Cross-modality Person re-identification with Shared-Specific Feature Transfer

Yan Lu[1,2], Yue Wu[3], Bin Liu[1,2,✉], Tianzhu Zhang[1], Baopu Li[4], Qi Chu[1,2], Nenghai Yu[1,2]
[1]School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[2]Key Laboratory of Electromagnetic Space Information, Chinese Academy of Science, Hefei, China
[3]Damo Academy, Alibaba Group, Beijing, China
[4]Baidu Research(USA), 1195 Baudeaux Dr, Sunnyvale, CA, USA
luyan17@mail.ustc.edu.cn, matthew.wy@alibaba-inc.com, baopuli@baidu.com
{flowice, tzzhang, qchu, ynh}@ustc.edu.cn

Abstract

Cross-modality person re-identification (cm-ReID) is a challenging but key technology for intelligent video analysis. Existing works mainly focus on learning modality-shared representation by embedding different modalities into a same feature space, lowering the upper bound of feature distinctiveness. In this paper, we tackle the above limitation by proposing a novel cross-modality shared-specific feature transfer algorithm (termed cm-SSFT) to explore the potential of both the modality-shared information and the modality-specific characteristics to boost the re-identification performance. We model the affinities of different modality samples according to the shared features and then transfer both shared and specific features among and across modalities. We also propose a complementary feature learning strategy including modality adaption, project adversarial learning and reconstruction enhancement to learn discriminative and complementary shared and specific features of each modality, respectively. The entire cm-SSFT algorithm can be trained in an end-to-end manner. We conducted comprehensive experiments to validate the superiority of the overall algorithm and the effectiveness of each component. The proposed algorithm significantly outperforms state-of-the-arts by 22.5% and 19.3% mAP on the two mainstream benchmark datasets SYSU-MM01 and RegDB, respectively.

1. Introduction

Person re-identification (ReID) aims to find out images of the same person to the query image from a large gallery. Many works focus on feature learning [17, 37] and metric learning [3, 24] on the RGB modality. These methods have achieved great success, especially with the most recent deep learning technology [39]. However, the dependency on bright lighting environments limits their applications in real complex scenarios. The performance of these methods degrades dramatically in dark environments where most cameras cannot work well [47]. Hence, other kinds of visual sensors like infrared cameras are now widely used as a complement to RGB cameras to overcome these difficulties, yielding popular research interest on RGB-Infrared cross-modality person ReID (cm-ReID).

Compared to conventional ReID task, the major difficulty of cm-ReID is the modality discrepancy resulting

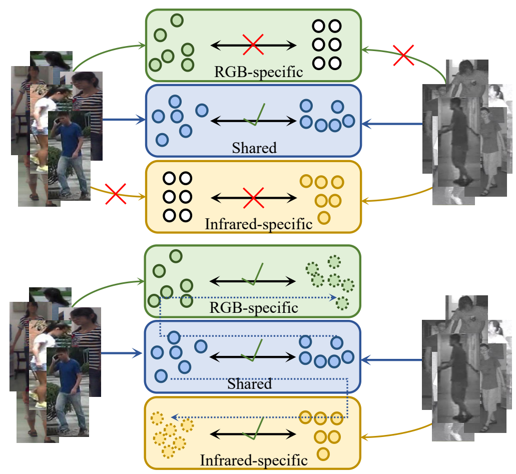

Figure 1. Illustration of the difference between our algorithm and modality-shared feature learning methods. The modality-shared feature learning methods abandon lots of useful specific cues because the modality-specific information cannot be extracted from the other modality. Our algorithm tries to introduce modality-specific features based on the cross-modality near neighbor affinity modeling, effectively utilizing both shared and specific information for each sample.

✉Bin Liu is the corresponding author.

from intrinsically distinct imaging processes of different cameras. Some discriminative cues like colors in RGB images are missing in infrared images. Previous methods can be summarized into two major categories to overcome the modality discrepancy: modality-shared feature learning and modality-specific feature compensation. The shared feature learning aims to embed images of whatever modality into a same feature space [47, 50, 51]. The specific information of different modalities such as colors of RGB images and thermal of infrared images are eliminated as redundant information [4]. However, the specific information like colors plays an important role in conventional ReID. With shared cues only, the upper bound of the discrimination ability of the feature representation is limited. As a result, modality-specific feature compensation methods try to make up the missing specific information from one modality to another. Dual-level Discrepancy Reduction Learning (D^2RL) [45] is the typical work to generate multi-spectral images to compensate for the lacking specific information by utilizing the generative adversarial network (GAN) [8]. However, a person in the infrared modality can have different colors of clothes in the RGB space. There can be multiple reasonable results for image generation. It's hard to decide which one is the correct target to be generated for re-identification without memorization of the limited gallery set.

In this paper, we tackle the above limitations by proposing a novel cross-modality shared-specific feature transfer algorithm (termed cm-SSFT) to explore the potential of both the modality-shared information and the modality-specific characteristics to boost the re-identification performance. It models the affinities between intra-modality and inter-modality samples and utilizes them to propagate information. Every sample accepts the information from its inter-modality and intra-modality near neighbors and meanwhile shares its own information with them. This scheme can compensate for the lack of specific information and enhance the robustness of the shared feature, thus improving the overall representation ability. Comparison with the shared feature learning methods are shown in Figure 1. Our method can exploit the specific information that is unavailable in traditional shared feature learning. Since our method is dependent on the affinity modeling of neighbors, the compensation process can also overcome the choice difficulty of generative methods. Experiments show that the proposed algorithm can significantly outperform state-of-the-arts by 22.5% and 19.3% mAP, as well as 19.2% and 14.4% Rank-1 accuracy on the two most popular benchmark datasets SYSU-MM01 and RegDB, respectively.

The main contributions of our work are as follows:

- We propose an end-to-end cross-modality shared-specific feature transfer (cm-SSFT) algorithm to utilize both the modality shared and specific information, achieving the state-of-the-art cross-modality person ReID performance.
- We put forward a feature transfer method by modeling the inter-modality and intra-modality affinity to propagate information among and across modalities according to near neighbors, which can effectively utilize the shared and specific information of each sample.
- We provide a novel complementary learning method to extract discriminative and complementary shared and specific features of each modality, respectively, which can further enhance the effectiveness of the cm-SSFT.

2. Related Work

Person ReID. Person ReID [53] aims to search target person images in a large gallery set with a query image. The recent works are mainly based on deep learning for more discriminative features [6, 18, 49, 56]. Some of them treat it as a partial feature learning task and pay much attention to more powerful network structures to better discover, align, and depict the body parts [10, 38, 39, 26]. Other methods are based on metric learning, focusing on proper loss functions, like the contrastive loss [40], triplet loss [17], quadruplet loss [2], etc. Both kinds of methods try to discard the unrelated cues, such as pose, viewpoint and illumination changing out of the features and the metric space. Recent disentangle based methods extend along this direction further by splitting each sample to identity-related and identitiy-unrelated features, obtaining purer representations without redundant cues [12, 54].

The aforementioned methods process each sample independently, ignoring the connections between person images. Recent self-attention [41, 29] and graph-based methods [1, 34, 35, 48] tried to model the relationship between sample pairs. Luo *et al.* proposed the spectral feature transformation method to fuse features between different identities [29]. Shen *et al.* proposed a similarity guided graph neural network [35] and deep group-shuffling random walk [34] to fuse the residual features of different samples to obtain more robust representation. Liu *et al.* utilized the near neighbors to tackle the unsupervised ReID [28].

Cross-modality matching. Cross-modality matching aims to match samples from different modalities, such as cross-modality retrieval [9, 16, 21, 27] and cross-modality tracking [57]. Cross-modality retrieval has been widely studied for heterogeneous face recognition [15] and text-to-image retrieval [9, 16, 21, 22, 27]. [15] proposed a two-stream based deep invariant feature representation learning method for heterogeneous face recognition.

Cross-modality person ReID. Cross-modality person ReID aims to match queries of one modality against a gallery set of another modality [44], such as text-image ReID [23, 32, 33], RGB-Depth ReID [11, 46] and RGB-Infrared (RGB-IR) ReID [4, 7, 13, 19, 20, 25, 42, 43, 45,

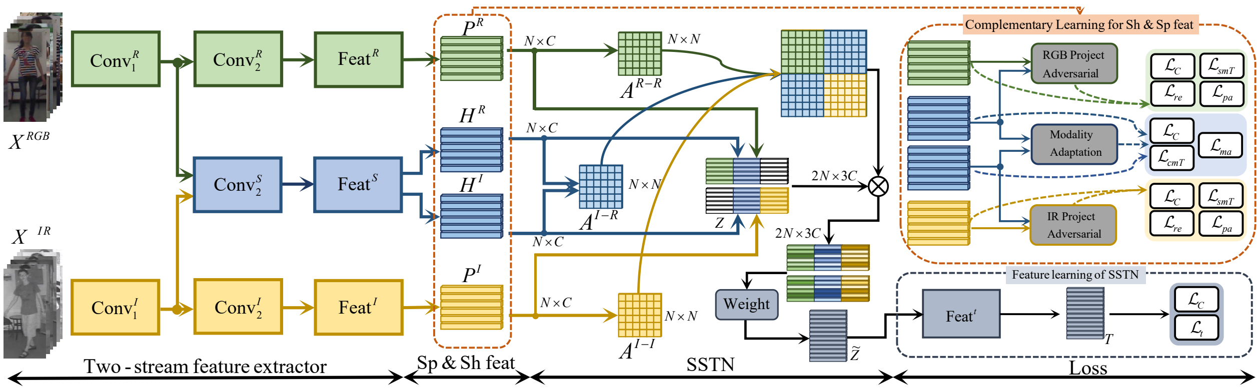

Figure 2. Framework of the cross-modality shared-specific feature transfer algorithm.

47, 50, 51, 52]. Wu *et al.* built the largest SYSU-MM01 dataset for RGB-IR person ReID evaluation [47]. Ye *et al.* advanced a two-stream based model and bi-directional top-ranking loss function for the shared feature embedding [50, 51]. To make the shared features purer, Dai *et al.* suggested a generative adversarial training method for the shared feature learning [4]. These methods only concentrate on the shared feature learning and ignore the potential values of specific features. Accordingly, some other works try to utilize modality-specific features and focus on cross-modality GAN. Kniaz *et al.* proposed ThermalGAN to transfer RGB images to IR images and extracted features in IR domain [20]. Wang *et al.* put forward dual-level discrepancy reduction learning based on a bi-directional cycle GAN to reduce the gap between different modalities [45]. More recently, Wang *et al.* [42] constructed a novel GAN model with the joint pixel-level and feature-level constraint, which achieved the state-of-the-art performance. However, it is hard to decide which one is the correct target to be generated from the multiple reasonable choices for ReID.

3. Cross-Modality Shared-Specific Feature Transfer

The framework of the proposed cross-modality shared-specific feature transfer algorithm (cm-SSFT) is shown in Figure 2. Input images are first fed into the two-stream feature extractor to obtain the shared and specific features. Then the shared-specific transfer network (SSTN) models the intra-modality and inter-modality affinities. It then propagates the shared and specific features across modalities to compensate for the lacked specific information and enhance the shared features. To obtain discriminative and complementary shared and specific features, two project adversarial and reconstruction blocks and one modality-adaptation module are added on the feature extractor. The overall algorithm is trained in an end-to-end manner.

To better illustrate how the proposed algorithm works, we distinguish the RGB modality, infrared modality and shared space with R, I and S in superscript. We use H and P to denote sHared and sPecific features, respectively.

3.1. Two-stream feature extractor

As shown in Figure 2, our two-stream feature extractor includes the modality-shared stream (in blue blocks) and the modality-specific stream (green blocks for RGB and yellow blocks for IR). Each input image X^m ($m \in \{R, I\}$) will pass the convolutional layers and the feature blocks to generate the shared feature and specific feature. For better performances, we separate the shared and specific stream at the shallow convolutional layers instead of the deeper fully-connected layers [50]:

$$\begin{aligned} H^m &= \text{Feat}^S(\text{Conv}_2^S(\text{Conv}_1^m(X^m))), \\ P^m &= \text{Feat}^m(\text{Conv}_2^m(\text{Conv}_1^m(X^m))). \end{aligned} \tag{1}$$

To make sure that the two kinds of features are both discriminative, we add the classification loss $\mathcal{L}_c$ on each kind of features respectively:

$$\begin{aligned} \mathcal{L}_c(H^m) &= \mathbb{E}_{i,m}[-\log(p(y_i^m|H_i^m))], \\ \mathcal{L}_c(P^m) &= \mathbb{E}_{i,m}[-\log(p(y_i^m|P_i^m))], \end{aligned} \tag{2}$$

where $p(y_i^m|*)$ is the predicted probability of belonging to the ground-truth class y_i^m for the input image X^m. The classification loss ensures that features can distinguish the identities of the inputs. Besides, we add a single modality triplet loss ($\mathcal{L}_{smT}$) [17] on specific features and a cross-modality triplet loss ($\mathcal{L}_{cmT}$) [4, 51] on shared features for better discriminability:

$$\begin{aligned} \mathcal{L}_{smT}(P) = &\sum_{i,j,k} \max[\rho_2 + ||P_i^R - P_j^R|| - ||P_i^R - P_k^R||, 0] \\ &+ \sum_{i,j,k} \max[\rho_2 + ||P_i^I - P_j^I|| - ||P_i^I - P_k^I||, 0], \end{aligned} \tag{3}$$

$$\begin{aligned} \mathcal{L}_{cmT}(H) = &\sum_{i,j,k} \max[\rho_1 + ||H_i^R - H_j^I|| - ||H_i^R - H_k^I||, 0] \\ &+ \sum_{i,j,k} \max[\rho_1 + ||H_i^I - H_j^R|| - ||H_i^I - H_k^R||, 0], \end{aligned} \tag{4}$$

where ρ_1 and ρ_2 are the margins of $\mathcal{L}_{cmT}$ and $\mathcal{L}_{smT}$, respectively. i, j, k represent indices of the anchor, positive of the anchor and negative of the anchor of triplet loss $(y_i = y_j, y_i \neq y_k)$.

3.2. Shared-Specific Transfer Network

The two-stream network extracts the shared and specific features for each modality. For unified feature representation, we pad and denote the features of each modality with a three-segment format: [RGB-specific; shared; Infrared-specific] as follows:

$$Z_i^R = [P_i^R; H_i^R; \mathbf{0}], \quad Z_i^I = [\mathbf{0}; H_j^I; P_j^I]. \tag{5}$$

Here, $\mathbf{0}$ denotes the padding zero vector, which means that samples of the RGB modality have no specific features of infrared modality, and vice versa. $[\bullet; \bullet]$ means concatenation in the columan dimension. For cross-modality retrieval, we need to transfer the specific features from one modality to another to compensate for these zero-padding vectors. Motivated by graph convolutional network (GCN), we utilize the near neighbors to propagate information and meanwhile maintain the context structure of the overall sample space. The proposed shared-specific transfer network can make up the lacking specific features and enhance the robustness of the overall representation jointly. As shown in Figure 2, SSTN first models the affinity of samples according to the two kinds of features. Then it propagates both intra-modality and inter-modality information with the affinity model. Finally, the feature learning stage guides the optimization of the whole process with classification and triplet losses.

Affinity modeling. We use the shared and specific features to model the pair-wise affinity. We take the specific features to compute the intra-modality affinity and the shared features for inter-modality as follows:

$$A_{ij}^{m,m} = d(P_i^m, P_j^m), \quad A_{ij}^{m,m'} = d(H_i^m, H_j^{m'}), \tag{6}$$

where $A_{ij}^{m,m}$ is the intra-modality affinity between the i-th sample and the j-th sample, both of which belong to the m modality. $A_{ij}^{m,m'}$ is the inter-affinity. $d(a,b)$ is the normalized euclidean distance metric function:

$$d(a,b) = 1 - 0.5 \cdot \left\| \frac{a}{||a||} - \frac{b}{||b||} \right\|. \tag{7}$$

The intra-similarity and inter-similarity represent the relation between each sample with others of both the same and different modalities. We define the final affinity matrix as:

$$A = \left[\begin{array}{c:c} \mathcal{T}(A^{R,R}, k) & \mathcal{T}(A^{R,I}, k) \\ \hdashline \mathcal{T}(A^{I,R}, k) & \mathcal{T}(A^{I,I}, k) \end{array} \right], \tag{8}$$

where $\mathcal{T}(\bullet, k)$ is the near neighbor chosen function. It keeps the top-k values for each row of a matrix and sets the others to zero.

Shared and specific information propagation. The affinity matrix represents the similarities across samples. SSTN utilizes this matrix to propagate features. Before this, features of the RGB and infrared modalities are concatenated in the row dimension, each row of which stores a feature of a sample:

$$Z = \left[\begin{array}{c} Z^R \\ \hdashline Z^I \end{array} \right]. \tag{9}$$

Following the GCN approach, we obtain the diagonal matrix D of the affinity matrix A with $d_{ii} = \sum_j A_{ij}$. The padded features are first propagated with the near neighbor structure $(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} Z)$ and then fused by a learnable non-linear transformation. After feature fusion, the propagated features will include shared features and specific features of both the two modalities. The propagated features $\widetilde{Z}$ are calculated as:

$$\widetilde{Z} = \left[\begin{array}{c} \widetilde{Z}^R \\ \hdashline \widetilde{Z}^I \end{array} \right] = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} Z W), \tag{10}$$

where σ is the activation function which is ReLU in our implementation. W is the learnable parameters of SSTN. These propagated features are finally fed into a feature learning stage to optimize the whole learning process. The transferred features T are denoted as:

$$T = \left[\begin{array}{c} T^R \\ \hdashline T^I \end{array} \right] = \mathrm{Feat}^t(\widetilde{Z}). \tag{11}$$

Following the common feature learning principle, we use the classification loss for feature learning:

$$\mathcal{L}_c(T^m) = \mathbb{E}_{i,m}[-\log(p(y_i^m | T_i^m))]. \tag{12}$$

In additoin, we use the triplet loss on the transferred feature to increase the discrimination ability. Since the transferred features include both shared features and specific features of two modalities. We add both the cm-triplet loss $\mathcal{L}_{cmT}(T)$ and sm-triplet loss $\mathcal{L}_{smT}(T)$ on it for better discrimination:

$$\begin{aligned} \mathcal{L}_t(T) &= \mathcal{L}_{cmT}(T) + \mathcal{L}_{smT}(T) \\ &= \sum_{i,j,k} \max[\rho_1 + ||T_i^R, T_j^I|| - ||T_i^R, T_k^I||, 0] \\ &+ \sum_{i,j,k} \max[\rho_1 + ||T_i^I, T_j^R|| - ||T_i^I, T_k^R||, 0] \\ &+ \sum_{i,j,k} \max[\rho_2 + ||T_i^R, T_j^R|| - ||T_i^R, T_k^R||, 0] \\ &+ \sum_{i,j,k} \max[\rho_2 + ||T_i^I, T_j^I|| - ||T_i^I, T_k^I||, 0]. \end{aligned} \tag{13}$$

3.3. Shared and specific complementary learning

SSTN explores a new way to utilize both shared the specific features to generate more discriminative representation. However, the overall performance may still suffer from the information overlap between shared and specific features. Firstly, if shared features contain much modality-specific information, the reliability of the inter-similarity matrix in equation (6) will be affected, leading to inaccurate feature transfer. Secondly, if the specific features are highly related to the shared features, the specific features can only provide little complement to the shared features. The redundant information in the specific features will also affect the sensitivity of the intra-modality similarity matrix in equation (6) due to the shared information. To alleviate these two problems, we utilize the modality adaptation [4] to filter out modality-specific information from the shared features. We also propose a project adversarial strategy and reconstruction enhancement for complementary modality-specific feature learning.

Modality adaptation for shared features. To purify the shared features to be unrelated to modalities, we utilize the modality discriminator [4] with three fully-connected layers to classify the modality of each shared feature:

$$\mathcal{L}_{ma} = \mathbb{E}_{i,m}[-\log(p(m|H_i^m, \Theta_D))], \tag{14}$$

where Θ_D represents parameters of the modality discriminator. $p(m|H_i^m)$ is the predicted probability of feature H_i^m belonging to modality m. In the discrimination stage, the modality discriminator will try to classify the modality of each shared feature. In the generation stage, the backbone network will generate features to fool the discriminator. This min-max game will make the shared features not contain any modality-related information.

Project adversarial learning for specific features. To make the specific features uncorrelated with the shared features, we propose the project adversarial strategy. In the training stage, we project the specific features to the shared features of the same sample. The projection error is used as the loss function

$$\mathcal{L}_{pa} = \mathbb{E}_{i,m}\left[\|\Theta^m \cdot P_i^m - H_i^m\|\right], \tag{15}$$

where Θ_p^m represents the projection matrix for modality m. In this equation, "$\cdot$" means matrix multiply. Similarly, in the discrimination stage, optimization of Θ_p^m will try to project the specific features to the corresponding shared features. While in the generation stage, the backbone network will generate specific features uncorrelated with shared features to fool the projection. This adversarial training can make the feature spaces of the two kinds of features linearly independent. Alternatively minimizing and maximizing the projection loss will lead the backbone network to learn specific patterns different from shared features.

Reconstruction enhancement. Modality adaption and project adversarial learning make sure that the shared and specific features do not contain correlated information between each other. To enhance both features to be complementary, we use a decoder network after features of each modality to reconstruct the inputs. We concatenate the shared and specific features and feed them to the decoder $\mathcal{D}e$:

$$\hat{X}^m = \mathcal{D}e^m([P^m; H^m]), \tag{16}$$

where $[\bullet;\bullet]$ means feature concatenation. The L_2 loss is used to evaluate the quality of the reconstructed images:

$$\mathcal{L}_{re} = \mathbb{E}_{i,m}[L_2(X_i^m, \hat{X}_i^m)]. \tag{17}$$

The reconstruction task makes a constraint on the overall information loss. Combined with project modality adaption and adversarial learning, shared and specific features are guided to be self-discriminate and mutual-complementary.

3.4. Optimization

Our proposed algorithm is trained in an end-to-end manner with the adversarial min-max games. We mix the loss function based on the principle that the classification and the triplet share the same importance. So the feature learning losses of each part are as follows:

$$\begin{aligned}
&\mathcal{L}(H) = \mathcal{L}_c(H^m) + 0.5 \cdot \mathcal{L}_{cmT}, \\
&\mathcal{L}(P) = 0.5 \cdot (\mathcal{L}_c(P^R) + \mathcal{L}_c(P^I)) + 0.5 \cdot \mathcal{L}_{smT}, \\
&\mathcal{L}(T) = \mathcal{L}_c(T) + 0.25 \cdot \mathcal{L}_t(T).
\end{aligned} \tag{18}$$

Furthermore, we think that the backbone feature extractor and SSTN share the same importance. Hence, the overall feature learning loss is as follows:

$$\mathcal{L}_{feat} = \mathcal{L}(H) + \mathcal{L}(P) + \mathcal{L}(T). \tag{19}$$

Therefore, the overall loss functions of the min and the max steps of each part are as follows:

$$\begin{aligned}
\mathcal{L}_{min} =& \mathcal{L}_{feat} + \lambda_1 \mathcal{L}_{re} - \lambda_2 \mathcal{L}_{ma} - \lambda_3 \mathcal{L}_{pa}, \\
\mathcal{L}_{max} =& -\lambda_2 \mathcal{L}_{ma} - \lambda_3 \mathcal{L}_{pa}.
\end{aligned} \tag{20}$$

The optimization process includes two sub-processes: (1) fix each discriminator and minimize $\mathcal{L}_{min}$. (2) fix all modules excluding the three discriminators and maximize the $\mathcal{L}_{max}$. Support Θ_N denotes the parameters of the overall networks except all the other discriminators. The alternative learning process is:

$$\begin{aligned}
\hat{\Theta}_N &= \arg\min_{\Theta_N} \mathcal{L}_{min}(\Theta_N, \hat{\Theta}_D, \hat{\Theta}^m), \\
\hat{\Theta}_D, \hat{\Theta}^m &= \arg\max_{\Theta_D, \Theta^m} \mathcal{L}_{max}(\hat{\Theta_N}, \Theta_D, \Theta^m).
\end{aligned} \tag{21}$$

In order to ensure the training effectiveness, every batch contains the equal number of RGB and infrared samples. The details of the sampling strategy are introduced in the implementation details. In the test stage, we utilize the two-stream network to extract disentangled features from the RGB set and the infrared set. We use SSTN to transfer modality-shared and modality-specific features. All features are L_2-normalized and we use the Euclidean distance to compute the final ReID performance.

Table 1. Comparison on SYSU-MM01. r1, r10, r20 denote Rank-1, 10, 20 accuracies (%) .

Method	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
HOG[5]	2.76	18.3	31.9	4.24	3.82	22.8	37.6	2.16	3.22	24.7	44.5	7.25	4.75	29.2	49.4	3.51
LOMO[24]	3.64	23.2	37.3	4.53	4.70	28.2	43.1	2.28	5.75	34.4	54.9	10.2	7.36	40.4	60.3	5.64
Zero-Padding[47]	14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
TONE+HCML[50]	14.3	53.2	69.2	16.2	-	-	-	-	-	-	-	-	-	-	-	-
BDTR[51]	17.0	55.4	72.0	19.7	-	-	-	-	-	-	-	-	-	-	-	-
D-HSME[13]	20.7	62.8	78.0	23.2	-	-	-	-	-	-	-	-	-	-	-	-
IPVT+MSR[19]	23.2	51.2	61.7	22.5	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN[4]	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
D^2RL[45]	28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
DGD+MSR[7]	37.4	83.4	93.3	38.1	43.9	86.9	95.7	30.5	39.6	89.3	97.7	50.9	46.6	93.6	98.8	40.1
JSIA-ReID[43]	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
AlignGAN[42]	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
cm-SSFT (Ours)	**61.6**	**89.2**	**93.9**	**63.2**	**63.4**	**91.2**	**95.7**	**62.0**	**70.5**	**94.9**	**97.7**	**72.6**	**73.0**	**96.3**	**99.1**	**72.4**

4. Experiments

In this section, we conduct comprehensive experiments to validate the effectiveness of the proposed cross-modality shared-specific feature transfer algorithm as well as each of its components.

4.1. Experimental settings

Datasets. SYSU-MM01 is a large-scale and frequently used RGB-IR cross-modality ReID dataset [47]. Images are collected from four RGB cameras and two IR cameras, in both indoor and outdoor environments. The training set contains 395 persons, with 22,258 RGB images and 11,909 IR images. The test set contains 96 persons, with 3,803 IR images for query and 301/3010 (one-shot/multi-shot) randomly selected RGB images as the gallery. There are two accordingly evaluation modes for RGB-IR ReID: *indoor-search* and *all-search* [47]. RegDB is collected by dual camera systems [31]. There are 412 identities and 8,240 images in total, with 206 identities for training and 206 identities for testing. Each identity has 10 different thermal (IR) images and 10 different visible (RGB) images. There are also two evaluation modes. One is Visible to Thermal to search IR images from a Visible image. The other mode is Thermal to Visible to search RGB images from a infrared image. This dataset has 10 trials with different splits of the dataset. We evaluate our model on the 10 trials to achieve statistically stable results.

Evaluation protocols. All the experiments follow the standard evaluation protocol in existing RGB-IR cross-modality ReID methods. Queries and galleries images are from different modalities. And then, the standard cumulated matching characteristics (CMC) curve and mean average precision (mAP) are adopted.

Implementation details. We use Resnet50 [14] as the backbone network, with the first convolutional layer, the 1st and 2nd bottlenecks as Conv_1. Conv_2 is the 3rd and 4th bottlenecks. k in Eq. (22) is set to 4. λ_1, λ_2 and λ_3 are set to 1.0, 0.2 and 0.2, respectively. We change the stride of the last convolutional layer in the backbone to 1 to benefit the learning of reconstruction decoders which are composed of 4 sub-pixel convolutional layers with channels all set to 64 [36]. We adopt the data and network augmentation methods in BoT for ReID [30] to enhance the performance. For fairness, we also give out results without any augmentation. The augmentations include: (1) the feature blocks are all set to BNNeck [30]; (2) the input images are augmented with random erasing [55]. The whole algorithm is optimized with Adam for 120 epochs with a batch size of 64 and a learning rate of 0.00035, decaying 10 times at 40, 70 epoch. Each mini-batch is comprised of 8 identities with 4 RGB images and 4 infrared images for each identity.

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

In this subsection, we compare our proposed algorithm with the baselines as well as the state-of-the-art methods, including Zero-Padding [47], TONE [50], BDTR [51], cmGAN [4], D^2RL[45], MSR[7], D-HSME[13], IPVT[19], JSIA-ReID[43] and AlignGAN[42].

The results on SYSU-MM01 are shown in Table 1. The proposed algorithm outperforms other methods by a large margin. Specifically, in all-search mode, our method surpasses AlignGAN by 19.2% on Rank-1 accuracy and 22.5% on mAP in the single-shot setting. The multi-shot setting exhibits a similar phenomenon. Compared with single-shot evaluation, mAP of most other methods drop significantly by about 5% or even more. But our method only drops 1.2%. This validates that the features extracted by our algorithm are much more discriminative, which can provide higher recall than other methods when the gallery size increases. For indoor-search mode, our method also gets the best performance on all the evaluation metrics, demonstrating the robustness of the proposed algorithm.

The results on RegDB are shown in 2. Our method always suppresses others by a large margin. For the Visible to Thermal mode, our method surpasses the state-of-the-art method by 14.4% on Rank-1 and 19.3% on mAP. For Thermal to Visible, the advantages are 14.7% on Rank-1 and 18.3% on mAP.

Table 2. Comparison on RegDB.

Method	*Visible to Thermal*		*Thermal to Visible*	
	r1	mAP	r1	mAP
HOG[5]	13.5	10.3	-	-
LOMO[24]	0.80	2.28	-	-
Zero-Padding[47]	17.8	18.9	16.7	17.9
TONE+HCML[50]	24.4	20.8	21.7	22.2
BDTR[51]	33.5	31.8	32.7	31.1
D^2RL[45]	43.4	44.1	-	-
DGD+MSR[7]	48.4	48.7	-	-
JSIA-ReID[43]	48.5	49.3	48.1	48.9
D-HSME[13]	50.9	47.0	50.2	46.2
IPVT+MSR[19]	58.8	47.6	-	-
AlignGAN[42]	57.9	53.6	56.3	53.4
cm-SSFT (Ours)	**72.3**	**72.9**	**71.0**	**71.7**

Table 3. Ablation study on RegDB.

	ShL	SpL	SaS	MoA	PA	RE	ShT	SpT	r1	mAP
1	$\checkmark$	-	-	-	-	-	-	-	42.4	45.0
2	$\checkmark$	$\checkmark$	-	-	-	-	-	-	48.1	49.3
3	$\checkmark$	$\checkmark$	$\checkmark$	-	-	-	-	-	52.3	53.1
4	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	-	-	-	56.1	57.2
5	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	-	-	58.7	57.9
6	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	-	60.3	59.4
7	$\checkmark$	$\checkmark$	$\checkmark$	-	-	-	$\checkmark$	$\checkmark$	60.8	60.1
8	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	-	$\checkmark$	$\checkmark$	67.5	67.6
9	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	$\checkmark$	$\checkmark$	71.1	71.2
10	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	65.8	66.1
11	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	-	$\checkmark$	64.9	65.3
12	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	72.3	72.9

4.3. Ablation study

In this subsection, we study the effectiveness of each component of the proposed algorithm.

Effectiveness of structure of feature extractor. We first evaluate how much improvement can be made by the structure of feature extractor. We ablate the specific feature extraction stream and evaluate the performance of the shared features only to see the influence. The results are shown in the 1st and 2nd row of Table 3, represented as ShL (shared feature learning) and SpL (specific feature learning). The specific streams can bring about 5.7% increment of Rank-1 accuracy because they can back-propagate modality-specific gradients to the low-level feature maps. We also test the influences casued by separating streams at shallow layers. The result in 3rd (SaS: Separating at Shallow) shows that it can make bring 4.2% gains for the more dicriminative features.

Influence of complementary learning. We evaluate the effectiveness of each module in the complementary learning. Since the complementary learning can affect both the features of the feature extractor and SSTN, we design two sets of experiments to observe the impact respectively. The influences on the feature extractor are shown in rows 4∼6 of Table 3. The results of SSTN are shown in rows 7∼9. We can see that all modules (the modality-adaptation (MoA), the project adversarial (PA) and reconstruction enhancement (RE)) can make both backbone shared features and SSTN features more discriminative. The whole complementary learning scheme can bring about 8% and 12% increments for the feature extractor and SSTN, respectively.

Table 4. Performances without data or network augmentation.

Settings	MM01		RegDB	
	r1	mAP	r1	mAP
SOTA(AlignGAN)'s baseline	29.6	33.0	32.7	34.9
SOTA(AlignGAN)	42.4	40.7	57.9	53.6
baseline (wo aug)	25.5	27.2	29.5	30.8
cm-SSFT (wo aug)	52.4	52.1	62.2	63.0
baseline (w aug)	38.2	39.8	42.4	45.0
cm-SSFT (w aug)	61.6	63.2	72.3	72.9

Table 5. Performances comparison with single query.

Method	MM01				RegDB			
	S-shot		M-shot		*V-T*		*T-V*	
	r1	mAP	r1	mAP	r1	mAP	r1	mAP
Single query	47.7	54.1	57.4	59.1	65.4	65.6	63.8	64.2
All queries	61.6	63.2	63.4	62.0	72.3	72.9	71.0	71.7

Effectiveness of feature transfer. We aim to quantify the contribution of the proposed feature transfer strategy. Firstly, we want to know whether the proposed transfer method itself only works on shared features. By comparing row 6 with row 10 (only transfer the shared feature, defined as ShT) in Table 3, we can see that feature transfer brings in 5.5% Rank-1 and 6.7% mAP improvements. Secondly, we want to verify whether modality-specific features can positively contribute valuable information to the final representation. According to row 10 and row 12 (transfer both two kinds of features. SpT means transferring specific features.) of Table 3, we can see that the overall performance gains 6.5% and 6.8% increments on Rank-1 and mAP. For further verifying the effectiveness of the specific feature transfer, we also try only transferring the specific features. The results are shown in row 11 and show that only transferring the specific features can also achieve satisfy performances. The feature transfer stage not only contributes an overall 12.0% Rank-1 and 13.5% mAP improvements but also verifies that modality-specific features can be well-explored for better re-identification.

Influence of data and network augmentation. For fair comparison, we also give results without random-erasing in data augmentation. For each feature block, we also use a commonly used fully-connected layer to replace the BN-Neck. The results are shown in Table 4. It can be seen that, without the augmentation, our baseline is weaker than the baseline of the state-of-the-art (AlignGAN [42]) method (because we don't use dropout). But our model still can suppress SOTA by 10.0% on Rank-1 and 12.1% on mAP on the SYSU-MM01 dataset. On the RegDB dataset, our method can suppress 4.3% on Rank-1 and 9.4% on mAP. The data and network augmentations can bring about 13% increments on the backbone and 9% on our method. Without them, our model still achieves the state-of-the-art performances, proving the effectiveness of our method.

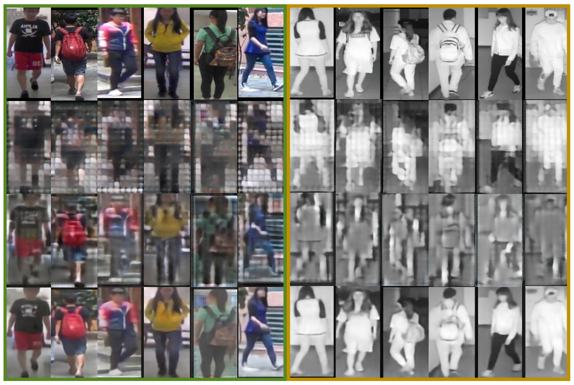

Figure 3. Reconstruction examples. The 1^{st} to 4^{th} rows correspond to original images, shared feature reconstructions, specific reconstructions and all feature reconstruction results, respectively.

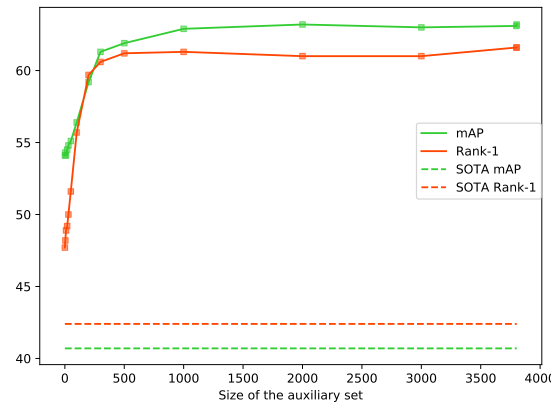

Figure 4. Influence of number of queries. Dashed lines correspond to the SOTA method. Solid lines correspond to ours.

4.4. Visualization of shared and specific features.

We take advantage of the reconstruction decoder to visualize the information of the modality shared and specific features. We remove P^m and H^m in Eq.(16) to observe the changes in the reconstructed images, respectively. The outputs are shown in Figure 3. We can see that shared feature reconstruction results are different and visually complementary to the specific features. For RGB images, the shared features contain less color information which is found in the images reconstructed by RGB-specific features. The specific features carried more color information but are less smooth. For infrared images, we can also observe that the specific features are different from the shared features. The combination of two kinds of features produces high-quality images. This proves that the shared and specific features produced by our feature extractor are complementary with each other.

4.5. Application in real scenarios

The SSTN in our cm-SSFT passes information between different modality samples. Every sample fuses the information from its inter-modality and intra-modality k near neighbors. Such setting hypothesizes that other query samples are treated as the auxiliary set. However, in some real application scenarios, there may be no or only a few auxiliary dates. In order to prove that our method is not limited in the experimental environments with some strong hypothesis, we show how to apply cm-SSFT to such single query scenarios, which also achieves state-of-the-art performances. We train the cm-SSFT algorithm exactly the same as illustrated in this paper. While in the testing stage, the SSTN only propagates information between only one query image with the gallery images. We slightly stabilize the affinity model A as follows:

$$Z = \left[\begin{array}{c} z^q \\ \hdashline Z^G \end{array}\right], A = \left[\begin{array}{c:c} k \cdot A^{q,q} & \mathcal{T}(A^{q,G}, k) \\ \hdashline k \cdot A^{G,q} & \mathcal{T}(A^{G,G}, k) \end{array}\right]. \tag{22}$$

It can be seen that we amplify k times the left column blocks of the affinity matrix, which is to balance the information of the two modalities. The experiments are shown in Table 5. The performance has dropped compared with all queries due to inadequate intra-modality specific information compensation. But our method still achieves better performances than state-of-the-arts and our baseline.

Besides, we also test the influence of the auxiliary set. The experiments are run on MM01 dataset for its large query set. We randomly sample n images from the query sets and watch the performance changing. For a specific n, we run 10 times to get the average performance. n is ranging from 1 (single query) to all query size. The results are shown in Figure 4. We can see that with the size of the auxiliary set growing, the performance saturates quickly.

5. Conclusion

In this paper, we proposed a cross-modality shared-specific feature transfer algorithm for cross-modality person ReID, which can utilize the specific features ignored by conventionally shared feature learning. It propagates information among and across modalities, which not only compensates for the lacking specific information but also enhances the overall discriminative. We also proposed a complementary learning strategy to learn self-discriminate and complementary feature. Extensive experiments validate the superior performance of the proposed algorithm, as well as the effectiveness of each component of the algorithm.

6. Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities (WK2100330002, WK3480000005) and the Major Scientific Research Project of Zhejiang Lab (No.2019DB0ZX01).

References

[1] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2017.

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[3] Ying-Cong Chen, Wei-Shi Zheng, and Jianhuang Lai. Mirror representation for modeling view-specific transform in person re-identification. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, pages 677–683, 2018.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[6] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8030–8039, 2019.

[7] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[10] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3642–3651, 2019.

[11] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018.

[12] Chanho Eom Ham et al. Learning disentangled representation for robust person re-identification. *arXiv preprint arXiv:1910.12003*, 2019.

[13] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8385–8392, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[15] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] Xiangteng He, Yuxin Peng, and Liu Xie. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1740–1748. ACM, 2019.

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[18] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019.

[19] Jin Kyu Kang, Toan Minh Hoang, and Kang Ryoung Park. Person re-identification between visible and thermal camera images based on deep residual cnn using single input. *IEEE Access*, 7:57972–57984, 2019.

[20] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[22] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4242–4251, 2018.

[23] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017.

[24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[25] Jian-Wu Lin and Hao Li. Hpiln: A feature learning framework for cross-modality person re-identification. *arXiv preprint arXiv:1906.03142*, 2019.

[26] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3810–3818, 2015.

[27] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[28] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 2429–2438, 2017.

[29] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4976–4985, 2019.

[30] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[31] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[32] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *arXiv preprint arXiv:1906.09610*, 2019.

[33] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5814–5824, 2019.

[34] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2265–2274, 2018.

[35] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 486–504, 2018.

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018.

[38] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019.

[39] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[40] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 791–808, 2016.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[42] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 3623–3632, 2019.

[43] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI-20 AAAI Conference on Artificial Intelligence*, 2020.

[44] Zheng Wang, Zhixiang Wang, Yang Wu, Jingdong Wang, and Shin'ichi Satoh. Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019.

[45] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.

[46] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017.

[47] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5380–5389, 2017.

[48] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. Unsupervised graph association for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8321–8330, 2019.

[49] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3760–3769, 2019.

[50] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[51] 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, 2018.

[52] Shizhou Zhang, Yifei Yang, Peng Wang, Xiuwei Zhang, and Yanning Zhang. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *arXiv preprint arXiv:1910.11656*, 2019.

[53] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[54] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[55] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[56] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8040–8049, 2019.

[57] Yabin Zhu, Chenglong Li, Yijuan Lu, Liang Lin, Bin Luo, and Jin Tang. Fanet: Quality-aware feature aggregation network for rgb-t tracking. *arXiv preprint arXiv:1811.09855*, 2018.