Joint Image and Feature Levels Disentanglement for Generalizable Vehicle Re-identification

Zhenyu Kuang, Chuchu He, Yue Huang, *Member, IEEE*, Xinghao Ding, and Huafeng Li

***Abstract*— Domain generalization (DG), which doesn't require any data from target domains during training, is more challenging but practical than unsupervised domain adaptation (UDA). Since different vehicles of the same type have a similar appearance, neural networks always rely on a small amount of useful information to distinguish them, meaning that is more significant to remove ID-unrelated information for vehicle re-Identification (re-ID). Therefore, it is the key to eliminating the interference of a large amount of redundant information for the generalizable vehicle re-ID method. To address this unique challenge, we propose a novel disentanglement learning method that encourages variational autoencoder (VAE) network to reduce ID-unrelated features of vehicles by minimizing image reconstruction errors and providing sufficient representation to vehicle labels. To capture the intrinsic characteristics associated with the DG task, our core idea is to build the identity information streaming framework to separate ID-related and ID-unrelated information at the image and feature levels. In contrast with the general decoupling methods, our method leverages the decoupling of joint image and feature levels to extract more generalizable features. Furthermore, we present a brand-new vehicle dataset of truck types named "Optimus Prime (Opri)", which includes multiple images of each truck captured by cameras at different high-speed toll gates. Experimental results on public datasets demonstrate that our method can achieve promising results and outperform several state-of-the-art approaches. Our codes and models are available at JIFD.**

***Index Terms*— Vehicle re-identification, variational autoencoder, domain generalization, representation disentanglement.**

I. Introduction

As an important branch of intelligent monitoring system, object re-ID has become an indispensable research topic in security management applications due to its non-contact advantages and the urgent need for intelligent video analysis. Supervised deep learning methods [1], [2], [3], [4], [5], [6], [7], [8] based on discriminative feature representation learning can effectively improve the recognition performance on re-ID tasks. Notably, the person re-ID task [9], [10] of UDA has recently gained increasing attention and its performance is even comparable to supervised methods. Although person re-ID methods based on UDA have achieved remarkable progress in recent years, it has not been fully explored for the same vehicle task, especially generalizable vehicle re-ID task.

Many current domain adaptation methods [11], [12], [13], [14] for vehicle re-ID mainly use clustering or pseudo-label strategies to improve recognition performance. However, these methods [15], [16], [17], [18], [19], [20] may be difficult to implement in practical applications, since unlabeled data in the target domain is required for training. In comparison to person re-ID task, vehicle re-ID encounters additional challenges. For example, vehicles with different identities may have the same brands and colors, humans with knowledge priors can still easily distinguish them in this realistic scenario. But it can be particularly difficult for neural networks to extract discriminant ID-related features from different vehicles that are highly similar in appearance [21]. This problem exists in a wide range of computer vision tasks, such as fine-grained recognition tasks [22], [23].

Facts have proved that achieving pedestrian matching does not need to extract all the characteristics related to pedestrian from the image, and only the ID-related features are sufficient [24]. Therefore, how to effectively separate the ID-related and ID-unrelated features from the images of a vehicle is the key to the re-ID methods. Recent disentangled feature representation methods [25], [26], [27] focus on learning invariant features to achieve this purpose. However, existing machine learning models are data-driven and usually require samples of various camera angles and illumination. As a result, they may rely on redundant or pseudo-correlation features [28], [29], [30], [31], [32], and lose generalizable performance [33]. Moreover, most of the existing decoupling methods, which use multiple different images of the same identity as the input of the backbone of the neural network, only suppress ID-unrelated features at the feature level through end-to-end learning. Because these methods do not consider the suppression of the ID-unrelated information at the image level before the feature extraction, it leads to a large amount of redundant information in the image directly flowing to the neural network training. Assuming that redundant information can be removed or decoupled at the image level, we can easily suppress ID-unrelated features at the feature level.

Manuscript received 9 January 2023; revised 18 May 2023 and 13 July 2023; accepted 23 August 2023. Date of publication 22 September 2023; date of current version 29 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant U19B2031, Grant 61971369, Grant 52105126, Grant 82272071, and Grant 62271430; and in part by the Fundamental Research Funds for the Central Universities under Grant 20720230104. The Associate Editor for this article was H. Huang. *(Corresponding author: Xinghao Ding.)*

Zhenyu Kuang, Chuchu He, Yue Huang, and Xinghao Ding are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: kmustkzy@126.com; hechuchu@stu.xmu.edu.cn; huangyue05@gmail.com; dxh@xmu.edu.cn).

Huafeng Li is with the Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: hfchina99@163.com).

Digital Object Identifier 10.1109/TITS.2023.3314213

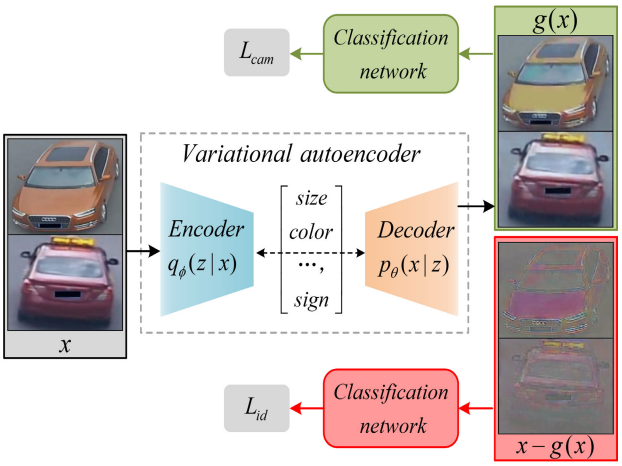

Fig. 1. Flowchart of the proposed method. The method extracts vehicle ID-related features and removes redundant information through a simple architecture. Since the task of the VAE network is to be as consistent as possible with x, the VAE network generates $g(x)$ containing a large amount of redundant information from the original image x. Meanwhile, the ID classification network can complete the re-ID DG task with reconstruction errors $x-g(x)$. That is, we squeeze the information that x provides through a bottleneck formed by the VAE network.

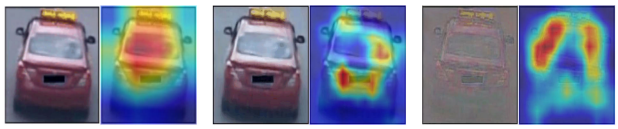

Fig. 2. Visualization results of grad-CAM [40] based on CNN method. $RC(x, g(x))$ represents the reconstruction accuracy between the original image x and the $g(x)$ generated by VAE (i.e., $RC(x, g(x))=1-\frac{\|x-g(x)\|_2}{\|x\|_2}$).

Inspired by the disentangled representation learning methods [34], [35], [36], our key idea is to control the reconstruction accuracy of VAE [37] from the perspective of information bottleneck theory [38], [39], so as to effectively separate out redundant information $g(x)$ which is unrelated to the vehicle, as shown in Fig.1. The goal of VAE network forces the two classification networks to pay attention to the most discriminating information for feature-level disentanglement, while the goal of the two classification networks forces VAE to reconstruct the ID-unrelated part of the vehicle for image-level disentanglement. In contrast, we believe that $x-g(x)$, which contains a small amount of information in the original image x, can provide sufficient ID-related feature representations. The advantage of establishing an information bottleneck is to reduce the interference of redundant information to the classification network and improve its generalization ability. Fig.2 presents the visualization results of our proposed method from the reconstruction perspective. For simplicity, we can represent $x-g(x)$ as $r(x)$ in this subsequent paper. In Section VI of this paper, a large number of ablation experiments demonstrate the positive effect of the information bottleneck on the DG task.

Overall, our work is driven by two aspects. On the one hand, our method is distinct from traditional domain-invariant feature methods [12], [42], [43], [44] in that it does not require image augmentation or feature reconstruction at all. Moreover, a simple but effective decoupling framework is

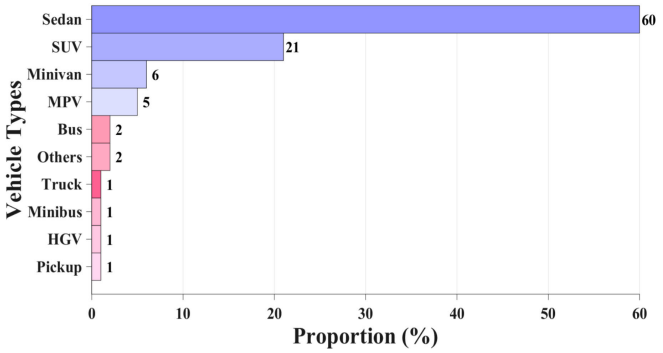

Fig. 3. The vehicle type distribution of a long-tailed dataset (e.g., the VeRI-Wild 2.0 dataset [41] with more than 40,000 IDs).

introduced for the first time, which includes both image-level disentanglement and feature-level disentanglement. To complete image-level disentanglement and form an information bottleneck, we consider the reconstruction quality of GAN [45] and VAE [46] for image-level disentanglement. We find that the GAN models attempt to rebuild fine-grained details, but it often introduces additional irrelevant information [47]. Compared with GAN, the disadvantage of VAE is that it does not use an adversarial network, so it tends to produce unrealistic and blurry images. To alleviate the problems caused by these two methods of image generation, we replace $g(x)$ generated by VAE with $r(x)$ to extract ID-related features by a supervised classification network. As a result, the proposed method can effectively reduce the irrelevant information introduced by the generated source domain images with poor qualities. And it can encourage the VAE network to reduce ID-unrelated features of vehicles by minimizing image reconstruction errors and providing sufficient representation to vehicle labels.

On the other hand, the exploration of domain adaptation methods for vehicles is limited by the collection of vehicle data types. Although the existing five commonly used datasets of the real-world VeRi–776 [48], VeRI-Wild [49], CityFlow [50], VehicleID [51], Vehicle-1M [52] and the synthetic VehicleX [53] have been widely used in vehicle domain adaptation tasks and achieved good performance [18], [19], [54], it may still be difficult to maintain the same level of recognition performance in a specific real-world scenario. Another reason is that the long-tailed problem [55] also exists in the existing public vehicle datasets [41], [52] (see Fig.3), but this problem has rarely been studied in the vehicle re-ID task. According to the number statistics of vehicle types in these datasets, the total number of other cars accounts for a large proportion compared with trucks. It limits the generalization ability of the re-ID models in real-world scenarios. In order to combine existing public datasets to study generalized vehicle re-ID, we build a fine-grained vehicle dataset "Optimus Prime (Opri)", which is composed entirely of truck images. Fig.4 shows sample images of the Opri dataset.

In summary, the contributions of this paper can be summarized as follows:

- To the best of our knowledge, a generalizable vehicle re-ID task is proposed for the first time in this paper. We design an information bottleneck scheme to reduce

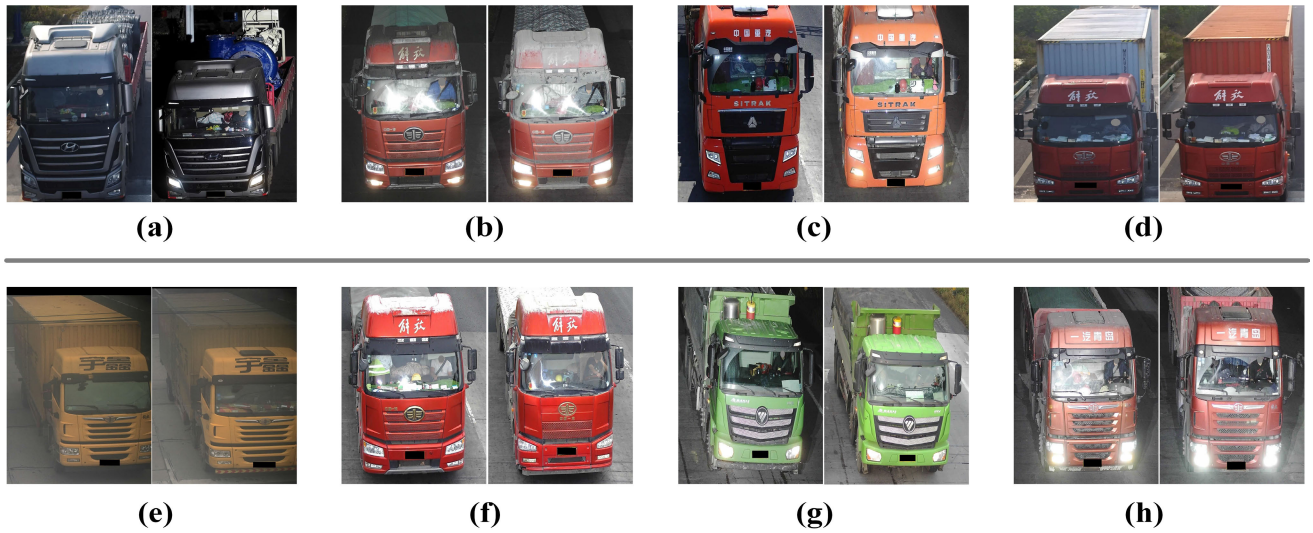

Fig. 4. Sample images of the Optimus Prime dataset. Each vehicle has at least two images in our dataset. The first row shows significant intra-instance differences for the same vehicles, which originate from changes in vehicle appearance and light intensity. Images in the second row all come from different vehicle identities with similar appearance and show slight inter-instance differences of similar vehicles. License plates and faces are obscured due to privacy.

the redundant information in the source domain dataset, so as to reduce the interference to ID-unrelated features and obtain a more generalized feature representation.

- We propose a novel disentanglement learning method called Joint Image and Feature levels Disentanglement (JIFD), which integrates the benefits of image-level disentanglement and feature-level disentanglement in one framework. We for the first time show that the reconstruction errors generated by the VAE network can be used as an information bottleneck to handle generalizable vehicle re-ID task. By minimizing the mutual information of ID-related and ID-unrelated features, feature-level disentanglement can be achieved and can guide image-level disentanglement to filter out redundant information of vehicle images.
- A large-scale truck dataset, Opri, is proposed for generalizable vehicle re-ID tasks in real-world highway scenarios. Opri is distinguished from existing datasets in three aspects, including data type, camera perspective, and practical application.

II. Related Work

In this section, we briefly introduce recent existing re-ID methods, which include unsupervised domain adaptation and domain generalization methods. Furthermore, we focus on methods related to representation disentanglement.

A. UDA for Vehicle Re-ID

Compared with supervised vehicle re-ID methods, completely unsupervised vehicle re-ID has not been fully explored. To achieve this goal, some studies [17], [56], [57] proposed to utilize camera information for unsupervised tasks. Zheng et al. [17] proposed a viewpoint-aware clustering algorithm, which assumed that images of vehicles under adjacent viewpoints shared the similarity in appearance. But it is still challenging for the scenes of only two larger differential perspectives. Yu et al. [56] achieved comparable performance through a clustering-based technique using only the label information of multiple cameras. However, the existing pseudo-label methods did not consider the distribution difference between cameras, but directly measured the feature similarity of different samples, which led to the reduction of the accuracy of cross-camera labeling calculation. To solve this problem, Xuan and Zhang [57] designed a two-stage pseudo-labels generation method to calculate sample similarities across cameras for obtaining high-confidence pseudo-labels and better optimization of the model. UDA methods are usually considered to improve the generalization capacity of CNN models. It is different from fully unsupervised methods. The goal of UDA is to transfer knowledge learned from one or more labeled source domains to an unlabeled target domain. MMT [18] proposed to refine the generation process of soft pseudo-labels by mutual mean-teaching. Ge et al. [19] proposed a self-paced learning framework to dynamically update hybrid memory for preserving the most reliable clusters. Wang et al. [58] proposed an uncertainty-aware clustering framework to alleviate the noise pseudo-label problem to a large extent for UDA tasks. Most domain adaption or clustering-based approaches try to alleviate the effects of unreliable pseudo-labels, or find the high-confidence samples from generated pseudo-labels. As unlabeled data in the target domain is required for training, the current UDA method still cannot solve the real-time problem.

B. DG for Re-ID

Despite its practical value in the deployment of real-world applications, generalizable re-ID has rarely been studied. Recent works [59], [60], [61], [62] on this topic have mainly focused on learning domain-invariant features by minimizing the inter-domain differences of the same identities. Choi et al. [63] proposed a meta batch-instance normalization (MetaBIN) approach. Their approach not only employs

normalization layers but also uses meta-learning to prevent overfitting of the model in labeled source domains. Yu et al. [64] proposed a multiple domain experts collaborative learning framework, which utilizes domain-specific experts for better exploiting all training domains. Semi-supervised knowledge distillation (SSKD) [65] proposed to learn domain-invariant representations for generalizable person re-ID. The above-mentioned methods alleviate the overfitting problem by learning identity-invariant features at the feature level, while our method aims to joint image and feature levels decoupling to make neural network have better generalization ability.

In order to further study the generalization performance of re-ID models, Liao and Shao proposed three different deep metric learning methods [66], [67], [68]. The Query-Adaptive Convolution (QAConv) method [66] constructed a class memory module to calculate the corresponding relationship between feature maps of different samples on the fly. Secondly, Liao and Shao found the vision transformer and the vanilla transformer methods lack the attention between image pairs and proposed TransMatcher [67]. Given pairs of images, they adapted the self-attention mechanism in transformers and applied it to metric learning for image matching. Furthermore, they proposed a graph sampling strategy and combined it with QAConv method (QAConv-GS) [68] to effectively improve the convergence rate of the model and generalization performance under unknown scenarios. The closest related work to ours was [25], which proposed a novel generative adversarial network called identity shuffle GAN (IS-GAN). This approach recombines decoupled identity-related and -unrelated features to generate new pairs of images. But this work could not effectively prevent from introducing noise information caused by generated images with poor qualities. Another work [26] proposed two orthogonal streams with disentangled feature representation, which encouraged the disentangled face features to encode a full representation of the input image. In contrast, we did not use the images generated by VAE network to extract ID-related features directly. In particular, to the best of our knowledge, domain generalization research work on vehicle re-ID has not progressed. This paper is the first to propose domain generalization for vehicle re-ID tasks.

C. Representation Disentanglement

Different from the traditional domain generalization approaches, representation disentanglement [41], [46], [69], [70] was designed to mine potential generative factors in data and be able to enable the manipulation of relevant representations for specific tasks [34], [71], [72], [73]. The motivation of β-VAE [46], [74] is to relax the information bottleneck [38], [39], so that it can obtain a more comprehensive representation ability and higher reconstruction accuracy. β-VAE added an additional hyperparameter β to the VAE [75], [76] target function, which effectively controlled the encoding capacity and obtained more potential representation factors. However, without the premise of inductive biases both on the models and the dataset [77], this is basically impossible to achieve the purpose of decoupling. Hence, it is necessary to introduce supervisory information to VAE models and the supervisory information may provide key guidance for representation disentanglement, as shown in the dual decoupling strategy in this paper as well. Hadad et al. [73] presented a simpler two-step adversarial architecture for representation disentanglement. Yang et al. [36] introduced a simple β-VAE classifier architecture to achieve the purpose of class-disentanglement and applied it to adversarial detection and adversarial defense tasks. Class-disentanglement [36] also provided a new perspective for understanding how neural network classifiers predict the class of images through the information bottleneck. These decoupling methods achieved satisfactory results in specific tasks.

In contrast to previous re-ID works, our work is motivated to design an information bottleneck to complete the decoupling task. JIFD integrates the benefits of image-level disentanglement and feature-level disentanglement in one framework. To better complete feature-level disentanglement, we minimize the mutual information between ID-related and ID-unrelated features. Moreover, our proposed method can be used as a plug-and-play training strategy and combined with other state-of-the-art methods to improve performance.

III. Proposed Method

Our method utilizes the reconstruction of VAE and two classification branches to perform the dual disentanglement. We exploit this simple but effective architecture to extract ID-related features while separating redundant ID-unrelated features. Then, feature-level disentanglement of mutual information is introduced to refine the ID-related information. We formulate our task as follows: let $\left\{\left(x^i, y^i_{id}, y^i_{cam}\right) \in S\right\}_{i=1}^{n_s}$ be the source domain with S denoting the source dataset, which contains the i-th instance x^i and its corresponding ID label y^i_{id}, camera label y^i_{cam}, where n_s is the total number of vehicle images. The architecture of our model is shown in Fig.5. Given an input image x^i, we use the VAE network to complete the image reconstruction and generate the image $g(x^i)$, and then obtain the image $x^i - g(x^i)$ by element-wise subtraction between images x^i and $g(x^i)$. For simplicity, we denote x^i as x. By the reconstruction of VAE, the input image x is decomposed into two parts at the image level, which are provided to the two branches explicitly as necessary information. The information shunting of the dual classification network can maximize the compression of the information of the input image x, and effectively preserve the ID-related features corresponding to the output label. Therefore, the extracted features are more discriminative for domain generalization re-ID tasks.

A. Image-Level Disentanglement

Vehicle re-ID faces the challenge that vehicles of the same type and color have highly similar appearances and only differ in subtle features that play a small part in the overall image [24]. Therefore, to extract ID-related features of these subtle differences, we use image reconstruction to achieve image-level disentanglement and form an information bottleneck. We have observed that GANs and VAEs serve different modeling purposes. GANs primarily focus on generating new images, while VAEs are better suited for modeling

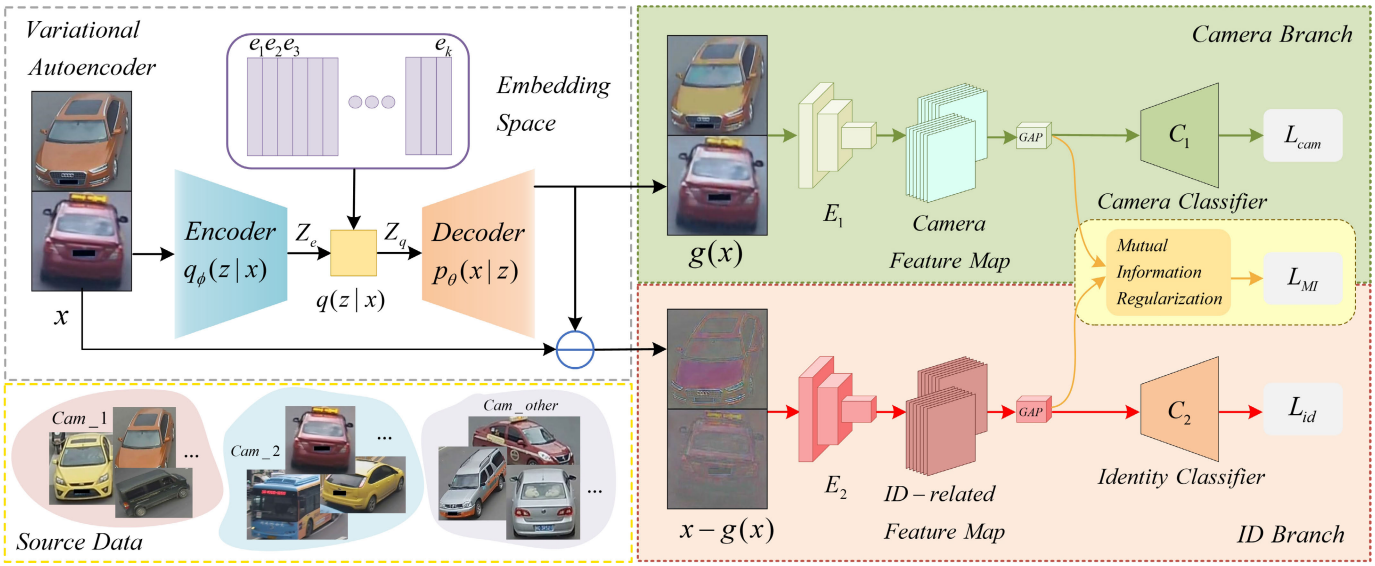

Fig. 5. Overview of the proposed architecture for vehicle re-ID. Firstly, the original image x are fed into VAE network to complete image reconstruction and $g(x)$ is obtained. $r(x)$ is further obtained by element-wise subtraction. Then, $g(x)$ and $r(x)$ are sent into two classification branches and trained according to camera and ID labels respectively. Camera classification branch is responsible for $g(x)$ to complete feature extraction of camera styles, that is, related domain styles and ID-unrelated features. ID classification branch utilizes the minimum necessary information from $r(x)$ to maximize ID-related features.

data distribution and learning implicit representations. VAEs belong to the category of deep generative models, training both a probabilistic encoder and a decoder simultaneously. On the other hand, GANs incorporate the concept of adversarial learning, where a discriminator is trained to differentiate between generated and real data, while the generator aims to minimize this distinction. Although GANs can produce natural-looking images, they often alter the logo and overall shape of the vehicle, which is not ideal for image reconstruction. Additionally, GANs lack an effective inference mechanism and only focus on data generation process estimation and the instability of GANs in the optimization process may cause the collapse of the model [78], [79]. Hence, we use VAE for image reconstruction.

Since the VAE aims to reconstruct the original image to the greatest extent possible, the reconstructed image retains plenty of features shared between categories, such as background color, illumination, and patterns, which are redundant for vehicle re-ID tasks. By the reconstruction of VAE, the input image is divided into two parts at the image level. ID-related and ID-unrelated features can be extracted by $r(x)$ and $g(x)$ in a supervised way respectively. VQ-VAE [37] is one of VAEs, which maintains a codebook and learns a series of discrete latent representations. Vector quantisation can learn discrete latent representations and avoid posterior collapse when pairing with a strong autoregressive decoder in the VAE frame. The latent codebook is not statics but learned and the coding range is more controllable. As we can see from Fig.5, the VQ-VAE consists of an encoder, parameterized by ϕ, that can produce the posterior $q_\phi(z|x)$ of the latent factors z, and a decoder, parameterized by θ, that can produce the data likelihood $p_\theta(x|z)$. It also defines a latent embedding space $e \in R^{k\times d}$ where k is the size of the discrete embedding space, and d is the dimensionality of each embedding vector $e_j \in R^d$, $j \in 1, 2, \ldots, k$. The shared embedding space e is then used to compute the discrete latent variables z by nearest neighbor search. The original image x is first passed into an encoder to acquire the continuous coding vector $z_e(x)$. Then, by the nearest neighbor search, $z_e(x)$ is mapped as one of the k vectors in the embedding space, and the corresponding indexes and the encoding vector $z_q(x)$ are obtained. Finally, $z_q(x)$ is passed into the decoder, to reconstruct the original image x and obtain $g(x)$. $r(x)$ is further obtained by element-wise subtraction. We apply a VQ-VAE with the following objective function:

$$L_{vae} = MSE(x, g(x)) + \gamma \left\| \text{sg}\left[z_e(x)\right] - e \right\|_2^2 + \delta \left\| z_e(x) - \text{sg}[e] \right\|_2^2, \tag{1}$$

where MSE refers to the mean squared error, e is the quantized code for the training example x, $g(x)$ is the reconstructed image, $z_e(x)$ is the output of the encoder, sg represents a stop-gradient operation that blocks gradients from flowing into its argument, γ and δ are hyperparameters. The first term is reconstruction loss, aiming at recovering the original image. The second term is codebook loss, to update the codebook variables, making the selected codebook e close to the output of the encoder. The third term is commitment loss, keeping the encoder's output consistent with the selected codebook vector to prevent it from fluctuating too frequently between different codebook vectors. The second term and the third term are to align the vector space of the codebook with the output of the encoder.

However, the reconstruction by VAE alone does not guarantee that $r(x)$ contains only ID-related features and filters out redundant features. The image-level disentanglement is not yet fully implemented. For the re-ID task, the VAE network needs to be guided by the classification network to achieve decoupling capability.

B. Feature-Level Disentanglement

If only ID labels are used to supervise the training of $r(x)$ where $g(x)$ is unconstrained, the ID classification network is enabled to extract basic ID-related features, but there is still feature interference that is not related to ID. For the vehicle re-ID generalization tasks, it is critical to suppress the extraction of the ID-unrelated features. Since re-ID is a cross-view task and camera labels are easily available, we consider the addition of camera labels. When $g(x)$ is supervised by camera labels, the same features between different identities from one camera view are extracted, that is, domain features and ID-unrelated features, which match with the characteristics of $g(x)$ itself. Therefore, it is reasonable to introduce the camera branch to process the ID-unrelated features and form a double-branch network together with the ID branch.

As shown in Fig.5, E_1 and E_2 are two encoders using ResNet-50 [80] pretrained on the ImageNet as backbones. C_1 and C_2 are two classifiers composed of BN (batch normalization) and FC (fully-connected layer). GAP denotes the global average pooling operation. The vehicle ID classification branch is supervised by the vehicle ID labels, enforcing $r(x)$ to contain as many ID-related features as possible. Meanwhile, $g(x)$ is sent to the camera classification branch to process the camera-related information. The two branches are implemented by corresponding cross-entropy loss respectively:

$$L_{id} = -\mathbb{E}_{(x-g(x),y_{id})} \sum_{k=1}^{K_1} \left[k = y_{id}\right] \log\left(C_1\left(f_{id}\right)\right), \tag{2}$$

$$L_{cam} = -\mathbb{E}_{(g(x),y_{cam})} \sum_{k=1}^{K_2} \left[k = y_{cam}\right] \log\left(C_2\left(f_{cam}\right)\right), \tag{3}$$

$$L_{cls} = L_{id} + L_{cam}, \tag{4}$$

where K_1-way class identifier C_1 is trained to correctly predict the vehicle ID labels and K_2-way class identifier C_2 is supervised by the camera ID labels, f_{id} means the feature representation after GAP operation in the ID branch and f_{cam} means the feature representation after GAP operation in the camera branch. The classification loss is the sum of the two.

The two branches are complementary to each other and undertake the processing of different information, which make them have different requirements for the inputs of the two classification networks. With the feedback of the feature-level disentanglement, $r(x)$ and $g(x)$ compete against each other and complete the image-level disentanglement. From the perspective of information bottleneck, the two branches form the information bottleneck constraints for the information flowing on each other, which reduces the burden of the classification task of each branch. We can extract the minimum necessary information required by the convolutional neural network to predict the class of original input, namely features on the ID branch. Since the features extracted from the ID branch are only related to vehicle ID and are not affected by ID-unrelated information, leveraging these features can provide better and more robust performance for vehicle re-ID in the domain generalization tasks.

C. Mutual Information Disentanglement

Although preliminary feature disentanglement has been accomplished intuitively by the two supervised networks mentioned above, some of their features are still entangled. In this section, we introduce a mutual information minimizer to enhance the feature disentanglement. We postulate that excellent feature disentanglement can guide the VAE network to better complete the task of image disentanglement, so as to actively separate ID-related features from the remaining information in the image. Mutual information is a measure of mutual dependence or mutual dependence between random variables based on Shannon entropy.

It can further realize the purpose of feature disentanglement. We minimize the mutual information between f_{id} and f_{cam}. This minimization forces f_{id} to be refined further so as to retain less ID-unrelated information. The mutual information between f_{id} and f_{cam} can be understood as the decrease of the uncertainty in f_{id} given f_{cam}:

$$I\left(\mathcal{D}_{f_{id}}; \mathcal{D}_{f_{cam}}\right) = H(\mathcal{D}_{f_{id}}) - H(\mathcal{D}_{f_{id}}|\mathcal{D}_{f_{cam}}), \tag{5}$$

where H is the Shannon entropy, and $H(\mathcal{D}_{f_{id}}|\mathcal{D}_{f_{cam}})$ is the conditional entropy of f_{cam} given f_{id}. For simplicity, mutual information quantifies the dependence of two random variables f_{id} given f_{cam}. It has the form,

$$I\left(\mathcal{D}_{f_{id}}; \mathcal{D}_{f_{cam}}\right) = \int_{X\times Z} \log \frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \otimes \mathbb{P}_Z} d\mathbb{P}_{XZ}, \tag{6}$$

where $\mathbb{P}_{XZ}$ is the joint probability distribution of $(\mathcal{D}_{f_{id}}; \mathcal{D}_{f_{cam}})$, $\mathbb{P}_X = \int_Z d\mathbb{P}_{XZ}$ and $\mathbb{P}_Z = \int_X d\mathbb{P}_{XZ}$ are marginal distributions of $\mathcal{D}_{f_{id}}$ and $\mathcal{D}_{f_{cam}}$ respectively, $\otimes$ denotes Kronecker product. Based on the above discussion, the mutual information is equivalent to the Kullback-Leibler (KL-) divergence between the joint, $\mathbb{P}_{XZ}$, and the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$:

$$I\left(\mathcal{D}_{f_{id}}, \mathcal{D}_{f_{cam}}\right) = D_{KL}\left(\mathbb{P}_{XZ} || \mathbb{P}_X \otimes \mathbb{P}_Z\right), \tag{7}$$

where D_{KL} is defined as,

$$D_{KL}\left(\mathbb{P}||\mathbb{Q}\right) := E_{\mathbb{P}}[\log \frac{d\mathbb{P}}{d\mathbb{Q}}], \tag{8}$$

whenever $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. The KL divergence can regularize the ID-unrelated features by comparing the distribution ranges of ID-related features. In the intuitive sense of Eq.7: the smaller the value of the objective function, the weaker the correlation between X and Z. We adopt the Mutual Information Neural Estimator (MINE) [72] to estimate the lower bound on n $i.i.d$ (independent and identically distributed) samples by a neural network T_θ.

$$\widehat{I\left(\mathcal{D}_{f_{id}}; \mathcal{D}_{f_{cam}}\right)}_n = \sup_{\theta\in\Theta} E_{\mathbb{P}_{XZ}^{(n)}}\left[T_\theta\right] - \log\left(E_{\mathbb{P}_X^{(n)}\otimes\widehat{\mathbb{P}}_Z^{(n)}}\left[e^{T_\theta}\right]\right). \tag{9}$$

It is worth noting that Eq.9 defines a new way to measure information, and the expression ability of the convolutional neural network can ensure that it can be approximately equal to mutual information with arbitrary accuracy. In practice, for simplicity of calculation, Monte Carlo integration [34]

is leveraged to estimate the expectations of Eq.9. The cost function of the mutual information disentanglement is as follows:

$$L_{MI} = I\left(\mathcal{D}_{f_{id}}, \mathcal{D}_{f_{cam}}\right) = \frac{1}{n}\sum_{i=1}^{n} T\left(f_{id}, f_{cam}, \theta\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{T\left(f_{id}, f'_{cam}, \theta\right)}\right) \tag{10}$$

where (f_{id}, f_{cam}) are sampled from the joint distribution and f'_{cam} is sampled from the marginal distribution.

D. Loss Function Design

By jointly taking all loss functions into account, the total loss function is formulated as:

$$L_{total} = \lambda_1 L_{vae} + \lambda_2 L_{cls} + \lambda_3 L_{MI} \tag{11}$$

where λ_1, λ_2 and λ_3 are hyperparameters to control the contributions of every part, which are 1, 0.2 and 0.01 respectively.

There is an empirical study of the trade-off between VAE reconstruction and classification. The weights setting should avoid the situation of strong recognition ability of classifiers and weak reconstruction, which may cause premature convergence of the whole model and poor generalization performance. Different from other methods to disentangle latent factors in latent space, our method integrates the disentanglement of image-level and feature-level, which is more intuitive and more interpretable. Each part of the model is indispensable and constitutes the whole of JIFD, which is simple but effective.

E. Training and Inference

We train JIFD in two stages. During the first stage, we first fix classifiers C_1, C_2 and train the VAE encoder E and the decoder D, the camera encoder E_1 and the ID encoder E_2 with the corresponding losses L_{vae}, L_{cam} and L_{id}. This ensures that the VAE network is consequently strong enough to reconstruct the original image and provides different image parts $\left(\text{i.e. } g(x) \text{ and } r(x)\right)$ for E_1 and E_2 respectively. Subsequently, the images $r(x)$ and $g(x)$ are used to extract the ID-related and ID-unrelated features f_{id} and f_{cam} through E_2 and E_1, as is shown in Fig.5. Finally, by minimizing the mutual information between disentangled feature pair (f_{id}, f_{cam}), we update mutual information estimator M by Eq.10. The first stage iterates for 20 epochs. In the second stage, we train the entire network and update all parameters in an end-to-end manner. The detailed training process is presented in Algorithm 1.

In the inference stage, each test image x is fed into the VAE network to complete image reconstruction and then $g(x)$ is obtained. By subtracting $g(x)$ from x element-by-element, we can obtain $r(x)$. We use the ID encoder E_2 to extract ID-related features from the image $r(x)$. This feature representation enables JIFD to efficiently perform generalizable vehicle re-ID by computing the cosine distance between a set of probe images and a set of gallery images.

Algorithm 1 Learning Algorithm for JIFD

Input: source labeled dataset $\left\{\left(x^i, y^i_{id}, y^i_{cam}\right) \in S\right\}_{i=1}^{n_s}$; VAE encoder E, VAE decoder D; camera encoder E_1, camera classifier C_1; ID encoder E_2, ID classifier C_2; mutual information estimator M.

Output: well-trained encoder $\hat{E}$, well-trained decoder $\hat{D}$; well-trained ID feature extractor $\hat{E}_2$, well-trained ID classifier $\hat{C}_2$.

while not converged **do**
  Sample mini-batch from $\left\{\left(x^i, y^i_{id}, y^i_{cam}\right) \in S\right\}_{i=1}^{n_s}$;
  if $epoch \leq iter_{pre}$ **then**
    Fix C_1, C_2;
    Variational Autoencoder:
    Update E, D by Eq.1;
    Feature-level Disentanglement:
    Update E_1 by Eq.3;
    Update E_2 by Eq.2;
  else
    Variational Autoencoder:
    Update E, D by Eq.1;
    Feature-level Disentanglement:
    Update E_1 by Eq.3;
    Update E_2 by Eq.2;
    Update C_1 by Eq.3;
    Update C_2 by Eq.2;
  Mutual Information Minimization:
  Calculate mutual information between the disentangled feature pair (f_{id}, f_{cam}) with M;
  Update M by Eq.10;
return $\hat{E} = E$, $\hat{D} = D$, $\hat{E}_2 = E_2$, $\hat{C}_2 = C_2$.

IV. Experiment

Domain generalization aims to transfer the learning task from a single or multiple source domains to unseen target domains. Currently, a leave-one-domain-out protocol is applied to most DG methods [81], [82], [83], [84], [85] to evaluate performance and verify their effectiveness. Since the person datasets have different scales and types (see Table I), this protocol is also widely used in most DG methods for person re-ID. For example, the previous works [63], [86], [87] conduct their experiments on five small re-ID datasets including PRID [88], GRID [89], VIPeR [90], i-LIDs [91] and CUHK01 [92]. Only a few methods [67], [68], [93] are designed to use a single domain to study generalization capabilities.

We try to follow this protocol for the vehicle domain generalization task. However, **(1)** the number of vehicle datasets is less than that of person datasets. There are significant differences in scale between different vehicle datasets, and experiment results in multi-source DG settings cannot reflect the data advantages of small datasets. **(2)** The VehicleX [94] dataset only provides a training set and these synthesized images of VehicleX use samples from the VeRi–776 [95] and VehicleID [51] datasets as references. Hence, we believe that VehicleX→VeRi–776 and VehicleX→VehicleID are not purely DG tasks. **(3)** Our method requires camera labels during

TABLE I

STATISTICS OF PERSON AND VEHICLE DATASETS. †:SYNTHESIZED DATA. '–' DENOTES THAT NO REPORTED RESULT IS AVAILABLE

Person Datasets	IDs	images	cameras	scale
i-LIDS [91]	119	476	2	small
VIPeR [90]	632	1,264	2	small
GRID [89]	251	1,275	6	small
PRID [88]	934	1,134	2	small
CUHK01 [92]	971	3,882	2	small
Market1501 [97]	1,501	32,668	6	medium
CUHK03 [98]	1,467	28,192	2	medium
DukeMTMC [99]	1,812	36,411	8	medium
MSMT17 [100]	4,101	126,411	15	large
Unreal†[101]	3,000	120,000	34	large
RandPerson†[102]	8,000	228,655	19	large
PersonX†[103]	1,266	273,456	6	large
MARS [104]	1,261	1,191,003	6	huge
Person30K [87]	30,000	1,384,940	6,497	huge
Vehicle Datasets	**IDs**	**images**	**cameras**	**scale**
VeRi–776 [95]	776	49360	20	medium
VehicleX†[94]	1,362	192,150	11	large
VehicleID [51]	26,267	221,763	12	large
CityFlow [50]	666	229,680	40	large
VeRI-Wild [49]	40,671	416,314	174	large
VeRI-Wild 2.0 [41]	42,790	825,042	274	huge
Vehicle-1M [52]	55,527	936,051	–	huge

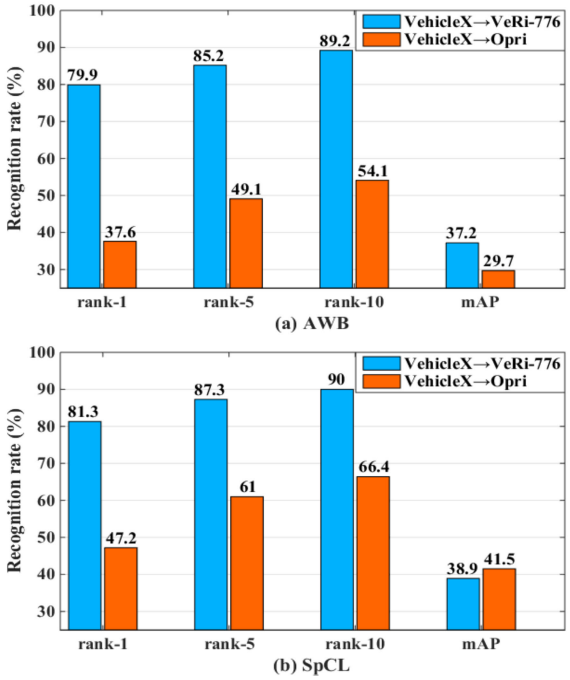

Fig. 6. The experimental results obtained by AWB [96] and SpCL [19] on VehicleX→VeRi–776 and VehicleX→Opri.

Fig. 7. Comparison of perspective changes in different datasets.

model training, but the large-scale vehicle re-ID benchmark of VehicleID and Vehicle-1M cannot provide this information. For the above reasons, we cannot follow the leave-one-domain-out protocol. Since this paper is the first to propose domain generalization for the vehicle re-ID task, we try our best to construct comprehensive experimental results for this task.

To demonstrate the performance gap between UDA and DG methods in terms of recognition accuracy, we compare two commonly used UDA methods: AWB [96] and SPCL [19], as illustrated in Fig. 6. While they achieve comparable performance on VehicleX→VeRi-776, their recognition performance significantly deteriorates on VehicleX→Opridue to the domain bias between the source and target domains. This performance disparity highlights that current UDA models primarily focus on aligning the feature distribution between the source and target domains, without adequately considering the inherent differences and relationships between different domains.

A. Datasets

We conduct experiments on VeRi–776 [95], VehicleX [94] and Opri. The details of these datasets are as follows and summarized in Table II. This paper puts forward three kinds of DG experimental settings, including *Real* (VeRi–776)→*Real* (Opri), *Virtual* (VehicleX)→*Real* (Opri) and *Real*+*Virtual* (VeRi–776+VehicleX)→*Real* (Opri). Fig.7 shows the perspective changes corresponding to different datasets. In addition, we also discuss the recognition performance of the current domain generalization methods under supervised settings.

1) VeRi–776 [95]: The VeRi–776 dataset consists of the training set with 37,778 images of 576 vehicles and the testing set with 11,579 images of 200 vehicles, which are captured by 20 cameras in a real-world traffic scenario. According to our data statistics, there are few images of truck types in this dataset and more images of other types of vehicles. Therefore, there is a long tail problem in VeRi–776.

2) VehicleX [94]: The VehicleX dataset is synthetic data, generated by a publicly available Unity engine [105], [106] and further converted to real-world style via SPGAN [107]. The dataset only provides a training set, which contains 192,150 images of 1,362 vehicles in total.

3) Opri: The Opri dataset is a collection of images of various types of trucks captured in real-world highway scenarios using 4K resolution cameras. This dataset is captured in a 24-hour-a-day uninterrupted manner by multiple surveillance cameras installed in several cities in China. It consists of 130,994 images of 17,835 trucks, which are cleaned by eight researchers over a two-month period. And all the different images of a truck are captured by different cameras. For example, when a truck appears under different cameras, each camera captures only one frontal image of this truck. The dataset is challenging for domain adaptation and domain

TABLE II
STATISTICS OF THE DATASETS USED IN THE PAPER. '–' DENOTES THAT NO REPORTED DATA IS AVAILABLE

Dataset	train IDs	train images	test IDs	query images	gallery images	cameras	total images
VeRi–776	576	37,746	200	1,678	11,579	20	51,003
VehicleX	1,362	192,150	–	–	–	11	192,150
Opri	8,918	67,321	8,917	8,917	54,756	–	130,994

Fig. 8. The Opri dataset is collected from in real-world highway scenarios. The three images from left to right represent vehicle images captured by the camera network under bright light, reflective paint, and license plate occlusion, respectively.

generalization tasks as it contains significant changes in truck appearance. In vehicle tracking and re-ID applications, most of the images captured by the camera network in the highway scene are frontal images of the truck in motion, as shown in Fig.8. The advantage of frontal truck images is that it can be combined with license plate recognition to realize intelligent traffic management automatically. Considering the above reasons, the Opri dataset is proposed. It is distinguished from existing datasets in three aspects, including data type, practical application, and fine-grained vehicle re-ID. For now, it is only open source for 131K images[1] Camera labels are not provided due to relevant policies. Efforts are being made to address this issue, and more data will be released soon.

B. Implementation and Evaluation Protocol

We adopt an ImageNet-pretrained ResNet-50 [80] as the backbone for two encoders. For a fair comparison, the experimental results of other state-of-the-art methods all use ResNet-50 as the backbone. Besides, in order to further verify the effectiveness of JIFD, we combine it with TransMatcher [67] method which has a stronger backbone and the same objective function. All our experiments are implemented on one NVIDIA GeForce RTX 3090Ti using the PyTorch toolbox. All vehicle images are resized to 384×128. The training images are augmented with random horizontal flipping, padding and random cropping. The batch size is 32 and the total number of training epochs is set to 50. The learning rate is set to $1e-5$. AdamW is employed as an optimizer. Mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-k are used to evaluate the experimental performance. VeRi–776→Opri indicates that VeRi–776 is a labeled source domain and Opri is an unseen target domain. VehicleX→Opri and VeRi–776+VehicleX→Opri have similar settings.

[1]The Opri dataset can be downloaded at https://xmu-smartdsp.github.io/dataset/OptimusPrime.html.

TABLE III
DG PERFORMANCE COMPARISON WITH OPRI AS THE TARGET DOMAIN. VERI–776 IS USED AS THE SOURCE DATASET IN OUR EXPERIMENT. THE CUMULATIVE MATCHING RATE(%) AND MEAN AVERAGE PRECISION (MAP)(%) ARE LISTED

UDA Methods		VeRi–776→Opri			
		mAP	rank-1	rank-5	rank-10
SpCL[19]	NIPS'2020	**40.8**	**46.8**	**59.8**	**65.5**
AWB-Pre[96]	TIP'2022	28.3	36.4	47.4	51.9
AWB-Post[96]	TIP'2022	25.2	32.9	43.2	48.1
DG Methods		**VeRi–776→Opri**			
		mAP	rank-1	rank-5	rank-10
Baseline[1]	TMM'2019	18.3	27.8	35.9	39.7
IS-GAN[25]	TPAMI'2019	26.4	37.8	47.1	51.4
MetaBIN[63]	CVPR'2021	22.0	32.5	41.9	45.9
Mixstyle[81]	ICLR'2021	26.4	36.2	47.1	51.6
OSNet-AIN[93]	TPAMI'2021	27.3	37.9	48.3	52.6
QAConv-GS[68]	CVPR'2022	35.2	44.6	56.4	61.0
TransMatcher[67]	NIPS'2021	34.7	42.8	56.0	61.1
JIFD	This work	25.3	35.0	45.5	49.9
TransMatcher[67]+JIFD		**41.2**	**50.2**	**63.1**	**67.4**

TABLE IV
DG PERFORMANCE COMPARISON WITH OPRI AS THE TARGET DOMAIN. VEHICLEX IS USED AS THE SOURCE DATASET IN OUR EXPERIMENT. THE CUMULATIVE MATCHING RATE(%) AND MEAN AVERAGE PRECISION (MAP)(%) ARE LISTED

UDA Methods		VehicleX→Opri			
		mAP	rank-1	rank-5	rank-10
SpCL[19]	NIPS'2020	**41.5**	**47.2**	**61.0**	**66.4**
AWB-Pre[96]	TIP'2022	29.7	37.6	49.1	54.1
AWB-Post[96]	TIP'2022	26.3	34.6	45.4	50.4
DG Methods		**VehicleX→Opri**			
		mAP	rank-1	rank-5	rank-10
Baseline[1]	TMM'2019	19.7	29.1	38.9	42.7
MetaBIN[63]	CVPR'2021	21.7	31.0	41.7	45.8
Mixstyle[81]	ICLR'2021	20.6	29.4	40.4	44.8
OSNet-AIN[93]	TPAMI'2021	14.0	21.5	30.2	34.1
TransMatcher[67]	NIPS'2021	28.9	36.7	50.4	55.7
JIFD	This work	23.0	32.8	43.1	47.4
TransMatcher[67]+JIFD		**30.9**	**37.5**	**52.1**	**57.9**

C. Comparison With Baseline

To conduct a fair and effective comparison with the baseline and state-of-the-art methods, we adopt cross-entropy loss as the basic objective function for all experiments. We train a weak baseline model [1] and a state-of-the-art model [67], respectively, and integrate the JIFD method to evaluate if it leads to further improvements. We compare our method with these two methods on three DG settings, as shown in Table III, IV and V. The experimental results demonstrate

TABLE V

DG PERFORMANCE COMPARISON WITH OPRI AS THE TARGET DOMAIN. VERI–776 AND VEHICLEX ARE USED AS SOURCE DATASETS IN OUR EXPERIMENT. THE CUMULATIVE MATCHING RATE(%) AND MEAN AVERAGE PRECISION (MAP)(%) ARE LISTED

DG Methods		VeRi–776+VehicleX→Opri			
		mAP	rank-1	rank-5	rank-10
Baseline[1]	TMM'2019	19.8	29.3	38.2	42.2
IS-GAN[25]	TPAMI'2019	27.1	37.2	47.4	52.3
Mixstyle[81]	ICLR'2021	25.2	34.5	46.0	50.7
OSNet-AIN[93]	TPAMI'2021	21.6	30.8	41.1	45.3
QAConv-GS[68]	CVPR'2022	**39.6**	**48.9**	**61.2**	**66.0**
TransMatcher[67]	NIPS'2021	28.8	34.8	49.5	55.7
JIFD	This work	25.5	35.4	44.4	48.8
TransMatcher[67]+JIFD		33.1	40.2	54.5	60.3

TABLE VI

PERFORMANCE COMPARISON OF DG METHODS IN SAME-DOMAIN RE-ID SETTINGS. THE CUMULATIVE MATCHING RATE(%) AND MEAN AVERAGE PRECISION (MAP)(%) ARE LISTED. '–' DENOTES THAT NO REPORTED RESULT IS AVAILABLE

DG Methods		VeRi–776		Opri	
		mAP	rank-1	mAP	rank-1
Baseline[1]	TMM'2019	**71.3**	93.3	89.6	90.3
IS-GAN[25]	TPAMI'2019	71.2	92.6	57.8	62.9
MetaBIN[63]	CVPR'2021	46.2	80.1	–	–
Mixstyle[81]	ICLR'2021	55.6	87.1	80.4	82.6
OSNet-AIN[93]	TPAMI'2021	70.8	**95.1**	91.8	92.4
TransMatcher[67]	TPAMI'2021	59.2	86.4	**94.4**	**94.9**
QAConv-GS[68]	CVPR'2022	63.2	86.4	93.9	94.4
JIFD	This work	46.3	82.4	–	–
TransMatcher[67]+JIFD		49.8	80.2	–	–

that our method achieves high recognition accuracy and outperforms the baseline method in terms of rank-1 and mAP. Moreover, when JIFD is applied to TransMatcher [67], it achieves the best recognition accuracy and improves TransMatcher by 7.2% rank-1 recognition rate and 6.5% mAP score on the VeRi–776→Opri setting. The TransMatcher [67]+JIFD method trains TransMatcher [67] on the source domain training set $r(x)$ of JIFD and then tests it on the unknown target dataset. It is even ahead of the UDA method in Table III on VeRi–776→Opri setting. But it has a limited amount of promotion on VehicleX→Opri setting. This is because VeRi–776 dataset is derived from real vehicle images, and these images may contain more redundant information compared to the images in the VehicleX dataset. JIFD can reduce the redundant information in images through image-level disentanglement and provide a more effective training strategy for other methods.

D. Comparison With State-of-the-Arts

First of all, we evaluate our JIFD method in single-source DG settings, which is shown in Table III and IV. Our method is compared with various recent state-of-the-art DG re-ID methods [25], [63], [67], [68], [81], [93]. For DG re-ID task, a limited number of works can report performance on VeRi–776→Opri and VehicleX→Opri, and even those methods are based on domain adaptation [19], [96]. Secondly, we aim to improve the generalization ability of deep re-ID models in unseen domains using data from multiple source domains (see Table V). Finally, we compare our method with current state-of-the-art DG re-ID methods in same-domain re-ID settings (see Table VI).

Experimental results show that the QAConv-GS [68] and TransMatcher [67] methods can achieve the best results in the domain generalization task. Compared with the JIFD method, QAConv-GS formulates a triplet-based ranking learning problem within mini batches provided by the graph sampling sampler. The TransMatcher method combines the ResNet-50 [80] with the Transformer architecture for more robust feature representations. As a result, it has a stronger backbone network than the baseline method. In order to further demonstrate the effectiveness of our method, we combine the component $r(x)$ of JIFD with the TransMatcher method to achieve better recognition performance. Interestingly, it can be found from Table III and V: **(1)** Compared with the performance of VeRi–776→Opri and VeRi–776+VehicleX→Opri, most methods have declined after introducing VehicleX datasets. It can be concluded that simply increasing the amount of data in multiple source domains may not always improve the generalization ability of the model in unseen domains. This is because these methods do not consider the distribution discrepancy and relations between different source domains. **(2)** Only QAConv-GS trained on VeRi–776+VehicleX still has obvious effect improvement, even better than the UDA method (SpCL) trained on Veri–776 in terms of rank-1. The excellent recognition performance of QAConv-GS is due to its effective sampling strategy. The larger the training set, the more effective the training effect of the model is. Therefore, we suggest that the data types and quantities of different source domains need to be treated differently in the process of training the models. But TransMatcher+JIFD on VeRi–776→Opri setting can still be ahead of method QAConv-GS on VeRi–776+VehicleX→Opri setting. This means that the interference caused by redundant information can significantly affect the performance of neural networks.

From Table IV, we have the following observations. **(1)** Our JIFD achieves the best overall performance on *Virtual*→*Real*, outperforming most recently published methods by a clear margin. **(2)** At the same time, in order to compare the performance gap between the current UDA and DG methods, we also conduct comparative experiments on VehicleX→Opri and VeRi→Opri using the state-of-the-arts UDA methods. From the analysis of experimental results, UDA methods are still ahead of DG methods. **(3)** It is clear that OSNet-AIN achieves the worst experimental results on the target datasets due to only using the *Virtual* dataset (VehicleX)–it's even inferior to the baseline method in terms of rank-1 and mAP.

Noteworthy, Table VI shows that none of the current DG methods can improve generalization performance while maintaining recognition performance in supervised settings. For the baseline [1] and OSNet-AIN [93] methods, enhanced performance on a single domain does not guarantee improved generalization, as it may rely on redundant or pseudo-correlated features and compromise the overall

TABLE VII

DG PERFORMANCE COMPARISON WITH OPRI AS THE TARGET DOMAIN. TRAINING ON ONE PART OF JIFD AND TEST ON ANOTHER PART. THE CUMULATIVE MATCHING RATE(%) AND MEAN AVERAGE PRECISION (MAP)(%) ARE LISTED

Baseline Method		VeRi–776→Opri			VehicleX→Opri		
Training	Test	mAP	rank-1	rank-5	mAP	rank-1	rank-5
x	x	18.3	27.8	35.9	19.7	29.1	38.9
x	$g(x)$	12.5	20.4	27.9	15.5	23.3	32.8
x	$r(x)$	11.7	18.8	26.6	10.9	16.6	25.4
$r(x)$	x	23.2	33.2	43.3	20.6	30.6	39.8
$r(x)$	$g(x)$	14.9	23.0	31.1	16.3	24.1	33.6
$r(x)$	$r(x)$	23.6	33.3	43.7	24.7	33.2	44.9
$g(x)$	x	16.3	25.7	32.9	11.5	19.9	26.7
$g(x)$	$g(x)$	15.8	24.2	33.0	16.7	25.5	34.8
$g(x)$	$r(x)$	7.1	11.7	18.0	9.1	15.6	22.9

TABLE VIII

ABLATION STUDIES OF OUR PROPOSED JIFD ON INDIVIDUAL COMPONENTS

Methods	VeRi–776→Opri			
	mAP	rank1	rank5	rank10
$L_{vae} + L_{id}$	23.9	34.1	43.6	47.7
$L_{vae} + L_{id} + L_{cam}$	24.6	34.5	44.6	48.7
$L_{vae} + L_{id} + L_{cam} + L_{MI}$(**Ours**)	**25.3**	**35.0**	**45.5**	**49.9**
Ours+RandomErase[108]	22.6	32.5	42.6	46.7
Methods	VehicleX→Opri			
	mAP	rank1	rank5	rank10
$L_{vae} + L_{id}$	22.1	31.5	42.0	45.9
$L_{vae} + L_{id} + L_{cam}$	22.4	31.7	42.3	46.5
$L_{vae} + L_{id} + L_{cam} + L_{MI}$(**Ours**)	**23.0**	**32.8**	**43.1**	**47.4**
Ours+RandomErase[108]	21.0	30.1	40.3	44.5

generalizability. For high-resolution and fine-grained Opri datasets, IS-GAN may fail to disentangle identity-related and -unrelated features from truck images. This is because it could not effectively prevent from introducing redundant information caused by generated images with poor qualities. For JIFD, although our proposed method discards a large amount of information in the original image to improve generalization performance, it is unable to prevent the loss of domain information and does not yield satisfactory performance in supervised experiments. Therefore, it is crucial not only to enhance the performance of re-ID methods within a single domain but also to ensure their effectiveness in domain generalization.

E. Ablation Studies

To demonstrate the effectiveness and contribution of each component of the JIFD, we perform comprehensive ablation studies on VeRi–776→Opri in Table VIII. We use the combination of the two loss functions L_{vae} and L_{id} as the baseline experiment to compare the influence of other loss functions on the results. We evaluate our complete model against three simplified versions, where we remove one component from each version. From the results, we also observe that the random erasing strategy [108] has a negative impact on the generalization performance of JIFD. Fig.10 and Fig.11show the visualization results after image-level disentanglement through the JIFD method.

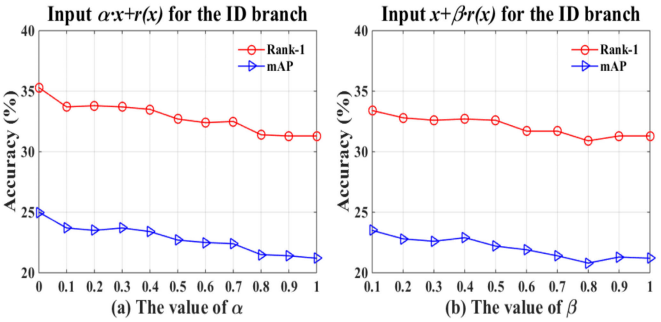

Fig. 9. Recognition performance with different inputs trained on VeRi–776.

Fig. 10. Image-level disentanglement results from JIFD on VeRi–776→Opri.

In Fig.9, we feed different inputs to the ID branch to see whether JIFD could provide further performance improvements. By adjusting the values of α and β to control the ratio between the original image x and $r(x)$ generated by the VAE network, the experimental results show that the neural network will deteriorate the generalization information with the increase of input redundant information. This also verifies our conjecture that the proposed method constitutes an information bottleneck. The information bottleneck here means the neural network uses only a small amount of information in the image to complete the re-ID task. The presence of redundant information in the input image will hinder the generalization performance of the convolutional neural network.

F. Extended Experimental Investigation

To illustrate how $g(x)$ and $r(x)$ affect the performance of the baseline method, our proposed JIFD method divides the original image x into two parts $g(x)$ and $r(x)$, where $r(x)$ only accounts for a small part of the information of the original image x. We combine the baseline method [1] to train on one part of JIFD and test on another part, as shown in Table VII. Experimental results show that baseline can significantly improve the recognition performance on the settings of VeRi–776($r(x)$)→Opri($r(x)$) and VehicleX($r(x)$)→Opri($r(x)$). Out of the three training strategies (x, $g(x)$ and $r(x)$) that we used, it appears that $r(x)$ has the beneficial impact on the domain generalization task. This indicates that our method is somewhat interpretable from an information bottleneck perspective.

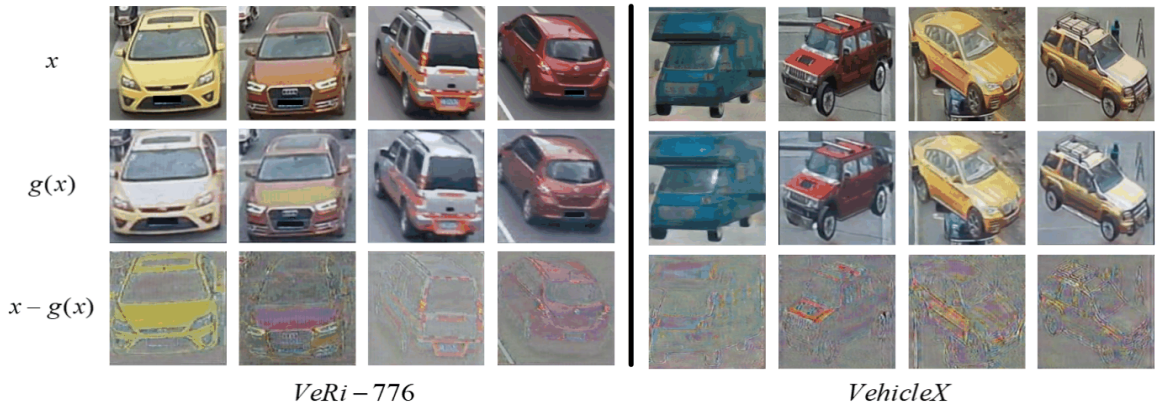

Fig. 11. More visualization results on VeRi–776 and VehicleX. When tested on Opri, VeRi–776 and vehicleX are used as the source dataset in our experiment respectively.

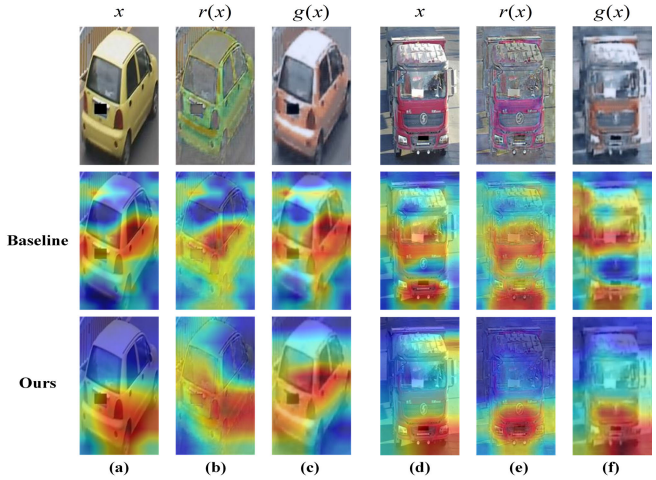

Fig. 12. Illustration of the activated regions for the car and truck. Visualization results of grad-CAM [40] based on baseline [1] and JIFD methods are shown in the second and third lines. The strong activation regions are marked in red. Columns (a), (b) and (c) show the same-domain visualization results of VeRi–776. When tested on opri$(x, r(x), g(x))$, VeRi–776(x) is used as the source dataset, as shown in columns (d), (e) and (f). JIFD shows the responses (i.e. heatmaps) of the ID encoder E_2.

In addition, we investigate the mechanism behind the effectiveness of our approach by analyzing the learned feature representations through visualization. Through these intuitions gained from the visualization regions of the car and truck in Fig.12, our method can still achieve better generalization results through small reconstruction error and reduce the interference of redundant information in the source domain images. It further explains the positive effectiveness of information bottleneck for DG task. From Fig.12 (d), (e) and (f), we can see that compared to the baseline method, our method focuses only on the most discriminative truck regions when applied to the domain generalization task. The baseline method tends to focus on the global features of the truck and does not perform the DG task well. This is because trucks of the same type and color may be highly similar in appearance. The shared but redundant features of these trucks may interfere with discriminant and ID-related features. So how to utilize the limited discriminant area is extremely important for vehicle re-ID task.

G. Limitations and Future Work

We notice that as image-level disentanglement is only evaluated based on reconstruction accuracy, it could not be able to fully retain vehicle identity information while reducing redundancy without other supervised information. So it is unable to prevent the loss of domain information and does not yield satisfactory performance in supervised experiments. Table VI shows the performance comparison and confirms the claim. Moreover, vehicles of the same make and color have very similar appearances, which can also be regarded as redundant information. Such redundant information is difficult to isolate effectively through classification networks. The analysis does not indicate any benefits from classification networks for image-level disentanglement. In future work, we envision creating and exploiting utilizing vehicle attribute information as a supervision way [109] to guide the VAE network to achieve better image-level disentanglement.

V. Conclusion

In this paper, we present a novel idea for tackling the task of domain generalization in vehicle re-ID. Our proposed method performs dual disentanglement at the image-level and feature-level, leveraging the perspective of information bottleneck to achieve the goal of domain generalization. To highlight the significant difference between current re-ID public datasets and specific scenarios in the real world, we collect frontal images of trucks from the highway and present a large-scale and challenging Opri dataset, which we combine with existing public datasets to provide a wealth of domain generalization experimental results. Moreover, our proposed JIFD method can be used as a plug-and-play training strategy and combined with other state-of-the-art methods to improve performance. Extensive comparative experiments demonstrate the effectiveness of our approach. Future research will aim to enhance the effectiveness of domain generalization while maintaining the performance of a single domain.

References

[1] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[2] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.

[3] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11173–11180.

[4] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 346–363.

[5] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2897–2906.

[6] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle instance retrieval," 2019, *arXiv:1909.06023*.

[7] D. Meng et al., "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7103–7112.

[8] Y. Huang et al., "Dual domain multi-task model for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 2991–2999, Apr. 2022.

[9] Z. Dai, G. Wang, W. Yuan, X. Liu, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1142–1160.

[10] H. Yao and C. Xu, "Dual cluster contrastive learning for object re-identification," 2021, *arXiv:2112.04662*.

[11] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3390–3399.

[12] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10981–10990.

[13] J. Yu and H. Oh, "Unsupervised vehicle re-identification via self-supervised metric learning using feature dictionary," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 3806–3813.

[14] S. He et al., "Multi-domain learning and identity mining for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2485–2493.

[15] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, pp. 1–18, Nov. 2018.

[16] Y. Fu et al., "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6112–6121.

[17] A. Zheng, X. Sun, C. Li, and J. Tang, "Viewpoint-aware progressive clustering for unsupervised vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11422–11435, Aug. 2022.

[18] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–15.

[19] Y. Ge, F. Zhu, D. Chen, and R. Zhao, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11309–11321.

[20] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, "Unsupervised vehicle re-identification with progressive adaptation," 2020, *arXiv:2006.11486*.

[21] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3997–4005.

[22] G. Bukchin et al., "Fine-grained angular contrastive learning with coarse labels," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8730–8740.

[23] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.

[24] Q. Meng et al., "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 722–734, Feb. 2021.

[25] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 5298–5309.

[26] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2080–2089.

[27] T. Zhang, K. Liang, R. Du, X. Sun, Z. Ma, and J. Guo, "Learning invariant visual representations for compositional zero-shot learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Tel-Aviv, Israel: Springer, 2022, pp. 339–355.

[28] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.

[29] S. Shahtalebi, J.-C. Gagnon-Audet, T. Laleh, M. Faramarzi, K. Ahuja, and I. Rish, "SAND-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization," 2021, *arXiv:2106.02266*.

[30] H. Zhang, Y.-F. Zhang, W. Liu, A. Weller, B. Schölkopf, and E. P. Xing, "Towards principled disentanglement for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8024–8034.

[31] K. Khezeli, A. Blaas, F. Soboczenski, N. Chia, and J. Kalantari, "On invariance penalties for risk minimization," 2021, *arXiv:2106.09777*.

[32] K. Ahuja et al., "Invariance principle meets information bottleneck for out-of-distribution generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3438–3450.

[33] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*.

[34] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5102–5112.

[35] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 35–51.

[36] K. Yang, T. Zhou, Y. Zhang, X. Tian, and D. Tao, "Class-disentanglement and applications in adversarial detection and defense," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 16051–16063.

[37] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 6306–6315.

[38] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[39] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," 2016, *arXiv:1612.00410*.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[41] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6854–6871, Oct. 2022.

[42] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.

[43] A. Wu, W.-S. Zheng, and J.-H. Lai, "Unsupervised person re-identification by camera-aware similarity consistency learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6922–6931.

[44] X. Chang, Y. Yang, T. Xiang, and T. M. Hospedales, "Disjoint label space transfer learning with common factorised space," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, vol. 33, no. 1, pp. 3288–3295.

[45] I. J. Goodfellow et al., "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[46] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–22.

[47] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3140–3149.

[48] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 869–884.

[49] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3230–3238.

[50] Z. Tang et al., "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8797–8806.

[51] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[52] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 6853–6860.

[53] M. Naphade et al., "The 4th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2665–2674.

[54] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8282–8291.

[55] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 943–952.

[56] J. Yu, J. Kim, M. Kim, and H. Oh, "Camera-tracklet-aware contrastive learning for unsupervised vehicle re-identification," 2021, *arXiv:2109.06401*.

[57] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11926–11935.

[58] P. Wang, C. Ding, W. Tan, M. Gong, K. Jia, and D. Tao, "Uncertainty-aware clustering for unsupervised domain adaptive object re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 2624–2635, 2022.

[59] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3143–3152.

[60] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.

[61] S. Lin, C.-T. Li, and A. C. Kot, "Multi-domain adversarial feature generalization for person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1596–1607, 2021.

[62] M. Tamura and T. Murakami, "Augmented hard example mining for generalizable person re-identification," 2019, *arXiv:1910.05280*.

[63] S. Choi, T. Kim, M. Jeong, H. Park, and C. Kim, "Meta batch-instance normalization for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3425–3435.

[64] S. Yu et al., "Multiple domain experts collaborative learning: Multi-source domain generalization for person re-identification," 2021, *arXiv:2105.12355*.

[65] L. He et al., "Semi-supervised domain generalizable person re-identification," 2021, *arXiv:2108.05045*.

[66] S. Liao and L. Shao, "Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 456–474.

[67] S. Liao and L. Shao, "TransMatcher: Deep image matching through transformers for generalizable person re-identification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 1992–2003.

[68] S. Liao and L. Shao, "Graph sampling based deep metric learning for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7359–7368.

[69] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[70] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Adv. Neural Inf. Process. Syst. (NLPS)*, vol. 29, 2016, pp. 5040–5048.

[71] A. H. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, "A unified feature disentangler for multi-domain image translation and manipulation," in *Proc. Adv. Neural Inf. Process. Syst. (NLPS)*, 2018, pp. 5040–5048.

[72] M. I. Belghazi et al., "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[73] N. Hadad, L. Wolf, and M. Shahar, "A two-step disentanglement method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 772–780.

[74] C. P. Burgess et al., "Understanding disentangling in β-VAE," 2018, *arXiv:1804.03599*.

[75] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[76] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1278–1286.

[77] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4114–4124.

[78] V. Dumoulin et al., "Adversarially learned inference," 2016, *arXiv:1606.00704*.

[79] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," 2016, *arXiv:1609.07093*.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[81] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–15.

[82] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst. (NLPS)*, vol. 31, 2018.

[83] A. T. Nguyen, T. Tran, Y. Gal, and A. G. Baydin, "Domain invariant representation learning with domain density transformations," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 5264–5275.

[84] M.-H. Bui, T. Tran, A. Tran, and D. Phung, "Exploiting domain-specific features to enhance domain generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21189–21201.

[85] G. Zhang, H. Zhao, Y. Yu, and P. Poupart, "Quantifying and improving transferability in domain generalization," in *Proc. Adv. Neural Inf. Process. Syst. (NLPS)*, vol. 34, 2021, pp. 10957–10970.

[86] Y. Dai, X. Li, J. Liu, Z. Tong, and L.-Y. Duan, "Generalizable person re-identification with relevance-aware mixture of experts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16145–16154.

[87] Y. Bai et al., "Person30K: A dual-meta generalization network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2123–2132.

[88] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scand. Conf. Image Anal.*, 2011, pp. 91–102.

[89] C. Change Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1988–1995.

[90] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Marseille, France: Springer, 2008, pp. 262–275.

[91] W. Zheng, S. Gong, and X. Tao, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, vol. 2, London, U.K., 2009, pp. 1–11.

[92] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2012, pp. 31–44.

[93] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning generalisable omni-scale representations for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5056–5069, Mar. 2021.

[94] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 775–791.

[95] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[96] W. Wang, F. Zhao, S. Liao, and L. Shao, "Attentive WaveBlock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 1532–1544, 2022.

[97] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[98] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[99] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[100] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[101] T. Zhang et al., "UnrealPerson: An adaptive pipeline towards costless person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11501–11510.

[102] Y. Wang, S. Liao, and L. Shao, "MARS: A video benchmark for large-scale person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp. 3422–3430.

[103] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.

[104] L. Zheng et al., "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 868–884.

[105] Z. Tang et al., "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 211–220.

[106] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," 2019, *arXiv:1912.08855*.

[107] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[108] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.

[109] H. Li, C. Li, A. Zheng, J. Tang, and B. Luo, "Attribute and state guided structural embedding network for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 5949–5962, 2022.

Zhenyu Kuang received the B.E. degree in electronics and communications engineering from the School of Information Engineering and Automation, Kunming University of Science and Technology, China. He is currently pursuing the Ph.D. degree with the School of Informatics, Xiamen University, China. His research interests include computer vision, person re-identification, and machine learning.

Chuchu He is currently pursuing the master's degree in informatics and communication engineering with the School of Information, Xiamen University. Her research interests include machine learning and computer vision.

Yue Huang (Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, Xiamen University, China, in 2005, and the Ph.D. degree from the Department of Biomedical Engineering, Tsinghua University, Beijing, China, in 2010.

She is currently an Associate Professor with the School of Informatics, Xiamen University. Her main research interests include image processing, sparse signal representation, and machine learning.

Xinghao Ding was born in Hefei, China, in 1977. He received the B.S. and Ph.D. degrees from the Department of Precision Instruments, Hefei University of Technology, Hefei, in 1998 and 2003, respectively.

He was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2009 to 2011. Since 2011, he has been a Professor with the School of Informatics, Xiamen University, Xiamen, China. His main research interests include machine learning, deep learning, computer vision, and signal processing.

Huafeng Li received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from Chongqing University in 2009 and 2012, respectively. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include image processing, computer vision, and information fusion.