Triplet Loss with Channel Attention for Person Re-identification

Daniel Organisciak, Chirine Riachy, Nauman Aslam, Hubert P. H. Shum†
Northumbria University
Department of Computer and Information Sciences
{daniel.organisciak, chirine.riachy, nauman.aslam, hubert.shum}@northumbria.ac.uk
† corresponding author

ABSTRACT

The triplet loss function has seen extensive use within person re-identification. Most works focus on either improving the mining algorithm or adding new terms to the loss function itself. Our work instead concentrates on two other core components of the triplet loss that have been under-researched. First, we improve the standard Euclidean distance with dynamic weights, which are selected based on the standard deviation of features across the batch. Second, we exploit channel attention via a squeeze and excitation unit in the backbone model to emphasise important features throughout all layers of the model. This ensures that the output feature vector is a better representation of the image, and is also more suitable to use within our dynamically weighted Euclidean distance function. We demonstrate that our alterations provide significant performance improvement across popular re-identification data sets, including almost 10% mAP improvement on the CUHK03 data set. The proposed model attains results competitive with many state-of-the-art person re-identification models.

Keywords

Person Re-identification, Squeeze and Excitation, Triplet Loss, Metric Learning, Siamese Network, Channel Attention, Weighted Euclidean

1 INTRODUCTION

Person re-identification (re-ID) is a core challenge within computer vision where an identity observed in one camera is required to be matched with another observation from a different viewpoint. This task has attracted a lot of interest due to the potential applications in the real-world as an increasing volume of large-scale urban surveillance data is collected.

In the past few years, convolutional neural networks (CNNs) have become ubiquitous within person re-ID due to significantly improving upon the state-of-the-art results [1, 2, 3]. Many person re-ID works make use of the standard convolutional neural networks with a classification loss [4]. More specific to re-ID, however, is the use of the triplet loss function [2, 5, 6, 7], either in place of or alongside the standard cross-entropy loss. The triplet loss, shown in Figure 1, enforces a distance margin, α, between the set of images of one person and all other images.

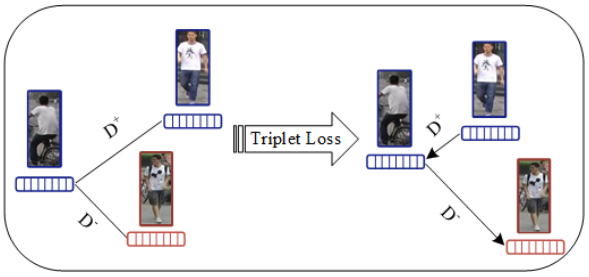

Figure 1: The triplet loss aims to reduce the distance of feature vectors from similar identities and increase the distance of feature vectors from dissimilar identities. We use channel attention in the form of squeeze and excitation units to get a better feature representation and improve the Euclidean distance by adding dynamic weights for each feature.

To date, most triplet loss works focus on mining better samples to improve the model generalisation [2, 10], or alter the loss function in order to increase the inter-class variance and decrease the intra-class variance [5, 7]. We identify two additional, under-researched lines of work to improve the triplet loss: improving the feature vectors obtained from the deep learning architecture by exploiting squeeze and excitation (SE) units, and adding dynamic weights to the distance function with which the triplet loss compares these feature vectors. We show that these alterations improve re-ID precision individually. When implemented together, these adjustments

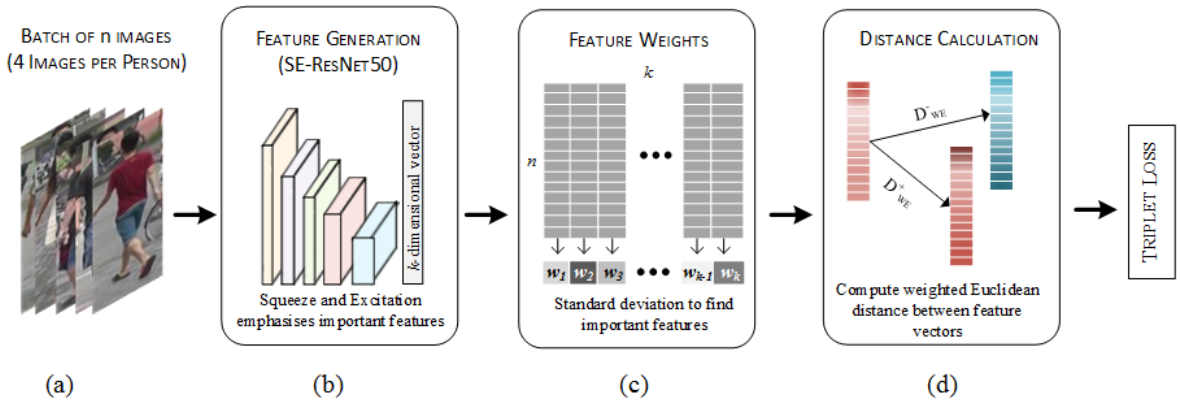

Figure 2: An overview of our architecture: (a) an input batch of n images is generated, (b) the batch is processed by SE-ResNet50 [8, 9] to generate one feature vector per image, (c) the standard deviation for each feature is computed then normalised to attain weights, (d) our improved triplet loss processes the mined triplets.

complement each other, resulting in a performance improvement of over 9% mAP on the CUHK03 data set compared to the regular triplet loss.

Distance: The triplet loss, by its nature, attempts to decrease the distance between positive pairs of images while increasing the distance of negative ones. However, to date, little research has been done to assess exactly how this distance should be formulated. The Euclidean distance has been shown to perform well within the triplet loss function, thus has not received much scrutiny. We show that adding dynamic weights to the Euclidean distance can deliver considerable benefit when applied to the task of person re-ID.

The standard Euclidean distance considers all features as equally important. As shown in Figure 2 (c), our dynamically weighted Euclidean distance assigns an importance score to each feature derived from a feature's batch-wise standard deviation. Features with higher variance are more informative, thus assist the model to distinguish between images of different identities. To conceptualise this idea, if everyone in a batch wears a plain, white t-shirt, it is impractical to consider this information for re-ID. We assess the batch-wise feature vectors for high-level features that act in this manner and diminish their importance while highlighting more useful features.

Features: We would like our backbone architecture to generate feature representations of images which can best be exploited by the dynamically weighted Euclidean distance. In order to achieve this, we use channel attention by adding SE units into our framework. These units act as weights to magnify important channels at each layer of the network while depreciating the value of less important channels. At deeper layers, these weights become more polarising to ensure salient features derived from the important channels are distinguishable from less important features.

As the less important features are mapped towards 0 by the SE units, they are more likely to have a low standard deviation and will therefore be assigned small weights by our dynamically weighted Euclidean distance.

The main contributions of this paper are as follows:

1. *Dynamically Weighted Euclidean Distance for Triplet Loss Feature Accentuation:* We introduce a weighted Euclidean distance which highlights features with high variation across the batch, in order to disregard features which are unimportant or susceptible to noise. This alone provides consistent performance improvement across all tested data sets.

2. *Feature Vector Generation with Channel Attention:* We are the first to adopt SE-ResNet 50 as the backbone architecture for the triplet loss. We demonstrate that the channel attention that SE units provide significantly boosts the performance of the triplet loss across a variety of data sets.

The rest of the paper is organised as follows: Section 2 contains an overview of work related to this paper. Section 3 details the formulation of our dynamically weighted Euclidean distance and explains how we use channel attention through SE units to boost the performance of the triplet loss. Section 4 contains our experimental results. Section 5 concludes the paper and discusses potential future directions that this work opens up.

2 RELATED WORK

Traditionally, popular methods for person re-ID comprised of two components: designing hand-crafted features [11] and learning distance metrics [12]. Hand-crafted features were required to be robust to variations in light, pose and viewpoint while using conventional distance metrics like the Mahalanobis distance [13],

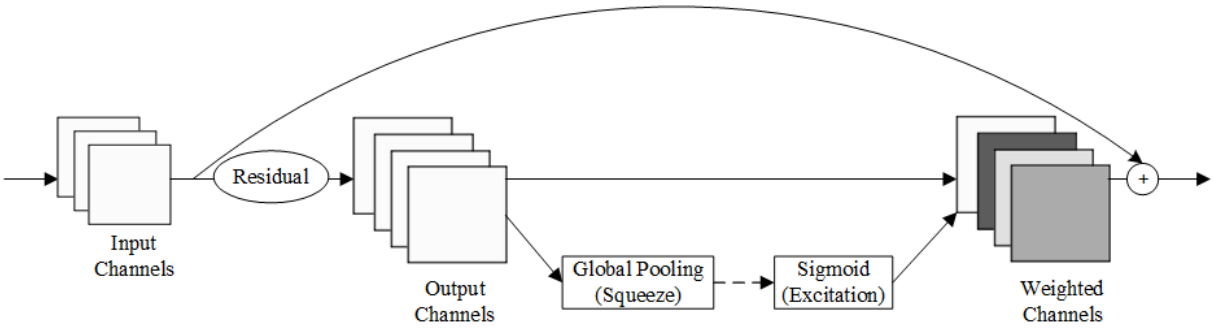

Figure 3: An overview of a ResNet block with a squeeze and excitation unit.

Bhattacharyya distance, and the l_1- and l_2-norms. In this respect, our work can be seen to be similar to this category of works, but within a deep learning context. We wish to improve the feature representations that are output by the backbone architecture, and also develop a better distance metric to compare these representations within a triplet loss setting.

2.1 Backbone Architecture

ResNet50: The majority of person re-ID works use the 50 layer variant of ResNet [9] as the backbone architecture. One possible reason for ResNet50 being ubiquitous in re-ID specifically is that mid-level features are relatively important compared with other fields. The skip connections in ResNet ensure that these important mid-level features have more presence in the features output by later levels compared with other popular backbones. Yu *et al.* [14] concatenate features from earlier ResNet layers with the final layer outputs to obtain a better representation. Zeng *et al.* [15] achieve state-of-the-art results with a hierarchical deep learning feature, which fuses features from several earlier layers, and define a new metric to best exploit this new feature.

Random Erasing [16] is a data augmentation technique that randomly removes a small area of each image in the input batch before processing them with ResNet. Sun *et al.* [17] use ResNet50 to learn discriminative features which are informed by 'parts' from the input image. Sun *et al.* [18] use Singular Vector Decomposition within ResNet to optimise the deep representation learning process. Due to its proven success, we also use ResNet50. We incorporate SE units to inform the network which channels are most important at each layer. This carries through to the output feature vector and allows us to tailor our loss function to identify the most salient features to assign more weight.

Squeeze and Excitation Networks: Squeeze and Excitation Networks [8] are frameworks that incorporate a squeeze and excitation unit at all or some layers of the architecture. Channels are obtained in a convolutional neural network for each filter learned by the CNN. Channel-wise spatial information is first 'squeezed' into per-channel descriptor to assess the relative importance of each channel. This information is then passed through a gating mechanism to 'excite' the descriptor. The original channels are then multiplied by their respective channel descriptor obtained from the squeeze and excitation process.

To date, SE units have seen limited use in person re-ID. Wang *et al.* [19] adapt SE Units as part of a fully attentional block. Li *et al.* [3] combine channel attention with spatial attention and part-wise attention to create their Harmonious Attention CNN. To the best of our knowledge, we are the first to use SE units purely as a backbone architecture to exploit important features that are generated as part of the output feature vector.

2.2 Triplet Loss

The triplet loss function has been used extensively for person re-ID due to its proven ability to attain state-of-the-art results [2, 10]. Traditional triplet models take three images as input: a query image, a positive image that has the same identity as the query, and a negative image that has a different identity to the query. A margin α is enforced to ensure a certain distance between positive and negative pairs.

Wang *et al.* [20] proposed to use the triplet loss function to learn image similarity. The triplet loss gained notoriety by significantly improving the state of the art for face verification [21, 22]. Since then, triplet loss research has typically focused on improving either the triplet mining algorithm or the loss function.

Building an effective triplet network is heavily reliant on the mining strategy. To challenge the framework to be able to handle tough cases, difficult triplets need to be mined, but choosing only the hardest triplets in the data set will result in a model that is not representative of the entire set of triplets. To strike the balance between finding difficult triplets while still generating a representative model, Hermans *et al.* [2] present *Batch Hard* mining, which selects only the hardest triplets

across each batch selected during training. In a similar manner, Almazan *et al.* [10] select triplets that start off relatively easy but get more difficult as training progresses.

Cheng *et al.* [5] introduce an improved triplet loss function that decreases the distance of images from the same identity whilst increasing the distance of images from a different identity. Chen *et al.* [7] add an additional term to the triplet loss to form a quadruplet loss. This term contains a second negative pair which helps to enlarge inter-class variations across the data set. Jiang *et al.* [23] demonstrate improved performance through adding a self-supervised attention loss to the quadruplet loss. While performance is enhanced by these works, they all focus on improving the same aspects of the triplet loss. We instead tackle the under-researched feature representation and the distance function.

3 METHODOLOGY

3.1 Triplet Loss Background

We formulate the triplet loss mathematically in order to provide motivation to investigate the distance metric and the feature representation.

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:

$$\begin{aligned}\mathscr{L}_{trp} = \sum_{t \in \mathscr{T}} \max((||f(x) - f(x^+)||_2 \\ - ||f(x) - f(x^-)||_2 + \alpha), 0),\end{aligned} \tag{1}$$

where the feature vector of an image x obtained from the convolutional neural network is denoted as $f(x)$, $\mathscr{T}$ is the set of mined triplets and $||\cdot||_2$ denotes the Euclidean distance. This loss will force negative images to be a distance of at least α away from the positive pair.

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

$$x^+ = \max_{x_{q,j} \in B} (||(f(x_{p,i}) - f(x_{q,j})||_2), \text{ where } p = q, \tag{2}$$

$$x^- = \min_{x_{q,j} \in B} (||(f(x_{p,i}) - f(x_{q,j})||_2), \text{ where } p \neq q. \tag{3}$$

From equations (1) - (3), we see that obtaining the feature representation $f(x)$, and computing the distance between feature representations of any two images, $||(f(x_1) - f(x_2)||$, are essential components of the triplet loss. We focus our research on improving these two aspects.

3.2 Dynamically Weighted Euclidean Distance

Although the triplet loss has seen extensive use in person re-ID, there has been little work to deviate from the standard Euclidean distance, despite it being a crucial element of the framework. We improve it by weighting each feature based on its importance.

To calculate which features are most discriminative, we use the $n \times k$ feature matrix output from the backbone of the network to calculate the standard deviation for each feature across the batch. This is shown in Figure 2 (c). The higher the standard deviation, the more variation in that feature, and the more effective it is at helping the framework to tell people apart. These more meaningful features should thus be assigned a greater weight.

We use a softmax function regularisation on the standard deviations, then multiply by the total number of features to obtain the final weights. Overall, the weight, w_i, for the i-th feature can be calculated as

$$w_i = \text{softmax}\left(\text{s.d.}(\mathbf{F_i})\right) \times k, \tag{4}$$

where $\text{s.d.}(\cdot)$ is the standard deviation and $F \in \mathbb{R}^{n \times k}$ is the batch-wise feature matrix output by the backbone of the model with features $i = 1, \ldots, k$.

To ensure that the more important features are more prominent when calculating the distance matrix, we use the weighted Euclidean distance, D_{WE}, between two feature vectors, $\mathbf{x}$ and $\mathbf{y}$:

$$D_{WE}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} w_i (x_i - y_i)^2}, \tag{5}$$

where w_i are the weights and k is the length of the feature vectors.

The standard triplet loss will separate embeddings to ensure that the distance between classes is greater than the hard margin α. Because we iteratively adjust the formulation of this distance, we are able to push classes apart even further, which leads to the model better representing the data.

3.3 Channel Attention Feature Embedding

The triplet loss evaluates the distance between feature representations, thus is very dependant on the quality of the feature vectors that are generated by the network. Furthermore, we would like these feature vectors to possess information which can be exploited by our dynamically weighted Euclidean distance.

We concentrate on improving the feature representations themselves by utilising channel attention via SE blocks [8] within ResNet50.

An SE unit is a mechanism that performs feature recalibration within the framework utilising it. By doing so, it selects features that are the most informative to the framework and accentuates them, while diminishing the importance of less useful features. These informative features then allow the re-id framework to create a better embedding which is more effective at separating classes.

In this regard, the SE unit acts as a process to determine the weight of each channel at each layer of the model, similarly to how our dynamically weighted Euclidean distance performs. The SE units perform different roles throughout the network, getting more polarising at deeper layers. As a consequence, unimportant channels are mapped near to 0 in the final block of ResNet, which has a large effect on the output feature representation of each image.

Note that as unimportant features are mapped towards 0 throughout the network, they will typically have a low standard deviation. On the contrary, important features will be less impacted by the SE units and are therefore more likely to have a higher standard deviation. This means that our dynamic weights will be much more likely give a large weight to features that are computed to be important by SE units, while still being able to identify features with high variance even though they are not determined to be salient by the network.

As we show in Figure 3, the unit first squeezes the channel-wise spatial information into a channel descriptor via Global Average Pooling. Formally, given a channel $u \in \mathbb{R}^{H \times W}$, we squeeze it to obtain its channel descriptor, c, as follows:

$$c = \text{squeeze}(u) := \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_{ij}, \qquad (6)$$

where H and W are the height and width of the channel respectively. These channel descriptors form a vector $\mathbf{z} = [c_1, \ldots, c_C]$ where C is the total number of channels.

Next, in order to calculate the channel-wise dependencies, this statistic needs to be excited. To achieve this, a simple gating mechanism with a sigmoid activation function is employed similarly to what is used within many spatial attention methods. The vector of squeezed channel descriptors $\mathbf{z}$ is passed through a dimensionality-reduction fully connected layer, a ReLU and then a dimensionality-increasing fully connected layer. This is then processed by a sigmoid activation to obtain the excited channel descriptors.

Formally, this excitation is written as:

$$\mathbf{s} = \text{excite}(\mathbf{z}) := \sigma\left(\mathbf{W_2}\delta(\mathbf{W_1}\mathbf{z})\right), \qquad (7)$$

where $\mathbf{W_1} \in \mathbb{R}^{\frac{C}{r} \times C}, \mathbf{W_2} \in \mathbb{R}^{C \times \frac{C}{r}}$ are the parameters of the dimensionality-reduction and dimensionality-increasing layers respectively, δ is the ReLU function and σ is the sigmoid activation function.

We show in Section 4 that this adaptation alone vastly improves the performance of the triplet loss on multiple data sets.

4 EXPERIMENTS

4.1 Evaluation Protocol

We perform experiments on the two most commonly used data sets to evaluate deep learning methods for person re-ID, CUHK03 [31] and Market-1501 [32]. In addition, we also provide results on VIPeR [33] to demonstrate that our method can considerably improve performance even on very small data sets, which many deep learning frameworks struggle on. We report the mean average precision (mAP) and top-1 matching rate (rank-1) scores for CUHK03 and Market-1501, and rank-1, rank-5 and rank-10 scores for VIPeR. Many works boost the performance of their framework with a post-processing technique such as re-ranking [34] or a data augmentation procedure like random erasing [16]. Although our results would improve, we do not use re-ranking or random erasing in any of our experiments as it does not help to evaluate the core performance of the network.

Throughout our experiments on CUHK03 and Market-1501, we use a batch size of $n = 96$ with four images per person, while on VIPeR we use a batch size of $n = 32$ with two images per person. The feature representations, $f(x)$, contain $k = 2048$ features in all of our experiments. We fix the margin $\alpha = 0.3$.

Note that, to fairly compare the effect of the dynamically weighted Euclidean distance, we use it within the triplet loss function but don't apply it during the mining phase.

All experiments are performed on a single NVIDIA GeForce GTX 1070 Ti GPU. Our model takes around 1 hour, 30 minutes to train on Market-1501 and around 35 minutes to train on CUHK03. We note that the system can further be optimised by tuning hyperparameters.

CUHK03: The CUHK03 data set contains 14297 bounding boxes of 1467 persons, with 767 identities used for training and 700 identities used for testing.

CUHK03 has two evaluation settings: the *labelled* setting contains bounding boxes that are manually annotated, and the *detected* setting contains bounding boxes that are automatically detected. We perform all of our experiments on the detected setting as it is a more realistic setup, which contains misplaced bounding boxes making the problem more challenging. It is more similar to what we would expect when applying re-ID to real-world tasks.

Market-1501: The Market-1501 data set contains images of 1501 people from six different cameras. The data set is split into 12936 images of 751 identities for

Comparison with baseline methods							
Data Set	**CUHK03**		**Market-1501**		**VIPeR**		
Method	mAP	rank-1	mAP	rank-1	rank-1	rank-5	rank-10
ResNet50	26.3	26.6	68.3	85.8	11.1	32.6	44.0
SE-ResNet50	37.8	38.6	72.4	87.9	17.4	40.8	51.3
TriNet*	48.8	51.4	67.9	83.4	38.3	67.7	80.4
Ours: DWE TriNet	54.8	56.1	69.7	84.2	39.2	73.7	83.2
Ours: SE TriNet	52.9	54.7	73.1	**88.1**	40.2	69.6	80.4
Ours: SE+DWE TriNet	**58.2**	**60.7**	**74.2**	88.0	**44.9**	**75.6**	**86.1**

Table 1: Comparison with baseline methods. *Trained with a hard margin $\alpha = 0.3$.

training and 19732 of 750 identities for testing. We use the single query setting throughout all of our experiments.

VIPeR: The VIPeR data set consists of 632 pedestrians captured by two cameras. Deep learning methods typically do not report performance for this data set so we train all models ourselves with a batch size of 32. As VIPeR only contains one image per person in each camera, we replace the mAP metric with rank-5 and rank-10 precision scores.

4.2 Comparison with Baseline Methods

We present our results with the baseline methods in Table 1. We select the baselines as ResNet50 [9] and SE-ResNet50 [8] with a cross-entropy loss, and TriNet [2] as our model is comprised of these elements. All triplet loss models in Table 1 are trained with the hard margin $\alpha = 0.3$ for direct comparison.

We comprehensively outperform baseline methods across all data sets. In particular, on CUHK03, we enhance the mAP of the triplet loss by 9.4% and the rank 1 performance by 9.3%. We also demonstrate considerable performance improvement on Market-1501 and VIPeR. The results show that both elements of our framework provide a significant contribution to enhance the re-ID precision.

CUHK03 (767/700) split		
Method	mAP	rank-1
DPFL [4]	37.0	40.7
SVDNet[18]	37.2	41.5
HACNN [3]	38.6	41.7
MLFN [24]	47.8	52.8
TriNet [2]	48.8	51.4
TriNet + RE [16]	50.7	55.5
DaRe [25]	51.3	55.1
PCB* [17]	57.5	63.7
HPM* [26]	57.5	63.9
MGN* [27]	**66.8**	**66.0**
Ours: DWE TriNet	54.8	56.1
Ours: SE TriNet	52.9	54.7
Ours: SE+DWE TriNet	58.2	60.7

Table 2: Comparison with baseline methods on the CUHK03 data set with the new split. *Use part-based information

Market-1501 (Single Query)		
Method	mAP	rank-1
DeepTransfer [28]	65.5	83.7
JLML [29]	65.5	85.1
TriNet [2]	67.9	83.4
TriNet + RE [16]	71.3	87.1
DaF [30]	72.4	82.3
DPFL [4]	73.1	88.9
HACNN [3]*	75.7	91.2
PCB* [17]*	**81.6**	**93.8**
Ours: DWE TriNet	69.7	84.2
Ours: SE TriNet	73.1	88.1
Ours: SE+DWE TriNet	74.2	88.0

Table 3: Comparison with baseline methods on the Market-1501 data set with the single query setting. For fair comparison, we don't include results which use re-ranking. *Use part-based information

4.3 Comparison with State of the Arts

We further compare with state-of-the-art models (without re-ranking or random erasing) on the selected three data sets. In particular we note that our simple alterations are enough to give us the second highest mAP score of any core framework on the CUHK03 data set. We also notice that the state-of-the-art deep learning methods struggle to compete with ours on a small data set such as VIPeR, which demonstrates the robustness of our model.

CUHK03: Our results on the CUHK03 data set can be found in Table 2. It can be observed that the weighted Euclidean significantly boosts the performance on CUHK03.

We attain the second highest performance across all models on mAP. Our simple alterations are shown to outperform very sophisticated, state-of-the-art models that exploit spatial attention. In particular, we outperform the state-of-the-art methods PCB [17] and HPM [26]. The only method that exceeds ours, MGN, is heavily engineered. It takes different sized portions of

VIPeR			
Method	R1	R5	R10
MLFN [24]	28.2	50.9	62.3
TriNet [2]	38.3	67.7	80.4
PCB* [17]	41.1	70.3	84.5
TriNet + RE [16]	41.8	71.2	83.5
Ours: DWE TriNet	39.2	73.7	83.2
Ours: SE TriNet	40.2	69.6	80.4
Ours: SE+DWE TriNet	**44.9**	**75.6**	**86.1**

Table 4: Comparison with popular deep learning methods on the VIPeR data set. *Use part-based information

the original image as input, which has been shown by multiple works to substantially improve performance. We note that (i) we can add this technique to our framework, (ii) they use a triplet loss in their model, which could be improved by adopting our formulation.

The most appropriate state-of-the-art method from Table 2 for comparison is Random Erasing [16], as it has become one of the most popular techniques within re-ID and also uses a triplet loss. Our method comprehensively outperforms it, improving on its rank-1 accuracy by 10%. We further note that even if we keep the backbone architecture as ResNet50, simply changing the Euclidean distance function to our dynamically weighted Euclidean distance boosts performance more than Random Erasing. This further demonstrates the significance of the enhancements we have implemented and that the distance formulation is a crucial component which should not be overlooked when developing a triplet loss framework.

Market-1501: The results on the Market-1501 data set are presented in Table 3. We see that including squeeze and excitation blocks within the backbone architecture and adding dynamic weights into the Euclidean distance both enhance the framework. Our modified framework exceeds many state-of-the-art methods.

We note that although DPFL [4] and HACNN [3] beat us on Market-1501, their results on CUHK03 are much weaker, which indicates their models are heavily optimised towards the Market-1501 data set and not capable of generalising well. PCB [17] uses a part-based method which, as previously discussed, substantially improves performance and is compatible with our framework.

VIPeR: We outperform the state-of-the-art deep learning methods by 3.1% on the rank-1 matching rate. This demonstrates that our enhancements are very robust, even on data sets that do not have enough data for deep learning. In particular, we see that methods such as MLFN [24], despite performing well on popular deep learning data sets, do not have the ability to generalise as well as ours.

5 CONCLUSION

In this paper, we have evaluated the effects of feature saliency on the triplet loss function. We achieved this in two different ways: via assigning dynamic weights into the distance function used by the triplet loss, and by incorporating a backbone architecture with channel attention to emphasise important features throughout training. We demonstrate that both alterations alone boost performance of the triplet loss and complement each other for a significant improvement in precision when used together.

It has been shown recently that spatial attention or part-based understanding can dramatically improve the performance of re-ID frameworks. Our method is complementary to these part-based approaches, in the sense that we can apply our weighted Euclidean distance to the part-based feature vector obtained from their framework. One of our future directions is to use spatial attention to improve the selected parts that are fed into these systems before processing them with the improvements that we have described in this paper.

This paper demonstrates that using a simple mechanism to determine distance function weights works very well. More sophisticated strategies such as learning the weights concurrently with the feature representations could be adopted and are planned as future work.

ACKNOWLEDGEMENT

This project was supported in part by the Royal Society (Ref: IES\R2\181024).

6 REFERENCES

[1] L. Zheng, Y. Yang, A. G. Hauptmann, Person Re-identification: Past, Present and Future (8) 1–20. arXiv:1610.02984.

[2] A. Hermans*, L. Beyer*, B. Leibe, In Defense of the Triplet Loss for Person Re-Identification, arXiv preprint arXiv:1703.07737.

[3] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[4] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, 2017.

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. 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), 2016.

[6] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, in:

Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, 2017, pp. 3988–3994.

[7] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: A deep quadruplet network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385.

[10] J. Almazan, B. Gajic, N. Murray, D. Larlus, Re-ID done right: towards good practices for person re-identificationarXiv:1801.05339, doi:10.1109/ICCV.2013.228.

[11] C. Riachy, A. Bouridane, Person re-identification: Attribute-based feature evaluation, SAMI 2018 - IEEE 16th World Symposium on Applied Machine Intelligence and Informatics Dedicated to the Memory of Pioneer of Robotics Antal (Tony) K. Bejczy, Proceedings 2018-February (2018) 85–90. doi:10.1109/SAMI.2018.8323991.

[12] W.-S. Zheng, G. Shaogang, X. Tao, Person re-identification by probabilistic relative distance comparison, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (2011) 649–656doi:10.1109/cvpr.2011.5995598.

[13] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, H. Bischof, Mahalanobis Distance Learning for Person Re-identification, Springer London, London, 2014, pp. 247–267.
URL https://doi.org/10.1007/978-1-4471-6296-4_12

[14] Q. Yu, X. Ching, Y.-Z. Song, T. Xiang, T. M. Hospedales, The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching (3). arXiv:arXiv:1711.08106v2.

[15] M. Zeng, C. Tian, Z. Wu, Person Re-identification with Hierarchical Deep Learning Feature and efficient XQDA Metric (2018) 1838–1846doi:10.1145/3240508.3240717.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data AugmentationarXiv:1708.04896.

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: The European Conference on Computer Vision (ECCV), 2018.

[18] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: The IEEE International Conference on Computer Vision (ICCV), 2017.

[19] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in: The European Conference on Computer Vision (ECCV), 2018.

[20] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[21] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[22] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference (BMVC), 2015.

[23] M. Jiang, Y. Yuan, Q. Wang, Self-attention Learning for Person Re-identification, Bmvc.

[24] X. Chang, T. M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[25] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, K. Q. Weinberger, Resource aware person re-identification across mutliple grane resolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[26] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, AAAI.

[27] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, MM '18, 2018, pp. 274–282.

[28] M. Geng, Y. Wang, T. Xiang, Y. Tian, Deep Transfer Learning for Person Re-identificationarXiv:1611.05244.

[29] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, 2017, pp. 2194–2200.

[30] R. Yu, Z. Zhou, S. Bai, X. Bai, Divide and Fuse: A Re-ranking Approach for Person Re-identification 1–13arXiv:1708.04169.

[31] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Computer Vision, IEEE International Conference on, 2015.

[33] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision – ECCV 2008, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 262–275.

[34] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.