

# Learning Memory-Augmented Unidirectional Metrics for Cross-modality Person Re-identification

Jialun Liu<sup>1,2\*</sup>, Yifan Sur<sup>2</sup>, Feng Zhu<sup>2</sup>, Hongbin Pei<sup>3</sup>, Yi Yang<sup>4</sup>, Wenhui Li<sup>1†</sup>  
<sup>1</sup> Jilin University, <sup>2</sup> Baidu Research<sup>3</sup>, Xi'an Jiaotong University<sup>4</sup>, Zhejiang University  
jialun18@mails.jlu.edu.cn      liwh@jlu.edu.cn

## Abstract

This paper tackles the cross-modality person re-identification (re-ID) problem by suppressing the modality discrepancy. In cross-modality re-ID, the query and gallery images are in different modalities. Given a training identity, the popular deep classification baseline shares the same proxy (i.e., a weight vector in the last classification layer) for two modalities. We find that it has considerable tolerance for the modality gap, because the shared proxy acts as an intermediate relay between two modalities. In response, we propose a Memory-Augmented Unidirectional Metric (MAUM) learning method consisting of two novel designs, i.e., unidirectional metrics, and memory-based augmentation. Specifically, MAUM first learns modality-specific proxies (MS-Proxies) independently under each modality. Afterward, MAUM uses the already-learned MS-Proxies as the static references for pulling close the features in the counterpart modality. These two unidirectional metrics (IR image to RGB proxy and RGB image to IR proxy) jointly alleviate the relay effect and benefit cross-modality association. The cross-modality association is further enhanced by storing the MS-Proxies into memory banks to increase the reference diversity. Importantly, we show that MAUM improves cross-modality re-ID under the modality-balanced setting and gains extra robustness against the modality-imbalance problem. Extensive experiments on SYSU-MM01 and RegDB datasets demonstrate the superiority of MAUM over the state-of-the-art. The code will be available.

Figure 1. Comparison between baseline and MAUM. We visualize the embedding space of baseline and MAUM with t-SNE [27], respectively. (a) In baseline, each identity has a modality-agnostic proxy for two modalities, which acts as a relay between IR and RGB features. The relay effect of baseline is revealed in the t-SNE visualization, where the modality gap between IR and RGB features is pretty large. (b) MAUM has two modality-specific proxies (MS-proxies, the orange solid dot for RGB and the blue solid dot for IR). Each MS-Proxy is fixed as a static reference for pulling close the features in the counterpart modality (dotted arrow). Further, MAUM stores historical MS-Proxies (void dot) into two memory banks, one for IR modality and one for RGB modality. Correspondingly, each identity has multiple IR and RGB proxies. The farthest MS-Proxies from the modality boundary become hard positive references and thus have a stronger “pulling close” effect (solid arrow). Consequently, as the visualization shows, MAUM suppresses the modality discrepancy.

## 1. Introduction

This paper considers cross-modality person re-identification (re-ID). Re-ID aims to retrieve images of the person-of-interest from the database. Real-world re-ID systems sometimes require recognizing the same person across daytime and night. To this end, they use two

different devices, i.e., the RGB camera at daytime and the Infra-Red (IR) camera at night. When the query and the gallery images are from different modalities, the significant modality discrepancy stands out as the most prominent challenge. In this paper, we try to improve cross-modality re-ID by addressing the modality discrepancy problem.

From the metric learning viewpoint, the keynote of re-ID is to learn an embedding space with both within-class com-

\* Work done during an internship at Baidu Research.

† Corresponding author.

pactness and between-class separability. A popular deep learning baseline [8, 17, 23, 23, 29, 41] for re-ID and face recognition task is based on deep classification learning. During training, it pulls all the features of the same identity toward a corresponding proxy, i.e., the weight vector in the classification layer.

When we apply this baseline to the cross-modality re-ID, we find that the modality discrepancy problem significantly hinders the within-class compactness, as illustrated in Fig. 1(a). In the baseline, all the instances of the same identity share a single proxy, regardless of the underlying modality. The modality-agnostic proxy strives to accommodate both the IR and RGB features and acts as an intermediate relay between them. Such relay effects result in considerable tolerance for the modality discrepancy. From the t-SNE [27] visualization in Fig. 1(a), we observe that there is an apparent modality discrepancy between the features of the two modalities. The features with different identities but same modality are even closer than those with same identity but different modality. For example, the between-class distance between ID-116 and ID-129 is smaller than the within-class distance of ID-116.

To suppress the modality discrepancy, we propose a Memory-Augmented Unidirectional Metric (MAUM) learning method. It is featured for two novel designs, 1) learning unidirectional metrics and 2) enhancing the unidirectional metrics with the memory bank.

First, we learn two unidirectional metrics (“IR to RGB” and “RGB to IR”) to alleviate the relay effect of the baseline. To this end, MAUM learns two modality-specific proxies (MS-Proxies) for each identity, as illustrated in Fig. 1(b). The RGB (IR) proxies only receive gradients from the RGB (IR) features and thus represent the dedicated modality. Afterward, we freeze them and use the RGB proxies as the static references for pulling IR features, and vice versa. These two unidirectional metrics promote the better cross-modality association.

Second, we further enhance these two unidirectional metrics through memory-based augmentation. MAUM stores the IR and RGB proxies into a respective memory bank after every iteration. Since the MS-proxies keep on all changing iteration by iteration, i.e., the “drift” phenomenon [30], each person has multiple diverse IR and RGB proxies in the memory bank, as illustrated in Fig. 1(b). Some historical MS-Proxies are farther away from the modality boundary (than the up-to-date MS-Proxies) and thus lay stronger “pulling close” effect on the counterpart-modality features. In a word, the memory bank enhances MAUM with hard positive references and consequentially promotes cross-modality association. We point out that memory-based learning in MAUM reveals a previously unknown yet important potential of the memory bank. Specifically, we employ the “drift” to enhance the references. In contrast,

the previous works [10, 15, 25, 30] consider the “drift” bringing negative impact and try to avoid it (as detailed in Section 2.2). As the visualization in Fig. 1(b) shows, the features with the same identity distribute compactly, which indicates that the modality discrepancy is suppressed. For example, the within-class embedding of ID-116 is significantly more compact than that in baseline (Fig. 1(a)).

In addition to the effectiveness of mitigating modality discrepancy, the proposed MAUM has a particular advantage under the modality imbalance scenario. In the training data, the IR images are usually scarcer than the RGB images because people have less movement at night, and the IR images are harder to annotate. In MAUM, both the unidirectional metrics and the memory-based augmentation are modality-specific. The augmentation on IR proxies is independent of that on RGB proxies and vice versa. Therefore, MAUM may re-balance the enhancement for the IR modalities. By re-balancing the augmentation, the MAUM compensates for the shortage of IR images and gains strong robustness against the modality imbalance.

Our main contributions are summarized as follows:

We propose a novel memory-augmented unidirectional metric learning method for cross-modality re-ID. It learns explicit cross-modality metrics in two uni-directions and further enhances them with memory-based augmentation.

We consider the modality imbalance, which is an essential realistic problem in cross-modality re-ID. By adjusting the modality-specific augmentation, MAUM shows strong robustness against modality imbalance.

We comprehensively evaluate our method under modality-balance and modality-imbalance scenarios. Experimental results confirm that MAUM improves cross-modality re-ID under both settings, surpassing the state-of-the-art significantly.

## 2. Related Work

### 2.1. Cross-modality metric learning

Cross-modality problem has been first studied in heterogeneous face recognition [12, 14, 21]. These early works use modality-agnostic proxies to enforce within-class compactness, i.e., the baseline in Section 1). [33] first introduce the cross-modality problem in person re-identification, and gradually draws the attention of re-ID community [18, 32, 35–40]. Among these works, we note the closest one to ours is [36]. Similar to our method, they also employ modality-specific classification layers. However, there are significant differences. [36] uses the ensemble of the modality-specific classifiers to generate an enhanced teacher model for collaborative ensemble learning. In contrast, MAUM uses modality-specific classifiers to learn the modality-specific proxies. Those proxies are fixed after convergence and used for learning unidirectional metrics.

Figure 2. The framework of the proposed MAUM. MAUM adopts ResNet50 as the backbone and shares the parameters from “conv2” to “conv4” for two modalities. The RGB and IR images are mapped into the deeply-embedded space to get the RGB and IR features, respectively. MAUM has three classifiers, the modality-agnostic classifier, the RGB, and the IR classifier, which are implemented with three Fully-Connected (FC) layers. Unlike the modality-agnostic classifier, the RGB (IR) classifier only accepts the RGB (IR) features so that the learned MS-Proxies are highly specific and alleviate the relay effect. Given the already-learned MS-Proxies, MAUM stores them into two corresponding memory banks after every iteration. The memory banks have three critical functions: bidirectional metric learning, augmentation through drift, and resisting modality-imbalance.

## 2.2. Memory-based learning

Memory bank has been extensively explored in supervised [30], semi-supervised [15,25] and unsupervised learning [10]. In semi-supervised learning, [15, 25] uses the memory bank to get the temporal ensemble of the historical predictions. It enforces consistency between the up-to-date prediction of the unlabeled sample and the temporal ensemble. Two important works in unsupervised learning (i.e., MOCO [10]) and supervised metric learning (i.e., XBM [30]) share a similar motivation for using the memory bank. Specifically, MOCO increases the quantity of the stored keys for better contrastive learning. XBM enhances the hard mining effect by storing historical features. To our understanding, they both benefit from the memory bank by the increase of negative features.

Against this background of memory-based learning, we point out that the novelty of MAUM lies in a new mechanism for cross-modality metric learning. In MAUM, the benefit of the memory bank is not due to the temporal consistency (as in the semi-supervised learning) or more negative samples (as in MOCO and XBM). The benefit in MAUM originates from the model drift, which helps MAUM to get hard positive references and promotes cross-modality association. This insight reveals a previously unknown yet important potential of the memory bank. Moreover, MAUM stores the proxies into the memory bank, which can be viewed as a novel augmentation for the metric learning task. In contrast, the previous works store only the feature vectors.

## 2.3. Imbalanced data learning

Data imbalance is an important challenge in deep learning. Most previous researches [4, 7, 22, 31] pay attention to the class imbalance problem and introduce two main approaches, i.e., re-sampling [7, 13, 31] and re-weighting [2-4, 22, 42]. Re-sampling over-sample the minority classes (with few samples) or under-sample the frequent class (with many samples) in training, aiming to balance the head and tail data in every iteration. Re-weighting assigns adaptive weights for different classes or even different samples in the loss function.

This paper notices a unique data imbalance problem in cross-modality tasks, i.e., the modality imbalance. It refers to the situation that one modality contains more samples than the other modality. In MAUM, the modality-specific augmentation is naturally disentangled and allows independent augmentation for a specified modality. It endows MAUM with strong robustness against the modality imbalance. As an essential contribution of this work, we hope it will inspire the community to pay attention to the modality imbalance problem.

## 3. Proposed Method

### 3.1. Learning MAUM

The framework of MAUM is illustrated in Fig 2. MAUM adopts ResNet50 [11] as the backbone network and accepts RGB and IR images as its input. MAUM splits the first convolutional block into two independent branches to

accommodate modality-specific low-level feature patterns, one for RGB and the other for IR modality. Two modalities for computation efficiency share all the following convolutional blocks. Given the convolutional feature maps, MAUM uses a global average pooling (GAP) to generate a deep embedding for each input image. Based on thising commonly-adopted backbone setting [37, 40], the proposed MAUM lays emphasis on its novel memory-augmented unidirectional metric learning approach. Specifically, MAUM learns two sets of modality-specific proxies (IR and RGB proxies) and stores them into the MS-Proxy memory banks. Given the MS-proxies in the memory banks and the features in the mini-batch, MAUM combines them to learn the unidirectional metrics. We elaborate the process of “learning modality-specific proxies” “constructing memory bank” and “learning unidirectional metrics” as follows.

### 3.1.1 Learning Modality-Specific Proxies

MAUM first supplements the modality-agnostic ID classifier in the baseline with two modality-specific (IR and RGB) ID classifiers to facilitate the unidirectional metrics. All these three ID classifiers are implemented with a respective Fully-Connected (FC) layer. The difference between them is that the modality-agnostic ID classifier accepts both RGB and IR features, while the IR (RGB) ID classifiers only accept IR (RGB) features for training. Correspondingly, the IR and RGB ID classifiers learn two sets of modality-specific proxies. Given the RGB features, the RGB classifier employs the widely-used cross-entropy loss as the optimization objective, which is formulated as:

$$L_{RGB} = -\frac{1}{N^R} \sum_{i=1}^{N^R} \log P_C = -\frac{1}{N^R} \sum_{i=1}^{N^R} \log \frac{\exp(w_{y_i}^R x_i^R)}{\sum_k \exp(w_k^R x_i^R)}; \quad (1)$$

where the superscript “R” indicates the RGB modality,  $N^R$  is the RGB instances number in current mini-batch,  $y_i$  is the ground-truth identity of  $x_i$ . We use the weight vector  $w_{y_i}$  as the proxy of  $y_i$  in RGB modality.

The loss function for the IR classifier and the modality-agnostic classifier is denoted as  $L_{IR}$  and  $L_{com}$ , respectively. The formulations are similar to Eq. 1 and omitted here.

In each modality-specific classifier, the MS-Proxies no longer struggle to accommodate the two opposite modalities and are thus highly representative for their dedicated modality.

### 3.1.2 Constructing Memory Bank

After the modality-specific proxies are fully trained, MAUM collects them into two corresponding memory banks. Specifically, we use a queue strategy for updating the memory bank. We set the memory bank sizes for RGB modality and IR modality as  $S_{RGB}$  and  $S_{IR}$ , respectively.

After the memory bank reaches its size limitation, we enqueue the newest proxies and dequeue the oldest ones. The memory banks have three critical functions for MAUM. First, they freeze the already-learned MS-Proxies and use them as static references for unidirectional metric learning. Second, they employ the model drift phenomenon [30] by accumulating the historical MS-proxies to increase the diversity of these MS-Proxies. Third, they help MAUM to gain extra robustness against modality-imbalance because the memory-based augmentation is modality-specific, which can be independently adjusted to re-balance the enhancement for the IR and RGB modalities (Section 3.2).

### 3.1.3 Learning Unidirectional Metrics

We freeze the MS-proxies in the memory bank. Then we use them as the static references for pulling close the features in the counterpart modality.

We note that although there is only a single IR and RGB proxy for each identity in the modality-specific classifier, storing historical MS-Proxies into the memory bank gradually increases their quantity. Consequently, in the RGB (IR) memory bank, there are multiple RGB (IR) proxies for every single identity, providing multiple positive references for a single IR (RGB) feature. Specifically, given a single RGB feature  $x^R$ , we assume there are  $N$  positive references  $u_1^I; u_2^I; \dots; u_N^I$  and  $M$  negative references  $v_1^I; v_2^I; \dots; v_M^I$  (the superscript “I” indicates the IR modality) in IR memory bank. Inspired by Circle Loss [23], we define the loss function for learning the unidirectional metric from RGB image to IR proxy as:

$$L_{R \rightarrow I} = \log 4 + \frac{2}{N} \sum_{j=1}^N \sum_{i=1}^M \exp(-\gamma (v_j^I x^R - u_i^I x^R + \Delta)); \quad (2)$$

where the feature and the proxies are normalized,  $\gamma$  is the scale factor, and  $\Delta$  is the margin parameter. We formulate  $L_{R \rightarrow I}$  with only a single feature  $x^R$  for simplicity. In practice, the loss function is averaged over all the RGB features in the current mini-batch.

The  $L_{I \rightarrow R}$  for learning IR-feature-to-RGB-proxy metric is symmetric to Eq. 2 and thus omitted here.

### 3.1.4 Optimization

We combine a modality-shared loss ( $L_{com}$ ), two modality-specific losses ( $L_{RGB}$  and  $L_{IR}$ ) and two unidirectional metric losses ( $L_{R \rightarrow I}$  and  $L_{I \rightarrow R}$ ) to get the overall loss function  $L_{Total}$  as follows:

$$L_{Total} = L_{com} + L_{RGB} + L_{IR} + \lambda (L_{R \rightarrow I} + L_{I \rightarrow R}); \quad (3)$$

where  $\lambda$  is the hyper-parameter which balances the contributions of the unidirectional metric loss.



### 3.1.5 MAUM with Part Features.

We note that part feature generally improves the visible re-ID [24], as well as the cross-modality re-ID [18, 39]. To validate that MAUM is compatible to part feature, we introduce a part-feature-based variant, MAUM<sup>P</sup>. According to a simple part feature baseline [24], MAUM evenly divides the last convolutional feature map into six-part features. During training, each part is supervised with a respective  $L_{Total}$ . During testing, all the six-part features are concatenated to form the final representation.

### 3.2. MAUM under Modality Imbalance Scenario

In cross-modality re-ID, the IR images are usually more scarce than the RGB images, yielding the modality-imbalance problem. It is because people usually have less movement at night time, and the IR images are inherently harder to annotate. When the modality-imbalance reaches an extreme, some identities may have only a single modality (i.e., RGB). We formally define these two cases as follows:

**Modality-imbalance scenario.** Each identity has two modalities. The IR images are fewer than the RGB images.

**Modality-fragmentary scenario.** Some identities have only a single modality (i.e., RGB images), and the other identities have two modalities.

So far as we know, MAUM is the first work to consider the modality-imbalance problem in cross-modality re-ID. Experiments show that this problem significantly deteriorates the re-ID accuracy. In MAUM, since the augmentation is based on two modality-specific memory banks, the ratio between them can be flexibly adjusted to compensate for the shortage of IR images. Consequently, MAUM could provide strong resistance against the modality-imbalance problem. Specifically, the IR proxy memory size is  $\frac{M_{IR}}{M_{RGB}}$  of the RGB proxy memory size  $M_{RGB}$  and  $M_{IR}$  are the amount of RGB images and IR images, respectively.

### 3.3. Mechanism Analysis

In this section, we analyze the mechanism of memory-based augmentation in MAUM. We show that the accumulated proxy drift in the memory bank is the reason for the enhancement of the unidirectional metric learning.

When we observe the proxy of the same identity at two different training iterations, these two observations naturally differ from each other. For quantitative analysis, we define the difference between two observations of the same proxy as the proxy drift, which is formulated by:

$$D(p; t, 4t) = \|p(t) - p(t + 4t)\|_2^2 \quad (4)$$

where  $p$  is the proxy under observation,  $t$  is the current iteration index, and  $4t$  is the sampling interval. Such definition is similar to [30], except that the object under observation is different. In [30], the drift is based on the features, while

Figure 3. (a) Larger interval brings larger drift. (b) The t-SNE [27] visualization of the proxy distribution in the memory bank. The drift increases the diversity of the historical proxies. The proxies far away from the modality boundary become hard positives for pulling close the features in the counterpart modality.

in MAUM, the drift is based on the weight vectors (i.e., the proxies).

Fig. 3 (a) visualizes the drift under different sampling intervals. It is observed that a larger interval brings a larger drift. Therefore, when storing the proxies into the memory bank, MAUM favors a relatively large sampling interval ( $4t = 10$  in our implementation) to promote diversity among historical proxies.

Fig. 3 (b) visualizes the distribution of the proxies in the memory bank with t-SNE [27]. Due to the proxy drift, the historical proxies scatter around the up-to-date proxy (of the same identity). Some historical proxies stay farther away from the modality boundary, which becomes hard positives for pulling close the feature in the counterpart modality. They facilitate the stronger cross-modality association and consequentially improve cross-modality re-ID.

**Discussion.** Recently, XBM [30] also noticed the drift phenomenon when using the memory bank to enhance metric learning. However, it considers the drift as a negative side effect accompanying the memory bank. Therefore, it begins the memory-based learning after the drift decays to a small range. Similarly, in MOCO [10], it applies exponential moving average operation on the deep model to smooth the drift of historical keys. In contrast, MAUM is fundamentally different from them. In MAUM, the drift is beneficial for augmentation. It uses the drift to increase the diversity of the historical proxies, which helps to learn the robust cross-modality association. This finding is contrary to the previous research and inspires a new understanding of the drift phenomenon and memory-based learning.

## 4. Experiments

### 4.1. Primary settings

**Datasets.** We evaluate our method on two public cross-modality re-ID datasets: SYSU-MM01 [33] and RegDB [19]. SYSU-MM01 consists of 91 identities from four RGB cameras and two IR cameras in indoor and out-

Methods		SYSU-MM01				RegDB			
		All-Search		Indoor-Search		Visible to Infrared		Infrared to Visible	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
G	MAC [35]	33.26	36.22	36.43	37.03	36.43	37.03	36.20	36.63
	Hi-CMD [6]	34.49	35.94	-	-	70.93	66.04	-	-
	MSR [9]	37.35	38.11	39.64	50.88	48.43	48.67	-	-
	AlignGAN [28]	42.40	40.70	45.90	54.30	57.90	53.60	56.30	53.40
	Xmodel [16]	49.92	50.73	-	-	62.21	60.28	-	-
	LZW [1]	45.00	45.94	49.66	59.81	-	-	-	-
	MACE [36]	51.64	50.11	57.35	64.79	72.27	69.09	72.12	68.57
	CMAAlign [20]	55.41	54.14	58.46	66.33	74.17	67.64	72.43	65.46
	FMI [26]	60.02	58.80	66.05	72.98	73.20	71.60	71.80	70.10
	Baseline <sup>G</sup>	49.45	47.21	55.23	62.86	64.22	58.21	61.18	59.76
P	MAUM <sup>G</sup>	61.59	59.96	67.07	73.58	83.39	78.75	81.07	78.89
	DDAG [39]	54.75	53.02	61.02	67.98	69.34	63.46	68.06	61.80
	cm-SSFT [18]	61.60	63.20	70.50	72.60	72.30	72.90	71.00	71.70
	NFS [5]	56.91	55.45	62.79	69.79	80.54	72.10	77.95	69.79
	MPANet [34]	70.58	68.24	76.74	80.95	82.80	80.70	83.70	80.90
	Baseline <sup>P</sup>	55.57	53.96	55.61	56.31	69.59	58.93	67.33	61.32
	MAUM <sup>P</sup>	71.68	68.79	76.97	81.94	87.87	85.09	86.95	84.34

Table 1. Comparison with the state-of-the-art methods on SYSU-MM01 and RegDB, respectively. Rank-1 and mAP are reported. For fair comparison, we divide the compared methods into two groups: “G” for the global-feature-based methods and “P” for the part-feature-based methods.

door environments. There are 2,258 RGB images and 11,909 IR images of 395 identities in the training set. The query set contains 3,803 IR images, while the gallery set contains 301 RGB images. Following [33], we employ two test protocols, i.e., all-search and indoor-search. RegDB is collected by dual-camera systems, including one visible and one infrared camera. There are 412 identities, with 206 identities for training and 206 identities for testing. Each identity contains 10 RGB and 10 IR images. There are also two evaluation modes. One is thermal to visible to search RGB images from an IR image. The other one is visible to thermal to search IR images from an RGB image.

**Evaluation metrics.** All experiments follow the standard evaluation protocol, i.e., the Cumulative Matching Characteristic (CMC) and mean average precision (mAP). All the reported results are the average of 10 trials.

**Implementation details.** For fair comparison, we use ResNet50 [11] pre-trained on ImageNet as the backbone model. The input images are resized to 256 × 144 × 3, for both RGB and IR images. MAUM with global feature (MAUM<sup>G</sup>) uses a 2048-d vector for feature representation, while MAUM with part features (MAUM<sup>P</sup>) uses 3072-d (512 × 6, 6 is the part number) vector for feature representation. The training batch size is set to 64, comprised of 8 identities with 4 RGB images and 4 infrared images for each identity. The scale factor and the margin parameter are set to 2 and 0.2, respectively. In Eq. 3 is set to 1.

In SYSU-MM01, the memory bank size is set to 3000 and 1500 for RGB and IR modality, respectively. In RegDB, the memory bank size of RGB and IR modality is set to 1500.

#### 4.2. Effectiveness of MAUM

We evaluate the effectiveness of MAUM with comparisons against the baseline and the state-of-the-art methods. For fair comparison, we divide the compared methods into two groups, i.e., the global-feature-based methods and the part-feature-based methods. Table 1 summarizes the results on RegDB and SYSU-MM01, from which we draw two observations.

First, comparing MAUM against Baseline, we observe that MAUM significantly improves the baseline. Specifically, 1) comparing MAUM<sup>G</sup> against Baseline<sup>G</sup>, MAUM<sup>G</sup> surpasses the Baseline<sup>G</sup> by +12:14% Rank-1 accuracy on SYSU-MM01 (all-search), and by +19:17% Rank-1 accuracy on RegDB (visible to infrared). 2) Comparing MAUM<sup>P</sup> against Baseline<sup>P</sup>, MAUM<sup>P</sup> is higher than the Baseline<sup>P</sup> by +16:11% Rank-1 accuracy on SYSU-MM01 (all-search), and by +18:28% Rank-1 accuracy on RegDB (visible to infrared).

Second, MAUM achieves competitive performance, under both global feature and part feature settings. Specifically, using global feature, MAUM outperforms all the other global-feature-based methods. It surpasses the strongest competitor (FMI) by +1:57% and +10:19%

Rank-1 accuracy on SYSU-MM01 (all-search) and RegDB (visible to infrared), respectively. Using part features, MAUM<sup>P</sup> surpasses the strongest competitor (MPANet) by +1:10% and +5:07% Rank-1 accuracy, on SYSU-MM01 (all-search) and RegDB (visible to infrared), respectively. In this paper, we report the new state-of-the-art. Specifically, based on the global feature, MAUM achieves 61:59% and 83:39% Rank-1 accuracy on SYSU-MM01 and RegDB, respectively. Based on the part feature, MAUM achieves 71:68% and 87:87% Rank-1 accuracy on SYSU-MM01 and RegDB, respectively.

#### 4.3. Ablation Study

Method	UM	MA	All Search	
			Rank-1	mAP
Baseline	%	%	49.45	47.21
MAUM	!	%	55.97	53.42
MAUM	!	!	61.59	59.96

Table 2. Ablation study on SYSU-MM01 (all search). UM: unidirectional metrics, MA: memory-based augmentation.

We investigate the two key components, unidirectional metrics (UM) and memory-based augmentation (MA) through ablation. The experimental results are summarized in Table 2, from which we draw two observations.

First, comparing “MAUM” (i.e., MAUM without memory-based augmentation) with “Baseline”, we observe that using only the UM component already brings 6:52% Rank-1 accuracy and 6:21% mAP improvement. It validates that UM effectively suppress the modality discrepancy and improve the cross-modality recognition.

Second, comparing “MAUM” with “MAUM”, we find that adding the memory-based augmentation further improves +5:62% Rank-1 accuracy and 6:54% mAP. It indicates that memory-based augmentation effectively reinforces the unidirectional metric learning.

Combining these two observations, we conclude that both the unidirectional metrics and the memory-based augmentation components are critical for MAUM.

#### 4.4. Modality Imbalance Scenario

We investigate MAUM under modality-imbalance scenario. For comprehensive investigation, we synthesize several different imbalance settings based on the original SYSU-MM01 dataset. Specifically, we set the number of IR images per identity  $N$  from 1 to 15, resulting in various imbalance ratios between RGB and IR modality. We compare MAUM with Baseline, MAUM (i.e., MAUM without memory-based augmentation) and FMI [26] (the strongest competitor). The results are summarized in Table 3, from which we have the following observations.

N	Acc.	Method			
		Baseline	FMI	MAUM	MAUM
Full	R-1	49.5	60.0	56.0	61.6
	mAP	47.2	58.8	53.4	59.0
15	R-1	36.3 (-26.7%)	44.7 (-25.5%)	40.3 (-28.7%)	47.3 (-23.2%)
	mAP	35.5 (-24.8%)	45.6 (-22.4%)	38.8 (-27.3%)	46.9 (-20.5%)
10	R-1	32.1 (-35.2%)	40.6 (-32.3%)	37.3 (-33.4%)	44.7 (-27.4%)
	mAP	32.8 (-30.5%)	41.6 (-29.2%)	37.1 (-30.5%)	43.2 (-26.8%)
5	R-1	29.3 (-40.8%)	37.4 (-37.7%)	34.1 (-39.1%)	40.2 (-34.7%)
	mAP	30.2 (-36.0%)	39.5 (-32.8%)	33.7 (-36.9%)	43.2 (-26.8%)
1	R-1	13.2 (-73.3%)	23.6 (-60.7%)	15.2 (-72.9%)	25.4 (-58.8%)
	mAP	16.2 (-65.7%)	24.9 (-57.7%)	20.5 (-61.6%)	27.7 (-53.1%)

Table 3. Evaluation on SYSU-MM01 under modality-imbalance scenario. We set the number of IR images from 1 to 15, resulting in various imbalance ratios between RGB and IR modality, the number of IR images per identity.

First, as  $N$  decreases, the re-ID accuracy of all methods dramatically drop. For example, the mAP accuracy of the baseline dramatically drops from 47:2% (all-search) to 35:5% ( $N = 15$ ), 32:8% ( $N = 10$ ), 30:2% ( $N = 5$ ) and 16:2% ( $N = 1$ ). It indicates that the modality-imbalance problem significantly challenges cross-modality re-ID.

Second, comparing the accuracy decrease ratio of Baseline and MAUM, MAUM is close to Baseline. It indicates that only unidirectional metrics also undergoes the challenge of modality-imbalance problem.

Third, compared with MAUM, MAUM presents higher robustness against the modality-imbalance problem. For example, when  $N = 5$ , the mAP accuracy decrease ratios of MAUM is 36:9%. In contrast, the mAP accuracy decrease ratio of MAUM is 26:8%. This indicates that the higher robustness of MAUM benefits from the independent modality-specific memory augmentation.

Moreover, compared with the SOTA method FMI [26], MAUM outperforms FMI under all the imbalanced settings and gains higher robustness against modality-imbalance.

#### 4.5. Modality Fragmentary Scenario

We investigate MAUM under the modality-fragmentary scenario, i.e., some training identities have only a single (RGB) modality. Based on the original SYSU-MM01, we synthesize several modality-fragmentary datasets by removing some identities' IR images. Specifically, there are only 50, 100 and 200 identities with IR images. For the RGB modality, we maintain the number of identities 395. The results are summarized in Table 4, from which we draw the following observations.

First, when there are 395 RGB identities and the number of IR identities gradually decreases from 395 to 200, 100 and 50, all methods' re-ID accuracy dramatically drops. It indicates that the modality fragmentary problem hinders cross-modality re-ID.

RGBs / IRs	Acc.	Method			
		Baseline	FMI	MAUM	MAUM
395 / 395	R-1	49.5	60.0	56.0	61.6
	mAP	47.2	58.8	53.4	59.0
395 / 200	R-1	42.8 (-13.5%)	53.2 (-11.3%)	47.7 (-14.8%)	56.2 (-8.8%)
	mAP	42.7 (-9.5%)	53.3 (-9.3%)	47.3 (-11.4%)	54.6 (-7.5%)
395 / 100	R-1	25.9 (-47.7%)	35.0 (-41.6%)	31.2 (-44.3%)	38.5 (-37.5%)
	mAP	26.3 (-44.3%)	36.0 (-38.8%)	30.6 (-42.7%)	39.2 (-33.6%)
395 / 50	R-1	18.2 (-63.2%)	24.2 (-59.7%)	21.1 (-62.3%)	28.8 (-53.2%)
	mAP	20.5 (-56.6%)	29.3 (-50.2%)	24.2 (-54.7%)	36.1 (-38.8%)

Table 4. Evaluation on SYSU-MM01 under modality-fragmentary scenario. RGBs / IRs denotes the quantity of RGB / IR identities.

Second, comparing the accuracy decrease ratio of Baseline, MAUM and MAUM, MAUM is close to Baseline. While, MAUM presents higher robustness against the modality-fragmentary problem. This further indicates that the higher robustness of MAUM benefits from the independent modality-specific memory augmentation. This observation is consistent with that of modality-imbalance scenario.

Third, under all modality-fragmentary setting, MAUM outperforms FMI [26] and gains the higher robustness.

## 5. Analysis of Hyper-parameters

Figure 4. The impact of memory size and sampling interval in Eq. 4). (a) As the memory size increases, the accuracy increases to the maximum (when memory is 3000) and then decreases. (b) As the sampling interval increases, the accuracy undergoes an increase and decrease, as well.

The drift-based augmentation of the memory bank is a key component of MAUM. According to Section 3.3, the augmentation strength is controlled by two hyper-parameters, i.e., the memory bank and the sampling interval. Fig. 4 experimentally analyze the impact of these two hyper-parameters on SYSU-MM01. In Fig. 4 (a), we fix the sampling interval (i.e., 10), and then increase the RGB memory size from 0 to 5000 (we set the IR memory size to be the half of RGB memory). In Fig. 4 (b), we fix the RGB memory size to 3000, and then adjust the sampling interval from 1 to 20. We make three observations as follows.

First, in Fig. 4 (a), when the memory size is 0, which means there is NO unidirectional learning, the achieved accuracy is even lower than the baseline. It indicates that the unidirectional metric learning is critical for MAUM. Removing it dramatically corrupts MAUM.

Second, in Fig. 4 (a), as the memory size increases, the accuracy of MAUM gradually increases to the maximum (when memory size is 3000) and then decreases. It indicates the memory bank has two-fold impacts on MAUM. On the one-hand, approximately increasing the quantity of stored proxies benefits the memory bank with higher diversity. On the other hand, over large memory might include some deeply-outdated historical proxies, which is not quite comparable with the up-to-date features [10, 30]. Based on the above observation, we select 3000 as the default memory size in all experiments.

Third, in Fig. 4 (b), we observe a similar “increase and decrease” phenomenon over the increase of sampling interval. It is reasonable because the sampling interval has a similar impact as the memory size. While increasing the interval enlarges the diversity of neighbouring proxies (which is beneficial), it also risks over-large divergence between the earliest proxies and the up-to-date proxies. In this paper, we recommend using 10 as the optimized sampling interval.

## 6. Conclusions

This paper proposes the Memory-augmented Unidirectional Metric (MAUM) learning method for cross-modality re-ID. MAUM has two advantages. First, instead of using a modality-agnostic proxy as the intermediate relay between two modalities, MAUM enforces explicit cross-modality association with two unidirectional metrics. Second, by exploring a novel potential of the model drift phenomenon, MAUM further enhances the cross-association through memory-based augmentation. Equipped with the two advantages, MAUM significantly suppresses the modality discrepancy and improves cross-modality re-ID. As another contribution, we bring the modality-imbalance problem into the cross-modality re-ID community, and demonstrate that MAUM presents high robustness and superiority on this problem.

**Limitation.** In MAUM, we employ two modality-specific memory banks to store the MS-Proxies. Although these proxies have no gradient, storing and employing them still need a few memory and computation cost. When the training set is large-scale, such as an industrial dataset, the memory and computation cost can not be ignored. How to optimize the memory and computation cost will be explored in our future work.

**Acknowledgments.** This work was supported in part by Postdoctoral Innovative Talent Support Program of China (BX2021240).



## References

- [1] Emrah Basaran, Muhittin Ökmen, and Mustafa E Kamasak. An efficient framework for visible-infrared cross modality person re-identification. *Signal Processing: Image Communication*, 87:115933, 2020. 6
- [2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106:249–259, 2018. 3
- [3] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? *International Conference on Machine Learning*, pages 872–881, 2019. 3
- [4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arachis, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413* 2019. 3
- [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*, pages 587–597, 2021. 6
- [6] 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. 6
- [7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 3
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2
- [9] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing* 29:579–590, 2019. 6
- [10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 5, 8
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6
- [12] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2
- [13] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, pages 5375–5384, 2016. 3
- [14] Jing Huo, Yang Gao, Yinghuan Shi, Wanqi Yang, and Hujun Yin. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE transactions on cybernetics* 48(6):1814–1826, 2017. 2
- [15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* 2016. 2, 3
- [16] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020. 6
- [17] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019. 2
- [18] 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. 2, 5, 6
- [19] 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. 5
- [20] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. *arXiv preprint arXiv:2108.07422* 2021. 6
- [21] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. Graphical representation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence* 39(2):301–312, 2016. 2
- [22] Li Shen, Zhouchen Lin, and Qingming Huang. Relay back-propagation for effective learning of deep convolutional neural networks. In *ECCV*, pages 467–482, 2016. 3
- [23] Yifan Sun, Changmao Cheng, Yuhang Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4
- [24] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with re-net part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 5
- [25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780* 2017. 2, 3
- [26] 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*, pages 1522–1531, 2021. 6, 7, 8
- [27] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research* 15(1):3221–3245, 2014. 1, 2, 5

- [28] 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. 6
- [29] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pages 5265–5274, 2018. 2
- [30] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 6388–6397, 2020. 2, 3, 4, 5, 8
- [31] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS* pages 7029–7039, 2017. 3
- [32] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 618–626, 2019. 2
- [33] 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. 2, 5, 6
- [34] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 4330–4339, 2021. 6
- [35] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia* pages 347–355, 2019. 2, 6
- [36] 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. 2, 6
- [37] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 4
- [38] 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, 2019. 2
- [39] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. *arXiv preprint arXiv:2007.09314* 2020. 2, 5, 6
- [40] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018. 2, 4
- [41] Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision* pages 1116–1124, 2015. 2
- [42] Q Zhong, C Li, Y Zhang, H Sun, S Yang, D Xie, and S Pu. Towards good practices for recognition & detection. In *CVPR workshops* volume 1, 2016. 3