Pattern Recognition Letters 142 (2021) 25–31

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Improving visible-thermal ReID with structural common space embedding and part models

Lingyan Ran[1], Yujun Hong[1], Shizhou Zhang*, Yifei Yang, Yanning Zhang

National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Article history:
Received 25 September 2020
Revised 24 November 2020
Accepted 30 November 2020
Available online 16 December 2020

Keywords:
Visible-thermal person re-identification
Spatial-structure-preserving feature embedding
Feature representation learning

ABSTRACT

With the emergence of large-scale datasets and deep learning systems, person re-identification(Re-ID) has made many significant breakthroughs. Meanwhile, Visible-Thermal person re-identification(V-T Re-ID) between visible and thermal images has also received ever-increasing attention. However, most of typical visible-visible person re-identification(V-V Re-ID) algorithms are difficult to be directly applied to the task of V-T Re-ID, due to the large cross-modality intra-class and inter-class variation. In this paper, we build an end-to-end dual-path spatial-structure-preserving common space network to transfer some V-V Re-ID methods to V-T Re-ID domain effectively. The framework mainly consists of two parts: a modility specific feature embedding network and a common feature space. Benefiting from the common space, our framework can abstract attentive common information by learning local feature representations for V-T Re-ID. We conduct extensive experiments on the publicly available RGB-IR re-ID benchmark datasets, SYSUMM01 and RegDB, for demonstration of the effectiveness of bridging the gap between V-V Re-ID and V-T Re-ID. Experimental results achieves the state-of-the-art performance.

1. Introduction

Person re-identification (Re-ID) aims to re-identify the same individual from non-overlapping camera views, which has great value in video surveillance. During the past few years, a large number of algorithms have been proposed to tackle visible-visible person re-identification (V-V Re-ID) problems [1–3] where both query images and gallery images are captured by RGB cameras. Meanwhile, many surveillance cameras support automatic switching between visible and infrared working modes to fit the surrounding illumination variations. Therefore, the need for cross-modality Re-ID methods [4,5], especially Visible-Thermal person re-identification (V-T Re-ID), to find the same person captured by other spectrum cameras are rising.

Although V-V Re-ID problems are properly studied, V-T Re-ID problems remain challenging. Recently, multifarious methods have been proposed for V-V Re-ID, including metric learning[3], feature learning [6] and GAN-based learning[7]. Meanwhile, due to the large cross-modality discrepancy resulting from imaging sensors and intra-modality appearance discrepancy influenced by illumination, background, pose and viewpoint variations, V-T Re-ID task is still a great challenge up to now. Researchers invest lots of energy in devising exquisite network to extract modality-invariant information to represent the feature while ignoring the relationship between V-V Re-ID and V-T Re-ID. It is urgent but promising to bridge the gap between the two similar tasks.

Generally speaking, the techniques to handle V-V Re-ID are quite mature and great performance has been demonstrated in the V-V Re-ID benchmark datasets. However, it lacks appropriate methods to transfer these powerful techniques to V-T Re-ID domain. In the field of cross-modal retrieval, the multi-path feature learning network, which contain two subnetworks linked at the joint layer for correlating the data of different modalities, has always been a common approach to bridge the gap between different modalities. [8] Motivated by the multi-path network, a popular pipeline which includes feature extraction phase and feature embedding phase is introduced to tackle the problem. For feature extraction phase, a multi-branch architecture is adopted to extract modality-specific feature vectors firstly, and for feature embedding phase, a mapping function is then adopted to project the modality specific features into a common feature space. Concretely, the model branches used to extract modality specific features are not required to have same architectures or share parameters necessarily, as optimal feature extraction model highly depends on the input data modalities. Note that in the feature extraction phase, the modality specific feature is usually processed into a 1D-shaped vector

* Corresponding author.
E-mail address: szzhang@nwpu.edu.cn (S. Zhang).

[1] First Author and Second Author contribute equally to this work.

https://doi.org/10.1016/j.patrec.2020.11.020
0167-8655/© 2020 Published by Elsevier B.V.

and the mapping function is typically devised as one or several fully-connected layers to project the modality specific feature vectors, leading to a common feature space which loses the spatial structure information as it is spanned by 1D-shaped feature vectors. By utilizing this pipeline, we drives the feature into a 3D tensor common space, which preserves structure information to reduce the gap between V-V Re-ID and V-T Re-ID.

Besides, we mainly exploit parted-based Re-ID methods, such as PCB [9], HPM [10], MGN [11] et al., to verify the effectiveness of the pipeline, since part-level features provide better person representations than a global one. Intuitively, these parted-based methods can project the features from different modalities into the same subspace to capture more detailed common information, which is easily applied to V-T Re-ID.

The main contributions of are summarized as follows:

- We bridge the gap between V-V Re-ID and V-T Re-ID and apply methods utilized for V-V Re-ID to V-T Re-ID effectively.
- We design a Dual-path Spatial-structure-preserving Common Space Network (DSCSN) to embedding cross-modality images into a 3D common feature space. With the 3D common space, which preserves the intrinsic spatial structure, we can reduce the modality gap easily.
- Part-based Re-ID methods are also explored to verify the effectiveness of our proposed model.

Extensive experiments on the popular SYSU-MM01 and RegDB datasets demonstrate that DSCSN is superior to traditional dual-path architectures for RGB-IR ReID and our proposed approach outperforms competitive algorithms.

2. Related works

2.1. V-V person re-identification

V-V Re-ID task addresses the problem of matching pedestrian RGB images across disjoint visible cameras, which suffers from the difficulties of large intra-class variation due to illumination, background, pose, and viewpoint variations. Exiting methods could be grouped into three categories: hand-craft feature representation, distance metric learning, and deep learning. An exquisite hand-craft feature is extracted to improve discrimination. For example, Yang et al. [2] introduced the salient color names based color descriptor (SCNCD) for global pedestrian color descriptions. The goal of metric learning is to keep all the vectors of the same class closer while pushing vectors of different classes further apart. Some endeavors learned a Mahalanobis distance function [1], or projection matrix [3]. Unlike traditional methods, deep learning based methods can automatically extract better pedestrian image features and obtain better similarity measurement in an end-to-end manner. Because there is a large gap between RGB domains and IR domains, most of exiting methods perform well in V-V Re-ID task may not directly achieve the corresponding performance in V-T Re-ID.

2.2. V-T person re-identification

Cross-modality retrieval [4] refers to searching instances across different modality data. Especially for V-T Re-ID task, it is quite challenging due to cross-modality variation between RGB and IR images and has attracted extensive research focus on to date. In [12], Wu et al. firstly defined the problem of cross-modality person Re-ID, and provided the first RGB-IR cross modality Re-ID dataset named SYSU-MM01 for the community. Based on the dataset, they explored three different network structures with zero-padding for automatically evolving domain-specific structure for RGB-IR matching. Ye et al. [5] proposed a hierarchical metric learning method called HCML to jointly optimize the modality-specific and modality-shared metrics. Ye et al. [13] further utilized a dual-path network with a bi-directional dual constrained top-ranking loss that ensures the learnt feature representations are discriminative enough. In [14], Dai et al. introduced a cross-modality generative adversarial network (cmGAN) to handle the lack of insufficient discriminative information. Hao et al. [15] introduced Sphere Softmax to learn a hypersphere manifold embedding and constrain the intra-modality variations and cross-modality variations on this hypersphere. Very recently, some GAN based domain adaptation methods were proposed to generate corresponding visible or infrared images. In [16], Wang et al. proposed a novel Alignment Generative Adversarial Network (AlignGAN) to simultaneously alleviate the cross-modality variation in the pixel space, the intra-modality variation in the feature space. Meanwhile, Wang et al. [7] generated cross-modality paired-images by disentangling features and decoding from exchanged features. Most above methods design exquisite framework to reduce the gap between the cross-modality data. In our method, we apply typical part-based V-V Re-ID algorithm to V-T Re-ID domain via building a 3D shaped tensor common space.

2.3. Local feature representation learning for person re-identification

Local feature representation learning aims to learn an effective feature extractor to capture abundant discriminative features of various body parts and has show promising performances. Generally, these methods can be classified into three categories: The first approach leverages explicit pose estimation to obtain a human body pose map [17]. Nevertheless, this requires an additional human pose estimation data to train an accurate estimator at first. The second approach utilizes an attention map [18,19] to leverage body parts implicitly, but the attended regions may not contain discriminative body parts. The third approach directly utilizes the predefined rigid parts (horizontal stripes or grids) [9,20,21] for fine-grained feature extraction, which may be less effective when the detectors do not localize the persons tightly. We adopt the 3rd approach to extract modality-invariant and modality-sharing local feature for V-T Re-ID due to its operability and practicability.

3. Method

In this section, we elaborate the framework of the proposed feature learning method for V-T person Re-ID. The framework learns common feature representations by projecting two modalities into a spatial-structure-preserving common space. Typical V-V Re-ID algorithms can be effectively applied to V-T Re-ID by building up the common space. Meanwhile, local feature extraction module is embedded to this framework to learn the modality-invariant feature representation captured by the common space to obtain discriminative information. As shown in Fig. 1, it comprises two main components: a Dual-path Spatial-structure-preserving Common Space Network (DSCSN) and a local feature learning network.

3.1. Dual-path spatial-structure-preserving common space network

A dual-path spatial-structure-preserving common space network is designed to extract common features with the shape of 3D convolution feature maps for the input RGB and IR images. It consists of two branches, a RGB-branch and an IR-branch, and both branches are designed with same network structures. Note that it mainly contains two steps: modality specific feature extraction and common feature embedding. The feature extraction step focuses

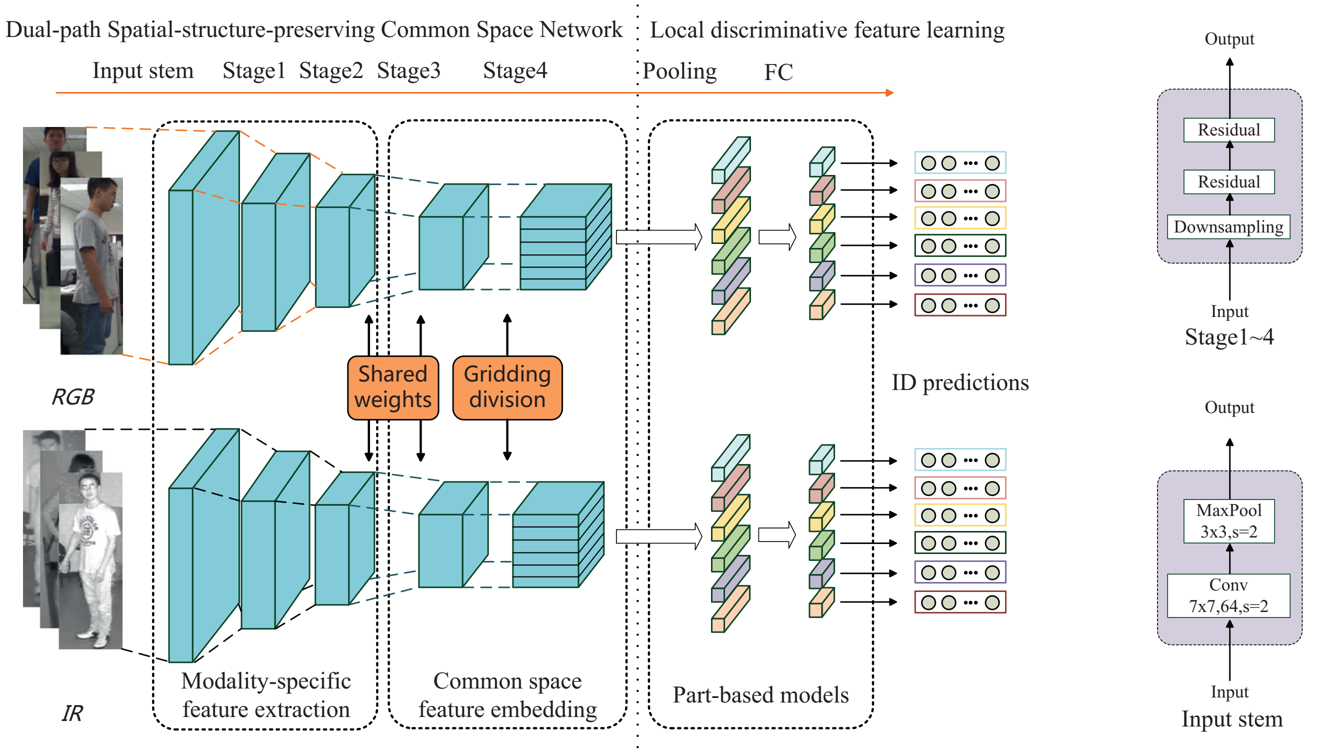

Fig. 1. The pipeline of our proposed framework for V-T person ReID. The end-to-end framework consists of two parts, DSCSN and local feature learning, with the former part embeds modility specific feature into common space and the latter one learns discriminative local features.

on extracting modality specific information for different modalities, and the feature embedding step aims to learn common features of RGB and IR image modalities.

As shown in Fig. 1, RGB and IR images are fed into the Dual-path Spatial-structure-preserving Common Space Network separately. The low-level layers without sharing parameters, which consists of the input stem, stage 1, and stage 2 of ResNet-50, are designed as the feature extraction part to extract the modality-specific features. After that, convolution architectures with shared parameters on top of feature extraction part are treated as a common feature embedding function to project the modality-specific inputs into a common space spanned by 3D convolution feature maps. Specially, stage 3 and stage 4 of ResNet-50 are treated as the convolution architectures. To clarify, $C_R(\cdot)$ and $C_I(\cdot)$ are denoted as transformation functions from the input images to common space features for RGB images and IR images respectively. Given an RGB image R and an IR image I, the extracted 3D common features F_R and F_I are represented by

$$F_R = C_R(R) \in \mathbb{R}^{h_F \times w_F \times c_F}$$
$$F_I = C_I(I) \in \mathbb{R}^{h_F \times w_F \times c_F} \tag{1}$$

where h_F, w_F and c_F are the height, width, and number of channels respectively.

It is worth mentioning that different from dual-path network proposed in [13], which introduces a shared fully connected layer as feature embedding function acting on 1D feature vectors, our feature embedding adopts convolutional architectures and acts on 3D feature tensors. The proposed network can preserve some spatial structure information for the common space. We will demonstrate the effectiveness of the 3D shaped tensor common space, where the gap between V-V Re-ID and V-T Re-ID can be easily bridged.

3.2. Local feature representation learning network

Local feature learning has shown great promising prospects for V-V ReID while not widely utilized for V-T ReID. A typical approach is to directly utilize the predefined rigid parts (horizontal stripes or grids) [9,20] for fine-grained feature extraction. Similar to these methods, we introduce part-based module after common feature embedding step to extract common local feature between RGB and IR images. Specifically, as shown in Fig. 1, after extracting modality-invariant common features with the shape of 3D convolution feature maps by DSCSN, we partition spatial-structure-preserving feature maps $F_m (m \in \{R, I\})$ into p horizontal grids $h_m^i (i \in 1, 2, \ldots p)$ to extract local feature which is concatenated to represent the body structure. Then each grid is averaged into a local feature vector with global average pooling(GAP). Afterwards, a convolutional layer is employed to reduce the dimension. Finally, each dimension-reduced column vector is put into a classifier, which is composed of a fully-connected (FC) layer with a following *Softmax* function. The classifier aims to transfer the feature vector into the class scores $s \in \mathbb{R}^{1\times C}$ to predict the identity of RGB or IR image individually. Here C is the ID number in training classes. The above procedure can be formulated as

$$s_m^i = Softmax(FC(CONV(GAP(h_m^i)))) \tag{2}$$

Since the model can obtain better discrimination ability in identifying the whole body considering the similarity of each grid, we compute the overall loss function via averaging the loss of different grids, which is defined as follows:

$$\mathcal{L} = \sum_{i=1}^{p} \mathcal{L}_i, \tag{3}$$

with

$$\mathcal{L}_i = -\frac{1}{2N}\left[\sum_{R_i}\sum_{c=1}^{C} y_{R_i}^c \log s_{R_i}^c + \sum_{I_i}\sum_{c=1}^{C} y_{I_i}^c \log s_{I_i}^c\right] \tag{4}$$

where N is the number of samples for each modality and p is the number of horizontal grids. y_R and y_I are one-hot coding ID labels for R and I respectively. s_R and s_I are the predicted label probability distributions of R and I.

Moreover, we also explore other part-based methods whose efficacy has been verified in V-V Re-ID domian. Horizontal Pyramid Matching (HPM) [10] approach learns to classify using partial feature representations at different horizontal pyramid scales. Multiple Granularity Network (MGN) [11] is a multi-branch deep network architecture consisting of two branches for local feature representations and one branch for global feature representations. Parameter-Free Spatial Attention Network (SA) [22] utilizes six independent losses for each local region and takes the summation of all six losses as the final loss. Most of models like these can be easily applied to V-T Re-ID benefitting from the common space explained in the previous section. Detailed experimental results can be found in next section.

4. Experimental results

In this section, we conduct comprehensive experiments to verify the efficacy of the proposed cross-modality framework as well as show the performance of other part-based methods.

4.1. Datasets and settings

SYSU-MM01 As a standard benchmark for cross-modality (RGB-IR) Re-ID, the SYSU-MM01 [23] RGB-IR Re-ID dataset is chosen to verify the efficacy of the proposed method. This dataset is collected by 6 cameras, including 4 visible cameras and 2 thermal ones. It contains 491 available identities with total 287,628 RGB images and 15,792 IR ones, and each person is captured by at least two different spectrum cameras. The dataset is separated into the training set and the test set, where images of the same person can only appear in either set. The training set contains 395 persons including 22,258 visible images and 11,909 thermal images and the testing images with 96 IDs. It brings great difficulty that some of the person images are captured in the indoor environments and some are in outdoor environments, and the variation between two modalities makes the dataset more challenging.

RegDB We also perform experiments on another publicly available dataset called RegDB [24], which consists of 4120 visible images and 4120 thermal images in total. A pair of aligned visible and infrared cameras are used to capture these paired images of 412 volunteers. Following the evaluation protocol adopted in Wang et al. [25], we randomly split this dataset into two halves, one for training and the other for testing. For testing, the query set consists of 2060 IR images and the gallery set contains 2060 RGB images.

4.2. Evaluation protocols

We follow the evaluation protocols used in Wu et al. [12] to evaluate our model on SYSU-MM01 dataset. There are two test modes, *all-search* mode and *indoor-search* mode. For the *all-search* mode, 3rd and 6th thermal cameras are for probe set and 1st, 2nd, 4th, 5th visible cameras are for gallery set. For the *indoor-search* mode, 3rd and 6th thermal cameras are for probe set and only 1st and 2nd visible cameras are for gallery set. Obviously, *all-search* mode is more challenging than *indoor-search* mode, due to the scene diversity. For both modes, the single-shot and multi-shot settings are adopted, where each identity contains 1 or 10 images randomly selected from the gallery set. Note that both modes use IR images as probe set and RGB images as gallery set.

Following [24], we compute the average of 10 times repeated random split of training and testing sets to obtain the results of RegDB. For each image in the probe set, we compute the feature similarities between the IR image and every RGB image in the gallery set to match the pedestrian.

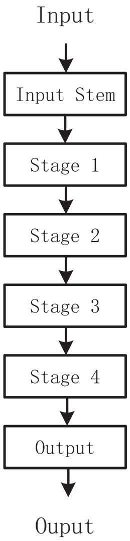

Fig. 2. The typical ResNet-50 model [26].

The Cumulative Matching Characteristic curve (CMC) and mean Average Precision(mAP) are adopted as evaluation metrics.

4.3. Implementation details

We first resize all the images to 384 × 128 pixels, then random cropping and horizontal flipping are used to augment the training data. At each iteration, we randomly select N person identities, and image pairs with one RGB and one IR image are splatted into mini-batches. Thus, totally $2 \times N$ images are fed into the network for training at each iteration. We set $N = 32$ in our experiments.

Then we choose pretrained ResNet-50 as the backbone architecture for the dual path network, which consists of multiple stages as in Fig. 2. Concretely, the parameters are not shared for the input stem, stage 1, and stage 2 of ResNet-50 during the modality-specific feature extraction step, while they are shared for the stage 3 and stage 4 which are treated as the feature embedding blocks. The output common feature of embedding blocks is equally split into $p = 6$ grids. The dimension of feature vector is reduced to 256 by the FC layer.

During training, the stochastic gradient descent (SGD) optimizer is utilized for optimization. We set the maximum number of training epochs to 40, and the initial learning rate to 0.1 which is then decayed by $1/10$ for the last 20 epochs.

During testing, the final descriptor of the input RGB or IR image are formed by concatenating different horizontal grids. Afterwards, we compute similarities as the final scores between the query image and gallery images using Cosine distance measurement.

Lastly, all the experiments are executed on NVIDIA GeForce 1080Ti graphics cards with the Pytorch development package.

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

In this subsection, we evaluate our proposed algorithm using the rank-1(r-1), r-10, r-20 accuracies of CMC and mAP metrics. Comparative methods are current popular methods, like Zero-padding [12], BDTR [13], cm-GAN [14], D^{2}RL [25], HSME [15], AlignGAN [16], JSIA-REID [7]. The results of comparison with state-of-the-art methods on SYSU-MM01 and RegDB are listed in Tables 1 and 2 respectively.

Table 1
Comparison with state-of-the-art methods on SYSU-MM01(%).

Method	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
	r-1	r-10	r-20	mAP	r-1	r-10	r-20	mAP	r-1	r-10	r-20	mAP	r-1	r-10	r-20	mAP
HOG + Euclidean	2.8	18.3	31.9	4.2	3.8	22.8	37.6	2.2	3.2	24.7	44.5	7.3	4.8	29.1	49.4	3.5
HOG + KISSME	2.1	16.2	29.1	3.5	2.8	18.2	31.3	2.0	3.1	25.5	46.5	7.4	4.1	29.3	50.6	3.6
HOG + LFDA	2.3	18.6	33.4	4.4	3.8	20.5	35.8	2.2	2.4	24.1	45.5	6.9	3.4	25.3	45.1	3.2
LOMO + CCA	2.4	18.2	32.5	4.2	2.6	19.7	34.8	2.2	4.1	30.6	52.5	8.8	4.9	34.4	57.3	4.5
LOMO + CDFE	3.6	23.2	37.3	4.5	4.7	28.2	43.1	2.3	5.8	34.4	54.9	10.2	7.4	40.4	60.3	5.6
LOMO + GAM	1.0	10.5	20.8	2.5	1.0	10.5	21.1	1.5	1.8	17.9	36.0	5.6	1.7	18.1	36.2	2.9
GSM [27]	5.3	33.7	53.0	8.0	6.2	37.2	55.7	4.4	9.5	49.0	72.1	15.6	11.4	51.3	73.4	9.0
One-stream Network [12]	12.0	49.7	66.7	13.7	16.3	58.1	75.1	8.6	16.9	63.6	82.1	23.0	22.6	71.7	87.8	15.0
Two-stream Network [12]	11.7	48.0	65.5	12.9	16.3	58.4	74.5	8.0	15.6	61.2	81.0	21.5	22.5	72.2	88.6	13.9
Zero-padding [12]	14.8	54.1	71.3	16.0	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 [5]	12.5	50.7	68.6	14.4	–	–	–	–	–	–	–	–	–	–	–	–
HCML [5]	14.3	53.2	69.2	16.2	–	–	–	–	–	–	–	–	–	–	–	–
BDTR [13]	27.3	67.0	81.0	27.3	–	–	–	–	31.9	77.2	89.3	41.9	–	–	–	–
cmGAN [14]	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	81.0	92.1	32.8
eBDTR [28]	27.8	67.3	81.3	28.4	–	–	–	–	32.5	77.4	89.6	42.5	–	–	–	–
D^{2}RL [25]	28.9	70.6	82.4	29.2	–	–	–	–	–	–	–	–	–	–	–	–
HSME [15]	18.0	58.3	74.4	20.0	–	–	–	–	–	–	–	–	–	–	–	–
D-HSME [15]	20.7	62.7	78.0	23.1	–	–	–	–	–	–	–	–	–	–	–	–
JSIA-REID [7]	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 [16]	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
Ours(Avg pool)	40.1	84.2	93.5	41.2	43.0	86.8	94.8	33.5	45.0	87.4	95.6	53.9	50.2	90.6	96.8	43.5
Ours(Max pool)	46.4	88.5	**95.8**	46.3	50.3	**90.5**	**97.1**	38.9	48.6	89.7	92.3	57.9	**58.2**	**94.5**	**98.6**	50.0
Ours(Avg&Max pool)	**47.2**	**89.1**	**95.8**	**47.1**	49.7	90.3	96.4	**39.4**	**51.0**	**91.7**	**97.1**	**59.7**	56.9	**94.5**	**98.6**	**50.8**

Table 2
Comparison with state-of-the-art methods on RegDB(%).

Method	r-1	r-10	r-20	mAP
Zero-padding [12]	17.8	34.2	44.4	18.9
TONE [5]	16.9	34.0	44.1	14.9
HCML [5]	24.4	47.5	56.8	20.8
BDTR [13]	33.5	58.4	67.5	31.8
eBDTR [28]	31.8	56.1	66.8	33.2
D^{2}RL [25]	43.4	66.1	76.3	44.1
HSME [15]	41.3	65.2	75.1	38.8
D-HSME [15]	50.9	**73.4**	**81.7**	47.0
JSIA-REID [7]	48.1	–	–	48.9
AlignGAN [16]	56.3	–	–	53.4
Ours	**59.0**	70.0	79.2	**62.5**

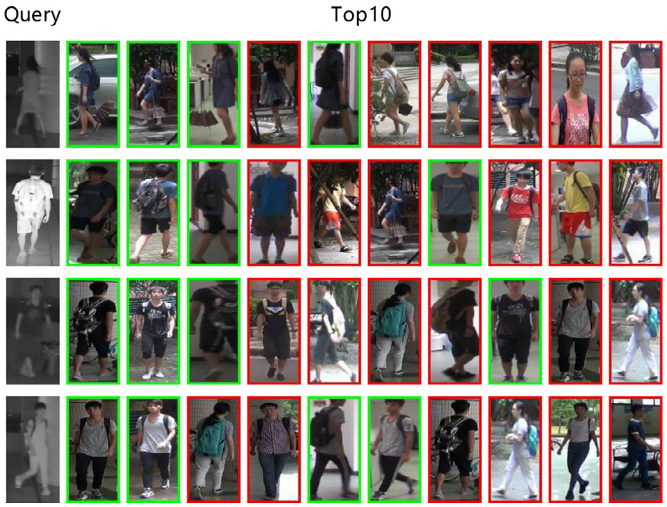

Fig. 3. Sample retrieval results of our proposed method on the test set of SYSU-MM01. Query images are listed first, followed by Top-10 matches with descending confidence score (green box for same ID and red box for mismatches).

In Table 1, The first column lists some mainstream methods on SYSU-MM01, including our methods on the bottom. Note that the bottom three rows show the performance of average pooling only, max pooling only and both pooling strategies. The last two columns refer to two test modes with two adopted shot settings mentioned in Section 4.2. *Evaluation Protocols*.

From Table 1, we can observe that traditional methods including HOG and LOMO are obviously surpassed by deep learning methods. Overall, Our method with both pooling strategies significantly outperforms all existing methods in any mode. Especially in the most difficult one, all-search single-shot mode, our method can outperform AlignGAN [16] by 4.8% rank-1 and 6.3% mAP respectively.

In addition, from the bottom three rows in Table 1, it can be observed that max pooling performs better than average pooling in most cases. As some researchers [10] have concluded, average pooling covers all locations of a particular parts, but it is easily distracted by background clutter and occlusion. Max pooling overcomes this problem by preserving the largest response values for a local view while discarding background clutter. Therefore, we also integrate these two strategies into a unified model to obtain better feature representations. Experimental results in Table 1 show that *Ours(Avg&Max pool)* outperforms *Ours(Avg pool)* by 7.1% rank-1 and 5.9% mAP respectively and outperforms *Ours(Avg pool)* by 0.8% rank-1 and 0.8% mAP respectively in all-search single-shot mode. Thus, it demonstrates that mixing the two pooling strategies performs better than using either of them.

In Table 2, our model still goes beyond many competitive models on RegDB dataset. Firstly, D^{2}RL [25] outperforms eBDTR [28] by 11.6% rank-1 and 10.9% mAP scores, which demonstrates the effectiveness of adversarial training. Secondly, D-HSME [15] outperforms D^{2}RL [25] by 7.5% Rank1 and 2.9% mAP scores, implying the effectiveness of metric learning. Additionally, AlignGAN [16] outperforms D-HSME [15] by 5.4% Rank1 and 5.6% mAP scores, verifying that generative methods can alleviate the cross-modality variation. Finally, Our method outperforms AlignGAN [16] by 2.7% Rank1 and 9.1% mAP scores, which demonstrates the effectiveness of our method for V-T Re-ID task.

In Fig. 3, we give some examples of the SYSU-MM01 testing case to demonstrate our method qualitatively. Some IR images

Table 3
Results of other part-based models on SYSU-MM01 dataset.

Method	r-1	r-10	r-20	mAP
HPM (Avg pool)	40.2	83.3	92.1	40.0
HPM (Max pool)	46.0	86.8	94.1	45.1
HPM (Avg&Max pool)	42.3	85.6	93.7	42.2
MGN (Rank)	37.9	81.5	91.4	38.6
MGN (ID)	36.5	80.2	91.0	39.6
MGN (Rank&ID)	39.0	82.3	92.3	41.9
SA	41.3	87.4	92.4	44.0
AlignGAN (Part-based)	39.3	84.0	93.1	39.3
Ours	**47.2**	**89.1**	**95.8**	**47.1**

Table 4
Effect of feature embedding on SYSU-MM01 and RegDB datasets.

	SYSU-MM01				RegDB			
Shared layers	r-1	r-10	r-20	mAP	r-1	r-10	r-20	mAP
None-stage	0.9	10.4	21.6	3.2	1.1	2.7	4.6	2.1
Stage 4	38.4	85.1	93.5	40.0	58.0	64.0	73.5	60.5
Stage 3–4	**47.2**	89.1	95.8	**47.1**	**59.0**	70.0	79.2	**62.5**
Stage 2–4	44.0	84.4	93.0	43.1	53.2	58.9	71.9	58.3
Stage 1–4	42.9	86.6	94.6	44.0	57.4	68.1	73.6	60.2
All	31.1	76.9	88.7	33.9	56.3	66.2	72.7	59.0

are selected as query images in first column, followed by Top-10 matches from RGB images with descending confidence score. The correct matching images are in the green rectangles, while the mismatches are in the red rectangles. From The retrieval results, we can see that our method gets great shots and reduce the modality gap dramatically.

4.5. Comparison with other part-based models

In addition to the method described above, we also attempt to adopt other part-based models to bridge the gap between V-V Re-ID and V-T Re-ID. The results are shown in Table 3, HPM [10] can achieve 42.3% rank-1 and 42.2% mAP by mixing avg pooling and max pooling strategies. MGN [11] can achieve 39.0% rank-1 and 41.9% mAP by integrating rank loss and id loss. SA [22] can achieve 41.3% rank-1 and 44.0% mAP respectively. Part-based AlignGAN [16], which the network learns local feature representation, can attain a lower accuracy compared with the methods in the table. Note that we adopt the most challenging single-shot all-search test mode to compare these methods.

In general, although these methods can not achieve the state-of-art accuracy, benefitting from the spatial-structure-preserving common space, they still perform well in V-T Re-ID domain.

4.6. Discussion

Effect of feature embedding function It is a key step for V-T Re-ID to choose appropriate feature embedding function. Previous feature embedding methods [5,13] aim to project the modality-specific feature vectors into the common feature space by utilizing a fully connected layer, which leads to lack of enough body structure information of final feature representation. However, we design the feature embedding function as convolution architectures to build a 3D tensor common space, which is beneficial to transferring models used for V-V Re-ID to V-T Re-ID due to more abundant information. Specially, we view parameter sharing parts of ResNet-50 as the feature embedding function. Table 4 demonstrate how the feature embedding function affects the V-T Re-ID performance. In Table 4, we can observe that when stage 3 and stage 4 of ResNet-50 are designed as feature embedding function, it achieves best results. Note that 'None-stage' means only parameters of fully-connected layer are shared, that is traditional dual-path architectures [13] and 'None' means no parameters are shared.

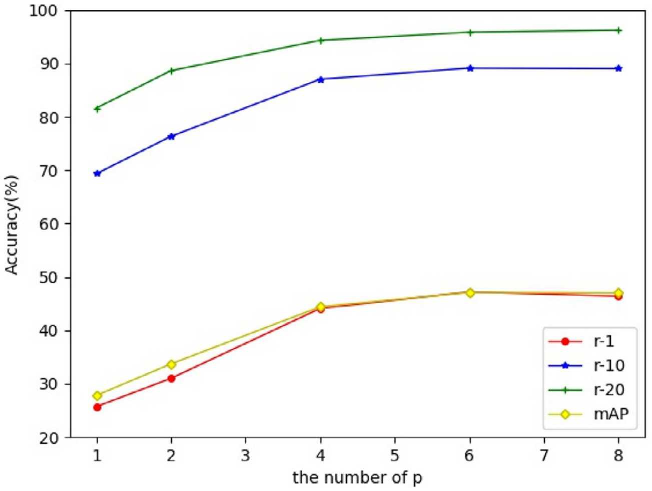

Fig. 4. Accuracy with different grid numbers. For different retrieval curves, varying p shows similar influence. And the performance comes best with $p = 5$.

The number of grids p For V-V Re-ID, the number of grids determines the granularity of local feature which affects the discrimination. We evaluate the number effect and the results are shown in Fig. 4. As p increases, retrieval accuracy improves at first since the network can capture more detailed information with narrower granularity. However, the accuracy tends to saturation when p is over 6.

5. Conclusion

In this paper, we introduce a Dual-path Spatial-structure-preserving Common Space Network, which effectively reduces the gap between V-V Re-ID and V-T Re-ID. The network projects the input images into a 3D tensor common space where abundant information is preserved. Most of part-based models utilized for V-V Re-ID can easily be applied to V-T Re-ID domain. Significant performance improvement is achieved on benchmark datasets, like SYSU-MM01 and RegDB dataset, with extensive experiments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61902322, U19B2037), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202000025, the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2019JQ158), China Postdoctoral Science Foundation funded project under Grant No. 2018M633577.

References

[1] W.S. Zheng, S. Gong, X. Tao, Person re-identification by probabilistic relative distance comparison, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[2] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S.Z. Li, Salient color names for person re-identification, in: Proceeings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 536–551.

[3] S. Liao, S.Z. Li, Efficient PSD constrained asymmetric metric learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[4] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385.

[5] M. Ye, X. Lan, J. Li, P.C. Yuen, Hierarchical discriminative learning for visible thermal person re-identification, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[6] S. Liao, Y. Hu, X. Zhu, S.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 (CVPR), 2015, pp. 2197–2206.

[7] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z. Hou, Cross–modality paired-images generation for RGB-infrared person re-identification, in: AAAI-20 AAAI Conference on Artificial Intelligence, 2020.

[8] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385.

[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceeings of the European Conference on Computer Vision (ECCV), 2018.

[10] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019.

[11] G. Wang, Y. Yuan, C. Xiong, J. Li, Z. Xi, Learning discriminative features with multiple granularities for person re-identification, in: 2018 ACM Multimedia Conference, 2018.

[12] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, RGB-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5380–5389.

[13] M. Ye, Z. Wang, X. Lan, P.C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking., in: IJCAI, 2018, pp. 1092–1099.

[14] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training., in: IJCAI, 2018, pp. 677–683.

[15] Y. Hao, N. Wang, J. Li, X. Gao, HSME: hypersphere manifold embedding for visible thermal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8385–8392, doi:10.1609/aaai.v33i01.33018385.

[16] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, RGB-infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-Net: attentive deep features for pedestrian analysis, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[19] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[20] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[21] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi--loss classification, in: Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[22] H. Wang, Y. Fan, Z. Wang, L. Jiao, B. Schiele, Parameter-free spatial attention network for person re-identification, arXiv preprint arXiv:1811.12150(2018).

[23] A. Wu, W.-S. Zheng, S. Gong, J. Lai, RGB-IR person re-identification by cross–modality similarity preservation, Int. J. Comput. Vis. (IJCV) (2020) 1–21.

[24] D. Nguyen, H. Hong, K. Kim, K. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605.

[25] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S. 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 (CVPR), 2019, pp. 618–626.

[26] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of tricks for image classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 558–567.

[27] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 39 (6) (2016) 1089–1102.

[28] M. Ye, X. Lan, Z. Wang, P.C. Yuen, Bi-directional center-constrained top-ranking for visible thermal person re-identification, IEEE Trans. Inf. Forensics Secur. (2019).