Learning Concordant Attention via Target-aware Alignment for Visible-Infrared Person Re-identification

Jianbing Wu[1,†] Hong Liu[2,†,*] Yuxin Su[3,†] Wei Shi[4,†] Hao Tang[5,‡]
[†] Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
[‡] Computer Vision Lab, ETH Zürich, Switzerland
[1,3]{kimbing.ng, yuxinsu}@stu.pku.edu.cn [2,4]{hongliu, pkusw}@pku.edu.cn [5]hao.tang@vision.ee.ethz.ch

Abstract

*Owing to the large distribution gap between the heterogeneous data in Visible-Infrared Person Re-identification (VI Re-ID), we point out that existing paradigms often suffer from the inter-modal semantic misalignment issue and thus fail to align and compare local details properly. In this paper, we present Concordant Attention Learning (CAL), a novel framework that learns semantic-aligned representations for VI Re-ID. Specifically, we design the Target-aware Concordant Alignment paradigm, which allows target-aware attention adaptation when aligning heterogeneous samples (*i.e.*, adaptive attention adjustment according to the target image being aligned). This is achieved by exploiting the discriminative clues from the modality counterpart and designing effective modality-agnostic correspondence searching strategies. To ensure semantic concordance during the cross-modal retrieval stage, we further propose MatchDistill, which matches the attention patterns across modalities and learns their underlying semantic correlations by bipartite-graph-based similarity modeling and cross-modal knowledge exchange. Extensive experiments on VI Re-ID benchmark datasets demonstrate the effectiveness and superiority of the proposed CAL.*

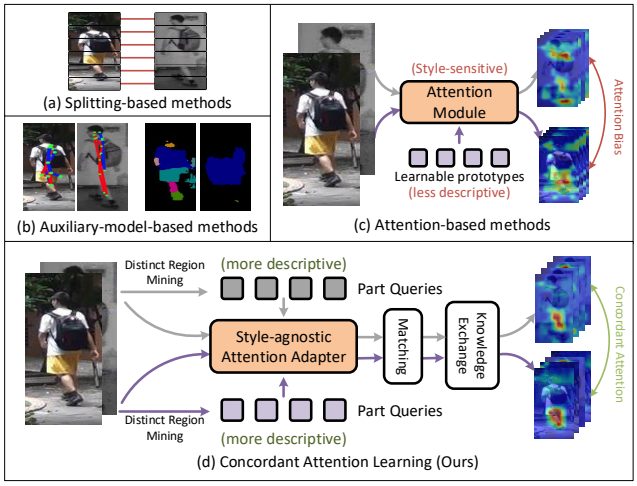

Figure 1. A high-level overview of typical local feature learning paradigms. (a) **Splitting-based**: The semantics of the hand-craft stripes are not always properly aligned. (b) **Auxiliary-model-based**: Pretrained auxiliary models are often error-prone due to the domain shifts (especially for infrared image). (c) **Attention-based**: Due to the inter-modal distribution gap, existing style-sensitive attention module with less descriptive learnable prototypes often fails to attend semantically consistent regions. (d) **CAL (Ours)**: Our method can learn concordant attention by discriminative region mining, target-aware style-agnostic attention, and part-aligned knowledge exchange.

1. Introduction

Person Re-identification (Re-ID) aims to associate person identities across non-overlapping cameras. It has gained increasing attention in recent years due to its practical applications in real-world surveillance systems. Conventional person Re-ID methods mainly focus on retrieving the same identity across visible (RGB) cameras [48, 29, 16, 23, 10]. Despite their remarkable success, they have limited applicability since visible cameras cannot capture discriminative information under poor-lighting conditions (*e.g.*, at night). To improve the illumination robustness, infrared (IR) cameras are widely applied to cooperate with visible ones in real-world surveillance systems. This increases the need to explore the Visible-Infrared Person Re-identification (VI Re-ID) problem, which aims to associate the person images taken by different spectrum cameras to achieve long-term person tracking in 24-hour surveillance systems.

Compared to traditional RGB-based re-identification, VI Re-ID is much more challenging due to the differences in spectral properties between visible and infrared images. This data heterogeneity can lead to severe misalignment in the feature space and large intra-class discrepancy, resulting in significant degradation in performance. In addition, similarly to RGB-based Re-ID tasks, VI Re-ID can also be

impacted by the changes in pose or background, resulting in increased difficulties.

Recent years have witnessed a surge of creative work on mitigating the modality gap for VI Re-ID [38, 43, 18, 44, 4, 20]. However, as most existing methods only consider learning global representations from the whole image, they fail to compare local details. In addition, owing to the shortcut learning characteristics [6], global-feature-based methods would tend to learn modality-specific shortcut patterns, making the learned features susceptible to concentrating on divergent regions in each modality and resulting in suboptimal performance. A seemingly straightforward solution is to learn semantic-aligned local features with either splitting-based [29, 8, 20], auxiliary-model-based methods [26, 19, 28, 1], or attention-base paradigms [18, 16, 33, 27] instead of the global ones. However, these methods also fail to achieve inter-modal semantic alignment, as demonstrated in Figure 1, and forcibly aligning these semantic-misaligned embeddings would inevitably injure the training process and compromise the performance.

Different from these deep learning methods, human visual systems can naturally avoid the misalignment issue thanks to their target-aware comparison strategy. Considering the scenario of comparing two images of persons (referred to as "base image" and "target image", respectively), human vision systems would first identify multiple distinctive key regions (such as facial and clothing details) in the target image and then direct their attention to corresponding regions in the base image. In this manner, humans can always make perfect part-to-part comparisons[1] since they are able to adaptively and accurately adjust their attention by referencing the key regions of the target image. This suggests that exploiting clues from the target image of the modality counterpart and exploring an effective cross-modal corresponding region-searching strategy can benefit in mitigating the inter-modal semantic misalignment issue.

In this paper, we present Concordant Attention Learning (CAL), a novel framework that mimics the target-aware comparison behavior of human vision systems to learn semantic-aligned representations for VI Re-ID. Firstly, we devise the Target-aware Concordant Alignment (TCA) paradigm, which aims to exploit discriminative local clues from the target modality (*i.e.*, the modality counterpart) when aligning heterogeneous embeddings. The proposed TCA consists of three components: (1) *Discriminative Region Mining*: identifying diverse and discriminative key regions from the feature maps of each training sample; (2) *Target-aware Style-agnostic Attention Adapter*: taking the selected key regions from the target modality as part queries, and applying target-aware refinement to adapt the feature attention and generate part-aligned embeddings; (3) *Part-aligned Metric Learning*: clustering or separating the part-aligned embeddings across modalities according to their identity labels. Even though this target-aware refinement scheme can mitigate the modality discrepancy by leveraging the clues from the target modality, it brings higher computational costs during inference. Because it needs to be carried out for all query-gallery pairs, and the gallery set is generally large. To this end, we further propose MatchDistill. First, we need to associate the generated part queries of different modalities to guarantee that the subsequent distillation is conducted on semantic-aligned features. This is achieved by modeling the correlations of their corresponding attention maps with bipartite graphs and conducting Cross-modal Query Matching (CQM) to find the optimal matches. After that, a dual-level knowledge distillation loss is designed to allow cross-modality knowledge exchange between the best-matched queries. This can facilitate the learning of underlying relationships between the visible and infrared modalities. After training with MatchDistill, each modality can learn knowledge from its modality counterpart, and no cross-modal interaction is needed during inference.

Overall, our contributions are summarized as follows:

- We propose Concordant Attention Learning (CAL), a target-aware training paradigm that mimics human behavior and learns concordant attention to alleviate the inter-modal attention bias issue for VI Re-ID.
- To enable part-aligned metric learning, we present Target-aware Concordant Alignment (TCA), which leverages cross-modal clues and allows adaptive attention adjustment when aligning heterogeneous embeddings.
- We propose MatchDistill, which matches the attention patterns across modalities and learns their underlying semantic correlations by bipartite-graph-based similarity modeling and cross-modal knowledge exchange.
- Extensive experiments demonstrate that the proposed CAL archives state-of-the-art performance on both the SYSU-MM01 [38] and RegDB [25] datasets.

2. Related Work

Visible-Infrared Person Re-ID. Visible-Infrared Person Re-ID (VI Re-ID) has drawn increasing attention in recent years [43, 17, 39, 12, 20]. Mainstream VI Re-ID paradigms focus on bridging the modality gap via designing metric learning constraints [45, 42, 40], normalizing feature statistics [14, 39, 13], synthesizing auxiliary training samples [34, 3, 49, 24], or developing modality-specific and modality-shared feature learning paradigms [5, 22, 30]. However, these methods only consider learning global representations from entire images without considering the local information of images, leading to limited expressiveness of learned features. A few works also attempt to learn local

[1]Note that we use the term "region" and "part" interchangeably to denote the same thing.

features for VI Re-ID [8, 20, 41, 47] by splitting the feature maps into several hand-craft stripes. However, due to inaccurate detection boxes, pose variations, and occlusion, the hand-craft stripes are not always well-aligned, resulting in unsatisfactory performance.

Local Feature Learning in Person Re-ID. Local feature learning is an important research direction for person re-identification since learning part aggregated features makes the model robust against misalignment [44]. It can be roughly divided into splitting-based methods, auxiliary-model-based methods, and attention-based methods. Splitting-based methods [29, 8, 20] focus on learning local features from horizontal-divided regions, which now serve as a strong part feature learning baseline. However, due to inaccurate detection boxes, pose variations, and occlusion, the semantics of the hand-craft stripes are not well aligned. The auxiliary-model-based methods [26, 19, 28, 1] often adopt off-the-shelf auxiliary human parsing or pose estimation networks to obtain semantically meaningful body parts. However, they require additional computational costs and are prone to noisy estimation, especially when the data distribution varies. The attention-based methods focus on exploiting attention mechanisms to localize discriminative human parts and have achieved great success [18, 16, 33, 27]. However, these methods cannot be directly applied to the cross-modality scenario of VI Re-ID due to the inter-modal semantic misalignment issue as discussed in Sec. 1.

3. Methodology

3.1. Intuition of Target-aware Alignment

Let $\boldsymbol{x}^k$ denote the training images of modality k, where $k \in \{V, I\}$ (V for visible modality and I for infrared modality). The visible and infrared samples in the dataset are denoted by $\mathcal{V} = \{\boldsymbol{x}_j^V, y_j^V\}_{j=0}^{N_v}$ and $\mathcal{I} = \{\boldsymbol{x}_j^I, y_j^I\}_{j=0}^{N_i}$, respectively, where N_v and N_i are the numbers of samples of each modality in the dataset, and y_j^k is the corresponding identity label of the j-*th* sample from modality k. The goal of Visible-Infrared Person Re-identification is to match the person identities across modalities according to feature similarities. Therefore, it is essential to reduce the large intra-class variation between heterogeneous samples. Existing paradigms often attempt to reduce the cross-modal intra-class variation by directly optimizing

$$\mathbb{E}_{i,j}\Big[\mathbb{1}(y_i^V = y_j^I)\cdot \mathrm{d}\big(f(\boldsymbol{x}_i^V), f(\boldsymbol{x}_j^I)\big)\Big], \tag{1}$$

where f denotes the feature extractor, and $\mathrm{d}(\cdot,\cdot)$ represents the distance between features. However, as discussed in Sec. 1, the attention misalignment issue could injure the training process and compromise performance. We thus endow the model with the ability to adaptively adjust the attention according to the target being compared to achieve attention consensus. First, the Discriminative Region Mining (DRM) module, formulated as $\gamma(\cdot)$, is introduced to disentangle the global features into diverse key body parts. We then devise the Target-aware Style-agnostic Attention Adapter (TSAA) to allow adaptive style-agnostic attention adjustment according to any given part queries regardless of the image styles. The feature extractor is then reformulated as $f(\boldsymbol{x}, \boldsymbol{p})$ to represent the above dynamic attention adjustment process, where $\boldsymbol{p}$ denotes the given body parts. The objective function to reduce the cross-modal intra-class variation can then be defined as

$$\begin{aligned}\mathbb{E}_{i,j,k}\Big[&\mathbb{1}(y_i^V = y_j^I)\cdot \mathrm{d}\big(f(\boldsymbol{x}_i^V, \gamma(\boldsymbol{x}_j^I)_k), f(\boldsymbol{x}_j^I, \gamma(\boldsymbol{x}_j^I)_k)\big) + \\ &\mathbb{1}(y_i^I = y_j^V)\cdot \mathrm{d}\big(f(\boldsymbol{x}_i^I, \gamma(\boldsymbol{x}_j^V)_k), f(\boldsymbol{x}_j^V, \gamma(\boldsymbol{x}_j^V)_k)\big)\Big].\end{aligned} \tag{2}$$

In this manner, the embeddings being optimized can achieve semantic concordance since the same queries are used to guide feature attention (for instance, $f(\boldsymbol{x}_i^V, \gamma(\boldsymbol{x}_j^I)_k)$ and $f(\boldsymbol{x}_j^I, \gamma(\boldsymbol{x}_j^I)_k)$ correspond to the same semantic region since they are refined using the same part query $\gamma(\boldsymbol{x}_j^I)_k$). This also allows cross-modal interaction to facilitate the learning of underline relationships between modalities. The overall pipeline is depicted in Figure 2, and more details will be introduced in the following subsections.

3.2. Target-aware Concordant Alignment

3.2.1 Discriminative Region Mining

The Discriminative Region Mining module aims to discover several diverse discriminative key regions from the feature maps $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ derived from the backbone network, where C, H, and W denote the number of channels, the height, and the width of the feature maps. Our DRM contains two stages: *Part Scoring* and *Token Selection and Aggregation*. The former aims to classify the tokens (*i.e.*, the spatial vectors of $\boldsymbol{X}$) into N_p different parts, while the latter selects the top-k tokens for each part based on the scores and aggregates them to derive the part features $\boldsymbol{P}$.

Part Scoring. To disentangle the feature maps into several diverse discriminative regions, we first develop a simple scorer network γ_θ to predict the scores of N_p discriminative parts for each token in the feature maps. Here, N_p is a hyperparameter representing the number of parts. Specifically, as illustrated in Figure 2, the feature maps $\boldsymbol{X}$ are first passed through multiple max-pooling layers with different scales (*i.e.*, 1×1 for identity mapping, 3×3, and 5×5 in this paper) to obtain features with different receptive fields. These features are then upsampled and concatenated along the channel dimension. Finally, the concatenated feature maps are fed into a point-wise convolutional layer with softmax activation along the spatial function to predict the scores of each part.

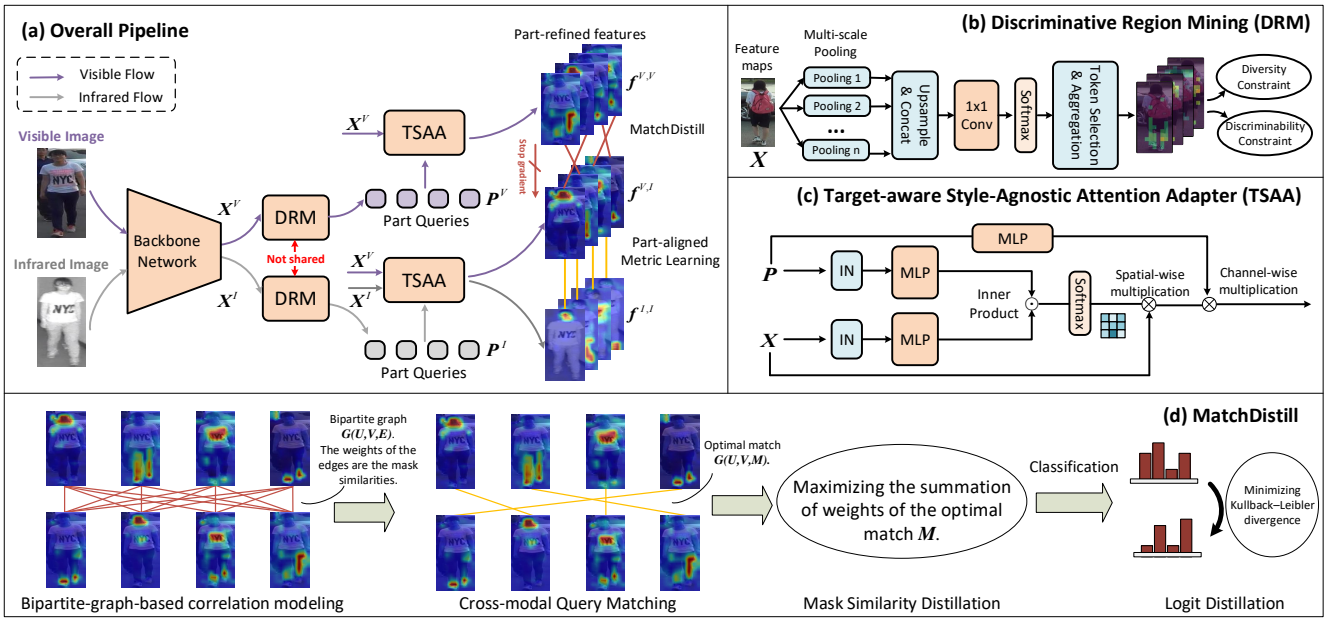

Figure 2. (a) The overall pipeline of the proposed Concordant Attention Learning (CAL), which consists of a deep CNN backbone, the Discriminative Region Mining (DRM), the Target-aware Style-agnostic Attention Adapter (TSAA), and the MatchDistill paradigm. Note that only half of the pipeline is depicted for clarity since the model is symmetric. (b) The pipeline of DRM. (c) The architecture of the proposed TSAA. (d) The pipeline of MatchDistill. Notations are introduced in Sec. 3.

Token Selection and Aggregation. Given the scores predicted by γ_θ, we then select the tokens with top-k scores for each part and aggregate them by weighted averaging to derive the aggregated part features. Specifically, let $\boldsymbol{S} \in \mathbb{R}^{(H\cdot W)\times N_p}$ denote the predicted score after the softmax layer produced by γ_θ, and let $\boldsymbol{T} \in \mathbb{R}^{(H\cdot W)\times C}$ denote the flattened feature maps of $\boldsymbol{X}$. The token selection and aggregation process can be formulated as

$$\boldsymbol{P} = \frac{\hat{\boldsymbol{S}}^T \boldsymbol{T}}{\sum_j \hat{\boldsymbol{S}}_j}, \text{ where } \hat{\boldsymbol{S}} = \mathcal{T}(\boldsymbol{S}, N_k), N_k = \lceil \alpha \cdot \frac{H\cdot W}{N_p} \rceil, \quad (3)$$

where $\mathcal{T}$ denotes the top-k operation; $\hat{\boldsymbol{S}} \in \mathbb{R}^{(H\cdot W)\times N_p}$ represents the selected regions, N_k is the number of selected tokens, α is a hyperparameter controlling the selection ratio, and $\boldsymbol{P} \in \mathbb{R}^{N_p \times C}$ represents the aggregated part features. There are two main advantages of selecting the top-k tokens instead of preserving all of them: (1) it helps filter out noisy tokens that could bring noisy clues to subsequent modules; (2) it can enhance the locality of the selected region and facilitate the learning of local representations.

Discriminability Constraint. To ensure the part features $\boldsymbol{P}$ contains discriminative information, we design the discriminative constraint loss, which is formulated as

$$\mathcal{L}_{dis} = -\sum_{i=1}^{N_p} \log(\Phi_i(\boldsymbol{P}_i)_y), \quad (4)$$

where Φ_k denotes part-specific classifiers implemented with a simple fully connected layer with softmax activation, and y is the label of the current sample.

Diversity Constraint. By training DRM with only Eq. (4), the model would tend to select several identical regions, resulting in suboptimal performance. This can be attributed to the shortcut learning characteristics [6] of deep learning systems that models attempt to find the simplest (but may be suboptimal) solution to solve a given task. To enhance the diversity of the selected regions, we further design the diversity constraint loss, which is defined as

$$\mathcal{L}_{div} = \sum_{i,j} \text{triu}(\hat{\boldsymbol{S}}^T \hat{\boldsymbol{S}}, 1)_{i,j}, \quad (5)$$

where $\text{triu}(\cdot, 1)$ denotes the upper triangular part (excluding the diagonal) of the given matrix. This diversity constraint loss can be regarded as imposing an explicit penalty on the model when the selected key regions become too similar.

As shown in Figure 2, the DRM module is not shared across modalities. This can help to learn better modality-specific part prototypes and select more discriminative regions, leading to better performance. This design will be empirically validated in Sec. 4.

3.2.2 Target-aware Style-agnostic Attention Adapter

To achieve target-aware attention, it is important to properly localize relevant regions in the feature maps by taking the aggregated part features $\boldsymbol{P}$ as the reference. A straightforward solution is to directly apply Scaled Dot-Product Attention [32] by taking $\boldsymbol{P}$ as the part queries and the flattened

feature maps $\boldsymbol{T}$ as the keys and the values, formulated as

$$\text{Attn}(\boldsymbol{T};\boldsymbol{P}) = \text{softmax}\left(\boldsymbol{P}\boldsymbol{T}^T/\sqrt{d}\right)\boldsymbol{T}, \tag{6}$$

where d is the feature dimension. However, this naive solution simply adopts dot product as the similarity metric, which could result in inaccurate attention scores. Because the features of different modalities suffer from large distributions shift, making them hard to be properly compared using a simple dot product operation. To this end, we design the Target-aware Style-agnostic Attention Adapter (TSAA), which includes a more reasonable similarity calculator. As shown in Figure 2 (b), we first apply Instance Normalization (IN) [31] to the part queries $\boldsymbol{P}$ and each token of the feature maps $\boldsymbol{X}$, formulated as

$$\text{IN}(\boldsymbol{P}) = \frac{\boldsymbol{P}-\mu(\boldsymbol{T}_P)}{\sigma(\boldsymbol{T}_P)}, \quad \text{IN}(\boldsymbol{T}) = \frac{\boldsymbol{T}-\mu(\boldsymbol{T})}{\sigma(\boldsymbol{T})}, \tag{7}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote calculating the instance-level mean and standard deviation; $\boldsymbol{T}_P$ is the flattened feature maps that $\boldsymbol{P}$ derives from (see Eq. (3)). The motivation of this design is that the modality gap between IR and RGB images can be interpreted as the large discrepancies in style, and IN has been proven to be effective in reducing the style variance [14, 39, 13]. After that, the normalized features are fed into multi-layer perceptrons (MLPs) to learn nonlinear projection and capture complex patterns of the features. We then calculate the inner product of the projected features after the MLPs and apply softmax function on the spatial dimension to derive the final similarity. We can then apply spatial attention by weighted averaging all tokens using the computed similarities. In addition, we also apply query-guided channel attention to emphasize important channels in the feature map and suppress less important ones, where scores of channel attention are generated by an MLP layer taking part queries as input. Formally, the TSAA module, denoted by $\mathcal{A}(\cdot;\cdot)$, can be formulated as

$$\mathcal{A}(\boldsymbol{T};\boldsymbol{P}) = \mathscr{S}(\boldsymbol{T},\boldsymbol{P})\,\boldsymbol{T} \otimes \mathcal{M}(\boldsymbol{P};\theta_3), \tag{8}$$

$$\mathscr{S}(\boldsymbol{T},\boldsymbol{P}) = \text{softmax}\left(\hat{\boldsymbol{P}}\hat{\boldsymbol{T}}^T/\sqrt{d}\right), \tag{9}$$

$$\hat{\boldsymbol{T}} = \mathcal{M}(\text{IN}(\boldsymbol{T});\theta_1), \quad \hat{\boldsymbol{P}} = \mathcal{M}(\text{IN}(\boldsymbol{P});\theta_2), \tag{10}$$

where $\mathscr{S}$ denotes the spatial similarity calculator, $\otimes$ denotes channel-wise multiplication, and $\mathcal{M}(\cdot;\theta)$ denotes MLP with parameter θ. The TSAA has two merits: (1) it allows style-agnostic attention adjustment based on any given part queries $\boldsymbol{P}$ regardless of the image styles, making it easier to achieve attention consensus; (2) it can naturally bridge the gap between cross-modal samples by attending to relevant regions and emphasizing relevant channels.

3.2.3 Part-aligned Metric Learning

Our part-aligned metric learning aims to reduce the cross-modal intra-class discrepancies between the part-aligned features. As shown in Figure 2 (a), we first employ TSAA to perform cross-modal refinement since it can effectively bridge the gap between cross-modal samples as discussed in Sec. 3.2.2. The cross-modal refinement is defined as

$$\boldsymbol{f}_i^{V,I} = \mathcal{A}(\boldsymbol{T}_j^V;\boldsymbol{P}_i^I), \quad \boldsymbol{f}_i^{I,V} = \mathcal{A}(\boldsymbol{T}_j^I;\boldsymbol{P}_i^V), \tag{11}$$

where $\boldsymbol{T}_j$ is randomly selected from the mini-batch and $y_i^I = y_j^V$; $\boldsymbol{f}^{V,I}$ and $\boldsymbol{f}^{I,V}$ are called cross-refined features. Similarly, we also apply self-refinement using

$$\boldsymbol{f}_i^{V,V} = \mathcal{A}(\boldsymbol{T}_i^V;\boldsymbol{P}_i^V), \quad \boldsymbol{f}_i^{I,I} = \mathcal{A}(\boldsymbol{T}_i^I;\boldsymbol{P}_i^I), \tag{12}$$

to obtain the self-refined features. Since $\boldsymbol{f}_i^{V,I}$ and $\boldsymbol{f}_i^{I,I}$ are refined using the same query (*i.e.*, $\boldsymbol{P}_i^I$), they would attend to the same regions and can thus be regarded as part-aligned features. We then design the Part-aligned Center Loss (PCL) to enhance the discriminability of these semantically aligned embeddings. Let $\boldsymbol{c}_j^k$ denotes the batch feature centroid of class y_j^k in modality k ($k \in \{V, I\}$) given by

$$\boldsymbol{c}_j^V = \frac{1}{|\mathcal{S}(j)|}\sum_{i\in\mathcal{S}(j)}\boldsymbol{f}_i^{V,V}, \quad \boldsymbol{c}_j^I = \frac{1}{|\mathcal{S}(j)|}\sum_{i\in\mathcal{S}(j)}\boldsymbol{f}_i^{I,I}, \tag{13}$$

where $\mathcal{S}(j) = \{i \mid y_i = y_j\}$. Our PCL is then defined as

$$\begin{aligned}\mathcal{L}_{pcl} = \sum\nolimits_{i,j}\big[&\mathbb{g}(\boldsymbol{f}_i^{VI},\boldsymbol{c}_j^I) + \mathbb{g}(\boldsymbol{f}_i^{IV},\boldsymbol{c}_j^V) + \\ &\mathbb{s}(\boldsymbol{c}_i^I,\boldsymbol{c}_j^V) + \mathbb{s}(\boldsymbol{c}_i^I,\boldsymbol{c}_j^I) + \mathbb{s}(\boldsymbol{c}_i^V,\boldsymbol{c}_j^V)\big].\end{aligned} \tag{14}$$

Here, $\mathbb{g}$ and $\mathbb{s}$ aim to gather up intra-class features and separate inter-class features, respectively, formulated as

$$\begin{aligned}\mathbb{g}(\boldsymbol{f},\boldsymbol{c}) &= \big[\mathbb{1}(y_{(\boldsymbol{f})} = y_{(\boldsymbol{c})})\cdot\|\boldsymbol{f}-\boldsymbol{c}\|_2\big], \\ \mathbb{s}(\boldsymbol{c}_i,\boldsymbol{c}_j) &= \Big[\mathbb{1}(y_{(\boldsymbol{c}_i)} \neq y_{(\boldsymbol{c}_j)})\cdot\big[\sigma-\|\boldsymbol{c}_i-\boldsymbol{c}_j\|_2\big]_+\Big],\end{aligned} \tag{15}$$

where σ is a marginal parameter adopted to avoid optimizing "already correct" centroid pairs; $[z]_+ = \max(z, 0)$; $y(\cdot)$ denotes the label of the corresponding embedding/centroid.

3.3. Cross-modal Knowledge Exchange

Even though the cross-refined embeddings yield smaller distribution gaps with respect to the target modality, it is infeasible to employ cross-refined embeddings for retrieval during the inference stage. Because the cross-modal refinement requires taking the part features $\boldsymbol{P}$ (derived from DRM using the images of modality counterpart) as queries and applying query-guided attention using Eq. (10). This process need to be carried out for all query-gallery pairs, which brings high computational costs since gallery sets are generally large. To this end, we propose a novel solution called MatchDistill, which allows DRM to directly generate target-modality-like part features. This is achieved by associating the heterogeneous queries $\boldsymbol{P}^V$ and $\boldsymbol{P}^I$ with our Cross-modal Query Matching algorithm and propagating

the modality-specific knowledge to the best-matched query of modality counterparts. For clarity, here we only take matching and propagating the knowledge from $\boldsymbol{P}^I$ to $\boldsymbol{P}^V$ as an example since MatchDistill has a symmetric pipeline. As shown in Figure 2, in MatchDistill, we first construct a complete bipartite graph $G(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{E})$, where the vertices $\boldsymbol{U}$ and $\boldsymbol{V}$ represent $\boldsymbol{P}^I$ and $\boldsymbol{P}^V$, respectively. Each edge in $\boldsymbol{E}$ connects a vertex in $\boldsymbol{U}$ to one in $\boldsymbol{V}$. The weights of the edges are the semantic similarities between queries, which are measured by the dot product of the attention mask calculated by Eq. (10), formulated as

$$w(\boldsymbol{E}^V_{k_1,k_2}) = \langle \mathscr{S}(\boldsymbol{X}^V, \boldsymbol{P}^I_{k_1}), \mathscr{S}(\boldsymbol{X}^V, \boldsymbol{P}^V_{k_2}) \rangle, \quad (16)$$

where $w(\cdot)$ denotes the weight of the edge. The idea behind this is straightforward: if $\boldsymbol{P}_{k1}$ and $\boldsymbol{P}_{k2}$ represent the same semantic region, their attention masks should have a high level of similarity. This could help effectively model the semantic correlations of the heterogeneous part queries. The part query matching problem can then be formulated as

$$\begin{aligned} &\arg\max_{\boldsymbol{M}\subseteq\boldsymbol{E}} \sum\nolimits_{e\in\boldsymbol{M}} w(e), \quad s.t.\ |\boldsymbol{M}| = N_p \text{ and} \\ &\forall (u_1,v_1),(u_2,v_2) \in \boldsymbol{M}, u_1 \neq u_2 \text{ and } v_1 \neq v_2, \end{aligned} \quad (17)$$

where $\boldsymbol{M}$ represents the edges of the optimal match (meaning that $\forall (u,v) \in \boldsymbol{M}$, $\boldsymbol{P}^V_u$ and $\boldsymbol{P}^I_v$ represent the same semantic region). We solve this part query matching problem (Eq. (17)) with the Kuhn-Munkres assignment algorithm [15]. After that, the mutual knowledge propagation process can be performed between all $(u,v) \in \boldsymbol{M}$ pairs since they are semantically aligned. Specifically, the proposed knowledge distillation is conducted in two levels, *i.e.*, the feature level and the logit level. For the feature level, we propose to align the similarity maps of the best-matched query pairs. This is equivalent to maximizing the weight of the optimal bipartite graph, formulated as

$$\mathcal{L}_{kd1} = -\sum\nolimits_{e\in\boldsymbol{M}} w(e). \quad (18)$$

For the logit level, we constrain the consistency of their softmax classification distribution by

$$\mathcal{L}_{kd2} = \sum\nolimits_{(u,v)\in\boldsymbol{M}} \mathcal{D}_{kl}(\boldsymbol{p}^{V,I}_u \parallel \boldsymbol{p}^{V,V}_v), \quad (19)$$

where $\boldsymbol{p}$ denotes the post-softmax classification probability, and $\mathcal{D}_{kl}$ denotes Kullback–Leibler divergence. With the above cross-modal knowledgable propagation, the self-refined features can also capture important information that can benefit cross-modal retrieval like the cross-refined ones. Therefore, there is no needs to perform cross-refinement during inference. In addition, by optimizing $\mathcal{L}_{kd1}$ and $\mathcal{L}_{kd2}$, we observe that the optimal matches $\boldsymbol{M}$ gradually get stabilized as the training progresses. This suggests that the optimal matches with the highest occurrence in the last training epoch, denoted as $\hat{\boldsymbol{M}}$, can be used for test-time matching.

3.4. Training and Inference

Training. The overall training objective is defined as

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{pcl} + \lambda_1\mathcal{L}_{kd1} + \lambda_2\mathcal{L}_{kd2} + \lambda_3\mathcal{L}_{dis} + \lambda_4\mathcal{L}_{div}, \quad (20)$$

where $\mathcal{L}_{id}$ is the routinely used identity classification loss; λ_1, λ_2,λ_3, and λ_4 are adopted to balance different losses.

Inference. During inference, we first reorder the learned prototypes of the DRM (*i.e.*, the parameters of the point-wise convolutional layer in DRM) according to $\hat{\boldsymbol{M}}$ to ensure that the self-refined embeddings can be semantically aligned. After that, the summation of the cosine similarity of each part is then adopted as the comparison metric. Formally, the similarity between the i^{th} visible and the j^{th} infrared sample is given by $\sum_p \cos(\boldsymbol{f}^{V,V}_{i,p}, \boldsymbol{f}^{I,I}_{j,p})$, where cos denotes cosine similarity computation, and the subscript p denotes indexing in the part dimension.

4. Experiments

4.1. Experimental Settings

Datasets. The proposed method is evaluated under the same protocols of existing work [20] on two widely-used VI Re-ID datasets, *i.e.*, SYSU-MM01 [38] and RegDB [25].

Implementation Details. We adopt ResNet-50 [9] as the backbone network following existing work. At each training iteration, 8 identities are randomly sampled. For each identity, 4 visible and 4 infrared images are selected to form a mini-batch. The model is trained for 100 epochs in total on RTX 3090 GPUs with SGD optimizer. The learning rate linearly increased from 0.01 to 0.1 in the first 10 epochs. The cosine annealing strategy [21] is adopted in the remaining 90 epochs to decay the learning rate to 10^{-3}. The hyperparameters are decided by cross-validation. Specifically, we set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$, $N_p = 4$, $\alpha = 0.6$, and $\sigma = 0.6$. All models are tested without using re-ranking algorithms or gallery set information for fair comparison.

4.2. Comparison with State-of-the-Art Methods

The comparison results on SYSU-MM01 and RegDB with state-of-the-art (SOTA) methods are shown in Table 1 and Table 2, respectively. It can be seen that the proposed CAL outperforms existing approaches by large margins on both datasets. Specifically, on the SYSU-MM01 dataset, our proposed method surpasses the state-of-the-art approaches [12, 20] on most of the evaluation metrics. As for the RegDB dataset, our method also achieves new state-of-the-art results. These experimental results verify the effectiveness and superiority of the proposed method.

4.3. Analysis and Discussion

Effectiveness of TCA and MatchDistill. To verify the effectiveness of the two key components of CAL (*i.e.*, TCA

Table 1. Comparison with SOTA methods on the SYSU-MM01 dataset. †: Following previous work [43], for cm-SSFT [22] and CIFT [17], we report the performance when no extra auxiliary information of the gallery set is introduced for fair comparison.

Method	Venue	All-search						Indoor-Search					
		Single-Shot			Multi-Shot			Single-Shot			Multi-Shot		
		R1 ↑	R10 ↑	mAP ↑	R1 ↑	R10 ↑	mAP ↑	R1 ↑	R10 ↑	mAP ↑	R1 ↑	R10 ↑	mAP ↑
Zero-Pad [38]	ICCV 2017	14.80	54.12	15.95	19.13	61.40	10.89	20.58	68.38	26.92	24.43	75.86	18.64
AGW [44]	TPAMI 2022	47.50	84.39	47.65	-	-	-	54.17	91.14	62.97	-	-	-
cm-SSFT† [22]	CVPR 2020	47.70	-	54.10	57.40	-	59.10	-	-	-	-	-	-
NFS [2]	CVPR 2021	56.91	91.34	55.45	63.51	94.42	48.56	62.79	96.53	69.79	70.03	97.70	61.45
SMCL [36]	ICCV 2021	67.39	92.87	61.78	72.15	90.66	54.93	68.84	96.55	75.56	79.57	95.33	66.57
MCLNet [7]	ICCV 2021	65.40	93.33	61.98	-	-	-	72.56	96.98	76.58	-	-	-
MPMN [35]	TMM 2021	48.98	90.33	62.41	60.88	88.70	51.90	64.89	96.85	76.47	74.42	92.93	66.98
FMCNet [46]	CVPR 2022	66.34	-	62.51	-	-	-	68.15	-	74.09	-	-	-
CAJ [43]	ICCV 2021	69.88	95.71	66.89	-	-	-	76.26	97.88	80.37	-	-	-
CIFT†[17]	ECCV 2022	71.77	-	67.64	78.00	-	62.46	78.65	-	82.11	86.97	-	77.03
MPANet [39]	CVPR 2021	70.58	96.21	68.24	75.58	97.91	62.91	76.74	98.21	80.95	84.22	99.66	75.11
CMT [12]	ECCV 2022	71.88	96.45	68.57	**80.23**	97.91	63.13	76.90	97.68	79.91	84.87	99.41	74.11
MAUM [20]	CVPR 2022	71.68	-	68.79	-	-	-	76.97	-	81.94	-	-	-
CAL (Ours)	**-**	**74.66**	**96.47**	**71.73**	77.05	**98.01**	**64.86**	**79.69**	**98.93**	**83.68**	**86.97**	**99.83**	**78.51**

Table 2. Comparison with SOTA methods under "visible to infrared" and "infrared to visible" modes on RegDB. †: see Table 1.

Method	Visible to Infrared			Infrared to Visible		
	R1 ↑	R10 ↑	mAP ↑	R1 ↑	R10 ↑	mAP ↑
Zero-Pad [38]	17.75	34.21	18.90	16.63	34.68	17.82
cm-SSFT†[22]	65.40	-	65.60	63.80	-	64.20
AGW [44]	70.05	86.21	66.37	75.93	90.93	69.49
NFS [2]	80.54	91.96	72.10	77.95	90.45	69.79
MCLNet [7]	80.31	92.70	73.07	75.93	90.93	69.49
CAJ [43]	85.03	95.49	79.14	84.75	95.33	77.82
SMCL [36]	83.93	-	79.83	83.05	-	78.57
MPANet [39]	83.70	-	80.90	82.80	-	80.70
MPMN [35]	86.56	96.86	82.91	84.62	95.51	79.49
FMCNet [46]	89.12	-	84.43	88.38	-	83.86
MAUM [20]	87.87	-	85.09	86.95	-	84.34
CIFT†[17]	92.17	-	86.96	90.12	-	84.81
CMT [12]	**95.17**	98.82	87.30	91.97	97.92	84.46
CAL (Ours)	94.51	**99.70**	**88.67**	**93.64**	**99.46**	**87.61**

and MatchDistill) and analyze their contribution to the overall performance, we conduct experiments on the more challenging SYSU-MM01 dataset. The experimental results are given in Table 3, where we adopt ResNet-50 as the baseline method. We can see that both TCA and MatchDistill are shown to be essential to the final performance and can significantly boost the performance of the baseline method. These findings demonstrate the effectiveness and importance of our proposed approach.

Analysis on the DRM Module. To demonstrate the superiority of the design of DRM, we conduct experiments by removing the key components of DRM or replacing them with other alternatives. The experimental results are shown in Table 4 (top). We can see that all modifications can lead to performance degradation. Specifically, we can see that when training DRM without the diversity constraint ($\mathcal{L}_{div}$) or discriminability constraint ($\mathcal{L}_{dis}$), the performance drops significantly, indicating that it is essential to guarantee both the discriminability and diversity of the selected regions. When replacing the modality-specific DRM with the modality-shared one, a performance drop can also be observed. This is because the modality-shared DRM cannot learn modality-specific patterns, making it hard to discover discriminative modality-specific clues. It can also be observed that using learnable vectors instead of DRM produces poor performance since they are less descriptive than the part features generated by DRM. Horizontal Splitting (HS), which divides the feature maps into horizontal chunks, achieves slightly better results than learnable vectors since the chunks are more informative than learnable vectors. However, HS is still inferior to our DRM as handcrafted stripes cannot properly disentangle the feature maps into fine-grained discriminative body parts. We also study the impact of the number of regions generated and the parameter N_k and α in DRM. The experimental results are shown in Figure 3. The best performance is achieved when $N_p = 4$ and $\alpha = 0.6$.

Analysis on the TSAA Module. To demonstrate the superiority of the TSAA design, we conduct experiments by replacing the spatial similarity calculation process (Eq. (9)) with the simple scaled dot-product operation and removing the query-guided channel attention mechanism. As shown in Table 4 (middle), the experimental results reveal significant drops in performance when adopting the scaled dot product as the similarity metric. This can be attributed to the style discrepancies between modalities, which make it difficult to effectively find the correlation between heterogeneous data with dot product. Our TSAA module, on the other hand, is adept at reducing style discrepancies between heterogeneous data when computing similarity scores. In addition, the experimental results also validate the effectiveness of our query-guided channel attention mechanism.

Analysis on PCL. The PCL aligns the cross-modal em-

Table 3. Effectiveness of the proposed components on the SYSU-MM01 dataset under the all-search single-shot mode.

Baseline	TCA	MatchDistill	SYSU-MM01		
			R1 ↑	R10 ↑	mAP ↑
✓			57.82	89.91	56.40
✓	✓		70.15	93.12	67.40
✓	✓	✓	**74.66**	**96.47**	**71.73**

Table 4. Ablation study of DRM and TSAA on SYSU-MM01 under the all-search single-shot mode. *Shared DRM*: the weights of DRM are shared for both modalities. *LV*: replacing the selected regions with N_p learnable vectors. *HS*: Horizontal Split (n chunks). *TSAA (SD)*: adopting scaled dot-product for similarity computation in TSAA. *CA*: the channel-wise attention in TSAA.

Method	SYSU-MM01		
	R1 ↑	R10 ↑	mAP ↑
DRM w/o $\mathcal{L}_{div}$	72.76	95.71	69.85
DRM w/o $\mathcal{L}_{dis}$	72.23	95.37	69.38
DRM → Shared DRM	72.57	94.95	69.51
DRM → LV	71.07	94.01	68.60
DRM → HS ($n = 2$)	72.70	95.50	70.02
DRM → HS ($n = 4$)	73.15	95.76	70.24
DRM → HS ($n = 6$)	72.57	94.95	69.51
TSAA (SD)	72.57	94.95	69.51
TSAA (SD) w/o CA	70.57	95.47	67.77
w/o PCL	66.94	93.42	64.32
PCL → Triplet loss [11]	68.09	93.24	65.41
PCL → Center loss [37]	71.48	94.60	69.13
PCL → ICA&CCA [7]	69.55	95.21	70.01
CAL (Ours)	**74.66**	**96.47**	**71.73**

beddings that have consistent semantics. We conduct experiments by removing or replacing PCL with other alternatives, including the prevailing metric learning losses [11, 37] and the one [7] designed specifically for VI Re-ID. The experimental results reported in Table 4 (bottom) show that our PCL contributes largely to the final performance and surpasses other competitors. This is mainly owing to the superior design of PCL that consider reducing intra-class variation using the part-aligned embeddings.

Analysis on MatchDistill. We conduct ablation experiments to evaluate the contribution of the proposed Cross-modal Query Matching algorithm and the two knowledge distillation losses ($\mathcal{L}_{kd1}$ and $\mathcal{L}_{kd2}$) in MatchDistill. As shown in Table 5, the experimental results demonstrate the significant role they play in enhancing retrieval accuracy. When aligning the attention maps with $\mathcal{L}_{kd1}$ and constraining the consistency of classification probability with $\mathcal{L}_{kd2}$, the performance boosts significantly, verifying the effectiveness of our knowledge exchange strategies. When incorporating our CQM, the accuracy can be further improved. This suggests that it is beneficial to ensure semantic consistency when conducting knowledge exchange by finding the best-matched part queries with our CQM algorithm.

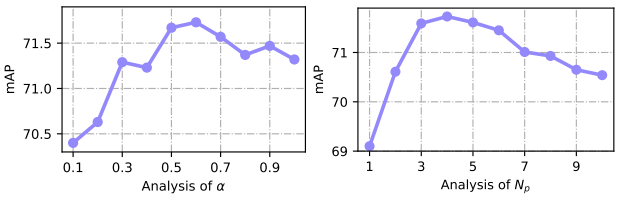

Figure 3. Parameter analysis of N_p and α.

Table 5. Effectiveness of each component of MatchDistill on the SYSU-MM01 dataset under the all-search single-shot mode.

$\mathcal{L}_{kd1}$	$\mathcal{L}_{kd2}$	CQM	SYSU-MM01		
			R1 ↑	R10 ↑	mAP ↑
			70.15	93.12	67.40
✓			72.75	95.71	69.86
	✓		72.26	95.16	68.50
✓	✓		73.63	95.79	70.61
✓	✓	✓	**74.66**	**96.47**	**71.73**

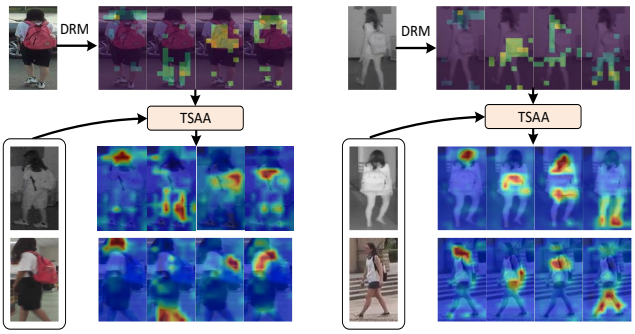

Figure 4. Qualitative analysis of the proposed DRM and TSAA.

Qualitative Results. We visualize the regions selected by the DRM module in Figure 4. It can be seen that the DRM can select diverse and discriminative regions, which could serve as strong guidance for the subsequent TSAA module. We also visualize the attention scores generated by the TSAA in Figure 4. We can see that the proposed TSAA has a strong capability in searching the corresponding regions given the key regions selected by DRM.

5. Conclusion

In this paper, we present the Concordant Attention Learning (CAL) framework, which learns concordant attention across the visible and infrared modalities and can effectively bridge the modality gap for VI Re-ID. We show that it is beneficial to design target-aware attention mechanisms to ensure attention concordance when aligning cross-modal samples. We also verify that exploiting cross-modal clues from the target modality and enabling cross-modal knowledge exchange in a match-and-distill manner can facilitate the learning of inter-modal correlations. Extensive experiments demonstrate the effectiveness and superiority of CAL over state-of-the-art methods, suggesting the potential of our approach for improving VI Re-ID research.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No.62073004), Science and Technology Plan of Shenzhen (No.JCYJ20200109140410340), and Shenzhen Fundamental Research Program (No.GXWD20201231165807007-20200807164903001).

References

[1] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-Temporal Attention-Aware Learning for Video-Based Person Re-Identification. *IEEE TIP*, 28(9):4192–4205, Sept. 2019. 2, 3

[2] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural Feature Search for RGB-Infrared Person Re-Identification. In *CVPR*, pages 587–597, June 2021. 7

[3] Xing Fan, Wei Jiang, Hao Luo, and Weijie Mao. Modality-Transfer Generative Adversarial Network and Dual-Level Unified Latent Representation for Visible Thermal Person Re-Identification. *The Visual Computer*, 38(1):279–294, Jan. 2022. 2

[4] Yujian Feng, Jian Yu, Feng Chen, Yimu Ji, Fei Wu, Shangdon Liu, and Xiao-Yuan Jing. Visible-Infrared Person Re-Identification via Cross-Modality Interaction Transformer. *IEEE TMM*, pages 1–13, 2022. 2

[5] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE TIP*, 29:579–590, 2020. 2

[6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov. 2020. 2, 4

[7] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In *ICCV*, pages 16403–16412, 2021. 7, 8

[8] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. Dual-Alignment Feature Embedding for Cross-Modality Person Re-Identification. In *ACM MM*, pages 57–65, 2019. 2, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, June 2016. 6

[10] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based Object Re-Identification. In *ICCV*, pages 15013–15022, Oct. 2021. 1

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv:1703.07737 [cs]*, Nov. 2017. 8

[12] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-Modality Transformer for Visible-Infrared Person Re-Identification. In *ECCV*, volume 13674, pages 480–496. 2022. 2, 6, 7

[13] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Style Normalization and Restitution for Domain Generalization and Adaptation. *IEEE TMM*, 24:3636–3651, 2022. 2, 5

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style Normalization and Restitution for Generalizable Person Re-Identification. In *CVPR*, pages 3143–3152, 2020. 2, 5

[15] Harold W Kuhn. The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-identification. In *CVPR*, pages 2285–2294, June 2018. 1, 2, 3

[17] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification. In *ECCV*, volume 13686, pages 381–398. 2022. 2, 7

[18] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. In *CVPR*, pages 2897–2906, Nashville, TN, USA, June 2021. 2, 3

[19] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose Transferrable Person Re-identification. In *CVPR*, pages 4099–4108, June 2018. 2, 3

[20] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Identification. In *CVPR*, pages 19366–19375, June 2022. 2, 3, 6, 7

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*, 2017. 6

[22] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-Modality Person Re-Identification with Shared-Specific Feature Transfer. In *CVPR*, pages 13376–13386, June 2020. 2, 7

[23] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In *ICCV*, pages 542–551, Oct. 2019. 1

[24] Ziling Miao, Hong Liu, Wei Shi, Wanlu Xu, and Hanrong Ye. Modality-Aware Style Adaptation for RGB-Infrared Person Re-Identification. In *IJCAI*, pages 916–922, 2021. 2

[25] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors*, 17(3):605, Mar. 2017. 2, 6

[26] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-Normalized Image Generation for Person Re-identification. In *ECCV*, volume 11213, pages 661–678. 2018. 2, 3

[27] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-identification. In *CVPR*, pages 5363–5372, June 2018. 2, 3

[28] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-Guided Contrastive Attention Model for Person Re-identification. In *CVPR*, pages 1179–1188, Salt Lake City, UT, June 2018. 2, 3

[29] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*, volume 11208, pages 501–518, 2018. 1, 2, 3

[30] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification. In *CVPR*, pages 1522–1531, June 2021. 2

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. In *CVPR*, pages 4105–4113, July 2017. 5

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, volume 30, 2017. 4

[33] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification. In *ECCV*, volume 11208, pages 384–400. 2018. 2, 3

[34] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *ICCV*, pages 3622–3631, Oct. 2019. 2

[35] Pingyu Wang, Zhicheng Zhao, Fei Su, Yanyun Zhao, Haiying Wang, Lei Yang, and Yang Li. Deep Multi-Patch Matching Network for Visible Thermal Person Re-Identification. *IEEE TMM*, 23:1474–1488, 2021. 7

[36] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic Modality Collaborative Learning for Visible Infrared Person Re-Identification. In *ICCV*, pages 225–234, Oct. 2021. 7

[37] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In *ECCV*, volume 9911, pages 499–515. 2016. 8

[38] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-Infrared Cross-Modality Person Re-identification. In *ICCV*, pages 5390–5399, Oct. 2017. 2, 6, 7

[39] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *CVPR*, pages 4330–4339, 2021. 2, 5, 7

[40] Hanrong Ye, Hong Liu, Fanyang Meng, and Xia Li. Bi-Directional Exponential Angular Triplet Loss for RGB-Infrared Person Re-Identification. *IEEE TIP*, 30:1583–1595, 2021. 2

[41] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic Tri-Level Relation Mining with Attentive Graph for Visible Infrared Re-Identification. *IEEE TIFS*, 2021. 3

[42] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE TIFS*, 15:407–419, 2019. 2

[43] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *ICCV*, pages 13567–13576, 2021. 2, 7

[44] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE TPAMI*, 44(6):2872–2893, June 2022. 2, 3, 7

[45] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *IJCAI*, pages 1092–1099, July 2018. 2

[46] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. FMCNet: Feature-level Modality Compensation for Visible-Infrared Person Re-Identification. In *CVPR*, pages 7349–7358, June 2022. 7

[47] Jiaqi Zhao, Hanzheng Wang, Yong Zhou, Rui Yao, Silin Chen, and Abdulmotaleb El Saddik. Spatial-Channel Enhanced Transformer for Visible-Infrared Person Re-Identification. *IEEE TMM*, pages 1–1, 2022. 3

[48] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person Re-identification in the Wild. In *CVPR*, pages 3346–3355, 2017. 1

[49] Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. Grayscale Enhancement Colorization Network for Visible-Infrared Person Re-Identification. *IEEE TCSVT*, 32(3):1418–1430, Mar. 2022. 2