Multi-Task Mutual Learning for Vehicle Re-Identification

Georgia Rajamanoharan
Vision Semantics Ltd
georgia@visionsemantics.com

Aytaç Kanacı Minxian Li Shaogang Gong
Queen Mary University of London
{a.kanaci,m.li,s.gong}@qmul.ac.uk

Abstract

Vehicle re-identification (Re-ID) aims to search a specific vehicle instance across non-overlapping camera views. The main challenge of vehicle Re-ID is that the visual appearance of vehicles may drastically changes according to diverse viewpoints and illumination. Most existing vehicle Re-ID models cannot make full use of various complementary vehicle information, e.g. vehicle type and orientation. In this paper, we propose a novel Multi-Task Mutual Learning *(MTML) deep model to learn discriminative features simultaneously from multiple branches. Specifically, we design a consensus learning loss function by fusing features from the final convolutional feature maps from all branches. Extensive comparative evaluations demonstrate the effectiveness of our proposed MTML method in comparison to the state-of-the-art vehicle Re-ID techniques on a large-scale benchmark dataset, VeRi-776. We also yield competitive performance on the NVIDIA 2019 AI City Challenge Track 2.*

1. Introduction

With the development of autonomous driving and smart city applications, the need to accurately analyze vehicles on urban streets via multiple computer vision tasks such as detection, classification and pose estimation, as well as re-identification, is ever-increasing. Specially, vehicle re-identification has attracted increasing attention in the research community [15, 16, 16, 26, 26, 35], as it can play an important role in intelligent transportation systems and public safety.

Vehicle re-identification (Re-ID) aims to search a specific vehicle instance across non-overlapping camera views. Due to the fact that the license plate is often not visible in a number of view angles (which are generally uncontrolled), vehicle Re-ID by visual appearance alone is of great practical value in real-world applications such as smart cities. This task is similar to a more popular task: person re-identification [7, 5, 13, 28, 12, 27, 22, 3, 18, 31, 19, 23], but with more challenges: (1) Unlike person Re-ID, the pose/orientation of vehicles results in occlusion and drastic visual geometry changes, since the vehicle is a kind of rigid body. This means that is difficult to infer the same identity from any given pose/orientation of a vehicle. (2) Even in the same orientation, vehicles of different identities may look very similar due to the being of the same, or similar, vehicle model. This requires vehicle Re-ID models to have a more discriminative fine-grained recognition ability.

Most previously proposed vehicle Re-ID methods [15, 16, 20, 35] focus on using a single branch to learn an embedded feature representation for vehicle instance re-identification from the original information (e.g. the original whole vehicle image). Due to the previously mentioned challenges for vehicle re-identification, this single branch structure can not take advantage of the diversity of vehicles. Moreover, most existing works [15, 16] train their vehicle Re-ID deep learning model using a single supervisory signal (e.g. vehicle ID). However, we argue that vehicle ID label *alone* can not differentiate the differences between vehicles due to the issues we raised above. But additionally we can make use of the fact that the orientation of a vehicle alters its view in a predictable manner. As a result, we suggest that imposing multiple and different supervisory signals simultaneously (e.g. the vehicle ID *and* vehicle orientation) allows the model to learn this variation in a more well-defined manner, and is thus more effective for learning the fine-grained discriminative features necessary for vehicle Re-ID. As orientation labels were provided by Wang *et al.* [26] for the VeRi dataset [16], is is possible to make use of these multiple signals for this purpose. Examples of the orientation labels provided by this paper can be seen in Figure 1.

In this work, we propose a novel *Multi-Task Mutual Learning* (MTML) based network architecture, that aims to simultaneously learn a number of recognition tasks from different supervisory signals, plus a consensus loss function, to build an improved representation for the purpose of vehicle re-identification.

We make two contributions in this work as follows: (1) We formulate a novel *Multi-Task Mutual Learning* (MTML) deep learning model by building each individual branch for

Index	Orientation	Colour
0	front	red
1	rear	-
2	left	-
3	left front	cyan
4	left rear	yellow
5	right	-
6	right front	green
7	right rear	black

Figure 1: Examples from the VeRi-776 dataset with the orientation labels provided in [26] (best viewed in colour).

a different recognition task relevant to vehicle Re-ID and by taking into consideration four different tasks: vehicle ID, multi-scale, grayscale, and orientation. Our model aims to discover and capture concurrently the complementary discriminative information. (2) We introduce a mutual learning mechanism for improving multi-task learning robustness. Our model benefits from multiple supervisory signals in order to enhance model learning of more discrimative features for vehicle Re-ID. Extensive comparative evaluations demonstrate the effectiveness of the proposed MTML method in comparison to the state-of-the-art vehicle Re-ID techniques on the a large-scale benchmark VeRi-776 [16]. We also yield competitive performance on the CityFlow [25] benchmark at the NVIDIA 2019 AI City Challenge.

2. Related Work

Vehicle Model Classification One closely related problem to re-identification is vehicle model classification [14, 30, 21, 10]. The two problems are usually studied independently. For example, Yang et al. [30] propose a part attributes driven vehicle model recognition. They also contribute a large comprehensive car dataset named "CompCars" with model class labels but without vehicle identity labels. More recently, Hu et al. [10] formulate a deep CNN framework capable of selecting spatial salient vehicle parts in order to learn more discriminative model representations without explicit parts annotations.

Vehicle Re-Identification. A number of deep learning techniques have been exploited for the purpose vehicle Re-ID. For instance, Liu *et al.* [16] explored a deep neural network to estimate the visual similarities between vehicle images. Liu *et al.* [15] also designed a Coupled Clusters Loss (CCL) to boost a multi-branch CNN model for vehicle Re-ID. All these methods utilize the global appearance features of vehicle images and ignore local discriminative regions. To explore local information motivated by the idea of landmark alignment [32] in both face recognition [24] and human body pose estimation [17], Wang *et al.* [26] considered 20 vehicle keypoints for learning and aligning local regions of a vehicle for Re-ID. Clearly, this approach comes with extra cost of exhaustively labelling these keypoints in a large number of vehicle images, and the implicit assumption of having sufficient image resolution/details for extracting these keypoints.

Additionally, space-time contextual knowledge has also been exploited for vehicle Re-ID subject to structured scenes [16, 20]. Liu *et al.* [16] proposed a spatio-temporal affinity approach for quantifying every pair of images. Shen *et al.* [20] further incorporated spatio-temporal path information of vehicles. Whilst this method improves the Re-ID performance on the VeRi-776 dataset, it may not generalize to complex scene structures when the number of visual spatio-temporal path proposals is very large with only weak contextual knowledge available to facilitate model decision.

Multi-Task Learning. Multi-task learning (MTL) is a machine learning strategy that learns several related tasks simultaneously for their mutual benefits [1]. A good MTL survey with focus on neural networks is provided in [2]. Deep CNNs are well suited for performing MTL as they are inherently designed to learn joint feature representations subject to multiple label objectives concurrently in multi-branch architectures. Joint learning of multiple related tasks has been proven to be effective in solving computer vision problems [6, 33]. Critically, our method is uniquely de-

signed to explore the potential of MTL in combining multiple diversities (e.g. scale and color) of the vehicle image and being supervised by multiple kinds of manual labels (e.g. ID and orientation) with each of them being associated with an individual branch of a single model.

3. Multi-modal Vehicle Re-identification

In order to perform Re-ID of previously unseen query vehicles, the aim of our model is to learn a feature embedding that allows for accurate retrivals based on distance (*e.g.* L1) from the query image representation. In order to perform this task, we utilise training data containing a number of different labels: identity class labels as well as vehicle orientation class labels. We assume two sets of training examples $\mathcal{I}_1 = \{\mathbf{I_i}\}_{i=1}^{N}$ and $\mathcal{I}_2 = \{\mathbf{I_i}\}_{i=1}^{M}$, containing N and M training images respectively. Both training sets contain the associated identity class labels $\mathcal{Y}_1 = \{y_i\}_{i=1}^{N}$ and $\mathcal{Y}_2 = \{y_i\}_{i=1}^{M}$, where $y_i \in [1, ..., N_{id}]$ for N_{id} distinct vehicle idenities spanning the two training sets. However, in addition, $\mathcal{I}_1$ also contains orientation labels, $\mathcal{O}_\infty = \{o_i\}_{i=1}^{N}$, where $o_i \in [1, ..., N_O]$ is the orientation (for N_O possible orientations).

In order to perform accurate Re-ID, we use this data to build a model constructed from multiple branches, each of which is tasked with learning a specific aspect of the data concurrently. The branches of the model are as follows: A) Identity classification B) Identity classification from a scaled image C) Identity from grayscale image D) Identity plus the vehicles' orienations. These individual branches then form a consensus prediction on the identity of the training examples, and this consensus is then employed for regulation of the individual branches.

3.1. Model Structure and Feature Learning

An overview of our proposed model can be seen in Figure 2. The model is composed of four sub-branches, each of which is simultaneously learning a representation to solve its own task. In addition, there is a single fusion branch, which allows feature selection to be performed from the entire collection of individual representations. It is the output from this branch that is taken during deployment. Each sub-branch will now be described in more detail.

(A) Vehicle Identity The root branch of our model is tasked with learning the best representation for vehicle identity discrimination, for both training sets $\mathcal{I}_1$ and $\mathcal{I}_2$. Here, we exploit the cross entropy classification loss function in order to train one branch to predict vehicle identity. Thus the branch calculates the softmax posterior probability of the class label y_i for a given training image $\mathbf{I_i}$:

$$p_i^{ID} = p(\hat{y_i} = y_i|\mathbf{I_i}) = \frac{\exp(\hat{y_i})}{\sum_{k=1}^{N_{id}} \exp(\hat{y_k})} \tag{1}$$

where $\hat{y_k} = \mathbf{w_k^T x_i}$, $\mathbf{x_i}$ is the feature vector for image $\mathbf{I_i}$ given by final layer of the branch, and $\mathbf{w_k}$ is the prediction function parameter for identity class *k*. The loss across a minibatch of N_B images can then be computed as:

$$L_{ID} = -\frac{1}{N_B} \sum_{i=1}^{N_B} \log p_i^{ID} \tag{2}$$

(B) Identity from Scaled Image Here we exploit the multi-scale analysis that has previously been shown to be of benefit for the task of re-identification, both for persons [4] and vehicles [11]. This is done by including a branch that is trained via cross entropy loss (Eq. 2) to predict the class identity from a rescaled version of the input image, in a similar way to branch A.

(C) Identity from Grayscale Image In order to encourage the model to focus on details of the vehicles, that allow for separation of highly similar identity classes, we ensure that one branch will be unable to use colour information for distinguishing between these classes. This is done by giving as input only the grayscale image, and again training the branch to predict identity via the cross entropy loss.

(D) Vehicle Orientation This branch is tasked with learning a representation to simultaneously predict the identity class and the orientation class when this is known. Both sets of labels are simultaneously employed in a joint loss function in order to optimise the branch for prediction of both identity and orientation. As orientation labels are not available for all training data, we employ a selective classifcation subset loss function, that allows the loss to be calculated across only the subset of the batch for which orientation labels are known.

Again, the cross entropy loss is exploited for this task. Hence, the branch calculates both Eq. (1), as well as the softmax posterior probability of the orientation label o_i for the images for which the orientation class is known:

$$p_i^O = p(\hat{o_i} = o_i|\mathbf{I_i}) = \frac{\exp(\hat{o_i})}{\sum_{j=1}^{N_O} \exp(\hat{o_j})} \tag{3}$$

where this time $\hat{o_j} = \mathbf{w_j^T x_j}$, and $\mathbf{w_j}$ is the prediction function parameter for orientation class *j*.

The loss for this branch is then calculated across the minibatch of images as:

$$L_O = -\frac{1}{N_B} \sum_{i=1}^{N_B} \log p_i^{ID} + \frac{1}{N_S} \sum_{i=1}^{N_B} q_i^O \tag{4}$$

where N_S is the size of the subset of the minibatch for which orientation labels are known, and

$$q_i^O = \begin{cases} \log p_i^O & \text{if } o_i \in \mathcal{O} \\ 0 & \text{otherwise} \end{cases}$$

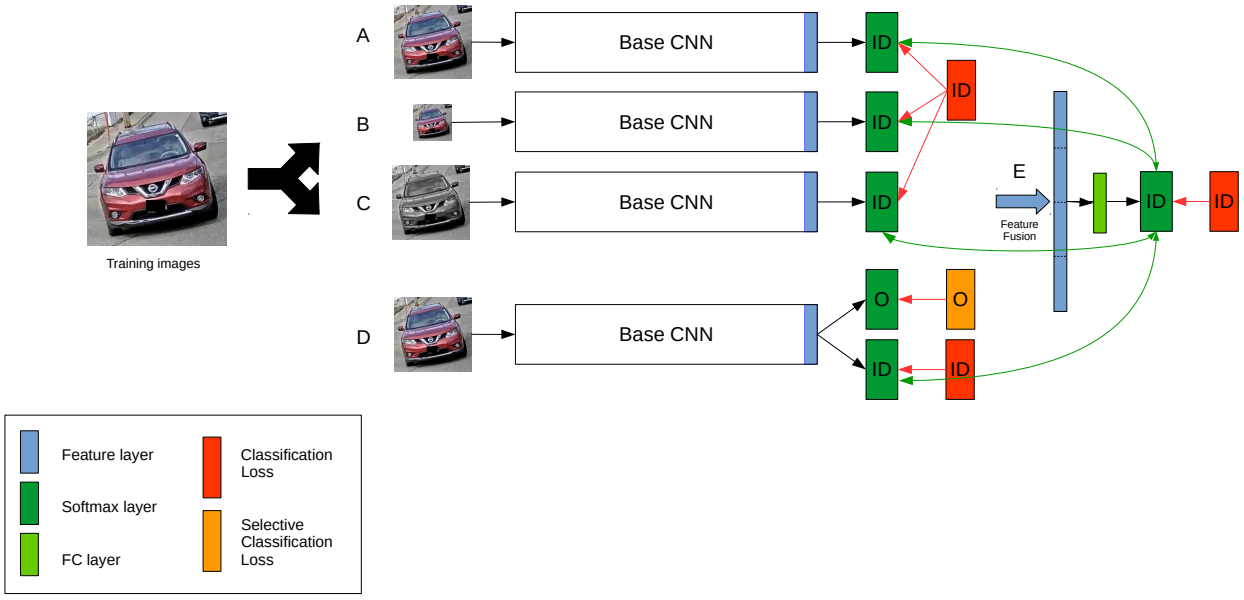

Figure 2: An overview of our proposed model (best viewed in colour). (A) Vehicle identity branch (B) Multi-scale analysis branch (C) Grayscale analysis branch (D) Vehicle orientation branch (E) Consensus learning through feature fusion. Feed-forward signals shown in black. Hard target (groundtruth) loss propagation shown in red. Soft target consensus feedback loss propagation shown in green.

(E) Consensus Learning and Feedback In order to harness the benefit of all branches for the purpose of vehicle re-identification, we employ consensus learning as proposed in [4] and previously harnessed for vehicle Re-ID in [11]. This is done via feature fusion of the final convolutional feature maps from all branches for consensus learning. As our branches are based on the ResNet50 architecture [8], these feature maps are formed via an average pooling operation which result in feature vectors of length 2048. Hence our fused features are of length 8192. We then add one additional fully connected layer, of size 1024, and the output of this passed to a final identity softmax classification layer, again employed with cross entropy loss. Hence:

$$p_i^C = p(\hat{y_i^C} = y_i|\mathbf{I_i}) = \frac{\exp(\hat{y_i^C})}{\sum_{k=1}^{N_{id}} \exp(\hat{y_k^C})} \quad (5)$$

Additionally, we also utilise a consensus propagation mechanism, similar to the previously proposed method [4, 11]. Here the consensus output is taken as 'soft targets' (as opposed to the groundtruth label 'hard targets') for the training data, and used to feedback information about the predictions made by the entire ensemble of branches. This is done concurrently with the training of the individual branches. Thie method is inspired by the idea of Knowledge Distillation (KD) [9], but is different in that here we employ the combined predictions from all the 'student' branches as a *virtual* teacher model, rather than utilising a pre-trainined powerful teacher model to provide the soft targets.

Specifically, the feedback mechanism employs the consensus probability predictions $P_i^C = [p_{i,1}^C, ..., p_{i,j}^C, ..., p_{i,N_{ID}}^C]$ given image $\mathbf{I_i}$, feeding these into the cross entropy loss between the two distributions to provide a consensus regularisation loss for the branch:

$$\mathcal{H}_i = \mathcal{H}(P_i^C, P_i) = -\frac{1}{N_{ID}} \sum_{j=1}^{N_{ID}} p_j^c \log p_j \quad (6)$$

The total consensus loss for a particular branch is then:

$$L_C = \frac{1}{N_B} \sum_{i=1}^{N_B} \mathcal{H}_i \quad (7)$$

This is added to each individual branch's loss functions. In addition, this mechanism provides regularisation of the whole network by propagating all of the consensus losses back through the feature fusion layer, which also boosts the learning of the ensemble.

3.2. Model Training

In order to train our model, we combine both training sets, $\mathcal{I}_1$ and $\mathcal{I}_2$, and employ batches that contain both im-

Algorithm 1 The MTML training algorithm.

Require: Training sets $\mathcal{I}_1$ $\mathcal{I}_2$, labels $\mathcal{Y}_1$ $\mathcal{Y}_2$ $\mathcal{O}_1$, model $\mathcal{M}$

- Initialise network branches with pre-trained ImageNet weights
- Initialise output layers of $\mathcal{M}$ randomly

for epoch $e \in (1, E)$ **do**

- Feed-forward through model to obtain all branch identity classification predictions on images in $\mathcal{I}_1$ and $\mathcal{I}_2$
- Feed-forward to obtain orientation classification predictions on $\mathcal{I}_1$
- Fuse features and perform consensus identity classification predictions on both training sets
- Calculate hard and soft losses identity losses for each branch and backpropagate to update weights
- Calculate orientation losses using labels for $\mathcal{I}_1$ and backpropagate to update weights on the orientation branch
- Calculate hard and soft identity losses for the consensus branch and backpropagate

end for

ages with and without orientation labelling. The full training algorithm can be seen in Algorithm 1.

3.3. Vehicle Re-ID deployment

During deployment, we employ the feature fusion layer from our trained model as the full feature representation in order to perform vehicle re-identification matching. As we do not necessarily have camera information about the query or gallery images, or timestamp information, which would allow the use of camera distance or time-based analysis, we use only a generic distance metric - the $L2$ metric - in order to match gallery images to the query. Hence, for each of the query image $\mathbf{I}^q$, and the gallery images $\{\mathbf{I}_i^g\}$, we compute our 6400 dimension fused feature representations, $\mathbf{x}^q$ and $\{\mathbf{x}_i^g\}$ respectively. We then calculate the $L2$ distance between the query representation and each of the gallery images, and rank the latter by increasing distance in order to calculate the Rank-1 and mAP performance scores.

4. Experiments

We conduct a number of experiments to explore the performance of our method. First we exploit a number of widely available vehicle ID benchmark datasets in order to assess the benefit of each of the branches of our model independently, and altogether. Then we compare the performance of our model to other current work by looking at our performance in the NVIDIA AI City Challenge 2019 Task 2 (Vehicle Re-identification). As our method includes a branch that predicts vehicle orientation in addition to identity, our model requires data that contains the orientation labels for training. As a result, we include the VeRI776 [16] dataset in the training set for all our experiments.

4.1. Datasets

We employ two Vehicle Re-ID datasets in our experiments, in order to train and test our method extensively. Firstly we conduct experiments on a benchmark dataset, VeRi-776 [16], which has been widely tested by the majority of recent works. And secondly we employ the new CityFlow dataset [25], a challenging dataset that has been shown to be more difficult than previous publicly available benchmarks. The VeRi-776 dataset [16] has 37,778 images of 576 IDs in the training set and 200 IDs in the test set. The standard probe and gallery sets consist of 1,678 and 11,579 images, respectively. There are also orientation labels, for 8 possible orientations, available for the VeRi-776 dataset, which were provided by [26]. The CityFlow dataset [25] has 36,935 images of 333 IDs in the training set and 333 different IDs in the test set. The standard probe and gallery sets consist of 1,052 and 18,290 images respectively. The data split statistics of both datasets are summarised in Table 1.

4.2. Implementation Details

We employ the ResNet50 [8] network architecture as the base of our model. We train the model with minibatches of size 8, using the Adam optimisation technique with a learning rate of 0.0001, exponential decay rates set as follows: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The two image sizes used were standard 224x224 and small (for the scaled branch) 160x160.

We measure the performance of our vehicle re-identification methods according to the standard Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). The CMC is computed on each individual rank k as the cumulative percentage of correct matches appearing at ranks $\leq k$. The mAP is calculated as the mean over all query images of the Average Precision, which itself is calculated as the precision cut-off at each correct recalled image position averaged over all possible correct gallery images.

4.3. Evaluation on Veri776 Dataset

Firstly, we train and test on the VeRi-776 dataset in order to compare with existing state-of-the-art methods with identical settings. In order to experiment the benefit of adding each of the separate branches of our model, we take branch A (vehicle identity) as our central branch, and all models have an E consensus branch. We then perform experiments where we include each of the other branches in turn. So MTML-S refers to a model built from branches A, B and E only, MTML-G has branches A, C and E only, and so on. We then test versions of the model with three branches

Dataset	Training			Probe		Gallery	
	#IDs	#Imgs	#Orients	#IDs	#Imgs	#IDs	#Imgs
VeRi-776 [16]	576	37778	8	200	1678	200	11579
CityFlow [25]	333	36935	-	333	1052	333	18290

Table 1: Details of the datasets employed for train and test.

Method	mAP	Rank-1	Rank-5
MSVF [11]	49.3	88.6	-
OIFE [26]	51.4	68.3	89.7
S-CNN+P-LSTM [20]	58.3	83.5	90.0
MTCRO [29]	61.6	87.2	94.2
MTCRO (ReRank) [29]	62.6	88.0	94.6
MTML-S	59.4	89.5	94.9
MTML-O	60.8	90.2	95.4
MTML-G	62.8	91.1	95.8
MTML-SG	63.7	90.6	95.8
MTML-OG	63.5	92.0	**96.4**
MTML-OSG	**64.6**	**92.3**	95.7
MTML-OSG (ReRank)	***68.3***	92.0	94.2

Table 2: Trained/tested on VeRi-776 only

plus E (e.g. MTML-SG includes branches A, B, C and E). MTML-OSG (branches A, B, C, D and E) is then our full model, and MTML-OSG (ReRank [34]) the results of the full model after additional re-ranking. Table 2 shows the results of this, with all experiments run for 150 epochs of training. As can be seen, even before re-ranking, our full model achieves state-of-the-art mAP and Rank-1 scores on this dataset, of 64.6% and 92.3% respectively. And after re-ranking the mAP score is increased to an impressive 68.3%.

The mAP results from the experiments also show how the individual branches contribute to the performance, with orientation (60.8%) improving over the result of scaled analysis alone (59.4%), and grayscale doing even better (62.8%). This demonstrates how allowing the model to learn about the orientation of the vehicle at the same time as identity can strengthen the performance. And that removal of the colour during learning - though obviously a useful indicator of identity at test time - allows for the model to focus on the more discriminatory features of the identity that ultimately boosts the re-id performance.

The combinations of three branches all show improvement over only two, with MTML-SG and MTML-OG achieving mAPs of 63.7% and 63.5% respectively. However they are still outperformed by combining all four branches in the MTML-OSG model. These results show that combining all the different signals for MTL does indeed allow for the overall model to perform better in the final task of vehicle re-identification.

4.4. Evaluation on CityFlow Dataset

Method	mAP	Rank-1	Rank-5
Resnet50 [25]	25.5	41.3	-
MTML-S	17.0	40.4	53.3
MTML-G	19.6	44.5	58.4
MTML-SG	20.6	44.1	55.8
MTML-SG (ReRank)	25.7	43.4	47.2

Table 3: Trained/tested on CityFlow

	VeRi-776			CityFlow		
Method	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
MTML-S	58.4	88.5	94.6	18.9	40.6	53.3
MTML-O	59.2	89.9	94.9	20.3	44.3	56.0
MTML-G	61.6	89.7	95.1	21.6	46.1	57.5
MTML-SG	62.6	90.8	**95.8**	22.1	45.8	56.5
MLML-OG	62.0	**91.2**	95.6	22.9	46.6	58.2
MTML-OSG	**62.6**	90.6	95.5	**23.6**	**48.9**	**59.7**
MTML-OSG (ReRank)	***66.4***	91.5	93.6	***29.2***	48.8	50.7

Table 4: Trained/tested on CityFlow+VeRi-776

We participated in Task 2 of the NVIDIA AI City Challenge 2019. The aim of this task was attempt city-scale multi-camera vehicle re-identification. Multiple cameras were placed at multiple intersections and no camera information was provided about the images.

Two sets of experiments are conducted on CityFlow benchmark: (1) Training on CityFlow, and (2) Training on CityFlow and VeRi-776. For the first set of experiments, we trained MTML-S (branch A, B and E), MTML-G (branch A, C and E) and MTML-SG model only on CityFlow training data. Table 3 shows that: (1) MTML-G branch combination is much better than MTML-S branch combination. The potential reason is that grayscale analysis is more useful than mult-scale in vehicle Re-ID. (2) Joint learning with MTML-SG is better than any individual one of them on mAP evaluation. Another interesting observation is that after re-ranking algorithm, mAP performance improves while CMC performace drops. The possible reason is that more similar image vehicles will get close after re-ranking. This means some false matching images at high rank will "at-

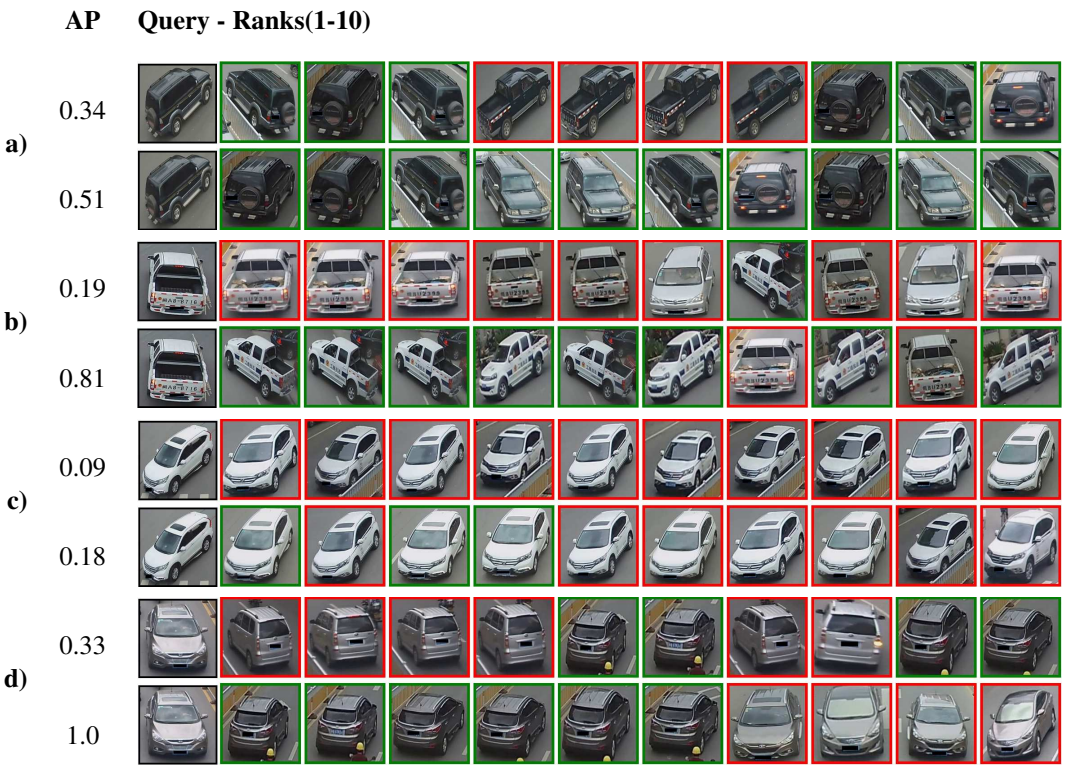

Figure 3: Qualitative comparison of example query images between experiments with and without inclusion of the orientation branch. Rank-1 to Rank-10 is shown. Each pair compares MTML-SG (upper) to MTML-OSG (below) trained and tested on the VeRi-776. AP refers to Average Precision of that query. Correct and incorrect identity matches are shown with green and red borders around images, respectively.

tract" more false matching images after re-ranking. In such a case, it will impact CMC performance.

Since the orientation label is only available in VeRi-776, for training the full MTML model including orientation supervisory signal, we did the second experiment which included both this database plus CityFlow. Table 4 shows that: (1) By adding VeRi-776 training data, with MTML-S, MTML-G and MTML-SG, we all obtain a better mAP and CMC performance on CityFlow than the model which was only trained on CityFlow. Meanwhile, the mAP and CMC performance is a slightly lower on VeRi-776 than the model only training on VeRi-776. We suspect this is due to training for a shorter period of time, as this experiment ran for only 100 epochs, compared to 150. (2) By adding the orientation branch, for the MLTML-OSG model, we obtain the best mAP performance 62.6% and 23.6% on VeRi-776 and CityFlow respectively. This is improved to 66.4% and 29.2% with reranking. This shows that our method of mutual learning between the orientation branch supervised by orientation labels and the other branches supervised by ID label is effective.

Qualitative results showing a comparison of rankings with or without orientation branch are shown in Figure 3. The advantage of learning both the orientation and ID signal can be seen in each pair where the MTML-OSG model is able to rank very different views of the same ID vehicle highly, which compares to the MTML-SG model which can only find images containing similar viewpoints, many of which are incorrect IDs (Figure 3(a,b,d)). It can also be observerd that similar viewpoints are better grouped together in the ranks (Figure 3(c,d)).

5. Conclusions

Vehicle Re-ID is a difficult problem due to the fact that the visual appearance of a vehicle instance may drastically change due to diverse viewpoints and illumination, whilst different vehicle instances of the same model type may have a very similar appearance. In this paper, we propose a novel *Multi-Task Mutual Learning* (MTML) deep model to learn discriminative feature simultaeously from multiple branches. Moreover, we design a consensus learning loss function by fusing feature of the final convolutional

feature maps from all branches. Extensive comparative evaluations demonstrate the effectiveness of the proposed MTML method in comparison to the state-of-the-art vehicle Re-ID techniques on the a existing large-scale benchmarks VeRi-776. We also yield a competitive performance on the NVIDIA 2019 AI City Challenge Track 2.

6. Acknowledgements

This work is partially supported by Vision Semantics Limited and the Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007.

[2] Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2590–2600, 2017.

[5] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):392–408, 2018.

[6] Qi Dong, Shaogang Gong, and Xiatian Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *IEEE Winter Conference on Applications of Computer Vision*, pages 520–529, 2017.

[7] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. Springer, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Qichang Hu, Huibing Wang, Teng Li, and Chunhua Shen. Deep cnns with spatially weighted pooling for fine-grained car recognition. *IEEE Transactions on Intelligent Transportation Systems*, 2017.

[11] Aytac Kanaci, Xiatian Zhu, and Shaogang Gong. Vehicle re-identification in context. In *In Proc. German Conference on Pattern Recognition*, 2018.

[12] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[14] Liang Liao, Ruimin Hu, Jun Xiao, Qi Wang, Jing Xiao, and Jun Chen. Exploiting effects of parts in fine-grained categorization of vehicles. In *IEEE International Conference on Image Processing*, 2015.

[15] 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.

[16] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European Conference on Computer Vision*, pages 869–884. Springer, 2016.

[17] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[18] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2265–2274, 2018.

[19] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *European Conference on Computer Vision*, pages 486–504, 2018.

[20] 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 *IEEE International Conference on Computer Vision*, 2017.

[21] Jakub Sochor, Adam Herout, and Jiri Havel. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3006–3015, 2016.

[22] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[23] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *European Conference on Computer Vision*, pages 402–419, 2018.

[24] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[25] 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. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[26] 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.

[27] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[28] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[29] Dongwu Xu, Congyan Lang, Songhe Feng, and Tao Wang. A framework with a multi-task cnn model joint with a re-ranking method for vehicle re-identification. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, ICIMCS '18, pages 1:1–1:7, New York, NY, USA, 2018. ACM.

[30] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[31] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.

[33] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016.

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

[35] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018.