Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification

Mang Ye[1], Jianbing Shen[1⋆], David J. Crandall[2], Ling Shao[1,4], and Jiebo Luo[3]

[1] Inception Institute of Artificial Intelligence, UAE
[2] Indiana University, USA [3]University of Rochester, USA
[4] Mohamed bin Zayed University of Artificial Intelligence, UAE
https://github.com/mangye16/DDAG

Abstract. Visible-infrared person re-identification (VI-ReID) is a challenging cross-modality pedestrian retrieval problem. Due to the large intra-class variations and cross-modality discrepancy with large amount of sample noise, it is difficult to learn discriminative part features. Existing VI-ReID methods instead tend to learn global representations, which have limited discriminability and weak robustness to noisy images. In this paper, we propose a novel dynamic dual-attentive aggregation (DDAG) learning method by mining both intra-modality part-level and cross-modality graph-level contextual cues for VI-ReID. We propose an intra-modality weighted-part attention module to extract discriminative part-aggregated features, by imposing the domain knowledge on the part relationship mining. To enhance robustness against noisy samples, we introduce cross-modality graph structured attention to reinforce the representation with the contextual relations across the two modalities. We also develop a parameter-free dynamic dual aggregation learning strategy to adaptively integrate the two components in a progressive joint training manner. Extensive experiments demonstrate that DDAG outperforms the state-of-the-art methods under various settings.

Keywords: Person Re-identification, Graph Attention, Cross-modality

1 Introduction

Person re-identification (Re-ID) techniques [59,68] have achieved human-level performance with part-level deep feature learning [4,40,67]. However, most of these techniques consider images of people collected by visible-spectrum cameras in the daytime, and thus are not applicable to night-time applications. Infrared cameras can be used to collect imagery in low light conditions [50], but matching this imagery to visible-spectrum images is a significant challenge.

Cross-modality visible-infrared person re-identification (VI-ReID) [50,58] aims to solve this problem by matching images of people captured by visible and infrared (including near- [50] and far-infrared (thermal) [29]) cameras. VI-ReID is

[⋆] Corresponding Author: *Jianbing Shen*

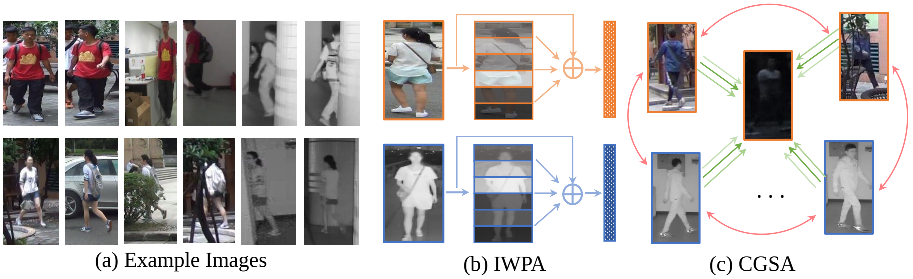

Fig. 1. Idea Illustration: (a) Example images from SYSU-MM01 dataset [50] with high sample noise due to data annotation/collection difficulty. Main components: (b) intra-modality weighted-part aggregation (IWPA), it learns discriminative part-aggregated features by mining the contextual part information within each modality. (c) cross-modality graph structured attention (CGSA), it enhances the representation by incorporating the neighborhood information from the two modalities.

challenging due to large visual differences between the two modalities and changing camera environments, leading to large intra- and cross-modality variations. Moreover, due to difficulties in data collection and annotation, VI-ReID usually suffers from high sample noise caused by inaccurate person detection results, *eg* extreme background clutter, as shown in Fig. 1 (a). Related cross-modality matching studies have been extensively conducted in visible near-infrared (VIS-NIR) face recognition [28,52]. However, the visual differences between images of people are much greater than those between face images, so those methods are not applicable for VI-ReID [50].

These challenges make it difficult to reliably learn discriminative part-level features using state-of-the-art single-modality Re-ID systems [40,45,55,67]. As a compromise, existing VI-ReID methods mainly focus on learning multi-modal sharable global features, either via one- [7,49,50] or two-stream networks [9,58]. Some work also integrates modality discriminant supervision [7,9] or GAN generated images [44,49] to handle the modality discrepancy. However, global feature learning methods are sensitive to background clutter and can not explicitly handle the modality discrepancy. In addition, part-based feature-learning methods [40,45,66,67] for single-modality Re-ID are typically incapable of capturing reliable part features under a large cross-domain gap [50]. Moreover, the learning is easily contaminated by noisy samples and destabilized when the appearance discrepancy is large across the two modalities. All of these challenges result in less discriminative cross-modality features and unstable training.

To address the above limitations, we propose a novel dynamic dual-attentive aggregation (DDAG) learning method with a two-stream network. DDAG includes two main components, as shown in Fig. 1: an intra-modality weighted-part aggregation (IWPA) and a cross-modality graph structured attention (CGSA). Our main idea is to mine contextual cues at both an intra-modality part-level and cross-modality graph-level, to enhance feature representation learning. I-WPA aims to learn discriminative part-aggregated features by simultaneously mining the contextual relations among the body parts within each modality and imposing the domain knowledge to handle the modality discrepancy. Our design is computationally efficient because we learn the modality-specific part-level attention rather than pixel-level attention [47,65], and it also results in stronger robustness against background clutter. We further develop a residual BatchNorm connection with weighted-part aggregation to reduce the impact of noisy body parts, and adpatively handle the part discrepancy in the aggregated features.

CGSA focuses on learning an enhanced node feature representation by incorporating the relationship between the person images across the two modalities. We eliminate the negative impact of samples with large variations by exploiting the contextual information in the cross-modality graph, assigning adaptive weights to both intra- and cross-modality neighbors with a mutli-head attentive graph scheme [42]. This strategy also reduces the modality discrepancy and smooths the training process. In addition, we introduce a parameter-free dynamic dual aggregation learning strategy to dynamically aggregate the two attentive modules in a mutli-task end-to-end learning manner, which enables complex dual-attentive network to converge stably, while simultaneously reinforcing each attentive component. Our main contributions are as follows:

- We propose a novel dynamic dual-attentive aggregation learning method to mine contextual information at both intra-modality part and cross-modality graph levels to facilitate feature learning for VI-ReID.
- We design an intra-modality weighted-part attention module to learn discriminative part-aggregated representation, adaptively assigning the weights of different body parts.
- We introduce a cross-modality graph structured attention scheme to enhance feature representations by mining the graphical relations between the person images across the two modalities, which smooths the training process and reduces the modality gap.
- We establish a new baseline on two VI-ReID datasets, outperforming the state-of-the-art by a large margin.

2 Related Work

Single-Modality Person Re-ID aims to match person images from visible cameras [18]. Existing works have achieved human-level performance on the widely-used datasets with end-to-end deep learning [1,14,15,17,39,54], either by global [17,64] or part-level feature learning [40,39,67]. However, these approaches are usually unable to handle the ambiguous modality discrepancy in VI-ReID [50], which limits their applicability in night-time surveillance scenarios.

Cross-Modality Person Re-ID addresses person re-identification across different types of images, such as between visible-spectrum and infrared [49,50,57], varying illuminations [62] or even between images and non-visual data like text descriptions [5,21]. For visible-Infrared-ReID (VI-ReID), Wu *et al.* [50] introduced a zero-padding strategy with a one-stream network for cross-modality feature representation learning. A two-stream network with dual-constrained top-ranking loss was proposed in [58] to handle both the intra- and cross-modality variations. In addition, Dai *et al.* [7] proposed an adversarial training framework with the triplet loss to jointly discriminate the identity and modality. Recently, Wang *et al.* [49] presented a dual-level discrepancy method with GAN to handle the modality difference at various levels. Similar technique was also adopted in [44]. Two modality-specific [9] and modality-aware learning [56] methods were proposed to handle the modality discrepancy at the classifier level. Meanwhile, other papers have investigated a better loss function [2,23] to handle the modality gap. However, these methods usually focus on learning global feature representations, which ignore the underlying relationship between different body parts and neighborhoods across two modalities.

Contemporaneously, some recent methods investigate the modality-aware collaborative ensemble learning [56] or grayscale augmented tri-modal learning [60]. An intermediate X-modality is designed in [19] to address the modality discrepancy. A powerful baseline with non-local attention is presented in [59].

Visible Near-Infrared Face Recognition addresses the cross-modality face recognition problem, which is also closely related to VI-ReID [13,28,32,46,52,30]. Early research mainly focused on learning modality-aware metrics [31] or dictionaries [16]. With the emergence of deep neural networks, most methods now focus on learning multi-modal sharable features [52], cross-modality matching models [34] or disentangled representations [51]. However, the modality difference of VI-ReID is much greater than that of face recognition due to the different camera environments and large visual appearances change, which limits the applicability of their methods to the VI-ReID [57,48].

Attention Mechanisms have been widely used in various applications to enhance the feature representation [37,42,53,3]. For person Re-ID, attention is used to combine the spatial-temporal information from different video frames [8,10,20,24]. Some work [22,26,41] has also investigated using multi-scale or different convolutional channels to capture the pixel-level/small-region-level attentions [35,36]. However, they are usually unstable for optimization in VI-ReID due to the large cross-modality discrepancy and noise.

Our part-attention module is also closely related to non-local networks [47,65]. However, the pixel-level design of these models is sensitive and inefficient for handling the noise encountered in VI-ReID task. In comparison, we design a learnable weighted part-level attention with a BatchNorm residual connection.

3 Proposed Method

Fig. 2 provides an overview of our proposed dynamic dual-attentive aggregation learning (DDAG) method. DDAG is developed on top of a two-stream network (§3.1), and contains an intra-modality weighted-part attention for discriminative part-aggregated features learning (§3.2) and a cross-modality graph structured attention for shared global feature learning (§3.3). Finally, we propose a parameter-free dynamic dual aggregation learning strategy to adaptively aggregate the two components for end-to-end joint training (§3.4).

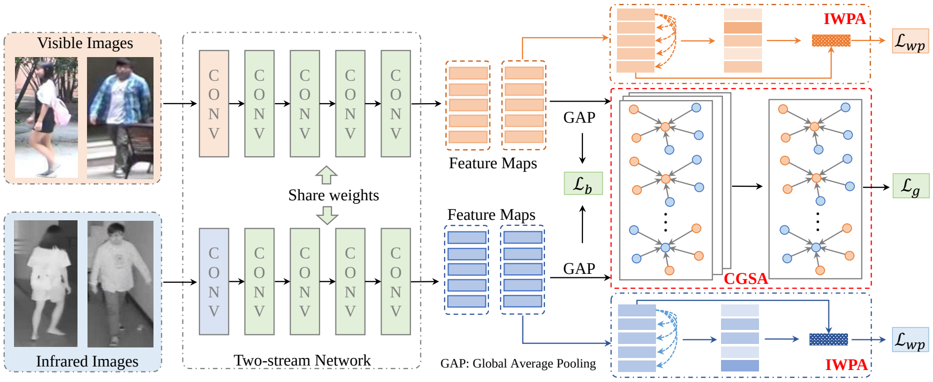

Fig. 2. The proposed DDAG learning framework for VI-ReID. **IWPA**: Intra-modality Weighted-Part Aggregation for discriminative part-aggregated features learning by mining the contextual part relations within each modality. **CGSA**: Cross-modality Graph Structured Attention for global feature learning by utilizing the neighborhood structural relations across two modalities. We further introduce a parameter-free dynamic dual aggregation learning strategy to adaptively aggregate two components.

3.1 Baseline Cross-Modality Re-ID

We first present our baseline cross-modality Re-ID model with a two-stream network for incorporating two different modalities. To handle the different properties of the two modalities, the network parameters of the first convolutional block[3] in each stream are different in order to capture modality-specific low-level feature patterns. Meanwhile, the network parameters of the deep convolutional blocks are shared for two modalities in order to learn modality-sharable middle-level feature representations. After the convolutional layers with adaptive pooling, a shared batch normalization layer is added to learn the shared feature embedding. Compared with the two-stream structures in [11,25,58,56], our design captures more discriminative features by mining sharable information in middle-level convolutional blocks rather than high-level embedding layers.

To learn discriminative features, we combine the identity loss $\mathcal{L}_{id}$ and online hard-mining triplet loss $\mathcal{L}_{tri}$ [61] as our baseline learning objective $\mathcal{L}_b$,

$$\mathcal{L}_b = \mathcal{L}_{id} + \mathcal{L}_{tri}. \tag{1}$$

The identity loss $\mathcal{L}_{id}$ encourages an identity-invariant feature representation. The triplet loss $\mathcal{L}_{tri}$ optimizes the triplet-wise relationships among different person images across the two modalities.

3.2 Intra-modality Weighted-Part Aggregation

As an alternative to the global feature learning in existing VI-ReID methods [7,49,50], this subsection presents a novel part-agggregated feature learning

[3] We adopt ResNet50 as the backbone network, following [44,49,58].

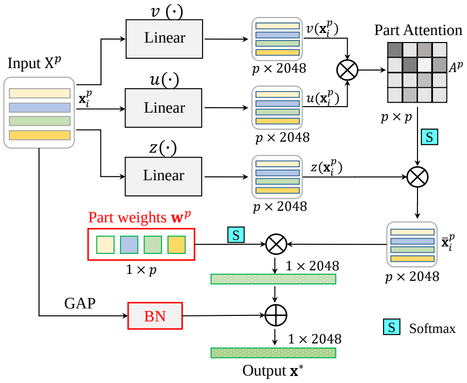

Fig. 3. Illustration of the proposed IWPA module, which mines the part-level relationships to learn the weighted-part aggregation with a residual BatchNorm connection.

method for VI-ReID, namely intra-modality weighted-part aggregation (IWPA, as shown in Fig. 3). IWPA mines the contextual information in local parts to formulate an enhanced part-aggregated representation to address the complex challenges. It first learns the within-modality part attention with a modified non-local module, and then uses a learnable weighted-part aggregation strategy with residual BatchNorm (RBN) to stabilize and reinforce the training process.

Part Attention. The input of our IWPA module is the extracted feature maps from the last residual block of the network, from which we extract the attention-enhanced part features. We denote the output feature maps of the last convolutional block as $\{X = \mathbf{x}_k \in \mathbb{R}^{C\times H\times W}\}_{k=1}^{K}$, where C represents the channel dimension ($C = 2048$ in our experiments), H and W represent the feature map size, and K represents the batch size. To obtain the part features, the feature maps are directly divided into p non-overlapping parts with a region pooling strategy. The part features of each input image are then represented by $X^p = \{\mathbf{x}_i^p \in \mathbb{R}^{C\times 1}\}_{i=1}^{p}$. Similar to [47], we feed each part into three 1×1 convolutional layers $u(\cdot)$, $v(\cdot)$ and $z(\cdot)$. The intra-modality part-based non-local attention $\alpha_{i,j}^p \in [0,1]^{p\times p}$ is then

$$\alpha_{i,j}^p = \frac{f(\mathbf{x}_i^p, \mathbf{x}_j^p)}{\sum_{\forall j} f(\mathbf{x}_i^p, \mathbf{x}_j^p)}, \tag{2}$$

where $f(\mathbf{x}_i^p, \mathbf{x}_j^p)$ represents the pairwise similarity between two part features. To enhance the discriminability, an exponential function is added to magnify the relationship, which enlarges the part attention discrepancy [63]. It is formulated by

$$f(\mathbf{x}_i^p, \mathbf{x}_j^p) = \exp(u(\mathbf{x}_i^p)^T v(\mathbf{x}_j^p)), \tag{3}$$

where $u(\mathbf{x}_i^p) = W_u\mathbf{x}_i^p$ and $v(\mathbf{x}_j^p) = W_v\mathbf{x}_j^p$ are two embeddings with convolutional operations $u(\cdot)$ and $v(\cdot)$. W_u and W_v are the corresponding weight parameters in u and v. With the exponential function, our attention calculation can be treated as a normalization with a softmax function. Note that our attention map is $p \times p$ to capture the part relationships, which is much smaller than that of pixel-level attention $HW \times HW$ in [47,65], making it more efficient. Meanwhile, the part relation is robust against noisy regions and local clutters in the person images.

With the learned part attention, attention-enhanced part features are then represented by the inner product of the embedded part features $z(\mathbf{x}_i^p)$ and the calculated attention A^p, which is formulated by

$$\bar{\mathbf{x}}_i^p = \mathbf{a}_i^p * z(\mathbf{x}_i^p), \tag{4}$$

where $\mathbf{a}_i^p \in A^p = \{\alpha_{i,j}^p\}^{p \times p}$ is the calculated part attention map. Therefore, the refined part features consider the relationship between different body parts. However, the simple average pooling or concatenation of part features is not powerful enough for fine-grained person Re-ID task, and may cause noisy parts accumulation. Meanwhile, it is inefficient to train multiple part-level classifiers, as in [40,56]. To address these issues, we design a residual BatchNorm (RBN) weighted-part aggregation strategy.

Residual BatchNorm Weighted-part Aggregation. This idea consists of two main parts: First, we use a residual BatchNorm connection of the original input feature map $\mathbf{x}^o$ after average pooling, and the residual learning strategy enables very deep neural networks to be trained and stabilizes the training process. Second, we use a learnable weighted combination of attention-enhanced part features to formulate a discriminative part-aggregated feature representation. In summary, it is formulated by

$$\mathbf{x}^* = \mathrm{BN}(\mathbf{x}^o) + \sum\nolimits_{i=1}^{p} w_i^p \bar{\mathbf{x}}_i^p, \tag{5}$$

where $\mathbf{x}^o \in \mathbb{R}^{C \times 1}$ represents the global adaptive pooling output of the input feature map X^p. BN indicates the batch normalization operation, and $\mathbf{w}^p = \{w_i^p\}_{i=1}^p$ represents a learnable weight vector of different parts to handle the modality discrepancy. Our design has three primary advantages: (1) it avoids multiple part-level classifier learning [40], making it computationally efficient for both training and testing, and it is more robust to background clutter compared to the pixel-level attention techniques [22,47]; (2) it enhances the discrimination power by adaptively aggregating attentive part features in the final feature representation; and (3) the residual BatchNorm (RBN) connection performs much better than the widely-used general residual connection with identity mapping [12,65] (as verified in §4.2), stabilizing the training process and enhancing the representational power for the cross-modality Re-ID under abundant noise. We use $\mathbf{x}^*$ as the representation of an input sample in the testing phase.

3.3 Cross-modality Graph Structured Attention

Another major challenge is that VI-ReID datasets often contain many incorrectly annotated images or image pairs with large visual differences across the two modalities (as shown in Fig. 1), making it difficult to mine the discriminative local part features and damaging the optimization process. In this subsection, we present our cross-modality graph structured attention, which incorporates the structural relations across two modalities to reinforce the feature representations. The main idea is that the feature representations of person images belonging to the same identity across the two modalities are mutually beneficial.

Graph Construction. At each training step, we adopt an identity-balanced sampling strategy for training [58]. Specifically, for each of n different randomly-selected identities, m visible and m infrared images are randomly sampled, resulting in $K = 2mn$ images in each training batch. We formulate an undirected graph $\mathcal{G}$ with a normalized adjacency matrix,

$$A^g = A^g_0 + \mathbb{I}_K, \tag{6}$$

where $A^g_0(i,j) = l_i * l_j$ (l_i and l_j are the corresponding one-hot labels of two graph nodes). $\mathbb{I}_K$ is an identity matrix, indicating that each node is connected to itself. The graph construction is efficiently computed by matrix multiplication between the one-hot labels in each training batch.

Graph Attention. This measures the importance of a node i to another node j within the graph, across two modalities. We denote the input node features by $X^o = \{\mathbf{x}^o_k \in \mathbb{R}^{C\times 1}\}_{k=1}^{K}$, which are outputs of the pooling layer. The graph attention coefficients $\alpha^g_{i,j} \in [0,1]^{K\times K}$ are then computed by

$$\alpha^g_{i,j} = \frac{\exp(\mathit{\Gamma}(\lceil h(\mathbf{x}^o_i), h(\mathbf{x}^o_j)\rfloor \cdot \mathbf{w}^g))}{\sum_{\forall A^g(i,k)>0} \exp(\mathit{\Gamma}(\lceil h(\mathbf{x}^o_i), h(\mathbf{x}^o_k)\rfloor \cdot \mathbf{w}^g))}, \tag{7}$$

where $\mathit{\Gamma}(\cdot)$ represents the LeakyRelu operation. $\lceil , \rfloor$ is the concatenation operation. $h(\cdot)$ is a transformation matrix to reduce the input node feature dimension C to d, and d is set to 256 in our experiments. $\mathbf{w}^g \in \mathbb{R}^{2d\times 1}$ represents a learnable weighting vector that measures the importance of different feature dimensions in the concatenated features, similar to [43]. Note that our design fully utilizes relations between all the images across two modalities, reinforcing the representation using context information of the same identity.

To enhance the discriminability and stabilize the graph attention learning, we employ a multi-head attention technique [38] by learning multiple $h^l(\cdot)$ and $\mathbf{w}^{l,g}$ ($l = 1, 2 \cdots, L$, L is the total number of heads) with the same structure and optimizing them separately. After concatenating the outputs of multiple heads, the graph structured attention-enhanced feature is then represented by

$$\mathbf{x}^g_i = \phi\lceil \sum_{\forall A^g(i,k)>0} \alpha^{g,l}_{i,j} \cdot h^l(\mathbf{x}^o_j)\rfloor_{l=1}^{L}, \tag{8}$$

and $\mathbf{x}^g_i$ is then robust to outlier samples, where ϕ is the ELU activation function. To guide the cross-modality graph structured attention learning, we introduce another graph attention layer with a one-head structure, where the final output node features are represented by $X^{g'} = \{\mathbf{x}^{g'}_i\}_{k=1}^{K}$. We adopt the negative log-likelihood (NLL) loss function for the graph attention learning, formulated by

$$\mathcal{L}_g = -\sum_{i}^{K} \log(\mathrm{softmax}(\mathbf{x}^{g'}_i)). \tag{9}$$

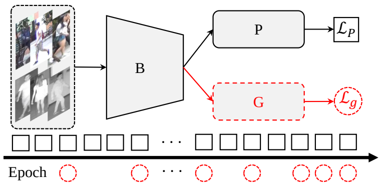

Fig. 4. Illustration of parameter-free dynamic dual aggregation learning. We decompose the overall training framework into two parts: instance-level part-aggregated feature learning $\mathcal{L}_P$ and graph-level global feature learning $\mathcal{L}_g$. We treat $\mathcal{L}_P$ as the dominant loss and progressively add $\mathcal{L}_g$ in the overall training process.

3.4 Dynamic Dual Aggregation Learning

Incorporating the above proposed intra-modality weighted-part attention and cross-modality graph-structured attention into an end-to-end joint learning framework is highly challenging. This is primarily because the two components focus on different learning objectives with very deep network structures, and directly combining them easily will result in *gradient explosion* problem after several steps. Moreover, the features from the same identity across two modalities are quite different in VI-ReID due to the large cross-modality variations, as demonstrated in Fig. 1. Therefore, the graph-structured attention would be unstable due to the large feature difference across the two modalities at the early stage.

To address the above issues, we introduce a dynamic dual aggregation learning strategy to adaptively integrate the above introduced two components. Specifically, we decompose the overall framework into two different tasks, instance-level part-aggregated feature learning $\mathcal{L}_P$ and graph-level global feature learning $\mathcal{L}_g$. The instance-level part-aggregated feature learning $\mathcal{L}_P$ is a combination of the baseline learning objective $\mathcal{L}_b$ and the intra-modality weighted-part attention loss $\mathcal{L}_{wp}$, represented by

$$\mathcal{L}_P = \mathcal{L}_b \underbrace{- \frac{1}{K}\sum\nolimits_{i=1}^{K} y_i \log(p(y_i|\mathbf{x}_i^*))}_{part\ attention\ loss\ \mathcal{L}_{wp}}, \tag{10}$$

where $p(y_i|\mathbf{x}_i^*)$ represents the probability of $\mathbf{x}_i^*$ being correctly classified into the groundtruth label y_i. The second term represents the instance-level part-aggregated feature learning with weighted-part attention within each modality. It is formulated by the identity loss on top of the aggregated part feature $\mathbf{x}^*$.

Dynamic Dual Aggregation Learning. Motivated by multi-task learning [6], our basic idea is that the instance-level part-aggregated feature learning $\mathcal{L}_P$ acts as the dominant loss, and then we progressively add the graph-level global feature learning loss $\mathcal{L}_g$ for optimization. The main reason for doing this is that it is easier to learn an instance-level feature representation with $\mathcal{L}_P$ at an early stage. With a better learned network, the graph-level global feature learning optimizes the features using the relationship between the person images across the two modalities, denoted by

$$\mathcal{L}^t = \mathcal{L}_P^t + \frac{1}{1+\mathbb{E}(\mathcal{L}_P^{t-1})}\mathcal{L}_g^t, \tag{11}$$

where t is the epoch number, and $\mathbb{E}(\mathcal{L}_P^{t-1})$ represents the average loss value in the previous epoch. In this dynamic updating framework (as shown in Fig. 4), the graph-level global loss $\mathcal{L}_g$ is progressively added into the overall learning process. This strategy shares a similar spirit to the gradient normalization in multi-task learning [6], but it does not introduce any additional hyper-parameter tuning.

When we optimize $\mathcal{L}_P$, the parameters of the identity classifier in the weighted-part attention loss $\mathcal{L}_{wp}$ are the same as those for the identity classifier in $\mathcal{L}_b$. Our motivation here is that this setting can guarantee that instance-level part-aggregated feature learning is directly performed on the part-aggregated features rather than additional classifiers, ensuring the discriminability of the learned features. Meanwhile, it avoids additional network parameters training.

4 Experimental Results

4.1 Experimental Settings

We use two publicly available VI-ReID datasets (SYSU-MM01 [50] and RegDB [29]) for the experiments. The rank-k matching accuracy and mean Average Precision (mAP) are used as evaluation metrics, following [50].

SYSU-MM01 [50] is a large-scale dataset collected by four RGB and two near-infrared cameras. The major challenge is that person images are captured in both indoor and outdoor environments. In total, the training set contains 22,258 visible and 11,909 near-infrared images of 395 identities. It contains two different testing settings, *all-search* and *indoor-search* mode. The query set contains 3,803 images of 96 identities captured from near-infrared cameras. The gallery set contains the images captured by all four RGB cameras in the *all-search* mode, while the *indoor-search* mode contains images of two indoor RGB cameras. Details on the experimental settings can be found in [50].

RegDB [29] is collected by a dual-camera system, including one visible and one far-infrared camera. In total, this dataset contains 412 person identities, each of which has 10 visible and 10 far-infrared images. Following [57], we randomly select 206 identities for training and the remaining 206 identities for testing. Thus the testing set contains 2,060 visible and 2,060 far-infrared images. We evaluate both visible-to-infrared and infrared-to-visible query settings. The performance is averaged over ten trials on random training/testing splits [49,57].

Implementation details. Our proposed method is implemented in PyTorch. Following existing VI-ReID works, ResNet50 [12] is adopted as our backbone network for fair comparison, following [59]. The first residual block is specific for each modality while the other four blocks are shared. The stride of the last convolutional block is set to 1, in order to obtain a fine-grained feature map. We initialize the convolutional blocks with the pre-trained ImageNet parameters, as

Table 1. Evaluation of each component on the large-scale SYSU-MM01 dataset. "B" represents the baseline results with a two-stream network trained by $\mathcal{L}_b$. "P" denotes the intra-modality weighted-part attention. "G" indicates the cross-modality graph structured attention. Dynamic dual-learning is adopted when aggregating two components. Rank at r accuracy(%) and mAP (%) are reported.

Datasets	*All Search*					*Indoor Search*				
Methods	$r=1$	$r=5$	$r=10$	$r=20$	mAP	$r=1$	$r=5$	$r=10$	$r=20$	mAP
B	48.18	75.81	85.73	93.52	47.64	49.52	78.86	88.70	95.27	58.12
$B+P$	53.69	81.16	88.38	94.56	51.37	58.08	84.91	92.37	97.26	65.07
$B+G$	50.75	78.43	86.71	93.62	49.73	52.90	83.50	92.65	97.75	62.26
$B+P+G$	54.75	82.31	90.39	95.81	53.02	61.02	87.13	94.06	98.41	67.98

done in [58]. All the input images are firstly resized to 288×144. We adopt random cropping with zero-padding and horizontal flipping for data augmentation. SGD optimizer is adopted for optimization, and the momentum parameter is set to 0.9. We set the initial learning rate to 0.1 with a warm-up strategy [27]. The learning rate decays by 0.1 at the 30th epoch and 0.01 at the 50th epoch, with a total of 80 training epochs. By default, we randomly select 8 identities, and then randomly select 4 visible and 4 infrared images to formulate a training batch. We set $p=3$ in Eq. 5, $L=4$ in Eq. 8.

4.2 Ablation Study

Evaluation of Each Component. This subsection evaluates the effectiveness of each component on the SYSU-MM01 dataset under both *all-search* and *indoor search* modes. Specifically, "B" represents the baseline results with a two-stream network trained by $\mathcal{L}_b$. "P" denotes the intra-modality weighted-part aggregation. "G" indicates the cross-modality graph structured attention.

We make several observations through the results shown in Table 1. *1) Effectiveness of baseline*: Using shared convolutional blocks, we achieve better performance than the two-stream network in [9,25,56,58]. Meanwhile, some training tricks taken from single-modality Re-ID [67] also contributes to this super baseline. *2) Effectiveness of P*: the intra-modality weighted-part aggregation significantly improves the performance. This experiment demonstrates that learning part-level weighted-attention features is beneficial for cross-modality Re-ID. *3) Effectiveness of G*: When we include the cross-modality graph structured attention ($B+G$), performance is improved by using the relationship between the person images across two modalities to reduce the modality discrepancy. *4) Effectiveness of dual-aggregation*: When aggregating two attention modules with the dynamic dual aggregation strategy, the performance is further improved, demonstrating that these two attentions are mutually beneficial to each other.

Why Weighted Part Attention with RBN? We next compare different part attention designs on the SYSU-MM01 dataset under the *all-search* mode. The results are shown in Table 2 and we make several observations. (1) *Effectiveness of weighted scheme.* We compare the *weighted* part features with average/concatenation schemes (termed as *weighted*, *avg* and *concat* in Table 2). We

Table 2. Evaluation of re-weighted part attention with different designs on the SYSU-MM01 dataset (*all-search mode*). Rank at r accuracy (%) and mAP (%) are reported. (*Setting: Baseline + Part attention.*)

Method	Res.	$r = 1$	$r = 10$	$r = 20$	mAP
B	N/A	48.18	85.73	93.52	47.64
avg	Res	48.34	86.03	93.72	48.43
concat	Res	50.34	86.43	94.19	49.77
weight	Res	51.06	86.78	94.39	49.92
weight	RBN	53.69	88.38	94.56	51.37

Table 3. Evaluation of graph attention on the SYSU-MM01 dataset (*all-search mode*). N_g represents the number of images selected for graph construction. Rank at r accuracy (%) and mAP (%) are reported. (*Setting: Baseline + Graph attention.*)

N_g	0	1	2	4	8
Rank-1	48.18	49.26	49.85	50.45	50.75
mAP	47.64	48.42	49.12	49.46	49.73

observe that the proposed learnable weighted-part scheme performs consistently better than its two counterparts. Another benefit of the weighted aggregation is that the feature dimension of final representation is much smaller than the concatenation strategy in [40], which is more suitable for real applications with resource-demanding scenarios. (2) *Effectiveness of residual BN (RBN) scheme.* We compare the general residual connection with the residual BN connection. Results demonstrate that RBN performs significantly better than the general residual connection. This suggests that the BN operation enhances the predictive and stable behavior of the training process [33], which is more suitable for VI-ReID with abundant noise. Note that the performance significantly drops without the residual connection.

Why Graph Structured Attention? We now evaluate the effect of different numbers (N_g) of selected images for graph attention calculation. The results are shown in Table 3. A larger N_g means that more neighbor images from the same identity are considered and the relationship is more reliable. Thus the accuracy is consistently improved with increasing N_g, demonstrating that the graph structured attention can largely reduce the modality discrepancy. Moreover, the infrared images capture less information than the visible images, but with much more noise. Mining the relation across two modalities, especially from the visible images, is thus beneficial for the cross-modality feature learning. The graph attention might also be applied in single-modality person re-identification.

Parameter Analysis We evaluate the effect of different body parts p and different numbers of graph attention heads L on the large-scale SYSU-MM01 dataset, under the challenging *all-search* mode. The results are shown in Fig. 5.

(1) As shown in the left figure, a larger p captures more fine-grained part features and improves the performance. However, when p is too large, the performance drops since small body parts cannot contain sufficient information for part attention learning. (2) As demonstrated in the right figure, a large L provides more reliable relationship mining, and thus consistently improves the

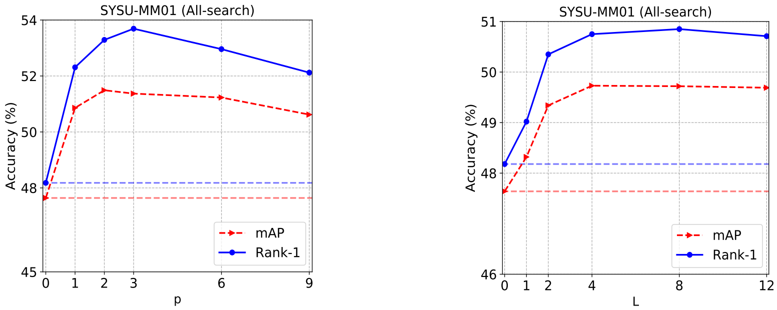

Fig. 5. Evaluation of different body parts p in Eq. 4 (left) and different numbers of graph attention heads L in Eq. 8 (right) on SYSU-MM01 dataset, under the challenging *all-search* mode. Rank-1 matching accuracy (%) and mAP (%) are reported.

Table 4. Comparison with the state-of-the-arts on SYSU-MM01 dataset on two different settings. Rank at r 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
One-stream [50]	ICCV17	12.04	49.68	66.74	13.67	16.94	63.55	82.10	22.95
Two-stream [50]	ICCV17	11.65	47.99	65.50	12.85	15.60	61.18	81.02	21.49
Zero-Pad [50]	ICCV17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
TONE [57]	AAAI18	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
HCML [57]	AAAI18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [7]	IJCAI18	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
BDTR [58]	IJCAI18	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
eBDTR [58]	TIFS19	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
HSME [11]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-
D^{2}RL [49]	CVPR19	28.9	70.6	82.4	29.2	-	-	-	-
MAC [56]	MM19	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
MSR [9]	TIP19	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN [44]	ICCV19	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
HPILN [23]	arXiv19	41.36	84.78	94.31	42.95	45.77	91.82	**98.46**	56.52
LZM [2]	arXiv19	45.00	89.06	-	45.94	49.66	92.47	-	59.81
AGW [59]	arXiv20	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
Xmodal [19]	AAAI20	49.92	89.79	**95.96**	50.73	-	-	-	-
DDAG (Ours)	-	**54.75**	**90.39**	**95.81**	**53.02**	**61.02**	**94.06**	98.41	**67.98**

performance. However, it also greatly increases the difficulty of optimization, which results in a slightly decreased performance with a too large L. Thus, we select $p = 3$ and $L = 4$ in all our experiments.

4.3 Comparison with State-of-the-Art Methods

This subsection presents a comparison with the current state-of-the-arts on two different datasets. The comparison includes eBDTR [58], D^{2}RL [49], MAC [56], MSR [9], AlignGAN [44] and Xmodal [19]. Note that AlignGAN [44] represents the state-of-the-art by aligning the features in both the feature level and pixel level with generated images. Xmodal generates an intermediate modality to bridge the gap. We also compare with several arXiv papers, including EDFL [25], HPILN [23], LZM [2] and AGW [59]. The results on two public datasets are shown in Tables 4 and 5.

Table 5. Comparison with the state-of-the-art methods on RegDB dataset on visible-infrared and infrared-visible settings. Rank at r accuracy (%) and mAP (%) are reported.

Setting	*Visible to Infrared*				*Infrared to Visible*			
Method	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
HCML [57]	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
Zero-Pad [50]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
BDTR [58]	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
eBDTR [58]	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
HSME [11]	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
D^{2}RL [49]	43.4	66.1	76.3	44.1	-	-	-	-
MAC [56]	36.43	62.36	71.63	37.03	36.20	61.68	70.99	36.63
MSR [9]	48.43	70.32	79.95	48.67	-	-	-	-
EDFL [25]	52.58	72.10	81.47	52.98	51.89	72.09	81.04	52.13
AlignGAN [44]	57.9	-	-	53.6	56.3	-	-	53.4
Xmodal [19]	62.21	83.13	91.72	60.18	-	-	-	-
DDAG (Ours)	**69.34**	**86.19**	**91.49**	**63.46**	**68.06**	**85.15**	**90.31**	**61.80**

The following observations can be made: 1) Methods with two-stream networks (*EDFL* [25], *MSR* [9], *LZM* [2] and our proposed *DDAG*) generally perform better than the one-stream network methods (*cmGAN* [7], *D^{2}RL* [49] and *Zero-Pad* [50]). We conjecture that the main reason is that two-stream networks can simultaneously learn modality-specific and modality-sharable features, which are more suitable for VI-ReID. 2) Our proposed DDAG significantly outperforms the current state-of-the-art AlignGAN [44] by a large margin on both datasets. Note that AlignGAN generates cross-modality image pairs to reduce the modality gap in both feature level and pixel level. In comparison, we do not require the time-consuming and resource-demanding image generation [44,49], and our training process is quite efficient without the adversarial training [7], or the additional modality generation [19].

Another experiment on the RegDB dataset (Table 5) shows that DDAG is robust to different query settings. We achieve much better performance under both *visible-to-infrared* and *infrared-to-visible* query settings, suggesting that DDAG can learn better modality-sharable features by utilizing the part relationship within each modality and graph-structured relations across two modalities.

5 Conclusion

We present a dynamic dual-attentive aggregation learning (DDAG) framework for VI-ReID. DDAG is innovative in two aspects: its IWPA component utilizes the part relationship within each modality to enhance the feature representation by simultaneously considering the part differences and relations; the CGSA module incorporates the neighborhood information across the two modalities to reduce the modality gap. We further design a dynamic dual aggregation learning strategy to seamlessly aggregate the two components. DDAG outperforms the state-of-the-art models on various settings, usually by a large margin. We believe the findings can also be applied in general single-modality person re-identification by mining the relation across multiple body parts, contextual images.

References

1. Bai, S., Tang, P., Torr, P.H., Latecki, L.J.: Re-ranking via metric fusion for object retrieval and person re-identification. In: CVPR. pp. 740–749 (2019)
2. Basaran, E., Gokmen, M., Kamasak, M.E.: An efficient framework for visible-infrared cross modality person re-identification. arXiv preprint arXiv:1907.06498 (2019)
3. Cao, J., Pang, Y., Han, J., Li, X.: Hierarchical shot detector. In: ICCV. pp. 9705–9714 (2019)
4. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV. pp. 371–381 (2019)
5. Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z., Wang, X.: Improving deep visual representation for person re-identification by global and local image-language association. In: ECCV. pp. 54–70 (2018)
6. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: ICML. pp. 793–802 (2018)
7. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. pp. 677–683 (2018)
8. Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M.: Bilinear attention networks for person retrieval. In: ICCV. pp. 8030–8039 (2019)
9. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE TIP **29**, 579–590 (2020)
10. Gong, Y., Zhang, Y., Poellabauer, C., et al.: Second-order non-local attention networks for person re-identification. In: ICCV. pp. 3760–3769 (2019)
11. Hao, Y., Wang, N., Li, J., Gao, X.: Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In: AAAI. pp. 8385–8392 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. He, R., Wu, X., Sun, Z., Tan, T.: Learning invariant deep representation for nir-vis face recognition. In: AAAI. pp. 2000–2006 (2017)
14. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: CVPR. pp. 9317–9326 (2019)
15. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: CVPR. pp. 7183–7192 (2019)
16. Huang, D.A., Frank Wang, Y.C.: Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In: ICCV. pp. 2496–2503 (2013)
17. Jingya, W., Xiatian, Z., Shaogang, G., Wei, L.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR. pp. 2275–2284 (2018)
18. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE TCSVT **30**(4), 1092–1108 (2019)
19. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: AAAI. pp. 4610–4617 (2020)
20. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR. pp. 369–378 (2018)
21. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: ICCV. pp. 1890–1899 (2017)

22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. pp. 2285–2294 (2018)
23. Lin, J.W., Li, H.: Hpiln: A feature learning framework for cross-modality person re-identification. arXiv preprint arXiv:1906.03142 (2019)
24. Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. In: BMVC (2019)
25. Liu, H., Cheng, J.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. arXiv preprint arXiv:1907.09659 (2019)
26. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV. pp. 350–359 (2017)
27. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. arXiv preprint arXiv:1906.08332 (2019)
28. Mudunuri, S.P., Venkataramanan, S., Biswas, S.: Dictionary alignment with re-ranking for low-resolution nir-vis face recognition. IEEE TIFS **14**(4), 886–896 (2019)
29. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
30. Pang, M., Cheung, Y.M., Shi, Q., Li, M.: Iterative dynamic generic learning for face recognition from a contaminated single-sample per person. IEEE TNNLS (2020)
31. Pang, M., Cheung, Y.M., Wang, B., Lou, J.: Synergistic generic learning for face recognition from a contaminated single sample per person. IEEE TIFS **15**, 195–209 (2019)
32. Peng, C., Wang, N., Li, J., Gao, X.: Re-ranking high-dimensional deep local representation for nir-vis face recognition. IEEE TIP (2019)
33. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? In: NeurIPS. pp. 2483–2493 (2018)
34. Sarfraz, M.S., Stiefelhagen, R.: Deep perceptual mapping for cross-modal face recognition. International Journal of Computer Vision **122**(3), 426–438 (2017)
35. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR. pp. 10023–10031 (2019)
36. Shao, R., Lan, X., Yuen, P.C.: Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. IEEE TIFS **14**(4), 923–938 (2018)
37. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR. pp. 5363–5372 (2018)
38. Song, G., Chai, W.: Collaborative learning for deep neural networks. In: NeurIPS. pp. 1837–1846 (2018)
39. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: CVPR. pp. 393–402 (2019)
40. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling. In: ECCV. pp. 480–496 (2018)
41. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: CVPR. pp. 7134–7143 (2019)

42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
43. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)
44. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: ICCV. pp. 3623–3632 (2019)
45. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM. pp. 274–282. ACM (2018)
46. Wang, N., Gao, X., Sun, L., Li, J.: Bayesian face sketch synthesis. IEEE TIP **26**(3), 1264–1274 (2017)
47. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
48. Wang, Z., Wang, Z., Zheng, Y., Wu, Y., Zeng, W., Satoh, S.: Beyond intra-modality: A survey of heterogeneous person re-identification. In: IJCAI (2020)
49. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: CVPR. pp. 618–626 (2019)
50. Wu, A., Zheng, W.s., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV. pp. 5380–5389 (2017)
51. Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z.: Disentangled variational representation for heterogeneous face recognition. In: AAAI. pp. 9005–9012 (2019)
52. Wu, X., Song, L., He, R., Tan, T.: Coupled deep learning for heterogeneous face recognition. In: AAAI. pp. 1679–1686 (2018)
53. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
54. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: CVPR. pp. 1389–1398 (2019)
55. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification. IEEE TIP **28**(6), 2860–2871 (2019)
56. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via a modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing (TIP) (2020)
57. Ye, M., Lan, X., Li, J., Yuen, P.C.: Hierarchical discriminative learning for visible thermal person re-identification. In: AAAI. pp. 7501–7508 (2018)
58. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE TIFS **15**, 407–419 (2020)
59. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020)
60. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE TIFS (2020)
61. Ye, M., Shen, J., Zhang, X., Yuen, P.C., Chang, S.F.: Augmentation invariant and instance spreading feature for softmax embedding. IEEE TPAMI (2020)
62. Zeng, Z., Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Illumination-adaptive person re-identification. IEEE TMM (2020)
63. Zhang, X., Yu, F.X., Karaman, S., Zhang, W., Chang, S.F.: Heated-up softmax embedding. arXiv preprint arXiv:1809.04157 (2018)

64. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
65. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: ICLR (2019)
66. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV. pp. 3219–3228 (2017)
67. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: CVPR. pp. 8514–8522 (2019)
68. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. pp. 1116–1124 (2015)