Implicit Modality Knowledge Alignment and Uncertainty Estimation for visible-infrared person re-identification

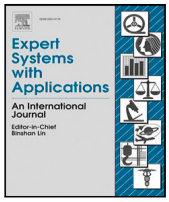
Song, W.; Shan, S.; Guoqiang, X.; Lew, M.S.K.; Gao, X.

Expert Systems With Applications 259 (2025) 125291

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Implicit Modality Knowledge Alignment and Uncertainty Estimation for visible-infrared person re-identification

Song Wu [a], Shihao Shan [a], Guoqiang Xiao [a], Michael S. Lew [c], Xinbo Gao [b,*]

[a] College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[b] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[c] LIACS Media Lab, Leiden University, Leiden, Netherlands

ARTICLE INFO

Keywords:
VI-ReID
Implicit Modality Generation
Distribution alignment
Uncertainty Estimation

ABSTRACT

Visible-infrared person re-identification (VI-ReID) is a crucial cross-modal matching task in computer vision. Existing research typically relies on a single auxiliary modality to address the semantic gap in cross-modal data, but this approach fails to fully exploit the information inherent in both modalities. To tackle this issue, we propose a novel framework called Implicit Modality Knowledge Alignment (IMKA) and Uncertainty Estimation (UE). This framework aims to enhance the robustness of the learned common embedding space and improve the prediction accuracy of the VI-ReID model. The IMKA module is designed to extract valuable supplementary information from the original modality data to generate implicit modality data. The SKA module then aligns the distribution of significant knowledge learned from this generated implicit modality data. Meanwhile, the UE strategy is implemented to mitigate the overconfidence in incorrect predictions. The uncertainty of predicted probabilities is evaluated in a Dirichlet space, and the designed evidence loss bolsters the confidence in uncertainty. This uncertain information can enhance retrieval performance. Extensive experimental results on two publicly available VI-ReID datasets demonstrate that our IMKA-UE framework significantly outperforms state-of-the-art methods. The code for our IMKA-UE framework is available at https://github.com/SWU-CS-MediaLab/IMKA-UE.

1. Introduction

Person Re-identification (He et al., 2021; Li et al., 2022; Liu, Zhang, Yu, Lu, & Yang, 2021c) is generally considered a retrieval task in computer vision, aiming to search for specific pedestrians across non-overlapping camera views (Liu, Zhang, Yu, Lu, & Yang, 2021a, 2021b; Wang, Zhang, Liang, Chen, & Zhou, 2021). Recently, significant progress has been made in single-modality Person Re-identification due to the high feature abstraction and representation learning capabilities of Deep Neural Networks (DNNs) (Hong, Chen, Liang, & Zhou, 2021; Zhang, Wang, Liang, Chen, & Zhou, 2021). However, current camera systems automatically switch to infrared mode when there is insufficient light at night, resulting in a large amount of cross-modal data. This has led to the emergence of the Visible-Infrared Person Re-identification (VI-ReID) task (Fu et al., 2021; Zhao, Liu, Chu, Lu, & Yu, 2021), which aims to identify pedestrians in the underlying database based on other modalities.

For the specific VI-ReID task, in addition to the intra- and inter-class differences in single-modality Person Re-identification, VI-ReID must address the substantial semantic domain gap between visible and infrared modalities. The visible and infrared modality data have different spectral wavelengths due to their heterogeneous physical properties (Pu, Chen, Liu, Bakker, & Lew, 2020). Therefore, unlike the single-modality Re-ID task, the substantial semantic domain gap makes VI-ReID more challenging. The key to VI-ReID lies in aligning the semantic information of different modality data and transferring them to a robust invariant feature common space that can bridge the semantic gap in cross-modal data.

The heterogeneous physical properties between visible and infrared modalities result in the loss of certain fundamental information in joint embedding space learning. Additionally, information loss between cross-modality data is unavoidable in the multi-modal learning process. This inevitable information loss further results in uncertainty in the final identity prediction. The recently proposed mechanism of implicit sample learning could achieve information augmentation and has made noticeable progress in VI-ReID. However, most existing implicit sample learning (Li, Wei, Hong, & Gong, 2020) can be summarized as data

* Corresponding author.
E-mail addresses: songwuswu@swu.edu (S. Wu), sshswu@email.swu.edu.cn (S. Shan), gqxiao@swu.edu.cn (G. Xiao), m.s.lew@liacs.leidenuniv.nl (M.S. Lew), gaoxb@cqupt.edu.cn (X. Gao).

https://doi.org/10.1016/j.eswa.2024.125291
Received 13 July 2023; Received in revised form 27 August 2024; Accepted 31 August 2024
Available online 3 September 2024

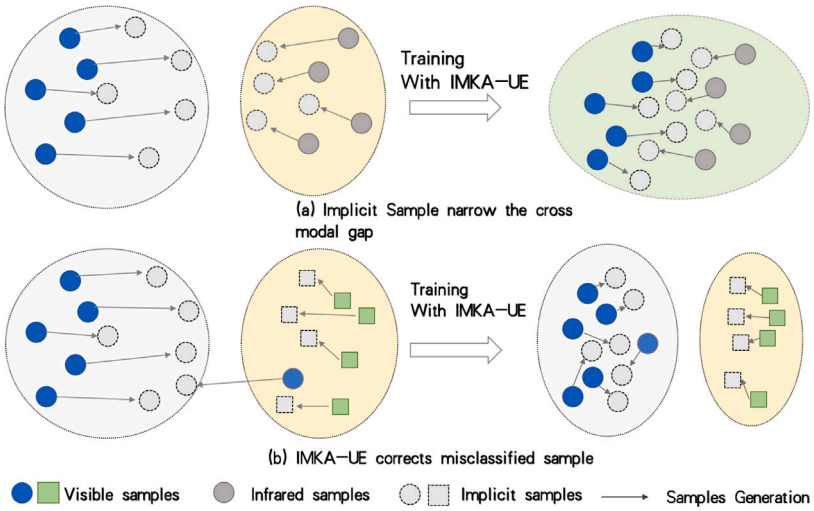


Fig. 1. Illustration of the Implicit Modality Knowledge Alignment (IMKA) mechanism and the Uncertainty Estimation (UE) strategy. Different shapes represent different ground-truth identities. As shown in Fig. 1(a), the blue and gray circles represent the original samples from the visible and infrared modes, respectively. The IMKA aims to effectively bridge the semantic gap between the two modalities by augmenting the generated implicit modality data during training. As shown in Fig. 1(b), the misclassified circles in the box class are intended to be re-clustered into the correct circle class using the UE strategy. UE helps models gain more evidence to support classification, which contributes to rectifying misclassifications and improving performance.

augmentation to support additional information, often neglecting the discovery of supplementary knowledge in each modality data to effectively generate implicit samples and reduce the semantic domain gap between disparate modalities (Wu, Yuan et al., 2024).

A straightforward solution for implicit sample generation is to generate data from one modality into another modality's style through generative adversarial networks (GANs) (Liu, Xiao, Lew, Gao, & Wu, 2024; Wang, Yang, Cheng, Chang, Liang et al., 2020). However, this approach is costly, and in addition to the large number of GAN parameters, the quality of the generated data (Liu, Xiao et al., 2022; Wang et al., 2020) cannot be guaranteed due to the lack of corresponding cross-modal original data for supervision. Meanwhile, existing implicit sample learning mainly focuses on the generation mechanisms for different modality data (Wu, Shan et al., 2024), while ignoring the redundant knowledge accumulation caused by extra implicit samples. To address these limitations, we propose a novel Implicit Modality Knowledge Alignment (IMKA) module. The IMKA module consists of an Implicit Modality Generation (IMG) module and a Significant Knowledge Alignment (SKA) module. The IMG module is a lightweight network architecture that learns valuable supplementary information from the original cross-modality data to generate identity-consistent high-quality implicit samples. Unlike existing implicit sample generation solutions, the designed IMG model utilizes progressive convolution to generate implicit samples and uses the corresponding modality data to supervise the style generation of the original modality. Thus, the IMG module has the advantage of learning useful supplementary information from the original modality to overcome the loss of fundamental information and the accumulation of redundant information, reducing the significant number of parameters in GANs and avoiding generation failures caused by complex GANs. Furthermore, the designed IMG consists of only two convolutional layers, which can replace GANs and play a role in solving the cross-modal semantic gap.

As shown in Fig. 1(a), the generated implicit samples help narrow the original semantic gap between visible light and infrared modalities, which is beneficial for subsequent training to extract cross-modal invariant semantic information. In Fig. 1(b), the generated implicit samples correct misclassifications and enhance performance. Misclassified blue circle samples can be clustered into the correct category through training with their implicit samples.

Moreover, due to the sparse distribution of the generated implicit modality data, which cannot bridge cross-modal data well, the SKA module is designed based on a novel distribution consistency loss. This loss aims to ensure that the significant knowledge extracted from the generated implicit samples aligns better in the learned latent common space. The distribution consistency loss guides and constrains the distribution of extracted significant knowledge in the generated implicit modalities to be consistent in the common latent space. Thus, the IMKA module effectively avoids redundant knowledge from extra implicit modality data, facilitates the accumulation of significant supplementary knowledge, and further enforces the alignment of significant supplementary knowledge in the learned latent embedding space. Most current VI-ReID methods based on generating auxiliary modalities use a single modality to assist in extracting cross-modality invariant pedestrian semantics. Our work is pioneering in using dual auxiliary modalities and achieves better performance with IMKA compared to previous methods.

Furthermore, most current VI-ReID methods calculate the probability of a sample belonging to each identity (Farooq, Awais, Kittler, & Khalid, 2022; Ge et al., 2022) using the Softmax function. For misclassified samples, the Softmax function enforces an incorrect classification probability, but the incorrect probability given by the network is still considered correct. Thus, improving the accuracy and confidence of the network's predictions remains a challenge. To address this, we propose a novel Uncertainty Estimation (UE) strategy to evaluate confidence and an evidence loss to enhance the accuracy of each identity prediction. Specifically, the output probability of the Softmax function is treated as evidence to support the identity prediction, and uncertainty is estimated by mapping the predicted values in the Dirichlet space. Based on the estimated uncertainty for each prediction, an evidence loss is constructed to enhance the quantified confidence of each epoch during the optimization process. Thus, in VI-ReID performance, the UE strategy can evaluate whether the model's predictions are plausible, effectively reduce the uncertainty of the prediction results, and achieve higher accuracy. Additionally, the UE strategy allows us to identify and exclude potentially corrupted or over-distributed data, thereby improving the model's robustness and generalization.

In summary, generating extra modality data, discovering and aligning significant knowledge of unseen domains as supplementary to the original data, and correcting the predictions of the learned model are the primary motivations of this paper.

In this paper, the main contributions of our proposed IMKA-UE-based VI-ReID are outlined as follows:

- We propose a novel Implicit Modality Knowledge Alignment (IMKA) mechanism, which consists of an Implicit Modality Generation (IMG) module and a Significant Knowledge Alignment (SKA) module. This mechanism effectively reduces the large semantic domain gap between visible and infrared modality data in the VI-ReID task. The IMG module aims to learn valuable supplementary information from the original modality data to generate implicit modality data. Meanwhile, the SKA module aims to constrain and align the distribution of significant knowledge learned from the generated implicit samples using a designed distribution consistency loss.
- We design an Uncertainty Estimation (UE) strategy for VI-ReID to correct the predictions of the learned model. This strategy treats the output probability of the Softmax function as evidence to support identity prediction, and estimates uncertainty by mapping the predicted values in the Dirichlet space. The UE strategy quantifies the confidence level of predicted probabilities in a Dirichlet space and combines this with a designed evidence loss to enhance quantified confidence, effectively reducing incorrect prediction results.
- Our evaluation results show that the proposed IMKA mechanism, together with the UE strategy, achieves state-of-the-art results on two commonly used datasets. Both IMKA and UE are effective in enhancing high-performance VI-ReID. To the best of our knowledge, this is the first application of the concepts of implicit modality knowledge alignment and uncertainty estimation in the VI-ReID task.

The structure of this paper is organized as follows: Section 2 reviews related work on visible-infrared person re-identification, organized by the starting point of the method. describes the details of our proposed IMKA mechanism and UE strategy. Section 4 presents extensive experiments where our method is compared with state-of-the-art methods and includes ablation experiments to validate the effectiveness of each module of our IMKA-UE. In Section 5, we discuss the strengths of our method and why it works effectively. Finally, Section 6 summarizes our work.

2. Related works

2.1. Visible-infrared person re-identification

Existing VI-ReID methods all start by addressing the semantic gap in cross-modal data. The first category focuses on learning mode-shared feature representations and aggregating specific visible and infrared features for better performance. Recent deep neural networks are designed to learn mode-aligned feature representations. Dual-level Discrepancy Reduction Learning (D^2RL) (Wang, Wang, Zheng, Chuang, & Satoh, 2019) proposes a two-path network that enhances the model's focus on cross-modal shared features through bi-directional double-constrained loss. Cross-Modal Align (cmAlign) (Park, Lee, Lee, & Ham, 2021) uses dense alignment to obtain salient local cross-modal shared features. Neural Feature Search (NFS) (Chen, Wan, Li, Jing, & Sun, 2021) constructs a search space of shared features to find discriminative cross-modal shared features. Structure-Aware Positional Transformer (SPOT) (Chen et al., 2022) learns semantic-aware shareable modal features using structure and location information. Mixup and Invariant Decomposition (MID) (Huang, Liu, Li, Zheng, & Zha, 2022) employs adaptive modal mixing and invariant decomposition to learn robust cross-modal shared features. The second approach mainly compensates for the lack of cross-modal information, such as GAN-based methods. Alignment Generative Adversarial Network (AlignGAN) (Wang et al., 2020) uses GANs for style migration to compensate for missing color information in infrared modal images. Feature-level Modality Compensation Network (FMCNet) (Zhang, Lai, Liu, Huang, & Han, 2022) constructs a feature-specific level compensation strategy that generates modally missing semantic information with modally shared features to mitigate gaps between cross-modal data. The main purpose of these approaches is to align the semantic information of cross-modal data, alleviate the semantic gap, and facilitate the extraction of semantic information invariant across modalities. Compared to these approaches, the IMKA module in this paper aims to learn supplementary information to generate implicit samples, further enhance the distribution consistency of the extracted discriminative knowledge in the learned common embedding space, and significantly reduce the complex parameters in GAN-based frameworks.

2.2. Uncertainty estimation

Uncertainty (Kendall & Gal, 2017) is an important concept in computer vision, which can be divided into two categories: epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty, also known as model uncertainty, reflects the uncertainty of the parameters in Bayesian inference. It can be reduced by collecting more data or using better models. For example, different neural networks with different weights may have different predictions for the same data, but they can all perform well. This means that the learned deep models can have different views of the data (Gal & Ghahramani, 2016). Epistemic uncertainty can affect the confidence and reliability of model predictions, especially in safety-critical applications such as autonomous driving or medical diagnosis. On the other hand, aleatoric uncertainty evaluates the uncertainty inherent in the data, such as noise or ambiguity in the input (Kong, Sun, & Zhang, 2020). It can be further divided into homoscedastic uncertainty, where all data have similar noise, and heteroscedastic uncertainty, where only some data have dissimilar noise. Aleatoric uncertainty can affect the quality and diversity of the data, and it may require more complex models to capture it.

Recently, some uncertainty estimation algorithms based on the Dirichlet distribution have been proposed (Han, Zhang, Fu, & Zhou, 2021; Sensoy, Kaplan, & Kandemir, 2018), which use evidence networks to model subjective logic through the Dirichlet distribution. The Dirichlet distribution (Kopetzki, Charpentier, Zügner, Giri, & Günnemann, 2020) is a generalization of the categorical distribution, which can represent not only the probability of each class but also the uncertainty of the probability. TMC (Han, Zhang, Fu, & Zhou, 2022) assesses the uncertainty of the results by mapping the classification evidence to the Dirichlet space and fusing the evidence to obtain a plausible classification result. The method of evidence fusion can combine multiple sources of information and handle conflicting or incomplete evidence. However, most existing methods of uncertainty estimation focus on accurately measuring the uncertainty of classification without considering how to use it for tasks. In this paper, we apply the measurement of uncertainty to the task of VI-ReID, which aims to correct predictions to improve the performance of matching person images across different modalities (visible and infrared). VI-ReID is challenging due to the large modality domain gap and intra-class variations. Thus, a specific evidence loss is designed to guide the network to obtain more evidence, enhancing the network's discriminative ability and modality alignment.

3. The proposed method

3.1. Problem definition and overview

The variables x^{Vis} and x^{In} denote original samples from the visible and infrared modalities, respectively, with their respective ground truth labels y^{Vis} and y^{In}. During the training process, cross-modal data with the same identity are fed into the network architecture. The training set is defined as $X_{train} = \left\{\left(x_i^{Vis}, y_i^{Vis}\right); \left(x_i^{In}, y_i^{In}\right)\right\}_{i=1}^{N}$, where $y_i^{Vis} = y_i^{In}$. Here, N represents the batch size.

The structure of our proposed IMKA-UE framework is illustrated in Fig. 2. It consists of four main phases: implicit modality generation (IMG), feature extraction, significant knowledge alignment (SKA),

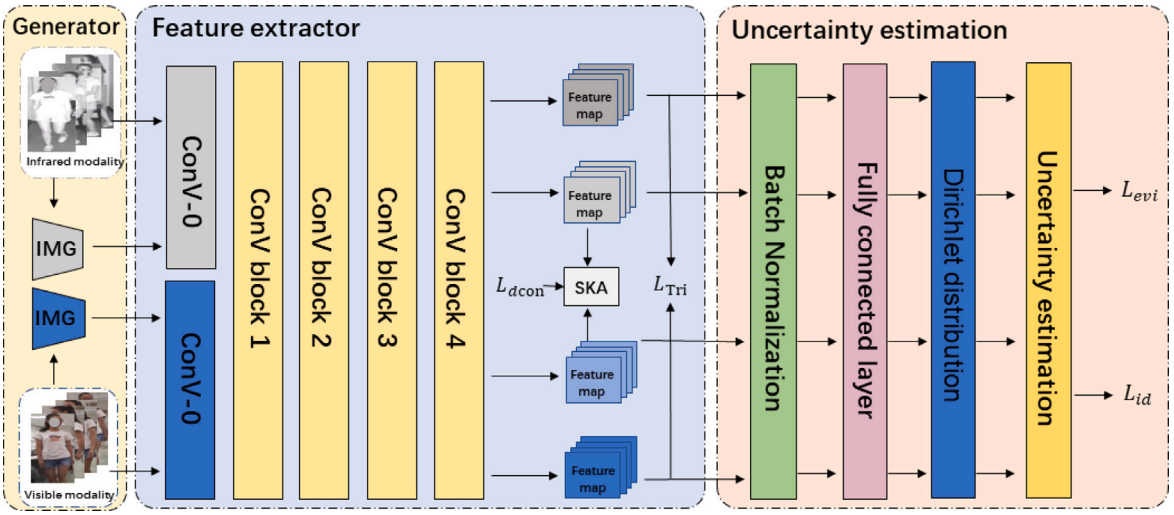


Fig. 2. Our proposed framework of IMKA-UE consists of a two-stream backbone network $F(\cdot)$, two Implicit Modality Generators (IMG) $G(\cdot)$, a Significant Knowledge Alignment (SKA), and an Uncertainty Estimation (UE) with a classifier $C(\cdot)$. The implicit modality samples of two different modalities are generated via IMG. As the feature extractor, ResNet50 (He, Zhang, Ren, & Sun, 2016) is used to extract the modality-special features of the two modalities with their respective modalities in Conv-0 and learn modality-sharing features in the remaining stages.

and uncertainty estimation (UE). In the implicit modality generation phase, the IMG module generates implicit samples for each of the two modalities. The backbone for the feature extractor is based on ResNet-50. Mode-specific features are extracted in the ConV-0 phase. The remaining convolutional blocks of the network extract cross-modal shared features with weight sharing. In the significant knowledge alignment phase, the distribution of the extracted significant knowledge representations is aligned using the distribution consistency loss. The uncertainty estimation module treats the output prediction probability of the network classifier as classification evidence. It maps the evidence to the Dirichlet space to quantify uncertainty and evaluate the accuracy of the classification results.

3.2. Implicit modality knowledge alignment

3.2.1. Implicit modality generation

It is a fact that heterogeneous physical properties between visible and infrared modalities result in losing fundamental information in the common embedding space learning, and certain lost fundamental information can enhance the discrimination and robustness of the learned common embedding space. Thus, the Implicit Modality Learning mechanism is designed to accumulate certain lost important information in the training process to generate implicit samples. The generated implicit samples can be used as beneficial supplementary information to the original modality data to supervise the optimization process and can also be used as an effective data augmentation way to robust the learning process.

The details of the designed Implicit Modality Generator (IMG) are shown in Fig. 3. It consists of a progressive convolution in which the significant supplementary information to the original samples can be extracted, and the implicit samples are generated based on this supplementary information to supervise the shared embedding space learning. For each original sample x_i, the corresponding implicit sample is generated by IMG $G(\cdot)$. The implicit samples can be obtained as follows,

$$x_i^{iVis} = G_1(x_i^{vis}); x_i^{iIn} = G_2(x_i^{in}) \tag{1}$$

Here G_1 and G_2 are two IMGs. x_i^{iVis} and x_i^{iIn} are implicit samples for the visible and infrared modality original samples, respectively. The generated implicit samples are used as supplementary information to the original samples to complement the knowledge accumulation of the backbone network. Compared with the GAN-based pixel-level image generation process, Our method does not have a complex generator discriminator, and all the generation process consists of only two convolutional layers. IMG generates implicit samples in a more lightweight way and easily preserves the identity information of the original samples. Based on the generated implicit samples, the training set can be further represented as follows,

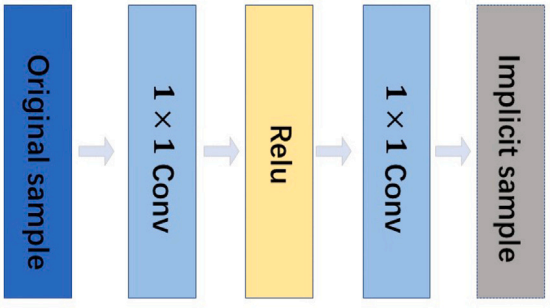


Fig. 3. This is the detailed architecture of the Implicit Modality Generator (IMG). Implicit modality samples are generated by progressive convolution.

$$X_{train} = cat\left\{X_i^{Vis}, X_i^{iVis}, X_i^{In}, X_i^{iIn}\right\} \tag{2}$$

$$y_i^{Vis} = y_i^{iVis} = y_i^{In} = y_i^{iIn} \tag{3}$$

Here y_i denotes the labels of their corresponding samples. For each modality and their corresponding implicit samples, cross-entropy loss L_{ce} is applied to supervise the network so that it fits the information about the identity attributes of the pedestrians,

$$\mathcal{L}_{ce}(p_i) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i^m \mid C(F(x_i^m))\right)\right) \tag{4}$$

where $m \epsilon \{Vis, iVis, In, iIn\}$ and n represents the number of training samples in a mini-batch, $F(\cdot)$ and $C(\cdot)$ are the feature extractor and the classifier. p_i and $p(y_i^m|C(F(x_i^m)))$ denotes the probability of the sample x_i^m being assigned to the ground truth class y_i^m. Also, considering the semantic gap between visible and infrared modalities, cross-modal

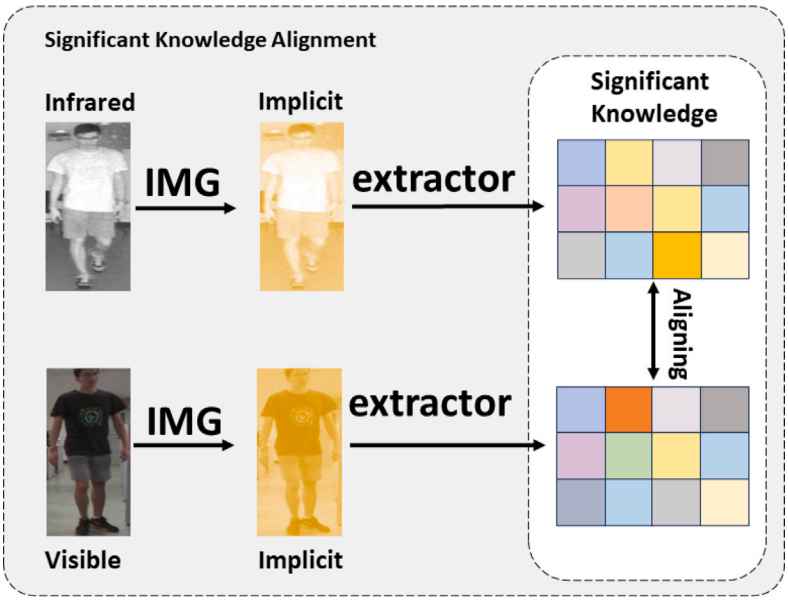


Fig. 4. This is a detailed illustration of the SKA. The distribution of learned significant knowledge in the generated implicit modality samples by the IMG will be enforced to be aligned by the distribution consistency loss.

triplet loss $\mathcal{L}_{Tri}$ is used as an inter-modal constraint.

$$\mathcal{L}_{tri}\left(m_j, n_k\right) = \sum_{i=1}^{N}\left[p + \max_{y_i=y_k}\left\|x_i - m_j\right\|_2 - \min_{y_i\neq y_k}\left\|x_i - n_k\right\|_2\right]_+ \tag{5}$$

$$\mathcal{L}_{Tri} = \mathcal{L}_{tri}\left(x^{Vis}, x^{In}\right) + \mathcal{L}_{tri}\left(x^{Vis}, x^{iIn}\right) + \mathcal{L}_{tri}\left(x^{iIn}, x^{iVis}\right) + \mathcal{L}_{tri}\left(x^{iVis}, x^{In}\right) \tag{6}$$

here p is a margin parameter. m_j and n_k denote the positive and negative samples of sample x_i, respectively.

3.2.2. Significant knowledge alignment

The excessive discrete distribution of the generated implicit sample may weaken the knowledge supplementary to the original sample. As shown in Fig. 4, to prevent the IMG model from accumulating redundant and useless knowledge and to better mediate the distribution of learned significant knowledge representations in the generated implicit samples corresponding to the original modalities, a distribution consistency loss $\mathcal{L}_{dcon}$ is imposed between the feature representations of the learned significant knowledge in the generated implicit samples. It is applied to control the distribution of the learned significant knowledge representations of two generated implicit sample modalities and better bridge the semantic information between visible and infrared modalities. $\mathcal{L}_{dcon}$ enforce the centroids of the cross-modal implicit sample significant knowledge representations close to each other, meanwhile ensuring significant knowledge representations in the generated implicit samples corresponding to the original modality data are better distributed in the learned latent space, which facilitates the training process of supplementary knowledge accumulation. $\mathcal{L}_{dcon}$ is defined as follows:

$$\mathcal{L}_{dcon} = \sum_{i=1}^{K}\left\|\frac{1}{M}\sum_{m=1}^{M}F(x_j^{iVis}) - \frac{1}{N}\sum_{n=1}^{N}F(x_j^{iIn})\right\|_2 \tag{7}$$

here K denotes the number of total classes. M and N denote the number of visible and infrared modality classes, respectively. $F(.)$ and $\|:\|_2$ function feature extractor and L2-norm to measure the distance of samples.

3.3. Uncertainty estimation

Considering the VI-ReID as a K-classification problem, a general classification model will follow a classifier based on the Softmax function to obtain the prediction probability of each person's identity. A disadvantage of this process is that the network's classification probability for identity is considered correct and reasonable even if the prediction results do not match the label. For this reason, the certainty of the network about the classification probability is quantified by introducing an uncertainty estimation mechanism and an evidence loss, which aims to quantify the network's confidence in the classification results.

The output of the classifier is represented as evidence, which represented by $e_v = \{e_1, \ldots, e_K\}$. K denotes the total number of categories of the identity samples, and e_i is the evidence supporting that the sample belongs to the ith identity class. Refer to the EDL (Sensoy et al., 2018) approach of inferring the Dirichlet distribution a_v from the evidence e_v, where a_v is designed by $a_v = e_v + 1$. Thus the classification probability b_i and uncertainty for the probability u_i can be obtained as follow,

$$b_i = \frac{e_i}{S_v} \tag{8}$$

$$u_i = 1 - \sum_{i=1}^{N} b_i = \frac{K}{S_v} \tag{9}$$

Where $S_v = \sum_{v=1}^{K} a_v = \sum_{v=1}^{K}\left(e_v + 1\right)$, and b_i represents the probability of the classification result. K is the number of classes. u_i represents our uncertainty about this classification result. Thus, the certainty of the predictions result of the network output can be effectively quantified.

A typical example under a triple classification task is provided to explain the uncertainty estimation mechina. It is assumed that the evidence for the sample obtained by the model classifier is $e = \{20, 1, 1\}$, so that the corresponding Dirichlet distribution can be denoted as $a = \{21, 2, 2\}$. Based on Eq. (8), the probability of a sample being in the first class $b_1 = \frac{20}{25}$ can be calculated, and the probability of a sample being in the second and third class are $b_2 = \frac{1}{25}$ and $b_3 = \frac{1}{25}$, respectively. Network's uncertainty of this classification result is $u_1 = \frac{3}{25}$, whcih can be obtained by Eq. (9). As shown in Fig. 5(a), a sharp distribution centered on one corner of the triangle's base is obtained. This distribution suggests that we have sufficient evidence to support

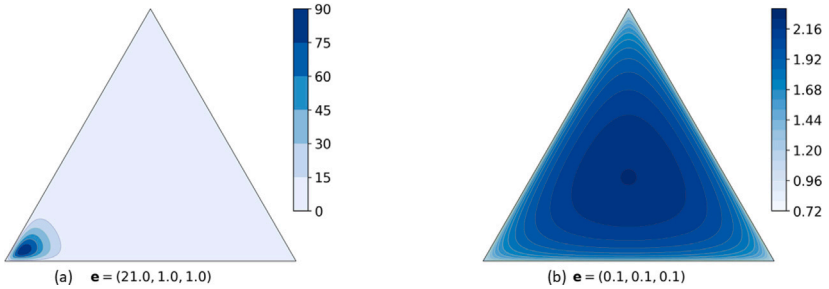


Fig. 5. This is an illustration of Dirichlet distribution. The probability density function of the Dirichlet distribution within an equilateral triangle is plotted. (a) represents the category where the evidence is mainly concentrated in the lower right corner; thus, it is a well-documented Dirichlet distribution with low uncertainty. (b) indicates that the evidence supporting the classification is evenly distributed in the three corners; thus, it is a Dirichlet distribution with high uncertainty and no clear evidence that the sample belongs to the definite class.

that the sample belongs to the first category, and we are confident about this classification result.

In contrast, let us assume that we have evidence $e = \{0.1, 0.1, 0.1\}$. This evidence is microscopic evidence for classification. Therefore, the Dirichlet distribution parameter $a = \{1.1, 1.1, 1.1\}$ can be obtained, and the uncertainty $u \approx 1$ for this classification can be calculated. Fig. 5(b) shows us a uniform evidence distribution over the three classes. By estimating the uncertainty u, it can be noted that we are unsure about the result of this classification, and little evidence can be obtained to support the classification. In this case, we cannot determine which category the sample belongs to, so it needs to be further optimized in the feature learning stage to obtain more classification evidence.

To sum up, we introduced the Dirichlet distribution and obtained the uncertainty of each classification without changing the classification results in the above way. Uncertainty is used for further optimization to help the model obtain more evidence supporting classification in order to ensure the accuracy of classification, thus helping VI-ReID solve the semantic gap.

By introducing the theory of uncertainty estimation, the evidence loss $\mathcal{L}_{evi}$ is designed to reduce the uncertainty of the network prediction, which, thus, can help the network to address the problem of overconfidence about the wrong classification results and obtain more evidence supporting the classification. The $\mathcal{L}_{evi}$ is designed as follow:

$$\mathcal{L}_{evi} = \sum_{i=1}^{n} \left(1 - \frac{e_v}{\sum_{v=1}^{K} (e_v + 1)} \right) = \sum_{i=1}^{n} u_i \tag{10}$$

$$\mathcal{L}_{id}(b_i) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{e_v}{\sum_{v=1}^{K} (e_v + 1)} \right) \tag{11}$$

$\mathcal{L}_{id}$ uses the sample classification probability evaluated by the uncertainty estimation mechanism. The evidence loss $\mathcal{L}_{evi}$ is used to guide the optimization of the network prediction values. By the optimization of $\mathcal{L}_{id}$ and $\mathcal{L}_{evi}$, it can be ensured that the classification results are more accurate and that higher confidence can be obtained in the classification results. The gradient of $\mathcal{L}_{evi}$ during back-propagation is as follows,

$$\begin{aligned} \frac{\partial \mathcal{L}_{evi}}{\partial e_v} &= \sum_{i=1}^{n} \frac{\partial u_i}{\partial e_v} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial e_v} \left(1 - \frac{e_v}{\sum_{v=1}^{K} (e_v + 1)} \right) \\ &= -\sum_{i=1}^{n} \frac{\sum_{v=1}^{K} (e_v + 1) - e_v}{(\sum_{v=1}^{K} (e_v + 1))^2} \end{aligned} \tag{12}$$

We can conclude that an increase in e_v will cause a decrease in $\mathcal{L}_{evi}$ and vice versa. Therefore, when optimizing $\mathcal{L}_{evi}$, we can obtain more sufficient evidence for judging the identity of pedestrians.

Algorithm 1 Algorithm of our IMKA-UE method

Input: Visible data x_i^{Vis} and infrared data x_i^{In} with their labels y_i^{Vis} and y_i^{In} ;
Output: The trained feature extractor $F(\cdot)$;
Initialization: The backbone ResNet-50 (He et al., 2016) with pre-trained weights on ImageNet serves as feature extractor $F(\cdot)$, totally iterations E and penalty factors λ_1, λ_2 and λ_3 for the overall loss function **Eq.** (13).
For $i = 1$ to W **do**

- 1. The IMG module generates implicit modality samples x^{iVis} and x^{iIn} by Eq. (1);
- 2. The ConV-0 of the feature extractor extracts specific features of the two modalities;
- 3. The remaining phase of the feature extractor extracts the shared features;
- 4. The SKA module aligns the significant knowledge distribution of generated implicit modality samples;
- 5. Classifier $C(\cdot)$ obtains classification evidence;
- 6. Uncertainty estimation is evaluated according to Eq. (9);
- 7. Update the parameters of the network with SGD according to Eq. (13);

End For
Return: The trained feature extractor $F(\cdot)$;

3.4. Overall training

The overall training loss $\mathcal{L}_{all}$ is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}_{dcon} + \lambda_2 \mathcal{L}_{evi} \tag{13}$$

where λ_1 is the weight of the loss function. L_{base} is the baseline loss function, which contains L_{id} and L_{Tri}. The algorithm of our method is given in Algorithm 1.

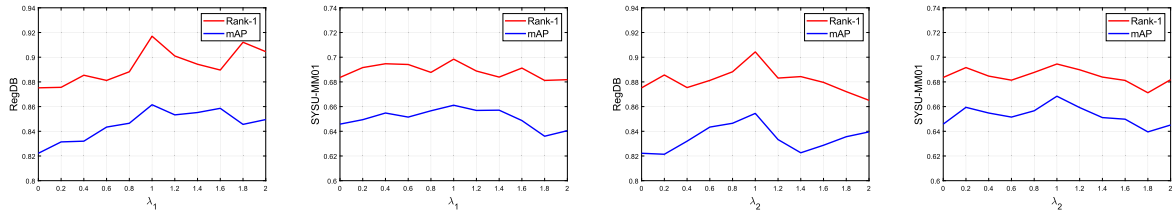
4. Experiments

4.1. Datasets and settings

4.1.1. Datasets

We use two public Visible-Infrared datasets, SYSU-MM01 (Wu, Zheng, Yu, Gong, & Lai, 2017) and RegDB (Dat, Hong, Ki, & Kang, 2017), to evaluate the proposed IMKA-UE. These datasets are widely used for the VI-ReID task and present different characteristics and challenges. We introduce them as follows:

- **SYSU-MM01:** This dataset was captured by six cameras in both indoor and outdoor scenes, including two indoor visible cameras,



(a) Performance impact of different λ_1 of Eq. (13 on RegDB (b) Performance impact of different λ_1 of Eq. (13) on SYSU-MM01 (c) Performance impact of different λ_2 of Eq. (13) on RegDB (d) Performance impact of different λ_2 of Eq. (13) on SYSU-MM01

Fig. 6. Performance impact of different penalty factors of Eq. (13).

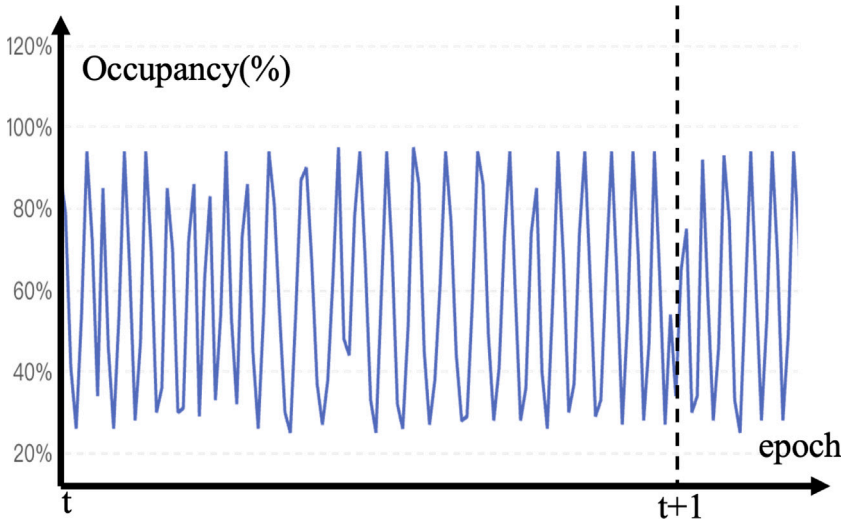


Fig. 7. GPU occupancy changes as the epoch progresses. As the GPU occupancy changes periodically with epoch, we take epoch iteration t to $t+1$ as an illustration to show the change of GPU occupancy.

two outdoor visible cameras, and two infrared cameras. It is a large-scale VI-ReID dataset containing a total of 491 pedestrian identities. Each identity has an average of 49.6 visible images and 24.3 infrared images. The dataset provides two test modes: all search and indoor search. In the all search mode, all the images from the six cameras are used as the gallery, while in the indoor search mode, only the images from the four indoor cameras are used as the gallery. The query set contains 2900 visible images and 2900 infrared images in both test modes. The training set contains 395 identities with 22,258 visible images and 11,909 infrared images. The remaining 96 identities are used for evaluation.

- **RegDB:** This dataset was acquired using a switchable mode camera that can capture both visible and infrared images. Ten photos were collected for each pedestrian in both modes under different illumination conditions and backgrounds. It contains a total of 412 pedestrian identities with 2060 visible images and 2060 infrared images. We randomly selected 206 identities to form the training set with 1030 visible images and 1030 infrared images. The remaining 206 identities were used for testing with another 1030 visible images and 1030 infrared images. Two test modes were adopted: Visible to Infrared and Infrared to Visible. In the Visible to Infrared mode, the visible images were used as queries and the infrared images were used as the gallery. In the Infrared to Visible mode, the infrared images were used as queries and the visible images were used as the gallery.

4.1.2. Experiments setting

The proposed method is implemented using PyTorch. The initial learning rate is set to 0.1, and the training epoch iteration t is set to 80. In the first ten epochs, the learning rate decreases gradually. From 10 to 20 epochs, the learning rate remains constant. From 20 to 50 epochs, it decreases to one-tenth, and from 50 to 80 epochs, it decreases to one percent. We use SGD for optimization, with p from Eq. (5) set to 0.3. The original samples and the corresponding implicit samples were trained on NVIDIA 3090 GPU in a collaborative training fashion. Training the SYSU-MM01 dataset took 35 h, while the RegDB dataset took 20 h. We show the change in GPU occupancy as epochs proceed in Fig. 7. We use the SGD optimizer with a weight decay of $5e^{-4}$. The batch size is set to 8 for each modality, and the training epoch is set to 80. During training, each image is resized to a length of 384 and a width of 192. For data augmentation, channel random erasing, random horizontal flipping, and random cropping are implemented. The $rank-k$ accuracy, mAP, and a new evaluation criterion mINP are utilized to evaluate the performance of our IMKA-UE and the compared methods. As depicted in Fig. 6, we demonstrate the difference in model performance with different weights. We can see that when the weights of the overall loss $\mathcal{L}_{all}$ in Eq. (13) are set to $\lambda_1 = \lambda_2 = 1$, the performance on both datasets is optimal.

4.1.3. Evaluation protocol

We follow the evaluation protocol described in Ye, Shen et al. (2021), which adopts three metrics to evaluate the performance of

Table 1
Comparison with the state-of-the-art methods on SYSU-MM01 dataset. The best results are highlighted in bold and sub-optimal results are underlined.

Method	Reference	All Search					Indoor Search				
		Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
SSFT (Lu et al., 2020)	CVPR20	47.7	–	–	54.1	–	–	–	–	–	–
Hi-CMD (Choi, Lee, Kim, Kim, & Kim, 2020)	CVPR20	34.9	77.6	–	35.9	–	–	–	–	–	–
X-Modal (Li et al., 2020)	AAAI20	49.9	89.8	96.0	50.7	–	–	–	–	–	–
DDAG (Ye, Shen, Crandall, Shao, & Luo, 2020)	ECCV20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.40	67.98	62.61
AGW (Ye, Shen et al., 2021)	TAPMI21	47.50	84.39	62.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
CICL (Zhao et al., 2021)	AAAI21	57.2	94.3	98.4	59.3	–	66.7	95.2	99.7	74.7	–
cmAlign (Park et al., 2021)	ICCV21	55.41	–	–	54.14	–	58.46	–	–	66.33	–
NFS (Chen et al., 2021)	CVPR21	56.90	91.30	96.50	55.50	–	62.80	96.50	99.10	69.80	–
KDEM (Shan, Xiong, Yuan, & Wu, 2022)	ICANN22	58.09	91.19	96.63	55.52	40.69	62.18	94.38	98.64	68.30	64.11
MID (Huang et al., 2022)	AAAI22	60.27	92.90	–	59.40	–	64.86	96.12	–	70.12	–
SPOT (Chen et al., 2022)	TIP22	65.34	92.73	97.04	62.25	–	69.42	96.22	99.12	74.63	–
DART (Yang et al., 2022)	CVPR22	68.72	96.39	98.96	66.29	53.26	72.52	97.84	99.46	78.17	74.94
FMCNet (Zhang et al., 2022)	CVPR22	66.34	–	–	62.51	–	68.15	–	–	74.09	–
DCLNet (Sun et al., 2022)	ACM MM22	70.79	–	–	65.18	–	73.51	–	–	76.80	–
DMiR (Hu, Liu, Zeng, Hou, & Hu, 2022)	TCSVT22	48.63	87.34	94.47	47.79	–	53.56	91.87	96.94	62.10	–
DMiR* (Hu et al., 2022)	TCSVT22	50.54	88.12	94.86	49.29	–	53.92	92.50	97.09	62.49	–
CmCen (Xu, Liu, Zhang, Xiao, & Wu, 2022)	KBS22	50.09	86.94	93.29	49.76	–	56.84	93.04	97.69	63.55	–
TSME (Liu, Wang et al., 2022)	TCSVT22	64.23	95.19	98.73	61.21	–	64.80	96.92	99.31	71.53	–
PMT (Lu, Zou, & Zhang, 2023)	AAAI23	67.53	95.36	98.64	64.98	51.86	71.66	96.73	99.25	76.52	72.74
IMKA-UE	ours	**71.94**	**96.78**	**99.21**	**67.69**	**53.81**	**75.82**	**97.86**	**99.57**	**79.58**	**75.71**

cross-modal person Re-ID methods: Cumulative Matching Characteristic (CMC), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). The CMC curve measures the probability that a correct match appears in the top-k ranked list for a given query image, reflecting the top retrieval accuracy of the cross-modal person Re-ID method. The mAP score calculates the average precision for each query image and then averages them over all queries, reflecting the overall retrieval quality of the cross-modal person Re-ID method by considering both precision and recall at different ranks. A higher CMC or mAP value indicates better performance of the cross-modal person Re-ID method. However, both CMC and mAP are biased towards easy queries and ignore hard queries, which may not reflect the robustness of the cross-modal person Re-ID method. Therefore, we also employ the mINP metric, a new evaluation metric for visible-infrared person re-identification (VI-ReID) that measures the cost of finding all the correct matches for a given query image. The mINP metric (Ye, Shen et al., 2021) considers both the rank and the number of true positives and false positives at different ranks, penalizing more for missing hard matches than easy matches. A higher mINP value indicates better performance and robustness of the VI-ReID method.

4.2. Comparison with state-of-the-art methods

Mean average precision (mAP) (%), Rank-1, 10, 20 accuracy (%) and mean Inverse Negative Penalty (mINP)(%) on RegDB and SYSU-MM01 are reported in Tables 1 and 2. Comprehensive comparison results with the state-of-the-art methods are shown in Tables 1 and 2, which include AlignGAN (Wang et al., 2020), SSFT (Lu et al., 2020), Hi-CMD (Choi et al., 2020), X-Modal (Li et al., 2020), DDAG (Ye et al., 2020), AGW (Ye, Shen et al., 2021), CICL (Zhao et al., 2021), cmAlign (Park et al., 2021), NFS (Chen et al., 2021), KDEM (Shan et al., 2022), MID (Huang et al., 2022), SPOT (Chen et al., 2022), DART (Yang et al., 2022),FMCNet (Zhang et al., 2022), DCLNet (Sun et al., 2022), DMiR (Hu et al., 2022), DMiR* (Hu et al., 2022), CmCen (Xu et al., 2022), TSME (Liu, Wang et al., 2022) and PMT (Lu et al., 2023). Here, DMiR uses an image size of (128, 256), and DMiR* uses an image size of (144, 288). It is clear that our proposed framework of IMKA-UE advances the performance of VI-ReID to a new benchmark. The experimental comparison results with the state-of-the-art methods illustrate that IMKA-UE has significant advantages in reducing the large semantic domain gap between visible and infrared modality data in the task of VI-ReID.

4.2.1. Results of SYSU-MM01 analysis

SYSU-MM01 is the first standard dataset for cross-modality person re-identification and is considered the authoritative dataset for VI-ReID. The results of IMKA-UE on the SYSU-MM01 dataset are shown in Table 1, where it achieves a new state-of-the-art performance in both all search and indoor search modes. In both test modes, IMKA-UE outperforms the previous best method, DCLNet (Sun et al., 2022), by 1.15% and 2.31% in terms of Rank-1 accuracy, respectively. For the mAP metric, which measures overall retrieval quality, IMKA-UE also surpasses the best-performing DART (Yang et al., 2022) by 1.40% and 1.41% in both test modes, respectively. Additionally, our IMKA-UE demonstrates state-of-the-art performance in both test modes for the mINP metric, reflecting the robustness of the retrieval system. Compared to the current state-of-the-art DART (Yang et al., 2022), the improvement is 0.55% and 0.77%, respectively. It is clear that IMKA-UE achieved a new state-of-the-art performance in both all search and indoor search modes on the SYSU-MM01 dataset, surpassing the previous best methods, DCLNet and DART.

4.2.2. Results of RegDB analysis

The performance of IMKA-UE on the RegDB dataset is presented in Table 2, where it achieves the best performance compared to state-of-the-art methods in any test mode with significant improvement. In the Visible to Infrared test mode, which uses visible images as queries and infrared images as the gallery, the performance of our IMKA-UE surpasses the previous best method, FMCNet, by 3.47% in terms of Rank-1 accuracy, indicating the top retrieval accuracy. Moreover, for the mAP metric, which evaluates overall retrieval quality, IMKA-UE improves by 4.22% over the second-best MID. Additionally, IMKA-UE outperformed DART (Yang et al., 2022) by 5.19% under the mINP metric, reflecting the robustness of the IMKA-UE-based retrieval system. In the Infrared to Visible test mode, which uses infrared images as queries and visible images as the gallery, IMKA-UE also shows superior performance over other methods. Compared to the second-best FMCNet in this test mode, IMKA-UE achieves performance improvements of 3.00% and 2.43% in Rank-1 and mAP metrics, respectively. For the mINP metric, compared to the current state-of-the-art DART, the improvement of IMKA-UE is 7.85%. These results demonstrate the effectiveness and superiority of our proposed IMKA-UE for cross-modal person re-identification.

4.2.3. Rank list of IMKA-UE

The VI-ReID task is a cross-modal retrieval task that uses pedestrian images in one modality as a query to retrieve pedestrian sequences

Table 2
Comparison with the state-of-the-art methods on RegDB dataset. The best results are highlighted in bold and sub-optimal results are underlined.

Method	Reference	Visible to Infrared			Infrared to Visible		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
SSFT (Lu et al., 2020)	CVPR20	65.4	65.6	–	63.8	64.2	–
Hi-CMD (Choi et al., 2020)	CVPR20	70.9	66.0	–	–	–	–
X-Modal (Li et al., 2020)	AAAI20	62.2	60.2	–	–	–	–
DDAG (Ye et al., 2020)	ECCV20	69.34	63.46	49.24	68.06	61.80	49.24
AGW (Ye, Shen et al., 2021)	TAPMI21	70.05	66.37	50.19	70.04	65.90	50.19
cmAlign (Park et al., 2021)	ICCV21	74.17	67.64	–	72.43	65.40	–
NFS (Chen et al., 2021)	CVPR21	80.50	72.10	–	78.00	69.80	–
CICL (Zhao et al., 2021)	AAAI21	78.8	69.4	–	77.9	69.4	–
CAJ (Ye, Ruan et al., 2021)	AAAI21	78.8	69.4	–	77.9	69.4	–
KDEM (Shan et al., 2022)	ICANN22	77.33	70.32	56.08	76.26	67.77	52.38
MID (Huang et al., 2022)	AAAI21	87.45	84.85	–	84.29	81.41	–
SPOT (Chen et al., 2022)	TIP22	80.35	72.46	–	79.37	72.26	–
DART (Yang et al., 2022)	CVPR22	83.6	75.67	60.6	81.97	73.78	56.7
FMCNet (Zhang et al., 2022)	CVPR22	89.12	84.43	–	88.38	83.86	–
DCLNet (Sun et al., 2022)	ACM MM22	81.2	74.3	–	78.0	70.6	–
DMiR (Hu et al., 2022)	TCSVT22	74.50	69.18	–	73.17	67.50	–
DMiR* (Hu et al., 2022)	TCSVT22	75.79	69.97	–	73.93	68.22	–
CmCen (Xu et al., 2022)	KBS22	74.03	67.52	–	74.22	67.36	–
TSME (Liu, Wang et al., 2022)	TCSVT22	87.35	76.94	–	86.41	75.70	–
PMT (Lu et al., 2023)	AAAI23	84.83	76.55	–	84.16	75.13	–
IMKA-UE	ours	**92.59**	**88.67**	**65.79**	**91.38**	**86.29**	**64.55**

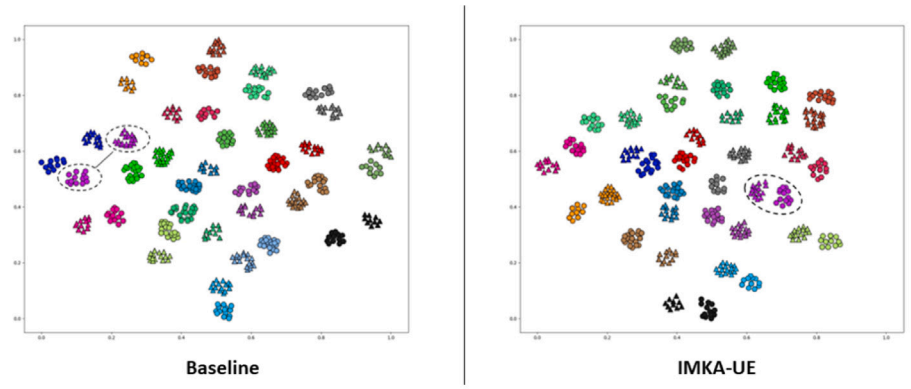
Fig. 8. Rank list visualization of IMKA-UE. Including yellow boxes that indicate the queries, green boxes that represent correctly retrieved images, and red boxes that refer to misclassified images.

in another modality. The IMKA-UE visual search sequence is shown in Fig. 8, arranged from top to bottom. The sequence corresponds to visible-to-infrared and infrared-to-visible test modes on RegDB, as well as All Search and Indoor Search on SYSU-MM01. This figure demonstrates that IMKA-UE achieves a high level of correctness, indicating its excellent retrieval capabilities in cross-modal person re-identification. Specifically, we achieved a perfect retrieval sequence for the small-scale RegDB dataset, with no retrieval errors in either query mode. However, for the SYSU-MM01 dataset, some retrieval errors still occurred. For example, in the third type, we retrieved two pictures of the same person wearing blue jeans with similar postures. These two pictures are similar in posture and clothing, leading to mismatches. In the fourth indoor search mode, the system mistakenly matched two pictures of women with similar clothing styles and body shapes, causing misjudgments.

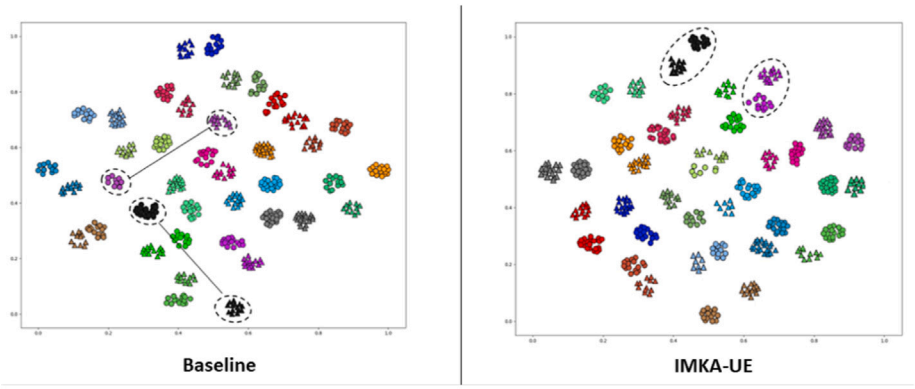
4.3. Ablation study

4.3.1. Ablation study of IMKA-UE

The ablation experiments were evaluated on the SYSU-MM01 and RegDB datasets. Rank-K accuracy (%) and mAP (%) are reported in



(a) T-SNE analysis of IMKA-UE on SYSU-MM01 dataset



(b) T-SNE analysis of IMKA-UE on RegDB dataset

Fig. 9. The distribution of feature embeddings in 2D feature space, where circles and triangles denote visible and infrared modalities. Different colors represent different categories.

Table 3
Ablation experiments on the SYSU-MM01(All Search).

Setting	Rank-1	mAP	mINP
Baseline	63.41	59.54	41.34
Baseline + x^{iVis}	67.55	62.25	46.54
Baseline + x^{iIn}	66.16	62.18	46.42
Baseline + x^{iVis} + x^{iIn}	68.36	64.58	49.87
Baseline + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$	69.84	66.11	51.58
Baseline + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$ + $\mathcal{L}_{evi}$	**71.94**	**67.69**	**53.81**

Tables 3 and 4. Experimental results demonstrate the effectiveness of the IMG module, the SKA module based on the distribution consistency loss, and the UE strategy based on the evidence loss. Firstly, the performance of VI-ReID is significantly improved based on the supervision from the generated implicit samples by the IMG module. It can be inferred that through the learning of implicit samples, the model accumulates valuable supplementary knowledge of the unseen domain to compensate for the loss of basic information in the embedding space caused by the natural differences between visible and infrared modalities, thus narrowing the semantic gap between modalities.

The SKA module, based on the distribution consistency loss $\mathcal{L}_{dcl}$, effectively constrains the distribution of learned significant knowledge in the generated implicit samples to align in the learned latent common embedding space. Promising experimental results illustrate that the implicit samples are concentrated as much as possible in the embedding space, improving the quality of visible and infrared modality clustering. Moreover, the UE strategy based on the evidence loss provides more evidence to support classification. With the evidence loss $\mathcal{L}_{evi}$, it can be effectively ensured that the model is more confident about each classification result. This eliminates situations where the model makes a high-confidence error due to missing classification evidence.

Fig. 10. This is the visualization of generated implicit modality samples. Images of each row are from the same identity. The left image in the box is the original modality sample, and the right one is generated implicit modality sample by IMG.

To further validate the effectiveness of our method, we implemented our algorithm on the VI-ReID method AGW using global features (Ye, Shen et al., 2021). The results on RegDB (Dat et al., 2017) are shown

Fig. 11. This is the visualization of class activation mapping of our trained IMAK-UE-based model, and images of each row are from the same identity.

Table 4
Ablation experiments on the RegDB(Visible to Infrared).

Setting	Rank-1	mAP	mINP
Baseline	76.99	69.98	53.22
Baseline + x^{iVis}	84.47	80.56	58.40
Baseline + x^{iIn}	82.18	79.12	58.21
Baseline + x^{iVis} + x^{iIn}	87.52	82.22	63.44
Baseline + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$	91.70	86.15	64.56
Baseline + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$ + $\mathcal{L}_{evi}$	**92.59**	**88.67**	**65.79**

Table 5
IMKA-UE performance on the AGW.

Setting	Rank-1	mAP	mINP
AGW	70.05	66.37	50.19
AGW + x^{iVis}	76.94	67.25	51.64
AGW + x^{iIn}	74.81	67.50	53.00
AGW + x^{iVis} + x^{iIn}	78.98	71.19	56.68
AGW + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$	82.18	73.12	57.08
AGW + x^{iVis} + x^{iIn} + $\mathcal{L}_{dcon}$ + $\mathcal{L}_{evi}$	**84.09**	**77.35**	**59.45**

in Table 5. Both IMKA and UE work well and demonstrate their effectiveness. Experimental results show that our method has good generalizability. In Fig. 9, we use the T-SNE (Laurens & Hinton, 2008) technique to show the feature distribution in 2D space. T-SNE (Laurens & Hinton, 2008) is a non-linear dimensionality reduction technique that maps high-dimensional data to a low-dimensional space while preserving the similarity between data points. The left side of the figure shows the baseline network, and the right side shows the proposed IMKA-UE. We mark significant improvements over the baseline with dashed black lines. It can be seen that the samples of the two modalities that were originally clustered far apart are clustered together based on IMKA-UE (see Fig. 9).

5. Discussion

5.1. Implicit modality learning analysis

The IMG module is a lightweight generator whose encoder–decoder is performed by a 1×1 convolution. The encoder convolves the three-channel image into a single channel, and the decoder reconstructs it into a three-channel image. The quality of the images generated by IMG is shown in Fig. 10. It can be seen that the implicit sample is consistent with the original sample in terms of identity. The crux of VI-ReID lies in extracting the common semantic information of two modalities. The difference in the inherent physical information of the two modalities makes it difficult to extract this common semantic information. Visible modality data has rich color information, while infrared modality data only has pedestrian texture information. The starting point of our paper is to bridge the inherent gap between the two modalities by learning

Table 6
Impact of metric methods.

Setting	Rank-1	Rank-10	Rank-20	mAP	mINP
L1-Norm	91.76	97.89	98.86	86.71	61.87
L2-Norm	**92.59**	**98.29**	**99.24**	**88.67**	**65.79**
cos-similarly	92.13	98.00	99.13	87.96	62.25

the implicit modalities generated by IMG. The model learns significant knowledge from both visible and infrared modalities, supplemented by knowledge from implicit modalities. Implicit modalities can be inserted between the cross-modal data as supplementary information, bridging the semantic gap between visible and infrared modalities to enhance the learned common embedding space.

5.2. Impact of metric methods

The distribution consistency loss $\mathcal{L}_{dcon}$ ensures the significant knowledge in the implicit modality distribution. Concentrated distributions of implicit modalities can better bridge visible and infrared modalities. Its role is to prevent the generated implicit modalities from being redundant, which is conducive to the process of model knowledge accumulation during training. This subsection evaluates the impact of using different metrics in the consistency loss $\mathcal{L}_{dcon}$ on performance. As shown in Table 6, the impact of L1-Norm, L2-Norm, and Cosine similarity is presented. According to the experimental results, L2-Norm exhibits better results.

5.3. Class activation mapping analysis

Extracting robust pedestrian features that are invariant across modalities is the top priority of VI-ReID. We further demonstrate that IMKA-UE focuses on the discriminative and significant areas in different pedestrian modality data using the Grad-CAM technique. Grad-CAM (Selvaraju, Cogswell, Das, Vedantam, Parikh et al., 2017) is a technique for generating visual interpretations of deep neural networks that can highlight regions in an image that are important for predicting object classes. It uses the gradient information of the correct target category to assign weights to each neuron in the last convolutional layer, then multiplies these weights with the feature map to obtain a heat map.

The class activation mapping of the learned model is shown in Fig. 11. The discriminative and significant areas in different pedestrian modality data focused on by IMKA-UE are shown for various acquisition modes on two datasets. It is clear that the attention regions of the two datasets are significantly different. For the RegDB dataset, the facial information of the image is missing. In this case, IMKA-UE mainly focuses on pedestrians' gait and silhouette information. These are used as semantic information for identifying pedestrian identities during cross-modal retrieval, which will not change under cross-modal settings. For the SYSU-MM01 dataset, we show the heatmap regions of images in three acquisition modes: visible light, infrared, and outdoor acquisition modes. We can see that for the SYSU-MM01 dataset, IMKA-UE uses facial information as an important area for identifying pedestrians while also paying attention to gait information. Additionally, we can observe that IMKA-UE uses the logo information on shoe styles and clothing logos as discrimination areas. This information can determine the identity of a pedestrian in both modes.

5.4. Methodological limitations and future prospects

Our starting point is to design a lightweight generator to replace the complex GAN generation network. It generates suitable latent samples to fill in the missing parts between cross-modal data. Through experiments, we can see that using two convolutional layers instead of GAN to generate samples improves performance. Like all generation methods, we generate additional samples, which undoubtedly increases training time. This is a common problem for all generation methods, but our method uses two convolutional layers to complete the generation process, greatly simplifying the generation complexity. Furthermore, while ensuring that the generated samples are conducive to solving the cross-modal semantics of VI-ReID, the lightweight IMG generation method reduces generation complexity, which is beneficial for the development and practical application of VI-ReID.

6. Conclusion

In this paper, we proposed a novel framework called IMKA-UE for high-performance VI-ReID. The experimental results demonstrated that our IMKA-UE achieves a new state-of-the-art performance on two benchmark datasets, surpassing existing methods by a large margin. The IMKA mechanism effectively bridges the semantic domain gap between visible and infrared modality data. Specifically, the IMG module excels at extracting valuable supplementary information from the original modality data to generate implicit modality data. The distribution of significant knowledge learned from the generated implicit modality data is effectively aligned by the SKA module, which is based on distribution consistency loss. Moreover, the UE strategy is powerful in evaluating the confidence level of the predicted probability in the Dirichlet space. Using the designed evidence loss, the UE strategy can learn valuable evidence to correct incorrect classifications, resulting in more accurate and credible predictions. Our extensive experiments and visualization analyses demonstrate the effectiveness and superiority of our proposed IMKA-UE for VI-ReID.

CRediT authorship contribution statement

Song Wu: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Shihao Shan:** Conceptualization, Methodology, Software, Investigation. **Guoqiang Xiao:** Writing - review & editing. **Michael S. Lew:** Writing - review & editing. **Xinbo Gao:** Methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, China (SWU-KT22032).

References

Chen, Y., Wan, L., Li, Z., Jing, Q., & Sun, Z. (2021). Neural feature search for RGB-infrared person re-identification. ArXiv.

Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., & Lin, C.-W. (2022). Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, *31*, 2352–2364. http://dx.doi.org/10.1109/TIP.2022.3141868.

Choi, S., Lee, S., Kim, Y., Kim, T., & Kim, C. (2020). Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *2020 IEEE/CVF conference on computer vision and pattern recognition*.

Dat, N., Hong, H., Ki, K., & Kang, P. (2017). Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, *17*(3), 605.

Farooq, A., Awais, M., Kittler, J., & Khalid, S. S. (2022). AXM-Net: Implicit cross-modal feature alignment for person re-identification.

Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., & He, R. (2021). CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11823–11832).

Gal, Y., & Ghahramani, Z. (2016). Uncertainty estimation in computer vision with application to medical image segmentation. arXiv preprint arXiv:1703.04977.

Ge, W., Du, J., Wu, A., Xian, Y., Yan, K., Huang, F., & Zheng, W.-S. (2022). Lifelong person re-identification by pseudo task knowledge preservation.

Han, Z., Zhang, C., Fu, H., & Zhou, J. T. (2021). Trusted multi-view classification. In *International conference on learning representations*. URL https://openreview.net/forum?id=OOsR8BzCnl5.

Han, Z., Zhang, C., Fu, H., & Zhou, J. T. (2022). Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15013–15022).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. IEEE.

Hong, Y., Chen, Y., Liang, X., & Zhou, P. (2021). Fine-grained shape-appearance mutual learning for cloth-changing person re-identifification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14792–14801).

Hu, W., Liu, B., Zeng, H., Hou, Y., & Hu, H. (2022). Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(8), 5095–5109. http://dx.doi.org/10.1109/TCSVT.2022.3147813.

Huang, Z., Liu, J., Li, L., Zheng, K., & Zha, Z. J. (2022). Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, *30*.

Kong, L., Sun, J., & Zhang, C. (2020). SDE-net: Equipping deep neural networks with uncertainty estimates. CoRR arXiv:2008.10546, URL https://arxiv.org/abs/2008.10546.

Kopetzki, A., Charpentier, B., Zügner, D., Giri, S., & Günnemann, S. (2020). Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? CoRR arXiv:2010.14986, URL https://arxiv.org/abs/2010.14986.

Laurens, V. D. M., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(2605), 2579–2605.

Li, Z., Shi, Y., Ling, H., Chen, J., Wang, Q., & Zhou, F. (2022). Reliability exploration with self-ensemble learning for domain adaptive person re-identification.

Li, D., Wei, X., Hong, X., & Gong, Y. (2020). Infrared-visible cross-modal person re-identification with an X modality. In *The thirty-fourth AAAI conference on artificial intelligence (AAAI-20)* (pp. 4610–4617).

Liu, J., Wang, J., Huang, N., Zhang, Q., & Han, J. (2022). Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(10), 7226–7240. http://dx.doi.org/10.1109/TCSVT.2022.3168999.

Liu, S., Xiao, G., Lew, M. S., Gao, X., & Wu, S. (2024). Core-attributes enhanced generative adversarial networks for robust image enhancement. *Engineering Applications of Artificial Intelligence*, *131*, Article 107799.

Liu, S., Xiao, G., Xu, X., & Wu, S. (2022). Bi-directional normalization and color attention-guided generative adversarial network for image enhancement. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 2205–2209). IEEE.

Liu, X., Zhang, P., Yu, C., Lu, H., & Yang, X. (2021a). Learning to learn from noisy labels for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14662–14671).

Liu, X., Zhang, P., Yu, C., Lu, H., & Yang, X. (2021b). Multi-task learning for person re-identification with multi-level features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 2286–2295).

Liu, X., Zhang, P., Yu, C., Lu, H., & Yang, X. (2021c). Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14662–14671).

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., & Yu, N. (2020). Cross-modality person re-identification with shared-specific feature transfer. In *2020 IEEE/CVF conference on computer vision and pattern recognition*.

Lu, H., Zou, X., & Zhang, P. (2023). Learning progressive modality-shared transformers for effective visible-infrared person re-identification. *vol. 37*, In *Proceedings of the AAAI conference on artificial intelligence* (2), (pp. 1835–1843).

Park, H., Lee, S., Lee, J., & Ham, B. (2021). Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences.

Pu, N., Chen, W., Liu, Y., Bakker, E. M., & Lew, M. S. (2020). Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2149–2158).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE international conference on computer vision*.

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty.

Shan, S., Xiong, E., Yuan, X., & Wu, S. (2022). A knowledge-driven enhanced module for visible-infrared person re-identification. In E. Pimenidis, P. P. Angelov, C. Jayne, A. Papaleonidas, & M. Aydin (Eds.), *Lecture notes in computer science*: *vol. 13529*, *Artificial neural networks and machine learning - ICANN 2022 - 31st international conference on artificial neural networks, Bristol, UK, September 6-9, 2022, proceedings, part I* (pp. 441–453). Springer, http://dx.doi.org/10.1007/978-3-031-15919-0_37.

Sun, H., Liu, J., Zhang, Z., Wang, C., Qu, Y., Xie, Y., & Ma, L. (2022). Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, & L. Toni (Eds.), *MM '22: the 30th ACM international conference on multimedia, lisboa, Portugal, October 10 - 14, 2022* (pp. 5333–5341). ACM, http://dx.doi.org/10.1145/3503161.3547970.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y. Y., & Satoh, S. (2019). Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *2019 IEEE/CVF conference on computer vision and pattern recognition*.

Wang, G. A., Yang, T., Cheng, J., Chang, J., Liang, X., & Hou, Z. (2020). Cross-modality paired-images generation for RGB-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*.

Wang, Y., Zhang, J., Liang, X., Chen, Y., & Zhou, P. (2021). Cross-modality person re-identification with camera style adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14772–14781).

Wu, S., Shan, S., Xiao, G., Lew, M. S., & Gao, X. (2024). Modality blur and batch alignment learning for twin noisy labels-based visible–infrared person re-identification. *Engineering Applications of Artificial Intelligence*, [ISSN: 0952-1976] *133*, Article 107990. http://dx.doi.org/10.1016/j.engappai.2024.107990, URL https://www.sciencedirect.com/science/article/pii/S0952197624001489.

Wu, S., Yuan, X., Xiao, G., Lew, M. S., & Gao, X. (2024). Deep cross-modal hashing with multi-task latent space learning. *Engineering Applications of Artificial Intelligence*, *136*, Article 108944.

Wu, A., Zheng, W. S., Yu, H. X., Gong, S., & Lai, J. (2017). RGB-infrared cross-modality person re-identification. In *2017 IEEE international conference on computer vision*.

Xu, X., Liu, S., Zhang, N., Xiao, G., & Wu, S. (2022). Channel exchange and adversarial learning guided cross-modal person re-identification. *Knowledge-Based Systems*, *257*, Article 109883.

Yang, M., Huang, Z., Hu, P., Li, T., Lv, J., & Peng, X. (2022). Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14308–14317).

Ye, M., Ruan, W., Du, B., & Shou, M. Z. (2021). Channel augmented joint learning for visible-infrared recognition. In *IEEE/CVF international conference on computer vision* (pp. 13567–13576).

Ye, M., Shen, J., Crandall, D. J., Shao, L., & Luo, J. (2020). Dynamic dual-attentive aggregation learning for visible-infrared person re-identification.

Ye, M., Shen, J., Lin, G., Xiang, T., & Hoi, S. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*(99), 1.

Zhang, Q., Lai, C., Liu, J., Huang, N., & Han, J. (2022). Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, USA, June 18-24, 2022* (pp. 7339–7348). IEEE, http://dx.doi.org/10.1109/CVPR52688.2022.00720.

Zhang, J., Wang, Y., Liang, X., Chen, Y., & Zhou, P. (2021). Person re-identification with deep similarity-guided graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 10581–10590).

Zhao, Z., Liu, B., Chu, Q., Lu, Y., & Yu, N. (2021). Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. *vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (4), (pp. 3520–3528). URL https://ojs.aaai.org/index.php/AAAI/article/view/16466.