Enhancing Visible-Infrared Person Re-Identification using Keypoint-based Global and Local Feature Representation

Soonyong Gwon[*]
Woon Chae[*]
Kisung Seo[†] KSSEO@SKUNIV.AC.KR
Department of Electronics & Computer Engineering, Seokyeong University, Seoul 02713, South Korea

Editors: Manuel Grana, Pawel Ksieniewicz, Leandro Minku, Pawel Zyblewski

Abstract

Visible-Infrared Person Re-Identification (VI-ReID) is a highly challenging classification task that involves matching identities between visible and infrared images. The datasets used for VI-ReID are characterized not only by typical difficulties such as imbalanced classes, numerous classes, high-dimensional features, and multi-view perspectives, but also by significant feature differences between the visible and infrared images of the same person, which complicate the matching process much more. The learning challenges for VI-ReId data are as follows. First, in matching visible and infrared images, key features related to color and clothing patterns are often changed or missing, and the addition of noise due to image degradation makes the images appear significantly different. Second, like zero-shot learning, all test data consists of the unseen classes. To solve these problems, it is necessary to develop a novel feature learning approach for all parts of the body by balancing global and local feature representations, and to learn meta-features that do not depend on a specific person identity. To overcome these challenges, the proposed method includes the following: 1) keypoint erasing-based global feature extraction and increasing the diversity of feature representations ensuring the diversity of representations for augmented images, 2) keypoint masking-based robust learning for local feature matching between visible and infrared body parts. These strategies ensure superior performance on highly challenging data. We compare our proposed methodology and various existing methods for the mAP and Rank-1 performances on the SYSU-MM01 datasets. Experimental results demonstrate that our proposed model shows superior performance and effectively solves the existing key and critical problems.

1. Introduction

In classification problems, the difficulty of the data has a significant impact on classification performance. Typical difficulties include the following - imbalanced classes, numerous classes, high-dimensional features, a small number of training examples, etc. In addition, special challenges may arise based on the specific nature of the problem. Beyond the inherent difficulty of the data itself, this can be intensified by the specific task, making classification more difficult. Visible and Infrared Re-Identification(VI-ReID) is a special case of Person Re-Identification, which aims to identify or distinguish specific person images. Person ReID

[*] Soonyong Gwon and Woon Chae contributed equally as first author
[†] Corresponding author.

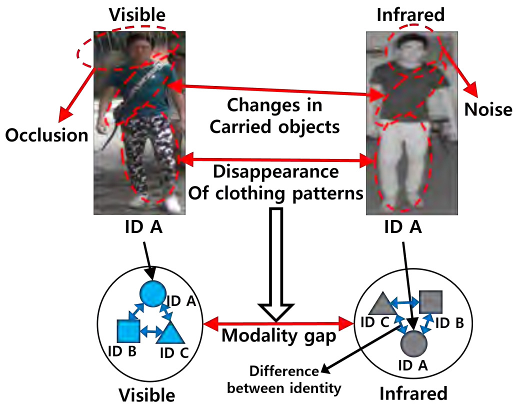

Figure 1: Feature differences between V-I and the significant modality gap with relatively small inter-ID differences

is a critical task for pedestrian identification in CCTV-based intelligent video surveillance systems (Kim et al., 2024). This task is particularly challenging due to various factors such as background variations over time and space, different camera angles, and occlusions caused by obstacles or other individuals (Kim et al., 2024; Sun et al., 2021; Liao et al., 2015; Zhong et al., 2021). To enhance video surveillance in night environments, a combination of infrared and visible cameras is often employed. However, matching visible and infrared images is much more challenging due to the substantial cross-modality gap (Zhang et al., 2022; Ye et al., 2021; Liu et al., 2018).

The VI-ReID task is a multi-class classification problem where each individual ID corresponds to a class. For instance, the well-known SYSU dataset consists of 96 IDs (classes). Beyond the typical challenges associated with standard training data, VI-ReID data introduces additional complexities: First, in matching visible and infrared images, key features related to color and pattern shape are often changed or missing, making the images appear significantly different. Additionally, infrared images often contain considerable noise. Figure 1 illustrates that for the same identity, VI features of the same region are very different, the modality gap is very large, and the differences between identities are very small. Second, unlike general classification tasks where some data from all classes are used for training, VI-ReID requires testing on entirely unseen classes. This means that the model cannot leverage any learned features from those classes, further increasing the difficulty.

The research trends for VI-ReID are as follows. Data augmentation methods for VI-ReID have been explored. Grey Transform (Hao et al., 2021) and Hue and Grey transform (Zhong et al., 2020) have been proposed to bridge the gap between visible and infrared images, but there is a limit to the performance improvement of color transforms alone. Random erasing (Wei et al., 2021b) is used for data augmentation but often leads to confusion in feature extraction due to inappropriate erasure areas. On the other hand, GAN-based generative methods (Liu et al., 2018, 2020b; Sun et al., 2019) are used to make the Visible image similar to the Infrared image. However, generative methods usually require additional intensive training and suffer from noise. More fundamentally, modality- based approaches have been proposed for focusing on the characteristics of different modalities. Modality-shared methods (Sun et al., 2019; Zhang et al., 2023; Wang et al., 2020; Ye et al., 2019) that extract modality-invariant features in a common embedding space. On the other hand, modality-specific feature extraction have been introduced (Liu et al., 2020b; Lu et al., 2023). Most of existing studies commonly focus on global features of the body and background, resulting in inefficiencies due to dispersed attention and problems with local feature mismatches. Therefore, these methods fail to adequately address the challenges presented by significant feature differences between visible and infrared images, particularly in cases where individuals have similar clothing or body shapes but are different persons, as well as in matching for unseen classes.

To address the aforementioned issues, this paper proposes a Keypoint-based Global-Local Feature Representation Network (KGLFR-Net) that not only exclusively separates and combines global and local features but also extracts robust meta-features that do not depend on specific person identity. Specifically, the Keypoint Erasing-based Global Feature Representation (KEGFR) method accurately removes specific body parts by erasing keypoints in both visible (V) and infrared (I) images, ensuring that the remaining body parts have the same components in both modalities. This process helps the network learn to represent global features of the same identity consistently. Similarly, the Keypoint Masking-based Local Feature Representation (KMLFR) method selects and masks identical body parts in both V and I images, ensuring that local features of the same parts are consistently represented. Additionally, extensive experiments on the SYSU-MM01 and RegDB datasets demonstrate the superiority and validate the effectiveness of our proposed method.

2. Related Work

Ordinary Person ReID. Ordinary person Re-identification is the task of retrieving pedestrian images captured by different visible cameras in different circumstances. Traditional computer vision techniques have been tried (Pang et al., 2023; Huang et al., 2022), and recently, deep learning-based human ReID methods (Liao et al., 2015; Wang et al., 2019; Tian et al., 2021) show excellent performance. As the main approaches, studies using the feature information of human body parts that are not occluded (Zhong et al., 2021; Lu et al., 2023; Gwon et al., 2024) and studies considering various hidden situations in which arbitrary areas are deleted and learned (Lu et al., 2023; Zhao et al., 2021) are being researched.

Visible-Infrared Person ReID. The deployment of infrared cameras with visible cameras is used to improve video surveillance in night environments. However, the matching of visible and infrared is suffer from the large gap between both modalities. To overcome this problem, visible-infrared person re-identification (VI-ReID) methods (Zhang et al., 2022; Ye et al., 2021; Liu et al., 2018) have been studied. (Dai et al., 2018) first introduced a large scale VI-ReID dataset, named SYSU-MM01, and proposed a deep zero-padding strategy to explore modality-specific structure in a one-stream network.

Data Augmentation in VI-ReID. The data augmentation methods focus on the generation of a third modality or the augmentation of each modality's data. Gray Transform (Hao et al., 2021) provides an intermediate modality by encompassing both the structural characteristics of visible images and the visual features of infrared images However, conventional gray scale images utilize limited information due to fixed weights. A Hue and Gray transform (Zhong et al., 2020) is proposed to enhance the diversity of augmented images. Random erasing (Wei et al., 2021b) is also used as one of the important data augmentation methods, but it confusingly extracts the features of person shape due to the inappropriateness of the erasing part. Unlike erasing of parts in image, There exist mixture-based methods for VI-ReID (Choi et al., 2020), but suffer from unnatural patterns. These data augmentation methods are mainly used as auxiliary methods in the VI-ReID task.

Generative Methods in VI-ReID. Generative methods aim to reduce differences in modality styles, primarily using Generative Adversarial Networks (GANs) for modality translation. Hi-CMD (Liu et al., 2018) and cmGAN (Tan et al., 2022) employ GANs to convert various modality features into a shared space. Similarly, AlignGAN (Wang et al., 2019) uses GANs to align cross-modality features at both pixel and feature levels. FM-CNet (Liu et al., 2020b) leverages GANs for feature-level modality compensation. These techniques effectively minimize differences between visible and infrared styles. The adversarial learning based FBP model (Sun et al., 2019) is used to automatically distinguish part representations according to the feature maps of pedestrian images. However, generative methods often require additional computation and are prone to introducing noise.

Modality Feature Extraction in VI-ReID. Modality invariant feature learning based VI-ReID has been proposed to project the features from different modalities into the same feature space. Main stream is modality-shared methods (Sun et al., 2019; Zhang et al., 2023; Wang et al., 2020; Ye et al., 2019) that extract modality-invariant features in a common embedding space. On the other hand, a few methods that incorporating modality-specific feature extraction have been introduced (Liu et al., 2020b; Lu et al., 2023) . However, the last few layers are difficult to map the specific representations of each modality to a shared space. Both approaches are biased towards specific modality characteristics, limiting their potential improvement for advantages of integration of both methods. To solve these problems, a Modality Restitution and Compensation Module (Zhao et al., 2021) is proposed to respectively distill modality-irrelevant and modality-relevant features. Also a balanced approach between Modality-Specific and Modality-Shared method (Liu et al., 2020a) is proposed. However, these methods rely more on the way each modality passes through the network rather than on features that are either common (or independent) across modalities or specific (or dependent) to each modality. Specifically, Modality-Shared feature is defined as the output of the network where different modalities converge and share the same parameters, while Modality-Specific feature is defined as the output of separate networks with different parameters for each modality. Therefore, this approach does not exactly represent the common features (like glasses) and specific features (like hair, which appears black in visible and white in infrared) that we generally assume.

Like the aforementioned studies, each branch shows remarkable improvements in terms of performance. However, the following critical issues remain unresolved in terms of accurate feature extraction and matching: individuals with similar clothing or body shapes but who are different persons, as well as matching for unseen classes during testing. This tendency becomes more pronounced with highly challenging data.

3. Visible-Infrared Person Re-identification

3.1. Challenges of Visible-Infrared Person Re-identification

Description of Visible-Infrared Person Re-identification Task. VI-ReID is a task where an infrared query image is compared with all images in the visible gallery set to find the same individual, and vice versa for a visible query image in an infrared gallery set. The images are ranked based on similarity, with the k-th most similar gallery image referred to as Rank-k (R-k). It is crucial for the same identity samples to appear frequently in the top ranks. Figure 2 illustrates examples of search and matching process. The VI-ReID task is a multi-class classification problem where each ID corresponds to a class, in the SYSU dataset, which comprises 96 IDs (classes). Matching visible and infrared images is significantly more challenging than general person ReID due to the substantial cross-modality gap.

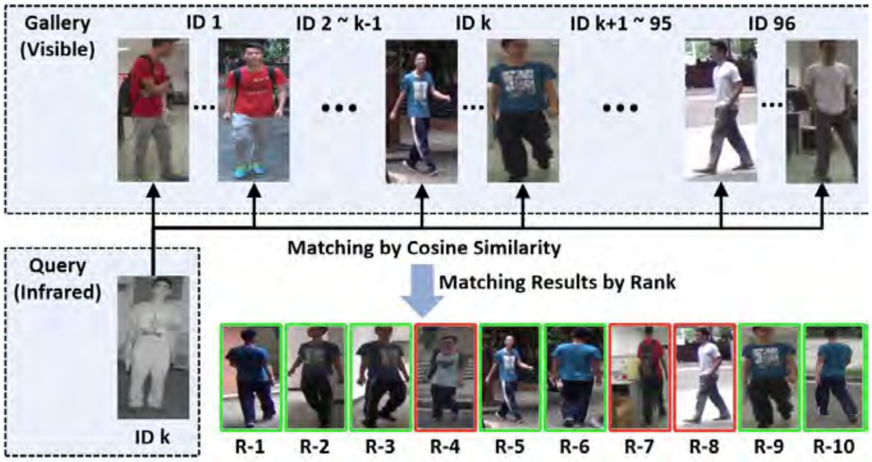

Figure 2: Illustration of Visible-Infrared Person Re-identification Task

Modality Discrepancy between Visible and Infrared. The most significant differences between visible and infrared images include the loss of color information, distortion in brightness and shading, disappearance of clothing patterns, the addition of noise due to image degradation, and a considerable reduction in three-dimensionality and texture information in infrared images compared to visible images as shown in Figure 3. It might also seem that infrared images are similar to grayscale images; however, as shown in Figure 3(a), there are distinct differences. Grayscale images are generated by converting color information through a transformation process, preserving brightness information related to color differences. Therefore, partial shapes and textures arising from differences in clothing patterns and colors remain, providing more crucial distinguishing features than infrared images. Moreover, in addition to the significant differences between V-I images, factors such as occlusion, posture, background, lighting, and changes in carried objects produce additional feature variations even for the same person as shown in Figure 3(b), further complicating re-identification.

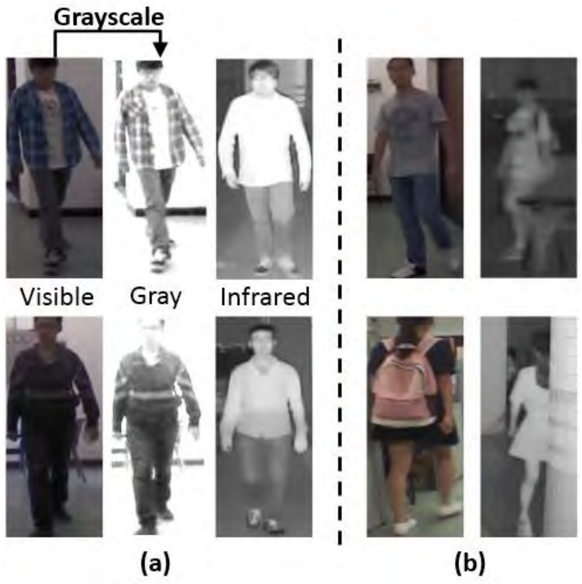

Figure 3: Modality Discrepancy between Visible and Infrared images

3.2. Problems of existing methods that focus on global features

To address the problem of focusing on dominant features, some methods have employed the strategy of training on images with randomly erased features to reduce reliance on these dominant characteristics. However, this has led to a mismatching problem where different features are represented closely, resulting in the incorrect identification of different body parts as the same feature. This issue is intensified in VI-ReID due to the substantial differences between the two modalities, making it difficult to match the same features across both types of images. Additionally, when matching global representations, local representations may be mismatched, leading to errors where different features are incorrectly aligned. For successful re-identification, it is crucial to develop meta feature learning that captures and matches the generality of various body features.

4. Method

4.1. Keypoint-based Elaborate Feature Extraction

Some existing methods (Choi et al., 2020; Lu et al., 2023; Gilroy et al., 2021) assume that similar features are present at similar vertical coordinates and use horizon part-based part extraction to leverage local features. However, these methods may fail when features are shifted vertically or when there is significant movement of body, as the same local region may not contain the same features. To accurately extract local features, we employ a pretrained keypoint estimation network to predict body parts. This approach enables precise matching of the same body parts. In Figure 4(a), different body parts can be present at the same location identified as the same parts by horizon-based part extraction. In contrast, 4(b) demonstrates that keypoint-based part extraction robustly matches the same body parts regardless of a person's position, ensuring accurate local feature matching. HRNet (Zhang et al., 2021) is used as a pretrained keypoint estimation network and outputs a total of 13 keypoint heatmaps consisting of head and both shoulders, elbows, wrists, hips, knees and ankles. Then, 11 body part coordinates are obtained by grouping the keypoints according to (Jia et al., 2021). The body part coordinates are used for global representation learning through keypoint erasing for the input image, and the keypoint heatmaps are used for local representation learning through keypoint masking for the feature maps.

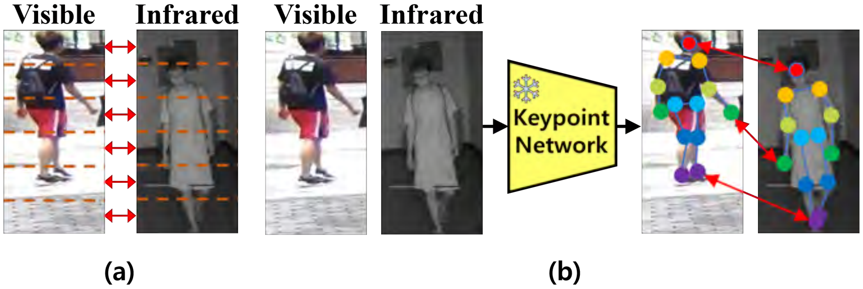

Figure 4: Comparison of horizontal part-based feature extraction and keypoint-based local feature extraction

4.2. Keypoint-based Combination of Global and Local Features

We propose a Keypoint-based Global-Local Feature Representation Network (KGLFR-Net) to separate and match global and local features. The proposed learning method includes keypoint deletion to diversify global features and keypoint masking to extract local features. As shown in Figure 6, keypoints are extracted for visible and infrared input images and used for keypoint erasing and masking. The image augmented by keypoint erasure is used as input to the ResNet-based Backbone network. Local feature maps are created by applying keypoint masking to the output global feature map and used for global and local representation learning, respectively. Global representation learning maintains sample diversity in the representation by data augmentation, and local representation learning learns various local features to perform mutually exclusive learning. Our framework improves the alignment of global features and the consistency of local features.

4.3. Keypoint Erasing-based Global Feature Representations

Representation learning aims to enable a model to recognise similarities be tween positive samples and differences between negative samples, with the aim of representing the same class closer together and different classes further apart. Typically, data augmentation is performed for model regularization, where some features are either removed or retained to train the model to distinguish individuals. However, this approach hinders discriminative feature learning by encouraging the model to represent different body parts uniformly.Consequently, incorrect feature learning may lead to reduced identification performance and hinder the learning of identity invariant representations, resulting in reduced adaptability to new features and potential overfitting. To address these problems, we propose the Keypoint Erasing-based Global Feature Representation Learning (KEGFR). Our training methodology augments intra-identity samples by removing identical features to ensure that the remaining features within the augmented images are uniform. Specifically, we randomly assign 11 body parts defined according to (Gilroy et al., 2021) to each identity and obtain the center coordinates (x_{ij}, y_{ij}) of each body part using keypoints for the i-th identity and j-th sample. For a training image area $A = W * H$, we sample random target width w and target height h satisfying $(0.02 * A < w * h < 0.4 * A)$. The proposed keypoint erasing is defined as follows:

$$\widetilde{Z}_{ij} = \begin{cases} \mu, (m,n) \in E_{ij} \\ \\ Z_{ij}(m,n), otherwise \end{cases} \tag{1}$$

Z represents the input image, and μ is the mean value of each channel. If pixel (m, n) belongs to the deletion area $E_{ij} = ([x_{ij} - (w_{ij}/2), x_{ij} + (w_{ij}/2)], [y_{ij} - (h_{ij}/2), y_{ij} + (h_{ij}/2)])$, then the pixel value is replaced by μ, otherwise, the pixel value is retained. The loss function of the proposed learning method is as follows:

$$\begin{aligned} L_{KEGFR} = \sum_{i=1}^{P} \sum_{j=1}^{K} \max \{[\max(d(r_{ij}, r_i)) \\ - \min_{i \neq l} \ (d(r_{ij}, r_l))] + m_{KEGFR}, 0\} \end{aligned} \tag{2}$$

P represents the number of identity, and K represents the number of samples constituting the identity. $d(.)$ denotes the Euclidean distance function, and r_{ij} represents the output of the model for $\widetilde{Z}_{ij}$. $r_i = [r_1, r_2, ..., r_K]$ denotes i-th identity samples and $d(r_{ij}, r_i) = [d(r_{ij}, r_{i1}), d(r_{ij}, r_{i2}), ..., d(r_{ij}, r_{iK})]$ represents the distance between the k-th sample and all the samples of the i-th identity. m_{KEGFR} is a hyper-parameter that controls the distance between positive and negative samples. Existing methods result in similar representations among images generated from random erasing due to the pulling effect on different parts during the training process. Consequently, the diversity effect of data augmentation is lost, and the problem arises where features among different parts become similar. KEGFR maintains the diversity of representations for augmented images by only pulling samples with similar features based on keypoints. Additionally, since identity-invariant representation is learned, it enhances inference performance by robustly re-identifying even unseen identities.

4.4. Keypoint Masking-based Local Feature Matching Representations

In training, existing methods only perform metric learning on global representations of the entire image, including the background, without matching local representations of individual body parts. Therefore, mismatching of local representations can occur, which can be amplified by the modality discrepancy between visible and infrared images. When matching

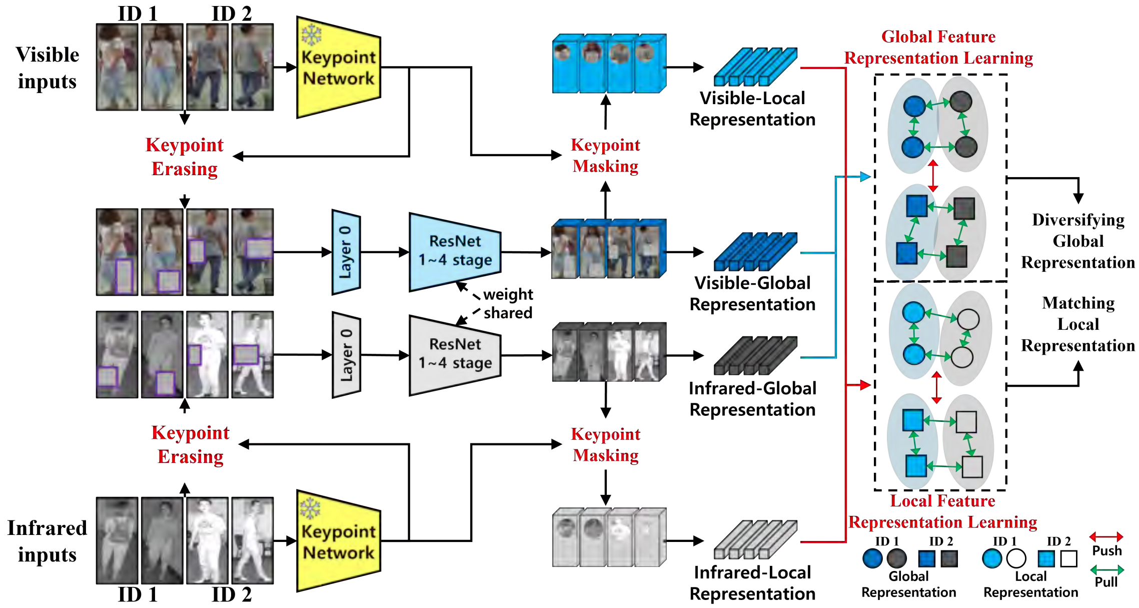
ID 1
ID 2
Visible inputs
Keypoint Network
Keypoint Erasing
Keypoint Masking
Visible-Local Representation
Global Feature Representation Learning
Layer 0
ResNet 1~4 stage
weight shared
Visible-Global Representation
Diversifying Global Representation
Infrared-Global Representation
Matching Local Representation
Keypoint Erasing
Keypoint Masking
Infrared inputs
Keypoint Network
ID 1
ID 2
Infrared-Local Representation
Local Feature Representation Learning
ID 1
ID 2
Global Representation
ID 1
ID 2
Local Representation
Push
Pull

Figure 5: Proposed Keypoint-based Global-Local Feature Representation Network (KGLFR-Net). The modality-specific module uses only one conv-layer at the front of ResNet-50(Ye et al., 2021) without weight sharing, on the other hand, the modality-shared module employs the remaining stages 1 to 4, with weight sharing.

global representations while local representations differ, attention may be dispersed to different body parts or the background. This reduces identification performance for samples with new feature combinations during testing, and this problem is more pronounced for hard samples. To achieve correct matching of local features, we propose Keypoint Masking-based Local Feature Representation learning (KMLFR). To extract local features, we define the keypoint mask as follows.

$$m_{ij} = max(k^c_{ij}/max(k^c_{ij}), 0) \tag{3}$$

k^c denotes the heatmap for the c-th keypoint out of 13 keypoints. The same keypoint heatmap is used for all samples within a mini-batch, and for stable learning, the maximum value of the mask is adjusted to 1, with values less than 0 being replaced by 0. The local feature representation is extracted as follows:

$$l_{ij} = G(m_{ij} * f_{ij}) \tag{4}$$

The global feature map is denoted by f, and $G(.)$ represents the Global Average Pooling operation. The loss function used in KMLFR is defined as follows:

$$L_{KMLFR} = \sum_{i=1}^{P} \sum_{j=1}^{K} \max \{[\max(d(l_{ij}, l_i)) \\ - \min_{i \neq l} (d(l_{ij}, l_l)) \Big] + m_{KMLFR}, 0 \Big\} \tag{5}$$

KMLFR focuses on learning local features, enabling the matching of local representations for the same identity while distinguishing different identities. This method ensures that feature representations for the same body parts across visible and infrared modalities become similar, enhancing matching performance on unseen data. Specifically, it improves recognition performance for challenging samples, such as those involving indistinguishable accessories or changes in pose. Additionally, by concentrating attention on body parts rather than the background, KMLFR facilitates robust identification.

4.5. Total Training Loss for Balancing of Global and Local Feature Representation

The final loss function is defined as follows and consists of the cross-entropy loss L_{id}, the Keypoint Erasing-based Global Feature Representations loss L_{KEGFR}, and the Keypoint Masking-based Local Feature Matching Representations loss L_{KMLFR}:

$$L_{Total} = L_{id} + \lambda_{KEGFR} * L_{KEGFR} + \lambda_{KMLFR} * L_{KMLFR} \tag{6}$$

λ_{KEGFR} and λ_{KMLFR} are hyperparameters set to 0.5 each to balance the backpropagation contributions of the respective losses. KEGFR enhances the discriminability of different features, while KMLFR ensures consistent matching of similar features, leading to robust identification even with novel or challenging data.

5. Experiments

5.1. Datasets

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. The SYSU-MM01 (Wu et al., 2017) dataset comprises two modes: the indoor-search mode, which includes images captured only with indoor cameras, and the all-search mode, which encompasses images from both indoor and outdoor cameras. The RegDB (Nguyen et al., 2017) dataset contains 412 IDs, with each ID having 10 visible images and 10 infrared images, providing a total of 8,240 images. The training and testing sets each consist of randomly selected, non-overlapping 206 IDs. This dataset offers two evaluation modes: Visible2Infrared, where infrared images are retrieved based on visible images, and Infrared2Visible, where visible images are retrieved based on infrared images. Performance metrics used for evaluation are Rank-1 accuracy and mean average precision (mAP).

Table 1: Comparison of our method with recent competitive methods on SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017) datasets. Rank-1 accuracy (%) and mAP (%) are reported. The best results and the second best are highlighted in red and blue, respectively.

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
FBP-AL (Wei et al., 2021a)	54.14	50.2	-	-	73.98	68.24	70.05	66.61
JCI (Zhao et al., 2021)	57.2	59.3	66.6	74.7	78.8	69.4	77.9	69.4
FMI (Tian et al., 2021)	60.02	58.8	66.05	72.98	73.2	71.6	71.8	70.1
MID (Huang et al., 2022)	60.27	59.40	64.86	70.12	87.45	84.85	84.29	81.41
cm-SSFT (Lu et al., 2020)	61.6	63.2	70.5	72.6	72.3	72.9	71.0	71.7
Hc-Tri (Liu et al., 2020b)	61.68	57.51	63.41	68.17	91.05	83.28	89.30	81.46
AGW-R[1] (Ye et al., 2021)	64.87	63.15	71.01	76.87	77.38	70.31	74.13	66.46
SPOT (Chen et al., 2022)	65.34	62.25	69.42	74.63	80.35	72.46	79.37	72.26
MCLNet (Hao et al., 2021)	65.4	61.98	72.56	76.58	80.31	73.07	75.93	69.49
FMCNet (Zhang et al., 2022)	66.34	62.51	68.15	74.09	89.12	84.43	88.38	83.86
SMCL (Wei et al., 2021b)	67.39	61.78	68.84	75.56	83.93	79.83	83.05	78.57
PMT (Lu et al., 2023)	67.53	64.98	71.66	76.52	84.83	76.55	84.16	75.13
MRCN (Zhang et al., 2023)	68.9	65.5	76.0	79.8	91.4	84.6	88.3	81.9
Ours (KGLFR)	70.44	67.59	79.39	82.78	91.94	88.13	90.83	86.77

5.2. Implementation Details

The keypoint extraction model uses the validated pre-trained model HRNet (Wang et al., 2020), requiring no additional training. Moreover, it does not increase extra computation during inference. The backbone network is the ResNet50(AGW (Ye et al., 2021)) pre-trained on ImageNet. Input images are resized to 288x144 and 320x160, respectively, for the backbone and keypoint extraction model. The batch size is set to 64, with each mini-batch consisting of 4 visible images and 4 infrared images from 8 different IDs. The network is trained using the SGD optimizer with a learning rate of 0.1, weight decay of 0.0005, and momentum of 0.9. The learning rate is reduced by a factor of 10 after 30 epochs and by a factor of 100 after 60 epochs, over a total of 100 epochs. Common data augmentation techniques such as RandomCrop and RandomHorizontalFlip are applied, but RandomErasing is not used unless explicitly stated. m_{KEGRF} is set to 0.3 to follow AGW (Ye et al., 2021), and m_{KMLFR} is set to 0.01 to account for the reduction in scale due to masking. All experiments are conducted on a single NVIDIA GeForce RTX 3090, but can also be performed on a single GPU with 8GB of memory, such as an RTX 2080.

5.3. Comparison with competitive Methods

The proposed KGLFR-Net is compared with recent competitive methods. Table 1 shows that KGLFR outperforms existing techniques on the SYSU-MM01 and RegDB datasets. On the SYSU-MM01 dataset, KGLFR achieves a Rank-1 accuracy of 70.44(%) and mAP of 67.59(%) in the all-search mode. This result surpasses the performance of MRCN (Zhang et al., 2023) in Table 1 by 1.54(%) in Rank-1 accuracy and 2.09(%) in mAP. In the indoor-search mode, KGLFR achieves a Rank-1 accuracy of 79.39(%) and mAP of 82.78(%), outperforming MRCN (Zhang et al., 2023) in Table 2 by 3.39(%) in Rank-1 accuracy and 2.98(%) in mAP. On the RegDB dataset, our method achieves a Rank-1 accuracy of 91.94(%) and mAP of 88.13(%) in the infrared-to-visible mode.

This represents an improvement of 0.54(%) in Rank-1 accuracy and 3.28(%) in mAP over MRCN (Zhang et al., 2023) and MID (Huang et al., 2022). In the visible-to-infrared mode, KGLFR-Net achieves a Rank-1 accuracy of 90.83(%) and mAP of 86.77(%), outperforming Hc-Tri (Liu et al., 2020b) and FMCNet (Zhang et al., 2022) by 1.53(%) in Rank-1 accuracy and 2.91(%) in mAP.

5.4. Ablation Study

We evaluate the efficiency of each component of the proposed KGLFR method, specifically KEGFR and KMLFR. All experiments were conducted on the SYSU-MM01 dataset in all-search mode, and the results of various configurations are presented in Table 2.

Table 2: Performance comparison of each component of our model on the SYSU-MM01 dataset. Index 1 represents the baseline, which employs AGW network, Identity loss and Hardest Triplet loss.

idx	Random erasing	KGLFR		R-1	mAP
		KEGFR	KMLFR		
1				52.88	51.45
2			✓	54.40	54.33
3	✓			64.87	63.15
4		✓		67.10	65.17
5		✓	✓	70.44	67.59

Index 1 corresponds to the BaseModel described in Section 4.2. Index 2 shows the results obtained by adding KMLFR to the baseline. These results indicate that KMLFR can learn modality-independent matching representations across cross-modalities, improving the Rank-1 performance by 1.52(%) and mAP by 2.88(%) compared to the baseline. Indexes 3 and 4 show the results of adding the existing data augmentation technique RandomErasing and the proposed KEGFR to the baseline, respectively. Index 4 demonstrates a greater performance improvement over index 3, with Rank-1 increasing by 1.86(%) and mAP by 1.92(%). This is because, unlike traditional RandomErasing, the proposed KEGFR allows for learning representations where all features within a sample are not solely dependent on identity. Index 5 presents the results of adding both the proposed KMLFR and KEGFR to the baseline, forming KGLFR. As a result, KGLFR improves the Rank-1 performance by 17.56(%) and mAP by 16.14(%) compared to the baseline.

5.5. Visualization

To further demonstrate the effectiveness of the proposed method, we visualize some of the results from Table 2 using retrieval results and GradCAM activation maps, as shown in

[1] trained AGW backbone with cross entropy loss, triplet loss and random erasing

Figure 6, 7 and 8. All experiments were conducted on the SYSU-MM01 dataset in all-search mode. The figure compares the top-10 retrieval results of PMT (Ye et al., 2019), the latest paper with official code available, and our proposed method. The first and second rows show the retrieval results for ID A, while the third and fourth rows show the results for ID B.

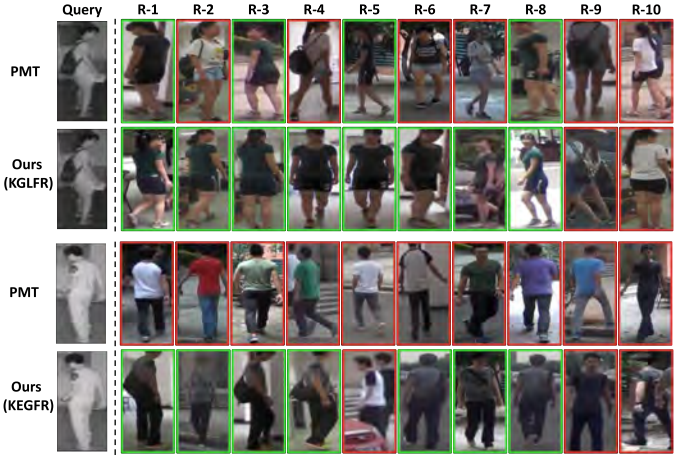

Figure 6: Comparison of the top-10 retrieval results of PMT (Ye et al., 2019) and our method. The first two rows correspond to ID A, and the third and fourth rows correspond to ID B. The green boxes denote correct matches, and the red boxes denote incorrect matches.

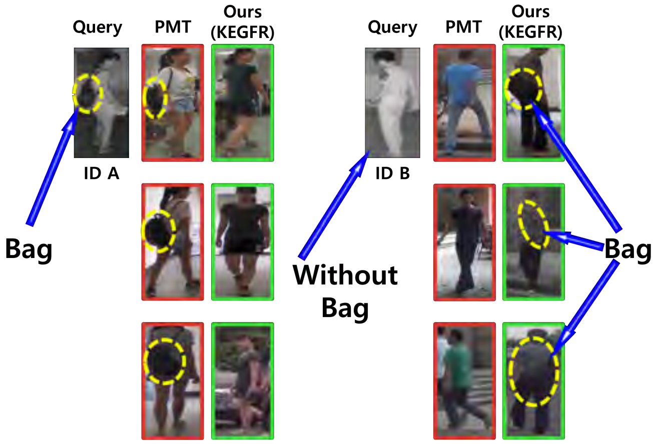

Figure 7: Detailed description of the errors in matching based on the presence or absence of a bag, along with the correct matching of the proposed method.

Retrieval results. In Figure 6, the proposed method demonstrates superior performance over the PMT (Ye et al., 2019) in retrieving samples with dominant features. In the

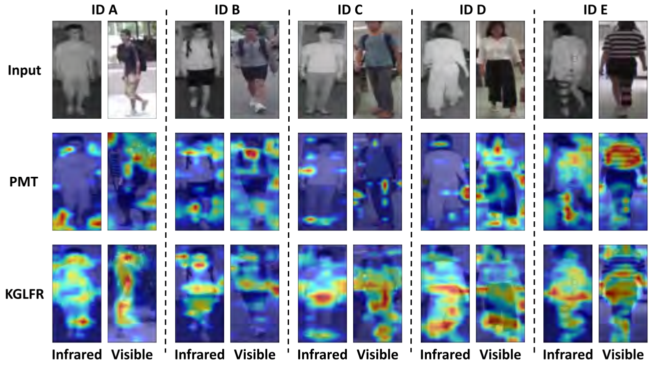

Figure 8: Comparison of the activation maps using Grad-CAM.

retrieval results for ID A and ID B by PMT, the focus is on the dominant feature of bags. If the query includes a bag, samples with bags are ranked higher, and if the query does not include a bag, samples without bags are prioritized. Additionally, PMT prioritizes searching for people with similar rear views to the query, leading to inaccurate matches being ranked higher. In contrast, the proposed approach concentrates on local features of human body parts and resolves the Identity dependent representation problem of global features. Consequently, scenes with more accurate views (frontal or rear views) are ranked higher regardless of dominant features. Figure 7 provides a detailed illustration of matching errors in the existing methods based on the presence or absence of a bag, alongside a depiction of accurate matching achieved by the proposed method.

Activation maps. In Figure 8, PMT (Ye et al., 2019) shows activation in the background, and different activations appear in different locations between Infrared and Visible images of the same ID. In contrast, KGLFR shows minimal activation in the background and consistent activation in the same body parts' locations. Therefore, KMLFR demonstrates its ability to address localized feature representation mismatching.

6. Conclusion

The Visible-Infrared Person Re-Identification task presents the challenges of mismatching global and local features of body parts and the difficulties associated with zero-shot learning. To address these crucial problems, we propose Keypoint Erasing-based Global Feature Representation (KEGFR), which erases identical body parts based on keypoints and aligns the global feature representations of the same identity across the remaining body parts, and Keypoint Masking-based Local Feature Representation (KMLFR), which accurately selects identical body parts using keypoint masking and trains the network to ensure consistent local feature representation for these parts. It also reduces modality discrepancy by matching cross-modality features. We compare the performance of our method and several existing approaches on mAP and Rank-1 measurements on the SYSU-MM01 and RegDB datasets and analyse the results of each module and visualisation in local feature matching. The experimental results demonstrate the superiority and effectiveness of our proposed method over recent competing approaches in visualisation.

Acknowledgments

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

References

Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022.

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

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

Shane Gilroy, Martin Glavin, Edward Jones, and Darragh Mullins. Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3833–3839, 2021.

Soonyong Gwon, Sejun Kim, and Kisung Seo. Balanced and essential modality-specific and modality-shared representations for visible-infrared person re-identification. *IEEE Signal Processing Letters*, 2024.

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

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*, volume 36, pages 1034–1042, 2022.

Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1673–1681, 2021.

Sejun Kim, Soonyong Gwon, and Kisung Seo. Enhancing diverse intra-identity representation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2522, 2024.

Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

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, 2020a.

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, 2020b.

Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4099–4108, 2018.

Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 1835–1843, 2023.

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.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

Zhiqi Pang, Chunyu Wang, Lingling Zhao, Yang Liu, and Gaurav Sharma. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

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.

Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 393–402, 2019.

Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. Mhsa-net: Multihead self-attention network for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1531, 2021.

Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 3623–3632, 2019.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

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, 2021a.

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, 2021b.

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.

Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE transactions on information forensics and security*, 15:407–419, 2019.

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.

Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: 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.

Yukang Zhang, Yan Yan, Jie Li, and Hanzi Wang. Mrcn: a novel modality restitution and compensation network for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3498–3506, 2023.

Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12136–12145, 2021.

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*, volume 35, pages 3520–3528, 2021.

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.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.