Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20217-8

Unified pre-training with pseudo infrared images for visible-infrared person re-identification

ZhiGang Liu[1,2] · Yan Hu[1]

Received: 9 June 2024 / Revised: 29 July 2024 / Accepted: 28 August 2024

Abstract
In the pre-training task of visible-infrared person re-identification(VI-ReID), two main challenges arise: i) Domain disparities. A significant domain gap exists between the ImageNet utilized in public pre-trained models and the specific person data in the VI-ReID task. ii) Insufficient sample. Due to the challenge of gathering cross-modal paired samples, there is currently a scarcity of large-scale datasets suitable for pretraininge. To address the aforementioned issues, we propose a new unified pre-training framework (UPPI). Firstly, we established a large-scale visible-pseudo infrared paired sample repository (UnitCP) based on the existing visible person dataset, encompassing nearly 170,000 sample pairs. Benefiting from this repository, not only are training samples significantly expanded, but pre-training on this foundation also effectively bridges the domain disparities. Simultaneously, to fully harness the potential of the repository, we devised an innovative feature fusion mechanism(CF^2) during pre-training. It leverages redundant features present in the paired images to steer the model towards cross-modal feature fusion. In addition, during fine-tuning, to adapt the model to datasets lacking paired images, we introduced a center contrast loss(C^2). This loss guides the model to prioritize cross-modal features with consistent identities. Extensive experimental results on two standard benchmarks (SYSU-MM01 and RegDB) demonstrate that the proposed UPPI performs favorably against state-of-the-art methods.

Keywords Cross-modality retrieval · Visible-infrared person re-identification · Pre-training tasks

Yan Hu contributed equally to this work

✉ Yan Hu
17628390919@163.com

ZhiGang Liu
ZhigangLiu@nepu.edu.cn

1 College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, Heilongjiang, China

2 Heilongjiang Petroleum Big data and Intelligent Analysis Key Laboratory, Northeast Petroleum University, Daqing 163318, Heilongjiang, China

1 Introduction

The significant progress [1, 2] achieved in object detection and image localization has established a substantial empirical foundation and theoretical framework for person re-identification (ReID) [3]. Currently, there is growing interest in visible-infrared person re-identification (VI-ReID) [4]. VI-ReID strives to effectively retrieve person images exhibiting maximal similarity amidst diverse illumination conditions. In light of diverse imaging principles, images captured under varying lighting conditions manifest significant discrepancies in their visual attributes. Moreover, within identical illumination settings, images portraying individuals of the same identity showcase considerable intra-class diversity owing to factors like posture and perspective.Hence,the pursuit of consistent inter modal embedding under the same identity, exemplified in Fig. 1, emerges as a pivotal challenge in VI-ReID.

This challenge has garnered considerable attention, leading to the emergence of two prominent approaches.One approach involves utilizing feature-based methods [5–12], which employ single-path or multi-path deep neural networks (DNNs) to acquire cross-modal consistent embeddings. These methods leverage sample labels to diminish the modal disparity between cross-modal features.The alternative approach centers on image-based methods [13–16], with the aim of bridging the substantial appearance disparities between different modalities. For instance, GANs [17] facilitate image conversion from one mode to another, mitigating cross-modality induced appearance differences. Despite their high effectiveness and frequently superior performance, they come with certain limitations: (1) due to the absence of involvement of the infrared mode, there are domain disparities in the pre-trained models mentioned on ImageNet; (2) the scarcity of labeled samples across modalities restricts the matching performance of these networks.

To address the aforementioned issues, we have made extensive efforts in terms of samples, training paradigm, and loss functions outside the network. Consequently, we propose UPPI,

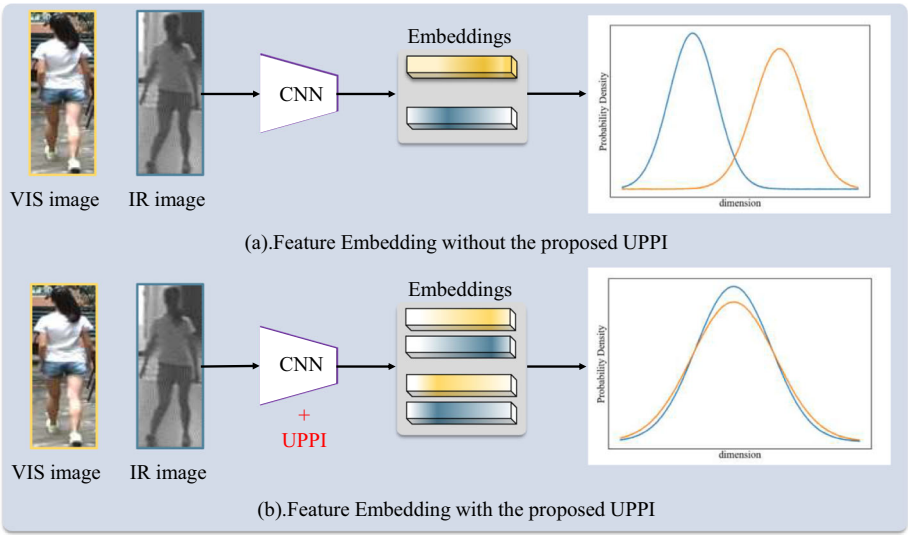

Fig. 1 The proposed UPPI framework generates consistent embeddings across diverse modal environments, enabling the network to learn identity-representative features. This approach effectively reduces the modal gap between VIS and IR images of the same identity

as depicted in Fig. 2. Firstly, alongside visible and infrared images, we have assembled a vast sample repository named UnitCP, which encompasses paired images possessing distinct appearance styles: visible and pseudo-infrared. The pseudo-infrared images exhibit comparable appearance styles to infrared images while maintaining consistent semantic information with visible images. This effectively bridges the substantial disparities between ImageNet and specific infrared datasets and also offers a solution for the scarcity of labeled infrared samples. Secondly, within the pre-training phase, we propose the integration of a cross-modal feature fusion module (CF^2), strategically designed to compel the network to discern shared features across diverse modalities by judiciously merging redundant features from paired samples in an explicit manner.

During the fine-tuning process, to facilitate the model's adaptation to the temporal absence of paired images, we introduce a novel central contrast loss (C^2). This loss function incentivizes the prioritization of identity-consistent cross-modal features in scenarios where paired samples are unavailable.

Our main contributions are summarized as follows. (1) To address the domain discrepancies between ImageNet and specific infrared datasets, as well as the absence of paired samples, we aim to investigate the feasibility of pretraining and propose a Unified Framework for Pre-training with Pseudo-Infrared Images (UPPI). (2) We establish a comprehensive cross-modal sample repository(UnitCP), serving as the foundation for our pre-training endeavors. (3) We propose a cross-modal feature fusion mechanism(CF^2), strategically devised to maximize the utilization of paired images. Moreover, to enhance the model's adaptability during fine-tuning in scenarios where paired images are lacking, we introduce a novel center contrast loss (C^2). (4) Extensive experimental results on two benchmark datasets, SYSU-MM01 and

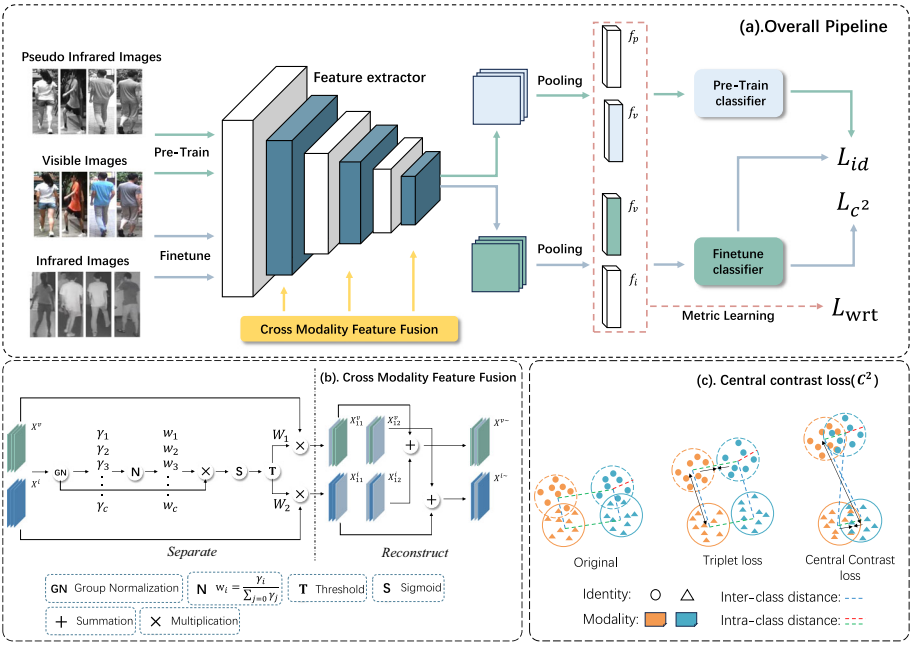

Fig. 2 Illustrates the UPPI framework pipeline, encompassing cross-modal feature fusion module (CF^2) and Central contrast loss (C^2) components. The distinctively colored arrows depict various training stages, including pre-training and fine-tuning

RegDB, validate the effectiveness and exceptional performance demonstrated by our proposed method.

2 Related work

2.1 VI-ReID tasks

In tackling this formidable cross-modal challenge, two distinct approaches arise. One class of methodologies endeavors to attain consistent embedding across diverse modalities [4–12]. For example, Wu et al. [4] constructed the largest visible near-infrared dataset SYSU-MM01 and proposed a zero-padding framework. TONE + HCML [5] have advanced the two-stream cross-modal Re-ID framework by concurrently optimizing shared and specific metrics. Lu et al. [6] endeavored to introduce modality-specific features. Mean-while, Wu et al. [7] proposed a modal-gated extractor that integrated a similarity preservation loss. Ye et al. [9] proposed an attention-based framework aimed at leveraging both local-level and image-level contextual cues. FMCNet [12] devised a feature compensation structure to extract additional discriminative features from shared ones. While these methods have achieved notable advancements in performance, the inherent domain distinctions between ImageNet and the specific datasets per-taining to VI-ReID impose constraints on further performance enhancements. Unlike these methods, which use pre-trained models on ImageNet for direct recognition, we first use pre-training tasks to reduce domain disparities for specific datasets in VI-ReID tasks and ImageNet, and then recognize.

An alternative set of methodologies endeavors to mitigate the visual disparities among cross-modal images through sample-based interventions [13–16], or through the development of effective sample augmentation strategies [18–20]. For instance, Wang et al. [15] decomposed the features extracted and decoded the shared modal features to produce high-quality cross-modal paired images. X-modality [16] devised a lightweight network that learns intermediate representations of visible and infrared images. Ye et al. [18] employed generated grayscale images for training more robust networks. Ye et al. [19] proposed a data augmentation strategy that randomly selected color channels to generate single-channel samples.Our pseudo-infrared image is similar to the latter category of methods; however, in order to avoid introducing additional noise, we employ a pseudo-infrared approach similar to that described in [19], while not being limited by it. We extensively leverage this pseudo-infrared technique on the existing visible person datasets, yielding a substantial number of cross-modal sample pairs.

2.2 Pre-training tasks

The practice of pre-training networks [21–23] on extensive corpora has exhibited significant advantages in natural language processing tasks. In computer vision,networks supervised pre-trained on ImageNet [24], such as AlexNet [25], ResNet [26], ViT [27], Swin Transformer [28], among others, have showcased substantial performance gains when applied to other tasks.Models pre-trained on ImageNet significantly boost domain disparty retrieval tasks within the same spectrum. However, this performance gain diminishes in cross-spectral settings.From the sample perspective, the lack of infrared samples appears to result in the pre-trained models lacking relevant prior knowledge. This absence is likely a key factor limiting performance in cross-spectral scenarios.Therefore, we propose a pre-training task on

a repository of visible-pseudo-infrared sample pairs, aiming to compensate for the missing infrared prior knowledge by introducing pseudo-infrared samples. Considering the similarities in the format of visible infrared samples, the unified pre-training framework resembles a single-stream network [29–33]. However, to further enhance the network's generalization performance, we have incorporated elements from the pre-training paradigm discussed in [34, 35].

3 Methodology

Let x_k represent a sample of mode k,where $k \in v, i$ (v denotes the visible mode, i denotes the infrared mode). The dataset consists of visible and infrared samples $V = \left\{x_v^j, y_v^j\right\}_{j=0}^{N_v}$,$I = \left\{x_i^j, y_i^j\right\}_{j=0}^{N_j}$, respectively, where N_v and N_i are the number of samples for each mode in the dataset. Here, y_k^j represents the identity label of the j th sample in mode k. VI-ReID aims to match samples with identical identities across modes. The proposed unified framework structure is illustrated in Fig. 2, with further details discussed in subsequent sections.

3.1 UnitCP: a repository of visible pseudo-infrared samples

In transfer learning, domain disparty refers to the difference between the data distribution learned by the model and the data distribution encountered during actual application or testing. Existing works [36–38] have explored the mechanisms and adjustment methods for domain disparty. In VI-ReID, due to the lack of cross-modality source datasets, many efforts [6–8] have focused on leveraging the large-scale single-spectrum dataset ImageNet to acquire prior knowledge. However, we argue that domain disparty remains a significant issue in these approaches. Standard domain disparty problems focus on inter-domain differences caused by factors such as different scenes and camera viewpoints within a single spectrum. In contrast, our identified domain disparty issue focuses on differences between datasets under different spectral settings. Specifically, since the target dataset comprises visible-infrared cross-spectrum samples, the source dataset should maintain a consistent setting. Therefore, we explored a cross-spectrum setting consistent with the target dataset in the source dataset.

Several GAN-based works [13–15] have successfully transformed visible images into infrared images, while some channel-related works [18, 19] have simulated infrared styles in visible images. As depicted in Fig. 3, the images generated by the former significantly differ in quality from those generated by the latter, even though the former successfully produced infrared-style images. Hence, we constructed UnitCP: a visible-to-pseudo-infrared sample repository based on the latter approach.

Initially, we gathered training and test samples from various datasets including Market1501 [39], DukeMTMC-reID [40], MSMT17 [41], and others. These samples were then organized based on labels to obtain approximately 180,000 visible person samples representing 7,250 unique identities. To ensure optimal generalization performance of the network, all these 180,000 samples were utilized for training while a distinct dataset called SYSU-MM01 was used as the test set. After thoroughly examining the merits and demerits of pseudo-infrared technologies, we synergistically integrated these techniques through a randomized selection process to establish a diversified sample-based pseudo-infrared scheme, as outlined

Fig. 3 Illustrates a comparison between pseudo-infrared images in UnitCP and infrared images generated by GANs.The leftmost column depicts the original image, while the upper and lower right columns showcase the images generated by the GAN network and the pseudo-infrared images, respectively. Our results demonstrate that the pseudo-infrared images exhibit superior visual credibility, semantic consistency, and detail compared to those generated by GANs

below:

$$\begin{cases} x_i^v = \left(x_i^R, x_i^R, x_i^R\right), n = 0 \\ x_i^v = \left(x_i^G, x_i^G, x_i^G\right), n = 1 \\ x_i^v = \left(x_i^B, x_i^B, x_i^B\right), n = 2 \\ x_i^v = \left(x_i^\delta, x_i^\delta, x_i^\delta\right), 2 < n \leq 5 \\ x_i^\delta = \alpha \cdot x_i^R + \beta \cdot x_i^B + \gamma \cdot x_i^G \end{cases} \tag{1}$$

Where v and r represent the visible mode and infrared mode respectively, R, G, B denote the three channels of the image. The variable n is a random number employed to enhance sample diversity, while α, β, γ are correlation coefficients utilized in gray technology.

Subsequently, we employed pseudo-infrared technology to convert 180,000 samples into an equivalent number of visible pseudo-infrared sample pairs. The abundance of visible pseudo-infrared sample pairs in UnitCP significantly mitigates the domain disparities between ImageNet and VI-ReID datasets, thereby maximizing the model's potential. To validate this, we visualize the performance curve of AGW across various scenarios. As illustrated in Fig. 4, the network pre-trained on UnitCP not only demonstrates faster convergence but also significantly enhances performance. It is noteworthy that the zero-shot mAP of AGW baseline on SYSU-MM01 surpasses the best performance without pre-training by more than 10 percentage points and exceeds that of AGW pre-trained on ImageNet by over 20 percentage points. This indicates that through exposure to simulated visible pseudo-infrared environments, the network effectively adapts to cross-modal settings.

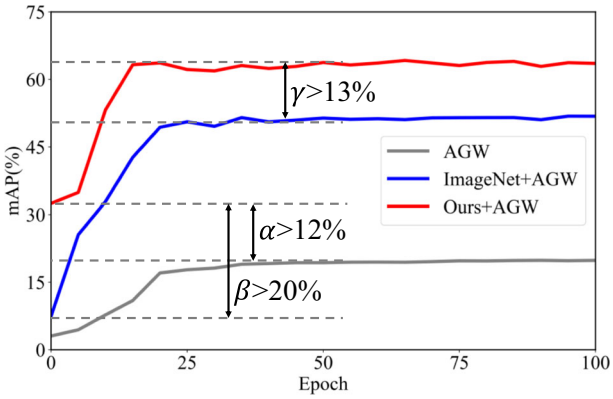

Fig. 4 Illustrates the performance curves of the network under different pre-training scenarios. α demonstrates the performance difference between zero-shot learning with UnitCP pre-training and the saturated training performance without pre-training. β represents the performance gap between zero-shot learning pre-trained on UnitCP and that pre-trained on ImageNet. Finally, γ indicate the difference in fine-tuning performance post pre-training on UnitCP versus pre-training on ImageNet

Notice that our proposed UnitCP serves as not only a static sample repository but also a dynamic cross-modal sample augmentation strategy. In future research, we intend to apply this sample augmentation strategy to larger datasets, such as ImageNet, which effectively addresses the domain disparities issue and significantly mitigates the scarcity of cross-modal samples.

3.2 CF^2: Cross-modal feature fusion module

After introducing a substantial collection of visible pseudo-infrared sample pairs, we thoroughly harness the potential of cross-modal pairwise semantics. Among the high-dimensional features extracted by the network, a significant portion of these features exhibit negligible impact on identity discrimination or even have adverse effects, as demonstrated in previous studies [42]. Motivated by this, we propose a module that exploits redundant features to alleviate disparities in cross-modal features. As il-lustrated in Fig. 2(b), we initially segregate the feature map into dense and sparse information components. Subsequently, we perform cross-reconstruction to establish fusion features for enhancing inter-feature imformation flow.

Specifically, the correlation coefficients in the group normalization (GN) layer are initially utilized to evaluate the information density of individual channels. Given a feature map $x \in \mathbb{R}^{N\times C\times H\times W}$ at an intermediate layer, where N represents batch size, C denotes channel count, and $H \times W$ denotes spatial dimensions of the feature, we proceed by normalizing input feature X using the subsequent formula:

$$X_{out} = GN\,(X) = \gamma \frac{X-\mu}{\sqrt{\sigma^2+\theta}} + \beta \tag{2}$$

where μ and σ are the mean and standard deviation in X, θ is a small positive number that guarantees division, and γ and β are trainable affine transforms.

The trainable parameter γ in the GN layer is utilized to quantify the spatial pixel variance of the channel during the standardization process, where a larger value of γ indicates a richer representation of spatial information. The weight W_γ, which determines the significance of

each channel, can be derived from (3).

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}, i, j = 1, 2, ..., C \tag{3}$$

The weights of the reweighted feature map W_γ are subsequently normalized to the range (0,1) through the application of the sigmoid function. We assign 0 to weights below the threshold and 1 to weights above the threshold, resulting in two weight matrices W_1 and W_2 that have the same scale as the feature map. In summary, the process for calculating weights to distinguish sparse feature maps from dense feature maps is as follows:

$$W = Gate\left(Sigmoid\left(W_\gamma\left(GN\left(X\right)\right)\right)\right) \tag{4}$$

We respectively replicate and weight the features of different modes to obtain a feature map that encompasses dense information, as well as a feature map that captures sparse information. Finally, the process for calculate ing the reconstructed features of both modes is outlined as follows:

$$\begin{cases} X_{11}^v = W_1 \otimes X^v, X_{12}^v = W_2 \otimes X^v \\ X_{11}^i = W_1 \otimes X^i, X_{12}^i = W_2 \otimes X^i \\ X^v = X_{11}^v \oplus X_{12}^i \\ X^i = X_{11}^i \oplus X_{12}^v \end{cases} \tag{5}$$

where $\oplus$ denotes element-wise addition and $\otimes$ denotes element-wise multiplication.

The effect comparison before and after CF^2 is illustrated in Fig. 5. The red-boxed area in (a) is severely affected by the background, leading to feature loss in the corresponding area in (b). In contrast, the corresponding area in (c) shows effective feature compensation. This improvement can be attributed to the efficient inter-modal information reciprocity channel constructed by CF^2, which significantly alleviates the issue of feature loss in infrared samples.

3.3 C^2: Central contrast loss

In the fine-tuning phase, the challenging conditions characterized by a scarcity of paired images and substantial intra-class variations compel the model to redirect its focus from images towards identity. Hence, it is crucial to establish intermodal identity consistency constraints for alignment purposes. Previous works [11, 43, 44] have introduced cross-modal distance constraints to effectively reduce modal distances between cross-modal features and achieved remarkable outcomes. However, these additional hard distance constraints may potentially result in the loss of essential discriminative features.

Instead, we propose a similarity-based center contrast loss. First, we use the Euclidean center of feature semantics to represent the same identity of each modality.

$$C_i^v = \frac{1}{N_v}\sum_{j=1}^{N_v} f_{j=1}^v, C_i^r = \frac{1}{N_r}\sum_{j=1}^{N_r} f_{j=1}^r \tag{6}$$

where N_v, N_r represent the number of visible samples and the number of infrared samples under the same batch of the same identity. Compared with the hard distance constraint based on samples, we use a soft identity constraint in order to maintain higher quality image discriminant features.

(a). Cross-modal original sample (b). Primary feature (c). Fusion feature

Fig. 5 Illustrates the comparative analysis of feature representations before and after the incorporation of the CF^2 framework. The integration of CF^2 has facilitated enhanced information exchange across multiple modalities, thereby significantly augmenting the fidelity and discriminative power of the resultant feature set. This improvement in feature quality is attributed to the synergistic interplay between the diverse data modalities, which collectively contribute to a more robust feature learning process

Assuming a batch size N_b and a sample size m for each modality and identity, the number of identities in a batch can be calculated as $N = N_b/2m$. The contrast loss plays a crucial role in maximizing the cross-modal center cosine similarity within the same identity while minimizing it between different identities. The corresponding formula is presented below.

$$
\begin{aligned}
L_{c^2}^{v} &= \frac{1}{N}\sum_{n=1}^{N} -\log\left(\frac{\exp(sim(C^{v},C^{r+}))}{\sum_{i=1}^{N_{c}^{r-}}\exp(sim(C^{v},C^{r-}))}\right), \\
L_{c^2}^{r} &= \frac{1}{N}\sum_{n=1}^{N} -\log\left(\frac{\exp(C^{r},C^{v+})}{\sum_{i=1}^{N_{c}^{v-}}\exp(sim(C^{r},C^{v-}))}\right), \\
L_{c^2} &= \left(L_{c^2}^{v} + L_{c^2}^{r}\right)/2
\end{aligned}
\tag{7}
$$

Where $+$ and $-$ respectively denote the feature centers sharing the same identity as the anchored identity center, and those with different identities; $sim\,(a, b) = a^{\top}b/\parallel a \parallel \cdot \parallel b \parallel$ represents the cosine similarity between a and b.

In addition, in order to provide the network with basic semantic signals, the two general loss functions in (8) are also adopted.

$$\begin{aligned} L_{id} &= -\frac{1}{N}\sum_{i=1}^{N}\log\left(P\left(\frac{y_i}{x_i}\right)\right) \\ L_{wrt} &= \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(\sum_{ij} w_{ij}^{p} d_{ij}^{p} - \sum_{ik} w_{ik}^{n} d_{ik}^{n}\right)\right) \\ w_{ij}^{p} &= \frac{\exp\left(d_{ij}^{p}\right)}{\sum_{d_{ij}^{p}\in p_i}\exp\left(d_{ij}^{p}\right)},\ w_{ik}^{n} = \frac{\exp(-d_{ik}^{n})}{\sum_{d_{ik}^{n}\in n_i}\exp(d_{ik}^{n})} \end{aligned} \tag{8}$$

where L_{id} denotes cross-entropy loss, L_{wrt} denotes weighted regularization triple loss, and (i, j, k) denotes a triple in each training batch for each anchored sample x_i. For x_i, P is the corresponding positive sample set and N is the corresponding negative sample set. d^p, d^n denotes pairwise distances of positive and negative sample pairs, respectively. d_{ij} is the Euclidean distance between two sample features. Using softmax weighting strategy to obtain two weights, w^p, w^n can force the network to pay more attention to distance optimization of difficult samples.

The overall objective function of UPPI can be expressed as (9), where ω represents the hyperparameter utilized for balancing L_{c^2} with the two base losses.

$$L = L_{id} + L_{wrt} + \omega L_{c^2} \tag{9}$$

4 Experiment

4.1 Datasets and implementation details

4.1.1 Datasets

We conducted simulation experiments on two general VI Re-ID datasets to evaluate our proposed method.

SYSU-MM01 consists of 491 identities captured by 6 cameras (4 RGB cameras and 2 IR cameras) at different times and under varying environmental conditions, resulting in a total of nearly 60,000 images. The training set comprises 395 identities with a total of 19,659 RGB images and 12,792 IR images, while the test set includes 96 identities. Following its standard evaluation protocol, the dataset offers global retrieval mode and indoor retrieval mode. To ensure fair comparison with state-of-the-art methods, we extracted respectively 301 and 3010 images to construct the gallery set for evaluation.

RegDB [45], created by a visible camera and a thermal camera, contains 412 identities, with 10 images for each identity under each mode, which contains several different perspectives, a total of 10,932 images. In the training phase, we randomly selected a subset consisting of 206 identity images as the training set while using the remaining 206 identity images as the test set.

4.1.2 Implementation details

We utilized the AGW [46] baseline as our backbone network and collected two versions of the UPPI sample repository, UniCP12 with 4100 identities and nearly 120,000 visible images, UniCP18 with 7,250 identities and almost 180,000 images. To account for the varying number

of classes, we set independent classifiers for both pre-training and fine-tuning processes. Our proposed CF^2 was inserted after the second and third convolution blocks; however, its parameters did not participate in updates due to a lack of cross-modal paired images during fine-tuning. After conducting numerous experiments, we determined that setting ω at 0.6 effectively balanced loss term C^2. During pre-training and fine-tuning stages, each batch randomly selected sixteen identities from which four visible images and four infrared images were chosen to form small batches. Each infrared sample was stacked into a three-channel image before being fed into the network. To expedite the training process, mixed precision training was employed. The Adam optimizer with a warm-up strategy was utilized, setting the learning rate to 3.5×10^{-4} for SYSU-MM01 and 8.25×10^{-4} for RegDB during the initial 10 epochs. Subsequently, at epoch 20 and epoch 40, the learning rate was reduced by factors of 0.1 and 0.01 respectively. Standard data augmentation techniques including random cropping, random horizontal flipping, color jittering, and random erasure were applied. A total of 120 epochs were conducted for network training; comprising of an initial pre-training phase spanning over 80 epochs followed by fine-tuning for an additional 40 epochs. To ensure fair comparison among different networks, any reordering algorithm was employed during evaluation.

4.2 Comparison with state-of-the-art methods

The comparison results with state-of-the-art methods on SYSU-MM01 and RegDB are shown in Tables 1 and 2. It can be seen that the performance indicators of the proposed UPPI

Table 1 Comparison with the state-of-the art methods on the SYSU-MM01 dataset

Method	Venue	All-search			Indoor-search		
		R1	R10	mAP	R1	R10	mAP
AGW [46]	TPAMI 2022	47.50	84.39	47.65	54.17	91.14	62.97
X-Modal [16]	AAAI 2020	49.92	89.79	50.73	–	–	–
DDAG [9]	ECCV 2020	54.75	90.39	53.02	61.02	94.06	67.98
NFS [47]	CVPR 2021	56.91	91.34	55.45	62.79	96.53	69.79
DFLN-ViT [48]	TMM 2022	59.84	92.49	57.70	62.13	94.83	69.03
MID [49]	AAAI 2022	60.27	92.90	59.40	64.86	96.12	70.12
SMCL [50]	ICCV 2021	67.39	92.87	61.78	68.84	96.55	75.56
MCLNet [51]	ICCV 2021	65.40	93.33	61.98	72.56	96.98	76.58
MPMN [52]	TMM 2021	48.98	90.33	62.41	64.89	96.85	76.47
FMCNet [12]	CVPR 2022	66.34	–	62.51	68.15	–	74.09
cm-SSFT [7]	CVPR 2020	61.60	89.20	63.20	70.50	94.90	72.60
PMT [53]	AAAI 2023	67.53	95.36	64.98	71.66	96.73	76.52
CMIT [54]	TMM 2022	70.94	94.93	65.51	73.28	95.20	77.18
CAJ [19]	ICCV 2021	69.88	95.71	66.89	76.26	97.88	80.37
MPANet [10]	CVPR 2021	70.58	96.21	68.24	76.74	98.21	80.95
MAUM [55]	CVPR 2022	71.68	–	68.79	76.97	–	81.94
CAL [56]	ICCV 2023	74.66	96.47	71.73	79.69	98.93	83.68
DEN [57]	WACV 2024	76.36	–	71.30	83.56	–	84.65
UPPI(Ours)	–	**76.58**	**97.25**	**72.76**	**84.23**	**97.96**	**85.66**

Table 2 Comparison with the state-of-the art methods on the RegDB dataset

Method	Venue	Visible to infrared			Infrared to visible		
		R1	R10	mAP	R1	R10	mAP
AGW [46]	TPAMI 2022	70.05	86.21	66.37	75.93	90.93	69.49
X-Modal [16]	AAAI 2020	62.21	83.13	60.18	–	–	–
DDAG [9]	ECCV 2020	69.34	86.19	63.46	68.06	85.15	61.8
NFS [47]	CVPR 2021	80.54	91.96	72.1	77.95	90.45	69.79
DFLN-ViT [48]	TMM 2022	92.1	97.97	82.11	91.21	98.2	81.62
MID [49]	AAAI 2022	87.45	95.73	84.29	84.29	93.44	81.41
SMCL [50]	ICCV 2021	83.93	–	79.83	83.05	–	78.57
MCLNet [51]	ICCV 2021	80.31	92.7	73.07	75.93	90.93	69.49
MPMN [52]	TMM 2021	86.56	96.89	82.91	84.62	95.51	79.49
FMCNet [12]	CVPR 2022	89.12	–	84.43	88.38	–	83.86
cm-SSFT [7]	CVPR 2020	72.3	–	72.9	71	–	71.7
PMT [53]	AAAI 2023	84.83	–	76.55	84.16	–	75.13
CMIT [54]	TMM 2022	88.78	94.76	88.49	84.55	93.72	83.64
CAJ [19]	ICCV 2021	85.03	95.49	79.14	84.75	95.33	77.82
MPANet [10]	CVPR 2021	83.7	–	80.9	82.8	–	80.7
MAUM [55]	CVPR 2022	87.87	–	85.9	86.95	–	84.34
CAL [56]	ICCV 2023	94.51	99.7	88.67	93.64	99.46	87.61
DEN [57]	WACV 2024	95.34	–	90.21	94.98	–	90.24
UPPI(Ours)	–	**97.28**	**99.16**	**91.11**	**95.26**	**98.39**	**90.48**

on the two datasets are mostly better than existing state-of-the-art methods. Most of the comparison methods employed in this study utilize pre-trained models on ImageNet and lack paired cross-modal images, thus limiting their untapped potential. Therefore, the UPPI proposed in this study is specifically designed to tackle these two significant challenges and offer our innovative solutions. (1) The construction of a large-scale visible pseudo-infrared paired sample repository(UnitCP), based on pseudo-infrared images, not only addresses the scarcity of cross-modal samples but also enables the pre-trained model to compensate for the lack of infrared experience while retaining prior knowledge acquired from ImageNet; (2) CF^2 facilitates the integration of cross-modal information by leveraging redundant features, thereby mitigating substantial visual disparities across different modalities; during the fine-tuning phase, C^2 significantly aids in redirecting the model's attention from images to identity.

4.3 Ablation study and analysis

4.3.1 Effectiveness of proposed components

The experimental results on the SYSU-MM01 dataset are presented in Table 3. It is evident that our pre-training weights significantly enhance the performance of the network on the cross-modal dataset, while the other two components also contribute to a portion of this improvement. To further validate our approach, we integrate these proposed components into well-established backbone networks such as Inception-V3 [58], ResNet50, EfficientNet-B3 [59], and Vit-Base [27]. As shown in Table 3, our components yield performance gains for these networks, demonstrating the versatility of our method.

Table 3 Effectiveness of the proposed components over different backbone networks on the SYSU-MM01 dataset under the all-search single-shot mode

Backbone	UnitCP18	CF^2	C^2	SYSU-MM01(All-search)		
				R1	R10	mAP
ResNet-50				48.16	87.47	50.41
	√			67.76	91.10	62.77
	√	√		69.18	94.92	64.31
	√	√	√	**70.75**	**96.43**	**66.89**
Inception-V3				40.22	74.28	44.36
	√			57.09	88.26	55.42
	√	√		60.66	89.90	58.17
	√	√	√	**61.91**	**92.26**	**60.15**
EfficientNet-B3				48.34	77.58	48.92
	√			64.88	92.30	61.16
	√	√		60.96	93.44	63.30
	√	√	√	**65.27**	**95.57**	**65.08**
AGW				47.50	84.39	47.65
	√			70.58	96.21	67.68
	√	√		74.66	96.46	70.71
	√	√	√	**76.58**	**97.25**	**72.76**
Vit-Base				40.23	83.63	39.90
	√			50.82	88.53	49.57
	√	√		53.77	90.94	52.73
	√	√	√	**57.06**	**92.41**	**55.08**

4.3.2 Comparative analysis of pre-training at different scales

we present in Fig. 6 the simulation results of pre-training weights obtained on ImageNet, UnitCP12 and UnitCP18 sample repositories loaded by four different backbone networks, ResNet-50, Inception-V3, EfficientNet-B3 and Vit-Base. It can be seen that the cross-modal matching performance of the model is significantly enhanced following saturated pre-training tasks. Notably, the zero-shot mAP in ResNet50 exhibits an increase of over 20%, while other

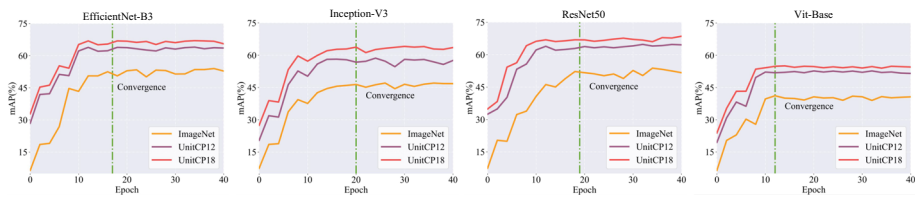

Fig. 6 Illustrates the performance curves for four feature extraction networks pre-trained on different datasets. The networks exhibit significantly improved performance and faster convergence on UnitCP12 compared to ImageNet. A moderate enhancement is also observed from UnitCP12 to UnitCP18. Notably, the substantial gap in zero-shot performance highlights the network's proficiency in cross-modal image environments, attributed to our specialized pre-training approach

Table 4 Comparison results of different feature fusion methods

Method	SYSU-MM01		
	R1	R10	mAP
UPPI w/o CF^2	70.58	96.21	67.68
CF^2 →DenseFuse	68.27	95.19	66.77
CF^2 →SeAFusion	71.02	96.08	68.05
CF^2 →FusionGAN	70.39	96.25	68.11
CF^2 →DIVFusion	73.79	96.37	69.52
UPPI(Ours)	**74.66**	**96.46**	**70.71**

Please note that the algorithms involved in the table do not include image reconstruction

backbone networks also demonstrate remarkable enhancements in their zero-shot matching capabilities. The substantial improvement in final retrieval performance can likely be attributed to the inclusion of a large number of samples. What factors, then, lead to the significant enhancement in zero-shot retrieval performance post-pretraining? Zero-shot performance is a crucial indicator of the extent of prior knowledge acquisition, and the sharp increase in zero-shot performance undoubtedly signifies that the model has acquired substantial prior knowledge through our tailored pretraining tasks. Crucially, prior knowledge from the visible spectrum has been effectively obtained using ImageNet. Therefore, our customized pretraining tasks compensate for the missing infrared prior knowledge, which is a key measure to mitigate domain discrepancy in cross-spectrum datasets.

4.3.3 Comparative analysis of different fusion methods

In Table 4, we substituted CF^2 in the pre-training phase with alternative fusion tech-niques, while keeping all other experimental conditions unchanged. The compared methods include DenseFuse [60], SeAFusion [61], FusionGAN [62], and DIVFusion [63]. In contrast to these approaches, we omitted a portion of the network responsible for generating the fused image in order to extract fusion features. The results demonstrate that our proposed feature fusion method outperforms the aforementioned techniques significantly in cross-modal retrieval tasks. We attribute this superiority to our method's emphasis on specific optimization

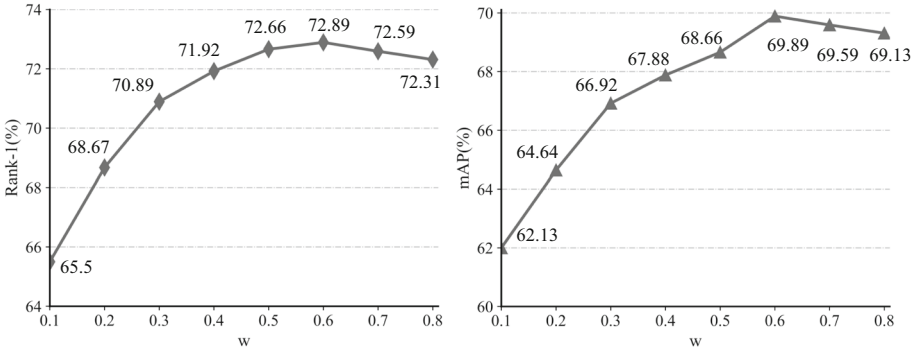

Fig. 7 The performance was systematically assessed across a range of parameter ω settings from 0.1 to 0.6, revealing a consistent enhancement that highlights the efficacy of C^2. Nevertheless, further increasing the parameter values may lead to feature distortion

Table 5 Comparison results of different identity constraints

Method	SYSU-MM01		
	R1	R10	mAP
UPPI w/o C^2	74.66	96.46	70.71
C^2 →CL	74.70	95.93	70.08
C^2 →CCL	73.58	96.27	70.86
C^2 →CPM	74.27	95.19	71.23
C^2 →CC	75.02	96.08	71.55
C^2 →DCL	75.39	96.65	72.11
UPPI(Ours)	**76.58**	**97.25**	**72.76**

processes that aid recognition rather than solely pursuing the visual quality of synthesized images. This observation further validates the suitability of CF^2 within the image environment established by UPPI.

4.3.4 Comparative analysis of different identity constraints

In (9), we set a parameter ω to control the tradeoff between the C^2 loss with identity loss and triplet loss. To explore the impact of the hyperparameter, we give an empirical analysis on the SYSU-MM01 datasets and report the results in Fig. 7. From the results, we can observe that

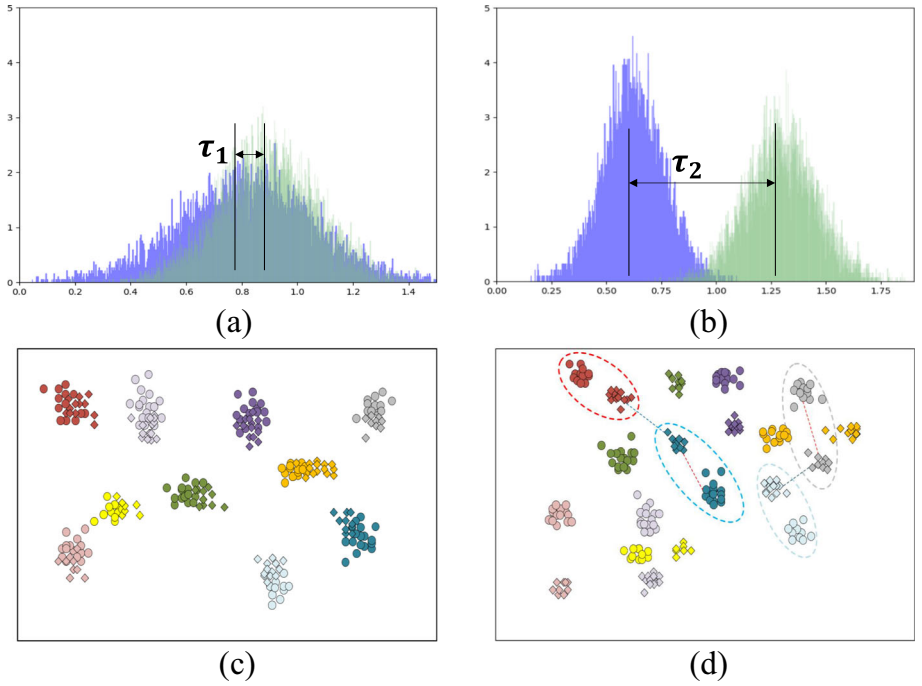

Fig. 8 The purple and green colors in Fig. 6(a) and (b) respectively denote intra-class distance and inter-class distance. In (c) and (d), distinct colors represent features of different identities, while diverse shapes indicate different modes. It is evident that our proposed UPPI method effectively reduces both modal distance and intra-class variation simultaneously

even adding the C^2 loss with a small weight (0.2), the final accuracy and mAP could improve significantly. The best performance is achieved when the parameter ω goes to 0.6. Although a bigger ω may obtain a more compact feature space for every single identity, the high weight of C^2 loss makes limited optimization for the feature from different identities. To compare C^2 with distance-based hard constraints, we substituted it with alternative hard distance constraints and conducted comparative experiments. The compared hard distance constraints include Center Loss (CL) [64], Center Cluster Loss (CCL) [10], Dual-Enhancement Center Loss (DCL) [11], Center-Guided Pair Mining Loss (CPL) [43], and Cross-center Loss (CC) [44]. The results presented in Table 5 demonstrate that proposed C^2, aiming to alleviate intra-class variations and promote the network's focus on identity features, yields superior performance improvements and significantly enhances accuracy.

4.4 Qualitative analysis

4.4.1 Feature visualization

We visualized the spatial distribution and distance distribution of different modal features through t-SNE [65]. Figure 8(a),(b) respectively present the intra-class distance and inter-class distance of Baseline and UPPI. It can be seen that the inter-class distance of UPPI features is significantly greater than the intra-class distance. As shown in Fig. 8(c), the inter-modal difference of some identities is very large, even greater than the difference between different identities in the same modal feature, which may lead to incorrect cross-modal retrieval results.

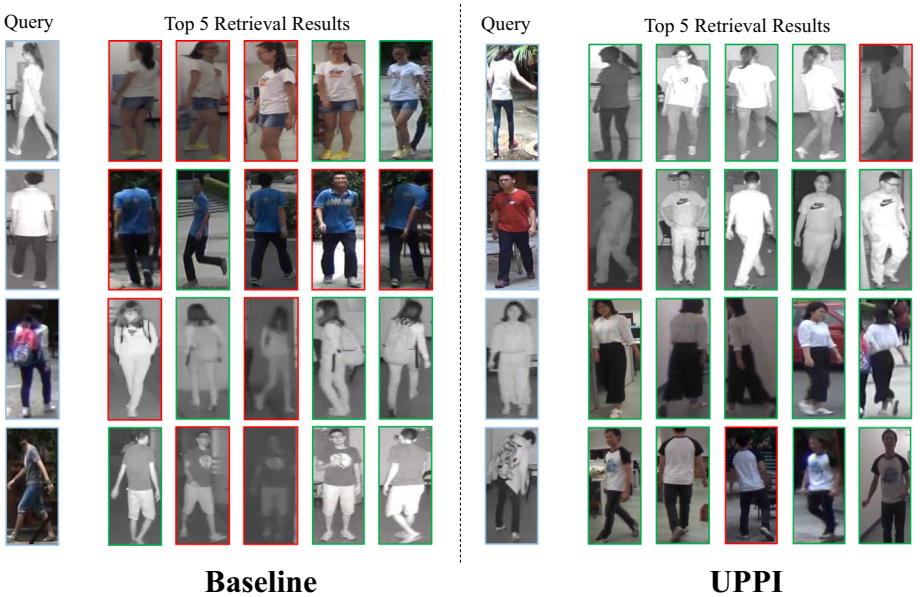

Fig. 9 The top-5 retrieval results of several hard queries obtained by the baseline method (AGW) and the proposed framework on the SYSU-MM01 dataset. The images with green bounding boxes have the same identity labels as the query images (i.e., correct matches), and those with red bounding boxes have different identity labels (i.e., wrong matches)

In contrast, the UPPI proposed by us makes the features more compact and the intra-class difference smaller, as shown in Fig. 8(d).

4.4.2 Analysis of retrieval results

Figure 9 shows the comparison of retrieval results between UPPI and baseline on SYSU-MM01, including single-shot and multi-shot settings. The findings demonstrate a significant enhancement in retrieval accuracy achieved by UPPI. Notably, certain queries depicted in Fig. 9 pose challenges even for human ; however, the network incorporating UPPI still manages to achieve accurate matches.

5 Conclusion

In this paper, we propose UPPI, the first unified pre-training framework for VI Re-ID. It enhances network performance from a training method perspective by incorporating three main components: (1) constructing a large-scale cross-modal sample warehouse (UnitCP) based on pseudo-infrared images and pre-training the network with it for cross-modal learning; (2) utilizing the cross-modal feature fusion module (CF^2) to identify potential redundant features and fuse them into cross-modal features; and (3) implementing the center contrast loss (C^2), which establishes flexible constraints of identity consistency that encourage the network to focus on identity-consistent cross-modal features while reducing differences in such features. A multitude of simulation experiments confirm significant performance gains resulting from UPPI.

Author Contributions Author 1(First Author): Conceptualization, Funding Acquisition, Resources, Supervision . Author 2(Corresponding Author): Methodology, Software, Investigation, Formal Analysis, Writing - Original Draft, Visualization, Investigation.

Funding This work was supported in part by the National Natural Science Foundation of China under Grant 51774090 and Grant 42002138, in part by the Heilongjiang Provincial Natural Science Foundation of China under Grant LH2020F003.

Data Availability and Access The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Competing interest The authors declare that they have no competing interest.

Ethical and Imformed Consent for Data Used All data and content we use is publicly available and does not violate any ethical guidelines.

References

1. Cui Y, Yan L, Cao Z, Liu D (2021) Tf-blender: temporal feature blender for video object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8138–8147
2. Liu D, Cui Y, Yan L, Mousas C, Yang B, Chen Y (2021) Densernet: weakly supervised visual localization using multi-scale feature aggregation. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 6101–6109

3. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv:1610.02984
4. Wu A, Zheng W-S, Yu H-X, Gong S, Lai J (2017) Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 5380–5389
5. Ye M, Wang Z, Lan X, Yuen PC (2018) Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI, vol 1, p 2
6. Ye M, Lan X, Li J, Yuen P (2018) Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
7. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13379–13389
8. Wu A, Zheng W-S, Gong S, Lai J (2020) Rgb-ir person re-identification by cross-modality similarity preservation. Int J Comput Vision 128(6):1765–1785
9. Ye M, Shen J, Crandall JD, Shao L, Luo J (2020) Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, pp 229–247
10. Wu Q, Dai P, Chen J, Lin C-W, Wu Y, Huang F, Zhong B, Ji R (2021) Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4330–4339
11. Gong J, Zhao S, Lam K-M, Gao X, Shen J (2023) Spectrum-irrelevant fine-grained representation for visible-infrared person re-identification. Comput Vis Image Underst 232:103703
12. Zhang Q, Lai C, Liu J, Huang N, Han J (2022) Fmcnet: feature-level modality compensation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7349–7358
13. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 618–626
14. Wang G, Zhang T, Cheng J, Liu S, Yang Y, Hou Z (2019) Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3623–3632
15. Wang G-A, Zhang T, Yang Y, Cheng J, Chang J, Liang X, Hou Z-G (2020) Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12144–12151
16. Li D, Wei X, Hong X, Gong Y (2020) Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 4610–4617
17. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Advan Neural Inform Process Syst 27
18. Ye M, Shen J, Shao L (2020) Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Trans Inf Forensics Secur 16:728–739
19. Ye M, Ruan W, Du B, Shou MZ (2021) Channel augmented joint learning for visible-infrared recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13567–13576
20. Alehdaghi M, Josi A, Cruz RM, Granger E (2022) Visible-infrared person re-identification using privileged intermediate information. In: European conference on computer vision, Springer, pp 720–737
21. Kenton JDM-WC, Toutanova LK (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol 1, p 2
22. Radford A, Narasimhan K, Salimans T, Sutskever I, et al (2018) Improving language understanding by generative pre-training
23. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv:1907.11692
24. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
25. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
26. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929

28. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
29. Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, Houlsby N (2020) Big transfer (bit): general visual representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, pp 491–507
30. Xie Q, Luong M-T, Hovy E, Le QV (2020) Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10687–10698
31. Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J (2019) Vl-bert: pre-training of generic visual-linguistic representations. arXiv:1908.08530
32. Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) Videobert: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7464–7473
33. Tan H, Bansal M (2019) Lxmert: learning cross-modality encoder representations from transformers. arXiv:1908.07490
34. Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2023) Supplementary material uniter: universal image-text representation learning. ReCALL 1(5):10
35. Li G, Duan N, Fang Y, Gong M, Jiang D (2020) Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11336–11344
36. Kynkäänniemi T, Karras T, Aittala M, Aila T, Lehtinen J (2022) The role of imagenet classes in fr\'echet inception distance. arXiv:2203.06026
37. Han C, Wang Q, Cui Y, Wang W, Huang L, Qi S, Liu D (2024) Facing the elephant in the room: visual prompt tuning or full finetuning? arXiv:2401.12902
38. Chong MJ, Forsyth D (2020) Effectively unbiased fid and inception score and where to find them. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6070–6079
39. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
40. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, Springer, pp 17–35
41. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
42. Li J, Wen Y, He L (2023) Scconv: spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6153–6162
43. Zhang Y, Wang H (2023) Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2153–2162
44. Tan L, Zhang Y, Shen S, Wang Y, Dai P, Lin X, Wu Y, Ji R (2023) Exploring invariant representation for visible-infrared person re-identification. arXiv:2302.00884
45. Nguyen DT, Hong HG, Kim KW, Park KR (2017) Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3):605
46. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC (2021) Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell 44(6):2872–2893
47. Chen Y, Wan L, Li Z, Jing Q, Sun Z (2021) Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 587–597
48. Zhao J, Wang H, Zhou Y, Yao R, Chen S, El Saddik A (2022) Spatial-channel enhanced transformer for visible-infrared person re-identification. IEEE Trans Multimedia
49. Huang Z, Liu J, Li L, Zheng K, Zha Z-J (2022) Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 1034–1042
50. Wei Z, Yang X, Wang N, Gao X (2021) Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 225–234
51. Hao X, Zhao S, Ye M, Shen J (2021) Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 16403–16412
52. Wang P, Zhao Z, Su F, Zhao Y, Wang H, Yang L, Li Y (2020) Deep multi-patch matching network for visible thermal person re-identification. IEEE Trans Multimedia 23:1474–1488

53. Lu H, Zou X, Zhang P (2023) Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 1835–1843
54. Feng Y, Yu J, Chen F, Ji Y, Wu F, Liu S, Jing X-Y (2022) Visible-infrared person re-identification via cross-modality interaction transformer. IEEE Trans Multimedia
55. Liu J, Sun Y, Zhu F, Pei H, Yang Y, Li W (2022) Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19366–19375
56. Wu J, Liu H, Su Y, Shi W, Tang H (2023) Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11122–11131
57. Kim S, Gwon S, Seo K (2024) Enhancing diverse intra-identity representation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2513–2522
58. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
59. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR, pp 6105–6114
60. Li H, Wu X-J (2018) Densefuse: a fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623
61. Tang L, Yuan J, Ma J (2022) Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. Inform Fusion 82:28–42
62. Ma J, Yu W, Liang P, Li C, Jiang J (2019) Fusiongan: a generative adversarial network for infrared and visible image fusion. Inform fusion 48:11–26
63. Tang L, Xiang X, Zhang H, Gong M, Ma J (2023) Divfusion: darkness-free infrared and visible image fusion. Inform Fusion 91:477–493
64. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Computer vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, pp 499–515
65. Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11)