Memory Regulation and Alignment toward Generalizer RGB-Infrared Person Re-identification

Feng Chen¹, Fei Wu¹, Qi Wu², Zhiguo Wan³

¹Nanjing University of Posts and Telecommunications, China
²The University of Adelaide, Australia
³Zhejiang Lab, China

chenfeng1271@gmail.com.

Abstract

The domain shift, coming from unneglectable modality gap and non-overlapped identity classes between training and test sets, is a major issue of RGB-Infrared person re-identification. A key to tackle the inherent issue - domain shift- is to enforce the data distributions of the two domains to be similar. However, RGB-IR ReID always demands discriminative features, leading to over-rely feature sensitivity of seen classes, e.g., via attention-based feature alignment or metric learning. Therefore, predicting the unseen query category from predefined training classes may not be accurate and leads to a sub-optimal adversarial gradient. In this paper, we uncover it in a more explainable way and propose a novel multi-granularity memory regulation and alignment module (MG-MRA) to solve this issue. By explicitly incorporating a latent variable attribute, from fine-grained to coarse semantic granularity, into intermediate features, our method could alleviate the over-confidence of the model about discriminative features of seen classes. Moreover, instead of matching discriminative features by traversing nearest neighbor, sparse attributes, i.e., global structural pattern, are recollected with respect to features and assigned to measure pairwise image similarity in hashing. Extensive experiments on RegDB [1] and SYSU-MM01 [2] show the superiority of the proposed method that outperforms existing state-of-the-art methods. Our code is available in https://github.com/Chenfeng1271/MGMRA.

1 Introduction

RGB-Infrared person re-identification (RGB-IR ReID) [3; 4; 5], regarded as the cross-modality image retrieval, aims to associate a person of interest across different scenes or disjoint camera views. As a vital support of surveillance system, deep learning based methods have become dominate components and achieved promising performance in recent years [6]. However, the open-set (zero-shot setting) and cross-modality nature of RGB-IR ReID is still challenging to the algorithm generalization in searching a person in real world.

Current RGB-IR ReID methods [5; 6; 7]mainly focus on mitigating the modality discrepancy by learning attention-based visual similarity with objective functions, e.g., cross-entropy loss for classification and triplet loss for clustering. Such kind of classification-based retrieval model [8] exploits category recognition as a proxy task to learn the projection function from data to the feature space, which is allured by the discriminative region of seen categories in training set [9; 10]. Therefore, this paradigm tailors for searching objects with same categories, instead of open-set retrieval like ReID.

Some early attempts [11; 12] already redirected the attribution decision of classification-based retrieval model toward generalizer and adaptable ReID. For example, SBSGAN [13] suppresses the background which behaves as noise in cross-camera image matching, so as to look closer to the foreground person content. SSG [11] harnesses the unlabeled samples to exploit the potential cross-domain similarity. We argue, however, that using extra GAN or data is suboptimal. Therefore, we look back on the details of classification-based ranking retrieval.

We simply denote the attribution decision of RGB-IR ReID
as \( p(y|x) = \text{softmax}(fc(GAP(f))) \) and attribution similarity as \( p(x^1, x^2) = \text{Similarity}(GAP(f^1), GAP(f^2)) \) where \( f = F(x) \) is the immediate features extracted from image \( x \). This pre-GAP (global average pooling) distills informative semantic features into fragile logit value. Moreover, the distilled semantic features sequentially is used for matching their discriminative regions after alignment. As shown in Figure 1 (a), the feature matching is simplified to traverse nearest neighbor via a \( 3 \times 3 \) similarity metric. It has been a popular choice as an attribution method with many follow-up variants [14, 15], but still lacks interpretability for domain shift in ReID case. What activates this inductive bias of over-relying discriminative feature maps and how to alleviate the strong focus on seen classes during the training phase?

In this paper, we build a probabilistic graphical model to reconsider alleviating domain shift in RGB-IR ReID, involving location-wise recognition cue \( z \in \mathbb{R}^{H \times W} \), category label \( y \in \{Y \mid \text{in exponential magnitude}, \text{resulting in heavy bias toward memory regulation and alignment (MG-MRA) module to shift in RGB-IR ReID, and then propose a multi-granularity} \text{tributions can be listed as follows:} \)

1) We analyze on the lack of interpretability for domain shift in RGB-IR ReID, and then propose a multi-granularity memory regulation and alignment (MG-MRA) module to solve this issue. The proposed MG-MRA is plug-and-play and more effective than previous GAN based or extra-data based methods.

2) The learned coarse-to-fine prototypes can consistently provide domain-level semantic templates with various granularity, meeting the requirement for multi-level semantic alignment.

3) Our proposed MG-MRA boosts the performance of baseline and existing state of the arts, e.g., AGW [5] and HCT [16] by a large margin with limited consumptions. We achieve a new state-of-the-art on RegDB [11] and SYSU-MM01 [2] with 94.59%/88.18% and 72.50%/68.94% Rank1/mAP respectively.

2 Related Work

Domain Shift in ReID: The domain discrepancy of ReID mainly comes from the intra-set gap (e.g., modality or cross-camera variance) and inter-set gap (e.g., category variance) [5]. One intuitional solution is to enforce the data distributions of the two domains to be similar in latent space by zero-shot learning (ZSL) [17; 18; 19]. A line of works [7; 3] focus on leveraging generative adversarial networks (GAN) to transfer the data from one domain to another. PTGAN [20] proposes a person transfer to take advantages of existing labeled data from different datasets. cmGAN [21] reduces the distribution divergence of RGB and IR features by explicitly transferring RGB images to IR. However, this unpaired GAN-based image synthesis is always costly and hard to optimize. Another line based on metric learning, e.g., triplet loss [22], pushes the samples belonging to different identities to be dissimilar and pulls that of the same identity to be close. [22] proposes a label distillation strategy to guide the model to focus on confident samples. Cycle loss [23] intentionally re-weights each similarity to highlight the less-optimized similarity scores. However, this kind of works mainly concentrates on solving intra-set domain shift, ignoring the intrinsic inter-set discrepancy. In this work, we aim to propose a plug-and-play method to handle both intra-set and inter-set domain discrepancy simultaneously without involving additional GAN or data.

Memory Network: Memory based methods have been explored for solving various problems [24; 25]. MemAE [26] proposes a memory-augmented autoencoder to retrieve the most relevant memory items for reconstruction in unsupervised anomaly detection. [27] uses a similar memory module to alleviate the forgetting issue by recording the patterns seen in mini-batch training. [28] proposes LF-GAA to interact jointly both low-level visual information and global class-level features for semantic disambiguation in zero-shot learning. However, these expertise of classification can not be directly applied in person re-identification which is re-graded as the open-set image retrieval with both inter-set and intra-set domain discrepancy. In this paper, we reformulate the classification-based ranking retrieval based on the interpretable probabilistic treatment of the last layers of CNNs. Then we design a hierarchical memory regulation and alignment module to alleviate domain shift.

Feature Regulation on ReID: A line of multi-task works
 delve to improve ReID performance by extra annotations or
expertise from general computer vision tasks. EANet [29]
and PDC [30] leverage part segmentation and pose estima-
tion to accurately align discriminative cues. Lin et. al [31]
explicitly investigated additional attributes of pedestrians to
assist ReID where attribute recognition logit is reweighed
and combined with ReID features for final prediction. These
multi-task based methods provide non-overlapped supervi-
sion beyond current ReID where we believe it as task regula-
tion to avoid over-emphasise on ReID. Another interesting
phenomenon lacking explanation is that PCB [32] module
provides significant improvement over other soft-attention
module, such as non-local attention, even though this hand-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
module, such as non-local attention, even though this hand-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted stripped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted stripped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted stripped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
gion than attention mask. For example, FBP-AL [33] intro-
crafted striped division seems to less care about salient re-
}
In the lower-level representation, we keep the lower-level representative patterns.

For intra-

\[ \sum_{j=1}^{N} w_{ij} \mathbf{m}_j \],

\[ d(\mathbf{f}_i, \mathbf{m}_j) = \frac{\mathbf{f}_i \mathbf{m}_j^T}{\|\mathbf{f}_i\| \|\mathbf{m}_j\|}, \]

where \( \mathbf{f}_i \) and \( \mathbf{m}_j \) are feature and prototype slice from input \( \mathbf{f} \) and prototype metric \( M \). \( w_{ij} \) is the normalized weight measuring the cosine similarity \( d(\cdot, \cdot) \) between \( \mathbf{f}_i \) and \( \mathbf{m}_j \). Thus, the assigned prototype \( h \) from feature \( \mathbf{f} \) could be calculated as:

Based on single-granularity memory module, we build a multi-granularity memory module (MG-MRA) as shown in Figure 2. MG-MRA consists of hierarchical semantic prototypes \( M \), i.e., part-instance-semantic, to avoid over-\n
\[ h = SG-MRA(\mathbf{f}, M) = \sum_{i=1}^{H \times W} \sum_{j=1}^{N} w_{ij} \mathbf{m}_j. \]

In implementation, MG-MRA is applied as an auxiliary branch which is only used to regulate training process. We adopt PCB module to achieve basic state-of-the-art performance where each striped feature also retrieves corresponding prototype from our memory module.

\[ h_i = MG-MRA(\mathbf{f}_i, M) + \mathbf{f}_i. \]

From Equation 6, the final output \( h_i \) is represented by a simple scalar · prototype index, which insights on how general prototype pattern shows in this image, rather than extracting salient features. Thus, we believe this memory module works like hashing for alignment, but is more flexible: the value is continuous, instead of binary, the index is learnable from the whole domain, instead of predefined.

### 4.3 Total Loss

We adopt all default loss objectives, i.e., hetero-center triplet loss \( L_{hc,tri} \) and identity loss \( L_{id} \), used in HCT [16] without
changing any hyper-parameters or structures. Therefore, we mainly introduce our memory loss function below.

Essentially, we expect the large prototype metric $M$ to be informative enough to encode diverse representative patterns, i.e., $\text{rank}(M) \rightarrow 2 \times (P \times I \times S \times N_r)$. However, directly forcing it would result in chaos for semantic grouping. Considering the metric-decomposition-like hierarchical summarization, it could be easily achieved by observing $h_{ins}$ and $h_{sem}$ in two stages: (1) let $\{h_{ins}\}_1$ of all striped features to be similar to make sure $m_{part}$ center at higher-level meaning; (2) let $\{h_{sem}\}$ of a mini-batch to be distinct to ensure the prototype index with semantic marginalization.

Specifically, we use Maximum Mean Discrepancy (MMD) with MSE Loss to measure part-level consistency. For the items in $\{h_{ins}\}_1$, we first randomly split it to two halves $\{h_{ins}\}_1$ and $\{h_{ins}\}_2$ to avoid setup bias, and then follow Equation 10 to achieve consistent instance prototype representation of each part.

$$
L_{ins} = \text{MSE}(\{h_{ins}\}_1, \{h_{ins}\}_2). \quad (10)
$$

Besides, for semantic marginalization, we adopt more flexible triplet loss $L_{sem}$, which pushes the embedding of samples sharing same ID to be close, and pull that of different ID to be far:

$$
L_{sem} = \max(d(a, p) - d(a, n) + \text{margin}, 0), \quad (11)
$$

where $a, p, n$ are the anchor, positive and negative samples respectively where $a$ could be arbitrary samples in $\{h_{sem}\}$. Finally the whole objectives could be assembled as:

$$
L_{total} = L_{nc, tri} + L_{id} + \lambda_1 L_M + \lambda_2 L_{ins} + \lambda_3 L_{sem}, \quad (12)
$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are predefined tradeoff parameters.

### 4.4 Theoretical Properties of MG-MRA

Now we build the probabilistic graphical model of our method and then revisit preliminary to examine its inherent properties on alleviating over-confidence of discriminative features of seen classes.

Our MG-MRA aims to retrieve memory prototypes $m$, which is independent from the attention-like module i.e., finding discriminative cues $z$. Therefore, as shown in Figure 1 (c), $p(y|x)$ in training could be reformulated as:

$$
p(y|x) = \sum_z \sum_m p(y|z, m|x) = \sum_z \sum_m p(y|z, m, x) \cdot p(z|x) \cdot p(m|x) \quad (13)
$$

Similar to Equation 4, the estimation of $z$ could be represented as:

$$
- \log p_{\theta'}(\hat{y}|\hat{x}) \leq - \sum_z \sum_m p_{\theta'}(z|\hat{x}, \hat{y}) p_{\theta''}(m|\hat{x}, \hat{y}) \cdot \log p_{\theta''}(\hat{y}, z|\hat{x}) \quad (14)
$$

where $\theta''$ is the model parameter of MG-MRA. Three reasons show MG-MRA is a work focusing on regulation. (1) Note that $p_{\theta''}(m|\hat{x}, \hat{y}) \leq 1$ and $p_{\theta''}(y, m|\hat{x}) \leq 1$, comparing to Equation 4, the model interest of expectation step is adjusted by $p_{\theta''}(m|\hat{x}, \hat{y})$ and pseudo-targets generation of maximization step is dissuaded by multiplying $p_{\theta''}(\hat{y}, m|\hat{x})$. (2) Moreover, due to the log operator, the gradient balance between $E$ and $M$ steps is also adjusted. (3) Different from discriminative cues $z$ which is dependent on $x$, memory prototype $z$ is inert and semi-independent to any specific sample but memorizes general pattern from all domain. Finally, during inference, we remove $m$ and recover model confidence on discriminative features, which is identical as the inference of classical paradigm without extra computation:

$$
p(y|x) = \sum_z p(y, z|x) = \sum_z p(y|x, z) \cdot p(z|x) \quad (15)
$$

### 5 Experiments

#### 5.1 Experimental Settings

**Datasets and evaluation metrics:** Two benchmark datasets, i.e., RegDB [1] and SYSU-MM01 [2] are used to evaluate the effectiveness of our method. RegDB contains 412 identities in which each identity correspondingly includes ten visible and ten far-infrared images. Following [16], we randomly split it to training set and test set equally, all of which has 2,060 visible images and 2,060 infrared images. SYSU-MM01 is a larger-scale dataset dedicated to RGB-IR ReID, containing in total 30,071 visible images and 15,792 infrared images of 491 identities. The training set has 22,258 visible images and 11,909 infrared images with 395 identities. Similarly, the query set involves 3,803 infrared images with 96 identities. We use the Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP) as our standard evaluation metrics.

**Implementation details:** Our basic framework is largely modified from HCT [16] where MG-MRA/SG-MRA acts as a plug module applied in HCT. Similarity, this manner is applicable to other frameworks with GAP as we done in ablation study. We follow HCT to adopt ResNet50 [35] pretrained on ImageNet [36] as our backbone whose stride of the last convolutional block is changed from 2 to 1. Then each immediate feature is split into 6 stripes along height axis. We adopt the stochastic gradient descent (SGD) optimizer for optimization with momentum 0.9, learning rate 0.1. For the triplet losses, i.e., $L_{sem}$ and $L_{tri}$, we set the relaxation margin to 0.3. For the $PK$ sampling strategy, we set $P = 8, K = 4$ for the RegDB dataset, and $P = 6, K = 8$ for the SYSU-MM01 dataset respectively. Per-prototype number of part, instance and semantic level is set to $P = 6, I = 5$ and $S = 1$. The trade-off constants $\beta, \lambda_1$, and $\lambda_3$ are set to 0.05, 0.1, 0.1, 1 respectively. The model is trained on a single NVIDIA P100 GPU with pytorch.

#### 5.2 Comparison with State-of-the-art Methods

We evaluate our method with other state-of-the-art (SOTA) methods on RegDB and SYSU-MM01, including AlignGAN [7], CMSP, Xmodel [4], GLMC [42], MPANet [41] etc. As shown in Table 1, our method achieves the best performance across most evaluation metrics in visible2infrared and infrared2visible setting. Besides, our method uses no bells and whistles to outperform existing methods with a large
margin: 2.75%/6.76% Rank1/mAP improvement on visible2infrared and 2.10%/6.13% Rank1/mAP improvement on infrared2visible over previous best SOTA GLMC [42]. Compared with HCT, the basic framework we modify from, our method could further boost 3.44%/4.90% and 3.92%/5.73% improvement of 1.92%/0.7% and 5.28%/1.96% Rank1/mAP improvement on All search and Indoor search. Moreover, our method provides promising performance margin over HCT, finally obtaining 72.50%/82.02% Rank1 for All search and Indoor search respectively.

5.3 Ablation Study

SG-MRA & MG-MRA: As shown in Table 3, we plug our SG-MRA and MG-MRA into one GAN-based method (AlignGAN [7]), one baseline model (AGW [5]), one attention-based alignment model (DDAG [6]) and one SOTA model (HCT [16]) to verify the generalization ability of our memory regulation module. Compared with original methods, SG-MRA and MG-MRA could provide 2%~5% Rank1 improvement. Moreover, the semantic hierarchy among MG-MRA explicitly eases the semantic abstraction by memorizing part-level and instance-level structural pattern. Therefore, MG-MRA could further boost the performance of different models by meeting different demands of semantic maintaining.

Per-prototype number: As shown in Table 4, we adjust the per-prototype number of part, instance and semantic metric to evaluate the sensitivity of MG-MRA for metric descriptive ability. Even we increase the metric size ten times larger ($P \times I \times S$), the performance of our MG-MRA seems to change trivially (within 1% on Rank and mAP). Therefore, considering both SG-MRA and MG-MRA, we believe the hierarchical semantic summarization would make more sense than simply enhancing prototype descriptive ability.

Robustness without re-tuning hyper-parameters: The training of RGB-IR ReID model is sometimes tricky [43].

| Method     | Source     | visible2infrared | infrared2visible |
|------------|------------|------------------|------------------|
| AlignGAN   | ICCV-2019  | 57.9             | 53.6             |
| Xmodel     | AAAI-2020  | 62.21            | 60.18            |
| CMSP [37]  | IJCV-2020  | 65.07            | 64.5             |
| DDAG [6]   | ECCV-2020  | 69.34            | 63.46            |
| AGW [5]    | TPAMI-2021 | 70.05            | 66.37            |
| cm-SSFT [38]| CVPR-2020  | 72.3             | 71.0             |
| CICL [39]  | AAAI-2021  | 78.8             | 77.9             |
| NFS [40]   | CVPR-2021  | 80.54            | 77.95            |
| MPANet [41]| CVPR-2021  | 83.7             | 80.93            |
| HCT [16]   | TMM-2020   | 91.05            | 89.30            |
| GLMC [42]  | TNNLS-2021 | 91.84            | 91.12            |
| Ours       |            | 94.59            | 92.32            |

| Method     | All search | Indoor search |
|------------|------------|---------------|
| AlignGAN   |            | 42.4          |
| CMSP       |            | 43.56         |
| AGW        |            | 47.50         |
| Xmodel     |            | 49.92         |
| DDAG       |            | 54.75         |
| NFS        |            | 56.91         |
| CICL       |            | 57.52         |
| cm-SSFT    |            | 61.66         |
| HCT        |            | 61.68         |
| GLMC       |            | 64.37         |
| MPANet     |            | 70.58         |
| Ours       |            | 72.50         |

| Method     | default | SG-MRA | MG-MRA |
|------------|---------|--------|--------|
| AlignGAN   |         |        |        |
| DDAG       | 58.32   | 54.17  | 57.30  |
| AGW        | 68.78   | 61.02  | 65.98  |
| HCT        | 71.21   | 68.37  | 73.96  |
| Ours       |         |        |        |

| $P \times I \times S$ | R1 | R5 | R10 | R20 | mAP |
|-----------------------|----|----|-----|-----|-----|
| 6 5 1                 | 58.74 | 84.01 | 89.95 | 95.22 | 53.84 |
| 3 5 1                 | 58.45 | 84.01 | 90.10 | 95.20 | 53.75 |
| 3 3 1                 | 57.80 | 82.30 | 89.55 | 95.28 | 53.56 |
| 1 1 1                 | 58.37 | 83.77 | 89.54 | 94.89 | 53.65 |
| 6 5 3                 | 58.20 | 83.31 | 90.12 | 94.90 | 53.50 |
| 3 5 3                 | 58.77 | 83.95 | 90.20 | 95.12 | 53.80 |
but our MG-MRA is robust to provide consistent improvement for setting within reasonable range, which means MG-MRA is an user-friendly plug without re-tuning their well-posed hyper-parameters. For example, all experiments reported in the paper use the default setting of basic structures. Moreover, as shown in Figure 3, we evaluate the potential of MG-MRA to adjust different parameters, i.e., batch size and learning rate. The HCT baseline model is quite sensitive to hyper-parameters but our MG-MRA could still boost 2-3% accuracy in most cases.

Regulation visualization: We visualize the attention heat map of training, query and gallery images in Figure 4 and t-SNE distribution of query and gallery images in Figure 5. As shown in the upper and bottom parts of Figure 4, our MG-MRA (third row) could alleviate the overwhelmed attention on training images, and correspondingly, look on more discriminative regions for searching persons of interest. Besides, we find the original distribution of AGW in Figure 5 locates more randomly and compactly which is harmful for ranking, while using our MG-MRA could achieve spaced distribution as the intuition of triplet loss. We also discuss the role of MG-MRA and the suggestions for implementing it on other frameworks in Appendix.

6 Discussions

1. Why MG-MRA is a regulation module, different from other memory variants?

As we analyzed in theoretical properties, our MG-MRA is designed for open-set RGB-IR ReID to divert the attention of the training process from recognition cues to general structural pattern. But other variants, e.g., MemAE [26] is applied in unsupervised generation task, mainly record close-set pattern to help sub-stream task. Empirically, we notice that adding MG-MRA and corresponding losses slower the decrease of ID loss and the increase of training accuracy, but improve the evaluation accuracy. We also illustrate the regulation effect in Figures 4 and 5. These findings are consist with general regulation term.

2. What the difference between MG-MRA and other regulation methods?

Our MG-MRA is a kind of feature regulation without any additional annotation or multi-task cooperation. For common kernel regulation term, such as $L_2$ term, it is unable to assess latent variable directly, e.g., recognition cue $z$. Moreover, we reconsider methods using multi-task learning and annotation as external task regulation. This external cues would semi-overlap with recognition cues, so as to enhance the category-invariance.

For the PCB based methods [32], we believe this kind of hard attention acts similarly as regulation: hand-crafted stripe division forces the model to look specific area without concession. It is really powerful proved by its improvement, however, still can’t flexibly adjust features.

3. What the implementation limit for applying MG-MRA on other methods?

Our MG-MRA is not suitable for methods containing pre-Global Maxpooling (GMP) or its variants, as shown in Equations 1 and 2. We notice that this setting would result in optimization collapse, e.g., loss would not decrease and performance are within 1% accuracy. It is an interesting phenomenon that maxpooling commonly outperforms average pooling (In DGTL [44], solely replacing GAP with GMP
would bring extra 10% improvement on RegDB), but it seems to be unsuitable in this case.

Another interesting phenomenon is that $L_{sem}$ must be the classical triplet version. We exchange it with semi-hard or hard ones for adopting our regulation module in other frameworks.

7 Conclusion

In this paper, we delve to analyze the over-confidence of seen class features of mainstream paradigm, which leads to domain shift in RGB-IR ReID, in a probabilistic explainable way. Then we propose a multi-granularity memory regulation module (MG-MRA) to ease this tendency in training process. The proposed MG-MRA is effective and plug-and-play method for generalizer RGB-IR ReID with pre-hoc self-explanation. Experiment results on RegDB and SYSU-MM01 amply demonstrate that our MG-MRA outperforms previous state-of-the-arts with a large boost. Besides, the ablation study and discussion on MG-MRA illustrate the optimization essence of proposed method and provide suggestions for adopting our regulation module in other frameworks.

References

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

[2] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In ICCV, pages 5380–5389, 2017.

[3] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In AAAI, pages 12144–12151, 2020.

[4] Diangang Li, Xing Wei, Xiaopeng Hong, and Yi-hong Gong. Infrared-visible cross-modal person re-identification with an x modality. In AAAI, pages 4610–4617, 2020.

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

[6] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In ECCV, pages 229–247, 2020.

[7] 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 ICCV, pages 3623–3632, 2019.

[8] Zhedong Zheng, Liang Zheng, Yi Yang, and Fei Wu. Query attack via opposite-direction feature: Towards robust image retrieval. arXiv preprint arXiv:1809.02681, 2018.

[9] Jae Myung Kim, Junsuk Choe, Zeynep Akata, and Seong Joon Oh. Keep calm and improve visual feature attribution. arXiv preprint arXiv:2106.07861, 2021.

[10] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 22(2):199–210, 2010.

[11] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross-domain adaptation approach for person re-identification. In ICCV, pages 6112–6121, 2019.

[12] Fabian Dubourvieux, Romaric Audigier, Angelique Loesch, Samia Ainouz, and Stephane Canu. Unsupervised domain adaptation for person re-identification through source-guided pseudo-labeling. In ICPR, pages 4957–4964, 2021.

[13] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In ICCV, pages 9527–9536, 2019.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, pages 618–626, 2017.

[15] Aditya Chattopadhay, Anirban Sarkar, Prantik Batra, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In WACV, pages 839–847, 2018.

[16] 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.

[17] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In ICCV, pages 3741–3749, 2019.

[18] Kun Wei, Cheng Deng, and Xu Yang. Lifelong zero-shot learning. In IJCAI, pages 551–557, 2020.

[19] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In AAAI, pages 1966–1974, 2021.

[20] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In CVPR, pages 79–88, 2018.

[21] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In IJCAI, volume 1, page 2, 2018.

[22] Ye Yuan, Wuyang Chen, Yang Yang, and Zhangyang Wang. In defense of the triplet loss again: Learning robust person re-identification with fast approximated
triplet loss and label distillation. In *CVPR Workshops*, pages 354–355, 2020.

[23] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020.

[24] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[25] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019.

[26] Dong Gong, Lingqiao Liu, Vuong Le, Budhadiya Saha, Moussa Reda Mansour, Svetla Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, pages 1705–1714, 2019.

[27] Tong He, Dong Gong, Zhi Tian, and Chunhua Shen. Learning and memorizing representative prototypes for 3d point cloud semantic and instance segmentation. In *ECCV*, pages 564–580, 2020.

[28] Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. Attribute attention for semantic disambiguation in zero-shot learning. In *ICCV*, pages 6698–6707, 2019.

[29] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. Eanet: Enhancing alignment for cross-domain person re-identification. *arXiv preprint arXiv:1812.11369*, 2018.

[30] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3960–3969, 2017.

[31] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhiyan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019.

[32] 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 *ECCV*, pages 480–496, 2018.

[33] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[34] Marek Kuczma. *An introduction to the theory of functional equations and inequalities: Cauchy’s equation and Jensen’s inequality*. 2009.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.