Infrared and Visible Cross-Modal Image Retrieval Through Shared Features

Fangcen Liu, Chenqiang Gao, Yongqing Sun, *Member, IEEE*, Yue Zhao, Feng Yang, Anyong Qin, and Deyu Meng, *Member, IEEE*

***Abstract*— Image retrieval is one of the key techniques of computer vision, and has been studied for a long time. Nevertheless, little attention is paid to infrared and visible cross-modal retrieval which can be widely used in various applications, e.g., infrared and visible surveillance systems. In this paper, we propose a shared features based infrared-visible cross-modal image retrieval method. The similar visual features are extracted from infrared and visible images as the shared features, and the Euclidean distance is used to measure the similarity between these features. The core of the proposed method comes from three aspects: 1) Feature separation network can separate image features into shared features and exclusive features; 2) Maximum Mean Discrepancy (MMD) loss is employed to constrain the distribution of shared features, which can reduce the retrieval error caused by different imaging angles and similarity of infrared images. 3) The cross-layer fusion encoder compensates for the context loss in the convolution of infrared images. Experimental results on the Infrared-Visible dataset demonstrate the proposed method is effective and outperforms the state-of-the-art approaches.**

***Index Terms*— Infrared-visible image retrieval, shared feature, cross-layer fusion encoder, maximum mean discrepancy.**

I. Introduction

MULTI-MODAL sensors have been becoming more and more popular in practical applications, especially in kinds of surveillance scenarios. The commonly-used image sensors for these applications are visible and infrared cameras, due to that their imaging characteristics are complementary. Visible images have sufficient color and texture information, yet they heavily rely on light conditions. On the contrary, infrared imaging works well on poor light conditions, even at night without any artificial lighting, but it has no color information and its texture information is also insufficient [1]–[5].

The applications of infrared and visible images require various computer vision techniques, one of which is infrared and visible cross-modal image retrieval. For example, if a suspect is captured at night via an infrared camera, we expect to quickly search him or her in visible videos during daytime for more information, and vice versa. Another example is that the first step of various infrared and visible image fusion or registration methods [6], [7] is to pair the infrared and visible images. However, the recording of infrared and visible images may not be synchronized, which causes ones not to directly obtain the paired infrared and visible images. In these typical examples, the infrared and visible cross-modal image retrieval can greatly reduce manual labors and time consumption.

Up to now, image retrieval has become an important research area in computer vision, but most of the attention is paid to the image retrieval in the same modality. Content-based image retrieval is the mainstream technique and has been rapidly developed [8]–[11]. Additionally, some cross-modal image retrieval studies also have been explored. The typical examples are the single-directional text to image retrieval and bi-directional text and image retrieval [12], [13]. Among cross-modal retrieval studies, however, little attention is paid to infrared and visible cross-modal image retrieval. To our best knowledge, there is not any specific research work on this task.

As shown in Fig. 1, there are many challenges for infrared and visible cross-modal image retrieval, summarized as follows:

(1) Different imaging effects: As mentioned previously, infrared images usually have better imaging quality than the visible images on the poor light conditions, while visible images have more texture information as shown in Fig. 1 (a). Besides, visible images have important color information [5]. Hence the context information of infrared images would be lost fast in convolution.

(2) Different imaging angles: Even though infrared and visible cameras are used to shoot the same object, there are usually imaging angle differences. As a result, image pixels between infrared and visible images are misaligned, as shown in Fig. 1 (b).

Manuscript received May 7, 2020; revised August 25, 2020 and October 28, 2020; accepted December 21, 2020. Date of publication January 4, 2021; date of current version October 28, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61571071, Grant 61906025, Grant 11690011, and Grant U1811461; in part by the Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2018jcyjAX0227 and Grant cstc2020jcyj-msxmX0835; and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN201900607 and Grant KJQN202000647. This article was recommended by Associate Editor Z.-J. Zha. *(Corresponding author: Chenqiang Gao.)*

Fangcen Liu, Chenqiang Gao, Yue Zhao, Feng Yang, and Anyong Qin are with the Chongqing Key Laboratory of Signal and Information Processing, Chongqing University of Posts and Telecommunication, Chongqing 400065, China, and also with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunication, Chongqing 400065, China (e-mail: liufc67@gmail.com; gaocq@cqupt.edu.cn; zhaoyue@cqupt.edu.cn; yangfeng@cqupt.edu.cn; qinay@cqupt.edu.cn).

Yongqing Sun is with NTT Media Intelligence Laboratories, Yokosuka 239-0847, Japan (e-mail: yongqing.sun.fb@hco.ntt.co.jp).

Deyu Meng is with the Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau, and also with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dymeng@mail.xjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2020.3048945.

Digital Object Identifier 10.1109/TCSVT.2020.3048945

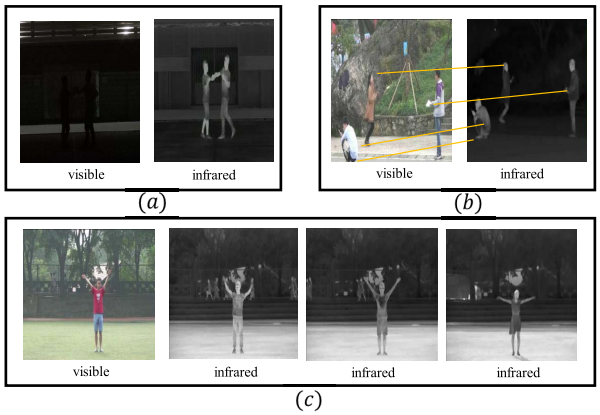


Fig. 1. Illustration of infrared and visible cross-modal image retrieval challenges. (a) Images are taken at night, where the infrared image has much more information than the visible image. (b) There are significant differences in terms of the angle and position between the infrared and visible images. (c) The left first image is the visible image, while the rest are infrared images. The discernibility of current infrared cameras is not strong and the resolution is low, resulting in infrared images that are sometimes similar.

(3) The similarity of infrared images: For infrared images, the gray value reflects the scene's infrared thermal radiation ability. Unfortunately, the discernibility of current infrared cameras on the thermal radiation difference is not strong, with low-resolution imaging. Thus, the similarity between infrared images is usually slightly obvious. For example, the arm swing of a single person occurs in many infrared scenes, as shown in Fig. 1 (c). When the infrared images are retrieved by a visible image, the arm wing features in the visible image are pretty similar to those in many infrared images, which easily leads to misjudgment.

However, infrared images and visible images still have potential similar features. It would be a good solution to utilize potential similar features to handle the cross-modal image retrieval problem. To this end, in this paper, we propose an infrared and visible cross-modal image retrieval method via shared features, inspired by the recent advance on the shared feature extraction studies [14], [15]. The shared features and exclusive features are extracted by utilizing a cross-layer fusion encoder structure. Besides, the Maximum Mean Discrepancy (MMD) method [16] is adopted to constrain the similarity distribution of shared features. The extracted shared features can eliminate the position information of the pixels and only contain similar visual information between two modalities.

The contributions of this paper are summarized as follows.

- We propose a novel infrared-visible cross-modal image retrieval method via shared feature extraction.
- Cross-layer fusion encoder is proposed in our work, which can mitigate the large context loss in the convolution of infrared images that have little details. Besides, MMD is adopted to constrain the shared features between two modalities to make them have the same distribution, and reduce the interference of pixel misalignment and similar infrared images.
- We evaluate the proposed method on the Infrared-Visible dataset and extensive experimental results show that the proposed method is effective and outperforms the baseline methods.

The rest of this paper is organized as follows. In Section II, we review related work briefly. The proposed method is described in detail in Section III. In Section IV, we report extensive experimental results, and conclusion is finally drawn in Section V.

II. Related Work

We briefly review the related work from three aspects: single-modal image retrieval, cross-modal retrieval and shared feature extraction.

A. Single-Modal Image Retrieval

The most widely used image retrieval techniques are content-based image retrieval and semantic-based image retrieval [17], [18]. The content-based image retrieval task analyzes and queries the content of the image, such as the color, texture, shape and other low-level features of the image. Early methods realized image retrieval by manually designing image color and texture features [19], [20]. In recent years, Convolutional Neural Network (CNN) based multiple feature fusion image retrieval methods were widely studied. Yue *et al.* [21] combined multi-scale CNN features with Vector Locally Aggregated Descriptors (VLAD) features to evaluate the performance of different convolutional layers for instance-level image retrieval. Tzelepi *et al.* [22] proposed a relevant feedback based retraining method to refine the high-level features and improved retrieval performance. However, this method requires a large amount of user feedback information. Yu *et al.* [23] used the complementary advantages of CNN features of different layers for image retrieval instead of simply splicing the output of different layers.

Semantic-based image retrieval considers the "semantic gap" between low-level image pixels and high-level semantics to map visual features to high-level semantic features, and searches for the similarity between semantic features of images [24], [25]. Early work took the advantage of manually designed local features of the image and aggregated them to produce a robust global description [26]. Several recent works have shown that deep neural networks performed well on image retrieval tasks. Babenko *et al.* [27] investigated possible ways to aggregate local deep features to produce compact global descriptors for image retrieval. This method is not only efficient, but also has few parameters, which bears little risk of overfitting. Gordo *et al.* [28] proposed a novel approach for instance-level image retrieval which produced a global and compacted fixed-length representation for each image by aggregating many region-wise descriptors. Korichi *et al.* [24] took into account the combination of positive and negative examples, which considerably decreased the noise and miss problems of image retrieval. This method well captured user's intention and improved significantly the performance. Besides, Song *et al.* [29] proposed two discriminative deep feature learning methods based on two loss functions which took the spatial distribution of features into account during learning discriminative features. The discriminative features of images are the real crux of the image retrieval task.

B. Cross-Modal Retrieval

In the field of cross-modal retrieval, there are two types of text-image retrieval: single-directional text to image retrieval and bi-directional text-image retrieval. The text to the image retrieval task annotates images with texts, and images can be organized by topics through text descriptions [30]. However, single-directional text to image retrieval relies heavily on manual annotation and has been gradually replaced by content-based image retrieval. Now bi-directional text-image retrieval is a hot research topic. The ubiquitous methods contain real-valued based retrieval and binary-valued based retrieval [31]. The real-valued approaches map the features of different modalities into a common real-valued space for similarity measurement. The binary methods aim at mapping the features into a common binary hamming space, which is more geared towards retrieval efficiency. This method needs to make a concession on the retrieval precision. However, real-valued approaches are just the opposite.

For real-valued based retrieval, in recent years neural networks were applied to cross-modal retrieval [32], and Restricted Boltzman Machine (RBM) was introduced into multi-modal data [33]. Inspired by these works, Feng *et al.* [34] proposed a Correspondence Auto-Encoder (Corr-AE) to retrieval. The Corr-AE focused more on the correlation across data than the complementarity learned from different modalities [35]. Based on the success of Generative Adversarial Networks (GANs) [36], [37] on the cross-modal retrieval task, Xu *et al.* [38] mapped texts and images into a shared subspace, minimizing intra-class differences and maximizing inter-class differences to retrieve. Wang *et al.* [31] and He *et al.* [39] mapped the different modal features into a common subspace in supervised and unsupervised ways, and measured the similarity between two modal samples in this common subspace. After that, Gu *et al.* [13] extracted the high-level global abstract features and the local grounded features of both modalities and made two modal extracted features closer. Other methods aimed to extract the features more efficiently. Wang *et al.* [40] developed an adaptive approach to control the flow of information across schema messaging. Thus text-image retrieval can carry out with both comprehensive and fine-grained cross-modal information. Dong *et al.* [41] adopted Bi-directional Gated Recurrent Unit (BiGRU) to extract contextual information from the text more efficiently. To bridge the heterogeneity gap between text and image, Peng *et al.* [42] treated the image as a special kind of language to provide visual description and effectively explored cross-media correlation in the feature space of each media type.

For binary-valued based retrieval, Jiang *et al.* [43] proposed a Deep Cross-Modal Hashing (DCMH) method. Unlike other relaxation-based methods [44], [45], this method directly learned the discrete hash codes without relaxation. They integrated feature learning and hash-code learning into the same framework. Wu *et al.* [46] used the hash code consistency to associate each input and output feature. Their method also regenerated the input features to minimize information loss. Xiao *et al.* [47] proposed a self-supervised adversarial network retrieval model based on an attention mechanism which has the ability to capture simultaneously global semantic information and local information of the text and image data. This method can accurately match the image and text description in complex content and improve the efficiency of computation time. Wang *et al.* [48] proposed a robust and flexible discrete hashing method which can efficiently learn robust discrete binary codes.

C. Shared Feature Extraction

Shared feature extraction is a general strategy which can be broadly applicable to many tasks [14], [15], [49]–[53]. In the field of transfer learning, Bousmalis *et al.* [15] employed the encoder to divide images into a domain exclusive space and a domain shared space which can improve the ability of the model to extract domain-invariant features. For image-to-image translation, Gonzalez *et al.* [14] proposed an encoder-decoder network to disentangle the shared representations and exclusive representations of images of two domains, and bidirectional image translation can be realized by combining one domain's share representations and another domain's exclusive representations. For person Re-identification (Re-ID), Eom *et al.* [49] designed an unrelated-identity encoder and a related-identity encoder to extract ID features, and finally made use of related-identity features for Re-ID. Recently, visible-infrared person re-identification has been rapidly developed [54], [55]. Kansal *et al.* [52] distilled identity features and dispelled spectrum features for such task. Lu *et al.* [50] paid more attention to specific features to extract shared-specific features for Re-ID work. Choi *et al.* [51] proposed a hierarchical cross-modality disentanglement method. In addition to dividing image features into ID-discriminative features and ID-excluded features, they improved the ability of Re-ID by providing a more refined disentanglement of ID-discriminative features.

III. METHODOLOGY

A. Overview

The framework of the proposed method is shown in Fig. 2. Given a query visible image x, our task is to retrieve the most similar image from an infrared image set $\{y_1, y_2, \cdots, y_n\}$. Firstly, we extract the shared features $\{x^s, y_1^s, \cdots, y_n^s\}$ of visible and infrared images through a shared feature extraction module containing two-stream networks, where the visible image and infrared images are fed into the visible stream and infrared stream, respectively. Then, we measure the similarity of x^s and y_i^s with the Euclidean distance, and obtain the retrieval rankings through sorting similarity measures. This process is the same for the situation that a query infrared image y retrieves from a visible image set $\{x_1, x_2, \cdots, x_n\}$.

In the proposed method, the shared feature extraction is the key module. To ensure the module can extract shared features of two different modalities, inspired by cross-domain disentanglement [14], we propose a feature separation network, as shown in Fig. 3. Different from [14], the proposed method

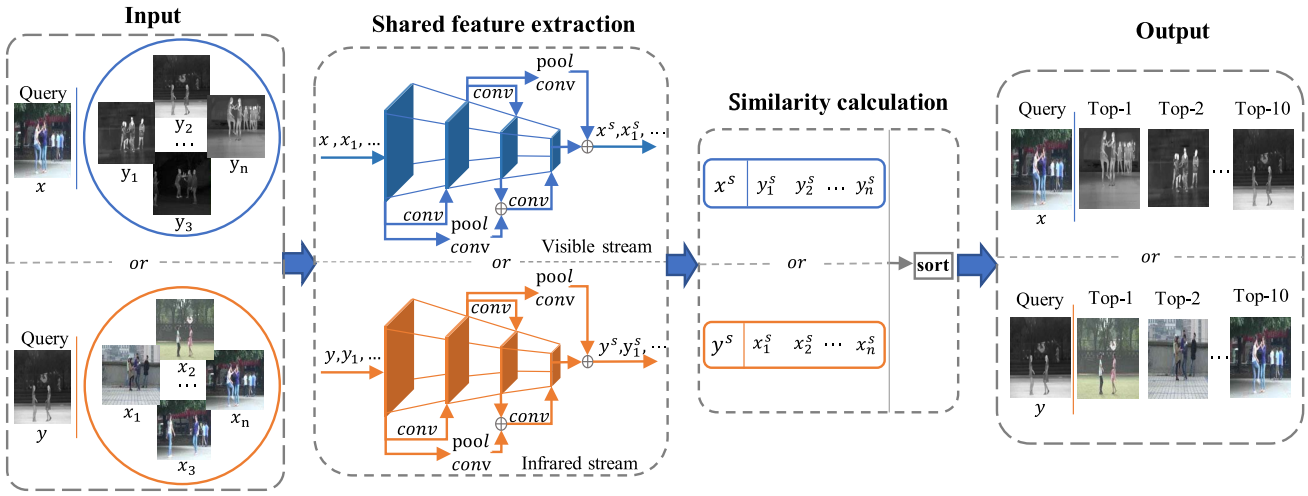


Fig. 2. The proposed framework of infrared and visible image retrieval in this paper.

focuses on solving the problems raised in Section I and does not require pixel-level alignment of the input images.

B. Preliminaries

We denote by X and Y the visible and infrared image modalities, respectively. $D = \{I_1, \cdots, I_i, \cdots, I_N\}$ is a collection of pairs of infrared and visible images, where $I_i = (x_i, y_i)$ includes a visible image x_i and an infrared image y_i. The shared feature extracted from the visible image is x_i^s, and the exclusive feature is x_i^e, then the total feature of a visible image can be expressed as: $x^t = \left(x_i^s, x_i^e\right)$. Similarly, $y^t = \left(y_i^s, y_i^e\right)$ represents the total feature of an infrared image.

As for feature separation network, it includes several encoders and decoders. Cross-layer fusion encoders G_e and F_e aim to separate image features into shared features and exclusive features. Among them, $G_e(x_i) = (x_i^s, x_i^e)$, $F_e(y_i) = (y_i^s, y_i^e)$. Then, we reconstruct images according to different feature maps by using combined feature decoders G_d, F_d or exclusive feature decoders G_d^e, F_d^e. Particularly, $G_d(x_i^s, \cdot) = fake(y_i)$, $F_d(y_i^s, \cdot) = fake(x_i)$, $G_d^e(x_i^e) = \dot{y}_i$ and $F_d^e(y_i^e) = \dot{x}_i$. The $fake(\cdot)$ means a fake image reconstructed by a decoder.

As can be seen in Fig. 3, the encoders G_e and F_e correspond to the visible stream and infrared stream of Fig. 2, respectively.

C. Feature Separation Network

The feature separation network contains two modules: feature extraction and feature reconstruction. The feature extraction module separates the image features into shared features and exclusive features, then constrains two modal shared features from the image's perspective, while the feature reconstruction module constrains shared features from the feature's perspective.

1) Feature Extraction: A pair of visible and infrared images are fed into the visible stream G_e and the infrared stream F_e, respectively. In the last step of the encoder, shared features x_i^s and y_i^s are obtained through convolution layers, while exclusive features x_i^e and y_i^e are obtained through full connection layers. The shared features x_i^s and y_i^s should contain the similar visual information of images from two modalities, and we constrain them by L1 loss and MMD loss. The L1 loss is as follows:

$$L_S = \mathbb{E}_{x\sim X, y\sim Y}\left[\left\|x_i^s - y_i^s\right\|\right]. \tag{1}$$

MMD is used to measure the distance between two different but related distributions, which is one of the most widely used loss functions in the transfer learning community. Essentially, it aims to find a transformation kernel function so that the distance between the transformed source domain and the target domain is minimized. MMD loss would measure and narrow the distance between two domain distributions [56]. In order to make the shared features of two modalities closer, we adopt the MMD loss to reduce the difference:

$$MMD(X, Y) = \left\|\mathbb{E}_{x\sim X}\left[f\left(x^s\right)\right] - \mathbb{E}_{y\sim Y}\left[f\left(y^s\right)\right]\right\|_{\mathcal{H}}^2, \tag{2}$$

where f is a mapping function which maps the sample features to the Reproducing Kernel Hilbert Spaces (RKHS) and then calculates the distribution distance.

Exclusive features x_i^e and y_i^e only contain independent features of their modalities. The real infrared image can not be generated by x_i^e, and the real visible image can also not be generated by y_i^e. To achieve these ends, we adopt the Gradient Reversal Layer (GRL) [57] which was originally introduced to learn domain-agnostic features. GRL is added in front of the G_d^e to generate a fake image $\dot{y}_i$ by the exclusive feature x_i^e. The fake image $\dot{y}_i$ should be completely different from the real image y_i, which enforces x_i^e not to contain any information about the infrared modality. We do the same operations for y_i^e.

The exclusive features generated by the two modalities are exchanged, and we send $\left(x_i^s, y_i^e\right)$ to G_d to generate the image $\hat{y}_i$ of the infrared modality, and $\left(y_i^s, x_i^e\right)$ to F_d to generate the image $\hat{x}_i$ of the visible modality. These operations and the L1 loss adopted here to align shared features from the image

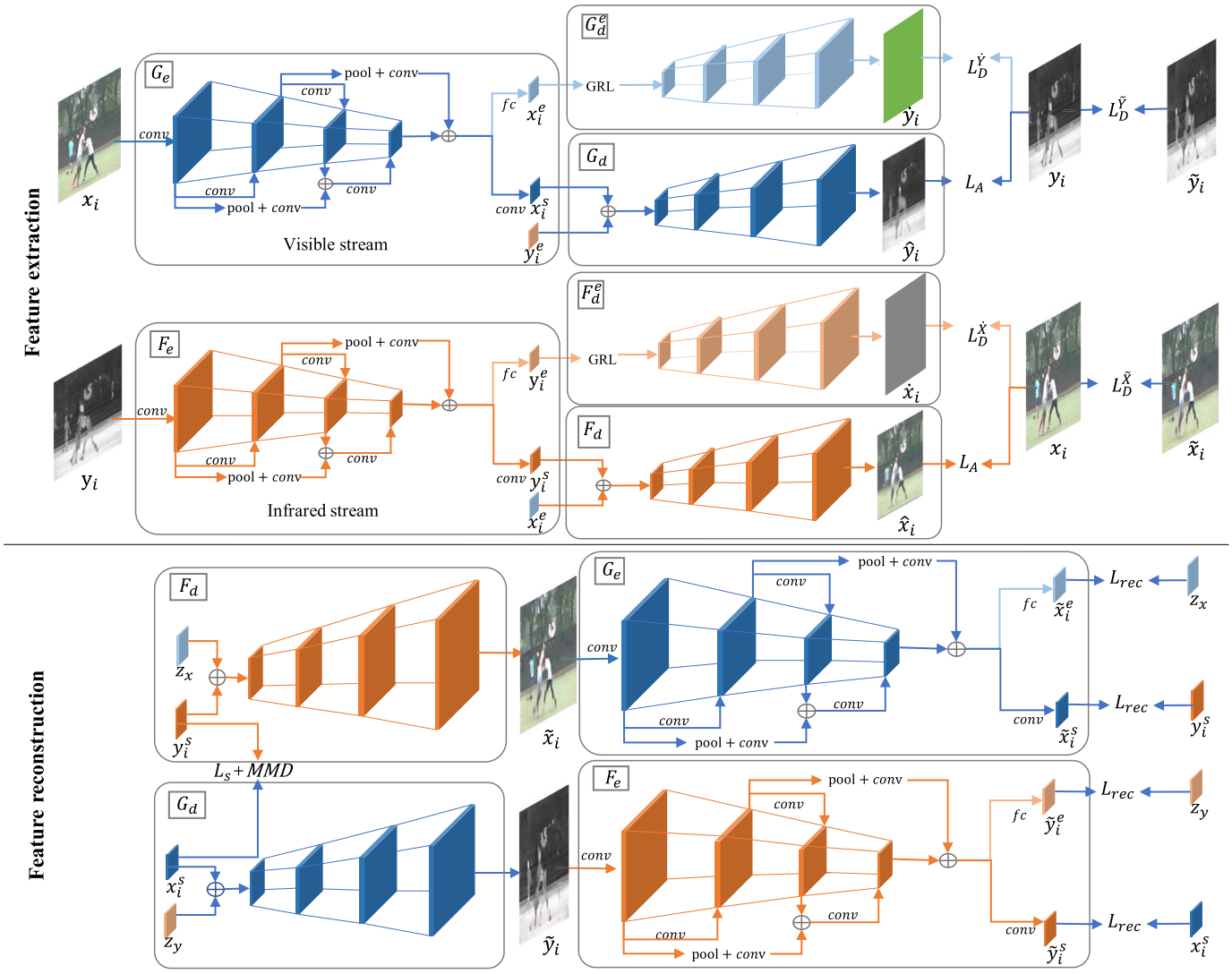


Fig. 3. The proposed feature separation network. The feature extraction module extracts shared features and exclusive features of a pair of images. For the feature reconstruction module, shared features and exclusive features are reconstructed to enhance the separation ability of the encoder.

perspective. The alignment loss function is defined as:

$$L_A = \mathbb{E}_{x\sim X, y\sim Y}[\|\hat{x}_i - x_i\| + \|\hat{y}_i - y_i\|]. \tag{3}$$

2) Feature Reconstruction: The concatenation features (x_i^s, z_y) and (y_i^s, z_x) are fed into G_d and F_d to generate the fake images $\tilde{y}_i$ and $\tilde{x}_i$. Gaussian noises z_x and z_y are sampled from $N(0, 0.1)$. These fake images should be similar to the real images.

The images $\tilde{y}_i$ and $\tilde{x}_i$ are re-input to the encoders F_e and G_e to obtain the generated shared features $\tilde{y}_i^s$, $\tilde{x}_i^s$ and exclusive features $\tilde{y}_i^e$, $\tilde{x}_i^e$ again. The reconstruction loss function is adopted here to encourage that the reconstructed shared features $\tilde{x}_i^s$, $\tilde{y}_i^s$ have the same distribution with x_i^s, y_i^s, and $\tilde{x}_i^e$, $\tilde{y}_i^e$ have the same distribution with z_x and z_y. This loss constrains similar information for two modalities from the feature perspective. More specifically, $L_{rec} = L_{rec}^X + L_{rec}^Y$ is the sum of reconstruction losses,

$$L_{rec}^X = \mathbb{E}_{x\sim X, y\sim Y}\left[\left\|\tilde{x}_i^s - y_i^s\right\| + \left\|\tilde{x}_i^e - z_x\right\|\right], \tag{4}$$

likewise for Y modality.

3) Adversarial Loss: As shown in Fig. 3, for aligning shared features from the image perspective, we introduce four decoders as generators: $G_d^e : X^e \to \dot{Y}_i$, $G_d : (X^e, z_y) \to \tilde{Y}$, $F_d^e : Y^e \to \dot{X}_i$, $F_d : (Y^e, z_x) \to \tilde{X}$. We adopt WGAN-GP [58] for our model to make the training more stable. As F_d for example, we define the loss as:

$$L_D^{\tilde{X}} = \mathbb{E}_{\tilde{x}_i\sim\tilde{X}}[D(\tilde{x}_i)] - \mathbb{E}_{x_i - X}[D(x_i)] + \lambda\mathbb{E}_{x_i'\sim X'}\left[\left(\left\|\nabla_{x_i'} D\left(x_k'\right)\right\|_2 - 1\right)^2\right], \tag{5}$$

where $D(\cdot)$ is a discriminator with a single scalar as output, λ is the penalty coefficient, X' is the distribution obtained by randomly interpolating between real images and fake images and $\|\cdot\|_2$ is 2 norm constraint. G_d^e, F_d^e and G_d possess the same loss configurations with F_d. The total adversary loss is defined as:

$$L_D = L_D^{\tilde{X}} + L_D^{\tilde{Y}} + L_D^{\dot{X}} + L_D^{\dot{Y}}. \tag{6}$$

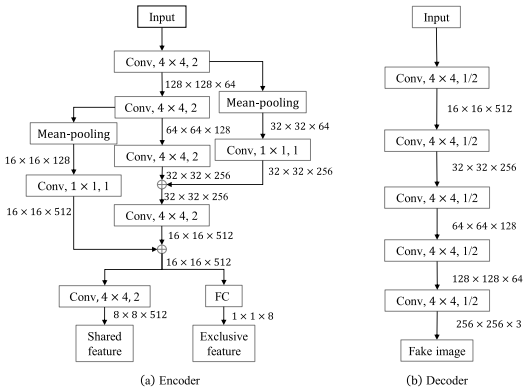


Fig. 4. Encoder-decoder architectures considered in the proposed method.

D. Overall Loss

The proposed method can be trained in an end-to-end fashion, and the full loss is defined as follows:

$$L = w_1 \times (L_{rec} + L_s + L_A) + w_G \times L_D + w_{\text{mmd}} \times MMD, \quad (7)$$

where w_1, w_G and w_{mmd} are trade-off hyper-parameters.

E. Similarity Measurement

After extracting the shared features, we calculate the Euclidean distance to measure the similarity. The Euclidean distance between the nth visible image and the mth infrared image is defined as:

$$D_{nm} = \|x_n^s - y_m^s\|_2. \quad (8)$$

Then we sort the results according to the Euclidean distances, and then choose the Top-10 results as the outputs.

F. Instantiation

The details of the encoder-decoder architectures used in this paper are shown in Fig. 4. As shown in Fig. 4 (a), the architecture of cross-layer fusion encoders G_e and F_e is composed of five convolution layers with the convolution kernel size of 4×4 and the stride of 2, two cross-layer connections and one fully connected layer. We obtain shared features and exclusive features through the last convolution layer and the fully connected layer, respectively. The input size of the encoder is $256 \times 256 \times 3$.

As shown in Fig. 4 (b), the architectures of decoders G_d F_d, G_d^e and F_d^e contain five convolution layers. The size of the convolution kernel is 4×4, and the stride is 1/2. Note that the input size of exclusive feature decoders G_d^e and F_d^e is $8 \times 8 \times 8$, while that of combined feature decoders G_d and F_d is $8 \times 8 \times 520$.

IV. Experiment

A. Experimental Setup

1) Dataset: The proposed method is evaluated on the public infrared-visible video dataset [3]. This dataset includes 12 actions: one hand wave, two hand wave, clap, walk, jog, jump, skip, shake hands, hug, push, punch and fight. Each action is filmed in 13 different scenes under different seasons. We extract 1 or 2 frames from each visible video and then the corresponding infrared frames are selected from the infrared video. Finally, 1363 pairs of images are obtained and we divide those images into training and testing sets with a split of 9:1. Each visible image only has one ground truth infrared image, and vice versa. The resolutions of the visible and infrared images are 293×256 and 480×720, respectively, and we fix the resolution of all images to be 256×256 in our experiment. Some samples of the dataset are shown in Fig. 5.

Fig. 5. Representative samples of the dataset. In each pair of images, the left one is the visible image, while the right one is the infrared image.

2) Evaluation Metrics: The proposed method adopts mean Average Precision (mAP@all) and Top-K precision for performance evaluation. For each infrared image, we rank the visible images according to the Euclidean distance. The Top-1 image is the most similar one. The Top-K precision is defined as:

$$Acc_{Top-K} = \frac{1}{N} \sum_{i=1}^{N} R(i), \quad (9)$$

where N is the number of queried images. $R(i) = 1$ if there is a ground truth image in the current K terms, otherwise, $R(i) = 0$. We can easily find out the ground truth for the query image from the Top-10 images, and thus we set the retrieval fault tolerance from Top-1 to Top-10.

3) Implementation Details: We explore the effect of the hyper-parameters w_1, w_G and w_{mmd}, the results can be seen in Fig. 6. As observed in these curves, the hyper-parameters w_1, w_G and w_{mmd} can be set experimentally to 100, 1 and 1, respectively. We set the initial learning rate as 0.0002. The whole network is trained with the Adam optimizer, and batch size of 8 on NVIDIA GeForce GTX 1080Ti GPU. The framework is implemented using TensorFlow 1.8.0 and accelerated by CUDA 9.0.

B. Comparisons With State-of-the-Art Methods

We compare the proposed method with seven state-of-the-art methods: Fine-tunning CNN image retrieval (FCIR)

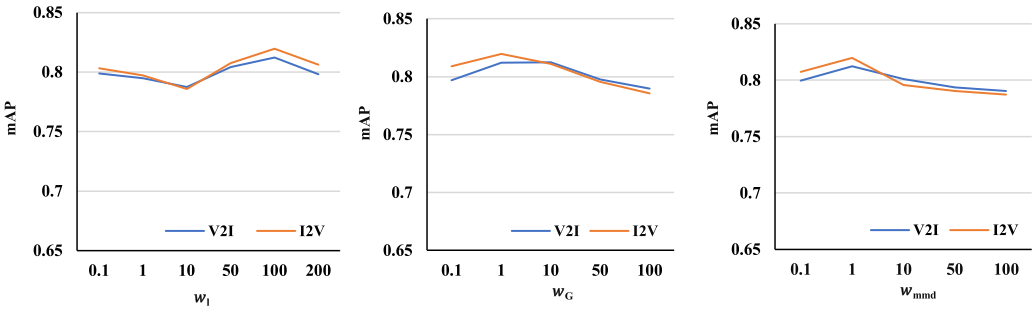


Fig. 6. The effect of hyper-parameters.

TABLE I
PERFORMANCE COMPARISONS OF DIFFERENT METHODS

Methods	V2I Precision (%)					I2V Precision (%)					Time
	mAP	Top-1	Top-2	Top-5	Top-10	mAP	Top-1	Top-2	Top-5	Top-10	(ms)
BicycleGAN [59]	3.48	0.61	0.61	3.06	6.13	4.34	0.61	3.06	5.52	7.36	24.1
DRIT++ [60]	12.17	4.92	9.02	14.75	21.31	10.89	4.09	6.56	13.93	20.49	9.6
FCLR [61]	39.18	24.53	36.81	55.82	67.48	35.89	21.47	31.28	53.89	66.25	89.8
UCAL [39]	46.41	30.06	48.46	66.84	71.77	48.38	33.74	46.62	66.25	77.91	91.2
DSN [15]	59.13	44.17	63.80	78.52	84.66	62.75	49.07	65.03	79.14	88.95	10.1
I2ITCDD [14]	77.48	71.16	78.52	82.82	88.34	74.39	63.19	78.52	86.50	93.25	8.9
CDIM [62]	78.71	72.39	78.52	85.89	92.02	80.91	73.62	82.21	88.34	90.71	92.0
Ours	**81.22**	**75.64**	**81.53**	**87.73**	**93.25**	**81.96**	**75.46**	**82.82**	**88.96**	**96.31**	9.4

with no human annotation [61], unsupervised cross-modal retrieval through adversarial learning (UCAL) [39], cross-domain image matching (CDIM) with deep feature maps [62], BicycleGAN [59], DRIT++ [60], domain separation networks (DSN) [15], and image-to-image translation for cross-domain disentanglement (I2ITCDD) [14].

For FCIR which is a single-modal retrieval method, we adopt a siamese architecture and train a two-branch network to extract visible and infrared image features. For UCAL, we extract the visible and infrared image features by the original image feature extraction branch, respectively. For CDIM, we adopt the proposed multi-channel normalized cross-correlation to measure the similarity between the features extracted from the two modal images. BicycleGAN is an image-to-image translation method, and we calculate the Euclidean distance between the visible and infrared image features output by the encoder. For DRIT++, we extract the content features for retrieval. I2ITCDD and DSN are both shared feature based methods, and we follow the original experimental settings to extract shared features.

1) Quantitative Evaluation: The retrieval precision and the test time (per image) of different methods are shown in Table I, in which V2I means that the infrared image retrieves the visible image, while I2V means the opposite.

It can be observed that the proposed method obviously outperforms all state-of-the-art methods in terms of all precision metrics. As observed, BicycleGAN and DRIT++ perform not well. The possible reason is that it is a pixel alignment method, which is not well met in our case. Comparing the result of UCAL to FCIR, we can find that the cross-modal method is better than the single-modal method. Furthermore, the performances of I2ITCDD and DSN methods is better than UCAL, which shows the importance of leaning shared features of different modalities. The CDIM method performs better than I2ITCDD due to the useful similarity measure way. However, I2ITCDD, CDIM and DSN fail to consider the characteristics of infrared images, so their performances are not as good as the proposed method. The proposed method not only concentrates on the similarity of the two-modal images but also compensates for the loss of infrared image context information via the cross-layer fusion encoder. Besides, the proposed method can address the issue of pixel misalignment by using MMD loss. As for time consumption, FCIR, CDIM and UCAL methods spend more time in extracting features. The proposed method and the rest methods are more efficient because of adopting a concise encoder to extract features.

2) Qualitative Evaluation: Fig. 7 and Fig. 8 show the qualitative evaluation comparisons. When the visible image retrieves infrared images, as shown in Fig. 7, the query image has relatively complex objects and scenes. For this situation, BicycleGAN and DRIT++ can capture similar objects and scenes. However, the pixel misalignment neglects some details that determine the correctness of the retrieval, resulting in the inability to find the correct one. The matched images can be found by UCAL and FCIR, but the rankings of ground truth images are far from the Top-1. For I2ITCDD, CDIM and DSN, we can see that the second-ranked images are ground truth images, and the first-ranked images are deceptive. Due to the lack of detail information on the infrared images, the context information is seriously lost in the convolution, which may lead to the retrieval error of Top-1. The proposed method can accurately retrieve the ground truth in Top-1, which shows that the cross-layer fusion encoder and MMD constraint are effective. When the infrared image retrieves visible images,

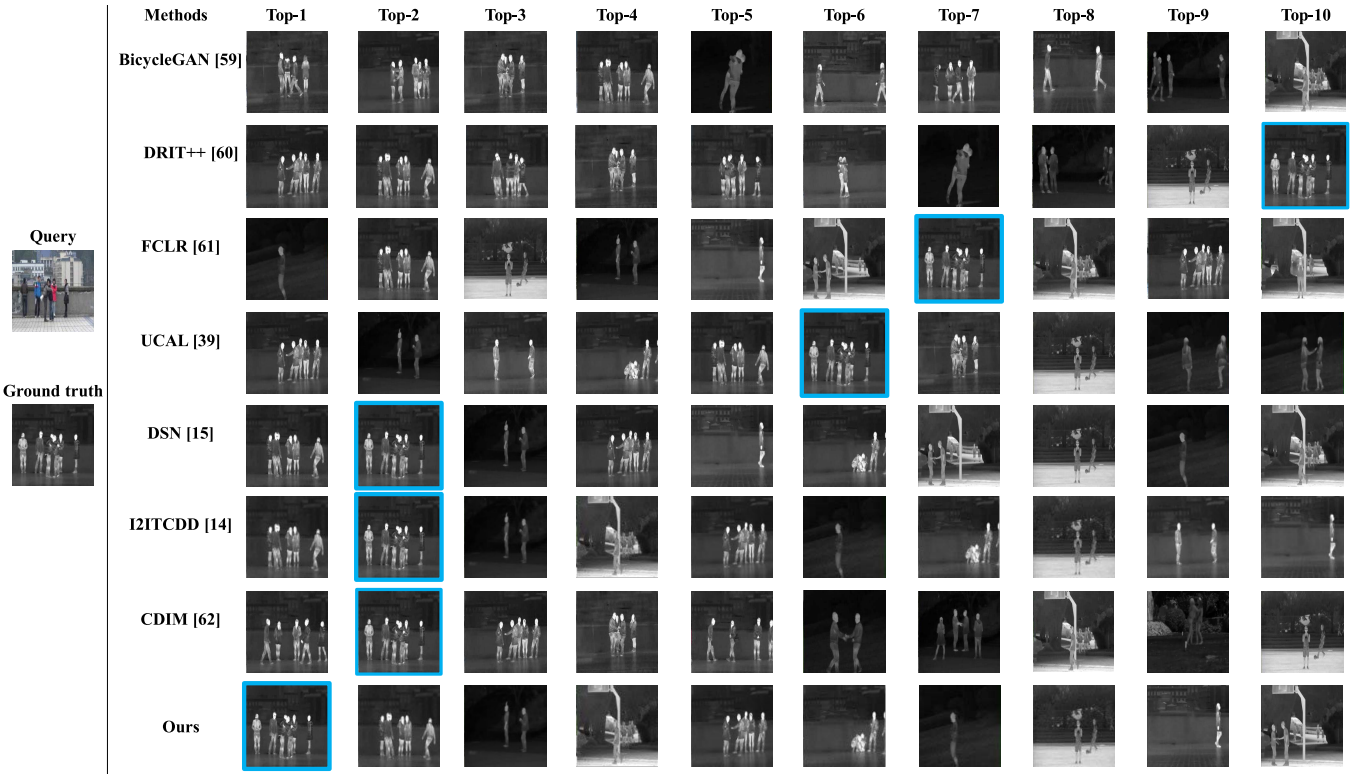


Fig. 7. Qualitative analysis of visible to infrared retrieval. The query visible image and the ground truth infrared image are in the first column and the following images are retrieval results. The similarity score of each result is reduced in turn from left to right, and those with blue boxes are ground truth images.

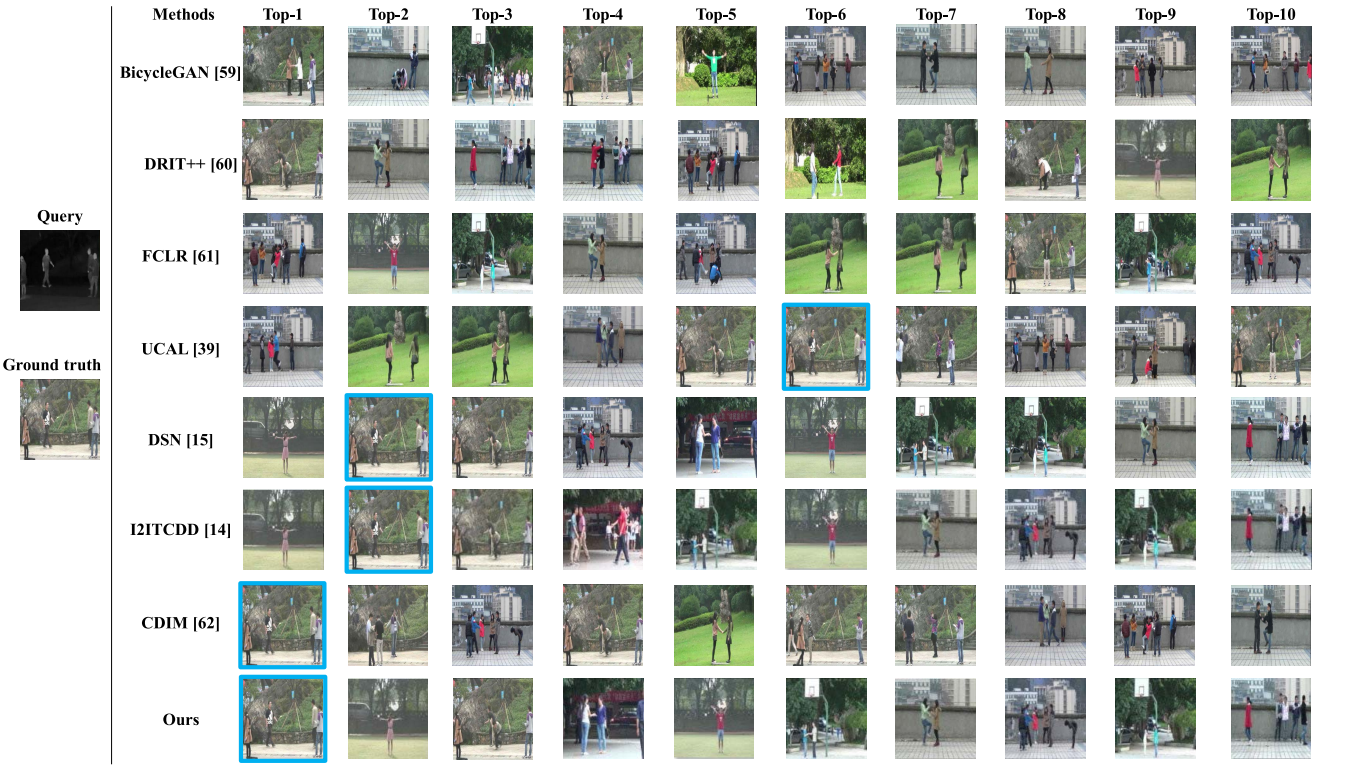


Fig. 8. Qualitative analysis of infrared to visible retrieval. The query infrared image and the ground truth visible image are in the first column and the matching results are shown in the other columns. From left to right, the similarity score of each result is reduced in turn, and those with blue boxes are ground truth images.

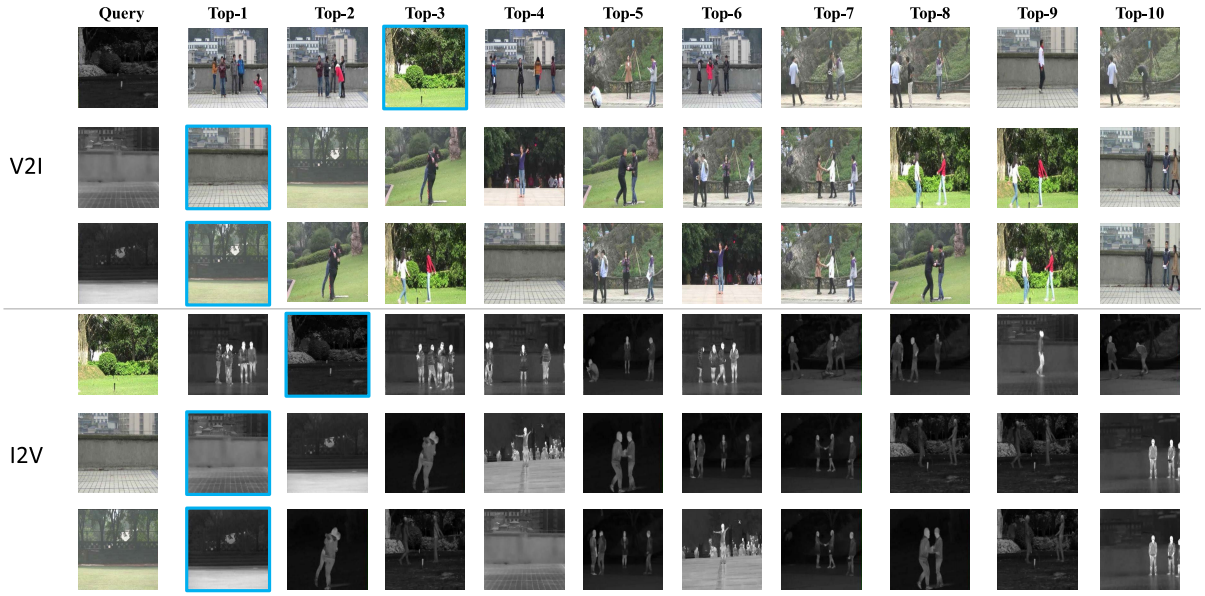


Fig. 9. Qualitative analysis results for the query images without any person, and those with blue boxes are ground truth images.

TABLE II
RESULTS OF mAP@ALL AND TOP-K PRECISION ON DIFFERENT CONFIGURATIONS

Tasks	Configurations	Precision (%)				
		mAP@all	Top-1	Top-2	Top-5	Top-10
V2I	No MMD	74.18	65.64	77.30	84.04	85.89
	No fusion	79.21	72.39	79.14	85.89	**93.25**
	Ours	**81.22**	**75.64**	**81.53**	**87.73**	**93.25**
I2V	No MMD	73.77	65.03	74.23	84.66	91.41
	No fusion	78.34	69.93	79.75	87.73	**96.31**
	Ours	**81.96**	**75.46**	**82.82**	**88.96**	**96.31**

as shown in Fig. 8, the observations are similar except the result of CDIM. We can find the correct image in Top-1 by using the proposed method and CDIM.

To further comprehensively explore the effectiveness of the proposed method, we randomly select three images without any foreground person for qualitative analysis, as shown in Fig. 9. When there is no foreground person, the proposed method can still perform well.

C. Ablation Study

In this section, we evaluate the effectiveness of key components used in the proposed method. We remove MMD loss or fusion layers to explore their effect on the performance of the proposed method. Specifically, when we remove fusion layers, the encoder has five convolution layers and one fully connected layer.

Experimental results in Table II obviously show the promotion effect of each component on the retrieval results. As can be seen, MMD loss gains more than 7% improvement on mAP@all, and Top-1 precision improved nearly by 10%. It shows that the distribution constraint on a pair of images features can greatly improve the retrieval precision. Meanwhile, the fusion method improves the Top-1 precision about by 4%, and the mAP@all about by 3%.

TABLE III
COMPARISON OF THE EUCLIDEAN DISTANCES WITH OR WITHOUT MMD LOSS. THE POSITIONS OF THE BOLD NUMBERS ARE WHERE THE GROUND TRUTH IMAGE APPEARS

Tasks	Methods	number	Euclidean distance			
			Top-1	Top-2	Top-5	Top-10
V2I	No MMD	12	6.7473	**6.8013**	7.3321	8.8226
		41	10.4330	**10.4884**	10.5891	10.9237
		50	10.7983	10.8077	11.0564	**11.1630**
	Ours	12	**4.0253**	4.0262	4.3607	5.3908
		41	**6.5588**	6.6148	6.6877	6.9233
		50	**6.1822**	6.3337	6.4501	6.6545
I2V	No MMD	12	5.7889	**6.0558**	6.5227	7.5095
		41	8.2315	8.4084	8.4991	8.6446
		50	7.9582	8.2962	8.4736	8.5039
	Ours	12	**4.0253**	4.0715	4.8688	5.4889
		41	6.4500	6.4702	6.5305	6.6284
		50	**6.1822**	6.2968	6.3971	6.5061

1) The Performance of the MMD: For the sake of confirming that MMD has the effect of shortening the distance between shared features, we randomly select three images, and compare the 1st, 2nd, 5th and 10th placed images' Euclidean distances with or without MMD loss. As shown in Table III, the best match images for No. 12, No. 41, No. 50 images appear at places 2nd, 2nd, and 10th if removing MMD loss, when the visible image retrieves the infrared image. After adding the MMD constraint, the positions of the most matched items are advanced, ranking in the 1st, 1st and 1st, respectively. Moreover, the Euclidean distance between the shared features is reduced. For the case that the infrared image retrieves the visible image, the observation is also the same. Note that without MMD loss, the ground truth images of the No. 41 and No. 50 infrared images rank in 9th and 4th, respectively, so the distance between these two images and the query images are not shown in Table III. With MMD loss, the ground truth image of the No. 41 infrared image ranked in 6th, so the distance between the image and the No. 41 image is also not shown in Table III.

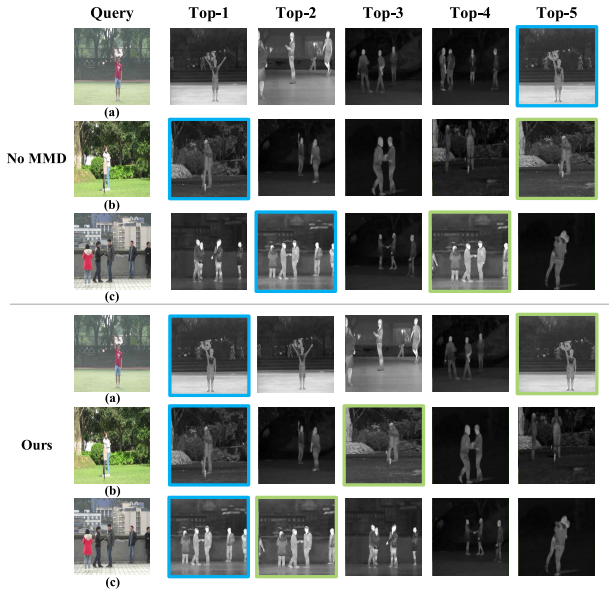


Fig. 10. Visualization of aligned and misaligned images retrieval results, and those with blue boxes are aligned ground truth images, while those with green boxes are misaligned ground truth images.

Furthermore, to investigate the performance of MMD loss in reducing mismatching error caused by different imaging angles and similar infrared images, we artificially produce several pairs of perfectly pixel-aligned images and misaligned images, then put them into the test set for testing. As V2I for example, we visualize the results of randomly selected three pairs of images, as shown in Fig. 10. As can be seen in these images, the aligned ground truth images can be found in Top-5 results in both the No-MMD method and the proposed method. As for misaligned images, we can not find the misaligned ground truth image of the image (a) by the No-MMD method. However, for the proposed method, we can exactly find the aligned ground truth images in Top-1 results and the misaligned ground truth images in Top-5. Besides, we can find that the Top-1 results of image (a) and (c) are pretty deceptive, due to the low resolution of infrared images. The No-MMD method can not distinguish between the deceptive image and the ground truth image. However, the proposed method can distinguish between the deceptive image and the ground truth image, and find the ground truth image in Top-1. Compared mismatched images with misaligned images, due to the coexistence of intra-modality and cross-modality discrepancies, other mismatched candidates may be closer to the query images than misaligned images. After adding the MMD constraint, the phenomenon is alleviated to some extent.

2) The Performance of the Cross-Layer Fusion Encoder: To explore the enhancement effect of the cross-layer fusion encoder on the context information of the infrared image, we feed some images into the fusion encoder and no-fusion encoder, respectively, and obtain the heat maps after the fourth convolution layer. The heat map can reflect the image feature intensity obtained from the feature maps, thus reflecting the feature extraction capability of the network. The heat map results can be found in Fig. 11, in which the first line is the

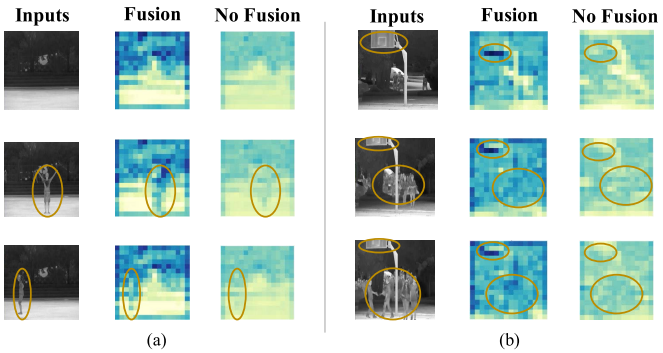


Fig. 11. Visualization of the heat maps output by convolution of the fourth layer before and after adding cross-layer fusion.

images with only background scenes, while the second and third lines are the images which are added different foreground objects. As can be seen from Fig. 11 (a) and (b), when some objects appear in the images, the corresponding position in the heat maps will change obviously with cross-layer fusion encoder. However, the heat maps will change less obviously with no-fusion encoder. It shows that our cross-layer fusion encoder can better extract the contour features of the infrared images and reduce the loss of context information in the process of convolution.

D. Relation to the Person Re-ID Task

Although the task handled in this paper is related to the person Re-ID task, they are of different tasks. The person Re-ID task considers one specific person in the image, while the proposed method focuses on image matching, which considering the whole image which could contains multiple persons or just background. Directly applying the proposed method to the person Re-ID task usually can not work well. We tested the proposed method on the person Re-ID dataset SYSU-MM01 [55], and the mAP is 1.69%.

V. Conclusion

In this paper, we propose a cross-modal image retrieval network based on feature separation, which utilizes the shared features of two modalities for retrieval. Considering that infrared images have fewer details inherently, cross-layer fusion is added to reduce the loss of context information during convolution, and then MMD is used reduce mismatching error and misjudgment of similar infrared images. Compared with state-of-the-art methods, the experimental results show that the proposed method has higher retrieval precision.

References

[1] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[2] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognit.*, vol. 76, pp. 463–475, Apr. 2018.

[3] L. Wang, C. Gao, Y. Zhao, T. Song, and Q. Feng, "Infrared and visible image registration using transformer adversarial network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1248–1252.

[4] L. Wang, C. Gao, L. Yang, Y. Zhao, W. Zuo, and D. Meng, "PM-GANs: Discriminative representation learning for action recognition using partial-modalities," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 384–401.

[5] C. Gao *et al.*, "InfAR dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36–47, Nov. 2016.

[6] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[7] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[8] R. R. Saritha, V. Paul, and P. G. Kumar, "Content based image retrieval using deep learning process," *Cluster Comput.*, vol. 22, no. S2, pp. 4187–4200, Mar. 2019.

[9] A. Latif *et al.*, "Content-based image retrieval and feature extraction: A comprehensive review," *Math. Problems Eng.*, vol. 2019, Aug. 2019, Art. no. 9658350.

[10] J. Zhang and Y. Peng, "SSDH: Semi-supervised deep hashing for large scale image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 212–225, Jan. 2019.

[11] A. Araujo and B. Girod, "Large-scale video retrieval using image queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1406–1420, Jun. 2018.

[12] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2049–2055.

[13] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[14] A. Gonzalez-Garcia, J. V. D. Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1287–1298.

[15] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and A. D. Erhan, "Domain separation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[16] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.

[17] A. Dureja and P. Pahwa, "Image retrieval techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 1.2, p. 215, Dec. 2017.

[18] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Apr. 29, 2020, doi: 10.1109/TCSVT.2020.2991171.

[19] C.-C. Lai and Y.-C. Chen, "A user-oriented image retrieval system based on interactive genetic algorithm," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 10, pp. 3318–3325, Oct. 2011.

[20] J. Wan *et al.*, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 157–166.

[21] J. Y.-H. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 53–61.

[22] M. Tzelepi and A. Tefas, "Relevance feedback in deep convolutional neural networks for content based image retrieval," in *Proc. 9th Hellenic Conf. Artif. Intell.*, May 2016, pp. 1–7.

[23] W. Yu, K. Yang, H. Yao, X. Sun, and P. Xu, "Exploiting the complementary strengths of multi-layer CNN features for image retrieval," *Neurocomputing*, vol. 237, pp. 235–241, May 2017.

[24] M. Korichi, M. L. Kherfi, M. Batouche, Z. Kaoudja, and A. Bencheikh, "Understanding user's intention in semantic based image retrieval: Combining positive and negative examples," in *Proc. IFIP Int. Conf. Comput. Intell. Appl.* Oran, Algeria: Springer, 2018, pp. 66–77.

[25] N. Ruan, N. Huang, and W. Hong, "Semantic-based image retrieval in remote sensing archive: An ontology approach," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, Jul. 2006, pp. 2903–2906.

[26] A. Lakdashti, M. S. Moin, and K. Badie, "A novel semantic-based image retrieval method," in *Proc. 10th Int. Conf. Adv. Commun. Technol.*, Feb. 2008, pp. 969–974.

[27] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1269–1277.

[28] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 241–257.

[29] K. Song, F. Li, F. Long, J. Wang, and Q. Ling, "Discriminative deep feature learning for semantic-based image retrieval," *IEEE Access*, vol. 6, pp. 44268–44280, 2018.

[30] S. Unar, X. Wang, C. Zhang, and C. Wang, "Detected text-based image retrieval approach for textual images," *IET Image Process.*, vol. 13, no. 3, pp. 515–521, Feb. 2019.

[31] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.

[32] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[33] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[34] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.

[35] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Yng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.* Washington, DC, USA: ACM, 2011, pp. 689–696.

[36] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[38] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.

[39] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1153–1158.

[40] Z. Wang *et al.*, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019.

[41] J. Dong *et al.*, "Dual encoding for zero-example video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9346–9355.

[42] Y. Peng and J. Qi, "Reinforced cross-media correlation learning by context-aware bidirectional translation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1718–1731, Jun. 2020.

[43] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3232–3240.

[44] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3419–3427.

[45] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2075–2082.

[46] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.

[47] J.-Q. Xiao, Z.-Y. Zhou, and X.-Q. Zhou, "Self-supervised adversarial hashing cross-modal retrieval with generative models based on attention mechanism," *DEStech Trans. Comput. Sci. Eng.*, May 2019, doi: 10.127 83/dtcse/icaic2019/29441.

[48] D. Wang, Q. Wang, and X. Gao, "Robust and flexible discrete hashing for cross-modal similarity search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2703–2715, Oct. 2018.

[49] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5298–5309.

[50] Y. Lu *et al.*, "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13379–13389.

[51] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10257–10266.

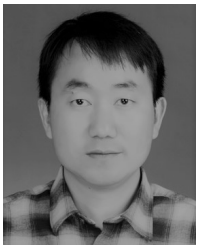
[52] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3422–3432, Oct. 2020.

[53] H. Yong, D. Meng, W. Zuo, and L. Zhang, "Robust online matrix factorization for dynamic background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1726–1740, Jul. 2018.

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

[55] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5380–5389.
[56] F. Zhou *et al.*, "Face anti-spoofing based on multi-layer domain adaptation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 192–197.
[57] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014, *arXiv:1409.7495*. [Online]. Available: http://arxiv.org/abs/1409.7495
[58] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[59] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.
[60] H.-Y. Lee *et al.*, "Drit++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, pp. 2402–2417, Feb. 2020.
[61] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.
[62] B. Kong, J. Supancic, D. Ramanan, and C. C. Fowlkes, "Cross-domain image matching with deep feature maps," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1738–1750, Dec. 2019.

Fangcen Liu received the bachelor's degree in information engineering from the Chongqing University of Posts and Telecommunications, China, in 2018, where she is currently pursuing the master's degree. Her research interests include image processing, deep learning, and cross-modal retrieval.

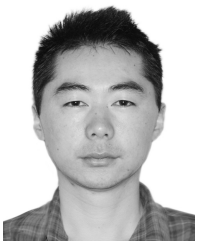


Chenqiang Gao received the B.S. degree in computer science from the China University of Geosciences, Wuhan, China, in 2004, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, in 2009. In August 2009, he joined the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China. In September 2012, he joined the Informedia Group, School of Computer Science, Carnegie Mellon University (CMU), working on Multimedia Event Detection (MED) and Surveillance Event Detection (SED) as a Visiting Scholar. In April 2013, he became a Post-Doctoral Fellow and continued work on MED and SED, till March 2014 when he returned to CQUPT. He is currently a Professor with CQUPT. His research interests include image processing, infrared target detection, action recognition, and event detection.

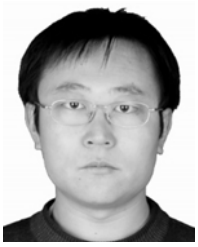
Yongqing Sun (Member, IEEE) received the B.E. and M.E. degrees in computer science from Xi'an Jiaotong University (XJTU), China, in 1998 and 2001, respectively, and the Ph.D. degree in information and computer science from Keio University, Japan, in 2005. In 2005, she joined NTT laboratories, where she has been engaged in research and development of video/image processing technologies. Her research interests include image processing, pattern recognition, machine learning, data mining, and multimedia retrieval.

Yue Zhao received the Ph.D. degree from Jilin University, China, in 2017. She is currently a Lecturer with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. Her research interests include image processing and analysis, pattern recognition, and medical image processing.

Feng Yang received the B.S. degree in science and technology of remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013 and 2018, respectively. She is currently a Lecturer with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. Her research interests include dynamic texture analysis, video processing, and remote sensing imaging.



Anyong Qin received the B.S. degree in information and computational science and the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2012 and 2019, respectively. He is currently an Associate Professor with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing. His research interests include pattern recognition, signal processing, image processing, machine learning, and hyperspectral images.



Deyu Meng (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. He is currently a Professor with the Institute for Information and System Sciences, Xi'an Jiaotong University. His current research interests include self-paced learning, noise modeling, and tensor sparsity.