IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 14, 2020, accepted June 17, 2020, date of publication June 22, 2020, date of current version July 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004092

Unifying Person and Vehicle Re-Identification

DANIEL ORGANISCIAK, DIMITRIOS SAKKOS, EDMOND S. L. HO, NAUMAN ASLAM, (Member, IEEE), AND HUBERT P. H. SHUM, (Senior Member, IEEE)
Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8QH, U.K.

Corresponding author: Hubert P. H. Shum (hubert.shum@northumbria.ac.uk)

This work was supported in part by the Royal Society under Grant IES\R2\181024 and Grant IES\R1\191147, and in part by the Defence and Security Accelerator under Grant DSTLX-1000140725.

ABSTRACT Person and vehicle re-identification (re-ID) are important challenges for the analysis of the burgeoning collection of urban surveillance videos. To efficiently evaluate such videos, which are populated with both vehicles and pedestrians, it would be preferable to have one unified framework with effective performance across both domains. Unfortunately, due to the contrasting composition of humans and vehicles, no architecture has yet been established that can adequately perform both tasks. We release a Person and Vehicle Unified Data Set (PVUD) comprising of both pedestrians and vehicles from popular existing re-ID data sets, in order to better model the data that we would expect to find in the real world. We exploit the generalisation ability of metric learning to propose a re-ID framework that can learn to re-identify humans and vehicles simultaneously. We design our network, *MidTriNet*, to harness the power of mid-level features to develop better representations for the re-ID tasks. We help the system to handle mixed data by appending unification terms with additional hard negative and hard positive mining to MidTriNet. We attain comparable accuracy training on PVUD to training on the comprising data sets separately, supporting the system's generalisation power. To further demonstrate the effectiveness of our framework, we also obtain results better than, or competitive with, the state-of-the-art on each of the Market-1501, CUHK03, VehicleID and VeRi data sets.

INDEX TERMS Person re-identification, vehicle re-identification, deep learning, triplet loss.

I. INTRODUCTION

Re-identification (re-ID) is a core challenge for the computer vision community whereby a detection is required to be matched with another detection of the same object, typically from a different viewpoint. With the increasing volume of large-scale urban surveillance data, re-ID has started to attract a large amount of attention. In the past few years, deep learning techniques have received increased popularity due to significantly improving the performance of both pedestrian [1]–[3] and vehicle [4]–[6] re-ID. In the real-world, person and vehicle re-ID often need to be used together, e.g. when a person of interest boards a vehicle and gets off somewhere else. We would prefer re-ID systems to be able to handle this occurrence for continuous tracking. For this reason, Wei *et al.* [7] attempt to develop an integrated application by using existing person and vehicle re-ID architectures. However, this does not truly unify the tasks, as the system accuracy depends on sub-systems, which is not ioptimal. It requires an additional component to classify between pedestrians and vehicles, which could introduce inaccuracies. We instead train person and vehicle re-ID in a unified manner. This approach allows us to discover underlying principles of re-ID, whereas handling the two systems separately does not allow us to explore this direction.

The challenges of re-identifying vehicles and persons have significant differences. For wide area video surveillance on humans, the same identity viewed from a different pose angle usually looks fairly alike. The shape of the detection remains upright and the colour information, predominantly extracted from articles of clothing, is of a similar pattern. The same condition cannot be satisfied for vehicles. Colour information can become far more distorted in different lighting due to the reflectiveness of the body of a car. The shape information of a car viewed from the front is significantly different than that viewed from a 45° or 90° angle. On the contrary, many high-end vehicle re-ID algorithms use license plate information [4], [8], [9], which is not applicable in the human domain. Moreover, pedestrians are more likely to undergo significant changes over time or viewpoint, e.g. a person's appearance is

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

greatly altered after they put on a coat. In general, changes to vehicles between viewpoints are high variance but predictable whereas the change in a person's colour representation is usually lower variance but prone to much more extreme outliers. We propose that there are *underlying principles of re-ID* that hold regardless of the composition of the object worked upon. Unifying person and vehicle re-ID allows us to explore and discover these underlying principles, precisely because they are so different.

Traditional works split the two tasks and design a network that can specifically target the individual task's respective challenges. However, we argue that this is inadequate. In the real world, urban surveillance videos provide a mixed stream of data, consisting of both vehicles and pedestrians, on which analysis is required. Mixing this data allows us to discover techniques and good practices, which are likely to extend to other re-ID tasks. For example, one may improve person re-ID performance by generating a better feature representation of pedestrians, e.g. by introducing squeeze and excitation modules [10]. This tells us nothing about the framework's actual ability to re-identify an object, despite the accuracy increasing. Our data set helps to solve this issue.

In this paper, we present an approach to unify the two tasks, summarised in Figure 1. We construct the Person and Vehicle Unified Data Set (PVUD) from other popular data sets, which is more representative of raw video surveillance data extracted from the real world. The data set is designed to be challenging and well-balanced, in order to prioritise re-ID systems that excel on both tasks. To the best of our knowledge, this is the first proposed re-identification data set containing both domains. We propose a triplet loss function that can be trained on either person or vehicle data and achieve state-of-the-art performance on each task. As the proposed framework is a form of metric learning, it does not require specific, domain-based design in order to re-identify objects. It is inherent in the framework to separate object classes in the same way that it separates different identities from one another, which makes it ideal to handle the challenges within the database that we design. We exploit information from mid-level layers which are more appropriate for the task of re-ID than the more abstract, final layer representations. In addition, we introduce hard negative and hard positive mining to our framework, with associated unification terms in the loss function, to improve its ability to handle multiple data streams.

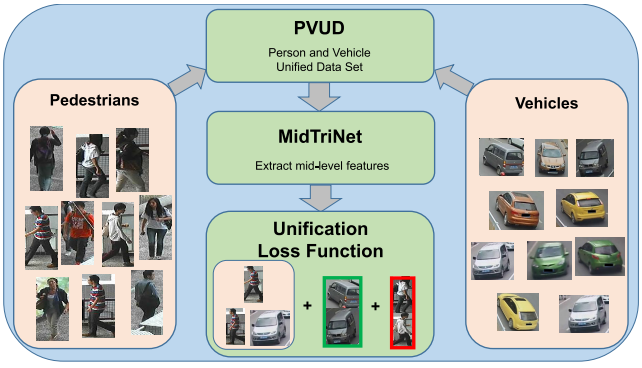


FIGURE 1. We propose a unified framework for pedestrians and vehicles re-identification using a new unified data set, PVUD, which challenges re-ID systems to be capable of handling both tasks simultaneously. Our framework includes MidTriNet to harness the power of mid-level features for re-ID, and a Unification Loss Function to better handle the mixed data stream.

We extensively test our proposed framework, attaining an 88.52% top-1 matching rate on PVUD, and competitive results with state-of-the-art methods on each of its components. The strong performance we obtain on the unified data shows that, contrary to discussion in [5], this is a realistic task on which to focus attention, particularly due to the presence of both pedestrians and vehicles within the vast majority of real-world, surveillance data.

This paper presents the following contributions. To facilitate research in this area, we open our data set and source code at https://github.com/PVUD:

1) *The Person and Vehicle Unified Data Set (PVUD)* - Motivated by the composition of large-scale, surveillance data, we compose a challenging data set containing pedestrian and vehicle information to encourage the re-identification community to pursue the development of frameworks which are applicable to real-world data.
2) *Harnessing information from earlier layers* - We propose *MidTriNet*, a triplet framework which exploits information from mid-level layers. This information is more valuable than features from the deepest layers across both person and vehicle re-identification tasks.
3) *A unified framework* - We append unification terms to the triplet loss to derive a unification triplet loss function. We also introduce term-specific mining algorithms to discover the most important data for the unification terms to focus on.

The rest of the paper is organised as follows: Section 2 contains an overview of related work. Section 3 introduces PVUD and provides details on how it is constructed and how data imbalance is avoided. Section 4 describes MidTriNet, the framework designed to improve re-ID performance and details the construction of the proposed unification terms to handle the unification task. Section 5 shows our experimental results and ablation studies. Section 6 concludes the paper and discusses future directions.

II. RELATED WORK

A. PERSON AND VEHICLE RE-IDENTIFICATION

Historically, popular methods for person re-ID were typically comprised of two components: designing hand-crafted features and learning distance metrics [11], [12]. Most works focused on developing features invariant to variations in light, pose and viewpoint while using conventional distance metrics like the Mahalanobis distance [13], Bhattacharyya distance, and the l_1- and l_2-norms. Research has also been performed on a post-processing technique called re-ranking [14], [15].

We do not include re-ranking in any of our experiments as it does not evaluate the core performance of the framework.

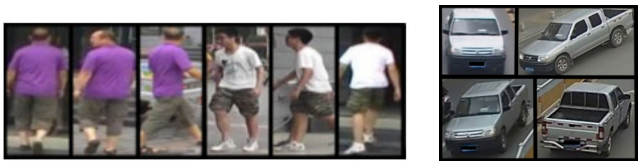
Although similar to person re-ID, vehicle re-ID has received comparatively little attention. This is inconsistent with other computer vision tasks in the vehicle domain, like detection and classification, which have received increased attention in recent years. This lack of popularity can be attributed to the inferiority of large-scale vehicle re-ID data sets compared with their human re-ID counterparts. This is beginning to change as two large-scale data sets, VeRi and VehicleID have more recently been released and have started to attract more research attention.

Research focus in re-identification has shifted towards deep learning methods, which are routinely used to obtain state-of-the-art results over a wide variety of challenges in computer vision and machine learning. Typically, two types of CNN model have been employed to solve the person re-ID task: the classification model that is used across a broad spectrum of computer vision problems [16] and, more commonly for re-ID, the Siamese model which takes multiple images as input, such as pairs [17], [18], triplets [10], [19], and quadruplets [2].

As there is typically more variance between viewpoints within vehicle re-ID compared to person re-ID (Figure 2), more creative methods have been proposed to obtain satisfactory results. Liu *et al.* [20] developed a two branch CNN to learn deep features and the distance metric simultaneously. Liu *et al.* [4] combined hand-crafted features and high-level attributes learned by a CNN with license-plate recognition and spatio-temporal information. Zhou *et al.* [6] trained a model on a toy car data set to in order to infer a multi-view vehicle representation from any input view. Due to the proficiency of deep learning at handling large-scale databases like the one we construct, we elect to utilise it in our experiments.

A recent trend in person re-ID has been to design attention modules that can extract colour information from clothing [21]–[25]. While these do attain strong performance, they are too heavily tailored towards being effective at re-identifying people. The most popular attention mechanisms for person re-ID are part-based systems [26], [27], which split the image into several parts, so the head, torso, legs, and feet are separated from each other. It is clear to see that, although these would perform well on the human proportion of our unified data set, these modules are not able to effectively re-identify vehicles. More generic attention modules also struggle. Colour and shape information remains consistent across different viewpoints in the person domain, but is incredibly inconsistent in the vehicle domain. Therefore, attention modules easily become confused when attempting to tackle these challenges simultaneously. For these reasons, we did not include attention within our unified system.

Recently, there has also been focus on unsupervised re-ID by domain adaptation [28], because traditional supervised re-ID cannot generalise to additional data sets. Fan *et al.* [29] develop a progressive unsupervised learning method that iterates between person clustering and CNN fine-tuning during training. Zhong *et al.* [30] explore three types of invariance that hinder the ability of the re-ID model to generalise to new domains: example invariance, camera invariance and neighbourhood invariance. Deng *et al.* [31] translate images to the target domain using CycleGAN [32] then enforce *domain-dissimilarity* between the translated image and other images in the data set. Ding *et al.* [33] use adaptive exploration to learn discriminative features in the target domain. Whereas unsupervised learning requires generalisation to unlabelled data, PVUD requires generalisation between data types.



(a) (b)

FIGURE 2. Matching people and vehicles contain different challenges: (a) Person shape and colour remains consistent across viewpoints; (b) Vehicle shape and colour changes drastically across viewpoints.

B. TRIPLETS

Triplets have been used extensively in the field of person re-ID. Triplets are generated by pairing query images with one image of the same identity and one with a different identity. Wang *et al.* [34] proposed to use the triplet loss function to learn image similarity. Cheng *et al.* [35] introduced an improved triplet loss function that decreases the distance of similar IDs and increases the distance of dissimilar IDs. Hermans *et al.* [1] proposed *Batch Hard* mining in order to find harder triplets to improve the efficacy of training, however the inter-class variance remains too close. Chen *et al.* [2] train with an additional negative pair and form a quadruplet loss to enlarge inter-class variations while Yuan *et al.* [36] attempt to get the same improvement by adding a similar term without needing to mine an additional image. Wu *et al.* [37] combine triplet loss with identification loss and centre loss. Tian *et al.* [38] mine more informative triplets via their re-weighting strategy.

Triplet-wise training has also been effectively applied for vehicle re-ID. Zhang *et al.* [39] combined the triplet loss with a classification loss and also ensured negative samples in one triplet act as positive samples in another triplet. Bai *et al.* [40] fed groups of images into their triplet network to mitigate inter-class variance and propose a mean-valued triplet loss to enhance learning. Due to the success of the triplet loss across both tasks, it is chosen as a strong backbone to our framework. We additionally mine batch-hard positive and negative examples and introduce respective positive-sample and negative-sample loss functions to further supplement the network in separating identity classes.

C. MID-LEVEL FEATURES

Yu *et al.* [41] concatenate features from earlier ResNet layers with the final layer representation for cross-domain image

TABLE 1. Individual data set characteristics.

Data Set	Train IDs	Test IDs	Images
Market-1501	750	751	32669
CUHK03	1372	95	14297
VeRI	576	200	40395
VehicleID	13134	13133	221763

matching. However, although their approach works well when it uses the triplet loss for sketch-based image retrieval, their approach does not work well with the triplet loss for re-ID so they switch to a classification loss. Zhu *et al.* [42] also fuse mid-level features with final level ones as part of a two stream posed-based and part-based architecture. Zeng *et al.* [43] perform an extensive analysis on the performance of each layer to develop a hierarchical deep learning feature, which fuses features from several earlier layers. Although their method works well with their newly defined metric, their model is heavily engineered for person re-ID, thus would struggle to adapt for vehicle data. Lin *et al.* [44] align mid-level features to boost the performance of unsupervised re-ID. This provides further evidence that mid-level features are an important tool for re-ID and not just specifically useful for supervised, person re-ID.

A number of works have been proposed to select features automatically [45]–[47] for machine learning classification. However, re-ID differs significantly from standard classification tasks in several ways: i) our model exploits metric learning rather than using a classification loss, ii) testing identities in re-ID do not appear in training, whereas all classes appear in training for regular classification tasks, iii) the evaluation metric is based on information retrieval rather than the classification accuracy. Therefore, we do not use any automatic feature selection method.

III. PVUD: PERSON AND VEHICLE UNIFIED DATA SET

There is no publicly available data set for re-identification that contains objects from both person and vehicle classes. As re-ID frameworks are mostly applicable to surveillance data, which generally consists of pedestrians and vehicles, it is imperative for re-ID to be able to handle both streams simultaneously if it is to be applicable to real-world data. Moreover, testing on multiple domains concurrently allows us to be more confident that any adjustments made to the network are beneficial for the re-identification task in general, rather than just for a specific domain. To facilitate the research in this direction, we release a unified data set based on existing ones in the field.

We select the two most popular data sets in each domain - Market-1501 [48] and CUHK03 [17] for pedestrians, and VeRi [49], [50] and VehicleID [20] for vehicles. An overview of the raw data sets, containing the number of identities for training and testing, along with the total number of images can be found in Table 1. The final composition of PVUD can be found in Table 2.

TABLE 2. The composition of PVUD - The number of person and vehicle images are balanced to ensure the data set remains unbiased.

Data Set	Train IDs	Train Images	Test IDs	Test Images
Market-1501	751	12936	200	3486
CUHK03	1372	13176	95	921
VeRI	676	14632	100	2116
VehicleID	1500	12964	500	2085
Person Total	2123	**26112**	295	**4407**
Vehicle Total	2176	**27596**	600	**4085**

A. SOURCE DATA SETS

CUHK03: The CUHK03 data set contains 14297 bounding boxes of 1467 persons. It contains two settings: one with manually annotated bounding boxes and one with automatically detected bounding boxes. We only consider the automatically detected setting as it contains some misplaced bounding boxes making it more challenging and more similar to what we would expect when applying re-identification to real-world tasks.

Market-1501: The Market-1501 data set contains 32668 automatically detected bounding boxes of 1501 individuals.

VeRi: The VeRi data set has 37,781 images of 576 vehicles for training and 11,579 images of 200 vehicles for testing. In order to obey the 'Balance' design principle, we move 100 vehicles from the test set to the train set. We also use a maximum of 20 images per vehicle. Rather than having standalone images from different viewpoints, VeRi contains 'tracks' of vehicles which are extracted as several consecutive frames from a video source. This means that images from all angles are available. Thus, VeRi usually requires image-to-track calculation rather than the standard image-to-image metric that other data sets use. To maintain consistency across data sets, we use the image-to-image testing on PVUD.

VehicleID: The VehicleID data set has 221763 images of 26267 identities. VehicleID contains 'Small', 'Medium', and 'Large' settings for testing. As Market-1501, CUHK03 and VeRi are much smaller, for easier integration, we only take data from the 'Small' set. Contrary to the VeRi data set, VehicleID only contains images from the front and back of the vehicle.

B. DESIGN PRINCIPLES

As discussed in [51], imbalanced data sets are inherently complex. When constructing this data set, it is important to ensure that person and vehicle data are equally balanced to accurately assess how strong a method is at re-identifying humans and vehicles simultaneously. As can be seen in Table 1, if we blindly conjoin the four data sets, there will be much more vehicle data than person data. This will result in the data set being biased towards vehicle re-identification methods, rather than methods which are effectively able to generalise across both tasks.

We lay out the following design principles to ensure the data set is as fair and balanced as possible without sacrificing

difficulty. We provide full details of our constructed data set in Table 2.

Balance: A critical property of the data set is balance between different domains. In this regard, we have two options. We may either equate the number of vehicle IDs with person IDs in the data set, or the number of vehicle images with person images. We find that equating IDs leads to too many vehicle test images, which may result in weak person re-ID frameworks attaining an artificially high result. Balancing the number of images will facilitate a more challenging data set that better represents the real-world. Our studies shows that balancing IDs gives an mAP of 84.83%, whereas balancing images reduces the mAP to 77.51%.

Size: A data set with both pedestrians and vehicles is already challenging. We wish to take this challenge further. A larger testing set means more negative images to compare against, i.e. more likelihood to find a negative with a high similarity score, which makes testing more challenging. We also want to ensure that the data set is large enough for deep learning models, which demonstrate much greater efficacy at handling large-scale, real-world surveillance data. As we release the data set to motivate the re-ID community to design frameworks which can handle real-world data, it is imperative that it is suitable for deep learning frameworks. For these reasons, we select the design principle of maximising the size of the data set.

Random Sampling: In the interests of fairness, we randomly sample from the four comprising data sets rather than hand selecting examples.

1) DESIGN CHOICES FOR VeRi:

VeRi requires image-to-track re-identification rather than image-to-image. Each track is composed of several consecutive frames from a video. We choose to include the entire track in order to more accurately model real-world surveillance videos. Many of the images used for training from VeRi are therefore similar to one another, so there is less effective information. This has two main consequences: (1) as VehicleID contains more effective training data, any framework must be capable of transferring knowledge between the data sets for accurate vehicle re-identification results, (2) models have to be robust against overfitting as VeRi training data can be very similar.

IV. A UNIFIED FRAMEWORK FOR PERSON AND VEHICLE RE-IDENTIFICATION

In this section, we will detail the implementation of MidTriNet and provide motivation for the design of our unification terms.

We present our architecture in Figure 3. The input batch is generated by taking four images of P identities, which are processed by MidTriNet and mapped into the embedding space. The distance matrix between all feature vectors is then calculated via a Euclidean distance and the hardest samples are mined. These samples are then fed into our novel loss function.

A. TRADITIONAL TRIPLET LOSS FUNCTION

The triplet loss function has seen extensive use for both person and vehicle re-ID due to its proven ability to attain state-of-the-art results and its efficacy in being able to handle difficult examples through specifically mining such examples during training. Traditional triplet models take three images as input: one query image, one image with the same identity as the query (positive), and one image with a different identity to the query (negative). The margin α is enforced to ensure distance between positive and negative pairs.

We denote a triplet, $t = (x, x^+, x^-)$, where x is the query image, x^+ is a positive image, and x^- is a negative image. The triplet loss function is formulated as follows:

$$\mathcal{L}_{trp} = \sum_{t \in \mathcal{T}} \max((||f(x) - f(x^+)||_2 - ||f(x) - f(x^-)||_2 + \alpha), 0), \quad (1)$$

where $f(x)$ is the feature vector of image x and $\mathcal{T}$ is the set of mined triplets.

B. MidTriNet

Contrary to most deep learning classification tasks, mid-level layers have been shown to have similar importance as higher-level layers for constructing effective feature embeddings for re-ID [41], [43]. Re-identification relies on matching human-understandable information such as colour of clothes, and features are required to be viewpoint invariant. Mid-level information such as colours and textures, which are robust to viewpoint changes, are extremely useful information to discern whether an individual in the gallery set is the same as that in the query image. The very abstract features in the final layers are therefore not necessarily optimal for comparison, particularly within a triplet loss framework which attempts to differentiate between identities by directly comparing the feature representations of each image. To exploit the important information generated by the mid-level layers, we develop *MidTriNet*, which contains two major design choices throughout our experiments. These choices are supported by the ablation studies on the stride length (Table 10) and ResNet blocks (Table 11) provided in Section V-D.

Layer Removal: We remove the final two `conv5` blocks to strike a balance between the powerful representation ability that is characteristic of conv5 blocks and the re-identification task-specific efficacy of mid-level layers. Not only does this improve the feature embedding for re-ID, but also helps to protect the model against overfitting, which is extremely important for this data set as described in Section III-B1. Through our extensive experimentation, we find that removing the final two ResNet blocks works best.

Stride: To best exploit mid-level features for re-ID, we reduce the stride length in the `conv4` block from 2 to 1. This ensures that we have more informative feature maps at the important mid-level layers, which enriches the output of those layers to improve the final feature representation. The `conv5` stride length is typically 1 for this reason. Reducing

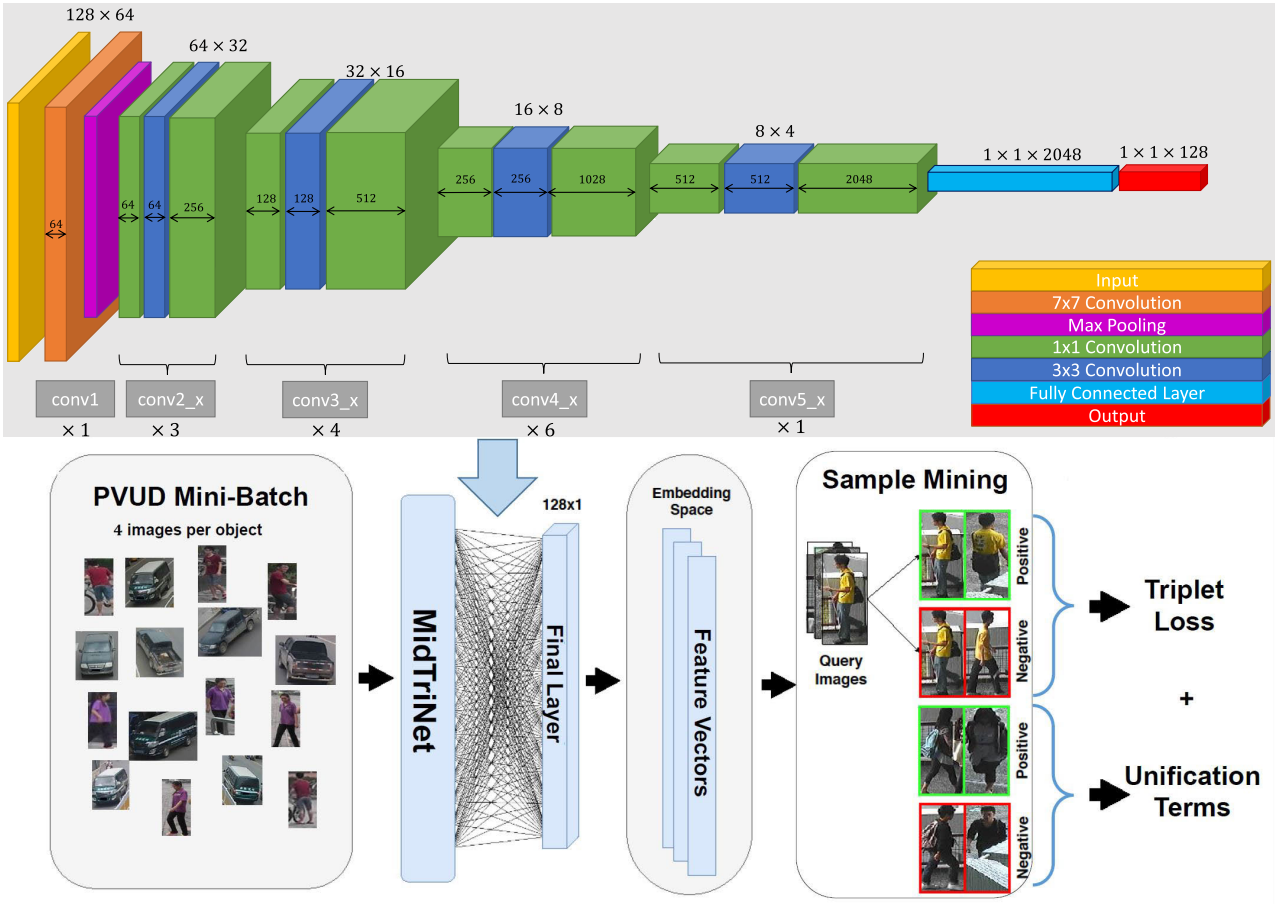


FIGURE 3. An overview of the architecture with unification terms. Each batch of images is processed with MidTriNet. We take the final layer of the network as the embedding space. We design unification terms specifically to make the network more robust against the mixed data that is present in PVUD and append them to the triplet loss function. Finally, we mine hard triplets, positive pairs and negative pairs to feed into our unification loss function.

the stride length of `conv4` to 1 allows us to focus on those features in the same way. This allows us to better compare the similarity between two images which benefits the network at all stages of training and testing.

C. UNIFICATION TERMS

The triplet loss aims to simultaneously pull images of the same identity closer together whilst pushing away an image of a different identity. This can be difficult when dealing with unified data. Different data sets have different characteristics (camera intrinsics, lighting conditions, etc.), so the feature representations are likely to be further apart from one another on average. This means that it is more difficult to find hard negatives (and thus hard triplets), so the model risks being unable to handle difficult situations when it comes to testing.

To counteract this, we mine the hardest negatives and positives across the batch. We design unification terms to separate hard negatives and compress hard positives, and append them to the loss function.

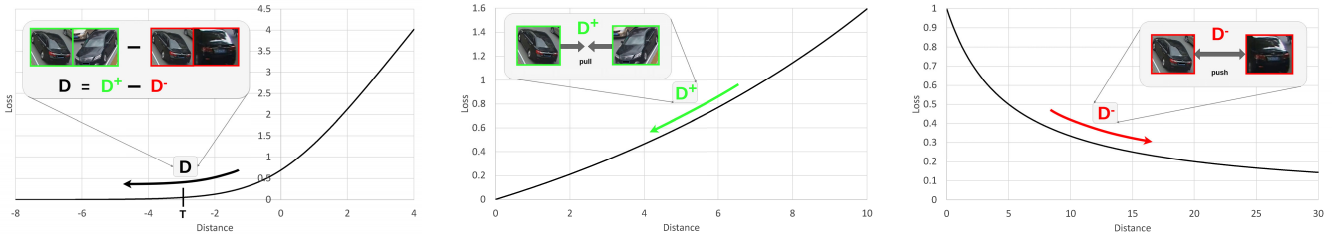
1) LOSS FUNCTION

Let $\mathcal{T}$ be the set of triplets, where $t = (x_0, x_0^+, x_0^-) \in \mathcal{T}$ is a triplet comprising of a query image x_0, a positive image x_0^+ from the same identity as x_0 and a negative image x_0^- from a different identity. Let $f(x)$ be the feature vector of an image x. $\mathcal{H}^+$ is the set of the hardest positive pairs $h^+ = (x_1, x_1^+)$ with lowest similarity and $\mathcal{H}^-$ is the set of negative pairs $h^- = (x_2, x_2^-)$ with highest similarity (likewise, the hardest negative pairs). We set $\mathcal{H}^+ = \mathcal{H}^- = \mathcal{T} = 4P$ where P is the number of identities in each batch. Throughout this section, we refer to D as the distance between two feature representations.

The first term we use is a modified triplet loss function presented in [1]. Their analysis shows that replacing the traditional hard margin α from (1) with the softplus function $\text{softplus}(D) = \log(1+e^D)$ is beneficial. The softplus function is shown in Figure 4(a). Overall, we have

$$\mathcal{L}_t = \sum_{t \in \mathcal{T}} \text{softplus}(||f(x_0) - f(x_0^+)||_2 - ||f(x_0) - f(x_0^-)||_2). \tag{2}$$

The second term focuses on pulling together the positive pairs. We design the function $\Psi(D) = \psi^D - 1$, where $\psi > 1$ is a constant, in order to heavily punish large distances between positive pairs as seen in Figure 4(b). This forces the network to pull images from the same class together during training in order to keep the loss minimal. In our experiments



(a) softplus(D) (b) $\Psi(D^+) = \psi^{D^+} - 1$ (c) $\Phi(D^-) = \frac{\phi}{\frac{\phi}{10}+D^-}$

FIGURE 4. Visualisations of the softplus, Ψ and Φ functions used to calculate the overall loss function found in Equation (5).

we use $\psi = 1.1$. The positive unification term is written as:

$$\mathcal{L}_p = \sum_{h^+ \in \mathcal{H}^+} \psi^{(||f(x_1)-f(x_1^+)||_2)} - 1. \tag{3}$$

Note that this is especially important in the vehicle domain. As discussed in Section I, vehicle shape can change drastically in different viewpoints. One of the reasons why our network is so robust to this shape deformation is that we force the model to learn from additional hard positives, of which a large proportion will typically be vehicles in a very different pose.

The third term works similarly but aims to push negative images away from each other. We adopt $\Phi(D) = \frac{\phi}{\frac{\phi}{10}+D}$ to punish negative pairs with small distances and reward pairs with large distances as seen in Figure 4 (c). Throughout our experiments, we set $\phi = 30$. The negative unification term is written as:

$$\mathcal{L}_n = \sum_{h^- \in \mathcal{H}^-} \frac{\phi}{\frac{\phi}{10} + \left(||f(x_2) - f(x_2^-)||_2\right)}. \tag{4}$$

From Equations (2), (3) and (4), we obtain our unification loss function:

$$\mathcal{L}_\mathcal{U} = \alpha_t \mathcal{L}_t + \alpha_p \mathcal{L}_p + \alpha_n \mathcal{L}_n, \tag{5}$$

where α_t, α_p and α_n, are weights for their relative losses. Empirically, we found that setting $\alpha_t = 0.05$, $\alpha_p = 0.5$ and $\alpha_n = 0.5$ performs best.

2) SAMPLE MINING

One of the most important elements of building a framework which utilises a triplet loss function is effective mining. We require it to effectively match vehicles with significant distortions in shape, thus it is imperative that the model is trained on the most difficult samples available. Likewise, we wish for it to be able to handle outlier scenarios, e.g. where a person is wearing a bag, thus having a highly different appearance in different viewpoints. To challenge the framework to be able to handle these tough cases, sufficiently difficult triplets need to be mined. However, if the model is only trained on the most difficult triplets, it will not be representative of the entire data set and could struggle on easier examples.

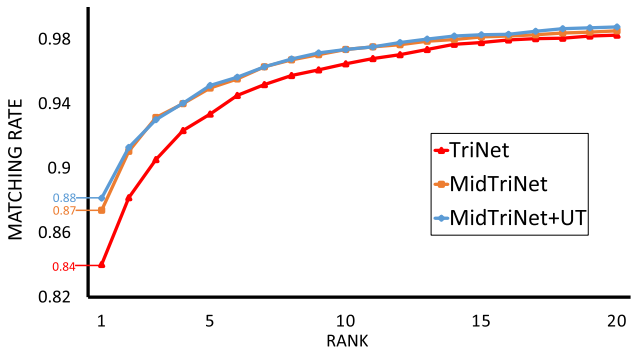


FIGURE 5. CMC curves for tested models on PVUD.

Let p be the identity of the image $x_{p,i}$ in the batch, B, and let $f(x_{p,i})$ be its feature vector, where $p = 1, \ldots, P$ and $i = 1, \ldots, 4$. Each query image $x_{p,i}$ is paired with its hardest positive image x^+ and hardest negative image x^-, where:

$$x^+ = \max_{x_{q,j} \in B} \left(||(f(x_{p,i}) - f(x_{q,j})||_2\right), \quad \text{such that } p = q, \tag{6}$$

$$x^- = \min_{x_{q,j} \in B} \left(||(f(x_{p,i}) - f(x_{q,j})||_2\right), \quad \text{such that } p \neq q. \tag{7}$$

Together, we obtain the triplet $t_{p,i} = (x_{p,i}, x^+, x^-)$ and these form the set of triplets, $\mathcal{T}$, where $\mathcal{T} = 4P$.

In a similar manner, we scan across the entire distance matrix to find the set of hardest positive pairs, $\mathcal{H}^+$, and the set of hardest negative pairs, $\mathcal{H}^-$, with $\mathcal{H}^+ = \mathcal{H}^- = 4P$.

V. EXPERIMENTAL RESULTS

In this section, the proposed architectures are exhaustively evaluated on the most popular modern data sets for person re-ID and vehicle re-ID. Our data set and source code can be found at https://github.com/PVUD

We give results for *MidTriNet* and *MidTriNet+UT* (Unification Terms). MidTriNet is the baseline TriNet model, with the addition of the design choices described in Section IV-B to harness mid-level features: reduction of the stride length in the `conv4` block, and removal of the final two `conv5` blocks. MidTriNet+UT includes the additional terms from Section IV-C which specifically help the system to handle the mixed data in our unified data set.

In addition to PVUD, we test our framework on individual data sets. For person re-ID, the Market-1501 [48] and CUHK03 [17] data sets are selected. Meanwhile for vehicle re-ID, we use the widely used VeRi [49], [50] and VehicleID [20] data sets.

A. EVALUATION PROTOCOL

For PVUD and person re-ID, we use the standard 'mean average precision' (mAP) and 'rank-1' metrics to evaluate our framework against the state-of-the-art methods. As many vehicle re-ID methods do not report mAP, we additionally report our 'rank-5' score for better comparison. For details on the individual data sets, see Section III-A.

The rank-x matching rate is defined as the percentage of query images with a correct match within the highest x ranks. The precision, P_x, at rank x is calculated via

$$P_x = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}. \tag{8}$$

The average precision for a given query, q, is calculated by taking the average of the precision scores at each *true* positive in the ranking list:

$$\text{AP}_q = \frac{1}{N}\sum_{i=1}^{N} P_i^+, \tag{9}$$

where N is the number of true positives in the gallery and P_i^+ denotes the precision at the i-th true positive in the ranking list. The mAP is then calculated via

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q} \text{AP}_q\,. \tag{10}$$

PVUD: For PVUD, we take subsets of the standard train/query/gallery splits of each of the four individual data sets. These are procured by following standard procedures as described below.

Note: as described in Section III-B, the training set of PVUD contains some instances from the VeRi test set. For fairness, we exclude these when we train on PVUD and test on VeRi to analyse the robustness of our system.

Market-1501: For the Market-1501 data set, either the single query or multiple query setting can be used. We evaluate on the single query setting as it is more challenging and applicable to real world scenarios.

CUHK03: Recently, a new train/test split has gained popularity for the CUHK03 data set. However, as discussed in Section III-B, to keep our unified data set balanced, we were required to use a similar train/test split as the initial CUHK03 split. For this reason, we conduct our tests on the original split.

VeRi: The VeRi data set differs from other re-ID data sets as it maps temporally close images in the gallery onto tracks. The re-identification is computed from the query to the entire track (image-to-track) rather than just to gallery images (image-to-image). We follow the standard procedure for computing the similarity between a query image and a track, by calculating the similarity between the query image and all images on the track and then to take the maximum.

VehicleID: For the VehicleID data set, we follow the standard procedure as described in [20]. Given an identity i with N_i images in the test set, $\max(6, N_i - 1)$ images of identity i are placed into the gallery set, and the remaining images are put into the query set.

TABLE 3. Results on our unified data set PVUD - The unification terms (UT) improve performance when the data is comprised from different domains due to the diversity of the data.

Method	mAP	rank-1
ResNet [52]	69.1	79.8
HA-CNN [21]	47.1	56.1
PCB [27]	64.7	73.3
TriNet [1]	74.5	85.2
MidTriNet (Ours)	76.6	87.4
MidTriNet + UT (Ours)	**77.5**	**88.5**

B. COMPARISON WITH BASELINES

PVUD: Our results on PVUD can be found in Table 3. It can be observed that the unification terms introduced in Section IV-C boost the mAP by 0.92% and increase the top-1 matching rate by 1.13%. Moreover, we attain significant improvement over the standard TriNet on both networks. We also compare with ResNet where the triplet loss is replaced with cross-entropy loss, and two other state-of-the-art re-ID methods: Parts-based Convolutional Baseline (PCB) and Harmonious Attention Network (HA-CNN) [21] [27]. PCB is designed to specifically handle person data, and was trained with a ResNet-50 backbone for fair comparison. HA-CNN uses an attention module that learns from the data that is provided, so it could be applied to vehicles as easily as it is to persons.

Both methods provide a significant reduction of performance compared to standard ResNet, implying that attention diminishes performance in both cases. For PCB, this is because the part models cannot adequately handle vehicle data as discussed in Section II-A. PCB encodes better feature representations of pedestrians but cannot generalise to vehicle data that it was not designed to handle. Therefore, PCB obtains moderately better performance at re-identifying pedestrians within PVUD but drastically worse performance at re-identifying vehicles. This results in a net performance decrease of 4.4% when applied to the overall data set. In contrast, HA-CNN has demonstrated strong performance when trained on each domain individually. However, the attention mechanism of HA-CNN becomes confused when trying to handle two drastically different data types simultaneously. This results in a sharp decrease in performance. We discuss the impact of the individual unification terms in more detail in Section V-D.

We compare against the baseline TriNet with two separate settings in Table 4. First, we train and test on individual data sets in the standard way. Secondly, we train all models on

TABLE 4. Comparison on individual data sets when trained with PVUD - MidTriNet significantly outperforms TriNet and the unification terms improve performance when the training data is comprised of both vehicles and pedestrians.

Data Set	Market-1501		CUHK03		VeRi		VehicleID	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
TriNet	65.95	82.39	83.24	84.69	65.63	76.96	66.15	80.05
MidTriNet (Ours)	68.57	84.17	**87.25**	89.29	68.08	81.20	69.55	84.65
MidTriNet + UT (Ours)	**68.90**	**84.20**	86.77	**90.31**	**69.25**	**82.32**	**70.56**	**85.33**

TABLE 5. Comparison on CUHK03 - MidTriNet outperforms state-of-the-art, while the unification terms have less effect in this less diverse, saturated data set. * TriNet did not converge on CUHK03.

Method	mAP	rank-1
DHSL [12]	-	60.0
JLML [53]	-	80.6
MLFN [54]	-	82.8
JLDE [36]	-	85.30
JAL [55]	-	88.4
CRAFT [56]	72.4	84.3
DPFL [3]	78.1	82.0
SAN [57]	82.2	84.3
SVDNet [58]	84.8	81.8
LSRO [59]	87.4	84.6
TriNet [1]	*	*
MidTriNet (Ours)	**88.5**	**91.0**
MidTriNet+UT (Ours)	87.2	88.5

TABLE 6. Comparison on Market-1501 - While our method has slightly lower mAP than MGCAM or HA-CNN, it can work complementary with them.

Method	mAP	rank-1
JLDE [36]	67.7	85.2
PAN [24]	69.3	86.7
SAN [57]	70.1	85.9
IC-TL [37]	70.1	86.6
ADV [38]	70.4	86.8
RE [60]	71.3	87.1
DaF [61]	72.4	82.3
DPFL [3]	73.1	88.9
SE+DWE [10]	74.2	88.0
MGCAM [62]	74.3	83.8
HA-CNN [21]	**75.7**	**91.2**
TriNet [1]	69.1	84.9
MidTriNet (Ours)	73.4	87.8
MidTriNet+UT (Ours)	74.0	88.9

PVUD and test on the individual data sets to analyse how robust models are at handling data from different sources. This is very challenging. The model is required to be able to use information from training on one data set to test on another. Despite this, we see very little performance loss on any of the data sets. Both of our models considerably outperform the baseline on both settings for all data sets.

Further, we can see that the unification terms benefit performance across all data sets when the models are trained on PVUD. This demonstrates that the additional sample mining for the unification terms helps to create a much stronger model for handling mixed data.

TABLE 7. Comparison on VeRi - Our method outperforms state-of-the-art and unification terms improve the rank-1 matching rate due to the diversity of the data set.

Method	rank-1	rank-5
XVGAN [6]	60.20	77.03
NuFACT [4]	76.76	92.79
VAMI [63]	77.03	90.83
PROVID [4]	81.56	95.11
HA-CNN [21]	83.00	92.41
TriNet [1]	83.25	95.23
MidTriNet (Ours)	88.56	**96.90**
MidTriNet + UT (Ours)	**89.15**	93.74

C. COMPARISON WITH STATE-OF-THE-ARTS

CUHK03: Our results on the CUHK03 data set are presented in Table 5. In particular, MidTriNet outperforms the mAP of the popular LSRO [59] by 1.1% on the detected data and exceeds the second best rank-1 score by 2.6%.

Market-1501: Our results on the Market-1501 data set can be found in Table 6. Our results are competitive with the state-of-the-arts, attaining an mAP of just 1.7% less than HA-CNN and achieving a 4.9% improvement on the original TriNet.

VeRi: We present our results on the VeRi data set in Table 7. Our method clearly outperforms the state-of-the-art at the vehicle re-ID task. We obtain a rank-1 score of almost 6% higher than the next best result.

VehicleID: Our results on the VehicleID data set can be found in Table 8. Our MidTriNet model consistently attains the highest mAP and rank-1 results across all three settings. Our methods considerably outperform state-of-the-arts, achieving 4% rank-1 improvement over the best method not to use a triplet loss.

On all five data sets presented in Tables 3 - 8, MidTriNet significantly outperforms the baseline TriNet model. This consistent performance enhancement across domains provides conclusive experimental evidence that harnessing mid-level information is an underlying principle of re-ID.

D. ABLATION STUDIES

In this section we present our ablation studies to demonstrate the benefits of a) mid-level information for re-ID, b) unification terms. All experiments in this section are performed on PVUD. We include confidence intervals at a 95% confidence level to demonstrate the significance of our design choices. We calculate these confidence intervals using the guidance for information retrieval tasks in [66].

TABLE 8. Comparison on VehicleID - Our method outperforms state-of-the-arts, while UT has minimal effect on this saturated data set.

	Test Size = 800		Test Size = 1600		Test Size = 2400	
Method	rank-1	rank-5	rank-1	rank-5	rank-1	rank-5
VAMI [63]	63.1	83.3	52.9	75.1	47.3	70.3
BIER [64]	82.6	90.6	79.3	88.3	76.0	86.4
DREML [65]	88.5	94.8	87.2	94.2	83.1	92.4
DRDL [20]	49.0	73.5	42.8	66.8	38.2	61.6
TriNet [1]	91.5	**97.9**	89.3	96.1	85.5	94.2
MidTriNet (Ours)	**92.5**	97.6	**90.6**	**96.6**	**86.5**	94.6
MidTriNet+UT (Ours)	91.7	97.7	90.1	96.4	86.1	**94.8**

TABLE 9. Ablation study on batch size - We see that the larger the batch, the stronger our performance.

Number of Identities	mAP	rank-1
18	72.10 ± 0.19	83.89 ± 0.30
32	76.78 ± 0.16	87.55 ± 0.24
36	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.22}$

TABLE 10. Ablation study on stride lengths of the `conv4` block - We see that reducing the stride length creates more informative mid-level features which boosts performance.

Block Strides	mAP	rank-1
2, 2, 2, 1	75.79 ± 0.17	86.77 ± 0.25
2, 2, 1, 1	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.22}$

TABLE 11. Ablation study on removing ResNet blocks - The final composition of MidTriNet (3,4,6,1), with two `conv5` blocks removed, significantly outperforms the others, validating our hypothesis that mid-level features perform best.

ResNet Blocks	mAP	rank-1
3, 4, 4, 1	76.59 ± 0.16	87.55 ± 0.30
3, 4, 6, 1	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.24}$
3, 4, 6, 3	76.11 ± 0.17	86.72 ± 0.25

Table 9 shows our ablation studies on the batch size. In particular, we find that larger batch sizes attain greater re-identification performance. This is because we can mine harder triplets, negatives and positives for our loss function so the framework learns more efficiently.

Our results with different stride sizes are presented in Table 10. We see that reducing the stride from 2 to 1 in the third ResNet block boosts both the mAP and rank-1 performance by over 1.7%. This shows that the more informative mid-level feature maps are very important in boosting re-ID performance.

Table 11 shows that removing the final two `conv5` blocks significantly boosts performance. This supports the notion that mid-level features are more suitable than final level features for a triplet loss re-ID framework.

We perform ablation studies on our unification terms in Table 12. We see that both the positive and the negative term contribute to the overall score. When the negative term is excluded, the positive term provides a performance improvement of 0.43% on the mAP metric. Likewise, when the positive term is excluded, the mAP is 0.34% higher than the standard MidTriNet. We arrived a the highest performance with the unification terms weighted equally and very large compared to the standard triplet loss term.

TABLE 12. Ablation study on unification terms when trained on PVUD - Both unification terms are effective and they have a complementary effect when used together.

α_t	α_p	α_n	mAP	rank-1
1	0	0	76.59 ± 0.16	87.39 ± 0.24
1	1	0	77.02 ± 0.15	88.03 ± 0.23
1	0	1	76.93 ± 0.16	87.60 ± 0.24
0.05	0.5	0.5	$\mathbf{77.51 \pm 0.15}$	$\mathbf{88.52 \pm 0.22}$

VI. CONCLUSION

In this paper, we have unified person and vehicle re-identification. Firstly, by constructing a balanced, challenging data set by combining the two most popular data sets in each domain; secondly, by designing a triplet loss framework that beats or is competitive with state-of-the-art methods on both tasks and also attains high performance on our newly designed data set. We propose MidTriNet, to demonstrate that utilising mid-level features is an underlying principle of re-ID. Our design to exploit them boosts performance across all data sets. Finally, we show the value of our data set by appending terms to the loss function, specifically to improve the accuracy when the data is merged.

The unification terms presented in this paper have been demonstrated to benefit the network, specifically when targeting mixed data streams. As a future work, we wish to explore this potential by deriving more complex mechanisms which target multi-domain data. Many popular re-ID approaches currently make use of attention modules and part-based representations to learn better feature representations by giving less weight to background pixels. One planned future work is to incorporate these ideas into our framework to further improve the network's robustness when dealing with data from different domains.

Transfer learning, whereby a model is trained on one data set and tested on another, is a largely under-researched area in re-ID. Our data set forces models to be able to use data from one data set to help the training of another set in the same domain. It also attains strong performance when the model is trained on the unified data set and tested on individual data

sets, suffering little performance loss. We wish to explore transfer learning in the future by training state-of-the-art methods on this data set and testing on a data set that is not a component of the unified one, such a DukeMTMC-reID. We hypothesise that the model is forced to be more robust and is less likely to overfit on our unified data, thus should perform stronger when applied to transfer learning.

We also plan to apply this for real world problems as future work, extending the architecture so that vehicle re-ID can be used to identify the area that a person of interest has travelled to via vehicle, which would significantly narrow down the number of cameras to evaluate, providing us with a much greater likelihood to re-identify the individual.

REFERENCES

[1] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[2] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.

[3] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2590–2600.

[4] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[5] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2017, pp. 1–12.

[6] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.

[7] L. Wei, X. Liu, J. Li, and S. Zhang, "VP-ReID: Vehicle and person re-identification system," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*. New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 501–504.

[8] M. A. Lalimi, S. Ghofrani, and D. McLernon, "A vehicle license plate detection method using region and edge based methods," *Comput. Electr. Eng.*, vol. 39, no. 3, pp. 834–845, Apr. 2013.

[9] W. Liu, X. Liu, H. Ma, and P. Cheng, "Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1618–1626.

[10] D. Organisciak, C. Riachy, N. Aslam, and H. P. H. Shum, "Triplet loss with channel attention for person re-identification," *J. WSCG*, vol. 27, no. 2, pp. 161–169, 2019.

[11] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, Jun. 2011, pp. 649–656.

[12] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3183–3193, Nov. 2018.

[13] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 247–267.

[14] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[15] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1056–1069, Mar. 2016.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

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

[18] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 9905, 2016, pp. 3–20.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

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

[21] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[22] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, Feb. 2019.

[23] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2119–2128.

[24] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[25] M. Jiang, Y. Yuan, and Q. Wang, "Self-attention learning for person re-identification," in *Proc. BMVC*, 2018, p. 204.

[26] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[27] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[28] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2110–2118.

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

[30] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.

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

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[33] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.

[34] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[35] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[36] C. Yuan, J. Guo, P. Feng, Z. Zhao, C. Xu, T. Wang, G. Choe, and K. Duan, "A jointly learned deep embedding for person re-identification," *Neurocomputing*, vol. 330, pp. 127–137, Feb. 2019.

[37] D. Wu, S.-J. Zheng, W.-Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69–75, Jan. 2019.

[38] H. Tian, X. Zhang, L. Lan, and Z. Luo, "Person re-identification via adaptive verification loss," *Neurocomputing*, vol. 359, pp. 93–101, Sep. 2019.

[39] Y. Zhang, D. Liu, and Z.-J. Zha, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1386–1391.

[40] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[41] Q. Yu, X. Ching, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The Devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," 2018, *arXiv:1711.08106*. [Online]. Available: https://arxiv.org/abs/1711.08106

[42] S. Zhu, X. Gong, Z. Kuang, and J. Du, "Partial person re-identification with two-stream network and reconstruction," *Neurocomputing*, vol. 398, pp. 453–459, Jul. 2020.

[43] M. Zeng, C. Tian, and Z. Wu, "Person re-identification with hierarchical deep learning feature and efficient XQDA metric," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1838–1846.
[44] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," 2018, *arXiv:1807.01440*. [Online]. Available: https://arxiv.org/abs/1807.01440
[45] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognit.*, vol. 64, pp. 141–158, Apr. 2017.
[46] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[47] J. Nalepa, G. Mrukwa, and M. Kawulok, "Evolvable deep features," in *Proc. Int. Conf. Appl. Evol. Comput.* Cham, Switzerland: Springer, 2018, pp. 497–505.
[48] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[49] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[50] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 869–884.
[51] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385
[53] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2194–2200.
[54] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.
[55] W. Xiang, J. Huang, X. Qi, X. Hua, and L. Zhang, "Homocentric hypersphere feature embedding for person re-identification," 2018, *arXiv:1804.08866*. [Online]. Available: https://arxiv.org/abs/1804.08866
[56] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.
[57] C. Shen, G.-J. Qi, R. Jiang, Z. Jin, H. Yong, Y. Chen, and X.-S. Hua, "Sharp attention network via adaptive sampling for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3016–3027, Oct. 2019.
[58] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.
[59] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.
[60] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: https://arxiv.org/abs/1708.04896
[61] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," 2017, *arXiv:1708.04169*. [Online]. Available: https://arxiv.org/abs/1708.04169
[62] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.
[63] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.
[64] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "BIER—Boosting independent embeddings robustly," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5189–5198.
[65] H. Xuan, R. Souvenir, and R. Pless, "Deep randomized ensembles for metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 723–734.
[66] I. Soboroff, "Computing confidence intervals for common ir measures," in *Proc. EVIA NTCIR*, 2014, pp. 25–28.

DANIEL ORGANISCIAK received the B.Sc. degree in mathematics and the M.Math. degree in mathematical logic from the University of Birmingham, U.K. He is currently pursuing the Ph.D. degree with Northumbria University, U.K. His research interests include computer vision and deep learning.

DIMITRIOS SAKKOS received the B.Sc. degree in mathematics from the Aristotle University of Thessaloniki, Greece, in 2012, and the M.Sc. degree in computer science from the University of Birmingham, U.K., in 2013. He is currently pursuing the Ph.D. degree with Northumbria University, Newcastle upon Tyne, U.K. His research interest includes image and video segmentation.

EDMOND S. L. HO received the B.Sc. degree in computer science from Hong Kong Baptist University, in 2003, and the M.Phil. degree in computer science and the Ph.D. degree in informatics from The University of Edinburgh, Scotland, in 2006 and 2011, respectively. He was a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, from 2011 to 2016. He is currently a Senior Lecturer with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, U.K. His current research interests include computer graphics and animation, computer vision, machine learning, and robotics.

NAUMAN ASLAM (Member, IEEE) received the Ph.D. degree in engineering mathematics from Dalhousie University, Halifax, NS, Canada, in 2008. He was an Assistant Professor with Dalhousie University, from 2008 to 2011. He joined Northumbria University, in August 2011. He is currently a Reader with the Department of Computer Science and Digital Technologies. He is also an Adjunct Assistant Professor with Dalhousie University.

HUBERT P. H. SHUM (Senior Member, IEEE) received the Ph.D. degree from the School of Informatics, The University of Edinburgh, U.K. He was a Senior Lecturer with Northumbria University, U.K., a Lecturer with the University of Worcester, U.K., a Postdoctoral Researcher with RIKEN, Japan, and a Research Assistant with the City University of Hong Kong. He is currently an Associate Professor (Reader) with Northumbria University and the Director of the Research and Innovation of the Computer and Information Sciences Department. His research interests include computer graphics, computer vision, motion analysis, and machine learning.