mathematics

Article

HOME: 3D Human–Object Mesh Topology-Enhanced Interaction Recognition in Images

Weilong Peng [1,†], Cong Li [1,†], Keke Tang [2,*], Xianyong Liu [3,*] and Meie Fang [1,†]

[1] School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511442, China
[2] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 511442, China
[3] Robotics Institute, Ningbo University of Technology, Ningbo 315100, China
* Correspondence: tangbohutbh@gmail.com (K.T.); liuxianyong@nbut.edu.cn (X.L.)
† These authors contributed equally to this work.

Abstract: Human–object interaction (HOI) recognition is a very challenging task due to the ambiguity brought by occlusions, viewpoints, and poses. Because of the limited interaction information in the image domain, extracting 3D features of a point cloud has been an important means to improve the recognition performance of HOI. However, the features neglect topological features of adjacent points at low level, and the deep topology relation between a human and an object at high level. In this paper, we present a 3D human–object mesh topology enhanced method (HOME) for HOI recognition in images. In the method, human–object mesh (HOM) is built by integrating the reconstructed human and object mesh from images firstly. Therefore, under the assumption that the interaction comes from the macroscopic pattern constructed by spatial position and microscopic topology of human–object, HOM is inputted into MeshCNN to extract the effective edge features by edge-based convolution from bottom to up, as the topological features that encode the invariance of the interaction relationship. At last, topological cues are fused with visual cues to enhance the recognition performance greatly. In the experiment, HOI recognition results have achieved an improvement of about 4.3% mean average precision (mAP) in the Rare cases of the HICO-DET dataset, which verifies the effectiveness of the proposed method.

Keywords: mesh topology; interaction recognition; HOME; HOI

MSC: 68T45

check for updates

Citation: Peng, W.; Li, C.; Tang, K.; Liu, X.; Fang, M. HOME: 3D Human–Object Mesh Topology-Enhanced Interaction Recognition in Images. *Mathematics* **2022**, *10*, 2841. https://doi.org/10.3390/math10162841

Academic Editors: Juan Cao, Li Zhang and Hongwei Lin

Received: 29 June 2022
Accepted: 4 August 2022
Published: 10 August 2022

1. Introduction

Human–object interaction detection (HOI) is a task to locate pairs of a human and an object in the scene and detect the interactions between them. It could be applied to many areas, e.g., video retrieval [1–3] and activity recognition [4–6] in videos. Recently, there are many methods proposed for object detection and recognition algorithms [7–9], but there are still some factors that affect the recognition performance seriously: (1) the same type of interaction may occur in different scenes; (2) multiple people may interact with the same object, and some different interactions are similar in perception. These factors make HOI detection a challenging task.

To handle multiple human–object pairs of interaction, many methods are developed to capture context information from images, using visual features of a human and an object, as well as their spatial features [10–12]. Generally, existing HOI detection methods could be roughly divided into three categories: attention mechanism-based methods [11,13], pose estimation-based methods [12,14,15], and scene graph-based methods [10,16]. First, some works adopt the idea of attention mechanism to obtain more abundant and effective features for interactive detection. Second, some works use human pose information to infer character interactions with more fine-grained human visual features. Third, some methods obtain context clues by building a scene graph, mainly by building a fully connected graph, in which nodes represent all detected objects. All in all, HOI is a multi-stage task, the success of other computer vision task research greatly promoted the development of HOI.

Obviously, most of the existing methods are based on the image level. However, understanding 2D vision-based behavior is difficult due to interference by perspective occlusions. For a certain interaction, it shows very different appearances in different human postures or from different views, e.g., different rider–horse appearances shown in the first row of Figure 1, which hinders extracting visual invariable information related to interaction action. Apart for the visual cues, the 3D spatial and topological information also provide the most important cues because they are similar for same action in geometry and topology, e.g., the similar human meshes with same riding posture shown in the second row of Figure 1. Therefore, the key to improving the HOI detection is to use the reconstructed connection information of human and object from the images and extract the topological from the bottom up during the perception process of HOI recognition. In our work, the effective edge information that represents the topological invariant features in the interaction relationship is exacted from the human–object mesh, which can greatly enhance the performance of the image-based HOI recognition task.

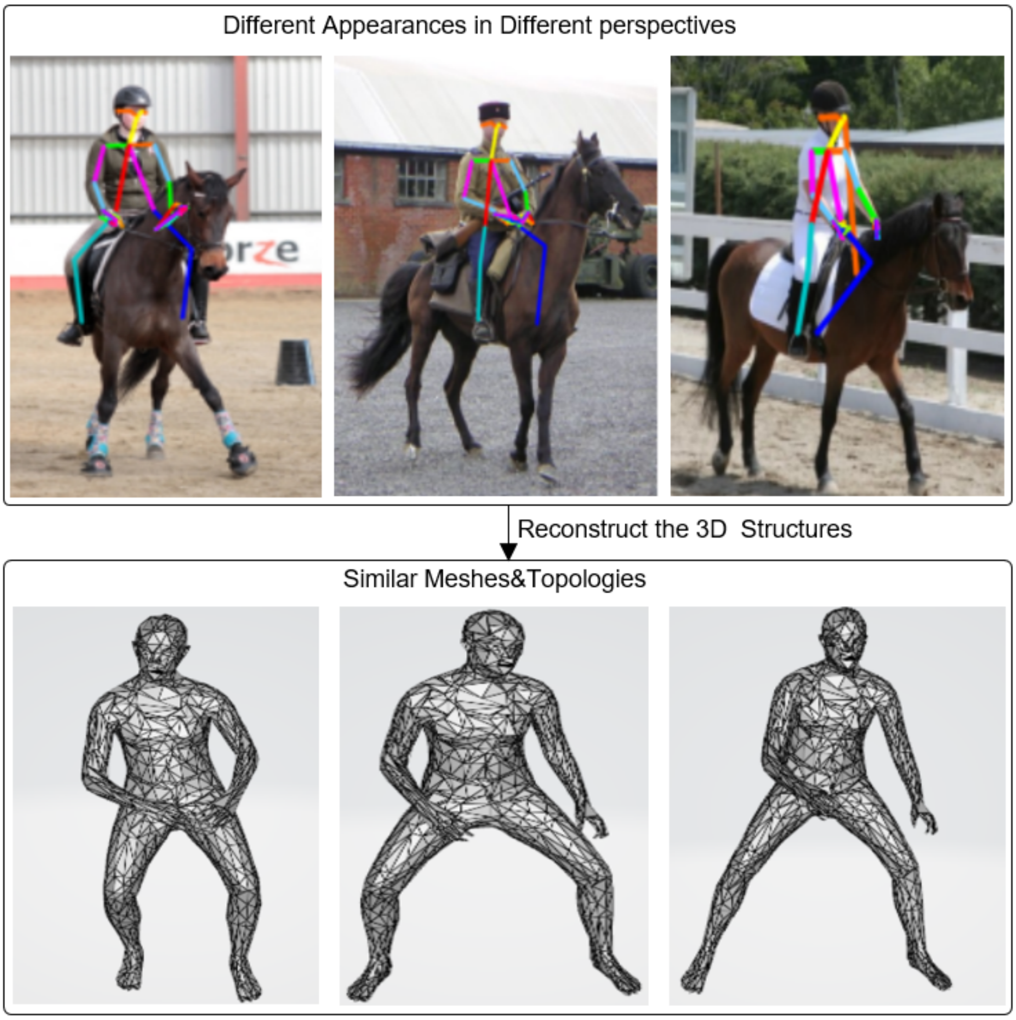

Different Appearances in Different perspectives
Reconstruct the 3D Structures
Similar Meshes&Topologies

Figure 1. Visual comparison of 3D meshes and topologies from different perspectives of images in the same HOI.

In this paper, we propose a 3D human–object mesh topology enhancement method (HOME) for HOI recognition. Firstly, we construct a human–object mesh (HOM) that contains human geometry and the 3D interaction relation between a human and an object.

Secondly, we build a framework that fuses both visual cues and topology cues for recognition. In particular, to find the interaction relation and to extract edge features in the HOM from bottom to top, we use MeshCNN [17], so as to construct invariant topological features that can represent interaction relationships from shallow to deep, which greatly promoted HOI discrimination performance.

In summary, our contributions are twofold:

- We provide the first perspective that human–object interaction is derived from HOM geometric topology and also propose a novel method of interaction detection that considers the bottom-up topological cues.
- We propose a HOME framework that fuses both visual cues and topological cues, respectively, from CNN and MeshCNN. It approaches the state-of-art level.

This paper first introduces the related work in Section 2. Then, this paper introduces the method in Section 3 and shows the experimental results in Section 4. Finally, we summarize our method in Section 5.

2. Related Work

Our work involves using mesh perception to enhance the recognition of HOI relation. Therefore, we will review three aspects of methods, including HOI detection, Graph Models, and 3D perception.

2.1. HOI Detection

HOI detection is very important for complex scene understanding and has been studied widely and intensively [18,19]. Human–Object interaction detection mainly includes two stages: object detection [8,9] and interaction recognition [19]. Previously, Chao et al. [19] proposed the HORCNN method, which firstly detects people and objects in the images, and then applies a deep neural network to extract visual and spatial features of people and objects, and finally, fuses them. The visual cues are quite limited for interaction relation discovery. Therefore, attention mechanism-based methods [11,13], pose estimation-based methods [14,15], and scene graph-based methods [10,16] are designed to strengthen the features. In terms of attention, pooling [20], spatial relations [16], body-part [21] , and instance [11] based attention are proposed from different levels to extract effective context information for action or interaction recognition. Considering that pose gives very relative cues to human activity, PMFNet [14] uses the pose-aware multi-level features to adaptively concentrate on the relevant areas of body parts. PastaNet [15] uses poses to guide part-state recognition of the body and reduces the representation of movements toward the human activity knowledge engine. To handle the many-to-many problem between human and object, GPNN [10] defines a fully connected graph in which the nodes represent all people and objects in the scene and parse the interaction relationship between people and objects. Recently, a lot of work has been done to add semantic information to assist interaction detection [22,23]. DRGNet [22] extends spatial features by embedding word features in each object. PastaNet [15] obtains the semantic information of each body part and object through BERT 's pre-trained model [24], and then learned the model through the fusion of semantic information and visual features.

2.2. Graph Models

To some extent, visual semantics is thought to be hidden behind the scene graph. Qi et al. [10] introduced the graph model for HOI detection, in which a fully connected graph is used to represent the relationship between humans and an object. Node features are initialized using appearance features and are updated during message passing. Wang et al. [25] improved the relation modeling with a heterogeneous graph that is better. Gao et al. [22] took advantage of the heterogeneity of nodes to construct two sparse subgraphs centered on people and objects. Based on these graph models, the pair of spatial relations is encoded into the node features. The currently known graph model focuses on the relation extraction from an appearance feature and a coarse spatial feature. In contrast, we build the graph relation based on mesh of a human and an object, which discovers the geometrical semantics of the interaction.

2.3. 3D Perception

In the recent years, the advance of geometry deep learning greatly promoted 3D perception tasks, e.g., point cloud classification [26–28], segmentation [29,30], and detection [31,32], mesh registration [33,34], etc. Previously, PointNet [30] adopted maxpooling to aggregate point set information to maintain the invariance of the point arrangement. Later, the variant PointNet++ [26] improved it by capturing hierarchical feature information. Since the perception on point cloud neglects the topology of geometry, mesh-based neural networks are proposed to perceive the inherent structure. For example, MeshCNN [17] develops a mesh pooling operation based on edge information for integrating the grid features from bottom to up that imitates the 2D convolution. In our opinion, the mesh perception reflects the topological semantics of a scene or object. If we build a triangular mesh that unifies a human and the corresponding object under an interactive relationship, the invariant features related to the interaction classification could be better captured.

3. HOME

In this section, we introduce the implemented HOME method in detail. We first describe how to build human–object mesh (HOM) from images. Then, we introduce the overall framework that considers HOM perception with respect to topological cues. Finally, we give the fusing loss for visual cues and topological cues to train HOI recognition model.

3.1. HOM Modeling

To learn interactive actions by using the topological information of a human and an object, we need to construct mesh data that represent the integration of the human and the object. Firstly, we detect all the human and object bounding boxes in images. Then, for each pair of human and object, we reconstruct their 3D shape, and merge them into an integrated human–object mesh (HOM) model, as shown in Figure 2.

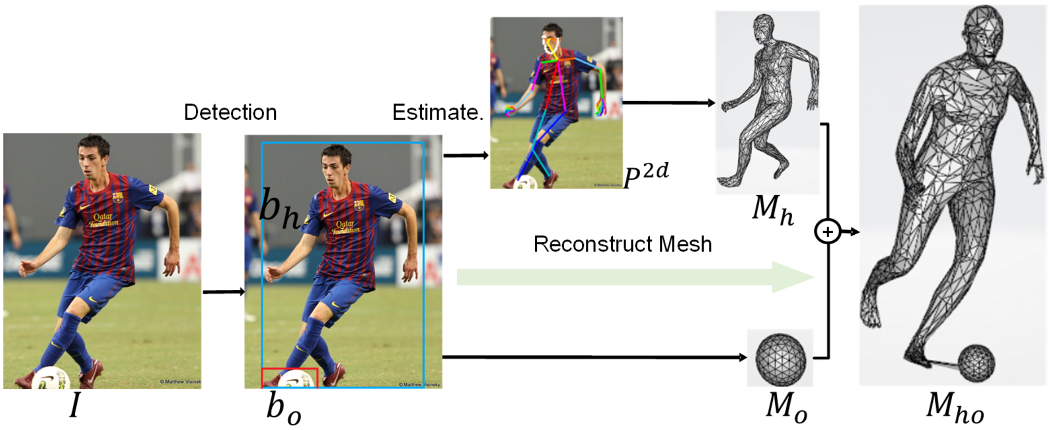


Figure 2. Building HOM from Image. Given the input image I, the bounding boxes of human and objects <b_h,b_o> are obtained from the detector firstly. Then, 2D human pose P^{2d}, human body mesh M_h, and object sphere mesh M_o are estimated and reconstructed subsequently. Finally, M_h and M_o are merged as HOM model M_{ho}. Note: the red rectangle and blue rectangle are the detection boxes of human and object respectively; the same below in other figures.

In detail, after the 2D object detection on image I, we obtain the bounding boxes <b_h, b_o> of human and object with potential interaction relation. The region of interest within b_h is fed into Openpose [35] to get the human 2D pose P^{2d} and camera parameters π. The human region is put into SMPLify-X [36] along with pose P^{2d} to recover the 3D body shape. The original reconstructed 3D body mesh has 10,475 vertices and 20,908 triangular facets. By considering the problem of GPU memory, the data are simplified to M_h with resolution of 3000 facets by considering economic GPU memory. According to π and b_o, we further estimate the object mesh represented by a hollow sphere using DJ-RN [37]: (1) locating the sphere center (o_1, o_2) in the image plane according to the camera perspective projection π, (2) using the prior object size and human–object distance to estimate the depth o_3; and then, (3) the position and size of the object are featured with parameter $O(o_1, o_2, o_3)$ and r. In particular, r is its prior radius of a certain object category, and the focal length f of the camera model is set to a fixed value of 5000 for all images. Finally, the 3D shape of object M_o is represented using a discrete mesh with 162 vertices and 320 facets.

We compute the HOM model by fusing M_h and M_o. To guarantee consistent HOM in a manifold, we choose three key points from the human and object, respectively, to define the connection relationship. However, the relative positions of people and objects are always changeable, because the interactions are in a dynamic manner in the actual scenes and they are contactless in some scenes. There are also some interaction relations in which there exist no physical contact between the human and object. Therefore, the key points are dynamically picked. This process is as follows:

(1) Calculate the space distance set from the object center O to all human body triangle center points.
(2) Find the closest human body mesh triangle F^h_{key} according to the distance set, and the corresponding three vertices (V'_1, V'_2, V'_3).
(3) Calculate the spatial distance set from the human body triangle F^h_{key} to the object center, and find the closest object triangle F^o_{key} and corresponding key points (V'_1, V'_2, V'_3) based on the same principle.
(4) Eliminate triangle F^h_{key} and M^o_{key}, and then merge M_h and M_o into M_{ho} as the final HOM model.

Some reconstructed HOM examples are shown in Figure 3. We can see that the HOMs in the second row illustrate the geometry topology of interactions clearly in the corresponding images in the first row.

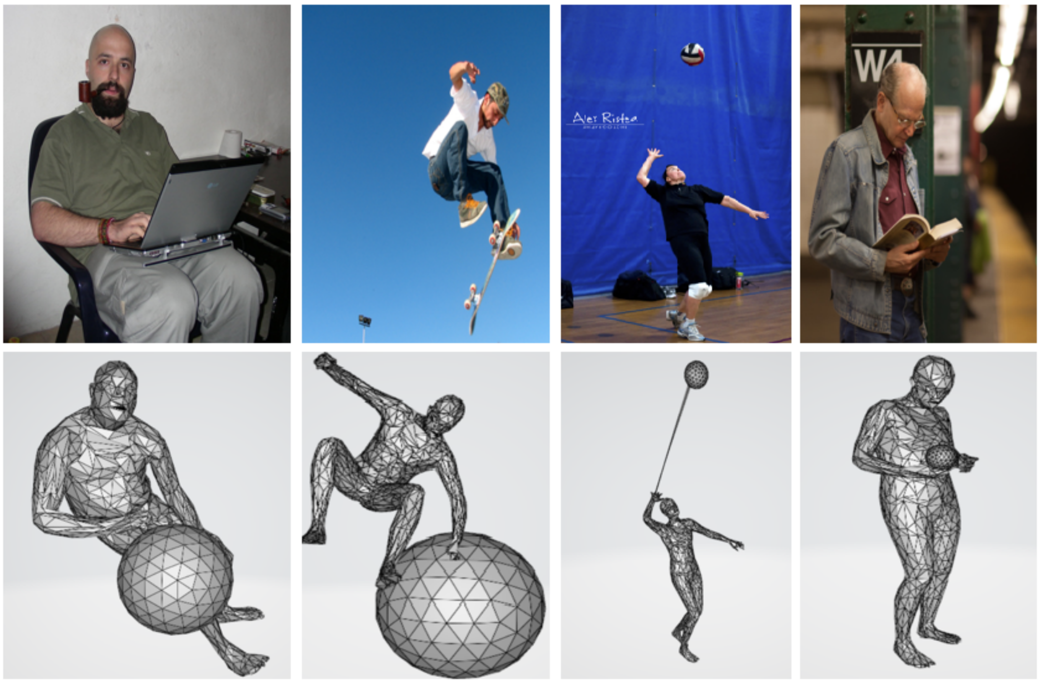


Figure 3. HOM examples (**the second row**) reconstructed from their corresponding images (**the first row**).

3.2. Framework of HOME

Figure 4 shows the framework human–object mesh topology-enhanced interaction recognition (HOME) framework clearly. On the whole, HOME includes several branches covering visual cues and topology cues to perceive the interaction in different aspects of information in the image. For visual cues, we extract the appearance features and the 2D spatial features, respectively, of the human and object. In addition, for topology cues, we extract the mesh features of HOM. At last, topology cues are fused into the visual branches to enhance the interaction classification.

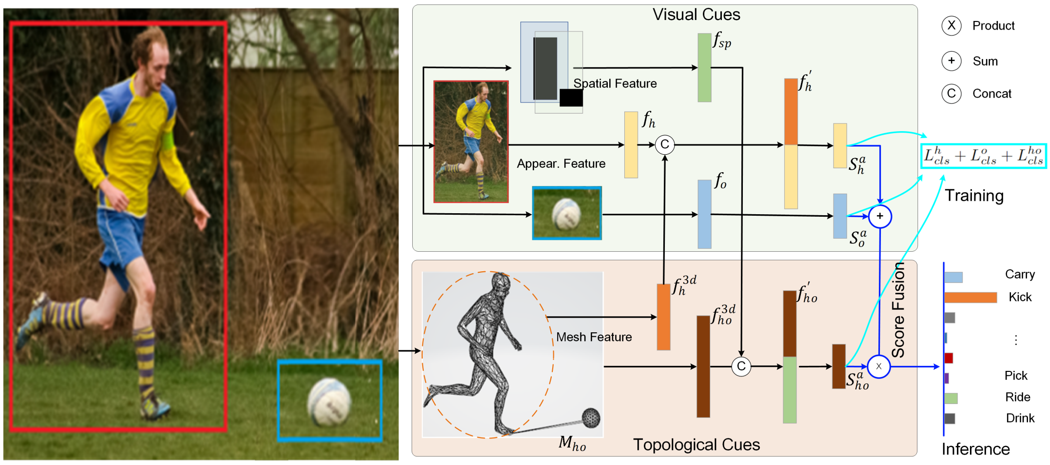


Figure 4. Framework of HOME. On the whole, our network contains two types of clues: 2D visual cues, including human and object appearance features (f_h,f_o), and spatial feature f_{sp}; topological cues, including human topological feature f_h^{3d} and human–object topological feature f_{ho}^{3d}. We fuse the corresponding visual clues f_h and topological feature f_h^{3d} as f_h', and fuse f_{sp} and $f_h^{3d}o$ as f_{ho}'. Finally, the network outputs the scores of three feature branches as S_H^a, S_O^a and S_{sp}^a, which are utilized for training and inference.

3.2.1. Visual Cues

Appearance Feature Extraction. In this section, we describe how to extract appearance features from instance human and object, respectively. After the detection step, we first extract the global feature of the entire image, and then, we use ROI Pooling to extract the instance features of human and objects, being noted as f_h and f_o, from the global feature heat map.

2D Spatial Feature Extraction. We use a double-channel binary map to represent the 2D spatial relation of a human and an object. In the channel for the human, the value is set to 1 in the location that is in the human body bounding box while it is set to 0 in other areas. In addition, in the channel for the object, the value is set to 1 in the region of object while it is set to 0 in other places. The double-channel map is fed into a convolutional neural network to extract the 2D spatial feature f_{sp}.

3.2.2. Topological Cues

Apart for the visual cues, the 3D spatial and topological information are also very import for interaction classification. Here, edge-based convolution is adopted to extract topological cues from HOM that represent the invariant features of the interaction relationship.

Bottom-up feature extraction. As the interaction semantics is related to body shape, pose, object orientation, and size, the topological cues locate at the low-dimensional manifold of the HOM space. So, the information extraction is carried out in a bottom-up manner. For each image I, HOM models are built for all detected persons and objects, obtaining $\{M_i^{3d}\}$. To feed it into MeshCNN, HOM should be translated to an edge-based feature in size of $n_e \times 5$, where n_e is the edge number of the mesh, while 5 means the number of features corresponding to a central edge. These features cover the angle between adjacent triangle faces, the respective vertex angle of the two adjacent triangles, and the ratio of edge length to height of the two adjacent triangles.

In particular, two branches of MeshCNN networks are used to extract the human feature f_h^{3d} and human–object feature f_{ho}^{3d}, respectively. Each network consists of four residual blocks, each of which contains three consecutive Mesh convolution operations, followed by a Mesh pooling, as shown in Figure 5. At the early stage of pooling, it captures the microscopic topology feature. At later stages, HOM topological information is simplified toward macroscopic semantic-related interaction. Figure 6 shows the meshes with different resolution after mesh pooling.

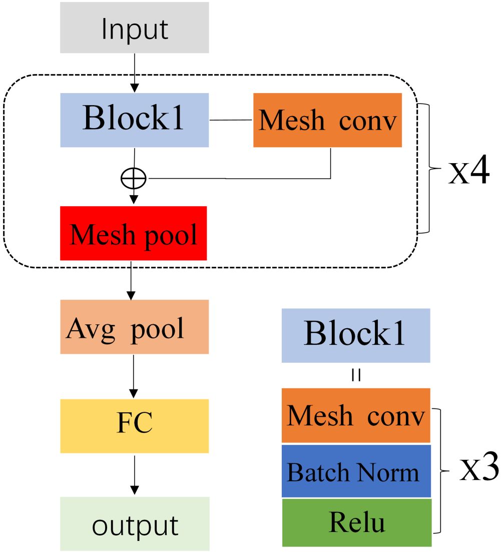


Figure 5. The network for perceiving the human mesh and HOM. The network consists of four residual blocks, each of which contains three consecutive Mesh convolution operations, followed by a Mesh pooling. Then, the network uses average pooling and fully connected network (FC) before outputting the final topological feature.

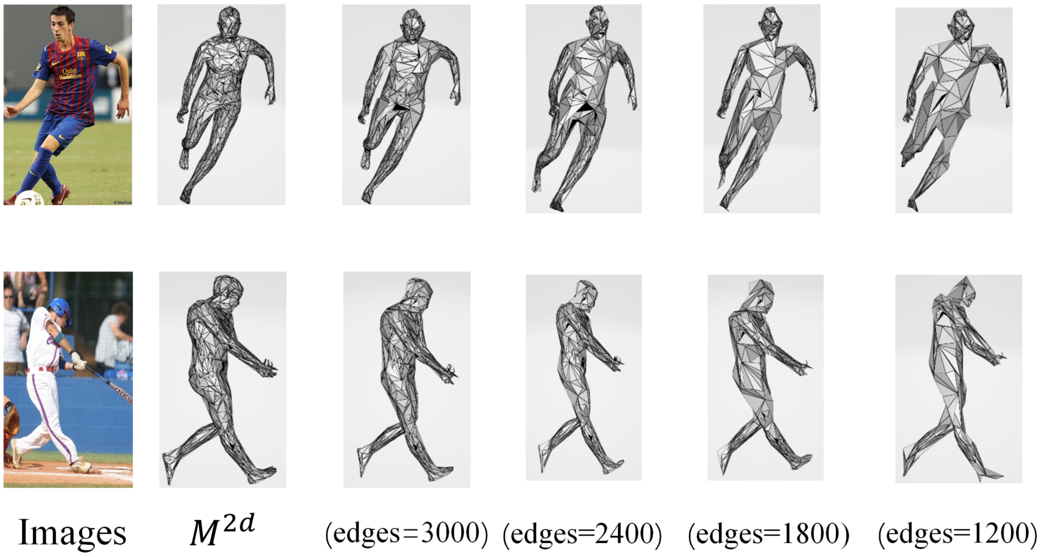


Figure 6. The human body meshes in different resolutions by downsampling of the mesh convolution. According to bottom-up feature extraction, the pooling operation simplifies the resolution from high to low, but the topological feature is maintained and encoded into the nodes, which reflects the invariant feature of different actions.

3.3. Topology-Enhanced Fusion

Through the above work, we obtain the visual cue features f_h, f_o and f_{sp}, and the topological cue features f_h^{3d}, f_{ho}^{3d}. We supervise interaction detection on the three branches of human, object, and space, respectively, but using topological cues to enhance them. In detail, we concatenate the human feature f_h with topological feature f_h^{3d}, and concatenate the spatial feature with topological feature f_{ho}^{3d} .

$$f_h' = \text{Concat}(f_h, f_h^{3d}) \tag{1}$$

$$f_{ho}' = \text{Concat}(f_{sp}, f_{ho}^{3d}) \tag{2}$$

At last, we obtain the enhanced human features f_h' and human–object feature f_{ho}'. In particular, the dimension of f_h, f_o, f_h^{3d} and f_{sp} is 1024, and Enhanced f_h' and f_{sp}' are of dimension 2048. Finally, f_h', f_{ho}', and f_o are fed into two fully connected layers and one sigmoid function followed, respectively, to obtain the corresponding confidence S_h, S_{ho}, and S_o for the classification of interaction.

3.4. Training and Inference

Loss for Training. Since HOI detection is a multi-label classification task, we choose the binary cross-entropy loss function during the training stage. The loss terms corresponding to human, object, and spatial map are L_{cls}^h, L_{cls}^o, and L_{cls}^{sp}, respectively. Finally, the total loss L_{total} for training detection is:

$$L_{total} = L_{cls}^h + L_{cls}^o + L_{cls}^{ho} \tag{3}$$

Inference. For a given input image, the final interaction classification is computed based on scores S_H^a, S_O^a, and S_{sp}^a from different branches. The score for inference is formulated as:

$$S_{HOI} = s_h * s_o * (S_h^a + S_o^a) * S_{ho}^a \tag{4}$$

where $<s_h, s_o>$ is the confidence pair of human and object from the detection result, indicating the probabilities of interactive subjects.

4. Experimental Results

In this section, we first describe the details of the implementation method. Then, we introduce the dataset and metrics adopted for the experiments. Finally, we evaluate the method by comparing with state-of-art methods and ablation studies.

4.1. Implementation Details

For a fair comparison, we use unified detection results, i.e., same bounding boxes and category prediction, as in ICAN [11]. Based on the Faster-RCNN detector with ResNet-50 [38] and FPN [39], the bounding boxes of persons and objects are predicted with score S_h and S_o, respectively. We only retain detection boxes, satisfying $S_h \leq 0.6$ and $S_o \leq 0.4$. To compute the visual cues, the feature maps of human and objects are scaled to a fixed size 7 × 7 for extracting the appearance features, and the feature maps to a fixed size 64 × 64 for extracting the spatial features. To compute the topological cues, the MeshCNN is pre-trained based on reconstructed HOM data.

Our deep learning model is run on Pytorch. We train the model on a NVIDIA 2080Ti GPU, and the initial learning rate for each branch is set to 0.0025 with optimizer SGD. The dropout rate in the layer before the last fully connected layer is set to 0.5. The gamma and weight delay are set to 0.1 and 0.0001, respectively. The batch size is set to 16. The algorithm converges after about the total number of 600 K iterations.

4.2. Dataset and Metric

Dataset. We adopt the widely-used HOI benchmark HICO-DET [10] to validate the effectiveness of HOME. HICO-DET is an instance-level benchmark consisting of 47,776 images (38,118 for training and 9658 for testing) and 600 HOI categories. It contains 80 object categories from COCO [40] 117 verbs and more than 150 thousand annotated HOI pairs. Based on previous work [11,16,22], we use 80% of the datasets as the training set, and the other 20% as validation. We reconstruct the meshes of humans and objects, and build HOM models for each pair of human and object. Particularly, the MeshCNN is pre-trained on the HOM training set before end-to-end training for fast convergence.

Evaluation Metrics. To evaluate the performance of the methods, we adopt the commonly used mAP (mean average precision) as in previous works [11,12,19,21]. Predicting is valid when it satisfies: (i) the predicted bounding boxes locate people and objects with IOU ≤ 0.5, and (ii) the interaction is classified correctly.

4.3. Quantitative Evaluation

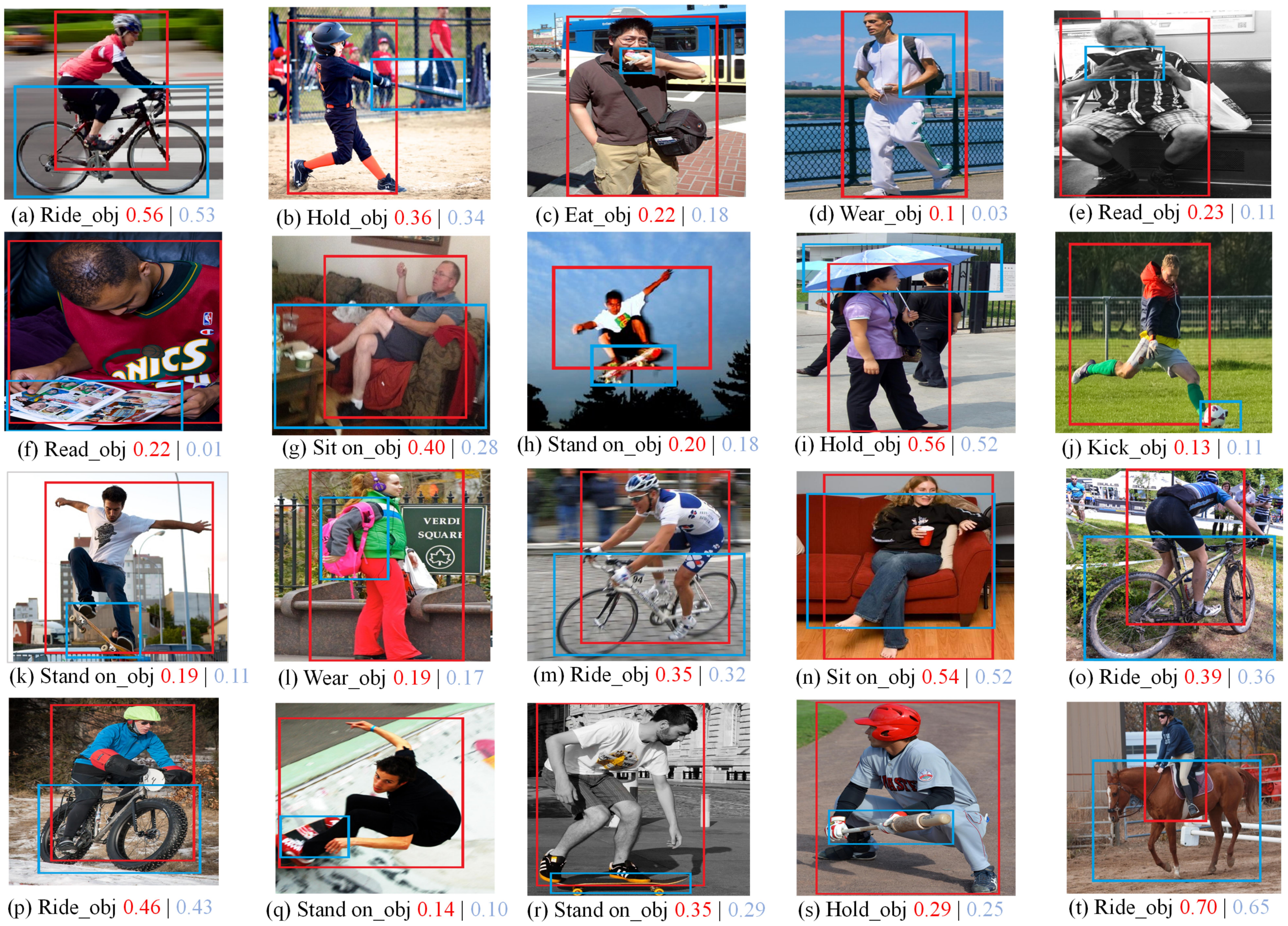
In order to explore the impact of human spatial topological information on detection results, we conduct experiments and report the mAP on Full, Rare, and Non-Rare parts of the HCIO-DET dataset, respectively. The comparison is mainly made through two types of interaction frameworks: (1) the interaction detection framework that only uses 2D images information, (2) 3D topology information is integrated into the interaction detection module of 2D image information.

Performance. For the datasets, we select several classical algorithms for quantitative comparison. Table 1 shows the results of our method compared with other methods [10,11,13,14,19,22,41]. In order to make a fair comparison, we adopt the same object detection results as DRG [22], ICAN [11], and PMFNet [14]. Unlike other approaches, GPNN [10] utilizes additional knowledge graphs to detect human–object pair interactions. We adopt the same backbone network ResNet50-FPN for visual feature extraction as in PMFNet [14] and DRG [22]. Among them, PMFNet uses human pose information to amplify local areas of the human body to obtain fine-grained information. DRG [22] made use of the heterogeneity of nodes to construct two sparse subgraphs centered on people and objects. The 2D-baseline method is a pruned DRG [22] that excludes two sparse subgraphs. HOME* is a HOME plus version that fusing HOM topological feature into DGR method by referring to the HOME framework. We can see that the HOME method shows the improvement of 0.37, 0.71, 0.33 mAP on Full, Rare, Non-Rare in contract with the 2d-baseline, and that HOME* shows 0.26, 0.32, 0.21 improvement in contrast to DGR. Both HOME and HOME* achieve state-of-art performance, validating the significance of the HOM topological cue to interaction recognition.

Table 1. Comparison of results in HICO-DET.

Method	Visual Feature Backbone	Full	Rare	Non-Rare
HORCNN [19]	CaffeNet	7.81	5.37	8.54
InteractNet [13]	ResNet50-FPN	9.94	7.16	10.77
GPNN [10]	ResNet101	13.11	9.34	14.23
ICAN [11]	ResNet50	14.84	10.45	16.15
No-Frills [41]	ResNet152	17.18	12.17	18.68
PMFNet [14]	ResNet50-FPN	17.46	15.65	18.00
DRG [22]	ResNet50-FPN	19.26	17.74	19.71
2d-baseline	ResNet50-FPN	18.78	16.52	19.32
HOME	ResNet50-FPN	19.15	17.23	19.66
HOME*	ResNet50-FPN	19.52	18.06	19.92

Quantitative results. We compute the predicted interaction scores and visualize them in Figure 7. Figure 7 shows the quantitative comparison between the 2d-baseline method and HOME. On the whole, HOME predicts a higher score for correct category than the 2d-baseline method.



(a) Ride_obj 0.56 | 0.53 (b) Hold_obj 0.36 | 0.34 (c) Eat_obj 0.22 | 0.18 (d) Wear_obj 0.1 | 0.03 (e) Read_obj 0.23 | 0.11

(f) Read_obj 0.22 | 0.01 (g) Sit on_obj 0.40 | 0.28 (h) Stand on_obj 0.20 | 0.18 (i) Hold_obj 0.56 | 0.52 (j) Kick_obj 0.13 | 0.11

(k) Stand on_obj 0.19 | 0.11 (l) Wear_obj 0.19 | 0.17 (m) Ride_obj 0.35 | 0.32 (n) Sit on_obj 0.54 | 0.52 (o) Ride_obj 0.39 | 0.36

(p) Ride_obj 0.46 | 0.43 (q) Stand on_obj 0.14 | 0.10 (r) Stand on_obj 0.35 | 0.29 (s) Hold_obj 0.29 | 0.25 (t) Ride_obj 0.70 | 0.65

Figure 7. Qualitative results. Blue labels score value output by the 2d-baseline model and red labels that of HOME. The prediction results and the correct action annotations are shown for the human–object pair located by the bounding boxes.

To further validate the effectiveness for different categories, we compute the mean of mAP values for each type of interaction, and report them in Figure 8. The results also show that topology information improves the performance in almost all categories.

4.4. Ablation Studies

In order to study the influence of different modules of the network model on the detection results, we completed a series of ablation experiments.

Multiple branches. To validate the effectiveness of multi-branch fusion, we ablate spatial features, topological features, and both of them for comparison. For the first, we delete the 2D spatial feature by extracting a branch during inference. For the second, we delete the 3D human–object topological cue branch. In addition, for the third, we delete both the spatial feature branch and the topological cue branch for testing.

In Table 2, we can observe that the detection results are reduced by 3.29, 4.17, and 2.96 mAP, respectively, if not utilizing both 2D spatial features and HOM topological features. In addition, the results are reduced by 0.37, 0.71, and 0.34 mAP, respectively, if not using HOM topological features. Moreover, without 2D spatial features, the detection results are reduced by 2.96, 2.99, and 2.79 mAP, respectively, in contrast with HOME. The above results validate the effectiveness of our multi-branch fusion.

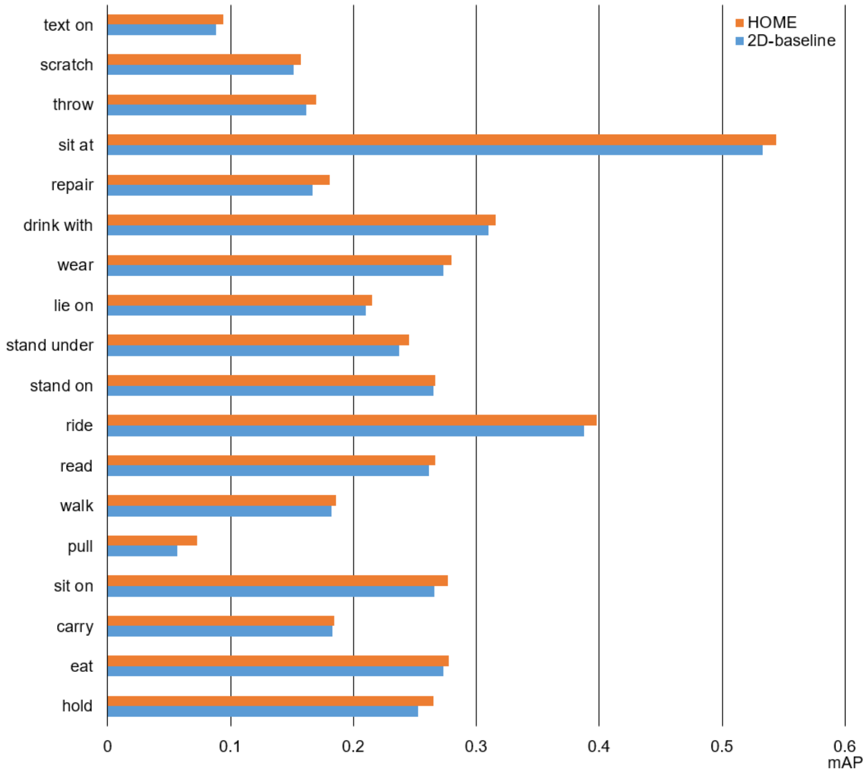


Figure 8. Performance (Mean of mAP values) comparison on different actions between our method and 2d-baseline on HICO-DET.

Table 2. Effects of different branches (mAP).

Method	Full	Rare	Non-Rare
HOME	19.15	17.23	19.66
w/o spatial & Topology	15.86	13.06	16.7
w/o spatial	16.19	14.24	16.87
w/o Topology	18.78	16.52	19.32

Fusion mode with HOM. We test two fusion methods: early fusion and late fusion of 2D human visual features and 3D human–object topological features. Early fusion firstly fuses two feature vectors and then sends them into the interaction detection. Late fusion is to fuse the detection results after obtaining the results from different branches. The evaluation results are reported in Table 3. We can see that both fusion modes achieve improvement in contrast with the 2D-baseline, and that the early fusion performs the best. The late fusion shows worse improvement than the early one due to the fact that human–object topologies of some different interactive actions may be similar and may result in ambiguity. To some extent, the early fusion method, which we used in HOME, complements appearance features with human–object topological features in a coarse-to-fine manner, showing its better generalization.

Table 3. Performance (mAP) comparison of different fusion modes.

Method	Full	Rare	Non-Rare
2D-baseline	18.78	16.52	19.32
HOME-late fusion	18.96	17.03	19.57
HOME-early fusion	19.15	17.23	19.66

Failure Case. Although HOME assists rough visual features and improves the judgment of interactive behavior by utilizing 3D human–object topology information, there also exists some specific scenes that could not be handled, as shown in Figure 9. Firstly, in the multi human–object scene, if two people are next to each other, the category may be fused since they have similar spatial relation to the object, as shown in Figure 9a,b. Secondly, due to lacking temporal information, it is difficult to understand if both of the two persons are washing one car, resulting in missed detection, as shown in Figure 9c. In addition, the object detector fails to detect objects that occluded by body, e.g., the phone cannot be detected in Figure 9d.

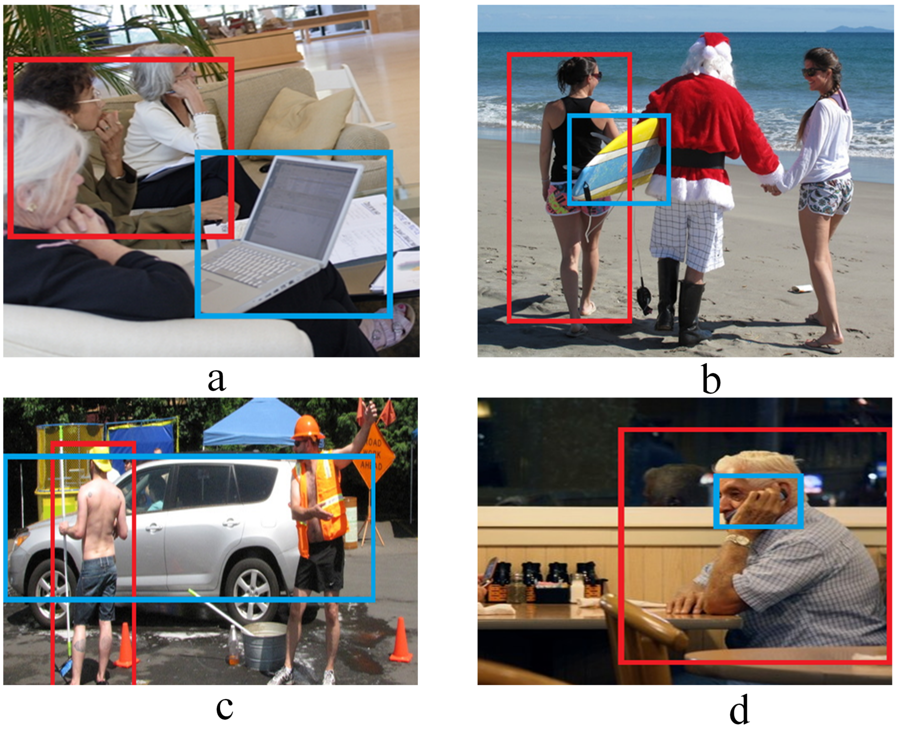


Figure 9. The failure cases: (**a**,**b**) not matching human and object when the people are next to each other under interaction; (**c**) missing detection when two persons are washing one car; (**d**) missing detection when the phone is covered by the human body.

5. Discussion

The advantage of the using the human–object mesh is that it suppresses interference of self-occlusion and external occlusion brought by pose. The improvement is obvious when the action is strongly related to the human posture. Although the topological cue could be used to cover the shortage of visual features for HOI recognition and shows improvements in most scenes, there still exists special action that is not easy to handle. For example, the fusion may be confused when the object is underfoot. The topological feature may be redundancy or unnecessary for "stand on" because the relation "on" is weakly relative to the human pose, resulting in an inconspicuous improvement on the verb "stand on", as shown in Figure 8. In addition, the method is limited by results of the object detector. When detection results are wrong or unfaithful, incorrect mesh reconstruction must have a negative impact on the recognition. At last, the inference time of our method is about 3fps when running on RTX 2080Ti GPU, which is a little time-consuming.

6. Conclusions

We propose a 3D human–object mesh topology enhancement method (HOME) for HOI recognition. In the method, topological cues of HOM integrated with human geometry and the 3D human–object interaction relation can be fused to the visual cues to enhance the HOI recognition. The two key contributions are: we provide the first perspective that human–object interaction is derived from HOM geometric topology; and the interaction relation is dug by extracting edge features of HOM from bottom to top, so as to construct invariant topological features for significant enhancement. The experiments validate that the topology-fused method greatly promotes HOI discrimination performance. However, there still exists a weak point: HOME relies on reconstruction of human–object meshes and MeshCNN perception, which is a little costly on storing edges and edge-based convolution. In the future, we will try using the geometry of a topology-guided graph network to accept visual feature, such that HOI recognition could be processed efficiently. Moreover, we hope that the idea of geometry topology could inspire more interesting work in the area of HOI recognition in the future.

Author Contributions: Conceptualization, W.P. and M.F.; methodology, W.P.; software, C.L. and X.L.; validation, C.L. and X.L.; formal analysis, W.P. and C.L.; investigation, W.P. and C.L.; resources, X.L.; data curation, X.L.; writing—original draft preparation, W.P. and K.T.; writing—review and editing, K.T. and M.F.; visualization, C.L.; supervision, M.F. and W.P.; project administration, X.L.; funding acquisition, W.P., K.T., X.L. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China (62072126, 62102105), the Science and Technology Program of Guangzhou (202201020229, 202002030263, 202102010419), the Open Project Program of the State Key Lab of CAD&CG (Grant No. A2218), Zhejiang University, Research start-up fund of Ningbo University of Technology, the Major Project of Science and Technology Innovation 2025 (Grant No. 2020Z062), the Guangdong Basic and Applied Basic Research Foundation (2020A1515110997, 2022A1515010138, 2022A1515011501).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Any data used to support the findings of this study are from previously reported studies and datasets, which have been cited.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DRG	Dual relation graph
GPNN	Graph parsing neural networks
HICO-DET	Human Interacting with Common objects for the HOI Detection Task
HOI	Human–Object Interaction
HOM	Human–Object Mesh
HOME	Human–Object Mesh Topology Enhanced Interaction Recognition Method
HORCNN	Human Object Region-based CNN
ICAN	Instance-centric attention network
InteractNet	Refer to the method in the work *Detecting and Recognizing Human–Object Interactions* [13]
IOU	Intersection over Union
mAP	mean average precision
No-Frills	Refer to the no-frills approach for HOI Detection [10]
PMFNet	Pose-aware Multi-level Feature Network

References

1. Yu, Y.; Ko, H.; Choi, J.; Kim, G. End-to-end concept word detection for video captioning, retrieval, and question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3165–3173.
2. Yu, Y.; Kim, J.; Kim, G. A joint sequence fusion model for video question answering and retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 471–487.

3. Dzabraev, M.; Kalashnikov, M.; Komkov, S.; Petiushko, A. Mdmmt: Multidomain multimodal transformer for video retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3354–3363.
4. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
5. Liu, K.; Liu, W.; Gan, C.; Tan, M.; Ma, H. T-C3D: Temporal convolutional 3D network for real-time action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
6. Yu, S.; Cheng, Y.; Xie, L.; Luo, Z.; Huang, M.; Li, S. A novel recurrent hybrid network for feature fusion in action recognition. *J. Vis. Commun. Image Represent.* **2017**, *49*, 192–203. [CrossRef]
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28, Montreal, QC, Canada, 7–12 December 2015.
8. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Qi, S.; Wang, W.; Jia, B.; Shen, J.; Zhu, S.C. Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 401–417.
11. Gao, C.; Zou, Y.; Huang, J.B. ican: Instance-centric attention network for human-object interaction detection. *arXiv* **2018**, arXiv:1808.10437.
12. Li, Y.L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.S.; Wang, Y.; Lu, C. Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3585–3594.
13. Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8359–8367.
14. Wan, B.; Zhou, D.; Liu, Y.; Li, R.; He, X. Pose-aware multi-level feature network for human object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9469–9478.
15. Li, Y.L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.S.; Ma, Z.; Chen, M.; Lu, C. Pastanet: Toward human activity knowledge engine. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 382–391.
16. Ulutan, O.; Iftekhar, A.; Manjunath, B.S. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13617–13626.
17. Hanocka, R.; Hertz, A.; Fish, N.; Giryes, R.; Fleishman, S.; Cohen-Or, D. Meshcnn: A network with an edge. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
18. Gupta, S.; Malik, J. Visual semantic role labeling. *arXiv* **2015**, arXiv:1505.04474.
19. Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 381–389.
20. Girdhar, R.; Ramanan, D. Attentional pooling for action recognition. In Proceedings of the Advances in Neural Information Processing Systems, 30, Long Beach, CA, USA, 4–9 December 2017.
21. Fang, H.S.; Cao, J.; Tai, Y.W.; Lu, C. Pairwise body-part attention for recognizing human-object interactions. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 51–67.
22. Gao, C.; Xu, J.; Zou, Y.; Huang, J.B. Drg: Dual relation graph for human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 696–712.
23. Zhong, X.; Ding, C.; Qu, X.; Tao, D. Polysemy deciphering network for robust human–object interaction detection. *Int. J. Comput. Vis.* **2021**, *129*, 1910–1929. [CrossRef]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Wang, H.; Zheng, W.s.; Yingbiao, L. Contextual heterogeneous graph network for human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 248–264.
26. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, 30, Long Beach, CA, USA, 4–9 December 2017.
27. Zhang, M.; Wang, Y.; Kadam, P.; Liu, S.; Kuo, C.C.J. Pointhop++: A lightweight learning model on point sets for 3d classification. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 3319–3323.
28. Zhang, M.; You, H.; Kadam, P.; Liu, S.; Kuo, C.C.J. Pointhop: An explainable machine learning method for point cloud classification. *IEEE Trans. Multimed.* **2020**, *22*, 1744–1755. [CrossRef]
29. Jiang, M.; Wu, Y.; Zhao, T.; Zhao, Z.; Lu, C. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. *arXiv* **2018**, arXiv:1807.00652.
30. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.

31. Qi, C.R.; Litany, O.; He, K.; Guibas, L.J. Deep hough voting for 3d object detection in point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9277–9286.
32. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27–28 October 2019; pp. 770–779.
33. Bogo, F.; Romero, J.; Loper, M.; Black, M.J. FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014; pp. 3794–3801.
34. Bucki, M.; Lobos, C.; Payan, Y. A fast and robust patient specific finite element mesh registration technique: Application to 60 clinical cases. *Med. Image Anal.* **2010**, *14*, 303–317. [CrossRef] [PubMed]
35. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
36. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27–28 October 2019; pp. 10975–10985.
37. Li, Y.L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; Lu, C. Detailed 2d-3d joint representation for human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10166–10175.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
41. Gupta, T.; Schwing, A.; Hoiem, D. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9677–9685.