# Augmented Dual-Contrastive Aggregation Learning for Unsupervised Visible-Infrared Person Re-Identification

Bin Yang
yangbin_cv@whu.edu.cn
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, China

Mang Ye*
yemang@whu.edu.cn
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, China

Jun Chen*
chenj@whu.edu.cn
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, China

Zesen Wu
zesenwu@whu.edu.cn
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, China

## ABSTRACT

Visible infrared person re-identification (VI-ReID) aims at searching out the corresponding infrared (visible) images from a gallery set captured by other spectrum cameras. Recent works mainly focus on supervised VI-ReID methods that require plenty of cross-modality (visible-infrared) identity labels which are more expensive than the annotations in single-modality person ReID. For the unsupervised learning visible infrared re-identification (USL-VI-ReID), the large cross-modality discrepancies lead to difficulties in generating reliable cross-modality labels and learning modality-invariant features without any annotations. To address this problem, we propose a novel Augmented Dual-Contrastive Aggregation (ADCA) learning framework. Specifically, a dual-path contrastive learning framework with two modality-specific memories is proposed to learn the intra-modality person representation. To associate positive cross-modality identities, we design a cross-modality memory aggregation module with count priority to select highly associated positive samples, and aggregate their corresponding memory features at the cluster level, ensuring that the optimization is explicitly concentrated on the modality-irrelevant perspective. Extensive experiments demonstrate that our proposed ADCA significantly outperforms existing unsupervised methods under various settings, and even surpasses some supervised counterparts, facilitating VI-ReID to real-world deployment. Code is available at https://github.com/yangbincv/ADCA.

## CCS CONCEPTS

• **Information systems → Information retrieval**.

[*] Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*MM '22, October 10–14, 2022, Lisboa, Portugal*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548198

## KEYWORDS

person re-identification, unsupervised learning, visible-infrared, cross-modality

**ACM Reference Format:**
Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. 2022. Augmented Dual-Contrastive Aggregation Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548198

## 1 INTRODUCTION

Person re-identification (ReID) targets at matching the same person across different cameras [14, 17]. In recent years, ReID has attracted increasing interest due to its importance in intelligent video surveillance applications [21, 36, 48]. However, most existing techniques consider images of people collected by visible cameras in the daytime, which cannot capture enough information about a person under poor lighting conditions, limiting the applicability of a single-modality ReID in practical surveillance [22, 38]. To address this problem, the cross-modality visible-infrared person re-identification (VI-ReID) is proposed to match images of people captured by visible and infrared (including near- [41] and far-infrared (thermal) [28] ) cameras. A few deep-learning methods have made initial attempts at VI-ReID to learn modality-sharable feature representations [11, 18, 40, 43, 46]. However, training VI-ReID models requires a large number of cross-modality (visible-infrared) identity annotations, which are more expensive than those in single-modality ReID, making supervised VI-ReID methods less scalable in real-world deployments.

With the above motivations, this paper studies the challenging unsupervised learning visible-infrared ReID task, named USL-VI-ReID, which aims at learning modality-invariant knowledge from the unlabeled visible-infrared dataset and matching the same person captured by visible and infrared cameras. Different from the existing unsupervised ReID work like H2H [24], our work does not use any labeled datasets for pre-training, while achieving higher accuracy.

State-of-the-art methods address the unsupervised single modality ReID (USL-ReID) problem with the contrastive learning [2, 16, 33, 35] . Specifically, they construct a memory module that contains the features of all identities. A pseudo ID is assigned to each training

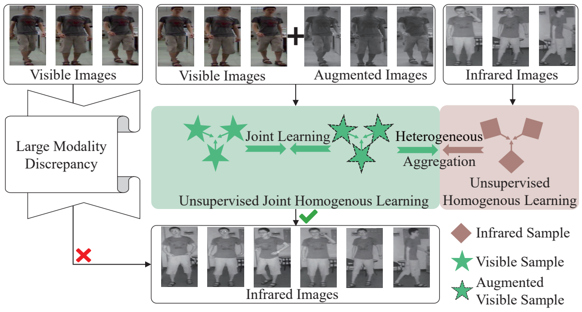

**Figure 1: Illustration of our main idea for USL-VI-ReID. The pentagram represents the visible sample and the pentagram with a dashed line denotes the augmented visible sample. The cube indicates the infrared sample. The large modality discrepancy hinders the retrieval of the same person across different modalities. Unsupervised joint homogenous learning and heterogeneous aggregation build a bridge between visible and infrared modalities without any annotations.**

image by a clustering algorithm. During training, the contrastive loss between training data features and memory module is minimized to train the feature extractor to learn a good representation. Such the pseudo-label-based USL pipeline has yielded good results on the single-modality USL ReID. However, for USL-VI-ReID, the differences between the visible and infrared cameras result in large cross-modality discrepancies, leading to difficulties in generating positive cross-modality labels. This undoubtedly results in limited retrieval performance for the same person across two modalities.

To handle the challenges in USL-VI-ReID, we propose a solution of unsupervised homogenous joint learning and heterogeneous aggregation. The rationale is that we can carry out unsupervised homogeneous learning to capture intra-modality features and jointly use the powerful color augmentation method like Channel Augmentation (CA) [46] to bridge the gap between visible modality and infrared modality, and at the same time take into account the heterogeneous aggregation of positive sample pairs across modalities for cross-modality label association, as shown in Figure 1.

Based on the above insights, we take inspiration from previous two-stream supervised VI-ReID methods to propose a novel Augmented Dual-Contrastive Aggregation (ADCA) learning framework. It builds a two-stream contrastive learning framework with two modality-specific memories to learn the intra-modality person representation. To guarantee the model modality generalizability, the visible stream incorporates the color augmentation method of random channel augmentation [46] for jointly contrastive learning. To handle the inter-modality variations, a Cross-modality Memory Aggregation (CMA) module is designed to select highly associated positive samples and aggregate their corresponding memory features for cross-modality contrastive learning. The CMA ensures that the optimization is explicitly concentrated on the modality-irrelevant perspective. The major advantage is that the CMA associates cross-modality information at the memory cluster level, enabling a tight coupling of the cross-modality label association for contrastive learning. This formulates a mutual reinforcement and efficient solution compared to the sample-to-sample association.

The main contributions can be summarized as follows:

- We propose a dual-stream contrastive learning framework with two modality-specific memory modules for USL-VI-ReID. To learn color-invariant features, the visible stream employs a powerful color augmentation method of random channel augmentation as a bridge to infrared modality for joint contrastive learning.
- We design a Cross-modality Memory Aggregation (CMA) module to select reliable positive samples and aggregate corresponding memory representations in a parameter-free manner, which enables the dual-stream framework to learn better modality-invariant knowledge, while simultaneously reinforcing each contrastive learning stream.
- We present extensive experiments on the SYSU-MM01 and RegDB datasets, which demonstrate that our method outperforms existing unsupervised methods under various settings, and even surpasses some supervised counterparts, providing a new baseline for USL-VI-ReID task and significantly pushing VI-ReID to real-world deployment.

# 2 RELATED WORK

## 2.1 Supervised Visible-Infrared Person ReID

Visible-infrared person ReID (VI-ReID) is a challenging cross-modality person recognition problem. The most problematic technique is to reduce the large modality gap of intra-class person images between visible and infrared modalities [6, 18, 31, 43–45, 47, 49]. [41] started the first attempt by directly utilizing grayscale images for training and testing with a zero-padding one-stream network. [37] proposed a dual-level discrepancy modeling method to generate the cross-modality images with the advancement of GANs, eliminating discrepancy at the pixel level. A joint modality and pattern alignment network in [42] was introduced to discover cross-modality nuances in different patterns for visible-infrared person ReID. Additionally, a channel augmentation (CA) method was introduced in [46] to homogenously generate color-irrelevant images by randomly exchanging the color channels for VI-ReID.

The above supervised VI-ReID methods require a large amount of cross-modality identity annotations, which hinders fast deployment to new scenes. In this work, we investigate purely unsupervised visible-infrared Person ReID (USL-VI-ReID) task where no identity annotations are required, with important implications for real-world VI-ReID deployments.

## 2.2 Unsupervised Single-Modality Person ReID

Unsupervised single-modality person ReID can be roughly categorized into unsupervised domain adaptation (UDA) and fully unsupervised learning (USL). There are two categories of UDA methods: pseudo label-based methods [1, 10, 12, 15, 16, 30], and domain translation-based methods [3, 4, 39, 52, 54], where the former achieve state-of-the-art performance. All these UDA-based methods require a labeled source dataset. Fully unsupervised learning ReID (USL ReID) methods have a better flexibility for deployment [25, 26, 33]. CAP [35] designed a camera-aware proxies learning framework to split each single cluster into multiple proxies. Cluster Contrast [7] was introduced to compute contrast loss at the cluster level to maintain cluster consistency. ICE [2] leveraged

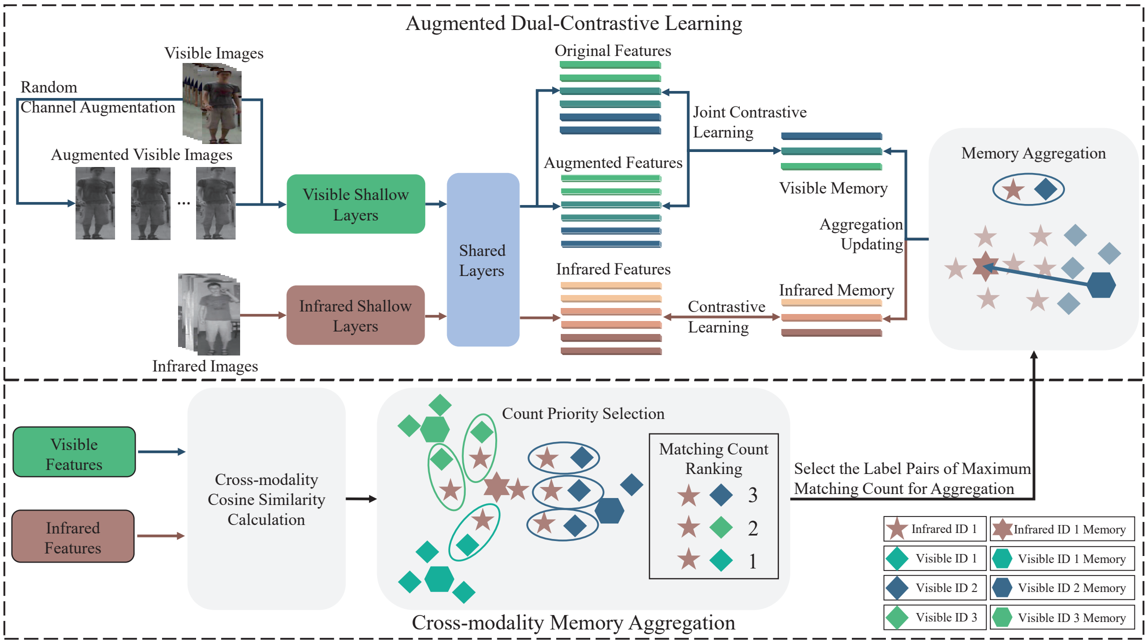

**Figure 2: Augmented Dual-Contrastive Learning Aggregation Framework. It comprises two components: Augmented Dual-Contrastive (ADC) learning and Cross-modality Memory Aggregation (CMA) module. ADC has a dual-path contrastive learning framework (one path for visible modality and the other for infrared modality). CMA selects the reliable positive cross-modality label pairs using the count priority selection strategy for aggregating the two modality-specific memories.**

inter-instance pairwise similarity scores to boost previous class-level contrastive ReID methods. However, the above approaches cannot learn effective modality-irrelevant knowledge against large modality discrepancy without any cross-modality annotation.

## 2.3 Unsupervised Visible-Infrared Person ReID

Unsupervised Learning Visible-Infrared Person ReID (USL-VI-ReID) problem has two challenges. First, different from single-modality person ReID, USL-VI-ReID has a large cross-modality discrepancy, which results in large intra-class variations and makes it difficult to generate reliable cross-modality labels. Second, there are no available cross-modality (visible-infrared) identity labels in USL-VI-ReID, leading to the difficulty in directly learning modality-invariant feature representations. H2H [24] started the first attempt by designing a two-stage method to solve the USL-VI-ReID task, including homogeneous learning and heterogeneous learning. However, H2H [24] used the Market-1501 dataset [51] as an extra labeled RGB dataset for pre-training, making the method less scalable in real-world deployments. No identity labels are available in our work, making the task more challenging in simultaneously learning discriminative and modality-invariant representations for cross-modality retrieval.

# 3 PROPOSED METHOD

## 3.1 Overview

We propose an Augmented Dual-Contrastive Aggregation (ADCA) learning framework for USL-VI-ReID, as shown in Figure 2. Our ADCA has two components including Augmented Dual-Contrastive (ADC) learning and Cross-modality Memory Aggregation (CMA) module. In ADC, two modality-specific shallow layers are utilized to capture modality-specific information for different modalities. Modality-shared layers focus on learning a multi-modality sharable space to bridge the gap between two heterogeneous modalities. Then, two modality-specific memories are constructed for mining inter- and intra-class information within each modality with contrastive learning. To ensure the modality generalization capability, we adopt random channel augmentation following [46] as an extra input to the visible stream for joint learning. To further explore the correlation between two modalities, we design a Cross-modality Memory Aggregation (CMA) module to select reliable positive cross-modality label pairs and aggregate corresponding memories, ensuring the discriminability of cross-modality features. These two modules reinforce each other. CMA module enables ADC framework to learn better modality-invariant features, while ADC's modality generalization capability allows CMA to better mine the knowledge of the association between two modalities.

## 3.2 Augmented Dual-Contrastive Learning Framework

The Augmented Dual-Contrastive (ADC) learning framework is shown at the top of Figure 2. The ADC has two input paths. Visible images and their augmented images obtained by random Channel Augmentation (CA) [46] are input into the visible path for learning visible-specific features while being able to learn color-invariant information. Infrared images are sent to the infrared path for obtaining infrared-specific features. The final embeddings are extracted by shared layers to learn modality-sharable information. Two modality-specific memories are updated by a momentum strategy. The parameters of feature extractors are updated with dual-contrastive learning.

To facilitate the description of our method, we first introduce the notations used in this paper. Let $X_i = \{x^i_1, x^i_2, \cdots, x^i_N\}$ denote the infrared images with $N$ instances. $X_v = \{x^v_1, x^v_2, \cdots, x^v_M\}$ and $X_{va} = \{x^{va}_1, x^{va}_2, \cdots, x^{va}_M\}$ indicate the visible and augmented visible training sets with $M$ instances, respectively. The $X_{va}$ follows the settings in CA [46]. $U_i = \{u^i_1, u^i_1, \cdots, u^i_N\}$ and $U_v = \{u^v_1, u^v_1, \cdots, u^v_M\}$ represent the corresponding features extracted by the infrared and visible feature extractor $f^i_\theta$ and $f^v_\theta$. $q_i$ and $q_v$ are query instance features extracted by $f^i_\theta$ and $f^v_\theta$, respectively.

**Modality-specific Memory Initialization.** The initialization of two modality-specific memory is illustrated in Figure 3. At the beginning of each training epoch, each cluster's representation $\{\phi^i_1, \ldots, \phi^i_K, \phi^v_1, \ldots, \phi^v_L\}$ of infrared and visible modalities are stored in infrared and visible memory dictionaries for modality-specific memory initialization, respectively. This process can be written as

$$\phi^i_k = \frac{1}{\left|\mathcal{H}^i_k\right|} \sum_{u^i_n \in \mathcal{H}^i_k} u^i_n, \tag{1}$$

$$\phi^v_l = \frac{1}{\left|\mathcal{H}^v_l\right|} \sum_{u^v_m \in \mathcal{H}^v_l} u^v_m, \tag{2}$$

where $\mathcal{H}^{i(v)}_k$ denotes the $k$-th cluster set in infrared or visible modality and $|\cdot|$ indicates the number of instances per cluster.

**Modality-specific Memory Updating.** During training, we sample $P$ person identities and $Z$ instances for each identity from each modality training set. Then, we obtain a total number of $3P \times Z$ query images including infrared, visible, and augmented visible person images in a batch. We update the two modality-specific memories by a momentum updating strategy:

$$\phi^{i(\delta)}_k \leftarrow \beta\phi^{i(\delta-1)}_k + (1-\beta)q_i, \tag{3}$$

$$\phi^{v(\delta)}_l \leftarrow \beta\phi^{v(\delta-1)}_l + (1-\beta)q_v, \tag{4}$$

$$\phi^{v(\delta)}_l \leftarrow \beta\phi^{v(\delta-1)}_l + (1-\beta)q_{va}, \tag{5}$$

where $q_{va}$ is the augmented query instance features. $\beta$ is the momentum updating factor. $\delta$ is the iteration number.

**Joint Learning Loss Function.** In each iteration, the modality-specific shallow layers and shared layers are jointly updated by three types of ClusterNCE [7] loss, including infrared, visible, and augmented visible loss by the following equations:

$$L_{q_i} = -\log \frac{\exp\left(q_i \cdot \phi^i_+/\tau\right)}{\sum_{k=0}^{K} \exp\left(q_i \cdot \phi^i_k/\tau\right)}, \tag{6}$$

$$L_{q_v} = -\log \frac{\exp\left(q_v \cdot \phi^v_+/\tau\right)}{\sum_{l=0}^{L} \exp\left(q_v \cdot \phi^v_k/\tau\right)}, \tag{7}$$

$$L_{q_{va}} = -\log \frac{\exp\left(q_{va} \cdot \phi^v_+/\tau\right)}{\sum_{l=0}^{L} \exp\left(q_{va} \cdot \phi^v_k/\tau\right)}, \tag{8}$$

where $\phi_+$ is the positive representation vector of the cluster corresponding to the pseudo label of the query and the $\tau$ is a temperature hyper-parameter following Cluster Contrast [7].

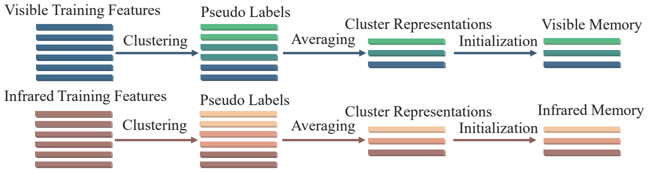

**Figure 3: Modality-specific memory initialization process.**

**Overall Loss.** Certainly, three types of ClusterNCE loss function are designed to learn discriminative representation:

$$L_{overall} = L_{q_i} + L_{q_v} + L_{q_{va}}. \tag{9}$$

The loss value is low when q is close to its positive cluster representation and dissimilar to all other cluster features. This process is equivalent to two modality-specific softmax-based classifiers that try to classify $q_i$ as $\phi^i_+$, and $q_v$ together with $q_{va}$ as $\phi^v_+$. $q_{va}$ is the query of channel augmented features for learning color-invariant information, and thus feature encoders have a certain modality generalization ability with the help of joint augmented learning.

**Disscusion.** Different from the learning approach in H2H [24], we integrate homogeneous learning of two modalities into a two-stream network which has a partially shared structure. Therefore, our approach can simultaneously learn modality-specific representation and modality-invariant features. In addition, joint augmented learning further aligns the feature spaces of the two modalities, giving the model a stronger modality generalization capability.

## 3.3 Cross-modality Memory Aggregation

Thanks to the augmented dual-contrastive learning framework, the model has certain modality generalization power and cross-modality retrieval capability, but the correlation between the samples of two modalities has not been fully explored and cannot handle the challenge of large cross-modality variations. To address this issue, we design a Cross-modality Memory Aggregation (CMA) module, which contains a count priority selection and a memory aggregation strategy. CMA can select reliable positive cross-modality label pairs and aggregate corresponding memories at the beginning of each training epoch, as shown in Figure 2.

For each cross-modality instance pairs $< x^v_m, x^i_n >$, we compute the similarity by:

$$sim(x^v_m, x^i_n) = \frac{u^v_m \cdot u^i_n}{||u^v_m|| \times ||u^i_n||}. \tag{10}$$

Then, we can match instance pairs by a threshold:

$$R = \{(y^v_m, y^i_n) | sim(x^v_m, x^i_n) \geqslant \Delta, x^v_m \in X_v, x^i_n \in X_i\}, \tag{11}$$

where the threshold $\Delta = 0.5$ in our work. $y^v_m$ and $y^i_n$ are the pseudo labels of $x^v_m$ and $x^i_n$, respectively. It is an initial similarity restraint, which can filter easy negative pairs for reducing calculations.

**Count Priority Selection.** For associating the reliable positive cross-modality cluster pairs, we argue that the number of each pseudo label pair represents the reliability of cross-modality label matching. Specifically, we count the label pairs in $R$ and rank them. One visible label can match many infrared labels, and the label pair corresponding to the maximum number can be regarded as the

**Algorithm 1:** Cross-modality Memory Aggregation

```
Data: Matched label pair set $R$
Data: The memories of two modality
      $\{\phi_1^i, \dots, \phi_K^i, \phi_1^v, \dots, \phi_L^v\}$
Result: Aggregate the reliable cross-modality memories
Count the set $R$ to obtain set $R_{cnt}$;
Sort $R_{cnt}$ from largest to smallest according to the number
  of label pairs;
for $< y_n^i, y_m^v >$ in $R_{cnt}$ do
    if $y_n^i = -1$ or $y_m^v = -1$ then
        Continue
    if $y_n^i$ or $y_m^v$ has appeared then
        Continue
    Update the infrared memory with Eq 12 ;
    Update the visible memory with Eq 13
```

**Table 1: The comparison of different memory updating strategies on SYSU-MM01. Rank at $r$ accuracy(%), mAP (%) and mINP (%) are reported.**

| | SYSU-MM01 (All Search) | | | | |
|---|---|---|---|---|---|
| Strategies | $r1$ | $r10$ | $r20$ | mAP | mINP |
| Average Memory | 44.39 | 82.91 | 91.11 | 41.02 | 25.94 |
| Visible Memory | 42.18 | 82.04 | 91.72 | 40.13 | 25.78 |
| Infrared Memory | **45.51** | **85.29** | **93.16** | **42.73** | **28.29** |

most reliable pseudo label pair, as shown in Figure 2. These pseudo label pairs are selected to aggregate the corresponding memories.

**Memory Aggregation.** We aggregate the selected memories using a momentum updating strategy by

$$\phi_{y_n^i}^{i(t)} \leftarrow \alpha \phi_{y_n^i}^{i(t)} + (1-\alpha)\phi_{y_n^i}^{i(t)}, \tag{12}$$

$$\phi_{y_m^v}^{v(t)} \leftarrow \alpha \phi_{y_m^v}^{v(t)} + (1-\alpha)\phi_{y_n^i}^{i(t)}, \tag{13}$$

where $\alpha$ is the momentum updating factor. $\phi_{y_n^i}^{i(t)}$ and $\phi_{y_m^v}^{v(t)}$ are the memories corresponding selected label pairs $< y_n^i, y_m^v >$ at epoch $t$. It is worth noting that we do not aggregate in both directions, and specifically we just aggregate the visible memory to the infrared memory. It means that the infrared memory is not changed as Eq 12 shows, and the visible memory is aggregated to infrared by Eq 13. This aggregation method allows the model to have a better ability to handle cross-modality variations. We will demonstrate it by comparing different aggregation strategies in Sec 4.5. The details are presented in Algorithm 1.

**Disscusion.** In contrast to the usual label refinement methods, we associate cross-modality samples by aggregating memory modules. This has two advantages: 1) In contrastive learning, the memory module provides supervised information for feature learning. The direct aggregation of memory modules enables a tight coupling of the cross-modality label association for contrastive learning, which is a mutual reinforcement. 2) The aggregation of memory is a fusion and association at the cluster level, which is more efficient than sample-to-sample association methods.

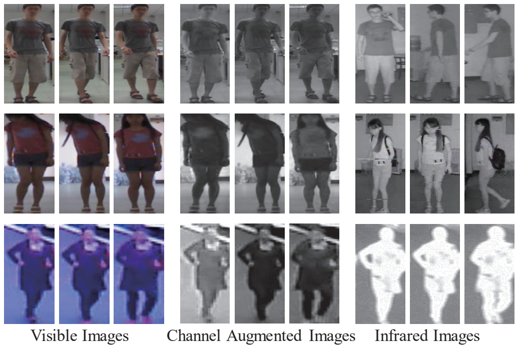

**Figure 4: Example images and random channel augmented images from SYSU-MM01 dataset (first two rows) and RegDB dataset (last row). Each row indicates images of the same identity from three different modalities.**

# 4 EXPERIMENTS

## 4.1 Datasets and Evaluation Protocol

We adopt two publicly available cross-modality person re-identification datasets (RegDB [28] and SYSU-MM01 datasets [41]) for evaluation. We plot some example images from two datasets in Figure 4, together with channel augmented images.

The SYSU-MM01 dataset is a cross-modality person image dataset captured by 2 near-infrared cameras and 4 visible cameras. The SYSU-MM01 contains 395 training identities, including 22,258 visible and 11,909 near-infrared images in outdoor and indoor environments. We adopt all-search and indoor-search evaluation modes.

RegDB dataset is captured by two aligned cameras (one visible and one thermal camera) system, containing 412 identities. We evaluate our method in the two test modes, including thermal to visible and visible to thermal. We strictly follow existing methods to perform ten trials of the gallery set selection [37, 47], and calculate the average performance.

**Evaluation Protocol.** Following existing works [44], mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) are calculated as the evaluation metrics.

## 4.2 Implementation Details

Our proposed framework is implemented in PyTorch. The settings of two shallow layers follow AGW [48], and we use ResNet50 [20] as the shared layers, which is initialized with ImageNet-pretrained weights [8]. During testing, we take the features of GeM [29] pooling layer to calculate the cosine similarity. At the beginning of each training epoch, DBSCAN [9] is adopted to generate pseudo labels in each modality independently.

During training, we sample 16 person identities and 16 instances for each identity from each modality training set. We apply the random horizontal flipping, random erasing, and random cropping images with size 288 × 144 for training. For augmented visible stream, we adopt random Channel Augmentation (CA) method [46]. Adam optimizer is used to train the model. The initial learning rate is set to $3.5e-4$ and reduced to $1/10$ of its previous value every 20 epochs. We train the model in total of 100 epochs, in

**Table 2: Ablation studies on the SYSU-MM01 dataset and RegDB. "DC" denotes the dual-contrastive learning without joint augmentation. "ADC" means the augmented joint dual-contrastive learning framework in Sec 3.2. "CMA" represents the cross-modality memory aggregation module in Sec 3.3. Rank at $r$ accuracy(%), mAP (%) and mINP (%) are reported.**

| | Components | | | | SYSU-MM01 (All Search) | | | | | SYSU-MM01 (Indoor Search) | | | | | RegDB (Visible to Infrared) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Index | Baseline | DC | ADC | CMA | $r1$ | $r10$ | $r20$ | mAP | mINP | $r1$ | $r10$ | $r20$ | mAP | mINP | $r1$ | $r10$ | $r20$ | mAP | mINP |
| 1 | √ | | | | 20.16 | 59.27 | 72.5 | 22.00 | 12.97 | 23.33 | 68.13 | 82.66 | 34.01 | 30.88 | 11.76 | 24.83 | 32.84 | 13.88 | 9.94 |
| 2 | | √ | | | 29.92 | 68.28 | 80.14 | 29.40 | 18.01 | 31.95 | 75.48 | 87.63 | 41.94 | 37.98 | 32.73 | 50.22 | 60.14 | 32.55 | 24.03 |
| 3 | | | √ | | 35.07 | 75.87 | 86.11 | 34.58 | 22.05 | 43.66 | 84.60 | 93.54 | 52.23 | 48.05 | 41.12 | 59.55 | 67.88 | 40.18 | 30.58 |
| 4 | | √ | | √ | 35.94 | 77.53 | 88.70 | 34.06 | 20.63 | 40.13 | 82.27 | 91.47 | 47.61 | 41.75 | 61.46 | 78.03 | 84.27 | 59.62 | 48.70 |
| 5 | | | √ | √ | **45.51** | **85.29** | **93.16** | **42.73** | **28.29** | **50.60** | **89.66** | **96.15** | **59.11** | **55.17** | **67.20** | **82.02** | **87.44** | **64.05** | **52.67** |

**Table 3: The comparison with the state-of-the-art methods on SYSU-MM01. It contains two groups, *i.e.*, unsupervised ReID methods and supervised VI-ReID methods. *cm-SSFT [22] reported a higher matching accuracy by using all gallery samples as auxiliary information, which is infeasible in many applications. Rank at $r$ accuracy(%), mAP (%) and mINP (%) are reported.**

| | SYSU-MM01 Settings | | All search | | | | | Indoor Search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Methods | Venue | $r1$(%) | $r10$(%) | $r20$(%) | mAP(%) | mINP(%) | $r1$(%) | $r10$(%) | $r20$(%) | mAP(%) | mINP(%) |
| Supervised | Zero-Padding [41] | ICCV-17 | 14.80 | 54.12 | 71.33 | 15.95 | - | 20.58 | 68.38 | 85.79 | 26.92 | - |
| | eBDTR [45] | TIFS-19 | 27.82 | 67.34 | 81.34 | 28.42 | - | 32.46 | 77.42 | 89.62 | 42.46 | - |
| | HSME [19] | AAAI-19 | 20.68 | 32.74 | 77.95 | 23.12 | - | - | - | - | - | - |
| | $D^2$RL [37] | CVPR-19 | 28.9 | 70.6 | 82.4 | 29.2 | - | - | - | - | - | - |
| | AlignGAN [34] | ICCV-19 | 42.4 | 85.0 | 93.7 | 40.7 | - | 45.9 | 87.6 | 94.4 | 54.3 | - |
| | X-Modal [23] | AAAI-20 | 49.9 | 89.8 | 96.0 | 50.7 | - | - | - | - | - | - |
| | Hi-CMD [5] | CVPR-20 | 34.9 | 77.6 | - | 35.9 | - | - | - | - | - | - |
| | cm-SSFT* [27] | CVPR-20 | 47.7 | - | - | 54.1 | - | - | - | - | - | - |
| | AGW [48] | TPAMI-21 | 47.50 | 84.39 | 92.14 | 47.65 | 35.30 | 54.17 | 91.14 | 95.98 | 62.97 | 59.23 |
| | DDAG [47] | ECCV-20 | 54.75 | 90.39 | 95.81 | 53.02 | 39.62 | 61.02 | 94.06 | 98.41 | 67.98 | 62.61 |
| | VCD+VML [31] | CVPR-21 | 60.02 | 94.18 | 98.14 | 58.80 | - | 66.05 | 96.59 | 99.38 | 72.98 | - |
| | CA [46] | ICCV-21 | 69.88 | 95.71 | 98.46 | 66.89 | 53.61 | 76.26 | 97.88 | 99.49 | 80.37 | 76.79 |
| | MPANet [42] | CVPR-21 | 70.58 | 96.21 | 98.80 | 68.24 | - | 76.74 | 98.21 | 99.57 | 80.95 | - |
| | MSO [13] | MM-21 | 58.70 | 92.06 | - | 56.42 | - | 63.09 | 96.61 | - | 70.31 | - |
| | AGM [50] | MM-21 | 69.63 | 96.27 | 98.82 | 66.11 | 52.24 | 74.68 | 97.51 | 99.14 | 78.30 | 74.00 |
| | MCLNet [18] | ICCV-21 | 65.40 | 93.33 | 97.14 | 61.98 | 47.39 | 72.56 | 96.98 | 99.20 | 76.58 | 72.10 |
| Unsupervised | HHL [52] | ECCV-18 | 2.86 | 21.41 | 35.36 | 7.37 | - | - | - | - | - | - |
| | SSG [12] | ICCV-19 | 2.32 | 17.23 | 28.88 | 5.00 | - | - | - | - | - | - |
| | ECN [53] | CVPR-19 | 8.07 | 32.49 | 45.95 | 12.68 | - | - | - | - | - | - |
| | SPCL [16] | NIPS-20 | 18.37 | 54.08 | 69.02 | 19.39 | 10.99 | 26.83 | 68.31 | 83.24 | 36.42 | 33.05 |
| | MMT [15] | ICLR-20 | 21.47 | 59.65 | 73.29 | 21.53 | 11.50 | 22.79 | 63.18 | 79.04 | 31.50 | 27.66 |
| | CAP [35] | AAAI-21 | 16.82 | 47.6 | 61.42 | 15.71 | 7.02 | 24.57 | 57.93 | 72.74 | 30.74 | 26.15 |
| | Cluster Contrast [7] | arXiv-21 | 20.16 | 59.27 | 72.5 | 22.00 | 12.97 | 23.33 | 68.13 | 82.66 | 34.01 | 30.88 |
| | ICE [2] | ICCV-21 | 20.54 | 57.5 | 70.89 | 20.39 | 10.24 | 29.81 | 69.41 | 82.66 | 38.35 | 34.32 |
| | H2H [24] | TIP-21 | 30.15 | 65.92 | 77.32 | 29.40 | - | - | - | - | - | - |
| | ADCA (Ours) | - | **45.51** | **85.29** | **93.16** | **42.73** | **28.29** | **50.60** | **89.66** | **96.15** | **59.11** | **55.17** |

which the previous 50 epochs are used for pre-training the ADC framework, and the CMA is executed in the last 50 epochs. The other settings of contrastive learning parameters follow [7]. The aggregation momentum value $\alpha$ is set to 0.1.

## 4.3 Comparison with State-of-the-Arts

There are limited corresponding state-of-the-art methods for USL-VI-ReID. To compensate for this, we refer to 3 unsupervised methods from H2H [24], and implement 6 advanced unsupervised methods on USL-VI-ReID. Meanwhile, we also report 16 state-of-the-art supervised methods for comparison. SPCL [16], MMT[15] and H2H [24] use Market1501 as an extra labeled RGB dataset for training. The comparisons on the SYSU-MM01 and RegDB are reported Table 3 and Table 4, respectively.

**Comparison with unsupervised methods.** The experiments demonstrate that our ADCA significantly outperforms existing unsupervised methods under various settings. Specifically, we surpass the existing single-modality unsupervised methods by considerable margins of about 20% and 40% mAP on SYSU-MM01 and RegDB

**Table 4: The comparison with the state-of-the-art methods on RegDB. It contains two groups, *i.e.*, unsupervised ReID methods and supervised VI-ReID methods. *cm-SSFT [22] reported a higher matching accuracy by using all gallery samples as auxiliary information, which is infeasible in many applications. Rank at *r* accuracy(%), mAP (%) and mINP (%) are reported.**

| | RegDB Settings | | Visible to Infrared | | | | | Infrared to Visible | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Methods | Venue | r1(%) | r10(%) | r20(%) | mAP(%) | mINP(%) | r1(%) | r10(%) | r20(%) | mAP(%) | mINP(%) |
| Supervised | Zero-Padding [41] | ICCV-17 | 17.75 | 34.21 | 44.35 | 18.90 | - | 16.63 | 34.68 | 44.25 | 17.82 | - |
| | eBDTR [45] | TIFS-19 | 34.62 | 58.96 | 68.72 | 33.46 | - | 34.21 | 58.74 | 68.64 | 32.49 | - |
| | HSME [19] | AAAI-19 | 50.85 | 73.36 | 81.66 | 47.00 | - | 50.15 | 72.40 | 81.07 | 46.16 | - |
| | D$^2$RL [37] | CVPR-19 | 43.4 | 66.1 | 76.3 | 44.1 | - | - | - | - | - | - |
| | AlignGAN [34] | ICCV-19 | 57.9 | - | - | 53.6 | - | 56.3 | - | - | 53.4 | - |
| | X-Modal [23] | AAAI-20 | 62.21 | 83.13 | 91.72 | 60.18 | - | - | - | - | - | - |
| | Hi-CMD [5] | CVPR-20 | 70.93 | 86.39 | - | 66.04 | - | - | - | - | - | - |
| | cm-SSFT* [27] | CVPR-20 | 72.3 | - | - | 72.9 | - | 71.0 | - | - | 71.7 | - |
| | AGW [48] | TPAMI-21 | 70.05 | 86.21 | 91.55 | 66.37 | 50.19 | 70.49 | 87.21 | 91.84 | 65.90 | 51.24 |
| | DDAG [47] | ECCV-20 | 69.34 | 86.19 | 91.49 | 63.46 | 49.24 | 68.06 | 85.15 | 90.31 | 61.80 | 48.62 |
| | VCD+VML [31] | CVPR-21 | 73.2 | - | - | 71.6 | - | 71.8 | - | - | 70.1 | - |
| | CA [46] | ICCV-21 | 85.03 | 95.49 | 97.54 | 79.14 | 65.33 | 84.75 | 95.33 | 97.51 | 77.82 | 61.56 |
| | MPANet [42] | CVPR-21 | 82.8 | - | - | 80.7 | - | 83.7 | - | - | 80.9 | - |
| | MSO [13] | MM-21 | 73.6 | 88.6 | - | 66.9 | - | 74.6 | 88.7 | - | 67.5 | - |
| | AGM [50] | MM-21 | 88.40 | 95.10 | 96.94 | 81.45 | 68.51 | 85.34 | 94.56 | 97.48 | 81.19 | 65.76 |
| | MCLNet [18] | ICCV-21 | 80.31 | 92.70 | 96.03 | 73.07 | 57.39 | 75.93 | 90.93 | 94.59 | 69.49 | 52.63 |
| Unsupervised | HHL [52] | ECCV-18 | 4.61 | 11.48 | 16.53 | 6.22 | - | - | - | - | - | - |
| | SSG [12] | ICCV-19 | 1.91 | 5.14 | 7.53 | 3.18 | - | - | - | - | - | - |
| | ECN [53] | CVPR-19 | 2.17 | 8.38 | 12.55 | 2.90 | - | - | - | - | - | - |
| | SPCL [16] | NIPS-20 | 13.59 | 26.98 | 34.88 | 14.86 | 10.36 | 11.70 | 25.53 | 32.82 | 13.56 | 10.09 |
| | MMT [15] | ICLR-20 | 25.68 | 42.23 | 54.03 | 26.51 | 19.56 | 24.42 | 41.21 | 51.89 | 25.59 | 18.66 |
| | CAP [35] | AAAI-21 | 9.71 | 19.27 | 25.6 | 11.56 | 8.74 | 10.21 | 19.91 | 26.38 | 11.34 | 7.92 |
| | Cluster Contrast [7] | arXiv-21 | 11.76 | 24.83 | 32.84 | 13.88 | 9.94 | 11.14 | 24.11 | 32.65 | 12.99 | 8.99 |
| | ICE [2] | ICCV-21 | 12.98 | 25.87 | 34.4 | 15.64 | 11.91 | 12.18 | 25.67 | 34.9 | 14.82 | 10.6 |
| | H2H [24] | TIP-21 | 23.81 | 45.31 | 54.00 | 18.87 | - | - | - | - | - | - |
| | ADCA (Ours) | - | **67.20** | **82.02** | **87.44** | **64.05** | **52.67** | **68.48** | **83.21** | **88.00** | **63.81** | **49.62** |

tasks, respectively. In addition, compared with H2H [24] which used an extra labeled RGB dataset for unsupervised cross-modality ReID, evident 13.33% and 45.18% mAP gains are achieved on SYSU-MM01 (all search) and RegDB (visible to infrared), respectively.

**Comparison with supervised methods.** Our ADCA surpasses some supervised methods including Zero-Padding [41], eBDTR [45], HSME [19], D$^2$RL [37], and AlignGAN [34]. This is an encouraging result, showing that unsupervised cross-modality ReID is promising in approximating the effects of supervised VI-ReID.

These considerable gains benefit from the insightful design of our method for USL-VI-ReID. There are three major advantages of our method: 1) We do not need any additional labeled data. This property makes our proposed framework more applicable to practical deployment. 2) Our solution is simple, efficient, and easy to implement. We assume that the performance would be further improved if advanced contrastive learning were introduced. 3) The learned feature is robust against different cross-modality datasets and matching settings.

## 4.4 Ablation Study

The performance boost of the ADCA framework in USL-VI-ReID mainly comes from the proposed Augmented Dual-Contrastive (ADC) learning and Cross-modality Memory Aggregation (CMA) module. We validate the effectiveness of each component by conducting ablation studies on SYSU-MM01 and RegDB datasets. Results are shown in Table 2.

**Baseline** in index 1 denotes that we directly train the model on SYSU-MM01 and RegDB tasks using one-stream contrastive learning with Cluster Contrast [7] method. Although Cluster Contrast has a promising performance on single-modality unsupervised ReID, it is observed that the baseline only achieves 22.00% mAP on SYSU-MM01 (all search) and 11.76% mAP on RegDB (visible to infrare). Therefore, directly using single-modality unsupervised ReID methods can hardly tackle the USL-VI-ReID problems.

**Effectiveness of DC.** Index 2 means the dual-constrastive learning without joint augmented learning. Compared with baseline, DC improves the performance of 7.40% and 18.67% mAP on SYSU-MM01 (all search) and RegDB (visible to infrared). The main gain is achieved by the design of the dual-path network which allows the model to learn homogeneous learning without interfering with each other, and capture certain modality-invariant features through the shared embedding layers.

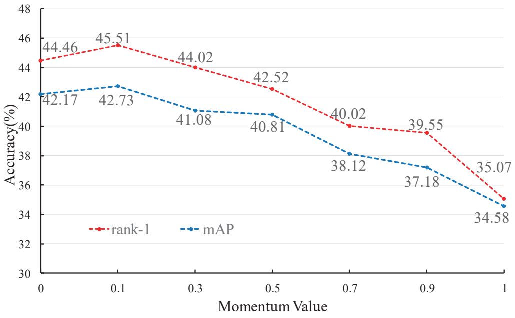

**Figure 5: The impact of memory aggregation momentum value on SYSU-MM01 dataset.**

**Effectiveness of ADC.** Index 3 represents the augmented dual-contrastive learning which incorporates random channel augmentation (CA) as an auxiliary branch for joint learning. Compared with DC, ADC brings consistent improvement in all settings. The improvements are 5.18% and 7.63% mAP on SYSU-MM01 (all search) and RegDB (visible to infrared). The reasons for improvements are two-fold. First, the random channel augmentation [46] encourages the model to learn the explicit relation between each color channel of the visible images and the single-channel infrared images. Second, the joint learning strategy fully utilizes the channel augmented images without modifying the network structures, which facilitates the learning of cross-modality features.

**Effectiveness of CMA.** Index 4 and 5 denote the DC and ADC with the CMA module. Compared with the results in DC and ADC, the significant improvements demonstrate the effectiveness of CMA. CMA can select reliable positive cross-modality label pairs and aggregate the corresponding memories to ensure the discriminability of cross-modality features at the cluster level, which is more efficient than point-to-point association methods.

## 4.5 Further Analysis

**Memory Aggregation Momentum Value.** We use the momentum update strategy to aggregate the two modality-specific memories. The $\alpha$ in Eq 12 and Eq 13 controls the speed of aggregating the memory. As shown in Figure 5, when $\alpha = 0.1$, ADCA achieves the best results.

**Memory Updating Strategy.** We present the different memory update strategies in Table 1. The average memory in Table 1 denotes that we update the two memories by:

$$\phi_{y_n^i}^{i(t)} \leftarrow \alpha\phi_{y_n^i}^{i(t)} + \frac{1}{2}(1-\alpha)(\phi_{y_n^i}^{i(t)} + \phi_{y_m^v}^{v(t)}), \tag{14}$$

$$\phi_{y_m^v}^{v(t)} \leftarrow \alpha\phi_{y_n^i}^{i(t)} + \frac{1}{2}(1-\alpha)(\phi_{y_n^i}^{i(t)} + \phi_{y_m^v}^{v(t)}). \tag{15}$$

The visible memory in Table 1 indicates that we use visible memory to update the two memories as follows:

$$\phi_{y_n^i}^{i(t)} \leftarrow \alpha\phi_{y_n^i}^{i(t)} + (1-\alpha)\phi_{y_m^v}^{v(t)}, \tag{16}$$

$$\phi_{y_m^v}^{v(t)} \leftarrow \alpha\phi_{y_m^v}^{v(t)} + (1-\alpha)\phi_{y_m^v}^{v(t)}. \tag{17}$$

The infrared memory in Table 1 means that we update the two memories by Eq 12 and Eq 13. The results demonstrate that the update strategy using infrared memory achieves the best performance.

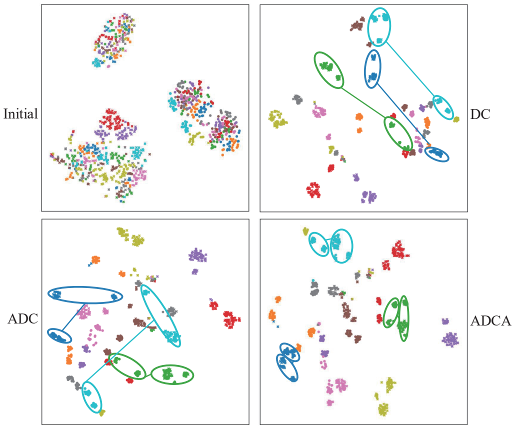

**Figure 6: The t-SNE visualization of 20 randomly selected identities. The color indicates the identity. Circle means visible modality and the cross means the infrared modality. The sample points of the two modalities in the oval are gradually drawn closer from DC to ADCA.**

**Visualization.** We plot the t-SNE [32] map of 20 identities randomly selected from the SYSU-MM01 dataset, as shown in Figure 6. From initial to ADCA, infrared and visible positive sample points are gradually drawn closer. Ultimately, the model learns a better distribution of cross-modality features in ADCA. This visualization further validates the effectiveness of our approach. We also note that some sample points of the same identity are not clustered together and there is still much room for improvement in the USL-VI-ReID task.

## 5 CONCLUSION

This paper introduces unsupervised learning visible infrared re-identification (USL-VI-ReID) task to alleviate the issue of expensive cross-modality annotations. To address the problem caused by large cross-modality discrepancies in USL-VI-ReID, we propose a novel Augmented Dual-Contrastive Aggregation (ADCA) learning framework which is based on the idea of homogenous joint learning and heterogeneous aggregation, improving the unsupervised cross-modality recognition. It has been validated on two different tasks, significantly outperforming current state-of-the-art unsupervised methods and even some supervised methods, pushing unsupervised VI-ReID to real-world deployment. We will further investigate our approach in other visible-infrared applications.

## ACKNOWLEDGMENTS

This work was supported in part by National Nature Science Foundation of China under Grants 62071338 and 62176188, and in part by Key Research and Development Program of Hubei Province (2021BAA187). The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

# REFERENCES

[1] Hao Chen, Benoit Lagadec, and Francois Bremond. 2021. Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification. In *WACV*. 1–10.

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*. 14960–14969.

[3] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. 2021. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*. 2004–2013.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2019. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*. 232–242.

[5] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*. 10257–10266.

[6] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality person re-identification with generative adversarial training.. In *IJCAI*. 6.

[7] Zuozhuo Dai, Guangyuan Wang, Siyu Zhu, Weihao Yuan, and Ping Tan. 2021. Cluster Contrast for Unsupervised Person Re-Identification. arXiv 2021. *arXiv preprint arXiv:2103.11568* (2021).

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

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*. 226–231.

[10] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *TOMM* (2018), 1–18.

[11] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. 2019. Learning modality-specific representations for visible-infrared person re-identification. *IEEE TIP* (2019), 579–590.

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

[13] Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu, Yidong Li, and Congyan Lang. 2021. MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5257–5265.

[14] Wenhang Ge, Chunyan Pan, Ancong Wu, Hongwei Zheng, and Wei-Shi Zheng. 2021. Cross-Camera Feature Prediction for Intra-Camera Supervised Person Re-identification across Distant Scenes. In *ACM MM*. 3644–3653.

[15] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526* (2020).

[16] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NeurIPS* (2020), 11309–11321.

[17] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*. 3642–3651.

[18] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. 2021. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In *ICCV*. 16403–16412.

[19] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*. 8385–8392.

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

[21] Xiangyuan Lan, Andy J Ma, Pong C Yuen, and Rama Chellappa. 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE TIP* 12 (2015), 5826–5841.

[22] Xiangyuan Lan, Shengping Zhang, Pong C Yuen, and Rama Chellappa. 2017. Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. *IEEE TIP* (2017), 2022–2037.

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

[24] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. 2021. Homogeneous-to-Heterogeneous: Unsupervised Learning for RGB-Infrared Person Re-Identification. *IEEE TIP* (2021), 6392–6407.

[25] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*. 8738–8745.

[26] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. 2020. Unsupervised person re-identification via softened similarity learning. In *CVPR*. 3390–3399.

[27] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*. 13379–13389.

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

[29] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE TPAMI* (2018), 1655–1668.

[30] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. 2020. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition* (2020), 107173.

[31] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*. 1522–1531.

[32] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[33] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised person re-identification via multi-label classification. In *CVPR*. 10981–10990.

[34] Guan’an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*. 3623–3632.

[35] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. 2021. Camera-aware proxies for unsupervised person re-identification. In *AAAI*. 4.

[36] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin’ichi Satoh. 2017. Person reidentification via discrepancy matrix and matrix metric. *IEEE TCYB* (2017), 3006–3020.

[37] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin’ichi Satoh. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*. 618–626.

[38] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin’ichi Satoh. 2019. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048* (2019).

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

[40] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. 2020. RGB-IR person re-identification by cross-modality similarity preservation. *IJCV* (2020), 1765–1785.

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

[42] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *CVPR*. 4330–4339.

[43] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. 2020. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE TIP* (2020), 9387–9399.

[44] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*. 7501–7508.

[45] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. 2019. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS* (2019), 407–419.

[46] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. 2021. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *ICCV*. 13567–13576.

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

[48] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI* (2021).

[49] Mang Ye, Jianbing Shen, and Ling Shao. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE TIFS* (2020), 728–739.

[50] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 788–796.

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*. 1116–1124.

[52] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. 2018. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*. 172–188.

[53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*. 598–607.

[54] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. 2020. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*. 87–104.