Supplementary Material for MMD-ReID: A Simple but Effective Solution for Visible-Thermal Person ReID

Chaitra Jambigi*
chaitraj@iisc.ac.in
Ruchit Rawal*
ruchitrawal@iisc.ac.in
Anirban Chakraborty
anirban@iisc.ac.in

Department of Computational and Data Sciences,
Indian Institute of Science
Bangalore, India

1 Architecture and Loss details

GEM pooling layer: To get fine grained features, our two stream network is terminated by a Generalised Mean Pooling layer [7, 10], which is defined as:

$$f = [f_1, f_2, ...f_K]^T, f_k = \frac{1}{|X_k|}(\sum_{x \in X_k} x^{p_k})^{1/p_k} \tag{1}$$

where f_k is a feature map, K is the number of feature maps input to GeM pooling, X_k is the set of pixels in a HxW shaped feature activation map say. The output of GeM layer is a 1-D vector with each component representing one feature map.

HC-Tri loss: Triplet loss [2] is a widely used metric learning loss in Person ReID. Each mini-batch sample is considered as an anchor, and the hardest positive and hardest negative sample is selected for this anchor. To effectively fetch positives in the mini-batch, the mini-batch is formed by randomly sampling P identiies and randomly sampling K images of each identity, resulting in a mini-batch with PK images. This loss compares each sample (anchor) to all other samples which is a strict constraint, perhaps too strict to constrain the pairwise distance if there exist some outliers (bad examples), which would form the adverse triplet to destroy other pairwise distances [3]. Therefore, [3] considers adopting the center of each person as the identity agent. In this manner, we can relax the strict constraint by replacing the comparison of the anchor to all the other samples by the anchor centre to all the other centres.

$$L_{hc\_tri}(C) = \sum_{i=1}^{P} \left[\rho + \left\| c_v^i - c_t^i \right\|_2 - \min_{n \in \{v,t\}, j \neq i} \left\| c_v^i - c_n^j \right\|_2 \right]_+ + \sum_{i=1}^{P} \left[\rho + \left\| c_t^i - c_v^i \right\|_2 - \min_{n \in \{v,t\}, j \neq i} \left\| c_t^i - c_n^j \right\|_2 \right]_+$$

 * denotes equal contribution.

where, $c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i, \quad c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i$

$\{c_v^i | i = 1,2,...P\}$ are the visible centres and $\{c_t^i | i = 1,2,...P\}$ are the thermal centres. L_{hc_tri} concentrates on only one cross-modality positive pair and the mined hardest negative pair in both the intra and inter-modality.

2 Implementation details

We adopt ResNet50 [1] as the backbone network. The stride of the last convolution layer is changed from 2 to 1 to get fine-grained features [8]. Input images are resized to 288x144 shape and padded with 10, followed by Data augmentation techniques like random cropping of 288x144 shape and Random Horizontal flipping. We also use Random erasing augmentation [11] with probability 0.5 for some experiments, which we discuss in the Results section of the main paper. We use a Stochastic Gradient descent optimizer (SGD) with momentum as 0.9 and 0.0005 weight decay. We set initial lr as 0.01 for ResNet50 parameters and 0.1 for BatchNorm layer and Classifier (FC layer) for both datasets (SYSU-MM01 and RegDB). Warmup learning rate strategy is applied to improve performance as [8]. For sampling, we choose P and K both as 4 for both datasets. Margin ρ for Margin MMD-ID loss is set as 1.4 for both the datasets and ρ_1 for HC-Tri loss is 0.3. The tradeoff parameters in total loss equation of main paper: $\lambda_1, \lambda_2, \lambda_3$ are set as 1, 0.25,2. We train our model on a single Nvidia GTX 1080Ti gpu card for 60 epochs which takes $\sim$ 6 hours to train for SYSU-MM01 and $\sim$ 1.3 hours for RegDB with all our losses.

3 Ablation Study

3.1 Effect of Random erasing augmentation

Random Erasing (RE) augmentation [11] is a well-known regularisation technique that helps in improving the generalisation ability of the model. We incorporate RE with our total loss formulation to get better performance. To ensure that the gain in performance is not because of adding RE, we perform a set of experiments with RE and without RE to see the net effect of adding RE. Table 1 shows the experiments with the corresponding rank-1 and mAP values. It is evident from the last two rows that even without adding RE, our final Margin MMD-ID loss along with Cross entropy and HC-Tri loss (row 7) performs comparably with the state of the art models. Adding RE (row 8) gives the boost hence we use RE in our final model. Also, adding RE with only Cross entropy loss (row 2) or with Cross entropy and HC-Tri loss (row 4) doesn't give much performance boost as RE on itself, cannot reduce the modality gap.

3.2 Dataset Complexity: RegDB

RegDB [6] is collected from two well-aligned cameras (one visible and one thermal), compared to six cameras for SYSU-MM01 (four visible and two thermal in both indoor and outdoor environments). For RegDB evaluation, the dataset is randomly split into two parts, one for training and one for testing. Thus, for each modality (e.g., visible), the samples during training and testing are captured using the same camera. This eliminates significant

Sr. No	Method	r1	mAP
1	C.E.	52.78	50.29
2	C.E. (w R.E.)	55.32	51.24
3	C.E. + HC-Tri	54.75	52.14
4	C.E. + HC-Tri (w R.E.)	60.94	55.39
5	C.E. + HC-Tri + MMD-ID	62.15	57.58
6	C.E. + HC-Tri + MMD-ID (w R.E.)	64.4	59.8
7	C.E. + HC-Tri + Margin MMD-ID	63.11	58.48
8	C.E. + HC-Tri + Margin MMD-ID (w R.E.)	**66.75**	**62.25**

Table 1: Effect of Random Erasing (R.E.) augmentation on different components in MMD-ReID. Results provided for All-Search mode in SYSU-MM01 dataset

intra-modality variations (such as viewpoint and pose changes), usually caused when images are captured using multiple cameras. Moreover, SYSU-MM01 (38,271) has more than four times the number of samples present in the RegDB dataset (8,240), further increasing the complexity of matching identities across modalities. The aforementioned reasons indicate that RegDB is a much simpler dataset to operate on with less vulnerability to overfitting due to train-test sampling similarities. Thus, applying MMD-ID on RegDB doesn't correspond to feature-degradation or overfitting and provides relatively decent performance compared to evaluation on SYSU-MM01 (Table-3 in the main paper, row-2;3). We also empirically verify this insight by generating the t-SNE plots for MMD-ID on the RegDB dataset. We observe that both train and test features demonstrate high inter-class separation and intra-class compactness (Fig.1). Fig.1 reveals that the features for each identity are easily separable and consequently have little chance of overfitting. Lastly, recent state-of-the-art works [3] have also observed a similar high performance on the RegDB dataset (compared to SYSU-MM01).

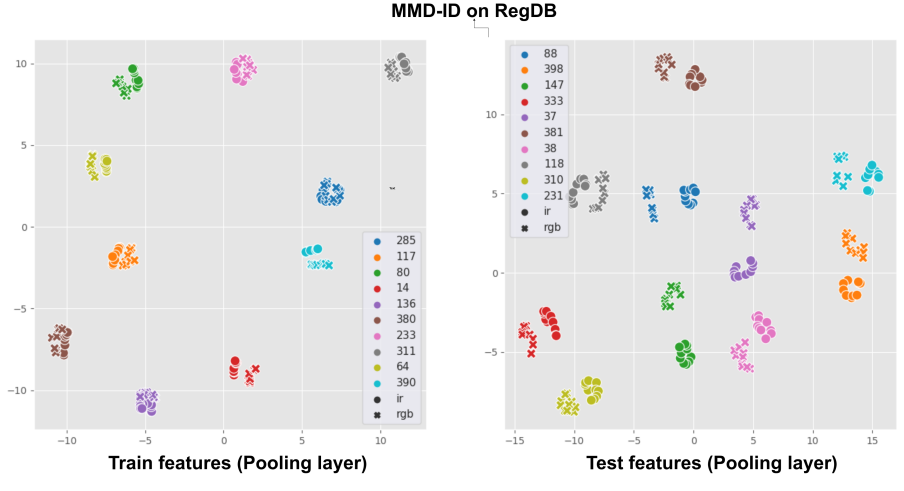

Figure 1: t-SNE visualisation on RegDB which shows the features are easily separable and less prone to overfitting.

3.3 Qualitative visualisation using T-SNE

Figure 2 shows the qualitative visualisation of the features after the BatchNorm layer, using T-SNE plots [9]. The left side plot is for the features (belonging to test-data) extracted by a model trained with only Cross entropy (CE) + HC-Tri loss, and the Right side plot is for the model trained with our MMD ReID framework. It is clear from the Left side plot that the visible and thermal features for a particular identity form separate clusters and are well separated, which is undesirable. The visible and thermal clusters ideally should be compact and as close as possible to avoid misclassifications. The right side plot has successfully achieved these properties by bringing the same identity visible and thermal features closer in feature space. Thus, the visual analysis also supports our MMD-ReID framework.

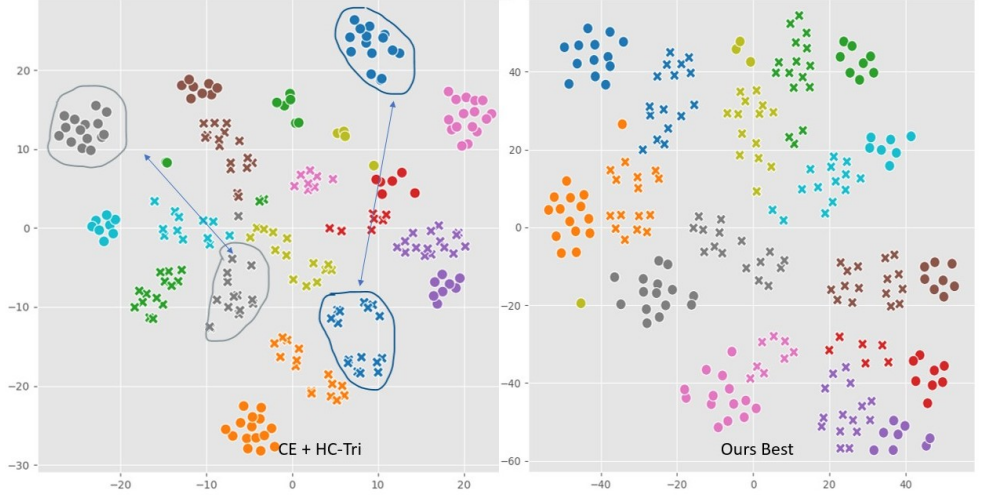

Figure 2: T-SNE visualisation on ten randomly sampled test identities (SYSU-MM01) for CE+HC-Tri loss trained model (baseline) Vs Our Best (MMD-ReID) model. Different color denotes different identities. Cross and circle marker denotes thermal and visible features respectively.

3.4 Implementation details for Margin MMD-ID with existing baselines:

To verify generalisation capability of MMD-ReID, we take three popular and open-sourced baselines and add MMD-ID and Margin MMD-ID losses on them. The details about the baselines and hyperparameters used are described below. Note that the table for accuracies with MMD on different baselines is presented in main paper, Table 4.

AGW (Average Generalized mean pooling with Weighted triplet loss): Ye et al. in their work [10] introduced a new powerful baseline for Person Re-ID. AGW proposed three major modifications on top of the best practices discussed in [5]: Non-local attention blocks, Generalized-mean (GeM) pooling layer, and Weighted regularized triplet loss. In line with the standard setup, the MMD-ID and Margin MMD-ID losses are computed on features ex-

tracted from the GeM layer while features extracted from the BatchNorm layer are used during inference time. The margin (ρ) in Margin MMD-ID is set as 0.4 whereas all other hyperparameters are kept the same as reported by [10].

DGTL (Dual-Granularity Triplet Loss): DGTL [2] utilizes sample-based and center-based triplet loss in a hierarchical manner to encourage intra-class compactness and inter-class discrimination at fine and coarse granularity levels simultaneously. This setup allows achieving competitive performance without the need for aggregating local-level features via architectural improvements. In accordance with previous experiments, we employ the MMD-ID and Margin MMD-ID loss on the features extracted from the pooling layer (in the fine granularity level branch). The margin (ρ) in Margin MMD-ID is set as 1.00 while all other hyperparameters remain unchanged.

HcTri (Hetero-center Triplet Loss): Since traditional triplet-loss is prone to outliers and often fails to converge, Liu et al. [3] in their work proposed a novel hetero-center triplet loss that operates on a coarse granularity level. The Hc-Tri loss in a part-based person feature learning framework leads to superior performance than the standard triplet loss. The Hc-Tri loss is computed for each part-level feature strip as well as the final concatenated global features. For a fair comparison with other baselines, we employ the MMD-ID and Margin MMD-ID loss only on the concatenated global feature vector. The margin (ρ) in Margin MMD-ID is set as 1.00 while all other hyperparameters are kept the same.

3.5 Computational cost analysis:

We create our batch with $2 \times PK$ images, where P is number of identities and K is number of visible and thermal images. $L_{Margin-MMD-ID}$ requires computing $PK(K-1)$ pairwise distances for same distribution term and $P \times K \times K$ distances for cross distribution term, which after summing makes a total of $PK[2K-1]$ computations. This is comparable with the computations needed for a batch with $2 \times PK$ images for a standard triplet loss [3] which is $2PK(2K-1)$ for hardest positive sample mining and $2PK \times 2(P-1)K$ for hardest negative sample mining. Also, Hc-Tri loss [3], for a batch requires P computations for positive and $2P \times 2(P-1)$ for negative term. Thus, computation wise, our loss is comparable to standard Triplet loss. We also do a training time analysis and observe the hours needed to train a model for 60 epochs for different setups. Table 2 shows that using MMD-ReID negligibly increases the training time over C.E. and C.E. + HC-Tri.

Setup	Training time
C.E.	5.45 hrs
C.E. + HC-Tri	5.81 hrs
MMD-ReID (All loss)	6 hrs

Table 2: Training time analysis on Nvidia GTX 1080Ti: SYSU-MM01

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and*

pattern recognition, pages 770–778, 2016.

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

[3] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 2020.

[4] Haijun Liu, Yanxia Chai, Xiaoheng Tan, Dong Li, and Xichuan Zhou. Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification. *IEEE Signal Processing Letters*, 28:653–657, 2021.

[5] Hao Luo, Youzhi Gu, Xingyu Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019.

[6] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[7] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[8] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[10] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.