Enhancing Diverse Intra-identity Representation for Visible-Infrared Person Re-Identification

Sejun Kim*, Soonyong Gwon*, Kisung Seo†
Seokyeong University, Seoul, Korea
{kimsejun5,gwonsy2,ksseo}@skuniv.ac.kr

Abstract

Visible-Infrared person Re-Identification (VI-ReID) is a challenging task due to modality discrepancy. To reduce modality-gap, existing methods primarily focus on sample diversity, such as data augmentation or generating intermediate modality between Visible and Infrared. However, these methods do not consider the increase in intra-instance variance caused by sample diversity, and they focus on dominant features, which results in a remaining modality gap for hard samples. This limitation hinders performance improvement. We propose Intra-identity Representation Diversification (IRD) based metric learning to handle the intra-instance variance. Specifically IRD method enlarge the Intra-modality Intra-identity Representation Space (IIRS) for each modality within the same identity to learn diverse feature representation abilities. This enables the formation of a shared space capable of representing common features across hetero-modality, thereby reducing the modality gap more effectively. In addition, we introduce a HueGray (HG) data augmentation method, which increases sample diversity simply and effectively. Finally, we propose the Diversity Enhancement Network (DEN) for robustly handling intra-instance variance. The proposed method demonstrates superior performance compared to the state-of-the-art methods on the SYSU-MM01 and RegDB datasets. Notably, on the challenging SYSU-MM01 dataset, our approach achieves remarkable results with a Rank-1 accuracy of 76.36% and a mean Average Precision (mAP) of 71.30%.

1. Introduction

Person Re-identification (ReID) [1, 16, 41, 44] is a retrieval task of matching the same person across multiple camera views. While most existing ReID methods focus on matching visible images captured during daylight, these methods may not perform well when poor conditions, such as at night or low-light environments. In order to improve video surveillance in poor conditions, infrared cameras are used in combination with visible cameras. However, matching visible and infrared is challenging due to the large gap of cross modality. To solve this problem, some visible-infrared person re-identification (VI-ReID) methods [10, 25, 29, 30, 34, 38] emerged for matching visible images with their corresponding infrared images, and vice versa.

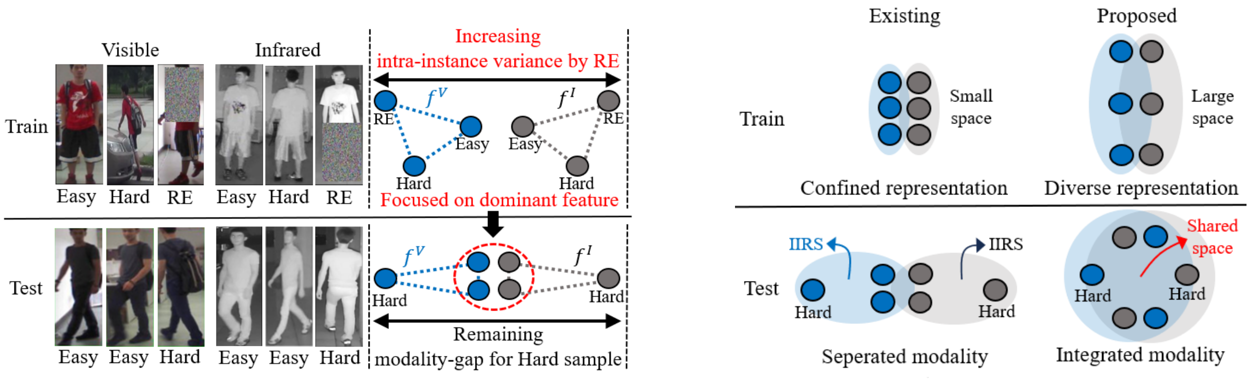
Various augmentation methods for VI-ReID have been used, such as gray transform [3, 7, 42] and generative based methods [4–6, 28] to make the Visible image similar to the Infrared image. However, these methods often result in damage to the visible texture due to artificial manipulation. Additionally, some studies have explored 3-modality learning methods [8, 13, 17, 31, 35, 37, 40] that generate middle-modality images to assist VI-ReID learning.

Nevertheless, previous studies did not consider the intra-instance variance caused by sample diversity. As a result, the trained model exhibits robust performance for easy samples, but still struggles with hard samples due to a significant modality gap, leading to decreased performance as shown in Figure 1a. To reduce the modality gap for hard samples, it is necessary to have metric learning that allows for diverse feature representation. This can be achieved by expanding the Intra-modality Intra-identity Representation Space (IIRS). Here, IIRS refers to the diversity of representations within the Intra-modality Intra-identity space. We propose Intra-identity Representation Diversification based metric learning, which effectively reduces the modality gap by training diverse feature representations that can handle all samples regardless of their difficulty level. Our approach, as shown in Figure 1b, achieves this by expanding the IIRS during train to learn various shared representations across opposite modalities, leading to an increased shared space at test and thus reducing the modality gap. Importantly, our method effectively solved the challenge of matching hard samples caused by intra-instance variance.

Only a few studies have focused on expanding the repre-

[*] Co-first Author, equally contribute
[†] Corresponding Author

This work was supported by National Research Foundation of Korea Grant RS-2023-00244355 funded by the Korea government.



(a) Example of intra-instance variance.

(b) Comparison of IIRS utilization.

Figure 1. Intra-instance variance and Intra-modality Intra-identity Representation Space (IIRS). (a) shows an example of the intra-instance variance problem by random erasing [43] in VI-ReID. It illustrates three visible and infrared sample images pairs for the same identity in both the train and test with the distribution of features in the embedding space. Note that different identities are used for train and test, data augmentation such as Random Erasing (RE) is not applied in the test. f^V and f^I and denote Visible and Infrared features. Due to the increased intra-instance variance during train, the model tends to focus on dominant features, which results in a stable performance for easy samples during test. However, the modality gap remains large, for hard samples. (b) illustrates the difference in the utilization of the Intra-modality Intra-identity Representation Space (IIRS) between the existing and the proposed method during train and test. The proposed method utilizes a broader IIRS during train compared to the existing method, enabling the learning of diverse representation abilities. As a result, in the test, the shared space between different modality is wider, effectively reducing the modality gap.

sentation space. For instance, [14] proposed Margin MMD-ReID to bring inter-modality features closer together while maintaining a small margin to prevent overfitting. However, despite achieving certain performance, this method has a limitation: if the distance between features is smaller than the margin, no backpropagation occurs, resulting in an inability to separate features that are too close to each other using a small margin. In contrast, our proposed method not only addresses this limitation but also expands the IIRS while considering intra-instance variance. In [9], the proposed Identity-aware marginal Center Aggregation (ICA) is conceptually similar to our proposed metric learning. ICA is designed to separate the center feature and sample feature by a small margin in order to maintain diversity. However, even with this diversity preservation, the modality gap still exists as it integrates visible and infrared features without specifying the corresponding modality, assuming modality-invariant features. In contrast, our method not only maintains diversity but also expands the IIRS to enable more discriminative representations.

In this paper, we highlight the significance of well-designed metric learning and enlarging the IIRS in addressing intra-instance variance in VI-ReID. Enlarging the IIRS enables diverse representations and reduces the modality gap. To achieve this, we introduce the Intra-identity Representation Diversification (IRD) strategy for discriminative representation learning. Furthermore, to enhance sample diversity, we employ a sophisticated Hue transform and Gray transform (HG) to generate infrared-like images from visible images, ensuring the variety of samples simultaneously. This preserves the image structure while providing diverse color patterns that enhance distinctive feature learning. Finally, we propose the Diversity Enhancement Network (DEN) for robustly handling intra-instance variance.

In summary, the main contributions are as follows: First, to the best of our knowledge, this is the first attempt to investigate the necessary to enlarge IIRS considering intra-instance variance by sample diversity for VI-ReID. Second, we propose Intra-identity Representation Diversification (IRD) to expand the IIRS for discriminative learning on intra-instance variance. This enlargement of IIRS enables diverse representations of intra-instances while effectively reduces the modality gap by providing a learnable space for counterpart modality learning. Third, we propose Diversity Enhancement Network (DEN) to solve the problem of intra-instance variance by sample diversity and flexibly learn diverse discriminative representations for VI-ReID.

2. Related Work

Metric learning. Metric learning aims to find similarity between the compressed two representation vectors. Person Re-identification requires distinguishing different individuals, Cross-Entropy loss is not effective for separating each person. The task resembles zero-shot learning because the identity label does not overlap in the train and test. Therefore, the model requires the ability to learn discriminative feature representations rather than learning the characteristics of each individual identity. To achieve this, the Hardest Triplet loss (HT) [12] was introduced, which had a significant impact on learning discriminative feature representations by pulling hard positive samples to become more similar and pushing hard negative samples to become less similar. Unlike conventional HT, Weighted Regularization Triplet loss (WRT) [36], which is usually used as a baseline in VI-ReID, gives a relative distance to each of positive pairs and negative pairs as a normalized softmax-based weight. This is effective for single modality because multiple samples can be handled simultaneously without additional hyperparameters. However, there is a limit to representing various and distinctive features in cross modality than HT. In particular, the increase in sample diversity by data augmentation constrains various representations because it pulls all samples belonging to intra-identity. We conduct a comparative analysis supported by the numerical evidence, and show that learning to enlarge the IIRS, as the sample varies, can effectively reduce the modality gap.

An inter-modality center loss [3, 19, 20, 26] is proposed to narrow the modality gap. These methods utilize a loss function that calculates the center feature for each modality and identity by average, aiming to bring the positive samples closer together while pushing the negative samples further apart. Although it is effective to group intra-identity samples to reduce the inter-modality gap, this approach still does not address intra-instance variance. As a result, the representation space of intra-identity becomes too confined, while the modality gap remains. This leads to the reinforcement of only dominant feature representations and makes it difficult to achieve discriminative representation for diverse features.

3. Methodology

3.1. Exploring Intra-modality Intra-identity Representation Space

In order to solve intra-instance variance problem, Intra-identity Representation Diversification (IRD) method is proposed to enlarge the Intra-modality Intra-identity Representation Space (IIRS). IIRS refers to the feature distribution space of samples for the same identity. We define the size of IIRS as follows.

$$IIRS(V,V) = \frac{1}{P \times K} \sum_{i=1}^{P} \sum_{k=1}^{K} d(f_i^V, f_k^V) \tag{1}$$

Where, P is the number of identity, K is the number of samples in each identity, $d(\cdot)$ is the euclidean distance value between features of intra-identity samples, and f^V denotes the visible feature. That is, the average distance among features of all samples from the same identity depicts the diversity that the corresponding identity can represent.

We analyze the necessity of enlarging IIRS for diverse samples in VI-ReID. In order to do that, we compare the variations in representation space size when employing Triplet Loss (WRT [36] and HT [12] that mentioned in Section 2) and Sample Diversity (Random Erasing(RE) [43]). WRT can be formulated as follows:

$$L_{WRT} = log(1 + exp(\sum_k w_{ik}^p d_{ik}^p - \sum_j w_{ij}^n d_{ij}^n)) \tag{2}$$

$$w_{ik}^p = \frac{exp(d_{ik}^p)}{\sum_{d_{ik}^p \in P_i} exp(d_{ik}^p)}, \quad w_{ij}^n = \frac{exp(-d_{ij}^n)}{\sum_{d_{ij}^n \in N_i} exp(-d_{ij}^n)} \tag{3}$$

Where, (i, k and j) denotes the anchor, positive, and negative samples, $d_{i,k}^p/d_{i,j}^n$ represents the pairwise distance of a positive/negative sample pair. In contrast to this approach, HT does not involve all samples. Instead, it selectively operates on the hardest samples and is performed as follows:

$$L_{HT} = max([max(d_{ik}^p) - min(d_{ij}^n) + m_{HT}], 0) \tag{4}$$

The notation $max(value, 0)$ indicates that no backpropagation occurs if the value is below zero. The margin m_{HT} is a hyper-parameter that ensures a certain distance between positive and negative samples.

Our observation reveals that the IIRS is increased as the feature representations become more diverse. During the test, the same person is matched using both visible and infrared modalities, retrieval performance improves when the feature representations within the intra-modality intra-identity exhibit greater diversity than similarity. We calculate the IIRS for intra-modality (V,V) and (I,I), and compare the results for WRT, HT with or without RE, as presented in Table 1.

idx	Loss	Sample Diversity	IIRS (V,V)	IIRS (I,I)	R1	mAP
1	WRT	-	0.441	0.472	47.75	47.79
2	**HT**	**-**	**0.442**	**0.478**	**53.12**	**52.09**
3	WRT	RE	0.449	0.459	63.4	60.78
4	**HT**	**RE**	**0.462**	**0.476**	**64.95**	**62.47**

Table 1. IIRS size for (V,V) and (I,I) using all test data from the SYSU-MM01 dataset. We use AGW [36] as a base model without the triplet loss. The largest value, indicated in bold, corresponds to the Sample Diversity.

In Table 1, WRT, which is primarily used in VI-ReID, exhibits lower performance than HT while also demonstrating smaller IIRS values for both (V,V) and (I,I). When comparing the variations in IIRS between idx 1 and 2, there is no significant gap between WRT and HT when sample diversity is not employed. However, in idx 3 and 4, the use of sample diversity widens the gap even further. Unlike HT, which learns only the hardest sample pair, WRT can not

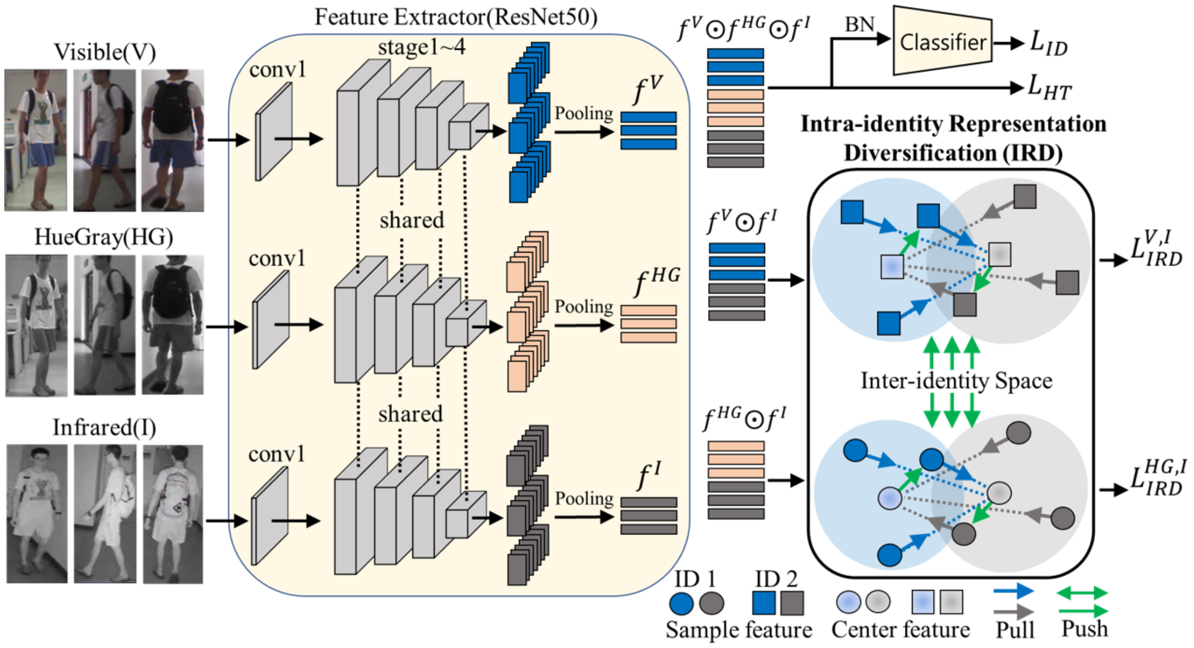


Figure 2. Proposed Diversity Enhancement Network (DEN). The Visible(V), Infrared(I) and HueGray(HG) images are fed into the feature extractor. The modality-specific module uses only one conv-layer at the front of ResNet-50 [11] without weight sharing, on the other hand, the modality-shared module employs the remaining stages 1 to 4, with weight sharing. $\odot$ is concatenation.f^V, f^{HG} and f^I are learned by using Identity Loss (L_{ID}), Hardest Triplet Loss (L_{HT}) and Intra-identity Representation Diversification Loss (L_{IRD}).

learn sufficient feature representation of increased samples through sample diversity because WRT tends to learns the dominant feature of all samples with the same identity, as shown in Figure 1a. Consequently, WRT has limitations in achieving discriminative representations for diverse samples compared to HT. Therefore, in order to cope with intra-instance variance by sample diversity, the metric learning capable of diverse feature representation for intra-identity is necessary. In the experimental section, we examine the variations in IIRS according to diverse samples (Hue Transform, Gray Transform) and analyze the results.

3.2. Intra-identity Representation Diversification

Based on Section 3.1, we propose an Intra-identity Representation Diversification (IRD) method that effectively narrows the modality gap while expanding the IIRS. The IRD loss consists of Positive Enhancement loss (PE) and Negative Enhancement loss (NE), formulated as follows:

$$L^{V,I}_{IRD} = L^{V,I}_{PE} + L^{V,I}_{NE} \tag{5}$$

$$\begin{aligned} L^{V,I}_{PE} = \frac{1}{P}\sum_{i=1}^{P} max([d(fc^{V,I}_i, fc^{I,V}_i) \\ -min(d(fc^{V,I}_i, f^{V,I}_i)) + m^{V,I}_{PE}], 0) \end{aligned} \tag{6}$$

$$\begin{aligned} L^{V,I}_{NE} = \frac{1}{P}\sum_{i=1}^{P} max([d(fc^{V,I}_i, fc^{I,V}_i) \\ -min(d(fc^{V,I}_i, fc^{I,V}_j)) + m^{V,I}_{NE}], 0) \end{aligned} \tag{7}$$

Where $fc^{V,I}_i$ represents the feature center, which is the average of the i-th identity features in the Visible or Infrared modality, $f^{V,I}_i$ is i-th identity sample feature. The function $d(\cdot)$ denotes the Euclidean distance, and margin $m^{V,I}$ is a hyper-parameter. The notation (i, j) signifies different identities. The Positive Enhancement loss (PE) aims to expand the IIRS, while the Negative Enhancement loss (NE) aims to enlarge the inter-identity representation space. In equation 6, the first term, $d(fc^{V,I}_i, fc^{I,V}_i)$, promotes the inter-modality center features of the i-th identity to be closer for learning the similarity of inter-modality intra-identity. The middle term, $-min(d(fc^{V,I}_i, f^{V,I}_i))$, aims to widen the small distance between the center feature and the nearest sample feature within the intra-modality for learning the diversity of intra-modality intra-identity. These mean that the representation space of the same identity from different modality becomes closer, leading to an expansion of the IIRS.

Since the center feature is derived from aggregating intra-modality features, it tends to capture the dominant features. However, in order to obtain discriminative feature representations that differ from these dominant features, we push the nearest sample features from the center feature by a small margin. Note that this process prevents concentration on dominant features and learns various feature representations. By performing this process for both visible and infrared modality, the features eventually get pushed towards the hetero-modality space within the same identity, as illustrated in Figure 2. Therefore, it is possible to effectively reduce the modality gap while learning diverse representations within the intra-identity.

However, in the process of pushing the intra-identity feature, there is a risk of it becoming closer to the representation space of other identities. To address this, Equation 7 is designed to ensure the separation of the intra-identity center feature from those of other identities, as depicted by green bidirectional arrows in the middle-right of Figure 2. Ultimately, the proposed Intra-identity Representation Diversification (IRD) approach, expands the representation space of each modality individually and is based on HT, rather than comparing all samples. This approach effectively reduces the modality gap while maintaining diverse representations.

3.3. Diversity Enhancement Network

Our proposed structure aims to learn diverse and discriminative representations while enlarging the representation space for VI-ReID, as depicted in Figure 2. In order to effectively learn from diverse samples, HueGray (HG) is generated using hue transform and gray transform on the Visible image. HG serves as a visually similar image to the Infrared (IR) modality, aiding in reducing the modality gap and significantly enhancing the diversity of the samples, enabling the model to learn diverse representations. Specifically, the hue transform randomly adjusts the hue degree within the range of 360 degrees, altering the color of the Visible image. This transformation allows the model to focus on extracting color-invariant modality-shared features, such as body shape. The gray transform applies a gray scale factor [0.299, 0.587, 0.114] to the R,G and B channels, resulting in a hue-transformed infrared-like image. In contrast to the channel exchange method used in the existing CAJ [35] approach, which replaced only one of the three channels (R, G, or B), our proposed HG can choose one among 360 different gray images by randomly varying the hue degree of a single image to ensure the variety of samples. This approach ensures that the shape of the image is maintained while improving the learning of discriminative feature representations. The ablation study conducted in the experiment section demonstrates a significant impact on performance by simply increasing the diversity of the generated HG images. To achieve effective discriminative learning on diverse samples, we utilize the IRD proposed in Section 3.2.

However, considering HG provide an intermediate modality between visible and infrared, applying equation 5 to the between HG and I(Infrared) may cause the learning direction of the model to be influenced by color differences. Hence, we introduce a Color Invariant loss (CI) between HueGray and Visible for emphasize shape-related features that are independent of color, formulated as follows:

$$L_{CI}^{HG,V} = \frac{1}{P \times K} \sum_{i=1}^{P \times K} d(f_i^{HG}, f_i^{V}) \tag{8}$$

Equation 8 represents an auxiliary loss function aimed at further reducing the modality gap between Visible and Infrared. Therefore, IRD (HG,I) is as follows:

$$L_{IRD}^{HG,I} = L_{PE}^{HG,I} + L_{NE}^{HG,I} + L_{CI}^{HG,V} \tag{9}$$

By learning the differences in features between HueGray and Infrared, which are closer to Infrared than Visible, the model can more effectively facilitate discriminative learning. The total Intra-identity Representation Diversification loss (L_{IRD}), is defined as follows:

$$L_{IRD} = L_{IRD}^{V,I} + L_{IRD}^{HG,I} \tag{10}$$

Finally, entire loss function consists of identity loss (L_{ID}) which using cross-entropy loss to each person identity, Hardest Triplet loss (L_{HT}) and Intra-identity Representation Diversification loss (L_{IRD}), as follows:

$$L = \lambda_{ID} \cdot L_{ID} + \lambda_{HT} \cdot L_{HT} + \lambda_{IRD} \cdot L_{IRD} \tag{11}$$

where λ_{ID},λ_{HT} and λ_{IRD} are hyperparameters to balance the contributions of individual loss terms.

4. Experiments

4.1. Datasets and Implementation details

Datasets. The two public VI-ReID datasets known by the name of SYSU-MM01 [32] and RegDB [23] are used to evaluate our method. SYSU-MM01 dataset contains 491 identities captured by 4 Visible cameras and 2 Infrared cameras in both indoor and outdoor environments. We utilize 395 identities consisting of 22,285 Visible images and 11,909 Infrared images for training. In test, 96 identities including 3,803 infrared query images and 301 Visible gallery images are used. Compared to RegDB, SYSU-MM01 is more difficult due to large variations. RegDB dataset includes 412 identities collected by one visible camera and one infrared camera. Each identity consist of 20 images involving 10 Visible images and 10 Infrared images. Half of the dataset including 206 identities is used for training and remaining 206 identities for test phase. **Evaluation metrics.** The public evaluation metrics, cumulative matching characteristics (CMC) at Rank-1 and mean average precision (mAP), are adopted to evaluate our method. Query and gallery images are from different modality. **Implementation details.** Please refer to the supplementary.

Method	SYSU-MM01				RegDB			
	All-search		Indoor-search		Visible to Infrared		Infrared to Visible	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
cm-SSFT [22]	61.6	63.2	70.5	72.6	72.3	72.9	71	71.7
Hc-Tri [26]	61.68	57.51	63.41	68.17	91.05	83.28	89.3	81.46
MCLNet [9]	65.4	61.98	72.56	76.58	80.31	73.07	75.93	69.49
FMCNet [38]	66.34	62.51	68.15	74.09	89.12	84.43	88.38	83.86
MMD-ReID [14]	66.75	62.25	71.64	75.95	95.06	**88.95**	**93.65**	**87.3**
SMCL [31]	67.39	61.78	68.84	75.56	83.93	79.83	83.05	78.57
CAJ [35]	69.88	66.89	76.26	80.37	85.03	79.14	84.75	77.82
MPANet [33]	70.58	68.24	76.74	80.95	83.7	80.9	82.8	80.7
MMN [39]	70.6	66.9	76.2	79.6	91.6	84.1	87.5	80.5
DCLNet [24]	70.8	65.3	73.5	76.8	81.2	74.3	78	70.6
MAUM [21]	71.68	68.79	76.97	81.94	87.87	85.09	86.95	84.34
CMT [15]	71.88	68.57	76.9	79.91	**95.17**	87.3	91.97	84.46
CM-EMD [18]	73.39	68.56	**80.53**	**82.71**	94.37	88.23	92.77	86.85
SEFL [8]	**75.18**	**70.12**	78.4	81.2	92.18	86.59	91.07	85.23
Ours (DEN)	**76.36**	**71.3**	**83.56**	**84.65**	**95.34**	**90.21**	**94.98**	**90.24**

Table 2. Comparisons of our method with state-of-the-art methods on SYSU-MM01 [32] and RegDB [23] datasets. Rank-1 accuracy (%) and mAP (%) are reported. The best results and the second are in Red and Blue, respectively.

4.2. Comparison with State-of-the-Art Methods

We compare the proposed DEN with state-of-the-art methods. Table 2 illustrates DEN achieves superior performance compared to the existing methods on the SYSU-MM01 and RegDB datasets. On the SYSU-MM01 dataset, DEN outperforms existing methods in both indoor-search and all-search, particularly achieving the Rank-1 accuracy of 83.56% and mAP of 84.65% in the indoor-search. The results presents 3.03% and 1.94% improvement over the performance of the previous best CM-EMD [18]. On the RegDB dataset, DEN achieves the Rank-1 accuracy of 94.98% and mAP of 90.24% in the infrared-to-visible, higher than the previous best MMD-ReID [14] by 1.33% and 2.94%. It's worth noting that MCLNet [9] and MMD-ReID [14] are methods that also consider the intra-identity representation space, similar to ours. Despite having a similar concept compared to the former methods, our method successfully reduces the modality gap and outperforms them. Particularly, DEN achieves higher performance than MMD-ReID by Rank-1 accuracy of 9.61% and mAP of 9.05% in the challenging All-search of the SYSU-MM01 dataset.

4.3. Ablation Study

In this section, we conduct ablation study to evaluate the effectiveness of each component of our model. As shown as Table 3, We analyze the key components of equation 5 and equation 9. Intra-identity Representation Diversification (IRD) consists of Positive Enhancement (PE) and Negative Enhancement (NE).

idx	IRD(V,I)		HG	IRD(HG,I)			R-1	mAP
	NE	PE		NE	PE	CI		
1							64.95	62.47
2	✓						66.29	63.72
3		✓					67.87	64.36
4	✓	✓					69.13	64.79
5			✓				71.02	67.46
6	✓	✓	✓				72.73	69.38
7	✓	✓	✓	✓	✓		73.92	69.68
8	✓	✓	✓			✓	73.15	69.21
9	✓	✓	✓	✓	✓	✓	**76.36**	**71.30**

Table 3. Performance comparison of each component of our model on the SYSU-MM01 dataset (All-search). Index 1 represents the baseline, which employs two-stream networks [36], Random Erasing [43], Identity loss and Hardest Triplet loss.

Visible-Infrared Intra-identity Representation Diversification Loss. Index 2 and 3 present the results obtained by adding NE(V,I) and PE(V,I) to the baseline, respectively. Index 2 shows less performance improvement than index 3. This is because NE(V,I) does not consider intra-instance variance, focus on dominant features. On the other hand, index 3, which considers intra-instance variance, achieves greater performance enhancement. In index 4, the inclusion of NE(V, I) and PE(V, I) enables the model to learn various discriminative representations effectively and reduce the modality gap. Consequently, Rank-1 accuracy increased by 4.18%, and mAP by 2.32% than baseline.

HueGray Transform. Index 5 and 6 represent the results obtained by adding HG to the baseline and index 4, respectively. These results demonstrate that performance is improved by increasing intra-instance variance through the application of HG. Specifically, in index 6, the achieved performance is 3.60% higher for Rank-1 accuracy and 4.59% higher for mAP compared to index 4. This highlights that IRD expands the representation space and enables the model to effectively learn the increased intra-instance variance.

HueGray-Infrared Intra-identity Representation Diversification Loss. Index 7 incorporates NE(HG, I) and PE(HG, I) on augmented samples using HG, resulting in an improvement of 1.19% for Rank-1 accuracy and 0.30% for mAP compared to index 6. Index 8 and 9 represent the results obtained by adding CI(HG,V) to index 6 and 7, respectively. Index 8, there is marginal improvement. However, index 9, both Rank-1 accuracy and mAP show notable improvements, with an increase of 3.63% for Rank-1 accuracy and 1.92% for mAP compared to index 6. This is because CI is designed to capture the missing relationships between Visible-HueGray samples, suppressing color-dominant features and enforcing shape-related features.

Summary. The Intra-identity Representation Diversification method expands the representation space to effectively learn the increased intra-instance variance. Additionally, it enables the model to learn the relationships between hetero-modality and successfully reduces the modality gap. Finally, compared to the baseline, our model increases Rank-1 accuracy by 11.41% and mAP by 8.83%, achieving 76.36% and 71.30%, respectively.

4.4. comparison of Intra Representation space size

In this section, we compare the variations in the size of the IIRS. We use all test data from the SYSU-MM01 dataset.

idx	Loss	Sample Diversity	IIRS (V,V)	IIRS (I,I)	R1	mAP
1	WRT	RE,Gray	0.411	0.431	59.93	59.2
2	**HT**	**RE,Gray**	**0.421**	**0.450**	**64.37**	**63.14**
3	WRT	RE,Hue	0.455	0.462	64.79	62.18
4	**HT**	**RE,Hue**	**0.458**	**0.467**	**67.29**	**65.34**
5	WRT	RE,HG	0.449	0.450	66.34	62.58
6	**HT**	**RE,HG**	**0.468**	**0.467**	**71.02**	**67.46**

Table 4. Comparison of IIRS with Sample Diversity consists of Random Erasing (RE) [43], Hue Transform (Hue), and Gray Transform (Gray). The bold mark indicates larger value between WRT and HT. We use AGW [36] as a base without the triplet loss.

Table 4 illustrates the differences when applying Sample Diversity to WRT and HT, as discussed in Section 3.1. Overall, HT exhibits a larger IIRS than WRT and achieves higher performance in both Rank-1 accuracy and mAP. Specifically, as the input samples become more diverse, the performance improvement is relatively small with WRT in from idx 3 to idx 5 (R-1: +1.55%, mAP: +0.4%), and the IIRS decreases. However, with HT in from idx 4 to idx 6, there is a significant performance improvement (R-1: +3.73%, mAP: +2.12%), and the IIRS is either increased or maintained. Therefore, to effectively reduce the modality gap for diverse samples, HT shows better results.

idx	Loss	SD	IIRS (V,V)	IIRS (I,I)	R1	mAP
1	HT	RE	0.462	0.476	64.95	62.47
2	HT,NE	RE	0.447	0.458	66.29	63.78
3	HT,PE	RE	0.504	0.505	67.87	64.38
4	HT,IRD	RE	0.520	0.523	69.13	64.79
5	**HT,IRD**	**RE,HG**	**0.531**	**0.528**	**72.73**	**69.38**

Table 5. Comparison of IIRS with Intra-identity Representation Diversification Loss (IRD), consisting of Negative Enhancement Loss (NE) and Positive Enhancement Loss (PE). SD denote the Sample Diversity.

Table 5 shows the variations in the IIRS when using the proposed Intra-identity Representation Diversification (IRD) loss described in Section 3.2. Compared to idx 1, which is the base model, idx 3 with PE shows greater performance improvement than idx 2 with NE. The reason for this is that NE alone is not as effective in reducing the modality gap compared to PE, as it only pulls the inter-modality intra-identity resulting in a smaller representation space and limited diversity in representations. However, when PE and NE are used together, such as in idx 4, NE plays an auxiliary role to PE, resulting in a larger IIRS and a boost in performance. As a result, it can be observed that both the performance and IIRS have increased. Specifically, when increasing Sample Diversity in idx 5, the IIRS expands, enabling the learning of diverse representations.

Method	IIRS (V,V)	IIRS (I,I)	HIMD (V,I)	R1	mAP
Base	0.462	0.476	0.849	64.95	62.47
CAJ	0.498	0.501	0.849	69.88	66.89
MMD-ReID	0.508	0.51	0.837	66.75	62.25
IRD(V,I)	**0.531**	**0.528**	**0.802**	**72.73**	**69.38**

Table 6. Comparison of IIRS and HIMD with existing methods of Base, CAJ [35] and MMD-ReID [14]. HIMD is Hetero-modality Intra-identity Max Distance. Base is AGW [36] method, which replaces WRT with HT.

As shown in Table 6, the proposed method effectively utilizes IIRS compared to existing methods while also achieving superior performance. Additionally, the proposed method exhibits the smallest HIMD, which calculates the average distance between hard features of hetero-modality intra-identities. This shows that the proposed method enables diverse representation with the expansion of IIRS, while also being robust to hard samples and effectively reducing modality-gap. CAJ, which employs channel augmentation, increases both IIRS and performance compared to Base. However, the HIMD remains the same, indicating its effectiveness for easy samples while not performing well for hard samples. Metric learning-based MMD-ReID shows a larger IIRS than Base and CAJ, but exhibits a smaller HIMD, leading to lower performance, especially. While this approach can cope with certain hard samples, its effectiveness is limited because it does not account for intra-instance variance. In contrast, our proposed method employs a carefully crafted metric that not only tackles intra-instance variance robustly but also performs well in handling hard samples and significantly reducing the modality gap.

4.5. Qualitative Visualization Comparison

In this section, we visualize the top-10 retrieval results, the feature distributions via t-SNE [27] and the activation maps using GradCAM++ [2] on SYSU-MM01 dataset.

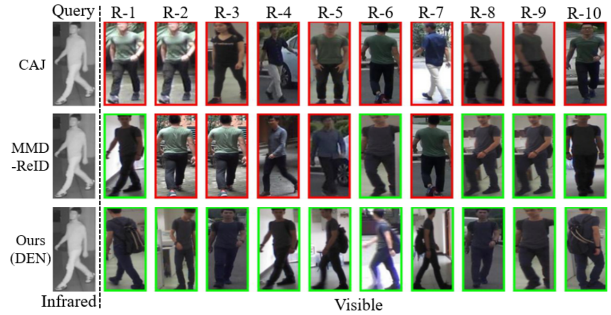


Figure 3. Comparison of the top-10 retrieval results of CAJ [35], MMD-ReID [14] and Ours. The green boxes denote correct matches, and the red boxes denote incorrect matches.

Retrieval results. Figure 3 illustrates that the proposed method outperforms existing methods in retrieval results for hard samples. Data augmentation based CAJ focuses only on dominant features due to metric learning without consider intra-instance variance, this retrievals a incorrect person scene (front appearance) similar to Query. While MMD-ReID proposes metric learning to promote the robustness of features, it cannot learn discrimination between features that have become too close. As a result, as shown in Figure 3, incorrect matches are placed at the front rankings. In contrast, the proposed method shows varied feature representation capabilities through metric learning, enabled by the flexible expansion of IIRS to address intra-instance variance. As a result, various correct scenes (front, back, and side appearances) are matched at the top of the retrieval rankings. Additional retrieval results are outlined in the supplementary section.

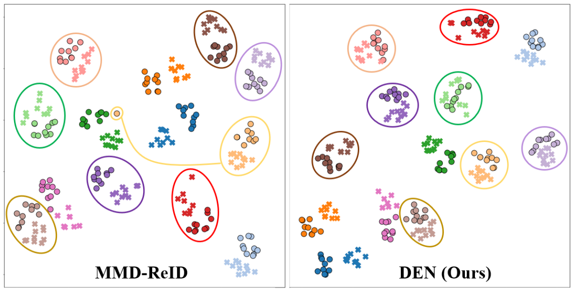


Figure 4. The feature distributions using t-SNE are displayed, where ● and **X** represent different modalities, and colors denote the same identity.

Feature distributions. In Figure 4, DEN exhibits a relatively reduced modality gap compared to MMD-ReID. Notably, the green circle of ours shows a complete overlap of cross-modalities within the same space, as opposed to distinct clustering based on their modalities.

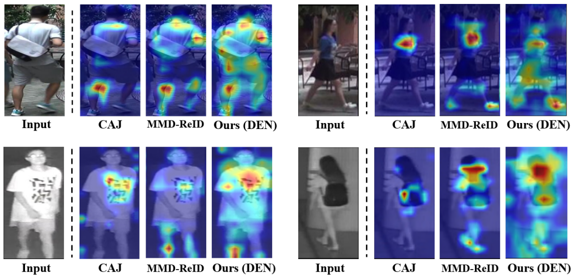


Figure 5. The activation maps using Grad-CAM++. Row 1 and 2 are images of visible and infrared, respectively.

Activation maps. In Figure 5, CAJ and MMD-ReID indicate activation primarily in specific body parts, while DEN shows activation across the entire body. Thus, our metric learning for enlarging IIRS can effectively represent diverse intra-identities.

5. Conclusion

Our study addresses the challenges of visible-infrared person re-identification (VI-ReID) by focusing on the expansion of the Intra-modality Intra-identity Representation Space (IIRS) and the learning of discriminative features for intra-instance variance. Our proposed Intra-identity Representation Diversification (IRD) loss and Diversity Enhancement Network (DEN) contribute to the enlargement of the IIRS and the efficient learning of diverse and discriminative representations. Experimental results on the SYSU-MM01 and RegDB datasets demonstrate the superiority of our method, achieving remarkable Rank-1 accuracy of 76.36% and mAP of 71.30% on the challenging SYSU-MM01 dataset. This highlights the significance of our approach in advancing the field of VI-ReID and provides a valuable foundation for future cross-modality research.

References

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

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018. 8

[3] Xiaozhou Cheng, Rui Li, Yanjing Sun, Yu Zhou, and Kaiwen Dong. Gray augmentation exploration with all-modality center-triplet loss for visible-infrared person reidentification. *IEICE TRANSACTIONS on Information and Systems*, 105(7):1356–1360, 2022. 1, 3

[4] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person reidentification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10257–10266, 2020. 1

[5] Pingyang Dai, Pingyang Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, page 6, 2018. 1

[6] Xing Fan, Wei Jiang, Hao Luo, and Weiji Mao. Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal person re-identification. *Visual Computer*, pages 1–16, 2022. 1

[7] Xing Fan, Hao Luo, Chi Zhang, and Wei Jiang. Cross-spectrum dual-subspace pairing for rgb-infrared cross-modality person re-identification. In *arXiv preprint arXiv:2003.00213*, 2020. 1

[8] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023. 1, 6

[9] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person reidentification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 16403–16412, 2021. 2, 6

[10] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8385–8392, 2019. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. In *arXiv preprint arXiv:1703.07737*, 2017. 3

[13] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1034–1042, 2022. 1

[14] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. Mmd-reid: A simple but effective solution for visible-thermal person reid. In *British Machine Vision Conference*, 2021. 2, 6, 7, 8

[15] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 480–496, 2022. 6

[16] Sejun Kim, Sungjae Kang, Hyomin Choi, Seong Soo Kim, and Kisung Seo. Keypoint aware robust representation for transformer-based re-identification of occluded person. *Proceedings of the IEEE Signal Processing Letters*, pages 65–69, 2023. 1

[17] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross modal person re-identification with an x modality. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4610–4617, 2020. 1

[18] Yongguo Ling, Zhun Zhong, Zhiming Luo, Fengxiang Yang, Donglin Cao, Yaojin Lin, Shaozi Li, and Nicu Sebe. Cross-modality earth mover's distance for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1631–1639, 2023. 6

[19] Haijun Liu, Jian Cheng, Wen Wang, Yanzhou Su, and Haiwei Bai. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*, 398:11–19, 2020. 3

[20] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23:4414–4425, 2020. 3

[21] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022. 6

[22] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020. 6

[23] Dattien Nguyen, Hyunggil Hong, Kiwan Kim, and Kangryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 5, 6

[24] Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5333–5341, 2022. 6

[25] Jia Sun, Yanfeng Li, Houjin Chen, Yahui Peng, Xiaodi Zhu, and Jinlei Zhu. Visible-infrared cross-modality person re-identification based on whole-individual training. *Neurocomputing*, 440:1–11, 2021. 1

[26] uanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020. 3, 6

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, pages 2579–2605, 2008. 8

[28] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12144–12151, 2020. 1

[29] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–626, 2019. 1

[30] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4676–4687, 2021. 1

[31] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021. 1, 6

[32] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 5, 6

[33] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 6

[34] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional centerconstrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019. 1

[35] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 1, 5, 6, 7, 8

[36] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3, 6, 7

[37] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020. 1

[38] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmc-net: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022. 1, 6

[39] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021. 6

[40] Zhiwei Zhao, Bin Liu, Qi Chu, Yan Lu, and Nenghai Yu. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3520–3528, 2021. 1

[41] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 1

[42] Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1418–1430, 2021. 1

[43] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 2, 3, 6, 7

[44] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2018. 1