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

Abstract

Learning modality invariant features is central to the problem of Visible-Thermal cross-modal Person Reidentification (VT-ReID), where query and gallery images come from different modalities. Existing works implicitly align the modalities in pixel and feature spaces by either using adversarial learning or carefully designing feature extraction modules that heavily rely on domain knowledge. We propose a simple but effective framework, MMD-ReID, that reduces the modality gap by an explicit discrepancy reduction constraint. MMD-ReID takes inspiration from Maximum Mean Discrepancy (MMD), a widely used statistical tool for hypothesis testing that determines the distance between two distributions. MMD-ReID uses a novel margin-based formulation to match class-conditional feature distributions of visible and thermal samples to minimize intra-class distances while maintaining feature discriminability. MMD-ReID is a simple framework in terms of architecture and loss formulation. We conduct extensive experiments to demonstrate both qualitatively and quantitatively the effectiveness of MMD-ReID in aligning the marginal and class conditional distributions, thus learning both modality-independent and identity-consistent features. The proposed framework significantly outperforms the state-of-the-art methods on SYSU-MM01 and RegDB datasets. Code will be released at https://github.com/vcl-iisc/MMD-ReID .

1 Introduction

Person re-identification (ReID) is widely studied in computer vision as a pedestrian matching problem between query and gallery images from different cameras [45, 46, 53]. Traditional methods focus on scenarios where single-modality cameras capture images: Visible-Visible ReID (VV-ReID), where the focus is on matching visible images. However, in 24-hour intelligent surveillance systems, we need to process data from infrared cameras at nighttime. Thus, there has been a significant interest in Visible-Thermal ReID (VT-ReID) which, given a visible image, aims to match it to the thermal image of the same person [8, 44, 48, 51]. VT-ReID is more challenging than VV-ReID as it suffers from both intra-modality variations (caused by pose, illumination, and viewpoint changes) as well as inter-modality variations (caused by a huge modality gap between visible and thermal images [13, 31, 44, 47]).

 * denotes equal contribution.

The quest to bridge the cross-modality discrepancy has pushed advancements in two significant directions: First, adversarial-learning based approaches have paved the way for joint pixel and feature space alignment [8, 22, 44, 47]. This is typically achieved by leveraging generative adversarial networks to translate an image from a heterogenous modality to the desired modality and using a mini-max setup to learn modality invariant feature representations. However, generative methods do not guarantee identity preservation across modality translation and often require excessive training tricks and additional computation. Second, shared feature learning techniques currently achieve state-of-the-art results for VT-ReID by projecting features from heterogeneous modalities into a common feature space [7, 26, 31, 52]. However, they heavily rely on carefully designed feature selection modules such as partition strips [26, 39, 54], semantic alignment [19], human landmarks [42] . Recent studies [5, 27, 32] have criticized the current state-of-the-art methods' overly complex and rigid nature, citing the need for new algorithmic ideas that are both simple and effective.

We approach the problem of learning modality invariant representations in the VT-ReID task from an explicit distribution discrepancy perspective. The centerpiece of such a formulation is the use of a statistical hypothesis testing framework called maximum mean discrepancy (MMD) [11] that measures the proximity between two distributions. MMD has been widely studied in unsupervised domain adaptation (UDA) literature to minimize marginal [29, 30] as well as (more recently the) class-conditional distribution discrepancy [49]. Inspired by this, we adopt MMD in the supervised VT-ReID task to align visible and infrared distributions for a particular identity. However, we (empirically) observed that this formulation is vulnerable to overfitting and feature degradation, leading to suboptimal results. To alleviate this problem, we introduce a novel margin-based MMD loss: Margin MMD-ID.

With the goal of providing a simple yet strong framework to achieve competitive performances, we propose MMD-ReID that utilizes Margin MMD-ID as its core training objective. MMD-ReID is simple, primarily as (1) it only uses the global features and does not rely on part-level features. (2) It is easily extendable since it's built on the traditional two-stream network that has enjoyed promising results in VT-ReID. Furthermore, (3) Margin MMD-ID loss is intuitive and easy to train in a deep learning setup. We demonstrate the effectiveness of MMD-ReID through extensive experimentations on two popular benchmark datasets: SYSU-MM01 and RegDB, outperforming the current state-of-the-art by 5.07% and 4% Rank1 accuracy, respectively. Moreover, we empirically observe that our modified loss: Margin MMD-ID is not only complementary to the current best practices in the ReID community but can also be easily adopted in existing baselines to further boost the performance.

In summary, the main contributions of our work are:

- We propose a simple but effective framework: MMD-ReID, which to the best of our knowledge, is the first work to explore the VT-ReID task from the perspective of explicit distribution discrepancy reduction constraint. MMD-ReID employs our novel margin-based modification: Margin MMD-ID loss to alleviate the problem of overfitting and feature degradation that occurs with standard MMD in supervised VT-ReID.

- Extensive experiments demonstrate that MMD-ReID achieves state-of-the-art results on two benchmark datasets: SYSU-MM01 and RegDB. It is worth mentioning that we achieve improvement in performance just by using global features.

- We empirically demonstrate that Margin MMD-ID can be used on top of existing baselines to improve their performance further. We verify our claim by performing experiments on three popular baselines: AGW [55], Hc-Tri [26], DGTL [27].

2 Related works

VV-ReID: Person re-identification problem has been primarily studied in a closed-world setting where images are acquired by single-modality cameras [33, 57, 60, 61]. Accordingly, researchers have focussed a great deal of attention on dealing with challenges of appearance changes pertaining to a single modality, such as variation in viewpoint [1, 21], pose [6, 37, 58], illumination [18], occlusions [17], and background clutter [38]. This is usually achieved by augmenting standard convolutional neural networks with powerful (manually-designed) feature selection modules such as partition-strips [26, 39], pose estimation [59] to handle occlusions and misalignment, etc. Another line of approaches utilizes deep metric learning [2, 9, 26, 41, 56] to design loss functions (such as triplet loss [15], quadruplet loss [4]) that ensure robust and discriminative feature representations.

VT-ReID: Recent VT-ReID methods primarily rely on either adversarial learning-based modality alignment or modality shared feature learning methods to alleviate cross-modality discrepancy. Inspired by the success of generative adversarial networks, adversarial learning-based approaches aim to perform cross-modality alignment in pixel and feature space. Dai *et al*. [8] utilized adversarial training strategies to learn modality invariant feature representations. Kniaz *et al*.[22] proposed a novel GAN framework ThermalGAN to translate a single visible probe image to a thermal probe set and perform conventional ReID in the thermal domain. Wang *et al*. [44] in their work, employed an end-to-end three-player mini-max setup to jointly optimize for pixel and feature space alignment across modalities. On similar lines, Wang *et al*. [47] proposed to decompose modality and appearance discrepancy and reduce them separately using a bi-directional cycleGAN and conventional feature level constraints, respectively.

Learning robust and discriminative shared feature representations is central to the success of VT-ReID systems. Most recent studies approach this through a two-stream network backbone (first proposed by Ye *et al*. [50, 51, 53]) that projects cross-modality embeddings in a common feature space. Ye *et al*. [52] in their work handle the modality discrepancy at both feature and classifier level by proposing an ensemble learning scheme to incorporate the modality shareable classifier and the modality-specific classifiers. Liu *et al*. [25] in pursuit of learning robust and discriminative person features, proposed a mid-level feature incorporation strategy using skip-connections. Shared features disregard modality-specific features reducing the discriminability of feature representation. To alleviate this problem, Lu *et al*. [31] proposed a novel shared-specific feature transform algorithm to utilize both modality-specific and modality-shared information by modeling the affinities between intra-modality and inter-modality samples. Liu *et al*. [26] in their work, proposed the hetero-center based triplet loss to provide a strong baseline for VT-ReID tasks utilizing both global and local feature extraction strategies.

MMD: In the scope of deep learning, MMD was first studied in the unsupervised domain adaptation (UDA) literature to align source and target distributions. Most notably, Long *et al*. [30] first introduced the idea of minimizing multi-kernel MMD between task-specific layers to enhance feature transferability across domains. To optimize conditional-distributions discrepancy, Long *et al*. [28] adopted a pseudo label refinement strategy to generate target domain labels and perform a joint adaptation of both marginal and conditional distributions between domains. Owing to its intuitive and strong foundations, MMD has been adopted by diverse emerging paradigms in deep learning such as generative adversarial networks, variational autoencoders, transfer-learning, noise-insensitive auto-encoders [24, 36].

3 Methodology

The rest of the paper is organised as follows: Section 3.1 briefly introduces MMD, Section 3.2 describes using MMD for VT-ReID task and margin-based modifications. Section 3.3 describes the architecture, batch sampling strategies and overall loss formulation. Section 4 describes in detail the datasets, experiments, results, and ablation studies. Section 5 concludes the work with future directions.

3.1 Maximum Mean Discrepancy (MMD)

The two sample test is one of the fundamental tests in statistics that tries to determine whether the given two datasets, $\{X_n\} \sim P$ and $\{Y_m\} \sim Q$ are generated from the same underlying distribution or not. This task is difficult since the distribution information is generally unknown apriori [2, 3, 10, 12]. MMD is a test statistic that measures the discrepancy of two distributions by embedding them in a Reproducing Kernel Hilbert space (RKHS) [11]. To simplify, MMD performs the two sample test by finding the difference between the mean function values of the two samples evaluated on a smooth function, where the function class for MMD is a unit ball in an RKHS. If the difference in mean values is large, then the samples are likely to be drawn from different distributions. The formulation of MMD is,

$$MMD^2(X,Y) = \left\| \frac{1}{N}\sum_{n=1}^{N} \phi(X_n) - \frac{1}{M}\sum_{m=1}^{M} \phi(Y_m) \right\|^2 \tag{1}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N} \phi(X_n)^\top \phi(X_{n'}) + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} \phi(Y_m)^\top \phi(Y_{m'}) - \frac{2}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} \phi(X_n)^\top \phi(Y_m) \tag{2}$$

$$= \mathbf{E}_{x,x'\sim P}\left[k(x,x')\right] + \mathbf{E}_{y,y'\sim Q}\left[k(y,y')\right] - 2\mathbf{E}_{x\sim P, y\sim Q}[k(x,y)] \tag{3}$$

where $\phi(.)$ is the feature mapping function.
Kernel trick can then be applied on the inner product in Eq.(2) to get Eq.(3)

3.2 MMD in VT-ReID

Let $\mathcal{P} = \left\{x_v^i ... x_v^{N_V}\right\}$ and $\mathcal{Q} = \left\{x_t^i ... x_t^{N_T}\right\}$ denote the visible and thermal images, respectively. N_V and N_T denote the total number of visible and thermal images in the dataset, respectively. To reduce the distribution discrepancy in the shared space, we use MMD distance as the criterion to explicitly learn representations such that the MMD loss between visible and thermal features is minimized.

$$L_{MMD}(P,Q) = \underbrace{\mathbf{E}_P\left[k(x_v,x_v')\right] + \mathbf{E}_Q\left[k(x_t,x_t')\right]}_{\textit{same modality distribution}} - \underbrace{2\mathbf{E}_{P,Q}[k(x_v,x_t)]}_{\textit{cross modality distribution}} \tag{4}$$

The first two terms are the kernel similarity between the same modality samples, which has a high value at the start of training. The last term is the similarity between cross-modality samples, which is low initially. When MMD loss is minimized, it eventually tries to bring the cross-modality similarity as close as possible to the same modality similarity, thereby aligning both the distributions. MMD aims to match infinite order moments with a Gaussian kernel [11]. Thus, reducing the MMD distance aligns the two distributions in a superior way compared to other implicit aligning methods discussed in the introduction section (Section-1).

MMD-ID: The above MMD loss formulation in Eq.(4) aligns the two modalities marginally without considering the class conditional distribution relationship between the two modalities. Thus, when modalities get aligned, the learned features may not preserve the class discriminative property. In order to align the modalities, respecting the class-wise distribution, we use a modified version of MMD, in which we precisely align the distributions on a per-identity basis and averaging over all possible identities. The modified loss is of the form,

$$MMD^2(P^c,Q^c) = \mathbf{E}_P\left[k\left(x_v^c, x_v^{c'}\right)\right] + \mathbf{E}_Q\left[k\left(x_t^c, x_t^{c'}\right)\right] - 2\mathbf{E}_{P,Q}[k(x_v^c, x_t^c)] \tag{5}$$

$$L_{MMD-ID}(P,Q) = \frac{1}{C}\sum_{c=1}^{C} MMD^2(P^c,Q^c) \tag{6}$$

P^c and Q^c denote visible and thermal sample distribution of a particular c^{th} identity.

Margin MMD-ID: Although MMD-ID is intuitive, it can suffer from the problem of overfitting, thus collapsing all the features of the same identity to a small region in feature space, as shown in Figure 1. To mitigate this effect and optimally use the strengths of MMD-ID, we propose a new margin-based loss as,

$$MMD'^2(P^c,Q^c) = \begin{cases} MMD^2(P^c,Q^c), & \text{if } MMD^2(P^c,Q^c) - \rho > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$L_{Margin-MMD-ID} = \frac{1}{C}\sum_{c=1}^{C} MMD'^2(P^c,Q^c) \tag{8}$$

We add a margin term ρ, which can control the amount of distribution alignment, thus keeping a balance between aligned and generalised model. Intuitively, we measure the averaged MMD-ID distance over the training and restrict the reduction to a certain value, i.e. ρ.

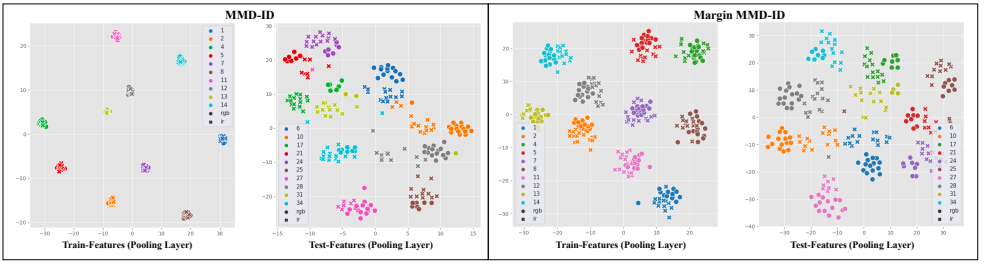

Figure 1: t-SNE plot for the last epoch of the model trained with MMD-ID and Margin MMD-ID. The intra-class compactness in train data doesn't translate to testing data indicating feature degradation and high overfitting.

3.3 MMD-ReID Framework

We introduce our proposed framework MMD-ReID as depicted in Fig 2. Our model mainly consists of two components: 1. Two stream backbone network to explore the shared and specific features 2. Our proposed Margin MMD-ID loss along with Identity softmax loss and triplet loss to get identity separable as well as discriminative features.

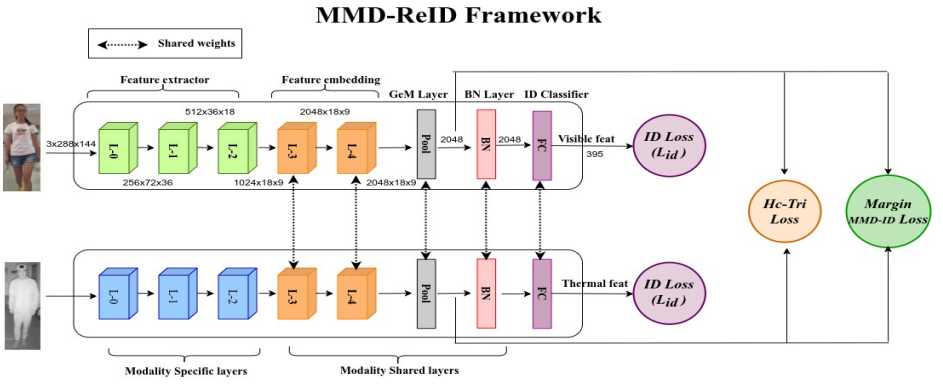

Figure 2: MMD-ReID: Structure of our two stream architecture for VT-ReID. Modality specific layers (L-0, L-1, L-2) have independent weights for each modality. Modality shared layers (L-3, L-4, Pool, BN, FC) have shared weights for both modalities, denoted by dotted by bi-directional arrows. Visible and Thermal features are extracted independently and ID loss is applied. Margin MMD-ID and Hc-Tri are applied on pooled features.

Two stream network: We adopt the conventional two-stream architecture as [26] which consists of feature extractor and feature embedding to extract modality-specific features and shared features, respectively. We use ResNet50 [14] as the backbone with initial shallow layers and first two res-convolution blocks as feature extractor (L-0,L-1,L-2 in Fig. 2) which have separate weights for each modality and last two res-convolution blocks (L-3,L-4) as feature embedding, followed by pooling and BN layers, which have shared weights for both modalities. To get fine-grained features, we use Generalized-mean (GeM) Pooling instead of average or max pooling [35, 55]. For details on GeM layer, refer supplementary material.

Batch sampling: We create our mini-batch by randomly sampling $P \times K$ images, where P is the number of identity in the batch and K is the number of images per identity. We randomly choose K visible and K thermal images, per identity to mitigate class imbalance issues, effectively making a batch size of $2 \times P \times K$.

Overall loss: We use our proposed Margin MMD-ID loss Eq.(8) along with the standard identity softmax loss to learn discriminative features. Our loss explicitly aligns the two modalities based on class conditional distributions, thereby reducing the intra-class discrepancy. However, inter-class separation is not guaranteed, which is needed for good representation learning in open-set problems. To tackle this, we use a variant of Triplet loss, called Hetero-center triplet loss (Hc-Tri) [26] to maximize inter-class distances. Hc-Tri is formulated in the same way as standard Triplet loss [16], but it takes centers of different modalities as input rather than individual samples. More details on Hc-Tri are provided in the supplementary material. Although Hc-Tri also reduces the intra-class distances, it is worth mentioning that the space in which MMD and Triplet losses work is different. Triplet loss formulation brings anchor and positive closer in euclidean space, whereas the MMD loss statistically matches all the higher-order moments. Thus, MMD is a stronger loss in terms of distribution alignment as compared to Hc-Tri loss. The total loss is of the form,

$$L = \lambda_1 L_{id} + \lambda_2 L_{Margin-MMD-ID} + \lambda_3 L_{Hc-Tri} \tag{9}$$

4 Experiments and Results

4.1 Datasets and settings

SYSU-MM01: SYSU-MM01 [48] is a large-scale dataset containing images captured by two thermal and four visible cameras. It contains 491 identities and we use 395/96 identities for training/testing, making 22,258 visible and 11,909 thermal images for training. The test set contains 3803 thermal images for Query and 301 randomly selected visible images as Gallery. We adopt the most challenging and commonly used evaluation mode: All search/Indoor search in Singleshot setting, where only one gallery image per identity is available. We follow the evaluation protocol as [26, 51, 55] to perform ten trials of gallery set selection and then report the average performance.

RegDB: The dataset [34] is collected by dual-camera systems (visible and thermal) and includes 412 identities. For each identity, ten visible and ten thermal images are captured. We follow the evaluation protocol as [7, 51] where the dataset is randomly split into two parts, one for training and one for testing. For testing, images from one modality are selected as gallery and images from other modality as probe set. The process is repeated for ten trials and averaged results are reported.

Evaluation metrics: Following standard protocol [48], Cumulative matching characteristics (CMC) and mean average precision (mAP) are adopted as evaluation metrics. Query and gallery are from different modalities. CMC (rank-k) measures whether correct identity from cross modality is retrieved in top-k results and mAP measures retrieval performance when the gallery set contains multiple matching images.

Implementation details: For implementation details please refer to the supplementary.

4.2 Results and Analysis

Comparison with state-of-the-art: The results on SYSU-MM01 and RegDB datasets is shown in Table 1, 2 respectively. All metrics for other methods are taken from their paper. In the All-search mode, our method surpasses the current state-of-the-art method: cm-SSFT by 5.15 %, 4.96 %, and 3.48 % in rank-1, rank-10, and rank-20 metrics respectively while achieving comparable mAP. A similar trend is observed in the Indoor search where we significantly outperform the state-of-the-art on all metrics. We observe that we marginally lag behind 'Farewell to Mutual Info.' on rank-10 and rank-20 in All-search mode, however considerably surpass them in rank-1 and mAP as well as on all metrics in Indoor-search. Our results on RegDB are better than the state-of-the-art: Hc-Tri by 4% on Rank-1 and by 5.67% on mAP for Visible to Thermal task and the gain for Thermal to Visible task is of 4.35% in Rank-1 and 5.84% in mAP.

Ablation study of different loss components: Table 3 shows the importance of each loss component in training. It is evident that using only cross-entropy loss (CE), or CE with Hc-Tri (row:1,6) loss gives sub-optimal results, and thus there is a scope for explicit modality alignment. We observe a boost in both rank-1 and mAP after adding MMD loss with CE (row 2) in both the datasets, which supports our claim that explicit discrepancy reduction helps in VT-ReID. We further see that replacing MMD with MMD-ID (row 3) rather drops the mAP and rank-1 by $\sim$ 2% for the SYSU-MM01 dataset, and the reason for this is the overfitting of the model leading to feature degradation as shown in Fig.1. To regularise this, we add a margin term in MMD-ID as per Eq.(7) and we see an increase in rank-1 and mAP indicating a reduction in misclassifications (row 4) which is in agreement with Fig.1.

Method	All Search				Indoor Search			
	r1	r10	r20	mAP	r1	r10	r20	mAP
BDTR [31]	17.01	55.43	71.96	19.66	-	-	-	-
SDL [20]	28.12	70.23	83.67	29.01	32.56	80.45	90.67	39.56
cmPIG [15]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
Hi-CMD [1]	34.94	77.58	-	35.94	-	-	-	-
CASE-Net [14]	42.9	85.7	94.0	41.5	44.1	87.3	93.7	53.2
AlignGAN [19]	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
Neural Feature Search [3]	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
Farewell to Mutual Info [18]	60.02	**94.18**	**98.14**	58.80	66.05	96.59	99.38	72.98
Hc-Tri [26]	61.68	93.10	97.17	57.51	63.41	91.69	95.28	68.17
cm-SSFT [21]	61.6	89.2	93.9	**63.2**	70.5	94.9	97.7	72.6
MACE [25]	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
MMD-ReID (Ours)	**66.75**	94.16	97.38	62.25	**71.64**	**97.75**	**99.52**	**75.95**

Table 1: Results on SYSU-MM01 dataset

Method	Visible to Thermal				Thermal to Visible			
	r1	r10	r20	mAP	r1	r10	r20	mAP
BDTR [31]	33.47	58.42	67.52	31.83	32.72	57.96	68.86	31.10
SDL [20]	26.47	51.34	61.22	23.58	25.74	50.23	59.66	22.89
cmPIG [15]	48.5	-	-	49.3	48.1	-	-	48.9
Hi-CMD [1]	70.93	86.39	-	66.04	-	-	-	-
AlignGAN [19]	57.9	-	-	53.6	56.3	-	-	53.4
Neural Feature Search [3]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
Farewell to Mutual Info [18]	73.2	-	-	71.6	71.8	-	-	70.1
Hc-Tri [26]	91.05	97.16	98.57	83.28	89.30	96.41	98.16	81.46
cm-SSFT [21]	72.3	-	-	72.9	71.0	-	-	71.7
MACE [25]	72.37	88.40	93.59	69.09	72.12	88.07	93.07	68.57
MMD-ReID (Ours)	**95.06**	**98.67**	**99.31**	**88.95**	**93.65**	**97.55**	**98.38**	**87.30**

Table 2: Results on RegDB dataset

Further adding Random erasing (RE) as augmentation helps in the overall generalization of our model giving the best accuracy in row 5. In a complementary sense, since Margin MMD-ID cannot increase inter-class distances, we adopt Hc-Tri loss for this purpose. As discussed in Section 3.3, although Hc-Tri loss reduces intra-class distances, MMD is a stronger loss in terms of distribution alignment, hence using Margin MMD-ID with Hc-Tri performs better than only Hc-Tri which can be shown from rows 6,9. Row 6-10 is similar to Row 1-5 but with added Hc-Tri loss and we see that we get the best performance (row 10) when we have all the four components of CE, Margin MMD-ID, Hc-Tri, and RE augmentation.

Components							SYSU-MM01		RegDB	
Sr.No	C.E.	HC-Tri	MMD	MMD-ID	Margin MMD-ID	R.E.	r1	mAP	r1	mAP
1	✓	✗	✗	✗	✗	✗	52.78	50.29	69.45 (72.94)	66.31 (69.53)
2	✓	✗	✓	✗	✗	✗	59.09	54.85	82.95 (84.66)	78.63 (80.17)
3	✓	✗	✗	✓	✗	✗	57.07	53.52	90.52 (91.02)	85.59 (86.74)
4	✓	✗	✗	✗	✓	✗	60.13	55.97	90.76 (91.33)	85.31 (85.51)
5	✓	✗	✗	✗	✓	✓	**64.86**	**60.12**	**93.57 (93.95)**	**86.54(88.74)**
6	✓	✓	✗	✗	✗	✗	54.75	52.14	86.18 (88.79)	80.80 (81.81)
7	✓	✓	✓	✗	✗	✗	59.25	55.32	89.94 (91.52)	84.70 (85.92)
8	✓	✓	✗	✓	✗	✗	62.15	57.58	90.85 (92.68)	86.53 (87.68)
9	✓	✓	✗	✗	✓	✗	63.11	58.48	92.44 (93.78)	87.76 (88.82)
10	✓	✓	✗	✗	✓	✓	**66.75**	**62.25**	**93.65 (95.06)**	**87.30 (88.95)**

Table 3: Ablation Study of different Components on SYSU-MM01 on RegDB datasets. For RegDB dataset, metrics reported as : Thermal to Visible (Visible to Thermal)

Using Margin MMD-ID with existing baselines: To further evaluate the generalisability of our Margin MMD-ID, we take three popular and open-sourced baselines: AGW ([55]), DGTL ([27]) and HcTri [26]. The top-row for each baseline in Table-4 corresponds to the metrics reported in their original work on the SYSU-MM01 dataset. We progressively add MMD-ID and Margin MMD-ID to evaluate their effects on the overall performance. Two goals of this experiment are we want the Margin MMD-ID to be easily integrated with ex-

isting baselines without many changes and to get an overall improvement by adding Margin MMD-ID loss in training. It is worth noting that adding Margin MMD-ID loss is not only compatible with the three baselines, but we also get a considerable improvement over baseline (top-row) as well as standard conditional MMD-ID (middle-row). For further details regarding each baseline experiment, please refer to the supplementary material.

Method	SYSU-MM01			
	r1	r10	r20	mAP
AGW	47.50 (54.17)	84.39 (91.14)	92.14 (95.98)	47.65 (62.97)
AGW + MMD-ID	53.10 (58.05)	89.97 (96.03)	95.83 **(99.32)**	51.12 (66.41)
AGW + Margin MMD-ID	**54.35 (59.17)**	**90.87 (96.09)**	**96.09** (99.27)	**51.91 (66.92)**
DGTL	57.34 (63.11)	-	-	55.13 (69.20)
DGTL + MMD-ID	58.77 (62.75)	90.94 (94.96)	96.01 (98.73)	55.59 (68.99)
DGTL + Margin MMD-ID	**59.63 (65.13)**	**92.10 (96.17)**	**96.84 (99.15)**	**56.50 (71.26)**
HcTri	61.68 (63.41)	**93.10** (91.69)	**97.17** (95.28)	57.51 (68.17)
HcTri + MMD-ID	63.50 (67.18)	92.11 (93.32)	96.47 (97.14)	59.69 (71.81)
HcTri + Margin MMD-ID	**64.35 (68.49)**	93.02 **(93.55)**	96.96 **(97.33)**	**60.11 (72.73)**

Table 4: Incorporating Margin MMD-ID on existing baselines (AGW [55], DGTL [27], HcTri [26]) for SYSU-MM01 dataset. For each setting metrics are reported as: All-Search (Indoor-Search)

Qualitative evaluation: To visualize the inter-class separation and intra-class compactness across the modalities (shown in Fig.3), we define a thermal and visible feature representative for each identity by calculating the centroid of image features belonging to that identity and modality. Thus, we have a visible and thermal feature vector for each identity. Ideally, discriminative yet modality-invariant features should give high intra-class and low inter-class similarity values. We calculate the intra-class similarity by finding the cosine distance between each identity's visible and thermal centroid features and calculate the mean and standard deviation, on which we fit a Gaussian distribution (Orange curve in Fig.3). Similarly, we calculate the inter-class similarity by finding the cosine distances between the visible and thermal centroid features of different identities and get the mean and standard deviation and fit a Gaussian distribution (Blue curve). Fig.3 shows that the intra-class similarity between visible and thermal pairs has increased, indicating the feature vectors of different modalities for same identity are more closer when we use Margin MMD-ID loss. As a result, the separation between the inter and intra class similarities has increased, which is needed to avoid misclassifications. To avoid outliers, we use centroids for each identity instead of individual samples. We choose this strategy of using all identities (then fitting a Gaussian over mean and standard deviation), instead of selecting few identities, so as to holistically visualise the inter-class and intra-class similarities.

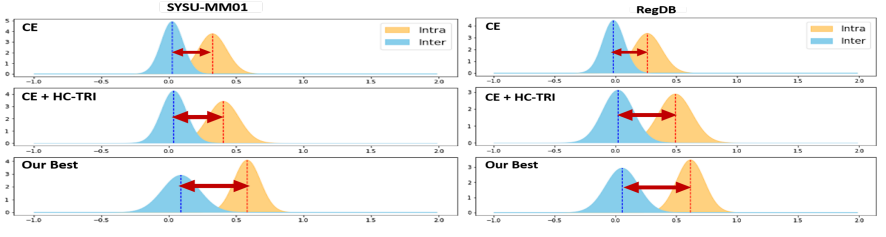

Figure 3: Plot for Gaussian fitted distributions over given mean(m) and std deviation(s) for Intra and Inter class similarities on Test identities.

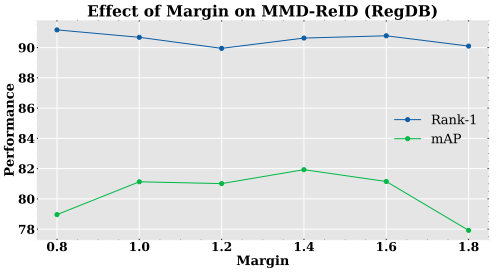

Figure 4: Sensitivity analysis for Margin on RegDB

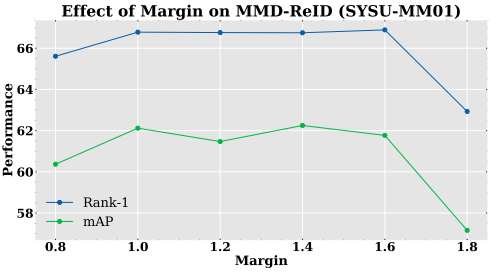

Figure 5: Sensitivity analysis for Margin on SYSU-MM01

Ablation study for Margin: We find the optimal margin value by following the similar strategy as employed by conventional methods [47], [8], [27] i.e., using validation data to tune the hyperparameters. Specifically, since 'ρ' is a hyper-parameter, we fine-tune it separately on both datasets. We perform a sensitivity analysis for the margin values (Fig. 4 for RegDB and Fig. 5 SYSU-MM01), which conveys that the performance is stable across a broad range of margins (ρ), around the optimal. Consequently, we choose ρ as 1.4 for both SYSU-MM01 and RegDB as it's the best performing margin for both datasets. It is worth noting that the stable nature of Margin MMD-ID for our configuration allowed us to keep same margin across both datasets.

Computational cost analysis: A detailed overview about the computations involved with Margin MMD-ID loss is given in the Supplementary (Section 3.5) We show that, computation wise, our loss is comparable to standard Triplet loss. We also do a training time analysis and report the hours needed to train the model for 60 epochs for different setups which confirms that the training time with MMD-ReID ($\sim$ 6 hrs) is almost same as the C.E. and C.E. + HC-Tri setup, thus making our method easily trainable.

5 Conclusion

Although the last few years have witnessed significant progress in the VT-ReID task, the current state-of-the-art methods aim to reduce the cross-modality discrepancy in an implicit fashion by aligning pixel and feature space representations using adversarial learning strategies or designing domain-knowledge reliant feature extraction modules. This paper provides a simple but effective framework for performing VT-ReID called MMD-ReID based on a margin-modification of the standard MMD. We empirically observed that using standard MMD to align identity-conditioned visible and thermal distributions in supervised VT-ReID task leads to overfitting and devise a simple margin-modification, Margin MMD-ID, to alleviate it. Extensive experimentations demonstrate the superiority of our proposed framework as well as validate the effectiveness of each component in it. We also evaluate the effect of incorporating Margin MMD-ID in existing baselines and observe that it leads to significant gains in performance. We thus urge the VT-ReID community to explore more simpler and stronger ways to solve this problem of VT-ReID.

6 Acknowledgments

This work is supported by a Young Scientist Research Award (Sanction no. 59/20/11/2020-BRNS) from DAE-BRNS, India. The Authors would like to thank the Visual Computing Lab (CDS, IISc) members for the insightful discussions and feedback on the project.

References

[1] Slawomir Bak, Sofia Zaidenberg, Bernard Boulay, and Francois Brémond. Improving person re-identification by viewpoint cues. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 175–180, 2014. doi: 10.1109/AVSS.2014.6918664.

[2] Gérard Biau and Laszlo Gyorfi. On the asymptotic properties of a nonparametric l/sub 1/-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11): 3965–3973, 2005.

[3] Peter J Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1320–1329, 2017.

[5] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 587–597, June 2021.

[6] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1362, 2016. doi: 10.1109/CVPR.2016.151.

[7] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020.

[8] Pingyang Dai, R. Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 2018.

[9] Weijian Deng, L. Zheng, Guoliang Kang, Yezhou Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018.

[10] Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[12] Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[13] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8385–8392, Jul. 2019. doi: 10.1609/aaai.v33i01.33018385. URL https://ojs.aaai.org/index.php/AAAI/article/view/4853.

[14] 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.

[15] Alexander Hermans, Lucas Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017.

[16] 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.

[17] Ruibing Hou, Bingpeng Ma, H. Chang, Xinqian Gu, S. Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7176–7185, 2019.

[18] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. Illumination-invariant person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 365–373, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350994. URL https://doi.org/10.1145/3343031.3350994.

[19] M. Kalayeh, Emrah Basaran, M. Gokmen, M. Kamasak, and M. Shah. Human semantic parsing for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[20] Kajal Kansal, AV Subramanyam, Zheng Wang, and Shin'ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[21] Srikrishna Karanam, Yang Li, and Richard J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4516–4524, 2015. doi: 10.1109/ICCV.2015.513.

[22] V. Kniaz, V. Knyaz, J. Hladůvka, W. Kropatsch, and V. Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *ECCV Workshops*, 2018.

[23] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. Learning shape representations for clothing variations in person re-identification. *arXiv preprint arXiv:2003.07340*, 2020.

[24] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/li15.html.

[25] Haijun Liu, Jian Cheng, Wen Wang, Yanzhou Su, and Haiwei Bai. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*, 398:11–19, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.01.089. URL https://www.sciencedirect.com/science/article/pii/S0925231220301478.

[26] 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.

[27] 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.

[28] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *2013 IEEE International Conference on Computer Vision*, pages 2200–2207, 2013. doi: 10.1109/ICCV.2013.274.

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016.

[30] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3071–3085, 2019. doi: 10.1109/TPAMI.2018.2868685.

[31] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

[32] 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.

[33] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Aggregating deep pyramidal representations for person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1544–1554, 2019. doi: 10.1109/CVPRW.2019.00196.

[34] 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.

[35] 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.

[36] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. *ArXiv*, abs/2003.06430, 2020.

[37] M. Sarfraz, Arne Schumann, Andreas Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[38] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. doi: 10.1109/CVPR.2018.00129.

[39] 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 Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 501–518, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01225-0.

[40] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1531, June 2021.

[41] R. Varior, Bing Shuai, Jiwen Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[42] G. Wang, S. Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and J. Sun. High-order information matters: Learning relation and topology for occluded person re-identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6448–6457, 2020.

[43] Guan-An Wang, Tianzhu Zhang Yang, Jian Cheng, Jianlong Chang, Xu Liang, Zengguang Hou, et al. Cross-modality paired-images generation for rgb-infrared person re-identification. *arXiv preprint arXiv:2002.04114*, 2020.

[44] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019.

[45] Taiqing Wang, S. Gong, Xiatian Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[46] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics*, 48(10):3006–3020, 2018. doi: 10.1109/TCYB.2017.2755044.

[47] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ich Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–626, 2019. doi: 10.1109/CVPR.2019.00071.

[48] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

[49] Hongliang Yan, Yukang Ding, P. Li, Qilong Wang, Yong Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 945–954, 2017.

[50] Mang Ye, X. Lan, Jiawei Li, and P. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[51] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1092–1099. International Joint Conferences on Artificial Intelligence Organization, 2018. doi: 10.24963/ijcai.2018/152. URL https://doi.org/10.24963/ijcai.2018/152.

[52] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020. doi: 10.1109/TIP.2020.2998275.

[53] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C. Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2020. doi: 10.1109/TIFS.2019.2921454.

[54] Mang Ye, J. Shen, David J. Crandall, L. Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. *ArXiv*, abs/2007.09314, 2020.

[55] 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.

[56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39, 2014. doi: 10.1109/ICPR.2014.16.

[57] Xuan Zhang, Hao Luo, X. Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, W. Jiang, C. Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *ArXiv*, abs/1711.08184, 2017.

[58] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 907–915, 2017. doi: 10.1109/CVPR.2017.103.

[59] L. Zheng, Yujiao Huang, H. Lu, and Y. Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28:4500–4509, 2019.

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. doi: 10.1109/ICCV.2015.133.

[61] Zhihui Zhu, X. Jiang, Feng Zheng, Xiao-Wei Guo, Feiyue Huang, W. Zheng, and Xing Sun. Viewpoint-aware loss with angular regularization for person re-identification. In *AAAI*, 2020.