A Meta-learning Approach for Domain Generalisation across Visual Modalities in Vehicle Re-identification

Eleni Kamenou
CSIT
Queen's University Belfast
ekamenou01@qub.ac.uk

Jesús Martínez del Rincón
CSIT
Queen's University Belfast
j.martinez-del-rincon@qub.ac.uk

Paul Miller
CSIT
Queen's University Belfast
p.miller@qub.ac.uk

Patricia Devlin-Hill
Thales
Belfast, United Kingdom
patricia.devlin-hill@uk.thalesgroup.com

Abstract

Recent advances in imaging technologies have enabled the usage of infrared spectrum data for computer vision tasks previously working with traditional RGB data, such as re-identification. Infrared spectrum data can provide complementary and consistent visual information in situations of low visibility such as night-time, or adverse environments. However, the main issue that prevents the training of multi-modal systems is the lack of available infrared spectrum data. To this end, it is important to create systems that can easily adapt to data of multiple modalities, at inference time. In this paper, we propose a domain generalisation approach for multi-modal vehicle re-identification based on the recent success of meta-learning training approaches, and evaluate the ability of the model to perform to unseen modality data at testing time. In our experiments we use RGB, near-infrared and thermal-infrared modalities using the RGBNT100 dataset and prove that our meta-learning training configuration can improve the generalisation ability of the trained model compared to traditional training settings.

1. Introduction

Vehicle Re-identification (ReID) is an important and challenging task in the computer vision literature for visual surveillance applications. A large number of models for ReID problem have been proposed [1, 16, 20, 26, 41], exploring various architectures, deep metric learning methods and image enhancement techniques. The vast majority of ReID approaches have been based on visible spectrum visual data, as traditional RGB sensors have been the most common ones for surveillance scenarios. However, one of the main factors preventing the applicability of such systems is their poor performance under low illumination conditions, at night time, in foggy weather or dark scenes. To this end, infrared spectrum imaging sensors -including near infrared and thermal infrared- have been recently deployed in surveillance applications. For instance, unlike RGB sensors, thermal sensors can provide consistent 24h high quality visual imagery, overcoming the aforementioned challenges.

Figure 1. Vehicle images selected from the RGBNT100 [19] dataset. Three vehicles are captured in visible, near-infrared and thermal-infrared spectrum modality.

Vehicle ReID for infrared or multi-modal data is still relatively understudied due to the scarcity of infrared imagery and lack of large training datasets in these modalities. This is ultimately due to the significantly higher cost of thermal sensors compared to normal visible light sensors, despite their advantages to provide more robust and consistent visual information. Training multi-modal ReID systems would ideally require availability of labelled data from all the modalities, which is usually not possible or ex-

pensive. To overcome this limitation, we aim to develop a system that can generalise to data coming from modalities that have not been used at training phase.

This generalisation across modalities is however challenging to be achieved in practice, due to the fundamental differences between the physical phenomena involved in image capturing in each modality. In particular, the visible light spectrum is the segment of the electromagnetic spectrum that the human eye can view, which corresponds to wavelengths from 0.4 to $0.7\mu m$ and this range is called visible light. Common RGB sensors visualise this information by measuring the reflected energy of the objects in the scene and provide information similar to what we as humans would process. On the other hand, infrared spectrum imagery depends on the emitted energy from objects and potential absorption/emission from the background. In particular near-infrared (NIR) sensors can capture near infrared light of 0.78 to $3\mu m$ wavelength, reflected by subjects. This type of representation is not affected by low illumination, shadows, and occlusions due to bad weathers. Similar to NIR, thermal-infrared (TIR) imagery measures the radiant temperature of the objects and refers to electromagnetic waves between 3 and $20\mu m$. The main difference between TIR and NIR is that TIR measures emitted energy, whereas the NIR measures reflected energy, closer to how to visible spectrum is visualised.

The significant differences in the ways of visualising the different modalities induce high domain discrepancy among them. In other words, the modality heterogeneity of the input visual data translates into significant domain gap between their distributions in the embedding space. Therefore, building a model that has such a strong generalisation ability to deal with multiple data modalities is an extremely challenging task. Model-agnostic meta-learning [7, 17] approaches have recently emerged to address the domain shift problem for visual data in deep neural architectures. Such a problem refers to the distribution shift between a set of training (source) data and a set of test (target) data. In this paper, we apply meta-learning training strategy to achieve generalisation across three visual modalities: visible (RGB), NIR and TIR. To do this, we iteratively use two of the three modalities as source domains, and keep one held-out to be used as target domain. The meta-learning training configuration separates the source domains into meta-train and meta-test domains and it is designed to simulate the train-test cross-domain generalisation functionality. We prove that meta-learning training settings indeed increase the out-of-distribution generalisation of the model compared to the typical training process baseline, keeping traditional single back-propagation.

The main contributions of this paper are summarised below:

- We propose a simple yet novel ReID framework for vehicle ReID in unseen visual modalities, that works effectively when no training data for the target modality is available.
- We apply for the first time a meta-learning training strategy for domain generalisation across visual modalities and we demonstrate that meta-learning can address their significant domain discrepancy.
- We provide a systematic examination of the performance of our ReID system using three different visual modalities, RGB, NIR and TIR which are common in video surveillance settings. This evaluation demonstrates that meta-learning provides and up to 4.9% improvement in rank-1 and 3.1% improvement in mAP against our baseline when performing on a completely unseen visual modality.

2. Related Work

As an emerging research topic, vehicle ReID has attracted great efforts [12, 14, 23, 25, 49] in the computer vision community with most works focusing on visible light spectrum ReID data. Below we provide literature review for multi-modal ReID, domain generalisation and meta-learning related work.

Multi-modal Visible-Infrared ReID As multi-modal ReID research evolves, several works have been proposed to deal with the heterogineity of RGB and infrared modalities [4, 10, 34, 36–40], focusing mainly on person ReID data. Few recent multi-modal approaches have also been proposed for vehicle data [8, 11, 13, 19, 27]. Generative adversarial network (GAN) architectures have been employed to create a unified generalised embedding space [4, 8, 10, 34, 36] aiming to produce modality-invariant representations. Recent approaches [11, 27] have also proposed transformer-based frameworks for multi-modal vehicle ReID aiming to reduce feature deviations towards modal variations by learning intra- and inter-modality information. Although these approaches propose systems working with diverse modalities simultaneously, they do not consider ReID applied to unseen modality data at inference time.

Domain Generalisation Domain generalisation is an essential research topic, as it examines the ability of the trained model to perform on data from different distributions than the training data, aka. out-of-distribution (OOD) generalisation [31, 42, 44]. Without access to target domain data, training a model that can work effectively in any unseen target domain data is arguably one of the hardest problems in research community. Researchers have approached it with a wide range of methods related to aligning

source domain distributions for domain-invariant representation learning [21,22], augmenting source data with image generation [8,46,47], or exposing the model to domain shift during training via meta-learning [6, 17, 30]. Particularly for ReID, models by default address heterogeneous settings since training and testing identities are completely different, which automatically imposes the domain shift issue. Further than that, cross-dataset ReID has also gained interest [3,43,45,48] with the objective to generalize a ReID model trained on source camera views to target camera views installed in a different environment. However, the application of generalisation methods to tackle the domain-shift across visual modalities [38] -for example across RGB, infrared, thermal, radar sensor data- is still widely understudied.

Meta-learning Recently, meta-learning has been a fast-growing area with applications to many computer vision tasks [6,7,17,18,24,28,30,43], and it is based on exposing the model to domain shift among available source domains during training, expecting that the trained model will then be able to deal with domain shift in unseen domains at testing time. Although this type of methodology has been discussed decades ago [33] in the literature, MAML [7] was a major paper that explicitly suggested the separation of training data into meta-train and meta-test tasks and trained a single shared source model using multiple source tasks for few-shot classification. Meta-learning has been also used for domain generalisation [2, 3, 18, 43], by making use of source domains to imitate the domain shift that the model is going to face at testing time. Meta Face Recognition (MFR) [9], proposes a loss function employing distances of hard samples, identity classification and the distances between domain centers. However, simply enforcing alignment of the centers of training domains does not necessarily align their distributions and may lead to undesirable effects, such as aligning different class samples from different domains. Meta-Learning Domain Generalization (MLDG) [17] proposes to generate domain shift during training by synthesizing virtual domains within each batch. Although promising results have been demonstrated in training generalisable models, to the best of our knowledge, meta-learning approaches have not yet been applied for multi-spectral visual domains.

3. Proposed Method

In this section, we describe the algorithmic pipeline of our meta-learning based framework for multi-modal domain generalisation for vehicle ReID. Figure 2 depicts the training and testing processes of the proposed framework. At training phase, we select two out of the three available modalities to be used as source domains, and iteratively assign each one of them as meta-test and meta-train domain. The meta-learning training configuration simulates the train-test cross-domain generalisation functionality. At testing phase, the trained model is exposed to data from a completely unknown visual modality. Our hypothesis is that the meta-learning training settings are going to create a model with the generalisation power to perform ReID on unseen modality imagery.

3.1. Backbone architecture

Regarding our architecture, a ResNet50 convolutional backbone network is used, followed by Adaptive Average Pooling (AAP) [13] and Batch normalization (B-norm) layers. At the top of the network there is an embedding fully connected FC_{emb} that project the input images into a multi-dimensional embedding space and a classification fully connected FC_{cls} layer which is detached at testing phase. The input channels of ResNet50 architecture are typically 3, which matches with the RGB imagery format. For NIR and TIR inputs, which consists of only one channel per pixel, the 1-channel information is copied to the 3 channels to fit the architecture rrequirements.

3.2. Loss Functions

In this section, we provide the mathematical analysis of the loss function components, first for the metric learning part and then for the classification part.

Given a set of input images, after passing them through the network, we get feature representation vectors and classification vectors X and C, respectively. During training, we employ both metric learning, using the Ranked-list Loss (RLL) function [35], applied to X embeddings, and cross-entropy [23] with label smoothing generalisation [32] for classification applied to C classification vectors.

Ranked-list Loss RLL is selected for metric learning due to its simplicity and state-of-art performance in ReID as structured metric learning loss function [13, 35].

In a batch of size B there are Z vehicle identity classes and M samples per class, $B = M \times Z$. Given a set of embeddings $X = \{x_c^i \quad | c = 1, ..., Z, \quad i = 1, ..., M\}$ each sample acts as anchor sample, having $M - 1$ positive pair samples and $(Z - 1) \times M$ negative pair samples in total. Given two distance margins a and m, RLL aims to ensure that the separation between negative samples is greater than 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:

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

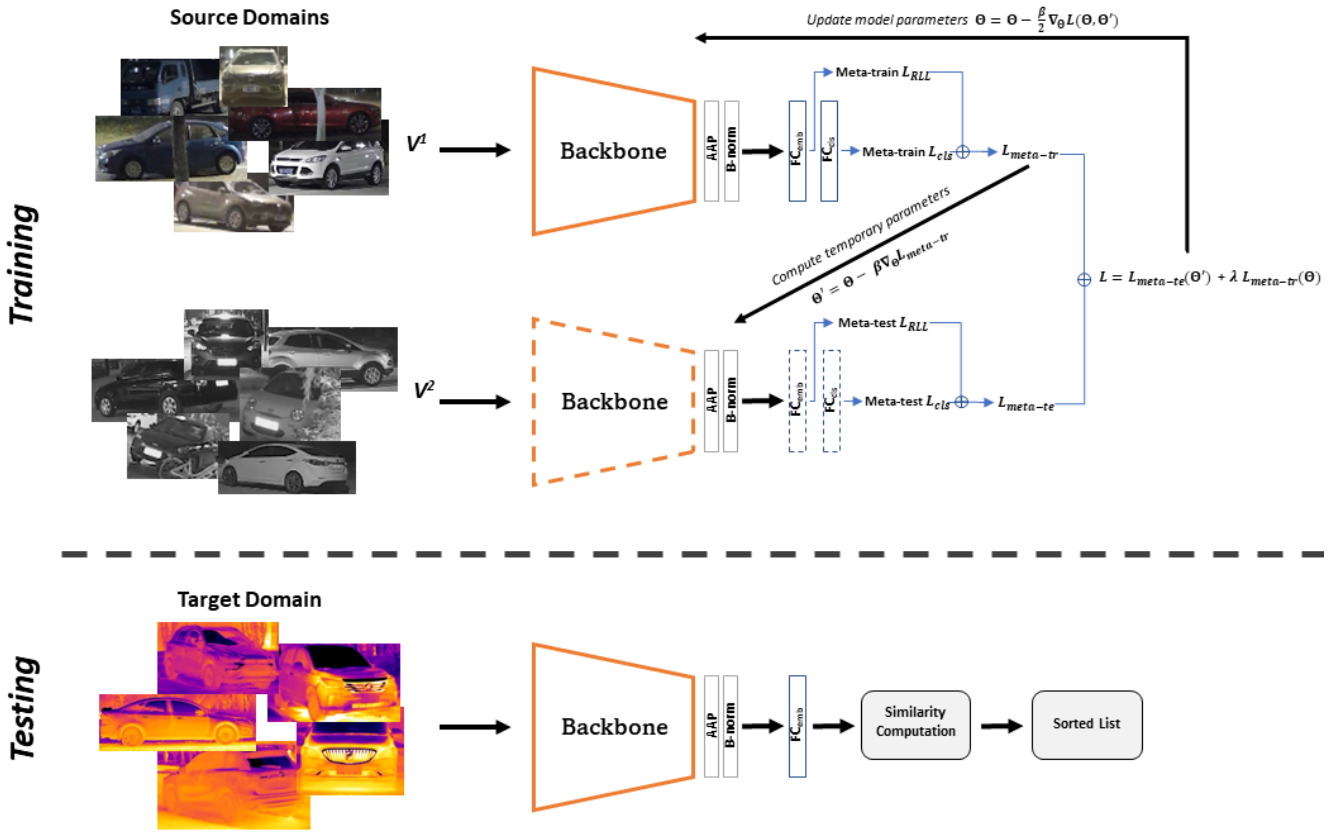

Figure 2. An overview of training and testing phases of the proposed framework. For the sake of visualisation, we have selected the RGB and NIR sets as source domains (meta-train V^1 and meta-test V^2, respectively) and the TIR set as target domain. During training, the model parameters are saved in a temporary copy and then updated using the loss computed from meta-train domain data, $L_{meta-tr}$. Then, we use the updated temporary model to compute the meta-test loss $L_{meta-te}$. Finally, the summation of the meta-train and meta-test loss L is used to update the original model towards a generalisable direction that performs well on meta-train and meta-test domains. This training configuration that simulates the train-test domain shift evaluation is expected to create a model which is able to generalise well to unseen target domains (in this case: TIR domain set) .

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

Finally the RLL loss, L_{rll}, is computed by summing L_p and L_n:

$$L_{rll}(X) = \frac{\sum_{c=1}^{C} \sum_{i=1}^{Z} L_p(x_c^i) + L_n(x_c^i)}{CZ} \tag{3}$$

Classification Loss As classification loss function, we employ cross-entropy [23] with label smoothing regularization [32], denoted as $L_{cls}(.)$ applied to C classification vectors. Adding classification has been proved effective in increasing the discriminative power of the model and leading to faster convergence in ReID frameworks [1].

3.3. Meta-learning Training

The training process of the proposed system is explained in Algorithm 1. During the training process, B input images are sampled from each of the K source domains, defined as $V^i, \quad i = 1, ..., K$. In our settings, domains correspond to different visual modalities. Passing the input images through the network, we get the sets of feature representation vectors and classification vectors for each domain: $X^i, C^i, \quad i = 1, ..., K$, respectively.

To perform the meta learning process, we first copy the original model parameters and update them using the loss computed from meta-train data, $L_{meta-tr}$. Then, we use the updated model with Θ' parameters to compute the meta-test loss $L_{meta-te}$. Finally, the summation of the meta-train and meta-test loss L is used to update the original model towards a generalisable direction that performs well on meta-train

Algorithm 1 Meta-learning Training for Generalisation across Visual Modalities.

Input
Source domains $D = [D_1, D_2, ..., D_K]$;
Architecture: $model$; Batch size B;
Hyper-parameters α, β, λ;

Output
Learned parameters: $\hat{\Theta}$

Initialize parameters Θ
repeat:
  Initialize the gradient accumulator: $G_\Theta \leftarrow 0$
  for each D_i, (meta-test domain) in D **do:**
    for each D_j, $i \neq j$ (meta-train domain) in D **do:**
      Sample $V^j = [v_1^j, v_2^j, ..., v_B^j]$ from D_j domain
      Compute $X^j, C^j \leftarrow model(V^j; \Theta)$
      Compute $L_{meta-tr} \leftarrow L_{rll}(X^j) + L_{cls}(C^j)$
      Compute $\Theta' \leftarrow \Theta - \beta \nabla_\Theta L_s$
      Sample $V^i = [v_1^i, v_2^i, ..., v_B^i]$ from D_i domain
      Compute $X^i, C^i \leftarrow model(V^i; \Theta')$
      Compute $L_{meta-te} \leftarrow L_{rll}(X^i) + L_{cls}(C^i)$
    end for
    $G_\Theta \leftarrow G_\Theta + \nabla_\Theta L_{meta-te} + \lambda \nabla_\Theta L_{meta-tr}$
  end for
  Update model parameters $\Theta \leftarrow \Theta - \alpha G_\Theta$
until convergence

and meta-test domains.

4. Experiments

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

4.1. Baseline

As baseline for our experiments, we consider the same model architecture, trained with a single back-propagation by the summation of metric learning and classification loss components ($L_{total} = L_{rll} + L_{cls}$) from all the source domains, without the integration of meta-learning training. By keeping the same source-target domain data separation, we ensure that the baseline is exposed to the same training set as our proposed approach. The baseline training pipeline is described in Algorithm 2. The purpose of setting this baseline is to evaluate the effect of meta-learning compared to conventional training for ReID when facing an unseen visual domain in the closest possible settings and model.

4.2. Datasets

For our experiments we use the RGBNT100 benchmark. RGBNT100 is a subset of RGBN300 [19] dataset which is currently the only dataset designed for vehicle ReID

Algorithm 2 Baseline Model Training.

Input
Source domains $D = [D_1, D_2, ..., D_K]$;
Architecture: $model$; Batch size B;
Hyper-parameters α;

Output
Learned parameters: $\hat{\Theta}$

Initialize parameters Θ
repeat:
  Initialize the gradient accumulator: $G_\Theta \leftarrow 0$
  for each D_i, in D **do:**
    Sample $V^i = [v_1^i, v_2^i, ..., v_B^i]$ from D_i domain
    Compute $X^i, C^i \leftarrow model(V^i; \Theta)$
    Compute $L_{total} \leftarrow L_{rll}(X^i) + L_{cls}(C^i)$
    $G_\Theta \leftarrow G_\Theta + \nabla_\Theta L_{total}$
  end for
  Update model parameters $\Theta \leftarrow \Theta - \alpha G_\Theta$
until convergence

that includes cropped vehicle images from diverse vision sensors. RGBNT100 is designed for three-spectral vehicle ReID among visible, near-infrared and thermal-infrared modalities. It contains aligned image triples from 100 vehicles for the three modalities captured by eight triples of RGB-NIR-TIR cameras. Each vehicle is captured by 2 to 8 sensor views and the number of captures of each vehicle varies from 50 to 200. 50 vehicles are used for training and the other 50 for testing, which translates into 8675 image triplets for training and another 8575 for testing along with 1715 queries.

In our experimental settings, since two modalities are used as source domains and the held-out third one as target domain, the parameter K in Algorithms 1 and 2 is equal to 2.

4.3. Evaluation Metrics

In ReID settings, the model is evaluated by ranking each gallery sample according to its similarity to the query sample, in the form of a sorted list. The two most widely used evaluation metrics in ReID are the mean average precision (mAP) and the rank-k scores. The rank-k score denotes the possibility that at least one true positive is ranked within the top k positions of the list. For mAP, the mean of all queries' average precision (AP) is computed, also known as the area under the Precision-Recall curve. In particular, the AP is computed for each query as:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}} \tag{4}$$

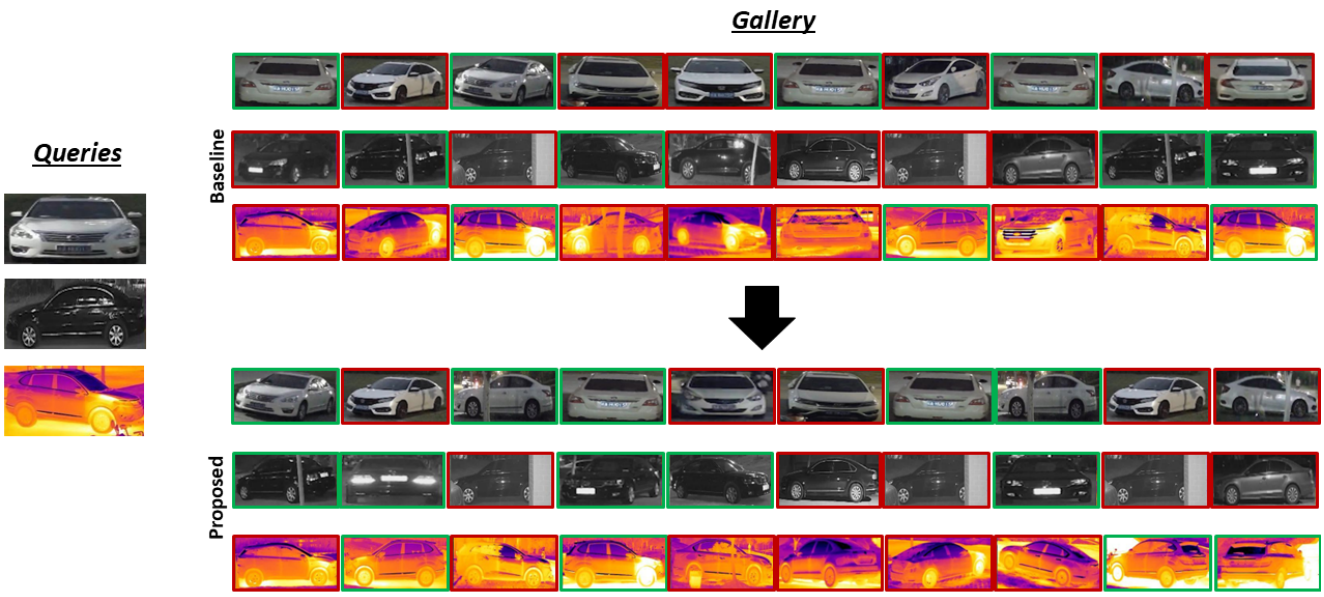

Figure 3. Qualitative results for one query image from each target domain modality from the RGBNT100 [19] dataset. For each query, there is a sequence of 10 gallery images corresponding to the ranking results produced by our models, using the baseline trained models (at the top) and the meta-learning trained models (at the bottom). Vehicles surrounded by green box denote the same vehicle as the probe (true positive), otherwise it is surrounded by a red box (false positive).

where n is the number of samples in the gallery set and N_{gt} is the number of positive samples in the gallery set. $P(k)$ is the fraction of true positives in the top k ranked gallery samples and corresponds to the precision at the k^{th} position of the results. $gt(k)$ is an indicator function that equals to 1 if the k^{th} result is correctly matched and 0 otherwise. The mAP is then computed over all queries' AP as:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

In the above equation, Q denotes the total number of queries.

4.4. Implementation Details

Our implementation was done in PyTorch [29]. As backbone, we adopt the ResNet50 architecture, pretrained on ImageNet [5]. The weights of FC_{emb} and FC_{cls} layers are randomly initialised. Also, the dimensionality of the embedding vectors is 512 and the embeddings are L-2 normalised before the loss computation. All images are resized to 128×128. Following the implementation details in [19], for data augmentation, standard random cropping and horizontal flipping are applied during training and the Adam [15] optimizer is used with weight decay equal to 0.0005 and a momentum of 0.9. The batch size is 32, consisting of $Z = 8$ identity classes and $M = 4$ images per class for each domain.

Learning rates, β and α, for the inner and the outer loop, are set to 10^{-4} and 3.5×10^{-5}, respectively. Finally, the λ weighting factor is set to 0.4 for the cases of RGB and TIR target domains and to 0.2 for the NIR target domain (see subsection 4.5.1).

4.5. Results

In order to simulate the unseen visual domain scenario, a leave-one-out domain setting is applied. In every experiment two domains are used for training and one is left as unseen for testing. This is repeated three times using the RGBNT100 dataset so as to examine all possible combinations. Each time the model is evaluated according to its ability to perform ReID using data from the target (unseen) domain. Tables 1, 2 and 3 show our system's performance against the baseline settings. Each table corresponds to a different source-target domain selection among the RGB, NIR and TIR data modalities.

It can be seen that meta-learning training configuration provides consistent improvement over the baseline under all source-target selection settings. As an example, in Figure 3, qualitative results are also provided for 3 different queries (3 different vehicles), one for each target domain. Given one query image for each target domain, we visualise the ranking results provided by the corresponding model.

Among the three source-target domain selection, TIR target domain shows the lowest performance and this re-

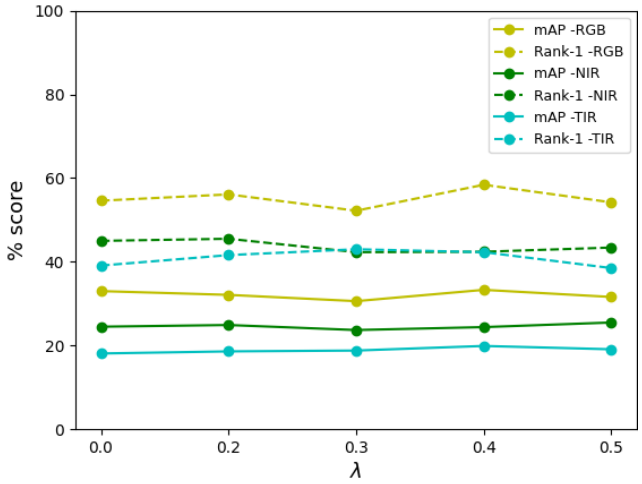

Figure 4. The mAP and rank-1 performance for different values of λ for the cases of RGB, NIR, and TIR target domains.

	Target: TIR			
Source: RGB, NIR	mAP	R-1	R-5	R-10
Baseline	18.6	40.9	47.3	50.8
Meta-learning Model	**19.9**	**42.3**	**50.7**	**56.2**

Table 1. The performance of our system on RGBNT100 [19] dataset. The model is trained on the training sets of RGB and NIR modalities (source domains) and then it is tested on the testing set of TIR modality (target domain). λ parameter is set to 0.4.

	Target: NIR			
Source: RGB, TIR	mAP	R-1	R-5	R-10
Baseline	23.7	40.6	45.7	49.0
Meta-learning Model	**24.9**	**45.5**	**49.9**	**52.9**

Table 2. The performance of our system on RGBNT100 [19] dataset. The model is trained on the training sets of RGB and TIR modalities (source domains) and then it is tested on the testing set of NIR modality (target domain). λ parameter is set to 0.2.

	Target: RGB			
Source: NIR, TIR	mAP	R-1	R-5	R-10
Baseline	30.2	53.9	59.3	61.8
Meta-learning Model	**33.3**	**58.4**	**63.9**	**66.9**

Table 3. The performance of our system on RGBNT100 [19] dataset. The model is trained on the training sets of NIR and NIR modalities (source domains) and then it is tested on the testing set of RGB modality (target domain). λ parameter is set to 0.4.

flects to the radical difference in visualising the TIR spectrum by measuring the emmitted object energy compared to RGB and NIR sensor data which measure the reflected object energy. This fundamental difference in the nature of the modalities renders the source-target generalisation even harder.

4.5.1 Tuning λ parameter

We are evaluating the effect of λ factor that controls the magnitude of meta-train loss $L_{meta-tr}$ in the final loss summation in Algorithm 1. Figure 4 demonstrates the mAP and rank-1 scores for different values of λ for each target domain settings. λ parameter is set to 0.4 for the RGB and TIR and to 0.2 for the NIR target domain. Observing the results in Figure 4, it seems that the addition of $L_{meta-tr}$ does not affect significantly the performance and λ is not a critical parameter, but provides a moderate improvement, especially in the rank-1 scores. This can be explained by the fact that even in the case of $\lambda = 0$ the meta-train domain has indirectly contributed in the $L_{meta-te}$ computation.

5. Conclusions & Future Work

In this work, we propose a domain generalisation framework for multi-modal vehicle ReID based on meta-learning training configuration. The visual modalities considered for this system involve RGB, NIR and TIR and we examine the ability of the ReID model to perform on a previously unseen modality, while evaluating the contribution of meta-learning techniques to achieve domain generalisation. The experimental analysis show that our proposed framework provides consistent improvement under all source-target domain selection settings in RGBNT100 benchmark. This proves the potential of meta-learning training to create a more generalisable ReID model compared to a model with conventional metric learning training.

Normally meta-learning methods require multiple source domains, that would allow multiple combinations of meta-train - meta-test domains. In our case that was not possible due to the dataset construction and only two source domains were available in our settings. The presence of more source modalities, even synthetic ones [8], would allow to examine the full potential of meta-learning training in across-modality generalisation. Also, another future direction for this work would be to experiment with different backbone models and more multi-modal ReID benchmarks, once available.

References

[1] Abner Ayala-Acevedo, Akash Devgun, Sadri Zahir, and Sid Askary. Vehicle re-identification: Pushing the limits of re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019*, pages 291–296. 1, 4

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Process-*

ing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS, pages 1006–1016, 2018. 3

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3425–3435, 2021. 3

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 677–683, 2018. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 248–255, 2009. 6

[6] Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, and Manmohan Chandraker. Cross-domain similarity learning for face recognition in unseen domains. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 15292–15301, 2021. 3

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. 2, 3

[8] Jinbo Guo, Xiaojing Zhang, Zhengyi Liu, and Yuan Wang. Generative and attentive fusion for multi-spectral vehicle re-identification. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1565–1572, 2022. 2, 3, 7

[9] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z. Li. Learning meta face recognition in unseen domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6162–6171, 2020. 3

[10] Yi Hao, Jie Li, Nannan Wang, and Xinbo Gao. Modality adversarial neural network for visible-thermal person re-identification. *Pattern Recognit.*, 107:107533, 2020. 2

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 14993–15002, 2021. 2

[12] Na Jiang, Yue Xu, Zhong Zhou, and Wei Wu. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking. In *2018 IEEE International Conference on Image Processing, ICIP 2018*, pages 858–862. 2

[13] Eleni Kamenou, Jesús Martínez del Rincón, Paul Miller, and Patricia Devlin-Hill. Closing the domain gap for cross-modal visible-infrared vehicle re-identification. In *26th International Conference on Pattern Recognition, ICPR*, pages 2728–2734, 2022. 2, 3

[14] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 6131–6140, 2019. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR*, 2015. 6

[16] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. A strong and efficient baseline for vehicle re-identification using deep triplet embedding. *Journal of Artificial Intelligence and Soft Computing Research*, 10, 2020. 1

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 3490–3497, 2018. 2, 3

[18] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 1446–1455. 3

[19] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 11345–11353, 2020. 1, 2, 5, 6, 7

[20] Hongchao Li, Xianmin Lin, Aihua Zheng, Chenglong Li, Bin Luo, Ran He, and Amir Hussain. Attributes guided feature learning for vehicle re-identification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–11, 2021. 1

[21] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5400–5409, 2018. 3

[22] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 3

[23] Chih-Ting Liu, Man-Yu Lee, Chih-Wei Wu, Bo-Ying Chen, Tsai-Shien Chen, Yao-Ting Hsu, and Shao-Yi Chien. Supervised joint domain learning for vehicle re-identification. In *CVPR Workshops*, pages 45–52, 2019. 2, 3, 4

[24] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In *Medical Image Computing and Computer Assisted Intervention - MICCAI*, volume 12262 of *Lecture Notes in Computer Science*, pages 475–485. Springer, 2020. 3

[25] Wu Liu, Xinchen Liu, Huadong Ma, and Peng Cheng. Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN. In *Multimedia Conference (MM) 2017*, pages 1618–1626. 2

[26] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(8):3794–3807, 2019. 1

[27] Wenjie Pan, Hanxiao Wu, Jianqing Zhu, Huanqiang Zeng, and Xiaobin Zhu. H-vit: Hybrid vision transformer for multi-modal vehicle re-identification. In *Artificial Intelligence - Second CAAI International Conference, CICAI*, volume 13604 of *Lecture Notes in Computer Science*, pages 255–267. Springer, 2022. 2

[28] Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In *4th International Conference on Learning Representations, ICLR*, 2016. 3

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[30] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12553–12562, 2020. 3

[31] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6816–6825, 2020. 2

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2818–2826. 3, 4

[33] Sebastian Thrun and Lorien Y. Pratt. Learning to learn: Introduction and overview. In Sebastian Thrun and Lorien Y. Pratt, editors, *Learning to Learn*, pages 3–17. Springer, 1998. 3

[34] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3622–3631, 2019. 2

[35] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil Martin Robertson. Ranked list loss for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 5207–5216. 3

[36] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 618–626, 2019. 2

[37] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *IEEE International Conference on Computer Vision, ICCV*, pages 5390–5399. IEEE Computer Society, 2017. 2

[38] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128:1765–1785, 2020. 2, 3

[39] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong 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*, pages 1092–1099, 2018. 2

[40] Shizhou Zhang, Yifei Yang, Peng Wang, Guoqiang Liang, Xiuwei Zhang, and Yanning Zhang. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *IEEE Trans. Image Process.*, 30:8861–8872, 2021. 2

[41] Jianan Zhao, Fengliang Qi, Guangyu Ren, and Lin 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)*, pages 2225–2235, 2021. 1

[42] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[43] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6277–6286, 2021. 3

[44] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *CoRR*, abs/2103.02503, 2021. 2

[45] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5056–5069, 2022. 3

[46] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 3

[47] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XVI 16*, pages 561–578. Springer, 2020. 3

[48] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR*, 2021. 3

[49] Jianqing Zhu, Huanqiang Zeng, Zhen Lei, Shengcai Liao, Lixin Zheng, and Canhui Cai. A shortly and densely connected convolutional neural network for vehicle re-identification. In *24th International Conference on Pattern Recognition, ICPR 2018*, pages 3285–3290. 2