Hindawi
Advances in Multimedia
Volume 2024, Article ID 4187991, 12 pages
https://doi.org/10.1155/2024/4187991

Research Article

Video Abnormal Action Recognition Based on Multimodal Heterogeneous Transfer Learning

Hong-Bo Huang[1,2] **Yao-Lin Zheng**[1] **and Zhi-Ying Hu**[1]

[1]*Computer School, Beijing Information Science and Technology University, Beijing 100101, China*
[2]*Institute of Computing Intelligence, Beijing Information Science and Technology University, Beijing 100192, China*

Correspondence should be addressed to Hong-Bo Huang; hhb@bistu.edu.cn

Received 3 January 2023; Revised 29 August 2023; Accepted 5 January 2024; Published 19 January 2024

Academic Editor: Yu-Chen Hu

Human abnormal action recognition is crucial for video understanding and intelligent surveillance. However, the scarcity of labeled data for abnormal human actions often hinders the development of high-performance models. Inspired by the multimodal approach, this paper proposes a novel approach that leverages text descriptions associated with abnormal human action videos. Our method exploits the correlation between the text domain and the video domain in the semantic feature space and introduces a multimodal heterogeneous transfer learning framework from the text domain to the video domain. The text of the videos is used for feature encoding and knowledge extraction, and knowledge transfer and sharing are realized in the feature space, which is used to assist in the training of the abnormal action recognition model. The proposed method reduces the reliance on labeled video data, improves the performance of the abnormal human action recognition algorithm, and outperforms the popular video-based models, particularly in scenarios with sparse data. Moreover, our framework contributes to the advancement of automatic video analysis and abnormal action recognition, providing insights for the application of multimodal methods in a broader context.

1. Introduction

The analysis of abnormal human actions is a critical task in many video surveillance applications. It involves detecting and classifying human body movements or crowd behavior, which can provide warning information to prevent or minimize the occurrence of injuries. Human action recognition has been a popular research topic in computer vision and artificial intelligence for many years [1].

In recent times, the use of deep learning driven by big data and high-performance parallel computing has gained significant attention in the computer vision. This approach has achieved remarkable results in various tasks, such as image classification, target detection, and image segmentation, with notable improvements in accuracy and speed [2–6]. Deep-learning-based methods have become the dominant approach in the field of human action recognition [7, 8].

Despite the success of deep-learning-based methods in various scenarios, they encounter certain challenges. These methods heavily rely on the models with a large number of parameters. According to the fundamental theory of machine learning, the complexity of a model directly affects the amount of labeled data required to train it. In the case of abnormal human action recognition, these actions occur less frequently in the real situations, making it challenging to collect video data. Moreover, cleaning and labeling these data are a laborious and expensive process. Although, various sample enhancement methods have been proposed to increase the size of the training data, the growth is still minimal compared to the vast sample space. Consequently, training complex deep-learning models with sparse samples remain a significant challenge.

Recently, a lot of work has been devoted to multimodal, cross-domain research [9]. Many studies of text and image intergeneration have also demonstrated that image data and text data expressing image-related information share common high-level semantic features [10, 11]. Inspired by these works, in this paper, we use the text data describing human abnormal action videos as the source domain data and the video data as the target domain, and propose a heterogeneous transfer learning framework for multimodal transfer learning.

In order to implement our method, we need some relevant video and text materials. We find that a considerable amount of video recordings depicting abnormal human actions, including news footage, user-generated content, and other forms of visual media, are often accompanied by textual annotations, such as news releases, comments, subtitles, and other forms of textual information. Based on the highly interdependent nature of text-based and video-based data, we have constructed a text-video dataset for recognizing abnormal human actions, named the abnormal action dataset (AAD). The dataset comprises 181 videos and 1,160 textual annotations, classified into eight distinct categories of actions.

In addition, using video data for human abnormal action recognition has some weaknesses, including the higher storage space requirement and computational complexity compared to joint sequences, as well as increased sensitivity to noise and occlusion. In view of these reasons, a sequential human action normalized descriptor (SHAND) is proposed in this paper to replace the video as the input of the model. The SHAND consists of temporal information of multiple human keypoints and has the same representation of human action in different camera views. Therefore, it is invariant to changing views and can accurately represent the changing pattern of human action.

In summary, for this paper, the main contributions are as follows:

(i) We propose a multimodal learning framework for heterogeneous transfer learning that enables the model to better learn features from both text and data domains for human abnormal action recognition.

(ii) We propose a SHAND, which offers a simpler and more robust representation, greatly accelerating the training and inference speed of the model.

(iii) We build a human abnormal action recognition dataset named AAD, which has a more realistic sample distribution and contains samples from both text and video domains for multimodal learning.

2. Related Work

Recently, multimodal learning has attracted extensive attention from the researchers. Among all the multimodal learning methods, transfer learning has become an important technique and idea due to its powerful transfer ability and remarkable enhancement effect.

Transfer learning can be categorized into four types based on the learning styles: instance-based transfer learning, feature representation-based transfer learning, model-based transfer learning, and knowledge-based transfer learning [12]. The most typical traditional domain adaptation algorithm is transfer component analysis (TCA) [13]. TCA aims to achieve feature transfer by minimizing the distance between source and target domain data distribution after feature mapping.

Many works have been improved based on TCA, such as Xu et al. [14] and Li et al. [15]. Yosinski et al. [16] pioneered the study of the method based on deep neural networks. Tzeng et al. [17] proposed deep domain confusion (DDC) to improve deep network adaptation by using the maximum mean discrepancy (MMD) criterion to measure the gap between two distributions and adding adaptive metric loss [18]. Long et al. [19] then improved DDC and proposed the deep adaptation network (DAN) network structure. They used a multikernel MMD metric (MK-MMD) instead of a single-kernel MMD to calculate the distance between the source and target domain feature spaces, achieving better results on several tasks [20].

After Goodfellow proposed generative adversarial networks, Ganin et al. [21] conducted research on transfer learning using adversarial networks and proposed domain-adversarial neural network (DANN). The learning goal of the DANN network is to generate features that are indistinguishable between the two domains as much as possible. Later, Bousmalis et al. [22] extended DANN and proposed the domain separation networks (DSN) architecture. DSN considers that both the source and target domains consist of a public part and a private part. The public part can learn the features of the public, and the private part is used to keep the independent features of each domain. Finally, multiple losses in the network are combined simultaneously as the final loss. In addition, the agile domain adaptation networks (ADANs) proposed by Chen et al. [23] and the method using Wasserstein GAN proposed by Shen et al. [24] have also achieved better results on DAN transfer algorithms.

When the source and target domain data are distributed in the different feature spaces, researchers have proposed the heterogeneous domain adaptation (HDA) scheme to build a bridge between two heterogeneous domains [25, 26]. Current HDA methods usually choose to project one distribution onto the other, such as Chen et al. [23] and [27, 28], or to find a common domain-invariant subspace for both domains, such as Hsieh et al. [29], Xiao and Guo [30], and Yao et al. [31]. Chen et al. [32] proposed the transfer neural trees (TNT) method, in which the random pruning method Transfer neural decision forest (Transfer-NDF) was used as the final prediction layer of the network, achieving promising results. Yao et al. [33] proposed an end-to-end joint learning algorithm soft transfer network (STN) for domain sharing classifier and domain invariant subspace, which achieved better performance for the first time using the scheme of giving soft labels to unlabeled data. Liu et al. [34] proposed an infrared human motion recognition framework using visible light assisted data to solve the problem of limited infrared motion data. Liu et al. [35] used joint sparse representation and distribution adaptation to hierarchically learn view invariant representation, achieving feature representation transfer across views. Deep image-to-video adaptation and fusion networks (DIVAFNs) [36] were also proposed by them, using video keyframes as a bridge to enhance action recognition in videos by transferring knowledge from images. Semantics-aware adaptive knowledge distillation networks (SAKDNs) [37] were proposed to use wearable sensors as the teacher mode and RGB videos as the student mode, the action recognition in the visual sensor mode is enhanced by adaptively transferring and distilling

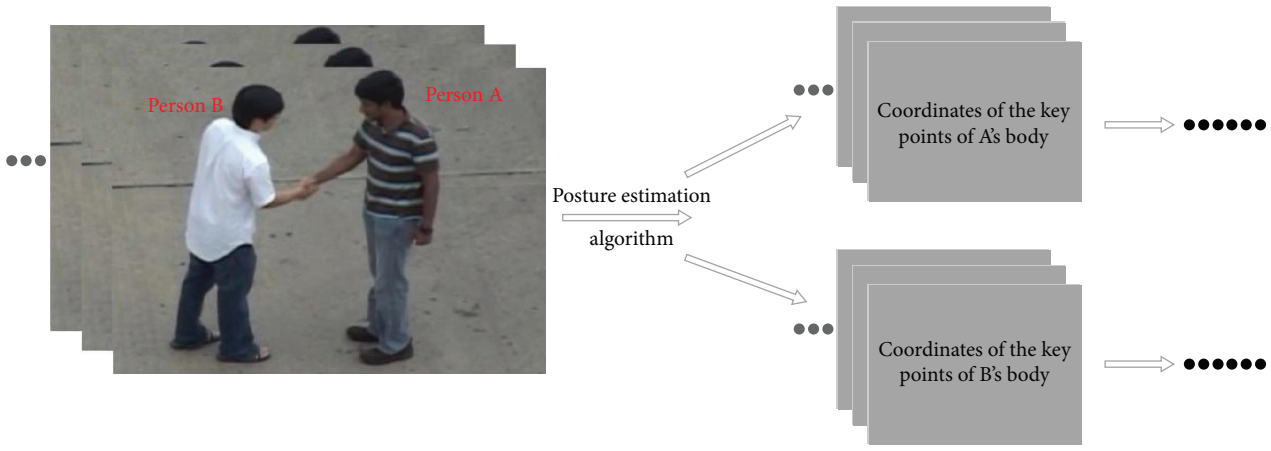

Figure 1: Diagram of double/multiperson action recognition.

knowledge from multiple wearable sensors. Liu et al. [38] proposed temporal contrastive graph for self-supervised video representation learning (TCGL), utilizing multiscale time dependencies in videos, a hybrid graph is used to compare learning strategies and jointly model time dependencies within and between segments for learning.

At present, domain adaptation methods have achieved promising effects on several experimental datasets, but the performance of the model will be generally poor when faced with more realistic scenarios and sparse sample distribution. When the source and target domains are only related in semantic space, designing a better feature mapping network to make the source domain better assist in the training is still a challenging task for heterogeneous transfer learning.

3. Human Abnormal Action Recognition Based on Multimodal Heterogeneous Transfer Learning

Video as a type of frame sequence has large amounts of data. It occupies excessive memory and storage space and has particularly high-information redundancy, resulting in significant computational overhead. Extracting representative and effective information can be helpful to solve this problem. Based on this consideration, in this paper, we use human keypoint estimation to describe action features, propose a descriptor named SHAND, and perform multimodal heterogeneous transfer learning based on the SHAND.

This section presents the categories and scope of involved abnormal actions, the construction method of the SHAND, the design method of the model, and the loss function for abnormal action recognition based on the multimodal heterogeneous transfer learning.

3.1. Categories and Scope of Involved Abnormal Actions. To facilitate the study, this paper focuses on eight classes of abnormal human movements in the common situations: fighting, falling, lying down, waving, shaking hands, walking, running, and hugging. It should be noted that while shaking hands and hugging may not be considered abnormal actions in daily life, they can hold special significance in different scenarios. As a result, they have been included in our recognition analysis.

Among these action categories, fighting, handshaking, and hugging are dual/multiplayer interaction actions, while the remaining five classes are single-person execution actions. Each video frame in the dual/multiplayer interaction action contains two or more target characters. Multiple target characters as a whole for input are not easy to extract to the interrelationship between targets. Therefore, we adopt a multistage discrimination method approach, i.e., when discriminating the actions of multitarget interactions, each target action is separately captured and extracted, and the interactions between the targets are then matched.

Viewing as a whole, if Person A and Person B in a video frame perform the same kind of action, their human keypoint coordinates will be simultaneously obtained by the posture estimation algorithm. These coordinates will be separately input to the abnormal action discriminator for recognition operation, as shown in Figure 1.

3.2. Constructing Sequential Human Action Normalized Descriptor. Human action in the video has two types of information: the position of the human joint points in the spatial domain and the mutual variation relationship of each joint point in the temporal domain. The former can be generally extracted by human pose estimation, while the latter is mostly represented by the temporal sequence of the regressed joint point coordinates.

To efficiently and accurately represent the change of the human joints in the video, we propose a SHAND, which consists of multiple human keypoint temporal information and could uniquely determine the relative position of each joint. By normalization, it has the same representation of human action under different camera views, thus it is invariant in changing views, and can accurately represent the changing pattern of human action, which has better representation capability theoretically.

Considering the higher accuracy and robustness of the OpenPose method in the human pose estimation algorithm, we use the OpenPose method in this paper for extracting the human keypoint coordinates, as shown in Figure 2 [39].

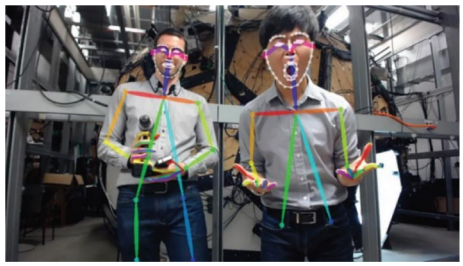

Figure 2: OpenPose human keypoints detection [39].

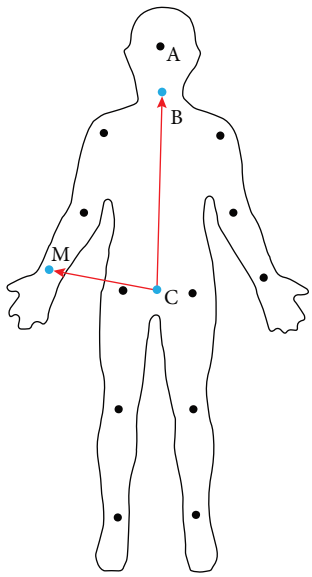

Figure 3: Construction of SHAND.

Total of 25 human keypoint coordinates are obtained by the OpenPose human pose estimation algorithm, and 15 of them are selected for constructing SHAND, which are: the nose (Point A), the center point of the neck (Point B), the midpoint of the left and right hip joints (Point C), the left and right shoulder joints, the left and right elbow joints, the left and right wrist joints, the left and right hip joints, the left and right knee joints, and the left and right ankle joints in the video, as shown in Figure 3.

When the coordinates of each keypoint position are obtained, the normalization of the relative position of each keypoint is performed with a reference vector. Since, the upper torso of the human body in general is not easily disturbed by the movements, the vector pointing from the midpoint of the left and right hip joints (Point C) to the neck point (Point B) is selected as the reference vector. The angle ($\leq 180°$) between each vector and the normalized module ratio is then obtained. Assuming that point M is the target point of the motion, then is the target vector. The normalized module ratio of the target vector is calculated by Equation (1), the cosine of the angle between the target vector and the reference vector is calculated by Equation (2), and the angle between the target vector and the reference vector can be calculated by Equation (3).

$$\ell\left|\overrightarrow{\mathrm{CM}}\right| = \frac{\sqrt{X_{\mathrm{CM}}{}^2 + Y_{\mathrm{CM}}{}^2}}{\sqrt{X_{\mathrm{CB}}{}^2 + Y_{\mathrm{CB}}{}^2}}, \tag{1}$$

$$\cos\theta_{<\overrightarrow{\mathrm{CM}},\overrightarrow{\mathrm{CB}}>} = \frac{X_{\mathrm{CB}} \cdot X_{\mathrm{CM}} + Y_{\mathrm{CB}} \cdot Y_{\mathrm{CM}}}{\sqrt{X_{\mathrm{CB}}{}^2 + Y_{\mathrm{CB}}{}^2} + \sqrt{X_{\mathrm{CM}}{}^2 + Y_{\mathrm{CM}}{}^2}}, \tag{2}$$

$$\theta_{<\overrightarrow{\mathrm{CM}},\overrightarrow{\mathrm{CB}}>} = \arccos\left(\cos\theta_{<\overrightarrow{\mathrm{CM}},\overrightarrow{\mathrm{CB}}>}\right), \tag{3}$$

where X and Y represent the horizontal and vertical coordinates of the keypoints, respectively. These normalized ratios of module length and the angles between the vectors will be used as the main features to construct the description vector.

The human action normalized descriptor is obtained by connecting point C with the other 13 keypoints, the pinch angle and normalized module are then respectively calculated by Equations (1) and (3). The size of the descriptor is 13×2 (without keeping any information). In the human abnormal action recognition task, we use n frames of the video to capture an action, and the operation is repeated for each frame to obtain a set of $n \times 13 \times 2$ sized temporal description vectors, i.e., the SHAND.

SHAND effectively captures changes in the human movements while filtering out variations due to different camera views. In addition, the size can be reduced by several orders of magnitude compared to raw video data due to the concise expression form that can effectively reduce redundant calculations. Although, it takes extra time to get the human keypoints, relying on the current efficient human pose estimation algorithm, our method could still significantly improve the speed of human action recognition, reducing the computing resource, and simplifying the recognition model.

3.3. Design of Multimodal Heterogeneous Transfer Learning Model. For video data that simultaneously contains text information, the text domain data and video domain data carry similar semantic information, and thus have common semantic features. By performing multimodal human abnormal action recognition with data from the two domains, the recognition performance and training stability can be improved, and the heavy dependence on labeled training data can be alleviated. In the light of this idea, we try to map the two data distributions to the same common feature space through a heterogeneous transfer learning method with the text domain as the source domain and the video domain as the target domain. Specifically, we obtain the human keypoints through the human pose estimation algorithm, calculate the SHAND, get the data input of the target domain, then process the text data using Word2Vec encoding (the specific process is described in detail in 4.1), and use the resulting word vector as the data input of the source domain.

The framework of the heterogeneous transfer learning is shown in Figure 4, where X_S represents the source domain data, i.e., the word vector extracted from the text domain

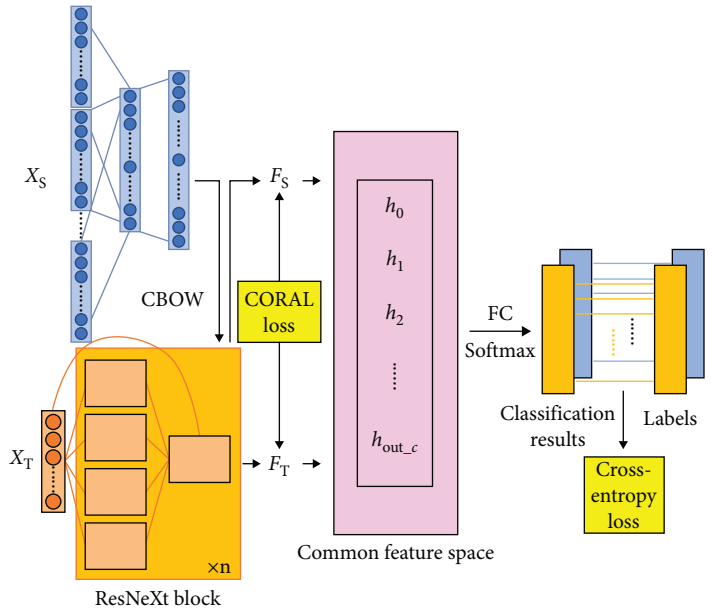

FIGURE 4: Framework of multimodal heterogeneous transfer learning.

data, and X_T represents the target domain data, i.e., the SHAND converted from the human action images. The two data distributions go through two feature mapping functions, namely F_S and F_T, and finally to the same common subspace.

The proposed framework extracts features from both the text and video domains. Specifically, it uses the Word2Vec model to extract word embeddings as input features from the text domain, and the ResNeXt−50 as a backbone network model to extract features from the video data and the word embeddings.

Initially, the network employs a convolutional layer with 64 output channels and a kernel size of 3×3 as the initial layer. Subsequently, the network constructs three residual blocks, each comprising eight branches. Each branch consists of two convolutional layers with 64 output channels. Within each residual block, a transition layer is employed, encompassing a convolutional layer with 64 output channels. In the final layer of each residual block, the outputs of the branches are combined with the residual connection and activated using the ReLU activation function, yielding the output of the residual block. This identical network architecture is applied to both the video data and word embeddings.

After the feature extraction and mapping to a common subspace process, we apply CORrelation ALignment (CORAL) as the loss function for computing the loss between the features of different domains. This loss function has been widely used in the domain adaptation tasks [40].

During the backpropagation process, we update the network parameters to simultaneously train the target and source domain data. Specifically, we use the gradient descent algorithm to minimize the loss function to improve the model's recognition accuracy. In updating the network parameters, we adopt learning rate decay and weight decay strategies to help the model converge faster.

In summary, our proposed framework uses the word2vec model to extract features from the text domain and use the ResNeXt−50 model as a backbone network to extract features from them. These features are mapped to a common subspace for comparing the feature distributions of different domains. The framework has excellent domain adaptation ability and can handle data distribution differences between the different domains.

3.4. Loss Function. In transfer learning, the loss function is a measure of the distance between the distribution of data in the source domain and that in the target domain, which is an important metric to guide transfer learning. Through designing and optimizing the loss function, the neural network could learn the features of the training data, which can be used for feature representation and classification. In this paper, the total loss function is shown as Equation (4).

$$\text{Loss} = K \times (\ell_{\text{CORAL}} + \ell_{\text{class\_T}}) + \ell_{\text{class\_S}} + \lambda \ell_2, \tag{4}$$

where $\ell_{\text{class\_S}}$ is the classification loss of the source domain data, $\ell_{\text{class\_T}}$ is the classification loss of the target domain data, ℓ_2 represents the L2 regularization of the weights as Equation (5), which can be considered as the penalty term of the loss function to avoid overfitting by limiting the values of parameters ω, and the space of the model. K is a hyperparameter to control the distance between the source and target domain distribution of the backpropagation process and the gradient size of the classification loss in the target domain. To make the gradient calculation primarily depend on the classification loss of the source domain data in the early stage of training, the initial value of K is set to a small value of 0.3. And ℓ_{CORAL} represents CORAL loss, proposed by Sun et al. [40], for aligning source and target domain data distribution. The equation is shown

as Equation (6) and the optimization objective is defined in Equation (7).

$$\ell_2 = \|\omega\|_2 = \sqrt{\sum_{i=1}^{n} (\omega_i)^2}, \tag{5}$$

$$\ell_{\mathrm{CORAL}} = \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{C}_{\mathrm{S}}\boldsymbol{A} - \boldsymbol{C}_{\mathrm{T}}\|_{\mathrm{F}}^2, \tag{6}$$

$$\min_{\boldsymbol{A}} \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{C}_{\mathrm{S}}\boldsymbol{A} - \boldsymbol{C}_{\mathrm{T}}\|_{\mathrm{F}}^2. \tag{7}$$

The ultimate goal of ℓ_{CORAL} is to find a second-order feature transformation matrix $\boldsymbol{A}$ that minimizes the distance between the source and target domain distribution, where $\boldsymbol{C}_{\mathrm{S}}$ and $\boldsymbol{C}_{\mathrm{T}}$, respectively, represent the covariance matrices of the source and target domains.

4. Experiment and Analysis

Human abnormal action data are often difficult to collect, hence fewer relative video datasets are publicly available. Therefore, by collecting and adding samples of video and text, we construct a video-text dataset named AAD. The keypoint information extracted from the abnormal action videos will be used to construct the SHAND vectors, and we will demonstrate the capability of better representation of the SHAND and its efficiency with experiments. In addition, we also conduct relevant experiments on the proposed multimodal heterogeneous transfer learning method to show the enhancement effect of the multimodal approach on abnormal action recognition.

4.1. Dataset and Preprocessing

4.1.1. Constructing the Dataset. Since few abnormal action video datasets are publicly available, in this paper, we collect some open human action datasets and add some new abnormal action videos captured by ourselves to build the target domain dataset. We name the dataset AAD. AAD contains eight classes of action, including "falling," "fighting," "lying down," "waving hand," "hugging" "running," "walking," and "shaking hands". Some samples are shown in Figure 5. The entire dataset consists of 181 videos, with video durations ranging from 1.5 to 7 s, and the frame rate is 30 fps.

To meet the demands of our multidomain transfer learning, corresponding textual descriptions are needed for each type of action. We manually add artificial textual descriptions to match each type of human abnormal action. Specifically, for each type of action, we add corresponding description sentences ranging from 71 to 226. In total, 1,160 sentences are constructed for the eight types of actions. By this means, we construct a video-related textual dataset suitable for heterogeneous transfer learning. In the experiments of this paper, 75% of the AAD is used as the training data and 25% as the test data.

4.1.2. Text Data Processing. The text data describing human abnormal actions constructed in this paper are processed by separation, alignment, and feature extraction, respectively. After the word separation operation, the Word2Vec model is used for word vector encoding. Word2Vec model has two types of training models: CBOW (Continuous Bag-of-Words) and Skip-Gram. The CBOW model is trained by predicting the central word using the context, while the Skip-Gram model uses the central word to predict the context. In this paper, the CBOW model is used to encode the word vector.

In the word vector training process, a corpus consisting of more than 8 -million Baidu online encyclopedia entries, more than 4 -million Sohu news items, and 229 GB of novels are used, which are collected from the Internet. The parameters of the pretraining process are set as follows: the word vector dimension is 128, the maximum distance of the word vector context is 5, and the words with occurrences below 10 are removed. The corresponding word vectors are then obtained by inputting the text data describing the human abnormal actions as the source domain data.

4.2. Experimental Settings. To fully verify the performance of the abnormal action recognition algorithm proposed, comprehensive experiments are conducted on AAD, including the following experimental steps:

(i) Extract the human keypoint coordinates by OpenPose human pose estimation algorithm.
(ii) Construct the SHAND obtained from the extracted keypoint coordinates.
(iii) Use the angle values of SHAND and normalized modulus length for action recognition.
(iv) Only use video frame images for input.
(v) Other popular video-based motion recognition methods are used to compare with the model proposed in this paper [41–45].
(vi) Perform abnormal action recognition based on multimodal heterogeneous transfer learning.

The video frame images are resized to 128×171 and cropped to 112×112 for matching the machine's computational power in the experiments. The network framework selected for all experiments is ResNeXt-50, except the other methods in the comparison experiment. The initial learning rate of training is set to 0.1 with exponential decay. The batch size is set to 20, the number of training epochs is set to 150, and the validation is performed once after each epoch. To ensure the fairness and reasonableness of the experiment, the network structure, parameters, and other underlying conditions will be guaranteed to be unchanged, and the training and testing sets used for the experimental process are the same.

The experiments are carried out under Linux Ubuntu 16.04 system. The programing language is Python 3.8 and the deep learning framework used is TensorFlow 2.4. The CPU is Intel i9-9900X and the GPU is NVIDIA RTX2080Ti 11 GB.

4.3. Experimental Results and Analysis

4.3.1. Experiments and Analysis of SHAND. All experiments are conducted on the AAD. The methods of human keypoint coordinates, angle vector, normalized ratios of module length, SHAND, and video frame images are trained and

Figure 5: Examples of human abnormal action dataset. (a) falling, (b) fighting, (c) lying down, (d) waving, (e) shaking hand, (f) running, (g) hugging, and (h) walking.

evaluated, respectively. The results are shown in Figure 6, where the horizontal coordinate represents the number of epochs and the vertical coordinate represents the human abnormal action recognition accuracy. The detailed accuracy of each method in the test is shown in Table 1.

Figure 6 shows that, except for the video-based method, the other methods achieve a fast convergence and high-training accuracy, attesting to the remarkable performance of the proposed algorithm's feature extraction and representation ability.

From the experiment, it can be concluded that the video-based method exhibits limited capability in learning the sample distribution, leading to weak feature representation and limited learning ability for human abnormal action recognition. In contrast, the SHAND-based methods have stronger representations and significant performance advantages. Using the coordinates of human keypoints and constructing SHAND can extract more distinctive features by deep convolutional neural networks, improving the human action recognition performance remarkably.

In terms of convergence speed, it is observed that both the method with SHAND as input and that with angle value or normalized ratios of module length as input have significantly faster convergence proceed than the original counterpart. This indicates that SHAND has a strong ability for expressing human action features, and thus enhances the convergence proceed noticeably.

We, respectively, use the angle values and the normalized ratios of module length in SHAND as input data for ablation experiments. As shown in Table 1, SHAND possesses excellent human action representation capability. The angle-based descriptors perform relatively better than the module-based descriptors. Intuitively, the ratios of module length are sensitive to factors such as pose, scale, and perspective, while the angle is relatively less sensitive to these factors and thus performs more robustly.

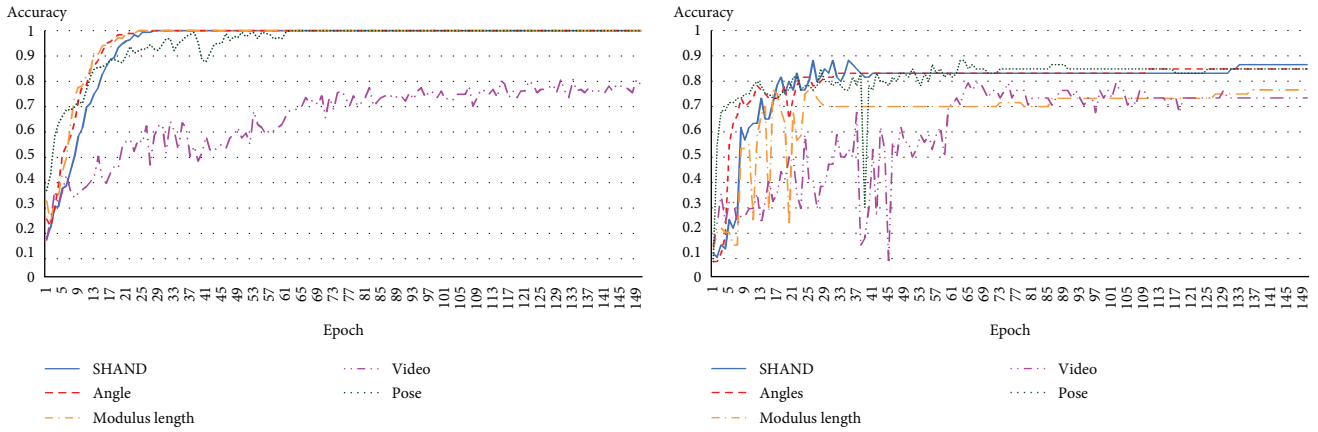

Figure 6: The training and the testing accuracy of the SHAND experiment. (a) Training accuracy variation curve of each method and (b) testing accuracy variation curve of each method.

Table 1: Test accuracy comparison of different methods.

Method	Testing accuracy (%)
Using the video frame images as input data	73.33
Using the human keypoints as input data	85.00
Using the normalized ratios of module length of SHAND as input data	76.67
Using the angles of SHAND as input data	85.00
Using the SHAND as input data	86.67
The multimodal heterogeneous transfer learning + SHAND	91.67

Table 2: Training time comparation of different methods.

Method	Training time (min)
Using the videos as input data	57.75
Using the human keypoints as input data	9.79
Using the module ratios of SHAND as input data	9.74
Using the angles of SHAND as input data	9.69
Using the SHAND as input data	9.81

The time taken to complete the model training for each of the above methods is shown in Table 2. It can be found that the time taken by all the methods is very close except for the video-based method. It takes about 5.9 times longer than the other methods since the input video frame image is reduced to 112×112, which is significantly larger than the SHAND vector with a size of only 2×13, indicating that the representation of the keypoints of the human body and its related features could replace the video frame image with a simpler and more direct form of the feature representation input, resulting in a faster action recognition task.

4.3.2. Experiments and Analysis of Multimodal Heterogeneous Transfer Learning. For verifying the enhancement effect of our method on abnormal action recognition, the text data domain constructed using the previously described method serves as the source domain, and the video domain serves as the target domain. Experiments are then conducted on the AAD using the multimodal heterogeneous transfer learning framework. The training and testing processes and results are, respectively, shown in Figure 7.

The training curves of the model demonstrate that with or without using heterogeneous transfer learning, the rising rate and training accuracy are higher than the video-based method when using SHAND as the input as shown in Figure 7 (a)). However, the slower rise rate of the transfer learning curve indicates that the convergence rate is slower when using text data as the source domain for knowledge transfer. This could be because using multiple networks and performing modal fusion could be more complicated than directly using one convolutional neural network.

The comparison results between our proposed method and several popular video-based models for human action recognition are presented in Figure 8 and Table 3, showing that our method outperforms them. Notably, we use text and SHAND as the mode of input data, which enables the model to train and reason much faster than other video-based models. By using text data to describe abnormal human action videos as source domain data and employing a multimodal heterogeneous transfer learning-based approach, the model's generalization ability is improved. This is likely due to the fact that the text and target domain data are mapped to the same common subspace, resulting in the presence of common features in the high-level semantic space that benefit the model's performance.

4.4. Visualization of the Examples. To gain a more comprehensive understanding of the strengths and weaknesses of the proposed method, we conducted visualizations of carefully selected common and error-prone samples using feature heat maps. Specifically, we transformed the preclassification feature maps and classification weights of our model into heat maps, which were subsequently combined with the samples to facilitate more detailed observations. Partial results are depicted in Figure 9. The failed samples are represented by the red boxes.

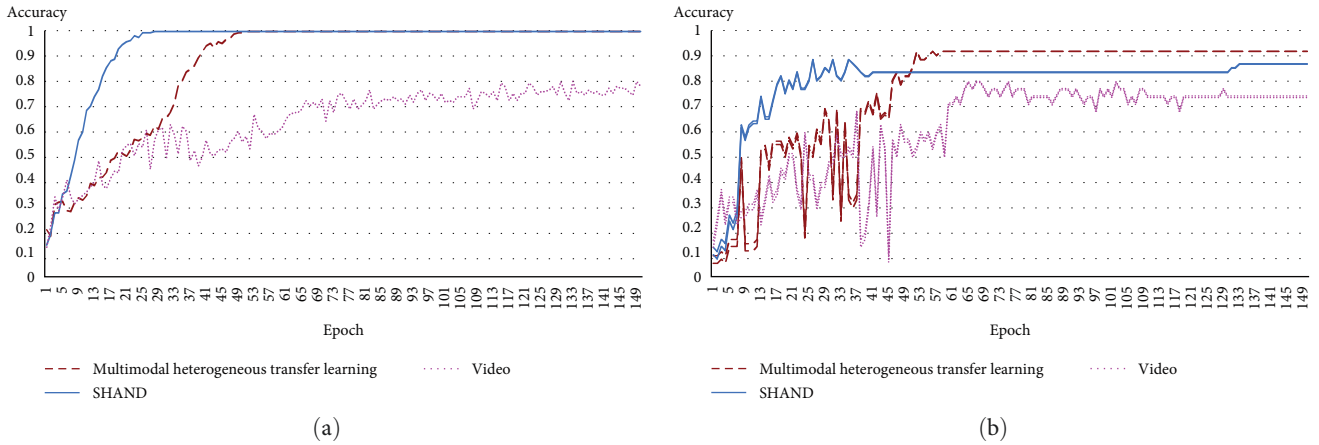

Figure 7: The training and the testing accuracy of the multimodal heterogeneous transfer learning experiment. (a) Training accuracy variation curves of the three methods and (b) testing accuracy variation curves of the three methods.

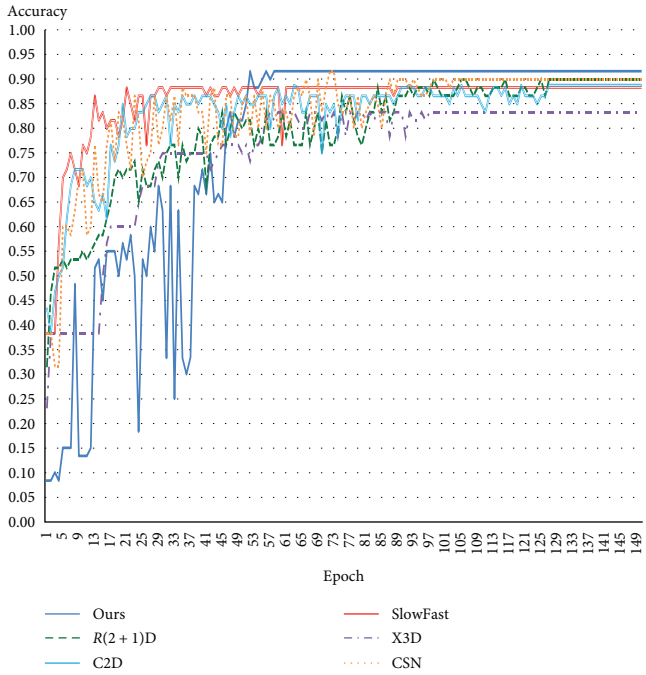

Figure 8: The testing accuracy of comparison with different popular video motion recognition models.

Table 3: Test accuracy comparation of multimodal and other methods.

Method	Parameters (M)	Testing accuracy (%)
$R(2+1)$D [41]	33.2	90.00
X3D [42]	3.8	83.33
SlowFast [43]	22.9	88.33
C2D [44]	25.6	88.33
CSN [45]	13.6	90.00
The multimodal heterogeneous transfer learning + SHAND (ours)	25.0	91.67

Remarkably, our model demonstrates a robust capability in accurately identifying several prevalent abnormal actions within the common category. Despite the presence of substantial background noise, the model effectively detects key regions of interest related to the subject in the video, enabling the determination of abnormality. Nevertheless, certain challenging samples, such as distinguishing between fighting and shaking hands, pose inherent difficulties and introduce a certain degree of error.

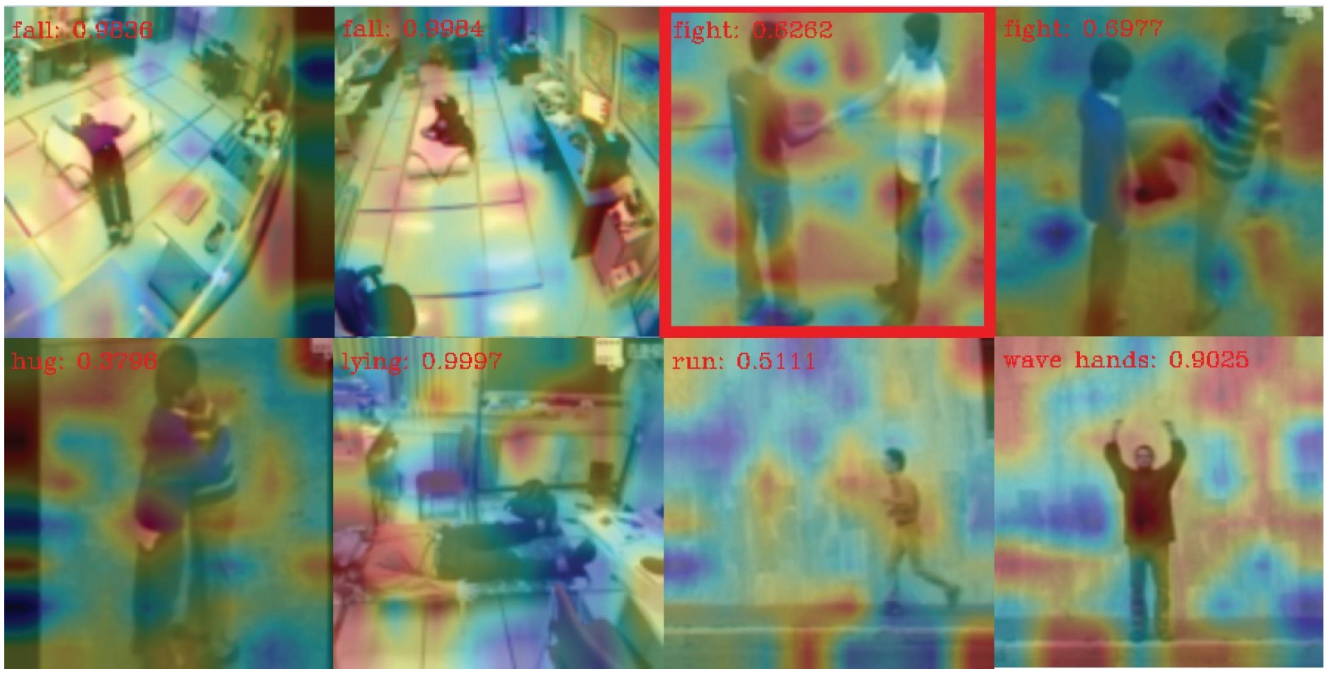

Figure 9: Feature heat maps of partial examples.

5. Conclusion

In this paper, we research the problem of video abnormal action recognition under sparse training data. The research is also an attempt to solve the difficult problems in multimodality and cross-domain, etc. We build a human abnormal action recognition dataset AAD, propose a SHAND with a simpler and more robust representation, and design a multimodal heterogeneous transfer learning framework, which maps the feature distribution in different domains to a common subspace and completes the knowledge transfer of common features. Our method makes the human abnormal action recognition model have better generalization performance and provides an idea for the practice and application of multimodal methods.

However, our framework still has much room for improvement, such as finding more suitable backbone networks for the migration learning and finding more suitable loss functions between the text and video modalities. Thus, future work can focus on improving the efficiency of multimodal heterogeneous transfer learning models as well as the generalization performance. In addition, more work exploring how features in the text and video domains map to a common subspace is urgently needed. Furthermore, we hope to investigate more effective ways to utilize more modal information, enabling human abnormal action recognition tasks to be applied more robustly.

Data Availability

The relevant abnormal action dataset and the model parameters data used to support the results and findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Beijing Municipal Education Committee Scientific and Technological Planning Project (KM201811232024), Beijing Information Science and Technology University Research Fund (2021XJJ30, 2021XJJ34), and the higher education research project of Beijing Information Science and Technology University (2019GJZD01). We are grateful for the support of these organizations.

References

[1] Y. Liu, Y.-S. Wei, H. Yan, G.-B. Li, and L. Lin, "Causal reasoning meets visual representation learning: a prospective study," *Machine Intelligence Research*, vol. 19, no. 6, pp. 485–511, 2022.

[2] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using unscented rauch-tung-striebel smoother and kernel correlation filter," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6008–6018, 2022.

[3] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFFN: image super-resolution via multi-level features fusion network," *The Visual Computer*, pp. 1–16, 2023.

[4] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, no. 7, pp. 4367–4380, 2021.

[5] Y. Chen, R. Xia, K. Zou, and K. Yang, "FFTI: image inpainting algorithm via features fusion and two-steps inpainting," *Journal of Visual Communication and Image Representation*, vol. 91, Article ID 103776, 2023.

[6] Y. Zhu, Y. Zhang, L. Liu et al., "Hybrid-order representation learning for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1248–1259, 2023.

[7] C. Liu, Z. Li, X. Shi, and C. Du, "Learning a mid-level representation for multiview action recognition," *Advances in Multimedia*, vol. 2018, Article ID 3508350, 10 pages, 2018.

[8] X. Liu and M. Wang, "Context-aware attention network for human emotion recognition in video," *Advances in Multimedia*, vol. 2020, Article ID 8843413, 10 pages, 2020.

[9] M. Y. Shakor and N. M. S. Surameery, "CNN-based transfer learning for 3D knuckle recognition," *Advances in Multimedia*, vol. 2023, Article ID 6147422, 12 pages, 2023.

[10] W. Li, P. Zhang, L. Zhang et al., "Object-Driven Text-To-Image Synthesis via Adversarial Training," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12166–12174, IEEE, 2019.

[11] C. Wu, L. Huang, Q. Zhang et al., "GODIVA: generating open-domain videos from natural descriptions," 2021.

[12] F. Zhuang, Z. Qi, K. Duan et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[13] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[14] W. Xu, Y. Wan, T.-Y. Zuo, and X.-M. Sha, "Transfer learning based data feature transfer for fault diagnosis," *IEEE Access*, vol. 8, pp. 76120–76129, 2020.

[15] X. Li, Y. Hu, M. Li, and J. Zheng, "Fault diagnostics between different type of components: a transfer learning approach," *Applied Soft Computing*, vol. 86, Article ID 105950, 2020.

[16] J. Yosinski, J. Clune, and Y. Bengio, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[17] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: maximizing for domain invariance," 2014.

[18] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[19] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 3071–3085, 2019.

[20] A. Gretton, D. Sejdinovic, and H. Strathmann, "Optimal kernel choice for large-scale two-sample tests," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[21] Y. Ganin, E. Ustinova, and H. Ajakan, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] K. Bousmalis, G. Trigeorgis, and N. Silberman, "Domain separation networks," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[23] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 557–570, 2018.

[24] J. Shen, Y. Qu, and W. Zhang, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings AAAI Conference on Artificial Intelligence*, pp. 4058–4065, AAAI Press, 2018.

[25] C. Shen and Y. Guo, "Unsupervised heterogeneous domain adaptation with sparse feature transformation," in *Proceedings Asian Conference on Machine Learning PMLR*, pp. 375–390, PMLR, 2018.

[26] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu, "Semi-supervised optimal transport for heterogeneous domain adaptation," in *Proceedings International Joint Conference on Artificial Intelligence-Pacific Rim International Conference on Artificial Intelligence (IJCAI)*, pp. 2969–2975, IJCAI, 2018.

[27] Y. H. H. Tsai, Y. R. Yeh, and Y. C. F. Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5081–5090, IEEE, 2016.

[28] Y. H. H. Tsai, Y. R. Yeh, and Y. C. F. Wang, "Heterogeneous domain adaptation with label and structure consistency," in *Proceedings 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2842–2846, IEEE, 2016.

[29] Y.-T. Hsieh, S.-Y. Tao, Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation," in *Proceedings 2016 Institute of Electrical and Electronics Engineers International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2016.

[30] M. Xiao and Y. Guo, "Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation," in *Proceedings Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pp. 525–540, Springer, 2015.

[31] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2142–2150, IEEE, 2015.

[32] W. Y. Chen, T. M. H. Hsu, Y. H. H. Tsai, Y. C. F. Wang, and M. S. Chen, "Transfer neural trees for heterogeneous domain adaptation," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9909 of *Lecture Notes in Computer Science*, pp. 399–414, Springer, 2016.

[33] Y. Yao, Y. Zhang, X. Li, and Y. Ye, "Heterogeneous domain adaptation via soft transfer network," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1578–1586, Association for Computing Machinery, 2019.

[34] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng, "Transferable feature representation for visible-to-infrared cross-dataset human action recognition," *Complexity*, vol. 2018, Article ID 5345241, 20 pages, 2018.

[35] Y. Liu, Z. Lu, J. Li, and T. Yang, "Hierarchically learned view-invariant representations for cross-view action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2416–2430.

[36] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2019.

[37] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5573–5588, 2021.

[38] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "TCGL: temporal contrastive graph for self-supervised video representation learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1978–1993, 2022.

[39] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1302–1310, IEEE, 2017.

[40] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings AAAI Conference on Artificial Intelligence*, pp. 2058–2065, AAAI, 2016.

[41] D. Tran, H. Wang, and L. Torresani, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6459, IEEE, 2018.

[42] C. Feichtenhofer, "X3d: expanding architectures for efficient video recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210, IEEE, 2020.

[43] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6202–6211, IEEE, 2019.

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, IEEE, 2018.

[45] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5552–5561, IEEE, 2019.