Person Re-Identification Method Based on Information-Balanced Identity Learning

Xin-Heng Li[1,2,3][0000−0001−8496−0049], Dan Chen[1,2,3][0009−0006−6777−6040]
Xin Wen[1,2,3][0009−0006−0735−5113], and Zhen-Tao Liu[1,2,3,*]

[1] School of Automation, China University of Geosciences, Wuhan 430074, China
[2] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China
[3] Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China
{Xin-Heng Li, Dan Chen, Xin Wen, Zhen-Tao Liu}@liuzhentao@cug.edu.cn

Abstract. Two key challenges remain unaddressed well in Visible - Infrared Person Re-Identification (VI-ReID). The first is the information imbalance between two modalities. As infrared modality contains much less information than visible modality, infrared images become hard examples for models to learn. This potentially leads to overfitting on visible modality and is harmful to model performance. The second is discriminative features. Conventional methods mainly concentrate on mitigating the modality gap while ignoring mining the identity-informative features. In addressing these challenges, we propose DI2L, a VI-ReID approach incorporated with advanced representation learning and metric learning techniques. DI2L first employs a channel augmentation module to generate auxiliary images for model robustness. Subsequently, it adopts a weighted part aggregation (DPA) module to obtain discriminative features by exploring the relationship between different parts of features. Then a information-balanced identity learning (I2L) module is proposed to address the information imbalance issue while searching for discriminative part features. Comprehensive experiments demonstrate the superiority and effectiveness of our approach against the SOTA methods on two commonly used datasets SYSU-MM01 and RegDB.

Subsequently, it adopts a weighted part aggregation (DPA) module to obtain discriminative features by exploring the relationship between different parts of features. Then a information-balanced identity learning (I2L) module is proposed to address the information imbalance issue while searching for discriminative part features. Comprehensive experiments demonstrate the superiority and effectiveness of our approach against the SOTA methods on two commonly used datasets SYSU-MM01 and RegDB.

Keywords: Cross-modality person re-identification, Information imbalance, Metric Learning, Representation Learning

1 Introduction

Person re-identification (ReID) aims at matching the images of a given person from a large gallery captured by multiple non-overlapping cameras. It is a core technology in numerous computer vision applications, including video surveillance analysis and intelligent transportation. During the past decade, most research efforts for person ReID have focused on the single-modality visible domain [1, 5, 8, 23], and have shown encouraging outcomes on many publicly available datasets. However, the assumption of a single visible modality inevitably limits the applicability of this technology.

Practical 24-hour surveillance systems usually deploy infrared cameras as compensation for visible cameras because visible cameras could not provide enough discriminative information under poor lighting conditions. This raises the important visible-infrared cross-modality person re-identification (VI-ReID) task. In the VI-ReID task, given a infrared (or visible) image of a specific person, the goal is to retrieve all visible (or infrared) images corresponding to the same identity.

Matching these infrared images directly to visible light images for ReID poses an additional challenge due to the inter-modality variation. Existing approaches mainly focus on mapping the images of different modalities into a shared feature space to mitigate the inter-modality variations [2, 3, 9], which ignore the important information imbalance issue. As the visible modality contains much more information and usually has more samples than the infrared modality, infrared images become hard examples to identify. This harmful information imbalance issue potentially leads to over-fitting on visible modality. Recent SOTA methods are based on part-based person representation that aims to extract discriminative human parts information. However, these methods usually focus only on the most discriminative part rather than the diverse parts which are helpful to distinguish different persons.

In addressing these challenges, we propose DI2L, a novel VI-ReID algorithm that improves the state-of-the-art backbone model from both representation learning and metric learning perspectives. It first employs a channel augmentation module to generate auxiliary images, mitigating both intra-modality and inter-modality variations. Then it incorporates a discriminative part feature aggregation (DPA) module and information-balanced identity learning (I2L) to address the information imbalance issue while searching for discriminative part features. The primary contributions of this paper are outlined below.

- DI2L adds a DPA module to a two-stream network model with a channel-augmented auxiliary modality. The DPA module explores the relationships between different local parts, improving the capability to extract discriminative features.
- DI2L incorporates an I2L module that imposes novel weighted-part focal loss, combining the weighted regularized triplet loss to simultaneously address information imbalance and modality variations.

- Comprehensive experiments on two popular datasets SYSU-MM01 and RegDB demonstrate the superiority and effectiveness of our approach against the state-of-the-art methods.

The remainder of the paper is organized as follows. Related works for Visible-Infrared Person Re-Identification (VI-ReID) are given in Section 2. Section 3 introduces a methodology for DI2L, i.e., modality-specific feature extrator, cross-modality feature extractor, and Information-balanced Learning (I2L). Section 4 introduces datasets, evaluation metrics, and implementation details of our experiments. Experiments and results analysis are described in Section 5.

2 Related Work

Conventional methods for VI-ReID mainly try to learn modality-relevant features. The key to solving the VI-ReID problem lies in learning shared features between the two modalities, which can reduce the differences between them. Two approaches i.e., representation learning and metric learning, have been commonly used, and with further research on the cross-modal pedestrian re-identification problem, modal inter-modal learning approaches have emerged.

The representation learning approach centers on the study of designing a suitable network model to extract discriminative and modality-independent pedestrian information between visible and infrared images, inputting the two modal images into the network, and comparing the similarity between them. To simultaneously handle the intra-modality and cross-modality variations, [16] proposed a bi-directional dual-constrained framework, combining both feature learning and metric learning.

It is difficult to effectively extract more image information by training the network using only global or local features, which causes the network to be overly dependent on certain information and affects the accuracy of the algorithm. The metric learning approach aims to train the network to learn the similarity between images, and focuses on designing a suitable loss function so that the distance spacing of the sample images in the same row of the human category in both modalities is smaller than the distance spacing of the sample images in different rows of the human category.

The learning method based on modal interconversion is to convert two modal images into the same modal image as a way to reduce the difference between the two modalities. To fully utilize the relations across two modalities, [18] designed a dual-attentive aggregation learning method. With the advancement of GANs, a dual-level discrepancy modeling method [13] generates cross-modality images, eliminating discrepancy at the pixel level. More recently, MAUM [9] has used modality-specific classifiers to learn the modality-specific proxies. However, the image generation process introduces unavoidable noise, which greatly affects the effectiveness of VI-ReID.

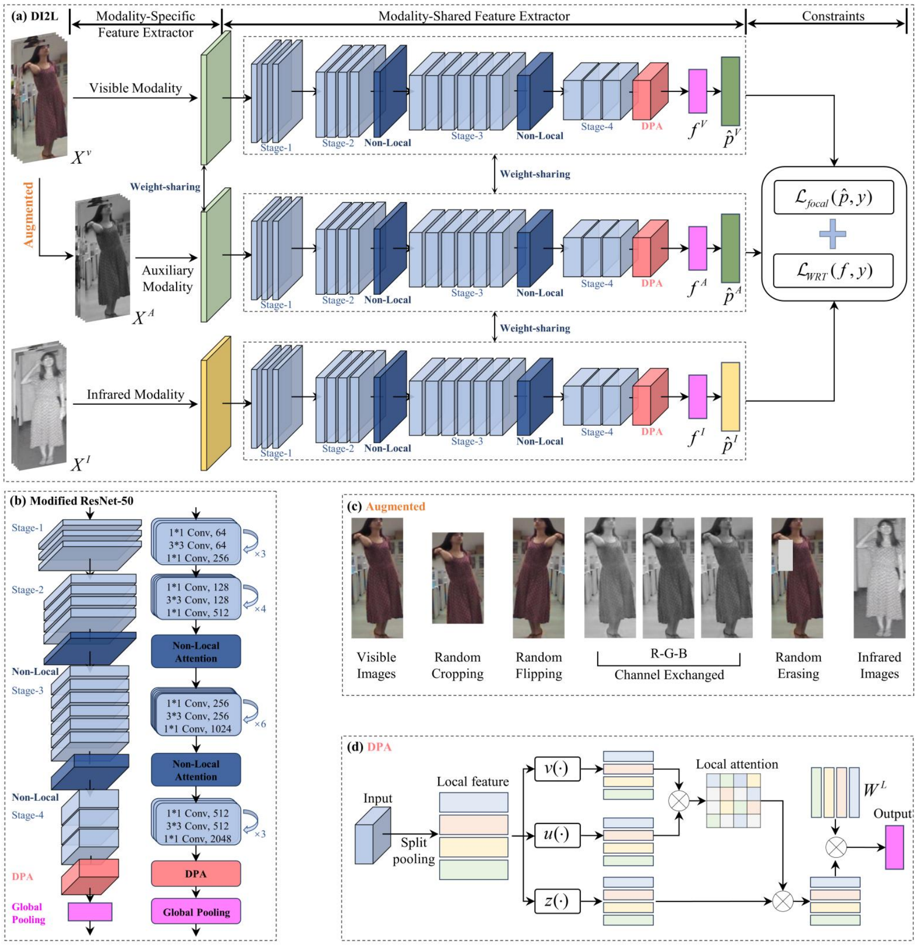


Fig. 1. (a) is the overall framework diagram of DI2L. (b) describes the detailed network structure of Modality-Shared Feature Extractor. (c) shows some examples of these augmentation operations. (d) illustrates the detailed structure of the DPA module.

3 Methodology

As shown in Fig. 1 (a), the input images, including the original visible images, the random channel augmented visible images, and infrared images, are first fed into the two-stream network to extract modality-specific features, as introduced in Section 3.1. Then these single-modality features are fed into a modality-shared feature extractor for the generation of effective discriminative features through a DPA module developed in Section 3.2. Finally, to better leverage the valuable identity information while addressing information imbalance, the I2L module combines a novel local attentive focal loss and weighted triplet loss to train the model, proposed in Section 3.3.

3.1 Modality-specific Feature Extrator

The three-channel color visible images contain rich and useful appearance information. However, directly recovering a three-channel visible image from a single-channel infrared image is quite challenging. Previous studies usually convert RGB images to grayscale images or leverage GANs to generate auxiliary images to tackle this challenge. However, GAN-based methods [13] introduce additional computational costs and unavoidable noises, while image graying may lose discriminative information.

Unlike these methods, as shown in Fig. 1 (c), DI2L applies random channel augmentation (RCA) to visible images instead of dropping color cues. Through decomposing the color channels of the three-channel visible images, RCA attempts to mine the relationship between each channel (R, G, or B) and the single-channel infrared images. Specifically, RCA first randomly selects one channel (R, G, or B) to replace the other channels, formulated as

$$\tilde{X}_i^{V,c} = \left(X_i^c, X_i^c, X_i^c\right), \quad c \in \{R, G, B\} \tag{1}$$

Afterward, these channel-exchanged images are applied with the following data augmentation operations with each probability $p = 0.5$ in order: random flipping, random cropping, and random erasing. Taking the original visible images, the channel augmented images, and the infrared images as inputs, DI2L first generates individual features for each single modality.

3.2 Cross-Modality Feature Extractor

DI2L then utilizes a share-weight cross-modality feature extractor to project the features from different modalities into the same subspace. Notably, aiming at thoroughly mining the identity information and abstracting discriminate features, DI2L incorporates a non-local attention mechanism with a DPA module to explore the relationship between different local features. As shown in Fig. 1 (b), we highlight the novelty of the proposed method compared to the popular ResNet-50 backbone using blocks with different colors: Firstly, we add a pixel-level non-local attention layer (dark-blue block) after the second and third convolution groups (stage-2 and stage-3) of the backbone. Secondly, a DPA module (pink block) is introduced after the fourth convolution group.

Non-local Attention Block Inspired by [21], we incorporate a pixel-level non-local attention mechanism after Stage 2 and Stage 3 of the backbone to obtain non-local mixed attention. It can be formulated as

$$a_i = \left(\sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right) \right) \Big/ \sum_{\forall j} f\left(x_i, x_j\right), \tag{2}$$

where i is the output feature position index and j is the index that enumerates all possible positions. x and a are the input and output. The function $g(x_j) = W^g{\cdot}x_j$ computes a representation of the input at the position j. The pairwise function $f(x_i, x_j)$ computes the relationship between x_i and x_j. In this paper, we adopt the Gaussian pairwise function formulated by

$$f\left(x_i, x_j\right) = \exp\left(u\left(x_i\right)^T v\left(x_j\right)\right) = \exp\left(\left(W^u x_i\right)^T W^v x_j\right),$$

where W^u and W^v are weight matrices. Then the output z at position i is calculated as

$$z_i = W^z \operatorname{softmax}\left(\left(W^u x_i\right)^T W^v x_j\right) g\left(x_j\right) + x_i, \tag{3}$$

where W^z is a weight matrix. For a given i, softmax computation along dimension j.

DPA Besides the pixel-level attention, as shown in Fig. 1 (d), DI2L attempts to mine the relationship between different body parts, introducing a DPA module. The input of DPA module is the extracted feature maps from the last residual block of the network, from which we extract the attention-enhanced part features. The output feature maps of the Stage-4 can be denoted as $\{X = x_k \in \mathbb{R}^{C\times H\times W}\}^K$ where C represents the channel dimension, H and W represent the feature map size, and K represents the batch size. To obtain the part features, the feature maps are directly divided into p non-overlapping parts with a region pooling strategy, denoted by $X^p = \{x_i^p \in \mathbb{R}^{C\times 1}\}_{i=1}^p$. The local features are fed into three convolutional layers $u(\cdot)$, $v(\cdot)$, and $z(\cdot)$. The local attention $\alpha_{i,j}^p \in [0,1]^p p$ is then calculated as

$$\alpha_{i,j}^p = \frac{f\left(\mathbf{x}_i^p, \mathbf{x}_j^p\right)}{\sum_{\forall j} f\left(\mathbf{x}_i^p, \mathbf{x}_j^p\right)}, \tag{4}$$

where $f\left(\mathbf{x}_i^p, \mathbf{x}_j^p\right) = \exp\left(u\left(\mathbf{x}_i^p\right)^T v\left(\mathbf{x}_j^p\right)\right)$ is an exponential function that enlarges the part attention discrepancy [19]. With the learned part attention, attention-enhanced part features are then represented by $\hat{x}_i^p = a_i^p \cdot z(x_i^p)$. These local features are aggregated together to generate the final feature $\hat{x}$. In this way, the refined part features consider the relationship between different body parts.

3.3 Information-balanced Learning (I2L)

In the VI-ReID literature, combining triplet loss and identity loss is one of the most popular solutions for deep Re-ID model learning [1,19]. In this section, we introduce I2L, which combines local attentive focal identity loss and a weighted regularization triplet loss to address the information imbalance between different modalities.

Focal identity loss We propose a local aggregated focal identity loss $\mathcal{L}_{focal}$ based on [7]. Specifically, $\mathcal{L}_{focal}$ reshapes the commonly used standard cross entropy (CE) loss [17, 25] by down-weighting the loss assigned to well-classified examples, denoted as

$$\mathcal{L}_{focal} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} -\big(1-p(y_j|\hat{x}_i)\big)^{\gamma}\log\big(p(y_j|\hat{x}_i)\big), \tag{5}$$

where N is the number of images of the mini-batch, K is the number of identities, $p(y_j|f_i)$ represents the probability of the i-th local aggregated feature f_i being correctly classified into the ground-truth label y_j, $(1-p_{ij})^{\gamma}$ is a modulating factor, with tunable focusing parameter $\gamma \leq 0$ When $\gamma = 0$, focal loss is equivalent to the CE loss.

Weighted triplet loss Following [17], we adopt the weighted regularization triplet loss (WRT) to optimize the relative distance between all the positive and negative pairs. Compared to the commonly used hard triplet loss, WRT loss sample triplets consider their contributions, improving the robustness against modality variations.

$$\mathcal{L}_{WRT} = \frac{1}{N}\sum_{i=1}^{N}\log\Big(1+\exp\Big(\phi\Big(\sum_{ij} w^p_{ij}d^p_{ij} - \sum_{ik} w^n_{ik}d^n_{ik}\Big)\Big)\Big), \tag{6}$$

$$w^p_{ij} = \frac{\exp\big(d^p_{ij}\big)}{\sum_{d^p_{ij}\in\mathcal{P}_i}\exp\big(d^p_{ij}\big)}, w^n_{ik} = \frac{\exp\big(-d^n_{ik}\big)}{\sum_{d^n_{ik}\in\mathcal{N}_i}\exp\big(-d^n_{ik}\big)},$$

$$\phi(x) = sgn(x)\cdot x^2,$$

where (i, j, k) represents a triplet within each training batch for each anchor sample x_i. For anchor x_i, $\mathcal{P}_i$ / $\mathcal{N}_i$ is the corresponding positive/negative set. d^p_{ij}/d^n_{ik} represents the pairwise distance of a positive/negative sample pair. d_{ij} is the Euclidean distance between two samples.

We mix the loss function based on the principle that the classification and the triplet share the same importance. Therefore, the overall loss functions are as follows.

$$\mathcal{L} = \mathcal{L}_{focal} + \mathcal{L}_{WRT}$$

4 Experimental Setup

4.1 Datasets

We conduct experiments on two public datasets, i.e., SYSU-MM01 and RegDB.

- **SYSU-MM01** is a challenging, large-scale dataset collected by four visible cameras and two near-infrared cameras [14]. It contains a total of 30,071 visible images and 15,792 infrared images of 491 identities. Each identity is captured by at least one visible camera and one near-infrared camera. The dataset is divided into a training set with 395 identities and a testing set with 96 identities.

- **RegDB** consists of 412 individuals and was captured using dual cameras [10]. Each person has 10 visible images taken by a visible camera and 10 infrared images captured by a far-infrared camera. Following [19], the dataset was randomly divided into two halves, one for training and the other for testing. The query set consists of 2,060 infrared images, and the gallery set contains 2,060 visible images.

4.2 Evaluation metrics

We adopted two widely used metrics, *Cumulative Matching Characteristics (CMC)* [12] and *mean Average Precision (mAP)* [24], to evaluate the performance. CMC-k (i.e., Rank-k matching accuracy) [12] represents the probability that a correct match appears in the top-k ranked retrieved results. mAP measures the average retrieval performance with multiple ground truths. It is originally widely used in image retrieval. For Re-ID evaluation, it can address the issue of two systems performing equally well in searching for the first ground truth. For reliable outcomes, the evaluation process is repeated 10 times with randomly sampled query and gallery sets [13, 17].

4.3 Implementation Details

The proposed approach is implemented with PyTorch a computer of the 64-bit Windows 10 system, equipped with two Intel Xeon Silver 4214 CPU, 64 GB memory and two NVIDIA RTX A6000 48G GPU. We adopt the ResNet50 [4] model pre-trained on ImageNet as the backbone Network. We adopt the SGD optimizer and the initial learning rate is set to be 0.01. The training epoch is set to 120, and the learning rate decays at the 20th and 50th epochs with a decay factor of 0.1. The images are resized to $3 \times 288 \times 144$. For infrared images, the three channels are the same. For each mini-batch, 8 identities are randomly selected, and for each selected identity, 4 visible images and 4 infrared images are sampled. As aforementioned, we additionally generate 4 auxiliary images from the original visible images using RCA. Consequently, we have $8 \times 4 \times 3$ images for each mini-batch.

5 Experimental Results

5.1 Comparison with State-of-the-art Methods

We compare our approach with state-of-the-art methods in VI-ReID published in the last three years. The comparison results of the SYSU-MM01 and RegDB datasets are listed in Table 1. In the challenging single-shot mode of SYSU-MM01 dataset, the proposed method achieves **70.42**% Rank-1 and **59.52**% mAP, significantly outperforming the state-of-the-art methods, including recently published

Table 1. Comparison results (%) with the SOTA on SYSU-MM01 and RegDB datasets.

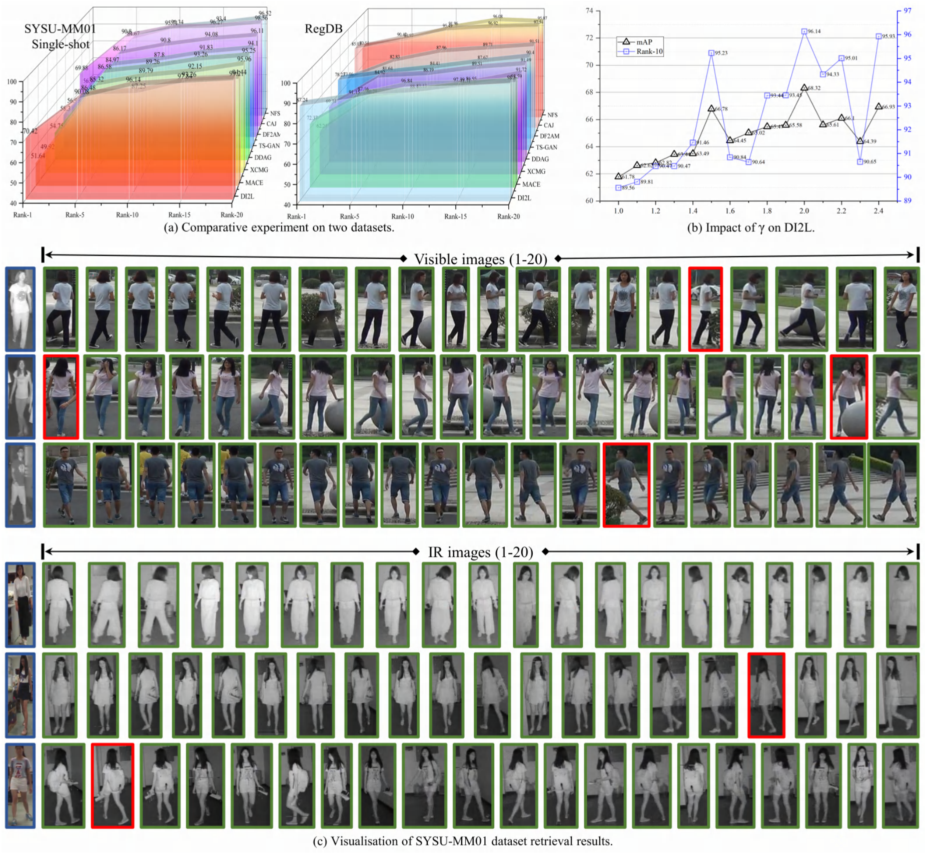
Method	SYSU-MM01								RegDB			
	Single-shot				Multi-shot							
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-padding [14]	14.8	54.12	71.33	15.95	19.13	61.4	78.41	10.89	17.75	54.12	71.33	15.95
D^{2}RL [13]	28.9	70.6	82.4	29.2	-	-	-	-	43.4	66.1	76.3	44.1
CDP+DHSM [3]	38	82.3	91.7	38.4	-	-	-	-	65.3	84.5	91	62.1
AlignGAN [11]	42.4	85	93.7	40.7	-	-	-	-	56.3	-	-	53.4
AGW [19]	47.5	-	-	47.65	-	-	-	-	70.05	-	-	66.37
MACE [15]	51.64	87.25	94.44	50.11	-	-	-	-	72.37	88.4	93.59	69.09
XCMG [6]	49.92	89.79	95.96	50.73	-	-	-	-	62.21	83.13	91.72	60.18
DDAG [18]	54.75	89.26	95.25	53.02	-	-	-	-	69.34	86.19	91.49	63.46
TS-GAN [22]	58.3	87.8	94.1	55.1	55.9	91.2	96.6	39.7	78.23	84.41	90.4	66.1
DF^{2}AM [20]	56.93	90.8	96.11	55.1	-	-	-	-	73.06	87.96	91.51	67.81
CAJ [17]	69.88	95.71	98.56	66.89	62.1	95.74	97.45	60.02	85.03	95.49	97.54	79.14
NFS [2]	56.91	91.34	96.52	55.45	63.51	94.42	97.81	48.56	80.54	91.96	95.07	72.1
MAUM [9]	61.59	-	-	59.96	-	-	-	-	83.39	-	-	78.75
DI2L(ours)	**70.42**	**96.14**	**99.21**	**68.32**	**63.71**	**96.22**	**98.31**	**61.54**	**87.24**	**96.84**	**98.50**	**81.76**

CAJ [17], NFS [22], and MAUM [9]. Our method also achieves the best performance in the multi-shot mode. On the RegDB dataset, our method outperforms other baseline methods by a large margin (at least 4% improvement on mPA and Rank-1). Specifically, we achieve **88.89**% Rank-1 and **83.38**% mAP. These results demonstrate the effectiveness of DI2L.

Table 2. Ablation study on the SYSU-MM01 dataset.

Method	Single Shot				Multi-shot			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Baseline	69.88	95.71	98.56	66.89	62.10	95.74	97.45	60.02
Baseline+DPA	71.46	96.07	98.79	68.18	63.15	96.15	98.03	61.26
Baseline+DPA+I2L	70.42	96.14	99.21	68.32	63.71	96.22	98.31	61.54

Compared to GAN-based methods (D^{2}RL [13], AlignGAN [11], and TS-GAN [22]), the proposed method achieves an absolute gain of at least 10% and 8% in terms of Rank-1 and mPA on the two datasets. It does not introduce additional learnable parameters nor extra computation cost in both the train and test phases, which is more effective and efficient. Compared to other two-stream feature network-based methods XCMG [6], DDAG [18], and CAJ [17], our approach further utilizes information behind local features, extracting more effective modality-shared features. With our modified backbone, our approach improves the mAP and Rank-1 accuracy on the two datasets. Compared with recently proposed NFS [2]] and MAUM [9], our method leverages focal loss to alleviate the modality information imbalance, bringing at least 9.56% and 4.63% mAP improvements on SYSU-MM01 and RegDB respectively.



(a) Comparative experiment on two datasets.

(b) Impact of γ on DI2L.

(c) Visualisation of SYSU-MM01 dataset retrieval results.

Fig. 2. (a) is comparative experiment on the retrieval accuracy of the top 20 images on two datasets. (b) describes the impact of γ on DI2L. (c) shows the visualisation of SYSU-MM01 dataset retrieval result, the red frame represents an incorrectly retrieved image and the green frame represents a correctly retrieved image.

5.2 Ablation Study

We evaluate the effectiveness of the key components introduced before, the DPA module and modality-balanced loss, on the large-scale SYSU-MM01 dataset. We adopt the state-of-the-art CAJ [17] network as our baseline model and show intermediate results as follows to highlight the contribution of these components toward the final performance.

- Baseline: The baseline model is trained using $\mathcal{L}_{CE} + \mathcal{L}_{WTR}$.
- Baseline+DPA: The baseline model is integrated with the proposed DPA module, and is trained using $\mathcal{L}_{CE} + \mathcal{L}_{WTR}$.
- Baseline+DPA+L$_{focal}$: The proposed whole framework, integrated with the proposed DPA module, is trained using $\mathcal{L}_{focal} + \mathcal{L}_{WTR}$.

As shown in Table 2, the DPA module brings a 1.5% mAP improvement to the baseline in the single-shot domain. Moreover, with I2L, the whole model brings an additional 1.08% Rank-1 improvement on the baseline. These results highlight the effectiveness of the proposed components of DI2L.

Parameter influence We evaluate the influence of the hyperparameter γ in the focal loss. Fig. 2 (b) shows the performance of Rank-1 and mAP on the SYSU-MM01 dataset by varying γ. We could find that when $\gamma = 2$, the model performs best. Moreover, we observe that the model performance is robust with different γ.

Visualization As shown in Fig. 2 (c) top, the retrieval results of the DI2L in the panoramic multi-camera search mode of the SYSU-MM01 dataset, an IR image is used as the query image, and other visible images under the pedestrian category of this IR image are found from the database. As shown in Fig. 2 (c) down, the retrieval result of DI2L algorithm in SYSU-MM01 dataset RGB-IR search mode.

It can be seen that the erroneous images retrieved by DI2L on the SYSU-MM01 dataset are mostly affected by background clutter or object occlusion, and are less affected by the pedestrian pose and silhouette information. On the other hand, on the RegDB dataset, DI2L pays attention to more detailed information to achieve accurate retrieval of the query image 2 with the detailed information of the pedestrian's backpack, coat and hat.

6 Conclusion

In this paper, DI2L is proposed for VI-ReID, which attempts to effectively learn the discriminative part features while handling the information imbalance between the two modalities. It first incorporates a DPA module to mine the relationship between different parts, then leverages an I2L module to focus on hard example learning. Extensive comparison experiments and ablation studies demonstrate the effectiveness of each component.

In future work, to address the problem of low-resolution of images, the improved resolution model is integrated in front of the cross-modal pedestrian re-recognition model, which enhances the information of pedestrian poses and contours in the image, and further improves the accuracy of the model. For the problem of occlusion and background interference, the existing uni-modal pedestrian re-identification method is used to solve the occlusion problem and enhance the model's anti-interference ability.

References

1. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1320–1329 (2017), https://api.semanticscholar.org/CorpusID:14795862
2. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 587–597 (2021), https://api.semanticscholar.org/CorpusID:233033720
3. Fan, X., Luo, H., Zhang, C., Jiang, W.: Cross-spectrum dual-subspace pairing for rgb-infrared cross-modality person re-identification. ArXiv **abs/2003.00213** (2020), https://api.semanticscholar.org/CorpusID:211677394
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), https://api.semanticscholar.org/CorpusID:206594692
5. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. ArXiv **abs/1703.07737** (2017), https://api.semanticscholar.org/CorpusID:1396647
6. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: AAAI Conference on Artificial Intelligence (2020), https://api.semanticscholar.org/CorpusID:214109021
7. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2999–3007 (2017), https://api.semanticscholar.org/CorpusID:47252984
8. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6638–6647 (2019), https://api.semanticscholar.org/CorpusID:203952464
9. Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., Li, W.: Learning memory-augmented unidirectional metrics for cross-modality person re-identification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19344–19353 (2022), https://api.semanticscholar.org/CorpusID:250520573
10. Nguyen, T.D., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors (Basel, Switzerland) **17** (2017), https://api.semanticscholar.org/CorpusID:3351302
11. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.H.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3622–3631 (2019), https://api.semanticscholar.org/CorpusID:204512264
12. Wang, X., Doretto, G., Sebastian, T.B., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. 2007 IEEE 11th International Conference on Computer Vision pp. 1–8 (2007), https://api.semanticscholar.org/CorpusID:14958192
13. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 618–626 (2019), https://api.semanticscholar.org/CorpusID:128356924
14. Wu, A., Zheng, W., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 5390–5399 (2017), https://api.semanticscholar.org/CorpusID:4796796

15. Ye, M., Lan, X., Leng, Q.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing **29**, 9387–9399 (2020), https://api.semanticscholar.org/CorpusID:220323031
16. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Transactions on Information Forensics and Security **15**, 407–419 (2020), https://api.semanticscholar.org/CorpusID:202699290
17. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 13547–13556 (2021), https://api.semanticscholar.org/CorpusID:244920021
18. Ye, M., Shen, J., Crandall, D.J., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. ArXiv **abs/2007.09314** (2020), https://api.semanticscholar.org/CorpusID:220647378
19. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 2872–2893 (2020), https://api.semanticscholar.org/CorpusID:210164273
20. Yin, J., Ma, Z., Xie, J., Nie, S., Liang, K., Guo, J.: Df^{2}am: Dual-level feature fusion and affinity modeling for rgb-infrared cross-modality person re-identification. ArXiv **abs/2104.00226** (2021), https://api.semanticscholar.org/CorpusID:232478698
21. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.R.: Residual non-local attention networks for image restoration. ArXiv **abs/1903.10082** (2019), https://api.semanticscholar.org/CorpusID:85501306
22. Zhang, Z., Jiang, S., Huang, C., Li, Y., Xu, R.Y.D.: Rgb-ir cross-modality person reid based on teacher-student gan model. Pattern Recognit. Lett. **150**, 155–161 (2020), https://api.semanticscholar.org/CorpusID:220525625
23. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing **28**, 4500–4509 (2019), https://api.semanticscholar.org/CorpusID:14685197
24. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 1116–1124 (2015), https://api.semanticscholar.org/CorpusID:14991802
25. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. ArXiv **abs/1910.09830** (2019), https://api.semanticscholar.org/CorpusID:204824181