DEEP VISUAL RE-IDENTIFICATION WITH CONFIDENCE

George Adaimi
VITA, EPFL
Switzerland
george.adaimi@epfl.ch

Sven Kreiss
VITA, EPFL
Switzerland
sven.kreiss@epfl.ch

Alexandre Alahi
VITA, EPFL
Switzerland
alexandre.alahi@epfl.ch

ABSTRACT

Transportation systems often rely on understanding the flow of vehicles or pedestrian. From traffic monitoring at the city scale, to commuters in train terminals, recent progress in sensing technology make it possible to use cameras to better understand the demand, *i.e.*, better track moving agents (*e.g.*, vehicles and pedestrians). Whether the cameras are mounted on drones, vehicles, or fixed in the built environments, they inevitably remain scatter. We need to develop the technology to re-identify the same agents across images captured from non-overlapping field-of-views, referred to as the visual re-identification task. State-of-the-art methods learn a neural network based representation trained with the cross-entropy loss function. We argue that such loss function is not suited for the visual re-identification task hence propose to model confidence in the representation learning framework. We show the impact of our confidence-based learning framework with three methods: label smoothing, confidence penalty, and deep variational information bottleneck. They all show a boost in performance validating our claim. Our contribution is generic to any agent of interest, *i.e.*, vehicles or pedestrians, and outperform highly specialized state-of-the-art methods across 5 datasets. The source code and models are shared towards an open science mission.

***K*eywords** Traffic monitoring · Person Re-Identification · Vehicle Re-Identification · Flow monitoring

1 Introduction

An important goal of transportation research is to improve and provide efficient public transportation systems that can accommodate many agents, whether it be vehicles or pedestrians, every day. This is especially important nowadays with the huge traffic congestion costing billions of dollars [1]. As a result, research efforts have been directed towards management and control of vehicle and pedestrian flows. Important prerequisites for such transportation network analysis are origin-destination (OD) matrices, which allow researchers to understand a population's trip demand. With the recent developments in new methods, such as data-driven methods [2], OD estimation have achieved

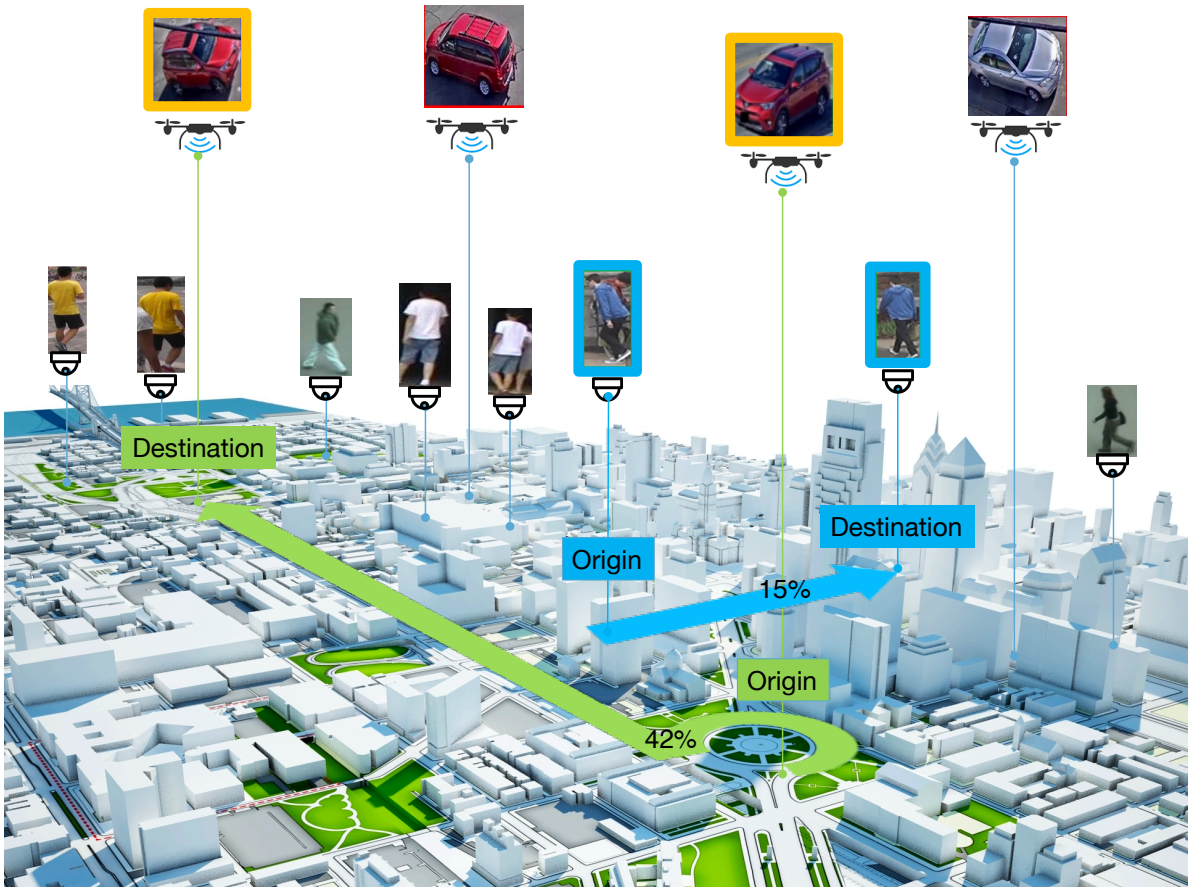

Figure 1: In this work, we present a new visual re-identification method, *i.e.*, whether agents (vehicles or pedestrians) captured in different images in non-overlapping views belong to the same agent. Agents with the orange bounding box belong to the same agent. Note different agents can be visually very similar.

good performance in various tasks of traffic management. However, it still faces the issue of how representative the chosen samples are of the population. One way to deal with this problem is to collect more data from the population, which is expensive and time-consuming when using traditional methods such as surveys or interviews. Another way is to make use of smartphones to collect such data [3]. Visual re-identification is a faster and cheaper way to collect these data. Visual re-identification is the task of associating images of the same agent taken from different cameras or from the same camera in different occasions. This is represented in Fig. 1. This task is nowadays possible due to the complex network of cameras already places in and around cities. Moreover, recent works have been pushing towards the use of drone technology to collect massive amounts of data in order to study the traffic phenomena, such as the recently released pNEUMA dataset [4]. All these collected data can be used as inputs to a visual re-identification model to associate different agents together and obtain their paths.

The task of re-identification (re-id) has long been a task of extracting features/representations from two observations and measuring how similar these features are. Since different variations affect these features, many works have introduced different methods to improve their extraction. Initially,

these features were hand-crafted and include spatio-temporal information such as color, width and height, and salient edge histograms. Some work have also tried to use different input modalities such as depth [5, 6, 7], infrared [8], LiDAR [9], or Inductive Loop Detectors for vehicles [10]. These features, however, fail drastically when dealing with unexpected scenarios. To remedy this problem and with the advent of machine learning, researchers are now benefiting from the strength of deep learning to be able to extract more general and more discriminative features allowing them to reach high performance. Since then, an arms race of methods was built on top of this by making use of different object-specific characteristics (*e.g.*, human semantic segmentation, pose) and by learning features through the supervision of a cross-entropy loss.

A main pitfall of learning with cross-entropy supervision is the fact that it separates the different inputs solely based on the labels without taking into consideration the actual similarity between the inputs. None of the recent methods have tackled the problem where, even though two very similar agents are distinct, their similarity score should encode information about how similar they appear while also distinguishing them. The network usually tries its best to find a boundary between the different classes even for inputs that are very similar. This leads the network to find unreasonable explanations for the differences in labels and thus would negatively affect its generalizability. Since re-id also deals with the problem of having a small set of images per class, it would aggravate this issue. Controlling the network's confidence in its predictions would alleviate this problem. To the best of your knowledge, this is the first paper to apply this concept in the field of re-identification.

In this paper, we propose to model confidence when learning representations appropriate for re-id. By inducing doubt while training a network, we are able to tackle the inherent problem discussed previously when cross-entropy is used in a distance metric and representation learning problem such as person or vehicle re-identification. Inspired by previous works that use uncertainty to regularize the network, we study three alternatives that aim at reducing the confidence of the network and show a gain of 6-7 % in mAP across 5 different datasets. Although these methods have shown only a small improvement in other image classification tasks [11, 12, 13], they drastically improve the performance of re-id models due to its innate problem (Section 3). By combining our methods with advanced ranking methods, we outperform state-of-the-art models without modeling additional characteristics specific to the object in question. The software is open-source and is made available online[1].

2 Related Work

Initially, re-identification's origins were based on multi-camera tracking. In this context, both visual and spatio-temporal information were used to predict whether the same agent was found in different cameras and at different times [14, 15, 16, 17]. Gheissari et al. [18] were the first to make use of only visual information to match different pedestrians. They made use of color and salient edge histograms to re-identify around 44 pedestrians recorded by 3 different cameras. This work clearly divided visual re-identification from multi-camera tracking. Alahi et al. [19] showed the benefits of visual re-identification in a network of fixed and mobile cameras. Fixed cameras installed at urban intersection can help the detection algorithms of cameras mounted on vehicles. Since then, researchers have proposed many methods to solve the visual re-identification tasks[20, 21]. In the

[1] https://git.io/deep-visual-reID-confidence

remaining of this section, we briefly present key methods tackling the problem for "person" and "vehicle".

2.1 Person Re-Identification

Even though person re-identification is highly beneficial to analyze how people make use of different transportation modes, it is not a widely studied matter in this context. Recent works in the transportation field have been using traditional and linear methods such as the Kalman filter [9, 22] and particle filters [23, 24, 25] to track pedestrians by using spatio-temporal information. These methods work well when tracking for a short period and while assuming that pedestrian's movement is linear. Applying the same methods over multiple cameras as well as in more complicated environments such as in a train station would lead to poor identification of passengers.

With the prevalence of deep neural networks in most computer vision tasks, person re-id followed this success when Li et al. [26] introduced a deep learning method for re-id that tried to overcome the problems of bounding box misalignment, photometric and geometric transforms while also introducing a new bigger dataset specifically for this task. This paved the way for new methods and datasets to emerge, causing the person re-id performance of machines to improve. Other work developed new methods that tackle specific challenges in person re-id by introducing different architectures and modules [27, 28, 29, 30].

Attention in Person reID. Recently, many methods have tried to improve the representation of the input by training multiple networks that extract global and local features and then combining these features to form the final representation. This is usually done by using either a deterministic way of dividing the different parts of the representation [31, 32, 33, 34] or making use of attention modules to separate the different parts [28, 35, 36]. Other works extracted intermediate representations to gain information about the input at different levels arguing that this allows the network to learn distinctive characteristics of the input at different scales [37, 38, 30]. Even though these methods showed improvement over their predecessors, they usually require separate networks to process each of the different features, leading to a more complex architecture and training procedure.

Human Characteristics. Another direction other researchers have taken is to make use of information and characteristics related specifically to humans in order to improve person re-id. The work by Xu et al. [39] aims at detecting three different types of pose information such as keypoints, rigid body parts (e.g., torso), and non-rigid parts (e.g., limbs). These information were extracted using an off-the-shelf human pose estimator. Then, with the help of these body parts, the features extracted by a feature extractor are refined and used to classify the different pedestrians. The use of third-party methods makes their model highly dependent on the performance of these methods. Another approach by Sarfraz et al [40] uses keypoint information, in addition to the input image, to train a ResNet-50 model as well as another connected module that detects the view (front, back or side). Kayaleh et al. [41] also made use of features extracted from different body parts and concatenated them to form a global feature which in turn was used to perform re-identification. The disadvantage of these methods is their high dependence on other methods and datasets that require annotation. Moreover, the fact that these models depend on specific human characteristics prevents them from being leveraged for other image-retrieval and clustering tasks.

Re-Ranking. In addition to learning better features, many works have tried to improve the ranking process of person re-id by including information about how the different galleries are related instead of just using the relationship between the pairs of queries and galleries [42, 43, 44, 42, 45, 46, 43, 47]. Zhong et al [42] introduced a method for refining the distances between the queries and galleries by making use of the k-reciprocal nearest neighbors. This is done as a post-processing step to improve the ranking process. Shen et al. [44] argued that this does not help in learning better features during training and introduced a new learnable module that performs a random walk on a graph connecting the different gallery images. By performing a random walk operation, gallery-to-gallery (G2G) information is taken into consideration while training the network, thus resulting in a more complete representation that provides a better ranking performance. Other methods also tried to include G2G information by using Graph Neural Networks [48] and Conditional Random Fields [49]. We will make use of G2G information by applying different re-ranking methods.

Metric Learning. Several previous works have tried to tackle the problem of person re-id by introducing new metric loss functions. Both contrastive [50] and binary loss functions have been employed in order to push apart negative image pairs while pulling positive image pairs together [51, 34]. Taking into consideration both the pull and push of contrastive loss, other methods[30, 52, 38] used triplet loss that simultaneously tackles negative and positive pairs leading to a less greedy method. Chen et al. [53] extended this loss to quadruplet inputs. The drawback of these methods is their high sensitivity to the sampling technique used. As a result, Yu et al.[54] introduced the HAP2S loss to tackle this drawback and showed improvement in performance. All the above methods try to encode metric information in the embedding space compared to cross-entropy which is considered as a representation learning method.

2.2 Vehicle Re-Identification

Vision-based methods for vehicle re-identification are rather new. Initially, vehicle datasets were small and mainly used for car color, model classification, or detection [55]. As a result, Liu et al. [56] built a large scale dataset for vehicle re-identification similar to person re-id datasets. They also made use of many techniques derived for person re-id and compared their performance. They later extended their work and added license plate verification as a way to improve re-identification [57]. This intuition comes from the fact that each car has a unique license plate. Metric learning techniques were also extended for this task. Liu et al [58] introduced coupled clusters loss, a variant of triplet loss, to deal with sensitivity to the choice of the triplet samples. Zhang et al. [59] also improved triplet loss by augmenting the training with a classification loss and modifying the dataset sampling method. Instead of randomly sampling triplets of anchors, positive, and negative samples, their method ensure that the negative sample is an anchor or positive sample in another triplet. This provides a way for negatives to be pushed towards similar images rather than being pushed away from the anchor randomly. Bai et al. [60] tried to deal with the problem of inter-class similarity and intra-class variance by introducing the group sensitive triplet embedding (GSTE). This is done by combining samples into intermediate "groups" at different granularity levels such as vehicle ID and vehicle model.

Other works made use of different attributes and modalities to improve vehicle re-identification. Li et al. [61] performed multi-task training that includes ID classification, attribute recognition, contrastive loss, and triplet loss. Tang el al. [62] made use of hand-crafted features such as color and

Figure 2: Pairs of images of different IDs but very similar appearance. None of these images belong to the same agent. Images taken from Market1501 [65] & VERI-Wild [66] datasets.

introduced a multi-modal metric learning method to fuse these features with deep features extracted using a neural network. Moreover, GANs are being increasingly adopted in vehicle re-identification. The main intuition is to transfer the query image to a domain that makes it more robust and efficient to compare with other images [63, 64]. Zhou et al. [63] proposed the generation of vehicle images from different views to deal with cross-view re-identification.

Large-scale datasets for vehicle re-identification are still recent and with the emergence of intelligent transportation, more research is being developed to improve this field. While it does introduce certain challenges different from person re-identification, the ability to find a re-identification method that performs well on any object of interest is important. In this paper, we do not make use of characteristics specific to the object of interest or feature division and show the importance of confidence when training a re-id model with a cross-entropy loss.

3 Problem Formulation

A re-id model's main task is to distinguish between different agents across images. As previously stated, this is a challenging task since it tries to relate images of agents across different cameras as well as at different times. The fact that the images are captured under different circumstances might lead to subtle differences in hue and image color that can drastically effect the performance of a re-id model. Moreover, the illumination, background clutter, occlusion, and observable object parts are usually dramatically different which might easily fool the network and render it unusable. Even images captured by the same camera can have many of these variations.

Due to the challenges explained above, there is not always a clear margin of separation between individual agents. Pedestrians or vehicles in some cases have very subtle differences that separate them from each other making the task of identifying them even more challenging for an observer. A good example is shown in Figure 2 which introduces the inherent challenge we are trying to tackle in this paper.

The pedestrians within the images in Figure 2 are very difficult to discern from one another even for a human eye. Each pair of images show two different pedestrians who share very similar appearances. When a model is trained to separate these images, it might face difficulties doing so. Since the images are very similar and a network's only main goal is to reduce its loss, it will learn to focus on the pose or even the illumination of the images to discern them. These two variations are some of the many variations that previous methods try to overcome. This problem is also shared with vehicle re-identification. Since different vehicles might share the same model and color, many variations such as illumination, viewing angle, and weather might be used by the network to separate them.

Current state-of-the-art re-id systems train their own models by using the cross-entropy loss function. The cross-entropy calculates the number of bits needed for an event, which in this case is the label given the input, using the estimated probability distribution instead of the true distribution. In the case of training a neural network, the cross-entropy is minimized so that the model distribution is the same as that of the ground-truth, which is usually a one-hot encoding. This means minimizing this loss pushes the distribution of the model to output a high probability for the correct label while outputting very low probabilities for the others. The fact that cross-entropy requires that the logits for the ground-truth label to be much bigger than other labels pushes the network to take into consideration certain destructive variations to separate the different classes and especially for images such as in Figure 2.

In order to modify the cross entropy in a way that solves the problem described above, we add a missing term to the loss function which allows it to not be confident about certain data points. Thus, the modified cross entropy loss function allows the network not to overfit on variations that are destructive for the re-id task and accept the fact that pedestrians or vehicles do sometimes look very similar. The idea of preventing the network from being very confident is not a new concept. However, its evaluation on other computer vision tasks, such as object detection, only leads to slight improvements in performance. From the reasoning based on Figure 2 as well as the characteristics of person and vehicle re-id datasets, we show in this paper that this concept, if applied to a simple baseline, can improve the results drastically and even outperform certain highly specialized state-of-the-art methods.

4 Method

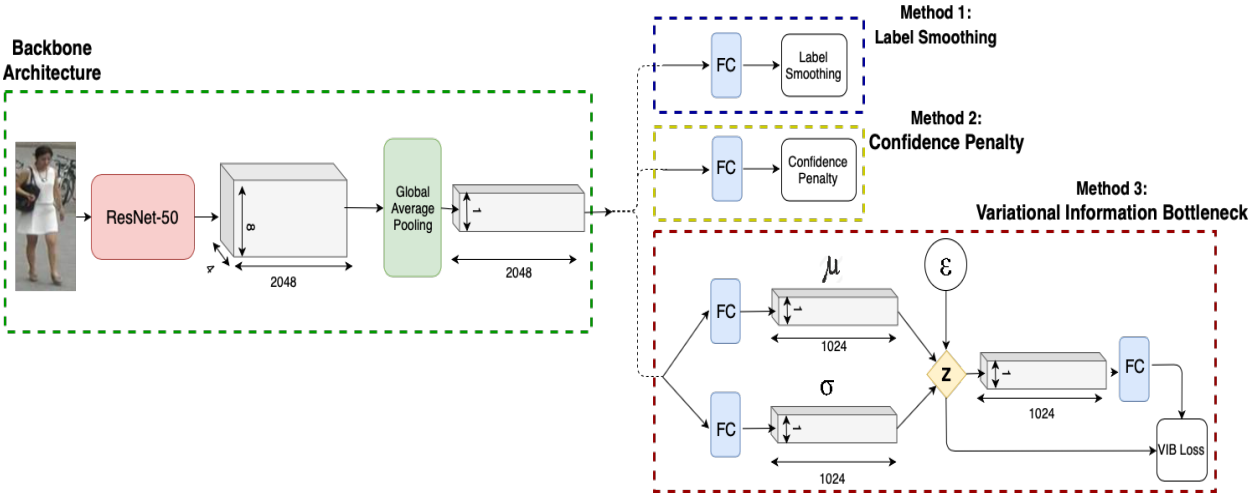

Figure 3: Network architecture including the three methods being studied: Label Smoothing (LS), Confidence Penalty (CP), Variational Information Bottleneck (VIB). ϵ is the Gaussian noise needed for the reparameterization trick, μ and σ are the mean and standard deviation respectively of the latent Z distribution.

Current re-id models face difficulties in distinguishing between different agents who share some visual similarities due to the model's objective of maximizing its confidence in its predictions, as previously discussed. In this section, we place the three common methods into a common framework of a cross-entropy term and a KL divergence term. This will allow us to investigate common properties and to identify their differences. We show how the three methods, LS, CP, and VIB, are specific instances of our common framework which forces the model to be less confident of its predictions than a plain cross entropy term. These methods usually show a small improvement when used during training in other computer vision tasks [67, 13, 12]; however, we show that, because of the problems specified in Section 3, these methods provide a drastic boost for the task of person and vehicle re-id.

We will bring all methods into the following common framework of a cross-entropy regularized with a KL divergence where the loss L is

$$L = \alpha H(q, p) + \beta KL \tag{1}$$

where $q = q(y|x)$ is the ground truth distribution of the output id y given the input image x, $p = p(y|x)$ is the predicted distribution, $H(q, p)$ is the cross-entropy between q and p and KL is the Kulback-Leibler divergence [68]. In ReID, we have multiple ids $y \in Y$ where we denote the number of Y – the number of classes – as C.

4.1 Label Smoothing

With Label Smoothing [67], the predicted distribution p is regularized towards the uniform distribution u with the KL divergence term:

$$L_{LS} = \alpha H(q, p) + \beta KL[u, p] \quad . \tag{2}$$

In this form, the KL divergence is the expectation over the uniform distribution of the logarithmic difference between u and p. Forming the convex mixture with $\alpha \equiv 1 - \beta$ and expanding this equation yields:

$$L_{LS} = -(1-\beta) \sum_{y \in Y} q \log p + \beta \sum_{y \in Y} u \log \frac{u}{p} \tag{3}$$

$$= -\sum_{y \in Y} [\beta u + (1-\beta) q] \log p \tag{4}$$

$$= H(q_{LS}, p) \tag{5}$$

where we dropped the constant term $u \log u$. We arrived at a single cross entropy without a KL divergence and defined a new label-smoothed ground truth distribution q_{LS}. The uniform distribution u is over the C classes, *i.e.*, it evaluates to $1/C$. For a sample n:

$$q_{LS}(y|x_n) = \begin{cases} 1 - \frac{(C-1)\beta}{C} & \text{true } y, \\ \frac{\beta}{C} & \text{otherwise.} \end{cases} \tag{6}$$

From a KL divergence between u and p, we obtained a loss function similar to the regularizer introduced by Szegedy et al. [67], which aims at allowing a model to be less confident about its prediction. It regularizes a softmax classifier by assigning a small value to all ground-truth labels.

This method makes sure that the label for the correct class does not become much larger than all other classes and thus prevents the network from over-fitting. When label smoothing was proposed and tested on ImageNet, it showed a small improvement of around 0.2% for top-1 error. Even though it did not show a huge improvement, we show in Section 7 that this method has a bigger effect on the task at hand based on the arguments stated in Section 3.

4.2 Confidence Penalty

Reversing the arguments of the KL divergence, we obtain an equation for confidence penalty [13]:

$$L_{CP} = \alpha H(q, p) + \beta KL[p, u] \tag{7}$$

Comparing Eq. 7 to Eq. 2, we can observe the main difference. The error calculation, in this case, between the uniform and predicted distribution is weighted by the predicted distribution. Expanding this equation:

$$L_{CP} = \alpha H(q, p) + \beta \sum_{y \in Y} p \log \frac{p}{u} \tag{8}$$

$$= \alpha H(q, p) - \beta H(p) \tag{9}$$

where $\sum_{y \in Y} p \log u$ is removed since it is a constant. We notice that the resulting loss function aims at maximizing the entropy of the predicted distribution. The increase in entropy forces the network to be less certain of its predictions. Pereyra et al. [13] reached a similar conclusion and showed that, by applying this method, they got a smoother predicted distribution as well as a small improvement on MNIST. This method, however, did not show an improvement on a more difficult dataset such as CIFAR-10. Similar to label smoothing, this method does not require any architecture modification as shown in Figure 3.

4.3 Deep Variational Information Bottleneck

Another way to increase the entropy of the output distribution is to force latent representations to be similar to each other irrespective of the input. This means that information distinguishing different samples is being lost. This idea can be derived from the information bottleneck (IB) principle [69] where the mutual information between the input and latent representations is minimized. The IB principle [69] is a technique that tries to find the best trade-off between accuracy and complexity of latent variables. Latent variables are hidden variables that describe a specific input while maintaining all the relevant information needed for a specific task. The information bottleneck method tries to maximize this objective:

$$\max_{p(z|x)} \alpha I(z; y) - \beta I(z; x) \tag{10}$$

where z is the latent variable. Based on the above equation, the objective is to learn a representation z that is very informative about y while compressive about x. In order to apply the IB objective to a neural network, Alemi et al. [12] approximated a lower bound to the information bottleneck by using variational inference and the reparameterization trick introduced by Kingma et al. [70] to introduce a new objective function referred to as Variational Information Bottleneck (VIB).

When applying this method, the model is divided into an encoder that takes the input x and maps it to a distribution describing the latent space z. The encoder outputs both the mean μ and standard deviation σ that describe this distribution. Then the predicted latent distribution is used to sample a specific latent representation. To force the second part of equation 10 to be maximized, this distribution should not depend on the input thus forcing the representation z to forget some information about it. This is done by minimizing the divergence between the encoder's distribution $w = w(z|x)$ and the prior r which is an isotropic multivariate Gaussian $r(z)$. The resulting objective function to minimize is:

$$L_{VIB} = \alpha H(q, \tilde{p}) + \beta KL[w, r] \tag{11}$$

where $\tilde{p} = p(y|f(x, \epsilon))$.

In order to compute the KL divergence analytically and back-propagate using its gradients, w is approximated by a multivariate Gaussian distribution with a diagonal covariance matrix. As can be seen in equation 11, if $\beta \to \infty$, the latent representation would follow a distribution independent of the input. This is somewhat similar to the effect of both confidence penalty and label smoothing where a single representation is forced to contain some information about more than one label. However, VIB applies this restriction directly to the latent space. Using this method while training, Alemi et al. [12] showed close results to state-of-the-art models while using less information about the input which is measured using mutual information $I(x; z)$.

Compared to previously mentioned methods, in order to use the VIB loss, a fully connected layer is added at the output of the ResNet-50 base model to compute the mean and standard deviation as shown in Figure 3.

5 Discussion

As can be seen in the equations above, all three methods try to increase the uncertainty of the model or in other words, decrease its confidence. On one hand, label smoothing and confidence penalty act on the output distribution while on the other hand, VIB acts on the latent representation directly. Thus this requires the original architecture to be modified to accommodate the VIB loss.

In addition, when expressed in terms of KL divergence, both label smoothing and confidence penalty are very similar except for the fact that the KL divergences are reversed. The forward KL divergence uses a constant represented by $u = \frac{1}{C}$ (Equation 2) to weight the log expression. However, the reverse KL divergence weighs using the output prediction of the model $p(y|x)$ which varies during the training process (Equation 7). In other words, confidence penalty does not equally penalize all the label predictions but implicitly gives more importance to predictions that the network incorrectly gives higher probabilities to and which are farther away than the uniform distribution. This weighing is adaptively changing as the network trains. Label smoothing however penalizes all label predictions equally. Moreover, label smoothing tries to prevent an output prediction of 0. This is because it is weighted by the uniform distribution u which is always greater than zero. Confidence penalty, however, might force certain output predictions to be zero although u is never zero.

By expressing all three methods in terms of KL divergence, we get more insight on why confidence penalty is able to outperform other methods. Compared to label smoothing, confidence penalty weighs the error by the network's current prediction and thus is more adaptable to the different inputs. Moreover, VIB provides the network with less degree of freedom compared to confidence penalty. This is because the latter acts on the output distribution allowing the representation, which is the main feature used for ranking in re-id, to move more freely in the feature space.

6 Experiments

To evaluate our proposed method, we use publicly available person and vehicle re-identification datasets which are Market-1501 [65], MSMT17 [71], DukeMTMC-reID [72], and VERI-Wild [66].

Market1501 [65]: The Market dataset is a well-known person re-identification dataset that contains 32,668 bounding boxes of 1,501 individuals captured using 6 cameras. These bounding boxes were obtained using the Deformable Part Model (DPM) [73]. The training set is made up of 751 identities with 12,936 images while the test set has 750 identities distinct from the one in the training set divided into query and gallery images.

MSMT17 [71]: This is a very recent dataset which was carried out over a long period of time. This benchmark contains a total of 126,441 bounding boxes of 4,101 identities captured using 15 cameras. The images vary in terms of location (outdoors, indoors), weather conditions (over a month), as well as different times of day (morning, noon, afternoon). The bounding boxes were

obtained using Faster RCNN and corrected using labelers. Containing many variations makes this dataset challenging as well as a good benchmark to use.

DukeMTMC-reID [72]: The DukeMTMC-reID dataset is a small part of the bigger DukeMTMC dataset that is usually used for multi-target multi-camera tracking. It is taken from 8 different cameras, and the person bounding box is manually labeled. It is made up of 1,404 different identities with 702 identities used for training and 702 other identities used for testing.

VERI-Wild [66]: The VERI-Wild dataset is a recently released large-scale vehicle re-identification dataset with 416, 314 vehicle images of 40, 671 IDs captured by 174 cameras. Vehicle recording is unconstrained and thus contains vehicles from different angles. This makes this dataset very challenging. The test set is divided into three subsets: small, medium, and large. We show our results on the small and medium subset since the large subset requires more memory.

Figure 4: An example from the EPFL Roundabout dataset and the vehicles that are detected. We make use of the crop of vehicles to build a re-identification dataset similar to VERI-Wild.

EPFL Roundabout: In order to validate the use of our methods on images recorded from drone view, we build the EPFL Roundabout dataset by recording vehicle flow of five different locations using a drone. The division of the dataset is similar to VERI-Wild. To encourage efficient training, we set the training and testing ratio as 1:3. As a result, we obtain 85,268 images with 3,479 IDS. An image example is shown in Figure 4 The dataset will be made publicly available.

6.1 Evaluation Protocol

For evaluation, we use the cumulative matching characteristic (CMC) and Mean Average Precision (mAP). These two metrics are the most popular evaluation metrics since re-identification systems should be able to output all the correct matches (mAP) in addition to having high accuracy at different ranks (CMC). During testing, for every query, there is a list of gallery images ordered in increasing order according to their L2 distance from this query.

Parameters	Market-1501	DukeMTMC	MSMT17	VERI-Wild
$\text{LR}_{cross},\ \alpha$	$5 \times 10^{-4},\ 2$	$2 \times 10^{-4},\ 1$	$3 \times 10^{-4},\ 4$	$1 \times 10^{-3},\ 6$
$\text{LR}_{LS},\ \alpha$	$5 \times 10^{-4},\ 2$	$5 \times 10^{-4},\ 5$	$3 \times 10^{-4},\ 5$	$1 \times 10^{-3},\ 6$
$\text{LR}_{CP},\ \alpha$	$6 \times 10^{-4},\ 3$	$6 \times 10^{-4},\ 3$	$4 \times 10^{-4},\ 5$	$1 \times 10^{-3},\ 6$
$\text{LR}_{VIB},\ \alpha$	$4 \times 10^{-4},\ 6$	$6 \times 10^{-4},\ 3$	$5 \times 10^{-4},\ 6$	$1 \times 10^{-3},\ 6$
β_{CP}	0.085	0.085	0.085	0.6
β_{VIB}	0.01	0.01	0.01	0.01

Table 1: Hyperparameters for the different datasets and methods. LR: learning rate, β: pre-factors for loss constraint, α: pre-factor for cross-entropy

6.2 Implementation Details

The model was pre-trained using ImageNet. We do not add any layer to ResNet-50 when training both using label smoothing and confidence penalty except for a fully connected layer that outputs the different labels. When training the VIB algorithm, a fully-connected layer was added before the classification layer to output the mean and standard deviation which describe the distribution of the latent representations. A latent variable is then sampled from the predicted distribution. For all methods and datasets, hyperparameter tuning was performed for ResNet-50 in order to get the best possible accuracy.

Data Augmentation. We follow methods of data augmentations that are commonly used in the field of person re-identification. Since Market1501 uses DPM to obtain the bounding boxes, the images are initially randomly cropped. For all datasets, the inputs are resized to 256x128. Before providing them to the network, a random rectangle, with pixel values randomly assigned between [0, 255], is erased [74] from the images, and the resulting images are flipped horizontally with a probability of 0.5. This makes the network more robust to the orientation of the agents in the image as well as occlusion. Each image is then normalized and standardized using the mean and standard deviation provided when using a model pretrained on ImageNet. These transformations were applied only for the training set.

Hyperparameter Tuning. Since the hyperparameters (e.g., learning rate, β, and α) we are trying to optimize have multiplicative effects on the training procedure, the best method is to perform a log-space search. This is due to two reasons. The parameter is not too sensitive such that there may not be too much difference with 10 and 15 compared to 10 and 1000. The other reason is that using logarithmic scales allows us to search over a bigger space quickly.

Training Procedure. The samples used to form the training batch are randomly sampled from the datasets. It does not require any special sampling such as the PK Sampling required by triplet loss[75], which randomly samples P identities and then randomly K images for each identity to form a batch. The mini-batch has a size of 32 images. The model is trained for 300 epochs using AMSGrad [76] for all datasets with the learning rate decaying by 10 at epoch 20 and 40. In order to make sure that all models were trained with the best parameters, we perform hyperparameter tuning, as discussed previously. The different hyperparameters for the different datasets are shown in Table 1.

	Market1501		DukeMTMC	
Model	mAP	Rank1	mAP	Rank1
ResNet-50 [52]	47.78	73.90	44.99	65.22
ResNet-50 [44]	59.8	81.4	55.5	75.3
ResNet-50 [40]	59.8	82.6	50.3	71.5
ResNet-50 [37]	66.0	84.3	48.6	71.6
ResNet-50 [77]	66.95	84.42	57.34	75.60
ResNet-50 [41]	66.32	85.10	54.77	73.70
Our ResNet-50	70.2	87.5	59.6	78.6

Table 2: Comparison with published ResNet-50 results on the Market-1501 and DukeMTM-reID dataset.

Evaluation Procedure. For testing, the features that are extracted just before the last classification layer are used for the ranking process. The features for the queries and galleries are extracted and then compared to rank the gallery images relative to each query image. This is done when label smoothing or confidence penalty is used. When using the VIB loss, the network has an additional fully connected layer that outputs the mean and standard deviation for every latent dimension and a reparameterization trick that depends on random Gaussian noise. For ranking, we use the mean produced by the model as features for each image since this represents the average of the distribution over which the input image is mapped to. This is also due to the fact that the standard deviations tend to 1. To the best of our knowledge, using a latent representation sampled from a Gaussian parametrized by the predicted mean and standard deviation has not been tackled before for the person ad vehicle re-id task.

7 Results

In order to show both qualitative and quantitative results, we split our results into three parts. In Sections 7.1 and 7.2, we compare our proposed methods to published baseline results and state-of-the-art methods respectively. In Section 7.3, we investigate the effect of the three methods on the ranking process of person and vehicle re-id. Although these methods were tested on ResNet-50, other re-id models can benefit from their positive effect on the performance especially when dealing with visually similar pedestrians.

7.1 Properly Trained Baseline

We compare our baseline to previously reported results of ResNet-50 on the Market-1501 and DukeMTMC-reID datasets. The published results reported in Table 2 correspond to pre-trained ResNet-50 that used the cross-entropy loss similar to our method. As can be observed in Table 2, there is a clear difference between our result and the results reported in published papers as well as amongst the published results themselves. Our properly trained baseline, which consists of a ResNet-50 model trained using a normal cross-entropy loss, was able to outperform all previous baselines. This table represents one of the many pitfalls that occurs when training a model. This is shown by the fact that papers that make use of exactly the same baseline have different results.

This is usually due to the hyperparameters chosen. Another pitfall is that to compare different baselines and losses, the same hyperparameters are set. This is somewhat unfair since different baselines and losses optimize different parameters and in different ways thus requiring distinct hyperparameters. This is why we employ different learning rates for different datasets and methods as shown in Table 1. As a result, we were able to achieve, using the baseline, around $\sim 3\%$ increase in mAP and rank-1 for both datasets.

7.2 Comparison with State-of-the-Art

We evaluate our proposed confidence-based methods against recently published papers in person and vehicle re-id. Each of our methods is evaluated on five datasets: Market1501, DukeMTMC-reID, MSMT17, VERI-Wild, and EPFL Roundabout. We are able to reach state-of-the-art performance without any human-specific design and added complexity thus showing the importance of penalizing the confidence of a network in the task of re-identification. We also do not make use of data augmentation during the evaluation stage like DuATM [79].

Evaluation on Market1501: As shown in Table 3, the models were able to reach state-of-the-art results. In order to better understand the importance of penalizing confidence compared to other methods, it is important to note some distinct differences. Confidence penalty was able to outperform HAP2S [54] which tried to deal with hard samples by giving them higher weights. Moreover, Mancs[38], which shows good performance, makes use of three different losses, attention layers, as well as a special sampling scheme. To compare our results with methods that include gallery-to-gallery information during inference, such as Deep Group RW [44] and SGGNN [48], we apply re-ranking to our three methods. We were able to outperform these methods with a significant increase in mAP($\sim 7.5\%$). As a result, we got state-of-the-art performance without the added complexity of learning new layers and parameters while tackling the problem stated in Section 3.

Evaluation on DukeMTMC-reID: Similar to the Market-1501 dataset, we achieved competitive results in all proposed methods with confidence penalty resulting in the best improvement (Table 3). In addition to that, using Sarfraz et al.'s [40] recent re-ranking method (ECN), we were able to get better results than PSE [40] in both mAP and rank-1. It is important to note that SPReID augments the training data of both DukeMTMC-reID and Market1501 with 10 datasets resulting in a large number of training samples which would improve the performance of the network.

Evaluation on MSMT17: Since this is a bigger dataset with many variations, it proved to be a challenging benchmark[71]. Nonetheless, we were able to show a notable improvement over previous methods as well as over our own baseline (Table 4). Similarly, confidence penalty performed the best by achieving 68.6% in rank-1 and 39.3% in mAP. By applying re-ranking, both rank-1 and mAP are further improved to 75.3% and 59.1% respectively.

Evaluation on VERI-Wild: This dataset contains a large amount of images and is divided into small, medium, and large subsets. We evaluate our model on the small and medium subset and are able to achieve state-of-the-art performance (Table 5). Compared to the person re-id datasets, using the VIB method achieves the best performance on VERI-Wild. This might be due to the fact that vehicles face the problems of similar appearance and different IDs more frequently. Thus, a more strict method of penalizing certainty is required compared to confidence penalty and label smoothing.

	Market1501		DukeMTMC	
Model	mAP	Rank1	mAP	Rank1
CamStyle (R)[78]	71.55	89.49	57.61	78.32
HAP2S_E+Xent(R)[54]	74.49	89.73	62.62	79.08
DuATM(!R)[79]	75.22	89.96	63.14	81.46
MLFN (!R)[37]	74.3	90.0	62.8	81.0
Shen et al.(R)[44]	75.3	90.1	63.2	80.3
PSE(R)[40] +ECN	84.0	90.3	79.8	85.2
DaRe(!R)[30] +RR	86.7	90.9	80.0	84.4
SPReID$^{w/fg}$(!R)[41]*	78.66	90.97	65.66	81.73
HA-CNN (!R) [28]	75.7	91.2	63.8	80.5
DuATM(!R)[79]**	76.62	91.42	64.58	81.82
SPReID^{comb}(!R)[41]*	79.67	91.45	68.78	83.3
P-Aligned (!R)[80]	79.6	91.7	69.3	84.4
SGGNN(R)[48]	82.8	92.3	68.2	81.1
Deep Group RW(R)[44]	82.5	92.7	66.4	80.7
Mancs(R)[38]	82.3	93.1	71.8	84.9
DNN+CRF(R) [49]	81.6	93.5	69.5	84.9
PCB+RPP [81]	81.6	**93.8**	69.2	83.3
P-Aligned (!R)[80]+RR	89.9	93.4	83.9	88.3
Our ResNet	70.7	87.2	59.6	78.6
Our ResNet(VIB)	76.1	90.2	62.4	80.7
Our ResNet(LS)	76.7	91.0	64.4	82.7
Our ResNet(CP)	78.2	91.4	66.8	83.9
Our ResNet+RR	85.7	89.7	78.5	83.4
Our ResNet(VIB)+RR	88.6	91.8	79.0	84.3
Our ResNet(LS)+RR	89.1	92.2	82.2	86.6
Our ResNet(CP)+RR	90.0	92.6	83.5	87.4
Our ResNet(VIB)+ECN	88.2	92.0	78.9	85.1
Our ResNet(LS)+ECN	89.4	92.7	83.2	86.9
Our ResNet(CP)+ECN	**90.1**	93.1	**84.1**	**88.5**

Table 3: Comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID. (!R): uses model different than ResNet, (R): uses ResNet-50, ECN: Expanded Cross Neighborhood Re-Ranking[40], "RR": k-reciprocal re-ranking[42], Xent: Softmax, *: uses combination of 10 datasets for training, **: uses data augmentation during evaluation stage.

Evaluation on EPFL Roundabout: Since the images are recorded from a drone view, vehicles are small and lack certain details. This makes it more challenging to discriminate between different vehicles and a a result suffers more from the problem discussed in Section 3. However, we were able to drastically improve the performance of the baseline by making use of confidence penalty, label smoothing, and VIB. Label smoothing, in this case shows, the biggest improvement of around 14% and 10% in both mAP and Rank1 respectively.(Table 6).

MSMT17			
Model	mAP	Rank1	Rank10
GoogleNet[71]	23.0	47.6	71.8
PDC[71]	29.7	58.0	79.4
GLAD[71]	34.0	61.4	81.6
Our ResNet	31.8	59.3	80.2
Our ResNet(VIB)	35.1	66.2	84.1
Our ResNet(LS)	36.9	66.8	84.9
Our ResNet(CP)	39.3	68.6	85.3
Our ResNet + RR	49.8	65.7	79.8
Our ResNet(VIB) +RR	55.4	73.3	84.7
Our ResNet(LS) + RR	57.1	73.7	85.3
Our ResNet(CP) + RR	**59.1**	**75.3**	**85.8**

Table 4: Comparison with state-of-the-art on the MSMT17 dataset.

VERI-Wild				
	Small		Medium	
Model	mAP	Rank1	mAP	Rank1
GSTE [82]	31.4	60.5	26.18	52.12
VERI-Wild [66]	35.1	64.0	29.8	57.8
Our ResNet	45.7	82.4	41.3	78.4
Our ResNet(LS)	57.7	84.6	57.2	85.5
Our ResNet(CP)	67.5	90.2	61.8	87.0
Our ResNet(VIB)	**74.1**	**92.1**	**68.5**	**89.7**

Table 5: Comparison with state-of-the-art on VERI-Wild dataset.

7.3 Effect of Proposed Methods

In addition to achieving state-of-the-art performance, it is also important to understand the effect of these three methods on the ranking process. All three methods aim at allowing the network to share some representation among different classes. This prevents the network from focusing on

	EPFL Roundabout ReID	
Model	mAP	Rank1
ResNet	41.5	75.0
Our ResNet(LS)	55.9	84.4
Our ResNet(CP)	56.1	85.2
Our ResNet(VIB)	**52.7**	**82.5**

Table 6: Effect of our methods on EPFL Roundabout ReID

Figure 5: Qualitative comparison of using confidence penalty on unseen Market1501 test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

undesirable information when separating very similar-looking pedestrians. To show this effect, we compare the confidence penalty model against the baseline model since it resulted in the best performance in person re-id. As can be seen, the test samples presented in Figure 5 and Figure 6 are difficult to rank even for an observer. This confirms the intrinsic difficulty of person and vehicle re-id stated in Section 3. When confidence penalty is not used for training, the network focuses on unimportant variations between the images. For instance, in both sets of samples (Figure 5), the incorrect gallery images are very similar to the query image despite belonging to a different

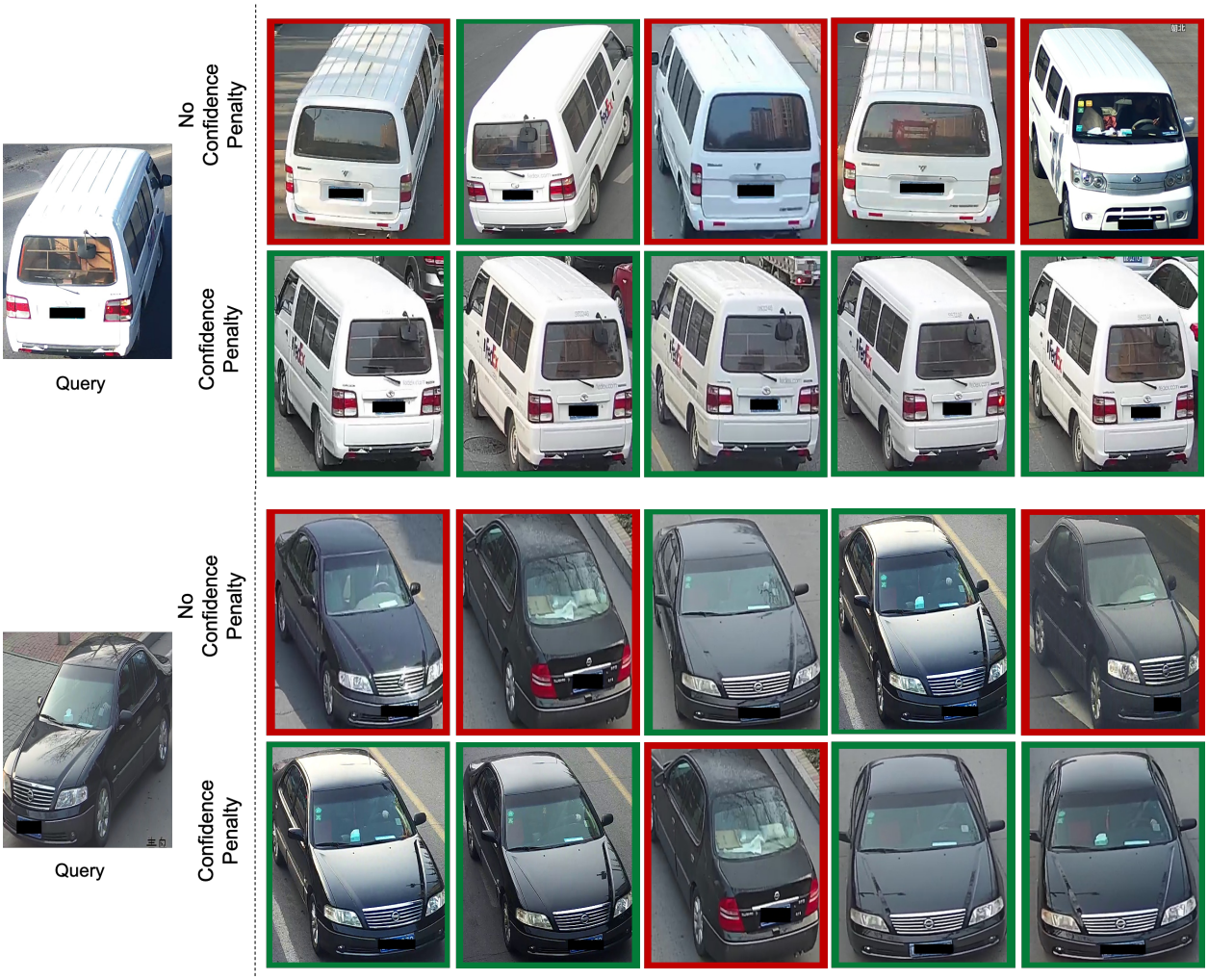

Figure 6: Qualitative comparison of using confidence penalty on unseen VERI-Wild test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

person. The baseline links the query image to the gallery images by possibly focusing on the background, shirt color, posture, and body rotation of the pedestrian in question. The same applies for vehicles. These characteristics are typically features that can confuse the model leading to wrong identification. Adding the confidence penalty is observed to remedy this challenge, as can be seen for all test samples provided. Adding the confidence penalty helps the model capture the subtle differences between multiple pedestrians that the baseline tends to misidentify. These are ideal examples of why confidence penalty drastically improved re-identification compared to less significant improvements in other computer vision tasks.

	Market1501		DukeMTMC	
Model	mAP	Rank1	mAP	Rank1
ResNet	70.7	87.2	59.6	78.6
Our ResNet(CP)	**78.2**	91.4	**66.8**	**83.9**
Our ResNet(VIB+LS)	76.6	90.4	63.9	82.0
Our ResNet(VIB+CP)	76.8	90.4	62.0	80.6
Our ResNet(LS+CP)	76.7	**91.7**	63.6	82.4
Our ResNet(VIB+LS+CP)	76.9	90.4	63.5	81.2

Table 7: Ablation study of different penalty combinations

8 Ablation Study

8.1 Combination of Penalties

Since the addition of the three methods to the baseline leads to an improvement in performance, one might wonder the effect of the different combination of these methods. As shown in Table 7, applying only confidence penalty (CP) leads to the best result. Intuitively, the combinations of the methods increases the restriction on the model preventing it from learning useful representations. Meanwhile, adding these methods together still leads an improved performance compared to the baseline which again affirms the beneficial effect that they have.

8.2 Penalties on State-of-the-Art methods

To study the effect of our best method on the different state-of-the-art methods, we test confidence penalty on PCB [81], HACNN [28], and BFE [83]. It is important to note that PCB divides the representation into multiple features that are then used for identification. These features are referred to as local features since they are spatially local to a certain region of the input. In addition to local features, HACNN also extracts a global representation from the whole input. In comparison, BFE performs person re-identification using only a global representation. Table 8 shows that the gain in performance for PCB is between +0.3 to +1.5%. One intuition behind the limited gain is that PCB divides its features into multiple parts (local features) before applying global average pooling. This allows the representations to be fine-grained focusing on the details that differentiate visually similar inputs. To analyze this, we test CP on HA-CNN that uses global and local features. Its performance is improved ($\sim$+3% mAP) to exceed PCB and to have on-par results with PCB+RPP. Also, CP has a bigger effect on global features than on local features in HA-CNN (Table 8) confirming our reasoning. The performance of another method, BFE, that uses only global features is also improved using CP compared to cross-entropy (Xent).

8.3 Vehicle Ego-centric Pedestrian Re-Identification

The dataset that were used to evaluate our methods are collected from cameras mounted in specific locations or using drones. In order to test the effectiveness of our method in a scenario where the camera is mounted on a car, we make use of the nuScenes dataset. This is the first dataset to contain all sensor information that a full autonomous vehicles requires from RGB camera, radars

	Market1501		DukeMTMC	
Model	mAP	Rank1	mAP	Rank1
PCB (own)[81]	78.5	92.4	69.6	83.8
Our PCB(LS)	77.1	91.7	69.0	**84.6**
Our PCB(CP)	**78.8**	**93.2**	**70.1**	84.3
PCB+RPP (own)[81]	81.5	93.3	71.3	84.4
Our PCB+RPP(LS)	80.9	92.8	71.3	**85.3**
Our PCB+RPP(CP)	**81.6**	93.3	**71.5**	84.2
HA-CNN[28]	75.7	91.2	63.8	80.5
HA-CNN(own)	78.6	91.6	67.3	83.1
Our HA-CNN(LS)	80.1	92.0	68.4	83.6
Our HA-CNN(CP)	**81.1**	**92.5**	**69.7**	**83.8**
HA-CNN(G)	72.3	87.9	60.0	78.7
Our HA-CNN(CP+G)	**77.3**	**91.5**	**63.4**	**82.0**
HA-CNN(L)	73.8	89.5	62.3	81.0
Our HA-CNN(CP+L)	**75.5**	**91.2**	**62.8**	81.0
BFE (Xent)[83]	83.7	93.5	73.5	86.4
Our BFE (CP)	**85.7**	**94.2**	**76.1**	**88.6**

Table 8: Effect of CP and LS on PCB, HA-CNN and BFE. Xent: Cross-Entropy, G: using global features, L: using local features

	nuScenes-ReID	
Model	mAP	Rank1
ResNet	61.7	66.3
Our ResNet(LS)	66.0	70.4
Our ResNet(CP)	**70.7**	**74.4**
Our ResNet(VIB)	67.7	70.4

Table 9: Effect of Confidence Penalty on NuScenes-ReID

to lidars. Since the pedestrians are tracked across images and different cameras, we can build a re-identification dataset similar to Market1501. We call the resulting dataset, nuScenes-ReiD. This is a challenging dataset since images are recorded from the view of a moving car resulting in different pedestrian sizes. As shown in Table 9, applying confidence penalty to the baseline significantly improves the performance (from 59.1%mAP to 70.9%). The code to build this re-id dataset will be made public.

9 Conclusions

An important task of transportation research is better analyzing and understanding traffic flow. Visual re-identification, the task of association similar agents, can aid in this goal. Thus, in this work we aim to deal with certain challenges that plague this task. First, we emphasize an intrinsic

characteristic of person and vehicle re-identification that poses a problem to the network being trained. The classes that these re-id task try to separate are not as easy as separating cats and dogs. Different agents with different identities can have very similar appearances. We have demonstrated that three methods, that reduce a model's confidence, are able to deal with this problem while achieving state-of-the-art results. Confidence penalty proved to be the best and most lightweight amongst the three different methods. In addition, it is interesting to note that VIB is able to achieve similar results while using smaller representations. Both label smoothing and confidence penalty use a representation of dimension 2048 while VIB uses a representation of dimension 1024. These three methods can be leveraged to improve the performance of previous re-id methods as well. It remains an exciting future work to study their effect on other image retrieval and clustering tasks.

With the ability to identify pedestrians and cars across both time and space, this allows us to better understand how they move from one place to the other without going through the manual and expensive way of using surveys or interviews. Some methods have also been developed to estimate the OD matrices from existing observed traffic flows. These methods, however, require the collected data to be large and representative of the real distribution. This is where re-identification can play a major role. CCTV cameras already placed around entry and exit of different transportation stops can be used to associate agents that pass through them. For instance, a person can be detected entering a specific train and then this detection can be associated to the same person exiting at a different station and at different times. This task provides an automatic method of collecting data about passenger movements and thus allowing us to build an accurate OD matrix that can be used for different transportation tasks.

References

[1] Benjamin Schneider. The high cost of global traffic jams, Feb 2018.

[2] Panchamy Krishnakumari, Hans van Lint, Tamara Djukic, and Oded Cats. A data driven method for od matrix estimation. *Transportation Research Part C: Emerging Technologies*, 2019.

[3] Fang Zhao, Ajinkya Ghorpade, Francisco Câmara Pereira, Christopher Zegras, and Moshe Ben-Akiva. Stop detection in smartphone-based travel surveys. *Transportation research procedia*, 11:218–226, 2015.

[4] Emmanouil Barmpounakis and Nikolas Geroliminis. On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment. *Transportation Research Part C: Emerging Technologies*, 111:50–71, 2020.

[5] A. Wu, W. Zheng, and J. Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, June 2017.

[6] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Person depth reid: Robust person re-identification with commodity depth sensors. *CoRR*, abs/1705.09882, 2017.

[7] A. Wu, W. Zheng, and J. Lai. Depth-based person re-identification. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 026–030, Nov 2015.

[8] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Junxuan Zhao, Hao Xu, Hongchao Liu, Jianqing Wu, Yichen Zheng, and Dayong Wu. Detection and tracking of pedestrians and vehicles using roadside lidar sensors. *Transportation Research Part C Emerging Technologies*, 100:68–87, 03 2019.

[10] Baher Abdulhai and Seyed M Tabib. Spatio-temporal inductance-pattern recognition for vehicle re-identification. *Transportation Research Part C: Emerging Technologies*, 11(3-4):223–239, 2003.

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *CoRR*, abs/1612.00410, 2016.

[13] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548, 2017.

[14] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3 – 19, 2013. Extracting Semantics from Multi-Spectrum Video.

[15] Timothy Huang and Stuart Russell. Object identification in a bayesian context. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2*, IJCAI'97, pages 1276–1282, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[16] Riccardo Mazzon, Syed Fahad Tahir, and Andrea Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, 33(14):1828–1837, 2012.

[17] David Held, Jesse Levinson, Sebastian Thrun, and Silvio Savarese. Combining 3d shape, color, and motion for robust anytime tracking. In *Robotics: science and systems*, 2014.

[18] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535, June 2006.

[19] Alexandre Alahi, Michel Bierlaire, and Pierre Vandergheynst. Robust real-time pedestrians detection in urban environments with low-resolution cameras. *Transportation research part C: emerging technologies*, 39:113–128, 2014.

[20] Etienne Corvee, Francois Bremond, Monique Thonnat, et al. Person re-identification using spatial covariance regions of human body parts. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435–440. IEEE, 2010.

[21] Etienne Corvee, Francois Bremond, Monique Thonnat, et al. Person re-identification using haar-based and dcd-based signature. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8. IEEE, 2010.

[22] Lie Guo, Linhui Li, Yibing Zhao, and Zongyan Zhao. Pedestrian tracking based on camshift with kalman prediction for autonomous vehicles. *International Journal of Advanced Robotic Systems*, 13(3):120, 2016.

[23] P. K. Gaddigoudar, T. R. Balihalli, S. S. Ijantkar, N. C. Iyer, and S. Maralappanavar. Pedestrian detection and tracking using particle filtering. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 110–115, May 2017.

[24] Hui Li, Yun Liu, Chuanxu Wang, Shujun Zhang, and Xuehong Cui. Tracking algorithm of multiple pedestrians based on particle filters in video sequences. *Computational Intelligence and Neuroscience*, 2016, 10 2016.

[25] Carine Hue, J-P Le Cadre, and Patrick Pérez. Tracking multiple objects with particle filtering. *IEEE transactions on aerospace and electronic systems*, 38(3):791–812, 2002.

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[27] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[29] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39, Aug 2014.

[32] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[33] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *Computer Vision – ECCV 2016*, pages 135–153. Springer International Publishing, 2016.

[34] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[35] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep representation learning with part loss for person re-identification. *CoRR*, abs/1707.00798, 2017.

[37] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[40] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[42] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[43] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, and Jun Chen. Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1239–1242, New York, NY, USA, 2015. ACM.

[44] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] J. García, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1305–1313, Dec 2015.

[46] Qingming Leng, Ruimin Hu, Chao Liang, Yimin Wang, and Jun Chen. Person re-identification with content and context re-ranking. *Multimedia Tools Appl.*, 74(17):6989–7014, September 2015.

[47] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, Dec 2016.

[48] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[49] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006.

[51] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dezhi Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. volume 9911, pages 135–153, 10 2016.

[52] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[53] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[54] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[55] Bin Tian, Ming Tang, and Fei-Yue Wang. Vehicle detection grammars with partial occlusion handling for traffic surveillance. *Transportation Research Part C: Emerging Technologies*, 56:80–93, 2015.

[56] 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.

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

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

[59] Yiheng Zhang, Dong Liu, and Zheng-Jun Zha. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1386–1391. IEEE, 2017.

[60] 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.

[61] Yuqi Li, Yanghao Li, Hongfei Yan, and Jiaying Liu. Deep joint discriminative learning for vehicle re-identification and retrieval. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 395–399. IEEE, 2017.

[62] Yi Tang, Di Wu, Zhi Jin, Wenbin Zou, and Xia Li. Multi-modal metric learning for vehicle re-identification in traffic surveillance environment. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258. IEEE, 2017.

[63] Yi Zhou and Ling Shao. Cross-view gan based vehicle generation for re-identification. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 186.1–186.12. BMVA Press, September 2017.

[64] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(8):3794–3807, 2019.

[65] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[66] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[67] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[68] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[69] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. pages 368–377, 1999.

[70] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

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

[72] 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.

[73] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010.

[74] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.

[75] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[76] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[77] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[78] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[79] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[80] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[81] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*, September 2018.

[82] 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.

[83] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch feature erasing for person re-identification and beyond. *ArXiv*, abs/1811.07130, 2018.

A Qualitative Results

Figure 7: Qualitative comparison of using label smoothing on unseen Market1501 test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

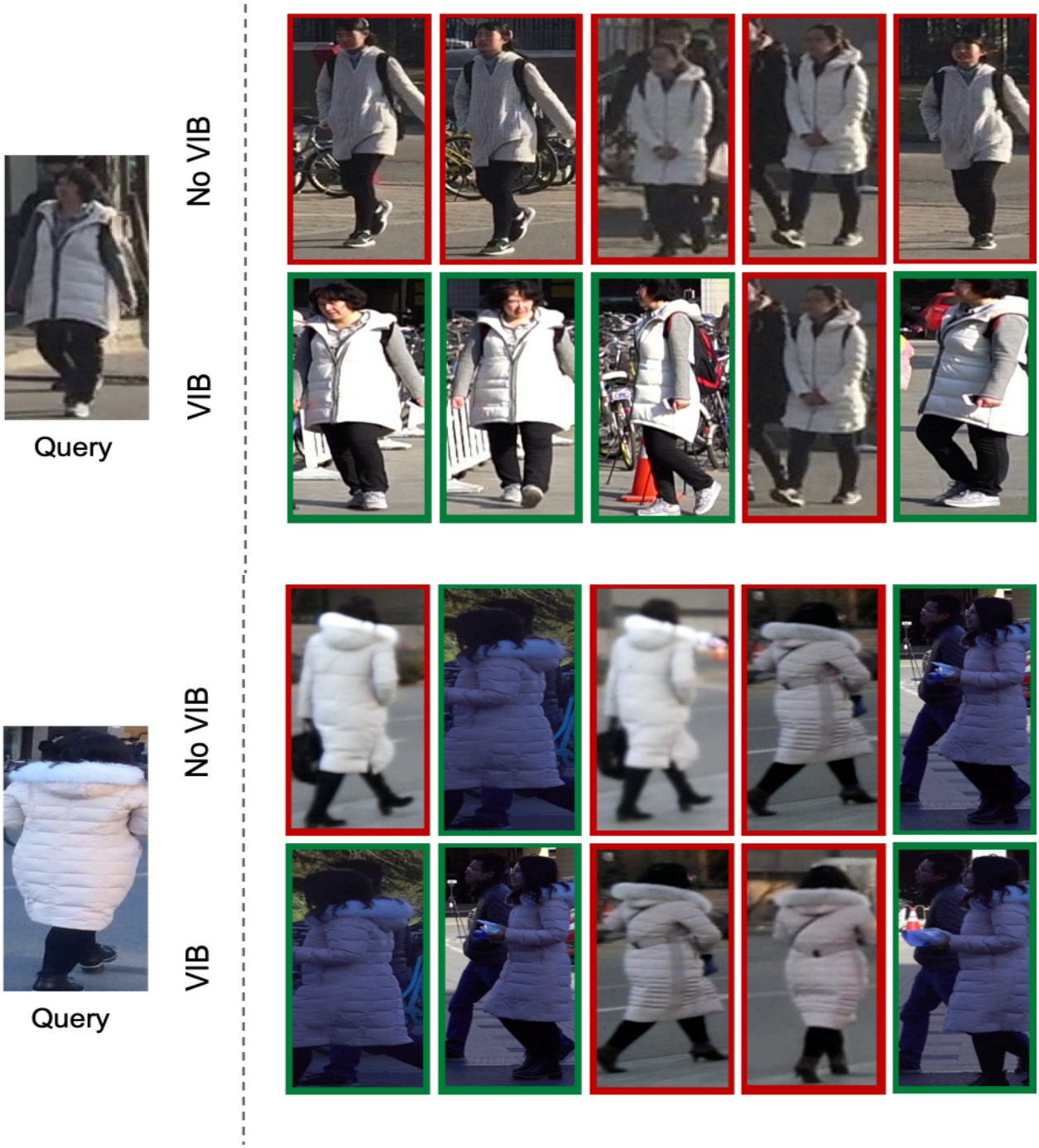

Figure 8: Qualitative comparison of using VIB on unseen Market1501 test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

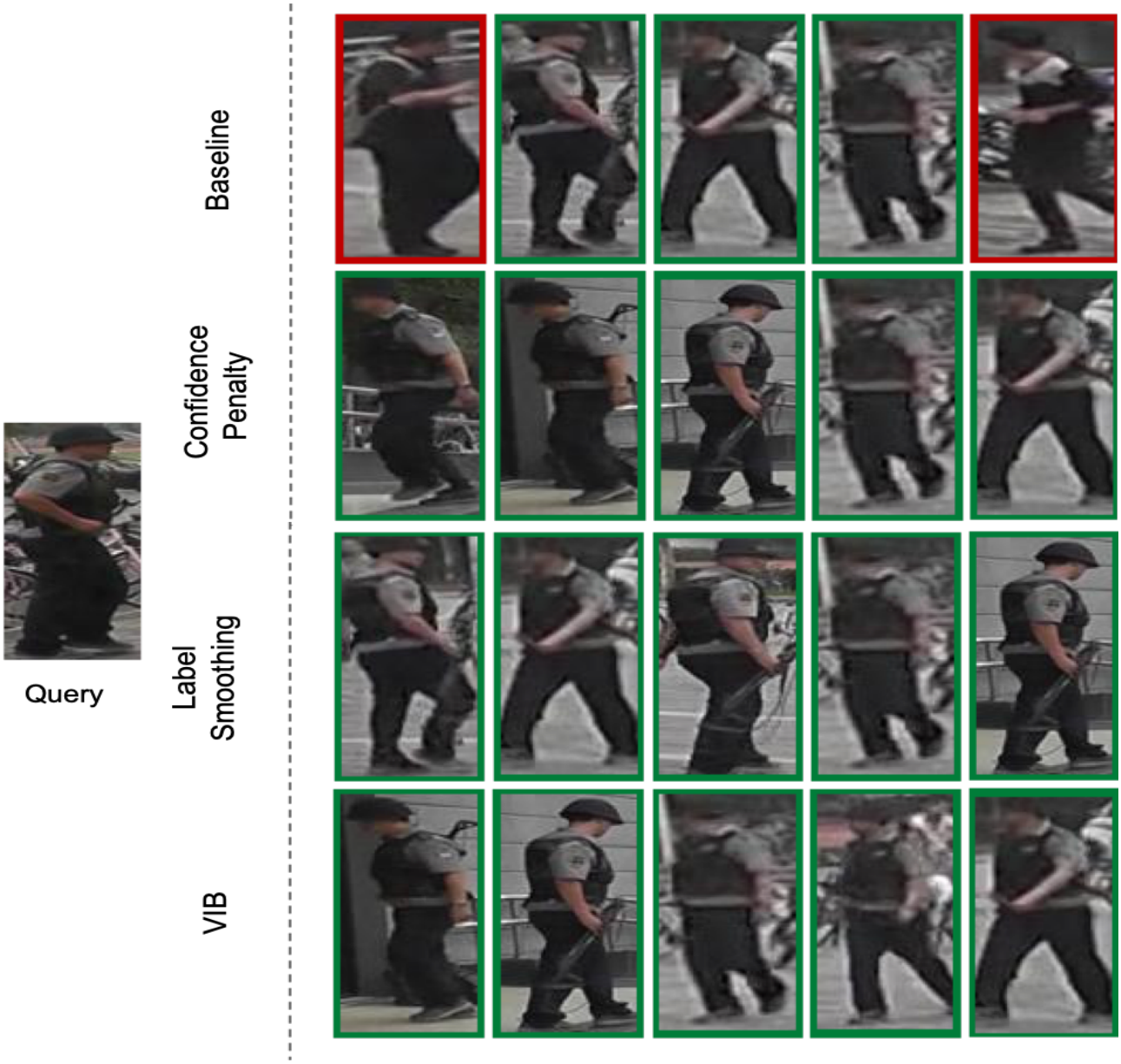

Figure 9: Qualitative comparison between the different methods on unseen Market1501 test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

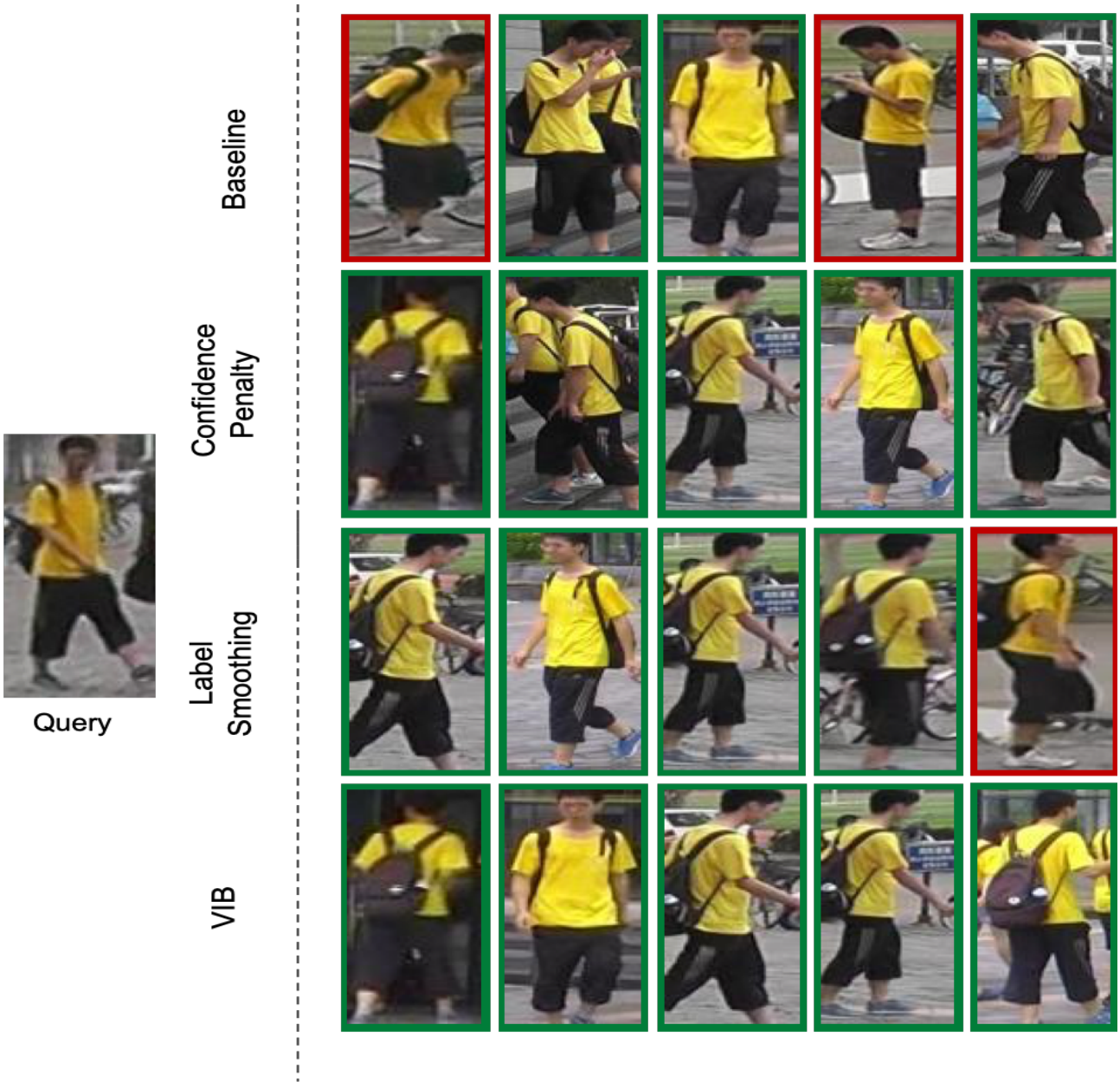

Figure 10: Qualitative comparison between the different methods on unseen Market1501 test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

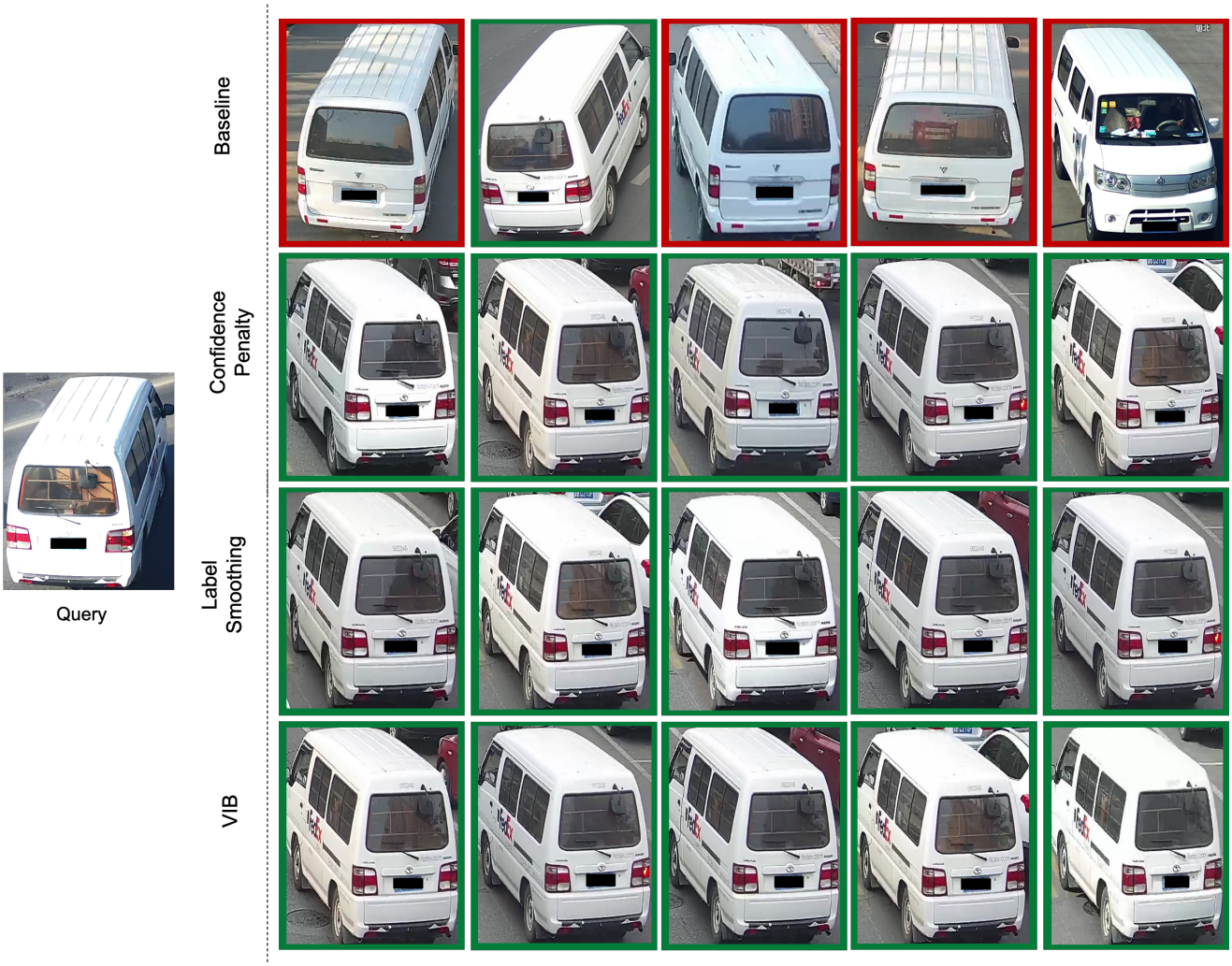

Figure 11: Qualitative comparison between the different methods on unseen VERI-Wild test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.

Figure 12: Qualitative comparison between the different methods on unseen VERI-Wild test samples. The gallery images are ranked according to L2 distance (top-5, left to right). Red frame indicates wrong ID while green frame indicates correct ID compared to the query. Best viewed in color.