Asymmetric Mutual Learning for Unsupervised Transferable Visible-Infrared Re-Identification

Ancong Wu, Chengzhi Lin, Wei-Shi Zheng

***Abstract*—Visible-infrared person re-identification (Re-ID) plays a crucial role in matching people across camera views in the darkness and normal lighting. To reduce annotation cost, it is advantageous to learn Re-ID model from unlabeled visible-infrared image pairs. However, large modality gap makes it difficult to discover the underlying cross-modality sample relations. Compared with cross-modality sample pairs in the target domain, it is easier to obtain more single-modality visible image samples from other domains. In this work, we study unsupervised transfer learning to extract modality-shared knowledge from auxiliary unlabeled visible images in a source domain and leverage this knowledge to learn cross-modality matching in the unlabeled target domain. Our framework consists of two stages: RGB-gray asymmetric mutual learning and unsupervised cross-modality self-training. In the first stage, to extract visible-infrared shared information from auxiliary unlabeled visible images, we regard RGB images and grayscale fake infrared images transformed from RGB images as two views to learn view-shared information and simultaneously preserve RGB-specific information. Based on information theoretic analysis, we learn an RGB-gray feature extractor and further introduce an auxiliary gray feature extractor to quantify RGB-gray shared knowledge. This knowledge is then transferred to the RGB-gray feature extractor without eliminating RGB-specific information. We call this process Cross-Modality Asymmetric Mutual Learning (CMAM). In the second stage, for unsupervised cross-modality self-training in the target domain, we fuse the complementary knowledge in two models by mutual learning and employ bipartite cross-modality pseudo labeling to alleviate modality gap. For a more extensive evaluation, we collected a new public multi-modality dataset, SYSU-MM02, constructed from untrimmed videos. Our method achieves the state-of-the-art performance on three benchmark datasets. Project page: https://www.isee-ai.cn/project/sysumm02.html.**

***Index Terms*—Person re-identification, cross-modality matching, unsupervised domain adaptation.**

I. Introduction

In recent years, with increasing demand of intelligent analysis of video surveillance, person re-identification (Re-ID) for matching pedestrians across non-overlapping camera views has attracted much attention. Based on deep learning models, supervised Re-ID methods [1]–[4] and unsupervised Re-ID methods [5]–[15] have achieved remarkable performance for visible image matching. However, current Re-ID methods are still unsatisfactory for visible-infrared cross-modality person re-identification [16] in the scenarios of pedestrian matching across cameras in the darkness and normal lighting, since the modality gap between visible images and infrared images caused by large visual differences is challenging.

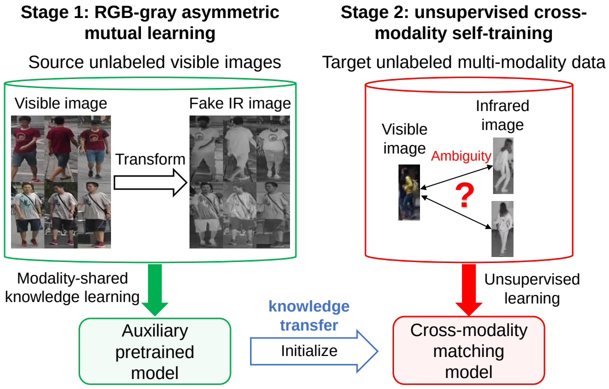

Stage 1: RGB-gray asymmetric mutual learning
Source unlabeled visible images
Visible image
Fake IR image
Transform
Modality-shared knowledge learning
Auxiliary pretrained model
knowledge transfer
Initialize
Stage 2: unsupervised cross-modality self-training
Target unlabeled multi-modality data
Visible image
Infrared image
Ambiguity
?
Unsupervised learning
Cross-modality matching model

Fig. 1. Unsupervised transferable visible-infrared person Re-ID framework. We learn modality-shared prior knowledge from source auxiliary unlabeled visible images that can be obtained more easily than visible-infrared sample pairs, and then transfer the knowledge to improve unsupervised cross-modality Re-ID in the target domain.

Most existing visible-infrared cross-modality person re-identification methods focus on supervised learning [17]–[26]. However, it is infeasible to annotate identities exhaustively for each unseen scene in large surveillance systems. It is significant to study unsupervised cross-modality Re-ID [27]–[32]. To discover the underlying visible-infrared sample relations under high visual ambiguity, modality-shared prior knowledge for bridging modality gap is required.

In this work, in order to learn such prior knowledge without annotating visible-infrared image pairs, we explore unsupervised transfer learning to extract modality-shared prior knowledge from auxiliary unlabeled visible images in a source domain and transfer the knowledge to cross-modality matching in the unlabeled target domain, because the underlying identity relation between visible image pairs can be discovered more easily than visible-infrared pairs with modality gap. To this end, we propose an unsupervised transferable visible-infrared person Re-ID framework as shown in Figure 1. The framework consists of two stages: RGB-gray asymmetric mutual learning and unsupervised cross-modality self-training.

This work was partially supported by the National Science Foundation for Young Scientists of China (62106288), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012974), the Guangzhou Basic and Applied Basic Research Scheme (2024A04J4066), the NSFC (U21A20471, U1911401, U1811461), Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085) and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (23ptpy77).

Ancong Wu and Chengzhi Lin are with the school of computer science and engineering, Sun Yat-sen University, Guangzhou 510006, China; Guangdong Key Laboratory of Information Security Technology (Email: wuanc@mail.sysu.edu.cn, linchzh3@mail2.sysu.edu.cn). Wei-Shi Zheng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China; Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education; Pazhou Laboratory (Huangpu), Guangzhou 510555, China (E-mail: wszheng@ieee.org).

Corresponding author: Wei-Shi Zheng.

In the first stage, to extract modality-shared prior knowledge from auxiliary unlabeled visible images, we transform the visible images (RGB) to grayscale fake infrared images (gray) by a diffusion model [33] trained on unlabeled infrared images in the target domain. The RGB images and the transformed grayscale images are regarded as two views. We assume that some RGB-infrared shared information can be extracted from the fake visible-infrared image pairs. Since the fake infrared images cannot perfectly represent the characteristics of real infrared images, some RGB-infrared shared knowledge is not contained in RGB-gray shared information. Hence, when extracting RGB-gray shared information, we expect to preserve the RGB-specific information that also contains RGB-infrared shared information. A tentative solution of extracting RGB-gray shared information is learning an RGB-gray feature extractor that can classify the pseudo labels of unlabeled samples and output view-consistent feature distributions. However, our information theoretic analysis in Section IV demonstrates that RGB-gray shared information and RGB-specific information cannot be simultaneously learned with a single view-shared model. To avoid this problem, we introduce an additional auxiliary gray feature extractor to quantify RGB-gray shared knowledge and transfer it to the RGB-gray feature extractor without eliminating RGB-specific knowledge. This process of knowledge transfer between two models forms a mutual learning framework, which is called the Cross-Modality Asymmetric Mutual Learning (CMAM).

In the second stage, to exploit the prior knowledge learned from visible images for unsupervised cross-modality Re-ID, we propose cross-modality mutual self-training. Complementary knowledge of RGB-gray matching and gray-gray matching in two models is fused and transferred to each other by mutual learning, so that only the RGB-gray feature extractor is required for inference without increasing computation costs. To alleviate modality gap for self-training, we utilize a bipartite cross-modality pseudo labeling strategy to separately perform intra-modality clustering and cross-modality cluster association based on bipartite graph matching.

For evaluating unsupervised learning in real-world scenario, we collected a new multi-modality pedestrian dataset SYSU-MM02, of which the training set is constructed by pedestrian detection on untrimmed videos. Such training set contains more noises than the training sets of existing visible-infrared pedestrian benchmark datasets, in which the samples are manually selected and annotated. SYSU-MM02 contains 25,774 images captured from one visible camera and two infrared cameras, in which 4,720 images of 118 identities are annotated for testing. This new dataset can be publicly available after data masking.

The contributions are summarized as follows:

1. We propose Cross-Modality Asymmetric Mutual Learning (CMAM) to learn modality-shared prior knowledge from auxiliary unlabeled visible images for visible-infrared matching.
2. To exploit prior knowledge for unsupervised cross-modality Re-ID, we propose cross-modality mutual self-training.
3. A new public visible-infrared pedestrian dataset SYSU-MM02 is constructed from real-world untrimmed videos.

II. Related Work

A. Visible-Infrared Person Re-Identification

Single-modality person re-identification methods for visible images have achieved remarkable performance in both supervised learning setting [1]–[4] and unsupervised learning setting [5]–[12], [14], [15]. However, visible-infrared person re-identification methods are still far from satisfaction.

Supervised Methods. The visible-infrared cross-modality person re-identification problem is first identified by Wu et al. [16]. Recently, supervised visible-infrared Re-ID has attracted increasing research interest. We categorize the methods into three groups: feature space alignment [17]–[26], local feature learning [34]–[38] and image style transformation [39]–[47].

Many approaches focus on learning consistent feature distributions for two modalities. Most methods focus on designing alignment loss functions [18], [21], [23]–[26]. Wu et al. [23] propose cross-modality similarity preservation for learning consistent cross-modality and same-modality ranking lists. Dai et al. [25] and Hao et al. [21] align cross-modality features by adversarial learning loss. Liu et al. [18] propose memory-augmented unidirectional metrics. Wu et al. [20] proposes a joint modality and pattern alignment network. Feng et al. [17] proposes a shaped-erased feature learning method to enrich modality-shared features. Chai et al. [48] learn two-stream transformers with distance distribution alignment for modality-specific class embeddings. CM-NAS [22] searches model architectures to find the optimal separation scheme of each BN layer for suppressing modality gap. The above feature alignment methods require cross-modality labeled data for alleviating modality gap, while our method can exploit auxiliary unlabeled RGB images to assist alleviating modality gap by asymmetric mutual learning.

Local feature learning methods mainly exploit attention module [37] or salience map [38] to enhance fine-grained feature representation. Zhang et al. [36] preserve spatial structures and attend to the differences of cross-modality image pairs. Chen et al. [35] propose structure-aware positional transformer network to utilize structural and positional information. The Semantic Alignment and Affinity Inference framework (SAAI) [34] aligns latent semantic part features with learnable prototypes and fuses local features and global features.

Image style transformation methods alleviate modality gap by transforming image styles to generate auxiliary training data. Wang et al. [47] perform cross-modality image generation by generative adversarial network. Li et al. [46] and Wei et al. [45] learn generator to convert images to a third modality. Huang et al. [42] design a modality-adaptive mixup scheme. To generate infrared-like grayscale images, Liu et al. [44] use linear combination of R, G, B channels and Ye et al. propose channel augmented joint learning (CAJL) [39], [43] that divides R, G, B channel for data augmentation. PartMix [40] method synthesizes augmented samples by mixing the part descriptors across the modalities. Zhong et al. [41] propose grayscale enhancement colorization network to colorize infrared images for bridging modality gap. The above image style transformation methods that transfer the style of visible images or infrared images in the target domain to augment training data. Although we also apply image style transformation for generating fake infrared images, our method can additionally exploit auxiliary unlabeled RGB images in a source domain for learning visible-infrared shared prior knowledge by asymmetric mutual learning.

Besides alleviating modality gap, some works take other factors of Re-ID into account. Yang et al. [49] focus on noisy label problem for visible-infrared Re-ID. Zhang et al. explore cross-modality Re-ID under low lighting [50]. For video-based visible-infrared Re-ID, Lin et al. [51] learn modality-invariant and motion-invariant features from videos and collect a new video multi-modality pedestrian dataset. Li et al. [52] leverage anaglyph data as intermediary for mitigating modality discrepancy and introduce a bidirectional spatial–temporal aggregation module to extract motion features from visible and infrared videos. Our method also generates fake infrared images as intermediary but does not rely on visible-infrared pairs of the same identity as the method of Li et al. [52]. Wang et al. [53] introduce near infrared and thermal infrared images to assist traditional Re-ID task on visible images by cross-modality interaction for multi-modality fusion. Our method explores modality interaction for alleviating modality gap in a different way derived from information theoretic analysis. We propose asymmetric mutual learning on the generated fake visible-infrared pairs for extracting modality-shared knowledge from auxiliary unlabeled visible images.

The above methods are supervised methods, while unsupervised cross-modality Re-ID remains a challenging problem.

Unsupervised Methods. Unsupervised visible-infrared Re-ID is still under-explored. Liang et al. [27] propose homogeneous-to-heterogeneous learning to perform pseudo labeling based on model pretrained on visible images. Wang et al. [28] propose an optimal-transport strategy to assign pseudo labels from visible to infrared modality. Yang et al. [29] propose an Augmented Dual-Contrastive Aggregation (ADCA) learning framework to learn intra-modality representation and associate cross-modality samples. Wu et al. [30] propose Progressive Graph Matching (PGM) to mine cross-modality correspondences under cluster imbalance scenarios. Yang et al. [31] propose a hierarchical framework to learn grand unified representation (GUR) for unlabeled samples by a bottom-up domain learning strategy. Pang et al. [32] propose cross-modality hierarchical clustering and refinement (CHCR), which separates clustering process into intra-modality stage and inter-modality stage and then refines pseudo labels by the consistency of clustering results of three channels in RGB images.

By contrast, our method focuses on learning modality-shared prior knowledge from auxiliary unlabeled visible images in a source domain to alleviate modality gap for cross-modality unsupervised domain adaptation, which complements existing unsupervised Re-ID techniques that focus on developing pseudo labeling strategies in the target domain.

B. Single-Modality Unsupervised Person Re-Identification

For extending Re-ID systems to new scenes without additional identity annotation, unsupervised learning for Re-ID has undergone fast development in recent years and achieves comparable performance as supervised Re-ID on visible images. Recent deep unsupervised Re-ID methods can be categorized into two types: pseudo labeling and style transfer.

The pseudo labeling methods obtain pseudo labels by clustering and progressively refine the noisy pseudo labels, such as soft multi-label learning [7], patch-based feature learning [8], asymmetric metric learning [9], camera-aware proxy [10], mutual mean-teaching [11], group-aware label transfer [12], inter-instance contrastive encoding [13], part-based pseudo label refinement [14], discrepant multi-instance proxies [54] and camera-driven representation learning [15].

The style transfer methods generally transfer camera-specific image styles by generative adversarial networks (GANs), such as HHL [6], CR-GAN [55] and JVTC [56]. The generated fake training data is used to alleviate the impact of cross-camera scene variations.

These unsupervised learning methods are developed for single-modality data and cannot alleviate modality gap for unsupervised visible-infrared Re-ID as our method.

C. Mutual Learning

Knowledge distillation [57] is a technique for transferring knowledge from a teacher model to a student model. By extending knowledge distillation from one-way transfer to two-way transfer, Zhang et al. [58] explore collaboratively learning multiple deep models that teach each other, which is called deep mutual learning (DML).

Deep mutual learning has also been applied for person re-identification. For unsupervised Re-ID, Ge et al. [11] propose mutual mean-teaching that use predictions of one model to guide the other model. For supervised visible-infrared person re-identification, Zhang et al. [59] propose a dual mutual learning method that transfers knowledge between a modality-shared branch and two modality-specific branches (RGB-specific branch and infrared-specific branch) to align features of different modalities.

For unsupervised transferable visible-infrared Re-ID, our asymmetric mutual learning method learns both RGB-gray shared information and RGB-specific information from auxiliary unlabeled visible images and the transformed grayscale fake infrared images to provide RGB-infrared shared prior knowledge for discovering visible-infrared sample relation. In contrast, the above mutual learning methods [11], [58], [59] are not suitable for our auxiliary generated RGB-gray image pairs, because they are limited to extracting only RGB-gray shared information with RGB-specific information eliminated, which also contains RGB-infrared shared information.

III. Problem Formulation

To learn modality-shared prior knowledge for visible-infrared matching, we utilize an auxiliary unlabeled visible image dataset $D_{RGBA} = \{\mathbf{I}_i^{RGBA}\}_{i=1}^{N_{RGBA}}$ that captures pedestrians by multiple cameras in a source domain, where N_{RGBA} is the number of samples. In comparison to discovering the underlying relations of visible-infrared image pairs, it is easier to discover the underlying relations of single-modality visible image pairs by current advanced unsupervised single-modality re-identification methods, such as PPLR [14]. For each image $\mathbf{I}_i^{RGBA}$, a pseudo identity label $\hat{y}_i^{RGBA}$ can be obtained based on the ImageNet pretrained ResNet-50 model [60]. For an unseen scene as target domain, we acquire unlabeled visible image set $D_{RGB} = \{\mathbf{I}_j^{RGB}\}_{j=1}^{N_{RGB}}$ and unlabeled infrared image set $D_{IR} = \{\mathbf{I}_k^{IR}\}_{k=1}^{N_{IR}}$. We assume that D_{RGB} and D_{IR} are captured in a local area in short term, so that most identities are captured in both visible cameras and infrared cameras.

Based on the training data, we develop a two-stage training framework. In the first stage, on the auxiliary unlabeled visible image dataset D_{RGBA}, we train an auxiliary pretrained model M_{pre} to learn modality-shared prior knowledge. In the second stage, based on the auxiliary pretrained model, we leverage the prior knowledge for unsupervised domain adaptation on unlabeled visible image sets D_{RGB} and D_{IR}.

IV. RGB-GRAY ASYMMETRIC MUTUAL LEARNING

To learn modality-shared prior knowledge for visible-infrared matching on unlabeled visible images in the first stage, we propose Cross-Modality Asymmetric Mutual Learning (CMAM), of which the overview is shown in Figure 2.

A. Grayscale Fake Infrared Image Transformation

To extract modality-shared knowledge on the auxiliary unlabeled visible images $\mathbf{I}_i^{RGBA}$, we learn to transform them to fake infrared images given the unlabeled infrared images $\mathbf{I}_k^{IR}$ in the target domain. To model the data distribution of infrared images, we take advantage of the strong distribution modeling ability of diffusion models [33] for generating fake infrared images from RGB images. To preserve the modality-invariant shape and texture information in RGB images for transformation, we exploit the edge map $\mathbf{E}_i^{RGBA}$ extracted from $\mathbf{I}_i^{RGBA}$ by Sobel operator [61] as condition for the diffusion-based generator. We adopt a convolutional network C_{con} following T2I-adapter [62] to encode the edge map $\mathbf{E}_i^{RGBA}$ into condition map for the generator. After training conditional diffusion model G_{IR} on unlabeled infrared images $\mathbf{I}_k^{IR}$ and the corresponding Sobel edge maps $\mathbf{E}_i^{IR}$ in the target domain, the auxiliary visible image $\mathbf{I}_i^{RGBA}$ is transformed to grayscale fake infrared image $\mathbf{I}_i^{grayA}$ by

$$\mathbf{I}_i^{grayA} = G_{IR}(\mathbf{N}, C_{con}(\mathbf{E}_i^{RGBA})), \tag{1}$$

where $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is Gaussian noise of the same size as visible image $\mathbf{I}_i^{RGBA}$, σ is the standard deviation of the Gaussian distribution.

We expect to extract modality-shared knowledge by exploring the relation between the fake visible-infrared pairs in the source domain, in order to facilitate discovering visible-infrared sample relations in the target domain.

B. RGB-Gray Shared Feature Extraction

For simplicity, we use V_1, V_2 and V_3 to represent RGB images (view 1), grayscale fake infrared images[1] (view 2) and infrared (IR) images (view 3), respectively.

RGB-Gray Feature Extractor. To learn discriminative features for RGB images and grayscale images, we introduce an RGB-gray feature extractor M_{joint} that takes both RGB images $\mathbf{I}_i^{RGBA}$ and the corresponding grayscale fake infrared images $\mathbf{I}_i^{grayA}$ as input. We assign the same pseudo label $\hat{y}_i^{RGBA}$ for visible image $\mathbf{I}_i^{RGBA}$ and its corresponding transformed grayscale image $\mathbf{I}_i^{grayA}$, by applying advanced unsupervised Re-ID method PPLR [14] on visible images.

We apply identification losses $\mathcal{L}_{ID-jA1}$ and $\mathcal{L}_{ID-jA2}$ for RGB image feature $\mathbf{f}_i^{jA1} = M_{joint}(\mathbf{I}_i^{RGBA})$ and grayscale image feature $\mathbf{f}_i^{jA2} = M_{joint}(\mathbf{I}_i^{grayA})$, respectively. With specific classifiers C_{j1} and C_{j2}, the classification probabilities are $\mathbf{p}_i^{jA1} = C_{j1}(\mathbf{f}_i^{jA1})$ and $\mathbf{p}_i^{jA2} = C_{j2}(\mathbf{f}_i^{jA2})$ for RGB image feature and grayscale image feature, respectively. The identification loss $\mathcal{L}_{PID-jA1}$ for RGB images is

$$\mathcal{L}_{PID-jA1} = \mathcal{L}_{ce-jA1} + \mathcal{L}_{tri-jA1}, \tag{2}$$

where $\mathcal{L}_{ce-jA1}$ is cross-entropy loss [63] on $\mathbf{p}_i^{jA1}$ and $\mathcal{L}_{tri-jA1}$ is soft-margin triplet loss [64] on $\mathbf{f}_i^{jA1}$. The identification loss $\mathcal{L}_{PID-jA2}$ for grayscale images is similar.

Intra-View Feature Relation Preservation. We assume that the fake visible-infrared pairs $(\mathbf{I}_i^{RGBA}, \mathbf{I}_i^{grayA})$ contain some modality-shared information that can facilitate visible-infrared matching. To learn view-shared features for RGB images and grayscale images, we expect that the intra-view feature relations are consistent, that is, RGB-RGB feature relations and gray-gray feature relations are consistent. To achieve this, we preserve intra-view feature relations in both class level and instance level.

In class level, we preserve the classification probabilities that represent relations between a sample and all classes. The probability preservation loss $\mathcal{L}_{PP}$ is

$$\mathcal{L}_{PP} = \sum_{i=1}^{N_{RGBA}} D_{SKL}(\mathbf{p}_i^{jA1}, \mathbf{p}_i^{jA2}), \tag{3}$$

where $\mathbf{p}_i^{jA1}$ and $\mathbf{p}_i^{jA2}$ are classification probability vectors of RGB images and grayscale images, respectively; $D_{SKL}(p,q) = D_{KL}(p\|q) + D_{KL}(q\|p)$ denotes the symmetric Kullback-Leibler (KL) divergence.

In instance level, we preserve the pairwise similarities between samples, since Re-ID is a similarity-based retrieval task. The similarity preservation loss $\mathcal{L}_{SP}$ is

$$\mathcal{L}_{SP} = \left\|\mathbf{F}_{jA1}^{\top}\mathbf{F}_{jA1} - \mathbf{F}_{jA2}^{\top}\mathbf{F}_{jA2}\right\|_F^2, \tag{4}$$

where $\mathbf{F}_{jA1} = [\mathbf{f}_1^{jA1}, \mathbf{f}_2^{jA1}, ..., \mathbf{f}_{N_{RGBA}}^{jA1}]$ and $\mathbf{F}_{jA2} = [\mathbf{f}_1^{jA2}, \mathbf{f}_2^{jA2}, ..., \mathbf{f}_{N_{RGBA}}^{jA2}]$ ($\mathbf{F}_{jA1}, \mathbf{F}_{jA2} \in \mathbb{R}^{d \times N_{RGBA}}$, d is the dimensionality of feature vector). The features are normalized by $\ell 2$-norm.

Information Theoretic Analysis. The losses $\mathcal{L}_{PP}$ in Eq. (3) and $\mathcal{L}_{SP}$ in Eq. (4) minimize the difference of feature distributions of two views. We analyze the effect of the losses from the perspective of mutual information.

For data of two views $\mathbf{v}_1, \mathbf{v}_2 \sim p(V_1, V_2)$, we assume that the corresponding representations $\mathbf{f}_1 \sim p_{\theta_1}(f_1|V_1)$ and $\mathbf{f}_2 \sim p_{\theta_2}(f_2|V_2)$ are in the same domain. Without loss of generality,

[1] In the following sections, "grayscale fake infrared image" is also called "grayscale image" for simplicity.

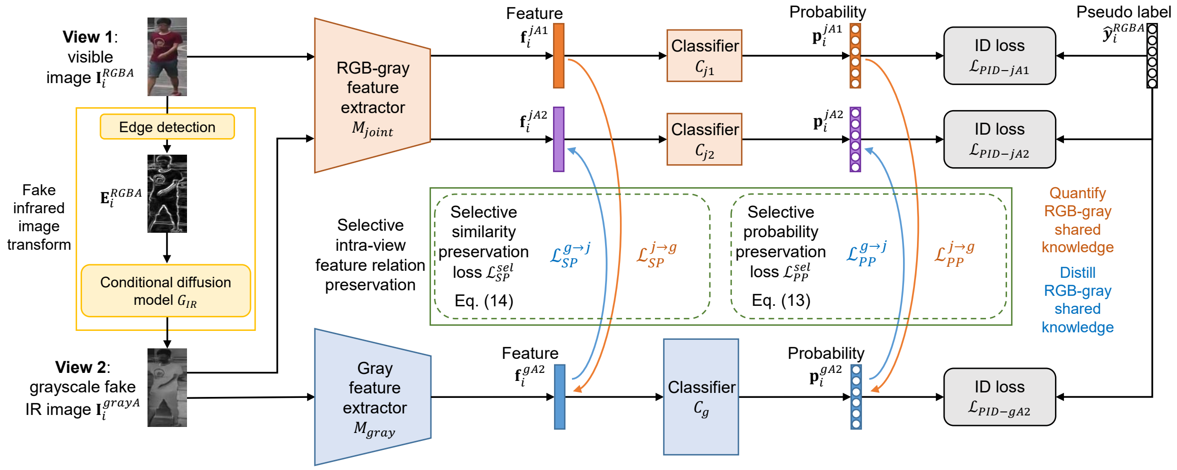

View 1: visible image $\mathbf{I}_i^{RGBA}$
Edge detection
$\mathbf{E}_i^{RGBA}$
Fake infrared image transform
Conditional diffusion model G_{IR}
View 2: grayscale fake IR image $\mathbf{I}_i^{grayA}$
RGB-gray feature extractor M_{joint}
Gray feature extractor M_{gray}
Feature $\mathbf{f}_i^{jA1}$
$\mathbf{f}_i^{jA2}$
Feature $\mathbf{f}_i^{gA2}$
Classifier C_{j1}
Classifier C_{j2}
Classifier C_g
Probability $\mathbf{p}_i^{jA1}$
$\mathbf{p}_i^{jA2}$
Probability $\mathbf{p}_i^{gA2}$
ID loss $\mathcal{L}_{PID-jA1}$
ID loss $\mathcal{L}_{PID-jA2}$
ID loss $\mathcal{L}_{PID-gA2}$
Pseudo label $\hat{\mathbf{y}}_i^{RGBA}$
Selective intra-view feature relation preservation
Selective similarity preservation loss $\mathcal{L}_{SP}^{sel}$ Eq. (14)
$\mathcal{L}_{SP}^{g\to j}$
$\mathcal{L}_{SP}^{j\to g}$
Selective probability preservation loss $\mathcal{L}_{PP}^{sel}$ Eq. (13)
$\mathcal{L}_{PP}^{g\to j}$
$\mathcal{L}_{PP}^{j\to g}$
Quantify RGB-gray shared knowledge
Distill RGB-gray shared knowledge

Fig. 2. Overview of Cross-Modality Asymmetric Mutual Learning (CMAM). In the first stage, CMAM learns modality-shared prior knowledge on auxiliary unlabeled visible images and the transformed grayscale fake infrared images in the source domain. The framework consists of a RGB-gray feature extractor M_{joint} and a gray feature extractor M_{gray}. We expect that, RGB image feature $\mathbf{f}_i^{jA1}$ contains view-shared information as well as RGB-specific information and grayscale image features $\mathbf{f}_i^{jA2}, \mathbf{f}_i^{gA2}$ contains view-shared information. Both the view-shared information and RGB-specific information extracted from the fake visible-infrared pairs are required to facilitate learning modality-shared information of real unlabeled visible-infrared pairs in the target domain.

the feature extractor parameters of two views are denoted by θ_1 and θ_2, respectively.

The mutual information between features f_1 and data V_1 of view 1 can be expressed by chain rule as

$$I_{\theta_1}(V_1;f_1) = I_{\theta_1}(V_1;f_1|V_2) + I_{\theta_1}(V_1;V_2;f_1), \tag{5}$$

where the first term is view-specific information and the second term is view-shared information.

The upper bound of conditional mutual information $I_{\theta_1}(V_1;f_1|V_2)$ can be expressed as:

$$\begin{aligned} I_{\theta_1}(V_1;f_1|V_2) &= \mathbb{E}_{\mathbf{v_1},\mathbf{v_2}}\mathbb{E}_{\mathbf{f_1}\sim p_{\theta_1}(f_1|V_1)}\log\frac{p_{\theta_1}(f_1|V_1)}{p_{\theta_1}(f_1|V_2)} \\ &= \mathbb{E}_{\mathbf{v_1},\mathbf{v_2}}\mathbb{E}_{\mathbf{f_1}\sim p_{\theta_1}(f_1|V_1)}\log\frac{p_{\theta_1}(f_1|V_1)}{p_{\theta_2}(f_2|V_2)}\frac{p_{\theta_2}(f_2|V_2)}{p_{\theta_1}(f_1|V_2)} \\ &= D_{KL}(p_{\theta_1}(f_1|V_1)||p_{\theta_2}(f_2|V_2)) - D_{KL}(p_{\theta_1}(f_1|V_2)||p_{\theta_2}(f_2|V_2)) \\ &\leq D_{KL}(p_{\theta_1}(f_1|V_1)||p_{\theta_2}(f_2|V_2)). \end{aligned} \tag{6}$$

The detailed derivation is in the supplementary material.

Hence, minimizing the KL divergence between feature distributions of two views by updating θ_1 eliminates the view-specific mutual information $I_{\theta_1}(V_1;f_1|V_2)$ in feature f_1.

Analogously, the upper bound of $I_{\theta_2}(V_2;f_2|V_1)$ is

$$I_{\theta_2}(V_2;f_2|V_1) \leq D_{KL}(p_{\theta_2}(f_2|V_2)||p_{\theta_1}(f_1|V_1)). \tag{7}$$

When $I_{\theta_1}(V_1;f_1|V_2)$ and $I_{\theta_2}(V_2;f_2|V_1)$ are decreased to 0, we have $I_{\theta_1}(V_1;f_1) = I_{\theta_1}(V_1;V_2;f_1)$ and $I_{\theta_2}(V_2;f_2) = I_{\theta_2}(V_1;V_2;f_2)$ according to Eq. (5), which means that the representations f_1 and f_2 are view-shared features.

For our RGB-gray feature extractor M_{joint}, the feature extractor parameters for two views θ_1 and θ_2 are shared, which is denoted by θ_j. According to Eq. (6) and Eq. (7), we have

$$\begin{aligned} I_{\theta_j}(V_1;f_1|V_2) \leq D_{KL}(p_{\theta_j}(f_1|V_1)||p_{\theta_j}(f_2|V_2)), \\ I_{\theta_j}(V_2;f_2|V_1) \leq D_{KL}(p_{\theta_j}(f_2|V_2)||p_{\theta_j}(f_1|V_1)). \end{aligned} \tag{8}$$

We use the probability preservation loss $\mathcal{L}_{PP}$ in Eq. (3) to approximate $D_{KL}(p_{\theta_j}(f_1|V_1)||p_{\theta_j}(f_2|V_2)) + D_{KL}(p_{\theta_j}(f_2|V_2)||p_{\theta_j}(f_1|V_1))$ as the approach of Tian et al. [65]. The similarity preservation loss $\mathcal{L}_{SP}$ in Eq. (4) can also minimize the difference of feature distributions in instance level as $\mathcal{L}_{PP}$ in class level. Thus, $\mathcal{L}_{PP}$ and $\mathcal{L}_{SP}$ simultaneously decrease $I_{\theta_j}(V_1;f_1|V_2)$ and $I_{\theta_j}(V_2;f_2|V_1)$ to extract RGB-gray view-shared features.

C. Asymmetric Mutual Learning

The RGB-gray view-shared features extracted by RGB-gray feature extractor M_{joint} in the source domain contain some RGB-infrared shared information in the target domain. The RGB-infrared shared information $I(V_1;V_3)$ is represented by

$$I(V_1;V_3) = I(V_1;V_3;V_2) + I(V_1;V_3|V_2). \tag{9}$$

The RGB-gray view-shared features extracted from $I(V_1;V_2)$ are expected to contain some information of $I(V_1;V_3;V_2)$ in the first term. However, since the conditional diffusion model G_{IR} cannot perfectly transform the visible images to infrared images, the generated fake infrared images (V_2) contain noises as compared with real infrared images (V_3). Moreover, there exists data distribution gap between fake visible-infrared pairs in the source domain and real visible-infrared pairs in the target domain. Thus, the second term $I(V_1;V_3|V_2)$ is not 0 and it is also a useful part of RGB-infrared shared information that is not contained in the RGB-gray view-shared features.

To make the extracted features contain information of $I(V_1;V_3|V_2)$, RGB-specific information $I(V_1|V_2)$ in the source domain should also be preserved for further extraction of RGB-infrared shared information in the target domain. However, minimizing $I_{\theta_j}(V_1;f_1|V_2)$ hinders learning RGB-specific features. When updating θ_j that is shared for two views, it is difficult to minimize $I_{\theta_j}(V_2;f_2|V_1)$ without decreasing $I_{\theta_j}(V_1;f_1|V_2)$.

Uni-Directional Intra-View Feature Relation Preservation. In order to simultaneously learn RGB-gray shared features and RGB-specfic features, we introduce an auxiliary gray feature extractor M_{gray} parameterized by θ_g to overcome the limitation of using shared parameter θ_j for two views. The gray feature extractor M_{gray} takes only grayscale images $\mathbf{I}_i^{grayA}$ as input to extract feature $\mathbf{f}_i^{gA2}$. With classifier C_g, the classification probability vector is $\mathbf{p}_i^{gA2} = C_g(\mathbf{f}_i^{gA2})$. Features $\mathbf{f}_i^{gA2}$ are learned by applying identification loss $\mathcal{L}_{PID-gA2}$, which is similar to $\mathcal{L}_{PID-jA1}$ in Eq. (2). By substituting $\theta_1 = \theta_j$ and $\theta_2 = \theta_g$ in Eq. (6) and Eq. (7), we have

$$\begin{aligned} I_{\theta_j}(V_1;f_1|V_2) \leq D_{KL}(p_{\theta_j}(f_1|V_1) \| p_{\theta_g}(f_2|V_2)), \\ I_{\theta_g}(V_2;f_2|V_1) \leq D_{KL}(p_{\theta_g}(f_2|V_2) \| p_{\theta_j}(f_1|V_1)). \end{aligned} \tag{10}$$

Our objective is that, parameter update of θ_g minimizes $I_{\theta_g}(V_2;f_2|V_1)$ for learning RGB-gray shared information in f_2; meanwhile, parameter update of θ_j does not decrease $I_{\theta_j}(V_1;f_1|V_2)$ for learning RGB-specific information in f_1.

To achieve this objective, we formulate a uni-directional probability preservation loss $\mathcal{L}_{PP}^{j\to g}$ as

$$\mathcal{L}_{PP}^{j\to g} = \sum_{i=1}^{N_{RGBA}} D_{SKL}(\mathbf{p}_i^{gA2}, \text{stopgrad}(\mathbf{p}_i^{jA1})), \tag{11}$$

where $\mathbf{p}_i^{jA1}$ is the classification probability vector of RGB image output by C_{j1} and $\mathbf{p}_i^{gA2}$ is the classification probability vector of grayscale image output by C_g; D_{SKL} denotes symmetric KL divergence as that in Eq. (3); $\text{stopgrad}(\cdot)$ denotes stop gradient for back propagation in optimization, which enables uni-directional knowledge transfer.

Similarly, the uni-directional similarity preservation loss $\mathcal{L}_{SP}^{j\to g}$ is formulated as

$$\mathcal{L}_{SP}^{j\to g} = \left\| \mathbf{F}_{gA2}^{\top}\mathbf{F}_{gA2} - \text{stopgrad}(\mathbf{F}_{jA1}^{\top}\mathbf{F}_{jA1}) \right\|_F^2, \tag{12}$$

where $\mathbf{F}_{gA2} = [\mathbf{f}_1^{gA2}, \mathbf{f}_2^{gA2}, ..., \mathbf{f}_{N_{RGBA}}^{gA2}]$ and $\mathbf{F}_{jA1} = [\mathbf{f}_1^{jA1}, \mathbf{f}_2^{jA1}, ..., \mathbf{f}_{N_{RGBA}}^{jA1}]$ are feature matrices.

Since the gradient with respect to θ_j is stopped for $\mathbf{p}_i^{jA1}$ and $\mathbf{f}_i^{jA1}$, minimizing $\mathcal{L}_{PP}^{j\to g}$ and $\mathcal{L}_{SP}^{j\to g}$ only updates θ_g to decrease $I_{\theta_g}(V_2;f_2|V_1)$, which is the useless gray-specific information of V_2 that should be eliminated. Thus, the gray feature extractor M_{gray} quantifies the RGB-gray shared knowledge.

Selective Intra-View Feature Relation Preservation. By using the losses $\mathcal{L}_{PP}^{j\to g}$ and $\mathcal{L}_{SP}^{j\to g}$, the gray feature extractor M_{gray} extracts view-shared information in features $\mathbf{f}_i^{gA2}$, which quantify the RGB-gray shared knowledge. To learn view-shared information for grayscale image features $\mathbf{f}_i^{jA2}$ extracted by RGB-gray feature extractor M_{joint}, the gray feature extractor M_{gray} is regarded as teacher model to transfer knowledge to the RGB-gray feature extractor M_{joint} by knowledge distillation.

To achieve this, we formulate a selective probability preservation loss $\mathcal{L}_{PP}^{sel}$ as

$$\begin{aligned} \mathcal{L}_{PP}^{sel} &= \mathcal{L}_{PP}^{j\to g} + \mathcal{L}_{PP}^{g\to j} \\ &= \sum_{i=1}^{N_{RGBA}} D_{SKL}(\mathbf{p}_i^{gA2}, \text{stopgrad}(\mathbf{p}_i^{jA1})) \\ &+ D_{SKL}(\mathbf{p}_i^{jA2}, \text{stopgrad}(\mathbf{p}_i^{gA2})), \end{aligned} \tag{13}$$

where the second term $\mathcal{L}_{PP}^{g\to j}$ is logit-based knowledge distillation loss from M_{gray} to M_{joint} for grayscale image features.

Similarly, selective similarity preservation loss $\mathcal{L}_{SP}^{j\to g}$ is

$$\begin{aligned} \mathcal{L}_{SP}^{sel} &= \mathcal{L}_{SP}^{j\to g} + \mathcal{L}_{SP}^{g\to j} \\ &= \left\| \mathbf{F}_{gA2}^{\top}\mathbf{F}_{gA2} - \text{stopgrad}(\mathbf{F}_{jA1}^{\top}\mathbf{F}_{jA1}) \right\|_F^2 \\ &+ \left\| \mathbf{F}_{jA2}^{\top}\mathbf{F}_{jA2} - \text{stopgrad}(\mathbf{F}_{gA2}^{\top}\mathbf{F}_{gA2}) \right\|_F^2, \end{aligned} \tag{14}$$

where the second term $\mathcal{L}_{SP}^{g\to j}$ is similarity-based knowledge distillation loss from M_{gray} to M_{joint} for grayscale image features.

The grayscale image features $\mathbf{f}_i^{gA2}$ extracted by M_{gray} can be regarded as intermediate features containing RGB-gray shared knowledge selected from RGB image features $\mathbf{f}_i^{jA1}$, and the intra-view feature relations for $\mathbf{f}_i^{jA1}$ and $\mathbf{f}_i^{jA2}$ are preserved indirectly without eliminating the useful RGB-specific knowledge in $\mathbf{f}_i^{jA1}$. Thus, we call $\mathcal{L}_{PP}^{sel}$ and $\mathcal{L}_{SP}^{sel}$ the selective intra-view feature relation preservation.

Compared with intra-view feature relation preservation using $\mathcal{L}_{PP}$, $\mathcal{L}_{SP}$ in Eq. (3) and Eq. (4) that guide the model to learn only view-shared information, $\mathcal{L}_{PP}^{sel}$ and $\mathcal{L}_{SP}^{sel}$ guide the models to extract view-shared information in features $\mathbf{f}_i^{jA2}$, $\mathbf{f}_i^{gA2}$ and extract RGB-specific information as well as view-shared information in features $\mathbf{f}_i^{jA1}$.

Cross-Modality Asymmetric Mutual Learning (CMAM). In the RGB-gray asymmetric mutual learning stage, the identification loss and the selective intra-view feature relation preservation losses are jointly optimized on the source domain to learn the auxiliary pretrained model by

$$\mathcal{L}_{pre}^{CMAM} = \lambda_{PP}\mathcal{L}_{PP}^{sel} + \lambda_{SP}\mathcal{L}_{SP}^{sel} + \mathcal{L}_{PID-pre}, \tag{15}$$

where $\mathcal{L}_{PID-pre} = \mathcal{L}_{PID-jA1} + \mathcal{L}_{PID-jA2} + \mathcal{L}_{PID-gA2}$ is the identification loss; λ_{PP} and λ_{SP} are trade-off parameters.

Since knowledge transfer between two models are different for two modalities, we call this framework Cross-Modality Asymmetric Mutual Learning (CMAM).

V. Unsupervised Cross-Modality Self-training

After training on auxiliary unlabeled visible images by Cross-Modality Asymmetric Mutual Learning (CMAM), RGB-gray view-shared features and RGB-specific features are learned by the auxiliary pretrained model. To exploit such prior knowledge for unsupervised visible-infrared Re-ID in the target domain, we propose cross-modality mutual self-training as shown in Figure 3.

Given RGB image $\mathbf{I}_j^{RGB}$ (view 1) and infrared image $\mathbf{I}_k^{IR}$ (view 3) in the target domain, the RGB-gray feature extractor M_{joint} takes visible image $\mathbf{I}_j^{RGB}$ and infrared image $\mathbf{I}_k^{IR}$ as input and outputs features $\mathbf{f}_j^{j1} \in \mathbb{R}^d$, $\mathbf{f}_k^{j3} \in \mathbb{R}^d$ and classification probability vectors $\mathbf{p}_j^{j1}$, $\mathbf{p}_k^{j3}$. The gray feature extractor M_{gray} takes the transformed grayscale image $\mathbf{I}_j^{gray}$ and infrared image $\mathbf{I}_k^{IR}$ as input and outputs features $\mathbf{f}_j^{g2} \in \mathbb{R}^d$, $\mathbf{f}_k^{g3} \in \mathbb{R}^d$ and classification probability vectors $\mathbf{p}_j^{g2}$, $\mathbf{p}_k^{g3}$.

A. Bipartite Cross-Modality Pseudo Labeling

To improve the discrimination ability of features, self-training by classifying pseudo labels is a favorable training

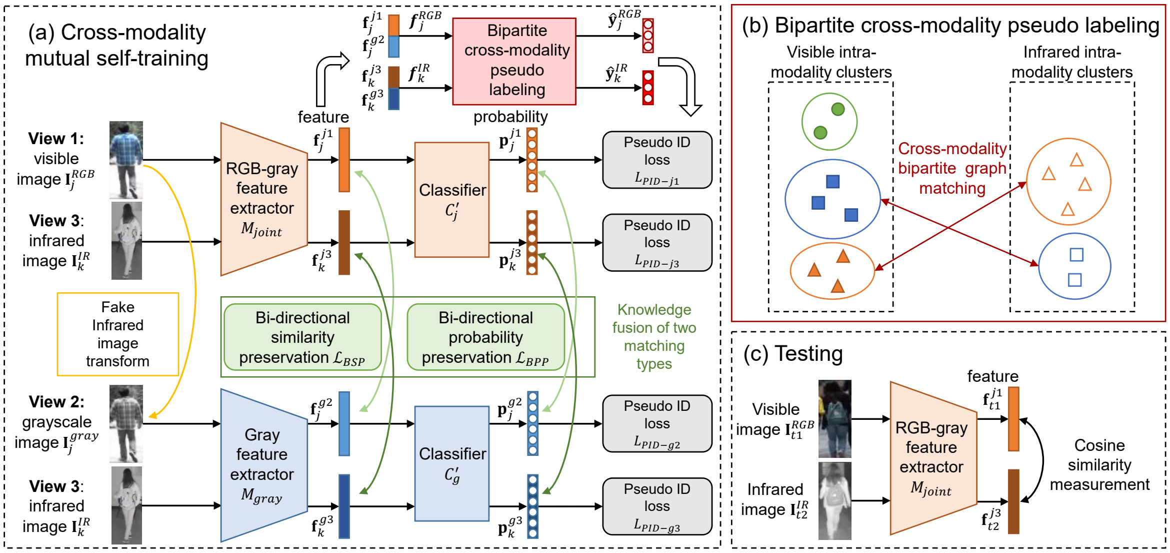

(a) Cross-modality mutual self-training
View 1: visible image $\mathbf{I}_j^{RGB}$
View 3: infrared image $\mathbf{I}_k^{IR}$
Fake Infrared image transform
View 2: grayscale image $\mathbf{I}_j^{gray}$
View 3: infrared image $\mathbf{I}_k^{IR}$
RGB-gray feature extractor M_{joint}
Gray feature extractor M_{gray}
feature
$\mathbf{f}_j^{j1}$
$\mathbf{f}_j^{g2}$
$\mathbf{f}_k^{j3}$
$\mathbf{f}_k^{g3}$
$\mathbf{f}_j^{RGB}$
$\mathbf{f}_k^{IR}$
Bipartite cross-modality pseudo labeling
$\hat{\mathbf{y}}_j^{RGB}$
$\hat{\mathbf{y}}_k^{IR}$
probability
Classifier C_j'
Classifier C_g'
$\mathbf{p}_j^{j1}$
$\mathbf{p}_k^{j3}$
$\mathbf{p}_j^{g2}$
$\mathbf{p}_k^{g3}$
Pseudo ID loss $\mathcal{L}_{PID-j1}$
Pseudo ID loss $\mathcal{L}_{PID-j3}$
Pseudo ID loss $\mathcal{L}_{PID-g2}$
Pseudo ID loss $\mathcal{L}_{PID-g3}$
Bi-directional similarity preservation $\mathcal{L}_{BSP}$
Bi-directional probability preservation $\mathcal{L}_{BPP}$
Knowledge fusion of two matching types
(b) Bipartite cross-modality pseudo labeling
Visible intra-modality clusters
Infrared intra-modality clusters
Cross-modality bipartite graph matching
(c) Testing
Visible image $\mathbf{I}_{t1}^{RGB}$
Infrared image $\mathbf{I}_{t2}^{IR}$
RGB-gray feature extractor M_{joint}
feature
$\mathbf{f}_{t1}^{j1}$
$\mathbf{f}_{t2}^{j3}$
Cosine similarity measurement

Fig. 3. Overview of unsupervised cross-modality mutual self-training. In the second stage, cross-modality mutual self-training is performed on multi-modality data in the target domain. To leverage the prior knowledge in auxiliary pretrained models, bi-directional knowledge transfer between RGB-gray feature extractor M_{joint} and gray feature extractor M_{gray} fuses the knowledge of RGB-gray matching and gray-gray matching. The double arrows indicates bi-directional knowledge transfer. Bipartite cross-modality pseudo labeling associates intra-modality clusters of two modalities by bipartite graph matching to alleviate the impact of modality gap. In testing, only M_{joint} is required for feature extraction.

strategy in advanced unsupervised Re-ID methods [13], [14]. When determining the pseudo labels $\hat{\mathbf{y}}_j^{RGB}$ and $\hat{\mathbf{y}}_k^{IR}$ for $\mathbf{I}_j^{RGB}$ and $\mathbf{I}_k^{IR}$, we use the features $\mathbf{f}_j^{RGB} = [\mathbf{f}_j^{j1}; \mathbf{f}_j^{g2}] \in \mathbb{R}^{2d}$ and $\mathbf{f}_k^{IR} = [\mathbf{f}_k^{j3}; \mathbf{f}_k^{g3}] \in \mathbb{R}^{2d}$ concatenated by features of two models, since the knowledge of RGB-gray matching and gray-gray matching learned in two models complement each other. The process of pseudo labeling is shown in Figure 3 (b).

When associating similar samples in the same pseudo class, modality gap incurs ambiguity for intra-modality similarity and cross-modality similarity. To alleviate the impact of modality gap, we separate intra-modality sample association and cross-modality sample association in two steps. First, we perform intra-modality clustering for features $\{\mathbf{f}_j^{RGB}\}_{j=1}^{N_{RGB}}$ and $\{\mathbf{f}_k^{IR}\}_{k=1}^{N_{IR}}$, of which the cluster centers are $\{\mathbf{c}_m^{RGB}\}_{m=1}^{N_{RGB}^{clu}}$ and $\{\mathbf{c}_n^{IR}\}_{n=1}^{N_{IR}^{clu}}$, respectively. Second, we cast cross-modality sample association as a maximal matching problem in a bipartite graph. To construct a bipartite graph $G = (A_{RGB}, A_{IR}, E)$, we regard the cluster centers $\mathbf{c}_m^{RGB}$, $\mathbf{c}_n^{IR}$ as two disjoint vertex sets A_{RGB}, A_{IR} and the similarities between all cross-modality cluster center pairs ($\mathbf{c}_m^{RGB}$, $\mathbf{c}_n^{IR}$) as edges E. The set of edges E_{sel} selected for achieving maximal matching is solved by

$$E_{sel} = BGM(A_{RGB}, A_{IR}, E), \tag{16}$$

where the bipartite graph matching operation BGM is implemented by Hungarian algorithm [66].

Samples in the same intra-modality cluster and the matched cross-modality cluster pairs are assigned the same pseudo label. During training, the pseudo labels are updated after every 200 iterations. As assumed in Section III, there are potential cross-modality positive pairs for most identities. Although there are noises in the clusters, when discriminating the pseudo classes, the effect of pulling potential cross-modality positive samples closer in the feature space is dominant.

B. Knowledge Fusion by Mutual Learning

With pseudo labels $\{\hat{\mathbf{y}}_j^{RGB}\}_{j=1}^{N_{RGB}}, \{\hat{\mathbf{y}}_k^{IR}\}_{k=1}^{N_{IR}}$, the objective of self-training is identifying the pseudo classes by a identification loss $\mathcal{L}_{PID}$. The classifiers C_j' and C_g' for RGB-gray feature extractor M_{joint} and gray feature extractor M_{gray} are constructed by proxies of pseudo classes stored in a memory bank [12]. Each proxy is cluster centers computed by moving average of features.

The identification loss $\mathcal{L}_{PID}$ is formulated as

$$\mathcal{L}_{PID} = \mathcal{L}_{PID-j1} + \mathcal{L}_{PID-j3} + \mathcal{L}_{PID-g2} + \mathcal{L}_{PID-g3}, \tag{17}$$

where each term is computed as in Eq. (2) for $(\mathbf{f}_j^{j1}, \mathbf{p}_j^{j1})$, $(\mathbf{f}_k^{j3}, \mathbf{p}_k^{j3})$, $(\mathbf{f}_j^{g2}, \mathbf{p}_j^{g2})$, $(\mathbf{f}_k^{g3}, \mathbf{p}_k^{g3})$, respectively.

To fuse the complementary knowledge of RGB-gray matching and gray-gray matching in two models, we exploit mutual learning to learn consistent classification probabilities and sample similarities for two models.

In class level, a bi-directional probability preservation loss $\mathcal{L}_{BPP}$ is formulated as

$$\mathcal{L}_{BPP} = \sum_{j=1}^{N_{RGB}} d_{SKL}(\mathbf{p}_j^{j1}, \mathbf{p}_j^{g2}) + \sum_{k=1}^{N_{IR}} d_{SKL}(\mathbf{p}_k^{j3}, \mathbf{p}_k^{g3}). \tag{18}$$

In instance level, a bi-directional similarity preservation loss $\mathcal{L}_{BSP}$ is formulated as

$$\mathcal{L}_{BSP} = \left\|\mathbf{F}_{j1}^{\top}\mathbf{F}_{j1} - \mathbf{F}_{g2}^{\top}\mathbf{F}_{g2}\right\|_F^2 + \left\|\mathbf{F}_{j3}^{\top}\mathbf{F}_{j3} - \mathbf{F}_{g3}^{\top}\mathbf{F}_{g3}\right\|_F^2, \tag{19}$$

where $\mathbf{F}_{j1}$, $\mathbf{F}_{j3}$ $\mathbf{F}_{g2}$, $\mathbf{F}_{g3}$ are feature matrices constructed by $\mathbf{f}_j^{j1}$, $\mathbf{f}_k^{j3}$, $\mathbf{f}_j^{g2}$, $\mathbf{f}_k^{g3}$, respectively.

Finally, we integrate mutual learning and bipartite cross-modality pseudo labeling by cross-modality mutual self-training in Figure 3 (a). The loss function is

$$\mathcal{L}_{uda} = \lambda_{BPP}\mathcal{L}_{BPP} + \lambda_{BSP}\mathcal{L}_{BSP} + \mathcal{L}_{PID}, \tag{20}$$

where λ_{BPP} and λ_{BSP} are trade-off parameters.

TABLE I
COMPARISON BETWEEN IMAGE-BASED VISIBLE-INFRARED RE-ID DATASETS AND SYSU-MM02. "NIR"/"TIR" DENOTES NEAR INFRARED/THERMAL INFRARED.

Name	#ID	#Sample	#Camera			**Untrimmed**
			RGB	NIR	TIR	**training set**
SYSU-MM01 [23]	491	44,745	4	2	0	no
RegDB [67]	412	8,240	1	0	1	no
RGBNT201 [68]	201	14,361	4	4	4	no
LLCM [50]	1,064	46,767	9	9	0	no
SYSU-MM02	118 (test)	25,774	1	2	0	yes

C. Inference

After knowledge fusion by mutual learning, both models learn the fused knowledge. Only M_{joint} is required for inference and the computation cost is not increased. Retrieval results are determined by cosine similarity as shown in Figure 3 (c).

VI. A Visible-Infrared Re-ID Dataset Constructed from Untrimmed Videos

For unsupervised visible-infrared person re-identification, although existing multi-modality datasets SYSU-MM01 [23], RegDB [67], RGBNT201 [68], HITSZ-VCM [51] and LLCM [50] can be applied for evaluation by removing the identity annotations in the training set, the images are selected manually during annotation. Such simulated training data on benchmark datasets collected for supervised learning is different from training data acquired in real-world scenarios, where only untrimmed videos are available. Potential cross-camera positive pairs may not exist for some unlabeled samples. Hence, we collect a new visible-infrared pedestrian dataset SYSU-MM02, of which the training set is constructed from untrimmed videos. The dataset will be publicly available after masking the faces.

Comparisons between different image-based visible-infrared person re-identification datasets are shown in Table I. Although the size of SYSU-MM02 is medium among existing public datasets, its distinctive characteristic is the untrimmed training set that complements existing datasets to facilitate research on real-world unsupervised cross-modality Re-ID.

A. Dataset Description

The dataset was captured by two near infrared cameras and one visible camera located in a neighborhood of a campus, as shown in Figure 4.

Our dataset SYSU-MM02 consists of 9,800 visible images and 15,974 near infrared images, which are captured from one visible camera (camera 1) and two near infrared cameras (camera 2 and camera 3). Visible camera resolution is 1920×1080 and infrared camera resolution is 704×576. Camera 1 is located at the gate of a building with lighting. Camera 2 and 3 are located at dark outdoor passages. There are lighting variations between visible images of camera 1 and contrast variations between infrared images of camera 2 and 3.

We capture videos by three cameras at night in three different days. The duration of each video is about 30 minutes in average. The time of videos is synchronized among three cameras. A majority of pedestrians come from a building and pass camera 1, 2 and 3 in order. There are multiple potential paths indicated by the dotted arrows, so that some pedestrians captured by camera 1 may not pass camera 2 and 3. In such a real-world situation, training data becomes more noisy than the unlabeled training data simulated on existing labeled datasets.

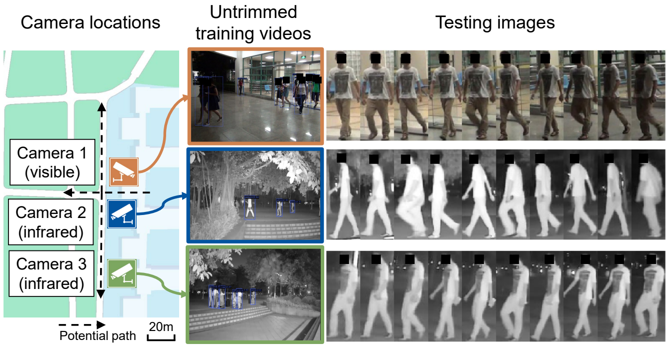


Fig. 4. Examples of our visible-infrared pedestrian dataset SYSU-MM02. Images are captured from one visible camera (camera 1) and two near infrared cameras (camera 2 and 3). Training images are automatically detected from untrimmed videos. The shown testing images are of the same identity. The dotted arrows indicate that there are multiple potential paths for pedestrians.

B. Data Processing

We use untrimmed videos in one day as training data and videos in the other two days for testing data annotation.

Training Set. The training set contains 7,440 visible images and 13,614 near infrared images. The pedestrian images are detected in one frame per second from untrimmed training videos by an off-the-shelf object detector YOLOX-x [69] trained on COCO [70]. Only images with areas over 1200 pixels and confidence score over 0.6 are selected.

Testing Set. The images in the testing set are detected and then annotated by human operators. The testing set contains 2,360 visible images and 2,360 near infrared images of 118 identities. For each identity in each camera, 10 images are annotated.

C. Evaluation Protocol

In training, the unlabeled images in the training set are used for unsupervised learning. In testing, we have two matching modes "visible to infrared" and "infrared to visible". The "visible to infrared" mode uses all visible images in camera 1 as query images and uses all near infrared images in camera 2 and camera 3 as gallery images; the "infrared to visible" mode exchanges the query images and gallery images as compared with the "visible to infrared" mode. Evaluation metrics are rank-k accuracy and mean average precision (mAP).

VII. Experiments

For unsupervised visible-infrared cross-modality Re-ID, we conducted extensive comparative evaluations of our method against the state-of-the-art unsupervised cross-modality Re-ID and unsupervised domain adaptation (UDA) methods on benchmark datasets SYSU-MM01 [23] and RegDB [67] as well as our newly collected SYSU-MM02. Furthermore, ablation study and hyper-parameter analysis were conducted. Our two-stage full model is denoted by CMAM-UDA.

A. Experiment Settings

The training process consists of RGB-gray asymmetric mutual learning stage and unsupervised cross-modality self-training stage. In the RGB-gray asymmetric mutual learning stage, we used all samples of Market-1501 [71] without label as the auxiliary training set of the source domain. In the unsupervised cross-modality self-training stage, we used the training set of the multi-modality benchmark datasets without label as the target domain. The testing sets and evaluation protocols were kept the same with the original datasets.

The datasets for evaluation are introduced as follows:

SYSU-MM01. This dataset contains 30,071 visible images captured by 4 visible cameras and 15,792 near infrared images captured by 2 infrared cameras. There are totally 491 identities, in which 395 identities are for training and 96 identities are for testing. There are two matching modes “all-search” and “indoor-search” for matching between all cameras and matching between indoor cameras, respectively.

RegDB. This dataset contains 8,240 images of 412 identities captured by a visible camera and a thermal camera. For each identity, there are 10 visible images and 10 thermal images. We followed the evaluation protocol of Ye et al. [26]. The 412 identities are split half and half for training and testing. There are two matching modes “VIS to TIR” and “TIR to VIS”, where “TIR” denotes thermal infrared images.

SYSU-MM02. To our best knowledge, our SYSU-MM02 is the first visible-infrared pedestrian dataset that automatically detects unlabeled training data from untrimmed videos, which is more realistic for evaluating unsupervised learning. As described in Section VI, two matching modes are denoted by “VIS to NIR” and “NIR to VIS”. For example, “VIS to NIR” denotes using visible (VIS) images as query and near infrared (NIR) images as gallery.

B. Implementation Details

1) Conditional Diffusion Model: We applied the model architecture and training strategy of EDM [33]. The attention resolutions were set 32, 16, 8. Residual blocks per resolution was set 1. The T2I-adapter network [62] for edge map was adopted for adding condition to the diffusion model. The training infrared images in the target domain were resized to 192×64. The training stage consisted of 50,000 iterations. The batch size was set 32 and the learning rate was set 0.0001 for all iterations. For image generation, the RGB image was converted to edge map as condition of the diffusion model. Then, the Heun solver [72] of 20 sampling steps was used for generating fake infrared images.

2) Backbone Model: We applied ResNet-50 [60] as backbone for RGB-gray feature extractor M_{joint} and gray feature extractor M_{gray}. The input image was resized to 384×128. In the last pooling layer, the original global average pooling was replaced by generalized mean pooling [73] and its output was used as feature. We used circle head [63] as classifier, of which the relaxation factor was set 0.35 and the scale factor was set 64. We applied label smoothing for classification probabilities and set smoothing parameter as 0.1.

3) Identification Loss: The identification loss $\mathcal{L}_{PID-jA1}$ in Eq. (2) is formulated as

$$\begin{aligned}\mathcal{L}_{PID-jA1} = &- \sum_{i=1}^{N_{RGBA}} \sum_{id=1}^{N_{RGBA}^{PID}} \hat{y}_{i,id}^{jA1} \log(p_{i,id}^{jA1}) + \\ &\sum_{(a,p,n)\in\mathcal{I}_{tri}} smargin(dist(\mathbf{f}_a^{jA1}, \mathbf{f}_p^{jA1}), dist(\mathbf{f}_a^{jA1}, \mathbf{f}_n^{jA1})),\end{aligned} \tag{21}$$

where the first term is cross entropy loss $\mathcal{L}_{ce-jA1}$ and the second term is triplet loss $\mathcal{L}_{tri-jA1}$. In the cross entropy loss, $\hat{y}_{i,id}^{RGBA}$ is the id-th element of the pseudo one-hot label $\hat{\mathbf{y}}_i^{RGBA} \in \mathbb{R}^{N_{RGBA}^{PID}}$ and $p_{i,id}^{jA1}$ is the id-th element of the classification probability vector $\mathbf{p}_i^{jA1}$. In the triplet loss, $dist(\cdot,\cdot)$ denotes Euclidean distance; $smargin(a,b) = \log(1+\exp(a-b))$ is the soft margin loss [64] for softening the margin between positive pairs and negative pairs; $\mathcal{I}_{tri}$ is the index set of triplets; (a,p,n) denotes the indices of anchor sample, positive sample and negative sample, which are selected by hard sample mining [64] to form a batch in training. The other identification losses are formulated similarly.

4) Training Strategy: In the first stage, the models were initialized by ImageNet [74] pretraining, which was also applied by the compared methods in our experiments. In the second stage, the models were initialized by the first stage. Following FastReID [75], we applied random flip and random erasing for data augmentation. The probability was set 0.5 for both random flip and random erasing. For random erasing, the minimum proportion, maximum proportion and minimum aspect ratio were set 0.02, 0.4, 0.3, respectively. Color jitter was also applied by setting the brightness range as [0.8, 1.2] and the contrast range as [0.85, 1.15].

In the loss $\mathcal{L}_{pre}^{CMAM}$ in Eq. (15) for RGB-gray asymmetric mutual learning stage, λ_{PP} was set 0.1 and λ_{SP} was set 0.5. In the loss $\mathcal{L}_{uda}$ in Eq. (20) for unsupervised cross-modality self-training, λ_{BPP} was set 0.01 and λ_{BSP} was set 0.05. In cross-modality pseudo labeling, k-means clustering [76] was applied. The number of cluster k was 400 for each modality.

For optimization, we adopted the ADAM optimizer [77]. The batch sizes used for RGB-gray asymmetric mutual learning stage and the unsupervised cross-modality self-training stage were 64 and 128, respectively. The optimization process of RGB-gray asymmetric mutual learning stage consisted of 30,000 iterations. The first 15,000 iterations was the first phase and the second 15,000 iterations was the second phase. The selective probability preservation loss $\mathcal{L}_{PP}^{sel}$ was not applied in the first phase. In the first 2,000 iterations of each phase, warmup strategy [78] was applied to increase the learning rate from 3.5×10^{-6} to 3.5×10^{-4}. In the last 6,000 iterations of each phase, Cosine annealing [79] was used for decreasing the learning rate. The optimization process of the unsupervised cross-modality self-training stage consisted of 12,000 iterations. The momentum for updating feature memory of classifiers was set 0.2. In the first 2,000 iterations, learning rate

TABLE II

COMPARISONS WITH THE STATE-OF-THE-ART UNSUPERVISED LEARNING METHODS. ROW 1 IS HAND-CRAFTED FEATURE; ROWS 2-3 CONTAIN SINGLE-MODALITY RE-ID METHODS; ROWS 4-5 CONTAIN MUTUAL LEARNING METHODS; ROWS 6-11 CONTAIN UNSUPERVISED CROSS-MODALITY RE-ID METHODS. "R-K" DENOTES THE RANK-K ACCURACY (%) AND MAP DENOTES THE MEAN AVERAGE PRECISION (%). THE BEST RESULTS ARE IN **BOLD**.

Method	SYSU-MM01						RegDB						SYSU-MM02					
	all-search			indoor-search			VIS to TIR			TIR to VIS			VIS to NIR			NIR to VIS		
	R-1	R-10	mAP	R-1	R-10	mAP	R-1	R-10	mAP	R-1	R-10	mAP	R-1	R-10	mAP	R-1	R-10	mAP
HOG [80]	3.7	19.8	5.0	3.3	24.7	9.1	52.9	71.1	31.0	57.5	80.4	33.9	5.7	25.1	2.4	4.0	20.6	2.9
GLT [12]	15.7	48.6	15.4	18.0	56.0	26.0	12.8	37.6	13.4	12.9	40.0	12.7	10.5	43.0	9.2	7.8	37.8	10.0
ICE [13]	23.3	60.5	21.9	26.6	69.3	35.0	11.6	40.3	14.3	14.5	43.8	15.3	21.5	56.2	15.9	17.8	57.1	17.9
MMT [11]	17.0	55.1	17.5	22.5	68.9	32.2	31.4	69.3	26.5	33.8	71.4	27.8	10.8	39.2	8.2	9.0	33.3	8.3
DML [58]	51.2	84.7	50.0	61.1	90.6	67.0	80.8	97.3	62.4	82.5	97.8	64.3	44.2	82.7	33.4	45.0	85.3	33.8
H2H [27]	30.2	65.9	29.4	-	-	-	23.8	45.3	18.9	14.1	31.9	12.3	-	-	-	-	-	-
OTLA [28]	29.9	-	27.1	29.8	-	38.8	32.9	-	29.7	32.1	-	28.6	-	-	-	-	-	-
ADCA [29]	45.5	85.3	42.7	50.6	89.7	59.1	67.2	82.0	64.1	68.5	83.2	63.8	-	-	-	-	-	-
CHCR [32]	47.7	87.3	45.3	50.1	90.8	42.2	68.2	81.5	63.8	69.1	83.7	64.0	-	-	-	-	-	-
PGM [30]	57.3	92.5	51.8	56.2	90.2	62.7	69.5	-	65.5	69.9	-	65.2	-	-	-	-	-	-
GUR* [31]	61.0	-	57.0	64.2	-	69.5	73.9	-	70.2	75.0	-	69.9	42.2	81.7	32.5	44.0	84.4	32.9
CMAM-UDA	**62.0**	**93.4**	**58.2**	**67.6**	**95.9**	**72.7**	**89.1**	**98.2**	**74.0**	**89.0**	**98.6**	**74.0**	**51.4**	**90.3**	**41.0**	**56.1**	**89.4**	**40.3**

was increased from 10^{-6} to 10^{-4} by warmup strategy [78]. Cosine annealing [79] was used in the last 10,000 iterations.

C. Compared Methods

Unsupervised Cross-Modality Re-ID. We compared with the state-of-the-art unsupervised visible-infrared Re-ID methods H2H [27], OTLA [28], ADCA [29], CHCR [32], PGM [30] and GUR [31]. The full model of GUR [31] utilize camera labels for intra-camera pseudo label learning and requires multiple cameras for each modality. Since infrared images on RegDB [67] and visible images on SYSU-MM02 are captured from only one camera and the other compared methods did not use camera label, we compared with the version of GUR without using camera label, which is denoted by GUR*.

Unsupervised Single-Modality Re-ID. We compared with some advanced unsupervised Re-ID methods MMT [11], GLT [12] and ICE [13]. MMT [11] requires two models for mutual teaching in training, which is technically related to our method.

Mutual Learning. Deep mutual learning (DML) [58] is a related work that takes the same images as input for two models and minimizes the difference of two predictions. MMT [11] also applied mutual learning to refine pseudo labels.

Hand-Crafted Feature. As some hand-crafted features can describe body shape without training, we compared with a representative descriptor HOG [80].

Implementation of the Compared Methods. For MMT [11], GLT [12] and ICE [13], we used the their released codes and adapted them for cross-modality Re-ID by converting the input visible images to grayscale images following channel exchanged augmentation [43]. For implementation of DML [58], we used the same backbone and unsupervised domain adaptation strategy in our method. For H2H [27], OTLA [28], ADCA [29], CHCR [32] and PGM [30], the results in their papers were used. For the best compared method GUR [31], we used the reported results for SYSU-MM01 [23], RegDB [67] and implemented GUR on SYSU-MM02.

D. Model Comparison and Analysis

Experiment results in Table II show that our method achieves the best performance and significantly outperforms the compared methods on SYSU-MM02 and RegDB.

These compared methods cannot learn visible-infrared shared prior knowledge from auxiliary unlabeled visible images as our method. Such prior knowledge facilitates cross-modality matching in the target domain.

E. Further Evaluations

1) Ablation Study: We carried out ablation study on SYSU-MM01 [23] for each component in the RGB-gray asymmetric mutual learning stage and the unsupervised cross-modality self-training stage, which are denoted by "RGB-gray mutual learning" and "cross-modality UDA", respectively. The notations of components and results are shown in Table III.

Comparison with Single-Model UDA. Rows 1-2 in Table III are results of using single model for unsupervised domain adaptation (UDA) by using identification loss $\mathcal{L}_{PID}$ in Eq. (17). "M_{joint}" and "M_{gray}" denote RGB-gray feature extractor and gray feature extractor trained separately by the identification loss $\mathcal{L}_{PID-pre}$ in Eq. (15).

Compared with "M_{joint}" and "M_{gray}", our full model "CMAM-UDA" shows improvement of over 10% rank-1 accuracy, which indicates the effectiveness of mutual learning.

Effect of $\mathcal{L}_{SP}^{sel}$ and $\mathcal{L}_{PP}^{sel}$ in RGB-gray Mutual Learning. In the RGB-gray asymmetric mutual learning stage, the loss function $\mathcal{L}_{pre}^{CMAM}$ in Eq. (15) consists of the selective similarity preservation loss $\mathcal{L}_{SP}^{sel}$ in Eq. (14) and the selective probability preservation loss $\mathcal{L}_{PP}^{sel}$ in Eq. (13). As shown by the comparison between "PRE1" and "CMAM-UDA", $\mathcal{L}_{SP}^{sel}$ and $\mathcal{L}_{PP}^{sel}$ significantly improve the performance, which indicates the effectiveness of asymmetric mutual learning. By comparing "PRE1", "PRE2" and "CMAM-UDA", we observe that $\mathcal{L}_{PP}^{sel}$ and $\mathcal{L}_{SP}^{sel}$ complement each other.

Effect of $\mathcal{L}_{BSP}$ and $\mathcal{L}_{BPP}$ in UDA. In cross-modality unsupervised domain adaptation stage, the bi-directional similarity preservation loss $\mathcal{L}_{BSP}$ in Eq. (19) and the bi-directional probability preservation loss $\mathcal{L}_{BPP}$ in Eq. (18) are minimized for fusing the complementary knowledge of two models. Comparisons of "UDA1", "UDA2" and "CMAM-UDA" verifies the effectiveness of $\mathcal{L}_{BSP}$ and $\mathcal{L}_{BPP}$ for knowledge fusion.

TABLE III
ABLATION STUDY OF OUR FRAMEWORK CMAM-UDA ON SYSU-MM01. “M_{joint}” DENOTES RGB-GRAY FEATURE EXTRACTOR. “M_{gray}” DENOTES GRAY FEATURE EXTRACTOR. “BGM” DENOTES BIPARTITE GRAPH MATCHING FOR CROSS-MODALITY BIPARTITE PSEUDO LABELING.

Model	RGB-Gray Mutual Learning				Cross-Modality UDA					all-search				indoor-search			
	M_{joint}	M_{gray}	$\mathcal{L}_{SP}^{sel}$	$\mathcal{L}_{PP}^{sel}$	M_{joint}	M_{gray}	$\mathcal{L}_{BSP}$	$\mathcal{L}_{BPP}$	BGM	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
M_{joint}	✓				✓				✓	47.5	85.3	92.9	47.3	56.5	93.0	96.7	64.0
M_{gray}		✓				✓			✓	51.5	89.3	95.7	49.2	59.3	91.5	96.9	65.0
PRE1	✓	✓			✓	✓	✓	✓	✓	57.5	91.6	97.0	54.7	64.1	94.7	98.3	69.9
PRE2	✓	✓	✓		✓	✓	✓	✓	✓	59.3	93.1	97.7	56.3	65.5	95.0	98.1	70.8
UDA1	✓	✓	✓	✓	✓	✓			✓	52.9	88.9	95.6	51.8	59.7	92.1	96.9	66.6
UDA2	✓	✓	✓	✓	✓	✓	✓		✓	60.3	92.9	97.5	57.0	67.3	95.5	98.2	71.7
UDA3	✓	✓	✓	✓	✓	✓	✓	✓		42.7	81.4	89.1	42.8	51.4	87.9	93.1	59.0
CMAM-UDA	✓	✓	✓	✓	✓	✓	✓	✓	✓	62.0	93.4	97.8	58.2	67.6	95.9	98.4	72.7

TABLE IV
PERFORMANCES (%) OF VARIANTS OF MUTUAL LEARNING FOR RGB-GRAY ASYMMETRIC MUTUAL LEARNING ON SYSU-MM01. “M_{joint}” AND “M_{gray}” DENOTE SINGLE-MODEL TRAINING. “RGB-RGB”, “GRAY-GRAY” AND “RGB-GRAY” DENOTE SYMMETRIC MUTUAL LEARNING BY BI-DIRECTIONAL KNOWLEDGE TRANSFER BETWEEN TWO TYPES OF IMAGES.

Model	all-search				indoor-search			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
M_{joint}	47.5	85.3	92.9	47.3	56.5	93.0	96.7	64.0
M_{gray}	51.5	89.3	95.7	49.2	59.3	91.5	96.9	65.0
RGB-RGB	51.2	84.7	92.5	50.0	61.1	90.6	95.7	67.0
GRAY-GRAY	52.9	89.0	95.7	50.3	57.3	90.8	96.4	63.8
RGB-GRAY	59.1	92.3	97.4	56.1	65.5	94.4	98.4	70.2
CMAM-UDA	62.0	93.4	97.8	58.2	67.6	95.9	98.4	72.7

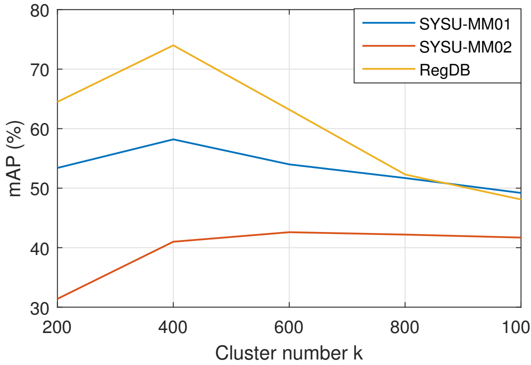


Fig. 5. Performances of using different cluster number k.

Effect of Bipartite Graph Matching (BGM) for Pseudo Labeling. In cross-modality bipartite pseudo labeling, bipartite graph matching (BGM) in Eq. (16) is for separating intra-modality and cross-modality sample association to alleviate modality gap. We compared with pseudo labeling by k-means clustering [76] regardless of modality of the samples. The cluster number $k = 400$ is the same as that in our method. We denote this method by “UDA3” in Table III. Compared with “UDA3”, applying BGM in our full model “CMAM-UDA” can bring significant improvement of over 10% mAP, which shows that modality gap can be alleviated better with this pseudo labeling strategy.

2) Variants of Mutual Learning: To justify the effectiveness of the RGB-gray asymmetric mutual learning, we compared with some variants of mutual learning on SYSU-MM01 [23]. The results are reported in Table IV. Rows 1-2 in Table IV are results of using single model for training. Rows 3-5 are results of variants of mutual learning, in which “X1-X2” denotes symmetric mutual learning using bi-directional knowledge transfer with two types of inputs X1 and X2. “RGB” denotes visible images and “GRAY” denotes the transformed grayscale images as illustrated in Section IV-A.

Asymmetric Mutual Learning v.s. Symmetric Mutual Learning. To show the effectiveness of asymmetric mutual learning in our CMAM, we compared with “RGB-GRAY” symmetric mutual learning, which replaced selective feature relation preservation losses $\mathcal{L}_{SP}^{sel}$, $\mathcal{L}_{PP}^{sel}$ by bi-directional feature relation preservation losses $\mathcal{L}_{BSP}$, $\mathcal{L}_{BPP}$ when training on the auxiliary unlabeled visible images. Our method outperformed the “RGB-GRAY” symmetric mutual learning, because CMAM can learn RGB-infrared modality-shared prior knowledge better by extracting both RGB-gray view-shared features and RGB-specific features, while “RGB-GRAY” mutual learning only extracts RGB-gray view-shared features and eliminates the useful RGB-specific features.

Mutual Learning Using the Same Inputs. “RGB-RGB” and “GRAY-GRAY” used the same input for symmetric mutual learning following deep mutual learning (DML) [58]. our method CMAM-UDA and “RGB-GRAY” mutual learning and can benefit from complementary knowledge learned from different input images and thus achieves better performance.

3) Impact of Cluster Number: For unsupervised domain adaptation, pseudo labeling is an important factor. The cluster number k is a significant hyperparameter for k-means [76] used in our method. The default value of cluster number is $k = 400$. We varied k from 200 to 1000 and evaluated our method on SYSU-MM01, RegDB and SYSU-MM02. The mAP of visible to infrared matching is shown in Figure 5. Generally, our method achieves better performances on all datasets when $k = 400$. When $k \in [200, 1000]$, our method is comparable to the state-of-the-art methods on SYSU-MM01 and SYSU-MM02, and performance variation is lower than 10 % mAP.

4) Impact of Fake Infrared Image Quality: As our method learns modality-shared prior knowledge from generated fake infrared images, image quality is an important factor. We simulated images from low quality to high quality by setting sampling step of diffusion model from 5 to 25. The default value is 20. The performances are reported in Table V. Examples of generated images are shown in Figure 6.

Our method can tolerate image quality variation to some extent. Using sampling step 25 is comparable to using default sampling step 20. When image quality is slightly degraded by using sampling step 15, the performance is also comparable

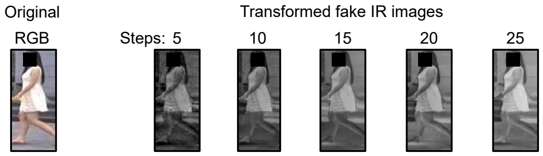


Fig. 6. Fake infrared images generated using different sampling steps.

TABLE V
PERFORMANCES (%) OF USING DIFFERENT SAMPLING STEPS FOR FAKE INFRARED IMAGE TRANSFORMATION ON SYSU-MM01.

Sampling steps	5	10	15	20 (default)	25
R-1	57.10	58.80	60.99	62.00	61.92
mAP	54.24	55.58	57.47	58.20	58.05

to that of using default sampling step 20. When the image quality is further degraded by using sampling steps 10 and 5, the performances become lower, but they are still better than the baseline method of RGB-RGB mutual learning by DML [58] (Rank-1=51.2%, mAP=50.0%) in Table II, which show the effectiveness of the generated images.

5) Increasing Labeled Data for Training: We varied the proportion of labeled identities from 0 to 50% on a target dataset SYSU-MM01 [23] to evaluate our method. Based on the models learned on unlabeled data by our method CMAM-UDA, we further finetuned our models by using labeled samples instead of pseudo labeled samples in unsupervised cross-modality self-training. We compared our asymmetric mutual learning method CMAM-UDA with a baseline deep mutual learning method (DML) [58] using the same finetuning strategy. Moreover, we compared with the state-of-the-art supervised cross-modality Re-ID method CAJ+ [39] for training model on limited labeled identities. The comparative evaluation results are presented in Table VI.

Our method can be further improved by using more labeled identities. Compared with DML [58], the performance gain of our method is remarkable for different proportions. Our method significantly outperforms the supervised method CAJ+ [39] when the labeled ID proportion is from 10% to 40%. The modality-shared prior knowledge learned by our method is effective when labeled data is limited.

6) Running Time Analysis: We evaluated the running time of each stage in our proposed framework using 1 Nvidia RTX 3090 GPU (24 GB) on a server with Intel(R) Xeon(R) Gold 6226R CPU and 256GB RAM. The source dataset was Market-1501 [71] and the target dataset was SYSU-MM01 [23]. The running time information is reported in Table VII.

The most time-consuming stage is transforming images in auxiliary dataset to fake infrared images by conditional diffusion model. Note that, only one GPU was used. To accelerate image transformation, we can use more GPUs to increase training batch size and generate images in parallel. Although our CMAM method and unsupervised cross-modality self-training use two ResNet-50 [60] models, only one model is used for inference. The inference speed 1.6ms/image can guarantee real-time performance in large-scale applications.

TABLE VI
PERFORMANCES (%) OF USING DIFFERENT PROPORTION OF LABELED IDENTITIES ON SYSU-MM01.

Labeled ID proportion		0	10%	20%	30%	40%	50%
CMAM-UDA	R-1	62.0	62.2	63.9	66.2	67.2	67.5
(ours, semi)	mAP	58.2	58.4	59.7	61.8	63.3	63.6
DML [58]	R-1	51.2	53.0	55.8	57.8	59.4	59.8
(semi)	mAP	50.0	51.4	54.0	56.0	57.3	57.8
CAJ+ [39]	R-1	-	13.2	23.0	37.8	60.0	64.7
(supervised)	mAP	-	16.0	24.9	35.1	55.0	61.4

TABLE VII
RUNNING TIME ANALYSIS ON 1 NVIDIA RTX 3090. "ITER" DENOTES ITERATION. "IMG" DENOTES IMAGE.

Stage	Image transformation		CMAM	Self-training		Re-ID
	Training	Inference	Training	Training	K-means	Inference
Speed	1.16s/iter	0.47s/img	0.60s/iter	0.50s/iter	144s/time	1.6ms/img
Quantity	50k iters	67k imgs	30k iters	12k iters	46 times	10k imgs
Time	16.17h	8.63h	5.03h	1.65h	1.84h	17s

VIII. CONCLUSION

In this work, we study unsupervised transferable visible-infrared cross-modality person re-identification. To overcome the modality gap for discovering unlabeled sample relations, we exploit auxiliary unlabeled visible images in a source domain to learn modality-shared prior knowledge and transfer it for unsupervised cross-modality self-training in the target domain. To achieve this, we propose Cross-Modality Asymmetric Mutual Learning (CMAM) for learning both RGB-gray view-shared information and RGB-specific information from pseudo labeled auxiliary visible images and the transformed fake infrared grayscale images in the source domain, which is developed based on our information theoretic analysis. The learned prior knowledge is further exploited for cross-modality self training in the target domain by bi-directional mutual learning for fusing complementary knowledge of RGB-gray matching and gray-gray matching. For evaluating unsupervised cross-modality Re-ID in a realistic scenario, we collected a new visible-infrared pedestrian dataset SYSU-MM02 from untrimmed videos. Our method achieved the state-of-the-art results in unsupervised setting on three benchmark datasets.

REFERENCES

[1] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 480–496.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 6, pp. 2872–2893, 2022.

[3] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 15 013–15 022.

[4] H. Gu, J. Li, G. Fu, C. Wong, X. Chen, and J. Zhu, "Autoloss-gms: Searching generalized margin-based softmax loss function for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4744–4753.

[5] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, pp. 1–18, 2018.

[6] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 172–188.

[7] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 2148–2157.

[8] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 3633–3642.

[9] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 4, pp. 956–973, Apr. 2020.

[10] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2021, pp. 2764–2772.

[11] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *International Conference on Learning Representations (ICLR)*, Apr. 2020.

[12] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha, "Group-aware label transfer for domain adaptive person re-identification." in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 5310–5319.

[13] H. Chen, B. Lagadec, and F. Bremond, "Ice: Inter-instance contrastive encoding for unsupervised person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 14 960–14 969.

[14] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 7308–7318.

[15] G. Lee, S. Lee, D. Kim, Y. Shin, Y. Yoon, and B. Ham, "Camera-driven representation learning for unsupervised domain adaptive person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 453–11 462.

[16] A. Wu, W. S. Zheng, H. X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 5390–5399.

[17] J. Feng, A. Wu, and W.-S. Zheng, "Shape-erased feature learning for visible-infrared person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 752–22 761.

[18] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 19 366–19 375.

[19] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 12 046–12 055.

[20] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji, "Discover cross-modality nuances for visible-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov. 2021, pp. 4330–4339.

[21] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *IEEE International Conference on Computer Vision (ICCV)*, Nov. 2021, pp. 16 403–16 412.

[22] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 823–11 832.

[23] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "Rgb-ir person re-identification by cross-modality similarity preservation," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 6, pp. 1765–1785, 2020.

[24] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 579–590, 2020.

[25] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, Jul. 2018, pp. 677–683.

[26] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Association for Advancement of Artificial Intelligence (AAAI)*, Feb. 2018, pp. 7501–7508.

[27] W. Liang, G. Wang, J. Lai, and X. Xie, "Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 6392–6407, 2021.

[28] J. Wang, Z. Zhang, M. Chen, Y. Zhang, C. Wang, B. Sheng, Y. Qu, and Y. Xie, "Optimal transport for label-efficient visible-infrared person re-identification," in *European Conference on Computer Vision (ECCV)*, Oct. 2022, pp. 93–109.

[29] B. Yang, M. Ye, J. Chen, and Z. Wu, "Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification," in *ACM International Conference on Multimedia (ACM MM)*, Oct. 2022, p. 2843–2851.

[30] Z. Wu and M. Ye, "Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9548–9558.

[31] B. Yang, J. Chen, and M. Ye, "Towards grand unified representation learning for unsupervised visible-infrared person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 069–11 079.

[32] Z. Pang, C. Wang, L. Zhao, Y. Liu, and G. Sharma, "Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, pp. 1–1, 2023.

[33] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 26 565–26 577, 2022.

[34] X. Fang, Y. Yang, and Y. Fu, "Visible-infrared person re-identification via semantic alignment and affinity inference," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 270–11 279.

[35] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 2352–2364, 2022.

[36] S. Zhang, Y. Yang, P. Wang, G. Liang, X. Zhang, and Y. Zhang, "Attend to the difference: Cross-modality person re-identification via contrastive correlation," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 8861–8872, 2021.

[37] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *European Conference on Computer Vision (ECCV)*, Aug. 2020, pp. 229–247.

[38] Z. Wei, X. Yang, N. Wang, B. Song, and X. Gao, "Abp: Adaptive body partition model for visible infrared person re-identification," in *IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2020, pp. 1–6.

[39] M. Ye, Z. Wu, C. Chen, and B. Du, "Channel augmentation for visible-infrared re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 46, no. 4, pp. 2299–2315, 2024.

[40] M. Kim, S. Kim, J. Park, S. Park, and K. Sohn, "Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 621–18 632.

[41] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 3, pp. 1418–1430, 2022.

[42] Z. Huang, J. Liu, L. Li, K. Zheng, and Z.-J. Zha, "Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2022.

[43] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 13 567–13 576.

[44] H. Liu, S. Ma, D. Xia, and S. Li, "Sfanet: A spectrum-aware feature augmentation network for visible-infrared person reidentification," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, pp. 1–14, 2021.

[45] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 225–234.

[46] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4610–4617, Feb. 2020.

[47] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3623–3632.

[48] Z. Chai, Y. Ling, Z. Luo, D. Lin, M. Jiang, and S. Li, "Dual-stream transformer with distribution alignment for visible-infrared person re-

identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 33, no. 11, pp. 6764–6776, 2023.

[49] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 14 308–14 317.

[50] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2153–2162.

[51] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang, "Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 973–20 982.

[52] H. Li, M. Liu, Z. Hu, F. Nie, and Z. Yu, "Intermediary-guided bidirectional spatial–temporal aggregation network for video-based visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 33, no. 9, pp. 4962–4972, 2023.

[53] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 3, 2022, pp. 2633–2641.

[54] C. Zou, Z. Chen, Z. Cui, Y. Liu, and C. Zhang, "Discrepant and multi-instance proxies for unsupervised person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 058–11 068.

[55] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 232–242.

[56] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 2004–2013.

[57] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[58] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 4320–4328.

[59] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, and Y. Qu, "Dual mutual learning for cross-modality person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 8, pp. 5361–5373, 2022.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[61] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.

[62] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[63] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 6398–6407.

[64] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[65] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 1522–1531.

[66] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[67] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[68] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 4, Feb. 2021, pp. 3529–3537.

[69] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision (ECCV)*, Sep. 2014, pp. 740–755.

[71] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1116–1124.

[72] U. M. Ascher and L. R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, 1998.

[73] Y. Ge, F. Zhu, D. Chen, R. Zhao, and hongsheng Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 11 309–11 321.

[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009, pp. 248–255.

[75] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.

[76] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 1, pp. 281–297, 1967.

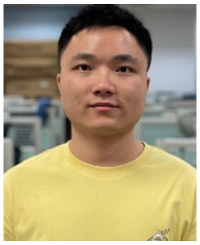
[77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, May 2014.

[78] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[79] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[80] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2005, pp. 886–893.

Ancong Wu received the bachelor's degree in intelligence science and technology from Sun Yat-sen University in 2015 and received Ph.D. degree in information and communication engineering from Sun Yat-sen University in 2020. He is now an associate researcher in Sun Yat-sen University. His research interests are computer vision algorithms and the applications for surveillance video analysis.



Chengzhi Lin is now a Master student in Sun Yat-sen University. He received the bachelor's degree in school of Computer Science and Engineering from Sun Yat-sen University in 2020. His research interests are deep learning algorithms and computer vision tasks such as person re-identification and text-video retrieval.

Wei-Shi Zheng received the PhD degree in applied mathematics from Sun Yat-sen University in 2008. He is currently a full Professor with Sun Yat-sen University. His research interests include person or object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He is an associate editor for the Pattern Recognition Journal and the area chair of a number of top conferences. He joined Microsoft Research Asia Young Faculty Visiting Programme. He was the recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China and a recipient of Royal Society-Newton Advanced Fellowship of United Kingdom.