# Relation Preserving Triplet Mining for Stabilising the Triplet Loss in Re-identification Systems

Adhiraj Ghosh[1,2], Kuruparan Shanmugalingam[1,3], and Wen-Yan Lin[1]

[1]Singapore Management University, [2]University of Tübingen, [3]University of New South Wales

## Abstract

*Object appearances change dramatically with pose variations. This creates a challenge for embedding schemes that seek to map instances with the same object ID to locations that are as close as possible. This issue becomes significantly heightened in complex computer vision tasks such as re-identification(reID). In this paper, we suggest that these dramatic appearance changes are indications that an object ID is composed of multiple natural groups, and it is counterproductive to forcefully map instances from different groups to a common location. This leads us to introduce Relation Preserving Triplet Mining (RPTM), a feature matching guided triplet mining scheme, that ensures that triplets will respect the natural subgroupings within an object ID. We use this triplet mining mechanism to establish a pose-aware, well-conditioned triplet loss by implicitly enforcing view consistency. This allows a single network to be trained with fixed parameters across datasets, while providing state-of-the-art results. Code is available at* https://github.com/adhirajghosh/RPTM_reid.

## 1. Introduction

Re-identification is the process of identifying images of the same object taken under different conditions. One of the main challenges of reID is pose-induced appearance changes [2, 9]. Not only does object appearance change with pose, different objects often look similar when viewed from the same pose, also known as inverse-variability. This paper suggests a new interpretation of the inverse-variability problem, one with the potential to significantly improve the effectiveness of reID algorithms. Although we focus on re-identification, the underlying principles developed here are not restricted to this task and have the potential to impact a wide range of other computer vision problems [1, 21, 30, 35]. Current reID frameworks deploy representation and metric learning methodologies in the attempt to learn embeddings that map semantically similar instances to relatively nearby locations; and semantically dissimilar images to relatively distant locations. This is typically achieved through a metric loss function such as triplet loss [37], which encourages a reference (anchor) input to be more similar to a positive (truthy) input than to a negative (falsy) input. The number of triplet combinations tend to grow polynomially with the number of instances in a dataset, as detailed by Hermans *et al.* [15]; however, most triplet combinations are redundant. This has led to the development of triplet mining, whose aim is to identify the most important triplets in a given sample set. While triplet mining is ubiquitous in reID algorithms [2, 13, 15, 39], it has an innate vulnerability.

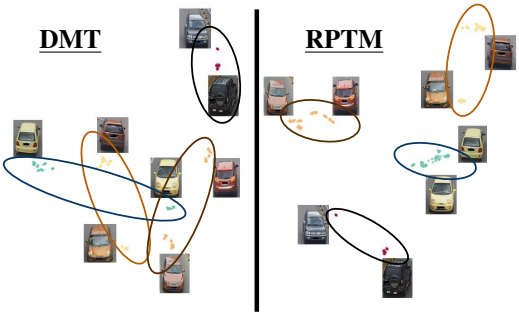
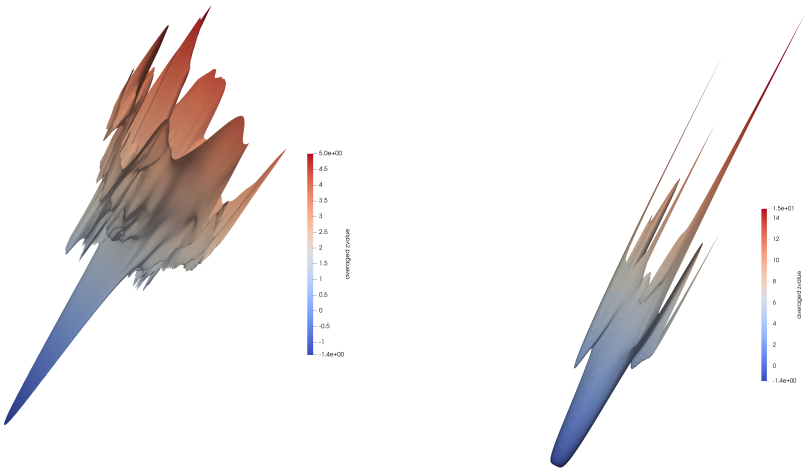



Figure 1: Comparing the features learned by DMT [13], a current state-of-the-art, with our proposed Triplet Mining scheme. Features correspond to the first four IDs of Veri-776 [25]. The distance preserving UMAP projection shows the RPTM feature transform is more intuitive.

Consider a hypothetical dataset containing instances of apple-the-phone and apple-the-fruit, both of which are classified as Apple. The dataset also has instances of phones made by Samsung, classified as Samsung-phone. This dataset will have many difficult triplets, for example, apple-the-phone (anchor), apple-the-fruit (positive) and Samsung-



(a) Without RPTM (b) With RPTM

Figure 2: Loss landscape visualisation of a ResNet-50 trained with SGD using Triplet Loss on Veri-776 with/without Relation Preserving Triplet Mining. RPTM demonstrates smoother loss surfaces, improved model generalisation and a wider minima, thus allowing better optimisation during training.

phone (negative), which triplet-mining techniques are encouraged to focus on. However, training with such triplets is counter-productive as they attempt to ensure that instances of apple-the-phone are mapped closer to instances of apple-the-fruit than to instances of Samsung-phone. Such a mapping mechanism violates the natural appearance relation between objects, and it can be seen that current metric learning systems enforce vastly different views of the same object to be coincident in feature space. It is unlikely that models trained on this hypothesis generalise adequately.

A similar phenomenon occurs in reID, where most datasets [23, 25, 34] group instances by ID. However, the appearance of a person or vehicle's front, rear and sides profiles are very different from each other and they appear to belong to physically different entities. This creates fallacious anchor-positive pairs, where the instances chosen to be anchor and positive do not share a natural group [2]. This fallacy in the triplet mining scheme can be further realised considering the fact that in [37], triplet loss was defined for face detection, where datasets only have the front view of the face, hence all anchor-positive pairs are semantically meaningful. Due to this, triplet mining does not generalise well to reID. This problem has been recognised in recent reID works [9, 19, 24, 28, 39], who incorporate pose awareness into the network, and in metric learning [35], in which latent characteristics shared within and between classes are explicitly learnt. Although this approach can be effective, it complicates network training and incurs an additional burden of training a new, dataset-specific, pose-aware layer.

We suggest a simple alternative, where feature matching [5, 26] is leveraged to discover natural groupings. Therefore, we propose *Relation Preserving Triplet Mining* (RPTM), a triplet mining scheme that respects natural appearance groupings. We further define our solution as *Implicitly Enforced View Consistency*, which we define as the process of exploiting internal, natural groupings within a class and mapping instances with the same view together as a semantic entity, to overcome intra-class separability. These groupings follow natural patterns referenced by semantics [21], and tend to be pose-related in the context of reID. Here, RPTM implicitly enforces pose-aware triplet mining, which prevents different poses from being mapped onto one another. This improves the conditioning of the triplet-cost, allowing for the same training parameters to be employed across a variety of different datasets. The resultant feature embeddings provide better reID results and are more intuitive, as shown in Figure 1. We observe that past triplet mining processes fail in terms of pose awareness and this may lead to poor ranking results, whereas RPTM not only shows pose awareness, better conditioned triplet mining also ensures accurate ranking results.

Our experiments are structured to demonstrate how a coherent triplet mining scheme can eliminate the largest vulnerability of using triplet loss in reID, without the requirement of key-point labels and pose estimation pipelines. One indicator of the effectiveness of our method is observing the loss optimisation landscape during training. Due to the smooth loss landscape for RPTM, shown in figure 2, we demonstrate how RPTM cleans up the triplet mining process with a triplet filtration step and prevents erroneous lo-

cal minimas. Thus, when trained with RPTM, models with larger parameters can optimise just as fast as smaller networks, which serves our main goal of achieving impressive retrieval results with self-imposed constraints on compute power as well as generalising parameter settings across tasks and datasets. As RPTM is robust to fluctuations of loss landscape, training deeper networks with SGD on a simple cost function is more accessible for object retrieval tasks.

In summary, our paper contributions are:

1. We explain how traditional triplet mining methods are ill-conditioned because it does not take into account natural groupings;
2. We propose a feature guided triplet mining scheme that we term Relation Preserving Triplet Mining (RPTM);
3. We show RPTM is well-conditioned enough to permit the use of constant training parameters across datasets and tasks. The resultant network is simultaneously capable of state-of-the-art in vehicle reID and competitive results for person reID.

## 2. Related Works

**Re-identification.** The demand for urban surveillance applications has led to a surge of interest in person and vehicle re-identification. Challenge benchmarks such as VehicleID [23], Veri-776 [25], DukeMTMC [34] and others have been established; and many new algorithms have been proposed [13, 19, 20, 24, 38, 46]. In reID, many algorithms achieve good results by estimating vehicular pose. Notably, Tang *et al.* [39] created a synthetic data set for pose estimation and Meng *et al.* [28] used a parser model to split vehicles into four parts for pose-aware feature embedding. Recently, Vision Transformers (ViT) for reID [14, 42, 50] were proposed for attention learning and [10] addressed person reID with noisy labels. We suggest that the root problem encountered by most of these techniques lies in their definition of triplet loss. By replacing traditional triplet losses with our RPTM technique, we show that it is possible to achieve state-of-the-art results by minimizing a simple cost function. This stands out from the trend towards ever more complex reID techniques.

**Triplet Loss.** The triplet loss was first introduced in the context of face identification [37]. Since then, it has undergone many refinements [2, 15, 44, 45]. Such triplet-based formulations implicitly assume that the given IDs correspond to meaningful groups. We suggest that this assumption is often wrong and that triplets should be defined with respect to naturally occurring groups rather than the given labels. This perspective on triplet loss differs significantly from that used in most papers. To our knowledge, the research most similar to ours is Bai *et al.* [2] who acknowledge the importance of naturally occurring groups within an ID. However, Bai *et al.* attempts to use the groups to force tighter mappings of an ID, fighting rather than harnessing the natural relationships. Another problem for clustering based works like Bai *et al.* [2]'s, is that variations often have no naturally occurring cluster boundaries. This is not a problem for RPTM which defines relations in a pairwise manner, rather than on the basis of shared clusters.

**Feature Matching.** RPTM uses feature matching to help establish triplets. Feature matching is a well-established field in computer vision, whose goal is to match key points between image pairs. Classic feature matching works include SIFT [26], SURF [3], ORB [36], *etc*. Recent developments include [4] for exploiting matching context information, and [27] for mismatch removal between two features sets. In this paper, we employ Grid-Based Motion Statistics (GMS) [5] as our feature matcher of choice. This is a newer algorithm which incorporates match coherence [22] to facilitate key-point matching. GMS outperforms most classic techniques while also being much faster.

## 3. Why Triplet Loss?

### 3.1. Neural Networks as Embedding Functions

Much of computer vision can be interpreted as an attempt to map image instances to a semantically meaningful embedding. Thus, if $\mathbf{x}_k$ represents an image instance and $\mathbf{y}_k$ its associated feature, the transformation from $\mathbf{x}_k$ to $\mathbf{y}_k$ can be denoted by $\mathbf{y}_k = f(\mathbf{x}_k)$, where $f : \mathbb{R}^{3\times w\times h} \rightarrow \mathbb{R}^d$; $w \times h$ denotes image dimension; and $d$ represents the embedding space's dimensions. In this scheme, the embedding function $f(.)$ is learnt by minimising the cross-entropy loss

$$E_{ent} = \sum_{k=1}^{m} \mathcal{L}_{ent}(\mathbf{x}_k), \tag{1}$$

where $m$ denotes the total number of training images.

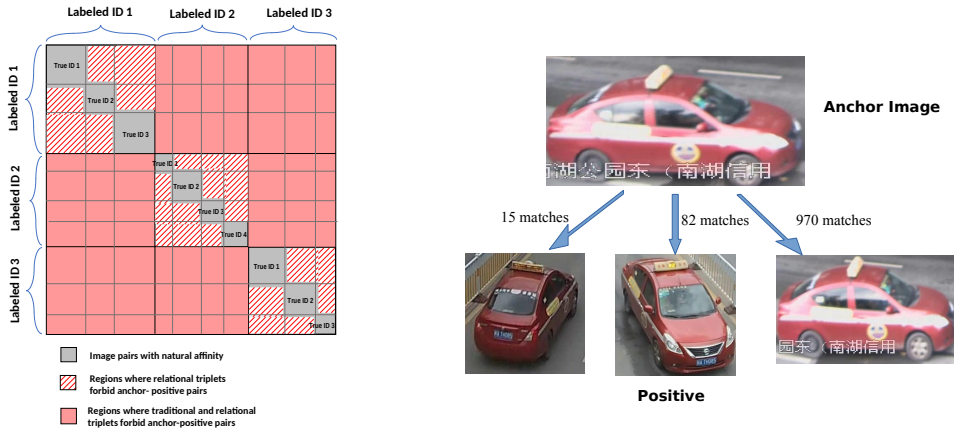
Minimising the cost in Eq. 1 provides an embedding that maximises classification accuracy. However, this does not ensure that the embedding is semantically meaningful. The retrieval problem requires an embedding in which semantically similar instances are mapped close to each other, leading to the development of triplet loss [37].

### 3.2. The Triplet Loss

A triplet loss is defined with respect to three image instances: Anchor (randomly chosen instance); Positive (instance that shares a common ID with the anchor); Negative (instance whose ID is different from the anchor). We denote these instances $\mathbf{x}_a, \mathbf{x}_p$ and $\mathbf{x}_n$, respectively. Given the anchor, positive and negative, the triplet loss is defined as [37]:

$$\mathcal{L}_{tri}(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) = \max(0, d_{ap} - d_{an} + \alpha), \tag{2}$$

where $\alpha$ is the desired margin separation between positive and negative instance, $d_{ap} = \|f(\mathbf{x}_a) - f(\mathbf{x}_p)\|$ and $d_{an} =$




(a) Affinity matrix

(b) Anchor-positive selection scheme

Figure 3: **Representational Schematic for Relation Preserving Triplet Mining**. In figure 3a, each ID contains a number of naturally occurring groups. Relational triplets are based on natural groups rather than IDs, thus preventing pathological anchor-positives. In figure 3b, observe that the positive shares clear similarities with the anchor(indicating they share a common natural group) but is not a near-duplicate.

$\|f(\mathbf{x}_a) - f(\mathbf{x}_n)\|$. The final triplet-cost is computed by summing the individual triplet losses:

$$E_{tri} = \sum_{c=1}^{t} \mathcal{L}_{tri}(\mathbf{x}_{ac}, \mathbf{x}_{pc}, \mathbf{x}_{nc}), \tag{3}$$

where $t$ is the total number of triplets. In general, triplet costs are not used in isolation. Instead, they are combined with the cross-entropy cost from Eq. 1, leading to the final cost function:

$$E = \lambda_{ent} E_{ent} + \lambda_{tri} E_{tri}, \tag{4}$$

where $\lambda_{ent}$ and $\lambda_{tri}$ control the weights given to the cross-entropy loss and triplet-cost respectively.

## 4. Relation Preserving Triplet Mining

To prevent training pipelines from stagnating, it is important to implement a good triplet mining scheme. Triplet mining is part of a larger framework which views features as the key to machine learning. For example, NetVLAD [1] and many other domain transfer works, show that adapting features significantly improves performance. Somewhat similarly, knowledge distillation [30] tries to compress unwieldy networks into more compact features for practical deployment. We focus on triplet mining, as most related works in reID use some form of triplet loss and also to effectively highlight *Implicitly Enforced View Consistency*.

Naïvely incorporating every possible triplet into the loss yields poor results [15]. Instead, training algorithms employ triplet-mining, a process which aims to incorporate only the most relevant triplets into the triplet-cost. Unfortunately, there is no consensus on how relevance can be measured; thus, triplet mining relies on heuristics. The two most popular heuristics are: hard-negative mining and semi-hard negative mining. Hard-negative mining focusses on triplets whose negatives are very similar to the anchor. Semi-hard negative mining shifts the focus from the hardest negatives to negatives close to the decision boundary. Both heuristics seem sensible and often perform well; however, closer inspection suggests something may be amiss.

Let us perform a thought experiment where we assign the IDs A and B, to similar car models. Hard or semi-hard mining finds the most confusing triplets, leading to the following triplet: front of car A as anchor, rear of car A as positive, and front of car B as negative. The triplet is indeed very hard; however, its incorporation into the training cost is counter-productive. This is because such a triplet encourages embedding mapping the rear of A to the front of A. The embedding is so counter-intuitive, it is unlikely to generalise well. To avoid such pathological cases, we introduce relational triplets, which address the problem of intraclass separability with greater attention than other methods.

### 4.1. Relational Triplets

Relational triplets change the triplet definition from one based on human assigned IDs to one based on naturally occurring groups. Formally, we denote the set of training images as $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_K\}$. We hypothesise that these images are members of naturally occurring (and possibly overlapping) subsets. The set of subsets is denoted by $\mathcal{N} = \{\mathcal{S}_m\}$, where

$$\mathcal{S} = \bigcup_{\mathcal{S}_m \in \mathcal{N}} \mathcal{S}_m. \tag{5}$$

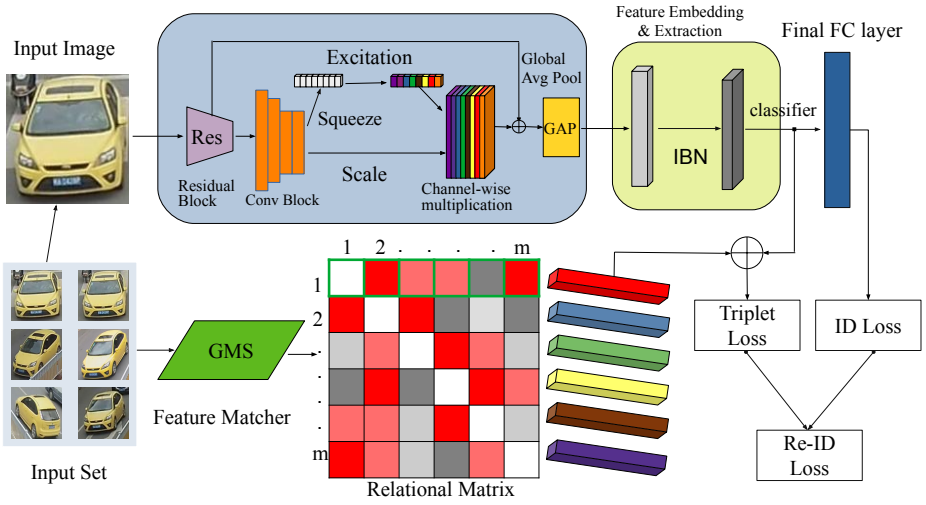



Figure 4: Schematic of a re-identification network deploying **Relation Preserving Triplet Mining**. The RPTM module includes Instance-Batch Normalisation (IBN) and Squeeze-Excitation (SE) to reduce channel inter-dependencies. The relational matrix is estimated using GMS matches and is used for triplet selection.

We use the relational indicator.

$$C(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1, & \text{if } \mathbf{x}_i, \mathbf{x}_j \text{ share a subset in } \mathcal{N}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

to denote whether two instances share a natural subset. A relational triplet is one where the anchor-positive pair shares a common natural subset, while the negative does not.

$$C(\mathbf{x}_a, \mathbf{x}_p) = 1, \quad C(\mathbf{x}_a, \mathbf{x}_n) = 0, \quad C(\mathbf{x}_p, \mathbf{x}_n) = 0. \quad (7)$$

Traditional triplets are a special case of relational triplets, where the given IDs mirror the natural subsets. This is not the case in reID, as we explained in the thought experiment and through the relational diagram in Figure 3a. Pathological triplets arise when anchor-positive pairs do not have natural affinity (share a common group). Observe that the traditional ID based triplet permits pathological anchor-positive pairs. In reID, the natural subsets likely correspond to object poses. This creates the possibility for identifying such subsets using a feature matching algorithm. The next section shows how this can be achieved.

### 4.2. Mining the Relation Preserving Triplets

GMS [5] is a modern feature matcher that uses coherence to validate hypothesised feature matches. The coherence scheme assumes that a true match hypothesis will be strongly supported by many other match hypotheses between neighbouring region pairs, while a false match hypothesis will not. The coherence-based validation is notably better than the traditional ratio test [26]. This allows GMS to reliably match features across significant viewpoint changes while simultaneously ensuring few matches between image pairs with nothing in common. As a result, the presence of GMS matches between image pairs provides a good approximation of the relational indicator in Eq. 5. GMS is quite effective in reID systems to quantify the innate relation between images and is crucial in establishing implicitly enforced view consistencies.

While GMS has few errors, errors do occur. To ensure an anchor-positive pair has a relational indicator of one, we set the positive instance of each anchor to be the image instance whose number of GMS matches with the anchor is closest to the threshold $\tau$. Here, we accept that setting similar anchor-positive pairs leads to poorer training. Hence, we use a middle-ground approach for anchor-positive selection, which we call $RPTM_{mean}$, in which $\tau$ is set as the average number of GMS matches in the set of nonzero pairwise GMS matches between the anchor and all other images. More formally, for two images, $\mathbf{x}_i, \mathbf{x}_j$ we predict that the natural relational indicator is true, $C(\mathbf{x}_i, \mathbf{x}_j) = 1$, if the number matches between them exceed $\tau$.

The above provides a semi-hard positive mining, that ensures anchor-positive pairs satisfy the relational indicator in Eq. 5, while also ensuring that the positive differs significantly from the anchor. An example is shown in Figure 3b.We define negatives using batch hard-triplet mining [15]. If $\mathcal{S}_b = \{\mathbf{x}_j\}$ denotes the set of instances in a batch that do not share an ID with $\mathbf{x}_a$, the negative is

$$\mathbf{x}_n = \underset{\mathbf{x}_j \in \mathcal{S}_b}{\operatorname{argmin}} \left( \|(f(\mathbf{x}_a) - f(\mathbf{x}_j)\| \right). \quad (8)$$

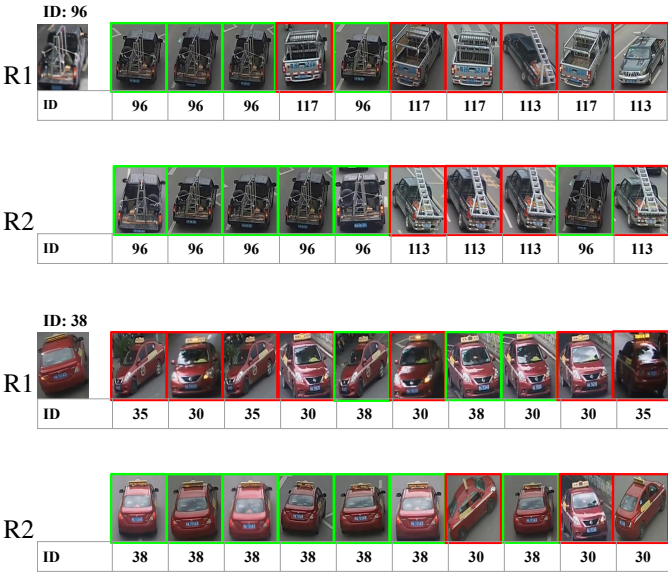



Figure 5: Qualitative retrieval results for bad targets without RPTM(R1) and with RPTM(R2). Correct identifications are outlined in green; wrong ones are outlined in red. RPTM clearly aligns backbone models with better pose awareness and provides fine-grained attention.

Observe that the triplets defined in this manner satisfy Eq. 7, making them relation preserving triplets. Given such triplets, the final embedding can be obtained by minimising the cost function in Eq. 4. As evidenced by the mining strategy, RPTM allows for an intrinsic understanding of the viewpoint and pose without hard coded pose estimation.

# 5. Implementation Details

A schematic of the network architecture is provided in Figure 4. In this section we discuss the model layout, elaborating on the comparative feature matching pipeline in Section 5.1 and the model structure with RPTM in Section 5.2. To test and highlight the universality of RPTM and its ability to generalise the training pipeline due to its novel triplet mining scheme, we put limitations on network and parameter tuning across all datasets.

## 5.1. Feature Matching

As discussed before, we use GMS feature matching to guide our triplet-mining process, in order to implement semi-hard positive mining. In theory, we need to establish GMS matches between an anchor and every other image in the dataset. In practise, we use image IDs as guides to the natural groupings and restrict the matching to only images that share a common ID with the anchor. This greatly reduces computational cost in triplet mining. Feature matching is performed on images that have been resized to (224, 224). The GMS feature matching parameters are: 10,000 ORB features whose orientation parameter is set to true and nearest neighbours are identified with the brute-force hamming distance. All other parameters are set according to the guidelines reported by [5]. After matching, the number of matches between image pairs is stored in a relational matrix $m \times m$, where $m$ is the number of training images.

## 5.2. Neural Network

For fair comparison of our results with established benchmarks, we chose ResNet-50 and ResNet-101 pretrained on ImageNet as our backbone. Our RPTM module includes instance-batch-normalization and a squeeze-excitation layer[16]. The weights of this network are trained by minimising the loss function in Eq. 4. This network is trained using triplets defined through our Relation Preserving Triplet Mining (RPTM) in Section 4.2.

The images are resized to (240,240) for vehicle reID and (300,150) for person reID. Data augmentation is applied, with random flipping, random padding, random erasing and colour jitter (randomly changing contrast, brightness, hue and saturation) all activated. Stochastic Gradient Descent(SGD) is used as the optimiser for the model. The initial learning rate is initialised at 0.005 and is set to decay by a factor of 0.1 every 20 epochs. The model is trained for 80 epochs with a batch size of 24. Training parameters are fixed for all datasets. [1]. Finally, Figure 5 provides qualitative comparisons showing that RPTM's top-k-ranked retrievals are significantly better than its backbone network (we showcase top-k results alternatively (top-1, top-3...top-19). We focus on demonstrating the quality of gallery image retrieval for query samples by RPTM by comparing top-k retrieval results with and without the RPTM pipeline.

# 6. Experiments

## 6.1. Datasets

**VehicleID** [23] allows us to test RPTM's scalability by offering multiple, progressively larger (and harder) testsets. We evaluate our algorithm with 800, 1600 and 2400 labels for testing. **Veri-776** [25] is a widely used benchmark with a diverse range of viewpoints for each vehicle and is designed to provide more constrained but highly realistic conditions. **DukeMTMC** [34] is a person re-identification benchmark with 1,404 distinct classes. While our focus is vehicle reID, we include this benchmark to show our algorithm can generalise to other problems.

## 6.2. Evaluation Metrics

Rankings are scored according to the protocols suggested in [23, 25] and all methods are reported with mean

[1]These parameters are significantly less computationally demanding than those used by recent state-of-the-art models [10, 14, 33, 41, 50]

| Model | Small (query size=800) | | | Medium (query size=1600) | | | Large (query size=2400) | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAP | r=1 | r=5 | mAP | r=1 | r=5 | mAP | r=1 | r=5 |
| C2F-Rank [11] | 63.50 | 61.10 | 81.70 | 60.00 | 56.20 | 76.20 | 53.00 | 51.40 | 72.20 |
| AGNet [47] | 76.06 | 73.14 | 86.25 | 73.39 | 70.77 | 81.75 | 71.75 | 69.10 | 80.40 |
| ANet [31] | - | 86.00 | 97.40 | - | 81.90 | 95.10 | - | 79.60 | 92.70 |
| VANet [9] | - | 88.12 | 97.29 | - | 83.10 | 95.14 | - | 80.35 | 92.97 |
| Smooth-AP [6] | - | 94.90 | **97.60** | - | **93.30** | **96.40** | - | 91.90 | **96.20** |
| RPTM (ResNet-50) | **82.30** | **95.00** | 96.70 | **79.90** | 92.50 | 96.20 | **78.60** | **92.10** | 95.70 |
| QD-DLP [51] | 76.54 | 72.32 | 92.48 | 74.63 | 70.66 | 88.90 | 68.41 | 64.14 | 83.37 |
| AAVER [18] | - | 74.69 | 93.82 | - | 68.62 | 89.95 | - | 63.54 | 85.64 |
| VehicleNet [48] | - | 83.64 | 96.86 | - | 81.35 | 93.61 | - | 79.46 | 92.04 |
| RPTM (ResNet-101) | **84.80** | **95.50** | **97.40** | **81.20** | **93.30** | **96.50** | **80.50** | **92.90** | **96.30** |

Table 1: Comparison with state-of-the-art methods on VehicleID. RPTM provides the best retrieval results in all three test sets, with notably better performance in the large test set.

| Model | mAP | r = 1 | r = 5 |
|---|---|---|---|
| SPAN [7] | 68.90 | 94.00 | 97.60 |
| PAMTRI [39] | 71.88 | 92.86 | 96.97 |
| PVEN [28] | 79.50 | 95.60 | 98.40 |
| TBE [38] | 79.50 | 96.00 | **98.50** |
| RPTM (ResNet-50) | **79.90** | **96.10** | **98.50** |
| GAN+LSRO* [43] | 64.78 | 88.62 | 94.52 |
| SAVER* [19] | 82.00 | **96.90** | 97.70 |
| RPTM (ResNet-50) * | **86.40** | 96.70 | **98.00** |
| CAL [33] | 74.30 | 95.40 | 97.90 |
| TransReID [14] | 80.60 | **96.80** | – |
| RPTM (ResNet-101) | **80.80** | 96.60 | **98.90** |
| AAVER* [18] | 66.35 | 90.17 | 94.34 |
| DMT* [13] | 82.00 | 96.90 | – |
| VehicleNet* [48] | 83.41 | 96.78 | – |
| Strong Baseline* [17] | 87.10 | 97.00 | – |
| RPTM (ResNet-101) * | **88.00** | **97.30** | **98.40** |

Table 2: Comparison with the state-of-the-art results on the Veri-776 dataset. The $*$ indicates the usage of re-ranking.

| Model | mAP | r = 1 | r = 5 |
|---|---|---|---|
| P2-Net [12] | 73.10 | 86.50 | 93.10 |
| GPS [29] | 78.70 | 88.20 | 95.20 |
| PNL [10] | 79.00 | 89.20 | – |
| SCSN [8] | 79.00 | 91.00 | – |
| RPTM (ResNet-50) | **80.20** | **91.40** | **95.80** |
| Top-DB-Net* [32] | 88.60 | 90.90 | – |
| NFormer* [41] | 83.40 | 89.50 | – |
| st-reID* [40] | **92.70** | **94.50** | **96.80** |
| RPTM (ResNet-50)* | 87.50 | 92.30 | 95.20 |
| PAT* [20] | 78.20 | 88.80 | – |
| LDS* [46] | **91.00** | 92.90 | – |
| RPTM (ResNet-101)* | 89.20 | **93.50** | **96.10** |

Table 3: Comparison on the DukeMTMC benchmark. RPTM provides competitive results even though it is not tuned for person reID. $*$ indicates re-ranking.

average precision (mAP) and Cumulative Matching Characteristics(CMC). For the Veri-776 and DukeMTMC datasets, we also use re-ranking [49], which refines the final rankings by considering the k-reciprocal nearest-neighbours of both the query and retrieved images, effectively improving upon the pairwise distance result that is used to quantify mAP and top-k ranking accuracies. Re-ranking is not adopted for VehicleID because there is often only one true match ID in the gallery set [18]. We split past works based on the complexity of the backbone network, with our results on ResNet-50 and ResNet-101 backbones.

### 6.3. Comparison with State-of-the-art

**VehicleID:** Table 1 shows that RPTM achieves state-of-the-art results on the challenging VehicleID dataset, indicating RPTM's scalability across vehicle datasets. Although not exceeding Smooth-AP[6], table 4 shows a drop in performance by Smooth-AP on Veri-776 and DukeMTMC.

**Veri-776:** As shown in table 2, RPTM surpasses the recent state-of-the-art vehicle reID models. These results are very respectable, especially if we consider the fact that well-performing algorithms like VehicleNet [48] uses supplementary data for training. We also edge out Strong Baseline [17], which uses deeper backbones and larger images. In addition, RPTM's training scheme is very simple, as it only requires gradient descent on a well-defined loss.

**DukeMTMC:**Table 3 shows RPTM achieves competitive results at person reID, despite training parameters tuned to vehicle datasets. With the exception of changing the image size to account for the aspect ratio of input images, no changes were made to the RPTM network or training parameters. These results are respectable for a network whose training parameters are tuned for a different task.

**Discussion:** Table 1, 2 and 3, show that incorporating RPTM to feature learning techniques make them more effective at re-identification. Performance improvements are especially notable on more difficult datasets like VehicleID and harsher evaluation metrics (mAP). These performances are quite remarkable when we take into account that RPTM uses constant training parameters for all three datasets. Most deep-learning algorithms require parameters to be tweaked from dataset to dataset, and RPTM's capability in this respect is an indication that relational aware triplet choice makes the triplet losses better conditioned.

To demonstrate the challenge of maintaining constant training parameters, we trained Smooth-AP [6] on two other datasets, while using the training parameters of Table 1, as shown in Table 4. We also acknowledge the use of Visual Transformers (ViT) in TransReID by He *et al.* [14], demonstrating impressive results, albeit using camera embeddings and viewpoint labelling. Although RPTM uses universal parameters that are compliant to low compute requirements, we still achieve state-of-the-art results compared to transformer-based ReID models. As an additional experiment, using the increased parameter settings defined in TransReID, we further improve our retrieval results, achieving an mAP of **82.5%(w/o re-ranking)** on Veri-776.

| Method | mAP | r = 1 | r = 5 |
|---|---|---|---|
| Smooth-AP (Veri-776) | 79.40 | 91.10 | 94.20 |
| RPTM (Veri-776) | **88.00** | **97.30** | **98.40** |
| Smooth-AP (DukeMTMC) | 65.70 | 79.90 | 88.40 |
| RPTM (DukeMTMC) | **89.20** | **93.50** | **96.10** |

Table 4: Performance of Smooth-AP [6] (ResNet-101 backbone) on Veri-776 and DukeMTMC, with re-ranking.

### 6.4. Ablation Study

**Image Size.** We begin by investigating how image size impacts re-identification. Table 5a shows that the evaluation metrics improve as the image size increases, a finding that is mirrored by many other reID algorithms, which often seek to use the largest possible image. However, we find that performance peaks at (240, 240) on Veri-776 and VehicleID, which validates RPTM's ability to achieve state-of-the-art results at lower resolutions compared to other benchmarks.

**Threshold for Positive Selection** Section 4.2 suggests positive images are chosen using a threshold, $\tau$, which is the mean number of non-zero matching results. We denote this scheme $RPTM_{mean}$ (semi-hard positive mining). There are a number of alternatives. One possibility is to fix $\tau$ on a low number of matches, such as 10. We term this scheme $RPTM_{min}$. The scheme ensures anchor-positive pairs are not near duplicates and corresponds to hard positive mining. The drawback is a vulnerability to occasional matching errors. Another possibility is to set $\tau$ to the largest number of matches that the anchor image has. We term this $RPTM_{max}$. This eliminates any vulnerability to GMS matching errors but sacrifices the positive image's distinctiveness. This corresponds to easy positive mining. Table 5b indicates that $RPTM_{mean}$ has the best performance; hence, it is adopted as our default mining scheme.

| Model | Veri-776 | | VehicleID(small) | |
|---|---|---|---|---|
| | mAP | r=1 | mAP | r=1 |
| $RPTM_{128\times128}$ | 56.5 | 84.5 | 72.5 | 89.0 |
| $RPTM_{160\times160}$ | 74.8 | 92.4 | 80.5 | 91.8 |
| $RPTM_{224\times224}$ | 85.1 | 95.2 | 83.1 | 92.9 |
| $RPTM_{240\times240}$ | **88.0** | **97.3** | **84.8** | **95.5** |

(a) Image size ablation

| Model | Veri-776 | | VehicleID(small) | |
|---|---|---|---|---|
| | mAP | r=1 | mAP | r=1 |
| $RPTM_{min}$ | 86.3 | 95.9 | 82.1 | 93.9 |
| $RPTM_{mean}$ | **88.0** | **97.3** | **84.8** | **95.5** |
| $RPTM_{max}$ | 82.2 | 95.6 | 79.8 | 93.1 |

(b) Thresholding ablation

Table 5: (a) ReID performance with increasing image size. mAP and rank-1 increase with image size until (240, 240), after which performance plateaus. (b) Comparing positive selection thresholds. $RPTM_{min,mean,max}$ correspond to hard positive, semi-hard positive and easy positive-mining.

## 7. Conclusion

In this work, we have shown that respecting natural data groupings within classes can help significantly improve triplet mining, not only facilitating the selection of better anchor-positive pairs but also, consequently, creating a more tractable optimisation procedure that leads to better generalisation. To that end, we introduced Relation Preserving Triplet Mining (RPTM), a triplet-alignment scheme to generate samples wary of the inverse-variability problem, proving that implicitly enforced view consistencies can significantly improve the reID pipeline. We showed how feature matches could be used to develop relation-aware triplet mining, leading to a better conditioned triplet loss, creating feature learners with enhanced training stability. Moreover, we highlighted that RPTM outperforms recent reID models while maintaining constant training parameters across datasets. Finally, we believe our research can be extended to Unsupervised Domain Adaptation for even better scalabilty across reID datasets due to RPTM's implicit ability to align similar features, Anomaly Detection in reID to see how noise affects performance, and also for general deep image retrieval to evaluate RPTM outside re-identification.

## References


[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 1, 4

[2] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399, 2018. 1, 2, 3

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 3

[4] Fabio Bellavia. Sift matching by context exposed. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[5] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4181–4190, 2017. 2, 3, 5, 6

[6] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020. 7, 8

[7] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *European Conference on Computer Vision*, pages 330–346. Springer, 2020. 7

[8] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3300–3310, 2020. 7

[9] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8282–8291, 2019. 1, 2, 7

[10] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2476–2486, June 2022. 3, 6, 7

[11] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 7

[12] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3642–3651, 2019. 7

[13] Shuting He, Hao Luo, Weihua Chen, Miao Zhang, Yuqi Zhang, Fan Wang, Hao Li, and Wei Jiang. Multi-domain learning and identity mining for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 582–583, 2020. 1, 3, 7

[14] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 3, 6, 7, 8

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 3, 4, 5

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[17] Su V Huynh. A strong baseline for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4147–4154, 2021. 7

[18] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6141, 2019. 7

[19] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*, pages 369–386. Springer, 2020. 2, 3, 7

[20] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. 3, 7

[21] Wen-Yan Lin, Siying Liu, Changhao Ren, Ngai-Man Cheung, Hongdong Li, and Yasuyuki Matsushita. Shell theory: A statistical model of reality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[22] Wen-Yan Daniel Lin, Ming-Ming Cheng, Jiangbo Lu, Hongsheng Yang, Minh N Do, and Philip Torr. Bilateral functions for global motion modeling. In *European Conference on Computer Vision*, pages 341–356. Springer, 2014. 3

[23] Hongye Liu, Yonghong Tian, Yaowei Wang, 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. 2, 3, 6

[24] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2021. 2, 3

[25] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016. 1, 2, 3, 6

[26] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 2, 3, 5

[27] Jiayi Ma, Aoxiang Fan, Xingyu Jiang, and Guobao Xiao. Feature matching via motion-consistency driven probabilistic graphical model. *International Journal of Computer Vision*, 130(9):2249–2264, 2022. 3

[28] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 2, 3, 7

[29] Binh X Nguyen, Binh D Nguyen, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Graph-based person signature for person re-identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3492–3501, 2021. 7

[30] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 1, 4

[31] Rodolfo Quispe, Cuiling Lan, Wenjun Zeng, and Helio Pedrini. Attributenet: Attribute enhanced vehicle re-identification. *arXiv preprint arXiv:2102.03898*, 2021. 7

[32] Rodolfo Quispe and Helio Pedrini. Top-db-net: Top dropblock for activation enhancement in person re-identification. *arXiv preprint arXiv:2010.05435*, 2020. 7

[33] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034, 2021. 6, 7

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2, 3, 6

[35] Karsten Roth, Biagio Brattoli, and Bjorn Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8000–8009, 2019. 1, 2

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2, 3

[38] Wei Sun, Guangzhao Dai, Xiaorui Zhang, Xiaozheng He, and Xuan Chen. Tbe-net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 3, 7

[39] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 211–220, 2019. 1, 2, 3, 7

[40] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8933–8940, 2019. 7

[41] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7307, 2022. 6, 7

[42] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2022. 3

[43] Fangyu Wu, Shiyang Yan, Jeremy S Smith, and Bailing Zhang. Joint semi-supervised learning and re-ranking for vehicle re-identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 278–283. IEEE, 2018. 7

[44] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*, pages 126–142. Springer, 2020. 3

[45] Hong Xuan, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2474–2482, 2020. 3

[46] Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to disentangle scenes for person re-identification. *Image and Vision Computing*, 116:104330, 2021. 3, 7

[47] Aihua Zheng, Xianmin Lin, Chenglong Li, Ran He, and Jin Tang. Attributes guided feature learning for vehicle re-identification. *arXiv preprint arXiv:1905.08997*, 2019. 7

[48] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*, 2020. 7

[49] 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. 7

[50] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022. 3, 6

[51] Jianqing Zhu, Huanqiang Zeng, Jingchang Huang, Shengcai Liao, Zhen Lei, Canhui Cai, and Lixin Zheng. Vehicle re-identification using quadruple directional deep learning features. *IEEE Transactions on Intelligent Transportation Systems*, 21(1):410–420, 2019. 7