Pair-dependent representations for Same-View Vehicle recognition

Anis Yassine BEN MABROUK[1] Antoine TADROS[1] Axel DAVY[2] Rafael GROMPONE VON GIOI[1] Gabriele FACCIOLO[1]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190, Gif-sur-Yvette, France

[2]HGH Infrared systems, 10 Rue Maryse Bastié, 91430 Igny, France

Résumé – Les méthodes d'apprentissage profond pour la reconnaissance d'objets rencontrent des difficultés pour généraliser. Pour savoir si deux images montrent le même objet, une représentation est calculée pour chaque image séparément. Cette approche est insuffisante pour des applications nécessitant une attention aux détails fins. Dans cet article, nous nous concentrons sur le problème de la reconnaissance de véhicules à partir d'images de même point de vue. Nous proposons une méthode basée sur l'appariement de points clés qui combine des informations indépendantes et dépendantes des paires pour prédire si deux images sont des instances du même objet. Nous proposons un protocole d'évaluation axé sur la généralisation à des types de véhicules non vus. Des expériences approfondies montrent que la méthode proposée se généralise mieux aux types de véhicules non vus que l'état de l'art indiquant ainsi une plus grande capacité de généralisation.

Abstract – Deep metric learning approaches for object instance recognition struggle with generalization. To infer if two images are of the same instance, a representation is computed for each image, independently of the other. This is insufficient for reliably finding distinctive fine-grained details. In this paper, we focus on the problem of same-view vehicle recognition. We propose a keypoint matching-based method that combines pair-independent and pair-dependent information to predict whether two images are instances of the same object. We propose an evaluation protocol focused on generalization to unseen types of vehicles. Extensive experiments show that the proposed method generalizes better to unseen vehicle types than the state of the art.

1 Introduction

Recognition in computer vision consists of determining if two images pertain to the same instance of an object while being robust to pose [7], viewpoint changes [16], occlusions [15], and degradations [18]. Nowadays, the most popular approaches to recognition rely on deep metric learning [8]. These approaches learn a mapping from the image space into a latent space that clusters objects of the same instance together. When employing a basic contrastive learning approach, networks learn to extract common features between images. However, without appropriate supervision, the network can overfit on basic, common patterns (such as color and general object shape) and fail to generalize well [8]. Additionally, obtaining a comprehensive representation requires large amounts of data [18]. When it comes to objects, they can be anything from faces, animals, people to vehicles. The vehicle setting is particularly challenging due to the very similar appearances of different vehicle instances. This is exacerbated by the lack of common information in cases when the matched objects are from different viewpoints. In this paper, we focus on the context of vehicles seen from the same point of view. We propose an evaluation protocol that puts forward generalization to unseen vehicle types. Additionally, to recognize vehicle instances, we propose a binary decision network that leverages both pair-specific descriptors through Lightglue [10] and pair-independent descriptors. Our method yields competitive performance and generalizes better than the state of the art.

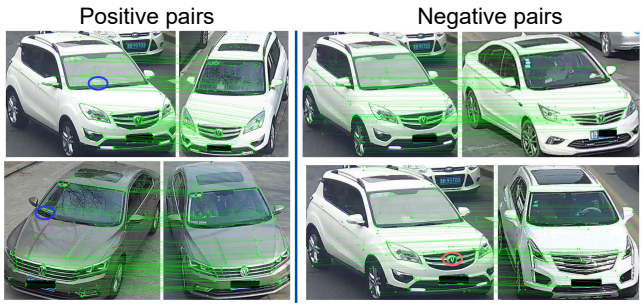

Figure 1 – When comparing image pairs for instance recognition, finding discriminative information in one image depends on its paired counterpart: the vehicle logo (in red) distinguishes the second negative pair but not the first. Different windshield elements (in blue) are relevant across the positive pairs illustrated. Lightglue [10] is a keypoint-matcher that associates spatially coherent details (in green) across images, as seen in the above figure, producing pair-dependent information in the process. We combine this information with pair-independent information to recognize vehicle instances.

2 Related works

Deep metric learning aims to learn an object representation such that the obtained mapping respects a similarity measure. For that, contrastive approaches are often used [18]. These consist of pulling samples of the same identity closer (i.e., positives) while pushing samples of different identities away (i.e., negatives) in a latent space. Various loss functions are used to achieve this, such as classification loss [21], triplet loss [6] and InfoNCE loss [14]. Sample selection has a major impact on what type of information the network learns [20].
Local feature matchers make correspondences between two sets of local features such as SuperPoint [3]. The goal is to filter matches from non-matches using spatial consistency: The relative position of the matched key points should be similar

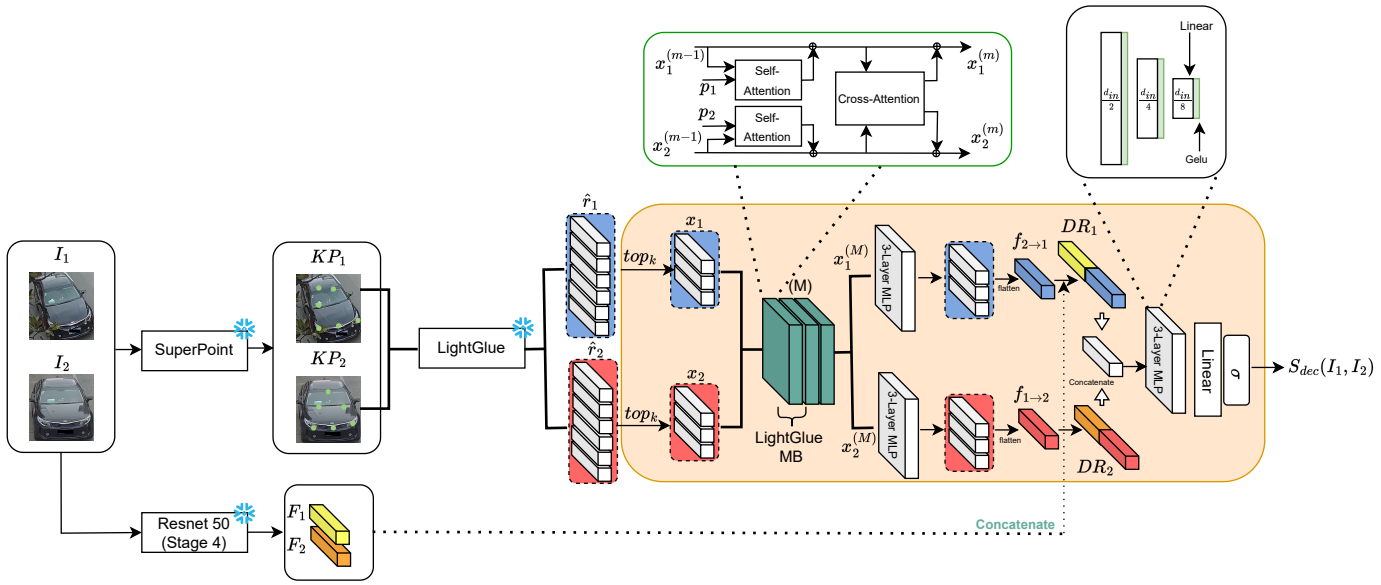

Figure 2 – An illustration of the proposed method. SuperPoint [3] features are extracted from image pairs and matched through Lightglue [10]. The top k descriptors (in terms of matching score) of each are fed together to M additional Lightglue matching blocks (MB) and are then separately processed through a 3-layer MLP to create pair-dependent representations $f_{2\to1}$ and $f_{1\to2}$. They are fed together with ResNet [4] features to another 3-layer MLP to produce a final decision score S_{dec}.

across images. Deep learning-based methods achieve state-of-the-art performance. Lightglue [10] proposes a transformer-based approach that solves the partial assignment problem with a network instead. The method is trained to associate a key point and its homography-warped counterpart. A confidence classifier and an exit criterion are added to reduce computations.

Vehicle Re-identification consists of recognizing vehicles through multiple views. Similar to the recognition task, the goal is to learn a similarity measure. Recent deep learning-based methods use metric learning to achieve state-of-the-art performance [8]. Some state-of-the-art approaches [13] use the early stages of the ResNet [4] backbone, adapting them with a classification and a triplet loss. Others like TransReID [5] use visual transformers instead to obtain an image representation out of local patch representations. They also propose a Jigsaw module that produces additional image representations out of different subsets of local representations. When using a contrastive approach, networks need proper supervision to learn relevant fine-grained information [9]. Hence, some methods complement their training with pose information [17]. Others fuse different types of representations [1], a view-dependent and a view-independent representation. That said, the relevance of information in an image changes depending on what that image is compared to. Hence, some methods like DCC [19] extract features from each image and compute an attention matrix to create co-dependent representations. They also use an LSTM-based comparator that mimics the foveation of human eyes to produce relative representations.

3 Proposed method: Dual representation network

A recognition method built only on pair-independent representations is not well-suited for finding fine-grained details [19]. The reason is that detecting discriminative regions depends on a given pair, as illustrated in Figure 1. Hence, we propose a method that combines pair-dependent representations obtained through Lightglue [10] with pair-independent representations. Contrary to [19], we use spatially coherent information complemented with pair-independent information.

As illustrated in Figure 2, n SuperPoint [3] keypoints $KP = (r, p, s)$ are extracted from each image, where $r \in \mathbb{R}^{n\times d}$ are the representations with $d = 256$, $p \in \mathbb{R}^{n\times 2}$ are coordinates and $s \in \mathbb{R}^n$ are detection scores. They are then matched with Lightglue (LG):

$$(\hat{r}_1, \hat{r}_2) = LG(KP_1, KP_2). \tag{1}$$

Out of the n transformed representations $\hat{r}_1$ and $\hat{r}_2$, we select the **top** k in terms of matching scores to obtain $x_1, x_2 \in \mathbb{R}^{k\times d}$. These representations are given as input to M additional Lightglue matching blocks (MB) which consist of 2 self-attention transformers applied separately to both inputs followed by a cross-attention transformer:

$$(x_1^{(m)}, x_2^{(m)}) = MB(x_1^{(m-1)}, x_2^{(m-1)}). \tag{2}$$

Their result is later fed into a 3-layer MLP (3 fully connected layers with GELU non-linearities) and flattened to create pair-dependent representations $f_{2\to1}, f_{1\to2} \in \mathbb{R}^{\frac{kd}{8}}$.

$$f_{2\to1} = flat(MLP_3(x_1^{(M)})),\ f_{1\to2} = flat(MLP_3(x_2^{(M)})). \tag{3}$$

As for the pair-independent representation, we use the earliest 4 stages of a ResNet-50 [4] backbone to extract representations $F \in \mathbb{R}^b, b = 2048$ that are normalized using a batchnorm1D. Both pair-dependent and pair-independent representations are then concatenated, producing dual representations:

$$DR_1 = [F_1, f_{2\to1}],\ DR_2 = [F_2, f_{1\to2}]. \tag{4}$$

Both dual representations are then concatenated and fed to another 3-layer MLP to create a final representation:

$$FR = MLP_3([DR_1, DR_2]). \tag{5}$$

The final decision scores are then computed as:

$$S_{dec}(I_1, I_2) = \sigma(P(FR)) = (S_0, S_1) \in \mathbb{R}^2, \quad (6)$$

where P is a linear projection and σ is a softmax function. S_0 and S_1 correspond to the predictions of labels 0 and 1, respectively.

The ResNet [4] backbone, SuperPoint [3] and Lightglue [10] networks are all **frozen**. The rest is trained and supervised with a negative log-likelihood loss:

$$\mathcal{L}_{nll} = -y \log(S_1) - (1-y) \log(S_0), \quad (7)$$

where y are the ground truth labels, taking a value of 1 for negatives and 0 otherwise. We refer to this combined approach as DRnet (dual representation network).

4 Experimental setup

We evaluate performances on the VeRI-Wild[12] and VeRi776[11] datasets. We create lists of image pairs I_1 and I_2, where a pair is either of the same identity (positives, labeled 0) or of different identities (negatives, labeled 1). Train and evaluation sets contain positives, difficult negatives (vehicles of the same type and color), and random negatives (of random type and color) with a 50-25-25% ratio. To study how well different approaches generalize, we split the dataset identities into 2 groups, one with common vehicle types such as Sedans, SUVs, minivans, and business multi-purpose vehicles (MPVs) and another with the remaining types. The former will be used for training and testing (Easy test), while the latter will only be used for testing (Hard test). The easy set identities are randomly split identity-wise into training, validation, and testing identities, with an 80-10-10 ratio. Lastly, we focus on vehicles seen from the same point of view as telling apart two images of similar vehicles with no viewpoint overlap can be ambiguous. Hence, all test set pairs are vehicles of the same view (front-front, rear-rear). We report the binary accuracy.

4.1 Training protocol

We retrain BoT, DCC and TransReID following the author's guidelines. When it comes to the DCC's number of glimpses, We use $T = 4$ for stability. We empirically fine-tune decision thresholds for the Lightglue$_{ratio}$, BoT, DCC, and TransReid. We randomly sample 100,000 pairs per epoch during training for 80 epochs and use a learning rate of 0,001, a weight decay of 4e-4, and a batch size of 256 with SING [2] optimizer. We resize the images to a 256×256 resolution when using ResNet-50 and to 1024×1024 to extract $n = 256$ keypoints with SuperPoint. Lightglue [10] and SuperPoint [3] are not retrained. We use imagenet weights for ResNet-50 [4]. All 8 layers of Lightglue are used without point pruning. We use $M = 3$ matching blocks and select the top $k = 32$ Lightglue features.

4.2 Results

We compare our proposed method to methods that use pair-independent representations (BoT [13] and TransReID [5]) and pair-dependent representations (DCC [19]). We also use Lightglue$_{ratio}$ [10], the ratio of the number of Lightglue matches (the number of matched keypoints out of available keypoints) as a baseline.

Test set \ Method		Lightglue$_{ratio}$ [10]	BoT [13]	DCC [19]	TransReID [5]	**DRnet**
VeRI-776	Easy	58,52	86,28	84,5	**90,18**	89,75
	Hard	65,64	76,1	65,92	73,8	**79,01**
VeRI-WILD	Easy	66,19	91,13	87,14	91,53	**96,61**
	Hard	73,05	85,32	71,68	87,44	**88,32**
VeRI-WILD$_{train}$ → VeRi776$_{test}$	Easy	-	65,74	70,3	70,97	**76,14**
	Hard	-	67,85	65,45	**71,78**	71,17

Table 1 – A comparison of the test accuracy (%) of various vehicle recognition methods on the VeRI-WILD and VeRi776 datasets. The best values are in bold, while the second best are underlined. The proposed method (DRnet) achieves state-of-the-art performance by combining pair-dependent and pair-independent representations.

Test set \ Method	Lightglue$_{ratio}$ [10]	DRnet (without ResNet features)	DRnet
Easy	58,52	83,5	**89,75**
Hard	65,64	75,35	**79,01**

Table 2 – Test accuracy (%) of the proposed method with and without the ResNet pair-independent representation on the VeRI776 dataset. The highest values are in bold. Adding additional processing with Lightglue matching blocks and using pair-independent representations complements the information contained in Lightglue features.

We notice in Table 1 that on the VeRi776 dataset, TransReID yields the best performance with DRnet slightly behind on the easy set. On the hard test set however, DRnet performs better than TransReID and the other methods. On the VeRI-Wild dataset, DRnet outperforms all other methods on both the easy and hard sets. We notice a similar behavior when cross-testing the previously mentioned methods (training on VeRI-Wild and testing on VeRi776) on the easy set, with the DRnet being marginally behind on the hard set. DRnet generalizes better overall to unseen vehicle types thanks to the added spatially-coherent information.

4.3 Ablation

On pair-dependent and pair-independent information. We study in Table 2 the impact of adding matching blocks and pair-independent representations. We see that compared to using the Lightglue representations as they are through the ratio of matches, further processing them (without adding ResNet features i.e, $DR_1 = f_{2\to1}$ and $DR_2 = f_{1\to2}$ in (4)) drastically improves performance. This shows that, as expected, fine-tuning the representations adapts them better to the recognition case. When adding ResNet features, we notice they drastically improve performance on both the easy and hard set. This is due to how Lightglue provides a sparse description of an image and is hence insufficient alone in a large number of cases.

The information contained in Lightglue. Lightglue yields two types of information: representations and keypoint matching scores, which are used to solve the partial assignment problem. To understand the importance of Lightglue [10] for our proposed method, we modify the representation fed to DRnet (without ResNet features) right before the matching blocks (MB) in (2) and report the results in Table 3.

First, instead of using the top $k = 32$ Lightglue representations in terms of matching scores (referred to as LG_{32}), we

	SP_{32}	$SP_{LG,32}$	LG_{32}
Easy	63,6	88,73	93,22
Hard	58,02	84,29	92,03

Table 3 – Test accuracy (%) of DRnet (without ResNet features) when using the top $k = 32$ SuperPoint features in terms of detection score, SuperPoint features in terms of Lightglue matching scores, and Lightglue features in terms of matching scores on the VeRI-Wild dataset. Both the partial assignment information and representations of Lightglue are important for DRnet. Adding the partial assignment information obtained from Lightglue vastly boosts performance.

use their SuperPoint counterparts: we pick out the SuperPoint representations associated with the $k = 32$ highest Lightglue matching scores. We refer to this as $SP_{LG,32}$. We notice a slight decrease in performance, showing that Lightglue representations were more adapted to the recognition task. Next, we use the top k representations of SuperPoint in terms of detection scores (referred to as SP_{32}) instead of the top matched with Lightglue. Compared to $SP_{LG,32}$, we notice a substantial drop in performance. This shows that, as expected, using the set of information shared between images is more adapted for recognizing vehicles. Both the assignment scores and representations of Lightglue are, hence, important for recognition.

5 Conclusion

In this paper, we present our binary decision method that leverages pair-dependent and pair-independent representations for same-view vehicle recognition and propose an evaluation protocol that focuses on generalization to unseen vehicle types. We conclude that pair-dependent and pair-independent representations contain complementary information that can be combined for better performance. Our proposed method leverages dual representations built on spatially consistent Lightglue features and generalizes better to unseen vehicle types than the state-of-the-art.

Acknowledgments: This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011013861R2 made by GENCI

References

[1] Y. Bai, J. Liu, Y. Lou, C. Wang, and L.-Y. Duan. Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification. *IEEE TR PARTS MATER*, 44(10), 2021.

[2] A. Courtois, D. Scieur, J.-M. Morel, P. Arias, and T. Eboli. Sing: A plug-and-play dnn learning technique. *arXiv preprint arXiv:2305.15997*, 2023.

[3] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. IEEE/CVF CVPR workshops*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF CVPR*, 2016.

[5] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. In *Proc. IEEE/CVF ICCV*, 2021.

[6] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[7] T. Hou, S. Wang, and H. Qin. Vehicle matching and recognition under large variations of pose and illumination. In *2009 IEEE COMP SOC ANN*, 2009.

[8] S. D. Khan and H. Ullah. A survey of advances in vision-based vehicle re-identification. *Comput. Vis. Image Underst.*, 182, 2019.

[9] Y. Li, K. Liu, Y. Jin, T. Wang, and W. Lin. Varid: Viewpoint-aware re-identification of vehicle based on triplet loss. *IEEE Trans. Intell. Transp. Syst.*, 23(2), 2020.

[10] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys. Lightglue: Local feature matching at light speed. In *IEEE/CVF ICCV*, 2023.

[11] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proc. ECCV*. Springer, 2016.

[12] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proc. IEEE/CVF CVPR*, 2019.

[13] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proc. IEEE/CVF CVPR Workshops*, 2019.

[14] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[15] C. Steger. Occlusion, clutter, and illumination invariant object recognition. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 34(3/A), 2002.

[16] X. Sun and L. Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proc. IEEE/CVF CVPR*, 2019.

[17] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proc. IEEE/CVF ICCV*, 2019.

[18] H. Wang, J. Hou, and N. Chen. A survey of vehicle re-identification based on deep learning. *IEEE Access*, 7, 2019.

[19] L. Wu, Y. Wang, J. Gao, M. Wang, Z.-J. Zha, and D. Tao. Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2), 2020.

[20] H. Xuan, A. Stylianou, and R. Pless. Improved embeddings with easy positive triplet mining. In *Proc. IEEE/CVF WACV*, 2020.

[21] Y. Zhai, X. Guo, Y. Lu, and H. Li. In defense of the classification loss for person re-identification. In *Proc. IEEE/CVF CVPR Workshops*, 2019.