Human-object Behavior Analysis Based on Interaction Feature Generation Algorithm

Qing Ye [1], Xiuju Xu [2], Rui Li[3]
School of Information Science and Technology, North China University of Technology, Beijing, China[1, 2]
SNBC, Shandong, China[3]

***Abstract*—Aiming at the problem of insufficient utilization of interactive feature information between human and object, this paper proposes a two-stream human-object behavior analysis network based on interaction feature generation algorithm. The network extracts human-object's feature information and interactive feature information respectively. When extracting human-object features information, considering that ResNeXt has powerful feature expression ability, the network is used to extract human-object features from images. When extracting interactive features information between human and object, an interaction feature generation algorithm is proposed, which uses the feature reasoning ability of graph convolutional neural networks. A graph model is constructed by taking human and objects as nodes and the interaction between them as edges. According to the interactive feature generation algorithm, the graph model is updated by traversing nodes, and new interactive features are generated during this process. Finally, the humans' and objects' features information and the human-object interaction feature information are fused and sent to the classification network for behavior recognition, so as to fully utilize the humans' and objects' feature information and the interaction feature information of human-objects. The human-object behavior analysis network is experimentally verified. The results show that the accuracy of the network has been significantly improved on HICO-DET and V-COCO datasets.**

***Keywords*—Two-stream human-object behavior analysis network; interaction feature generation algorithm; interactive feature information; ResNeXt; graph convolutional neural networks; graph model**

I. Introduction

With the rapid development of science and technology, artificial intelligence (AI) has caused a profound impact in many fields [1], and its continuous improvement of related technologies [3] has penetrated into the application of various fields. In this dynamic context, research on human-object behavior analysis technology [5] has received increasing attention. How to effectively advance development of this technology will directly affect the application space of AI technology in real life. Whether it is possible to accurately and timely determine and analyze various interactions between people and objects in daily life will provide a solid and reliable foundation for further scene understanding and analysis, which will be of great value in theoretical research and engineering implementation.

However, due to the large amount of information, fast action and multiple interaction in the interaction process of human-objects, the specific interaction behaviors between human and object cannot be accurately analyzed by utilizing all features, resulting in a low analysis rate of human-objects' behaviors.

In order to solve this problem, this paper researches on human-object behavior analysis, proposes an interactive feature generation network based on graph convolutional neural networks (GCNNs), and uses this network to analyze and identify the interaction between human and objects, trying to improve the accuracy of human-object behavior analysis on the basis of previous technologies.

The rest of this paper is organized as follows: Section II briefly reviews related work. Section III introduces the related models and algorithms of this paper. Section IV introduces the datasets used in the experiments, the experimental settings, and analyzes and discusses the experimental results. Finally, conclusions are drawn in Section V.

II. Related Work

The traditional feature extraction method, which is to extract the behavioral features of human through the traditional manual method, has the advantages of simple implementation and strong operability. However, due to the relatively fixed templates used, it is difficult to perform fast and effective behavior analysis in facing complex environments and large datasets. Its application scenarios are limited, so more suitable for datasets with few types of behaviors and small scales. In order to improve the accuracy rate of human-object behavior analysis, it is often necessary to train on large datasets. Therefore, facing with such a huge amount of computation, graph convolution has certain advantages.

Simonyan et al. [7] proposed a feature extraction network algorithm based on two-stream convolutional neural networks (CNN). By extracting optical flow features from continuous images, and extracting image features from each frame, temporal and spatial features of continuous frame images are obtained, and two-channel feature information is fused and classified and identified. The proposed method breaks the concept of single-channel feature extraction, breaks through the limitations of feature extraction algorithms, and has a milestone significance. After that, based on multi-channel feature extraction, many researchers continued to improve and conduct in-depth research on human-object behavior analysis. However, these methods also increase a certain amount of computational burden in feature extraction.

Wang et al. [8] proposed a three-channel feature extraction network. In addition, some researchers proposed that human skeleton points, as a means of posture prediction, can be used to supplement features. The multi-channel feature extraction algorithm formed by this method has gradually become research mainstream in the field of human-object behavior analysis. Wang et al. [9] continue to improve two-stream network architecture, and put forward advanced network architectures such as Temporal Segment Networks (TSN) for human-object behavior recognition.

He et al. [10] proposed Residual Network (ResNet) on previous Network research. This network proposes and optimizes the residual module, which not only deepens network depth, but also avoids training network degradation caused by network depth. The proposal of ResNet is of great significance in the field of image analysis. In the following years, researchers have continued to optimize and improve the ResNet, and proposed improved residual networks such as ResNeXt [26].

Yan et al. [11] and Mohamed et al. [12] have proposed spatial temporal graph convolutional networks (ST-GCN), which are used for general representation of skeleton sequences, realize human behavior analysis and recognition based on key points of human skeleton. This method constructs skeleton graph sequence by in space and time connecting the joint nodes of human skeleton, and builds multi-layer ST-GCN network based on this.

He et al. [13] proposed difficulty movement recognition method of calisthenics based on graph convolutional neural network (GCNN). This method constructs the multi-layer pyramid structure [14] of action images to complete the preprocessing of difficult movements in aerobics, and performs the training of samples to realize the difficult movement recognition of calisthenics. GCNN not only has more powerful expression ability and higher behavior recognition accuracy, but also has a stronger generalization ability, which makes it possible that the GCNN can be used for image feature extraction.

Thacker et al. [15] proposed through facial expression recognition to analyze human behavior. Aurangzeb et al. [16] proposed a human behavior analysis and recognition method based on multi-types features fusion and irrelevant features reduction, which initially selects a luminance channel and calculates motion estimation using optical flow. Afterwards, the moving regions are extracted through background subtraction approach. In the features extraction step, shape, color, and Gabor wavelet features are extracted and fused based on serial method. Thereafter, reduced irrelevant and redundant features are removed by Von Neuman entropy approach. The selected reduced features are finally recognized by One-Against-All (OAA) Multi-class SVM classifier.

Degardin et al. [17] provided a comprehensive overview of human behavior analysis over the past decade, presenting state-of-the-art and must-know methods in this field. Lanovaz [18] described machine behavior analysis. He proposed that machine behavior analysis is a science, which can study artificial behavior through its replicability, behavioral terms and philosophical assumptions of human behavior analysis, and study how machines interact with external environment and produce relevant changes. Lin et al. [19] built a framework of external and internal factors to analyze human behavior, and game theory is used to simulate the conflicting interests of human behavior.

Bhatnagar et al. [20] proposed BEHAVE dataset, which is the first full body human-object interaction dataset with multi-view RGBD frames and corresponding 3D SMPL and object fits along with the annotated contacts between them. They also proposed a method that can record and track not just the humans and objects but also their interactions. Peng et al. [21] proposed a 3D max residual feature map convolution network (3D-MRCNN). The model can solve the deficiencies of the network degradation and gradient disappearance caused by convolution calculation, and achieve the improvement of classification accuracy without reducing the training efficiency, which is of value in the field of human behavior analysis in intelligent sensor networks.

In the field of human-object behavior analysis, there are various interactions between human and objects, the interaction between the same human and object may be different. For example, there are a variety of behavioral relations between human and bicycle, such as "riding", "pushing" and "resisting". In addition, it is also common to have multiple simultaneous interactions with the same human-object. For example, when a person picks up a glass to drink, he or she has both "lifting" and "drinking" interactive behaviors. In these cases, if interactive features between human-object aren't fully utilized, even if the human and object are identified, it is difficult to accurately and comprehensively analyze interactive behavior between them, resulting in inaccurate or omission analysis of human-object behavior. Obviously, if interaction features among human-objects aren't fully utilized, the accuracy rate of human-object analysis will be greatly reduced.

However, there are still some limitations in the use and reasoning of interactive behavior feature information in the existing methods for human behavior analysis. Therefore, in order to solve the above problems, this paper proposes a two-stream human-object behavior analysis network based on interactive feature generation algorithm. Although the method in this paper has made remarkable progress in the field of human-object behavior analysis, it still has some limitations in the behavior analysis of fuzzy or obscured human-objects, and its accuracy needs to be further improved.

III. METHOD

A. System Overview

In this paper, aiming at the problem of insufficient utilization of human-objects' interactive behavior features, a two-stream human-object behavior analysis network based on interaction feature generation algorithm is proposed. The network takes human-object features information obtained from the object detector as the input of the interaction feature generation network, and the interaction feature generation algorithm is used to generate new human-object behavior interaction features, thereby enhancing the full utilization of the interaction behavior features. The network block diagram is shown in Fig. 1. Drawing on the idea of "two-stream network" [7], a two-stream human-object behavior analysis network based on CNNs and GCNNs is constructed. This network is divided into two streams to extract feature information of human and objects and the interaction feature information between the human-object from images. First of all, the input image is preprocessed. Then, the preprocessed image and the original image are fed into the object detector (e.g., Faster R-CNN), and the human-object features information in the image are extracted through the backbone of the object detector, such as ResNet. Human-object features information are input into the interaction feature generation network, the interactive features information of them are generated through this network. The specific method are as follows. Firstly, the human-object feature sequence is extracted by using CNN and input into the interactive feature generation network. Secondly, the detected people and objects are taken as nodes and the interactions between them are taken as edges to build an interactive feature graph model, which is used as the input of GCNN. Finally, the interactive feature generation algorithm is used to traverse the nodes of the graph model to update the interactive relationship, and then generate new interactive features of human-object behaviors, so as to make full use of interactive behavioral features. Moreover, the humans' and objects' features information and the human-object interaction feature information are fused and sent to SoftMax classifier [22] for identification and classification, so the recognition result is obtained, which realizes the full use of human-object feature information and human-object interaction feature information.

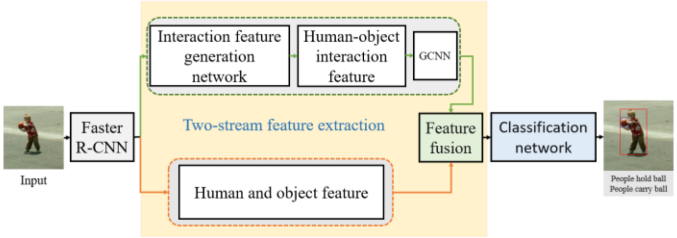


Fig. 1. Overall framework of model.

B. Image Preprocessing

In this paper, image preprocessing involves applying self-transformations to the images, generating new images that are different from the original ones, while preserving the original image category labels. Fig. 2 shows some examples of image preprocessing methods. That is, the image dataset can be augmented by operations such as horizontal flipping, vertical flipping, scaling, random rotation, and generative adversarial networks (GANs) [23] to expand the data samples.

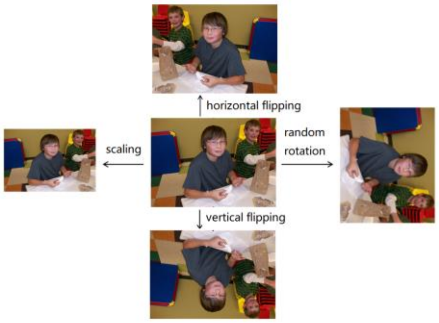


Fig. 2. Example of image preprocessing method.

To reduce computational complexity, increase the training data samples, and improve the model's generalization ability, the original input image is preprocessed before object detection. Specifically, the original image is first input. Next, apply random rotation and flip to generate a new image by randomly rotating and flipping the original input image vertically, horizontally, or both. Then, the image is randomly cropped. Finally, adjust the height and width of all images to 400 x 400. In order to ensure the accuracy and adequacy of the original image and annotation data, no other processing or enhancement operations are applied to the images in this study.

C. Object Detection Network

This study adopts a two-stage approach to detect and analyze human-object behavior, which offers the advantage of reducing irrelevant background and interference from other human, while accurately utilizing and extracting features of the target human and object. Specifically, an object detector is first used for preliminary screening, and regions of interest for human and objects in image are proposed, thereby extracting effective bounding boxes for human and objects. Then, behavior analysis and inference are conducted on the selected humans and objects. To quickly and accurately locate and identify humans and objects in the image, this paper adopts Faster R-CNN [24] as the object detector.

Faster R-CNN is proposed based on the improved Fast R-CNN algorithm [25], which uses Region Proposal Networks (RPN) to replace original Selective Search (SS) method to generate proposal box, and CNN that generates proposal box and object detection are shared, which effectively improves the speed of detection and achieves good results. The whole Faster R-CNN system is a single and unified network, and RPN module is its "attention".

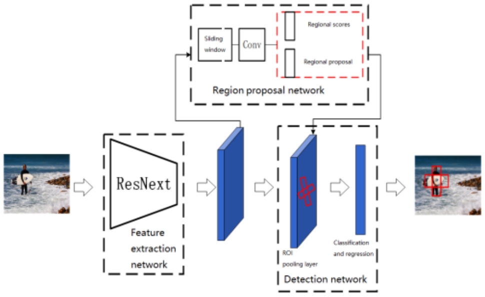


Fig. 3. Network framework of Faster R-CNN.

The network framework of Faster R-CNN is shown in Fig. 3. First, the pre-processed image is input into the feature extraction network (e.g., ResNeXt) to extract the human-object feature information and obtain the image feature map, which is shared by the RPN and the detection network. Second, the extracted feature map is input into RPN, and the sliding window is used to perform convolution operation with the input feature map. According to different scales and different basic sizes, a point on the feature map corresponds to k proposals on the original image. Classification judgment and regression operation are performed on the generated boxes respectively, and $2k$ regional scores and $4k$ regional proposals are obtained. After screening, a relatively accurate region candidate box is finally obtained. Finally, the feature map output by the feature extraction network and the region candidate box information output by the region proposal network are integrated. And a fixed-size proposal feature map is generated through the RoI pooling layer, and that is input into the fully connected layer. The classification and regression network are used for classification and regression training to obtain the accurate position of the predicted object, so as to realize human-objects' identification and detection. The following sections will analyze each part of the Faster R-CNN in detail.

1) Feature extraction network: ResNext [26] is selected as the feature extraction network of Faster R-CNN. Its input is a pre-processed image x, and its output is a convolution feature map containing global features. Its main algorithm is on convolution operations. Convolution is a linear operation, and the specific operation process is shown in Formula (1):

$$s(t) = (x * w)(t) \tag{1}$$

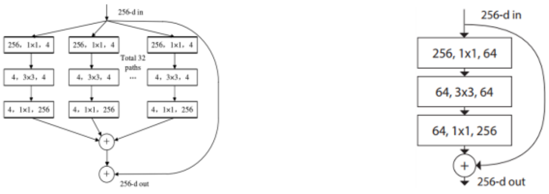
Where x is the input of convolution layer, w is the kernel function, t is the time, and $s(t)$ is the output of convolution layer. When t is set to discrete time, the formula can be expressed as Formula (2):

$$s(t) = \sum_{-\infty}^{\infty} x(a)w(t-a) \tag{2}$$

a represents the discrete time point, $x(a)$ represents the picture x obtained at time a. In this paper, we need to perform convolution operation on image data. The computer sees the input image as a collection of pixels and transforms it into a two-dimensional digital matrix. When the convolution object is two-dimensional, the convolution kernel is also two-dimensional. The process of the convolution operation on the image is expressed by Formula (3):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m,k-n) \tag{3}$$

In this paper, ResNeXt is selected as the backbone network of the object detector, whose network topology is shown in Fig. 4(a). ResNet [10] introduces a residual network topology to alleviate the gradient explosion problem during deep network training. However, as a deep network, although it can better extract features, the increase of network parameters will inevitably bring heavy computational burden, as shown in Fig.4(b). ResNeXt improves the topology of ResNet. It refers to the idea of replacing large convolution kernels with several small convolution kernels in VGG [27] and the idea of splitting in GoogLeNet. It adopts the algorithm of grouping convolution, which realizes that number of network layers and parameters are basically unchanged compared with ResNet, but the recognition accuracy is significantly improved.



(a) ResNeXt (b) ResNet

Fig. 4. Network topology.

The channel in Fig. 4(a) actually changes from 256 to 128 and then to 256, and the number of parameters is 70,144. In Fig. 4(b), the number of ResNet channels changes from 256 to 64 and then back to 256, and the number of parameters is 69,632. However, if the ResNet channel is also changed from 256 to 128 and then to 256, the number of parameters is 212,992, which is much larger than the above two values. Therefore, ResNeXt deepens the depth of feature maps in the convolution layers by grouping convolutions without increasing the number of parameters. ResNeXt combines advantages of GoogLeNet and residual network to improve accuracy without changing model complexity. ResNeXt is easy to train and can effectively avoid exploding gradient problem caused by network deepening. So, it is suitable for feature extraction.

2) Region proposal network: Region Proposal Network (RPN) is a fully convolutional network, which can demarcate human and objects in images and generate bounding boxes. RPN introduces the anchor mechanism, which uses a sliding window to map the anchor points of the feature map to the original image, and each anchor point is mapped to generate k region boxes with different scales and sizes. Intersection over union (IoU) between each region box and the desired region is calculated. When the IoU is greater than 0.7, it is marked as a positive sample; when the IoU is less than 0.3, it is marked as a negative sample. Regression training is performed on these samples, and then the calibrated region proposal boxes are screened and modified. RPN performs a binary classification judgment on whether there are human-objects in the generated k region boxes. If the result is that there are human-objects, the position of the region proposal box is modified according to the regression result. After iterative training, a number of bounding boxes are generated, and the non-maximum suppression algorithm is used to remove the overlapping bounding boxes to get the final effective human and object bounding boxes in images.

3) Detection network: The detection network consists of RoI pooling layer and classification regression network. The main function of the RoI pooling layer is to unify the corresponding features of the bounding boxes through pooling. The feature generated by the ROI pooling layer with a dimension of c × 7 × 7 (c is the number of channels, generally the number of object classes in the dataset) is reshaped into a vector. Send the vector to two full connected layers, and then conduct Softmax to obtain the probability of the object for different classes.

D. Interaction Feature Generation Algorithm

In traditional deep learning, data samples are often considered to be independent. But in graph neural networks (GNNs), each sample node will establish a connection with other data samples through edges, and this connection can be used to form interdependence between different instances [28]. So GNNs has powerful reasoning ability and feature propagation ability.

In this paper, to solve the problem of insufficient use of interactive information between human and objects in the images, an interactive feature generation algorithm based on GCN is proposed. Based on this algorithm, an interactive feature generation network is constructed, and the network framework is shown in Fig. 5. First of all, the pre-processed image is sent to the object detector (e.g., Faster R-CNN), which is used to conduct preliminary detection and segmentation of the image, and obtain the feature sequences of human and object respectively. Secondly, the graph model is constructed by taking people and objects as nodes and the interaction between them as edges, and the graph model is input into the interactive feature generation network in the form of data structure. Finally, each node in the graph model is traversed by the interactive feature generation algorithm, its interaction relationship is updated, and then new interactive features are generated.

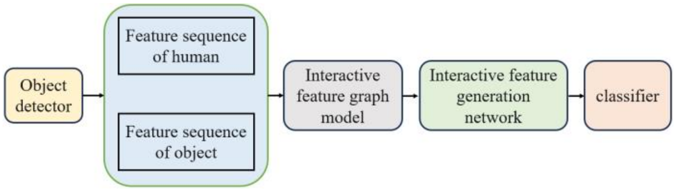

Object detector
Feature sequence of human
Feature sequence of object
Interactive feature graph model
Interactive feature generation network
classifier

Fig. 5. Interaction feature generation network.

A graph model is a data structure composed of nodes and edges. Nodes refer to human and objects instances in the image, and edges refer to the relationship between them. Each node has its own features and structure information. The graph model is generally represented by an adjacency matrix, and the process is shown in Fig. 6. The calculation Formula (4) of the graph convolution operator is as follows, and the central node is set as i:

$$h_i^{l+1} = \sigma\left(\sum_{j\in N_j}\frac{1}{C_{ij}} h_j^l W_{R_j}^l\right) \tag{4}$$

Where h_i^{l+1} represents the feature representation of node i at the $l+1$ layer; C_{ij} represents normalized factors, such as taking the reciprocal of the node degree; N_j represents the neighbor of node j, and contains its own information; R_j represents the type of node j;$W_{R_j}^l$ represents the transformation weight parameter of a node of type R_j.

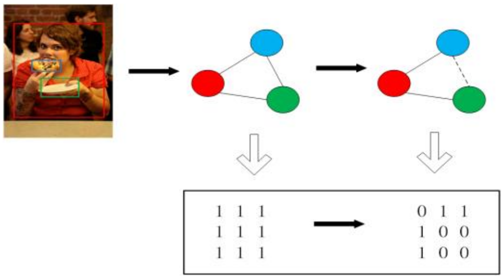


Fig. 6. Construction of graph model.

The construction process of the graph model is shown in Fig. 6. Specifically, humans and objects in the image are taken as nodes, and the interactions between them are taken as edges. Based on the interaction between human and objects, a fully connected undirected graph G is constructed according to Formula (5), and undirected graph is initialized as an adjacency matrix with all elements being 1. Then, each node in the undirected graph is traversed through the interaction feature generation network. During the traversal process, according to whether there is the interaction between human and objects, the undirected graph and the corresponding adjacency matrix are updated.

$$G = (V, E) \tag{5}$$

Where V represents all nodes in the image, E represents all edges. Each object represents a node $v \in V$ of the undirected graph, and the interactive between different target objects is regarded as an edge $e \in E$. If there are n objects in the image, the undirected graph has $n(n-1)/2$ edges. Since the graph model is converted into the form of adjacency matrix for calculation operation, the undirected graph is represented as an adjacency matrix $A_{n\times n}$ of $n \times n$ size, whose matrix elements $i, j \in \{0,1\}$, 0 indicates that there is no interaction between node i and j, and 1 indicates that there is interaction.

Aiming at the problem of underutilization of human-object interaction behavior features, a two-stream human-object behavior analysis network based on interactive feature generation algorithm is proposed. A network example is shown in Fig. 7. Specifically, assume that there is an interaction between the person in the image and all objects in the image. However, in reality, there is generally no interaction between human and objects without overlap in space. Hence, in the interactive feature network, graph model and the corresponding adjacency matrix are updated according to the principle that there is no interaction between people and objects with non-overlapping bounding boxes in space. The updated graph model contains only the possible interactions between human and objects in the image. This not only saves computation, but also facilitates the subsequent graph reasoning.

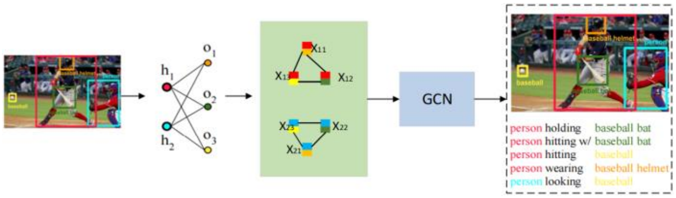


Fig. 7. Example of interaction feature generation algorithm.

The interaction feature generation algorithm learns the structural connection between human and objects, which is shown in Fig. 7. In the process of traversing nodes, better interactive features are generated according to the interaction between human and object, and the update process is shown in Formula (6):

$$x_{\text{ho}}^{(n+1)} = \sigma\left(x_{\text{ho}}^{(n)} + \sum_{o'\in O} W\, x_{\text{ho'}}^{(n)}\right) \tag{6}$$

Where h represents object judged to be human; o represents object judged as object, O represents the total number of objects; W is the projection matrix; $x_{ho}^{(n)}$ is the interaction relation between h and o, $x_{ho'}^{(n)}$ is the interaction between h and o' ($o' \in O$ and $o' \neq o$), $x_{ho}^{(n+1)}$ is the updated interaction between h and o.

The specific interaction feature generation algorithm formula is as follows:

$$F_h = f_h + \sum_{o=1}^{O} X_{ho} W_{oh}(f_o) \qquad (7)$$

$$F_o = f_o + \sum_{h=1}^{H} X_{oh} W_{ho}(f_h) \qquad (8)$$

As shown in Formulas (7) and (8), F represents the newly generated interactive feature vector, f represents the feature vector extracted from the original image, H represents total number of humans in image, and X represents the interaction between objects and human, and $X_{ho} = X_{oh} = x_{ho}^{(n+1)}$.

The generated human-object interaction features and human and object features generated by the object detector are sent to MLP for feature fusion. The fused features are fed into SoftMax classifier to obtain the final human-object behavior analysis result.

IV. Experiment

In this section, the experimental datasets and implementation details are first described. Then, our model is evaluated by quantitative comparison with state-of-the-art methods, followed by ablation studies to validate the components in our framework. Finally, several visualization results are shown to demonstrate effectiveness of our method.

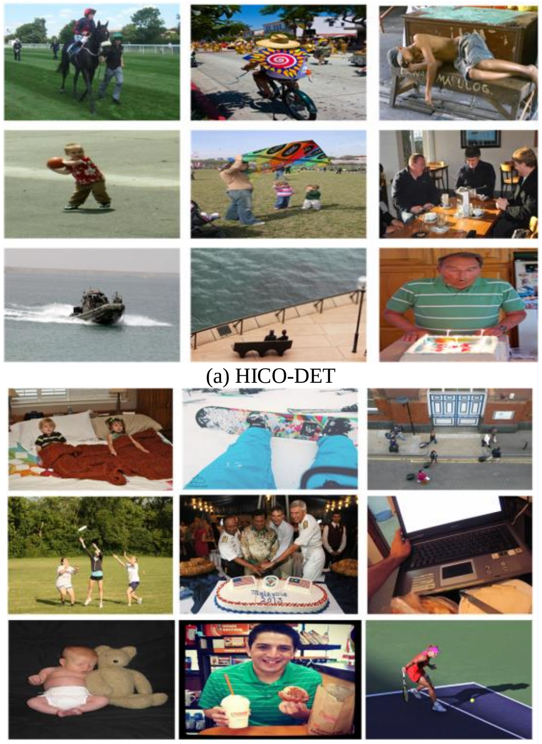
A. Datasets

1) HICO-DET [29]: The HICO-DET dataset, introduced in 2018, is a large dataset for studying human-object behavior. It consists of 117 common behaviors involving 80 different objects, along with their associated behaviors. The sample are shown in Fig. 8 (a). The dataset contains a total of 47,776 images, with 38,118 images allocated to the training set and 9,658 images in the test set. A noteworthy aspect of this dataset is that many images feature multiple pairs of human-object behavior annotations, resulting in over 150,000 human-object pairs and 600 distinct human-object behavior categories in total.

2) V-COCO [30]: V-COCO is a specialized dataset in the field of human-object behavior analysis, which is derived from MS-COCO. The dataset is divided into three main parts: the training set comprising 2,533 images, the validation set containing 2,867 images, and the test set with 4,946 images. It not only inherits all the annotations from the MS-COCO dataset, but also incorporates additional semantic extension markup of human-object pairs through the use of AMT (Amazon Mechanical Turk) proposed by Gupta et al. This extension results in a more professional and refined sub-dataset. The sample are shown in Fig. 8 (b).

B. Implementation Details and Evaluation Metric

In this paper, Faster R-CNN [24] is used as object detector, its backbone module uses ResNeXt50 trained in ImageNet-1 K dataset as feature extractor. The SGD optimizer is used for training with an initial learning rate of 0.01, weight decay of 0.0001, and momentum of 0.9. For V-COCO, the learning rate is reduced to 0.001 at 80 iterations. For HICO-DET, the learning rate is reduced to 0.001 at 60 iterations. All experiments in this paper are conducted on GeForce RTX 2080Ti GPU and CUDA 11.4 with a batch size of 4.



(a) HICO-DET

(b) V-COCO

Fig. 8. Example images of datasets.

This paper uses accuracy to measure the performance of human-object behavior analysis. During accuracy calculation of the metrics, the prediction of the human-object pair is considered correct (1) if the IoU of the human-object bounding box and the ground-truth box is greater than 0.5, and (2) the interaction class label of the prediction of the human-object pair is correct.

C. Experimental Results

To analyze the effectiveness of our method in human-object behavior analysis in more detail, our proposed framework was compared with several existing human-object interaction detection methods on the V-COCO and HCO-DET datasets, and the results are shown in Table I.

TABLE I. Comparison of Accuracy Rate of Human-Object Behavior Analysis Methods

Human-object behavior analysis methods	V-COCO (%)	HICO-DET (%)
Wang[31]	47.3	21.08
DRG[32]	51.0	23.89
TIN[33]	47.8	20.26
PMFNet[34]	52.0	21.20
PFNet[35]	52.8	24.89
PPDM[36]	-	26.84
IP-Net[37]	51.0	23.92
UnionDet[38]	47.5	21.27
Ours method	52.4	27.89

By comparison and analysis of the data in Table I, it can be seen that in recent years, the accuracy of most human-object behavior analysis methods on the V-COCO database has been generally more than 50%, among which the PFNet method has the highest accuracy of 52.8%, which is 0.4% higher than our method. One important reason is that PFNet uses more fine-grained body part level information. However, because this paper uses a graph model structure with strong inference ability, and design a graph node adaptive interactive update algorithm, so the results are not very different. Although the accuracy of our method is lower than that of PFNet method, it has a significant increase compared with other methods.

According to the data in Table I, the accuracy of human-object behavior analysis on the HICO-DET database is generally low. However, our method achieves the state-of-the-art results on the HICO-DET database, which is 1.05% higher than the previous state-of-the-art the PPDM method. It is inferred that the reason may be the difference in the difficulty and the focus of dataset. Compared with the existing human-object behavior analysis, the accuracy rate of the two-stream network has been significantly improved.

D. *Ablation Study*

In this paper, a two-stream human-object behavior analysis network based on interaction feature generation algorithm is proposed and its effectiveness is tested. The first experiment is to test the feature extraction networks in object detector. ResNeXt 50 and ResNet 50 are used as the feature extraction networks separately. Training and testing are carried out on the HICO-DET and the V-COCO datasets. The suitable extraction network is selected by comparing the test results on each dataset. The second experiment is to verify the effectiveness of network proposed in this paper. In this experiment, three networks are proposed: (1) a single-stream human-object behavior analysis network without using interaction feature generation algorithm; (2) a single-stream human-object behavior analysis network with interaction feature generation algorithm; (3) a two-stream human-object behavior analysis network based on interaction feature generation algorithm. Training and testing are carried out on the HICO-DET and V-COCO datasets respectively, and through the test results on each dataset, it is proved whether the interaction feature generation algorithm and the two-stream human-object behavior analysis network based on the interaction feature generation algorithm are beneficial to improve the accuracy rate of human-object behavior analysis.

1) Comparison experiment between ResNeXt 50 and ResNet 50: In this experiment, ResNeXt 50 and ResNet 50 are used as the backbone of the object detector, respectively, training and testing are carried out in two databases. ResNeXt 50 and ResNet 50 pre-trained models are tested in the new database by means of transfer learning. Fig. 9 and Fig. 10 show the experiments performed on V-COCO and HICO-DET databases, respectively.

As shown in Fig. 9, the horizontal axis represents the number of training epochs, and the vertical axis represents the accuracy of the test set for 29 kinds of human-object behavior analysis in the V-COCO database. The green and red curves represent the accuracy test results of ResNeXt 50 and ResNet 50 as the number of training epochs increases, respectively.

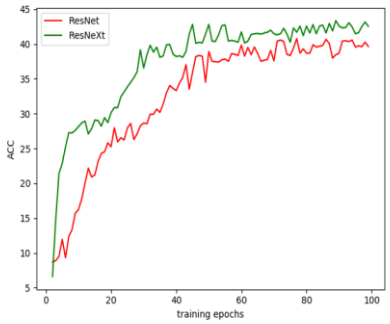


Fig. 9. Comparison of accuracy of classification results in V-COCO database.

As shown in Fig. 10, the horizontal axis represents the number of training epochs, the vertical axis represents the test sets accuracy of 117 kinds of human-object behavior analysis in HICO-DET database. The green and red curves represent the accuracy test results of ResNeXt 50 and ResNet 50 as the number of training epochs increases, respectively.

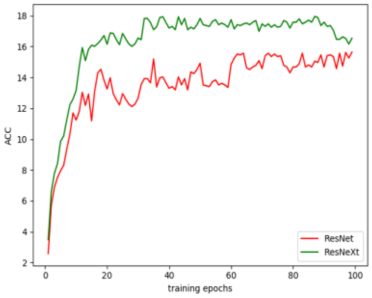


Fig. 10. Comparison of accuracy of analysis results in HICO-DET database.

It can be seen from the experimental results in Fig. 9 and Fig. 10, the number of training epochs of the network model is the same, and the accuracy rate of the human-object behavior analysis result of ResNeXt 50 is better than that of ResNet 50. In HICO-DET database, when ResNeXt 50 is used as the backbone of the object detector, the convergence is better, and its accuracy is significantly higher than that ResNet 50. The use of ResNeXt 50 as backbone for object detector in human-object behavior analysis can be concluded to significantly improve accuracy.

2) Verify the effectiveness of the two-stream human-object behavior analysis network based on interaction feature generation algorithm: Based on the above experimental analysis, this study decided to use ResNeXt 50 as the feature extraction network of object detector. Therefore, in the second experiment, in order to verify the effectiveness of the two-stream network, training and accuracy test are carried out on the single-stream network without using the interactive feature generation algorithm, the single-stream network using interactive feature generation algorithm, and the two-stream network based on the interactive feature generation algorithm on the two databases respectively.

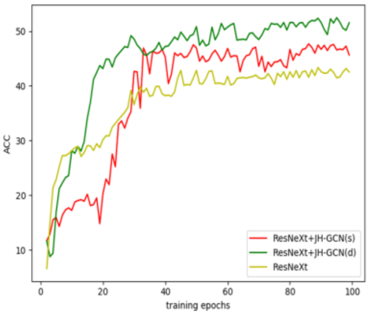


Fig. 11. Effectiveness comparison experiment in V-COCO database.

As shown in Fig. 11, the three proposed network structures are trained and tested in the V-COCO database. The horizontal axis represents training epochs, and the vertical axis represents the accuracy rate of the 29 human-object behavior test sets in the V-COCO database. The green, red and yellow curves respectively represent the training and accuracy testing process on the two-stream network, the single-stream network using the interactive feature generation algorithm, and single-stream network without using interactive feature generation algorithm.

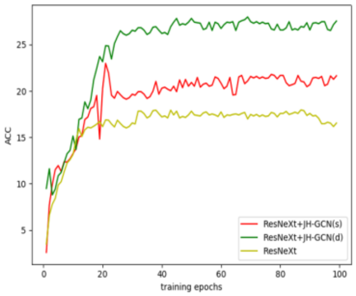


Fig. 12. Effectiveness comparison experiment in HICO-DET database.

As shown in Fig. 12, the horizontal axis represents training epochs, the vertical axis represents accuracy of the analysis of 117 common behaviors in the HICO-DET database. The green, red and yellow curves respectively represent the training and accuracy testing process on the two-stream network, the single-stream network using interaction feature generation algorithm, and single-stream network without using interaction feature generation algorithm.

The experimental results of training and testing on two datasets are analyzed in Fig. 11 and Fig. 12. The overall trend of the curve is as follows. The green curve is above the red and yellow curve, and the red curve is above the yellow. In the V-COCO database, there are coincident points before the curves converge. But after convergence, the convergence accuracy of the two-stream network is higher than the other two networks, and the convergence speed is faster. The accuracy rate of the single-stream network using the interaction feature generation algorithm is higher than that of the single-stream network without using interaction feature generation algorithm. In the HICO-DET database, all three network curves converge quickly. It is obvious that the accuracy of the two-stream network after convergence is higher than that of other two networks. And the accuracy of the single-stream network with interaction feature generation algorithm is higher than that of the single-stream network without using the interaction feature generation algorithm.

From the above analysis, it can be concluded that the interaction feature generation algorithm can effectively improve the accuracy of human-object behavior analysis, and accuracy of the two-stream network is higher than that of other two networks, which indicates the effectiveness of the two-stream network in human-object behavior analysis.

In this paper, four network architectures are trained and tested in the V-COCO and HICO-DET databases. The single-stream network without the interaction feature generation algorithm and the feature extraction network is ResNet 50, the single-stream network without interaction feature generation algorithm and the feature extraction network is ResNeXt 50, the single-stream network with interaction feature generation algorithm and the extraction network is ResNeXt 50, and the two-stream network based on interaction feature generation algorithm (the feature extraction network is ResNeXt 50).

TABLE II. COMPARISON OF ACCURACY OF FOUR NETWORKS

The network architecture	V-COCO	HICO-DET
ResNet 50	40.3	15.73
ResNeXt 50	43.2	17.71
ResNeXt 50+ Interaction feature generation algorithm	47.6	21.54
A two-stream human-object behavior analysis network based on interaction feature generation algorithm	52.4	27.89

The specific accuracy the human-object behavior analysis methods of the four network architectures are shown in Table II. When ResNet 50 is used as the feature extraction network of the object detector, the model obtained after training on the V-COCO database is used to test, its accuracy is 40.3%; the model obtained after training on the HICO-DET database is used to test, its accuracy is 15.73%. When ResNet 50 is used as the feature extraction network of the object detector, the model obtained after training on the V-COCO database is used to test, its accuracy is 43.2%, which is 2.9% higher than ResNet 50; the model obtained after training on the HICO-DET database is used to test, its accuracy is 17.71%, which is 1.98% higher than ResNet 50. It can be concluded that ResNeXt 50 is more suitable as the feature extraction network of the object detector.

On the basis of determining that the feature extraction network is ResNeXt 50, according to the proposed interaction feature generation algorithm, the single-stream and the two-stream human-object behavior analysis network based on the interaction feature generation algorithm are proposed. The test results are shown in Table II. For V-COCO database, the model obtained after training of the single-stream network using the interaction feature generation algorithm is tested, its accuracy rate is 47.6%, which is 4.4% higher than that without the interaction feature generation algorithm. The model obtained after training on the HICO-DET database is used to test, its accuracy is 21.54%, which is 3.83% higher than that without the interaction feature generation algorithm. The experimental data show that interaction feature generation algorithm is beneficial to improve the accuracy of human-object behavior analysis.

For V-COCO database, the model obtained after training of the two-stream network is tested, its accuracy is 52.4%, which is 4.8% higher than the single-stream network using interaction feature generation algorithm. The model obtained after training on HICO-DET database is used to test, its accuracy is 27.89%, which is 6.35% higher than that the single-stream network using the interaction feature generation algorithm. As shown in Table II, the accuracy rate of the two-stream network based on interaction feature generation algorithm is higher than that of the other three networks, which shows that the network is effective in human-object behavior analysis.

E. Visualized Results

To qualitatively visualize the detection effect of our model, Fig. 13 shows the visualization results of human-object behavior analysis on the test images of our model.

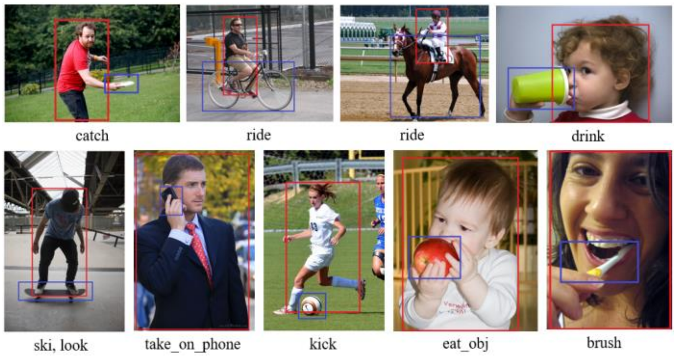


Fig. 13. Example detections of human-object behavior analysis method based on interaction feature generation algorithm.

In the nine example detections images in Fig. 13, after the human and objects are detected in each image, the behavior analysis results are successively, "catch", "ride", "ride", "drink", "ski" and "look", "talk_on_phone", "kick", "eat_obj", "brush". Human and objects as background aren’t selected. It can be concluded that the human-object behavior analysis method based on the interaction feature generation algorithm proposed in this paper is effective.

As can be seen from the above experiments, the experimental results validate the effectiveness of our proposed model in enhancing human-object behavior analysis performance. The comparison ablation study and qualitative results collectively demonstrate the superior performance and robustness of our proposed method on V-COCO and HICO-DET datasets. Therefore, the algorithm can be used in areas such as intelligent surveillance video and intelligent robots.

V. Conclusion

Aiming at the problem of insufficient utilization of interactive behavior feature information between human and object, a two-stream human-object behavior analysis network based on interaction feature generation algorithm is proposed. The network uses interactive feature generation algorithm to generate new behavior interaction features, so as to make the full use of interactive behavior features. After experimental verification on datasets, the recognition accuracy of the two-stream network based on interaction feature generation algorithm is higher, compared with the single-stream network that only uses the image feature extraction network and only uses the interaction feature generation algorithm.

However, the proposed algorithm still has some limitations when it comes to the behavior analysis of fuzzy or obscured human-objects in images. Therefore, before conducting human behavior analysis, additional techniques are required to accurately process and identify blurred or obscured people or objects. Future research efforts will focus on improving the processing of difficult and fuzzy human-object recognition processes, aiming to optimize human-object behavior analysis techniques and improve the accuracy of human-object behavior analysis.

Acknowledgment

This paper is supported by Key Research and Development Program (No.2020YFC0811004), Key Research and Development Program (No.2020YFB1600702), Technology Project of Beijing Municipal Education Commission (No. SQKM201810009002), Beijing Innovation Team, Key scientific research direction construction project of North China University of Technology. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Y. Xu, The Vigorous Development of Artificial Intelligence Innovation Application[N]. People’s Posts and Telecommunications News, 2022-05-10(001).

[2] B. Zhang, J. Zhu, H. Su, "Toward the third generation artificial intelligence." Science China Information Sciences 66.2 (2023): 121101.

[3] X. Zhang, "Application of artificial intelligence recognition technology in digital image processing." Wireless Communications and Mobile Computing 2022 (2022): 1-10.

[4] J.H. Tao, J.T. Gong, N. Gao, S.W. Fu, S. Liang, C. Yu, Human-computer interaction oriented to virtual-real integration. Chinese Journal of Image and Graphics, 2023, 28(06):1513-1542.

[5] X. Yun, H.S. Song, H.X. Liang, et al., Behavior Analysis System for Key Positions Based on Deep Learning[J]. Computer Engineering and Applications, 2021, 57(06): 225-231.

[6] M.L. Deng, Z.D. Gao, L. Li, et al., A review of human behavior recognition based on deep learning[J]. Computer Engineering and Applications, 2022, 58(13): 14-26.

[7] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 1(4): 568–576.

[8] H. Wang, C. Schmid, Action recognition with improved trajectories[C]. Proceedings of the IEEE international conference on computer vision, 2013: 3551-3558.

[9] L. Wang, Y. Xiong, Z. Wang, et al., Temporal segment networks for action recognition in videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2740-2755.

[10] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778..

[11] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. 32nd AAAI Conference on Artificial Intelligence, 2018: 7444-7452.

[12] A. Mohamed, K. Qian, M. Elhoseiny, et al., Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14424-14432.

[13] L. He, H.M. Li, J.G. Sun, et al., Difficulty movement recognition method of calisthenics based on graph convolutional neural network[J]. Journal of West Anhui University, 2022, 38(02): 136-141.

[14] S.M. Pan, Y.J. Wang, Y.W. Zhong, Cross-domain pedestrian re-identification based on graph convolutional neural network[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2020, 48(09): 44-49.

[15] C.B. Thacker, R.M. Makwana, Human behavior analysis through facial expression recognition in images using deep learning[J]. International Journal of Innovative Technology and Exploring Engineering, 2019, 9(2): 391-397.

[16] K. Aurangzeb, I. Haider, M.A. Khan, et al., "Human behavior analysis based on multi-types features fusion and Von Nauman entropy based features reduction." Journal of Medical Imaging and Health Informatics 9.4 (2019): 662-669.

[17] B. Degardin, H. Proença, Human behavior analysis: a survey on action recognition[J]. Applied Sciences, 2021, 11(18): 8324.

[18] M.J. Lanovaz, Some Characteristics and Arguments in Favor of a Science of Machine Behavior Analysis[J]. Perspectives on behavior science, 2022, 45(2): 399-419.

[19] A. Lin, Y. Xu, H. Shen, Quantitative Analysis of Human Behavior in Environmental Protection[J]. Journal of the Knowledge Economy, 2022: 1-28.

[20] B.L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, (2022). Behave: Dataset and method for tracking human object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15935-15946.

[21] B. Peng, Z. Yao, Q. Wu, H. Sun and G. Zhou, (2022). 3D convolutional neural network for human behavior analysis in intelligent sensor network. Mobile Networks and Applications, 27(4), 1559-1568.

[22] G.X. Liu, J. Huang, Transfer learning technology of machine vision detection discriminative semantic segmentation based on label reservation Softmax algorithm[J]. Optical Precision Instruments, 2022, 30(01): 117-125.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[24] H. Duan, J. Huang, W. Liu and F. Shu, (2022, August). Defective Surface Detection based on Improved Faster R-CNN. In 2022 IEEE International Conference on Industrial Technology (ICIT), pp. 1-6. IEEE.

[25] R. Girshick, Fast R-CNN [C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[26] Z. Zhang, Research on face recognition method and application based on single sample[D]. University of Electronic Science and Technology of China, 2021.

[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409-1556, 2014.

[28] X.R. Wang, H. Zhang, Small sample classification network based on attention mechanism and graph convolution[J]. Computer Engineering and Applications, 2021, 57(19): 164-170.

[29] Y.W. Chao, Y. Liu, X. Liu, H. Zeng and J. Deng, (2018, March). Learning to detect human-object interactions. In 2018 ieee winter conference on applications of computer vision (wacv), pp. 381-389. IEEE.

[30] S. Gupta, J. Malik, Visual semantic role labeling[J]. arXiv preprint arXiv:1505-04474, 2015.

[31] H. Wang, W.S. Zheng and Y.B. Ling, "Contextual heterogeneous graph network for human-object interaction detection." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020.

[32] C. Gao, J. Xu, Y. Zou and J.B. Huang, "Drg: Dual relation graph for human-object interaction detection." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020.

[33] Y. L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. S. Fang, et al., (2019). Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3585-3594.

[34] B. Wan, D. Zhou, Y. Liu, R. Li and X. He, (2019). Pose-aware multi-level feature network for human object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9469-9478.

[35] H. Liu, T. J. Mu and X. Huang, (2021). Detecting human-object interaction with multi-level pairwise feature network. Computational Visual Media, 7, 229-239.

[36] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian and J. Feng, (2020). PPDM: Parallel point detection and matching for real-time human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 482-490.

[37] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, J. Sun, (2020). Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4116-4125.

[38] B. Kim, T. Choi, J. Kang and H.J. Kim, (2020). Uniondet: Union-level detector towards real-time human-object interaction detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16 (pp. 498-514). Springer International Publishing.