Geometry-Informed Graph Neural Networks for Multi-Person Human-Object Interaction Recognition in Videos

Tanqiu Qiao

A thesis presented for the degree of
Doctor of Philosophy at Durham University

Department of Computer Science
Durham University
United Kingdom
February 17, 2025

Abstract

Human-Object Interaction (HOI) recognition in videos is a fundamental task in computer vision with wide-ranging applications, including robotics, surveillance, and autonomous systems. Accurately modeling the complex interactions between multiple humans and objects in dynamic environments is crucial for developing intelligent systems that can understand and recognize human behavior.

HOI recognition in multi-person scenarios presents unique challenges that surpass traditional action recognition and single-person HOI tasks. With multiple individuals interacting simultaneously with various objects, complexities such as occlusions and overlapping interactions become prevalent. Video-based analysis is crucial, as static images fail to capture the temporal dynamics necessary for understanding these interactions. To tackle these challenges, integrating geometric cues like human poses and object keypoints with visual features such as appearance and motion is essential. Geometric understanding is inherently more robust to occlusions and can provide additional spatial information that visual features alone may miss. The primary aim of this research is to develop a robust and accurate multi-person HOI recognition framework that effectively fuses geometric and visual features, addressing these complexities through three objectives: (1) designing advanced multimodal feature fusion methods, (2) collecting comprehensive multi-person HOI datasets, and (3) creating a generalizable framework suited for diverse scenarios.

The motivation behind this research direction stems from the limitations of current visual-based approaches, which often fail to generalize in complex real-world scenarios. Extracting geometric is inspired by skeleton-based action recognition, as they are less affected by challenges like partial occlusions. Effective fusion of geometric and visual features is critical for creating a holistic representation that enhances the model's understanding of interactions. Additionally, the success of this framework hinges on the availability of high-quality datasets that reflect the diversity of real-world MPHOI situations. Therefore, we also collect multi-person HOI datasets that not only aid in training and validating the proposed model but also contribute to the broader research community. This comprehensive approach ensures that our framework is well-equipped to handle the intricate nature of MPHOI recognition in dynamic video environments.

This research introduces a series of novel frameworks designed to enhance the robustness and accuracy of multi-person HOI recognition in videos. We start with the Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN), the first attempt to complement visual features with geometric features learned from geometric understanding via graph-based deep learning methods. We also introduce MPHOI-72, a novel two-person HOI dataset specifically designed to evaluate the effectiveness of 2G-GCN in

multi-person HOI scenarios, thereby advancing the field from single-person to multi-person HOI recognition.

Building on the insight from 2G-GCN that the geometric cues offer extensive complementary information, the need for a more effective fusion of geometric and visual features is identified. We propose the CATS framework to advance HOI recognition from category-level to scenery-level understanding. This framework fuses geometric and visual features for each human and object category, and subsequently constructs a scenery interactive graph to learn the relationships among these categories, providing a more structured and comprehensive understanding of the interactions within a scene.

Recognizing the need for further improvements in multimodal feature fusion and dynamic interaction modeling, we propose the Geometric Visual Fusion Graph Neural Networks (GeoVis-GNN). It further refines the fusion of geometric and visual features at the entity level via a dual-attention mechanism and enhances HOI modeling by an interdependent entity graph. To better represent realistic multi-person HOI scenarios, we introduce MPHOI-120, a challenging dataset collecting three-person HOI activities with frequent occlusions and exponentially increasing interaction complexity.

We validate the effectiveness of our methods through extensive experiments and qualitative analysis, demonstrating that our approaches outperform state-of-the-art techniques in HOI recognition across both multi-person and single-person scenarios in videos.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

In the pursuit of knowledge, supervisors lead,
Through tangled paths, they plant the seed.
Prof Hubert, with patience, cleared the haze,
Guiding my steps through complex maze.

To labmates, friends, who shared long nights,
With coffee, code, and hopeful sights.
Each late-night laugh, each shared despair,
Your strength in numbers, ever rare.

My family stood through silent years,
Unseen, yet felt, through doubts and fears.
To you, the roots beneath my ground,
In all success, your love is found.

This journey ends, yet you remain,
My compass through the endless plane.

In my PhD journey at Durham, where late nights are plentiful and bubble tea is often in short supply, there exists a guide, Prof Hubert Shum, my supervisor. Like a master pâtissier transforming simple ingredients into something extraordinary, he finds the hidden potential in my early, unrefined efforts, seeing value amidst the mess. Hubert is not just the architect of my academic growth; he is the steady hand that keeps my project from sinking, providing vital support and resources whenever I feel overwhelmed. For his endless patience and guidance, my gratitude is as rich and layered as the finest dessert, a trust no words can fully express.

My deep thanks also go to Dr Frederick Li, my co-supervisor, the perfect complement in this recipe of my journey, who skillfully enhances the foundation that Hubert has crafted. Frederick refines my drafts and ideas, guiding me with wisdom through the layered complexity of academia. His advice is the perfect recipe card, helping me turn complex challenges into simple, manageable steps.

To my dear friends in our group—Ruochen Li, Shuang Chen, Haozheng Zhang, Manli Zhu, Xiatian Zhang, Ziyi Chang, Xiaotang Zhang—and my favorite food scientist, Mingze Hou: you are the essential ingredients, the cherries and chocolate chips that make this

journey truly flavorful. You bring zest and laughter to my life, adding the sweet distractions that keep me from burning out in the endless mix of deadlines. You are the decorations on my cake, bringing happiness and joy, making sure I do not lose myself to the rigorous layers of academia.

To my research and life partner, Luis Li, my lovest Tiramisu and my companion through every step of this journey: you are the calm that soothes my anxious moments, the voice of reassurance when I doubt, the gentle reminder of stability and love, my truest confidant and most valued partner.

To my parents, the foundation and compass of my life—you are the roots that keep me grounded. Your unwavering love and support are the forces that have brought me this far, even when I am far away pursuing my dreams abroad. The sacrifices you have made and the wisdom you shared have guided me through each challenge and new chapter. I carry forward the strength you've given me, and I am grateful beyond measure for everything you've done. For all that you are, I am in perpetual gratitude.

This journey ends, and from here I can go anywhere I wish to be.

.

Dedication

To my parents.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Human-Object Interaction (HOI) recognition in videos is a crucial area of research in computer vision, where multi-person scenarios introduce unique challenges that go beyond the complexities of traditional action recognition and single-person HOI tasks. In multi-person environments, multiple individuals often interact with various objects simultaneously, leading to overlapping and intertwined actions that are difficult to disentangle. This creates ambiguity in associating actions with entities and interpreting interactions. Furthermore, frequent occlusions complicate the accurate tracking and recognition of interactions, challenges that traditional action recognition models, designed for single individuals or isolated actions, are not equipped to handle.

While image-based HOI detection provides a snapshot of interactions, it lacks the temporal context needed to understand the dynamics and evolution of interactions over time. Video-based HOI recognition captures the continuity and sequence of human actions, allowing for a more comprehensive understanding of human behavior. This temporal dimension is essential for accurately predicting and interpreting complex interactions in real-world scenarios, such as surveillance [1, 2] and human-computer interaction [3, 4].

Real-world HOIs frequently involve multiple interacting individuals and objects,

which introduce occlusions and dynamic complexities. In these situations, methods with visual cues alone [5–9] may be insufficient, as they can struggle to accurately represent interactions when parts of the scene are obscured. Geometric understanding, such as human poses, tends to be more robust to partial occlusions [10–13]. This research explores the application of graph neural networks to effectively incorporate such geometric cues and fuse them with visual data, enabling a more comprehensive understanding of interactions and improving the accuracy of HOI recognition in complex, multi-person scenarios.

1.1 Motivations

Despite significant advancements in HOI recognition, most existing approaches [6, 13–18] have primarily focused on recognizing interactions in static images. While these methods have laid the groundwork for understanding human-object dynamics, they inherently lack the temporal context that video-based analysis provides. Recognizing interactions in videos introduces a temporal dimension that captures the progression and evolution of actions over time, offering a more comprehensive understanding of human behavior. However, this temporal aspect also brings about new challenges, particularly when dealing with real-world HOIs that involve multiple interacting humans and objects. Such scenarios often lead to occlusions and dynamic interaction changes that are difficult to model using traditional visual features alone.

One of the primary motivations for this research is the realization that visual cues, while powerful in providing detailed appearance information (*e.g.*, color, texture, and shape) [5, 8, 9, 19–21], have limitations in handling the complexity of real-world HOIs, especially in crowded or cluttered environments. Visual features can struggle to accurately capture interactions when objects or body parts are partially obscured, leading to potential misinterpretations of the scene. This is particularly problematic in scenarios where multiple individuals are interacting with various objects, as the occlusions and overlaps can cause significant challenges for traditional visual-based recognition methods.

Geometric features, such as human poses and object keypoints, offer a promising

solution to these limitations. Unlike visual features, geometric data is inherently more robust to partial occlusions and can provide a consistent representation of the spatial relationships between humans and objects [10, 11, 13, 22, 23]. Accurately modeling the spatial and functional relationships between humans and objects, such as how a person is positioned relative to a cup they are holding, is crucial for understanding the broader context of the interaction. For example, recognizing that a person is reaching for a cup suggests they are likely about to lift it to drink. This understanding is essential for predicting subsequent actions. This research is motivated by the potential to enhance HOI recognition by integrating geometric features to complement visual data, thereby leveraging the strengths of both modalities.

Graph neural networks have demonstrated significant potential in modeling complex patterns and relationships within data [6–9, 24–28]. However, the integration of geometric and visual features in the context of HOI recognition remains relatively unexplored. This research aims to bridge this gap by proposing a novel framework that fuses these multimodal features, enabling a more comprehensive representation of HOIs. Our approach is designed to enhance the accuracy and robustness of HOI recognition, particularly in challenging scenarios with multiple interacting individuals and objects.

1.2 Research Aims

The primary aim of this research is to advance the field of multi-person HOI recognition in videos by developing a novel framework that effectively integrates geometric and visual features. Recognizing that traditional visual-based approaches often struggle with challenges such as occlusion and dynamic complexities, particularly in scenarios involving multiple interacting individuals and objects, this research seeks to create a more robust and accurate model for HOI recognition. To achieve this, our research focuses on the following objectives:

1. **Integration of Geometric and Visual Features**: This research aims to develop advanced methodologies for fusing geometric features, such as human poses and object keypoints, with visual features like appearance and

motion. By complementing visual data with geometric data, which is inherently more resilient to occlusions, this multimodal approach is designed to improve robustness in occluded and dynamic scenarios.

2. **Collection of Multi-Person Datasets**: Another critical aim of this research is to collect multi-person HOI datasets to enhance the study and understanding of complex interactions involving multiple humans and objects. These datasets will be essential for training and evaluating the proposed framework, and they are expected to contribute significantly to the broader research community by providing a rich resource for future HOI recognition studies.

3. **Development of a Comprehensive HOI Recognition Framework**: The final aim is to create a framework that not only integrates these multimodal features but also ensures that the model can generalize effectively across different scenarios. This involves testing and validating the model on diverse datasets to ensure its applicability to various real-world contexts.

These research objectives are crucial for advancing HOI recognition in videos by addressing limitations of traditional visual-based methods. Integrating geometric and visual features enhances robustness against occlusions and dynamic complexities, while the collection of multi-person datasets supports comprehensive model training and evaluation, contributing valuable resources to the field. Developing a generalizable framework ensures applicability across diverse scenarios, making the model suitable for practical applications such as robotics and surveillance technologies. These aims collectively provide more efficient and reliable solutions for understanding and recognizing HOIs in videos.

1.3 Contributions

This research presents a progressive journey through the challenges and advancements in multi-person HOI recognition, systematically building upon each contribution to address increasingly complex scenarios.

Our journey begins with developing the Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN), as presented in Chapter 3. Recognizing that traditional visual-based methods often struggle with occlusions and dynamic complexities in multi-person HOIs, our initial contribution focuses on improving HOI recognition by learning geometric features from graph-based methods and fusing geometric and visual data in a feature-level graph. To support this research, we introduce the MPHOI-72 dataset, a novel two-person HOI dataset designed to address the challenges of multi-person HOIs and to validate the effectiveness of the 2G-GCN framework in more complex scenarios.

Building on the observations from 2G-GCN, where the integration of geometric features shows promise but also reveals the need for a more meaningful manner to fuse geometric and visual features. In Chapter 4, we propose an end-to-end category to scenery framework, CATS, starting by generating geometric features for various categories through graphs respectively, then fusing them with corresponding visual features. Subsequently, we construct a scenery interactive graph with these enhanced geometric-visual features as nodes to learn the relationships among human and object categories. This methodological advance facilitates a deeper, more structured comprehension of interactions, bridging category-specific insights with broad scenery dynamics.

Finally, recognizing that even with the advancements made in CATS, there remains a challenge in more effectively fusing multimodal features and accurately capturing the intricate dynamics of multi-person interactions in complex environments. We then propose a Geometric Visual Fusion Graph Neural Network (GeoVis-GNN) in Chapter 5. GeoVis-GNN further refines the fusion of geometric and visual features in the entity level via an attention mechanism and enhances HOI modeling by an interdependent entity graph. This work not only improves the accuracy of HOI recognition in challenging scenarios but also introduces the MPHOI-120 dataset, which showcases three-person HOI activities in daily life to evaluate the ability of GeoVis-GNN to handle frequent occlusions and multiple interacting entities.

Together, these contributions form a cohesive narrative of incremental advancements in HOI recognition, each building upon the insights gained from the previous

work, ultimately leading to a more robust, accurate, and context-aware understanding of HOIs in videos. The main contributions of this thesis are summarized as follows:

- A novel Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN) consists of a two-level graph structure that models geometric features between humans and objects, together with the corresponding visual features. Additionally, a novel two-person HOI dataset (MPHOI-72) is proposed to advance the area from single-person HOI to multi-person HOI (Chapter 3).

- A novel end-to-end framework, CATS, ranging from category-level feature fusion to scenery-level graph. It integrates multi-category, multi-modality fusion of visual and graph-based geometric features with an attention-based scenery interactive graph to recognize multi-person HOIs in videos (Chapter 4).

- A novel bottom-up framework, GeoVis-GNN, fuses geometric and visual features at the feature level while learning entity interactions at the entity level. The framework includes a dual-attention feature fusion module to optimize entity-specific representation and an interdependent entity graph to model both explicit and implicit interactions among entities. Additionally, a new challenging dataset (MPHOI-120) collects HOI activities of three humans interacting with multiple objects in daily life with frequent occlusions (Chapter 5).

1.4 Publications

The research conducted for this thesis has resulted in publications or is currently undergoing peer review as outlined below:

- **Qiao, T.**, Men, Q., Li, F. W., Kubotani, Y., Morishima, S., & Shum, H. P., "Geometric Features Informed Multi-person Human-object Interaction Recognition in Videos." In European Conference on Computer Vision (ECCV). Springer, 2022. ..... (Chapter 3)

- **Qiao, T.**, Li, R., Li, F. W., & Shum, H. P., "From Category to Scenery: An End-to-End Framework for Multi-Person Human-Object Interaction Recogni-

tion in Videos." In International Conference on Pattern Recognition (ICPR), 2024. ..... (Chapter 4)

- **Qiao, T.**, Li, R., Li, F. W., Kubotani, Y., Morishima, S., & Shum, H. P., "GeoVis-GNN: Geometric Visual Fusion Graph Neural Networks for Multi-Person Human-Object Interaction Recognition in Videos." 2024. (*Under Review at Expert Systems with Applications*) ..... (Chapter 5)

1.5 Thesis Structure

This thesis is structured to systematically explore the advancements in Multi-Person Human-Object Interaction (MPHOI) recognition using a combination of geometric and visual features. The chapters are organized to guide the reader through the motivations, existing literature, methodologies and findings of this research in a coherent manner.

In Chapter 1, the introduction lays the foundation by discussing the importance of HOI recognition in videos, particularly highlighting the challenges posed by multi-person HOI activities and occlusions. It also outlines the research aims and key contributions of the thesis.

Chapter 2 delves into the foundational research in HOI recognition, covering topics such as the development of HOI datasets as well as methodologies, and the challenges posed by geometry-informed HOI analysis. This chapter also discusses the role of multimodal feature fusion in human activity recognition, and the relevance of these aspects to the contributions of this thesis.

In Chapter 3, the focus shifts to the development of a novel Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN) for multi-person HOI recognition. This chapter introduces the MPHOI-72 dataset, detailing its design and the challenges it addresses. It also presents the experimental setup and results that demonstrate the effectiveness of the proposed model.

Chapter 4 introduces the CATS framework, which extends the scope from category-level feature fusion to scenery-level graph modeling for multi-person HOI recognition. The chapter provides an in-depth explanation of the methodology, including the

multi-category, multi-modality fusion module and the scenery interactive graph, along with a thorough evaluation of the framework.

Chapter 5 presents a Geometric Visual Fusion Graph Neural Network (GeoVis-GNN), which integrates geometric and visual features at the feature level while learning entity interactions at the entity level. This chapter also introduces the MPHOI-120 dataset, which features three humans interacting with multiple objects, and discusses the performance of GeoVis-GNN in various challenging scenarios.

Finally, Chapter 6 summarizes the key contributions of the thesis, reflects on the advancements made in HOI recognition, and outlines potential directions for future research, emphasizing areas where further innovation is required.

CHAPTER 2

Literature Review

HOI recognition in videos encompasses human action analysis [29–31] and skeleton-based activity recognition [32–34] by integrating the detection of human movements and postures with the contextual understanding of interactions between humans and objects, thereby offering a more holistic approach to activity recognition in complex environments.

In this section, we provide a comprehensive review of key areas in HOI research. It begins with HOI datasets (Section 2.1), distinguishing between image-based and video-based datasets to outline the data foundations supporting HOI studies. Next, the HOI recognition task is examined (Section 2.2), covering HOI detection methods in images, and then extending to HOI recognition in videos, where temporal dynamics are central. Section 2.3 explores the utilization of visual and geometric information in human activity tasks, especially HOI and action recognition, emphasizing the use of human skeleton data, and further discusses challenges specific to geometry-informed HOI recognition. Section 2.4 then investigates methods of fusing multimodal features, highlighting the importance of fusing diverse data types for improved HOI understanding. Finally, Section 2.5 illustrates the metrics and evaluation techniques essential for assessing model performance, ensuring robust and reliable outputs in

practical applications. Our contributions are deeply informed by and build upon the discussed literature, underscoring their relevance and impact in the field of HOI recognition.

2.1 HOI Datasets

Object detection, HOI detection, and HOI recognition represent a progressive relationship in understanding visual scenes. Object detection focuses on identifying and localizing individual objects in an image, classifying them into categories like "person" or "bicycle." HOI detection extends this by identifying interactions between humans and objects within a static image, adding relational complexity by recognizing actions such as "holding a cup" or "sitting on a chair," but is limited to spatial cues available in a single frame. HOI recognition in videos further advances this by incorporating temporal information, capturing interactions that evolve over time, such as "picking up a cup" or "putting down a bag," allowing for a more dynamic understanding of human-object interactions.

To explore the current landscape of HOI datasets, we structure it into two main sections. Section 2.1.1 covers image-based HOI datasets, which focus on static relational understanding within single images. Section 2.1.2 then explores video-based HOI datasets, emphasizing temporal dynamics and the ability to capture evolving HOIs across frames. Together, these sections provide a comprehensive overview of resources supporting both spatial and spatio-temporal HOI understanding.

2.1.1 Image-Based HOI Datasets

Datasets have been crucial in driving advancements in computer vision research by providing data for training and evaluation, and inspiring new research directions. ImageNet [35], a large-scale dataset with millions of images, enabled significant progress in the object detection task using deep learning techniques. Other important datasets include PASCAL VOC [36], which focuses on object detection, and SUN [37], which emphasizes scene understanding. MS COCO [38] is designed for object detection and segmentation in natural contexts and provides annotations for objects,

instances, and captions. Visual Genome [39], with its extensive annotations, including objects, attributes, and relationships, enables the study of more complex visual relationships.

HOI detection has gained significant attention as an extension of object detection. Early work focuses on smaller datasets with limited HOI categories, such as PASCAL VOC [36] and Stanford 40 Actions [40]. The HICO [41] dataset, featuring a broad set of 600 HOI categories, addresses this limitation and provides a benchmark for image-level HOI classification. This task focuses on determining whether a particular HOI is present in an image by answering a simple yes/no question for each predefined HOI category. Building on HICO, the HICO-DET [42] dataset introduces instance annotations, making it suitable for the more detailed task of HOI detection. Unlike classification, HOI detection requires localizing interactions by predicting bounding boxes for both the human and the object involved and assigning an interaction label. HICO-DET [42] thus serves as a benchmark for HOI detection, providing over 150,000 annotated instances of human-object pairs across the 600 HOI categories, with an average of 250 instances per category.

The V-COCO [43] dataset also focuses on HOI detection, providing similar annotations but with fewer action categories. It includes a total of 10,346 images containing 16,199 people instances, with each annotated person having binary labels for 26 different actions. The HCVRD [44] dataset, built on Visual Genome [39], offers a large-scale dataset for human-centric visual relationship detection. It features fine-grained labels for human sub-categories, a wider range of predicates compared to previous datasets like HICO [41], and includes zero-shot relationships.

Recent research has increasingly focused on 3D HOI, aiming to identify interactions by reasoning about which body parts are in contact with objects, assessing their proximity to these objects, and considering the context of the surrounding scene [45]. This approach goes beyond traditional 2D analysis by introducing a spatial depth that allows models to better understand the geometry and relative positioning of HOIs [45, 46]. However, contact annotation presents significant challenges, as the areas of contact are occluded in images, requiring reasoning about body parts and scene elements.

The HOT [47] dataset uses 2D contact area heatmaps and associated body-part labels to address full-body human-object contact detection in images. The DAMON [45] dataset provides a large-scale set of paired images and accurate vertex-level 3D contact annotations directly on the 3D SMPL [48] mesh, facilitating research into dense 3D contact estimation in the wild. The 3DIR [46] dataset contains natural HOI images, including object point clouds and dense 3D human contact annotations on the SMPL-H [49] mesh. 3DIR also includes annotations for object affordance and human-object spatial relations. DAMON and 3DIR represent a shift from 2D contact annotations to dense 3D annotations that can support a wider range of applications such as human activity understanding [50, 51], affordance detection [52, 53] and augmented or virtual reality [54, 55].

2.1.2 Video-Based HOI Datasets

Early efforts in video-based action recognition often focus on simple actions performed against uncluttered backgrounds, as exemplified by the KTH [56] action dataset. Datasets like Hollywood2Tubes [57] introduce the challenge of localizing actions within untrimmed videos, utilizing movie clips, but with a limited action vocabulary. The UCF Sports [58] dataset is composed of website videos, further illustrating the early focus on easily accessible online footage, especially of sporting events.

A notable development is the MPII Cooking [59] dataset, specifically designed to investigate fine-grained activities in a controlled kitchen setting. Alongside the Breakfast [60] dataset, it facilitates the analysis of intricate hand-object interactions, including aspects of bimanual manipulation. The 50 Salads [61] dataset also focuses on food preparation, providing annotated video and accelerometer data of individuals preparing salads in the kitchen. It is designed to investigate a range of recognition problems, including cross-subject and intra-subject generalization. Multi-modal recordings of bimanual tasks carried out in a kitchen or workshop environment are included in the KIT Bimanual Actions [62] and Bimanual Manipulation [63] datasets. They helps create algorithms that can comprehend and mimic intricate bimanual manipulations.

The CAD-120 [19] and AVA [64, 65] datasets represent significant advancements in

video-based HOI recognition. CAD-120 [19] introduces complex, multi-step activities involving object interactions in a controlled environment. This allows researchers to investigate high-level activities broken down into human sub-activities and object affordance to examine HOI patterns. The AVA [64, 65] dataset, on the other hand, moves towards densely annotating atomic actions in diverse movie clips. This focus on atomic actions, the fundamental building blocks of human behavior, contrasted with earlier datasets' focus on composite actions. Notably, AVA also includes labels for interactions with objects and other individuals, broadening the scope of HOI recognition research.

The Something-Something [66] dataset takes a unique approach by utilizing crowd-sourced videos, enabling large-scale data collection focused on fundamental physical interactions rather than high-level activities. While not explicitly designed for HOI, it provides valuable insights into human interactions with everyday objects. This dataset is designed to encourage models to develop a deeper understanding of physical concepts and actions rather than relying on superficial cues like object appearance. Importantly, the effectiveness of models trained on this dataset should be assessed based on their ability to generalize to Something-Else [67], meaning objects they have not previously encountered.

Egocentric datasets further broaden the study of HOI. Early egocentric datasets such as BEOID [68], GTEA Gaze+ [69] and ADL [70], capture human natural behavior in daily activities and encompass a broader spectrum of interactions. The EPIC-KITCHENS [71] dataset comprises unscripted cooking activities recorded from a first-person perspective, offering a wealth of data for understanding natural human behavior. Charades-ego [72] features paired third-person and first-person videos, further enhancing the investigation of egocentric HOI. EPIC-KITCHENS-100 [73], an extension of EPIC-KITCHENS [71], contains even denser annotations and introduces a novel focus on unsupervised domain adaptation for HOI recognition in egocentric environments.

The recent Ego-Exo4D [74] dataset is a large-scale, multimodal, multiview video dataset focusing on skilled human activities captured from both egocentric (first-person) and exocentric (third-person) perspectives. The dataset features over 800

participants performing activities like sports, music, dance, and bike repair in diverse real-world settings. It uniquely includes synchronized egocentric and exocentric videos, multichannel audio, eye gaze, 3D point clouds, and multiple language descriptions, including expert commentary. Notably, Ego-Exo4D [74] aims to advance research in understanding and modeling skilled actions, pushing beyond the limitations of existing egocentric datasets focused on daily-life or procedural activities.

HOI4D [75] is also an egocentric HOI dataset but focuses on capturing category-level interactions with a broader range of objects across various indoor environments. It is designed to encourage models to recognize interactions with objects unseen during training, promoting generalization beyond instance-level understanding. HOI4D provides detailed 4D annotations, including panoptic and motion segmentation, 3D hand poses, object poses, and reconstructed meshes, aiming to enable a more comprehensive understanding of the spatial and temporal dynamics of interactions. BEHAVE [76] also strives to capture 3D instance-level interactions in more realistic settings. It provides 3D human and object meshes and explicitly annotates surface contacts between humans and objects, facilitating research on fine-grained interaction modeling, which is not commonly addressed in earlier datasets.

Human activities in real-world settings often involve interactions among multiple people and objects. In this research, we focus on multi-person HOI recognition in videos, specifically aiming to capture the progression and evolution of human sub-activities over time. As the number of entities in a scene increases, the interactions expand exponentially, leading to heavy occlusion and complex dynamic relationships. To effectively understand these interactions, geometric information such as human poses and object bounding boxes are essential to complement visual cues for a more comprehensive interpretation of HOIs.

However, most existing HOI datasets for sub-activity recognition, such as CAD-120 [19] and Bimanual Actions [62], are limited to single-person scenarios. To address this gap, we introduce two novel multi-person HOI datasets (in Chapter 3 and Chapter 5), respectively collecting activities in which two and three people interact with multiple objects in daily life to better reflect real-world HOIs. Both datasets provide framewise annotations of geometric representations and sub-activity labels,

supporting detailed, multi-person HOI analysis.

2.2 HOI Recognition Task

This section provides an overview of key advancements in HOI recognition, divided into two primary areas: HOI Detection in Images (Section 2.2.1) and HOI Recognition in Videos (Section 2.2.2). HOI detection in images focuses on identifying interactions within a single static picture, combining object localization with interaction classification to provide a spatially grounded understanding of HOIs. In contrast, HOI recognition in videos introduces a temporal dimension, capturing interactions as they evolve over time, thus requiring models to handle dynamic, sequential data for improved interaction context. Together, these sections highlight the progression from static image-based detection to temporally-aware recognition in videos, reflecting the expanding scope and depth of research in HOI recognition.

2.2.1 HOI Detection in Images

The HOI detection task in images aims to identify sets of triplets $\langle human, verb, object \rangle$, effectively localizing and classifying the interactions between humans and objects within a given image [17, 77, 78]. This involves not only detecting the presence of humans and objects but also understanding the specific action or verb that connects them, providing a richer understanding of the scene's semantics. Based on network architecture design, existing solutions for HOI detection can be broadly divided into two categories: two-stage and one-stage approaches.

Two-Stage HOI Detection

Two-stage methods [42, 43, 79–82] typically use a pre-trained object detector such as Faster R-CNN [83] and Mask R-CNN [84] to first detect humans and objects in an image. Then in the second stage, a separate network classifies the interaction between each possible human-object pair. Most work [42, 80–82] focuses on the improvement of the second stage, employing a multi-stream architecture with parallel

streams for human, object and pairwise features, which are fused to produce final interaction predictions for each human-object pair.

The human and object streams typically encode visual features from human and object bounding boxes, respectively [80]. In FCMNet [85], object visual features are replaced with word embeddings, as the detailed visual appearance of the object is often less crucial to determining the interaction category. Besides visual features, Bansal *et al.* [86] introduce word embeddings to the human stream for feature augmentation. PDNet [87] incorporates word embeddings across all streams to provide language-guided channel attention and enhanced feature representation. Extensive research has also focused on the pairwise stream, which encodes the relationship between humans and objects. iCAN [80] proposes a two-channel binary image representation to capture spatial relations, while FCMNet [85] presents a fine-grained version based on human parsing to highlight critical cues.

In addition to spatial relationships, Graph Neural Networks (GNNs) [88], particularly Graph Convolutional Networks (GCNs) [24] and Graph Attention Networks (GATs) [89], have been developed to assimilate valuable expressions of graph-structured data. While GCNs assign equal weights to all neighbors of a given node, GATs can assign different weights to nodes within the same neighborhood. DRG [90] models the interaction between humans and objects as edges in a graph, using GCNs to refine the features of these edges and predict interaction classes. RPNN [91] utilizes a relation parsing neural network based on GCNs to model the structured relationships between human and object pairs, enabling it to learn complex interaction patterns. CHG [27] constructs a heterogeneous graph with human and object nodes, employing GATs to learn attention weights for each node and its neighbors, effectively capturing the contextual relationships between entities in the scene.

Auxiliary models, such as human pose features, body-part cues [92], language models [93], and graph models [26], can be easily incorporated into the two-stage pipeline to improve HOI. Notably, Bansal *et al.* [86] and Hou *et al.* [94] introduce feature-level augmentation, which has proven effective for HOI. However, these methods face high complexity and low efficiency due to the sequential, separate nature of the two-stage architecture.

One-Stage HOI Detection

One-stage methods [77, 78, 95–101], on the other hand, usually perform object detection and interaction prediction in parallel. In the absence of explicit object locations, these methods rely on predefined interaction areas for interaction prediction, which can achieve faster inference times but may require heuristic-based post-processing steps to associate interactions with object pairs. Depending on the definition of interaction area employed, existing approaches can be classified into (i) point-based methods and (ii) union region-based methods.

Point-based methods such as PPDM [77] and IPNet [98] treat HOI as a point detection task, using a one-stage method to directly detect interactions through a new definition of interaction points. Moreover, PPDM predicts object detection and HOI detection in a unified CenterNet-based framework [96]. GGNet [78] infers a set of action-aware points via glance and gaze steps to address the semantic ambiguity problem of predefined interaction areas.

Union region-based methods such as UnionDet [97] introduces a novel union-level detector to directly detect the interaction region, enabling real-time performance by eliminating the need for post-processing grouping steps common in other one-stage methods. DIRV [99] focuses on densely sampled interaction regions to directly predict interactions and their corresponding human-object pairs, and uses a voting strategy to achieve accurate and efficient HOI detection. However, point-based and union region-based methods struggle in scenarios where the interacting human and object are distant or when multiple, overlapping interactions occur within the same scene, such as in crowded environments [102, 103].

Recently, a new trend of one-stage approaches utilizing Transformer architectures [104–106] has emerged to address these challenges and enhance HOI detection performance. These Transformer-based models use query-driven mechanisms to focus on relevant regions and directly predict interaction classes, eliminating hand-crafted post-processing like non-maximum suppression. This shift is inspired by the success of DETR [104], a Transformer-based object detection model that demonstrates the effectiveness of set prediction for object detection tasks.

A notable advantage of Transformers in HOI detection is their ability to model the

intricate relationships between humans, objects and their interactions through self-attention mechanisms [103, 107]. This capability facilitates a deeper understanding of the scene and allows for a more holistic prediction of HOI triplets. For instance, QPIC [103], one of the first Transformer-based HOI detectors, utilizes a query-based approach to aggregate image-wide contextual information, enhancing the representation of each HOI instance.

Several subsequent works have expanded on this foundation, exploring different architectural designs and incorporating additional information to improve performance. AS-Net [102] reformulates HOI detection as an adaptive set prediction problem, employing parallel instance and interaction branches to dynamically focus on relevant image regions. HOTR [107], on the other hand, concentrates on streamlining the HOI pipeline by directly predicting a set of HOI triplets using an encoder-decoder architecture, thereby eliminating the need for post-processing steps.

Despite their success, initial Transformer-based HOI detectors are often restricted to single-scale feature maps, limiting their ability to capture interactions across varying object sizes and distances. MSTR [108] addresses this limitation by introducing multi-scale processing through HOI-aware deformable attentions, which enables the model to selectively sample features at different resolutions based on each HOI query. This multi-scale approach allows MSTR to capture finer details and improve performance.

Furthermore, recent studies have investigated leveraging semantic information to enhance the representation learning capabilities of Transformer-based HOI detectors. SSRT [101] incorporates a support feature generator to create object-action prediction candidates and uses these candidates to generate spatial and semantic features, refining the model's understanding of the scene. CATN [18] focuses on enriching the object query with category-aware semantic information, thereby improving the initialization of the model and leading to enhanced performance. In addition to semantic information, GeoHOI [13] exploits fine-grained geometric features, such as keypoint positions of humans and objects, which can also provide valuable information for HOI detection, particularly in cases of occlusion.

Another direction of research focuses on disentangling different aspects of HOI

prediction within the Transformer architecture. This approach aims to improve the model's ability to handle complex scenarios and address challenges related to mis-grouping and the independent nature of decoding sub-tasks. For instance, the Disentangled Transformer (DisTR) [109] separates the triplet prediction into human-object pair detection and interaction classification through distinct instance and interaction streams. This separation allows for specialized processing and enhances the model's capacity to capture the compositional nature of HOI.

2.2.2 HOI Recognition in Videos

Recognizing HOI in videos presents unique challenges compared to image-based detection due to the temporal dimension. The task requires understanding not only the spatial relationships between humans and objects but also how these interactions evolve over time.

Initial efforts on video-based HOI understanding focus on activity detection. Researchers explore probabilistic graphical models like Hidden Markov Models (HMMs) [110, 111], Dynamic Bayesian Networks (DBNs) [112], Conditional Random Fields (CRFs) [113, 114], and semi-CRFs [115] to capture temporal structures, but these methods are limited in their ability to anticipate future actions or model complex interactions. Koppula *et al.* [19] introduce the CAD-120 dataset, a benchmark for HOI recognition, and propose a Markov Random Fields (MRF) to model entities in videos with fully connected spatial and temporal edges. It also starts a trend to use sub-activity segments as temporal time units.

Their work is extended into the Anticipatory Temporal Conditional Random Field (ATCRF) model [21], which anticipates future human sub-activities or object affordances and gathers features from frame-level nodes. ATCRF is further advanced into GP-LCRF [20] to reduce the dimensionality of the frame-level human representation. Another extension of ATCRF is the Recursive CRF [116], in which the CRF is placed under a Bayesian filtering with an efficient belief computation. Based on the progress of spatio-temporal relation modules, MRF-like models advance into more efficient implementations with Convolutional Neural Networks (CNNs) [117] and Recurrent Neural Networks (RNNs) [118]. Nevertheless, these architectures

capture spatio-temporal context across the entire video scene but lack focus on key instances and fail to directly model spatio-temporal dependencies between humans and objects [6].

Graphs, commonly used to represent non-grid structures, naturally suit HOI recognition based on human and object instances by effectively modeling these dependencies [6, 14, 15, 88]. Typically, these methods first detect humans and objects, then recognize HOIs by analyzing the spatio-temporal evolution of graphs, where nodes represent the detected human and object instances.

Recently, numerous methods have attempted to model video-based HOI scenes using graph structures. Jain *et al.* [14] propose a model for integrating the strength of spatio-temporal graphs with RNNs in sequence learning. The Graph Parsing Neural Network (GPNN) [15] has been particularly successful in this domain, representing HOI structures with graphs and automatically parsing the optimal graph structure in an end-to-end manner. This model has proven effective in both static and dynamic scenes, demonstrating their ability to handle the complex relationships between humans and objects, especially in scenarios with an uncertain number of interaction pairs.

Further advancements focus on exploring more sophisticated graph structures and temporal modeling techniques. For instance, Wang *et al.* [6] utilizes spatio-temporal graphs that directly model the global relationship between the human and the object to be interacted with, capturing the state change of the interacting objects across frames. Other methods like LIGHTEN [7] explore the use of graph sequences to encode videos, capturing spatio-temporal relationship evolutions over the temporal dimension and spatial graph topologies. Additionally, ASSIGN [5] leverages the close coupling between the structure and content of events, allowing them to support each other in a joint discovery framework to achieve optimal solutions. It is a pioneering approach in learning the autonomous behavior of video entities, including their dynamic structure and interactions with coexisting neighbors. PGCN [8] investigates the exploit of temporal pyramid pooling modules to extend the capabilities of GCNs for action segmentation, particularly in the context of HOI recognition. STIGPN [9] exploits spatio-temporal graph convolutions to enhance the detection of salient

human-object interactions and efficiently modeling long-term dynamics.

Multi-person HOI recognition introduces additional challenges due to the heavy occlusion and complex interactions among multiple humans and objects. Most prior methods struggle with these scenarios, as they rely solely on visual features while neglecting the valuable information in geometric features. Our research explores deep neural networks to integrate geometric information of humans and objects, aiming to improve HOI recognition performance in multi-person scenes.

2.3 Features for Human Activity

Visual and geometric information are commonly used in human activity recognition tasks. Section 2.3.1 examines the exploitation of visual and geometric features for HOI detection, showing how combining appearance-based cues with geometric information like human pose and object positioning can improve detection performance. Section 2.3.2 explores how to use human skeletons in video-based human action recognition, highlighting their ability to capture motion dynamics and spatial relationships across frames. Section 2.3.3 discusses the challenges of incorporating geometric features for HOI recognition in videos, including issues related to ambiguous interactions and dynamic human-object relationships over time.

2.3.1 Visual and Geometric Features for HOI Detection

Some HOI detection methods relying solely on visual features [42, 43, 79, 80, 85, 95, 101, 119], such as those based on appearance, often use deep convolutional neural networks to extract features from detected human and object regions. For instance, Gkioxari *et al.* [95] introduce a human-centric branch within the Fast R-CNN [83] framework to predict HOIs using visual features from human and object regions. Fang *et al.* [119] employ a base network incorporating visual features of the whole person and the scene to capture the global context for HOI detection.

However, visual features alone are often insufficient to handle complex relations, as many interaction types involve fine-grained actions that are difficult to distinguish based on similar object-level features [120]. For instance, differentiating between

actions like "hold" and "catch" in sports scenes requires detailed, localized features to capture subtle distinctions. Methods relying solely on visual cues may struggle to capture these nuances, leading to potential misclassification of human-object pair interactions.

Geometric features, like human pose, object keypoints, and spatial configurations, provide valuable complementary information for HOI detection [13, 120]. Fang *et al.* [119] and Wan *et al.* [120] investigate semantic cues from human body parts, employing an attention module to identify the most informative body regions for HOI recognition. Wu *et al.* [121] propose extracting cross-person cues from body parts, providing valuable supplementary information for interactiveness discovery. Park *et al.* [122] introduce a graph with a pose-conditioned self-loop structure, enabling human node embeddings to be updated based on local features of human joints. Zhu *et al.* [13] incorporate human pose information into visual feature extraction to guide the identification of relevant body parts, and further use human and object keypoints to measure the likelihood of HOIs, enhancing interaction query representation. Such methods highlight the significance of geometric features in capturing fine-grained spatial information for HOI detection, improving detection accuracy, particularly in scenarios with occlusion.

2.3.2 Geometric Features for Human Action Recognition

With the advancement of geometric features in HOI detection, we aim to explore geometry-informed approaches for HOI recognition. Specifically, we employ human skeleton data, inspired by its successful application in video-based human action recognition, as it effectively captures motion dynamics.

Skeleton-based action recognition has attracted considerable interest, largely due to the ease of obtaining human skeleton data and its robustness to variations in viewpoint and appearance [123, 124]. The rise of deep learning has led to data-driven approaches, where RNNs, particularly LSTMs [125], are widely used for their capacity to model the temporal dynamics of skeleton sequences effectively [126–129]. CNN-based approaches typically transform skeleton sequences into pseudo-images using predefined transformation rules [130–133].

Graph Convolutional Networks (GCNs) [24] are designed to process graph-structured data. The human skeleton naturally lends itself to representation as a graph, with joints as nodes and bones as edges. This graph structure encodes important relationships between body parts, reflecting the kinematic constraints of human movement. GCNs can directly operate on this graph structure, leveraging the connectivity information to learn meaningful representations [134]. This is in contrast to methods that treat skeleton data as vector sequences or pseudo-images, which may not fully exploit the inherent spatial relationships between joints.

Human actions involve coordinated movements of multiple body parts over time. GCNs can capture these complex spatio-temporal dependencies by propagating information across the graph through message-passing mechanisms [32, 33]. This allows the model to learn how the movements of different joints relate to each other, both within a single frame and across consecutive frames. The hierarchical nature of GCNs further enables the model to learn increasingly abstract representations of the skeleton sequence, capturing higher-level motion patterns [32].

Yan *et al.* [32] propose ST-GCN and pioneer this approach by modelling the human skeleton as a graph and applying spatio-temporal graph convolutions to extract features for classification. The STGR network [135] enhances the skeleton graph by adding edges through frame-wise attention and global self-attention mechanisms. Similarly, 2s-AGCN [134] introduces adaptive graph structures with self-attention and a learnable residual mask, employing a two-stream ensemble with skeleton bone features to improve performance. SGN [124] explicitly encodes joint semantics, such as joint type and frame index, to strengthen feature representation. EfficientGCN [136] adopts techniques from CNN architectures, including separable convolutions and compound scaling, to create lightweight yet effective GCN models. Duan *et al.* [23] model human skeletons by sampling short skeleton sequences and using GCNs to extract spatio-temporal features from each sequence, capturing individual motion patterns. Li *et al.* [33] employ a GCN to model the entire two-person skeleton graph, where joints are nodes connected by edges representing both natural body connectivity and learned relationships from graph diffusion.

Based on the strengths of GCNs in modeling graph-structured data, we incorporate

geometric features in HOI videos. We choose human skeletons and object bounding boxes for humans and objects respectively since they provide essential geometric cues that are robust to visual variations, such as changes in appearance, lighting, or viewpoint. Human skeletons effectively capture the pose and motion of individuals, enabling finer-grained action understanding, while object bounding boxes define spatial context, essential for modeling interactions. Together, these features can enhance the model's ability to represent dynamic relationships in video-based HOI recognition.

2.3.3 Challenges in Geometry-Informed HOI Recognition

Introducing geometric features such as the keypoints of human pose and objects to HOI learning in videos is challenging and underexplored for a few reasons. On the one hand, in a video, interaction definitions might be ambiguous, such as "lift a cup" *vs.* "place a cup", "approaching" *vs.* "retreating" *vs.* "reaching". These actions might be detected as the same image label due to their visual similarity. Videos allow the use of temporal visual cues that are not presented in images [5].

On the other hand, the model needs to consider human dynamics throughout the video, as well as the shifting orientations and spatial arrangements of objects relative to humans [15]. This makes it difficult to directly extend image-based models to video that exploit the ROI features of human-object union [16].

Our work first attempts to introduce geometric features to HOI recognition architecture in Chapter 3 and propose a novel two-level graph to address the challenges. To resolve ambiguity in fine-grained actions, the first graph captures the spatial and temporal evolution of human and object keypoints, allowing for nuanced action differentiation across frames. The second graph bridges these geometric features with visual representations, enabling the model to capture shifting human dynamics and object orientations throughout the video. This two-level graph design allows our model to seamlessly incorporate temporal and spatial cues, overcoming the limitations of image-based methods in handling complex, dynamic interactions in videos.

2.4 Multimodal Fusion for Human Activities

Integrating diverse data modalities provides unique and complementary perspectives, contributing to a more comprehensive understanding of complex subjects, especially in tasks related to human activity recognition. Multimodal fusion combines visual, geometric, linguistic, and other feature types, each adding distinct contextual information [137–139]. Effective fusion strategies are particularly valuable in HOI tasks, where capturing the subtle nuances of human behavior requires robust integration of these varied feature types. Consequently, developing methods to fuse multimodal features remains a key research area, as it directly impacts the accuracy and interpretability of human activity recognition models.

In multimodal research of human action recognition, attention has been directed towards key human body areas, particularly the hands [140–142]. These studies employ attention-based methods to improve the overall accuracy of models that integrate skeletal and visual modalities. Building on this, Bruce *et al.* [143] expand the focus to include additional body regions such as the head, hands, and feet, adopting a temporal approach, where a fused representation is derived by multiplying spatial attention weights with the appearance features. In addition, Boulahia *et al.* [137] investigate the integration of various image modalities (RGB, Depth, Skeleton, and InfraRed) at different stages of the action recognition pipeline, encompassing early, intermediate, and late fusion techniques, to enhance the robustness of recognition.

In human interaction analysis, Wan *et al.* [120] concatenate human skeletal embeddings with visual embeddings from other branches like human, object and union to obtain the final holistic feature in the HOI scene. Zhou *et al.* [109] combine embedded visual and human pose features through element-wise addition. Wang *et al.* [9] directly concatenate multimodalities to obtain visual-spatial and spatial-semantic feature sequences, which are then input into a two-stream network. However, these direct operations overlook the fundamental representation and scale discrepancies between multimodal features, potentially causing misalignment in entity representation and less effective learning outcomes. Zhang *et al.* [144] initially concatenate appearance, spatial, and linguistic features to represent each human-

object pair and construct an interaction-centric graph for multimodal fusion, followed by a structure-aware Transformer for image-based HOI predictions. This graph-based fusion approach is computationally intensive for videos.

For the multi-person HOI recognition task involving multiple entities, effectively fusing multimodal features at the feature level to enhance the learning of entity-level HOIs remains a compelling research challenge. Chapter 4 attempts to fuse visual and geometric features according to different categories (*i.e.*, human and object), then embedded to a scenery graph to learn HOIs. Chapter 5 introduces another bottom-up framework that explores a dual-attention feature fusion mechanism, providing greater insight into the fusion process by highlighting important features in each entity. These refined features are then processed through an interdependent entity graph to model HOIs more effectively. This method achieves state-of-the-art performance in multi-person HOI scenarios.

2.5 Evaluation and Metric

In computer vision, especially in HOI recognition, systematic analysis of evaluation metrics is essential to improve model performance and applicability. Analyzing these metrics allows researchers to pinpoint model strengths and weaknesses, enabling iterative refinements that ensure robust and dependable outputs in real-world applications.

The F_1 score is a commonly used metric to evaluate the performance of machine learning models, especially for classification problems. It provides a balanced assessment of a model's accuracy, taking into account both precision and recall.

- Precision: This measures how many of the positive predictions made by the model are actually correct. In other words, it assesses the accuracy of the positive predictions.
- Recall: This measures how many of the actual positive instances in the dataset are correctly identified by the model. It focuses on the model's ability to capture all positive cases.

The F_1 score is calculated as the harmonic mean of precision and recall, ensuring that both metrics contribute equally to the final score. This means that to achieve a high F_1 score, a model must perform well in both identifying true positives and avoiding false positives.

Mathematically, the F_1 score can be expressed as:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.1}$$

Precision and Recall Formulae:

- Precision = True Positives / (True Positives + False Positives)
- Recall = True Positives / (True Positives + False Negatives)

Where:

- True Positives (TP): Correctly predicted positive instances.
- False Positives (FP): Incorrectly predicted positive instances.
- False Negatives (FN): Incorrectly predicted negative instances.

Joint segmentation and label recognition is our main task where models need to segment the timeline for each entity in the video and label those segments. $F_1@k$ metric [145] is designed specifically for this task, which evaluates the correctness of a predicted segment by comparing its Intersection over Union (IoU) with the ground-truth segment. A predicted segment is only considered correct if its IoU with the ground truth reaches a certain threshold, denoted as k, with common values 0.10, 0.25 and 0.50. The $F_1@k$ metric is particularly valuable in HOI recognition because (1) it penalizes over-segmentation errors, (2) it is tolerant to minor temporal shifts between predictions and ground truth, which may arise from annotator variability, and (3) it depends on the number of actions rather than the duration of each action instance.

CHAPTER 3

Geometric Features Informed Multi-person HOI Recognition

Portions of this chapter have previously been published in the following peer-reviewed publication [146]:

- **Qiao, T.**, Men, Q., Li, F. W., Kubotani, Y., Morishima, S., & Shum, H. P., "Geometric Features Informed Multi-person Human-object Interaction Recognition in Videos." In European Conference on Computer Vision (ECCV). Springer, 2022.

Human-Object Interaction (HOI) recognition in videos is important for analyzing human activity. Most existing approaches that focus on visual features often suffer from occlusion in real-world scenarios, a problem further complicated when multiple people and objects are involved in HOIs. Consider that geometric features such as human pose and object position provide meaningful information to understand HOIs, we argue to combine the benefits of both visual and geometric features in HOI recognition, and propose a novel Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN). The geometric-level graph models the interdependency between the geometric features of humans and objects, while the

fusion-level graph further fuses them with the visual features of humans and objects. To demonstrate the novelty and effectiveness of our method in challenging scenarios, we propose a new multi-person HOI dataset (MPHOI-72). Extensive experiments on MPHOI-72 (two-person HOI), CAD-120 (single-human HOI) and Bimanual Actions (two-hand HOI) datasets demonstrate our superior performance compared to state-of-the-arts.

3.1 Introduction

The real-world human activities are often closely associated with surrounding objects. HOI recognition focuses on learning and analyzing the interaction between human and object entities for activity recognition. HOI recognition involves the segmentation and recognition of individual human sub-activities/object affordances in videos, such as drinking and placing, to gain an insight into the overall human activities [5]. Based on this, downstream applications such as security surveillance, healthcare monitoring and human-robot interactions can be developed.

Earlier work in HOI detection is limited to detecting interactions in one image [79,80,147]. With HOI video datasets proposed, models have been developed to learn the action representations over the spatio-temporal domain for HOI recognition [14,15]. Notably, ASSIGN [5] proposes a visual feature attention model to learn asynchronous and sparse HOI in videos, achieving state-of-the-art results.

A main challenge of video-based HOI recognition is that visual features usually suffer from occlusion. This is particularly problematic in real-world scenarios when multiple people and objects are involved. Recent research has shown that extracted pose features are more robust to partial occlusions than visual features [10, 11]. Bottom-up pose estimators can extract body poses as long as the local image patches of joints are not occluded [22]. With advanced frameworks such as Graph Convolutional Networks (GCNs), geometric pipelines generally perform better than visual ones on datasets with heavy occlusion [12]. Therefore, geometric features provide complementary information to visual ones [11, 148].

In this work, we propose to fuse geometric and visual features for HOI recogni-

Figure 3.1: Two examples (*Cheering* and *Co-working*) of our collected multi-person HOI dataset. Geometric features such as skeletons and bounding boxes are annotated.

tion in videos. Our research insight is that geometric features enrich fine-grained human-object interactions, as evidenced by previous research on image-based HOI detection [149, 150]. We present a novel Two-level Geometric feature informed Graph Convolutional Network (2G-GCN) that extracts geometric features and fuses them with visual ones for HOI recognition in videos. We implement the network by using the geometric-level graph to model representative geometric features among humans and objects, and fusing the visual features through the fusion-level graph.

To showcase the effectiveness of our model, we further propose a Multi-Person dataset for Human-Object Interaction (MPHOI), which closely ensembles real-world activities that contain multiple people interacting with multiple objects. Our dataset includes common multi-person activities and natural occlusions in daily life (Fig. 3.1). It is annotated with the geometric features of human skeletal poses, human and object bounding boxes, and ground-truth HOI activity labels, which can be used as a versatile benchmark for multiple tasks such as visual-based or skeleton-based human activity analysis or hybrid.

We outperform state-of-the-arts in multiple datasets, including our novel MPHOI-72 dataset, the single-human HOI CAD-120 [19] dataset, and the two-hand Bimanual Actions [62] dataset. We also extensively evaluate core components of 2G-GCN in ablation studies. Our main contributions are as follows:

- We propose a novel geometry-informed 2G-GCN network for HOI recognition in videos. The network consists of a two-level graph structure that models geometric features between humans and objects, together with the corresponding visual features.

- We present the novel problem of MPHOI in videos with a new MPHOI-72 dataset[1], showcasing new challenges that cannot be directly resolved by existing methods.
- We outperform state-of-the-art HOI recognition networks in our MPHOI-72 dataset, the CAD-120 [19] dataset and the Bimanual Actions [62] dataset.

3.2 The Multi-Person HOI Dataset (MPHOI-72)

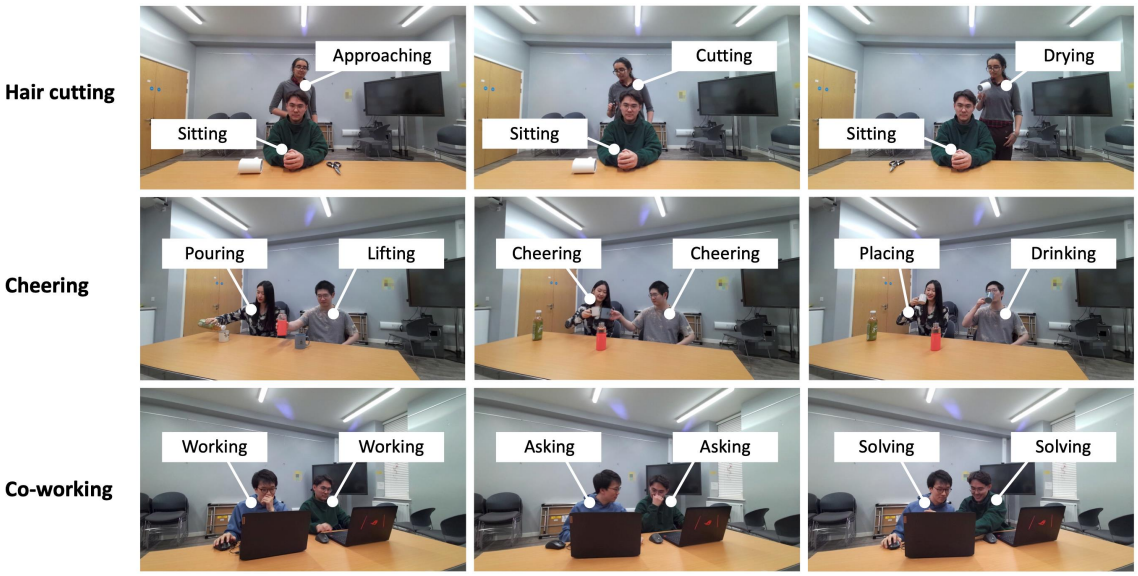

Figure 3.2: Sample video frames of three different MPHOI activities in MPHOI-72.

We propose an HOI dataset with multi-person activities (MPHOI-72), which is challenging due to many body occlusions among humans and objects. We have 3 males and 2 females, aged 23-27, who are randomly combined into 8 groups with 2 people per group and perform 3 different HOI activities interacting with 2-4 objects. We also prepared 6 objects: cup, bottle, scissors, hair dryer, mouse and laptop. 3 activities = {*Cheering*, *Hair cutting*, *Co-working*} and 13 sub-activities = {*Sit*, *Approach*, *Retreat*, *Place*, *Lift*, *Pour*, *Drink*, *Cheers*, *Cut*, *Dry*, *Work*, *Ask*, *Solve*} are

[1]Data collection performed in the UK, under Durham University Ethics Approval Ref: COMP-2022-06-03T19_29_22-cbmw62. The dataset can be downloaded at https://collections.durham.ac.uk/collections/r19g54xh706.

defined. The dataset consists of 72 videos captured from 3 different angles at 30 fps, with a total of 26,383 frames and an average length of 12 seconds.

Fig. 3.2 shows some sample video frames of the three activities in our MPHOI-72 dataset, and the sub-activity label of each subject is annotated frame-wise. The top row presents *Hair cutting* from the front view, where one subject is sitting and another subject interacts with a pair of scissors and a hair dryer. Most part of the body of the subject standing at the back is invisible. The second row presents a popular human activity, *Cheering*, in which two subjects pour water from their own bottles, lift cups to cheer, and drink. The high-level occlusion exists between humans, cups and bottles during the entire activity. The bottom row presents *Co-working*, which simulates the situation of two co-workers asking and solving questions. Besides, we also consider distinct human sizes, skin colors and a balance of gender. These samples illustrate the diversity of our dataset.

We use Azure Kinect SDK to collect RGB-D videos with 3840×2160 resolution, and employ their Body Tracking SDK [151] to capture the full dynamics of two subject skeletons. Object bounding boxes are manually annotated frame-wise. For each video, we provide such geometric features: 2D human skeletons and bounding boxes of the subjects and objects involved in the activity (Fig. 3.1).

3.3 Two-level Geometric Features Informed Graph Convolutional Network (2G-GCN)

To learn the correlations during human-object interaction, we propose a two-level graph structure to model the interdependency of the geometric features, known as 2G-GCN. The model consists of two key components: a geometry-level graph for modeling geometry and object features to facilitate graph convolution learning, and a fusion-level graph for fusing geometric and visual features (Fig. 3.3).

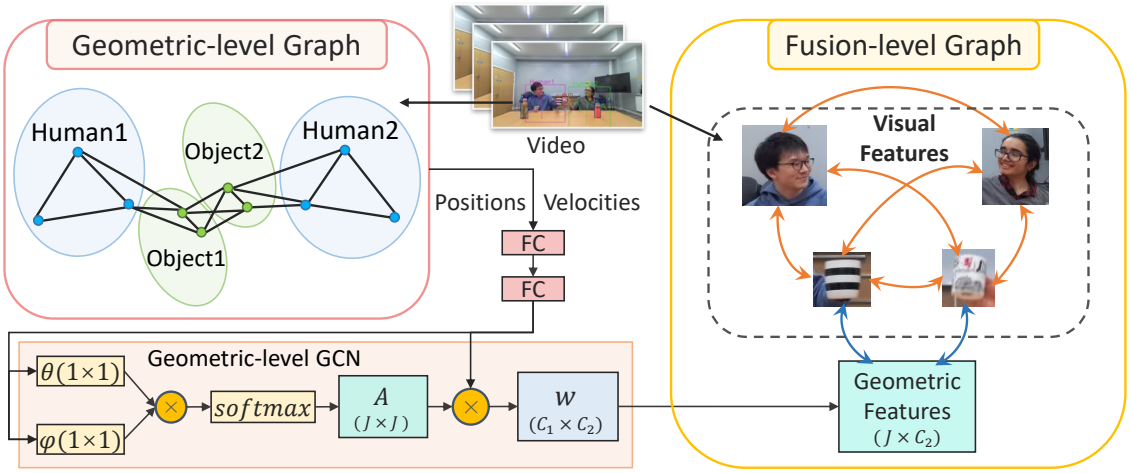

Figure 3.3: Our 2G-GCN framework comprises a geometric-level and a fusion-level graph.

3.3.1 Geometric Features

The geometric features of humans can be represented in various ways. Human skeletons contain an explicit graph structure with joints as nodes and bones as edges. The joint position and velocity offer fine-grained dynamics in human motion [124], while the joint angle also provides spatial cues in 3D skeleton data [134]. Alternatively, body shapes and how they deform during movement can be represented by surface models [152] or implicit models [153]. We employ human skeletons with joint position and velocity, because they are essential cues to human motion. Also, unlike body shapes, they are invariant to human appearance.

We represent human poses in an effective representation to inform HOI recognition. For human skeleton, we select specific body keypoints and denote them as a set $\mathcal{S} = \{M_t^{h,k}\}_{t=1,h=1,k=1}^{T,H,K}$, where $M_t^{h,k}$ denotes the body joint of type k in human h at time t, T denotes the total number of frames in the video, H and K denote the total number of humans and keypoints of a human body in a frame, respectively. For a given human body keypoint $M_t^{h,k}$, we define its position as $\mathbf{p}_{t,h,k} = (x_{t,h,k}, y_{t,h,k})^T \in \mathbb{R}^2$ in 2D, and the velocity as $\mathbf{v}_{t,h,k} = \mathbf{p}_{t+1,h,k} - \mathbf{p}_{t,h,k}$, which is the forward difference of neighbor frames. In the channel of each human skeleton keypoint, we concatenate its position $\mathbf{p}_{t,h,k}$, and velocity $\mathbf{v}_{t,h,k}$ in the channel domain, forming the human geometric context $\mathbf{h}_{t,h,k} = [\mathbf{p}_{t,h,k}, \mathbf{v}_{t,h,k}] \in \mathbb{R}^4$.

As objects play a crucial role in the HOI videos, we also consider their geometric features. The two diagonal points of the object bounding box are utilised to represent

the object position. We define all object keypoints as $\mathcal{O} = \{B_t^{f,u}\}_{t=1,f=1,u=1}^{T,F,2}$, where $B_t^{f,u}$ denotes the object keypoint of type u in object f at time t. F denotes the maximum number of objects in a video and $u = \{1, 2\}$ is the index of the top-left and the bottom-right points of the object bounding box, respectively. The object geometric context $\mathbf{o}_{t,f,u} = [\mathbf{p}_{t,f,u}, \mathbf{v}_{t,f,u}] \in \mathbb{R}^4$ can be obtained by the same process as the human skeleton.

3.3.2 The Geometric-Level Graph

We design a novel geometric-level graph that involves both human skeleton and object keypoints to explore their correlations in an activity (Fig. 3.3 left). We use $\mathbf{g}_t$ to denote a graph node with geometric features of a keypoint from either a human $h_{t,h,k}$ or an object $o_{t,f,u}$ at frame t. Therefore, all keypoints of the frame t are denoted by $G_t = (\mathbf{g}_{t,1}; \cdots; \mathbf{g}_{t,J})$, where $J = H \times K + F \times 2$ joints, each with 4 channel dimensions including its 2D position and velocity. This enables us to enhance the ability of GCN to capture correlations between human and object keypoints in HOI activities by learning their dynamic spatial cues. We embed $\mathbf{g}_t$ using two fully connected (FC) layers following [124] as:

$$\widetilde{\mathbf{g}}_t = \sigma(W_2(\sigma(W_1\mathbf{g}_t + \mathbf{b}_1)) + \mathbf{b}_2) \in \mathbb{R}^{C_1}, \tag{3.1}$$

where C_1 is the dimension of the joint representation, $W_1 \in \mathbb{R}^{C_1 \times 4}$ and $W_2 \in \mathbb{R}^{C_1 \times C_1}$ are weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias vectors, and σ is the ReLU activation function.

We propose an adaptive adjacency matrix exploiting the similarity of the geometric features in the GCN. We employ the dot-product similarity in $\widetilde{\mathbf{g}}_t$, as it allows us to determine if and how strong a connection exists between two keypoints in the same frame t [124, 134, 154]. This is a better choice for our problem comparing to other strategies, *e.g.* the traditional adjacency matrix only represents the physical structure of the human body [32] or a fully-learned adjacency matrix without the supervision of graph representations [134]. We represent the adjacency matrix A_t

with ${j_1}^{th}$ and ${j_2}^{th}$ keypoints as:

$$A_t(j_1, j_2) = \theta(\widetilde{\mathbf{g}}_{t,j_1})^T \phi(\widetilde{\mathbf{g}}_{t,j_2}), \tag{3.2}$$

where $\theta, \phi \in \mathbb{R}^{C_2}$ denote two transformation functions, each implemented by a 1×1 convolutional layer. Then, SoftMax activation is conducted on each row of A_t to ensure the integration of all edge weights of a node equal to 1. We subsequently obtain the output of the geometry-level graph from the GCN as:

$$Y_t = A_t \widetilde{\mathbf{G}}_t W_g, \tag{3.3}$$

where $\widetilde{\mathbf{G}}_t = (\widetilde{\mathbf{g}}_{t,1}; \cdots; \widetilde{\mathbf{g}}_{t,J}) \in \mathbb{R}^{J \times C_1}$ and $W_g \in \mathbb{R}^{C_1 \times C_2}$ is the transformation matrix. The output size is $T \times J \times C_2$.

3.3.3 The Fusion-Level Graph

We propose a fusion-level graph to connect the geometric features learned from GCN with visual features. Previous works on CNN-based HOI recognition in videos overemphasize visual features and neglect geometric features of humans and objects [155, 156]. State-of-the-arts like ASSIGN [5] also exclude geometric features. In contrast, we first extract visual features for each human or object entity by ROI pooling, and then introduce the geometric output Y_t from the GCN as the auxiliary feature to complement the visual representation. The feature vectors for all entities are then embedded by a two-layer MLP with ReLU activation function to the same hidden size.

A key design of the fusion-level graph is an attention mechanism to estimate the relevance of the interacted neighboring entity. As illustrated in the fusion-level graph of Fig. 3.3, each person and object denote an entity through time, while Y_t forms an additional entity joining the graph. All connections between the visual features of all humans and objects in the video are captured, represented by orange arrows. The blue arrows denote the connection between geometric and object visual features. Empirically, connecting the geometry-object pairs consistently performs

better than applying a fully-connected graph with geometry-human connections. A possible reason is that humans are generally bigger in size and therefore have a larger chance of occlusion. Correlating such relatively noisy human visual and geometry features is a harder problem than the objects' equivalent. The fusion strategy is evaluated in the ablation studies.

The attention mechanism employed in the fusion-level graph calculates a weighted average of the contributions from neighbouring nodes, implemented by a variant of scaled dot-product attention [157] with identical keys and values:

$$\text{Att}\left(q, \{z_i\}_{i=1\ldots n}\right) = \sum_{i=1}^{n} \text{softmax}\left(\frac{q^T z_i}{\sqrt{d}}\right) z_i, \tag{3.4}$$

where q is a query vector, $\{z_i\}$ is a set of keys/values vectors of size n, and d is the feature dimension.

Once the fusion-level graph is constructed, we employ ASSIGN [5] as the backbone for HOI recognition. ASSIGN is a recurrent graph network that automatically detects the structure of HOI associated with asynchronous and sparse entities in videos. Our fusion-level graph is compatible with the HOI graph structure in ASSIGN, allowing us to employ the network to predict sub-activities for humans and object-affordances for objects depending on the dataset.

3.4 Experiments

3.4.1 Datasets

We have performed experiments on our MPHOI-72 dataset, the CAD-120 [19] dataset and the Bimanual Actions [62] dataset, showcasing the superior results of 2G-GCN on multi-person, single-human and two-hand HOI recognition.

CAD-120 is widely used for HOI recognition. It consists of 120 RGB-D videos of 10 different activities performed individually by 4 participants, with each activity replicated 3 times. A participant interacts with 1-5 objects in each video. There are 10 human sub-activities (*e.g.*, *eating*, *drinking*), and 12 object affordances (*e.g.*, *stationary*, *drinkable*) in total, which are annotated per frame.

Bimanual Actions is the first HOI activity dataset where subjects use two hands to interact with objects (*e.g.*, the left hand holding a piece of wood, while the right hand sawing it). It contains 540 RGB-D videos of 6 subjects performing 9 different activities, with each repeated 10 times. There are a total of 14 action labels for each hand, and each entity in the video is annotated frame by frame.

3.4.2 Implementation Details

Network Settings

We implement 2048-dimensional ROI pooling features extracted from the 2D bounding boxes of humans and objects in the video detected by a Faster R-CNN [83] module, which is pre-trained [158] on the Visual Genome dataset [39] for entity visual features. We set the number of neurons to 64 and 128 for both FC layers for the embedding and the transformation functions of Eq. 3.2 in the geometric-level graph, respectively (*i.e.*, $C_1 = 64$, $C_2 = 128$).

Experimental Settings

2G-GCN is evaluated on the task of joint segmentation and label recognition. It requires models to segment the timeline for each entity in a video and assign labels to each segment. For evaluation, we report the $F_1@k$ metric [145] with the commonly used thresholds $k = 10\%$, 25% and 50%, which is frequently adopted in prior segmentation researches [5, 145, 159].

For the CAD-120 and Bimanual Actions datasets, we use leave-one-subject cross-validation to evaluate the generalization effort of 2G-GCN in unknown subjects. On MPHOI-72, we define a cross-validation scheme that chooses two subjects not present in the training set as the test set.

With four Nvidia Titan RTX GPUs, training MPHOI-72, CAD-120 and Bimanual Actions takes 2 hours, 8 hours and 5 days, respectively. Testing the whole test set takes 2 minutes, 3 minutes and 20 minutes, respectively.

3.4.3 Quantitative Comparison

Multi-person HOIs

In our challenging MPHOI-72 dataset, 2G-GCN beats ASSIGN [5] by a considerable gap (Table 3.1). 2G-GCN significantly outperforms ASSIGN and has smaller standard deviation values in every F_1 configurations, reaching 68.6% in F_1@10 score, which is approximately 9.5% higher than ASSIGN. The performance of visual-based methods such as ASSIGN is generally ineffective, since remarkable occlusions in MPHOI typically invalids visual features to HOI recognition task. The significant gaps between the results of 2G-GCN and ASSIGN demonstrate that the application of geometric features and its fusion with visual features can motivate our model to learn stable and essential features even when significant occlusion appears in HOIs.

Table 3.1: Joined segmentation and label recognition on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN [5]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN	**68.6 ± 10.4**	**60.8 ± 10.3**	**45.2 ± 6.5**

Table 3.2: Joined segmentation and label recognition on CAD-120.

Model	Sub-activity			Object Affordance		
	F_1@10	F_1@25	F_1@50	F_1@10	F_1@25	F_1@50
rCRF [116]	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3	72.1 ± 2.5	69.1 ± 3.3	57.0 ± 3.5
Independent BiRNN	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8	84.6 ± 2.1	81.5 ± 2.7	71.4 ± 4.9
ATCRF [160]	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3	79.9 ± 3.1	77.0 ± 4.1	63.3 ± 4.9
Relational BiRNN	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5	82.3 ± 2.3	78.5 ± 2.7	68.9 ± 4.9
ASSIGN [5]	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8	92.0 ± 1.1	90.2 ± 1.8	82.4 ± 3.5
2G-GCN	**89.5 ± 1.6**	**87.1 ± 1.8**	**76.2 ± 2.8**	**92.4** ± 1.7	**90.4** ± 2.3	**82.7 ± 2.9**

Single-person HOIs

The generic formulation of 2G-GCN results in excellent performance in single-person HOI recognition. Table 3.2 presents the results of 2G-GCN with state-of-the-arts and

two BiRNN-based baselines on CAD-120. Bidirectional GRU is used as a baseline in both cases: The Independent BiRNN models each entity individually (*i.e.*, there are no spatial messages), but the Relational BiRNN incorporates extensive spatial relations between entities. Three previous works, ATCRF [160], rCRF [116] and ASSIGN [5], are fully capable of performing this task, where ASSIGN is relatively new and can improve the scores to higher levels. For both human sub-activity and object affordance labelling, 2G-GCN beats ASSIGN in every configuration of the $F_1@k$ metric. Especially for the sub-activity labelling, 2G-GCN improves 1.5% over ASSIGN in $F_1@10$, and more than 2% in $F_1@\{25, 50\}$ with lower standard deviation values. These findings demonstrate the benefits of using geometric features from human skeletons and object bounding boxes, rather than only using visual features like ASSIGN.

Two-hand HOIs

Table 3.3: Joined segmentation and label recognition on Bimanual Actions.

Model	Sub-activity		
	$F_1@10$	$F_1@25$	$F_1@50$
Dreher *et al.* [62]	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
Independent BiRNN	74.8 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN [5]	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN	**85.0** ± 2.2	**82.0** ± 2.6	**69.2** ± **3.1**

For two-hand HOI recognition on the Bimanual Actions dataset, 2G-GCN outperforms ASSIGN [5] by about 1%. We compare the performance on the joined segmentation and labelling task with Dreher *et al.* [62], ASSIGN [5] and BiRNN baselines (Table 3.3). Dreher *et al.* [62] have the worst results due to their fairly basic graph network, which ignores hand interactions and does not account for long-term temporal context. By taking into account a larger temporal context, the BiRNN baselines outperform Dreher *et al.* [62]. Our 2G-GCN has made a small improvement over ASSIGN [5]. This is partly because the hand skeletons provided by the Bimanual Actions dataset are extracted by OpenPose [22], which is relatively weak on hand pose estimation.

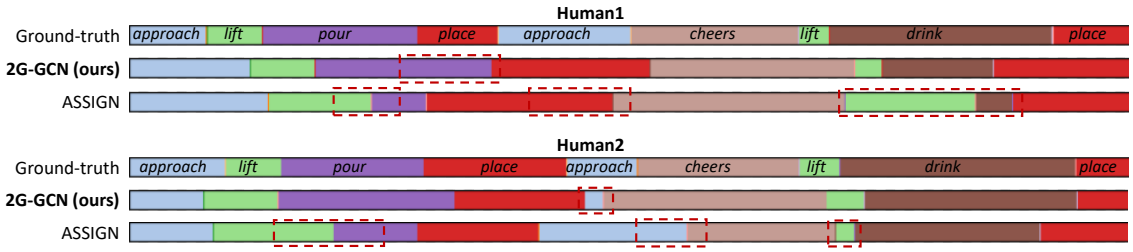

Figure 3.4: Visualizing the segmentation and labels on MPHOI-72 for *Cheering*. Red dashed boxes highlights major segmentation errors.

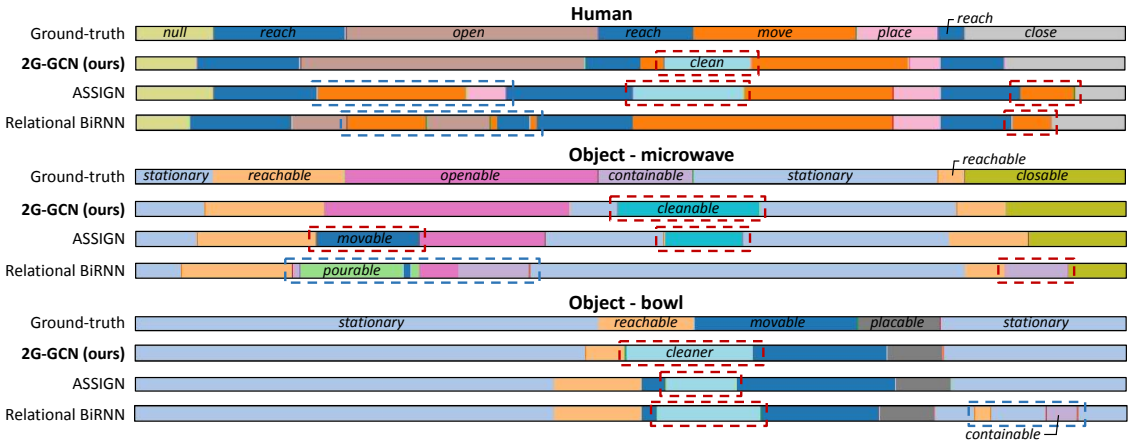

Figure 3.5: Visualizing the segmentation and labels on CAD-120 for *taking food*. Red dashed boxes highlight over-segmentation. Blue ones highlight chaotic segmentation.

3.4.4 Qualitative Comparison

We compare the visualization of 2G-GCN and relevant methods on our challenging MPHOI-72 dataset. Fig. 3.4 shows an example of segmentation and labeling results with 2G-GCN and ASSIGN [5] approaches compared with the ground-truth for a *Cheering* activity. We highlight some major segmentation errors with red dashed boxes. Although both models have some errors, 2G-GCN is generally more robust to varying segmentation periods and activity progression than ASSIGN. 2G-GCN is not particularly sensitive to the timeline of *place* and *approach*, while ASSIGN crashes for most sub-activities.

Fig. 3.5 displays an example of a *taking food* activity on the CAD-120 dataset. We highlight over-segmentation with the red dashed box and chaotic segmentation with the blue dashed box. From the figure, our 2G-GCN is able to segment and recognize both human sub-activities and object affordances more accurately than

the other two models. ASSIGN [5] and Relational BiRNN fail to predict when the human opens or closes the microwave (*e.g.* the *open* and *close* sub-activities for the human, and the *openable* and *closable* affordances for the microwave).

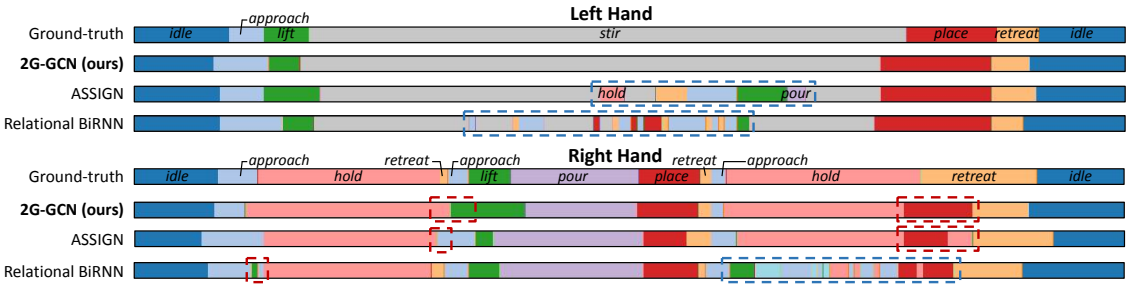

Figure 3.6: Visualizing the segmentation and labels on Bimanual Actions for *cooking*. Red dashed boxes highlight extra or missing segmentation. Blue ones highlights chaotic segmentation.

Fig. 3.6 depicts the qualitative visualization of a *cooking* activity on the Bimanual Actions dataset. Here, 2G-GCN performs outstandingly with precise segmentation and labelling results for the left hand, while ASSIGN [5] and Relational BiRNN have a chaotic performance when segmenting the long *stir* action. In contrast, the right hand has more complex actions, which confuses the models a lot. 2G-GCN generally performs better than ASSIGN, although both of them have some additional and missing segmentations. Relational BiRNN has the worst performance with chaotic segmentation errors in the *hold* action.

3.4.5 Ablation Studies

The two proposed graphs in our method contain important structural information. We ablate various essential modules and evaluate them on our MPHOI-72 and CAD-120 datasets to demonstrate the role of different 2G-GCN components as shown in Table 3.4 and Table 3.5, where GG and FG denote the geometric-level graph and fusion-level graph, respectively.

Table 3.5: Ablation study on CAD-120. GG and FG denote the geometric-level graph and the fusion-level graph, respectively.

	Model	Sub-activity		Object Affordance	
		F_1@10	F_1@25	F_1@10	F_1@25
(1)	GG (w/o skeletons) & FG	87.7	84.9	91.0	88.3
(2)	GG (w/o objects) & FG	88.3	85.6	90.4	88.5
(3)	GG (w/o embedding) & FG	89.4	86.4	91.5	90.0
(4)	GG (w/o similarity) & FG	88.7	85.0	90.6	89.0
(5)	GG & FG (w/o human-object)	73.4	68.8	90.3	88.4
(6)	GG & FG (w/o object-object)	88.3	84.5	90.9	88.5
(7)	GG & FG (w human-geometry)	89.0	86.6	91.4	89.3
(8)	2G-GCN	**89.5**	**87.1**	**92.4**	**90.4**

Table 3.4: Ablation study on MPHOI-72. GG and FG denote the geometric-level graph and the fusion-level graph, respectively.

	Model	Sub-activity	
		F_1@10	F_1@25
(1)	GG (w/o skeletons) & FG	66.8	60.2
(2)	GG (w/o objects) & FG	66.7	59.8
(3)	GG (w/o embedding) & FG	62.2	56.5
(4)	GG (w/o similarity) & FG	66.1	58.9
(5)	GG & FG (w/o human-human)	67.2	59.6
(6)	GG & FG (w/o human-object)	58.6	51.7
(7)	GG & FG (w/o object-object)	65.7	60.2
(8)	GG & FG (w human-geometry)	65.6	60.7
(9)	2G-GCN	**68.6**	**60.8**

We first investigate the importance of geometric features of the human and objects. The experiments in row (1) drops the human skeleton features in the geometric-level graph, while row (2) drops the object keypoint features. Row (3) explores the effect of the embedding function on geometric features. The last component we ablated is the similarity matrix used in the GCN, the result comparison between row (4) and (8) demonstrates its significance in the model.

We further ablate different components in the fusion-level graph as shown in

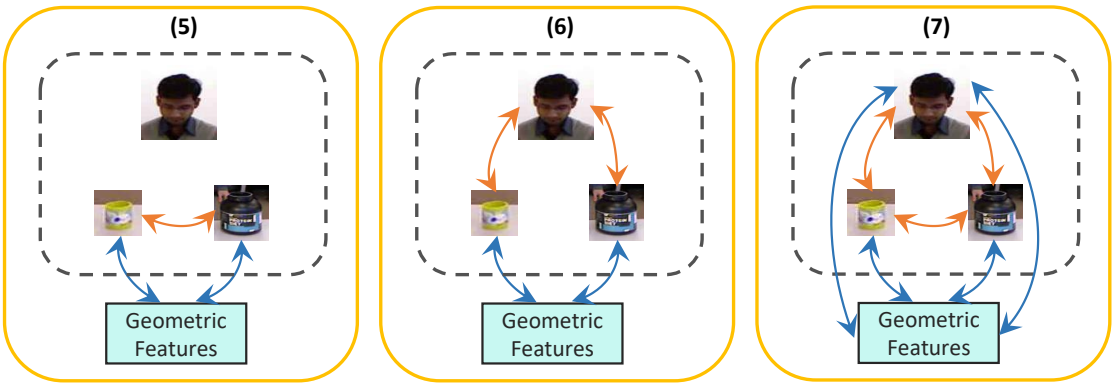

Figure 3.7: Ablation study of the fusion-level graph. Human-object, object-object and geometry-human relations are ablated (rows (5), (6), (7) in Table 3.5 respectively).

Fig. 3.7. We disable the attention connection between the pair of human-object and object-object in row (5) and (6), respectively, and also supplement the human-geometry connection in row (7). The inferior results reported in rows (5) and (6) verify the significance of incorporating all these pair connections in our full 2G-GCN.

3.5 Summary

We propose a two-level graph GCN for tackling HOIs in videos, which consists of a geometric-level graph using human skeletons and object bounding boxes, and a fusion-level graph fusing the geometric features with traditional visual features. We also propose a novel MPHOI-72 dataset to enable and motivate research in multi-person HOI recognition. Our 2G-GCN outperforms state-of-the-art HOI recognition networks in single-person, two-hand and multi-person HOI domains.

Our method is not limited to two humans; the geometric-level graph can represent multiple humans and objects. To handle an arbitrary number of entities, a graph can be constructed by only considering the k-nearest humans and objects, allowing better generalization [161]. If there are a large number of entities, to avoid handling a large fully-connected graph, we can apply an attention mechanism to learn what nodes are related [162], thereby better recognizing HOIs.

Building on the promising results from 2G-GCN, where the integration of geometric features demonstrated effectiveness, the next chapter will investigate novel approaches to achieve a more meaningful fusion of geometric and visual features.

CHAPTER 4

Learning Multi-Person HOI From Category to Scenery

Portions of this chapter have previously been published in the following peer-reviewed publication [163]:

- **Qiao, T.**, Li, R., Li, F. W., & Shum, H. P., "From Category to Scenery: An End-to-End Framework for Multi-Person Human-Object Interaction Recognition in Videos." In International Conference on Pattern Recognition (ICPR), 2024.

Video-based HOI recognition explores the intricate dynamics between humans and objects, which are essential for a comprehensive understanding of human behavior and intentions. While previous work has made significant strides, effectively integrating geometric and visual features to model dynamic relationships between humans and objects in a graph framework remains a challenge. In this work, we propose a novel end-to-end category to scenery framework, CATS, starting by generating geometric features for various categories through graphs respectively, then fusing them with corresponding visual features. Subsequently, we construct a scenery interactive graph with these enhanced geometric-visual features as nodes to learn the relationships among human and object categories. This methodological advance facilitates a deeper,

more structured comprehension of interactions, bridging category-specific insights with broad scenery dynamics. Our method demonstrates superior performance on the two-person HOI MPHOI-72 dataset and comparable performance on the single-person HOI CAD-120 dataset.

4.1 Introduction

HOI recognition delves into the subtle dynamics between humans and objects, aiming to capture the breadth of their interactions from basic actions to complex activities. This field transcends mere identification to explore the depth of their interactions, from elementary actions to intricate sequences, which are essential for a comprehensive understanding of human behavior and intentions [5, 146, 164]. Accurate HOI recognition is crucial across various domains, serving as a cornerstone for developing sophisticated surveillance [1, 2], enhancing video analysis techniques [165–167], and facilitating effective human-robot collaboration [3, 4].

Prior work on single-person HOI video datasets marks a significant advancement [19, 62, 63], enabling the development of models that understand spatio-temporal actions through visual cues [5, 14, 15]. A notable progression is presented by 2G-GCN [146], which leverages geometric features informed networks for HOI recognition in videos, broadening the scope to encompass two-person HOIs with the introduction of a novel dataset.

While fusing geometric and visual features achieves remarkable performance, video-based HOI recognition still faces challenges in effectively fusing these features and learning dynamic relationships between humans and objects in a graph model. 2G-GCN [146] attempts to enrich visual data with geometric information via a graph-based network. However, merging geometric features of all humans and objects with individual visual features in a single graph leads to a critical flaw by neglecting category-specific characteristics. This fusion difficulty hampers accurate and specific HOI learning, especially in complex multi-person scenes.

Categorization simplifies learning and improves behavior discrimination by grouping similar features, enhancing model accuracy in identifying diverse interactions.

In this work, we follow natural cognitive processes [168, 169] to learn HOIs from category-level feature fusion to scenery-level graph representation, facilitating a structured and comprehensive understanding. This strategy enables a more sophisticated integration of varied feature types, ensuring each level is fully leveraged for enhanced representational efficacy. We propose a novel end-to-end CATegory to Scenery framework (CATS), which initially generates geometric features via a graph for different categories, integrating them with corresponding visual features. Subsequently, a scenery interactive graph is constructed using these enriched geometric-visual features as nodes, to deeply understand the interaction dynamics among all humans and objects.

Our approach achieves superior performance on the two-person HOI MPHOI-72 [146] dataset and comparable performance on the single-person HOI CAD-120 [19] dataset. Additionally, we conduct ablation studies to evaluate the core components of our model. Our main contributions are:

- We propose an end-to-end framework CATS ranging from category-level feature fusion to scenery-level graph for multi-person HOI recognition in videos.
- We propose a multi-category multi-modality fusion module that fuses visual features and graph-based geometric features for human and object categories, respectively.
- We propose a scenery interactive graph to learn the relationships among human and object categories via an attention-based graph.

4.2 An End-to-End Category to Scenery Framework (CATS)

We propose an end-to-end framework CATS (Fig. 4.1) to learn HOIs from category-level to scenery-level, which first focuses on the inherent characteristics of different categories, capturing their physical properties and contextual visual cues to achieve a rich feature representation. It then adopts a graph attention neural network to

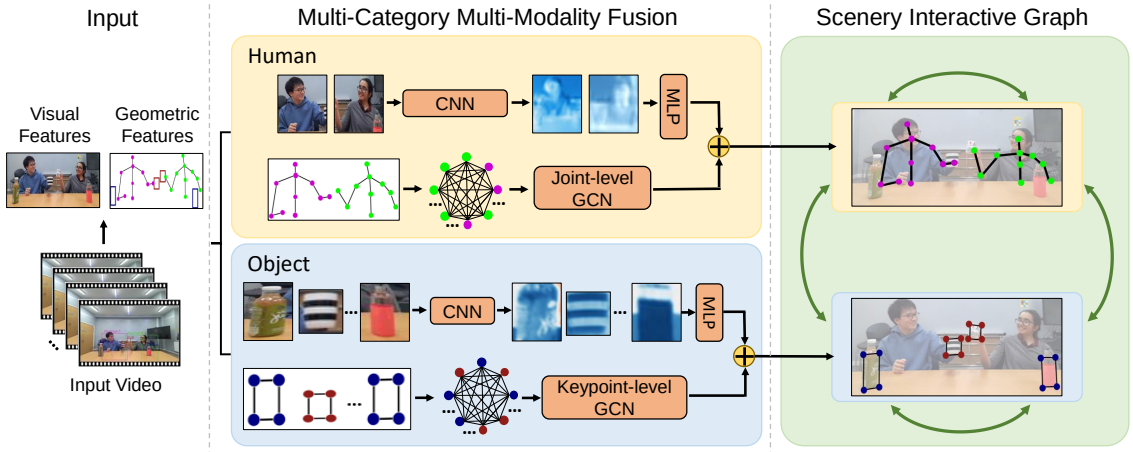

Figure 4.1: Overview of our end-to-end framework CATS. We first learn geometric features via a graph for human and object categories, fusing them with corresponding visual features. Subsequently, a scenery interactive graph is constructed to deeply understand the interaction dynamics between multi-categories.

learn multi-category features as a scenery graph representation, which represents the true HOI. This approach mirrors natural cognitive processes [168, 169] facilitating a structured and comprehensive understanding of interactions within various contexts.

Alternative architecture performs suboptimally, an approach treats each human and object as an entity independently, ignoring the correlation between the same category and compromising the model's ability to understand complex dynamics. An alternative method [146] groups all human poses and object bounding boxes into a single category for geometric feature learning, and then combines these geometric features with visual features in a single graph learning, which complicates entity representation and hampers explicit HOI learning. We compare these alternative architectures with our method in Experimental Results 4.3.

4.2.1 Multi-Category Multi-Modality Fusion

Previous CNN-based methods for HOI recognition in videos have predominantly focused on visual features [5, 155, 156], which may not be sufficient in cases of occlusion. While more advanced approaches like 2G-GCN [146] have attempted to incorporate geometric features to complement visual features, they categorize all human skeletons and object bounding boxes under a single category for geometric feature learning, thereby neglecting the distinct characteristics unique to each category and potentially

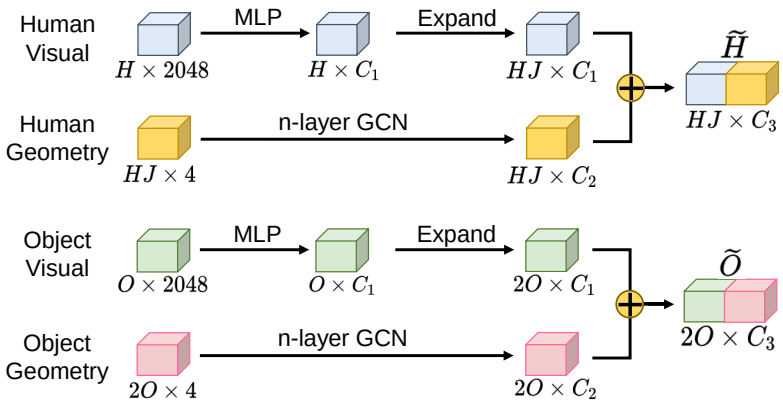

Figure 4.2: The process of learning and fusing geometric and visual features for human and object categories.

generating skewed geometric features.

To this end, we propose a multi-category multi-modality fusion module that first learns geometric features via a graph for human and object two categories and then fuses them with corresponding visual features (Fig. 4.1). These category-specific features establish a rich multimodal context, providing a solid foundation for subsequent accurate interaction recognition.

Geometric Features

For feature representation in human category and other related tasks, following previous successes [146, 170], we concatenate the position and velocity of all humans into keypoint channels, forming human geometric features $\mathcal{HG} = \{hg_{t,h,j}\}_{t=1,h=1,j=1}^{T,H,J} \in \mathbb{R}^4$, where $hg_{t,h,j}$ denotes the body joint of type j in human h at time t, T denotes the total number of frames in the video, H and J denote the total number of humans and keypoints of a human body in a frame, respectively. Similar to humans, object geometric features $\mathcal{OG} = \{og_{t,o,u}\}_{t=1,o=1,u=1}^{T,O,2} \in \mathbb{R}^4$, where $og_{t,o,u}$ denotes the bounding box diagonal points u in object o at time t and O denotes the total number of objects.

As shown in Fig. 4.2, human and object geometric features are adopted n-layer GCNs to capture spatial dynamics and interactions in each category. This enables deeper analysis through successive transformations, allowing the graph-based network to learn intricate patterns of spatial dynamic interactions at multiple levels of abstraction [171, 172]. Here, taking human geometric features as an example, the

operation of each GCN layer is formalized as:

$$H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right), \tag{4.1}$$

where $H^{(l)}$ represents the activation matrix at the lth layer ($H^{(0)} = \mathcal{HG}$ for the initial layer), A is the adjacency matrix defining the graph structure, $W^{(l)}$ is the weight matrix for the lth layer, and σ is the Tanh activation function.

For an n-layer GCN, this transformation is applied iteratively to obtain the final embedded human geometric features:

$$HG' = H^{(n)} = \sigma\left(AH^{(n-1)}W^{(n-1)}\right) \tag{4.2}$$

where n is the total number of GCN layers, iterating the process from $l = 0$ to $n - 1$. We choose $n = 4$ based on empirical experimental results. Through this operation, we can obtain the embedded human and object geometric features: $HG' \in \mathbb{R}^{T\times HJ\times C_2}$ and $OG' \in \mathbb{R}^{T\times 2O\times C_2}$.

Visual Features

In contrast to geometric features, visual features in videos offer a wealth of contextual information and essential feature representations. Following [5, 146], we derive 2048-dimensional visual features of entities from Region of Interest (ROI) pooled 2D bounding boxes around humans and objects in video frames. As shown in Fig. 4.2, they are subsequently reduced dimensionally to C_1 through an MLP with learnable embeddings and aligned dimensionally with geometric features. This process results in the embedded human and object visual features: $HV' \in \mathbb{R}^{T\times HJ\times C_1}$ and $OV' \in \mathbb{R}^{T\times 2O\times C_1}$.

Multi-Modality Fusion

Finally, we fuse embedded geometric and visual features in the human and object keypoint channel, producing new enriched human and object feature representations,

respectively:

$$\widetilde{H} = HG' \oplus HV' \in \mathbb{R}^{T \times HJ \times C_3}; \tag{4.3}$$

$$\widetilde{O} = OG' \oplus OV' \in \mathbb{R}^{T \times 2O \times C_3}, \tag{4.4}$$

where $\oplus$ represents concatenate operation and $C_3 = C_1 + C_2$. This refined fusion of geometric and visual cues creates a richly contextualized blend, laying a solid foundation for enhanced scenery graph learning of HOIs.

4.2.2 Scenery Interactive Graph

To effectively model the interactions between humans and objects, the existing method [5] focuses exclusively on their visual features to construct an interaction graph. This approach taps into the visual aspect of interactions, which is essential but insufficient for grasping the dynamic spatial relationships critical to understanding the complexities of HOI. Furthermore, 2G-GCN [146] offers a more comprehensive view but fuse geometric features representing all entities with visual features representing individuals, which results in hierarchical misalignment and fails to explicitly learn HOIs.

To overcome the constraints of prior approaches, we propose a scenery interactive graph that adopts a graph attention neural network to learn interactions between different categories with enriched feature representation (Fig. 4.1), to deeply understand the interaction dynamics among all humans and objects. This structured approach facilitates a comprehensive understanding of interactions within various contexts.

GAT for Learning Scenery Graph

Specifically, we adopt Graph Attention Networks (GAT) [173] in learning scenery graph interactions is particularly advantageous due to their ability to dynamically adjust to rapid changes in human and object interactions within scenery graphs, thanks to their adaptive edge weighting and handling of non-static features. This ensures a precise focus on relevant entities and their evolving relationships, optimizing

the model's responsiveness to the complex dynamics of interactions.

We construct the HOI scenery graph $\mathcal{G}_{s-t} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \mathbb{R}^{T\times(HJ+2O)\times C_3}$ represents the node features, which is obtained by concatenating the local human feature representation $\widetilde{H}$ and object feature representation $\widetilde{O}$, and $\mathcal{E} \in \mathbb{R}^{T\times(HJ+2O)\times(HJ+2O)}$ denotes the initialized fully-connected adjacency matrix. For each node $\mathcal{V}_i$ at time step $t \in [1, \dots T]$, the feature representation is:

$$\mathcal{V}_i^t = \sigma\left(\sum_{j\in\mathcal{N}_{(i)\cup i}} \alpha_{i,j}^t \mathbf{\Theta}\mathcal{V}_j^t\right), \tag{4.5}$$

and the attention coefficients $\alpha_{i,j}$ are computed as:

$$\alpha_{i,j}^t = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{W}^\top[\mathbf{\Theta}\mathcal{V}_i^t, \|, \mathbf{\Theta}\mathcal{V}_j^t]\right)\right)}{\sum_{n\in\mathcal{N}_{(i)\cup i}}\exp\left(\text{LeakyReLU}\left(\mathbf{W}^\top[\mathbf{\Theta}\mathcal{V}_i^t, \|, \mathbf{\Theta}\mathcal{V}_n^t]\right)\right)}, \tag{4.6}$$

where $\mathbf{\Theta}(\cdot)$ is the transformation function, $\mathcal{N}(\cdot)$ is the neighbor set of node i and $\mathbf{W}$ represents learnable parameters. This dynamic weighting is crucial as it allows the model to adaptively focus on the most relevant nodes and edges, reflecting the changing nature of interactions and relationships within the scene.

RNN-based Network for Learning Temporal Dependency

After obtaining the learned HOI scenery graph representations at each time step t, we employ an RNN-based network to learn the temporal dependencies across all the time steps. Specifically, we utilize a Bi-direction Gated Recurrent Unit (Bi-GRU) [174] that enables our model to integrate both past and future contexts, enhancing its understanding of the sequential dynamics in HOIs. The GRU's gating mechanisms effectively manage long-term dependencies, ensuring robust temporal modeling. For the learned step-wise feature representations, we utilize a Gumbel-Softmax module [175], enabling precise and adaptable delineation of sub-event lengths in video sequences. This module is instrumental in enabling gradient-based optimization while maintaining probabilistic integrity in segmenting actions, a crucial aspect when dealing with the inherently fluctuating characteristics of video content. Subsequently, we employ another Bi-GRU to discern the temporal relations among

segmented sub-actions. The processed features are then leveraged to identify specific sub-activities/affordances associated with entities, with the granularity of recognition tailored to suit the requirements of the specific dataset.

4.3 Experiments

4.3.1 Datasets

We evaluate CATS on two datasets: MPHOI-72 [146] and CAD-120 [19], showcasing the superior results on multi-person and single-person HOI recognition.

The MPHOI-72 dataset is valuable for two-person HOI tasks. It contains 72 videos of 8 pairs of people performing 3 distinct activities (*Cheering*, *Hair cutting* and *Co-working*) with 13 human sub-activities (e.g., *Sit*, *Pour*). Each video showcases two participants interacting with 2-4 objects from 3 unique angles. Geometric features and human sub-activities labels are frame-wise annotated.

CAD-120 is a prominent dataset for single-person HOI recognition. It contains 120 RGB-D videos, capturing 10 distinct activities executed by 4 participants, each repeated three times. In each video, a participant interacts with 1-5 objects. The dataset provides frame-wise annotations for 10 human sub-activities (e.g., *opening*, *placing*).

4.3.2 Evaluation Protocol

Following the evaluation protocol of [5, 146], we assess CATS on the task of joint segmentation and label recognition, which requires the model to segment the timeline for each entity in a video and assign labels to each segment. The $F_1@k$ metrics [145] with the commonly used thresholds $k = 10\%$, 25% and 50% are reported. For the MPHOI-72 and CAD-120 datasets, we use leave-two-subject and leave-one-subject cross-validation to evaluate the generalization effort of CATS in unknown subjects, respectively.

Table 4.1: Joined segmentation and label recognition on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN [5]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN [146]	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5
CATS	**71.3 ± 5.0**	**65.8 ± 3.9**	**48.8 ± 5.3**

4.3.3 Network Setting

The visual features of humans and objects are extracted from 2D bounding boxes within the video using a Faster R-CNN module [83] that has been pre-trained [158] on the Visual Genome dataset [39]. For multi-modality fusion, we set $C_1 = 512$ and $C_2 = 256$, resulting in a fused dimension of $C_3 = 768$, which supports varied feature dimensions as shown in Fig. 4.2.

4.3.4 Quantitative Comparison

Multi-person HOIs

In the MPHOI-72 dataset, results in Table 4.1 demonstrate CATS surpasses the previous state-of-the-art models, ASSIGN [5] and 2G-GCN [146], showcasing significant performance improvements. This is highlighted by CATS's superior performance across all F1 configurations coupled with substantially lower standard deviations. Specifically, in the F1@10 score, CATS achieves 71.3%, which is approximately 3% and 12% higher than 2G-GCN and ASSIGN, respectively, marking a clear advancement in both predictive accuracy and consistency in the domain of HOI recognition. These experimental outcomes further underscore the significance of geometric features in the Multi-Person Human-Object Interaction (MPHOI) domain.

Single-person HOIs

In the CAD-120 dataset, as presented in Table 4.2, CATS does not perform as reliably as in the multi-person HOI scenario. For human sub-activity labeling task, CATS performs competitively and surpasses various prior methods, including those reliant on visual features like ATCRF [160] and [5], as well as the more sophisticated visual-

Table 4.2: Joined segmentation and label recognition on CAD-120.

Model	Sub-activity			Object Affordance		
	F_1@10	F_1@25	F_1@50	F_1@10	F_1@25	F_1@50
rCRF [116]	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3	72.1 ± 2.5	69.1 ± 3.3	57.0 ± 3.5
Independent BiRNN	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8	84.6 ± 2.1	81.5 ± 2.7	71.4 ± 4.9
ATCRF [160]	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3	79.9 ± 3.1	77.0 ± 4.1	63.3 ± 4.9
Relational BiRNN	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5	82.3 ± 2.3	78.5 ± 2.7	68.9 ± 4.9
ASSIGN [5]	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8	92.0 ± 1.1	90.2 ± 1.8	82.4 ± 3.5
2G-GCN [146]	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8	**92.4** ± 1.7	**90.4** ± 2.3	**82.7** ± 2.9
CATS	**89.6** ± 2.1	**87.3** ± **1.5**	76.0 ± 3.5	90.2 ± 1.5	89.1 ± 2.4	80.5 ± 2.8

geometric approach offered by 2G-GCN [146]. Notably, CATS secures competitive performance in both F1@10 and F1@25 metrics, registering improvements of 1.6% and 0.1% over ASSIGN and 2G-GCN, respectively. This achievement underscores CATS's capability to accurately model and predict the dynamics of human actions.

However, it cannot accurately recognize object affordances in the single-person HOI, which drops by about 1%-2% in all F1 configurations. This outcome may stem from two primary factors. First, there is an imbalance in feature representation, with fewer keypoints representing objects compared to humans. For instance, while a human entity may be defined by nine keypoints, an object is often represented by only two. This disparity may reduce the model's emphasis on objects within the scene graph, potentially leading to suboptimal attention to object affordances. Second, our task involves both segmentation and label recognition, a two-stage process requiring the model to first divide the timeline and then assign appropriate labels. Such a task may not align well with an end-to-end framework, as it demands distinct stages of processing that the current setup may struggle to accommodate effectively.

4.3.5 Qualitative Comparison

In this section, we present a qualitative comparison of CATS with the state-of-the-art method across the MPHOI-72 and CAD-120 datasets.

Fig. 4.3 and Fig. 4.4 illustrate *Cheering* and *Hair Cutting* activities within the MPHOI-72 dataset, comparing the segmentation and labeling tasks performed by CATS and 2G-GCN [146] against the ground truth. Significant segmentation errors are marked with red dashed boxes. Although both methods exhibit some discrepancies

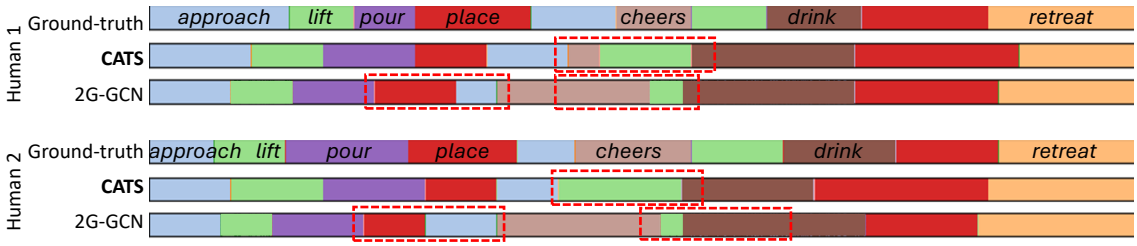

Figure 4.3: Visualization of segmentation on MPHOI-72 for *Cheering* activity. Red dashed boxes highlight major segmentation errors.

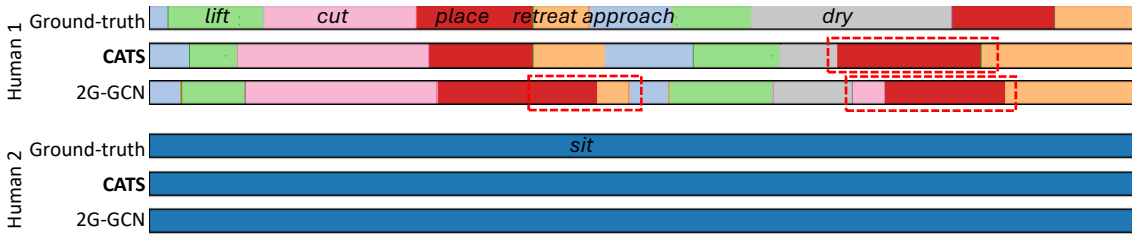

Figure 4.4: Visualization of segmentation on MPHOI-72 for *Hair cutting* activity. Red dashed boxes highlight major segmentation errors.

in their predictions, CATS more closely aligns with the ground truth, offering a more precise and stable visualization across a variety of actions. Conversely, 2G-GCN is prone to generating inappropriate sub-activities such as *cheers* and *lift* in the *Cheering* activity. Moreover, in the *Hair Cutting* activity, 2G-GCN oversegments the *cut* sub-activity into *dry* sub-activity, further deviating from the expected interaction dynamics. This segmentation error may occur due to the model's inability to effectively leverage contextual clues, causing misclassification of fine-grained actions when subtle motion and interaction changes are not recognized as interconnected within the overall activity. While CATS provides category-to-scenery information, it can handle these challenges by leveraging contextual awareness to maintain coherence across fine-grained actions.

Fig. 4.5 and Fig. 4.6 illustrate the *Cleaning Objects* and *Making Cereal* activities from the single-person CAD-120 dataset, with abnormal segmentation instances accentuated by red dashed boxes. For the *Cleaning Objects* activity, both methods effectively match the overall ground truth. However, CATS provides a visualization that more closely approximates the ground truth. In the *Making Cereal* activity, CATS significantly outperforms 2G-GCN, particularly in sub-activities such as

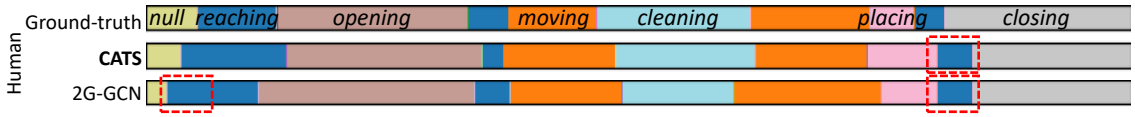

Figure 4.5: Visualization of segmentation on CAD-120 for *Cleaning objects* activity. Red dashed boxes highlight major segmentation errors.

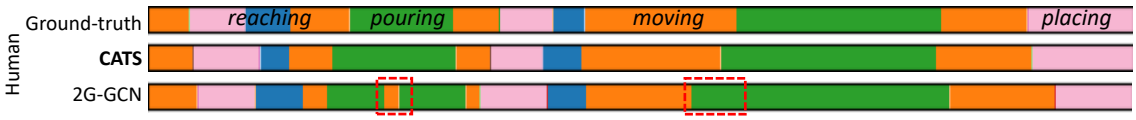

Figure 4.6: Visualization of segmentation on CAD-120 for *Making Cereal* activity. Red dashed boxes highlight major segmentation errors.

pouring, *moving*, and *reaching*, while 2G-GCN yields some inaccurate segmentations and over-segments.

4.3.6 Alternative Architectures and Ablation Studies

Architecture Alternatives Comparison

We evaluate the HOI recognition performance on the MPHOI-72 and CAD-120 datasets by conducting tests on various alternative model structures. The experimental outcomes, as detailed in Tables 4.3 and 4.4, reveal that our model consistently delivers superior results compared to these alternatives. This superior performance is likely attributable to the unique consideration our model gives to category-level interactions, specifically the distinct analysis of human-human and object-object interactions. Unlike other approaches that might treat interactions generically or overlook the nuanced distinctions between different types of interactions, our model maintains a comprehensive view.

Table 4.3: Comparison between architecture alternatives and CATS on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
Independent-entity architecture	65.1 ± 3.3	58.7 ± 1.7	40.4 ± 3.9
2G-GCN [146]	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5
CATS	71.3 ± 5.0	65.8 ± 3.9	48.8 ± 5.3

Table 4.4: Comparison between architecture alternatives and CATS on CAD-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
Independent-entity architecture	85.9 ± 4.0	84.1 ± 4.9	72.8 ± 5.2
2G-GCN [146]	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8
CATS	89.6 ± 2.1	87.3 ± 1.5	76.0 ± 3.5

Table 4.5: Results of different GCN layers in multi-category multi-modality fusion on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
1-layer GCN	70.4 ± 1.7	62.0 ± 2.5	43.9 ± 3.8
2-layer GCN	68.8 ± 4.3	62.1 ± 4.3	44.0 ± 3.3
3-layer GCN	67.4 ± 4.2	63.3 ± 3.4	44.2 ± 1.3
5-layer GCN	70.4 ± 5.7	60.0 ± 2.3	43.7 ± 2.2
4-layer GCN (Ours)	**71.3 ± 5.0**	**65.8 ± 3.9**	**48.8 ± 5.3**

GCN Layers for Geometric Feature Learning

In this section, we conduct ablation studies to elucidate the impact of the depth of GCN layers on the geometric learning of human joints and object keypoints within our network, results are shown in Table 4.5. To assess the influence of GCN layer depth on model performance, we explore configurations with 1, 2, 3, 4, and 5 GCN layers. Through this comparative analysis, we aim to identify the most effective layer depth that balances computational efficiency with the nuanced understanding of spatial relationships essential for interpreting complex interactions between humans and objects. The results indicate that a configuration of 4-layer GCN offers the optimal balance, providing the best performance in terms of both accuracy and computational efficiency. This depth allows for sufficient complexity to understand and model the geometric relationships critical for accurate interaction recognition, without incurring the diminishing returns or increased computational demand associated with additional layers.

4.4 Summary

In summary, we propose CATS, an end-to-end framework that enhances video-based HOI recognition through sophisticated integration of category-level and scenery-level analyses. It first fuses multi-modal features of different categories, and then constructs a scenery interactive graph to learn the relationships between these categories. CATS achieves superior results on the multi-person HOI benchmark and delivers comparable performance on the single-person HOI benchmark.

Based on insights gained from CATS, we decide to move away from the end-to-end framework and return to a two-stage approach, as seen in models like ASSIGN [5] and 2G-GCN [146]. We recognize a persistent challenge in achieving more effective multimodal feature fusion and accurately capturing the complex dynamics of multi-person interactions in intricate environments. In the next chapter, we focus on maximizing the representational capabilities of features for each entity, while enabling simultaneous representation of pairwise interactions and complex interdependencies among entities.

CHAPTER 5

Geometric and Visual Feature Fusion in Multi-Person HOI

Portions of this chapter have been submitted to the peer-reviewed venue in the following publication:

- **Qiao, T.**, Li, R., Li, F. W., Kubotani, Y., Morishima, S., & Shum, H. P., "GeoVis-GNN: Geometric Visual Fusion Graph Neural Networks for Multi-Person Human-Object Interaction Recognition in Videos." 2024.

Video-based multi-person HOI recognition is crucial for understanding complex human behaviors in real-world scenarios. Prior work has largely underutilized the potential of effectively fusing geometric and visual features of humans and objects. Recognizing the multimodality of features across various entities within HOI scenes, we propose GeoVis-GNN to learn representative features following a bottom-up approach. It first optimizes the feature representation of each entity by a dual-attention feature fusion mechanism at the feature level, then models explicit interactions and implicit interdependencies among entities and neighbors through an interdependent entity graph. To demonstrate the novelty and effectiveness of our method, we propose a new open-access multi-person HOI dataset (MPHOI-

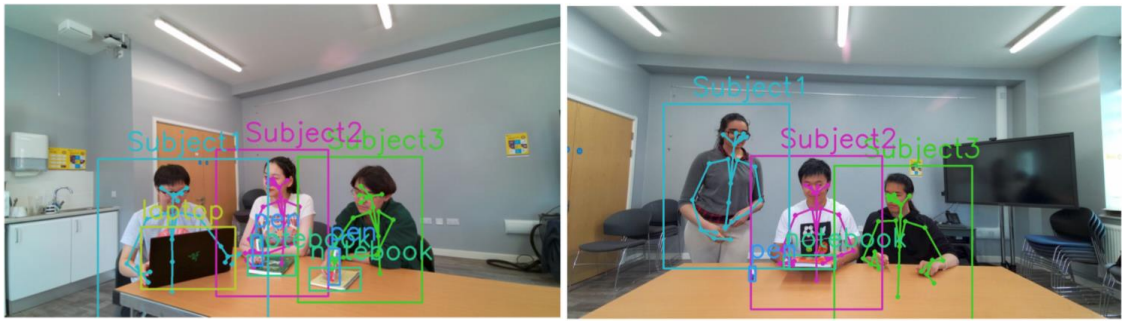

Figure 5.1: Two examples (*Teaching* and *Signing*) of our collected three-person HOI datasets. Geometric features such as skeletons and bounding boxes are annotated.

120), which includes three humans interacting with multiple objects in daily life, surpassing existing datasets in terms of complexity and human participation. Our method outperforms state-of-the-arts across multiple datasets, including MPHOI-120 (three-person HOI), MPHOI-72 (two-person HOI), CAD-120 (single-person HOI) and Bimanual Actions (two-hand HOI) datasets.

5.1 Introduction

Video-based multi-person HOI recognition focuses on understanding the dynamic interactions between multiple individuals and various objects within the scene. This task requires models to interpret complex temporal and spatial relationships, discerning various actions despite challenges like frequent occlusions, overlapping actions, and diverse interaction patterns. The previous effort of 2G-GCN [146] first introduces graph-based geometric features to video-based HOI recognition, extending beyond single-person scenarios to two-person HOI with a new dataset. CATS [163] presents an end-to-end framework to learn geometry-informed HOI from category to scenery.

Video-based HOI recognition faces challenges with aligning and fusing multimodal features in a meaningful way. 2G-GCN [146] merges geometric features representing all entities, with visual features for individual entities. This may lead to misalignment of feature hierarchies, as high-level spatial information may not align well with detailed, entity-specific visual data. Consequently, this inefficiency in feature fusion compromises the model's ability to distinguish between entities effectively. Directly

concatenating visual and geometric features in CATS [163] treats all feature types uniformly, potentially diluting the unique contributions of each. This approach may overemphasize dominant features while underestimating subtle geometric cues, reducing accuracy in recognizing fine-grained interactions.

In this work, we propose to learn HOI features following a bottom-up approach ranging from the feature-level fusion to the entity-level graph. Our insight is that it allows for more nuanced and effective integration of diverse feature types, and ensures that each level is optimally exploited for maximum representational power. We present a novel Geometric Visual Fusion Graph Neural Network (GeoVis-GNN) that first optimizes the feature representation of each entity by a dual-attention geometry-visual fusion. These enriched entity-specific representations are then fed into the interdependent entity graph to further model explicit interactions and implicit interdependencies.

To provide the context of truly multi-person multi-object interactions, we introduce a novel multi-person human-object interaction dataset, MPHOI-120. This dataset surpasses existing ones [19, 146] in terms of the number of humans and objects involved, capturing intricate scenarios where three individuals interact with multiple objects in daily life (Fig. 5.1). The frequent occlusions between humans and objects add to its complexity. Frame-wise annotations include geometric features of human skeletal poses, bounding boxes for humans and objects, and ground-truth HOI sub-activity labels.

Our approach surpasses state-of-the-art performance on multiple datasets, including the proposed three-person HOI MPHOI-120 dataset, the two-person MPHOI-72 [146] dataset and the single-person HOI CAD-120 [19] dataset and the two-hand Bimanual Actions [62] dataset. Additionally, we conduct ablation studies to evaluate the core components of our model. Our main contributions are:

- A bottom-up framework GeoVis-GNN for multi-person HOI recognition in videos, which first optimizes the feature representation of each entity by fusing geometric and visual features at the feature level while learning entity interactions at the entity level.

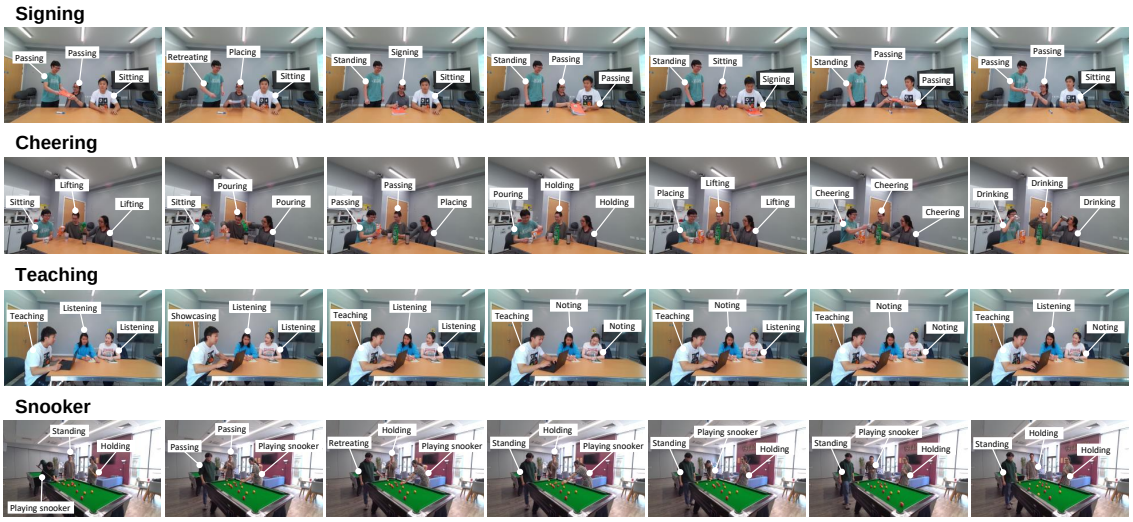

Figure 5.2: Sample video screenshots from our new MPHOI-120 dataset, displaying annotated labels for sub-activities along the timeline of four different multi-person HOI activities.

- A dual-attention fusion mechanism optimizes multimodal feature integration by embedding and adaptively fusing visual and geometric features to create a rich, entity-specific representation.
- A interdependent entity graph simultaneously models explicit interactions between independent entities and implicit interdependencies among neighboring entities surrounding a specific entity.
- A new MPHOI-120 showcases three humans interacting with multiple objects in daily life[1]. It contains more humans and objects than any other existing HOI datasets, making it especially challenging due to the frequent body occlusions between humans and objects.

5.2 The Three-Person HOI Dataset

Most video-based HOI datasets mainly record single-person HOIs, although they provide different perspectives [19, 73, 75, 76]. Efforts to portray multiple human interactions remain nascent, with the UCLA HHOI [176, 177] dataset capturing

[1]Data collection performed in the UK, under Durham University Ethics Approval Ref: COMP-2022-06-03T19_29_22-cbmw62.

interactions of up to two people and one object, and MPHOI-72 [146] slightly extending this to two people and several objects. However, these collections do not fully capture the full complexity of multi-person and multi-object interactions in different real-world scenarios.

We propose a new multi-person HOI dataset MPHOI-120 to address this gap by capturing more intricate daily activities involving three humans and various objects, approximating real-world HOIs. The inclusion of an additional person and more objects exponentially escalates the complexity, not only multiplying the possibilities of human-human, human-object, and object-object interactions but also significantly increasing occlusions among the entities. Thus, MPHOI-120 presents a far more intricate challenge than traditional two-person HOI datasets.

5.2.1 Dataset Details

MPHOI-120 includes 120 videos captured from 3 different angles at 30 fps, with 1920×1080 resolution and an average length of 15 seconds. We have 4 males and 3 females randomly combined into 10 groups with 3 people per group and perform 4 different HOI activities interacting with 2 to 5 objects. There are 6 objects: notebook, pen, bottle, cup, laptop and snooker cue. In total, 4 activities = {*Signing*, *Cheering*, *Teaching* and *Snooker*} and 17 sub-activities = {*Sit*, *Stand*, *Pass*, *Approach*, *Retreat*, *Lift*, *Place*, *Note*, *Shake hands*, *Hold*, *Pour*, *Drink*, *Cheers*, *Showcase*, *Teach*, *Listen*, *Play snooker*} are defined. The sub-activity label of each subject is annotated frame-by-frame. Fig. 5.2 shows some sample video screenshots with annotated sub-activities along the timeline of all activities.

Signing is a fairly complex activity that simulates three people signing a contract. It represents that Subject1 first stands holding a notebook and passes the notebook to Subject 2; Subject 2 signs with a pen then passes the notebook and pen to Subject 3; Subject 3 further signs on the notebook, and then passes them back from Subject 2 to Subject 1; Subject 2 and Subject 3 finally shake hands. Massive high-level occlusions exist between humans, notebooks and pens during the entire activity.

Cheering is one of the most common multi-person activities. We provide two bottles and three cups for the three subjects to enhance the interaction between

human-human and human-object. In the beginning, two randomly chosen subjects each pour juice from bottles into their cups. One of these subjects then hands the bottle to a third subject without a bottle, who also pours juice into his/her cup. Finally, all subjects raise their cups in a collective cheering gesture and drink the juice.

Teaching simulates the interaction between a teacher and students, where a randomly selected subject plays the teacher, using a laptop to demonstrate and talk with the other two subjects (students), each equipped with a pen and a notebook. This activity requires the system to carefully distinguish the dynamic state of whether the subject is holding a pen and writing in a notebook to determine whether the subject is *noting* or *listening*.

Snooker is a rather interesting activity in this work, three subjects are designed to play snooker together with only 2 snooker cues in one pool table. Three subjects take turns to serve, and after a subject finishes, he/she will pass the snooker cue to the subject who is without one. This setup results in significant body occlusion among the subjects, particularly when a player navigates around the pool table to position themselves for their shot.

Leveraging the Azure Kinect SDK along with the Body Tracking SDK [151], we acquire RGB-D videos to capture the comprehensive dynamics of multiple individual skeletons. We offer 2D human skeletal data and bounding boxes for both subjects and objects within each video, serving as geometric characteristics. The integration of depth information within our dataset further broadens its utility, such as versatile benchmarks for 3D human pose estimation [178, 179] and 3D object estimation [180, 181], among other applications.

5.2.2 Statistical Comparison of Datasets

We perform a statistical comparison between MPHOI-120 and existing HOI datasets, as shown in Tab. 5.1. MPHOI-120 includes scenarios with three people interacting and 17 sub-activities, which is higher than any other listed dataset, standing out for its complexity and richness. With a substantial 53,604 frames across 120 videos, it provides ample data for training robust models. Additionally, the high video

Table 5.1: A statistical comparison between MPHOI-120 and existing HOI datasets.

Datasets	MPHOI-120	MPHOI-72 [146]	CAD-120 [19]	Bimanual Actions [62]
No. people interacting	3	2	1	1
Total videos	120	72	120	540
Total frames	53604	26383	61585	221000
Video average length	15s	12s	17s	15s
No. sub-activities	17	13	10	14
No. subjects/objects	7/6	5/6	4/10	6/12
Total activities	4	3	10	9
Fps	30	30	30	30
Resolution	1920×1080	3840×2160	640×480	640×480

resolution (1920 × 1080) ensures detailed feature capture, essential for advanced HOI analysis. In contrast, although Bimanual Actions [62] is large, it is limited to dual-hand movements of an individual, leading to a more monotonic data distribution.

5.3 Methodology

We propose a bottom-up approach to design GeoVis-GNN, which first optimizes the feature representation of each entity by fusing geometric and visual features at the feature level, then proceeds to learn interactions in the entity-level graph. The bottom-up approach has been widely used in pose estimation [182–184] and object detection [185–187] tasks with considerable performances. It ensures a thorough understanding of the fundamental aspects of each entity before delving into complex entity-level interactions. This approach, starting from basic features and building upwards, enables detailed feature integration to achieve more effective entity interaction analysis.

Alternative design performs suboptimally, a top-down approach begins with a broad view of entity-level relationships before focusing on specific entity features can lead to missed crucial interaction details and a misalignment between overarching patterns and individual interaction nuances. An alternative method [146] that combines feature fusions and entity graph learning within a single graph entangles the entity concept, lacking a specific feature to represent each entity, which fails to learn HOIs explicitly. We compare these alternative architectures with our method in Experimental Results 5.4.

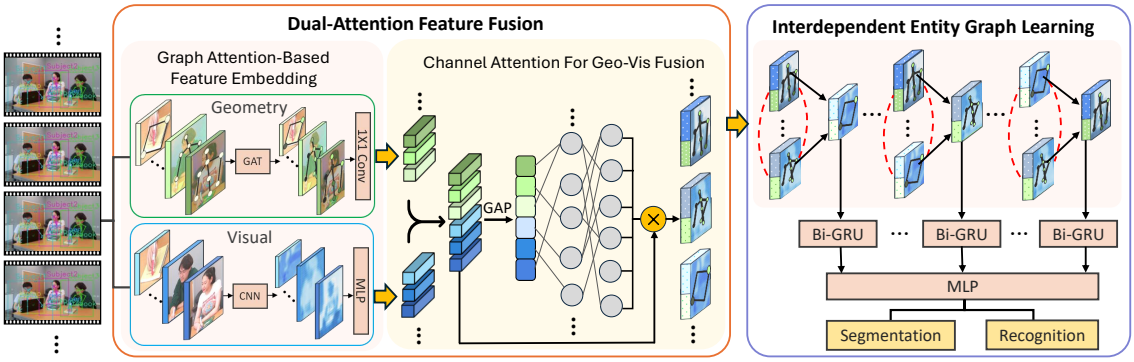

Figure 5.3: Overview of our bottom-up framework GeoVis-GNN. We first design a dual-attention fusion for feature optimization, which embeds and fuses visual and geometric features in a graph attention-based mechanism and channel attention module, respectively. The enriched entity-specific representations are then inputted into the interdependent entity graph to further model explicit interactions and implicit interdependencies. Finally, we apply a BiGRU to capture the temporal dependencies to obtain segmentation and recognition results.

5.3.1 Dual-Attention Fusion for Feature Optimization

Prior approaches of CNN-based HOI recognition in videos mainly focused on visual features [5, 155, 156], which may prove inadequate in occluded scenarios. Advanced methods such as 2G-GCN [146] attempt to integrate geometric features within a GCN framework to augment visual data. However, their fusion of collective geometric features with individual visual features risks hierarchical misalignment, fusion inefficiencies, and difficulties in entity distinction. CATS [163] also employs GCN to model geometric features but directly combines them with visual features, which may dilute their distinct contributions.

We propose a dual-attention fusion mechanism at the feature level to optimize multimodal feature integration (Fig. 5.3). It first embeds visual and geometric features of each entity to establish a rich multimodal context, with geometric features learned through a graph attention-based mechanism. A second channel attention module then adaptively emphasizes informative features while suppressing less relevant ones and fuses them, ensuring a targeted amalgamation of the corresponding multimodal features per entity. This results in a well-contextualized feature representation that effectively blends geometric and visual cues, providing a robust foundation for subsequent entity-level graph learning.

Graph Attention-Based Feature Embedding

Previous research [109, 146, 163] learns geometric features by GCNs, which typically apply the same convolution operation to all neighbors of a node, without distinguishing between the different roles or importance that different neighbors might play in the context of multi-person HOIs. This can lead to a homogenization of features that fails to capture the complexity of multi-entity dynamics.

We propose a graph attention-based [188] embedding to learn multi-entity geometric features, adaptively weighting the importance of each entity's geometric features through an attention mechanism to capture the evolving significance of interactions. This enables the model to expertly handle occlusions and dynamic environments for multi-person HOI recognition.

For feature representation, following previous successes [146, 163], we concatenate the position and velocity of all entities into keypoint channels, forming geometric features $\mathcal{G} = \{g_t^{e,k}\}_{t=1,e=1,k=1}^{T,E,K} \in \mathbb{R}^4$ with $g_t^{e,k}$ as the k-th type features for entity e at frame t, where T denotes the total number of frames in the video, E and K denote the total number of entities and keypoints of an entity in a frame, respectively. Human joints and object bounding box diagonals are extracted as keypoints.

We adaptively infer spatial correlations with a GAT among keypoints k_1 and k_2 for a single timestep among entities as follows:

$$\mathbf{g}_t^s = \alpha_{k_1,k_1}\mathbf{\Theta}\mathbf{g}_{t,k_1} + \sum_{k_2 \in \mathcal{K}} \alpha_{k_1,k_2}\mathbf{\Theta}\mathbf{g}_{t,k_2}, \tag{5.1}$$

and the attention coefficients α_{k_1,k_2} are computed as:

$$\alpha_{k_1,k_2} = \frac{\exp\left(\Gamma\left(\mathbf{a}^\top[\mathbf{\Theta}\mathbf{g}_{k_1} \parallel \mathbf{\Theta}\mathbf{g}_{k_2}]\right)\right)}{\sum_{k_3 \in \mathcal{K}\cup\{k_3\}} \exp\left(\Gamma\left(\mathbf{a}^\top[\mathbf{\Theta}\mathbf{g}_{k_1} \parallel \mathbf{\Theta}\mathbf{g}_{k_3}]\right)\right)}, \tag{5.2}$$

where $\mathbf{\Theta}$ and Γ are the transformation function and LeakyReLU activation, respectively.

$\mathbf{g}_t^s$ is then fused with a 1×1 convolution along the temporal channel to form spatial-temporal geometric features $\mathbf{g}_t^{st} \in \mathbb{R}^{T \times NK \times C_1}$, effectively summarizing temporal dynamics while avoiding the complexities of 3D convolutions. It is then reshaped

to $\mathbf{g}_t^{st} \in \mathbb{R}^{T\times N\times KC_1}$ and embedded by a Multi-Layer Perceptron (MLP) to get low-level entity geometric features $\mathbf{g}'_t \in \mathbb{R}^{T\times N\times C_2}$.

Unlike geometric features, visual features in videos contain rich contextual information and fundamental feature representations. Following [5, 146], we extract entity visual features $\mathbf{v}_{t,n} \in \mathbb{R}^{2048}$ from ROI pooled 2D bounding boxes of humans and objects in videos, utilizing a pre-trained Faster R-CNN [83] module on the Visual Genome [39]. They are subsequently aligned dimensionally with geometric features to $\mathbf{v}'_t \in \mathbb{R}^{T\times N\times C_2}$ through an MLP with learnable embeddings.

Geo-Vis Channel Attention-Based Feature Fusion

Incorporating geometric and visual features poses a significant challenge due to their inherent representation and scale discrepancies. Prior approaches have attempted multimodal fusion by element-wise addition [109] or feature concatenation [120, 163]. However, such direct operations are infeasible for our task as they do not account for the disparate nature of feature spaces, leading to suboptimal learning outcomes.

We propose a novel geometry-visual channel attention-based feature fusion to effectively integrate geometric and visual features of all humans and objects, which achieves selective feature enhancement and encourages complementarity between multimodal features. We exploit channel attention mechanisms [189] in geometry-visual channels of all entities, allowing adaptively emphasizing informative features while suppressing less relevant ones across different channels, which is particularly advantageous for learning more representative visual and geometric features in diverse HOI scenarios. For instance, visual features often suffer in noisy backgrounds but thrive in scenarios with small backgrounds. Geometric features demonstrate strength in addressing partial occlusions [146], which is a common situation in multi-person HOI scenarios.

Specifically, as shown in Fig. 5.3, our attention-based feature fusion module first concatenates $\mathbf{g}'_t$ and $\mathbf{v}'_t$ along the entity dimension to entity geometry-visual features $\mathbf{gv}_t \in \mathbb{R}^{T\times 2N\times C_2}$, and compute a channel attention A as:

$$A = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\left(GAP\left(\mathbf{gv}_t\right)\right)\right)\right), \tag{5.3}$$

where GAP denotes Global Average Pooling [190], $\mathbf{W}_1$ and $\mathbf{W}_2$ are weights of Fully-Connected (FC) layers, δ and σ represent the ReLU and Sigmoid activation. Apply these values to original features for attended geometry-visual fusion features:

$$\mathbf{gv}'_t = A \cdot \mathbf{gv}_t. \tag{5.4}$$

After assigning distinct weights to each geometry-visual channel of an entity, the weighted features are strategically split into separate geometric and visual streams. These are then adeptly fused back together, producing a new enriched entity representation $\widetilde{\mathbf{gv}}_t \in \mathbb{R}^{T\times N\times C_3}$. This refined feature fusion set, being a weighted and well-contextualized blend of geometric and visual cues, sets the stage for more discerning entity-level graph learning.

Compared to our attention-based feature fusion, [144, 191] apply Transformer to fuse geometric and visual features in image-based HOI detection, which is constrained in processing video data due to memory inefficiency. Graph-based feature fusion treats multimodal features as graph nodes [192, 193], which is heavily reliant on the design of graph representation. As HOI is a dynamic process, it is non-trivial to manually define an appropriate representation.

5.3.2 Interdependent Entity Graph

In HOI analysis, most approaches [5, 6, 9] construct an independent entity graph that assumes a fixed structure to decipher spatial interactions between entities focusing solely on visual features. 2G-GCN [146] represents geometric features of all entities as a single entity linked with visual features of object entities, failing to explicitly model interactions between all entities. CATS [163] learns interactions between human and object categories but neglects relationships between entities within the same category, which is particularly limiting in multi-person HOI scenarios.

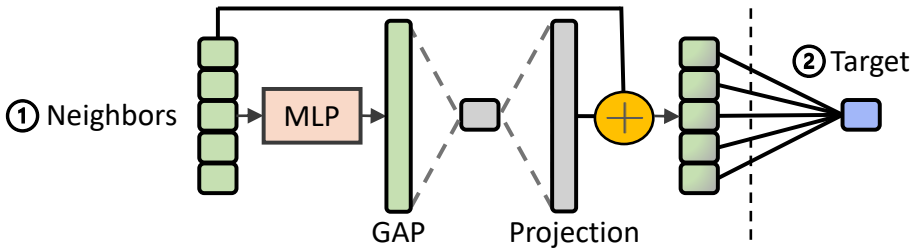

Figure 5.4: In the interdependent entity graph, we model neighbor features ① before aggregating them to the target entity ②.

Our insight is that an effective entity-level graph should not only capture explicit interactions among independent entities but also concurrently discern the implicit interdependencies that exist among neighboring entities surrounding a specific entity. This dual focus is crucial for understanding the intricate graph network of relations that exist around any specific entity within the scene.

To this end, we propose an interdependent entity graph to capture the interdependencies among all neighboring nodes around a particular entity with fused geometric and visual features, then refine it by applying attention weights between the entity in focus and its neighbors (Fig. 5.3 right). This entity-level graph offers a richer representation of spatial interactions, advancing the understanding of complex behavioral patterns beyond the reach of previous methods.

Specifically, as illustrated in Fig. 5.4, given a specific entity e at each frame t, we first calculate the features from its neighbor u to itself as follows:

$$S_t^u = \lambda \times \widetilde{\mathbf{gv}}_t^u + (1-\lambda) \times \frac{(\mathrm{GAP}(\mathbf{W}_3(\widetilde{\mathbf{gv}}_t^u)))}{N-1}, \tag{5.5}$$

where λ controls the contextual fusion threshold, and $\mathbf{W}_3$ is the weight of a FC layer. These neighboring features are then aggregated into a robust representation that encapsulates the collective attributes of the neighboring group:

$$S_t^e = STACK_{u \in N, u \neq e}(S_t^u \odot M(S_t^u)), \tag{5.6}$$

where $M(\cdot)$ is the mask indicator for valid neighbors and $\odot$ denotes element-wise multiplication. Meanwhile, we apply a dot-product attention mechanism [5, 194] to

obtain the attention weights between node e and its neighbors as:

$$W_t^e = \sum_{u \in N, u \neq e} Softmax(\frac{S_t^e (S_t^u)^T}{\sqrt{d}}), \tag{5.7}$$

where d is the feature dimension. Finally, the refined feature representation of the entity is $F_t^e = W_t^e \odot S_t^e$, ensuring a contextually aware integration of features.

After obtaining the fused features of each entity at each time step, we apply a Gumbel-Softmax module [175] to F_t^e, enabling precise and adaptable delineation of sub-event lengths in video sequences. It efficiently facilitates gradient-based learning and ensures probabilistically coherent segmentation, essential for handling the dynamic nature of video data. Finally, we apply a Bi-directional Gated Recurrent Unit (BiGRU) [174] to capture the temporal dependencies between each sub-action and then use the output features to recognize sub-activities for humans and object affordances for objects, varying according to the dataset.

5.4 Experimental Results

5.4.1 Datasets

We evaluate GeoVis-GNN on multiple datasets: MPHOI-120, MPHOI-72 [146], CAD-120 [19], and Bimanual Actions [62], showcasing the superior results on three-person, two-person, single-person and two-hand HOI recognition.

The MPHOI-72 dataset is valuable for two-person HOI tasks. It contains 72 videos of 8 pairs of people performing 3 distinct activities (*Cheering*, *Hair cutting* and *Co-working*) with 13 human sub-activities (*e.g.*, *Sit*, *Approach*, *Pour*). Each video showcases two participants interacting with 2-4 objects from 3 unique angles. Geometric features and human sub-activities labels are frame-wise annotated.

CAD-120 is a prominent dataset for single-person HOI recognition. It contains 120 RGB-D videos, capturing 10 distinct activities executed by 4 participants, each repeated three times. In each video, a participant interacts with 1-5 objects. The dataset provides frame-wise annotations for 10 human sub-activities (*e.g.*, *opening*, *cleaning*, *placing*) and 12 object affordances (*e.g.*, *openable*, *cleanable*, *placeable*).

The Bimanual Actions dataset is a large-scale collection of 540 RGB-D videos capturing HOIs using both hands. It documents the actions of 6 subjects who engage in 9 varied bimanual tasks, with each task performed 10 times. The dataset assigns 14 unique action labels to each hand, with frame-wise annotations for each entity within the videos.

5.4.2 Implementation Details

We follow [5, 146, 163] to evaluate GeoVis-GNN on the task: joined segmentation and label recognition, and report $F_1@k$ score [145] with standard thresholds of $k = 10\%$, 25%, and 50%.

For dataset evaluation, we use a leave-one-subject-out cross-validation method for CAD-120 and Bimanual Actions, and a leave-two-subjects-out approach for MPHOI-72. For MPHOI-120, our cross-validation scheme specifies three subjects not present in the training set as the test set. Training MPHOI-120, MPHOI-72, CAD-120 and Bimanual Actions on four Nvidia Titan RTX GPUs take 6, 4, 8 hours and 7 days respectively, while testing the entire set takes approximately 2, 2, 6 and 20 minutes respectively.

In the network configuration, we set $C_1 = 128$, $C_2 = 256$, and $C_3 = 512$ to support varied feature dimensions. As Bimanual Actions has a significantly more monotonic data distribution, we set $C_2 = 32$, $C_3 = 64$ and $[S_t^u = 0]$.

5.4.3 Quantitative and Qualitative Comparison with SOTAs

Three-person HOIs

In the MPHOI-120 dataset, GeoVis-GNN beats ASSIGN [5], 2G-GCN [146] and CATS [163] by a considerable gap (Tab. 5.2). Especially under multi-person HOI conditions, ASSIGN drops below 60% in F_1 metrics due to occlusions affecting visual features in HOI tasks. GeoVis-GNN shows an improvement of about 2% to 4% in $F_1@\{10, 25, 50\}$ over SOTA, highlighting that the effective dual-attention fusion strategy and interdependent entity graph enable our model to learn essential features and stable interactions even when unexpected occlusion and more complex

interactions appear.

Table 5.2: Joined segmentation and label recognition results on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN [5]	58.0 ± 8.5	53.7 ± 7.9	39.1 ± 7.4
2G-GCN [146]	60.7 ± 6.5	55.3 ± 6.9	39.6 ± 6.5
CATS [163]	62.8 ± 2.7	56.7 ± 4.2	42.8 ± 3.9
GeoVis-GNN	**65.1** ± 5.2	**59.8** ± 4.7	**46.6** ± 5.1

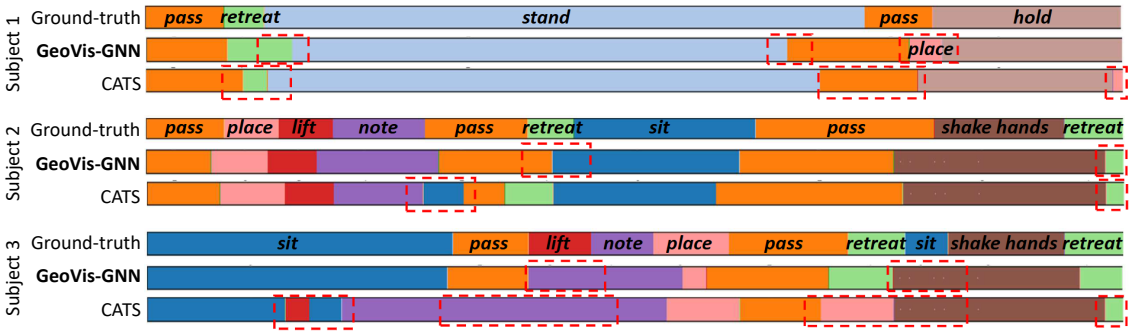

Figure 5.5: Visualization of segmentation on MPHOI-120 for *Signing* activity. Red dashed boxes highlight major segmentation errors.

Fig. 5.5 illustrates the visualization results of GeoVis-GNN and CATS on MPHOI-120 comparing with Ground-truth for the *Signing* activity, where red dashed boxes highlight major segmentation errors. Although both GeoVis-GNN and CATS make errors compared to Ground-truth, GeoVis-GNN can contribute relatively plausible segmentation results in all three subjects. For example, in subject 3, CATS over-segments *sit* in the beginning and then completely misses *pass* and *lift* before *note*, while our GeoVis-GNN can accurately segment *sit* and *pass* but miss *lift*. This is likely due to the *lift* action of the subject being very fast and closely resembles the *note* action, leading our model to misclassify *lift* as *note*. Incorporating temporal attention mechanisms could potentially enhance performance in the short duration of the action and its overlapping features with subsequent actions.

Two-person HOIs

GeoVis-GNN achieves an impressive performance on the MPHOI-72 dataset (Tab. 5.3), with an F_1@10 score of 84.3%, significantly outstripping the 71.3% scored by CATS [163]. Across all F_1 configurations, GeoVis-GNN exhibits substantial improvements of 13.0%, 10.8%, and 10.6%, respectively. The advanced technique for fusing geometric and visual features allows to capture more complex patterns in the data, while CATS and 2G-GCN cannot leverage it due to its inefficient fusion.

Table 5.3: Joined segmentation and label recognition results on MPHOI-72.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN [5]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN [146]	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5
CATS [163]	71.3 ± 5.0	65.8 ± 3.9	48.8 ± 5.3
GeoVis-GNN	**84.3** ± 5.5	**76.6** ± 4.5	**59.4** ± **4.9**

Fig. 5.6 shows the visualization of segmentation and labeling on the MPHOI-72 dataset with the two advanced models for the *Cheering* activity comparing with Ground-truth. GeoVis-GNN presents more reasonable and robust segmentation results in all sub-activities, while CATS provides some unexpected abnormal results in certain sub-activities, such as *pour* and *place*. Interestingly, CATS directly recognizes the static action *sit* rather than the ongoing action *retreat* following *place* at the end of the activity for subject 1. This may result from the dominant role of visual features, as these two actions appear similar in the front view.

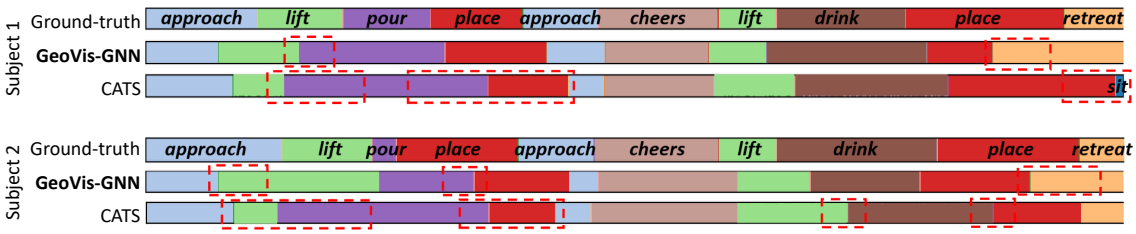

Figure 5.6: Visualization of segmentation on MPHOI-72 for *Cheering* activity. Red dashed boxes highlight major segmentation errors.

Single-person HOIs

Table 5.4: Joined segmentation and label recognition results on CAD-120.

Model	Sub-activity			Object Affordance		
	F_1@10	F_1@25	F_1@50	F_1@10	F_1@25	F_1@50
rCRF [116]	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3	72.1 ± 2.5	69.1 ± 3.3	57.0 ± 3.5
Independent BiRNN	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8	84.6 ± 2.1	81.5 ± 2.7	71.4 ± 4.9
ATCRF [160]	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3	79.9 ± 3.1	77.0 ± 4.1	63.3 ± 4.9
Relational BiRNN	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5	82.3 ± 2.3	78.5 ± 2.7	68.9 ± 4.9
ASSIGN [5]	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8	92.0 ± 1.1	90.2 ± 1.8	82.4 ± 3.5
2G-GCN [146]	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8	92.4 ± 1.7	**90.4** ± 2.3	82.7 ± 2.9
CATS [163]	89.6 ± 2.1	87.3 ± 1.5	76.0 ± 3.5	90.2 ± 1.5	89.1 ± 2.4	80.5 ± 2.8
GeoVis-GNN	**89.9** ± 2.0	**87.8** ± 1.9	**76.7** ± 3.1	**92.7 ± 0.4**	**90.4 ± 0.6**	**83.3 ± 1.8**

Tab. 5.4 shows the effectiveness of GeoVis-GNN in CAD-120 evaluated by sub-activity and object affordance labels. GeoVis-GNN beats previous visual-based [5, 116, 160] and geometry-informed [146, 163] networks for both labels and achieves the highest F_1 scores of mean in every configuration. Notably, the two geometry-informed networks show comparable performance in human sub-activity recognition, but CATS performs poorly in object affordance recognition. This may be due to two main factors: an imbalance in feature representation, with fewer keypoints for objects than humans, reducing object emphasis in the scene graph, while the dual-attention feature fusion in GeoVis-GNN helps mitigate this. Additionally, our task requires both segmentation and label recognition, a two-stage process that does not align well with the end-to-end framework of CATS, which may struggle with such distinct processing stages. Although CATS performs well in multi-person HOI scenarios, empirical results indicate that it is less suited for single-person HOI tasks. Therefore, in the subsequent HOI recognition comparisons involving a single individual, we use 2G-GCN as the state-of-the-art benchmark.

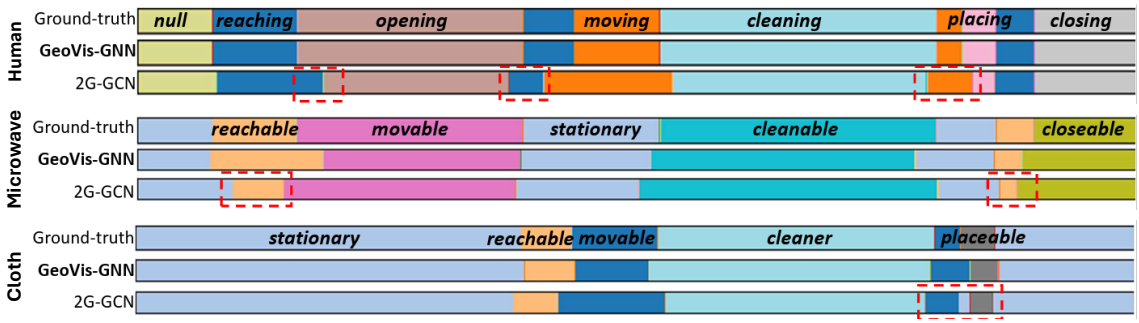

Figure 5.7: Visualization of segmentation on CAD-120 for *Cleaning objects* activity. Red dashed boxes highlight major segmentation errors.

Fig. 5.7 presents the visualization outcomes for the *Cleaning Objects* activity in CAD-120, depicting a scene where a person uses a cloth to clean a microwave. The qualitative analysis shows that GeoVis-GNN surpasses 2G-GCN in recognizing human sub-activities and object affordances, notably *reachable* and *movable* for the microwave, closely matching the Ground-truth.

Two-hand HOIs

GeoVis-GNN achieves the superior performance on the large-scale Bimanual Actions dataset (Tab. 5.5), with near 1% improvement in the same standard deviation at F_1@10. The slight improvement is partly due to the limited hand pose estimation that OpenPose [22] uses for the hand skeleton of the dataset, which may introduce noise, especially in occlusions. Fig. 5.8 presents the visualization outcomes for the *Pouring* activity in Bimanual Actions. The qualitative analysis demonstrates that GeoVis-GNN has outstanding performance in segmenting and recognizing actions of both hands, which almost overlaps the Ground-truth, while 2G-GCN over-segments some sub-activities like *pour*.

Table 5.5: Joined segmentation and label recognition results on Bimanual Actions.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
Dreher *et al.* [62]	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
Independent BiRNN	74.8 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN [5]	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN [146]	85.0 ± 2.2	82.0 ± 2.6	69.2 ± 3.1
GeoVis-GNN	**85.8 ± 2.2**	**82.7** ± 2.8	**69.7 ± 3.0**

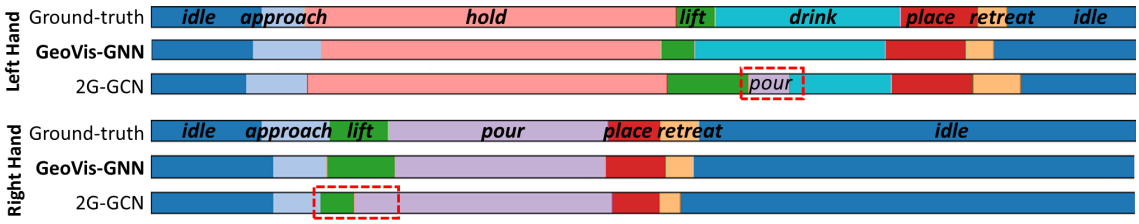

Figure 5.8: Visualization of segmentation on Bimanual Actions for *Pouring* activity. Red dashed boxes highlight major segmentation errors.

5.4.4 Ablation Study and Alternative Architecture

We extensively evaluate the design of channel attention-based feature fusion. Fig. 5.9 shows four design strategies, in which: **(a)**: Separately concatenate human features *hv, hg* and object features *ov, og* on feature-channel with attentions; **(b)**: Separately concatenate human features *hv, hg* and object features *ov, og* on entity-channel with attentions; **(c)**: Separately concatenate visual features *hv, ov* and geometric features *hg, og* on entity-channel with attentions; **(d) Ours**: Concatenate all features *hv, hg, ov, og* on entity-channel with a unified attention. The results of the comparison are shown in Tab. 5.6. Our design **(d)** presents the highest F_1 score with a significant improvement gap compared to the other designs. Notably, design **(a)** shows the lowest score, indicating the importance of entity-channel fusion. Although **(b)** and **(c)** contribute relatively high score, they still show 3.7% and 6% performance degradation in F_1@10, respectively. This demonstrates the efficiency of our holistic

entity-channel attention in selectively enhancing the most crucial visual or geometric features among all entities.

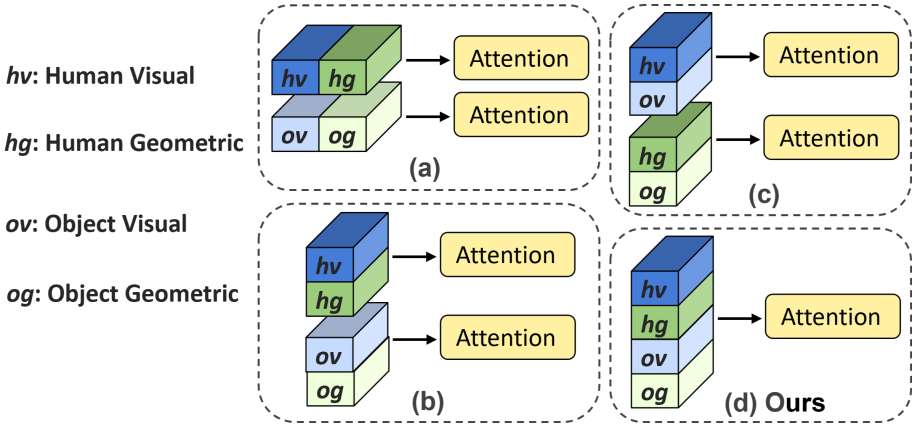

Figure 5.9: Different designs to combine geometric and visual features in channel attention-based feature fusion.

Table 5.6: Results of different strategies in channel attention-based feature fusion on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
(a) ho feature-channel attention	58.2 ± 4.0	50.7 ± 4.2	38.4 ± 3.6
(b) ho entity-channel attention	61.4 ± 5.7	56.5 ± 5.3	40.4 ± 4.7
(c) vg entity-channel attention	59.1 ± 5.3	50.4 ± 6.0	39.9 ± 4.8
(d) **GeoVis-GNN** (ours)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

To further validate the effectiveness of our proposed modules in GeoVis-GNN, we perform ablation studies on MPHOI-120, where CAF and IEG denote the attention-based feature fusion and the interdependent entity graph, respectively. Detailed results are presented in Tab. 5.7, in which: variant (1) denotes the IEG module is removed in our method; variant (2) indicates that both CAF and IEG modules are removed; variant (3) represents not only the proposed CAF and IEG modules being removed, but also the GAT for geometric feature embedding being replaced by GCN; variant (4) denotes the alternative architecture with top-down design. From

Table 5.7: Architecture alternative and ablation study on MPHOI-120. CAF and IEG denote the channel attention-based feature fusion and the interdependent entity graph, respectively.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
(1) GAT, w CAF, w/o IEG	61.2 ± 6.0	55.7 ± 5.2	45.4 ± 4.6
(2) GAT, w/o CAF&IEG	59.3 ± 6.1	52.5 ± 5.7	39.4 ± 4.3
(3) GCN, w/o CAF&IEG	58.6 ± 6.4	51.5 ± 5.3	38.3 ± 5.7
(4) Top-down architecture	62.8 ± 5.7	56.7 ± 5.2	42.8 ± 4.9
(5) **GeoVis-GNN** (ours)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

the results, removing any component from our model will result in a significant performance reduction. In particular, (1) shows 3.9% performance degradation in the F_1@10 metric, clearly validating the contribution of IEG to the final performance. Besides, (2) and (3) show 5.8% and 6.5% performance degradation in F_1@10, which indicates that CAF for feature fusion and GAT embedding are important for segmentation and recognition tasks. The comparative results between (4) and (5) highlight the strengths of our bottom-up architecture, exhibiting its enhanced ability for detailed feature integration and a more refined understanding of interactions.

5.4.5 HOI Attention Analysis

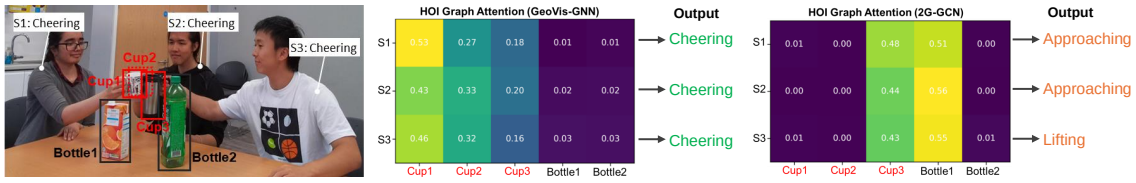

Figure 5.10: Visualization of HOI attention maps for GeoVis-GNN and 2G-GCN during a *Cheering* activity. Correct and incorrect recognition results are highlighted in green and orange, respectively.

To enhance the interpretability of our model, we deep into the attention analysis in the HOI graph. We compare GeoVis-GNN with the recent advanced method that constructs entity-level HOI graphs. Fig. 5.10 presents a comparative analysis of HOI attention maps in entity-level graphs generated by GeoVis-GNN and 2G-GCN for a *Cheering* activity involving three subjects, each holding a cup, with two bottles placed on the table. In the left attention map, our GeoVis-GNN model demonstrates

its superior interpretability by accurately focusing on all three cups, even effectively handling occlusions, such as Cup2 being partially hidden behind Cup1. This targeted attention enables the model to correctly recognize the *Cheering* sub-activity for all three subjects (highlighted in green).

In contrast, the 2G-GCN model exhibits less precise attention, incorrectly focusing on Cup3 and Bottle1, leading to erroneous sub-activity predictions such as *Approaching* and *Lifting* (highlighted in orange). This comparison highlights GeoVis-GNN's ability to maintain robust attention across relevant entities, even in occluded or cluttered environments, thereby ensuring more accurate HOI recognition. The clear distinction in attention focus between the two models underscores the effectiveness of our bottom-up approach in capturing the essential elements of complex interactions, which is critical for accurate activity recognition in multi-person scenarios.

5.4.6 Analysis of Varying Number of Objects

Table 5.8: Results of different number of object usage on MPHOI-120.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
2 objects only	61.4 ± 3.4	55.4 ± 2.0	40.1 ± 3.2
3 objects only	62.6 ± 6.9	56.2 ± 8.2	41.8 ± 9.1
4 objects only	63.1 ± 6.4	56.7 ± 7.5	43.2 ± 8.7
GeoVis-GNN (5 objects)	65.1 ± 5.2	59.8 ± 4.7	46.6 ± 5.1

Tab. 5.8 presents a comprehensive analysis of our model's performance when varying the number of objects considered on the MPHOI-120 dataset. Notably, MPHOI-120 contains 2-5 objects in total, even when using only 2 objects, our model outperforms the 2G-GCN baseline, demonstrating its robustness and highlighting its capability to extract meaningful interactions even from a limited set of objects.

Increasing the number of objects from 2 to 5 improves performance across all F_1 metrics, but also increases memory cost. This trade-off suggests that while more

objects provide richer interaction contexts, leading to better recognition accuracy, the memory requirements scale with the number of objects included. However, in highly cluttered environments with potentially hundreds of objects, our design offers an advantage by enabling the selection of a fixed number of objects to avoid a linear increase in memory consumption.

5.4.7 Cross-Dataset Zero-Shot Study

In real-world applications, models usually perform reliably on unseen data distributions without the luxury of extensive retraining or domain-specific adaptations. To demonstrate the robustness and generalization capabilities of our proposed GeoVis-GNN, we conduct a cross-dataset zero-shot evaluation, as detailed in Table 5.9. This study involves training GeoVis-GNN exclusively on the three-person MPHOI-120 dataset and subsequently testing it on the two-person MPHOI-72 dataset.

Table 5.9: Zero-shot results of training on three-person HOI dataset (MPHOI-120) and testing on two-person HOI dataset (MPHOI-72).

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
ASSIGN [5]	33.7	31.5	28.2
2G-GCN [146]	36.2	33.3	30.4
CATS [163]	38.5	35.6	33.2
GeoVis-GNN	42.1	40.3	34.5

Our results show that GeoVis-GNN significantly outperforms the existing baselines, ASSIGN, 2G-GCN and CATS, achieving an improvement of over 3.6% in the F_1@10 score. This substantial performance gain underscores the stronger generalization ability of GeoVis-GNN compared to state-of-the-art methods. The ability to effectively transfer learned features from a more complex three-person HOI scenario to a simpler two-person setting highlights the model's adaptability and transferability across diverse multi-person HOI datasets.

Additionally, in many real-world scenarios, target domain fine-tuning or transfer learning is often employed to adapt models to specific environments. However, our zero-shot results, while not reaching the performance levels achievable when training and testing on the same two-person dataset, are achieved without any such fine-tuning, relying solely on training with one dataset and testing on another that do not necessarily share a direct relationship. This suggests that GeoVis-GNN has the potential to generalize across different datasets with varying characteristics, even without extensive retraining. Although there is room for improvement, the results are promising and indicate that our approach can still be valuable in scenarios where labeled data for every possible situation may not be readily available.

5.5 Summary

Our bottom-up GeoVis-GNN framework for video-based multi-person HOI recognition innovates by a novel dual-attention fusion, optimizing feature integration by embedding and fusing visual and geometric features through a graph attention mechanism and a channel attention module. These enhanced entity-specific representations are then fed into an interdependent entity graph, enabling the modeling of both explicit interactions and implicit interdependencies for a more comprehensive understanding of multi-person HOI. Additionally, we propose a challenging three-person HOI dataset (MPHOI-120), and GeoVis-GNN sets new benchmarks across three-person, two-person, single-person, and bimanual HOI scenarios.

Our fusion-based bottom-up approach is not limited to the HOI recognition domain, but also shows promise in other similar fields. For trajectory prediction tasks [167, 195], unlike traditional approaches that simply merge contextual and geometric features [196, 197], our methodology can enhance their integration. In action quality assessment, our approach uniquely fuses visual and geometric features, providing a richer contextual feature representation than previous studies which focused on only one of these aspects [198–200].

CHAPTER 6

Conclusion

In the domain of multi-person HOI recognition, the integration of geometric and visual features, along with innovative graph neural networks, is crucial for improving accuracy in complex interactions. The contributions made during this doctoral research include novel graph neural network architectures (Chapters 3 to 5) and high-quality MPHOI datasets (Chapter 3 and Chapter 5). These advancements collectively enhance the understanding and modeling of multi-person interactions, addressing challenges like occlusions and dynamic complexities in real-world scenarios. The research contributions are reviewed in Section 6.1, and potential directions for future research are discussed in Section 6.2.

6.1 Review of Contributions

This thesis specifically focuses on the segmentation and recognition of distinct human sub-activities along the video timeline. Our research makes multiple key contributions to multi-person HOI recognition:

First, in Chapter 3, we introduce a Two-level Graph Convolutional Network (2G-GCN) that leverages a geometric-level graph to incorporate human skeletons

and object bounding boxes, and a fusion-level graph that integrates these geometric features with traditional visual features. To support this framework, we also developed the MPHOI-72 dataset, designed to facilitate research in multi-person HOI recognition.

Building upon these insights, in Chapter 4, we present CATS, an end-to-end framework that advances HOI recognition by integrating category-level and scenery-level analyses. CATS fuses multi-modal features across different categories and constructs a scenery interactive graph to learn relationships among these categories, enhancing the model's ability to recognize complex interactions.

Finally, in Chapter 5, we propose the Geometric Visual Fusion Graph Neural Network (GeoVis-GNN), a bottom-up framework that uses a dual-attention mechanism to fuse multimodal features of each entity before feeding them into an entity-level graph for comprehensive interaction analysis. This enables a more precise and effective integration of diverse feature types, ensuring that each level is optimally utilized to maximize representational power. To further validate this approach, we develop a challenging three-person HOI dataset, MPHOI-120, demonstrating that GeoVis-GNN sets new benchmarks across various HOI scenarios. Collectively, these contributions advance the field by providing robust and scalable solutions for complex, real-world HOI recognition tasks.

6.2 Future Research Directions

The methodologies presented in Chapters 3 to 5, while demonstrating promising performance, also have certain limitations. In this section, we will critically discuss these limitations and suggest potential directions for future research to address these challenges.

6.2.1 Object Geometric Representation

All methods in Chapters 3 to 5 utilize human skeletons and object bounding boxes as geometric features for human bodies and objects, respectively. While human skeletons are inherently more informative, the coarse nature of object bounding

boxes limits the ability to fully capture the complexity of HOIs. Unlike human body features, which offer rich spatial and temporal insights, bounding boxes only indicate an object's approximate location and size, failing to capture finer details like shape, orientation, or interaction points [201].

To address this limitation, it is necessary to adopt a more informative geometric representation of objects and allow for a better understanding of how objects are manipulated or interacted with by humans. The rotation-equivariant detector [202] enriches the object representation with rotated bounding boxes, resulting in improved object detection performance. The convex-hull features [203] allow representing objects of irregular shapes and layouts. The recently proposed self-supervised keypoints learning method [13] uses segmentation masks to detect adaptive keypoints for different object categories including humans in a unified manner.

Adopting more informative object geometric representations aligns better with the complexity of human geometric features, ensuring a more balanced representation in HOI recognition [13]. By providing both humans and objects with similarly rich geometric details, the model can capture interactions more accurately and fairly. This helps improve recognition accuracy and provides a clearer understanding of the spatial and structural relationships in complex HOI tasks.

6.2.2 Non-Contact HOIs

A characteristic of human activity in the real world is the presence of a large number of entities. For example, in a scenario where a person is surrounded by various objects but focuses on watching TV, this key non-contact interaction can be captured by the attention-based feature fusion module proposed in Chapter 5. This module is adept at handling scenes with multiple entities by discerning the dynamic relevance and underlying connections among individuals.

However, dealing with noisy environments, which may include hundreds of humans and objects, presents a significant challenge. Jiang *et al.* [204] propose a method that hallucinates human configurations in scenes where humans are not directly observed, providing a more meaningful context for understanding object arrangements and interactions. By leveraging infinite factored topic models, the system can infer hidden

human-object relationships based on object affordances and configurations, even in complex environments.

Nie *et al.* [205] handles noisy environments by focusing on human pose trajectories and using a probabilistic model to generate multiple hypotheses for object configurations, handling uncertainty and noise. A voting mechanism ensures that predictions focus on relevant HOIs, minimizing the impact of distant or irrelevant objects. Therefore, it is a research direction to efficiently extract both in-contact and non-contact interactions, identifying the most probable HOIs in such complex scenarios.

6.2.3 In-the-Wild HOI

In-the-wild HOI studies may face various challenges, such as data ambiguity and mutual occlusion, which are common in uncontrolled environments. Data ambiguity stems from the fact that a 2D image can represent multiple 3D configurations of humans and objects, making it difficult to determine the true spatial arrangement [206]. For example, a large surfboard far away and a small one closer to the camera can have the same 2D projection. This ambiguity in scale and depth poses a challenge for accurately reconstructing the scene. Mutual occlusion further complicates this issue, as the interaction regions of the human and object can be partially or fully obscured, making it challenging to fully perceive the interaction [207]. This is especially problematic in uncontrolled environments where objects and body parts can overlap in complex ways [206].

One key insight is that considering humans and objects jointly can provide valuable contextual cues and constraints to resolve ambiguity. Zhang *et al.* [206] leverage "3D common sense" constraints to improve the reconstruction of human-object arrangements in the wild. They tackle ambiguity by using a scale loss, incorporating object size priors, and employing an occlusion-aware silhouette reprojection loss to optimize object poses using 2D segmentation masks. Cao *et al.* [207] address ambiguity stemming from 2D images by employing 3D contact priors learned from motion capture data and utilizing differentiable rendering with depth information to refine hand and object poses in the presence of occlusions. Huang *et al.* [208] employ

multi-view data to resolve occlusions by reconstructing human bodies from multiple viewpoints to accurately capture human-scene contact. DECO [45] reasons about contact in occluded regions using cross-attention between scene context and body-part features and employs a 2D pixel-anchoring loss to ground 3D contact estimations to 2D image evidence. LEMON [46] infers contact regions by integrating semantic information with geometric correlations between human and object shapes and disambiguates interactions by leveraging multi-branch attention to derive interaction intention features from geometric and image features.

Although these methods showcase the significant progress being made in addressing the challenges in in-the-wild HOI studies, these challenges are primarily addressed in image-based HOI studies and remain largely unexplored in video-based studies. One future direction aims to extend our work into in-the-wild HOI recognition in videos, tackling the unique obstacles that arise in these less structured, real-world scenarios.

6.2.4 Weakly-Supervised Learning in HOI

In the current landscape of HOI recognition, precise action label annotations are a common requirement for training deep learning models [209]. For instance, labeling specific sub-activities, such as "lifting" or "drinking", requires meticulous frame-by-frame annotation, which is both time-consuming and expensive. Additionally, there is often variability in how different annotators interpret and label sub-activities, leading to inconsistent annotations. This issue is especially pronounced in complex multi-person HOI scenarios, where interactions are more dynamic and less clearly defined. As a result, the reliance on fully-supervised approaches that depend on detailed, accurate labels limits the scalability of HOI recognition systems and their ability to generalize to real-world scenarios.

To overcome the limitations of expensive and inconsistent label annotations, weakly-supervised learning offers a promising solution. Ren *et al.* [210] train a temporal action localization model using only video-level category labels, generating pseudo-labels by thresholding the attention sequence of an S-MIL model trained in the first stage of a two-stage pipeline. Rizve *et al.* [211] similarly exploit only video-

level labels for training, but it generates confidence-aware pseudo-action snippets, incorporating priors into the training process. Li *et al.* [209] eliminate the need for precise annotations of interaction types or spatiotemporal locations. Instead, a contrastive weakly supervised training loss is used to associate spatiotemporal regions with actions and objects while ensuring temporal continuity through self-supervision. By utilizing weakly-labeled data, models can generalize better to varied environments and complex HOIs, reducing the dependency on costly annotation efforts. These approaches open new avenues for improving the robustness and efficiency of HOI recognition systems in real-world applications.

Bibliography

[1] M. Dogariu, L.-D. Stefan, M. G. Constantin, and B. Ionescu, "Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios," in *2020 13th International Conference on Communications (COMM)*, pp. 157–160, IEEE, 2020. 1, 45

[2] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, 2024. 1, 45

[3] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 271–280, 2013. 1, 45

[4] D. Mukherjee, K. Gupta, L. H. Chang, and H. Najjaran, "A survey of robot learning strategies for human-robot collaboration in industrial settings," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102231, 2022. 1, 45

[5] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16041–16050, 2021. 2, 20, 24, 29, 35, 36, 37, 38, 39, 40, 41, 45, 47, 49, 50, 52, 53, 54, 58, 66, 68, 69, 70, 72, 73, 74, 75, 77, 81

[6] N. Wang, G. Zhu, L. Zhang, P. Shen, H. Li, and C. Hua, "Spatio-temporal interaction graph parsing networks for human-object interaction recognition," pp. 4985–4993, 2021. 2, 3, 20, 69

[7] S. P. R. Sunkesula, R. Dabral, and G. Ramakrishnan, "Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos," pp. 691–699, 2020. 2, 3, 20

[8] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5195–5201, 2022. 2, 3, 20

[9] N. Wang, G. Zhu, H. Li, M. Feng, X. Zhao, L. Ni, P. Shen, L. Mei, and L. Zhang, "Exploring spatio–temporal graph convolution for video-based human–object interaction recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5814–5827, 2023. 2, 3, 20, 25, 69

[10] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018. 2, 3, 29

[11] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *European Conference on Computer Vision*, pp. 488–504, 2020. 2, 3, 29

[12] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *European Conference on Computer Vision*, pp. 72–90, 2020. 2, 29

[13] M. Zhu, E. S. Ho, S. Chen, L. Yang, and H. P. Shum, "Geometric features enhanced human-object interaction detection," *IEEE Transactions on Instrumentation and Measurement*, 2024. 2, 3, 18, 22, 85

[14] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317, 2016. 2, 20, 29, 45

[15] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *European Conference on Computer Vision*, pp. 401–417, 2018. 2, 20, 24, 29, 45

[16] R. Dabral, S. Sarkar, S. P. Reddy, and G. Ramakrishnan, "Exploration of spatial and temporal modeling alternatives for hoi," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2281–2290, 2021. 2, 24

[17] X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao, "Distillation using oracle queries for transformer-based human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19558–19567, 2022. 2, 15

[18] L. Dong, Z. Li, K. Xu, Z. Zhang, L. Yan, S. Zhong, and X. Zou, "Category-aware transformer network for better human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19538–19547, 2022. 2, 18

[19] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013. 2, 12, 13, 14, 19, 30, 31, 36, 45, 46, 52, 61, 62, 65, 71

[20] Y. Jiang and A. Saxena, "Modeling high-dimensional humans for activity anticipation using gaussian process latent crfs.," in *Robotics: Science and systems*, pp. 1–8, Berkeley, CA, 2014. 2, 19

[21] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14–29, 2015. 2, 19

[22] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv e-prints*, pp. arXiv–1812, 2018. 3, 29, 39, 76

[23] H. Duan, M. Xu, B. Shuai, D. Modolo, Z. Tu, J. Tighe, and A. Bergamo, "Skeletr: Towards skeleton-based action recognition in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13634–13644, 2023. 3, 23

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 3, 16, 23

[25] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *European Conference on Computer Vision*, pp. 234–251, 2018. 3

[26] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3, 16

[27] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *European Conference on Computer Vision*, pp. 248–264, 2020. 3, 16

[28] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "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*, pp. 13617–13626, 2020. 3

[29] H. Wang, B. Yu, J. Li, L. Zhang, and D. Chen, "Multi-stream interaction networks for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3050–3060, 2021. 9

[30] H. Luo, G. Lin, Y. Yao, Z. Tang, Q. Wu, and X. Hua, "Dense semantics-assisted networks for video action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3073–3084, 2021. 9

[31] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13778–13790, 2023. 9

[32] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI Conference on Artificial Intelligence*, 2018. 9, 23, 34

[33] S. Li, X. He, W. Song, A. Hao, and H. Qin, "Graph diffusion convolutional network for skeleton based semantic recognition of two-person actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8477–8493, 2023. 9, 23

[34] Q. Yu, M. Tanaka, and K. Fujiwara, "Exploring vision transformers for 3d human motion-language models with motion patches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 937–946, 2024. 9

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009. 10

[36] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010. 10, 11

[37] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, IEEE, 2010. 10

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014. 10

[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017. 11, 37, 53, 68

[40] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *2011 International conference on computer vision*, pp. 1331–1338, IEEE, 2011. 11

[41] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1017–1025, 2015. 11

[42] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (WACV)*, pp. 381–389, 2018. 11, 15, 21

[43] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015. 11, 15, 21

[44] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel, "Care about you: towards large-scale human-centric visual relationship detection," *arXiv preprint arXiv:1705.09892*, 2017. 11

[45] S. Tripathi, A. Chatterjee, J.-C. Passy, H. Yi, D. Tzionas, and M. J. Black, "Deco: Dense estimation of 3d human-scene contact in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8001–8013, 2023. 11, 12, 87

[46] Y. Yang, W. Zhai, H. Luo, Y. Cao, and Z.-J. Zha, "Lemon: Learning 3d human-object interaction relation from 2d images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16284–16295, 2024. 11, 12, 87

[47] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas, "Detecting human-object contact in images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17100–17110, 2023. 12

[48] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866, 2023. 12

[49] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022. 12

[50] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193, 2021. 12

[51] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-c. Zhu, "Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities," in *European Conference on Computer Vision*, pp. 767–786, Springer, 2020. 12

[52] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3823–3833, 2016. 12

[53] T. Nagarajan and K. Grauman, "Learning affordance landscapes for interaction exploration in 3d environments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2005–2015, 2020. 12

[54] J. I. Lipton, A. J. Fay, and D. Rus, "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 179–186, 2017. 12

[55] S. Grover, K. Sidana, and V. Jain, "Pipeline for 3d reconstruction of the human body from ar/vr headset mounted egocentric cameras," *arXiv preprint arXiv:2111.05409*, 2021. 12

[56] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 32–36, IEEE, 2004. 12

[57] P. Mettes, J. C. Van Gemert, and C. G. Snoek, "Spot on: Action localization from pointly-supervised proposals," in *Computer Vision–European Conference on Computer Vision 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 437–453, Springer, 2016. 12

[58] S. Khurram, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv: 1212.0402*, 2012. 12

[59] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194–1201, IEEE, 2012. 12

[60] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014. 12

[61] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738, 2013. 12

[62] C. R. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2020. 12, 14, 30, 31, 36, 39, 45, 61, 65, 71, 77

[63] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pp. 0–0, 2021. 12, 45

[64] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018. 12, 13

[65] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020. 12, 13

[66] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5842–5850, 2017. 13

[67] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1049–1059, 2020. 13

[68] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas, "You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video.," in *BMVC*, vol. 2, p. 3, Citeseer, 2014. 13

[69] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Computer Vision–European Conference on Computer Vision 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 314–327, Springer, 2012. 13

[70] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2847–2854, IEEE, 2012. 13

[71] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European conference on computer vision (European Conference on Computer Vision)*, pp. 720–736, 2018. 13

[72] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018. 13

[73] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," 2021. 13, 62

[74] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 19383–19400, 2024. 13, 14

[75] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022. 14, 62

[76] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "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, 2022. 14, 62

[77] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "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, 2020. 15, 17

[78] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13234–13243, 2021. 15, 17

[79] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *European Conference on Computer Vision*, pp. 414–428, 2016. 15, 21, 29

[80] C. Gao, Y. Zou, and J.-B. Huang, "ican: Instance-centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018. 15, 16, 21, 29

[81] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3585–3594, 2019. 15

[82] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5694–5702, 2019. 15

[83] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016. 15, 21, 37, 53, 68

[84] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017. 15

[85] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *European Conference on Computer Vision*, pp. 248–265, Springer, 2020. 16, 21

[86] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10460–10469, 2020. 16

[87] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *European Conference on Computer Vision*, pp. 69–85, Springer, 2020. 16

[88] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008. 16, 20

[89] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017. 16

[90] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *European Conference on Computer Vision*, pp. 696–712, Springer, 2020. 16

[91] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 843–851, 2019. 16

[92] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu, "Pastanet: Toward human activity knowledge engine," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 382–391, 2020. 16

[93] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1981–1990, 2019. 16

[94] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *European Conference on Computer Vision*, pp. 584–600, Springer, 2020. 16

[95] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8359–8367, 2018. 17, 21

[96] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569–6578, 2019. 17

[97] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *European Conference on Computer Vision*, pp. 498–514, Springer, 2020. 17

[98] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4116–4125, 2020. 17

[99] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1291–1299, 2021. 17

[100] V. O. Maraghi and K. Faez, "Scaling human-object interaction recognition in the video through zero-shot learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 9922697, 2021. 17

[101] A. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, "What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5353–5363, 2022. 17, 18, 21

[102] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating hoi detection as adaptive set prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9004–9013, 2021. 17, 18

[103] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10410–10419, 2021. 17, 18

[104] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020. 17

[105] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020. 17

[106] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021. 17

[107] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 74–83, 2021. 18

[108] B. Kim, J. Mun, K.-W. On, M. Shin, J. Lee, and E.-S. Kim, "Mstr: Multi-scale transformer for end-to-end human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19578–19587, 2022. 18

[109] J. Zhou, Z. Wang, J. Meng, S. Liu, J. Zhang, and S. Chen, "Human interaction recognition with skeletal attention and shift graph convolution," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2022. 19, 25, 67, 68

[110] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden markov models," in *Computer Vision, IEEE International Conference on*, vol. 3, pp. 1455–1455, IEEE Computer Society, 2003. 19

[111] P. Natarajan and R. Nevatia, "Coupled hidden semi markov models for activity recognition," in *2007 IEEE Workshop on Motion and Video Computing (WMVC)*, pp. 10–10, IEEE, 2007. 19

[112] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Proceedings ninth IEEE International Conference on Computer Vision*, pp. 742–749, IEEE, 2003. 19

[113] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, and M. Csail, "Hidden-state conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007. 19

[114] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 210–220, 2006. 19

[115] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," *Advances in neural information processing systems*, vol. 17, 2004. 19

[116] O. Sener and A. Saxena, "rcrf: Recursive belief estimation over crfs in rgb-d activity videos.," in *Robotics: Science and systems*, Citeseer, 2015. 19, 38, 39, 54, 75

[117] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995. 19

[118] L. Medsker and L. C. Jain, *Recurrent neural networks: design and applications*. CRC press, 1999. 19

[119] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *European Conference on Computer Vision*, pp. 51–67, 2018. 21, 22

[120] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "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, 2019. 21, 22, 25, 68

[121] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu, "Mining cross-person cues for body-part interactiveness learning in hoi detection," in *European Conference on Computer Vision*, pp. 121–136, Springer, 2022. 22

[122] J. Park, J.-W. Park, and J.-S. Lee, "Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17152–17162, 2023. 22

[123] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 183–192, 2020. 22

[124] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1112–1121, 2020. 22, 23, 33, 34

[125] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997. 22

[126] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016. 22

[127] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017. 22

[128] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 148–157, IEEE, 2017. 22

[129] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*, pp. 816–833, Springer, 2016. 22

[130] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *arXiv preprint arXiv:1705.08106*, 2017. 22

[131] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28, 2017. 22

[132] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3288–3297, 2017. 22

[133] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017. 22

[134] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2019. 23, 33, 34

[135] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8561–8568, 2019. 23

[136] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2022. 23

[137] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, p. 121, 2021. 25

[138] L. Wang and P. Koniusz, "Self-supervising action recognition by statistical moment and subspace descriptors," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4324–4333, 2021. 25

[139] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A.-A. Khoo, "Tevad: Improved video anomaly detection with captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5559, 2023. 25

[140] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pp. 604–613, 2017. 25

[141] F. Baradel, C. Wolf, and J. Mille, "Human activity recognition with pose-driven attention to rgb," pp. 1–14, 2018. 25

[142] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 469–478, 2018. 25

[143] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," vol. 35, pp. 3199–3207, 2021. 25

[144] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19548–19557, 2022. 25, 69

[145] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 156–165, 2017. 27, 37, 52, 72

[146] T. Qiao, Q. Men, F. W. B. Li, Y. Kubotani, S. Morishima, and H. P. H. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *European Conference on Computer Vision*, 2022. 28, 45, 46, 47, 48, 49, 50, 52, 53, 54, 56, 57, 58, 60, 61, 63, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 77, 81

[147] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2470–2478, 2015. 29

[148] N. Bodla, G. Shrivastava, R. Chellappa, and A. Shrivastava, "Hierarchical video prediction using relational layouts for human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12146–12155, 2021. 29

[149] S. Zheng, S. Chen, and Q. Jin, "Skeleton-based interactive graph network for human object interaction detection," pp. 1–6, 2020. 30

[150] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Pose-based modular network for human-object interaction detection," *arXiv preprint arXiv:2008.02042*, 2020. 30

[151] "Quickstart: Set up azure kinect body tracking," 2022. 32, 64

[152] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graphics*, vol. 34, no. 6, pp. 1–16, 2015. 33

[153] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 33

[154] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018. 34

[155] V. O. Maraghi and K. Faez, "Zero-shot learning on human-object interaction recognition in video," in *2019 5th Iranian conference on signal processing and intelligent systems (ICSPIS)*, pp. 1–7, 2019. 35, 47, 66

[156] H. Le, D. Sahoo, N. F. Chen, and S. C. Hoi, "Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues," *arXiv preprint arXiv:2010.10095*, 2020. 35, 47, 66

[157] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017. 36

[158] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018. 37, 53

[159] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019. 37

[160] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14–29, 2016. 38, 39, 53, 54, 75

[161] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14424–14432, 2020. 43

[162] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "Sgcn: Sparse graph convolution network for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8994–9003, 2021. 43

[163] T. Qiao, R. Li, F. W. B. Li, and H. P. Shum, "From category to scenery: An end-to-end framework for multi-person human-object interaction recognition in videos," in *International Conference of Pattern Recognition*, 2024. 44, 60, 61, 66, 67, 68, 69, 72, 73, 74, 75, 81

[164] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Explainable video action reasoning via prior knowledge and state transitions," in *ACM MM*, pp. 521–529, 2019. 45

[165] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8688–8697, 2019. 45

[166] M. Liu, S. Tang, Y. Li, and J. M. Rehg, "Forecasting human-object interaction: joint prediction of motor attention and actions in first person video," in *European Conference on Computer Vision*, pp. 704–721, Springer, 2020. 45

[167] R. Li, S. Katsigiannis, and H. P. Shum, "Multiclass-sgcn: Sparse graph-based trajectory prediction with agent class embedding," pp. 2346–2350, IEEE, 2022. 45, 82

[168] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," vol. 33, pp. 5011–5022, 2020. 46, 47

[169] C. Baldassano, D. M. Beck, and L. Fei-Fei, "Human–object interactions are more than the sum of their parts," *Cerebral Cortex*, vol. 27, no. 3, pp. 2276–2288, 2017. 46, 47

[170] L. Li, H. P. H. Shum, and T. P. Breckon, "RAPiD-Seg: Range-Aware Pointwise Distance Distribution Networks for 3D LiDAR Segmentation," in *European Conference on Computer Vision*, 2024. 48

[171] Y. You, T. Chen, Z. Wang, and Y. Shen, "L2-gcn: Layer-wise and learned efficient training of graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2127–2135, 2020. 48

[172] S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu, "Graph neural tangent kernel: Fusing graph neural networks with graph kernels," vol. 32, 2019. 48

[173] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6272–6281, 2019. 50

[174] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. 51, 71

[175] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016. 51, 71

[176] T. Shu, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance for human-robot interaction," *arXiv preprint arXiv:1604.03692*, 2016. 62

[177] T. Shu, X. Gao, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1669–1676, 2017. 62

[178] Y. You, H. Liu, T. Wang, W. Li, R. Ding, and X. Li, "Co-evolution of pose and mesh for 3d human body estimation from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14963–14973, 2023. 64

[179] K. Zhai, Q. Nie, B. Ouyang, X. Li, and S. Yang, "Hopfir: Hop-wise graphformer with intragroup joint refinement for 3d human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14985–14995, 2023. 64

[180] M. Fan, M. Chen, C. Hu, and S. Zhou, "Occˆ 2net: Robust image matching based on 3d occupancy estimation for occluded regions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9652–9662, 2023. 64

[181] T. Heitzinger and M. Kampel, "A fast unified system for 3d object detection and tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17044–17054, 2023. 64

[182] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017. 65

[183] M. Kresović and T. D. Nguyen, "Bottom-up approaches for multi-person pose estimation and it's applications: A brief review," *arXiv preprint arXiv:2112.11834*, 2021. 65

[184] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14676–14686, 2021. 65

[185] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 850–859, 2019. 65

[186] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5977, 2019. 65

[187] N. Samet, S. Hicsonmez, and E. Akbas, "Houghnet: Integrating near and long-range evidence for bottom-up object detection," in *European Conference on Computer Vision*, pp. 406–423, Springer, 2020. 65

[188] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," *arXiv preprint arXiv:2105.14491*, 2021. 67

[189] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. 68

[190] M. Lin, Q. Chen, and S. Yan, "Network In Network," in *ICLR*, 2014. 69

[191] D. Tu, W. Sun, G. Zhai, and W. Shen, "Agglomerative transformer for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21614–21624, 2023. 69

[192] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *European Conference on Computer Vision*, pp. 541–556, Springer, 2020. 69

[193] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11525–11533, 2020. 69

[194] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. 70

[195] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang, "Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1410–1420, 2023. 82

[196] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1349–1358, 2019. 82

[197] K. Guo, W. Liu, and J. Pan, "End-to-end trajectory distribution prediction based on occupancy grid maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2242–2251, 2022. 82

[198] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 280–291, 2019. 82

[199] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9839–9848, 2020. 82

[200] H. Li, Q. Lei, H. Zhang, J. Du, and S. Gao, "Skeleton-based deep pose feature learning for action quality assessment on figure skating videos," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103625, 2022. 82

[201] M. Zhu, E. S. L. Ho, and H. P. H. Shum, "A skeleton-aware graph convolutional network for human-object interaction detection," in *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics*, SMC '22, 2022. 85

[202] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2786–2795, June 2021. 85

[203] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8792–8801, June 2021. 85

[204] Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3d scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2993–3000, 2013. 85

[205] Y. Nie, A. Dai, X. Han, and M. Nießner, "Pose2room: understanding 3d scenes from human activities," in *Proc. Eur. Conf. Comput. Vis.*, pp. 425–443, Springer, 2022. 86

[206] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3d human-object spatial arrangements from a single image in the wild," in *European Conference on Computer Vision*, pp. 34–51, Springer, 2020. 86

[207] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12417–12426, 2021. 86

[208] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black, "Capturing and inferring dense full-body human-scene contact," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13285, 2022. 86

[209] S. Li, Y. Du, A. Torralba, J. Sivic, and B. Russell, "Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1845–1855, 2021. 87, 88

[210] H. Ren, W. Yang, T. Zhang, and Y. Zhang, "Proposal-based multiple instance learning for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2394–2404, 2023. 87

[211] M. N. Rizve, G. Mittal, Y. Yu, M. Hall, S. Sajeev, M. Shah, and M. Chen, "Pivotal: Prior-driven supervision for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22992–23002, 2023. 87

APPENDIX A

Hardware Acknowledgements

In addition to the individuals acknowledged for their contributions to this thesis, we would like to recognize the essential hardware support that made this research possible.

We extend our sincere gratitude to Durham University's NVIDIA CUDA Center (NCC) GPU system [1], whose computational resources were instrumental in carrying out the research and experiments presented here. The NCC cluster, established through Durham University's strategic investment funds and managed by the Department of Computer Science, provided the high-performance computing environment necessary to process extensive datasets and run complex deep learning models efficiently. This infrastructure enabled the rigorous testing and refinement of the methodologies developed in this thesis, significantly contributing to the research outcomes. We are grateful for the access to this advanced resource, which has been critical to the success of this work.

[1] https://nccadmin.webspace.durham.ac.uk