cd-rll} component by setting $w = 0$ to view the clear impact of L_{cos} alone. As can be seen in Table III, a small λ value does not apply the domain alignment properly while a high λ value also decreases the multi-modal ReID accuracy. Setting $\lambda = 0.5$ balances the domain alignment with the other loss components and achieves the highest overall accuracy.

	IV - IV		V - I		I - V	
	mAP	R-1	mAP	R-1	mAP	R-1
$\lambda = 0$	69.5	88.6	35.3	47.1	34.5	44.4
$\lambda = 0.2$	**70.1**	89.2	39.7	53.7	39.6	52.0
$\lambda = 0.4$	69.2	88.6	41.8	56.5	41.0	53.0
$\lambda = 0.5$	69.7	**89.3**	**44.4**	**61.7**	**43.8**	**59.6**
$\lambda = 0.6$	67.7	87.9	41.1	58.0	41.6	56.0
$\lambda = 0.8$	62.7	84.6	34.0	48.8	34.9	46.5

Table III: The impact of Domain Alignment L_{cos}

2) Impact of embedding size: When inserting the shared network layers, a critical parameter to decide was the size of the embeddings, D_e, that FC_e generates. Table IV shows the performance for different embedding sizes, while keeping $w = 0$ and $\lambda = 0.5$. After tuning D_e we set it to 512.

	IV - IV		V - I		I - V	
	mAP	R-1	mAP	R-1	mAP	R-1
D_e = 256	69.0	87.5	39.3	52.7	39.4	52.7
D_e = 512	**69.7**	89.3	**44.4**	**61.7**	**43.8**	**59.6**
D_e = 1024	69.6	**89.8**	39.6	53.2	39.2	51.6

Table IV: The impact of embedding size D_e.

3) Impact of Inter-modality Learning L_{cd-rll} : By tuning the weighting factor, w of Equation 6 while keeping $\lambda = 0.5$ and $D_e = 512$, we test the effectiveness of cross-modal learning. According to Table V, L_{cd-rll} offers a 2% surge in mAP under any testing settings, but not clear improvement in Rank-1. Therefore, we can claim that L_{cd-rll} provides small contribution to the system's accuracy under cross-modal settings but helps to keep state-of-art performance in multi-modal settings.

	IV - IV		V - I		I - V	
	mAP	R-1	mAP	R-1	mAP	R-1
$w = 0.0$	69.7	89.3	44.4	**61.7**	43.8	**59.6**
$w = 1.0$	69.6	88.4	45.4	60.7	43.7	58.2
$w = 2.0$	70.1	88.4	45.1	60.8	44.0	57.3
$w = 3.0$	**71.0**	**89.9**	46.0	59.6	45.1	57.7
$w = 4.0$	70.3	87.8	**46.3**	60.8	**45.4**	58.5
$w = 5.0$	69.0	87.6	45.5	58.6	42.9	54.0

Table V: The impact of Inter-modality Learning L_{cd-rll}.

V. Conclusion

In this work, we build a system that is able to apply ReID by using data from both visible and infrared spectrums, while addressing both the multi-modal and cross-modal scenarios. The large differences between the way of visualising the two light spectrums induces high domain discrepancy in the feature space which makes the cross-modal matching a very challenging problem. To the best of our knowledge, this work is the first to explore the visible-infrared cross-modal settings for vehicle ReID. Our method consists of a 2-stream architecture with a shared network part and an incorporated domain alignment technique, along with inter-modal and intra-modal metric learning, to minimise the domain shift in the embedding space. The experiments have shown that our system achieves state-of-the-art performance on RGBN300 dataset in the multi-modal settings and improves against other works and our own baseline by up to $+11\%$ mAP and $+16\%$ Rank-1 score in the cross-modal settings.

Manuscript 852 submitted to 26th International Conference
on Pattern Recognition. Received January 24, 2022.

REFERENCES

[1] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, 2020, pp. 11 345–11 353.

[2] A. Ayala-Acevedo, A. Devgun, S. Zahir, and S. Askary, "Vehicle re-identification: Pushing the limits of re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019*, pp. 291–296.

[3] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, 2019.

[4] H. Li, X. Lin, A. Zheng, C. Li, B. Luo, R. He, and A. Hussain, "Attributes guided feature learning for vehicle re-identification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2021.

[5] ——, "Attributes guided feature learning for vehicle re-identification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2021.

[6] J. Zhao, F. Qi, G. Ren, and L. Xu, "Phd learning: Learning with pompeiu-hausdorff distances for video-based vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2225–2235.

[7] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram, "A strong and efficient baseline for vehicle re-identification using deep triplet embedding," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 10, 2020.

[8] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2017, pp. 5390–5399.

[9] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, 2018, pp. 677–683.

[10] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, 2019, pp. 3622–3631. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00372

[11] Y. Hao, J. Li, N. Wang, and X. Gao, "Modality adversarial neural network for visible-thermal person re-identification," *Pattern Recognit.*, vol. 107, p. 107533, 2020. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107533

[12] Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 618–626.

[13] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, 2018, pp. 1092–1099. [Online]. Available: https://doi.org/10.24963/ijcai.2018/152

[14] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,*, S. A. McIlraith and K. Q. Weinberger, Eds., 2018, pp. 7501–7508. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16734

[15] S. Zhang, Y. Yang, P. Wang, G. Liang, X. Zhang, and Y. Zhang, "Attend to the difference: Cross-modality person re-identification via contrastive correlation," *IEEE Trans. Image Process.*, vol. 30, pp. 8861–8872, 2021. [Online]. Available: https://doi.org/10.1109/TIP.2021.3120881

[16] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 5207–5216.

[17] C. Liu, M. Lee, C. Wu, B. Chen, T. Chen, Y. Hsu, and S. Chien, "Supervised joint domain learning for vehicle re-identification," in *CVPR Workshops*, 2019, pp. 45–52.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, pp. 2818–2826.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1116–1124. [Online]. Available: https://doi.org/10.1109/ICCV.2015.133

[22] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *Computer Vision - ECCV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11208, 2018, pp. 501–518. [Online]. Available: https://doi.org/10.1007/978-3-030-01225-0_30

[23] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference, MM*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds., 2018, pp. 274–282. [Online]. Available: https://doi.org/10.1145/3240508.3240552

[24] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *International Conference on Computer Vision, ICCV 2019*. IEEE, 2019, pp. 8350–8360.