Supervised Joint Domain Learning for Vehicle Re-Identification

Chih-Ting Liu* [1], Man-Yu Lee* [1], Chih-Wei Wu* [1],
Bo-Ying Chen[1], Tsai-Shien Chen[1], Yao-Ting Hsu[2], Shao-Yi Chien[1]
[1] NTU IoX Center, National Taiwan University
[2]Department of Electrical Engineering, National Taiwan University
{jackieliu, leemanyu, cwwu, byc, tschen}@media.ee.ntu.edu.tw
{b05901128, sychien}@ntu.edu.tw

Abstract

Vehicle Re-Identification (Re-ID), which aims at matching vehicle identities across different cameras, is a critical technique for traffic analysis in a smart city. It suffers from varying image quality and challenging visual appearance characteristics. A solution for enhancing the feature robustness is by training Convolutional Neural Networks on multiple datasets simultaneously. However, the larger set of training data does not guarantee performance improvement due to misaligned feature distribution between domains. To mitigate the domain gap, we propose a Joint Domain Re-Identification Network (JDRN) to improve the feature by disentangling domain-invariant information and encourage a shared feature space between domains. With our JDRN, we perform favorably against state-of-the-arts methods on the public VeRi-776 dataset and obtain promising results on the 2019 AI City Challenge.

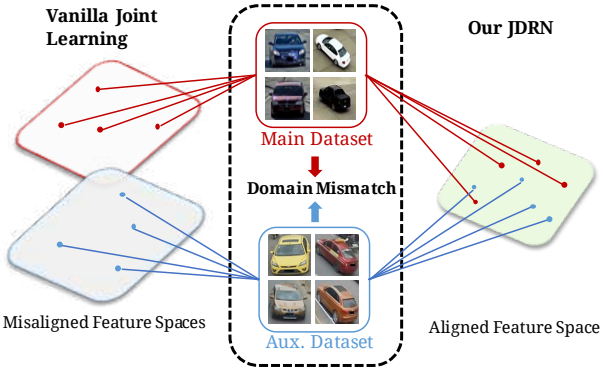

Figure 1: **Joint domain learning.** Owing to the domain mismatch between datasets, vanilla joint learning of Re-ID features on multiple datasets often results in misaligned feature spaces, which in turns degrade the Re-ID performance. With our proposed JDRN, we teach the network to project images to an aligned domain-invariant feature space to enhance the Re-ID ability.

1. Introduction

Vehicle re-identification (Re-ID) is the problem of tracking and identifying moving vehicles across videos captured at multiple locations. It is a crucial technology for analyzing and predicting traffic flow in an envisioned smart city.

Vehicle Re-ID is fundamentally challenging due to a number of reasons. First, vehicle Re-ID data may come in varying image quality due to low camera resolution, motion blur, hostile weather condition, etc. This prevents us from using license plates for identifying vehicles such as the winner system demonstrated in last year's AI City Challenge [21]. Also, privacy can be a concern regarding license plates. Second, the nature of vehicle appearance is difficult to model. Images of the same vehicle can be visually different due to large view point variation, while two different vehicles of the same car model and color may differs only in stickers or scratch marks. Such characteristics makes it hard for classical hand-crafted features [10, 8, 28] to solve vehicle Re-ID. Recent approaches adopt Convolutional Neural Network (CNN) to learn the features and distance metrics in an end-to-end manner [25, 19, 34, 23, 15]. However, we discover that typical CNN features trained on a single dataset is not sufficiently robust due to the limitation of existing datasets. For instance, VeRi-776 [14] and CityFlow-ReID [20] datasets provide abundant vehicle images of diverse view points but fall short in the number of labeled identities. On the other hand, VehicleID [12] dataset contains enormous amount of identities but its images are only taken from front and rear of vehicles. One solution to alleviate this issue is to augment the training set with the combination of other datasets. However, shown by Xiao *et al*. [26], mixing multiple datasets for training may degrade the overall Re-ID performance due to domain mis-

*denotes equal contribution

match between datasets. Such training protocol does not guarantee image features are optimized in an aligned space, where the augmented dataset could benefit the performance, as demonstrated in Figure 1.

To mitigate the impact of domain mismatch, we take inspiration from the literature of domain transfer. Li *et al*. [9] proposed to learn aligned features between different domains to facilitate Re-ID in an unlabeled environment. However, similar techniques have not been explored for supervised setting, which aims to learn more robust features among multiple labeled datasets.

In this paper, we exploit the benefit of learning Re-ID features on multiple datasets simultaneously. The key to alleviate the domain mismatch is to learn a domain-invariant feature for Re-ID since the knowledge of how to discriminate different vehicles should be irrelevant to domains. To this end, we propose a Joint Domain Re-identification Network (JDRN) for solving the vehicle Re-ID problem. We guide the CNN to learn a domain-invariant feature by enforcing the CNN to reconstruct from a domain-shared feature and a domain-specific feature. By separating the information shared by different domains and the one specific to each domain, the CNN is able to disentangle the common knowledge shared across domains to enhance the Re-ID performance. As shown in Fig 1, we can learn a more robust representation on a more diverse set of data. Extensive experiments on VeRi-776 [14] and CityFlow-ReID [20] datasets prove that reinforcing the model with domain-invariant information on multiple datasets can greatly boost the Re-ID performance. Moreover, our proposed JDRN outperforms state-of-the-arts on VeRi-776 and achieves promising results in the 2019 Nvidia AI City Challenge[1]. The contributions of our work can be summarized as follows:

- We learn a robust representation for vehicle Re-ID by jointly learning on multiple datasets to mitigate the limitation of a single dataset.
- We propose a Joint Domain Re-identification Network (JDRN) to further address the domain mismatch problem when learning Re-ID features on multiple domains simultaneously.
- Our JDRN demonstrates promising performance on two vehicle Re-ID datasets and validates the value of joint domain learning.

2. Related Work

Vehicle Re-Identification. Vehicle Re-ID has attracted more and more attention over the years due to the rise of deep learning. Liu *et al*. [14] pioneered in learning visual features for vehicles with CNNs on large-scale dataset, VeRi-776. Typical approaches trained the network using the classification loss (i.e. cross-entropy loss) [14, 15, 23, 24] or the contrastive loss (i.e. triplet loss, siamese loss) [19, 34, 12]. Several works [27, 24, 15, 19] focused on enhancing the feature by teaching the network to model specific vehicle attributes, such as car color and model information. Some others designed multi-branch structure [15] or attention mechanisms [24] to extract discriminative feature from local regions of vehicles. Still others addressed the viewpoints variation of vehicle images by augmenting data with generative model [34] or learning viewpoint-invariant features with the guidance of keypoint information [23]. While the work mentioned above built novel network architectures and loss functions to learn better features for vehicles, the robustness of their feature were still limited by training on single dataset. As suggested by Xiao *et al*. [26], we explore the extra robustness provided by jointly training on multiple datasets in this work.

Domain Invariant Feature Learning. Although learning on multiple datasets could potentially benefit feature learning, it is sometimes harder to learn a better universal feature due to domain gap between datasets. Such problem has been exhaustively studied in the literature of domain transfer. Shown by several works [9, 17, 22], learning a domain-invariant feature could generalize the feature for unlabeled domains. DICA [17] proposed a kernel-based optimization to learn an invariant transformation by minimizing the dissimilarity across domains. Tzeng *et al*. [22] introduced a domain confusion loss to enforce network to minimize the Maximum Mean Discrepancy between domains. Li *et al*. [9] designed the domain separation network and Lin *et al*. [11] constructed the MMFA network to leverage information from different domains for person Re-ID. All works mentioned above focused on learning domain-invariant features for transferring knowledge from labeled domain to unlabeled domain. In contrast, we demonstrate with the proposed JDRN that such concept is also critical for joint learning a stronger feature on multiple labeled datasets. Particularly, while Li *et al*. [9] aim to transfer knowledge from labeled domains to unlabled domains, our work differs from theirs by trying to harvest a general feature for multiple labeled domains and reinforcing the alignment between feature spaces of different domains by introducing an additional loss.

3. Proposed Methods

Given the image-label pairs $\{I_i^m, y_i^m\}_{i=1}^{N_m}$ in our main target dataset and another set of pairs $\{I_i^a, y_i^a\}_{i=1}^{N_a}$ in an auxiliary dataset, where N_m and N_a denote the total number of images of the two dataset respectively, the goal of our model is to learn the discriminative ability to perform vehicle Re-ID on the main dataset with the aid of the labeled auxiliary

[1] https://www.aicitychallenge.org/

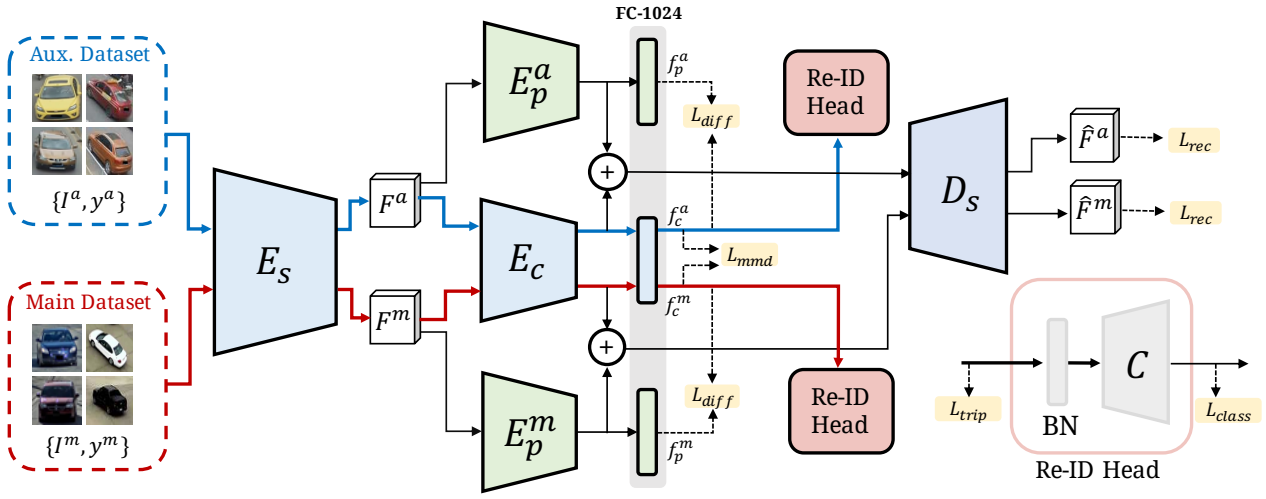

Figure 2: **Overview of our Joint Domain Re-identification Network.** The shared encoder E_S first extract feature maps F^m and F^a from main and auxiliary datasets, which are fed into the content encoder E_c, private encoder E_p^m and E_p^a to learn domain-invariant and domain-specific features respectively, while a 1024-dim fully-connected layers (FC-1024) is applied after each encoders to generate the feature vectors f_c^a, f_c^m, f_p^m and f_p^a. We employed a shared decoder D_s to reconstruct F^m and F^a from the concatenation of feature maps after E_c and E_p^m/E_p^a. Finally, two Re-ID Heads including a batch normalization layer (BN) and a classifier C are configured after f_c^a and f_c^m to perform Re-ID learning from both domains.

dataset.

In Sec. 3.1, we present the pipeline of learning Re-ID features with metric learning and discriminative learning. In Sec. 3.2, we describe our Joint Domain Re-identification Network (JDRN) to deal with the problem of domain mismatch between datasets. This architecture facilitates features with domain-invariant characteristics by explicitly modeling domain-specific information. The final inference stage for vehicle Re-ID is demonstrated in Sec. 3.3.

3.1. Re-Identification Feature Learning

In the vehicle Re-ID task, we aim to learn a Encoder E with a CNN model to extract robust vehicle features. For the CNN model, we adopt the DenseNet-161 [6] as the backbone network. We apply an additional fully connected layer after the global average pooling layer to obtain a 1024-dim feature vector $f \in \mathbb{R}^{1024}$. Inspired by the recent advancement in person Re-ID [31], we employ both metric learning [18] and discriminative learning [30] to facilitate the model with Re-ID ability on top of f, as illustrated as Re-ID head in Fig. 2. For metric learning, it is used to ensure the separation of feature distance between positive and negative pairs. Instead of using the batch-hard triplet loss [5] to train the network, we choose to optimize the soft-margin weighted triplet loss recently proposed by Ristani *et al*. [18]. The formulation is as follows:

$$\mathcal{L}_{trip} = \sum_{a,p,n} F(w_p d(E(I_a), E(I_p)) - w_n d(E(I_a), E(I_n))). \tag{1}$$

Note that I_a indicates an anchor sample, with its associated positive and negative images I_p and I_n , respectively. The function $d(x, y)$ calculates the Euclidean distance between x and y, where $E(I)$ is the feature vector f of image I. We use a soft-plus function $F(x) = \log(1 + e^x)$ as a penalty function instead of a typical hinge function. The w_p and w_n are adaptive weights calculated according to the normalized feature distances, which are as follows:

$$w_p = \frac{e^{d(f_a, f_p)}}{\sum\limits_{a,p} e^{d(f_a, f_p)}}, w_n = \frac{e^{-d(f_a, f_n)}}{\sum\limits_{a,n} e^{-d(f_a, f_p)}} \tag{2}$$

Instead of using binary weights to consider the most difficult positive and negative pairs in the batch-hard triplet loss [5], the adaptive weights re-weights the importance of different feature pairs depending on its difficulty. To be specific, positive pairs with larger feature distance should be considered as harder examples, which could be more beneficial for training the network. On the other hand, positive pairs with smaller feature distance should have smaller weighting since it is a much easier sample for the network. And vice versa for negative pairs. Applying the adaptive weights can alleviate the impact of noisy outliers [18], which sometimes serve as the hard samples in batch-hard triplet loss training. With the weighted triplet loss, the model can effectively learn to gather positive pairs and separate negative pairs in the embedding space.

As for discriminative learning, we optimize the cross-

entropy loss, which is formulated as follows:

$$\mathcal{L}_{class} = -\sum_{i=1}^{n} y_i \cdot log\hat{y}_i \tag{3}$$

Note that y_i is the one-hot encoded label vector and n denotes the batch size. We obtain $\hat{y}_i$ by applying a Batch Normalization (BN) layer [7] on f followed by a classifier C which is a fully-connected layer to perform identity classification [29, 30]. Notice that we empirically separate triplet loss and cross-entropy loss with a BN layer, which coincide with the observation in some work on person Re-ID [16]. It is believed that embedding space without normalization is more suitable for distance metric learning while normalized feature space forces the model to classify samples on a more constraint angular space with cross-entropy loss [5, 13, 2]. We apply the triplet loss and cross-entropy loss on both main and auxiliary datasets.

3.2. Joint Domain Feature Learning

While mixing the target dataset and auxiliary dataset for training provide a more diverse set of training samples, jointly optimizing the backbone network on multiple datasets may lead to inferior Re-ID performance. It is possible that the domain bias between datasets hinders the model to learn an universal representation for vehicle Re-ID regardless of domain. To mitigate the domain gap between target and auxiliary dataset, we design the Joint Domain Re-Identification Network (JDRN) to learn domain-invariant Re-ID features in an aligned embedding space between the two domains. The idea is to encourage domain-invariant feature learning by modeling domain-specific and domain-shared attributes separately. As illustrated in Fig. 2, we first divide the backbone network E into two components E_s and E_c, and add two private encoders E_p^m/E_p^a with the same architecture of E_c but without sharing the parameters. We also add a fully-connected layers after E_p^m/E_p^a to generate the 1024-dim feature vectors. The two private modules E_p^m/E_p^a of the main and auxiliary datasets and the shared content encoder E_c are used to decompose the mid-level feature maps F^m/F^a, which are the outputs of the sharing encoder E_s, into domain-specific feature vectors f_p^m/f_p^a and the desired domain-invariant features f_c^m/f_c^a.

To ensure that the content and private encoders extract different attributes of the input feature maps, difference loss $\mathcal{L}_{diff}$ is applied to encourage the orthogonality between f_c and f_p:

$$\mathcal{L}_{diff} = \parallel \mathbf{H}_c^{m\top}\mathbf{H}_p^m \parallel_F^2 + \parallel \mathbf{H}_c^{a\top}\mathbf{H}_p^a \parallel_F^2 \tag{4}$$

$\mathbf{H}_c^m$ and $\mathbf{H}_c^a$ are matrices with dimension $\mathbb{R}^{n\times 1024}$, whose rows are the content features f_c^m and f_c^a respectively. $\mathbf{H}_p^m$ and $\mathbf{H}_p^a$ are obtained in a similar manner. Note that $\parallel \cdot \parallel_F^2$ denotes the square Frobenius norm.

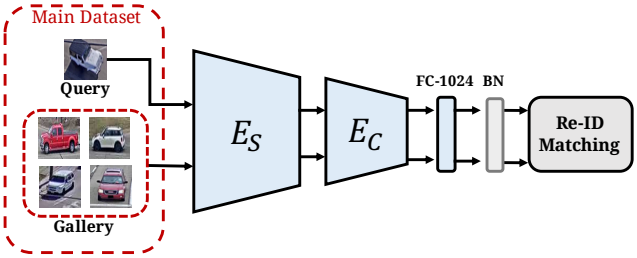

Figure 3: **Re-ID inference stage.** In testing phase, each image in query and gallery set is passed forward to the Es, Ec, the FC-1024 layer and the BN layer in Re-ID Head to obtain the final 1024-dim feature for Re-ID.

While applying $\mathcal{L}_{diff}$ aims at disentangling the feature, there are no constraints on the private encoder to extract information related to the original mid-level feature maps F. The reconstruction loss is a desirable solution. We employ a share decoder D_s to reconstruct the feature maps F^m and F^a for both domains. The input of D_s is the concatenation of the feature maps after E_c and E_p, which are the features before the global average pooling and fully-connected layer. The reconstruction loss $\mathcal{L}_{rec}$ is formulated as:

$$\mathcal{L}_{rec} = \sum_{i=1}^{n_m} \parallel F_i^m - \hat{F}_i^m \parallel_2^2 + \sum_{i=1}^{n_a} \parallel F_i^a - \hat{F}_i^a \parallel_2^2, \tag{5}$$

where $\hat{F}^m/\hat{F}^a$ are the outputs of D_s and n_m/n_a are the batch size of images from main and auxiliary datasets.

Besides sharing the content encode E_c for main and auxiliary domain to encourage the learning of aligned feature distribution, we also adopt the Maximum Mean Discrepancy (MMD) measure [4] inspired by [1] to minimize the feature distribution difference between two domains. We can formulate the loss $\mathcal{L}_{mmd}$ as follows:

$$\mathcal{L}_{mmd} = \parallel \frac{1}{n_m}\sum_{i=1}^{n_m}\phi(f_{c,i}^m) - \frac{1}{n_a}\sum_{j=1}^{n_a}\phi(f_{c,j}^a) \parallel_{\mathcal{H}}^2, \tag{6}$$

where ϕ is the mapping operation which projects the distribution into a reproducing kernel Hilbert space $\mathcal{H}$ [3].

In sum, for training our Joint Domain Re-identification Network (JDRN), the total objective can be written as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{trip} + \mathcal{L}_{class} + \alpha\cdot\mathcal{L}_{diff} + \beta\cdot\mathcal{L}_{rec} + \gamma\cdot\mathcal{L}_{mmd}, \tag{7}$$

where α, β and γ are the hyper-parameters for adjusting the influence of different loss functions. We optimize the model with the $\mathcal{L}_{total}$ in an end-to-end manner on main and auxiliary dataset simultaneously.

3.3. Inference Stage for Vehicle Re-Identification

Once the model is trained, the combination of encoders $\{E_S, E_C\}$ has the ability to extract more general vehicle

features thanks to the aid of auxiliary dataset. For the inference of vehicle Re-ID features on main dataset, as shown in Fig. 3, we feed the query and gallery images into our $\{E_S, E_C\}$ encoders and the Re-ID head to obtain the final feature vectors for Re-ID matching. The matching is done with comparing cosine similarities between features.

4. Experiments

4.1. Datasets and Evaluation Protocol

We evaluate our approach on the CityFlow-ReID [20] and VeRi-766 [14] dataset. CityFlow-ReID is a subset of images sampled from the CityFlow dataset [20], which also serves as the competition dataset for Track 2 of 2019 AI City Challenge. It consists of 36,935 images of 333 identities in the training set and 18,290 images of another 333 identities in the testing set. It has the largest scale of spatial coverage and number of cameras among all the existing vehicle datasets. Each vehicle is captured by 4.55 camera on average. The dataset exhibits some challenging scenarios, such as radial distortion and inconsistent resolution of images owing to the use of various traffic cameras. Veri-766 contains 766 vehicles with each vehicle captured by 2-18 cameras. The whole dataset is split into 37,778 images of 576 identities for training and 11,579 images of 200 identities for testing. In the following sections, we use "CF-ReID" and "VeRi" as the abbreviation of CityFlow-ReID and VeRi-776 dataset.

In our experiments, we adopt the standard train/test split and report the Mean Average Precision (mAP) [28] to evaluate the Re-ID performance. Note that on CF-ReID, the mAP results are reported with rank list of size 100 on 50% of the testing set displayed by the AI City Challenge Evaluation System. For VeRi, the mAP results are evaluated with full size rank list, which are of the size of gallery images.

4.2. Implementation Detail

Following Section 3, we use DenseNet-161 as the backbone for our JDRN and we split it into two parts. The last block $DenseBlock4$ is used and duplicated as E_c, E_p^m and E_p^a while the rest of the blocks are used as share encoder E_s. Decoder D_s is implemented with fully convolution networks which contains three layers. The input images are resized to 224×224 and augmented with random horizontal flip to form a batch of size 32, consisting 8 randomly sampled identities, each with 4 sampled images. We choose SGD optimizer with the initial learning rate starting from 0.005 and decay it by 10 times every 15000 iterations to train network for 40000 iterations. For the $\mathcal{L}_{total}$ in Equation 7, we empirically set the hyper-parameters α, β and γ to 0.1.

Table 1: **Comparisons of different baselines with two training set configuration.** Self: Trained on individual dataset. Joint: Jointly trained on both datasets.

Method	Training set	CF-ReID mAP	VeRi mAP
Baseline (E)	Self	36.26	59.94
Baseline (E)	Joint	35.81	56.98
Baseline w/ $\mathcal{L}_{mmd}$	Joint	32.91	56.97
Our JDRN	Joint	**44.14**	**69.08**

4.3. Ablation Studies

We conduct experiments on both CF-ReID and VeRi to analyze the effectiveness of our JDRN design. Table 1 shows the results. First, we train a baseline model by optimizing the model E introduced in Sec. 3.1 with $\mathcal{L}_{trip}$ and $\mathcal{L}_{class}$ on the main dataset alone. $\mathcal{L}_{trip}$ and $\mathcal{L}_{class}$. As presented in the first row, the baseline performs worse than our final JDRN since the training data exhibits much less diversity when using only one dataset for learning. Next, we perform the vanilla joint training on two datasts with or without the $\mathcal{L}_{mmd}$, which is used to align the feature distributions of two domains. It can be seen that without carefully designed network, E cannot learn more discriminative features despite using more training data. We attribute this observation to the lack of ability to distill domain-invariant information. Therefore, the model is greatly influenced by the domain bias when performing vehicle Re-ID, resulting in inferior performance. Note that if we add $\mathcal{L}_{mmd}$ on the baseline encoder E, the performance are even worse than those without it, revealing that forcing the model to map features into the similar embedding space without separating the domain-specific characteristics will degrade the original capability to perform Re-ID. Last, we improve the performance on both datasets significantly by taking advantage of another auxiliary dataset with our JDRN design. The superior performance is not possible without designing private encoders E_p^m/E_p^a and $\mathcal{L}_{diff}, \mathcal{L}_{rec}$ to disentangle domain-invariant information for learning a robust Re-ID feature across different domains.

4.4. Comparison with State-of-the-arts

We compare our method with existing state-of-the-art methods on VeRi dataset in Table 2. Our JDRN outperforms all state-of-the-art methods thanks to the addition of auxiliary dataset, CF-ReID, and the meticulously designed joint domain feature learning scheme. It is also worth noting that our method even performs better than [19] and [34] that explores spatio-temporal information in addition to the visual features. Furthermore, we boost our performance even more by applying common re-ranking techniques [32]. The experiments on VeRi show that learning domain-invariant

Table 2: **Comparison with state-of-the-arts Re-ID methods on VeRi-776 dataset.** Note that [19] and [34] utilize the spatial-temporal information for re-identification.

Method	Source	VeRi
		mAP
XVGAN [33]	BMVC17	24.65
FACT+Plate-SNN [14]	ECCV16	25.88
OIFE [23]	ICCV17	51.42
RNN-HA [24]	ACCV18	56.80
S-CNN+Path-LSTM [19]	ICCV17	58.27
VAMI+STR [34]	CVPR18	61.32
Our JDRN	-	69.08
Our JDRN + re-ranking	-	**73.10**

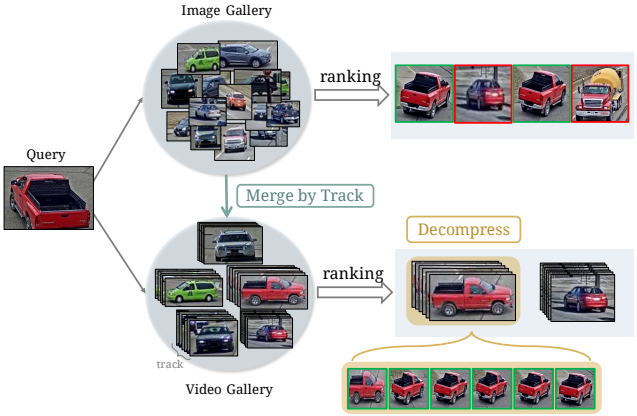

Figure 4: **Illustration of "Merge and Decompress" algorithm.** The upper part shows a typical framework of image-based Re-ID. The lower part illustrates the process of our video-based inference scheme, including merging image features in the gallery set into video features according to track information and decompressing the video-based ranking results into image rank list.

features on multiple datasets with our proposed JDRN can reinforce the model with more discriminative ability for vehicle Re-ID.

4.5. Submission on the 2019 AI City Challenge

We also submit our proposed method to the 2019 AI City Challenge, which holds competition for vehicle Re-ID with the CF-ReID dataset. As a supplement to our JDRN model, we employed some additional techniques for the final submission. First, we design a "Merge and Decompress" algorithm for the video-based inference scheme according to the additional tracking information of each image. As illustrated in Fig. 4, we first merge image features of the same track in the gallery by average pooling. That is to say, a video track is represented by only one summarized feature vector. Then, we perform typical Re-ID scheme to rank the videos features in the gallery according to the query image feature. After obtaining the ranking list, we then decompress the images belonging to each video tracks for aggregating the final rank list. This video-based inference scheme is designed to refine the rank list with the help of tracking information. For example, if a query image ranks a pair of positive gallery images of the same track differently, our video-based inference scheme would refine the ranking of the two images by ranking both equally with the summarized video feature. Second, we adopt the k-reciprocal re-ranking method in [32] to re-rank our Re-ID results. Finally, we ensemble several results from our previous submissions with a voting mechanism to achieve a more reliable score.

Different from Table 1, the score of our final submission to 2019 AI City Challenge Track2 is calculated with 100% testing set. With our JDRN and the tricks mentioned above, we achieve $\mathbf{49.98}\%$ in mAP at the rank list size of 100 (rank100-mAP), which is significantly better than the baseline score 26.3% presented by CityFlow-ReID [20]. We rank 18^{th} among all 84 participated teams.

4.6. Qualitative Analysis on CF-ReID

To better understand the advantage of our proposed JDRN, we compare the qualitative results on CF-ReID between the JDRN and the baseline model trained only on CF-ReID in Fig. 5. We observe that JDRN is able to make more accurate rank-1 prediction by discriminating finer appearance detail, such as the shape of front/tail lights, the color of bumpers, stickers on the side of vehicles. We attribute the improvement to the introduction of auxiliary datasets, which provides more data samples to learn more discriminative vehicle features. Note that the improvement is only made possible by our meticulously designed JDRN as demonstrated in previous experiments.

5. Conclusion

In this paper, we introduce a Joint Domain Re-ID Network (JDRN) to utilize multiple labeled datasets for learning a robust feature representation for vehicle Re-ID. We discover that simultaneously training different datasets often results in undesirable performance due to domain mismatch between datasets. To alleviate the impact of it, we design a novel network architecture to disentangle domain-invariant and domain-specific information. The domain-invariant feature brings the knowledge of identifying vehicles in different domains together. As a result, a more universal yet discriminative feature can be learned on multiple datasets using our JDRN. Extensive experiments on VeRi and the recently introduced CityFlow-ReID daetasets prove the effectiveness of our JDRN and performs favorably against a number of state-of-the-arts.

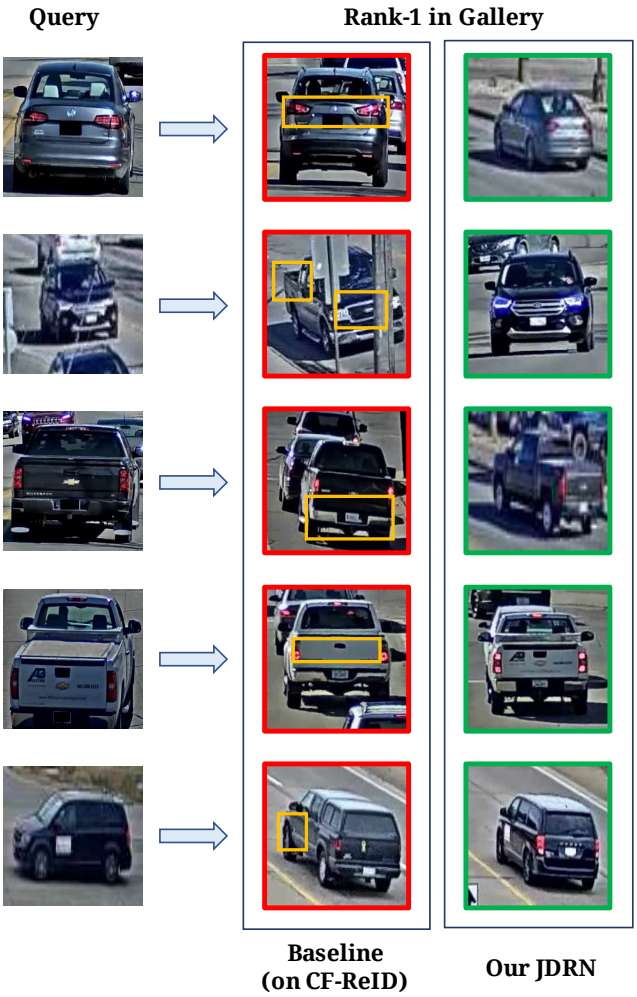

Figure 5: **Visualization of Re-ID results on CF-ReID.** We visualize the rank-1 candidate suggested by the baseline model trained only on CF-ReID (first row in Table 1) and our final JDRN model. Images with green outlines represent correct matches, and those with red outlines are not. The yellow boxes highlight the subtle detail that differentiate from correct matches.

Acknowledgment

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 108-2633-E-002-001), National Taiwan University(NTU-108L104039), Intel Corporation, Delta Electronics and Compal Electronics.

References

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[3] Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*, 2008.

[4] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295. IEEE, 2012.

[9] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–178, 2018.

[10] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[11] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*, 2018.

[12] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[14] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018.

[15] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bags of tricks and a strong baseline for deep person re-identification. *arXiv preprint arXiv:1903.07071*, 2019.

[17] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[18] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046, 2018.

[19] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1900–1909, 2017.

[20] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. *arXiv preprint arXiv:1903.09254*, 2019.

[21] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 108–115, 2018.

[22] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[23] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2017.

[24] Xiu-Shen Wei, Chen-Lin Zhang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Coarse-to-fine: A rnn-based hierarchical attention model for vehicle re-identification. *arXiv preprint arXiv:1812.04239*, 2018.

[25] Chih-Wei Wu, Chih-Ting Liu, Cheng-En Chiang, Wei-Chih Tu, and Shao-Yi Chien. Vehicle re-identification with the space-time prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 121–128, 2018.

[26] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[27] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 562–570, 2017.

[28] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[29] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[30] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017.

[31] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.

[32] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[33] Y Zhou and L Shao. Cross-view gan based vehicle generation for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[34] Y Zhou, L Shao, and A Dhabi. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn*, volume 2, 2018.