# PM-GANs: Discriminative Representation Learning for Action Recognition Using Partial-Modalities

Lan Wang[1,2], Chenqiang Gao[1,2](✉), Luyu Yang[3], Yue Zhao[1,2], Wangmeng Zuo[4], and Deyu Meng[5]

[1] School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
gaocq@cqupt.edu.cn
[2] Chongqing Key Laboratory of Signal and Information Processing,
Chongqing 400065, China
[3] University of Maryland College Park, College Park, MD 20742, USA
[4] Harbin Institute of Technology, Harbin 150001, China
[5] Xi'an Jiaotong University, Xi'an 710049, China

**Abstract.** Data of different modalities generally convey complimentary but heterogeneous information, and a more discriminative representation is often preferred by combining multiple data modalities like the RGB and infrared features. However in reality, obtaining both data channels is challenging due to many limitations. For example, the RGB surveillance cameras are often restricted from private spaces, which is in conflict with the need of abnormal activity detection for personal security. As a result, using partial data channels to build a full representation of multi-modalities is clearly desired. In this paper, we propose a novel Partial-modal Generative Adversarial Networks (PM-GANs) that learns a full-modal representation using data from only partial modalities. The full representation is achieved by a generated representation in place of the missing data channel. Extensive experiments are conducted to verify the performance of our proposed method on action recognition, compared with four state-of-the-art methods. Meanwhile, a new Infrared-Visible Dataset for action recognition is introduced, and will be the first publicly available action dataset that contains paired infrared and visible spectrum. (The dataset will be available at http://www.escience.cn/people/gaochenqiang/Publications.html).

**Keywords:** Cross-modal representation
Generative adversarial networks · Infrared action recognition
Infrared dataset

## 1 Introduction

Human action recognition [11,31,48,51,55,59] aims to recognize the ongoing action from a video clip. As one of the most important tasks in computer vision,

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11210, pp. 389–406, 2018.
https://doi.org/10.1007/978-3-030-01231-1_24

action recognition plays a significant role in many useful applications like video surveillance [24,49], human-computer interaction [32,43] and content retrieval [2,60], with great potentials in artificial intelligence. As a result, massive attention has been dedicated to this area which made large progress over the past decades. Most state-of-the-art methods have contributed to the tasks in visible imaging videos, and show saturated performances among the widely-used benchmark datasets including KTH [52] and UCF101 [14]. Generally speaking, the task of action recognition is quite well-addressed and has already been applied to real-world problems.

However, there are still many occasions where visible imaging is limited. First, the RGB cameras rely heavily on the light conditions, and perform poorly when light is insufficient or over-abundant. Action recognition from night-view RGB data remains a rather difficult task. Moreover, as an act to protect the fundamental human dignity–Privacy, RGB cameras are strictly restricted from most private areas including the personal residential area, public washroom where abnormal human activities are likely to threat personal security. Infrared cameras, that capture the heat radiation of objects, are excellent alternatives in these occasions [13]. The application of thermal imaging in military affairs and police surveillance has continued for years, and has more potentials beyond the government use. With many advantages over the RGB camera, it is predicted that infrared cameras will become more common in public spaces like hospitals, nursing centers for elderly and home security systems [35].

While infrared cameras can fill the limited spots of RGB cameras, many visible features are nevertheless lost in the infrared spectrum due to their similarity in temperature [58,62]. Visible features like color, texture are effective clues in activity representations. Since the two are complementary to each other, it is desired to utilize both visible and infrared features to benefit the task of action recognition. Furthermore, it will be more desired to utilize both feature domains when ONLY infrared data is available. In the previous cases when the demand of abnormal action recognition and the demand of privacy conflicted, it will be great if we can obtain both infrared and visible features, while use only the infrared data. The question is, how can one obtain visible features when the visible data is missing? The situation is not unique to the task of action recognition. In fact, data with different modalities of complementary benefits widely exists in multimedia such as systems with multiple sensors, product details with combined information of text description and images [39]. Here we are inspired by the intra-modal feature representations to make up for the missing data using adversarial learning with the available part of the data channel.

Recently, much attention has been given to cross-modal feature representations [6,10,16,23,57] dealing with unpaired data, which maps multiple feature spaces onto a common one, or generates a different representation via adversarial training. The basic model of generative adversarial networks (GANs) [8,15,41] consists of a generative model $G$ and a discriminative model $D$. Many interesting image-to-image translations such as genre translation, face and pose transformation indicate the broader potentials of GANs to explore the hidden correlations in cross-modal representations [35,38]. Inspired by this, we therefore seek an algorithm that can translate from the infrared representation to the visible domain, which allows us to further exploit the benefits of both feature spaces with only part of the data modalities. More generally speaking, we aim at an architecture that learns a full representation for data of different modalities, using partial modalities. Different from the existing works of cross-modal which seeks a common representation from different data spaces, our goal is to exploit the transferable ability among different modalities, which is further utilized to construct a full-modal representation when only partial data modalities are available.

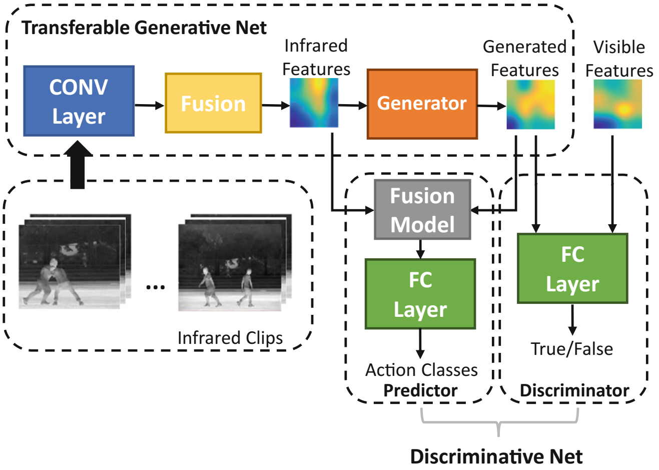


**Fig. 1.** The framework of the proposed Partial-modal Generative Adversarial Networks (PM-GANs). Infrared video clips are sent to the transferable generative net to produce fake feature representation of the visible spectrum. And the discriminator attempts to distinguish between the generated features and the real ones. The predictor construct a full representation using the generated features and infrared features to conduct classification

With a completely different target, in this paper we propose a novel Partial-modal Generative Adversarial Networks (PM-GANs), which aims to learn the transferable representation among data of heterogeneous modalities using cross-modal adversarial mechanism and build discriminative full-modal representation architecture using data of one/partial modalities. The main contributions are summarized as follows.

- **Partial-modal representation** is proposed to deal with missing data modalities. Specifically, the partial-modal representation aims to obtain the transferable representation among data with different modalities. And when only partial-modal representations are accessible, the model can still generate a comprehensive description, as if constructed with data of all modalities.

- **Partial-modal GANs architecture** is proposed that can exploit the complementary benefits of all data channels with heterogeneous features using only one/partial channels. The generative model learns to fit the transferable distribution that characterizes the feature representation in the specific data channels that are likely to be missing in practice. Meanwhile, the discriminative model learns to judge whether the translated distribution is representative enough for the full modalities. Extensive experiment results reveal the effectiveness of the PM-GANs architecture, which outperforms four state-of-the-art methods in the task of action recognition.
- **Partial-modal evaluation dataset** is newly introduced, which provides paired data of two different modalities–visible and infrared spectrum of human actions. Researchers can evaluate the transferable ability of the algorithms between the two modalities, as well as the discriminative ability of the generated representation by comparing with a series of baselines we provided in this paper. Meanwhile, the dataset can be used as a benchmark for bi-channel action recognition, since it is also carefully designed to serve for this purpose. The dataset contains more than 2,000 videos, 12 different actions, and to the best of our knowledge, is the first publicly available action recognition dataset that contains both infrared and visible spectrum.

The rest of the paper is organized as follows. In Sect. 2, we review the background and related works. In Sect. 3, we elaborate the details of our proposed method. Section 4 presents the newly-introduced dataset, its evaluations, and the experimental results on it. Finally, Sect. 5 draws the conclusion.

## 2 Related Work

**Transfer Learning and Cross-Modal Representation:** In the classical pattern recognition and machine learning tasks, sufficient training data that has variations in modality is clearly a desired but unrealistic goal [46,47], thus restricting the representative ability of the model. Among the studies to address this problem, transfer learning attempts to transfer the feature space from a source domain to a target domain, and to lessen the adaption conflicts via domain adaption [12,36,37,50]. The transferred knowledge type is not restricted to feature representation or instance, and it also contains modality-correlation. With different aims, cross-dataset and cross-modal feature representation fall into feature-representation transfer by adapting the representations from different domains to a single common latent space, where features of multiple modalities are jointly learned and combined. Among these algorithms, Canonical Correlation Analysis (CCA) [19,61] is a widely used one, which seeks to maximize the correlation between the projected vectors of two modalities. Another classical algorithm is Data Fusion Hashing (DFH) [3] that embeds the input data from two arbitrary spaces into a Hamming space in a supervised way. Differently, Cross-View Hashing (CVH) [27] maximizes the weighted cumulative correlation and can be viewed as the general representation of CCA.

In recent years, with the renaissance of neural networks, many deep learning based transfer learning and cross-modal representation methods have been proposed as well. Cross Modal Distillation (CMD) [16] learns representations for modalities with limited labeled data which are not able to be directly trained on deep networks. Bishifting Autoencoder Network [23] attempts to alleviate the discrepancy between the source and target datasets to the same space. To further take the feature alignment and auxiliary domain data into consideration, Aligned-to-generalized encoder (AGE) [35] is proposed to map the aligned feature representations to the same generalized feature space with low intraclass variation and high interclass variation. Since GANs have been proposed by Goodfellow *et al.* [15] in 2014, a series of GANs-based methods have arisen for a wide variety of problems. Recently, a Cross-modal Generative Adversarial Networks for Common Representation Learning (CM-GANs) [38] is proposed. CM-GANs seeks to unify the inconsistent distribution and representation of different modalities by filling the heterogeneity of knowledge types like image and text. In contrast, we have completely different goal, which aims to use only partial data modalities to obtain a full-modal representation. Our focus is beyond the jointly-learned representation of multiple feature spaces, and takes one step further to achieve a discriminative partial-modality representation, which corresponds to our original aim of handling the problem of insufficient training data and data types.

**Infrared Action Recognition and Dataset:** Most previous contributions [33,42] to the progress of action recognition have been made to the visible spectrum. Early approaches utilized the hand-crafted representation followed by classifiers, such as 3D Histogram of Gradient (HOG3D) [26], Histogram of Optical Flow (HOF) [29], Space Time Interest Points (STIP) [28] and Trajectories [53]. Wang et al. [54] proposed the Improved Dense Trajectories (iDT) representation, making breakthroughs among hand-crafted features. In hand-crafted representation scheme, encoding methods such as Bag of Words (BoW) [30], Fisher vector [44], VLAD [7] are applied to aggregate the descriptors into video-level representation. Benefiting from the success of Convolutional Neural Networks (CNNs) in image classification, several deep network architectures have been proposed for action recognition. Simonyan *et al.* [48] proposed a two-stream CNNs architecture which simultaneously captured appearance and motion information by spatial and temporal nets. Tran *et al.* [49] investigated 3D ConvNets [21,24] in large-scale supervised training datasets and effective deep architectures, achieving significant improvements. Wang *et al.* [56] proposed a temporal segment network to investigate long-term temporal information. Carreira *et al.* [5] designed a two-stream inflated 3D ConvNet, inflating filters and pooling kernels into 3D to learn seamless spatiotemporal feature extractors.

Recently, increasing efforts have been devoted to infrared action recognition [13]. Corresponding to the classical methods employed in visible spectrum, spatiotemporal representation for human action recognition is also used under thermal imaging scenarios [17]. The combination of both visible and thermal imaging to improve human silhouette detection is also introduced by Han *et al.* [18]. However, the scenario has not been studied where infrared data is available while the RGB channel is missing. The scenario has great potential in real-world of protecting privacy while benefiting the task of action recognition, and is meaningful to both the study of pattern recognition and the welfare of the community at large. Therefore, we are motivated to dedicate to improving the situation by constructing a robust and discriminative partial-modal representations, and to specify action recognition as the case in this paper.

# 3 Proposed Approach

The overall pipeline of the proposed PM-GANs for action recognition is shown in Fig. 1. Our goal is to generate a full-modal representation using only the partial modalities. The framework learns the transferable representation among different data channels based on conditional adversarial networks. Based on the transferred representation, the framework builds a discriminative full-modal representation network using only part of the data channels.

## 3.1 Transferable Basis for Partial Modality

The transferable ability with the PM-GANs architecture is the basis for the construction of full-modal representation with partial modality. We assume that there exists a mapping from an observed distribution $f_{Vis}$ and an input distribution $f_{Inf}$, producing an output representation which shares the feature of the observed $f_{Vis}$. Therefore, we attempt to learn a generator $G$ to generate the feature distribution of the missing data channel $f_{Vis}$ from the partially available distribution denoted as $f_{Inf}$. Based on the scheme of conditional adversarial networks, the generator immediately transforms the partially available distribution $f_{Inf}$ and noise $z$ to output the missing distribution via the following equation:

$$\begin{aligned}\min_G \max_D \mathcal{L}(G, D) =& \mathbb{E}_{f_{Vis} \sim P_{data}(f_{Vis})}[\log D(f_{Vis})] + \\ & \mathbb{E}_{f_{Inf} \sim P_{data}(f_{Inf}), z \sim P_{data}(z)}[\log(1 - D(G(f_{Inf}, z)))],\end{aligned} \tag{1}$$

where $G(f_{Inf}, z)$ denotes the output distribution. The input distribution $f_{Inf}$ and observed distribution $f_{Vis}$ denote the data of infrared and RGB channels respectively in our action recognition task. The generator $G$ is designed to minimize this objective to fake the generated distribution as well as possible, while the real output feature discriminator $D$ tries to maximize its accuracy of telling the real from the fake one.

In this work, the discriminator is also designed for pattern recognition. Thus, another prediction loss is explored:

$$\mathcal{L}_p(G, D_p) = \mathbb{E}_{f_{Inf} \sim P_{data}(f_{Inf}), z \sim P_{data}(z)}[L_{cls}(l, D_p(f_{Inf}, G(f_{Inf}, z)))], \tag{2}$$

where $l$ denotes the correct label of partially available data samples, in the form of one-hot vector, and $L_{cls}$ is log loss over the predicted class confidences vector and the ground truth label. For convenience, we denote the discriminator part and the predictor part of discriminative net as $D_d$ and $D_p$ respectively. Finally, the objective function can be formulated as:

$$\begin{aligned}\mathcal{L}_{PM-GANs}(G, D_d, D_p) = &-\mathbb{E}_{f_{Vis}\sim P_{data}(f_{Vis})}[\log D_d(f_{Vis})]\\ &-\mathbb{E}_{f_{Inf}\sim P_{data}(f_{Inf}),z\sim P_{data}(z)}[\log(1-D_d(G(f_{Inf},z)))]\\ &+\mathbb{E}_{f_{Inf}\sim P_{data}(f_{Inf}),z\sim P_{data}(z)}[L_{cls}(l, D_p(f_{Inf}, G(f_{Inf},z)))].\end{aligned} \tag{3}$$

### 3.2 Transferable Net

The transferable net simulates the target distribution from the convolutional feature map of the partially-available data distribution, which, as shown in Fig. 2, is made as an input of the generator to obtain feature maps of missing distribution. The input clips are denoted as $\{f_{Inf}^{(1)}, f_{Inf}^{(2)}, \dots, f_{Inf}^{(T)}\}$, where $f_{Inf}^{(i)} \in \mathbb{R}^{H\times W\times D}$ and $H, W, D$ denote the height, width and number of channels of feature maps. To incorporate all feature maps into a high-level representation, the sum fusion model in [9] is applied to compute the sum of $T$ feature maps at the same spatial location $i, j$ and feature channel $d$:

$$f_{Inf}^{sum}(i,j,d) = f_{Inf}^{(1)}(i,j,d) + f_{Inf}^{(2)}(i,j,d) + \dots + f_{Inf}^{(T)}(i,j,d), \tag{4}$$

where $1 \leq i \leq H$, $1 \leq j \leq W$, $1 \leq d \leq D$. The final feature map of the input distribution is computed as the average value of sum feature map $f_{Inf}^{sum}$ in each

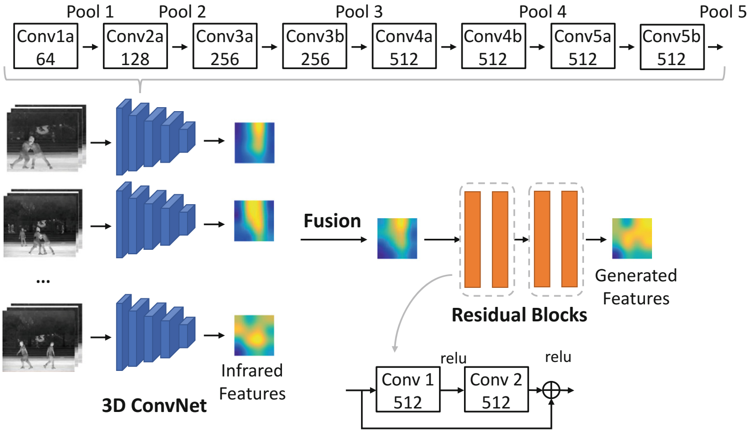


**Fig. 2.** The proposed transferable generative net is built upon the C3D network [49]. Video clips are sent to 3D ConvNet to obtain feature maps of each clip $f_{Inf}^{(i)}$ and all feature maps are fused as $f_{Inf}$ to represent the whole action video. Then, residual blocks are added to this net to produce fake feature maps $f_g$ similar to the visible spectrum

location, denoted as $f_{Inf}$. Then the generator takes the final input feature map and generates the fake target feature map, $G(f_{Inf}, z)$. The generator consists of two residual blocks [20] to produce feature map with the same size as infrared feature map. Thus, the generative loss $L_G$ is expressed as:

$$L_G = -\log(D_d(G(f_{Inf}, z))). \tag{5}$$

### 3.3 Discriminative Net Using Partial Modality

To enable the generative net to produce full-modal representation which incorporates the complementary benefits among data of different modalities, a two-part discriminative net is designed, as shown in Fig. 1. The discriminative net contains a discriminator part and a predictor part. The discriminator part follows the scheme of conventional discriminator in GAN which is applied to distinguish between real and fake visible feature map in order to boost the quality of generated fake feature. Specifically, the discriminator part consists of a fully-connected layer followed by a sigmoid function, which produces an adversarial loss. Thus, the adversarial loss $L_a$ is defined as:

$$L_a = -\log D_d(f_{Vis}) - \log(1 - D_d(G(f_{Inf,z}))), \tag{6}$$

where $L_a$ encourages the discriminator network to distinguish the generated target feature representation from real one.

The predictor aims to boost the accuracy of assigning the right label to each feature distribution. It consists of a fully-connected layer followed by a softmax layer which takes the fusion of the feature map of both the partially-available data channel and generated missing channel and finally outputs the category-level confidences. To fuse these two feature maps, a convolutional fusion model in [9] is applied to automatically learn the fusion weights:

$$f_{conv} = f_{cat} * \mathbf{f} + b, \tag{7}$$

where $\mathbf{f}$ are filters with dimensions $1 \times 1 \times 2D \times D$, and $f_{cat}$ denotes the stack of two feature maps at the same spatial locations $(i, j)$ across the feature channels $d$:

$$\begin{aligned} f_{cat}^{(i,j,2d)} &= f_{Inf}^{(i,j,d)}, \\ f_{cat}^{(i,j,2d-1)} &= f_{g}^{(i,j,d)}, \end{aligned} \tag{8}$$

where $f_g$ denotes the generated fake feature map $G(f_{Inf}, z)$.

Thus, the predictive loss $L_p$ can be formulated as:

$$L_p = -\log l \cdot D_p(f_{Inf}, G(f_{Inf}, z)). \tag{9}$$

The final discriminative loss $L_D$ can be defined as the weighted sum of adversarial loss and predictive loss:

$$L_D = w_1 \cdot L_a + w_2 \cdot L_p. \tag{10}$$

In the training process, the transferable net and the full-modal discriminative net are alternatively trained until the generated feature of missing channel becomes close to real and the discriminative net achieves precise recognition. In the testing process, we only need to send one/part of the data modality into the PM-GANs framework, and the generative net will automatically generate a transferred feature representation for the missing modality, and the predictor of discriminative net constructs a full-modal representation and predicts the label.

## 4 Experiments

In this section, we first introduce our new dataset for partial-modality infrared action recognition. In detail, the specifications and a complete evaluation of the dataset will be elaborated. For the experiment part, we introduce the configurations of the experiments and show the results and analyses corresponding to our method. Specifically our experiments are in three folds. First, we assess the effectiveness of the transferable net by comparing the generated feature representations with the real ones. Second, we evaluate the discriminative net ability using partial data modality. Finally, we compare our approach with four state-of-the-art methods to verify the effectiveness of the PM-GANs.

### 4.1 Cross-Modal Infrared-Visible Dataset for Action Recognition

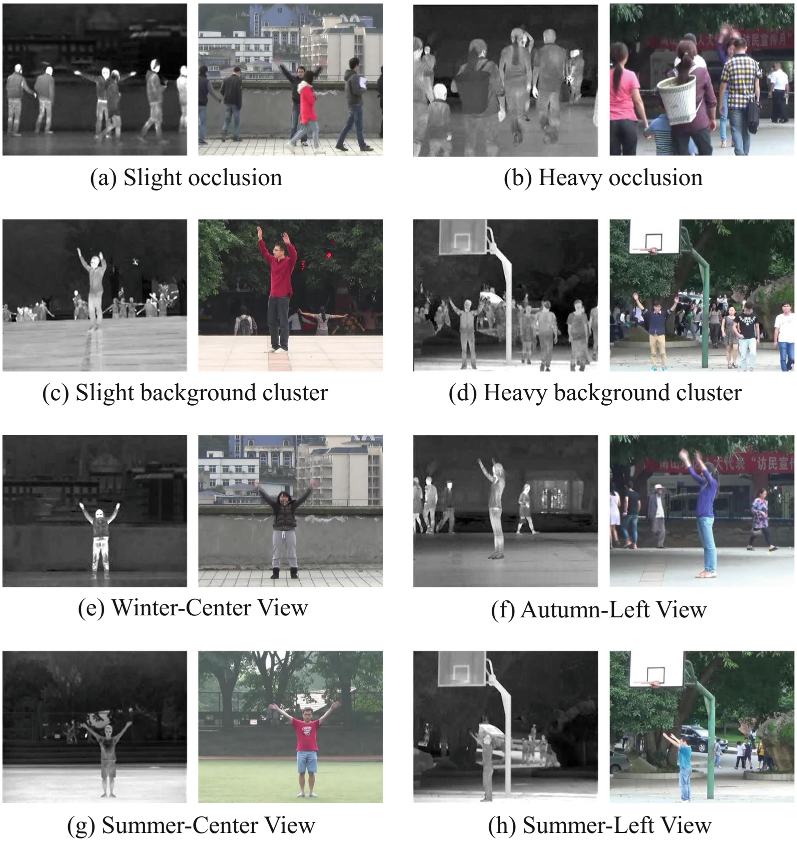
We introduce a new action recognition dataset, which is constructed by paired videos of both RGB and infrared data channels. Each action class contains a singular action type, and each video sample contains one action class. In total there are 12 classes of both individual actions and person-person interactive actions. For individual actions: one hand wave (wave1), multiple hands wave (wave2), handclap, walk, jog, jump, skip, and interactive actions: handshake, hug, push, punch and fight. For each action class, there are 100 paired videos, with a frame rate of 25 fps. The frame resolutions are 256 × 293 for infrared channel and 480 × 720 for RGB channel. Each action is performed by around 50 different volunteers. A sample of the frames is illustrated in Fig. 3. The duration of videos varies from several seconds to more than 10 s.

In order to simulate the real-world variations, four scenario variables are considered: the background complexity, season, occlusion, and viewpoint.

**Background Complexity:** In our newly-introduced dataset, the background varies from relatively simple scenes (plain background) to complex ones (with moving objects). For simple background, there are only one or two people performing actions, as shown in Fig. 3(c). While for complex background, interrupting pedestrian activities concur with the objective action in different degrees, as shown in Fig. 3(d).

**Season:** The infrared channel is heavily effected by the seasons, because it reflects the heat radiation of objects. In winter, when ambient temperature is in a low value, the imaging of human body is salient and clear. However, in summer, the contrast between human and background is ambiguous. Thus, we divide the seasons into three categories: winter, spring/autumn, summer, as shown in Fig. 3(e)–(h). The video number proportions of these three seasons are 30%, 50%, and 20%, respectively. All actions were performed in these three seasons.

**Occlusion:** Specific videos with occlusions from 0% to over 50% are arranged in each action class to promote the diversity and complexity of dataset, as shown in Fig. 3(a)–(b).


(a) Slight occlusion

(b) Heavy occlusion

(c) Slight background cluster

(d) Heavy background cluster

(e) Winter-Center View

(f) Autumn-Left View

(g) Summer-Center View

(h) Summer-Left View

**Fig. 3.** Example paired frames for the action "wave2" in the newly introduced multi-modal dataset for action recognition. The left ones are in infrared channel and the right ones are in RGB channel

**Viewpoint:** The variation of different viewpoints is also an important factor considered. The video clips under the front-view, left-side-view, right-side-view are all included in the dataset, as shown in Fig. 3(e)–(h).

We split 75% of the paired video clip couples as the training set, and the rest as the testing set. To investigate the suitable representations for each spectrum and the most complimentary representation couples, we select several effective representations to test their discriminative ability on RGB channel, infrared channel, and the combined channels.

We feed the original video clips, the MHI image clips [1] and the optical flow clips [29], denoted as "Org", "MHI", and "Optical Flow", into the 3D-CNN [49] to obtain spatiotemporal features. The 3D-CNN takes a 16-frame clip as inputs and performs 3D convolution and 3D pooling, which calculates the appearance and motion information simultaneously. Specifically, we extract the output of the last fully connected layer and conduct a max pooling to all clip features of one video. In the case of two-modality fusion, we directly concatenate the features of infrared channel and RGB channel. After that, a linear SVM classifier is trained to obtain the final recognition results. The 3D-CNN is fine-tuned by the corresponding maps of our training set.

**Table 1.** The evaluation results of different features on different channels and their fusion on the proposed dataset

| Method | Descriptor | Accuracy (%) |
|---|---|---|
| Infrared Channel | Org | 55% |
| | Optical flow | 69.67% |
| | MHI | 61% |
| RGB Channel | Org | 49% |
| | Optical flow | 78.66% |
| | MHI | 65.33% |
| Fusion | Org | 55.33% |
| | Optical flow | 80.67% |
| | MHI | 68.67% |

As shown in Table 1, the performances of different representations for both infrared and RGB channels and their combined results are listed. It is clearly observed that for both infrared and RGB channels, the 3D-CNN features after optical flows achieve the best performance. In two modalities fusion, the 3D-CNN features after optical flows in RGB channel can effectively boost the performance of using the infrared channel only. Thus, in the following experiments of transferable nets and discriminative nets [50], the optical flows are selected as input for representation learning via PM-GANs.

### 4.2 Implementation Details

For the input data, we compute optical flows using the toolbox of Liu [34]. The 3D ConvNet in transferable generative net is fine-tuned on the infrared optical flows of training set. And the adversarial visible feature maps are extracted from a 3D ConvNet fine-tuned on the visible optical flows of training set. The sampled numbers of clips $T$ is set as 5, and each clip has a duration of 16 frames. The loss weights $w_1$ and $w_2$ are set as 0.1 and 0.9 respectively. We set the initial learning rate at $2 \times 10^{-5}$. The whole network is trained with ADAM optimization algorithm [25] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, batch size of 30 on a single NVIDIA GeForce GTX TITAN X GPU with 12 GB memory. The framework is implemented using TensorFlow library and accelerated by CUDA 8.0.

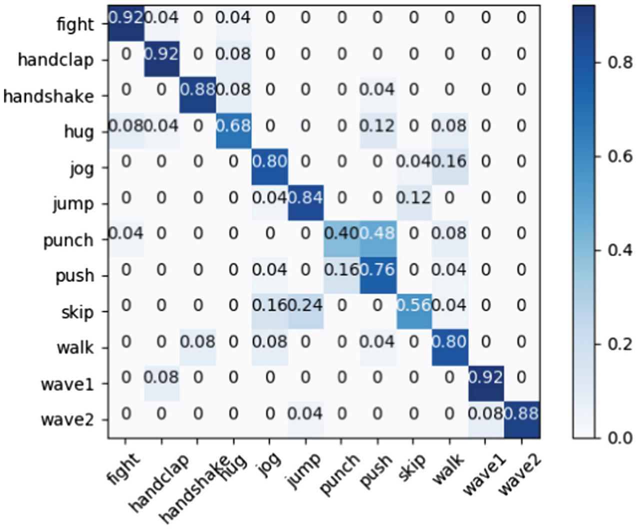
### 4.3 Transferable Net Evaluations

The PM-GANs model is evaluated on the proposed action recognition dataset. We present the results of five different modalities as shown in Table 2. For single modality, we utilize the 3D ConvNet part and the predictor part without fusion model for training and testing. And for the case of real infrared and RGB channel fusion, we directly input the real feature map of RGB channel to the fusion model instead of using generated ones. From Table 2, we can observe the generated RGB representations perform better than the original infrared ones, which shows that the PM-GANs have indeed discovered useful information through modality transfer. Moreover, the fusion of infrared and generated RGB representations achieves an Accuracy of 78%. Although it performs worse than the original RGB channels and the fusion of infrared and RGB channels, it only utilizes the information of infrared channel in the testing process.

**Table 2.** Evaluation results on the discriminative ability of transferable modality

| Data modalities | Accuracy (%) |
|---|---|
| Infrared channel | 71.67% |
| RGB channel | 79.33% |
| Generated RGB | 76.67% |
| Infrared + RGB channels | 82.33% |
| Infrared channel + Generated RGB | 78% |

In order to analyze the intra-class performance, the confusion matrices are drawn in Fig. 4. As observed, the proposed method generally shows good performance in action classification: in most classes, the testing samples are assigned the correct label. However, we notice that the "punch", "skip" action samples are likely to be classified as "push" and "jump" respectively. One likely reason



| | fight | handclap | handshake | hug | jog | jump | punch | push | skip | walk | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fight | 0.92 | 0.04 | 0 | 0.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| handclap | 0 | 0.92 | 0 | 0.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| handshake | 0 | 0 | 0.88 | 0.08 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0 |
| hug | 0.08 | 0.04 | 0 | 0.68 | 0 | 0 | 0 | 0.12 | 0 | 0.08 | 0 | 0 |
| jog | 0 | 0 | 0 | 0 | 0.80 | 0 | 0 | 0 | 0.04 | 0.16 | 0 | 0 |
| jump | 0 | 0 | 0 | 0 | 0.04 | 0.84 | 0 | 0 | 0.12 | 0 | 0 | 0 |
| punch | 0.04 | 0 | 0 | 0 | 0 | 0 | 0.40 | 0.48 | 0 | 0.08 | 0 | 0 |
| push | 0 | 0 | 0 | 0 | 0.04 | 0 | 0.16 | 0.76 | 0 | 0.04 | 0 | 0 |
| skip | 0 | 0 | 0 | 0 | 0.16 | 0.24 | 0 | 0 | 0.56 | 0.04 | 0 | 0 |
| walk | 0 | 0 | 0.08 | 0 | 0.08 | 0 | 0 | 0.04 | 0 | 0.80 | 0 | 0 |
| wave1 | 0 | 0.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.92 | 0 |
| wave2 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0 | 0.08 | 0.88 |

**Fig. 4.** The results illustrated in confusion matrices using the proposed method

is that two sets of actions are similar in both movement and process, sometimes even hard to distinguish for human eyes.

To get insight into how effective the transferable ability of PM-GANs are, we rearrange the training and testing splits. Specifically, we utilize the scenes of Spring/Autumn and Summer for training, and Winter for testing. We use this split to examine the generalization ability of the proposed model. As can be seen in Table 3, the generated fake RGB representations outperform the original infrared ones, which shows the robust transferability of PM-GANs.

**Table 3.** Evaluation of the models generalization ability using a separate dataset

| Modalities of a separate dataset | Accuracy (%) |
|---|---|
| Infrared channel | 74.17% |
| RGB channel | 79.44% |
| Generated RGB | 77.78% |
| Infrared + RGB channels | 82.78% |
| Infrared channel + Generated RGB | 80.28% |

### 4.4 Comparisons with Other Methods

To evaluate the effectiveness of PM-GANs, we compare our method with four state-of-the-art methods, including the most effective handcrafted features iDT

[54], and the state-of-the-art deep architecture [49]. In addition, we also compare our methods with two state-of-the-art framework for infrared action recognition [13, 22]. For iDT features, Fisher Vector [40] is applied to encode and then a linear SVM classifier [45] is trained for action classification. As for the C3D architecture, the network is fine-tuned by the proposed training dataset. Then max pooling followed by a SVM classifier is applied as the evaluation in Table 1. For [13], we follow the original experimental settings provided by the author. For [22], we implement and select the configuration with the optimal results based on the original submission. We apply the discriminative code layer and the second fusion strategy for feature extraction, and train a K-nearest neighbor classifier (KNN) [4] using the provided Gaussian kernel function for classification. Note that all of the results are achieved using unified optical flows as the input.

**Table 4.** Comparisons with four state-of-the-art approaches

| Method | Accuracy (%) |
|---|---|
| iDT [54] | 72.33% |
| C3D [49] | 69.67% |
| Two-stream CNN [13] | 68% |
| Two-stream 3D-CNN [22] | 74.67% |
| PM-GANs | 78% |

Table 4 presents the accuracy of the competing approaches. As observed, the handcrafted iDT method achieves comparable results with some high-level architecture. Methods using 3D-CNN outperform the method with 2D-CNN architecture. One reason to explain is that the 3D-CNN architecture is better in modeling temporal variations. The two-stream 3D-CNNs outperform the conventional iDT framework and robust C3D models, showing effective strength of the proposed discriminative code layer. Our proposed PM-GANs achieve the highest accuracy, which shows the effectiveness of the transferred feature representation and the robustness of our constructed model using only part of the data modalities.

## 5 Conclusions

In this paper, we proposed a novel Partial-modal Generative Adversarial Networks to construct a discriminative full-modal representation with only part of the data modalities being available. Our method learns the transferable representation among heterogeneous data modalities using adversarial learning, and build a discriminative net that represents all modalities. Our method is evaluated in the task of action recognition and outperforms four state-of-the-art methods on the newly-introduced dataset. The dataset, which contains paired videos in both infrared and visible spectrum, will be made as the first publicly available visible-infrared dataset for action recognition.

**Acknowledgment.** This work is supported by the National Natural Science Foundation of China (No. 61571071, 61661166011, 61721002), the Natural Science Foundation of Chongqing Science and Technology Commission (No. cstc2018jcyjAX0227) and the Research Innovation Program for Postgraduate of Chongqing (No. CYS17222).

## References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
2. Bouwmans, T., Zahzah, E.H.: Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance. Comput. Vis. Image Underst. **122**, 22–34 (2014)
3. Bronstein, M.M., Bronstein, A.M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3594–3601. IEEE (2010)
4. Bui, D.T., Nguyen, Q.P., Hoang, N.D., Klempe, H.: A novel fuzzy k-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. Landslides **14**(1), 1–17 (2017)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE (2017)
6. Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2940–2949. IEEE (2016)
7. Delhumeau, J., Gosselin, P.H., Jégou, H., Pérez, P.: Revisiting the VLAD image representation. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 653–656. ACM (2013)
8. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1486–1494. MIT Press (2015)
9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1933–1941. IEEE (2016)
10. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 7–16. ACM (2014)
11. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T.: Rank pooling for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 773–787 (2017)
12. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning (ICML), pp. 1180–1189. ACM (2015)
13. Gao, C., et al.: Infar dataset: infrared action recognition at different times. Neurocomputing **212**, 36–47 (2016)
14. van Gemert, J.C., Jain, M., Gati, E., Snoek, C.G., et al.: Apt: Action localization proposals from dense trajectories. In: British Machine Vision Conference (BMVC), pp. 177.1–177.12. British Machine Vision Association (2015)

15. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680. MIT Press (2014)
16. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2827–2836. IEEE (2016)
17. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshops), p. 17. IEEE (2005)
18. Han, J., Bhanu, B.: Fusion of color and infrared video for moving human detection. Pattern Recogn. **40**(6), 1771–1784 (2007)
19. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)
21. Ji, S., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
22. Jiang, Z., Rozgic, V., Adali, S.: Learning spatiotemporal features for infrared action recognition with 3D convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshops). IEEE (2017)
23. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimed. **17**(3), 370–381 (2015)
24. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725–1732. IEEE (2014)
25. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference (BMVC), pp. 275-1–10. British Machine Vision Association (2008)
27. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: International Koint Conference on Artificial Intelligence (IJCAI), pp. 1360–1365. AAAI Press (2011)
28. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2–3), 107–123 (2005)
29. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2008)
30. Li, T., Mei, T., Kweon, I.S., Hua, X.S.: Contextual bag-of-words for visual categorization. IEEE Trans. Circ. Syst. Video Technol. **21**(4), 381–392 (2011)
31. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.: Videolstm convolves, attends and flows for action recognition. Comput. Vis. Image Underst. **166**, 41–50 (2018)
32. Lindtner, S., Hertz, G.D., Dourish, P.: Emerging sites of HCI innovation: hackerspaces, hardware startups & incubators. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 439–448. ACM (2014)

33. Liu, A.A., Xu, N., Nie, W.Z., Su, Y.T., Wong, Y., Kankanhalli, M.: Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans. Cybern. **47**(7), 1781–1794 (2017)
34. Liu, C., et al.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009)
35. Liu, Y., Lu, Z., Li, J., Yao, C., Deng, Y.: Transferable feature representation for visible-to-infrared cross-dataset human action recognition. Complexity **2018**, Article ID 5345241, 20 p. (2018)
36. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML), pp. 97–105. ACM (2015)
37. Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: a survey of recent advances. IEEE Sig. Process. Mag. **32**(3), 53–69 (2015)
38. Peng, Y., Qi, J., Yuan, Y.: CM-GANs: cross-modal generative adversarial networks for common representation learning. arXiv preprint arXiv:1710.05106 (2017)
39. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 521–535 (2014)
40. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
41. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
42. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 667–681 (2018)
43. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015)
44. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
45. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: 2004 International Conference on Pattern Recognition, (ICPR), vol. 3, pp. 32–36. IEEE (2004)
46. Shao, L., Zhu, F., Li, X.: Transfer learning for visual categorization: a survey. IEEE Trans. Neural Netw. Learn. Syst. **26**(5), 1019–1034 (2015)
47. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
48. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NIPS), pp. 568–576. MIT Press (2014)
49. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497. IEEE (2015)
50. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2962–2971. IEEE (2017)
51. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1510–1517 (2018)

52. Veeriah, V., Zhuang, N., Qi, G.J.: Differential recurrent neural networks for action recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4041–4049. IEEE (2015)
53. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
54. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. Int. J. Comput. Vis. **119**(3), 219–238 (2016)
55. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4305–4314. IEEE (2015)
56. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
57. Wei, Y., et al.: Cross-modal retrieval with CNN visual features: a new baseline. IEEE Trans. Cybern. **47**(2), 449–460 (2017)
58. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5380–5389. IEEE (2017)
59. Yang, L., Gao, C., Meng, D., Jiang, L.: A novel group-sparsity-optimization-based feature selection model for complex interaction recognition. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9007, pp. 508–521. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16814-2_33
60. Yang, Y., Zha, Z.J., Gao, Y., Zhu, X., Chua, T.S.: Exploiting web images for semantic video indexing via robust sample-specific loss. IEEE Trans. Multimed. **16**(6), 1677–1689 (2014)
61. Yeh, Y.R., Huang, C.H., Wang, Y.C.F.: Heterogeneous domain adaptation and classification by exploiting the correlation subspace. IEEE Trans. Image Process. **23**(5), 2009–2018 (2014)
62. Zollhöfer, M., et al.: Real-time non-rigid reconstruction using an RGB-D camera. ACM Trans. Graph. (TOG) **33**(4), 156 (2014)