Dual-Semantic Consistency Learning for Visible-Infrared Person Re-Identification

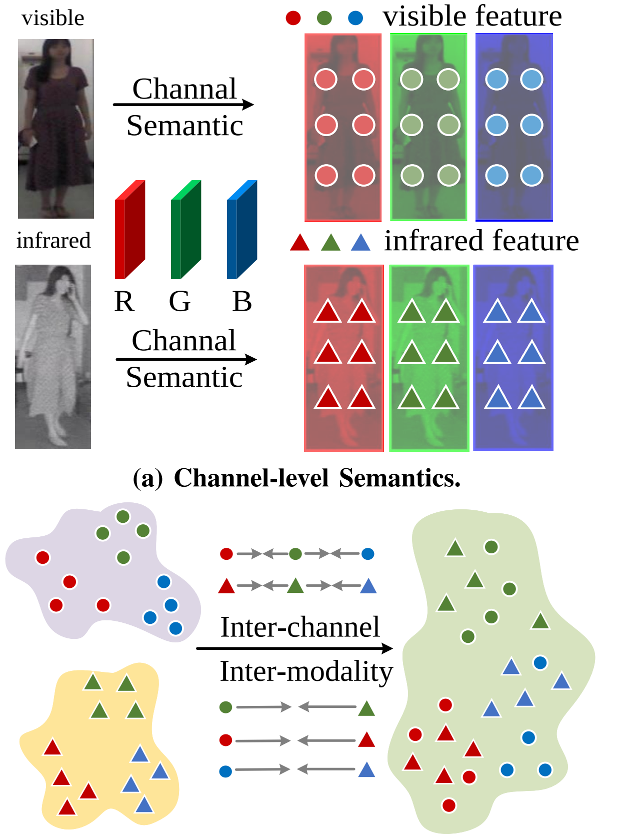
Yiyuan Zhang*, Yuhao Kang*, Sanyuan Zhao†, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*—Visible-Infrared person Re-Identification (VI-ReID) conducts comprehensive identity analysis on non-overlapping visible and infrared camera sets for intelligent surveillance systems, which introduces the modality discrepancy based on instance variations. At present, most of the existing methods focus on reducing modality discrepancy and extracting modality-shared features on the instance level. Differently, we propose a novel framework, named Dual-Semantic Consistency Learning Network (DSCNet), which attributes modality discrepancy to channel-level semantic inconsistency. DSCNet optimizes channel consistency from two aspects, the fine-grained inter-channel semantics, and the comprehensive inter-modality semantics. Furthermore, we propose Joint Semantics Metric Learning to optimize the channel and modality feature distribution together. It jointly exploits the correlation between channel-specific semantics and modality-specific semantics in a fine-grained manner. Experimental results on the SYSU-MM01 and the RegDB benchmarks show that the proposed DSCNet presents remarkable superiority compared with current state-of-the-art methods. On the more challenging SYSU-MM01 dataset, our network can achieve 73.89% Rank-1 accuracy and 69.47% mAP value. Our code is available at https://github.com/bitreidgroup/DSCNet .**

***Index Terms*—visible-infared person re-identification, person re-identification, semantic consistency**

I. Introduction

PERSON re-identification (ReID) is of great importance for public safety. It is a branch of image retrieval technique in computer vision that determines the presence of a specific person in an image or a video sequence. Precise person ReID is challenging because of the variability of objective environment (shooting perspective, occlusion, background noise) and the appearances of pedestrians themselves. Given a query person image captured by a surveillance equipment to retrieve the same one under multiple cameras [18], person ReID technique makes up for the visual limitations of the camera itself, and can also be combined with person detection and tracking tasks [39]. The development of person ReID technique has a prominent impact on the fields of intelligent video surveillance and public security. As the demand for public security evolves, more and more infrared cameras are integrated into the surveillance systems, which aims to enhance the ability of accurate retrieving the specific person day and night. Since identity-relevant infrared information should be combined with visible information during retrieving, it places a new technical requirement, named Visible Infrared person Re-Identification (VI-ReID). In addition to the considerations in visible person ReID [3], [54], such as the difference of camera internal parameters, viewpoint, pedestrian clothing, occlusion, illuminations, and so on, VI-ReID faces more significant intra-class varieties due to the discrepancy between the visible and the infrared modalities [45], which make it more difficult to handle.



(a) Channel-level Semantics.

(b) Inter-channel and inter-modality semantic consistency.

Fig. 1: Motivation. Exiting VI-ReID methods are suffering from modality discrepancy. We think that modality discrepancy derives from the inconsistency of original channel-level semantics (a), and we propose dual-semantic consistency learning (b).

Most of the existing VI-ReID methods can be roughly divided into two categories: the first one solves the cross-modality problem by maximizing the modality-invariance and minimizing the dissimilarity of the features across modalities [53], [8], [11], [10]; the second one generates images

Y. Zhang, Y. Kang, and S. Zhao are with the School of Computer Science, Beijing Institute of Technology, China. (Email: zhaosanyuan@bit.edu.cn)

J. Shen is with the State Key Laboratory of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau, Macau, China. (Email: shenjianbingcg@gmail.com)

* Equal Contribution, † Corresponding author: *Sanyuan Zhao*

of intermediate modalities or images of opposite modalities, so that the cross-modality matching problem can be converted into intra-modality matching task to enhance the retrieval accuracy [16], [41]. However, there are still some shortcomings existing in these two major categories. The first approach extracts modality-invariant features, but modality-invariance is too abstract to be directly represented. It is often difficult to guarantee their qualities. Worse still, the pursuit of certain properties of the features, like modality-invariance or modality-irrelevance, may easily overlook the diversity of the image inherent semantics, and the specificity of the modalities. These lead to indirectly information loss in person image representation. The second GAN-based approach suffers from such expensive computational complexity in training procedure and unavoidable noise introduction, that the accuracy of identity discrimination may be not satisfactory enough.

Infrared photography is the same as visible photography technology. The difference lies in that it uses the reflection, refraction, and transmission generated by the interaction between the infrared light and the objects, as well as the infrared light emitted by the object itself. As visible light is filtered out, infrared images can reveal the scenery and camouflage invisible to the eyes, explore the phenomena that can not be competent by visible light, and identify the indistinguishable objects. Thus, an infrared image cannot be simply considered a normal image consisting of R/G/B channels, which leads to semantic differences on the channel level. Moreover, channel semantics in visible and infrared modalities essentially represents diverse identity relevance from different views, which greatly affects the performance of specific person retrieval. Therefore, the modality discrepancy can be attributed to the heterogeneity of channel semantics between visible and infrared modalities, which motivates us to settle the existing channel-level problem and propose the Dual-Semantic Consistency Learning Network (DSCNet).

As shown in Fig. 1 (a), we apply the gray-to-color method for the infrared modality. The circles and triangles in Red, Green, Blue colors represent the feature extracted on the R/G/B channels. The proposed DSCNet, learns channel sementic consistency from two aspects, *i.e.* the Inter-Channel Semantic Consistency learning (ICSC) and Inter-Modality Semantic Consistency learning (IMSC). Fig. 1 (b) presents the designs of ICSC and IMSC. ICSC maximizes intra-modality channel semantic consistency by boosting the similarity of numerical distribution between the R/G/B channels. Compared with ICSC which works on fine-grained level, IMSC minimizes inter-modality channel semantic inconsistency on a comprehensive level by reducing the distance of modality-specific features at the meantime.

This is the first work that reinforces channel-level semantic consistency to relieve the infrared and visible modality discrepancy, which helps to extract identity-relevant and discriminative features. In addition, we propose Joint Semantic Metric Learning (JS) to optimize the joint cross-modality features in a fine-grained manner. The basic idea is that channel semantic consistency of intra-and inter-modality should be jointly utilized to narrow the gap between the visible and the infrared modalities, as well as to boost the discriminative representation of identities. JS constrains the distance between feature representations of identities both on the modality and the channel levels, which makes the semantics of the same identity much easier to match. For one thing, we reduce the variations between intra-class instances so that the generalized semantics across modalities of the same identity will be more centralized. For another, by simultaneously reinforcing the correlation between channel semantics and modality semantics of the same identity, and enhancing the identity discrimination, we avoid the difficulties to represent modality semantic discrepancy. The proposed DSCNet effectively optimizes the distribution of instance representations across modalities and prominently boosts the ability to generalization.

Our main contributions can be summarized as follows:

- We propose a novel learning framework named Dual-Semantic Consistency learning Network (DSCNet) for VI-ReID, which attributes the discrepancy between visible and infrared modalities to channel semantic heterogeneity.
- DSCNet is the first attempt to learn cross-modal identity discrimination on the channel level, which is comprehensively different from existing VI-ReID methods on the instance level.
- Extensive experimental results validate that DSCNet presents superiority over current state-of-the-art methods by a surprising margin on two mainstream benchmarks of VI-ReID.

II. Related Work

A. Visible Person Re-Identification

Single modality person re-identification finds the same person across different visible cameras and has achieve prominent performances on existing public datasets [62], [63], [64], [71], [72], [73]. To solve the misalignment human parts and color differences, [38] designs a cascaded WConv module that can extract comparison features for two input images. [60] considers camera style variation and solves it by camera-aware style transfer. [31] proposes a Part-based Convolutional Baseline and Refined part pooling. For spatial localization, [55] aggregates local and global features and the gradual information between them with dynamic training. [57] keeps attention consistency among images of the same person, by a Siamese framework which can incorporate attention and attention consistency. ABD-Net [3] treats the orthogonality regularization diversity as a complementary cue to channel-wise and position-wise attention. [21] explores connections between samples for dataset-level observation, and builds a similarity graph inside a data batch. [48] gives effort to long-range relationships of the image, and makes second-order statistics for the features. To deal with occluded person images, [36] estimates person key points, designs adaptive direction graph convolutional layer which takes the local features as nodes and matches graphs for different images for retrieval. For video person re-identification, [46] develops multi-level Context-aware Part Attention (CPA) model for discriminative and robust local part features. [61] makes a matching between the image and the video by a joint feature projection matrix.

Provided a video with an appearing person without further instance labels in frames during training, [23] designs a weakly supervised method named develop deep graph metric learning (DGML), which measures the consistency of spatial graphs for successive frames and distinguishes spatial graphs between videos. [17] uses network architecture searching to combine pattern information and search for light weighted network. [19] proposes an end-to-end Part-Aware Transformer (PAT) to deal with occluded person via an transformer encoder-decoder structure and achieves satisfied results. For unsupervised person re-identification task, [49] studies intra-inter camera similarity to generate pseudo-labels by supervising with cameras. [52] tackles pseudo label noise by comparing pseudo label similarities during different training stages and refining them accordingly. [29] solves unsupervised domain adaption problem by mapping camera style between different cameras and lets the network learn target camera-invariant features. [1] applies hypothesis transfer learning which can transfer information from the source models and data. For generalized person re-identification, [14] proposes Style Normalization and Restitution (SNR) module which filters out the style relative features by instance normalization and restitute discriminative information.

B. Visible-Infrared Person Re-Identification

Bridging the gap between the heterogeneous features of visible and infrared images is challenging for VI-ReID task [65], [66], [67], [68], [34], [69], [70]. Given a visible or an infrared query image, the task aims to retrieving person in the opposite modality gallery. At the beginning, [26] takes the visible and the infrared images to decrease the affect of noise in human body recognition. [45] analyses popular cross-domain methods and proposes deep zero-padding. [59] applies two-stream network structure and designs a hierarchical cross-modality matching metric learning strategy to fetch modality-specific and modality-shared features. After that, many works give efforts to decrease modality discrepancy via modality-invariant information. [6] uses cutting-edge generative adversarial network to extract discriminative features, and combines the ID loss and the cross-modality triplet loss to minimize inter-class ambiguity and maximize cross-modality similarity. [12] utilizes Sphere Softmax to deal with the correlation between classification subspace and feature subspace, and designs a two-stage training scheme to acquire non-correlated features. To further explore the shared features subspace, [15] segregates the identity and spectrum-related features and designs a two-branch network, one for identity-related features and the other for spectrum-relevant features. [39] proposes Dual-level Discrepancy Reduction Learning to reduce the modality gap which converts a visible or an infrared image to its opposite modality. [32] uses Alignment GAN to incorporate pixel alignment and feature alignment. [13] considers the intra-modality similarities among gallery samples and presents a similarity inference metric to optimize cross-modality image matching. [33][20][5] also focus on exploring the modality shared and identity specific features, and generate cross-modality images. [43] proposes dynamic dual-attentive aggregation (DDAG) which adopts intra-modality part-level and cross-modality graph-level features. [40] designs bi-directional dual-constrained top ranking loss to learn discriminative features. Instead of manually designed learning architectures, [9] finds that appropriately separating Batch Normalization layers is the key to boosting the performance, and designs the BN-oriented network searching strategy. [4] automates the feature selection process by network architecture searching, too. [11] presents MCLNet to fool the modality classifier to concentrate on the modality's irrelevant features. [50] generates an auxiliary gray-scale modality from visible images to approximate the infrared images and solve the tri-modal learning problem. [27] applies pixel-level correspondences across modalities to suppress modality-related information. [41] presents syncretic modality collaborative learning, which generates auxiliary modality that aggregates visible and infrared image features and learns via the three modalities, too. [47] exploits nuanced but discriminative information by a proposed pattern alignment module and a modality alleviation module. [41] designs an information bottleneck strategy (VSD) for representation learning to preserve sufficient features and suppress irrelevant information.

However, existing methods overlook the heterogeneity on channel level, and do not arrange the numerical distribution on channel level. The visible images consist of R/G/B color channels, and the one-channel infrared images mostly get tranformed into R/G/B color channel representations. We explore to constrain the channel semantic consistency on intra-modality channel-level and inter-modality comprehensive level, which promotes performance to accurately distinguish identities on a large margin.

III. Dual-Semantic Consistency Learning

In this section, we firstly formulate the cross-modality task (§ III-A), then introduce the proposed framework of Dual-Semantic Consistency Learning Network (DSCNet). It consists of three major components: Inter-Channel Semantic Consistency learning (ICSC, § III-B), Inter-Modality Semantic Consistency learning (IMSC, § III-C), and Joint Semantic metric learning (JS, §III-D). In the end, we summarize the objective function and algorithm (§ III-E).

A. Formulation

Formally, the visible and the infrared images can be formulated as $\mathcal{V} = \{x_i^v | x_i^v \in \mathcal{V}\}$, $\mathcal{R} = \{x_i^r | x_i^r \in \mathcal{R}\}$, respectively. The corresponding ground-truth labels are denoted as $\mathcal{Y}_v = \{y_{x_i^v} | x_i^v \in \mathcal{V}\}$ and $\mathcal{Y}_r = \{y_{x_i^r} | x_i^r \in \mathcal{R}\}$. We denote $y_{x_i^v}$ as y_i^v, and $y_{x_i^r}$ as y_i^r for ease of representation. VI-ReID matches the visible image x_i^v with the infrared image x_j^r of the same identity in a mutual manner. Therefore, the optimization objective of VI-ReID is to maximize the mapping similarity between the visible image x_i^v and the infrared image x_j^r if they belong to the same identity and keep discrimination between different identities. A visible image contains three color channels and can be formulated as $x_i^v = x_i^v(R_i^v, G_i^v, B_i^v)$. And an one-channel infrared image can also get transferred into three-channel representation as $x_i^r = x_i^r(R_i^r, G_i^r, B_i^r)$ through an inverse operation of color-to-gray method. The feature extractor θ_e extracts feature representation of visible

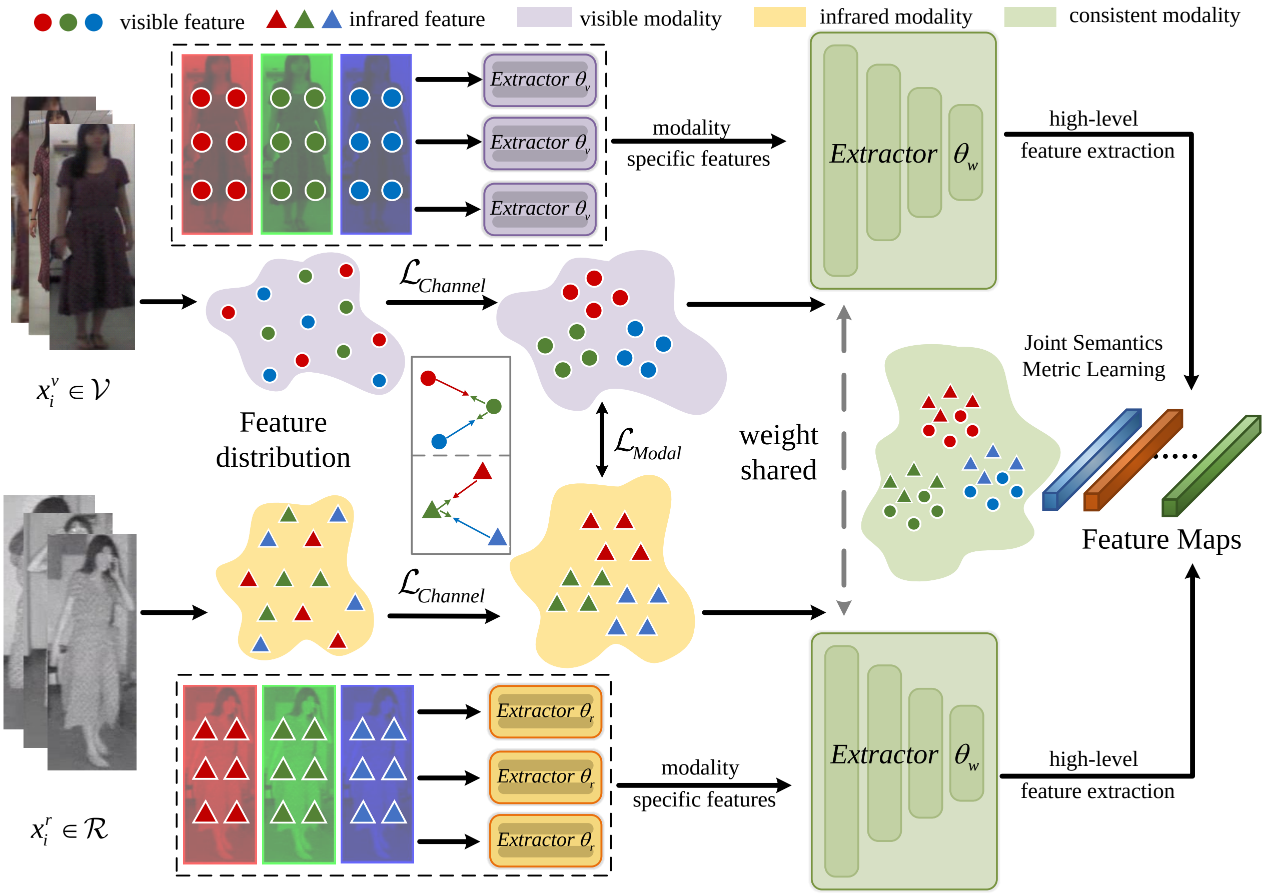


Fig. 2: Illustration of the Dual-Semantic Consistency Learning network (DSCNet). We first extract features through the extractors θ_v and θ_r (1st Residual 1ayer of pretrained ResNet-50) from visible and infrared images. Then we design the inter-channel and inter-modality semantic consistency learning. The semantic-consistent features are fed into weight-shared layers for learning higher-level representations. We supervise the cross-modal retrieval by joint semantics metric learning.

and infrared images f_i^v, f_i^r. Thus the optimization objective can be formulated as :

$$\mathcal{L} = \sum \ell(\theta_e(x_i^v(R_i^v, G_i^v, B_i^v), x_i^r(R_i^r, G_i^r, B_i^r)); y_i^v, y_i^r), \quad (1)$$

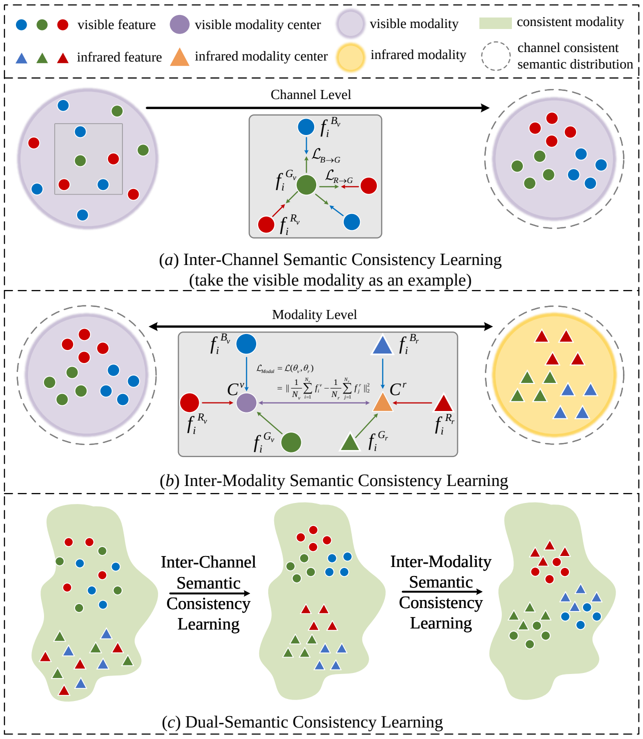
where $\ell(\cdot)$ indicates the mapping computation of the variants.

Fig. 2 illustrates the framework of Dual-Semantic Consistency Learning (DSCNet). DCSNet takes a two-stream network as our feature extractor. Two-stream network utilized for feature extraction consists of two parts, *i.e.* the modality-shared layers θ_w, and the modality-specific layers θ_v, θ_r. Fed with visible and infrared images, it utilizes modality-specific layer θ_v, θ_r to extract the visible and the infrared modality representations, respectively. Then it applies the weight-shared mechanism to get modality-shared features. In DSCNet, we improve this procedure by adding a Dual-Semantive Consistency Learning scheme inside the two-stream network. Specifically, with the visible and infrared input images, we first optimize the consistency of channel semantics inside the modality to balance the distribution of the features on the R/G/B channels. This step is named Inter-Channel Semantic Consistency Learning (ICSC). Then we comprehensively optimize the consistency of inter-modality channel semantics to narrow the gap of identity semantics between the visible and the infrared modalities, which we name as Inter-Modality Semantic Consistency Learning (IMSC). Last but not least, we jointly optimize the semantics to achieve identity-level discrimination across modalities via embedding the instances into highly related identities.

B. Inter-Channel Semantic Consistency Learning

A piece of basic knowledge is that, for an image containing R/G/B channels, the different channel has independent semantics from each other and the semantics of R/G/B channels has a significant correlation with each other to represent the comprehensive instance semantics. As shown in Fig. 2, we acquire highly modality-specific features f^v and f^r after modality-specific layers θ_v, θ_r. For the same identity, modality-specific features correspond to inherent but different semantics, due to the visible and infrared imaging principles. Channel semantics intrinsically represent the fine-grained and diverse identity-relevant information.

Since the infrared images are captured according to the amount of heat radiated from the surface of the objects, they can not be regarded as common images consisting of three channels. The variations between channel images R^v, G^v, B^v, R^r, G^r, and B^r significantly contribute to the modality discrepancy. However, most existing methods mainly reduce the modality discrepancy on the instance level and



(*a*) Inter-Channel Semantic Consistency Learning (take the visible modality as an example)

(*b*) Inter-Modality Semantic Consistency Learning

(*c*) Dual-Semantic Consistency Learning

Fig. 3: Dual-Semantic Consistency Learning. (a) represents the process of the inter-channel semantic consistency learning, (b) represents the process of the inter-modality semantic consistency learning and (c) represents the composition of the dual-semantic consistency learning.

focus on modality-shared feature extraction. They treat the infrared images as common images consisting of three channels, which retain channel-level semantic discrepancy during processing. The ignorance of the channel semantic alignment significantly leads to the loss of the fine-grained channel-level identity relevance for person re-identification.

In this work, we attribute the main modality discrepancy to the channel semantic heterogeneity between the visible modality channels (R^v, G^v, B^v) and the infrared modality channels (R^r, G^r, B^r). We are inspired to eliminate modality discrepancy as much as possible from the perspective of channels. The key to this problem lies in that how can we maintain the identity relevance of these channel features, while reducing the variations of different channel semantics. Since the extended three-channel IR images are fake R/G/B images, we try to make the network learn similar R/G/B channel distributions as the visible images. Formally, we consider the pairwise channel semantics of the visible images $x_i^v(R_i^v, G_i^v, B_i^v) \in \mathbb{R}^{B\times C\times H\times W}$ and the infrared images $x_i^r(R_i^r, G_i^r, B_i^r) \in \mathbb{R}^{B\times C\times H\times W}$. The modality-specific feature $f_i^v, f_i^r \in \mathbb{R}^{B\times C'\times H'\times W'}$ are obtained from the modality-specific extractor θ_v, θ_r. We split f_i^v and f_i^r on the channel dimension denoted as $f_i^v = [f_i^{R_v}, f_i^{G_v}, f_i^{B_v}]$, $f_i^r = [f_i^{R_r}, f_i^{G_r}, f_i^{B_r}]$, where $f_i^{R_v}, f_i^{G_v}, f_i^{B_v}, f_i^{R_r}, f_i^{G_r}, f_i^{B_r} \in \mathbb{R}^{B\times C''\times H'\times W'}$, $C'' = C'/3$. Accordingly, the objective is to align the semantic distribution of $f_i^{R_v}$ with $f_i^{R_r}$, $f_i^{G_v}$ with $f_i^{G_r}$, and $f_i^{B_v}$ with $f_i^{B_r}$.

Under this circumstance, our semantic consistency learning achieves this goal mainly with two modules, as Fig. 3 shows. The first one is the Inter-Channel Semantic Consistency Learning (ICMC). The channel semantic consistency indicates the similarity of the numerical distribution of the Red, Green, and Blue channels. As Fig. 3 (a) shows, for each modality, we minimize the inter-channel semantic difference by the alignment of channel semantics and maximize intra-image channel semantics consistency at the same time. We consider the semantic consistency as similarity of logistic distribution between channel features $f_i^{R_v}, f_i^{G_v}, f_i^{B_v}$, as well as $f_i^{R_r}, f_i^{G_r}, f_i^{B_r}$. According to the experiments, we find that choosing the Green channel as a center and constraining Red Blue to consist with the Green channel obtains the most satisfactory results, rather than the other schemes. Both visible and infrared features need refining the channel-level consistency, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{ICSC}(\theta_v,\theta_r) = \frac{1}{N}\sum_{i=1}^{N}(f_i^{R_v}\cdot\log\frac{f_i^{R_v}}{f_i^{G_v}} + f_i^{R_r}\cdot\log\frac{f^{R_r}}{f^{G_r}})\\ + \frac{1}{N}\sum_{i=1}^{N}(f_i^{B_v}\cdot\log\frac{f_i^{B_v}}{f_i^{G_v}} + f_i^{B_r}\cdot\log\frac{f_i^{B_r}}{f_i^{G_r}}),\end{aligned} \tag{2}$$

where $\mathcal{L}_{ICSC}$ denotes the semantic consistency between color channels of Red and Green, and $\mathcal{L}_{ICSC}$ represents the semantic consistency between color channels of Blue and Green.

In ICMC, we focus on updating the parameters of modality-specific feature extractors θ_v and θ_r alternatively until they reach equilibrium. This ensures maximized intra-modality channel semantics consistency and minimizes the inter-channel semantic discrepancy. The parameters of θ_v, θ_r can be optimized as:

$$\begin{aligned}\hat{\theta}_v &= \arg\min_{\theta_v}(\mathcal{L}_{ICSC}(\theta_v,\hat{\theta}_r) + \mathcal{L}_{ICSC}(\theta_v,\hat{\theta}_r))\\ \hat{\theta}_r &= \arg\min_{\theta_r}(\mathcal{L}_{ICSC}(\hat{\theta}_v,\theta_r) + \mathcal{L}_{ICSC}(\hat{\theta}_v,\theta_r))\end{aligned} \tag{3}$$

By reinforcing the modality-specific feature extractors learn channel level consistent information, ICMC reliefs discrepancy of internal modality to a large extent.

C. Inter-Modality Semantic Consistency Learning

In most of the cross-domain person re-identification works, the performance depends on how to map the cross-modality representations of the same identity compactly and keep discrimination between identities. However, many factors like noise and occlusion in single modality easily contribute to the feature variations across modalities which makes the VI-ReID task more challenging [51], [30]. To address the limitations derived from modality discrepancy, it is important to construct cross-modality alignment. Therefore, we utilize inter-modality semantic consistency to indicate the similarity between visible and infrared features distributions on the modality-level. Even each modality achieves consistency on the R/G/B channels, the visible and the infrared modality semantics of the same identity are still independent from each other. Thus we maximize inter-modality channel semantics consistency at the meantime.

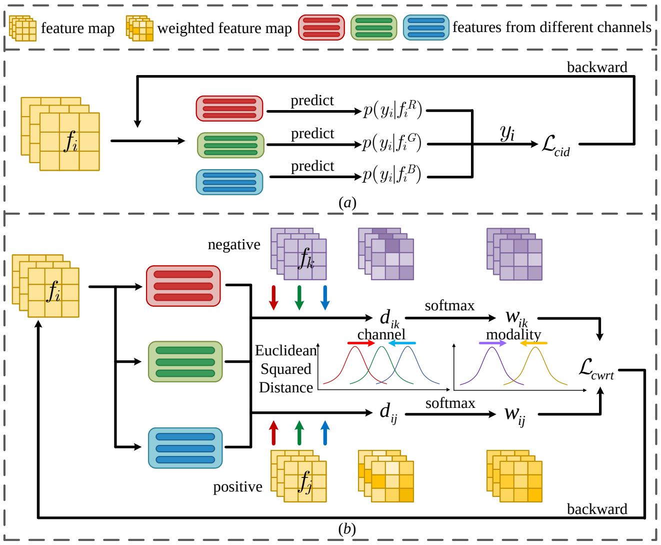


Fig. 4: Illustration of the Joint Semantics Metric Learning. (a) The process of predicting identity according to features extracted from different channels. (b) The process of hard sample mining.

Fig. 3 (b) shows the proposed Inter-Modality Semantic Consistency Learning (IMSC) processing. By obtaining modality-specific feature f_i^v, f_i^r with intra-modality channel semantic consistency, we further eliminate channel semantic discrepancy across modalities. Since the modality-specific extractor θ_v, θ_r extracts independent and intuitive features across modalities, we represent each modality with the Euclidean centers C^v, C^r of the feature semantics:

$$C^v = \frac{1}{N_v}\sum_{i=1}^{N_v} f_i^v, \quad C^r \quad = \frac{1}{N_r}\sum_{i=1}^{N_r} f_i^r, \tag{4}$$

where N_v and N_r denotes the number of samples in the visible and the infrared modalities, respectively. C^v and C^r are computed in a batch. IMSC takes the heterogeneous semantics to learn the representation for modality level channel semantic consistency, regardless of identities. The distance between the Euclidean centers C^v and C^r can be aligned according to metric learning, so that features extracted by the evolved modality-specific extractor θ_v and θ_r will represents more modality consistency. In IMSC, $C^v, C^r, f_i^v, f_i^r \in \mathbb{R}^{B\times C'\times H'\times W'}$. The objective is to maximize cross-modality semantics consistency and minimize the visible and infrared feature divergence:

$$\mathcal{L}_{Modal} = \mathcal{L}(\theta_v, \theta_r) = ||C^v - C^r||_2^2. \tag{5}$$

The parameters of modality-specific feature extractors θ_v and θ_r are updated accordingly, which can be optimized alternately as:

$$\begin{aligned}\hat{\theta}_v &= \arg\min_{\theta_v}(\mathcal{L}_{Modal}(\theta_v, \hat{\theta}_r), \\ \hat{\theta}_r &= \arg\min_{\theta_r}(\mathcal{L}_{Modal}(\hat{\theta}_v, \theta_r).\end{aligned} \tag{6}$$

Therefore, the proposed IMSC can comprehensively improve the consistency of the modality-shared identity semantic and reduce the channel discrepancy on the modality level. Fig. 3 (c) demonstrates the collaboration of ICSC and IMSC. The advantages of Dual-Semantic Consistency learning are two folds. On the one hand, we desire to extract representative modality-specific semantics, which inherently represents the discrimination of identities in a single modality. On the other hand, we can effectively maintain the modality-specific features and control the comprehensive cross-modality matching.

D. Joint Semantics Metric Learning

Most existing metric learning methods attach great importance on dealing with the distances between feature semantics of identities, like ID loss [58]:

$$\mathcal{L}_{id} = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \log(p(y_i|x_i)), \tag{7}$$

and weighted regularized triplet loss [2]:

$$\begin{aligned}&\mathcal{L}_{wrt}(i,j,k) = \log(1+\exp(w_i^p d_{ij}^p - w_i^n d_{ik}^n)), \\ &w_i^p = \frac{\exp(d_{ij}^p)}{\sum_{d^p\in\mathcal{P}}\exp(d^p)}, w_i^n = \frac{\exp(d_{ik}^n)}{\sum_{d^n\in\mathcal{N}}\exp(d^n)},\end{aligned} \tag{8}$$

where d denotes the distance between two samples. ID loss and Triplet loss optimize the distribution of features on the instance level. Some methods adopt center loss [22] to reduce the variations and learn representative features. But they are still on instance level. For one thing, instance semantics consisting of channel semantics determines that optimization staying on the instance level will be a coarse-grained approach. For another, features extracted from instances always easily get influenced by the factors like noise and shielding, which leads to confusion in terms of modality discrepancy and instance variations.

In this paper, we enhance the semantic consistency in the representation space from two perspectives. To boost the identity discrimination of the semantics, we design the inter-modality and inter-channel semantic consistency learning. Furthermore, to fully exploit the advantages of the semantic-consistent representations across modalities, we propose Joint Semantic Metric Learning (JS) to deal with the problem. The strategy of JS is shown in Fig. 4. Formally, we obtain semantic consistent features f^v and f^r from the modality specific extractor θ_v and θ_r. f^v and f^r provide the corresponding channel semantic representations. Then we utilize weight-shared feature extractor θ_w to obtain high-dimensional representations $[f^{R_v}, f^{G_v}, f^{B_v}], [f^{R_r}, f^{G_r}, f^{B_r}] \in \mathbb{R}^{B\times C'\times H'\times W'}$ and modality-shared discrimination between identities (Fig. 4 (a)). The relationship between these channel-level features and the ground-truth labels can be formulate with information entropy and supervised by the Channel-level ID loss $\mathcal{L}_{cid}$

$$\begin{aligned}&\mathcal{L}_{cid} = \\ &-\frac{1}{N_v}\sum\nolimits_{i=1}^{N_v}(\log(p(y_i^v|f_i^{R_v})) + \log(p(y_i^v|f_i^{G_v})) + \log(p(y_i^v|f_i^{B_v}))) \\ &-\frac{1}{N_r}\sum\nolimits_{i=1}^{N_r}(\log(p(y_i^r|f_i^{R_r}) + \log(p(y_i^r|f_i^{G_r})) + \log(p(y_i^r|f_i^{B_r}))),\end{aligned} \tag{9}$$

where $p(\cdot)$ denotes the prediction probability of the visible channel features $f_i^{R_v}$, $f_i^{G_v}$, $f_i^{B_v}$ belongs to identity y_i^v, or the infrared channel features $f_i^{R_r}$, $f_i^{G_r}$, $f_i^{B_r}$ belongs to identity y_i^r. p is calculated by cross-entropy. In addition, we constrain the distribution of channel-level features to more fine-grained optimizing the cross-modal person retrieval, and propose the Channel-level Weighted Regularized Triplet Loss

$$\begin{aligned}
&\mathcal{L}_{cwrt} = \log(1+\exp(\sum w_{ij}^{p} d_{ij}^{p} - \sum w_{ik}^{n} d_{ik}^{n})),\\
&w_{ij}^{p} = \frac{\exp(d_{ij}^{p})}{\sum_{d_{ij}^{p}\in\mathcal{P}}\exp(d_{ij}^{p})}, w_{ik}^{n} = \frac{\exp(d_{ik}^{n})}{\sum_{d_{ik}^{n}\in\mathcal{N}}\exp(d_{ik}^{n})},\\
&d_{ij} = ||f_i^R - f_j^R||_2^2 + ||f_i^G - f_j^G||_2^2 + ||f_i^B - f_j^B||_2^2,
\end{aligned} \tag{10}$$

where (i, j, k) represent the hard triplet samples that are mined during training progress, the superscript p and n denote the positive samples and the negative samples, respectively. The loss function of Joint Semantic Metric Learning can be represented as:

$$\begin{aligned}
\mathcal{L}_{Joint}(\theta_w) &= \mathcal{L}_{cid} + \mathcal{L}_{cwrt}\\
&= -\frac{1}{N}\sum\nolimits_{i=1}^{N}\log(p(y_i|f_i^R) + p(y_i|f_i^G) + p(y_i|f_i^B))\\
&+ \log(1+\exp(\sum w_{ij}^{p} d_{ij}^{p} - \sum w_{ik}^{n} d_{ik}^{n})).
\end{aligned} \tag{11}$$

In the Joint Semantics Metric Learning progress, we focus on updating the parameters of weight-shared feature extractors θ_w. It ensures to optimization of the distribution of channel feature embeddings on the channel level. Besides, the model can avoid getting confused about the instance variations and modality discrepancy. The parameters θ_w can be optimized as:

$$\hat{\theta}_w = \arg\min_{\theta_w}(\mathcal{L}_{Joint}(\theta_w)). \tag{12}$$

E. Objective Function

The proposed DSCNet contains the ICSC, IMSC and JS structure. The objective function of DSCNet is improved with the following terms.

- $\mathcal{L}_{ICSC}(\theta_v, \theta_r)$. We reduce the semantic divergence between color channels.
- $\mathcal{L}_{Modal}(\theta_v, \theta_r)$.We reinforce the cross-modality representation on semantic consistency and eliminate the modality discrepancy to a large extent.
- $\mathcal{L}_{Joint}(\theta_w)$. We optimize the distribution of channel feature embeddings on the channel level and utilize the joint semantic consistency of channels and modalities.

Considering the above terms, the objective function of DSCNet can be represented as:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{Modal}(\theta_v, \theta_r) + \lambda_2 \cdot \mathcal{L}_{ICSC}(\theta_v, \theta_r) + \lambda_3 \mathcal{L}_{Joint}(\theta_w) \tag{13}$$

It is worth noting that, the objective functions target different modules in the whole network. $\mathcal{L}_{ICSC-I}, \mathcal{L}_{ICSC-II}, \mathcal{L}_{Modal}$ optimize the modality-specific extractor θ_r and θ_v, and the term $\mathcal{L}_{Joint}$ optimizes the weight-shared extractors θ_w. They are combined to supervise the network to extract modality-irrelevant features with prominent modality and channel semantic consistency.

IV. Experimental Results

We first apply two popular VI-ReID datasets, the SYSU-MM01 [45] and the RegDB datasets [26] (§ IV-A) to verify the proposed DSCNet. We then give the implementation details (§ IV-B), conduct the ablation study with an analysis of the performance of each network module (§ IV-D), and provide parameter analysis by visualization. Finally, we make comparison with state-of-the-art methods (§ IV-F).

A. Datasets and Evaluation Metrics

SYSU-MM01 [45] is the largest public available dataset for VI-ReID, which provides 287,628 visible images and 5,792 infrared images in total. It collects images from 491 identities by 6 cameras. Camera 1, 2, 4, and 5 are visible cameras, and Camera 3 and 6 are infrared cameras. SYSU-MM01 provides 296 identities for training, 99 for validation, and 96 for testing. Especially, it provides two modes during testing, *i.e.* the *all-search* mode and the *indoor-search* mode. In the *all-search* mode, the visible images captured by cameras 1, 2, 4 and 5 are set as gallery set, and in the *indoor-search* mode, visible images captured by cameras 1 and 2 are adopted. In both modes, infrared images captured by cameras 3 and 6 are utilized as the probe sets. The collected images variate greatly in terms of perspective, aspect ration, brightness, person appearances. As a result, SYSU-MM01 is very challenging for VI-ReID task.

RegDB [26] is another popular VI-ReID dataset, and the infrared images are initially used to decrease the noise in visible person re-identification. It contains 412 identities, and 10 visible and 10 infrared samples for each person. The persons have 254 females and 158 males. 412 persons are photoed from the front view, and 256 of them are captured from the back view. For testing, RegDB provides two modes, *visible-to-infrared* mode and *infrared-to-visible* mode. When one modality sample is treated for a gallery setting, the other modality samples are for a probe set. Because the RegDB dataset is collected by a dual camera system, the relative person positions and backgrounds do not change a lot. Besides, the images in RegDB are of small size, low resolution, and little appearance or pose changes, so that identity always represents very similar in different images. We randomly select 206 identities for training, and the other 206 identities for testing, referring to the evaluation protocol of [59]. We test for 10 trials to obtain stable results [39].

Evaluation Protocol. We use the cumulative matching characteristics (CMC) [24] and the Mean average precision (mAP) evaluation metrics. CMC indicates the accuracy of top-K predictions. the mAP is the mean integral of recall concerning precision. CMC and the mAP standard metrics reflect the effectiveness of the identification prediction.

B. Implementation Details

Training. We implement DSCNet on a single NVIDIA 2080Ti GPU with PyTorch. Firstly, a ResNet-50 pre-trained on ImageNet is adopted as the backbone network. The modality-specific extractors θ_v and θ_r are initialized independently by the basic ResNet. We apply weight-shared network θ_w to extract high-dimensional features and take the two-stream network in AGW [2] with channel-level random erasing (CRE) [25] as the backbone. The mini-batch size of instances is set to 48. During the training stage, there are 24 visible and 24 infrared images captured with 6 people in a mini-batch. We utilize the popular data augmentation operations, including random cropping, random horizontal flipping, and channel random erasing. For channel random erasing, each image is cropped into 288×144 and flipped, and then erased on a channel level. The SGD optimizers is set with a momentum $p = 0.9$ and a decay $d = 5 \times 10^{-4}$. The learning rates of the feature extractor $\theta_v, \theta_r, \theta_w$ are scheduled differently and are set to $1/10$ of the classifiers. We design the warm-up learning rate, with an initial setting of 0.1. It decays to 0.01 between 20 and 39 epochs, 0.003 between 40 and 49 epochs, and 0.001 after 50 epochs.

Testing. We use the trained two-steam network to extract features of the images from the query set and gallery set and take the classifier for re-identification. In this procedure, there is no need to utilize the ICSC and the IMSC modules.

C. Parameter Analysis

Eq. 13 introduces hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ to balance the contribution of different loss functions. So we analyze the hyper-parameters of network by testing each hyper-parameter on different values which varies from 1 to 8. As illustrated in Fig. 7, with the increased value of λ_1, λ_2 and λ_3, the Rank-1 score shows varying degrees of decreasing trend. It can be found that when parameter λ_3 increases, there is a significant decline in Rank-1 score. According to quantities of experiment results at the current learning rate, it is the most effective when λ_1, λ_2, and λ_3 are all set to 1. We can conclude that these loss functions focus on different perspectives of optimization and their impacts can be weighted by the hyper-parameters for a better performance.

D. Ablation Study

To verify the function of ICSC, IMSC, and JS, we evaluate each of the three components and their combination by conducting different experiments on the SYSU-MM01 dataset in both of the *all-search* and the *in-door* modes and make analyses accordingly. The results of the ablation study are shown in Tab. I. In the setting of merely *Base* (the first row), we utilize the loss function $\mathcal{L}_{id}$ and $\mathcal{L}_{wrt}$ for identity discrimination.

Effectiveness of Joint Semantics Metric Learning (JS). In *all-search* mode, the setting *Base*+$\mathcal{L}_{cid}$ adopts $\mathcal{L}_{cid}$ instead of $\mathcal{L}_{id}$. It achieves a bonus of 2.08% on Rank-1 and 2.26% on mAP. Similarly, *Base*+$\mathcal{L}_{cwrt}$ utilizes $\mathcal{L}_{cwrd}$ instead of $\mathcal{L}_{wrt}$. Compared with *Base*, it obtains 2.81% improvement in Rank-1 and 1.55% in mAP. *Base*+$\mathcal{L}_{cid}$ + $\mathcal{L}_{cwrt}$ works better than individually adopt $\mathcal{L}_{cid}$ and $\mathcal{L}_{cwrt}$. Joint Semantic Metric Learning lies based on modality and channel semantic consistency. Therefore, we adopt it with ICSC and IMSC in the following settings. From the fifth row to the eleventh row, $\mathcal{L}_{cid}$ and $\mathcal{L}_{cwrt}$ are adopted instead of $\mathcal{L}_{id}$ and $\mathcal{L}_{wrt}$.

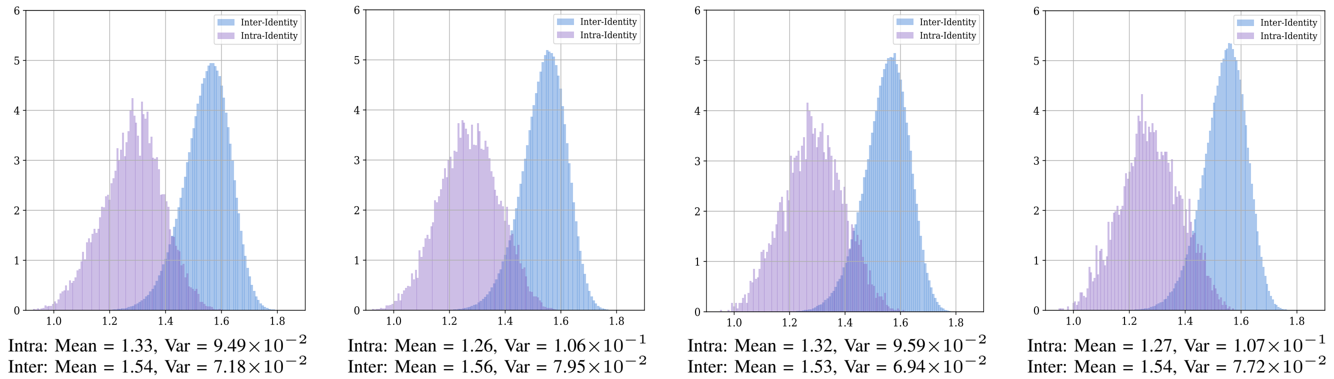
Effectiveness of Inter-Channel Semantic Consistency Learning (ICSC). In ICSC, there are two major losses, $\mathcal{L}_{ICSC\text{-}I}$ and $\mathcal{L}_{ICSC\text{-}II}$. Since we find that keeping the Red and the Blue channels consistent with the Green channel overcomes the other channel settings, we merely conduct experiments to evaluate the performance of $\mathcal{L}_{ICSC\text{-}I}$ and $\mathcal{L}_{ICSC\text{-}II}$ rather than another channel semantic consistency learning strategy. As shown in Tab. I, taking the *all-search* mode as an example, the baseline with the JS which is denoted as *base*+$\mathcal{L}_{cid}$+$\mathcal{L}_{cwrt}$, achieves a Rank-1 score of 62.58% and a mAP score of 57.98%. When it is trained with an additional loss $\mathcal{L}_{ICSC\text{-}II}$, the Rank-1 score is improved by 1.87% and the mAP by 3.35%. When the baseline with the JS is trained with $\mathcal{L}_{ICSC\text{-}I}$, the Rank-1 and the mAP are improved by 2.11% and 4.10%. Baseline with the JS works with both $\mathcal{L}_{ICSC\text{-}II}$ and $\mathcal{L}_{ICSC\text{-}I}$ brings a further performance boost, with a 69.13% Rank-1 score and a 65.54% mAP, higher than the setting of *base*+$\mathcal{L}_{cid}$+$\mathcal{L}_{cwrt}$+$\mathcal{L}_{ICSC\text{-}II}$ and *base*+$\mathcal{L}_{cid}$+$\mathcal{L}_{cwrt}$+$\mathcal{L}_{ICSC\text{-}I}$. It reveals that constraining the Red channel and the Blue channel to be close to the Green channel in terms of the feature distribution consistency are independent and complementary. By aligning them according to color channels, the features of the R/G/B three channels represent more compactly and regularly.

Effectiveness of Inter-modality Semantic Consistency Learning (IMSC). We conduct more experiments to testify IMSC on the SYSU-MM01 dataset. Different from ICSC, IMSC constrains the centers of the two modality features to be close. As shown in Tab. I, $\mathcal{L}_{Modal}$ represents the inter-modality loss function. We take the *all-search* mode as an example, too. It can be found that with the IMSC constraints, the baseline with $\mathcal{L}_{cid}$ and $\mathcal{L}_{cwrt}$ is enhanced by a Rank-1 accuracy of 3.42% and a mAP of 5.52%. In the setting of *base*+$\mathcal{L}_{cid}$+$\mathcal{L}_{cwrt}$+$\mathcal{L}_{ICSC\text{-}II}$+$\mathcal{L}_{Modal}$, the metric scores are enhanced by 6.60% and 6.85% in Rank-1 and mAP, respectively. The setting of *base*+$\mathcal{L}_{cid}$+$\mathcal{L}_{cwrt}$+$\mathcal{L}_{ICSC\text{-}I}$+$\mathcal{L}_{Modal}$ also improves the two metric scores, 8.15% in Rank-1 and 7.53% in mAP. When baseline, ICSC, IMSC, and JS operate together, as shown in the last line of Tab. I, the network achieves the best performance, a 73.89% Rank-1 score, and a 69.47% mAP score. Adopting ICSC and IMSC simultaneously can obtain better performance than individually using them. It indicates that ICSC and IMSC adjust the feature distribution in different ways. The former adjusts the distribution inside the modality on a fine-grained channel level, while the latter adjusts channel semantics comprehensively between modalities. ICSC constrains the feature distribution according to inter-channel alignment, and IMSC impels the centers of feature semantics of the two modalities to be close. ICSC and IMSC are also complementary and have a relationship of mutual promotion.

E. Visualization analysis

Visualization of t-SNE for distribution of feature semantics. In Fig. 6a, we utilize dash line to circle each identity. For the baseline, each identity has a large dash circle and contains samples of the two modalities with the dispersed distribution. For DSCNet, since we design the ICSC and the IMSC learning strategy to align the semantic distribution from the fine-grained inter-channel and the comprehensive inter-modality perspectives, each identity can be cast to a more compact distribution.

Visualization of Intra-class feature distance. The visualization results are shown in Fig. 5. In the all and indoor-search modes, mean values for the feature distances of intra-class decline prove that DSCNet successfully reduces the modality divergence compared with baseline. Meanwhile, the mean value of the feature distance for the inter-class is becoming larger. It proves that DSCNet learns better identity discrimination between different classes compared with baseline.



Intra: Mean = 1.33, Var = 9.49×10^{-2}
Inter: Mean = 1.54, Var = 7.18×10^{-2}

(a) Base-all

Intra: Mean = 1.26, Var = 1.06×10^{-1}
Inter: Mean = 1.56, Var = 7.95×10^{-2}

(b) DSCNet-all

Intra: Mean = 1.32, Var = 9.59×10^{-2}
Inter: Mean = 1.53, Var = 6.94×10^{-2}

(c) Base-indoor

Intra: Mean = 1.27, Var = 1.07×10^{-1}
Inter: Mean = 1.54, Var = 7.72×10^{-2}

(d) DSCNet-indoor

Fig. 5: Feature Distance. We visualize the feature distances of intra-and-inter classes and we provide the statistical variables of the curves. Accordingly, In the all and indoor-search modes, mean values for the feature distances of intra-and-inter classes obviously decline, which proves that DSCNet successfully reduces the modality divergence compared with baseline.

Methods						SYSU-MM01 *all-Search*					SYSU-MM01 *indoor-Search*				
Base	ICSC § III-B		IMSC § III-C	JS § III-D											
	$\mathcal{L}_{ICSC-I}$	$\mathcal{L}_{ICSC-II}$	$\mathcal{L}_{Modal}$	$\mathcal{L}_{cid}$	$\mathcal{L}_{cwrt}$	r=1	r=5	r=10	r=20	mAP	r=1	r=5	r=10	r=20	mAP
✓						59.11	84.93	92.22	96.74	54.03	62.41	85.14	90.62	96.56	67.98
✓	✓			✓	✓	64.45	88.43	94.77	98.26	61.33	67.07	91.17	96.42	99.41	72.68
✓		✓		✓	✓	64.69	88.93	95.21	98.66	62.08	68.98	91.03	97.15	99.41	74.73
✓			✓	✓	✓	66.00	89.09	94.71	98.37	63.50	72.74	92.98	97.28	99.59	77.60
✓	✓		✓	✓	✓	69.18	90.90	96.06	98.95	64.83	75.09	93.07	96.78	99.23	78.55
✓		✓	✓	✓	✓	70.73	91.43	95.58	98.40	65.51	75.82	93.66	97.60	99.50	79.42
✓	✓	✓		✓	✓	69.13	91.01	95.69	**98.95**	65.54	74.18	94.75	**98.73**	99.77	77.87
✓				✓		61.19	84.99	91.06	95.82	56.29	63.32	87.95	93.61	96.83	67.90
✓					✓	61.92	84.38	90.98	95.79	55.58	64.18	87.64	93.93	98.28	69.55
✓				✓	✓	62.58	84.54	91.53	96.42	57.98	65.13	87.91	93.25	97.74	70.08
✓	✓	✓	✓	✓	✓	**73.89**	**92.09**	**96.27**	98.84	**69.47**	**79.35**	**95.74**	98.32	**99.77**	**82.68**

TABLE I: Ablation study of ICSC, IMSC, JS on the *all-search* mode of SYSU-MM01 dataset. Rank-r accuracy(%) and mAP(%)are reported. Where ”Base” indicates the AGW [2] with random erasing supervised by $\mathcal{L}_{id}, \mathcal{L}_{wrt}$.

Visualization of Heat Map for DSCNet. As shown in Fig. 6b and Fig. 6c, each heat map corresponds to the output features of the well learns two-stream network. The brighter the region, the higher weights are assigned to the corresponding places. The highlight regions in the heatmaps are mostly focused on the faces, shoulders, and feet. Compared with the heat maps extracted from the baseline shown on the top, the heat maps on the bottom from the DSCNet are more focused on the identity-relevant features. This reveals that DSCNet has a stronger anti-interference ability to factors such as illumination, human gestures, occlusions, and so on. It can pay attention to human bodies during retrieval and be discriminative between identities.

Visualization of Visible-Infrared Retrieval Results. We randomly selected 6 infrared images and 6 visible images from the SYSU-MM01 dataset, and provide top 10 images with the highest similarities which are predicted by the DSCNet, as shown in Fig. 8. The retrieval results marked with green boxes are correct, and those marked with red boxes are wrong. Even the query images are difficult for human visual system, it can be seen from the retrieval results that the DSCNet is able to discriminate identities and correct results usually have a high matching degree with the images to be retrieved.

F. Comparison With the State-of-the art Methods

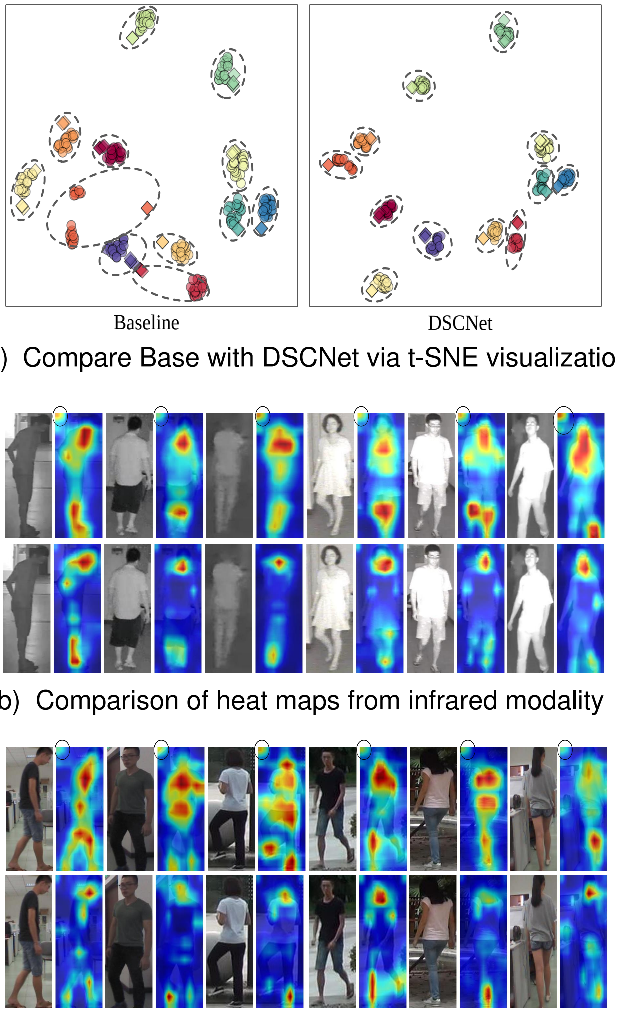
In this subsection, we verifies the proposed DSCNet by comparing with state-of-the-art methods, including Zero-Pad [45], HCML [59], cmGAN [6], HSME [12], AliGAN [37], CMSP [44], JSIA [35], XIV [16], MACE [56], MSR [8], Hi-CMD [5], cm-SSFT [20], AGW [2], VSD [66],CoAL [65] MCLNet [11], SMCL [41], CM-NAS [10], MPANet [47], DG-VAE [28], HAT [42], Hi-CMD [5], DDAG [43], NFS [4], which are proposed lately. Tab. II and Tab. III demonstrate the performance results of all the methods on both the SYSU-MM01 and the RegDB datasets, respectively.

From the experiments on SYSU-MM01 dataset (Tab. II), it can be found that the proposed DSCNet outperforms the the other well-known works, and achieves Rank-1 score of 73.89% and a mAP score of 69.47% in *All Search* mode, and a 79.35% Rank-1 and a 82.65% mAP in *Indoor Search* modes, respectively. Although in the SYSU-MM01 dataset, samples vary heavily accross modality in terms of aspect-ratio, illumination, occlusion, background, perspective, relative position of human, and gesture, the DSCNet can solve the cross-modality human retrieval problem by pursuing channel semantic consistency and inter-modality semantic consistency.

In most of the VI Re-ID works, like cm-SSFT [20], CM-NAS [10], and AGW [2], they focus on how to extract modality-invariant features with different designed learning methods. Instead, we adjust the semantics in the R/G/B between channels and across modalities, so that the Red and Blue channels will consist of the Green channel, and the feature

TABLE II: Comparison with the state-of-the-arts on SYSU-MM01 dataset. Rank-k accuracy (%) and mAP (%) are reported.

Settings		*All Search*				*Indoor Search*			
Method	Venue	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP
Zero-Pad [45]	ICCV17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [59]	AAAI18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [6]	IJCAI18	26.97	67.51	80.56	27.80	31.63	77.23	89.18	42.19
HSME [12]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-
AliGAN [37]	ICCV19	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
CMSP [44]	IJCV20	43.56	86.25	-	44.98	48.62	89.50	-	57.50
JSIA [35]	AAAI20	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
XIV [16]	AAAI20	49.92	89.79	95.96	50.73	-	-	-	-
MACE [56]	TIP20	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
MSR [8]	TIP20	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Hi-CMD [5]	CVPR20	34.94	77.58	-	35.94	-	-	-	-
cm-SSFT [20]	CVPR20	47.70	-	-	54.10	-	-	-	-
CoAL [65]	MM20	57.22	92.29	97.57	57.20	63.86	95.41	98.79	70.84
AGW [2]	TPAMI21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
MCLNet [11]	ICCV21	65.40	93.33	97.14	61.98	72.56	96.88	99.20	76.58
SMCL [41]	ICCV21	67.39	92.87	96.76	61.78	68.84	96.55	98.77	75.56
NFS [4]	CVPR21	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
CM-NAS [10]	CVPR21	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
MPANet [47]	CVPR21	70.58	96.21	98.80	68.24	76.74	98.21	99.57	80.95
DSCNet	**Ours**	**73.89**	**96.27**	**98.84**	**69.47**	**79.35**	**98.32**	**99.77**	**82.65**



(a) Compare Base with DSCNet via t-SNE visualization

(b) Comparison of heat maps from infrared modality

(c) Comparison of heat maps from visible modality

Fig. 6: Heat maps extracted by Base and DSCNet are shown in the top and bottom. DSCNet is able to resist background interference compared with our baseline.

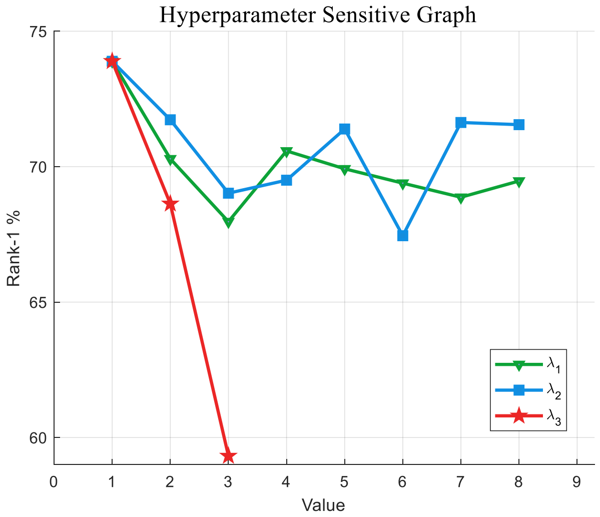
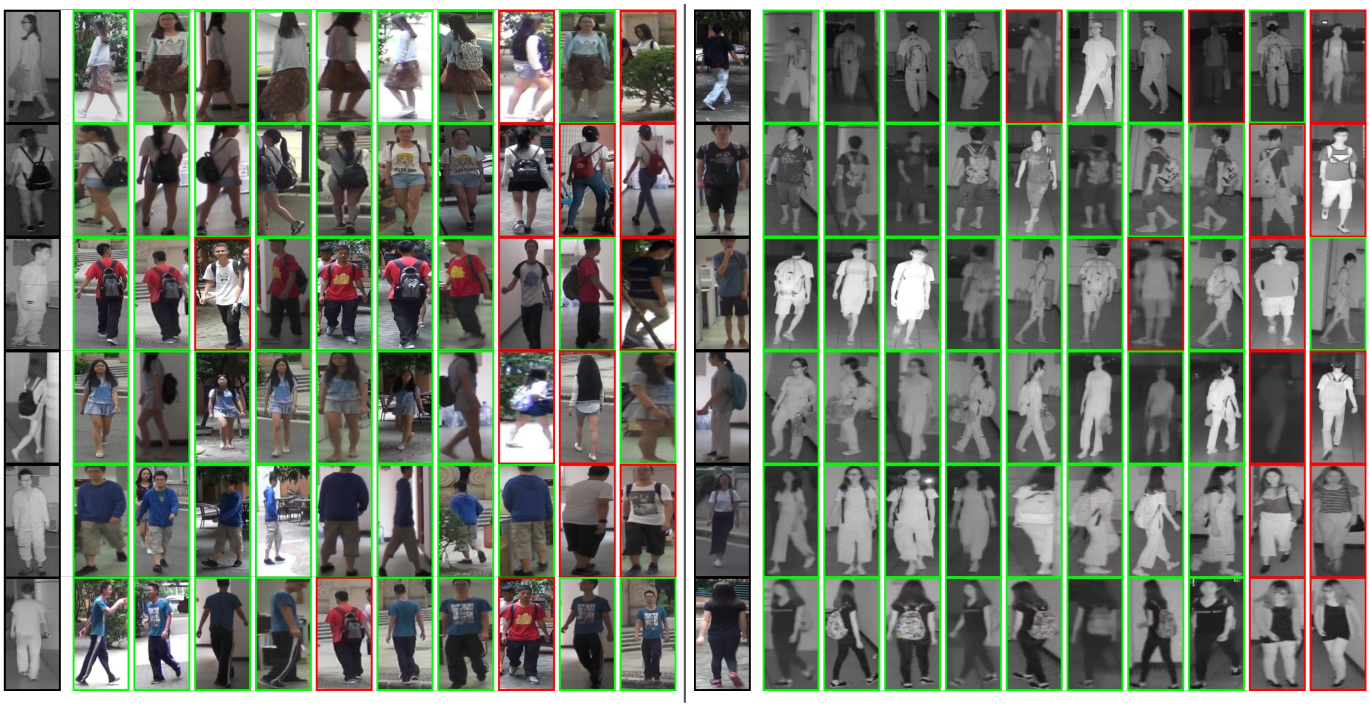


Fig. 7: Parameter Analysis for our Objective Function. We varied the value of λ_1, λ_2 and λ_3 from 1 to 8 to test the performance of the network.

center of the two modalities will be close. DSCNet realizes alignment inside modality and between modalities. Since the two kinds of alignment operation are different, they will not disturb each other, but help to boost prediction accuracy mutually. Compared with the generating based cross-modality methods, such as cmGAN [6], AliGAN [37], XIV [16], Hi-CMD [5], SMCL [41]) and so on, DSCNet does not introduce intermediate steps like produce images of features. Thus, it does not require expensive computing costs in time and space and does not suffer from noise. Methods like MPANet [47] utilize both holistic and fine-grained spatial features. Although DSCNet also adopts fine-grained features, the difference is that we do not explore nuance information in spatial space. Besides, the feature alignment strategies are different. MPANet compares features between samples to align human parts.



(a) Infrared to visible images

(b) Visible to infrared images

Fig. 8: Illustration of person retrieval results. We separately visualize infrared and visible queries which correspond the top-10 re-rank images. Green indicates ”True”, and Red indicates ”False”.

TABLE III: Comparison with the state-of-the-arts on RegDB dataset. Rank-r accuracy (%) and mAP(%) are reported.

Settings		*Visible to Infrared*		*Infrared to Visible*	
Method	Venue	r=1	mAP	r=1	mAP
Zero-Pad [45]	ICCV17	17.75	18.90	16.63	17.82
HCML [59]	AAAI18	24.44	20.08	21.70	22.24
HSME [12]	AAAI19	50.85	47.00	50.15	46.16
AliGAN [37]	ICCV19	57.90	53.60	56.30	53.40
CMSP [44]	IJCV20	65.07	64.50	-	-
JSIA [35]	AAAI20	48.10	48.90	48.50	49.30
XIV [16]	AAAI20	62.21	60.18	-	-
HAT [42]	TIFS20	71.83	67.56	70.02	66.30
MSR [8]	TIP20	48.43	48.67	-	-
MACE [56]	TIP20	72.37	69.09	72.12	68.57
DDAG [43]	ECCV20	69.34	63.46	68.06	61.80
Hi-CMD [5]	CVPR20	70.93	66.04	-	-
AGW [2]	TPAMI21	70.05	66.37	70.49	65.90
CM-NAS [10]	CVPR21	84.54	80.32	82.57	78.31
VSD [66]	CVPR21	73.2	71.6	71.8	70.1
MCLNet [11]	ICCV21	80.31	73.07	75.93	69.49
SMCL[41]	ICCV21	83.93	**79.83**	83.05	**78.57**
DSCNet	**Ours**	**85.39**	77.30	**83.50**	75.19

However, we align the channel semantics inside a sample image.

From Tab. III, DSCNet also performs better than any other recent works in the RegDB dataset, with a 85.39% Rank-1 score and a 77.30 mAP score in *Visible to Infrared* modes, and a 83.50% Rank-1 score and a 75.19 mAP score in *Infreared to Visible* mode.Sample images in RegDB show similarities, such as in perspective, background, human gestures, and so on. Thus, it is easier to retrieve a person in this dataset. RegDB and SYSU-MM01 are quite different in sample representation variance, it is hard for a method to achieve the highest results in both of the datasets. The results on the RegDB dataset (Tab. III) also illustrate that the proposed DSCNet benefits better performance compared to other methods when the samples are very similar. It reveals that the semantic consistency learning strategy aligns features appropriately and is generous in the cross-modality person retrieval problem.

V. Conclusion and Future Work

In this paper, we have proposed a novel VI-ReID framework named DSCNet. It focuses on eliminating the modality discrepancy via reinforce the semantic consistency between the R/G/B channels and between the visible/infrared modalities. This approach ensures the extracted features become more identity-relevant and modality-invariant. DSCNet explores channel-level identity relevance and discrimination. Meanwhile, it significantly reveals that channel-level semantic consistency prominently influences the performance of this cross-modality retrieval task. It is also worth noting that our Dual-Semantic Consistency Learning structure can be further assembled with other advanced existing VI-ReID methods. Extensive experimental results validate the outstanding performance of DSCNet, as well as the effectiveness of all the components in this network.

References

[1] S. M. Ahmed, A. R. Lejbolle, R. Panda, and A. K. Roy-Chowdhury, “Camera on-boarding for person re-identification using hypothesis transfer learning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 144–12 153.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[3] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, “Abd-net: Attentive but diverse person re-identification,” in *IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019, pp. 8351–8361.

[4] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for rgb-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 587-597, 2021.

[5] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 677–683.

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 994–1003.

[8] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 579–590, 2019.

[9] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "Cm-nas: Rethinking cross-modality neural architectures for visible-infrared person re-identification," arxiv.org/abs/2101.08467, 2021.

[10] C. Fu, Y. Hu, X. Hu, H. Shi, T. Mei, and R. He, "Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification," in *IEEE Conference on International Conference on Computer Vision (ICCV)*, October 2021, pp. 11 823–11 832.

[11] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *IEEE Conference on International Conference on Computer Vision (ICCV)* , 2021, pp. 16 403–16 412.

[12] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 8385–8392.

[13] M. Jia, Y. Zhai, S. Lu, S. Ma, and J. Zhang, "A similarity inference metric for rgb-infrared cross-modality person re-identification," in *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20*, 2020.

[14] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3143–3152.

[15] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3422–3432, 2020.

[16] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 4610–4617.

[17] H. Li, G. Wu, and W. S. Zheng, "Combined depth space based architecture search for person re-identification," 2021.

[18] H. Li, M. Ye, and B. Du, "Weperson: Learning a generalized re-identification model from all-weather virtual data," in *ACM Multimedia*, 2021.

[19] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," 2021.

[20] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019, pp. 4976–4985.

[22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 0–0.

[23] J. Meng, W. S. Zheng, J. H. Lai, and L. Wang, "Deep graph metric learning for weakly supervised person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2021.

[24] H. Moon and P. J. Phillips, "Computational and performance aspects of pca-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[25] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 567–13 576.

[26] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[27] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *2021 IEEE International Conference on Computer Vision (ICCV)*, 2021.

[28] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification," in *ACM Multimedia*, 2020, pp. 2149–2158.

[29] C. X. Ren, B. H. Liang, and Z. Lei, "Domain adaptive person re-identification via camera style generation and label propagation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1290–1302, 2019.

[30] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 608–617.

[31] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," *European Conference on Computer Vision (ECCV)*, 2018.

[32] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," 2019.

[33] G. A. Wang, T. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[34] Y. Zhang, S. Zhao, Y. Kang, and J. Shen, "Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification," in *European Conference on Computer Vision (ECCV)*, 2022.

[35] G.-A. Wang, T. Z. Yang, J. Cheng, J. Chang, X. Liang, Z. Hou *et al.*, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 12 144–12 151.

[36] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6449–6458.

[37] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019, pp. 3623–3632.

[38] Y. Wang, Z. Chen, W. Feng, and W. Gang, "Person re-identification with cascaded pairwise convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Z. Wang, Z. Wang, Y. Zheng, Y. Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2020.

[41] Z. Wei, X. Yang, N. Wang, and G. X., "Syncretic modality collaborative learning for visible infrared person re-identification," in *2021 IEEE International Conference on Computer Vision (ICCV)*, 2021.

[42] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Transactions on Information Forensics and Security*, 2020.

[43] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," European Conference on Computer Vision (ECCV), pp. 229-247, 2020.

[44] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "Rgb-ir person re-identification by cross-modality similarity preservation," *International Journal of Computer Vision*, pp. 1–21, 2020.

[45] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *IEEE Conference on International Conference on Computer Vision (ICCV)*, 2017, pp. 5380–5389.

[46] D. Wu, M. Ye, G. Lin, X. Gao, and J. Shen, "Person re-identification by context-aware part attention and multi-head collaborative learning," *IEEE Transactions on Information Forensics and Security*, 2021.

[47] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji, "Discover cross-modality nuances for visible-infrared person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4330–4339.

[48] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order nonlocal attention networks for person re-identification," in *IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019, pp. 3760–3769.

[49] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," 2021.

[50] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Transactions on Information Forensics and Security*, vol. PP, no. 99, pp. 1–1, 2020.

[51] S. Yu, S. Li, D. Chen, R. Zhao, J. Yan, and Y. Qiao, "Cocas: A large-scale clothes changing person dataset for re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3400–3409.

[52] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," 2021.

[53] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2019.

[54] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3186–3195.

[55] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8514–8522.

[56] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing* , vol. 29, pp. 9387–9399, 2020.

[57] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5735–5744.

[58] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM TOMM*, vol. 14, no. 1, pp. 1–20, 2017.

[59] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7501–7508.

[60] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5157–5166.

[61] X. Zhu, X. Y. Jing, X. You, W. Zuo, S. Shan, and W. S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2017.

[62] G. Zhang, Y. Chen, W. Lin, A. Chandran, and X. Jing, "Low resolution information also matters: Learning multi-resolution representations for person re-identification," *arXiv preprint arXiv:2105.12684*, 2021.

[63] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8913–8925, 2021.

[64] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *European conference on computer vision*. Springer, 2020, pp. 275–292.

[65] X. Wei, D. Li, X. Hong, W. Ke, and Y. Gong, "Co-attentive lifting for infrared-visible person re-identification," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1028–1037.

[66] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1522–1531.

[67] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, and Y. Qu, "Dual mutual learning for cross-modality person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[68] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "Fmcnet: Feature-level modality compensation for visible-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7349–7358.

[69] J. Sun, Y. Li, H. Chen, Y. Peng, X. Zhu, and J. Zhu, "Visible-infrared cross-modality person re-identification based on whole-individual training," *Neurocomputing*, vol. 440, pp. 1–11, 2021.

[70] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, "Global-local multiple granularity learning for cross-modality visible-infrared person reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[71] Y. Liu, W. Zhou, M. Xi, S. Shen, and H. Li, "Multi-modal context propagation for person re-identification with wireless positioning," *IEEE Transactions on Multimedia*, 2021.

[72] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and C. Petrioli, "Person re-identification through wi-fi extracted radio biometric signatures," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1145–1158, 2022.

[73] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, "Learning longterm representations for person re-identification using radio signals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 699–10 709.