A Triplet-Learnt Coarse-to-Fine Reranking for Vehicle Re-identification

Efklidis Katsaros[1,2], Henri Bouma[2][a], Arthur van Rooijen [2], and Elise Dusseldorp [1]
[1]*Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands*
[2]*TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, Netherlands*
henri.bouma@tno.nl, elise.dusseldorp@fsw.leidenuniv.nl

Keywords: Vehicle Re-identification, Deep Learning, Coarse-to-Fine, Computer Vision

Abstract: Vehicle re-identification refers to the task of matching the same query vehicle across non-overlapping cameras and diverse viewpoints. Research interest on the field emerged with intelligent transportation systems and the necessity for public security maintenance. Compared to person, vehicle re-identification is more intricate, facing the challenges of lower intra-class and higher inter-class similarities. Motivated by deep metric learning advances, we propose a novel, triplet-learnt coarse-to-fine reranking scheme (C2F-TriRe) to address vehicle re-identification. Coarse vehicle features conduct the baseline ranking. Thereafter, a fully connected network maps features to viewpoints. Simultaneously, windshields are detected and respective fine features are extracted to capture custom vehicle characteristics. Conditional to the viewpoint, coarse and fine features are combined to yield a robust reranking. The proposed scheme achieves state-of-the-art performance on the VehicleID dataset and outperforms our baselines by a large margin.

1 INTRODUCTION

Vehicle re-identification (re-id) is a recent studied research task in the computer vision community. Given a query vehicle tracked on a camera, we aim to find its re-appearances on different ones. Optimization of transportation, traffic management and mobility increasingly grows research interest on vehicle re-id. Relevant applications consist of automatic toll collection and traffic analysis. Vehicle re-id is addressed independently of license plates as they tend to be indiscernible in low-resolution videos or faked.

Vehicle re-identification is inspired by the extensively studied person re-id field (Zheng et al., 2016) but faces substantially different challenges. Given a static scene, it is easier for human vision to "identify" a query person in the crowd than a vehicle in a parking lot. Vehicles exhibit larger intra-class semantic differences compared to persons under diverse viewpoints and across different cameras. Simultaneously, persons tend to bear discriminative parts robust to illumination. Accordingly, human outfit and skin color are more discriminative than vehicle color. The latter is more prone to vary considerably under different lightings.

[a] https://orcid.org/0000-0002-9363-6870

The proposed method consists of a novel two-fold and triplet-learnt coarse-to-fine scheme. Our approach relies on the simple re-identification assumption that if a vehicle is captured twice by the same view, windshields should be similar. If they are not (different logos or stickers), the vehicles compared are not the same. Prompted by deep metric learning advances, we employ the triplet loss to learn coarse and fine vehicle similarities. Given an input vehicle image, we first extract the whole-image features and use them to perform both the baseline ranking and viewpoint classification. In the second phase, we detect and crop the windshields. Subsequently, we extract the fine windshield features. Conditional to the viewpoints, we use the windshield distances to refine the baseline ranking. Windshield information is explicitly incorporated into our reranking mechanism to further pull same-ID images closer and push different-ID ones further apart.

The rest of this work is structured as follows. In Section 2, we discuss related vehicle re-id studies. In Section 3 we propose our "C2F-TriRe" scheme and describe its components. In Section 4, we introduce the dataset and the evaluation metrics. Thereafter, in Section 5, we present the experiments and compare the results of our study. Conclusions are drawn in Section 6.

E. Katsaros, H. Bouma, A. van Rooijen, E. Dusseldorp, "A triplet-learnt coarse-to-fine reranking for vehicle re-identification," Proc. Int. Conf. Pattern Recognition Applications and Methods ICPRAM, 518-525 (2020).
https://doi.org/10.5220/0008974005180525
Copyright 2020 SCITEPRESS. http://www.icpram.org

2 RELATED WORK

Softmax loss retains its popularity in the re-id literature. Zhu et al. (Zhu et al., 2019) propose quadruple directional deep learning features to learn view-invariant representations. The authors introduce quadruple directional average pooling layers and a spatial normalization layer in the end of four convolutional networks. The latter are optimized with the softmax over samples and classes. In another work, Guo et al. (Guo et al., 2018) combine the softmax with other partial loss terms into their coarse-to-fine ranking loss function. The partial loss terms impose larger differences on vehicles of different models (coarse) and images of vehicles belonging to the same model but different vehicle identities (fine). A last pairwise loss term is then added to shrink the (intra-class) variance of images belonging to the same vehicle identities.

Given the recent advances in deep metric learning, many studies employ the triplet loss in various settings. Kumar et al. (Kumar et al., 2019) transfer the batch-sampling idea from person (Hermans et al., 2017) to vehicle re-id and introduce new sampling schemes for the triplet loss. Bai et al. (Bai et al., 2018) account for the intra-class variance by formulating subgroups using k-means clustering in the feature space during training. The weighted average of jointly optimized triplet and softmax loss functions is minimized to learn the space. For the triplet loss, the anchor is selected as the center of the positive instances, whereas the negative sample is picked as the closest-to-the-anchor negative instance.

Other works focus explicitly on regional information. Such is the region aware deep model (RAM) (Liu et al., 2018). Liu et al. introduce a multi-branch network trained with both attributes (color, model) and identities supervision. Features extracted from global and local branches are fused and a joint loss function is optimized to account for attributes, local and global information simultaneously. Zhao et al. (Zhao et al., 2019) assume a hierarchical structure in three levels, that is, whole-image, vehicle model and personalized traits like annual service signs, tissue boxes. An SSD detector is employed to learn such objects. Therefore, images are manually annotated with bounding boxes. A boosting model is then trained to aggregate features from all levels and rank images based on classification scores. Learning personalized information likewise, however, depends heavily on the plenitude of annotated objects and the detection performance.

3 PROPOSED METHOD

Motivated by recent advances in deep metric learning, we adopt the triplet loss to learn vehicle features on coarse and fine stages. We propose a novel coarse-to-fine reranking scheme (C2F-TriRe). The proposed scheme relies on a simplistic re-identification assumption; conditional to the viewpoint, same-identity vehicle images should bear similar windshields. Personalized characteristics (stickers, hangings, driver's figure, interior decoration, boxes, etc.) should recur on both reappearances of the same vehicle. The structural components of the proposed scheme are shown in Figure 1. Our coarse network is a DenseNet-121 (Huang et al., 2017) optimized to learn vehicle (dis)similarities. On top of the former a fully connected and an output layer are attached and optimized to classify viewpoints. We then crop the windshields and – in the same spirit with the coarse – train another fine DenseNet-121 to learn semantic similarities between windshields. Accordingly, we use the coarse whole-image features and their L2 distances to conduct the baseline ranking. Thereafter, the fine features are introduced to refine the coarse distances for all query-gallery pairs sharing the same viewpoint. Different-viewpoint pairs retain their initial baseline ranking.

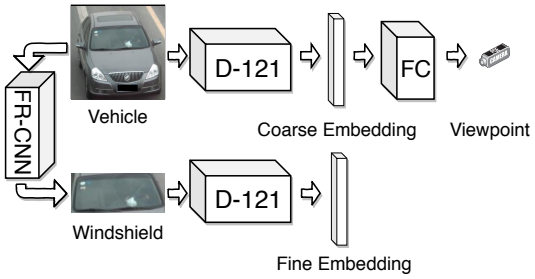

Figure 1: Proposed coarse-to-fine reranking pipeline, with DenseNet-121 (D-121), fully connected layer (FC) and faster R-CNN (FR-CNN).

Prior to the proposed coarse-to-fine reranking, fine windshields need to be detected. Windshield detection is a single-object detection task, that is, interest is on one class. Performance-wisely we resort to the Faster R-CNN (Ren et al., 2015). The detector scans the input image to detect candidate windshield patches. Patches deemed as windshields are further propagated into the detector architecture, for bounding box regression and classification. The coarse network is denoted as f and operates on whole-vehicle images, projecting semantically similar ones to metrically closer d-dimensional embeddings and dissimilar ones further apart, such that:

$$f : \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^{d}, \tag{1}$$

where $\mathbb{R}^{h\times w}$ is the space spanning images of height h and width w. To learn such a mapping, the vision community resorts to the triplet loss (Schroff et al., 2015; Hermans et al., 2017; Kumar et al., 2019). The latter considers three images, an anchor x_n, one positive and one negative sample, as (x_n, x_n^+, x_n^-). The triplet loss for an input triplet is positive when the following constraint is violated:

$$\left\|f(x_n) - f(x_n^+)\right\|^2 \geq \left\|f(x_n) - f(x_n^-)\right\|^2 - a, \quad (2)$$

where α is a pre-defined margin determining the extend up to which the triplet constraint is violated. The triplet loss is then defined as follows:

$$L(f \mid x_n, x_n^+, x_n^-) = max(\left\|f(x_n) - f(x_n^+)\right\|^2 - \left\|f(x_n) - f(x_n^-)\right\|^2 + a, 0) \quad (3)$$

In the next step, we train a fully connected network for vehicle viewpoint classification. Triplet coarse embeddings encode viewpoint information. As such, the fully connected network is attached on top of the coarse DenseNet-121 one. Likewise, no further convolutional operations are required for neither the training nor the inference phase. The classifier consists of a 256-nodes fully connected layer, followed by a REctified Linear Unit (ReLU) and the output one. The viewpoint classifier is denoted as:

$$v : \mathbb{R}^d \rightarrow \{back, front\}, \quad (4)$$

Thereafter, in the same fashion with the coarse DenseNet-121, the fine f_w one is trained with the triplet loss to operate on windshields, projecting them on a d-dimensional space which retains their semantic similarities. Assuming an image x_n, we omit the detection part and denote fine embeddings as $f_w(x_n)$. Fine distances are meaningful only for same-viewpoint vehicle images. Assuming a query image q and a set of gallery images $G = \{g_1, g_2, ..., g_n\}$, coarse and fine distances are – for simplicity – denoted as $d(q, g_i)$ and $d_w(q, g_i)$. We introduce a reranking rule as follows; If a pair (q, g_i) does not share the same viewpoint, we retain their initial baseline distance as per the the coarse model. Otherwise, we swift their distance through a weighted average of coarse and fine distances. The reranking equation is formulated as:

$$d'(q, g_i) = \begin{cases} \lambda d(q, g_i) + (1-\lambda) d_w(q, g_i), & v(q) = v(g_i) \\ d(q, g_i), & otherwise \end{cases} \quad (5)$$

The optimal hyper-parameter λ is determined as the scalar that maximizes performance on the validation set. Our proposed reranking scheme results in a two-fold improvement, as illustrated in Figure 2. When both query and its same-identity expected match (reappearance) are captured from the same viewpoint, combined coarse and fine features lead essentially to more accurate vehicle re-identification. Distracting gallery vehicles of same viewpoint, color and model will be pushed further away due to the dissimilarity of their windshields whereas for the same reason, the expected image will be pulled closer to the query. When query and its reappearance are captured from different viewpoints, distracting same-view gallery vehicles are pushed further away; simultaneously, the expected same-identity one preserves its baseline distance from the query. Accordingly, it is more likely to be retrieved.

4 EXPERIMENTAL SETUP

4.1 Dataset

VehicleID (Liu et al., 2016) is one of the largest and most popular vehicle datasets. It contains 221763 images of 26267 vehicles (8.7 images/ID) split equally on the vehicle identity-level to form the train and test sets consisting of 13164 (110178 images) and 13103 (111585 images) identities respectively. As most real-world re-id datasets, VehicleID is captured by cameras initially installed for plate verification. Subsequently, vehicles appear in two viewpoints, frontal and backside. In contrast to others, VehicleID is not yet saturated.

The 13103 identities form three different testing splits of 800, 1600 and 2400 identities (not all 13103 ones are used) specified by the evaluation protocol. For each of the splits, the gallery set is formulated with one image per identity. The rest of the images are allocated to the corresponding query set. The amount of images in the query and gallery sets for each of the splits is shown in Table 1. Query sizes in the table are slightly different than the ones reported in the original paper, as the version of the dataset employed for this work is corrected for fake license plates.

Table 1: Test set splits on the VehicleID dataset.

Number of images	**Small**	**Medium**	**Large**
Query size	5693	11777	17377
Gallery size	800	1600	2400

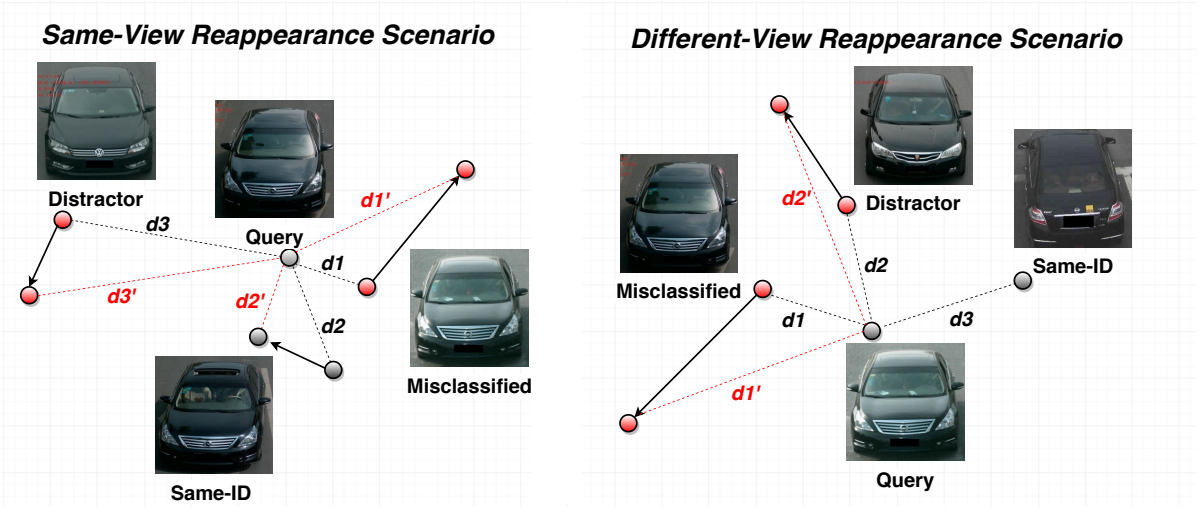

Figure 2: Schematic representation of the proposed C2F-TriRe reranking mechanism. On the right-hand side, our scheme can accurately match identical stickers on the frontal windshields of both query and its reappearance (same-ID), pulling them closer. On the left-hand side, dissimilar windshields push distractors further away; the expected same-ID image retains its initial distance and is thus the correct match after reranking.

4.2 Evaluation Metrics

Following the evaluation protocol, we compute the Cumulative Matching Characteristic (CMC) curve along with the respective Top-1 and Top-5 match rates (accuracy). CMC@k (at rank k) is defined as follows:

$$CMC@k = \frac{\sum_{q=1}^{Q} gt(q,k)}{Q} \tag{6}$$

where $gt(q,k)$ is an indicator function returning 1 when the ground-truth match for query q belongs in the first k ranks, 0 otherwise. Q denotes the total number of queries. Top-1 and Top-5 match rates are equivalent to CMC@1 and CMC@5 respectively. To account for query-gallery splitting variability, results are averaged over 20 replications for all experiments performed.

5 EXPERIMENTS AND RESULTS

All experiments are performed with PyTorch 1.1.0 (Paszke et al., 2017) utilizing two Tesla K80 GPUs and an Intel Xeon E5-265ov3 @2.30GHz processor.

5.1 Softmax-Learnt Ranking

Similarly to Zhu et al. (Zhu et al., 2019), we implement a baseline approach, optimizing DenseNet-121 with the softmax loss function. Softmax guides training towards learning features upon which the feature space is separated into sub-spaces wherein each class belongs (decision areas). Learning the convolutional weights simultaneously with the decision boundaries results in similar identities being clustered closer together and dissimilar ones further apart. Upon inference we rank images based on the L2 distances of the extracted embeddings.

The training setup is similar to Zhu et al., except for the validation strategy. Instead of the classification-based validation approach "Leave-One-Image-Out" (L-O-O), we resort to the "Leave-N-Identities-Out" (L-N-IDs-O). Out of the total 13164 identities in the training set, we randomly allocate 12052 for actual training and the rest 1112 for validation. Likewise, we measure Top-1 and Top-5 accuracy on the validation identities after each training epoch. For training, softmax loss is augmented with an L2 regularization term set at 0.001. Batch size is set to 128, that is, 32 positive and 32 negative image pairs. Images are resized to 224x224, augmented by horizontal mirroring and random rotations within the ranges of [-3 °, 0°] and [0°, 3°]. The network is trained for 80 epochs. The optimizer employed is Stochastic Gradient Descent (SGD) with a momentum of 0.9. Learning rate starts from 0.01 and is multiplied by 0.1 on epochs 45 and 70.

Results of softmax-trained models are depicted in Table 2. Shifting to the Leave-N-Identities-Out validation strategy, we observe an 1.5-6% and 8-10% increase on Top-1 and Top-5 accuracy respectively, compared to match rates reported by Zhu et al.. An appropriate validation strategy is essential to guide softmax optimization towards learning embeddings

Table 2: Performance evaluation of softmax-trained DenseNet-121 models under different validation setups.

Validation Strategy		Small Subset		Medium Subset		Large Subset	
Method	Implementation	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
L-O-O	(Zhu et al., 2019)	66.10	77.87	67.39	75.49	63.07	72.57
L-N-IDs-O	Ours	**72.26**	**89.25**	**68.57**	**84.09**	**65.11**	**80.44**

for re-id. Likewise, a decent baseline is achieved.

5.2 Triplet-Learnt Coarse-to-Fine Reranking Scheme

5.2.1 Coarse Triplet Learning

While softmax preserves inter-class distances to retain efficacy of decision boundaries, the triplet loss learns inter-class differences and intra-class similarities simultaneously. Accordingly, triplet-learnt vehicle embeddings are more representative for re-id as intra-class variance is further minimized. The coarse proposed network is a DenseNet-121 trained with the triplet loss. To address triplet optimization limitations, we construct the triplets "within-the-batch". Therefore, we resort to the batch-all (BA) (Hermans et al., 2017) batch-sampling strategy as it hardly depends on the choice of the margin α (see Eq. 3).

The experimental setup for triplet optimization is described as follows. For each update, we sample P identities comprising of K images each, composing a batch of PK images. Out of these, we construct all possible $PK(PK-K)(K-1)$ triplets and retain the ones that violate the triplet constraint, that is, the ones that yield non-zero loss values. Batch size is set to 72 ($P=12,\ K=4$) and the margin α is set at either 0.3 or 0.7. Following common practices in the literature (Hermans et al., 2017; Kumar et al., 2019), Adam is used as the optimizer, further augmented with an L2 regularization term fixed at 0.001. The learning rate is set to 0.0002 and is multiplied by 0.85 every 30 epochs for 300 in total. One epoch is one pass over all the identities (but not all images) and both our validation and data augmentation frameworks are identical to our classification approach described in subsection 5.1.

Results are shown in Table 3. Top-1 and Top-5 accuracy for DenseNet-121 initialized on ImageNet and trained with a 0.3 margin is 76.68% and 94.70% respectively on the small subset, that is, a minor upsurge compared to the 0.7 margin. Deviations on the larger subsets are of similar magnitude.

Compared to softmax, the triplet-trained coarse DenseNet-121 leads to higher results. Metric learning of parameters f propagates more meaningful gradients for the re-identification task. Accordingly, triplet-learnt vehicle embeddings are robust, that is, they accommodate the viewpoint-caused variability and are thus semantically more representative.

5.2.2 Viewpoint Classification

Conditional to our coarse network, we show that viewpoint classification can be reliably performed with only minor mistakes. Therefore, we design a feed-forward fully connected network to classify vehicle images in two classes, namely, front and backside view.

We annotate viewpoints on 5600 randomly selected vehicle images. The train, validation and test sets consist of 3250, 950 and 1400 images respectively. A fully connected network is attached on our coarse triplet network. The learning rate is set to 0 for the latter and 0.01 for the former. Accordingly, only the newly-introduced fully connected weights are learnt. For all experiments, the learning rate is multiplied by 0.9, every 30 epochs, for 150 in total. Four experiments are performed, connecting the triplet embedding layer (1024) with a pre-last layer entailing 32, 128, 256 or 512 nodes followed by a RELU and a two-nodes output layer followed by a Sigmoid activation. A last sub-experiment is performed connecting directly the embedding with the output layer. The optimization objective loss minimized is the softmax.

The results for our viewpoint classification model are illustrated in Table 4. A network with 256 hidden nodes is the best-performing fully connected component. Respective accuracy on the test set is 98.34%. Architectures with less connections yield inferior results. Therefore, viewpoint classification can be accurately performed and error propagation into the C2F-TriRe scheme is minimized.

5.2.3 Windshield Detection

A Faster R-CNN (Ren et al., 2015) object detector is trained to extract windshields from images of vehicles. Regarding windshield detection, two classes are of interest, windshield or not (background). Faster R-CNN bears a ResNet-51 backbone network and is pre-trained on the COCO dataset (Lin et al., 2014). The latter consists of 81 different classes (80 or background) of instances. The detector is now fine-tuned

Table 3: Performance evaluation of triplet-trained DenseNet-121 models under different margins.

Training Details		**Small Subset**		**Medium Subset**		**Large Subset**	
Loss	Margin	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Triplet	$\alpha = 0.7$	75.93	93.12	71.03	89.94	66.92	86.98
Triplet	$\alpha = 0.3$	**76.68**	**94.52**	**71.76**	**90.97**	**67.78**	**87.45**

Table 4: Fully connected networks performance on viewpoint classification. Validation column refers to the accuracy value evaluated on the validation set upon convergence.

Training Details	**Validation**	**Test**
Connections	Accuracy	
(1024, 2)	77.06	75.13
(1024, 32, 2)	97.58	97.35
(1024, 128, 2)	97.91	97.90
(1024, 256, 2)	98.24	**98.34**
(1024, 512, 2)	98.24	98.19

to learn the two classes.

The experimental setup is the following. The 81-nodes classification layer is substituted by a two-nodes one. The regression layer is retained as such. To fine-tune the detector on learning windshields, we annotate 700 images, 500 for training and 200 for validation. The bounding boxes are manually drawn so as to encompass the whole windshields. The detector is trained for 50 epochs with a batch size of 4. Optimizer used is SGD with a learning rate, a momentum and an L2-penalty of 0.005, 0.9 and 0.0005 respectively. In accordance with the latest detection research we adopt the COCO evaluation metrics. Model weights are saved upon the highest noted AP@IoU of 0.75 on the validation set.

Observed validation statistics converge in 10 epochs, as, first, the detector is pre-trained on a larger and challenging dataset and second, due to the small amount of our manually annotated windshields. Values for AP@IoU of 0.5 and 0.75 are 0.99 and 0.97. Qualitative evaluation of bounding boxes yields no detection errors on inspected images. Boxes predicted on unseen vehicles encompass the whole windshields, identically to our annotations. Faster R-CNN is clearly able to detect windshields, regardless of scale, location and vehicle type.

5.2.4 Fine Triplet Learning

Similarly to the coarse triplet network, we optimize a fine DenseNet-121 to learn semantic similarities between windshields. We infer windshields and viewpoints over the VehicleID training set to construct a custom "windshields" dataset. Then, we group windshield images based on vehicle identity and inferred viewpoint. Each new windshield identity contains all images of a vehicle identity sharing the same viewpoint. Accordingly, a vehicle identity which contained 16 images out of which 10 were front and 6 were backside-viewed, now yields two new windshield identities. We retain only identities with five images or more. The windshields dataset consists of 7362 identities (74658 images) used for training and another 691 (6462 images) for validation.

We train the fine triplet DenseNet-121 under the following setup. Windshields are resized to mean height and width (62×152). We resort to the same BA sampling strategy, using a batch size (*P*K*, with various combinations of *P. K*, of 140 and margins α of 0.2 and 0.5. One pass over the training dataset loops, again, over all windshield classes. We retain random rotations, but we do not flip images horizontally, as discriminative spatial features are subject to be confounded. Optimizer is again Adam, with a learning rate and an L2-penalty of 0.0001 and 0.001. The former is multiplied with 0.9 every 30 for 240 epochs in total. Identically to the coarse network, weights are now saved upon highest validation Top-1 accuracy on matching windshields.

Results for the fine triplet DenseNet-121 on our custom validation set are illustrated in Table 5. A 0.5 margin performs slightly higher. Choice of P,K does not severely impact learning of windshield (dis)similarities. The model optimized with a 0.5 margin over batches of *P*=17 identities and *K*=8 images each is utilized as the fine-grained feature extraction component of the proposed method.

Table 5: Triplet fine DenseNet-121 performance on matching windshields. Validation columns refer to Top-1 and Top-5 match rates observed on the custom validation set upon convergence.

Training Details		**Validation**	
Margin	P×K	Top-1	Top-5
$\alpha = 0.2$	20×7	90.47	95.61
$\alpha = 0.5$	17×8	**90.91**	**96.03**
$\alpha = 0.2$	35×4	90.32	95.57
$\alpha = 0.5$	35×4	90.87	95.79

5.2.5 Coarse-to-fine reranking

Prior to evaluating our coarse-to-fine reranking scheme, the hyper-parameter λ is chosen based on

Table 6: Comparative results of the proposed C2F-TriRe versus baselines and various related works on the VehicleID dataset.

Method	Small Subset		Medium Subset		Large Subset	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
QD-DLF(Zhu et al., 2019)	72.3	92.5	70.6	88.9	69.0	88.3
C2F-Rank(Guo et al., 2018)	61.1	81.7	56.2	76.2	51.4	72.2
XG-6 (Zhao et al., 2019)	76.1	92.0	73.1	88.1	71.2	84.4
RAM (Liu et al., 2018)	75.2	91.5	72.3	87.0	67.7	84.5
GS-TRE (Bai et al., 2018)	75.9	84.2	74.8	83.6	**74.0**	82.7
Triplet BS (Kumar et al., 2019)	78.8	96.2	73.4	92.6	69.3	89.4
Softmax Ranking	72.26	89.25	68.57	84.09	65.11	80.44
Triplet Ranking (Coarse)	76.68	94.52	71.76	90.97	67.78	87.45
C2F-TriRe (k=10)	79.63	95.82	76.06	93.03	70.44	88.62
C2F-TriRe (k=20)	79.51	95.70	75.64	93.11	71.57	89.94
C2F-TriRe (k=30)	79.87	95.78	75.82	93.08	71.77	90.01
C2F-TriRe (k=50)	**80.28**	**96.34**	**76.09**	**93.13**	72.23	**90.08**

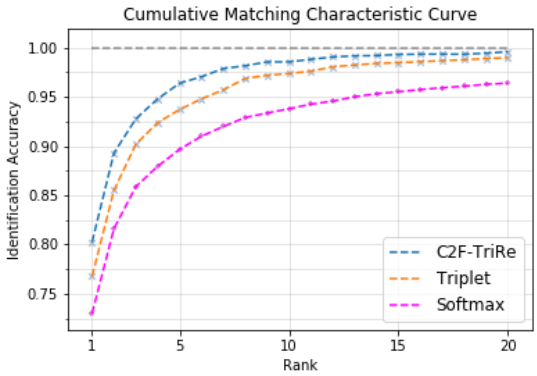

Figure 3: CMC curve of C2F-TriRe versus triplet-learnt (coarse) and softmax-learnt ranking on the small VehicleID subset.

Top-1 match rate on the L-N-IDs-O validation set (1112 vehicle identities). Therefore, we evaluate coarse and fine features over the latter and re-rank gallery candidates conditional to inferred viewpoints (see Eq. 5) for each λ. We perform additional experiments for the parameter k - referring to the number of the nearest (as per the coarse ranking) vehicles to be reranked for each query.

Results of our C2F-TriRe scheme are depicted in Table 6. The CMC curve is shown in Figure 3. The optimal hyper parameter λ was found to be 0.85. C2F-TriRe outperforms the triplet (coarse) and softmax models by Top-1 and Top-5 margins of 3-4.5% and 8-9% respectively on the small VehicleID subset. The margin is consistent over all test sets of VehicleID for both match rates. Moreover, we show that, with a negligible accuracy trade-off one can minimize computational time by restricting windshield detections and evaluations to only 10 (k-closest as per the coarse ranking) for each query. The additional computational cost of C2F-TriRe more than the coarse embedding is negligible when k is much smaller than the gallery set (e.g. $k = 10$, $n = 2400$). Note that, due to sampling variability, match-rates are not necessarily monotonically increasing with k on the small and medium subsets. Compared to the top-performing approaches, C2F-TriRe notes the highest performance on all subsets and ranks of the VehicleID dataset, except for Top-1 on the large version, where GS-TRE reports an 1.8% higher match rate. Regarding Top-5 rates, C2F-TriRe outperforms GS-TRE by a margin of 12%, 10% and 7%, on the small, medium and large subsets, respectively. Furthermore, the proposed method achieves an 1.5-3% upsurge on Top-1 match rate compared to Triplet BS.

6 CONCLUSION

In this paper, we propose C2F-TriRe, a novel coarse-to-fine reranking scheme for vehicle re-identification. The introduced method involves a triplet coarse network operating on whole-vehicle images, a fully connected one for viewpoint classification, a windshield detector and one triplet fine network operating on the windshield area. The whole-image features are extracted with the coarse DenseNet-121; they are utilized for the baseline ranking before inputted for viewpoint classification. Conditional to the viewpoint, a fine DenseNet-121 is then used to extract windshield features. Upon inference, a different-viewpoint query-gallery pair retains its coarse, baseline distance whereas same-viewpoint ones are reranked.

The proposed C2F-TriRe exploits metric learn-

ing advances to incorporate personalized vehicle information into a robust vehicle re-id system. Our C2F-TriRe scheme outperforms both baselines and related works. We show that, extending end-to-end deep learning architectures to a coarse-to-fine vision system leads to increased vehicle re-id performance. Moreover, restricting reranking to a small number k, – referring to the coarsely-ranked nearest neighbors – reduces dramatically computational expenses with only a minor performance trade-off. Future work could include generalization to other viewpoints, such the side view.

ACKNOWLEDGEMENTS

Portions of the research in this paper use the VehicleID dataset collected under the sponsor of the National Natural Science Foundation of China.

REFERENCES

Bai, Y., Lou, Y., Gao, F., Wang, S., Wu, Y., and Duan, L.-Y. (2018). Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399.

Guo, H., Zhao, C., Liu, Z., Wang, J., and Lu, H. (2018). Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Kumar, R., Weill, E., Aghdasi, F., and Sriram, P. (2019). Vehicle re-identification: an efficient baseline using triplet embedding. *arXiv preprint arXiv:1901.01015*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, H., Tian, Y., Yang, Y., Pang, L., and Huang, T. (2016). 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.

Liu, X., Zhang, S., Huang, Q., and Gao, W. (2018). Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Zhao, Y., Shen, C., Wang, H., and Chen, S. (2019). Structural analysis of attributes for vehicle re-identification and retrieval. *IEEE Transactions on Intelligent Transportation Systems*.

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zhu, J., Zeng, H., Huang, J., Liao, S., Lei, Z., Cai, C., and Zheng, L. (2019). Vehicle re-identification using quadruple directional deep learning features. *IEEE Transactions on Intelligent Transportation Systems*.