MAENet: Boosting Feature Representation for Cross-Modal Person Re-Identification with Pairwise Supervision

ABSTRACT

Person re-identification aims at successfully retrieving the images of a specific person in the gallery dataset given a probe image. Among all the existing research areas related to person re-identification, visible to thermal person re-identification (VT-REID) has gained proliferating momentum. VT-REID is deemed to be a rather challenging task owing to the large cross-modality gap [23], cross-modality variation and intra-modality variation. Existing techniques generally tackle this problem by embedding cross-modality data with convolutional neural networks into shared feature space to bridge the cross-modality discrepancy, and subsequently, devise hinge losses on similarity learning to alleviate the variation. However, feature extraction methods based simply on convolutional neural networks may fail to capture the distinctive and modality-invariant features, resulting in noises for further re-identification techniques. In this work, we present a novel modality and appearance invariant embedding learning framework equipped with maximum likelihood learning to perform cross-modal person re-identification. Extensive and comprehensive experiments are conducted to test the effectiveness of our framework. Results demonstrated that the proposed framework yields state-of-the-art Re-ID accuracy on RegDB and SYSU-MM01 datasets.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Image search**.

KEYWORDS

Information retrieval, cross-modal retrieval, person re-identification, deep learning

ACM Reference Format:

. 2020. MAENet: Boosting Feature Representation for Cross-Modal Person Re-Identification with Pairwise Supervision. In *Woodstock '20: ACM Symposium on Neural Gaze Detection, June 03–05, 2020, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

While video surveillance cameras of different types and distinctive resolutions are prevalent across the world to ensure security, it has attracted growing attention to person re-identification (Re-ID) research which is targeted at accurate retrieval of a specific query person from a gallery of person images taken by disjoint surveillance cameras in different views [28][44]. Most of the current progress on Re-ID primarily focuses on only visible images, which means both the probe image and the set of gallery images are RGB images [1][9][12][14][15][16][33][34].

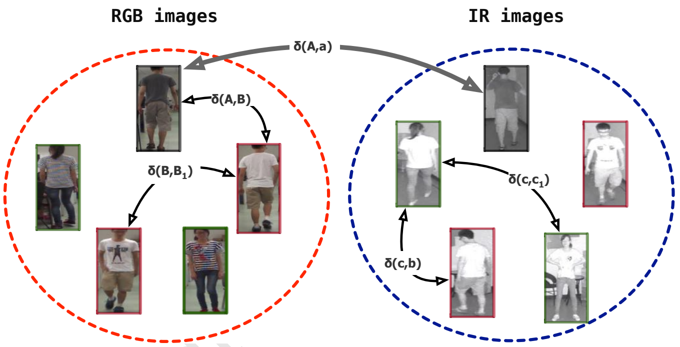

Figure 1: Illustration of the difficulty of VT-REID resulted from the modality discrepancy, cross-modality variation and intra-modality variation. Boxes with the same color denote pictures from the same person. Owing to the cross-modality discrepancy and variation, the intra-person distance $\delta(A, a)$ is larger than the inter-person distance $\delta(A, B)$. Further, as demonstrated in the right part of the picture, intra-person distance $\delta(c, c_1)$ is larger than $\delta(c, b)$ caused by intra-modality variation.

In real-world settings, however, a large proportion of criminal acts are committed at low illumination environments where traditional visible cameras usually fail to capture the distinctive appearance information. In such a case, thermal cameras which utilize infrared lights are widely adopted to capture informative human features, which highlights the necessity of directing research attention to study the cross-modality person Re-ID problem, i.e., the probe image being visible while the gallery image set being thermal. This problem is formally defined as Visible-Thermal Person Re-identification (VT-REID) by [35].

Compared with visible-to-visible person re-identification, VT-REID is particularly complicated due to the cross-modality discrepancy [22], as depicted in Fig. 1, which is caused by the distinct sensor spectrums for visible and thermal cameras. Meanwhile, as illustrated in Fig. 1, the additional cross-modal variation resulted from cross-camera views and the notable intra-modality variation caused by changing human appearances and camera viewpoints further adds to the difficulty of VT-REID.

The general process for existing methods can be summarized into two separate parts: feature embedding and similarity learning. Existing methods generally simply employ convolutional neural networks like AlexNet and ResNet to embed the features, which, we argue, does not ensure that the extracted feature vectors capture

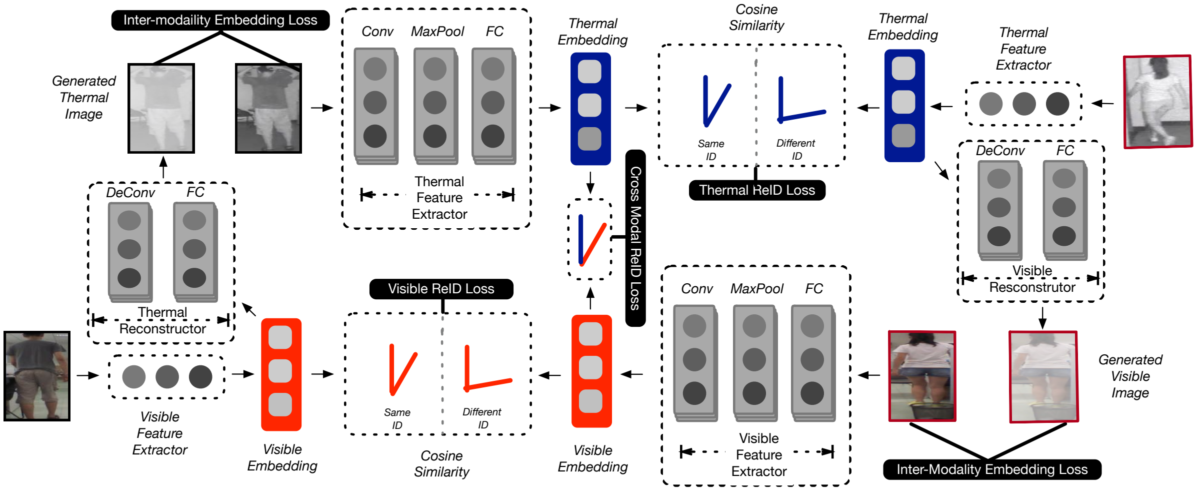

Figure 2: The framework of the proposed model. It contains two separate modality-specific embedding networks and two reconstruction networks.

the informative features which are beneficial for the subsequent similarity learning process. For example, pictures of the same person with different appearances display distinct features and thus, leading to utterly different embedding vectors for pictures from the same person. Since accurate person re-identification requires the embedding vectors of the same person being close, it inspires us to ponder the possibility of designing a framework to distill the modality and appearance invariant features from person pictures. Consider the two green boxes in the right part of Fig. 1, though they exhibit distinct visual features, the outline of both figures inside of the pictures remain alike. Therefore, to successfully identify one with another, we need to recognize the features that are invariant to appearance changes. Inspired by these observations, in this paper, we propose MAENet - an effective framework for visible to thermal person Re-identification. Specifically, we develop a novel modality and appearance invariant embedding method to extract the shared features for person images captured from different sets of cameras and distinct appearances. Following the novel embedding method, we design two pairwise constraints to the previously learned embedding vectors to preserve the similarity relationships conveyed in the training data. We summarize our contributions in the following paragraphs :

- We present a novel modality and appearance invariant feature embedding framework which can extract the shared features across different modalities and appearances. The proposed embedding method could apply to other re-identification works which requires feature embedding learning.
- We devise and derive the theoretical formulation of weighted maximum likelihood for preserving the cross-modal similarity, as well as maintaining intra-modal discriminability. To the best of our knowledge, this is the first attempt to apply maximum likelihood learning in this literature.
- We performed extensive experiments to evaluate our proposed method, and results have illustrated our superiority over previously published works.

2 RELATED WORK

2.1 Person Re-identification

We divide the related works into two aspects: handcrafted features-based methods and deep learning-based methods.

Handcrafted Features-based Methods. [17] proposes a Local Maximal Occurrence (LOMO), and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) to learn a robust feature representation and distance metric. LOMO is later adopted by [42] and [43] to further boost its performance.

Deep Learning based Methods. The state-of-the-art performances on the majority of widely-studied person Re-ID datasets are maintained by deep learning based models. The first two works which integrate deep learning into person Re-ID tasks are [39] and [13]. [31] later deepens the neural network by utilizing smaller convolutional weights. [26] further integrates long short-term memory to process image parts sequentially to enhance the discriminative ability of the deep features. In 2018, [25] introduced a part-based convolutional baseline (PCB) to produce part-level features for person retrieval. [5] presented a neural network which utilizes view information in feature extraction stage so as to mitigate the intra-class variation. [41] further explores the possibility of applying unsupervised learning in this domain. Though these methods have achieved impressive results on many datasets, the vast majority of them only tackles the visible-to-visible person Re-id and do not apply to cross-modality person re-id due to the existence of a modality gap between different image modalities.

2.2 Cross-Modal Person Re-identification

For cross-modality Re-ID tasks, researchers have investigated RGB-Depth [6][20][29], text-image [10][11][37][40]. Specifically, for VT-REID, [30] initially presents the problem of Visible-thermal Re-ID in 2017 and proposed a deep zero-padding network for shared feature

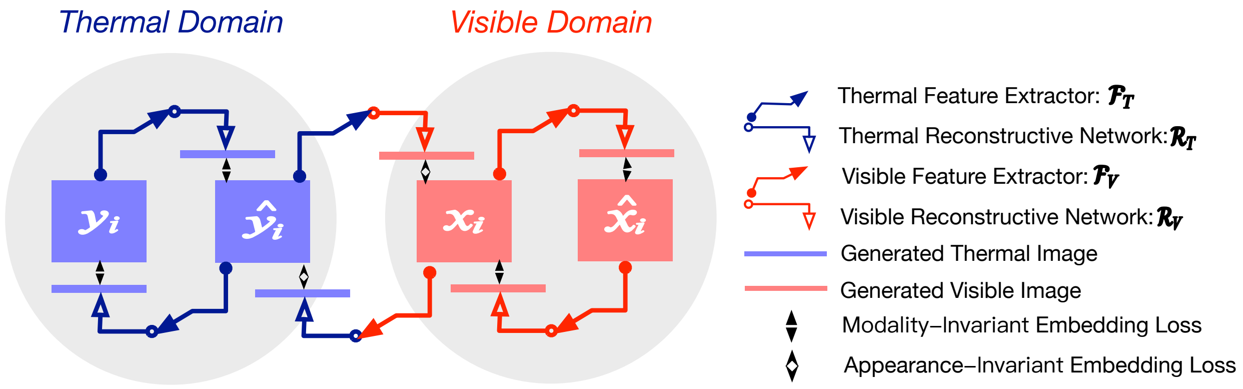
Thermal Domain
Visible Domain
y_i
$\hat{y}_i$
x_i
$\hat{x}_i$
Thermal Feature Extractor: $\mathcal{F}_T$
Thermal Reconstructive Network: $\mathcal{R}_T$
Visible Feature Extractor: $\mathcal{F}_V$
Visible Reconstructive Network: $\mathcal{R}_V$
Generated Thermal Image
Generated Visible Image
Modality–Invariant Embedding Loss
Appearance–Invariant Embedding Loss

Figure 3: The illustration for the modality and apperance invariant embedding module.

learning, which only integrates identity information, thus, contributing to the loss of discriminability of the learned representation[38]. Ye et al. [35] introduces a two-stage learning framework which incorporates feature learning and metric learning. However, [32] suggests that a two-stage framework involving human intervention may not apply to large-scale usage scenarios. Ye et al [38] later presents a dual-path end-to-end VT-REID learning framework with a bi-directional dual-constrained top-ranking loss to learn the cross-modal similarities while preserving the intra-modal discriminability. However, most of the aforementioned VT-REID mainly focus on similarity learning to perform re-identification, ignoring the importance for the extracted embedding to capture the most distinctive features in the images.

3 NOTATIONS AND PROBLEM DEFINITION

3.1 Notations

In the whole paper, we denote mappings with a calligraphic uppercase letter, such as $\mathcal{X}$; Bold-face uppercase letters such as **B** stand for sets while bold-face lowercase letters like **b** represent vectors. **U** represents vector space.

3.2 Problem Definition

Assume that $\mathbf{X} = \{k\}_{i=1}^{N} = \mathbf{V} \cup \mathbf{T}$ is the set of N training instances which contain person images from two modalities: visible images **V**, and thermal images **T**. Let $I_{k_i} = i$ be a character function to map the image to its index. $\mathcal{S} : N * N \longrightarrow \{0, 1\}$ is a similarity mapping. $\mathcal{S}_{ij} = 1$ if k_i and k_j are pictures of the same person. Meanwhile, $\mathcal{S}_{ij} = 0$ if k_i and k_j are from different people. $\mathbf{P} = \{\{k_p, k_q : S_{pq} = 1\} : p, q \in N\}$ is a set comprised of image pairs with the same identity. The target of supervised visible-thermal person Re-ID is to learn two embedding mappings $\mathcal{F}_V : \mathbf{V} \longrightarrow \mathbf{U}, \mathcal{F}_T : \mathbf{T} \longrightarrow \mathbf{U}$, to satisfy : given any v in the probe visible image set $\mathbf{V}' \subset \mathbf{V}$ and the gallery $\mathbf{T}' \subset \mathbf{T}$, find $k \in \mathbf{T}'$, where $S_{I_v I_k} = 1$ and $k = argmin_t d(\mathcal{F}_V(v), \mathcal{F}_T(t))$. d is a pre-defined distance measure, e.g. the cosine similarity. The probe images set could also be thermal, while the gallery set being visible.

4 PROPOSED MODEL

This paper presents a deep learning-based framework for VT-REID, which incorporates enhanced feature representation learning and similarity learning and optimizes them in an end to end manner. The proposed architecture is illustrated in Fig. 2.

4.1 Hybrid Deep Feature Extraction Architecture

The hybrid deep feature extraction architecture is shown in Fig. 2, which constitutes two modality-specific deep convolutional neural networks ($\mathcal{F}_V$ and $\mathcal{F}_T$) to extract features from visible and thermal image, respectively. Both feature extractor networks employ identical network structures, which are extended from Resnet-50 [7] with a fully connected layer to project the features into a shared embedding space. Note that the fully connected layers for two feature-extracting networks share model parameters to enhance the learning of modality-shareable information.

4.2 Modality and Appearance Invariant Reconstructive Embedding

To further ensure the above embedding vectors for data from both modalities incorporate the necessary discriminative information, we ponder the possibility of adding additional constraints on the embedding vectors. Inspired by [3], which employs an AutoEncoder-like network structure to distill the correlation between cross-modal data, we design the network structure as illustrated in Fig. 3. The loss function for modality invariant embedding is formulated as :

$$L_{MI} = \sum_{\mathbf{Q} \in \mathbf{P}} \sum_{x_i, y_i \in \mathbf{Q}} (||x_i - \mathcal{R}_V(\mathcal{F}_T(y_i))||_2^2 + ||y_i - \mathcal{R}_T(\mathcal{F}_V(x_i))||_2^2) \quad (1)$$

where $\mathcal{R}_V$ and $\mathcal{R}_T$ are two reconstructive neural networks which take embedding vectors as inputs and reconstruct the corresponding input visible image $x_i \in \mathbf{V}$ and thermal image $y_i \in \mathbf{T}$. x_i, y_i are from the same person identity. By minimizing the cross-modality reconstruction error L_{MI}, intuitively, the embedding vectors are pushed to capture the modality-invariant features.

Since VT-REID also suffers from large intra-modality variation, we further devise an appearance-invariant constraint to ensure the embedding vectors capture the shared information across different pictures for the same person in one modality. The demonstration is in the right part of Fig. 3. Below we formulate the formal appearance-invariant embedding as :

$$L_{AI} = \sum_{\mathbf{Q}\in\mathbf{P}} \sum_{x_i,\hat{x}_i,y_i,\hat{y}_i\in\mathbf{Q}} (||x_i - \mathcal{R}_V(\mathcal{F}_V(\hat{x}_i))||_2^2 + ||\hat{x}_i - \mathcal{R}_V(\mathcal{F}_V(x_i))||_2^2 + ||y_i - \mathcal{R}_T(\mathcal{F}_T(\hat{y}_i))||_2^2 + ||\hat{y}_i - \mathcal{R}_T(\mathcal{F}_T(y_i))||_2^2) \tag{2}$$

where L_{AI} denotes the integrated intra-modality appearance-invariant embedding loss for visible and thermal modality. $\hat{x}_i \in \mathbf{V}, \hat{y}_i \in \mathbf{T}$. Similarly, $x_i, \hat{x}_i$ are images from the same person in visible domain while $y_i, \hat{y}_i$ are images for the same identity in thermal domain. By minimizing L_{AI}, the embedding vectors are pushed to preserve the appearance-invariant information conveyed in multiple pictures for the single person in one modality. Below, we formulate the total reconstructive embedding loss as:

$$L_{RE} = L_{MI} + L_{AI} \tag{3}$$

By minimizing L_{RE} , the learned embedding vector for one image will capture modality and appearance invariant features simultaneously.

4.3 Pairwise Relationship Guided Similarity Learning

For efficient and effective cross-modal person re-identification, assume that we have two image instances x_i and x_j which belong to the same identity, their corresponding feature vectors **a** and **b** should have short cosine distance, and vice versa. To better ensure the above statement, we devise our objective function by incorporating three separate kinds of loss functions: the inter-modal pairwise re-identification loss, the intra-modal pairwise re-identification loss, and the identity loss. The pairwise re-identification loss, intuitively, is to reinforce the similarity of the pairwise feature vectors extracted from image instances of the same identity and diminish the similarity of those pairwise vectors from different identities.

Suppose we have a training set $\mathbf{X} = \{x_i\}_{i=1}^B \cup \{y_j\}_{j=1}^B, B < N$, we can get the features $\mathbf{H} = \{h_i^V\}_{i=1}^B \cup \{h_j^T\}_{j=1}^B$ where $h_i^V = \mathcal{F}_V(x_i)$ since x_i is a visible image otherwise $h_j^T = \mathcal{F}_T(y_j)$ since y_j belongs to thermal images. The similarity label is written as $s_{ij}^{mn} = S_{I_{k_i^m}\ I_{k_j^n}}$ where m and n represent the modality. $I_{k_i^m}$ denotes the index of i_{th} image in the modality of m. Basically, $s_{ij}^{VT} = 1$ means i_{th} and j_{th} image pair (k_i^V, k_j^T) are from visible and thermal modality respectively and they are of the same person identity. The likelihood function is defined as follows:

$$p(s_{ij}^{mn}|h_i^m, h_j^n) = \sigma(d(h_i^m, h_j^n))^{s_{ij}^{mn}}(1 - \sigma(d(h_i^m, h_j^n)))^{(1-s_{ij}^{mn})} \tag{4}$$

where $d(h_i^m, h_j^n) = \cos(h_i^m, h_j^n)$, which is the cosine similarity between h_i^m and h_j^n. Meanwhile, $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. $p(s_{ij}^{mn}|h_i^m, h_j^n)$ denotes the conditional probability of similarity label s_{ij}^{mn} given the extracted feature vectors h_i^m and h_j^n. Meanwhile, we define a weight parameter w_{ij}^{mn} to tackle the data imbalance problem [2] as:

$$w_{ij}^{mn} = \begin{cases} |\mathrm{S}|/|\mathrm{S}_1^{mn}|, & s_{ij}^{mn} = 1 \\ |\mathrm{S}|/|\mathrm{S}_0^{mn}|, & s_{ij}^{mn} = 0 \end{cases} \tag{5}$$

where $\mathrm{S}_1^{mn} = \{(i,j) : s_{ij}^{mn} = 1\}$ is the set of pairs where each pair belongs to the same identity and $\mathrm{S}_0^{mn} = \{(i,j) : s_{ij}^{mn} = 0\}$ is the set of pairs where each pair belongs to different persons.

Below, we will formally derive and demonstrate the formulation of the inter-modal pairwise re-identification loss. The weighted logarithm maximum likelihood estimation loss for cross-modal re-identification L_{Inter} is formulated as

$$\begin{aligned} L_{Inter} &= -\log p(\mathrm{S}|\mathbf{H}) = -\sum_{i,j} w_{ij}^{VT} \log p(s_{ij}^{VT}|h_i^V, h_j^T) \\ &= -\sum_{i,j} w_{ij}^{VT}(s_{ij}^{VT} d(h_i^V, h_j^T) - log(1 + e^{d(h_i^V, h_j^T)})) \end{aligned} \tag{6}$$

It is rather obvious that optimizing the above loss will lead to the increment of cosine similarity of similar pairs and the reduction of cosine similarity between dissimilar pairs. Since the inter-modal loss does not consider the intra-modal variation problem, it is natural and necessary to add two intra-modal pairwise re-identification losses for thermal images and visible images, respectively.

For thermal image modality, the training pairs (y_i, y_j) are both thermal images. Analogously, the intra-modal embedding loss for thermal image modality can be formulated as:

$$L_{Intra\_Thermal} = -\sum_{i,j} w_{ij}^{TT}(s_{ij}^{TT} d(h_i^T, h_j^T) - log(1 + e^{d(h_i^T, h_j^T)})) \tag{7}$$

In a same manner, the pairwise re-identification loss for visible image modalities is derived as:

$$L_{Intra\_Visible} = -\sum_{i,j} w_{ij}^{VV}(s_{ij}^{VV} d(h_i^V, h_j^V) - log(1 + e^{d(h_i^V, h_j^V)})) \tag{8}$$

The total intra-modal pairwise re-identification loss L_{Intra} is formulated as:

$$L_{Intra} = L_{Intra\_Visible} + L_{Intra\_Thermal} \tag{9}$$

To further boost the discriminability of learned representations, as suggested by [38], we further integrate the identity loss $L_{Identity}$ for both modalities by treating each identity as a class and utilizing softmax cross-entropy to guide the classification process. Intuitively, the identity loss could propel the learned representation to preserve the identity-related information so as to correctly classify each identity, thus, alleviating the headache brought by the considerable modality variation.

The overall learning objective could be obtained by summing up the above separate losses: the reconstructive embedding loss, the cross-modality pairwise embedding loss, two separate intra-modal embedding loss, and the identity loss for each person. The total loss is then formally defined as:

$$L_{total} = L_{Inter} + \lambda_1 L_{Intra} + \lambda_2 L_{Identity} + \lambda_3 L_{RE} \tag{10}$$

where λ_1, λ_2, and λ_3 are three pre-defined coefficients.

4.4 Optimization

The whole convolutional neural network structure is optimized with stochastic gradient descent (SGD). The detailed optimization process for the entire framework is illustrated below in Alg. 1.

Table 1: Vsible to Thermal Test Results on State of The Art Methods

Datasets	RegDB				SYSU-MM01*(Single-Shot all search)*			
Methods	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP
HOG	13.49	33.22	43.66	10.31	2.76	18.25	31.91	4.24
MLBP	2.02	7.33	10.90	6.77	2.12	16.23	28.32	3.86
LOMO[16]	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
GSM[19]	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
SVDNet[24]	17.24	34.12	44.51	19.04	14.64	53.28	64.24	15.17
PCB[25]	18.32	36.42	46.51	20.13	16.43	54.06	65.24	16.26
TONE+HCML	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
Zero-Padding[30]	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
cmGAN[4]	-	-	-	-	26.97	67.51	80.56	27.80
BDTR (ResNet50)[38]	30.56	54.62	65.42	32.45	27.32	66.96	81.07	27.32
MAENet (ResNet50)	44.32	63.98	72.67	45.55	29.79	77.24	89.87	36.11

Algorithm 1: The training algorithm for our framework

```
Input: Training image set X, similarity labels $\mathcal{S}$, learning
       rate $\alpha$ and parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ maximum training
       iteration $T$
Output: network parameters $\theta_V$, $\theta_T$, $\theta_{RV}$, $\theta_{RT}$;
Initialize the network parameters $\theta_V$ and $\theta_T$ with pretrained
  weights on ImageNet , $\theta_{RV}$ and $\theta_{RT}$ randomly, t = 0
while not converged and $t < T$ do
    $t = t + 1$;
    Sample $Q_1, Q_2 \in \mathbf{P}, Q_1 \cap Q_2 = \emptyset$ from X to get
      $x_i, \hat{x}_i, y_i, \hat{y}_i \in$Q1,$x_j, \hat{x}_j, y_j, \hat{y}_j \in Q_2$ ;
    Calculate $L_{RE}$
      with $\mathcal{F}_V(\bullet;\theta_V)$, $\mathcal{F}_T(\bullet;\theta_T)$, $\mathcal{R}_V(\bullet;\theta_{RV})$, $\mathcal{R}_T(\bullet;\theta_{RT})$ by
      Eq. 1, 2, 3 ;
    For $p \in \{i,j\}$, $h_p^V = \mathcal{F}_V(x_p;\theta_V)$, $h_p^T = \mathcal{F}_T(y_p;\theta_T)$ ;
    Calculate $L_{Inter}$ given $s_{ij}^{VT}, w_{ij}^{VT}, h_i^V, h_j^T$ by Eq. 9;
    Calculate $L_{Intra\_Thermal}$ and $L_{Intra\_Visible}$ similarly
      to get $L_{total}$ by Eq. 10;
    Back propagate the $L_{total}$ to get gradients
      $\nabla_{\theta_*} * \in \{V, T, RV, RT\}$ ;
    Update $\theta_* = \alpha(\theta_* - \nabla_{\theta_*})$ ;
```

Table 2: Cross-modal Re-Identification Results on the SYSU-MM01 Dataset with Indoor-Search Single-Shot Mode

Methods	r=1	r=10	r=20	mAP
HOG	3.22	24.68	44.52	7.25
MLBP	3.43	26.42	45.36	7.72
LOMO	2.24	22.53	41.53	6.64
GSM	9.46	48.98	72.06	15.57
SVDNet[24]]	20.24	64.32	83.62	28.74
PCB[25]	22.63	65.24	83.92	30.46
Zero-Padding[30]	20.58	68.38	85.79	26.92
TONE[35]	20.82	68.86	84.46	26.38
cmGAN[4]	31.63	77.23	89.18	42.19
BDTR (ResNet50)[38]	31.92	77.18	89.28	41.86
MAENet(ResNet50)	38.54	86.10	95.61	49.72

5 EXPERIMENTS

In this section, we conduct comprehensive experiments to demonstrate the efficacy of our proposed network. In the subsequent paragraphs, we first describe the dataset settings, and the evaluation metrics. Then, we compare the test performance with state-of-the-art models to corroborate its effectiveness. At last, we conduct ablation analysis to further verify the necessity and efficacy of different parts of the whole integrated model.

5.1 Datasets and Evaluation Protocols

Datasets. There exists two standard datasets for this scenario: SYSU-MM01 [30] and RegDB [21]. RegDB contains images captured by two different cameras for 412 person identities, where each person has ten visible images and ten thermal images. SYSU-MM01 is a large-scale VT-REID dataset captured by six cameras: four RGB cameras and two IR cameras in a campus setting. It contains 491 identities in total: 395 for training and 96 for testing. As described in [30], two test modes are adopted for this dataset, i.e. *all-search* mode and *indoor-search* mode. *all-search* consider all the images in the dataset while *indoor-search* mode excludes the visible images captures by outdoor cameras 4 and 5. *indoor-search* is deemed to be easier since it only considers visible images taken inside of the room.

Evaluation Metrics. We adopt two widely-used evaluation metrics for this task: the Cumulative Matching Characteristic (CMC) and mean average precision (MAP). We adopt the single-shot test setting, as suggested by [36], through randomly selecting one image for each person to form the gallery set.

5.2 Experiment Results on Two Datasets

We compare our model with state-of-the-art methods targeted at solving VT-REID problems. Specifically, we include **TONE+HCML** [35], **cmGAN** [4], **Zero-Padding** [30], **BDTR** [36] for detailed analysis and comparison. We also include include several other cross-modality learning methods for comparison. Most of the test results are from [36]. The selected competing learning methods can

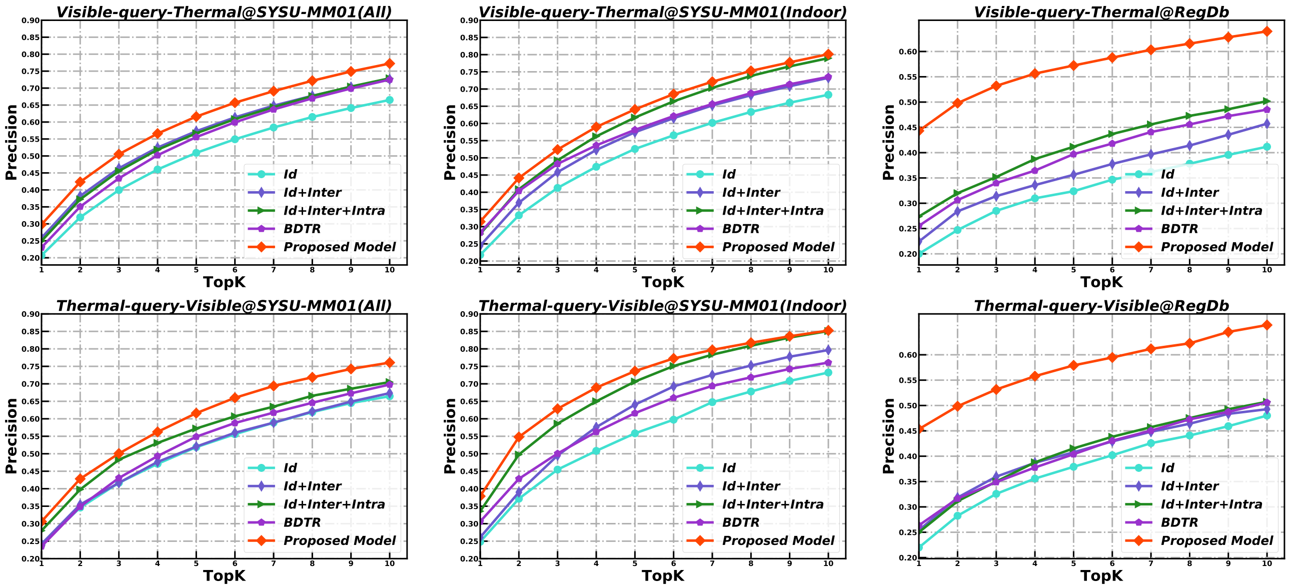

Figure 4: The Re-Identification Rates on Two Settings

be mainly categorized into two classes: feature extraction based methods: (**HOG**, **MLBP**, **LOMO** [16]), and matching model learning methods: **XQDA** [16], **MLAPG** [18], **GSM** [19], **SCDL** [27], **rCDL** [8]. As suggested by [38], we further include state-of-the-art single-modality ReID methods([24], [25]) for comparison. As displayed in Table 1 and Table 2, It is easy to spot that all the competing methods have gained significant performance improvements from *all-search mode* to *indoor-search mode* in SYSU-MM01. The increased test results for *indoor search* could be attributed to the diminished intra-modality variation since the *indoor-search* mode only incorporates visible images captured indoor. It is also rather evident from the table that the traditional cross-modality learning methods and methods designed for single-modality re-identification scenarios did not exhibit satisfactory performance since they failed to consider the discrepancy between data from distinct modalities. When compared with aforementioned **VT-REID** methods, our model is a clear winner in each evaluation metric across both datasets. Note that our method surpasses **TONE+HCML** and **Zero-Padding** in a noticeable margin in both datasets. **BDTR** achieves comparable results in this scenario. In RegDB, We exceed **BDTR** by 13.1% in mean average precision(MAP), and in SYSU, we surpass **BDTR** by near 9%. Meanwhile, our proposed model also outperforms all of these methods in the rank-1 re-identification rate.

5.3 Ablation Study

In this section, we thoroughly investigate the effectiveness of different components of our final integrated model. The results are presented in Fig. 4. Specifically, we separately test the effectiveness of our model with pure identity loss (denoted in the figure as *Solo Id*), identity loss plus inter-modal pairwise re-identification loss(denoted as *Id + Inter*) , identity loss plus inter and intra-modal re-identification loss (denoted as *Id + Inter +Intra*) and our full model. Note that we also include *BDTR* for ablation comparison since we share the same backbone feature extraction network structure. Clearly, from Fig. 4, the model with pure identity loss exhibits the lowest results in all datasets and across both re-identification scenarios. When combined further with *Inter and Intra* re-identification loss, as depicted by the green line, it has achieved notable performance increases, outperforming **BDTR** with a notable margin. Finally, when we further integrate the novel modality and appearance invariant embedding module into the framework, a marked surge is spotted. Specifically, it leads to additional increases in rank-1 re-identification accuracy in visible-to-thermal setting from 27.28%, 24.93% to 44.32%, 32.33% in RegDB and SYSU-MM01(all) respectively. Similar improvements can also be discovered in thermal to visible settings with rank-1 re-identification accuracy jumping from 25.00% to 45.34% in RegDB and 33.56% to 37.86% in SYSU-MM01(*indoor*). The notable performance surge manifests the validity and effectiveness of our novel reconstructive embedding method for the entire VT-REID framework.

6 CONCLUSION

In this paper, we introduce a novel modality and appearance invariant embedding module to distill the features that are common across two modalities and different appearances for images from the same person. Further, we devise two kinds of re-identification constraints based on maximum likelihood estimation to perform similarity learning. Extensive experimental results have illustrated our superiority against competing methods.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3908–3916.

[2] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang. 2018. Deep cauchy hashing for hamming space retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1229–1237.

[3] Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu. 2016. Correlation autoencoder hashing for supervised cross-modal search. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 197–204.

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training.. In *IJCAI*. 677–683.

[5] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. 2018. Learning view-specific deep networks for person re-identification. *IEEE Transactions on Image Processing* 27, 7 (2018), 3472–3483.

[6] Albert Haque, Alexandre Alahi, and Li Fei-Fei. 2016. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1229–1238.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] De-An Huang and Yu-Chiang Frank Wang. 2013. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Proceedings of the IEEE international conference on computer vision*. 2496–2503.

[9] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2288–2295.

[10] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.

[11] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 152–159.

[13] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 152–159.

[14] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. 2015. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3765–3773.

[15] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. 2013. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3610–3617.

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.

[17] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.

[18] Shengcai Liao and Stan Z Li. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3685–3693.

[19] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. 2016. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1089–1102.

[20] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 2014. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4512–4519.

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

[22] M Saquib Sarfraz and Rainer Stiefelhagen. 2017. Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision* 122, 3 (2017), 426–438.

[23] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 12 (2000), 1349–1380.

[24] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. 2017. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 3800–3808.

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.

[26] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. 2016. A siamese long short-term memory architecture for human re-identification. In *European conference on computer vision*. Springer, 135–153.

[27] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2216–2223.

[28] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. 2016. Scale-Adaptive Low-Resolution Person Re-Identification via Learning a Discriminating Surface.. In *IJCAI*. 2669–2675.

[29] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing* 26, 6 (2017), 2588–2603.

[30] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 5380–5389.

[31] Lin Wu, Chunhua Shen, and Anton van den Hengel. 2016. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255* (2016).

[32] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. 2018. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. *Pattern Recognition* 76 (2018), 727–738.

[33] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. 2016. An enhanced deep feature representation for person re-identification. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–8.

[34] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1249–1258.

[35] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[36] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. 2019. Bi-directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE Transactions on Information Forensics and Security* (2019).

[37] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. 2015. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 547–550.

[38] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. 2018. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking.. In *IJCAI*. 1092–1099.

[39] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 34–39.

[40] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. 2017. Adversarial attribute-image person re-identification. *arXiv preprint arXiv:1712.01493* (2017).

[41] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. 2019. Unsupervised Person Re-identification by Soft Multilabel Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2148–2157.

[42] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1239–1248.

[43] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. 2016. Sample-specific svm learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1278–1287.

[44] Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence* 40, 5 (2017), 1224–1244.