A Tale of Two Losses: Discriminative Deep Feature Learning for Person Re-Identification

A Tale of Two Losses: Discriminative Deep Feature Learning for Person Re-Identification.

Alessandro Borgia[1,2], Yang Hua[3,4], Neil M. Robertson[3,4]

[1] *ISSS/EPS, Heriot-Watt University, UK*
[2] *SIP-JRI, University of Edinburgh, UK*
[3] *ECIT, Queen's University Belfast, UK*
[4] *AnyVision, Belfast, UK*

Abstract

The changing camera viewpoint on full-body pedestrians in a multi-camera scenario may be problematic to handle, above all if the fields of view are non-overlapping. A direct effect of the viewpoint variability is that a pair of images of the same person shot by different cameras may appear to be more distant from each other in the feature space than one of them from an image of a different identity captured by the same camera. In order to tackle this problem, we propose to train a state-of-the-art CNN by two new loss functions that jointly increase the inter-class discriminative power of the deep features and their intra-class compactness. In particular, one loss function promotes the aggregation of the feature points around the centres of the view they belong to, within the scope of their own identity. The second loss encourages to push away from each other the feature clusters corresponding simultaneously to different views and different identities. Under the supervision of the two new objectives we achieve state-of-the-art accuracy with ResNet50 on Market-1501 and CUHK03 datasets, beating the performance of the softmax loss.

Keywords: Person re-id, Loss function, Multi-camera net, Changing viewpoint, Discriminative deep features.

1 Introduction

In this paper we investigate the problem of how the changing viewpoint in a multi-camera network with non-overlapping fields of view affects the performance of the person re-identification task. Traditionally, the person re-id problem is tackled using three kinds of approaches: **(1)** *feature design*, dealing either with hand-crafted feature modelling [Zhao et al., 2013, Zhao et al., 2014] or with deep feature extraction approaches [Varior et al., 2016b, Xiao et al., 2016b, Xiao et al., 2016a]. Hybrid solutions are also possible [Wu et al., 2016]. **(2)** *metric learning* [Yi et al., 2014a, Chen et al., 2012, Kulis, 2013, Hoffer and Ailon, 2015], that relies on learning a distance function tuned to a particular task in the feature space. **(3)** *side information*, that is hidden information encoded into the input data. Common strategies include target alignment and full-body pose estimation [Zheng et al., 2017, Pishchulin et al., 2016]. In some cases, strategies of different kind are used together as in [Bak et al., 2015] where target alignment and pose estimation are combined with metric learning to produce better performing pose-driven metric learning schemes. Our method belongs to the first group, embracing the deep learning (DL) approach, by virtue of the relevant impact that it has taken in person re-id. In order to enhance the discriminative power of the deep features, we assert the importance of *influencing the construction itself of the feature space*: we aim to do that by a training loss capable of reproducing the semantic similarity of images in the input space. This is different from using a metric learning approach where a metric is learned in the feature space already generated. The target of making the features extracted by a Convolutional Neural Network (CNN) more discriminative has already been pursued by [Wen et al., 2016] in face recognition. It proposes a centre loss function that, promotes the compactness of the clusters of features around the centre point of each face, thus achieving state-of-the-art results on many benchmarks.

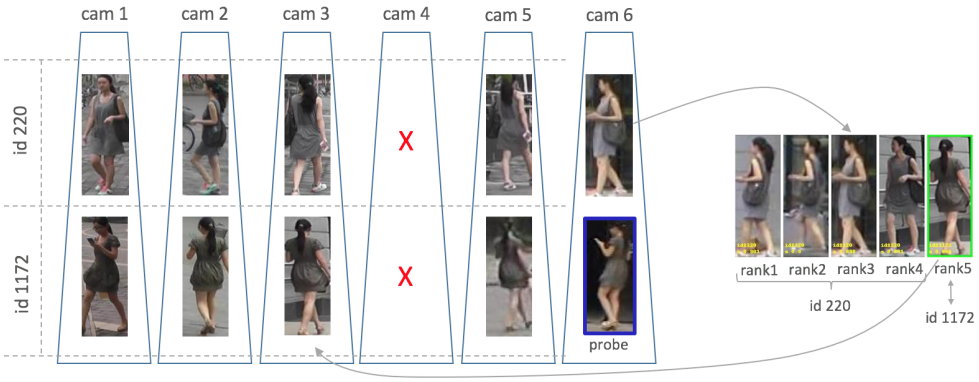

Figure 1: Illustration of the changing viewpoint problem in the multi-camera setting (Market-1501, 6 cameras) when re-id is addressed as a ranking problem. We can see that the image ranked 3^{rd} (which is a false positive since it belongs to a different identity to the probe's) is ranked by the CNN higher than the first right match (ranked only 5^{th}). This happens because the false positive is captured under the same field of view of the probe (cam.6), differently from what happens for the true positive (cam.3).

Respect to face recognition, the person re-identification task, though, is somewhat different: person re-id performance are heavily affected by deceptive viewpoint changes due to the disjoint nature of the multi-camera network for pedestrians, which we do not observe in face recognition. The viewpoint variability may cause images of two different pedestrians observed under the same camera to look (and be ranked as) more similar between each other than respect to their correspondent shots taken under other views. In order to disentangle this kind of ambiguity in person re-id, the capability we need to equip our CNN with is *learning intra/inter-camera relationships* and exploiting them to discriminate better on the base of pedestrian identity. Hence, producing highly discriminative features is our ultimate goal that we pursue by introducing new training objectives. A direct application of the centre loss to the person re-id problem, would not be that beneficial because it does not take into account the information of the field of view under which pedestrian images are captured, which is available in the person-re-id datasets. Starting from this observation, we propose to adapt the centre loss for person re-id by introducing two new loss functions, the *intra-Group Centre Loss* (**intra-GCL**) and the *inter-Group Centre Loss* (**inter-GCL**). The first one represents a direct extension of the centre loss that exploits the field of view information: it penalizes a large distance of each feature point from its related centre point but, instead of considering only one centre point per identity (like centre loss does). It refers as many centres per identity as the number of views available for one identity. Hence, it encourages the intra-identity view-based feature clusters to be compact. On the other side, the second loss function encourages a large inter-subclass separation of the view-field-based clusters belonging to different identities respect to the inter-subclass separability within a single identity, which is aligned with the requirement of person re-id evaluation. The main contributions of this paper can be summarized in the following:

- We propose two new loss functions for learning discriminative deep features that prove to be effective in mitigating the changing viewpoint problem, in the multi-camera setting with disjoint fields of view.
- We achieve state-of-the-art performance on Market-1501 and CUHK03, without employing extra training data (like by training on multiple-datasets as in [Xiao et al., 2016a]) and other side information.

2 Related Work

Hand-Crafted and Hybrid Feature. Before the DL approach became popular, the field of person re-id was dominated by approaches based on designing hand-crafted features that defined the state-of-the-art. LAB colour histograms are extracted from image patches in [Zhao et al., 2013, Zhao et al., 2014] and combined with SIFT

descriptor as a complementary feature. A Bag-of-Words (BoW)-based descriptor is used in [Zheng et al., 2015] in order to bridge the gap between person re-id and image search. Some strengths of this method are that it well accommodates local features and enables fast global feature matching.

Deep Features. Recently, lots of works addressing person re-identification have adopted the DL paradigm, exploiting the availability of new large-scale datasets like Market-1501 ([Zheng et al., 2015]) and CUHK03 ([Wang, 2014]). [Yi et al., 2014b] is the first work to apply DL to person re-id by designing a "siamese" CNN for deep metric learning relying on the cosine similarity as connecting function. Later approaches, following the same direction as [Varior et al., 2016a, Varior et al., 2016b], confirm the success of the siamese CNNs intuitively due to the fact that they force to learn the relationship existing between different camera viewpoints. A new promising view addressing person re-identification in an end-to-end framework is presented in [Xiao et al., 2016b] that jointly handles pedestrian detection and searching in one unified trainable net. Despite it achieves impressive results on real-world street snaps and movies its generic Faster R-CNN-based detector brings some problems because it limits the depth of the following re-id feature extraction subnet. A remarkable approach is introduced by [Zheng et al., 2017] which performs a misalignment correction by Convolutional Pose Machines and combines the corrected and the original features together to enhance the overall discriminative power. The feature extraction part relies entirely two parallel CNNs. In these works the training is performed for identity classification but the learned network is employed for re-id feature extraction according to the transfer learning principle, borrowing the idea from [Sun et al., 2014] in face verification/recognition.

Hybrid Feature. Hybrid approaches combine hand-crafted features with the "learning from data" paradigm: [Wang, 2014] and [Ahmed et al., 2015], for example, design siamese CNNs with constraints on the shape of the objective to learn by adding hand-crafted layers. A different hybrid strategy is applied in [Wu et al., 2016] where a new feature extraction model produces more discriminative features by fusing together convolutional and hand-crafted histogram features, complementary to some degree.

Training Losses. One of the most effective losses used for training CNNs in person re-id is represented by the softmax function when included into the siamese network model. Two recent works exploiting this configuration are [Varior et al., 2016b] and [Varior et al., 2016a] that consider ways of exploiting spatial relations of images (within a single image or between image pairs). A different loss to the one used in siamese CNNs, the "triplet loss" is adopted by the triplet network model in [Hoffer and Ailon, 2015] with the effect of gaining insensitivity (differently from siamese CNNs) to context calibration. On the other side, this is paid in terms of training complexity and slow convergence caused by the explosion of the number of samples. One limitation of both the siamese and triplet models is that they only rely on weak re-ID labels (same id or different id). A modification of the softmax loss, the "random sampling softmax loss, is proposed in "[Xiao et al., 2016b] and allows to supervise the training with sparse and unbalanced labels.

3 Proposed Method

Aiming to mitigate the impact of viewpoint changes on the performance by enhancing the discriminative features of the pedestrian images, we propose the intra-GCL and the inter-GCL. In order to clarify the role that the two objectives play in the training process, let us call *sub-class* each of the feature clusters belonging to an identity that corresponds to a particular field of view (e.g the blue triangles in Fig.2). There are up to 6 sub-classes for an identity in Market-1501 and always 2 in CUHK03).

$$L_{intra} = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i^{g_i} - \boldsymbol{c}_{y_i}^{g_i}\|_2^2 \qquad (1)$$

$$L_{inter} = \sum_{i}^{m} \frac{\sum\limits_{\substack{g=1 \\ g\neq g_i}}^{s} \|\boldsymbol{x}_i^{g_i} - \boldsymbol{c}_{y_i}^{g}\|_2^2}{\sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1 \\ g\neq g_i}}^{s} \|\boldsymbol{x}_i^{g_i} - \boldsymbol{c}_{t}^{g}\|_2^2} \qquad (2)$$

Intra-Group Centre Loss. It aims to encourage all the multi-dimensional points belonging to a sub-class of an identity to get near to the centre of that sub-class itself, gaining in compactness. It is formulated in Eq. (1) where $\boldsymbol{x}_i^{g_i}$ is a multi-dimensional point in the feature space, y_i is the ground-truth label corresponding to the i^{th} mini-batch point, g_i is the sub-class to which belongs the i^{th} mini-batch point, m is the training mini-batch size. The intra-GCL needs to be applied in joint supervision with the softmax loss: the presence of the softmax avoids that the centres degenerate to the null solution, while the presence of the intra-GCL avoids that the deep features contain too much intra-class variations. A week aspect of the intra-GCL is that, in the multi-camera scenario where the changing viewpoint problem may occur (Fig. 1), it may bring together images of different people sharing the same camera viewpoint, which is the reason why it requires to be balanced by the inter-GCL.

Inter-Group Centre Loss. The inter-GCL aims to penalize the distances of the image representation currently contributing to the training from all the centres of the sub-classes belonging to its identity, except its own. On the other side, it pushes away from the current feature point all the sub-classes referring to a different view and belonging to the rest of the training set identities. The inter-GCL is formulated in Eq. (2) where $\boldsymbol{c}_{y_i}^{g}$ is the centre of the sub-class g of identity y_i, s is the maximum number of cameras in the dataset and n is the number of identities in the training set. The reason why in the summation at the numerator in Eq. (2) we do not include all the terms $\|\boldsymbol{x}_i^{g_i} - \boldsymbol{c}_t^{g_i}\|_2^2$ is because they are already accounted separately in Eq. (1). The underlying assumption about this formulation is that pedestrians captured by the same camera present similar poses: the more it is true for a specific dataset the larger improvement is expected from the joint training supervision. One situation that this loss might not handle properly, though, is a possible initialization where a few image-centre distances (Fig. 2) contributing to the summation in the denominator of Eq. (2) take much larger values than the rest of the summation terms.

Combined With Softmax Loss. The two losses work in combination with the softmax loss. In particular, training involving the inter-GCL requires the concurrent supervision of both the intra-GCL and the softmax loss to avoid the dispersion of the view-field-based sub-classes. Indeed, the inter-GCL only expresses a relative constraint on the intra-class compactness (numerator in Eq. (2)) respect to the inter-class distance (denominator) so it needs to be combined to an absolute distance-based constraint provided by the intra-GCL. The linear combination of the three losses is expressed by the two parameters λ_{intra} and λ_{inter} in Eq. 3 where the first term represents the softmax loss with $\boldsymbol{W}_j$ denoting the j^{th} column of the weights $\boldsymbol{W}$ in the last fully connected layer and $\boldsymbol{b}$ the bias term.

$$L = -\sum_{i=1}^{m} log\frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + \boldsymbol{b}_{y_i}}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j^T \boldsymbol{x}_i + \boldsymbol{b}_j}} \quad + \lambda_{intra} L_{intra} \quad + \lambda_{inter} L_{inter} \tag{3}$$

The Stochastic Gradient Descent (SGD) is used for back-propagating the error of the derivatives of the loss with regards to the data $\boldsymbol{x}_i^{g_i}$, for $i = 1, ..., m$ and also respect to the centres of all the sub-classes of all the identities (Appendix A). The centres too, as the rest of the CNN parameters are updated in mini-batches according to the equation $c_j^{t+1} = c_j^t - \alpha \frac{\partial L_{(*)}}{\partial c_j^t}, \ \forall j$, where t is the iteration step, α is a scalar with values in $[0, 1]$ controlling the learning rate of the centres and $L_{(*)}$ denotes either the L_{intra} or L_{inter}, depending on which loss function we are focusing on.

4 Experiment

Database. We perform our experiments against two of the largest datasets for the person re-id task, **CUHK03** ([Wang, 2014]) and **Market-1501** ([Zheng et al., 2015]), since they allow to learn a CNN on a significant number of different views of the same identity. The "labelled" subset of CUHK03 is made up of 1360 different identities, each with up to 10 images, the first 5 seen under one camera and the remaining 5 under a different field of view. We reproduces the setting used in [Wang, 2014]: each of the 20 testsets counts 100 images and

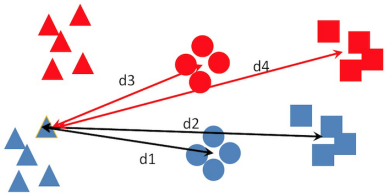

Figure 2: Conceptual scheme of how the inter-GCL function operates. Colour represents identity membership, shape indicates camera view membership. The arrows pointed blue triangle is the image currently contributing to the loss. The function may be re-written like this: $L_{inter}(\boldsymbol{x}_i) = \frac{d_1+d_2}{d_1+d_2+d_3+d_4}$. Best viewed in colour.

the validation set includes 100 identities. Market-1501 has got a larger depth than CUHK03. It exposes each identity up to 6 views and for each view several tens of instances of the same person are present. It consists of a train set of 751 identities shot in 12936 images and a testing set of 750 identities, corresponding to 13115 images. 2798 "distractors" (heavily misaligned detections) are also added to the test set.

Evaluation Metric and Protocols. It is worth noting that in order to measure the performance against the two datasets, the Cumulative Matching Curve (**CMC**) is employed for both datasets. In the case of Market-1501, which provides many ground truths for each query, the mean-Average Precision (**mAP**) is also computed to take into account both precision and recall. For CUHK03 we have adopted the evaluation protocol in [Wang, 2014]. Following this, the first 5 images of each identity (shot by camera 1, 3, or 5) represent view A, the remaining 5 (shot by camera 2, 4 or 6) represent view B. The probe set consists of the images seen in view A. The gallery-set of each probe is made up of 100 randomly chosen images seen in view B, one per each of the 100 identities in the testset. The gallery images selection is repeated 100 times, the CMC is calculated each time and, finally, the mean CMC curve is reported. The evaluation protocol we have implemented in Python for Market-1501 is compliant with the one used in [Zheng et al., 2015] according to which each of the 3368 queries is to be tested against its own gallery-set. The gallery set is formed of all the testing images except the ones having a file-name starting with '-1' (ID identifier) and the ones belonging to the probe's "junk set" comprising all the test images having the probe's same identity and field of view.

Implementation Details. One Caffe layer has been implemented separately for each loss and concatenated to the softmax loss layer. The SGD proceeds on the base of mini-batches and, since the two losses are added separately as two different layers to the Caffe training prototxt file (defining the structure of the CNN at trainaing time), two independent systems of centres are generated for the two layers which progressively converge. We experimented the two training objectives to train ResNet50 ([Zheng et al., 2017]), a residual learning-based state-of-the-art CNN ([He et al., 2016]) formed by 53 convolutional layers, feeding it with RGB images resized to 224x224 pixels. It is first trained for the identity classification task and, at testing stage, the deep features are extracted from the fully connected layer *pool5* for ResNet50, with dimension 2048. The training was extended up to 15000 iterations in all our simulations in the configuration softmax + intra-GCL + inter-GCL. The better identity classification performance are reached, the better re-id accuracy is achieved (Table 2). We ran our deep learning experiments on a single machine equipped with one NVIDIA GeForce GTX Titan X GPU and an Intel Core i7-5960X CPU @ 3.00GHz, 64.0 GB RAM. The training takes 4 hours for 15000 iterations.

Experimented Results. We report in Table 2 (with related graphs in Fig. 3) the results achieved on CUHK03 and Market-1501 with ResNet50 supervised by the softmax loss in linear combination with the intra-GCL and inter-GCL against the results achieved by the usual training relying only on the softmax loss. The study has been carried out by parametrizing the performance with regards to the two scaling factors λ_{intra} of the intra-GCL and λ_{inter} of the inter-GCL. By varying $(\lambda_{intra}, \lambda_{inter})$ in the range $[10^{-5}, 10^{-2}]$, it comes out that the point of maximum for the rank 1 re-id accuracy is $(\lambda_{intra}, \lambda_{inter}) = (5 * 10^{-4}, 10^{-4})$. Table 1 reports the performance by

	CUHK03	Market-1501	
L	rank1	rank1	mAP
0	51.60	73.02	47.62
0.00001	63.66	76.22	53.39
0.00005	62.35	76.48	53.95
0.0001	60.85	76.51	53.43
0.0005	61.57	75.89	53.26
0.001	59.27	75.83	53.46
0.005	57.30	75.27	52.92
0.01	51.25	74.61	50.86

Table 1: Performance (%) under the combined losses supervision for $\lambda_{intra} = 0.0005$ and λ_{inter} changing.

	CUHK03		Market-1501		
	rank1	id acc	rank1	mAP	id acc
Bow+Kissme [Zheng et al., 2015]	-	-	44.42	20.76	-
Null Space [Zhang et al., 2016]	54.7	-	55.43	29.87	-
LSTM Siamese [Varior et al., 2016b]	57.3	-	61.6	35.3	-
Gated Siamese [Varior et al., 2016a]	61.8	-	65.88	39.55	-
Baseline(R,pool5) [Zheng et al., 2017]	51.60	94.23	73.02	47.62	91.19
ours	**63.66**	96.79	**76.51**	**53.43**	93.59

Table 2: Softmax vs combined losses supervision for ResNet50. Results (%) at the point of maximum in the $(\lambda_{intra}, \lambda_{inter})$ plane. The accuracy (*id acc*) of the identity classification task is measured at iteration #15000.

varying λ_{inter} when λ_{intra} is fixed at 0.0005. The same data are plotted in Fig. 3. For Market-1501 a rank 1 accuracy of **76.51**% is achieved, improving the baseline result (72.41%) in [Zheng et al., 2017] of 5.66% as shown in Fig.3. The correspondent mAP value is **53.43**%, that improves the $Baseline(R, pool5)$ result (46.79%) of 14.19%. For CUHK03 the improvement is 23.37%. Furthermore, Table 2 shows that our method achieves better performance than many state-of-the-art approaches like [Varior et al., 2016b] or [Varior et al., 2016a].

5 Conclusion

In this paper we have proposed two new loss functions for training a state-of-the-art CNN, re-formulating the centre loss for the person re-identification task, in order to get more discriminative features that could mitigate the effects of camera viewpoint changes on pedestrians. The experiments presented showed that the supervision of the two losses, combined with the softmax loss, helps significantly the performance in the disjoint multi-camera scenario, beating several state-of-the-art approaches on CUHK03 and Market-1501.

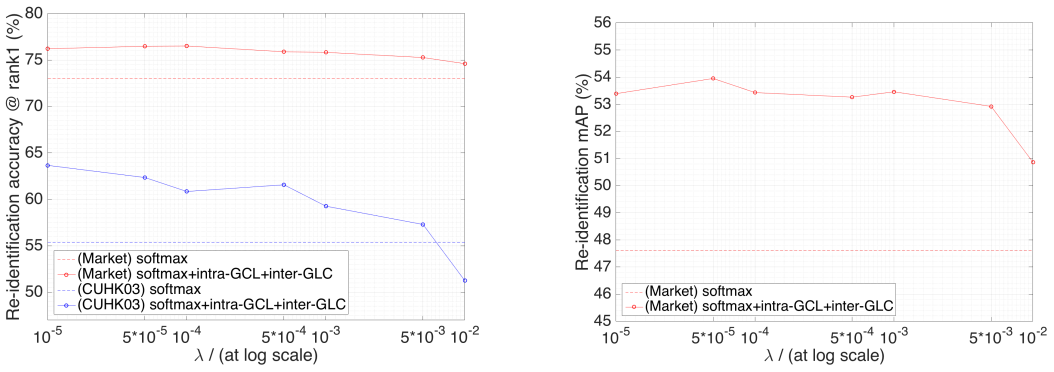

Figure 3: Performance improvement that our method achieves on ResNet50. Best viewed in colour.

A APPENDIX

$$\frac{\partial L_{inter}}{\partial \boldsymbol{x}_i^{g_i}} = 2\frac{\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}(\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_j^{g}) * \sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g}\|_2^2 - \sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_j^{g}\|_2^2 * \sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}(\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g})}{(\sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g}\|_2^2)^2}\delta(y_i=j) \tag{4}$$

$$\frac{\partial L_{inter}}{\partial \boldsymbol{c}_q^{k}} = \begin{cases} 2\sum\limits_{i=1}^{m}\frac{(-\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_j^{g}\|_2^2)*(\boldsymbol{c}_q^{k}-\boldsymbol{x}_i^{g_i})}{(\sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g}\|_2^2)^2}\delta(y_i=j), & \text{for } q\neq j, k\neq g_i \\ 2\sum\limits_{i=1}^{m}\frac{(\sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g}\|_2^2-\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_j^{g}\|_2^2)*(\boldsymbol{c}_q^{k}-\boldsymbol{x}_i^{g_i})}{(\sum\limits_{t=1}^{n}\sum\limits_{\substack{g=1\\g\neq g_i}}^{s}\|\boldsymbol{x}_i^{g_i}-\boldsymbol{c}_t^{g}\|_2^2)^2}\delta(y_i=j), & \text{for } q=j, k\neq g_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

References

[Ahmed et al., 2015] Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916.

[Bak et al., 2015] Bak, S., Martins, F., and Bremond, F. (2015). Person re-identification by pose priors. In *SPIE/IS&T Electronic Imaging*, pages 93990H–93990H. International Society for Optics and Photonics.

[Chen et al., 2012] Chen, D., Cao, X., Wang, L., Wen, F., and Sun, J. (2012). Bayesian face revisited: A joint formulation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7574 LNCS(PART 3):566–579.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[Hoffer and Ailon, 2015] Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

[Kulis, 2013] Kulis, B. (2013). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364.

[Pishchulin et al., 2016] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., and Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937.

[Sun et al., 2014] Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898.

[Varior et al., 2016a] Varior, R. R., Haloi, M., and Wang, G. (2016a). Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer.

[Varior et al., 2016b] Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016b). A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer.

[Wang, 2014] Wang, X. (2014). DeepReID : Deep Filter Pairing Neural Network for Person Re-Identification. *Cvpr*, pages 1–8.

[Wen et al., 2016] Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer.

[Wu et al., 2016] Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., and Zheng, W.-S. (2016). An enhanced deep feature representation for person re-identification. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE.

[Xiao et al., 2016a] Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016a). Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258.

[Xiao et al., 2016b] Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2016b). End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*.

[Yi et al., 2014a] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014a). Constrained Deep Metric Learning for Person Re-identification. *2014 22nd International Conference on Pattern Recognition*, (1):34–39.

[Yi et al., 2014b] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014b). Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE.

[Zhang et al., 2016] Zhang, L., Xiang, T., and Gong, S. (2016). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248.

[Zhao et al., 2013] Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

[Zhao et al., 2014] Zhao, R., Ouyang, W., and Wang, X. (2014). Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151.

[Zheng et al., 2017] Zheng, L., Huang, Y., Lu, H., and Yang, Y. (2017). Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*.

[Zheng et al., 2015] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124.