Learning Transferable Representations for Hierarchical Relationship Exploration

ZHI HOU

Ph.D.

Supervisor: Dacheng Tao
Auxiliary Supervisor: Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
The University of Sydney
Australia

13 July 2023

Statement of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgment has been made in the text

Zhi Hou	Date

Authorship Attribution Statement

In addition to the statement above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Zhi Hou
School of Computer Science
Faculty of Engineering
The University of Sydney 11 April 2023

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Dacheng Tao
School of Computer Science
Faculty of Engineering
The University of Sydney 11 April 2023

Acknowledgements

During my Ph.D. study, I received enormous support from my supervisors, colleagues, friends, and family. Without their help, this thesis can never be completed.

First and foremost, I would like to show my deepest appreciation to my primary supervisor, Prof. Dacheng Tao, for his meticulous guidance, insightful advice, and kind support throughout my Ph.D. study. It is your detailed suggestions and funding support that give me the opportunity to pursue a research career and succeed my Ph.D. at the University of Sydney.

I am also particularly grateful to my auxiliary supervisor, Dr. Baosheng Yu (Postdoc), for his countless support and suggestions in both study and life throughout my Ph.D. candidature, including but not limited to paper revision and idea discussion.

Meanwhile, I am so appreciative of my previous supervisors, Prof. Yu Qiao and Prof. Xiaojiang Peng, who provided extensive support and guidance in academic research before my Ph.D. study. I am grateful to have had the opportunity to learn from them. I would like to thank my Master's supervisor, Prof. Guihua Wen, for his considerable advice and support.

I would like to thank my collaborators, friends, and colleagues: Mr. Junhao Zhang, Dr. Chaoyue Wang, Dr. Yibin Zhang, Dr. Xie Di, Mr. Haiming Sun, Mr. Chao Li, Mr. Jinlong Fan, Dr. Xinqi Zhu, Dr. Hao Guan, Mr. Weijie Chen, Mr. Kai Wang, Dr. Jiaxian Guo, Dr. Zhen Wang, Dr. Shen Zhang, Mr. Lewei, Lu, Mr. Haibo Qiu and so on. Besides, I want to thank Dr. Yandong Wen, Dr. Zhiding Yu and Dr. Weiyang Liu, whose academic experiences significantly encourage me to follow my interest in pursuing a Ph.D. degree.

Besides, I would like to show my gratitude to my family, including my grandparents, my parents, my uncles, my aunts, and my cousins, for their continuous support throughout my whole life. Particularly thanks to my aunts, who provide a good environment for my study and growth in the winter and summer vacations in primary school.

Last but not least, my greatest appreciation is for my fiancée, Ms. Xiaoying Huang. Thanks for her support from all aspects.

Publication List

(1) Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. "Detecting human-object interaction via fabricated compositional learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14646-14655. 2021. (**Chapter 2**)

(2) Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. "Affordance transfer learning for human-object interaction detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 495-504. 2021. (**Chapter 2**)

(3) Zhi Hou, Baosheng Yu, and Dacheng Tao. "BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7256-7266. 2022. (**Chapter 3**)

(4) Zhi Hou, Baosheng Yu, and Dacheng Tao. "Discovering Human-Object Interaction Concepts via Self-Compositional Learning." In European Conference on Computer Vision, pp. 461-478. Springer, Cham, 2022. (**Chapter 2**)

(5) Zhi Hou, Baosheng Yu, and Dacheng Tao. "Compositional 3D Human-Object Neural Animation." Under review. (**Chapter 4**)

Abstract

The visual scenes are composed of basic elements, such as objects, parts, and other semantic regions. It is well-acknowledged that humans perceive the world in a compositional and hierarchical way in which visual scenes are treated as a layout of distinct semantic objects/attributes/parts. Those separated objects/attributes/parts are linked together via different relationships, including visual relationships and semantic relationships. Particularly, the shared parts/attributes/objects of the visual concepts (object, visual relationships), are shared and thus transferable among different visual concepts. Humans can easily imagine a new composite concept from the shared parts of different concepts, while one of the important shortcomings of current deep neural networks is the compositional perception ability and thus it requires a large scale of data to optimize the deep neural networks. From the perspective of compositional perception, this thesis thinks one of the limitations of typical neural networks is that the factor representations of deep neural networks are not sharable and transferable among different concepts. Therefore, the thesis introduces various techniques, including compositional learning framework, compositional invariant learning, and BatchFormer module, to enable the factor representations of deep neural networks sharable and transferable among different concepts for hierarchical relationship exploration, involving human-object interaction, 3D human-object interaction and sample relationships.

Contents

List of Figures

CHAPTER 1

Introduction

The visual scenes are hierarchically layout in a compositional way, composed of parts, objects, and other semantic regions, while those separated objects/parts are linked together via different relationships, and transferable among different scenes. Visual relationship is crucial for visual scene understanding, and understanding interactions between human and object is crucial for Intelligent driving, Human-centric Content Generation, Argument Reality, Robotics and Embody AI. Meanwhile, semantic relationships among different samples is beneficial for achieving better representations. Humans perceive the visual world in a compositional and hierarchical way, and can usually achieve excellent generalization according to the visual compositionality. However, it is challenging for deep neural networks to achieve the compositional perception ability. Humans usually treat the shared element/factor of different concepts as similar, however the element or factor representations of typical deep neural networks are not always shared among different concepts and are representative for the concept, especially for those visual relationships and visual categories. Meanwhile, the visual relationships and semantic relationships in the visual world usually demonstrate hierarchical structure, and also crucial to understand the visual world. Besides, the visual relationships have broad application potential in real world.

Current approaches [175, 224] usually decouple the representations of the visual relationships or visual concepts into several factor representations or several factor predictions according to explicit supervisions. The decoupled representations or predictions significantly benefit the compositional generation of the visual relationships detection. Previous approaches also introduce to utilize the language embedding to enable the transferability of decoupled representations among different categories by aligning the representations to corresponding

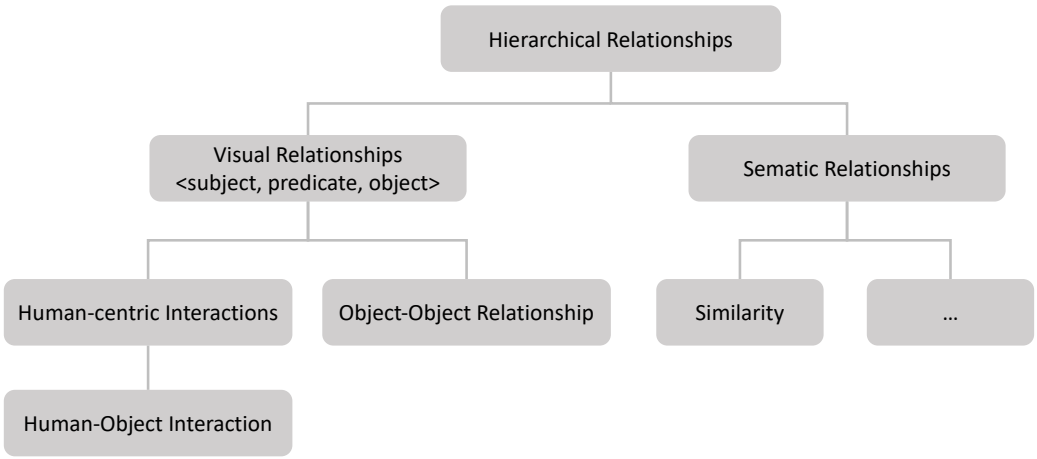

FIGURE 1.1. Hierarchical relationship terminologies.

language embeddings. However, the language embedding limits the representation of the visual features and visual features are not fully transferable among different concepts. Therefore, this thesis introduces to learn transferable factor/element representations for hierarchical visual relationships from pure visual perspective without external knowledge. Besides, typical decomposition and composition approaches require the explicit supervisions, which limits the application of those approaches for typical methods. The thesis further presents a BatchFormer module to implicitly enable the feature transferability among different samples.

There are two important relationship expressions: one is visual relationships composed of a triple $\langle subject, predicate, object \rangle$ (e.g., a person plays basketball), which explicitly and spatially exist in the visual scenes (e.g., Human-centric interactions and Object-Object Relationships); another is semantic relationships that are not spatially layout in the visual scenes, such as similarity (e.g. porpoises and dolphins). Particularly, the two kinds of relationships always demonstrate the hierarchy. The pair-wise visual relationships in the scenes hierarchically build the scene structure, while the semantic relationships are usually built by the class hierarchies (e.g., similar classes usually belong to the same super-category, porpoises and dolphins are fish and share similar body shapes.). Meanwhile, semantic relationships are also existing in visual relationships in which the element of different triples might be the same, i.e., different concepts of visual relationship might also be related. Those shared elements or semantics of different concepts are usually transferable, i.e., we can transfer

the shared elements from one concept to another concept. For deep neural networks, we can achieve better generalization by enabling element/attribute feature transferability among different concepts or visual relationships. In detail, if we can transfer the elements/attributes from the concepts with extensive training samples to the concepts with a few training samples, we can significantly facilitate the learning for few-shot concepts.

There are various kinds of challenges in visual relationship understanding. On the one hand, the distribution of HOI dataset is usually long-tailed, which significantly hampers the optimization of HOI models, and limits the performance of few-shot HOI categories. Meanwhile, HOI requires fine-grained action recognition and 3D understanding. Particularly, the self-occlusion is serious in 3D Human-Object Interaction. In this thesis, we mainly focus on 2D human-object interaction in Chapter 2 and present a neural human-object deformation approach to address the self-occlusion for 3D human-object interaction reconstruction and rendering in Chapter 4.

Human-Object Interaction (HOI) is widely popular among visual relationships, and demonstrates great potential applicability for robotics and content generation. Compositionality is one of the significant characteristics of Human-Object Interaction (HOI), i.e., a Human-Object Interaction can be decomposed into a person with a specific action and a corresponding object. Whist massive deep neural approaches are introduced to improve the generalization, the compositional generalization has been poorly investigated, especially for Human-Object Interaction. Meanwhile, massive verb samples and object samples form two long-tailed distributions for verb and object, as a result of which the long-tailed HOI distribution is more serious according to the compositional long-tailed distribution in HOI. Furthermore, the compositionality of HOI also provides an effective way to reason the object attribute/affordances, i.e., compositional object attribute/affordance reasoning. Specifically, each factor in the pair, e.g., verb-object in HOI, has individual semantics. Human-Object Interaction does not only demonstrate the interaction between human and object, but also illustrates the affordances of the object, further implying the clues for discovering novel-object pairs. Based on this observation, the thesis presents two additional relevant tasks, i.e., Object affordance recognition and HOI concept discovery, and introduces a novel framework, named as visual

compositional learning or VCL, to mimic the ability of human-level compositional perception for Human-Object Interaction. The thesis further introduces building pseudo labels from the verb-pair confidence of the composite HOIs, and devises a self-training strategy with pseudo labels to optimize all the composite HOIs in an end-to-end way, and finally, significantly improve few- and zero-shot HOI detection, object affordance recognition, and HOI concept discovery.

Except for image-level interaction understanding, 3D vision understanding is highly required for real-world applications. Neural rendering of animatable 3D human avatars has been intensively explored by implicit neural representations, while the rich human-object interactions (HOIs) are crucial for numerous human-centric scene capturing/understanding applications such as AR/VR and robotics. Therefore, the thesis also addresses the challenge of HOI animation in a compositional manner, i.e., animating novel HOIs including the novel interaction, human and/or object via a sequence of novel driving poses. Specifically, the thesis first adopts the neural human-object deformation to model and render HOI dynamics based on the neural representations. Next, the thesis devises new compositional conditional neural radiance fields (or CC-NeRF), which decomposes the interdependence between the human and object latent codes to enable compositionally controlling the animation of novel HOIs.

Not only perceiving existing visual relationships in the visual scene but also exploring the implicit semantic relationships among different samples are important for hierarchical visual understanding. However, the semantic relationships are implicit, and we usually do not have any annotations for the categories of semantic relationships, which is challenging for neural networks to mine the relationships among different samples. Meanwhile, the relationship graph among different samples is large and it is difficult to mine the relationships from the whole of the dataset directly. The thesis carefully analyzes previous approaches that mine the relationships, and introduces a unified way to implicitly transfer the representations among different samples for exploring the sample relationships.

Existing methods mainly explore sample relationships in a vanilla way from the perspectives of either the input or the loss function for data-scarcity tasks. Differently, the thesis

proposes a batch transformer module, BatchFormerV1, to equip deep neural networks themselves with the ability to explore sample relationships in a learnable way. Basically, the proposed method enables data collaboration, e.g., head-class samples will also contribute to the learning of tail classes. Considering that exploring instance-level relationships has very limited impacts on dense prediction and is impractical to dense prediction tasks, the thesis presents BatchFormerV2 to enable exploring sample relationships for pixel-/patch-level dense representations. In addition, to address the train-test inconsistency where a mini-batch of data samples are *neither necessary nor desirable* during inference, the thesis also devises a two-stream training pipeline, i.e., a shared model is first jointly optimized with and without BatchFormerV2 which is then removed during testing. The proposed module is plug-and-play without requiring any extra inference cost.

1.1 Background

1.1.1 Knowledge Transfer

Knowledge Transfer is very popular among deep neural networks, especially for data scarcity tasks. It is well-acknowledged that different samples share similar semantics, and we can transfer the shared semantics for novel class learning. Therefore, extensive approaches investigate the transfer learning for few-shot learning [272, 185], in which common paradigms usually learn to transfer the knowledge from the base classes or pre-trained models to novel classes. Meanwhile, knowledge transfer is also a significant technical route in the long-tailed recognition. The long-tailed learning approaches implicitly or explicitly transfer the meta knowledge between head and tail classes for long-tailed recognition [263, 171]. Out-of-distribution generalization is a significant challenge, and recently attracts extensive interest from the community. Those methods aim to transfer the invariant knowledge for domain generalization [201, 4, 116]. This thesis thinks learning transferable representations are significant to achieve robust and reliable deep neural networks.

1.1.2 Hierarchical visual relationship

The visual scenes are composed of basic elements [103], such as objects, parts, and other semantic regions. Different visual elements in the scene are combined together via different relationships, including explicit relations and implicit relations. Visual relationships, e.g., spatial relations [64, 179] and semantic relations [104], are widely ubiquitous in the visual world. Currently, visual relationship understanding [175, 78] receives increasing interest from the community. Those visual relationship approaches [175, 78] usually treat the visual scene as a hierarchical structure, but rarely aim to improve the compositional generalization in a learning way. Objects can be treated as a set of attributes [156], and extensive approaches [185, 178] investigate the hierarchical structure among object attributes to improve the compositional zero-shot learning. The thesis devises to leverage the relationship (e.g., sharing attributes among different objects) between samples for data-scarcity tasks. The hierarchical visual relationship is a way to achieve human-level intelligence since human beings understand the visual scene in a hierarchical and compositional way.

1.1.3 2D Human-Object Interaction

HOI understanding [78] is of great importance for visual relationship reasoning [175, 278] and action understanding [24, 301]. Different approaches have been investigated for HOI understanding from various aspects, including HOI detection [28, 157, 161, 305, 131, 33, 318, 234, 290], HOI recognition [26, 128, 117], video HOI [48, 123], compositional action recognition [180], video generation [187], and object affordance reasoning [60]. Compositionality is a significant characteristic of HOI, which can be decomposed into a verb and an object. Meanwhile, there is also a popular challenge based on compositionality, i.e., compositional generalization. For Human-object interaction, compositional zero-shot detection/recognition and compositional long-tailed detection/recognition have attracted massive interests from the community. To address the compositional generalization, The thesis presents a novel framework, visual compositional learning, to compose HOI features from pair-wise images. Besides, Human-Object Interaction is firmly related to object affordance understanding since the human-object interaction actually presents what action can be applied for the object. The

thesis leverages the compositional object affordance reasoning approach to recognize the object affordance for the novel objects. Meanwhile, DETR-based methods (e.g., Qpic [234]) achieve superior performance on HOI detection. However, these approaches mainly consider the perception of known HOI concepts and pay no attention to HOI concept discovery. To fulfill the gap between learning on known and unknown concepts, a novel task, i.e., HOI concept discovery, is explored in this thesis. Currently, zero-shot HOI detection also attracts massive interest from the community [224, 9, 204, 111, 110]. However, those approaches merely consider known concepts and are unable to discover HOI concepts. Some HOI approaches [204, 9, 258, 257] expand the known concepts via leveraging language priors. However, that is limited to existing knowledge and can not discover concepts that never appear in the language prior knowledge. HOI concept discovery is able to address the problem, and enable unknown HOI concept detection.

1.1.4 Sample Relationship Exploration

The relationships among different samples are rich and ubiquitous in the visual world, and the sample relationships have been implicitly explored for various vision tasks [293, 168, 105, 184] in the community. On the one hand, those popular data augmentation strategies, including mixup [292], copy-paste [71] and crass-grad [222]. For example, Zhang *et al*. [293] propose to regularize the model to favor simple linear behavior in-between training samples with mixup. However, mixup [293] merely considers a linear transformation between data samples, while the thesis aims to investigate the non-linear relationship among samples in a more powerful way. On the other hand, the sample relationship exploration is also active among those data scarcity tasks, e.g., few/zero-shot learning, long-tailed recognition, and domain generalization. The compositionality of samples has also inspired many approaches to improve few/zero-shot generalization [237, 94, 111, 185], where the parts/attributes shared among different samples have been explored via the prior knowledge on label relationships. Several approaches also use sample/class relationships to conduct transductive inference [168, 105, 171, 184], e.g., transductive few-shot classification [168], meta embedding [171, 313],

and non-parametric transformer [139]. However, those approaches usually require inference with multiple samples (e.g., query set, or bank features). Meanwhile, many recent domain generalization methods [201, 4, 116] aim to find casual/invariant representations across domains, which we think internally utilize the relationship among samples of the same class but different domains. However, those methods usually investigate the sample relationships separately. The thesis proposes to facilitate representation learning by exploring the relationships among different samples.

1.1.5 3D Human-Object Interaction

The visual scene understanding not only requires to perceive the world in the images, but also the geometry of the objects and scenes. Recently, implicit neural representations [195, 182, 39] dominate the 3D visual scene representation. Meanwhile, NeRF [183] represents 3D points in the scene with density and color, and renders the scene with volumetric rendering techniques, achieving photorealistic novel rendering. Human avatar generation [199, 165, 143, 190, 230, 158, 303, 229] has achieved significant progress, especially in novel view rendering for novel poses. More recently, [151, 255] demonstrate appealing avatar generation under out-of-distribution poses. For 3D Human-object interaction, early work mainly investigates synthesizing human pose and object [120], human body reconstruction [62], object recognition [266], or human 3D pose estimation [2, 133, 38, 160] under the interaction with objects or environments. Recently, increasing approaches [16, 233, 232, 84, 125, 46, 280, 294, 115, 262, 276] focus on 3D Interactions between Human and its surrounding objects. Zhang *et al*. [294] present to reconstruct the spatial arrangements of Human-Object Interaction. [280, 46] reconstruct the meshes of human-object interactions, while recent work [125] introduces the neural representations to human-object interaction and significantly advances the novel view synthesis performance. Particularly, a real HOI dataset, BEHAVE [16], consisting of 8 subjects and diverse objects, is introduced with spare views of HD videos and the poses of humans and objects. The thesis mainly conducts experiments based on BEHAVE. Concurrent works [85, 262, 276, 115] focus on reconstruction or 3D tracking, significantly ignoring interaction animations. Besides, though current compositional approaches on human-centric

interactions have studied the recognition [128], detection [111], object affordance [108], 2D generation [187], and 3D human-scene synthesis [298], the compositional 3D animation remains unsolved. The thesis investigates 3D human-object interaction from the novel challenge, compositional human-object neural animation.

1.2 Contributions

The works during my Ph.D. study mainly focus on learning transferable representations for the hierarchical visual relationship understanding based on Human-Object Interaction, sample relationships, and 3D Human-Object Interaction. According to the hierarchical characteristics of the visual world, this thesis has deeply explored the compositional learning for both 2D Human-Object Interaction and 3D Human-Object Interaction. Meanwhile, sample relationships are also investigated according to the similarity and dissimilarity between different samples. In a nutshell, the contributions of this thesis can be summarized as follows,

- **Chapter 2** introduces a series of compositional approaches, including visual compositional learning, fabricated compositional learning, affordance transfer learning and self-compositional learning for exploring human-object interaction. A core idea behind those methods is transferring verb/object representations among different HOI samples. Visual compositional learning effectively disentangles the verb and object representations, and thus significantly improves the compositional generalization for HOI detection, including Long-tailed HOI detection and zero-shot HOI detection. Fabricated compositional learning re-balances the distribution for HOI dataset to facilitate HOI compositional generalization, and thus effectively improves previous approaches on compositional HOI detection benchmarks. Affordance transfer learning (ATL) enables object affordance recognition with HOI model by a novel compositional affordance reasoning strategy. Meanwhile, ATL transfers the verb representation to the novel object such that it enables the human-novel-object interaction detection. Lastly, this thesis presents a self-compositional learning strategy via leveraging a concept confidence matrix to build pseudo labels for all the

composite HOI features, and thus make use of all the composite features to facilitate the optimization on object affordance recognition and HOI concept discovery.

- **Chapter 3** extends the compositional approaches from pair-wise images to mini-batch images, and presents a simple yet effective module, named as Batch Transformer or BatchFormer, to implicitly explore the sample relationships for robust representation learning. BatchFormer implicitly enables representation transfer among different samples. Meanwhile, the chapter introduces a shared classifier strategy to maintain batch-invariant learning, that is invariant to batch size, as a result of which the Batchformer module can be removed during inference and thus increases no additional computation budget. Nevertheless, the proposed Batchformer module is limited to image classification tasks. This chapter further extends the BatchFormer module as BatchFormerV2 for dense prediction tasks, in which the thesis applies the BatchFormerV2 module in the vision Transformer networks in the batch dimension. Meanwhile, the thesis shares the proposed Batchformerv2 module among different spatial positions. To maintain batch-invariant learning, the thesis further devises a two-stream pipeline, in which one stream utilizes the Batchformerv2 modules to explore the samples relationships, and another stream keeps the typical networks. Among the two streams, the other modules except for Batchformerv2 module are shared. Extensive experiments demonstrate the proposed method on over ten popular datasets, including 1) different data scarcity settings such as long-tailed recognition, zero-shot learning, domain generalization, and contrastive learning; and 2) different visual recognition tasks ranging from image classification to object detection and panoptic segmentation.
- **Chapter 4** proposes a compositional 3D Human-Object neural animation approach, which introduces a neural human-object deformation method and compositional conditional NeRF to enable the compositionally pose-driven control of human-object interactions for novel humans and objects. Specifically, the proposed method treats the object as an additional pseudo bone compared to body bones, and leverages the typical deformation methods to animate the human-object interaction. Besides, a compositional conditional NeRF is introduced with two latent codes to control

the identity of human and object. This chapter further presents a compositional invariant learning to disentangle the interdependence between the two latent codes, such that we can animate the interaction for novel person and novel objects. The compositional invariant learning facilitates the animation for novel human-object pairs, including novel human body and static objects, and improves the pose transfer among different objects. Extensive experiments show the considerable generalization on human-object neural animation for novel poses, novel person and novel objects.

CHAPTER 2

Compositional Learning for Human-Object Interaction Exploration

---

The visual scenes are naturally hierarchical and compositional, which has attracted extensive interest in compositional generalization. Compositionality is one of the significant characteristics in Human-Object Interaction (HOI), i.e., a Human-Object Interaction can be decomposed into a person with a specific action and a corresponding object. Meanwhile, the decomposed elements (e.g. verb or object) are transferable among similar objects, and are able to compose novel HOI samples. Whilst current compositional approaches focus on the generalization of recognition/detection, they ignore the compositional reasoning for attribute/affordances recognition behind the compositional generalization. Specifically, each factor in the pair, e.g., verb-object in HOI, has individual semantics, and the compositionality does a significant favor in reasoning the attributes (e.g., affordances for the object) of the factors. In this chapter, we explore the compositional generalization and reasoning under Human-Object Interaction scenarios. Human-Object Interaction does not only demonstrate the interaction between human and object, but also illustrates the affordances of the object, further implying the clues for discovering novel-object pairs. Based on this observation, we present two additional relevant tasks, i.e., Object affordance recognition and HOI concept discovery. To ease those challenges simultaneously, we introduce a novel framework, named as visual compositional learning or VCL, to mimic the ability of human-beings in compositional perception for Human-Object Interaction. Specifically, VCL first decomposes an HOI representation into object and verb specific features, and then composes new interaction samples in the feature space via stitching the decomposed features. The integration of decomposition and composition enables VCL to share object and verb features among different HOI samples and images, and to generate new interaction samples and new types of HOI. Furthermore, we build pseudo labels from the verb-pair confidence of the composite HOIs, and devise a self-training strategy with pseudo

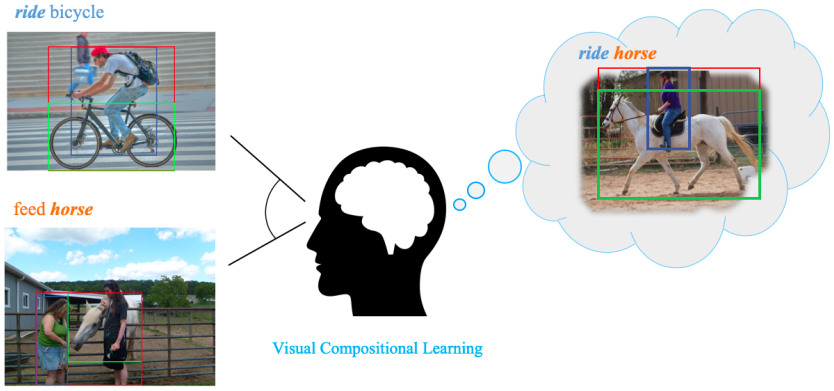

FIGURE 2.1. An illustration of Visual Compositional Learning (VCL). VCL achieves the new concept feature of $\langle ride, horse \rangle$ from $\langle feed, horse \rangle$ and $\langle ride, bicycle \rangle$ via visual compositional learning without external prior knowledge.

labels to optimize all the composite HOIs in an end-to-end way, and finally significantly improve few- and zero-shot HOI detection, object affordance recognition, and HOI concept discovery. With extensive experiments, we demonstrate the compositional approaches not only significantly facilitate the generalization of HOI detection, but also enable the model to recognize object affordance and discover HOI concepts.

2.1 Motivations and Contributions

Human-Object Interaction (HOI) [78, 283] is increasingly popular among the community due to its broad ranges of potential applications in visual scene understanding [175], action understanding [77], and meta universe [16]. Although current HOI approaches have achieved significant progress in common vision tasks, such as recognition [59, 155] and detection [28, 66, 235], the internal structure of HOI, i.e., compositionality, and the compositional reasoning have been poorly explored. A few approaches have investigated the compositional approaches [224, 128, 176] for HOI recognition/detection. However, the compositional reasoning for object affordance and HOI concept discovery is significantly ignored.

The visual scenes are composed of basic elements, such as objects, parts, and other semantic regions. It is well-acknowledged that humans perceive the world in a compositional way in

which visual scenes are treated as a layout of distinct semantic objects [103, 226]. People usually exhibit the capacity to understand and produce a potentially infinite number of novel combinations of known components [42], while current deep neural approaches fail to perceive the visual scenes compositionally. There are massive compositional generalization challenges [144, 132, 5] and methods [36, 166, 189, 130, 178, 5], ranging from natural language to computer vision. Following those compositional approaches, we can understand HOIs by decomposing them into objects and human interaction (verb) types. This decomposition helps to solve the rare Human-Object Interactions with compositional generalization, including compositional zero-shot learning and few-shot learning. For example, in HICO-DET dataset [28], $\langle hug, suitcase \rangle$ is a rare case with only one example, while we have more than 1000 HOI samples including object "suitcase", and 500 samples including the verb "hug". Obviously, object representations can be shared among different HOIs. And samples with the same verb usually exhibit similar human poses or action characteristics [91]. By combining the concepts of "suitcase" and "hug" learned from these large number samples, one can handle the rare case $\langle hug, suitcase \rangle$. This inspires to reduce the complexity of HOI detection and handle unseen/rare categories via learning compositional components, i.e., human verbs and objects from visual scenes, as illustrated in Figure 2.1.

Inspired by the above analysis, this thesis proposes a conceptually simple yet effective framework, Visual Compositional Learning (VCL), for Human-Object Interaction Detection, which performs compositional learning on the visual verb and object representations. However, it is non-trivial to effectively compose new valid HOI samples due to the limited HOI instances in each image. Meanwhile, combining verbs and objects from the HOI images might also hamper the annotated examples in the image. Therefore, we present a novel compositional learning approach by composing HOI samples in the feature space with verb and object features from different images and different HOI types. In this way, we can significantly augment the training samples, and thus relieve the few- and zero-shot challenges in HOI detection/recognition. Moreover, our VCL encourages the model to learn the shared and distinctive verb and object representations that are insensitive to variations (i.e., the specific images and interactions), and achieve better generalization. Compared to other compositional

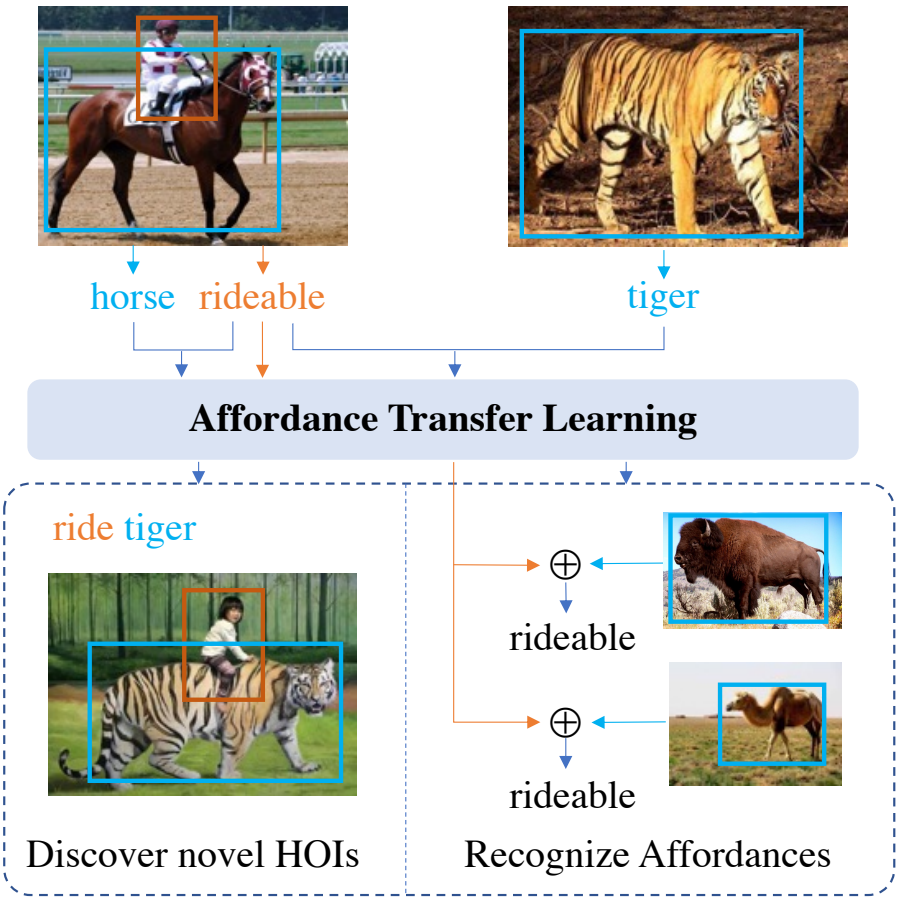

FIGURE 2.2. An intuitive example to demonstrate affordance transfer approach for jointly exploring human interactions with novel objects (e.g., "tiger"), and recognizing the affordance of novel objects. The proposed method is able to learn from the unseen interaction samples (e.g., "ride tiger") that are composed of affordance representations and novel object representations, which meanwhile transfers the affordance to novel objects and enables the object affordance recognition.

approaches [128, 208], VCL is an end-to-end framework without the requirements on language knowledge priors.

However, in real-world scenarios, long-tailed distributions are common for the data perceived by human vision system, e.g., actions/verbs and objects [171], which poses a significant challenge for HOI compositional learning. The combinatorial nature of HOI further highlights the issues of long-tailed distributions in HOI detection, while human can efficiently learn to recognize seen and even unseen HOIs from limited samples. An intuitive example of

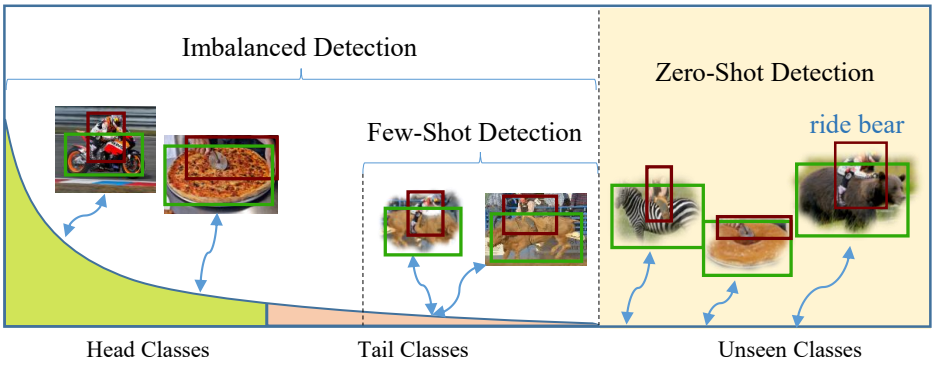

FIGURE 2.3. Open long-tailed HOI detection addresses the problem of imbalanced learning and zero-shot learning in a unified way. We propose to compose new HOIs for open long-tailed HOI detection. Specifically, the blurred HOIs, e.g., “ride bear", are composite.

open long-tailed HOI detection is shown in Figure A.2, in which one can easily recognize the unseen action “ride bear”, nevertheless it never even happened. However, existing HOI detection approaches usually focus on either the head [66, 161, 264], the tail [277] or unseen categories [224, 204], leaving the the problem of open long-tailed HOI detection poorly investigated.

Open long-tailed HOI detection falls into the category of the long-tailed zero-shot learning problem, which is usually referred into several isolated problems, including long-tailed learning [122, 93], few-shot learning [61, 246], zero-shot learning [146]. To address the problem of imbalanced training data, existing methods mainly focus on three strategies: 1) re-sampling [83, 126]; 2) re-weighted loss functions [45, 22, 92]; and 3) knowledge transfer [263, 171, 61, 146, 216, 65]. Specifically, re-sampling and re-weighted loss functions are usually designed for imbalance problems, while knowledge transfer is introduced to relieve all the long-tailed [263], few-shot [225], and zero-shot problem [271, 65]. Recently, two popular knowledge transfer methods have received increasing attention from the community, data generation [263, 264, 271, 171, 61, 146, 216, 129] (transferring head/base classes to tail/unseen classes) and visual-semantic embedding [65] (transferring from language knowledge). Along the first way, we address the problem of open long-tailed HOI detection from the perspective of HOI generation.

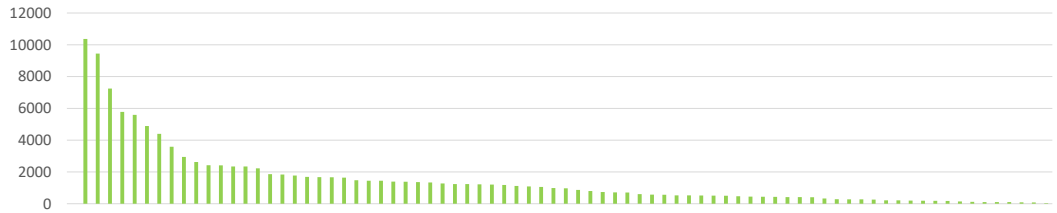

FIGURE 2.4. Illustration of the distribution of the number of object box in HICO-DET dataset. The categories are sorted by the number of instances.

Inspired by the compositionality of HOI, several zero-and few-shot HOI detection approaches have been proposed to enforce the factored primitive (verb and object) representation of the same primitive class to be similar among different HOIs, such as factorized model [224, 9] and factor visual-language model [277, 204, 9]. However, regularizing factor representation, i.e., enforcing the same verb/object representation to be similar among different HOIs, is only sub-optimal for HOI detection. Our previous work [111] presents to compose novel HOI samples via combining decomposed verbs and objects between pair-wise images and within images. Nevertheless, it still remains a great challenge to compose massive HOI samples in each minibatch from images due to the limited number of HOIs in each image, especially when the distribution of objects/verbs is also long-tailed. We demonstrate the distribution of the number of objects in Figure 2.4.

The long-tailed distribution of objects/verbs makes it difficult to compose new HOIs from each mini-batch, significantly degrading the performance of compositional learning-based methods for rare and zero-shot HOI detection [111]. Inspired by the recent success of visual object representation generation [271, 86, 264], we thus apply fabricated object representation, instead of fabricated verb representation, to compose more balanced HOIs. We referred to the proposed compositional learning framework with fabricated object representation as Fabricated Compositional Learning or FCL. Specifically, we first extract verb representations from input images, and then design a simple yet efficient object fabricator to generate object representation. Next, the generated visual object features are further combined with the verb features to compose new HOI samples. With the proposed object fabricator, we are able to generate balanced objects for each verb within the mini-batch of training data as well as compose massive balanced HOI training samples.

Furthermore, HOI is different from other compositional generalization challenges, the factors (i.e., verbs and objects) of HOI also exhibit explicit semantics. In detail, the action in the HOI also illustrates the affordances of the objects, i.e., what actions can be applied to a particular object [73]. Different from other compositional methods [128, 208, 176] that merely focus on compositional generalization, VCL is also capable of compositionally reasoning the object affordances. VCL treats each HOI into a verb and an object, in which the verb also indicates one of the possible affordances (or functionalities) of the object [75, 91]. Meanwhile, the integration of different verb and object representations during optimization actually transfers the verb (affordance) representations to different objects, which we termed as affordance transfer learning (ATL), and thus empowers the model to discriminate whether a specific verb-object pair is possible or not, i.e., object affordance recognition. In other words, VCL makes it possible to transfer the shared verb (that indicates also affordance) representations to be semantically combinable with objects. We thus devise a compositional object affordance reasoning method as follows: 1) we maintain a feature bank of decoupled affordance representations from the HOI detection dataset; 2) we extract object representations from additional object detection datasets using the same HOI backbone network; and 3) we combine the object representations with all affordance representations in the feature bank as the input of the HOI classifier. Finally, we are able to obtain a set of HOI predictions, which are further used to infer the object affordances. Moreover, as illustrated in Figure 4.1, with the combination of verb representations from HOI images and object representations from object images, the affordance transfer approach also enables the HOI detection with novel objects, and improves the affordance recognition for novel objects in the new domain. For example, with the shared affordance representation (e.g., “rideable”) between “tiger” and “horse” as illustrated in Figure 4.1, we are able to compose new HOIs (i.e., “ride tiger”), and thus enable the detection of unseen HOIs.

Though VCL can facilitate the compositional generalization and enable the compositional reasoning for object affordance recognition, it merely considers the predefined HOI concepts/categories and object affordances in the dataset, the same as current HOI approaches [224, 111, 235]. For instance, the preprint work [111] simply removes the composite HOIs that are out of label space, ignoring a large number of composite samples with unknown categories

(unlabeled composite samples). As a result, the model is inevitably biased to known object affordances/verbs, and leads to a similar inferior performance to the one in Positive-Unlabeled learning [49, 56, 218]. That is, without negative samples for training, the network will tend to predict high confidence on those impossible verb-object combinations or overfit verb patterns.

Besides, there are still massive reasonable HOI concepts/categories that can be inferred from decoupled verbs and objects from the dataset. For example, there are only 600 HOI categories known in HICO-DET [28], while we can find 9,360 possible verb-object combinations from 117 verbs and 80 objects. Meanwhile, given that the distribution of HOI samples is naturally long-tailed in real-world scenarios, it is quite laborious and challenging to collect all possible HOI categories, especially for categories composed of rare actions and objects. Therefore, we further propose to discover the novel reasonable HOI categories/concepts from known HOI categories and their instances, named as HOI concept discovery.

Object affordance [73] indicates whether each action can be applied into an object, i.e., if a verb-object combination is reasonable, we then discover a novel HOI concept/category. At the same time, two objects with similar attributes usually share the same affordance, i.e., humans usually interact with similar objects in a similar way [73]. For example, cup, bowl, and bottle share the same attributes (e.g., hollow), and all of these objects can be used to "drink with". We can thus infer novel HOI categories, i.e., novel affordances for object classes, from an object affordance perspective. An illustration of unknown HOI detection via concept discovery is shown in Figure 2.5.

Nonetheless, the affordance prediction approach is for each object instance, while HOI concept discovery is for the object class. Though we can estimate the possibility via collecting the average affordance predictions of a large number of object instances, it is time-consuming to predict affordances for a large number of objects. By contrast, we introduce an online HOI concept discovery method, which is able to collect concept confidence in a running mean manner with verb scores of all composite features in mini-batches during training. with the online concept confidence, we can then construct pseudo labels [148] for all composite HOIs belonging to either known or unknown categories. Inspired by this, we further utilize self-training to improve the visual compositional learning framework, dubbed self-compositional

FIGURE 2.5. An illustration of unknown HOI detection via concept discovery. Given some known HOI concepts (e.g., “drink_with cup", “drink_with bottle", and “hold bowl"), the task of concept discovery aims to identify novel HOI concepts (i.e., reasonable combinations between verbs and objects). For example, here we have some novel HOI concepts, “drink_with wine_glass", “fill bowl", and “fill bottle". Specifically, the proposed self-compositional learning framework jointly optimizes HOI concept discovery and HOI detection on unknown concepts in an end-to-end manner.

learning (or SCL), via jointly optimizing all composite representations and improving concept predictions in an iterative manner. Specifically, SCL combines the object representations with different verb representations to compose new samples for optimization, and thus implicitly pays attention to the object representations and improves the discrimination of composite representations. By doing this, we can improve object affordance learning, and then facilitate the HOI concept discovery. Moreover, with the discovered HOI concepts, we can detect HOIs with unknown concepts.

In a nutshell, we introduce a compositional learning framework to explore Human-Object Interaction: 1) improve the compositional generalization, including long-tailed HOI detection and zero-shot HOI detection; 2) enable object affordance recognition with HOI model and

significantly facilitate the performance; 3) present a novel task HOI concept discovery, and largely improves the performance.

2.2 Related Work

2.2.1 Human-Object Interaction

HOI understanding [78] is of great importance for visual relationship reasoning [175, 278] and action understanding [24, 301]. Different approaches have been investigated for HOI understanding from various aspects, including HOI detection [28, 157, 161, 305, 131, 33, 318, 234, 290], HOI recognition [26, 128, 117], video HOI [48, 123], compositional action recognition [180], 3D scene reconstruction [294, 47], video generation [187], and object affordance reasoning [60]. Recently, compositional approaches have been intensively proposed for HOI understanding using the structural characteristic [128, 111, 187, 154]. Meanwhile, DETR-based methods (e.g., Qpic [234]) achieve superior performance on HOI detection. However, these approaches mainly consider the perception of known HOI concepts, and pay no attention to HOI concept discovery. To fulfill the gap between learning on known and unknown concepts, a novel task, i.e., HOI concept discovery, is explored in this chapter. Currently, zero-shot HOI detection also attracts massive interest from the community [224, 9, 204, 111, 110]. However, those approaches merely consider known concepts and are unable to discover HOI concepts. Some HOI approaches [204, 9, 258, 257] expand the known concepts via leveraging language priors. However, that is limited to existing knowledge and can not discover concepts that never appear in the language prior knowledge. HOI concept discovery is able to address the problem, and enable unknown HOI concept detection.

2.2.2 Object Affordance Learning

The notation of affordance is formally introduced in [73], where object affordances are usually those action possibilities that are perceivable by an actor [191, 73, 72]. Noticeably, the action possibilities of an object also indicate the HOI concepts related to the object. Therefore,

object affordance can also represent the existence of HOI concepts. Recent object affordance approaches mainly focus on the pixel-level affordance learning from human interaction demonstration [138, 63, 60, 87, 186, 51, 288]. Yao *et al*. [284] present a weakly supervised approach to discover object functionalities from HOI data in the musical instrument environment. Zhu *et al*. [316] introduce to reason affordances in knowledge-based representation. Recent approaches propose to generalize HOI detection to unseen HOIs via functionality generalization [9] or analogies [204]. However those approaches focus on HOI detection, ignoring object affordance recognition. Specifically, our preliminary work *et al*. [107] introduces an affordance transfer learning (ATL) framework to enable HOI model to not only detect interactions but also recognize object affordances. Inspired by this, we further develop a self-compositional learning framework to facilitate object affordance recognition with HOI model to discover novel HOI concepts for downstream HOI tasks.

2.2.3 Compositional Learning

Disentangled representation learning has attracted increasing attention in various kinds of visual tasks [15, 100, 172, 101, 19] and the importance of Compositional Learning to build intelligent machines is acknowledged [15, 145, 69, 13]. Higgins *et al*.[101] proposed Symbol-Concept Association Network (SCAN) to learn hierarchical visual concepts. Recently, [19] proposed Multi-Object network (MONet) to decompose scenes by training a Variational Autoencoder together with a recurrent attention network. However, both SCAN [101] and MONet [19] only validate their methods on the virtual datasets or simple scenes.

Besides, Compositional GAN [6] was introduced to generate new images from a pair of objects. Recently, Label-Set Operations network (LaSO) [1] combined features of image pairs to synthesize feature vectors of new label sets according to certain set operations on the label sets of image pairs for multi-label few-shot learning. Both Compositional GAN [6] and LaSO [1], however, compose the features from two whole images and depend on generative network or reconstruct loss. In addition, Kato *et al*.[128] introduced a compositional learning method for HOI classification [26] that utilizes the visual-language joint embedding model to the feature of the whole of the image. But [128] did not involve multiple objects detection in the scene.

Ma *et al*. [176] present a concept-guided vision transformer with concept-feature dictionary, to promote relational reasoning and facilitate semantic object-centric correspondence learning. Our visual compositional learning framework differs from them in following aspects: i) it composes interaction features from regions of images, ii) it simultaneously encourages *discriminative* and *shared* verb and object representations.

Recently, there are considerable works [153, 208, 185, 5, 178, 127] investigating compositional zero-shot learning. Those approaches usually incorporate language priors or language embedding to improve the compositional generalization. They do not explore the combinable between attributes and factors, or the compositional reasoning for the attributes recognition (e.g., affordance recognition on HOI).

Generalized Zero/Few-Shot Learning. Different from typical zero/few-shot learning [61, 146, 246], generalized zero/few-shot learning [272] is a more realistic variant, since the performance is evaluated on both seen and unseen classes [216, 29]. The distribution of HOIs is naturally long-tailed [28], i.e., most classes have a few training examples. Moreover, the open long-tailed HOI detection aims to handle the long-tailed, low-shot and zero-shot issues in a unified way. The long-tailed data distribution [122, 93, 114] is one of the challenging problems in visual recognition. Currently, re-sampling [79, 126], specific loss [163, 45, 22, 92], knowledge transfer [263, 171], and data generation [264, 142, 271, 1] are major strategies for imbalanced learning [122, 93, 114]. To make full use of the composition characteristic of HOI, we aim to compose HOI samples by visual feature generation to relieve the open long-tailed issue in HOI detection. Recent feature generation methods [142, 271] mainly depend on Variational Autoencoder [135] and Generative Adversarial Network [76], which usually suffer from the problem of model collapse [214]. Wang *et al*.[264] present a new method for low-shot learning that directly learns to hallucinate examples that are useful for classification. Similar to [264], we compose HOI samples with an object fabricator in an end-to-end optimization without using the adversarial loss.

2.2.4 Semi-Supervised Learning

Semi-supervised learning is a learning paradigm for constructing models that use both labeled and unlabeled data [282]. There are a wide variety of Deep Semi-Supervised Learning methods, such as Generative Networks [136, 227], Graph-Based methods [249, 74], Pseudo-Labeling methods [148, 275, 102]. HOI concept discovery shares a similar characteristic to semi-supervised learning approaches. *H*OI concept discovery has instances of labeled HOI concepts, but no instances of unknown concepts. We thus compose HOI representations for unknown concepts according to [219]. With composite HOIs, concept discovery and object affordance recognition can be treated as PU learning [49]. Moreover, HOI concept discovery requires discriminating whether the combinations (possible HOI concepts) are reasonable and existing. Considering each value of the concept confidences also represents the possibility of the composite HOI, we construct pseudo labels [148, 219] for composite features from the concept confidence matrix, and optimize the composite HOIs in an end-to-end way.

2.3 Methods

In this section, we present the proposed Compositional Learning framework for HOI detection. We first provide an overview of the proposed compositional approaches. Then, we introduce vanilla compositional learning. Next, we introduce a Fabricated Compositional Learning method to address the compositional long-tailed challenge, and further leverage a pretrained HOI model to infer the affordances of an object. Last, we present a method to predict HOI concept confidence with compositional learning, and build pseudo labels for optimizing all composite verb-object pairs with self-training.

2.3.1 Overview

To explore Human-Object Interaction from compositional generalization and compositional reasoning perspectives, we propose the Visual Compositional Learning (VCL) framework to compose HOI samples in the feature space and optimize the samples in an end-to-end way.

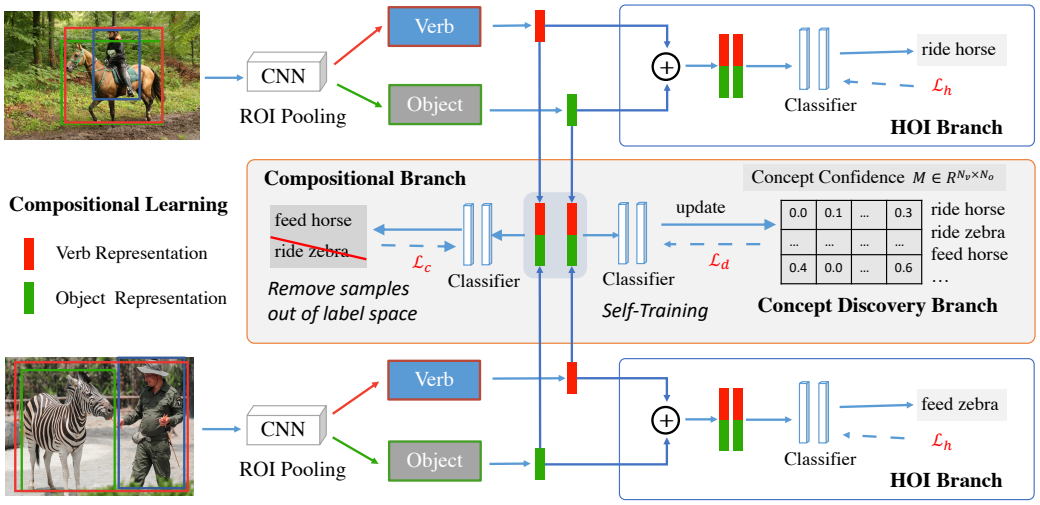

FIGURE 2.6. Overview of the proposed Compositional Learning framework. Given two images, we first detect human and objects with Faster-RCNN [213]. Next, with ROI-Pooling and Residual CNN blocks, we extract verb features (i.e.the union box of human and object) and object features. Then, these features are fed into the following branches: HOI branch, compositional branch and concept discovery branch. In compositional branch and concept discovery branch, verb and object features are further mutually combined to generate composite HOI features, while the verb and object features are combined according to the annotation in HOI branch. Meanwhile, we update the concept confidence $\mathbf{M} \in R^{N_v \times N_o}$, where N_v and N_o are the numbers of verb classes and object classes respectively, with the predictions of all composite HOI features. The concept discovery branch is optimized via a self-training approach to learning from composite HOI features with the concept confidence $\mathbf{M}$. *Note that all the parameters are shared across images and the newly composited HOI instances can be from a single image if the image includes multiple HOIs*

The proposed visual compositional learning framework is able to facilitate HOI compositional generalization, e.g., long-tail learning and compositional zero-shot learning, enables object affordance recognition via compositional reasoning, and discover novel HOI concepts/categories. As shown in Figure 4.2, to perform compositional learning, our method takes as input a randomly selected image pair. We first employ a Faster R-CNN [213] to detect human and objects in images. Subsequently, we use ROI-Pooling and Residual CNN blocks to obtain features of verbs and objects individually. We input the annotated HOI pairs into the HOI branch, and compose new HOI samples for compositional branch among image pairs. The predictions of all verbs and objects in each mini-batch are utilized to update the confidence

matrix $\mathbf{M} \in R^{N_v \times N_o}$, where N_v and N_o are the numbers of verb classes and object classes respectively for HOI concept discovery. Next, the confidence matrix $\mathbf{M}$ is leveraged to build pseudo labels for those composite verb-object pairs that we do not know whether they are reasonable or not. Finally, we optimize all the composite verb-object pairs with pseudo labels in a self-traing manner. The classifier and network are shared among different branches.

2.3.2 Vanilla Compositional Learning

We first devise a vanilla compositional learning approach for HOI compositional generalization based on a popular two-stage HOI detection pipeline. Stitching the verb and object representations among different images or annotated HOIs generates a large number of composite verb-object pairs. With the composite HOI samples in the feature space, we significantly augment the dataset and diversify the sample space for HOI recognition. Meanwhile, the combination approach can also generate samples for those unseen categories that do not have training instances, and thus relieve the unseen distribution challenges.

Given another verb representation $\hat{x}_v$ (sharing the same label $\mathbf{l}_v$ with x_v), and another object representation $\hat{x}_o$ (sharing the same label $\mathbf{l}_o$ with x_o), regardless of the sources of the verb and object representations, an effective composition of verb and object should be

$$g_{hoi}(\hat{x}_v, \hat{x}_o) \approx g_{hoi}(x_v, x_o), \tag{2.1}$$

where g_{hoi} indicates the HOI classification network. By doing this, we can compose new verb-object pair $\langle \hat{x}_v, \hat{x}_o \rangle$, which have similar semantic type $\mathbf{y}$ to the real pair $\langle x_v, x_o \rangle$, to relieve the scarcity of rare and unseen HOI categories. To generate effective verb-object pair $\langle \hat{x}_v, \hat{x}_o \rangle$, we regularize the verb representation $\hat{x}_v$ and object representation $\hat{x}_o$ such that same verbs/objects have similar feature representations. Specifically, the compositional loss is defined as follows:

$$\mathcal{L}_c = \mathcal{L}_{BCE}(g_{hoi}(\hat{x}_v, \hat{x}_o), \hat{\mathbf{y}}), \tag{2.2}$$

Where $\hat{\mathbf{y}}$ indicates the labels for the composite HOIs.

Nevertheless, the composite verb-object pairs might also include infeasible pairs, e.g., $\langle feed, bicycle \rangle$. Therefore, it requires to select the feasible HOI samples. We first introduce a vanilla strategy to select the composite samples. Specifically, we only keep the composite verb-object pairs in the label space and remove the composite pairs out of label space. Though simplicity, this strategy also effectively improves the compositional generalization for those categories that are in the label space.

2.3.3 Fabricated Compositional Learning

The motivation of compositional learning is to decompose a model/concept into several sub-models/concepts, in which each sub-model/concept focuses on a specific task, and then all responses are coordinated and aggregated to make the final prediction [17]. Recent compositional learning method for HOI detection considers each HOI as the combination of a verb and an object to compose new HOIs from objects and verbs within the mini-batch of training samples [128, 111]. However, existing compositional learning methods fail to address the problem of long-tailed distribution on objects.

To address the open long-tailed issue, we propose to generate balanced objects for each decoupled visual verb as follows. Similar to previous approaches, such as factor visual-language joint embedding [277, 204] and factorized model [224, 81], when $\hat{x}_v$ is similar to x_v and $\hat{x}_o$ is similar to x_o, we then have that Equation (2.1) can be generalized to HOI detection via the compositional branch. We refer to the proposed compositional learning framework with fabricated object representation as Fabricated Compositional Learning or FCL. We train the proposed fabricated compositional method with composited HOI samples $\langle \hat{x}_v, \hat{x}_o \rangle$ in an end-to-end manner.

2.3.3.1 Object Generation

The HOI is composed of a verb and an object, in which the verb is usually a very abstract notation compared to the object, making it difficult to directly generate verb features. Recent visual feature generation methods have demonstrated the effectiveness of feature generation

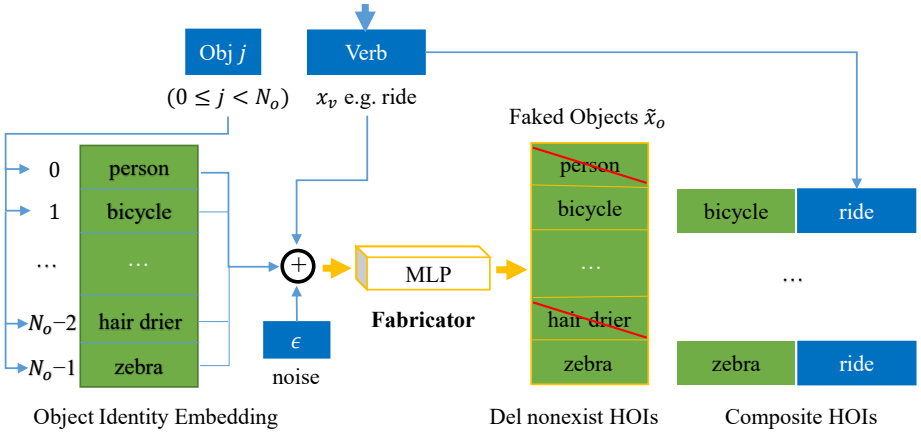

FIGURE 2.7. For a given visual verb feature and each j_{th} $(0 \leq j < N_o)$, we firstly select the j_{th} object identity embedding. Then, we concatenate verb feature, object embedding and Gaussian noise to input to fabricator for generating a fake object feature. We can fabricate N_o objects for a verb feature. We finally remove nonexisting HOIs as described in Section 3.2.2.

for visual object recognition [264, 271]. Therefore, we devise an object fabricator to generate object feature representations for composing novel HOI samples.

The overall framework of object generation is shown in Figure 2.7. Specifically, we maintain a pool of object identity embeddings, i.e., v_{id}. In each HOI, the pose of the object is usually influenced by the human who is interacting the object [294], and the person who is interacting with the object is firmly related to verb feature representation. Thus, for each extracted verb and the j_{th} object $(0 \leq j < N_o$ and N_o is the number of all different objects), we concatenate the j_{th} object identity embedding v_{id}^j, the verb feature x_v and a noise vector $\epsilon \sim \mathcal{N}(0, 1)$, as the input of the object fabricator, i.e.,

$$\hat{x}_o = f_{obj}(\{v_{id}^j, x_v, \epsilon\}), \tag{2.3}$$

where $\hat{x}_o$ is the fake object feature and f indicates the object fabricator network. Here, the noise ϵ is used to increase the diversity of generated objects. We then combine the fake object feature $\hat{x}_o$ and the verb x_v to compose a new HOI sample $\langle x_v, \hat{x}_o \rangle$. Specifically, during training, both real HOIs and composite HOIs share the same HOI classification network g_{hoi}.

We further devise a regularization loss $\mathcal{L}_{reg}$ to regularize the verb representations as follow,

$$\mathcal{L}_{reg} = \mathcal{L}_{BCE}(x_v, \mathbf{l_v}) \tag{2.4}$$

where $\mathbf{l_v}$ is the label (verb category) of the verb representations. $\mathcal{L}_{reg}$ aims to regularize verb features. Specifically, object features extracted from a pre-trained object detector backbone network (i.e.Faster-RCNN [213]) are usually discriminative. Thus, we only regularize verb representation.

2.3.4 Label Decomposition and Composition

In the proposed method, we should decompose the annotated HOIs into verbs and objects, and further compose novel HOIs. For the feature decomposition and composition, we use bounding boxes to decompose and the concatenation operation to compose. For the corresponding HOI label process, we devise a simple way as follows.

Given two images I_1 and I_2, we compose new interaction samples within a single image and between images by first considering all possible verb-object pairs and then removing infeasible interactions in the HOI label space.

Existing HOI labels mainly contain one object and at least one verb, which set the HOI detection as a multi-label problem. To avoid frequently checking verb-object pairs, we design an efficient composing and removing strategy. First, we decouple the HOI label space into a verb-HOI matrix $\mathbf{A}_v \in R^{N_v \times C}$ and an object-HOI matrix $\mathbf{A}_o \in R^{N_o \times C}$, where N_v, N_o, and C denote the number of verbs, objects and HOI categories respectively. $\mathbf{A}_v$ ($\mathbf{A}_o$) can be viewed as the co-occurrence matrix between verbs (objects) and HOIs. Then, given binary HOI label vectors $\mathbf{y} \in R^{N \times C}$, where N, C denote the number of interactions and HOI categories (Annotated in dataset) respectively. We can obtain the object label vector and verb label vector as follows,

$$\mathbf{l}_o = \mathbf{y}\mathbf{A}_o^{\mathsf{T}},\ \mathbf{l}_v = \mathbf{y}\mathbf{A}_v^{\mathsf{T}}, \tag{2.5}$$

where $\mathbf{l}_o \in R^{N \times N_o}$ is usually one-hot vectors meaning one object of a HOI example, and $\mathbf{l}_v \in R^{N \times N_v}$ is possibly multi-hot vectors meaning multiple verbs. e.g.$\langle \{hold, sip\}, cup \rangle$). Similarly, we can generate new interactions from arbitrary $\mathbf{l}_o$ and $\mathbf{l}_v$ as follows,

$$\hat{\mathbf{y}} = (\mathbf{l}_o \mathbf{A}_o) \& (\mathbf{l}_v \mathbf{A}_v), \tag{2.6}$$

where $\&$ denotes the "and" logical operation. The infeasible HOI labels that do not exist in the given label space are all-zero vectors after the logical operation. And then, we can filter out those infeasible HOIs. In the implementation, we obtain verbs and objects from two images by ROI pooling and treat them within and between images as the same. Therefore, we do not treat two levels of composition differently during composing HOIs.

2.3.5 Affordance Transfer Learning

In addition to demonstrating what the interaction looks like, Human-Object Interaction also illustrates how to apply the actions to the target objects, i.e., Object Affordance. In this section, we devise an affordance transfer approach based on the visual compositional learning framework to transfer the object affordances from HOI images to object images. Specifically, we first decompose the verb representations from HOI images and the object representation from object images, and then combine the object representation and verb representation to compose novel HOI features, and finally optimize the composite features and real features in an end-to-end way, as illustrated in Figure 2.8.

2.3.5.1 Compositional Object Affordance Reasoning

The compositional approach not only facilitates compositional generalization for HOI detection, but also enables compositional object affordance reasoning. In this subsection, we introduce how to infer the object affordance during the testing phase. Considering that we jointly optimize the decoupled components (i.e., object features and affordance features from object and HOI images) in HOI samples and novel object samples with affordance transfer learning, the proposed method thus is able to distinguish whether a novel object is combinable

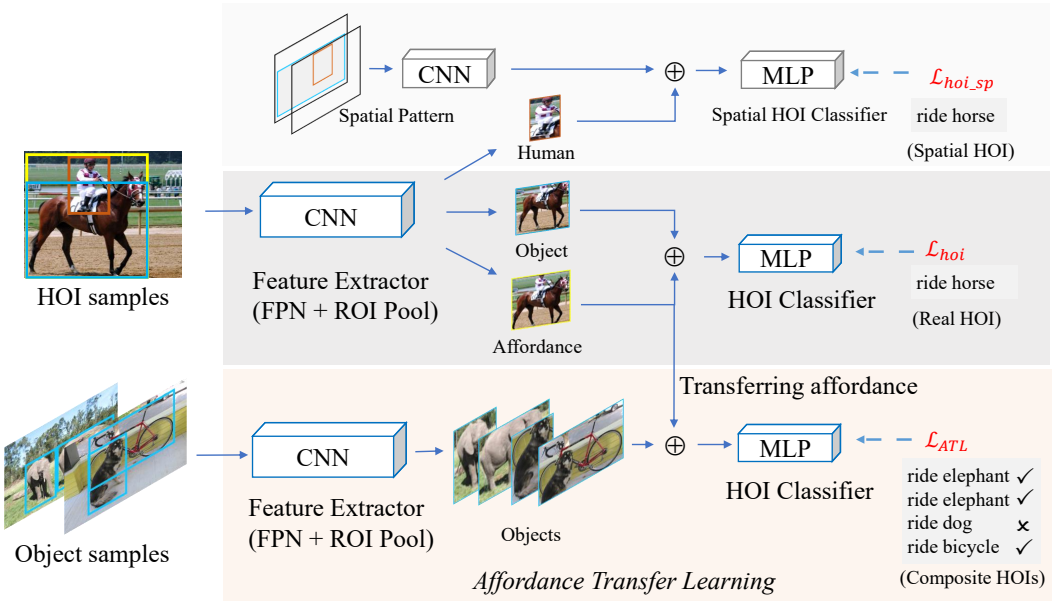

FIGURE 2.8. An overview of affordance transfer learning or ATL for HOI detection. We first extract the human, object, and affordance features via the ROI-Pooling from the feature pyramids [213], respectively. Meanwhile, we also extract new object features from an additional object dataset using the same backbone network. After that, we concatenate the affordance and the object features (from HOI datasets) as the real HOIs. We also compose new HOIs using the affordance features and the object features extracted from additional object datasets, which transfer the affordance to novel objects. Both the composite HOIs and real HOIs share the same HOI classifier. In addition, human features and spatial pattern features are combined to construct the spatial HOI branch.

or not with a specific affordance (i.e., valid HOIs). Therefore, we design a simple yet effective object affordance recognition method using the HOI detection model. Specifically, we first build an affordance feature bank as follows.

Affordance Feature Bank. We construct the affordance feature bank from HOI datasets (e.g.HICO-DET and HOI-COCO). In order to reduce storage space and computation, we randomly choose a maximum of L instances for each affordance in HICO-DET. Then, we extract the features of those affordances to construct an off-the-shelf affordance feature bank.

Given an object feature extracted from the object image, we combine it with all affordances in the feature bank to obtain a set of HOIs. As illustrated in Figure 2.9, we obtain all HOI predictions from the HOI classifier. After that, we are able to convert all HOI predictions to

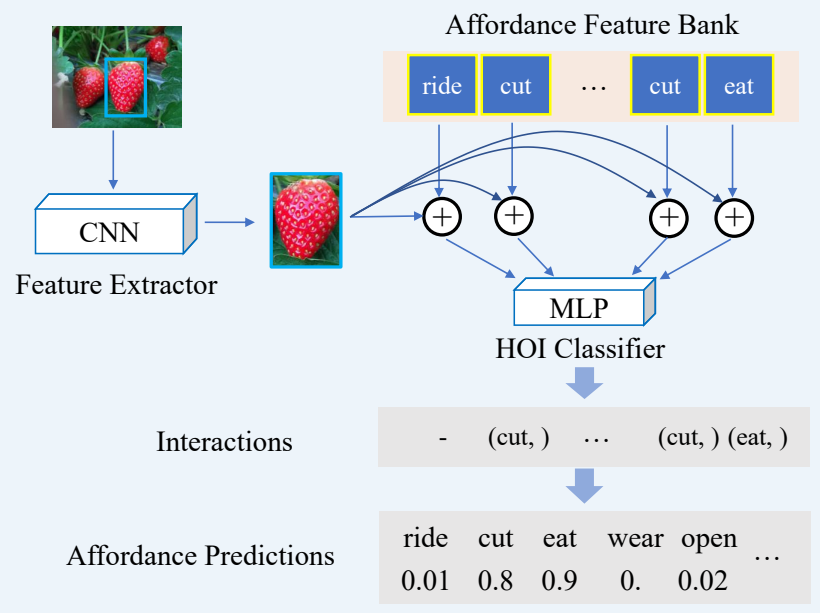

FIGURE 2.9. An illustration of compositional object affordance reasoning with HOI network. Here, we use verb to represent affordance. We first construct an affordance feature bank from the decoupled affordance representations. For any object (e.g.strawberry), we extract the object feature by the Feature Extractor according to the bounding box. Then, the object feature is combined with all affordances in the bank to input into HOI classifier for obtaining predicted interactions. The interactions are further converted into affordances (e.g.eatable).

affordance predictions according to the HOI-verb co-occurrence matrix $\mathbf{A}_v$. Specifically, we remove the predicted affordances whose label is not the same as the corresponding affordance labels in the feature bank. As a result, we obtain a list of affordances with many repeated elements. Let F_i denotes the frequency (count) of the affordance i and S_i indicate the number of affordance (or verb) i in the feature bank, we evaluate the probability of the affordance i as $\frac{F_i}{S_i}$. Via averaging the prediction for each object class, we can obtain concept discovery confidence.

2.3.6 Concept Discovery

As shown in Figure 4.2, we keep an HOI concept confidence vector during training, $\mathbf{M} \in R^{N_v N_o}$, where each value represents the concept confidence of the corresponding combination

between a verb and an object. To achieve this, we first extract all verb and object representations among pair-wise images in each batch as $\mathbf{x}_v$ and $\mathbf{x}_o$. We then combine each verb representation and all object representations to generate the composite HOI representations $\mathbf{x}_h$. After that, we use the composite HOI representations as the input to the verb classifier and obtain the corresponding verb predictions $\hat{\mathbf{Y}}_v \in R^{NN \times N_v}$, where N indicates the number of real HOI instances (i.e., verb-object pair) in each mini-batch and NN is then the number of all composite verb-object pairs (including unknown HOI concepts). Let $\mathbf{Y}_v \in R^{N \times N_v}$ and $\mathbf{Y}_o \in R^{N \times N_o}$ denote the label of verb representations $\mathbf{x}_v$ and object representations $\mathbf{x}_o$, respectively. We then have all composite HOI labels $\mathbf{Y}_h = \mathbf{Y}_v \otimes \mathbf{Y}_o$, where $\mathbf{Y}_h \in R^{NN \times N_v N_o}$, and the superscripts h, v, and o indicate HOI, verb, and object, respectively, $\otimes$ indicate kronecker product. Similar to affordance prediction, we repeat $\hat{\mathbf{Y}}_v$ by N_o times to obtain concept predictions $\hat{\mathbf{Y}}_h \in R^{NN \times N_v N_o}$. Finally, we update $\mathbf{M}$ in a running mean manner [118] as follows,

$$\mathbf{M} \leftarrow \frac{\mathbf{M} \odot \mathbf{C} + \sum_i^{NN} \hat{\mathbf{Y}}_h(i,:) \odot \mathbf{Y}_h(i,:)}{\mathbf{C} + \sum_i^{NN} \mathbf{Y}_h(i,:)}, \tag{2.7}$$

$$\mathbf{C} \leftarrow \mathbf{C} + \sum_i^{NN} \mathbf{Y}_h(i,:), \tag{2.8}$$

where $\odot$ indicates the element-wise multiplication, $\hat{\mathbf{Y}}_h(i,:) \odot \mathbf{Y}_h(i,:)$ aims to filter out predictions whose labels are not $\mathbf{Y}_h(i,:)$, each value of $\mathbf{C} \in R^{N_v N_o}$ indicates the total number of composite HOI instances in each verb-object pair (including unknown HOI categories). Actually, $\hat{\mathbf{Y}}_h(i,:) \odot \mathbf{Y}_h(i,:)$ follows the affordance prediction process [107]. The normalization with $\mathbf{C}$ is to avoid the model bias to frequent categories. Specifically, both $\mathbf{M}$ and $\mathbf{C}$ are zero-initialized. With the optimization of HOI detection, we can obtain the vector $\mathbf{M}$ to indicate the HOI concept confidence of each combination between verbs and objects.

2.3.7 Self-Training

Existing HOI compositional learning approaches [111, 110, 107] usually only consider the known HOI concepts and simply discard the composite HOIs out of label space during optimization. Therefore, there are only positive data for object affordance learning, leaving a

large number of unlabeled composite HOIs ignored. Considering that the concept confidence on HOI concept discovery also demonstrates the confidence of affordances (verbs) that can be applied to an object category, we thus try to explore the potential of all composite HOIs, i.e., both labeled and unlabeled composite HOIs, in a semi-supervised way. Inspired by the way used in PU learning [49] and pseudo-label learning [148], we devise a self-training strategy by assigning the pseudo labels to each verb-object combination instance using the concept confidence matrix $\mathbf{M}$, and optimize the network with the pseudo labels in an end-to-end way. With the self-training, the online concept discovery can gradually improve the concept confidence $\mathbf{M}$, and in turn, optimize the HOI model for object affordance learning with the concept confidence. Specifically, we construct the pseudo labels $\tilde{\mathbf{Y}}_v \in R^{NN \times N_v}$ from the concept confidence matrix $\mathbf{M} \in R^{N_v \times N_o}$ for composite HOIs $\mathbf{x}_h$ as follows,

$$\tilde{\mathbf{Y}}_v(i,:) = \sum_{j}^{N_o} \frac{\mathbf{M}(:,j)}{\max(\mathbf{M})} \odot \mathbf{Y}_h(i,:,j), \tag{2.9}$$

where $0 \leq j < N_o$ indicates the index of object category, $0 \leq i < NN$ is the index of HOI representations. Here, N is the number of HOIs in each mini-batch, and is usually very small on HICO-DET and V-COCO. Thus the time complexity of Equation 2.9 is small. The labels of composite HOIs are reshaped as $\mathbf{Y}_h \in R^{NN \times N_v \times N_o}$. Noticeably, in each label $\mathbf{Y}_h(i,:,:)$, there is only one vector $\mathbf{Y}_h(i,:,j)$ larger than 0 because each HOI has only one object. As a result, we obtain pseudo verb label $\tilde{\mathbf{Y}}_v(i,:)$ for HOI $\mathbf{x}_{h_i}$. Finally, we use composite HOIs with pseudo labels to train the models, and the loss function is defined as follows,

$$\mathcal{L}_d = \frac{1}{NN} \sum_{i}^{NN} (\frac{1}{N_v} \sum_{k}^{N_v} \mathcal{L}_{\text{BCE}}(\frac{\mathbf{Z}(i,k)}{T}, \tilde{\mathbf{Y}}_v(i,k))), \tag{2.10}$$

where $\mathbf{Z}(i,:)$ is the prediction of the i-th composite HOI, $0 \leq k < N_v$ means the index of predictions, T is the temperature hyper-parameter to smooth the predictions (the default value is 1 in the experiment), $\mathcal{L}_{\text{BCE}}$ indicates the binary cross entropy loss.

2.3.8 Optimization

During the training stage, we train the proposed method with typical loss L_h for the annotated HOIs in HOI branch, a loss L_c for compositional branch, which removes infeasible composite HOIs, and L_d for all composite HOIs with pseudo labels in concept discovery branch. Meanwhile, similar to [66, 157], we can also incorporate spatial patterns via a unique branch to facilitate HOI detection. We use L_{other} to indicate this loss. Lastly, the overall training loss function is defined as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_h + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_d + \mathcal{L}_{other} + \lambda_4 \mathcal{L}_{reg}, \tag{2.11}$$

where λ_1, λ_2, λ_3 and λ_4 are four hyper-parameters to balance different losses. Both the feature extractors and the classifier modules are jointly trained in an end-to-end manner. $\mathcal{L}_h$, $\mathcal{L}_h$, $\mathcal{L}_c$, $\mathcal{L}_d$ are binary cross entropy losses. $\mathcal{L}_{other}$ can be the additional losses, e.g., the loss for spatial features, which enhances the HOI representations for classification.

During the testing stage, the compositional learning modules are not necessary.

2.4 Experiments

2.4.1 Datasets and Evaluation Metrics

2.4.1.1 Dataset

We evaluate the proposed method on HICO-DET [28] and V-COCO [80] for HOI detection. For object affordance recognition, we build HOI-COCO and use HICO-DET for training. We evaluate the performance of object affordance recognition on HICO-DET test set, HOI-COCO test set, COCO [164] validation set, and Object365 [223] subset. Those datasets are listed as follows,

HICO-DET [28] dataset consists of 38,118 images in the training set and 9,658 test images over 600 types of interactions (80 object categories in COCO dataset and 117 unique verbs) with over 90,000 HOI instances.

HOI-COCO is built from the V-COCO dataset [80], which contains 10,346 images with 16,199 person instances. Each annotated person in V-COCO has binary labels for 26 different actions. V-COCO mainly focuses on verb recognition, and has limited object categories (only two). Thus we construct a new benchmark HOI-COCO for the evaluation of verb-object pairs as follows. We use 21 actions from all 26 actions in V-COCO (i.e., five non-interaction actions, “walk", “run, “smile", “stand" and “point ", are removed). As a result, we build HOI-COCO benchmark with 222 HOI categories over 21 verbs and 80 objects. Meanwhile, we use the same train/val split in V-COCO for HOI-COCO. Similar to HICO-DET [28], we evaluate the performance on HOI-COCO under three different settings: Full (222 types), Rare (97 types), and NonRare (115 types). The HOI type in Rare category contains less than 10 training instances, and the distribution of HOI categories is long-tailed.

COCO [164] dataset is a widely-used benchmark for common object detection with 80 different object classes. Considering that both HICO-DET [28] and HOI-COCO consist of the same object label sets to COCO, we thus directly incorporate the COCO dataset as the additional object dataset in our experiments.

Object365 [223] is a recently proposed large-scale common object detection dataset with 365 object categories. The domain of Object365 is different from COCO [164]. In detail, we select objects that are labeled as COCO classes from Object365 validation dataset to evaluate the affordance recognition of objects on new domain. Meanwhile, we choose 12 new types of objects and label manually the affordance of those objects according to the HICO-DET and HOI-COCO, respectively. Those objects are used to evaluate affordance recognition on new types of objects.

Moreover, We extend two popular HOI detection datasets, HICO-DET [28] and V-COCO [80], to evaluate the performance of different methods for HOI concept discovery. Specifically, we first manually annotate all the possible verb-object combinations on HICO-DET (117 verbs and 80 objects) and V-COCO (24 verbs and 80 objects). As a result, we obtain 1,681 concepts on HICO-DET and 401 concepts on V-COCO, i.e., 1,681 of 9,360 verb-object combinations on HICO-DET and 401 of 1,920 verb-object combinations on V-COCO are reasonable. Besides, 600 of 1,681 HOI concepts on HICO-DET and 222 of 401 HOI concepts

on V-COCO are known according to existing annotations. Thus, the HOI concept discovery task requires discovering the other 1,081 concepts on HICO-DET and 179 concepts on V-COCO.

2.4.1.2 Evaluation Metrics

We follow the standard evaluation metric [66, 277] and report mean average precision for HICO-DET dataset [28], V-COCO and HOI-COCO. A prediction is a true positive only when the detected human and object bounding boxes have IoUs larger than 0.5 with reference to ground truth, and the HOI category is accurately predicted. Object affordance recognition is a multi-label classification problem (i.e.an object usually has multiple affordances). Thus, we compare mean Average Precision for evaluating object affordance recognition. HOI concept discovery aims to discover all reasonable combinations between verbs and objects according to existing HOI training samples. We report the performance by using the average precision (AP) for concept discovery and mean AP (or mAP) for object affordance recognition. For HOI detection, we also report the performance using mAP.

2.4.2 Implementation Details

For HICO-DET, similar to recent methods [10], we use the object detector fine-tuned on HICO-DET. For HOI-COCO, we directly use the object detector pre-trained on COCO. Besides, all HOI classifiers consist of two fully-connected layers with 1024 hidden units. To compare with recent methods on HICO-DET, we use two object images in each mini-batch. On HOI-COCO, we only use one object image for evaluation. Following [110], we also include a sigmoid loss for verb representation and the loss weight is 0.3 on HICO-DET. During training, following [66, 157, 111], we augment the ground truth boxes via random crop and random shift. During inference, we keep human and objects with the score larger than 0.3 and 0.1 on HICO-DET respectively. $\lambda_1 = 2$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ on HICO-DET, and $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ on V-COCO/HOI-COCO, respectively. $\lambda_4 = 0.3$ is only used for HICO-DET. To prevent composite interactions from dominating the training of the model, we keep the number of composite interactions not more than the number of objects

in each mini-batch by randomly sampling composite HOIs. We train the model for 1.2M iterations on HICO-DET dataset and 300K iterations on HOI-COCO with an initial learning rare of 0.01. For object affordance recognition, we use the actions of each HOI dataset as affordances and remove the “no interaction” categories on HICO-DET dataset. We keep the object affordance predictions if the affordance score is large than 0.5. For self-training on affordance recognition and concept discovery, we use a modified HOI compositional learning framework, i.e., we directly predict the verb classes and optimize the composite HOIs using SCL. For self-training, we remove the composite HOIs when its corresponding concept confidence is 0, i.e., the concept confidence has not been updated. If not stated, the backbone is ResNet-101. The Classifier is a two-layer MLP. We train the model for 3.0M iterations on HICO-DET and 300K iterations on HOI-COCO with an initial learning rate of 0.01. For zero-shot HOI detection, we keep human and objects with the score larger than 0.3 and 0.1 on HICO-DET, respectively. In our experiment, L is 100 in consideration of the computation and efficiency. Experiments are conducted using a single Tesla V100 GPU (16GB), except for experiments on Qpic [234], which uses four V100 GPUs with PyTorch [197].

2.4.3 HOI detection

2.4.3.1 Fabricated Compositional Learning

We compare FCL with recent state-of-the-art HOI detection approaches [259, 161, 9, 111, 67] using fine-tuned object detector on HICO-DET to validate its effectiveness on long-tailed HOI detection. For fair comparison, we use the same fine-tuned object detector provided by [111]. For evaluation, we follow the settings in [28]: Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) in “Default" and “Known Object" on HICO-DET.

HICO-DET In Table 2.1, we find that the proposed method achieves new state-of-the-art performance, **24.68%** and **26.80%** mAP on “Default" and “Known Object". Meanwhile, we achieve a significant performance improvement of **2.82%** over the contemporary best rare performance model [111] under the same object detector, which indicates the effectiveness of the proposed compositional learning for the long-tailed HOI detection. Furthermore, with the

TABLE 2.1. Comparison to the state-of-the-art approaches on HICO-DET dataset [28]. FCL ^{DRG} is FCL with object detector provided by [67]. FCL + VCL means we fuse the result provided in [111] with FCL. VCL^{DRG} uses the released model of VCL.

Method	Default			Known Object		
	Full	Rare	NonRare	Full	Rare	NonRare
FG [9]	21.96	16.43	23.62	-	-	-
IP-Net [259]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [161]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [111]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [67]	24.53	19.47	26.04	27.98	23.11	29.43
Baseline	23.35	17.08	25.22	25.44	18.78	27.43
FCL	**24.68**	**20.03**	**26.07**	**26.80**	**21.61**	**28.35**
FCL + VCL	**25.27**	**20.57**	**26.67**	**27.71**	**22.34**	**28.93**
VCL [111] ^{DRG}	28.33	20.69	30.62	30.59	22.40	33.04
Baseline^{DRG}	28.12	21.07	30.23	30.13	22.30	32.47
FCL ^{DRG}	**29.12**	**23.67**	**30.75**	**31.31**	**25.62**	**33.02**
(FCL + VCL) ^{DRG}	**30.11**	**24.46**	**31.80**	**32.17**	**26.00**	**34.02**
VCL [111] ^{GT}	43.09	32.56	46.24	-	-	-
FCL^{GT}	**44.26**	**35.46**	**46.88**	-	-	-
(FCL + VCL)^{GT}	**45.25**	**36.27**	**47.94**	-	-	-

TABLE 2.2. Illustration of Fabricated Compositional Learning on V-COCO based on PMFNet [248]

Method	AP_{role}
PMFNet [248]	52.0
Baseline	51.85
FCL	**52.35**

same object detection result to [67], our results surprisingly increase to **29.12%** on "Default" mode. Here, we merely change the detection result provided in [111] to that provided in [67] during inference. Particularly, we find our method is complementary to compose HOIs between images [111]. By simply fusing the result provided by [111] with FCL, we can further largely improve the results under different object detectors.

V-COCO We also evaluate FCL on V-COCO. Although the data on V-COCO is balanced, FCL still improves the baseline (reproduced PMFNet [248]) in Table A.8.

TABLE 2.3. Comparison to recent state-of-the-art methods with fine-tuned detector on HICO-DET dataset [28]. The content in brackets indicates the source of the object images. The last two rows are one-stage HOI detection results.

Method	Default			Known Object		
	Full	Rare	NonRare	Full	Rare	NonRare
FG [10]	21.96	16.43	23.62	-	-	-
IP-Net [259]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [161]	21.73	13.78	24.10	24.58	16.65	26.84
DRG [67]	24.53	19.47	26.04	27.98	23.11	29.43
VCL [111]	23.63	17.21	25.55	25.98	19.12	28.03
ATL (HICO-DET)	23.67	17.64	25.47	26.01	19.60	27.93
ATL (COCO)	**24.50**	**18.53**	**26.28**	**27.23**	**21.27**	**29.00**
ATL (HICO-DET) ^{DRG}	27.68	20.31	29.89	30.05	22.40	32.34
ATL (COCO) ^{DRG}	**28.53**	**21.64**	**30.59**	**31.18**	**24.15**	**33.29**
Baseline (One-Stage)	22.77	16.54	24.63	26.31	21.60	27.72
ATL (One-Stage)	**23.81**	**17.43**	**25.72**	**27.38**	**22.09**	**28.96**

2.4.3.2 Affordance Transfer Learning

HICO-DET. We report the performance on three different settings: Full (600 categories), Rare (138 categories) and NonRare (462 categories) in "Default" and "Known" modes on HICO-DET. As shown in Table 2.3, the proposed method outperforms recent state-of-the-art methods among all categories. Furthermore, with better object detection results provided in [67], the performance of ATL dramatically increases to **28.53%**. Meanwhile, we find ATL is more effective on the Rare category. Specifically, when using the objects from the training set of HICO-DET, the proposed method is similar to VCL [111] as shown in Table 2.3. ATL also improves the baseline effectively based on the One-Stage method. Here, the baseline is the model without compositional learning.

HOI-COCO. We find the proposed method has similar performance to VCL when using HOI-COCO as the source of object images in Table 2.4. Here, we evaluate the performance of VCL on HOI-COCO dataset using the official code from [111]. When using the COCO object dataset, the proposed method signifcantly improves the performance, especially on Rare categories, e.g., over **1.5%** than VCL and **2.9%** than the baseline, respectively. Meanwhile, the proposed method also gives a larger improvement than baseline in NonRare category

TABLE 2.4. Comparison to recent state-of-the-art methods on HOI-COCO dataset.

Method	object data	Full	Rare	NonRare
Baseline	-	22.86	6.87	35.27
VCL [111]	HOI-COCO	23.53	8.29	35.36
ATL	HOI-COCO	23.40	8.01	35.34
ATL	COCO	24.84	9.79	36.51
ATL	COCO, HICO-DET	**25.29**	**9.85**	**37.27**

comparing with VCL, suggesting that ATL also increases the diversity of HOIs via composing new samples. Furthermore, when using both HICO-DET and COCO to provide object images, we further improve the performance to **25.29%**.

2.4.4 Zero-Shot HOI detection

The proposed affordance transfer learning enables the detection of HOIs with novel objects due to the mechanism of composing HOI samples of unseen classes. Therefore, we evaluate the proposed method for zero-shot HOI detection on HICO-DET [28]. We report the performance on two settings: 1) Unseen Composition and 2) Novel Object. Specifically, Unseen Composition means there are unseen HOIs in the test but the verbs and objects of the unseen HOIs exist in training data, while the objects of unseen HOIs in novel object HOI detection do not exist in training data. For compositional zero-shot learning, we follow [111] to evaluate on rare-first unseen HOIs (firstly select tail HOIs in HICO-DET as unseen data) and non-rare first unseen HOIs (firstly select head HOIs in HICO-DET as unseen data). We evaluate zero-shot HOI detection on three categories: Unseen (120 categories), Seen (480 categories) and Full (600 categories). For novel object HOI detection, similar to [10], we choose 100 unseen categories (including 12 unseen objects) and 500 seen categories. We choose the object detector provided in [111] to compare fairly with [111].

2.4.4.1 Fabricated Compositional Learning

There are different settings [9] for zero-shot HOI detection: 1) unseen composition; and 2) unseen object. Specifically, for the unseen composition setting, it indicates that the training

TABLE 2.5. Comparison of zero-shot detection results of our proposed method. UC indicates unseen composition zero-shot HOI detection. UO indicates unseen object zero-shot HOI detection. For better illustration, we choose the mean UC result of [9].

Method	Type	Unseen	Seen	Full
Shen *et al.*[224]	UC	5.62	-	6.26
FG [9]	UC	11.31	12.74	12.45
VCL [111] (rare first)	UC	10.06	24.28	21.43
Baseline (rare first)	UC	8.94	24.18	21.13
Factorized (rare first)	UC	7.35	22.19	19.22
FCL (rare first)	UC	**13.16**	24.23	**22.01**
VCL [111] (non-rare first)	UC	16.22	18.52	18.06
Baseline (non-rare first)	UC	13.47	19.22	18.07
Factorized (non-rare first)	UC	15.72	16.95	16.71
FCL (non-rare first)	UC	**18.66**	**19.55**	**19.37**
FG [9]	UO	11.22	14.36	13.84
Baseline	UO	12.86	20.77	19.45
FCL	UO	**15.54**	20.74	**19.87**

data contains all factors (i.e., verbs and objects) but misses the verb-object pairs; for the unseen object setting, it requires to detect unseen HOIs, in which the object do not appear in the training data. For unseen composition HOI detection, similar to [111], we select two groups of 120 unseen HOIs from tail preferentially (rare first) and from head preferentially (non-rare first) separately, which roughly compares the lowest and highest performances. As a result, we report our result in the following settings: Unseen (120 HOIs), Seen (480 HOIs), Full (600 HOIs) in the “Default" mode on HICO-DET dataset. For a better comparison, we implement the factorized model [224] under our framework for unseen composition zero-shot HOI detection. For unseen object HOI detection, we use the same HOI categories for unseen data as [9] (i.e.randomly selecting 12 objects from the 80 objects and picking all HOIs containing those objects as unseen HOIs). Then, we report our results in the setting: Unseen (100 HOIs), Seen (500 HOIs), Full (600 HOIs). To compare with the contemporary work [111], we use the same object detection result released by [111]. Here, our baseline method is the model without object fabricator, i.e., the compositional branch.

Unseen composition. Table 2.5 shows that FCL achieves large improvement on Unseen category by **4.22%** and **5.19%** than baseline, and by **3.10% and 2.44%** compared to previous works [9, 111] on the two selection strategies respectively. Meanwhile, the two selection

strategies witness a consistent improvement with FCL on nearly all categories, which indicates that composing novel HOI samples contributes to overcome the scarcity of HOI samples. In rare first selection, FCL has a similar result to baseline and VCL [111] on Seen category. But step-wise optimization can improve the result on Seen category and Full category (See Table 2.14). In addition, the factorized model has a very poor performance in the head classes compared to our baseline. Noticeably, factorized model achieves better performance on Unseen category than the baseline in non-rare first selection while has worse results on Unseen category in the rare first selection. FCL witnesses a consistent improvement in different evaluation settings. In the remaining data, unseen HOIs of rare first zero-shot have more rare verbs (less than 10 instances) than that of non-rare first zero-shot.

Unseen object. We further evaluate FCL in novel object zero-shot HOI detection, which requires to detect HOIs that is interacting with novel objects. Table 2.5 shows FCL effectively improves the baseline by 2.68% on Unseen Category, although there are no real objects of unseen HOIs in training set. This illustrates the ability of FCL for detecting unseen HOIs with novel objects. Here, the same as [9], we also use a generic detector to enable unseen object detection.

2.4.4.2 Affordance Transfer Learning

Compositional Zero-Shot HOI Detection. In Table 2.6, we find our approach effectively improves the non-rare first zero-shot HOI detection. Meanwhile, our approach achieves better result on seen category in rare first zero-shot HOI detection. Particularly, the affordances in tail part of HOIs are usually rare, the composite samples of tail HOIs with additional objects are much less than that of head HOIs. Therefore, our approach achieves even worse result on unseen category.

Novel Object HOI Detection. Table 2.6 demonstrates that transferring affordance representation to novel objects effectively facilitates the detection of unseen HOIs with novel objects. Here we use the network without affordance transfer learning as our baseline. We find using HICO-DET (remove HOIs with unseen objects) as object images even degrades the performance on unseen categories compared to the baseline because we compose massive seen HOI

TABLE 2.6. Comparison of Zero Shot Detection results of our proposed method. UC means unseen composition HOI detection. NO means novel object HOI detection. * means we only use the boxes of the detection results. Here, the baseline means we do not use affordance transfer learning.

Method	Type	Unseen	Seen	Full
Shen *et al.*[224]	UC	5.62	-	6.26
FG [10]	UC	10.93	12.60	12.26
VCL [111] (rare first)	UC	10.06	24.28	21.43
ATL (rare first)	UC	9.18	**24.67**	**21.57**
VCL [111] (non-rare first)	UC	16.22	18.52	18.06
ATL (non-rare first)	UC	**18.25**	**18.78**	**18.67**
FG [10]	NO	11.22	14.36	13.84
Baseline	NO	12.84	20.63	19.33
ATL (HICO-DET)	NO	11.35	20.96	19.36
ATL (COCO)	NO	**15.11**	**21.54**	**20.47**
Baseline*	NO	0.00	14.13	11.77
ATL (HICO-DET)*	NO	0.00	13.67	11.39
ATL (COCO)*	NO	**5.05**	**14.69**	**13.08**

samples but not unseen HOI samples with HICO-DET. Besides, similar to [10], we use a generic object detector to enable HOI detection with novel objects, which provides a strong baseline. While we only use the boxes of the detector (not use the object label predicted by detector), the performances of baseline and ATL (HICO-DET) on unseen categories decrease to 0. However, ATL (COCO) still achieves 5.05% on the unseen category.

2.4.5 Object Affordance Recognition

Following [107] that has discussed average precision (AP) is more robust for evaluating object affordance, we evaluate object affordance recognition with AP on HICO-DET. Table A.34 illustrates SCL largely improves SCL− (without self-training) by **over 9%** on Val2017, Object365, HICO-DET under the same training iterations. SCL requires more iterations to converge, and SCL greatly improves previous methods on all datasets with 3M iterations (Please refer to Appendix for convergence analysis). Noticeably, SCL directly predicts verb rather than HOI categories, and removes the spatial branch. Thus, SCL without self-training (SCL−) is a bit worse than ATL. Previous approaches ignore the unknown affordance recognition. We use the released models of [107] to evaluate the results on novel affordance recognition. Here, affordances of novel classes (annotated by hand [107]) are the same in the

TABLE 2.7. Comparison of object affordance recognition with HOI network (trained on HICO-DET) among different datasets. Val2017 is the validation 2017 of COCO [164]. Obj365 is the validation of Object365 [223] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. ATL$*$ means ATL optimized with COCO data. Unknown affordances indicate we evaluate with our annotated affordances. Previous approaches [111, 107] are usually trained by less 0.8M iterations (Please refer to the released checkpoint in [111, 107]). We thus also illustrate SCL under 0.8M iterations by default. SCL$-$ means SCL without self-training. Results are reported by Mean Average Precision (%).

Method	Known Affordances				Unknown Affordances			
	Val2017	Obj365	HICO	Novel	Val2017	Obj365	HICO	Novel
FCL [110]	25.11	25.21	37.32	6.80	-	-	-	-
VCL [111]	36.74	35.73	43.15	12.05	28.71	27.58	32.76	12.05
ATL [107]	52.01	50.94	59.44	15.64	36.80	34.38	42.00	15.64
ATL$*$ [107]	56.05	40.83	57.41	8.52	37.01	30.21	43.29	8.52
SCL$-$	50.51	43.52	57.29	14.46	44.21	41.37	48.68	14.46
SCL	**59.64**	**52.70**	**67.05**	14.90	**47.68**	**42.05**	**52.95**	14.90
SCL (3M iters)	**72.08**	**57.53**	**82.47**	**18.55**	**56.19**	**46.32**	**64.50**	**18.55**

TABLE 2.8. Comparison of object affordance recognition with HOI network among different datasets (based on Mean average Precision). Val2017 is the validation 2017 of COCO [164]. Subset of Object365 is the validation of Object365 [223] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. Object means what object dataset we use. ATL^{ZS} means novel object zero-shot HOI detection model in Table 3 on HICO-DET. For ATL^{ZS}, we show the results of the 12 classes of novel objects in Val2017, Subset of Object365 and HICO-DET.

Method	HOI Data	Object	Val2017 of COCO			Subset of Object365		
			Rec	Prec	F1	Rec	Prec	F1
Baseline	HOI	-	28.62	32.34	27.08	21.75	22.20	19.83
VCL [111]	HOI	HOI	76.93	71.79	72.15	68.60	67.52	65.82
ATL	HOI	HOI	80.71	72.79	74.44	71.76	67.34	67.13
ATL	HOI	COCO	**90.94**	**87.33**	**87.65**	**82.95**	**82.13**	**80.80**

two settings. We find SCL improves the performance considerably by **over 10%** on Val2017 and HICO-DET.

TABLE 2.9. Comparison of object affordance recognition with HOI network among different datasets (based on Mean average Precision). Val2017 is the validation 2017 of COCO [164]. Subset of Object365 is the validation of Object365 [223] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. Object means what object dataset we use. ATL^{ZS} means novel object zero-shot HOI detection model in Table 3 on HICO-DET. For ATL^{ZS}, we show the results of the 12 classes of novel objects in Val2017, Subset of Object365 and HICO-DET.

Method	HOI Data	Object	HICO-DET			Novel classes		
			Rec	Prec	F1	Rec	Prec	F1
Baseline	HOI	-	36.64	49.83	37.67	12.39	8.63	9.62
VCL [111]	HOI	HOI	87.98	82.59	83.84	54.75	35.85	40.43
ATL	HOI	HOI	90.29	83.21	85.30	**58.73**	37.75	42.75
ATL	HOI	COCO	**93.35**	**90.77**	**91.02**	53.65	**40.94**	**43.57**

2.4.6 HOI Concept Discovery

Baseline and Methods. We perform experiments to evaluate the effectiveness of our proposed method for HOI concept discovery. For a fair comparison, we build several baselines and methods as follows,

- **Random**: we randomly generate the concept confidence to evaluate the performance.
- **Affordance**: discover concepts via affordance prediction [107] as described in Sec 2.3.5.1.
- **GAT** [244]: build a graph attention network to mine the relationship among verbs during HOI detection, and discover concepts via affordance prediction.
- **Qpic*** [234]: convert verb and object predictions of [234] to concept confidence similar as online discovery.
- **Qpic* +SCL**: utilize concept confidence to update verb labels, and optimize the network (Self-Training). Here, we have no composite HOIs.

Please refer to the Appendix for more details, comparisons (e.g., re-training, language embedding), and qualitative discovered concepts with analysis.

Results Comparison. Table 2.10 shows affordance prediction is capable of HOI concept discovery since affordance transfer learning [107] also transfers affordances to novel objects.

TABLE 2.10. The performance of the proposed method for HOI concept discovery. We report all performance using the average precision (AP) (%). SCL means self-compositional learning. SCL− means online concept discovery without self-training. K indicates Known, while UK indicates UnKnown.

Method	HICO-DET		V-COCO	
	UK (%)	K (%)	UK (%)	K (%)
Random	12.52	6.56	12.53	13.54
Affordance [107]	24.38	57.92	20.91	95.71
GAT [244]	26.35	76.05	18.35	98.09
Qpic* [234]	27.53	87.68	15.03	13.21
SCL−	22.25	83.04	24.89	96.70
Qpic* [234] + SCL	28.44	88.91	15.48	13.34
SCL	**33.58**	**92.65**	**28.77**	**98.95**

Affordance prediction achieves 24.38% mAP on HICO-DET and 21.36% mAP on V-COCO, respectively, significantly better than the random baseline. With graph attention network, the performance is further improved a bit. Noticeably, [107] completely ignores the possibility of HOI concept discovery via affordance prediction. Due to the strong ability of verb and object prediction, Qpic achieves 27.42% on HICO-DET, better than affordance prediction. However, Qpic has poor performance on V-COCO. The inference process of affordance prediction for concept discovery is time-consuming (over 8 hours with one GPU). Thus we devise an efficient online concept discovery method which directly predicts all concept confidences. Specifically, the online concept discovery method (SCL−) achieves 22.25% mAP on HICO-DET, which is slightly worse than the result of affordance prediction. On V-COCO, the online concept discovery method improves the performance of concept discovery by **3.98%** compared to the affordance prediction. The main reason for the above observation might be due to that V-COCO is a small dataset and the HOI model can easily overfit known concepts on V-COCO. Particularly, SCL significantly improves the performance of HOI concept discovery from 22.36% to **33.58%** on HICO-DET and from 24.89% to **28.77%** on V-COCO, respectively. We find we can also utilize self-training to improve concept discovery on Qpic [234] (ResNet-50) though the improvement is limited, which might be because verbs and objects are entangled with Qpic. Lastly, we meanwhile find SCL largely improves concept discovery of known concepts on both HICO-DET and V-COCO.

2.4.7 HOI Detection with Unknown Concepts

HOI concept discovery enables zero-shot HOI detection with unknown concepts by first discovering unknown concepts and then performing HOI detection. The experimental results of HOI detection with unknown concepts are shown in Table 2.11. We follow [111] to evaluate HOI detection with 120 unknown concepts in two settings: rare first selection and non-rare first selection, i.e., we select 120 unknown concepts from head and tail classes respectively. Different from [111, 110] where the existence of unseen categories is known and the HOI samples for unseen categories are composed during optimization, HOI detection with unknown concepts does not know the existence of unseen categories. Therefore, we select top-K concepts according to the confidence score during inference to evaluate the performance of HOI detection with unknown concepts (that is also zero-shot) in the default mode [28].

As shown in Table 2.11, with more selected unknown concepts according to concept confidence, the proposed approach further improves the performance on unseen categories on both rare first and non-rare first settings. Specifically, it demonstrates a large difference between rare first unknown concepts HOI detection and non-rare first unknown concepts HOI detection in Table 2.11. Considering that the factors (verbs and objects) of rare-first unknown concepts are rare in the training set [110], the recall is very low and thus degrades the performance on unknown categories. However, with concept discovery, the results with top 120 concepts on unknown categories are improved by relatively **34.52%** (absolutely 0.58%) on rare first unknown concepts setting and by relatively **20.31%** (absolutely 1.19%) on non-rare first setting, respectively. with more concepts, the performance on unknown categories is also increasingly improved.

We also utilize the discovered concept confidences with SCL to evaluate HOI detection with unknown concepts on Qpic [234]. For a fair comparison, we use the same concept confidences to SCL. Without concept discovery, the performance of Qpic [234] degrades to 0 on Unseen categories though Qpic significantly improves zero-shot HOI detection. Lastly, we show zero-shot HOI detection (the unseen categories are known) in Table 2.11 (Those rows where

TABLE 2.11. Illustration of HOI detection with unknown concepts and zero-shot HOI detection with SCL. K is the number of selected unknown concepts. HOI detection results are reported by mean average precision (mAP)(%). We also report the recall rate of the unseen categories in the top-K novel concepts. "K = all" indicates the results of selecting all concepts, i.e., common zero-shot. $*$ means we train Qpic [234](ResNet-50) with the released code in zero-shot setting and use the discovered concepts of SCL to evaluate HOI detection with unknown concepts. Rec indicates Recall. U indicates Unknown/Unseen. S indicates Known/Seen.

Method	K	Rare First				Non-rare First			
		U	S	Full	Rec (%)	U	S	Full	Rec (%)
SCL	0	1.68	22.72	18.52	0.00	5.86	16.70	14.53	0.00
SCL	120	2.26	22.72	18.71	10.83	7.05	16.70	14.77	21.67
SCL	240	3.66	22.72	18.91	15.00	7.17	16.70	14.80	25.00
SCL	360	4.09	22.72	19.00	15.83	7.91	16.70	14.94	30.83
SCL	all	9.64	22.72	19.78	100.00	13.30	16.70	16.02	100.00
Qpic$*$ [234]	0	0.0	**30.47**	24.37	0.00	0.0	23.73	18.98	0.0
Qpic$*$ [234]	120	2.32	30.47	24.84	10.83	14.90	22.19	20.58	21.67
Qpic$*$ [234]	240	3.35	30.47	25.04	15.00	14.90	22.79	21.22	25.00
Qpic$*$ [234]	360	3.72	30.47	25.12	15.83	14.91	23.13	21.48	30.83
Qpic$*$ [234]	all	15.24	30.44	27.40	100.00	21.03	23.73	23.19	100.00
ATL [107]	all	9.18	24.67	21.57	100.00	18.25	18.78	18.67	100.00
FCL [110]	all	13.16	24.23	22.01	100.00	18.66	19.55	19.37	100.00
Qpic + SCL	all	**19.07**	30.39	**28.08**	100.00	**21.73**	**25.00**	**24.34**	100.00

TABLE 2.12. The branches ablation study of the model on HICO-DET test set. Verb-object branch only means we train the model without spatial-human branch.

Method	Full	Rare	NonRare
Two branches	19.43	16.55	20.29
Verb-Object branch only	15.77	13.35	16.49
Verb-Object branch only (w/o VCL)	15.33	10.85	16.67

K is all). We find that SCL significantly improves Qpic, and *forms a new state-of-the-art* on zero-shot setting though we merely use ResNet-50 as backbone in Qpic. We consider SCL improves the detection of rare classes (including unseen categories in rare first and seen categories in non-rare first) via stating the distribution of verb and object. See Appendix for more analysis, e.g., SCL improves Qpic particularly for rare categories on Full HICO-DET.

TABLE 2.13. Composing Strategies study of VCL on HICO-DET test set Mean average precision (mAP) (%) are reported.

Method	Full	Rare	NonRare
Baseline (w/o VCL)	18.43	14.14	19.71
Within images	18.48	14.46	19.69
Between images	19.06	14.33	20.47
Between and within images	19.43	16.55	20.29

2.4.8 Ablation Studies

Branches. There are two branches in our method and we evaluate their contributions in Table 2.12. Noticeably, we apply VCL to Verb-Object branch during training, while we do not apply VCL to Spatial-Human. By keeping one branch each time on HICO-DET dataset during inference, we can find the verb-object branch makes the larger contribution, particularly for rare category (**3%**). This efficiently illustrates the advantage of VCL for rare categories. But we can improve the performance dramatically from 16.89% to 19.43% with Spatial-Human Branch. Meanwhile, we can find the proposed VCL is orthogonal to spatial-human branch from the last two rows in Table 2.12. Noticeably, *by comparing verb-object branch only during inference and verb-object branch only from training, we can find the spatial-human branch can facilitate the optimization of verb-object branch (improving the mAP from 15.77% to 16.89%).*

Composing interactions within and/or between images. In Table 2.13, we can find composing interaction samples between images is beneficial for HOI detection, whose performance in the Full category increases to 19.06% mAP, while composing interaction samples within images has similar results to baseline. It might be because the number of images including multiple interactions is few on HICO-DET dataset. Remarkably, composing interaction samples within and between images notably improves the performance up to 19.43% mAP in Full and **16.55%** mAP in Rare respectively. Those results mean composing interactions within images and between images is more beneficial for HOI detection.

Verb Feature Regularization. We use a simple auxiliary verb loss to regularize verb features. Although verb regularization loss can slightly improve the rare and unseen category

TABLE 2.14. Comparison between step-wise optimization and one step optimization. ZS is the setting in our ablation study.

Method	Full	Rare	NonRare	Unseen
one step (long-tailed)	24.03	18.42	25.70	-
step-wise (long-tailed)	**24.68**	**20.03**	**26.07**	-
one step (ZS)	**19.69**	18.22	20.82	**17.64**
step-wise (ZS)	19.61	**18.69**	**21.13**	15.86
one step (rare first ZS)	22.01	15.55	24.56	**13.16**
step-wise (rare first ZS)	**22.45**	**17.19**	**25.34**	12.12
one step (non-rare ZS)	**19.37**	15.39	20.56	**18.66**
step-wise (non-rare ZS)	19.11	**17.12**	**21.02**	15.97

TABLE 2.15. Illustration of the effect of fine-tuned detectors on FCL. The COCO detector is trained on COCO dataset provided in [269]. We fine-tune the ResNet-101 Faster R-CNN detector based on Detectron2 [269]. Here, the baseline is our model without fabricator. The last column is object detection result on HICO-DET test.

Method	Detector	Full	Rare	NonRare	Object mAP
Baseline	COCO	21.24	17.44	22.37	20.82
FCL	COCO	**21.80**	**18.73**	**22.71**	20.82
Baseline	HICO-DET	23.94	17.48	25.87	30.79
FCL	HICO-DET	**24.68**	**20.03**	**26.07**	30.79
Baseline	GT	43.63	34.23	46.43	100.00
FCL	GT	**44.26**	**35.46**	**46.88**	100.00

performance (See row 1 and row 3 in Table 2.16), FCL further achieves better performance. This indicates that regularizing factor features is suboptimal compared to the proposed method. Semantic verb regularization like [277] has a similar result.

Verb and Noise for Fabricator. Table 2.17 demonstrates that performance drops without verb representation or noise. This shows verb representations can provide useful information for generating objects and noise efficiently improves the performance by increasing feature diversity. We meanwhile find the fabricator still effectively improves the baseline without verb in Table 2.17, which indicates the efficiency of FCL.

Verb Fabricator. The result of fabricating verb features (from verb identity embedding, object features and noise) is even worse as in Table 2.17. This verifies that it is difficult to directly generate useful verb or HOI samples due to the complexity and abstraction.

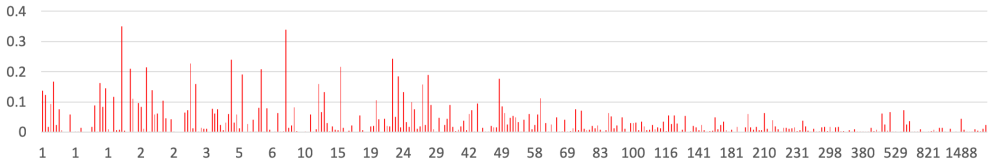

FIGURE 2.10. Illustration of the improvement in those improved categories between FCL and baseline on HICO-DET dataset under default setting. The graph is sorted by the frequency of category samples and the horizontal axis is the number of training samples for each category. The result is reported in mAP (%).

TABLE 2.16. Illustration of proposed modules under step-wise optimization. FCL means proposed Fabricated Compositional Learning. V indicates the verb regularization loss.

FCL	V	Full	Rare	NonRare	Unseen
-	-	18.12	15.99	20.65	12.41
✓	-	19.08	17.47	20.95	14.90
-	✓	18.32	16.73	20.82	12.23
✓	✓	**19.61**	**18.69**	**21.13**	**15.86**

TABLE 2.17. Ablation study of fabricator under step-wise optimization. FCL within image means we compose HOIs within image. + verb fabricator is we fabricate verb and object features.

Method	Full	Rare	NonRare	Unseen
FCL	**19.61**	**18.69**	21.13	**15.86**
FCL w/o noise	19.45	17.69	21.22	15.74
FCL w/o verb	19.20	18.02	21.04	14.71
FCL + verb fabricator	19.47	16.93	21.43	15.89

Step-wise Optimization. Table 2.14 illustrates that step-wise training has better performance in rare and non-rare categories while has worse performance in unseen categories. We think it might be because the model with the step-wise training has the bias to seen categories in the first step since there are no training data for unseen categories.

Object Detector. The quality of detected objects has important effect on two-stage HOI Detection methods [111]. Table 2.15 shows that the improvement of FCL over baseline is higher with the fine-tuned detector on HOI data. COCO detector without finetuning on HICO-DET contains a large number of false positive and false negative boxes on HICO-DET due to domain shift, which is in fact less useful to evaluate the effectiveness of modeling human interactions for HOI detection. If the detected boxes during inference are false, the

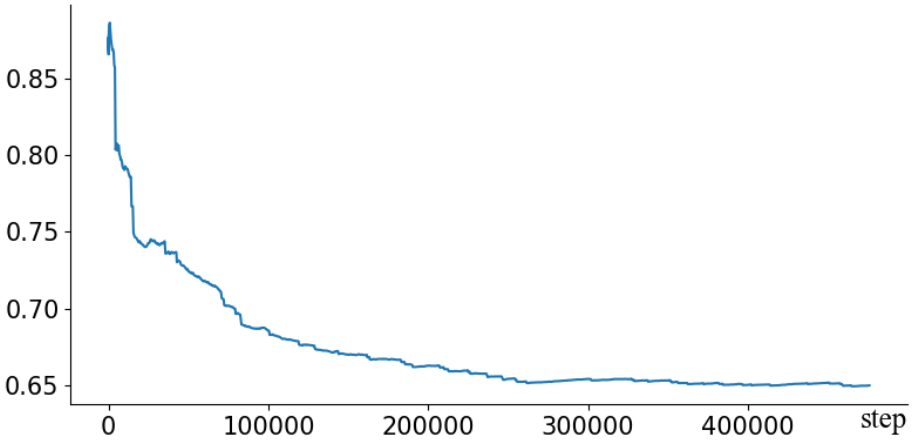

FIGURE 2.11. The changing trend of cosine similarity between fabricated object features and real object features during optimization in long-tailed HOI detection in step-wise training.

TABLE 2.18. Illustration of the number of object images in each batch on HICO-DET dataset.

#Images	Full	Rare	NonRare
1	24.07	18.17	25.83
2	**24.50**	**18.53**	**26.28**
3	24.19	17.33	26.24

features extracted from the false boxes are also unreal and have large shift to the fabricated objects during training. This causes that fabricated objects are less useful for inferring HOIs during inference. Besides, GT boxes provide a strong object label prior for verb recognition.

The number of object images in each batch. Table 2.18 shows ATL achieves best performance with 2 object images. We think more object images increase the diversity of object features and balance the object distribution. However, too many object images also hamper the performance.

Object detector. Due to the domain shift between HICO-DET and COCO, COCO detector usually achieves worse result. we thus use the same fine-tuned object detector as [111]. Table 2.19 illustrates better detected object boxes improve the performance largely. Meanwhile, we find ATL is apparently sensitive to worse boxes. Under a worse object detector (i.e.COCO detector), ATL does not improve the result. It might be because composing

TABLE 2.19. Illustration of the effect of different object detectors on HOI detection in HICO-DET. The fine-tuned detector is provided in [111]. GT means ground truth boxes. The last column is the detection mAP on HICO-DET test dataset.

Model	Detector	Full	Rare	NonRare	mAP
Baseline	COCO	**21.07**	**16.79**	**22.35**	20.82
ATL	COCO	20.08	15.57	21.43	20.82
Baseline	Fine-tuned	23.44	16.80	25.43	30.79
ATL	Fine-tuned	**24.50**	**18.53**	**26.28**	30.79
Baseline	GT	43.32	33.84	46.15	100
ATL	GT	**44.27**	**35.52**	**46.89**	100

TABLE 2.20. Illustration of the effect of domain shift on ATL between object images and HOI images on HOI-COCO dataset. Sub-COCO is a subset of COCO images that we randomly choose the same number of object instances to the objects of HICO-DET from COCO dataset.

Method	Object images	Full	Rare	NonRare
ATL	HICO-DET	24.21	9.52	35.61
ATL	Sub-COCO	**24.74**	**9.60**	**36.50**

affordance features and object features from additional images results in poor generalization to worse boxes. When we transfer affordance representation to objects from a large number of additional images via composing novel HOI samples, we improve the scene generalization (i.e.the model generalizes to novel scenes) of the affordance representation learning, while degrading the generalization to worse object boxes on HICO-DET test set. The object affordance recognition in Table 2.8 illustrates the scene generalization of affordance and object representations. Noticeably, a worse object detector largely hampers HOI detection in two-stage method. Thus, it is necessary to utilize a better object detector for evaluating HOI detection, and ATL further improves HOI detection effectively with a better object detector.

Domain difference. From the large performance gap between different object detectors in Table 2.19, we find the HICO-DET dataset has a different domain to COCO. Table 2.20 shows with the same number of object instances, COCO dataset improves the performance larger than HICO-DET dataset due to the domain difference on HOI-COCO. There is a similar trend in Table 2.3 and Table 2.4. With the same COCO dataset, our method facilitates HOI detection on HOI-COCO dataset better than that on HICO-DET.

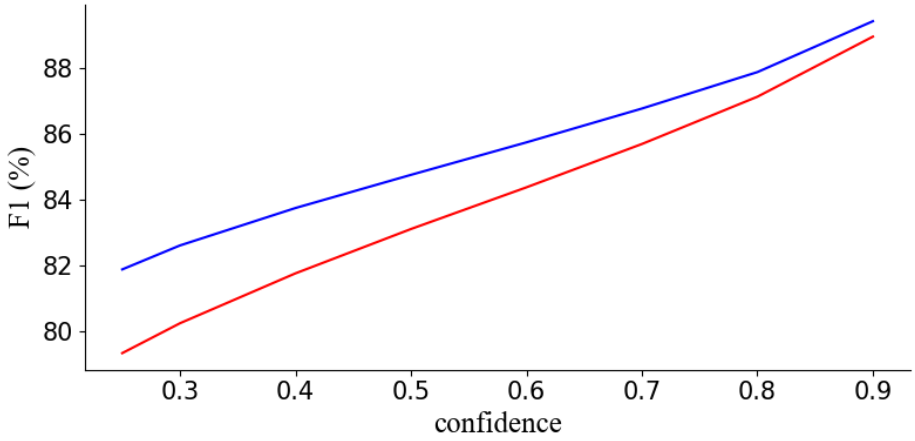

FIGURE 2.12. Comparison of object affordance recognition (F1) between ATL and the conversion from object detection results on HICO-DET. Confidence is the object detection confidence for choosing object boxes. Red is our method and Blue is the conversion from object detection results.

TABLE 2.21. The effect of different number of verbs in affordance feature bank. Mean average Precision (mAP) (%) is reported. Dataset means the evaluation object dataset. HICO-DET means the test set of HICO-DET. Val2017 means the validation set of COCO2017.

#M	Dataset	1	5	10	20	40	80	100
Baseline	Val2017	13.39	15.90	17.69	18.74	19.25	19.67	**19.71**
ATL (COCO)	Val2017	52.98	53.74	55.40	55.19	54.88	55.77	**56.05**
Baseline	HICO-DET	14.77	18.30	20.22	21.70	22.21	23.00	**23.18**
ATL (COCO)	HICO-DET	56.04	58.03	**59.14**	57.84	56.61	57.23	57.41

Affordance comparison with object detection results. Our method can also be applied to detected boxes of an object detector. For a robust comparison, we directly compare ATL with the object affordance result converted from object detection results according to the object affordance annotation (i.e.the ground truth affordances of an object category) on HICO-DET test set. Here we use the detected box of a COCO pretrained Faster-RCNN. We train our model on HOI-COCO dataset and COCO (2014) dataset, which has the same training set as COCO pretrained Faster-RCNN. Figure 2.12 illustrates ATL achieves better affordance recognition results among different confidences. Meanwhile, ATL has better performance than object affordance detection when the confidence of the detected box is lower.

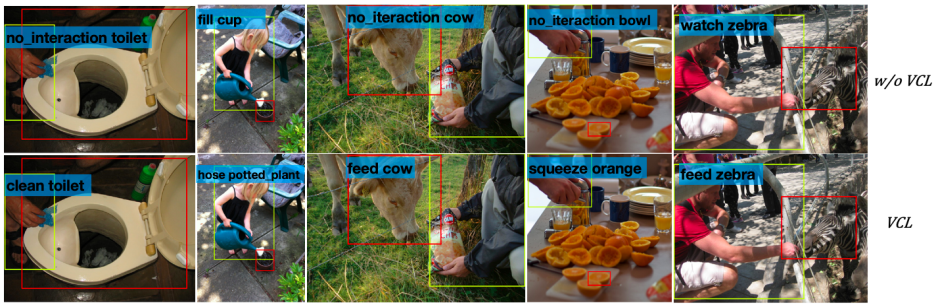

FIGURE 2.13. Some rare HOI detections (Top 1 result) detected by the proposed Compositional Learning and the model without Compositional Learning. The first row is the results of the baseline model without VCL. The second row is the results of VCL

The effect of the number of verbs on affordance recognition. In affordance recognition, we randomly choose M instances for those affordances with more than M instances in dataset and all instances for other affordances. We ablate M in Table 2.21 under the ATL model with COCO objects and our baseline. The baseline is the model without compositional learning. Besides, when we use different M, we also update S_i. If we keep S_i the same as the number when $M = 100$, all results will be very small when $M < 100$.

Table 2.21 shows the number goes stable after 20. This means we do not need to store a large number of templates of affordance representation.

2.5 Qualitative Analysis

Figure 2.13, we qualitatively show that our proposed Visual Compositional Learning framework can detect those rare interactions correctly while the baseline model without VCL misclassifies on HICO-DET. The results demonstrate that our proposed Visual Compositional Learning framework is significantly beneficial for rare categories.

2.5.1 Visualization

Figure 2.14 illustrates the Grad-CAM under different methods. We find the proposed SCL focus on the details of objects and small objects, while the baseline and VCL mainly highlight

the region of human and the interaction region, e.g., SCL highlights the details of the motorbike, particularly the front-wheel (last row). Besides, SCL also helps the model via emphasizing the learning of small objects (e.g., frisbee and bottle in the last two columns), while previous works ignore the small objects. This demonstrates SCL facilitates affordance recognition and HOI concept discovery via exploring more details of objects.

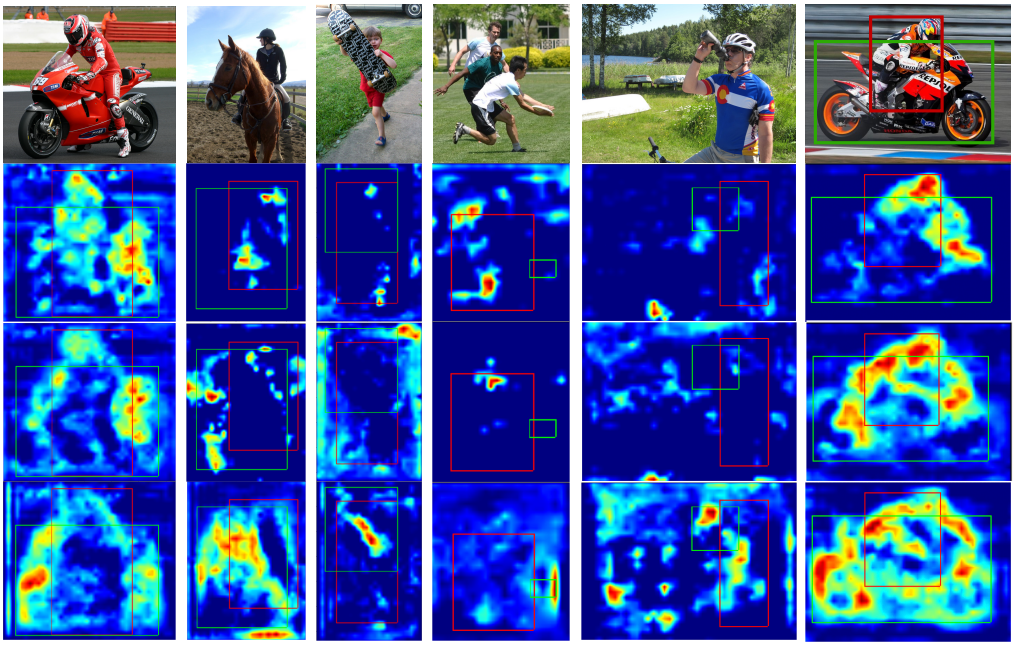

FIGURE 2.14. A visual comparison of recent methods using the Grad-CAM [221] tool. The first row is input image, the second row is baseline without compositional approach, the third row is vanilla VCL [111] and the last row is the proposed SCL. Here, we compare all models using the same dataset.

We also illustrate the verb and object features by t-SNE visualization [177]. Figure 2.15 illustrates that VCL overall improves the the discrimination of verb and object features. There are many noisy points (see black circle region) in Figure 2.15 without VCL and verb presentation. Meanwhile, we can find the proposed verb representation learning is helpful for verb feature learning by comparing verb t-SNE graph between the left and the middle. Besides, the object features are more discriminative than verb. We think it is because the verb feature is more abstract and complex and the verb representation requires further exploration.

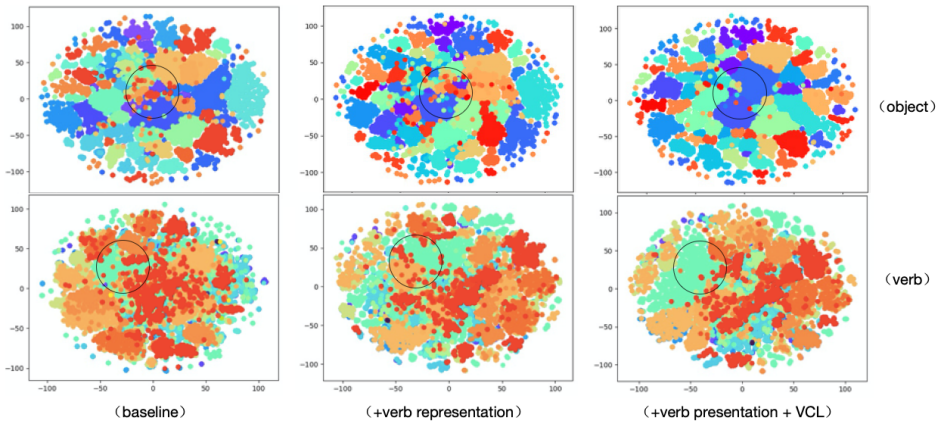

FIGURE 2.15. Visual illustration of object features (80 classes) (up) and verb features (117 classes) (bottom) on HICO-DET dataset (20000 samples) via t-SNE visualization [177]. Left is the visual illustration of baseline, the middle includes verb representation and the right uses both VCL and verb representation

Illustration of improvement among categories. In Figure 2.10, we find that *the rarer the category is, the more the proposed method can improve*. The result illustrates the benefit of FCL for long-tailed issues in HOI Detection.

Visualized Analysis between fabricated and real object features. Figure 2.11 presents that cosine similarity between fabricated and real object features gradually goes down to stability in step-wise training. This demonstrates that end-to-end optimization with a shared HOI classifier helps fabricate efficient and similar objects during the optimization process.

2.5.2 Fabricated Object Representations

We analyze the real object features and fabricated object features in detail in Figure 2.16, 2.17 by selecting the top 10 frequent classes in HICO-DET. 1) In Figure 2.16 (a) and Figure 2.17 (a), we find the fake object features of the same class are close to each other, while the features from different classes are separable although they might share the same verb. 2) Figure 2.16 (b) and Figure 2.17 (c) show features of different verbs slightly clustered together within each object class. **We can find there are outliers in some object classes because those outliers have different verbs**. 3) for unseen object ZSL, Figure 2.17 shows all fake object features of

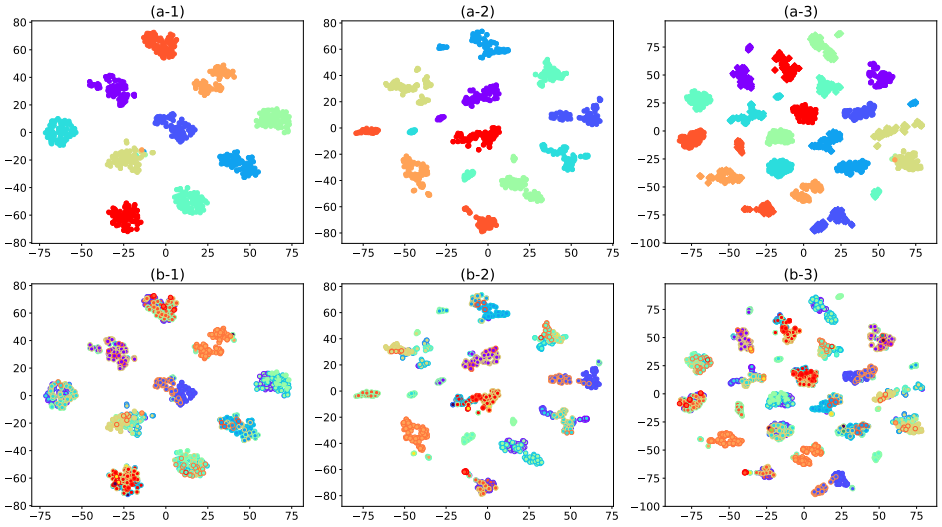

FIGURE 2.16. The illustration of real object representations, fabricated object representations and joint representations extracted from long-tailed HOI detection model. We select the top 10 frequent object classes from HICO-DET training data. For each class, we randomly select 100 instances. Column 1 is real object representations, Column 2 is fabricated object representations and Column 3 is the joint representations. In Column 3, a diamond point means fabricated object representations. Raw a is the base t-SNE figure. In raw b, we label different verbs with different edges (color) in Raw b.

the same class are also closer to each other. Particularly, the unseen objects (red edge in row b) are also separable from others. 4) The Column 3 in Figure 2.16 and Figure 2.17 illustrate fake object features are still separable from its real objects of the same class. However, there are still some fabricated features that are closer to it's corresponding real features (e.g.the dark blue class in Figure 2.16 and the jade-green class in Figure 2.17). We think Column 3 in the two Figures also shows a future direction for fabricating objects, i.e.generate more realistic objects.

e **The Effect of Objects on HOI Detection** In the nature, different types of objects form a long-tail distribution. Then, all those actions that people perform on those objects are inevitably long-tailed. As a result, those HOIs that we observed are long-tailed. This motivates us to fabricate balanced objects for composing HOI samples with visual verbs. We have demonstrated the long-tailed distribution of objects in Figure 2 in the paper and the effect of different object detector on HOI detection in Table 7 in paper. We further illustrate HOI detection has roughly similar performance to object detection among most object categories

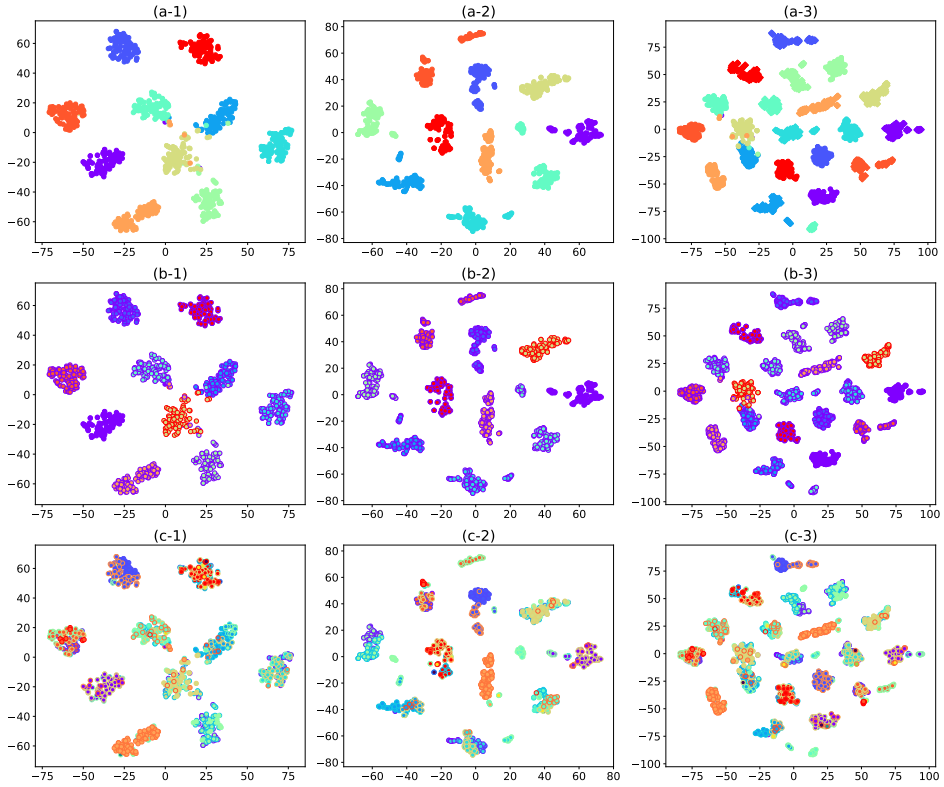

FIGURE 2.17. The illustration of real object representations, fabricated object representations and joint representations extracted from unseen object zero-shot model. Column 1 is real object representations, Column 2 is fabricated object representations and Column 3 is the joint representations. In Column 3, a diamond point means fabricated object representations. Raw a is the base t-SNE figure. In raw b, we point out the unseen objects with red edges. In Raw c, we label different verbs with different edges (color).

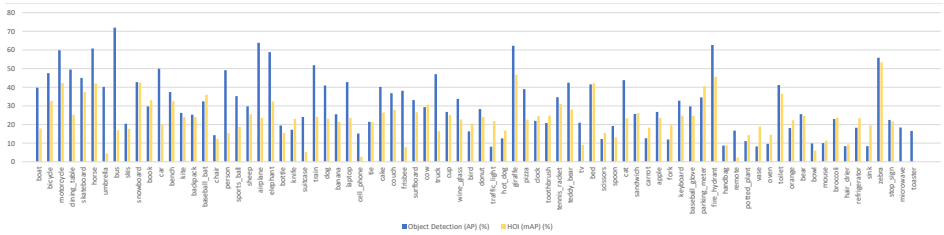

FIGURE 2.18. Illustration of Object detection result and HOI detection result in HICO-DET dataset. Blue is Object result. Yellow is HOI result. We average HOI detection AP according to the object categories for a direct comparison.

in Figure A.3, which also illustrates the importance of object detectors for HOI detection at the same time. Meanwhile, it is necessary to balance the distribution of objects.

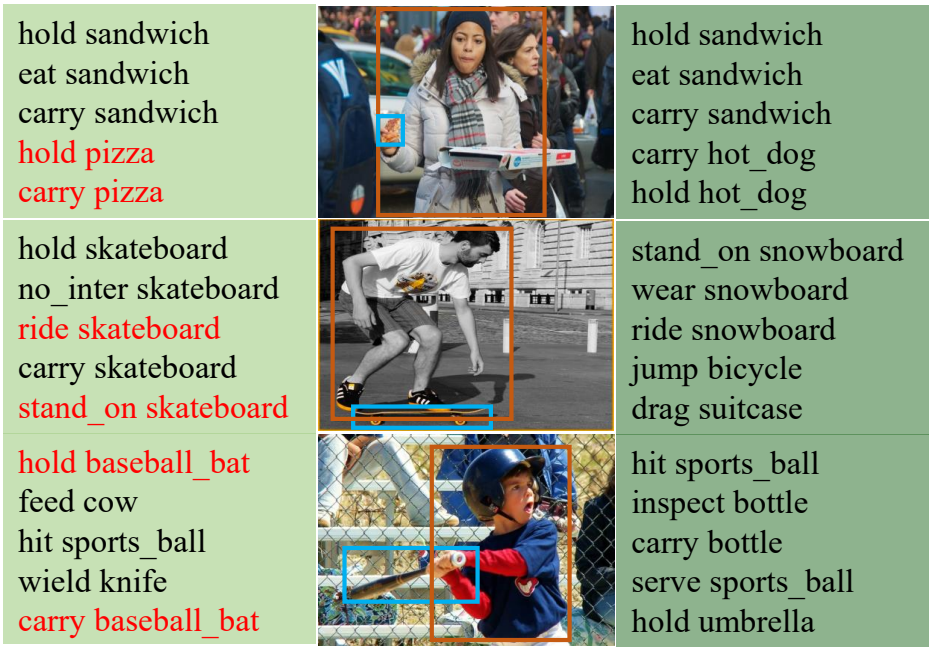

FIGURE 2.19. Illustration of unseen object zero-shot detection result (top 5) between the proposed method and Baseline. The correct results are highlighted in red.

Visualized Object Affordances

We demonstrate the result of exploring unseen HOIs with novel objects in Figure 2.19. We find the baseline can not recognize the object at all, while the proposed method effectively detects the HOI with unseen objects.

2.6 Discussions

Nowadays, the deep learning community has made significant progress on language models, particularly large language models or LLM [192]. The large language model is able to provide prior knowledge about the visual world, including the object affordances, action categories, and their correlations. Current approaches have made use of large language models or visual-language models [209, 150] to facilitate human-object interaction detection, especially open-vocabulary HOI detection. While the large language models have included prior knowledge about object categories and verb categories, it is still challenging to understand how to interact

with a novel object. Compositional generation is also important when incorporating language models for visual understanding.

2.7 Summary

In this chapter, we first introduce two HOI-relevant tasks, i.e., compositional object affordance recognition and Human-Object Interaction Concept Discovery, to enrich the exploration of Human-Object Interaction. Compositional object affordance reasoning aims to recognize the affordances of an object instance from HOI model, while HOI concept discovery requires discovering all reasonable combinations (i.e., HOI concepts) between verbs and objects according to a few training samples of known HOI concepts/categories. Next, we propose a simple yet efficient deep Visual Compositional Learning framework, which composes the interactions from the shared verb and object latent representations between images and within images, to address HOI compositional generalization, compositional object affordance reasoning, and HOI concept discovery. The compositional approach also implies transferring of the verb/affordance representation to object representations, and thus enables compositional object affordance reasoning. We thus devise a simple yet effective compositional method to incorporate HOI detection model for object affordance reasoning. Moreover, we introduce self-compositional learning to facilitate the compositional approach, which maintains an online updated concept confidence matrix, and assigns pseudo labels according to the matrix for all composite HOI features, and thus optimize both known and unknown composite HOI features via self-training. Extensive experiments demonstrate the proposed methods improve HOI compositional generalization, facilitate compositional object affordance reasoning, and enable HOI detection with unknown concepts on multiple datasets.

CHAPTER 3

Sample Relationship Exploration

In the last Chapter, the thesis comprehensively investigated the existing relationships (e.g.. Human-Object Interactions) in the visual scenes by transferring verb/object representations among different HOIs. However, there are also visual relationships, e.g., the similarity, among different samples, which we named as sample relationship. Despite the great success achieved, deep learning technologies usually suffer from data scarcity issues in real-world applications, where existing methods mainly explore sample relationships in a vanilla way from the perspectives of either the input or the loss function. In this chapter, we propose a batch transformer module, BatchFormerV1, to equip deep neural networks themselves with the abilities to explore sample relationships in a learnable way. Basically, the proposed method enables data collaboration, e.g., head-class samples will also contribute to the learning of tail classes, and provides an unified way to transfer the representations among different samples to facilitate the challenging sample recognition (e.g. tail classes). Considering that exploring instance-level relationships has very limited impacts on dense prediction, we generalize and refer to the proposed module as BatchFormerV2, which further enables exploring sample relationships for pixel-/patch-level dense representations. In addition, to address the train-test inconsistency where a mini-batch of data samples are *neither necessary nor desirable* during inference, we also devise a two-stream training pipeline, i.e., a shared model is first jointly optimized with and without BatchFormerV2 which is then removed during testing. The proposed module is plug-and-play without requiring any extra inference cost. Lastly, we evaluate the proposed method on over ten popular datasets, including 1) different data scarcity settings such as long-tailed recognition, zero-shot learning, domain generalization, and contrastive learning; and 2) different visual recognition tasks ranging from image classification to object detection and panoptic segmentation.

3.1 Motivations and Contributions

The recent success of deep learning heavily relies on collecting and annotating large-scale training data [95, 97], while data scarcity issues have been repeatedly found in real-world applications. Among the methods for handling data scarcity issues, sample relationship has received much attention from the community, especially for the tasks without proper training data distributions to guarantee good generalization performances, such as long-tailed recognition [265], zero-shot learning [185], and domain generalization [25]. An intuitive example revealing the effectiveness of sample relationships is shown in Figure 3.1.

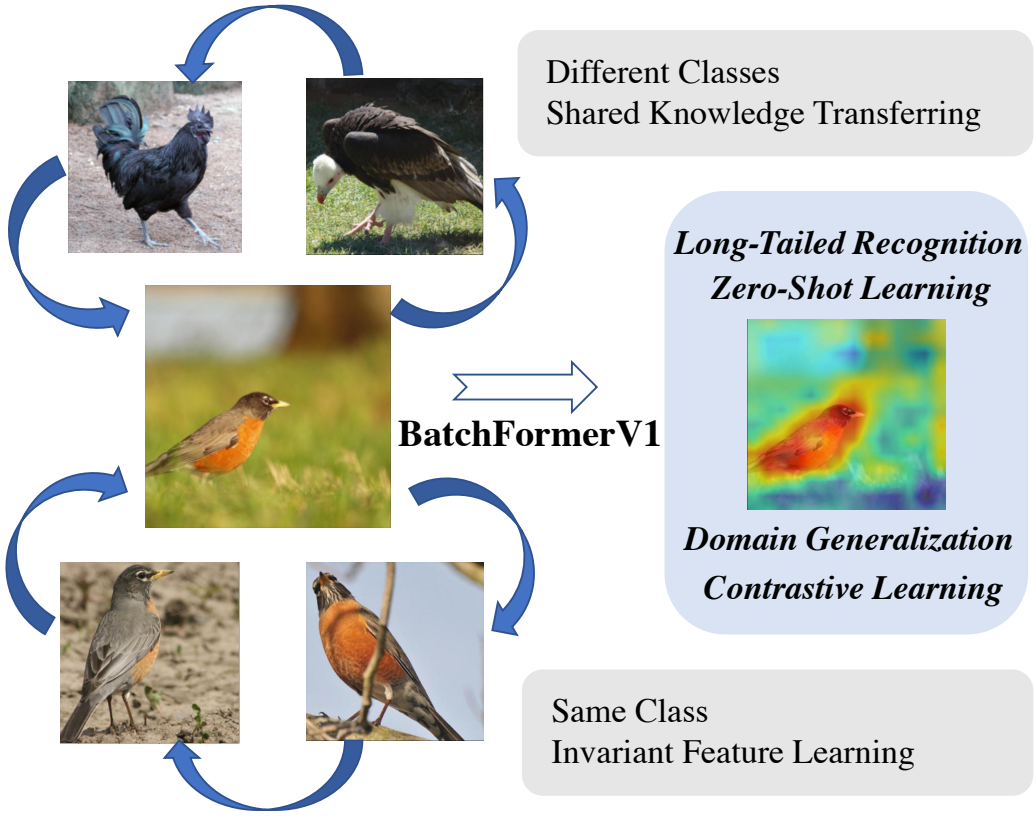

FIGURE 3.1. An illustration of sample relationships. Specifically, similar classes tend to share some parts (e.g., cock, robin, vulture share body shape, and claw shape), and transferring the shared knowledge from head to tail classes thus facilitates learning with long-tailed distributions.

Sample relationships have been intensively explored using an explicit scheme [222, 293, 152, 71, 263, 250, 313]. Specifically, a simple yet very effective way is to generate new samples

from existing training data [152], such as mixup [293], copy-paste [71], crossgrad [222], and compositional learning [128, 111, 5, 185]. Another way is to transfer knowledge between different samples, e.g., 1) transferring the meta knowledge between head and tail classes for long-tailed recognition [263, 171]; 2) transferring the knowledge from seen to unseen classes for zero-shot learning [272, 185]; and 3) transferring the invariant knowledge for domain generalization [201, 4, 116]. However, all above-mentioned methods explore sample relationships from either the input or the loss function, failing to enable deep neural networks themselves with the abilities to automatically explore sample relationships, i.e., no interaction appears in the batch dimension during representation learning.

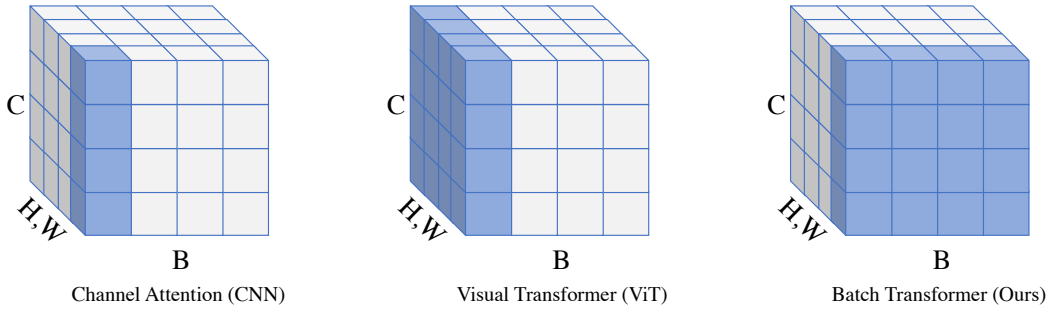

FIGURE 3.2. An illustration of the attention mechanisms on channel, spatial, and batch dimensions.

The transformer architecture has been very widely used to explore spatial-temporal relationships [242, 53, 170, 286, 43, 32], while it is non-trivial to apply attention mechanism in the batch dimension due to the train-test inconsistency [118, 171, 313]. For example, a mini-batch of data samples is neither necessary nor desirable during inference, while it always requires to track training statistics as the running mean and variance when applying the batch normalization [118]. Another example keeps all category centers during training, which is then used to enhance the tail/unseen categories during testing [171, 313]. To empower deep neural networks with structural advances for sample relationship learning, we introduce the attention mechanism to the batch dimension and address the train-test inconsistency as follows. Specifically, we capture the sample relationships within each mini-batch of training data by applying a transformer encoder into the batch dimension, and refer to it as Batch Transformer or BatchFormerV1. An intuitive example showing the differences between the channel, spatial, and batch attentions is illustrated in Figure 3.2. To address the train-test

inconsistency, we then utilize a shared classifier before and after the proposed BatchFormerV1 module to achieve the batch-invariant learning, i.e., learn representations invariant to with and without mini-batch. By doing this, BatchFormerV1 is only required during training, and we do not need to change the inference structure of the original model.

Inspired by BatchFormerV1 for instance-level visual recognition, we further propose to enable exploring the sample relationships for pixel-/patch-level dense representation learning. That is, exploring sample relationships with only instance-level representations that are usually from the last fully connected layer has very limited influences on dense representation learning. Therefore, we generalize BatchFormerV1 such that it can be used in different intermediate layers to explore semantics at different scales, ranging from the pixel to the patch and the whole image. Nevertheless, it is difficult to setup a shared classifier before and after each layer for the batch-invariant learning. We thus introduce a generalized module, BatchFormerV2, to enable the information propagation between different data samples at multiple semantic levels by learning the batch-invariant representations. By doing this, the proposed BatchFormerV2 module can be applied in different intermediate layers of popular vision transformers. The train-test inconsistency is addressed by a two-stream training pipeline, i.e., one stream with BatchFormerV2 and the other stream without, while all other layers are shared by both streams. All shared layers will be jointly optimized to generalize on the inputs with and without BatchFormerV2. During testing, we can directly remove all BatchFormerV2 modules without sacrificing model performance.

During training, the proposed method enables the information propagation of all samples in each mini-batch. Therefore, all samples can contribute to the learning on each categories. Intuitively, it implicitly enriches current training samples with hallucinated features from the whole mini-batch (e.g., the shared parts between two categories). For example, in long-tailed recognition, such hallucinated features can also improve the feature space of the tail classes. Meanwhile, the loss function also emphasizes on the rare classes via propagating larger gradients of rare classes in each mini-batch. In addition, we also find two obvious changes on the learned representations with the proposed module, i.e., it effectively facilitates the model to learn 1) comprehensive representations by focusing on almost all different object parts; and

2) invariant representations by focusing on the object itself rather than complex backgrounds (See more results in Section 3.3.6, Section A4.3, and Appendix).

A preliminary version of BatchFormerV1 has been published in [106] on the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR2022). The description of BatchFormerV2 also appears in the preprint [109], which is *not* submitted to any other conference or journal. Our main contributions can be summarized as follows:

- We propose to learn sample relationships from the perspective of improving the internal structure of deep neural networks during training.
- We devise a simple yet effective module termed as BatchFormerV1 to explore sample relationships for different data scarcity tasks.
- We introduce a generalized module, BatchFormerV2, to enable sample relationship learning for dense prediction.
- We develop a two-stream training strategy to address the train-test inconsistency, such that the proposed module can improve model performance without any extra inference cost.
- We perform extensive experiments in different settings and tasks, which show the effectiveness of the proposed method, including long-tailed recognition, zero-shot learning, domain generalization, self-supervised representation learning, image classification, object detection, and panoptic segmentation.

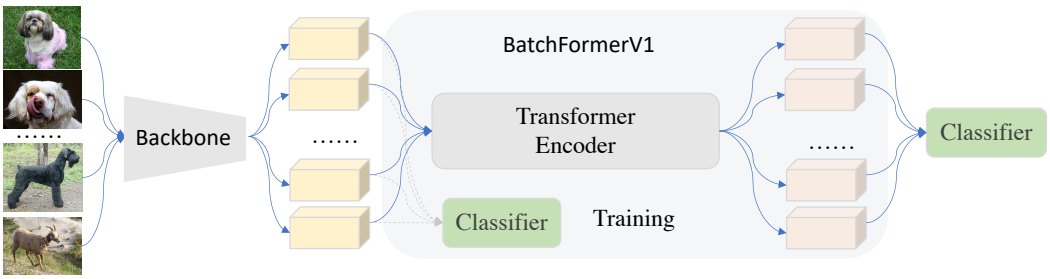

FIGURE 3.3. The main deep representation learning framework with BatchFormerV1. Specifically, we apply BatchFormerV1 between the backbone network (e.g., ResNet) and the classifier to explore sample relationships. With a shared classifier for training, we can remove BatchFormerV1 during testing.

3.2 Related Work

3.2.1 Sample Relationship

There are rich relationships among different samples, which have been widely used via various strategies [293, 168, 105, 184]. Specifically, Zhang *et al*. [293] propose to regularize the model to favor simple linear behavior in-between training samples with mixup. However, mixup [293] merely considers a linear transformation between data samples, while we aim to investigate the non-linear relationship among samples in a more powerful way. The compositionality of samples has also inspired many approaches to improve few/zero-shot generalization [237, 94, 111, 185], where the parts/attributes shared among different samples have been explored via the prior knowledge on label relationships. Several approaches also use sample/class relationships to conduct transductive inference [168, 105, 171, 184], e.g., transductive few-shot classification [168], meta embedding [171, 313], and non-parametric transformer [139]. However, those approaches usually require to inference with multiple samples (e.g., query set, or bank features). Meanwhile, many recent domain generalization methods [201, 4, 116] aim to find casual/invariant representations across domains, which we think them internally utilize the relationship among samples of the same class but different domains.

3.2.2 Data Scarcity Learning

. Learning with limited and imperfect data has turned out to be very challenging in a variety of data scarcity tasks. For example, in many real-world applications, the data from different classes usually follow a long-tailed distribution where a large portion of classes have very few instances. Current long-tailed learning approaches can be roughly categorized into 1) distribution re-balancing (e.g., re-sampling [31, 83, 93], re-weighting [20, 45, 236, 306, 121]); 2) ensemble of diverse experts [308, 265, 21]; and 3) knowledge transfer [263, 171, 313, 99, 250]. Other data scarcity tasks include zero-shot learning and domain generalization [18, 310, 252]. Specifically, zero-shot learning aims to recognize unseen classes without training

samples, while domain generalization aims at generalizing from seen to unseen domains. Current zero-shot learning approaches [146, 194, 272, 260] usually transfer knowledge of seen classes to unseen classes via modern techniques (e.g., graph network [185], data generalization [315], and compositional learning [185, 111]). Recently, compositional zero-shot learning has been widely explored in different tasks [128, 111, 5, 185], and we thus mainly evaluate the proposed module on compositional zero-shot learning [185]. In addition, recent domain generalization methods usually include data augmentation [311, 310], meta learning [8, 300], and disentangled/invariant representation learning [201, 4, 30, 206]. The proposed method facilitates robust representation learning without any assumptions on the dataset distributions and can thus be used for different tasks and datasets.

3.2.3 Vision Transformer

Transformer was first introduced by Vaswani *et al.*[242] for machine translation. As a core part, attention mechanism aggregates information from the entire input sequence and then update it [7]. In the past few years, transformer-based architectures have dominated in natural language processing (NLP), e.g., BERT [52] has shown superior performance among massive downstream NLP tasks. Besides, self-attention mechanism also demonstrates powerful modeling ability for non-grid data [243], and thus improves graph representation learning. In addition, transformer models have presented a new paradigm for computer vision tasks, including classification [53, 170], detection [23, 314, 181, 68], segmentation [23, 41, 302, 228, 261, 274], and representation learning [34, 11, 96]. Specifically, Dosovitskiy *et al.* [53] introduced a pure vision transformer model, termed as ViT, by applying a sequence of image patches, which achieves impressive performance on image classification. Recently, vision transformer [53, 170] has gradually become the new backbone for vision tasks, and massive large models based on transformers have emerged in computer vision, including CLIP [209], MoCo [98], DALL-E [210], and MAE [96]. Except for the backbone, transformer-based models, e.g., DETR [23], have also reformed the pipeline of detection and segmentation. Specifically, DETR [23], constructed upon the encoder-decoder transformer architecture, demonstrates a clear set-based method for detection and greatly simplifies the traditional

pipeline which includes many hand-designed components. Recently, Zhu *et al*. [314] present Deformable-DETR, which largely accelerates the convergence and improves the performance. Though the great success of transformer in computer vision, current approaches merely investigate the spatial self-attention, while ignoring the pixel-/patch-level relationships among different data samples.

3.3 Methodology

In this section, we first provide an overview of the main deep representation learning framework with BatchFormerV1. We then review vanilla vision transformer architecture and introduce BatchFormerV1 in detail. After that, we present a generalized BatchFormerV2 with the two-stream training strategy. Lastly, we provide some insights and give a discussion on the proposed method.

3.3.1 Overview

We aim to enable deep neural networks themselves with the ability to learn the relationships from a mini-batch of data samples during the end-to-end deep representation learning. The main deep representation learning framework with BatchFormerV1 is shown in Figure 3.3. Specifically, a backbone network is first used to learn representations for individual data samples, i.e., there is no interaction between different samples in each mini-batch. After this, a new module is introduced to capture the relationships between different samples by using the cross-attention mechanism in transformer, and we refer to it as Batch Transformer or BatchFormerV1. The output of BatchFormerV1 is then used as the input of the final classifier. To address the train-test inconsistency, we utilize an auxiliary classifier before BatchFormerV1, i.e., by sharing its weights with the final classifier, we can transfer the knowledge learned from sample relationships to the backbone and the auxiliary classifier. During testing, we can thus remove BatchFormerV1 and directly use the auxiliary classifier for classification.

3.3.2 Revisiting Vision Transformer

The great success of transformer architecture in NLP has recently spread to almost every region of computer vision known as vision transformers [53, 170, 23]. The transformer architecture not only achieves superior performance in different vision tasks [170] (e.g., image classification, object detection, and semantic segmentation), but also brings novel paradigms for some fundamental tasks (e.g., DETR [23] for object detection and MAE [96] for self-supervised learning). Among typical vision Transformers, the overall model usually consists of a stack of multiple Transformer encoder blocks [242], where each transformer encoder block contains a multi-head self-attention layer (MSA) followed by a feed-forward network (FFN). Specifically, the self-attention mechanism used in vision Transformer can be described as follows. Given $\mathbf{Q},\ \mathbf{K},\ \mathbf{V} \in R^{N\times C}$ as the query, key, and value, respectively, where N is the number of image patches (or tokens) and C is the embedding dimension. We then have the output $\mathbf{Z}$ for the self-attention module:

$$\mathbf{Z} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}})\mathbf{V}, \tag{3.1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are learned from the same input. Specifically, multi-head self-attention module applies attention by splitting the input into multiple representation subspaces and then concatenates the representations from different heads. From a perspective of information propagation, the transformer architecture aggregates the feature of the tokens via spatial attention. Different from attention layers in a typical vision transformer, the proposed method performs self-attention on the batch dimension, i.e., it aggregates the feature of the tokens from different samples within each mini-batch in an end-to-end learning way.

3.3.3 BatchFormerV1

We introduce the detailed structures of BatchFormerV1, including the transformer encoder and the shared classifier, as follows.

Transformer Encoder. The transformer encoder includes multi-head self-attention (MSA) and MLP blocks. A LayerNorm (LN) is used after each block. Let $X \in R^{N\times C}$ denote

a sequence of input features, where N is the length of the sequence and C indicates the dimension of input features. We then have the output of the transformer encoder as follows,

$$\hat{X}_l = LN(MSA(X_{l-1}) + X_{l-1}), \tag{3.2}$$

$$X_l = LN(MLP(\hat{X}_l) + \hat{X}_l), \tag{3.3}$$

where l indicates the index of layers in the transformer encoder. The multi-head self-attention layers have been widely used to model the relationships from channel and spatial dimensions [242, 53, 112]. Therefore, we argue that it can also be extended to explore the relationships in the batch dimension. As a result, different from typical usage of transformer layers, the input of BatchFormerV1 will be first reshaped to enable the transformer layers working on the batch dimension of the input data. By doing this, the self-attention mechanism in transformer layers then becomes the cross-attention between different samples for BatchFormerV1.

Shared Classifier. Since we can not assume batch statistics for testing, such as sample relationships, there might be a gap between the features before and after the BatchFormerV1 module. That is, we can not perform inference on new samples by directly removing BatchFormerV1. Therefore, apart from the final classifier, we also introduce a new auxiliary classifier to not only learn from the final classifier but also keep consistent with the features before BatchFormerV1. To achieves this, we simply share the parameters/weights between the auxiliary classifier and the final classifier. We refer to this simple yet effective strategy as "shared classifier". With the proposed "shared classifier", we can thus remove BatchFormerV1 during testing. BatchFormerV1 is thus a plug-and-play module for robust deep representation learning. BatchFormerV1 is also easy to implement using typical deep learning packages, e.g., Figure 3.4 shows how to implement BatchFormerV1 with several lines of python code based on Pytorch [198].

```
def batchformer_v1(x, y, encoder, is_training):
    # encoder: TransformerEncoderLayer(C,4,C)
    if not is_training:
        return x, y
    orig_x = x
    x = encoder(x.unsqueeze(1)).squeeze(1)
    x = torch.cat([orig_x, x], dim=0)
    y = torch.cat([y, y], dim=0)
    return x, y
```

FIGURE 3.4. Python Code of BatchFormerV1 based on Pytorch.

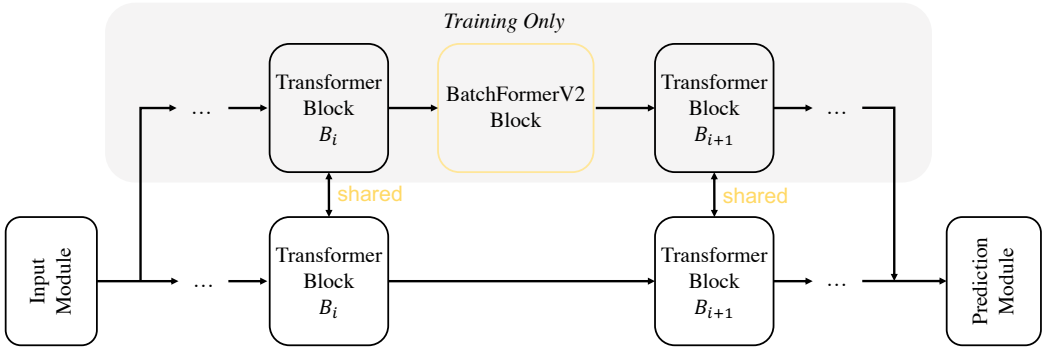

FIGURE 3.5. The two-stream training pipeline for BatchFormerV2. For example, the input indicates the feature map from backbone for DETR [23], while the input is the feature map after the patch embedding layer for ViT [53]. The outputs of two streams are the input of the shared prediction module. In addition, the transformer blocks and the prediction module, e.g., the transformer decoder in DETR [23] and the classification head in ViT [53], are shared by two streams. During inference, the stream with BatchFormerV2 is removed.

3.3.4 BatchFormerV2

To generalize the batch attention mechanism into pixel/patch level feature maps for dense representation learning, we devise BatchFormerV2 as follows. Given $\mathbf{Q},\ \mathbf{K},\ \mathbf{V} \in R^{B\times N\times C}$, we then have

$$\mathbf{Z}_i = \text{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{C}})\mathbf{V}_i, \quad \mathbf{Z} = \text{concat}(\mathbf{Z}_1, \ldots, \mathbf{Z}_N), \tag{3.4}$$

where $\mathbf{Q}_i,\ \mathbf{K}_i,\ \mathbf{V}_i \in R^{B\times C}$ and $\mathbf{Z} \in R^{B\times N\times C}$. Given the input for a specific layer/block with the spatial dimensions $H,\ W$, i.e., the number of image patches is $N = H \times W$. During training, at each spatial position $i = 1, \ldots, N$, we treat the batch of patch features

in current position as a sequence, i.e., we have N sequences each with the length of B. All above-mentioned sequences are then fed into a shared transformer block.

Two reasons that we use a shared transformer block are as follows: 1) it will increase the computation and memory consumption considerably if we use different blocks at different spatial positions; 2) it will be difficult to dense prediction with different sizes of input images, which is in line with the motivation of fully convolutional networks (FCNs) for dense prediction [173] as well as the convolution operations [147] and channel-wise attentions [113]. Therefore, we share the transformer block among the spatial dimensions in BatchFormerV2. By doing this, the proposed BatchFormerV2 can be implemented by simply transposing the spatial and batch dimensions before the standard multi-head self-attention layer. Noticeably, as illustrated in Figure 3.6, BatchFormerV2 can also be easily implemented with a few lines of code using popular deep learning packages such as PyTorch [198].

```
def batchformer_v2(x, encoder, is_training, is_first):
    # x: input with the shape (B, N, C).
    # encoder: TransformerEncoderLayer(C, nhead, C, batch_first=False)
    if not is_training:
        return x
    orig_x = x
    if not is_first:
        orig_x, x = torch.split(x, len(x)//2)
    x = encoder(x)
    x = torch.cat([orig_x, x], dim=0)
    return x
```

FIGURE 3.6. Python code of BatchFormerV2 based on PyTorch.

3.3.5 Two-Stream Training

One significant challenge for applying batch attention is the train-test inconsistency. Specifically, in BatchFormerV1, we address the train-test inconsistency by introducing a shared classifier, which enables removing the proposed module during inference. Inspired by this, we generalize this solution to dense representation learning in a similar way, i.e., learning batch-invariant representations. Therefore, we introduce a new two-stream training strategy as follows. When applying the proposed BatchFormerV2 module to a specific block of vision transformers, we create a new siamese stream followed by a BatchFormerV2 module, leaving

the original stream unchanged. That is, both two streams share the same Transformer block. By doing this, during training, all shared blocks are trained on a mixture of the distributions with or without a BatchFormerV2 module. Therefore, during testing, the original stream can work well for both with and without a mini-batch of testing data available. To avoid introducing any extra inference load, we thus remove the BatchFormerV2 module for testing. In addition, from the perspective of regularization, BatchFormerV2 also serves as a strong regularization during training, which has turned out to be very useful in vision transformers. Lastly, with the proposed two-stream training strategy, BatchFormerV2 can be easily integrated into existing transformer architectures for different visual applications in a plug-and-play manner, such as ViT [53] for image classification and DETR [23]/Deformable-DETR [314] for object detection.

3.3.6 Discussion

To better understand how the proposed method helps representation learning by exploring sample relationships, we also provide an intuitive explanation from the perspective of gradient propagation for optimization. For simplicity, we use BatchFormerV1 as an example.

Intuitively, without BatchFormerV1, all losses only propagate gradients on the corresponding samples and categories, i.e., one-to-one, while there are gradients on other samples with BatchFormerV1 (the dashed line) as illustrated in Figure 3.7. Specifically, given samples $X(X = X_0, X_1, X_i, ..., X_{N-1})$ and the corresponding losses $L_0, L_1, L_i, ..., L_{N-1}$ in the mini-batch, with BatchFormerV1, we then have

$$\frac{\partial L_i}{\partial X} = \frac{\partial L_i}{\partial X_i} + \sum_{j \neq i}^{N-1} \frac{\partial L_i}{\partial X_j}. \tag{3.5}$$

That is, BatchFormerV1 brings new gradient terms $\frac{\partial L_i}{\partial X_j}$, where $i \neq j$. From a perspective of gradient optimization, L_i also optimizes the network according to sample $X_j (j \neq i)$, that is a significant difference compared to the model without BatchFormerV1. In other word, $X_j (j \neq i)$ can be regarded as a virtual sample [293, 99] of y_i, where y_i is the label of X_i. We consider that both BatchFormerV1 and Mixup [293] can be regarded as data-dependent

augmentations. BatchFormerV1 implicitly draws virtual examples [293] from the vicinity distribution of samples via cross-attention module. From this perspective, BatchFormerV1 has implicitly augmented $N-1$ virtual samples for each label y_i via the relationship modeling among samples in the mini-batch. Previous approaches [8, 315] have demonstrated data augmentation is helpful for long-tailed recognition [99, 250], zero-shot learning [315], and domain generalization [311, 310]. Our gradients analysis in Section 3.4.5 also demonstrates the tail classes have larger gradients on other samples compared to head classes.

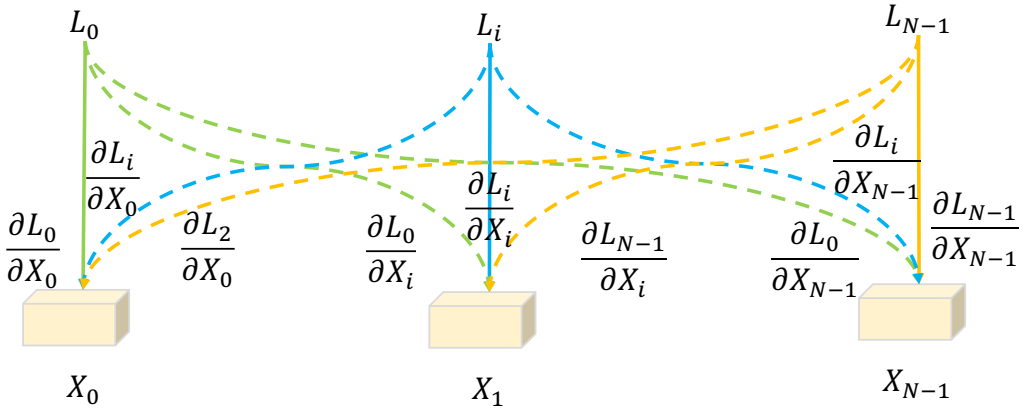

FIGURE 3.7. The gradient propagation scheme with the proposed BatchFormerV1 module. Dashed lines represent the new gradient propagation among data samples.

3.4 Experiments

In this section, we perform extensive experiments to show the effectiveness of the proposed method in 1) a variety of data scarcity learning settings, including long-tailed recognition, zero-shot learning, domain generalization, and contrastive learning; and 2) different visual recognition tasks, including image classification, object detection, and panoptic segmentation. For simplicity, we use BFV1 for BatchFormerV1 and BFV2 for BatchFormerV2 in all Tables.

3.4.1 Long-Tailed Recognition

In this subsection, we evaluate BatchFormerV1 on long-tailed recognition by using recent state-of-the-art methods, Balanced-Softmax [212], RIDE [265], and PaCo [44], as our baseline.

We use four popular datasets as follows: 1) CIFAR-100-LT has 50,000 training images and 10,000 validation images with 100 categories; 2) ImageNet-LT [171] contains 115.8K images of 1000 classes from ImageNet2012. The number of images in each class ranges from 5 to 1,280; 3) iNaturalist2018 [241] is a large-scale fine-grained dataset with 437.5K images from 8,142 categories; and 4) Places-LT [171] is a long-tailed scene classification dataset derived from [307] with 184.5K images from 365 categories with the number of per-category images ranging from 5 to 4,980. If not otherwise stated, we follow the same settings used in baseline methods. Particularly, there is a small difference for RIDE [265]. We train the model with the batch size 400 on four V100 GPUs for 100 epochs with an initial learning rate of 0.1 and 0.2 on ImageNet-LT and iNaturalist2018, respectively. The learning rate is decayed with cosine schedule on iNaturalist2018.

Results on CIFAR-100-LT. Table 3.1 shows that the proposed BatchFormerV1 is orthogonal to recent state-of-the-art methods such as Balanced-Softmax [212] and Paco [44]. We notice that BatchFormerV1 improves Balanced-Softmax by 2.4% for Few classes when the imbalance ratio is 100, and by 1.8% on Medium classes and by 1.2% on Few classes, respectively, when the imbalance ratio is 200. Besides, the performance of PaCo on imbalance ratio 200 also increases by 1.5% on Medium classes and 0.7% on Few classes, respectively. For ratio 100, BatchFormerV1 mainly improves Many classes since Paco has achieved good performance on Few classes. Overall, BatchFormerV1 improves the recognition of tail classes while maintaining the performance of head classes.

TABLE 3.1. Illustration of imbalance ratio 100 and 200 on CIFAR-100-LT. $*$ indicates the model is trained with the official code in one stage (e.g., [212]).

Method	100				200			
	All	Many	Med	Few	All	Many	Med	Few
RIDE [265]	48.0	68.1	49.2	23.9	-	-	-	-
Balanced [212]*	50.7	68.0	49.7	31.9	46.4	70.0	51.5	24.3
+ BFV1	**51.7**	**68.4**	49.3	**34.3**	**47.5**	**70.2**	**53.3**	**25.5**
Paco [44]	51.9	63.9	**53.0**	**36.5**	47.1	68.1	51.5	27.5
+ BFV1	**52.4**	**68.4**	52.1	34.0	**47.8**	68.1	**53.0**	**28.2**

Results on ImageNet-LT. In Table 3.2, BatchFormerV1 improves Balanced-Softmax by 2.4% on medium classes and **6.9%** on few classes respectively. Meanwhile, when using a ResNet-50

TABLE 3.2. Illustration of ResNet-10/50 on ImageNet-LT. $*$ indicates the model is trained with released code in one stage. RIDE-3e indicates three experts used in RIDE.

Method	ResNet-10				ResNet-50			
	All	Many	Med	Few	All	Many	Med	Few
OLTR [171]	35.6	43.2	35.1	18.5	-	-	-	
LFME [273]	38.8	47.0	37.9	19.2	-	-	-	-
Balanced [212]*	41.0	52.6	38.3	18.0	50.1	61.1	47.5	27.6
+ BFV1	**43.2**	**52.8**	**40.7**	**24.9**	**51.1**	**61.4**	**47.8**	**33.6**
RIDE-3e [265]*	44.7	57.0	40.3	25.5	53.6	64.9	50.4	33.2
+ BFV1	**45.7**	56.3	**42.1**	**28.3**	**54.1**	64.3	**51.4**	**35.1**
PaCo [44]	-	-	-	-	57.0	64.8	55.9	39.1
+ BFV1	-	-	-	-	**57.4**	62.7	**56.7**	**42.1**
two stage								
RIDE-3e [265]	45.9	57.6	41.7	28.0	54.9	66.2	51.7	34.9
+ BFV1	**47.6**	55.3	**45.5**	**33.3**	**55.7**	64.6	**53.4**	**39.0**

TABLE 3.3. Ablation studies of batch size. The backbone is ResNet-10.

	B=16	B=32	B=64	B=128	B=256	B=512
All	43.2	**43.6**	43.3	43.2	42.4	42.5
Many	**53.8**	53.7	52.9	52.8	52.2	52.0
Medium	40.2	40.7	**40.8**	40.7	39.5	39.6
Few	23.9	24.9	25.3	24.9	25.2	**25.8**

TABLE 3.4. Ablation studies on the shared classifier. The backbone is ResNet-50.

	All	Many	Medium	Few
Balanced [212] + BFV1 (shared)	**50.9**	**60.7**	**47.7**	34.1
Balanced [212] + BFV1	42.4	41.3	43.3	**42.2**

backbone, BatchFormerV1 improves Balanced-Softmax on the Few category by **5%**. When BatchFormerV1 is applied in RIDE [265], the results on Medium classes and Few classes increase by 1.8% and 2.8%, respectively. With a ResNet-50 backbone, BatchFormerV1 also improves Medium and Few classes by 1% and 1.9%, respectively. BatchFormerV1 achieves clear improvement overall, while the performance on Many classes drops. Furthermore, BatchFormerV1 also effectively improves RIDE under two-stage training strategy (RIDE uses a larger model to teach small model with distilling loss). Here, different from RIDE, we use a pre-trained model (the same model) to initialize the model and train the model

again with BatchFormerV1. PaCo [44] is recently introduced for long-tailed recognition with supervised contrastive learning. We also find BatchFormerV1 can facilitate ImageNet-LT on Medium classes and Few classes. Noticeably, PaCo uses a strong data augmentation strategy from supervised contrastive learning with 400 training epochs. We consider that data augmentation degrades the improvement of BatchFormerV1 on ImageNet-LT. We also show that BatchFormerV1 achieves comparable results on balanced ImageNet in the appendix.

In addition, we show the influences of different batch size and the shared classifier in Table 3.3 and Table A.24, respectively. Specifically, in Table 3.3, we find that BatchFormerV1 is less sensitive to the batch size that is small than 128, while the best performance on Few category is achieved with the batch size 512; In Table A.24, we see that BatchFormerV1 without a shared classifier achieves the similar performance on three categories, while BatchFormerV1 with a shared classifier maintains the performance on Many category. This demonstrates the effectiveness of a shared classifier and the re-balancing ability of BatchFormerV1.

Results on iNaturalist2018. We only evaluate BatchFormerV1 on RIDE since Balanced-Softmax has very limited performance on iNaturalist2018 while training PaCo requires over 36 GPU days. We train RIDE with cosine decay learning-rate scheduler in one stage and obtain a baseline slightly better than reported with 72.5%. As illustrated in Table 3.5, BatchFormerV1 improves the Medium and Few by 1.8% and 2.6%, respectively.

TABLE 3.5. Results on iNaturalist2018.

Method	All	Many	Medium	Few
BBN [308]	66.3	49.4	70.8	65.3
cRT [126]	65.2	69.0	66.0	63.2
RIDE-3e [265]	72.2	70.2	72.2	72.7
PaCo [44]	73.2	-	-	-
RIDE-3e [265]*	72.5	68.1	72.7	73.2
+ BFV1	**74.1**	65.5	**74.5**	**75.8**

Results on Places-LT. Table 3.6 illustrates that BatchFormerV1 improves BALMS on the Few category. Besides, BatchFormerV1 effectively improves PaCo [44], which also delivers a new state-of-the-art on Places-LT. Different from the results on CIFAR-100-LT and ImageNet-LT, here BatchFormerV1 mainly improves the Many category, which is possibly because PaCo

has much worse performance on the Many category and BatchFormerV1 can re-balance the learning on imbalanced training data.

TABLE 3.6. Results on Places-LT. The backbone is ResNet-152.

Method	All	Many	Medium	Few
OLTR [171]	35.9	44.7	37.0	25.3
τ-normalized [126]	37.9	37.8	40.7	31.8
BALMS [212]	37.8	41.4	38.8	29.1
+ BFV1	**38.2**	39.5	38.3	**35.7**
PaCo [44]	41.2	37.5	**47.2**	33.9
+ BFV1	**41.6**	**44.0**	43.1	33.7

TABLE 3.7. Results on compositional zero-shot learning. * means we use the released code to reproduce the results. S means seen, U means unseen, s means state and o means object.

Method	MIT-States						UT-Zap50K						C-GQA					
	AUC	HM	S	U	s	o	AUC	HM	S	U	s	o	AUC	HM	S	U	s	o
CompCos [178]	4.5	16.4	25.3	24.6	27.9	31.8	28.7	43.1	59.8	62.5	44.7	73.5	-	-	-	-	-	-
CGE [185]	6.5	21.4	32.8	28.0	30.1	34.7	33.5	60.5	64.5	71.5	48.7	76.2	3.6	14.5	31.4	14.0	15.2	30.4
CGE [185]*	6.3	20.0	31.6	27.3	30.3	34.5	31.5	46.5	60.3	64.5	46.3	74.4	3.7	14.9	30.8	14.7	**15.8**	29.0
+ BFV1	**6.7**	**20.6**	**33.2**	**27.7**	**30.8**	**34.7**	**34.6**	**49.0**	**62.5**	**69.2**	**49.7**	**75.6**	**3.8**	**15.5**	**31.3**	14.7	15.3	**30.0**

3.4.2 Zero-Shot Learning

We evaluate BatchFormerV1 for compositional zero-shot learning on three popular datasets: 1) MIT-States [119] consists of 30,000 training images of natural objects with 1,262 seen compositions (23.8 image per composition, 115 states and 245 objects), and 13,000 test images with 400 seen compositions and 400 unseen compositions; 2) UT-Zappos [285] includes 23,000 training images of shoes catalogue and we use the splits from [208]. UT-Zappos has 83 seen compositions for training (277.1 images per composition, 16 states, 12 objects) and 18 seen compositions and unseen compositions in test set; 3) C-GQA [185] provides 26,000 training images with 6,963 seen compositions (3.7 images per composition, 453 states, 870 objects) and 3,000 test images with 18 seen compositions and unseen compositions. All experiments are evaluated following the same settings in [185]. We adopt the evaluation protocol of [208] and report the Area Under the Curve (AUC) (in %) between the accuracy on seen and unseen compositions. Similar to [185], we also report unseen accuracy, seen accuracy, and the best harmonic mean.

Results. Table 3.7 shows that BatchFormerV1 effectively improves the AUC and HM among all datasets compared to the baseline. For fair comparison, we use the official code and the same setting to reproduce [185] as our baseline. We notice that BatchFormerV1 mainly improves the Seen category on MIT-States and C-GQA, while it largely improves the Unseen category by nearly **5%** on UT-Zap50K. This might be because the number of seen composition instances on MIT-States and C-GQA is few, e.g.23.8 image per seen composition on MIT-States and 3.7 images per seen composition. In other words, the recognition of seen compositions on MIT-States and C-GQA is few-shot learning. We think BatchFormerV1 can find invariant features among images of the same class on two datasets.

3.4.3 Domain Generalization

In this subsection, we first show that BatchFormerV1 effectively improves the baseline without other domain generalization techniques on PACS [149]. We then apply BatchFormerV1 to a recent popular domain generalization method, e.g., SWAD [25].

We perform experiments on four popular domain generalization datasets: 1) PACS [149] covers 7 object categories and 4 domains (Photo, Art Paintings, Cartoon and Sketches); 2) VLCS [58] contains 10,729 images from 4 domains and 5 classes; 3) OfficeHome [245] contains 15,588 images from 4 domains and 65 classes; and 4) TerraIncognita [14] contains 24,788 images from 4 domains and 10 classes. For the baseline, we use ResNet-18 as the backbone. The model is then trained with the SGD optimizer for 30 epochs using an initial learning rate of 0.001 for the first 24 epochs and 0.0001 for the last 6 epochs. For fair comparison, we only use common data augmentations, e.g., flip, color-jiter, and scale. All models are trained for five times and the average performance is applied. When comparing with others methods, the implementations are based on [124]. For SWAD [25], we use the official code and always follow the same setting (see more details in the appendix).

Results. We show the experimental results on the domain generation in Table 3.8 and Table 3.9. Specifically, Table 3.8 illustrates that BatchFormerV1 consistently improves the

baseline methods: CORAL [231] and MixStyle [309]. In Table 3.9, we see that BatchFormerV1 clearly improves recent state-of-the-art method [25] on all four datasets. BatchFormerV1 improves [25] by over 2% on both OfficeHome and TerraIncognita, indicating that it can facilitate invariant representation learning to improve cross-domain generalization.

TABLE 3.8. Illustration of BatchFormerV1 for domain generalization on PACS [149].

Method	art_paint	cartoon	sketches	photo	Avg.
Baseline	79.9± 1.0	73.0±1.5	67.7± 3.0	95.7±0.4	79.1
+ BFV1	**80.4**±0.2	**73.8**±2.0	**68.6**±1.8	**96.3**±0.2	**79.8**
CORAL [231]	79.2±1.7	**75.5** ±1.1	71.4±3.1	94.7±0.3	80.2
+ BFV1	**80.6**±0.9	74.7±1.9	**73.1**±0.3	**95.1**±0.3	**80.9**
MixStyle [309]	81.7±0.1	**76.8**±0.0	80.8±0.0	93.1±0.0	83.1
+ BFV1	**84.8** ±0.4	75.3±0.0	**81.1** ±0.4	**93.6**±0.0	**83.7**

TABLE 3.9. Illustration of BatchFormerV1 for domain generalization based on SWAD [25]. The backbone is ResNet-18.

Method	PACS	VLCS	OfficeHome	TerraIncognita
SWAD [25]	82.9	76.3	62.1	42.1
+ BFV1	**83.7**	**76.9**	**64.3**	**44.8**

3.4.4 Contrastive Learning

Contrastive learning aims to learn representations that attract similar samples and dispel different samples, while BatchFormerV1 builds a transformer network among samples to implicitly explore sample relationships for representation learning. Therefore, BatchFormerV1 can also be easily applied to contrastive learning. We mainly evaluate BatchFormerV1 with MoCo-v2 [35] and MoCo-v3 [40] using linear classification protocol. We also show object detection results based on these pretrained backbone models in the appendix. As shown in Table 3.10, when using ResNet-50 as backbone, BatchFormerV1 can consistently improve self-supervised learning via contrastive learning, e.g., BatchFormerV1 improves both MoCo-v2 and MoCo-v3 by around 1.% on ImageNet-1k.

TABLE 3.10. Illustration of BatchFormerV1 for contrastive learning.

Method	Epochs	Top-1	Top-5
MoCo-v2 [35]	200	67.5	-
+ BFV1	200	**68.4**	88.5
MoCo-v3 [40]	100	68.9	-
+ BFV1	100	**69.8**	89.5

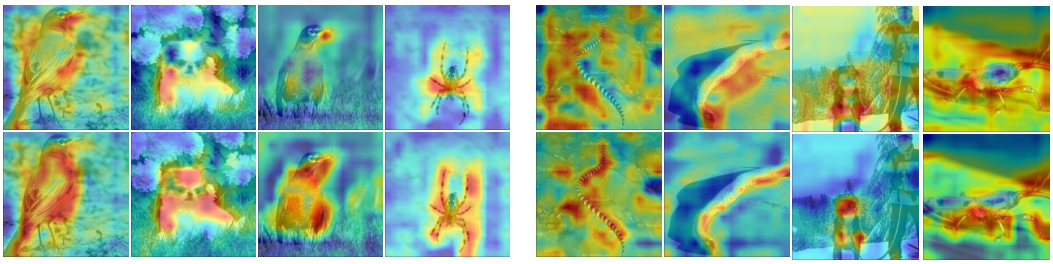

FIGURE 3.8. Visualization of BatchFormerV1 on low-shot test images using Grad-Cam [212]. The first row is baseline, and the second row is BatchFormerV1. The left part of the figure shows that BatchFormerV1 enables the model to pay attention to more details when the scene is simple and clean, while the right part of the figure shows that BatchFormerV1 facilitates to ignore the spurious correlation in the image. More figures are shown in the appendix.

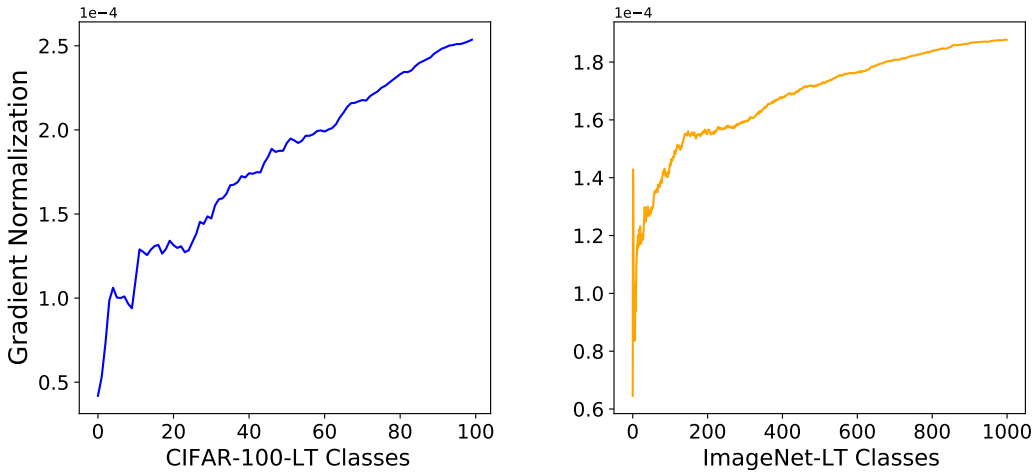

FIGURE 3.9. The gradient of each class to other images in mini-batch on CIFAR-100-LT and ImageNet-LT (based on [212]). For each class, we obtain the gradient norm to other images in all mini-batches, and then average the gradients of each class. The classes are sorted by descending order according to the number of instances.

3.4.5 Analysis

In this subsection, we provide some analyses of BatchFormerV1, including visual and gradient analyses.

Visual Analysis. We illustrate the visualized comparison with Grad-CAM [220] between the baseline and BatchFormerV1 in Figure 3.8. We find BatchFormerV1 focuses on more details of objects and ignores spurious correlations. On the one hand, when the image includes complex scenes with many disturbing factors, BatchFormerV1 effectively improves the attention of the network on the corresponding object regions (e.g., the sea snake on the sandbeach, the dog on the snow and the insect on the leaf in Figure 3.8). On the other hand, BatchFormerV1 also pays more attention on regions of the object when the scene is clear (e.g., the bird, dog, and spider). See more results in the appendix.

Gradients Analysis. BatchFormerV1 has increased new gradient backward: the loss of each label has the gradients on other images. In other words, we have implicitly augmented the samples for the class of each image in the mini-batches. The other images in the mini-batch can also be regarded as the virtual instances of current image class. The gradient is firmly related to the effect of each image label on other images. Figure 3.9 illustrates the rarer the class is, the larger the gradients of the class have on other images in the mini-batch. Thus, BatchFormerV1 actually utilizes the other images to facilitate low-shot recognition via increasing gradients of few-shot labels on other images.

3.4.6 Image Classification

In this subsection, we evaluate BatchFormerV2 for image classification using vanilla vision transformer or ViT [53].

To compare with baseline methods, we follow the same training strategy with DeiT [238]. We perform image classification experiments on two popular datasets, CIFAR-100 [140] and ImageNet [50]. For CIFAR-100, we follow the setups in [34] and train all models with an initial learning rate 6e-4 and the batch size 1024. For ImageNet, we train all models for

300 epochs with an initial learning rate 1e-3 and the batch size 1024. All experiments are conducted on a cluster with eight NVIDIA A100 GPUs (40GB). When applying the proposed BatchFormerV2 module, we always use the same number of heads with the corresponding baseline model. For CIFAR-100, we insert BatchFormerV2 for all layers. For ImageNet, we insert BatchFormerV2 in the eighth layer. Empirically, we observe frequent crashes during training if we apply BatchFormerV2 in very early layers on ImageNet. More details and discussions are provided in the appendix.

Results on CIFAR-100. Current vision transformer architectures (e.g., ViT [53]) usually require large-scale training data or strong regularization to avoid the overfitting problem. Therefore, it still remains challenging to train vision transformers from scratch on small datasets. In this chapter, we also find that the proposed BatchFormerV2 module can significantly improves the performance of vision Transformer on small datasets. As illustrated in Table 3.11, BatchFormerV2 significantly improves the performance of DeiT-B from 52.2% to 66.6 % by **14.4%**, DeiT-S from 57.5% to 68.5% by **11%**, DeiT-Ti from 49.2% to 58.7% by **9.5%**. When we train all models with more epochs, i.e., 300 epochs, the improvement is also considerable. The possible reason is that BatchFormerV2 enables the information propagation among patches in different images, which benefits the optimization and generalization when learning on small datasets. Specifically, we may find that DeiT-B does not achieve better performance compared to DeiT-S. This is possibly because DeiT-B is a too large model for a small dataset such as CIFAR-100.

TABLE 3.11. Image classification results on CIFAR-100. Specifically, following the experimental setups in [34], we train all models from scratch and report the top-1 accuracy (%).

Model	#Params	Input	Epochs=100	Epochs=300
DeiT-Ti [238]	5M	224^2	49.2	69.2
+ BFV2	5M	224^2	**58.7**	**73.4**
DeiT-S [238]	22M	224^2	57.5	72.5
+ BFV2	22M	224^2	**68.5**	**75.2**
DeiT-B [238]	86M	224^2	52.2	71.8
+ BFV2	86M	224^2	**66.6**	**74.8**

TABLE 3.12. Illustration of BatchFormerV2 on common object detection using DETR [23] and Deformable-DETR [314] as baselines. Following [314], we train all models with 50 epochs using the official code. * indicates using iterative bounding box refinement.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR [23]	ResNet-50	34.8	55.6	35.8	14.0	37.2	54.6
+ BFV2	ResNet-50	**36.9**	**57.9**	**38.5**	**15.6**	**40.0**	**55.9**
Conditional-DETR [181]	ResNet-50	40.9	61.8	43.3	20.8	44.6	59.2
+ BFV2	ResNet-50	**42.3**	**63.2**	**45.1**	**21.9**	**46.0**	**60.7**
SMCA (single scale) [68]	ResNet-50	41.0	-	-	21.9	44.3	59.1
+ BFV2	ResNet-50	**42.3**	**63.5**	**45.4**	**22.5**	**45.7**	**60.1**
Deformable-DETR [314]	ResNet-50	43.8	62.6	47.7	26.4	47.1	58.0
+ BFV2	ResNet-50	**45.5**	**64.3**	**49.8**	**28.3**	**48.6**	**59.4**
Deformable-DETR* [314]	ResNet-50	45.4	64.7	49.0	26.8	48.3	61.7
+ BFV2	ResNet-50	**46.7**	**65.6**	**50.5**	**28.8**	**49.7**	**61.8**
Deformable-DETR [314]	ResNet-101	44.5	63.7	48.3	25.8	48.6	59.6
+ BFV2	ResNet-101	**46.0**	**65.2**	**50.5**	**28.4**	**49.8**	**60.7**

Results on ImageNet. Table 3.13 shows that BatchFormerV2 consistently improves the performance among different ViT models. We observe BatchFormerV2 achieves similar improvement, i.e., around 0.5%, compared to the baseline. Compared to the improvement on object detection and panoptic segmentation, the improvement on image classification is relatively small. It might be because dense prediction requires to localize the objects in the images, i.e., there are multiple targets in the image, while image classification treats the whole image as the target and requires to recognize the image. Furthermore, we think the strong data augmentation in classification might be also a limitation for BatchFormerV2 on ImageNet.

TABLE 3.13. Image classification results on ImageNet.

Model	#Params	Input	Top-1	Top-5
DeiT-Ti [238]	5M	224^2	72.2	91.1
+ BFV2	5M	224^2	**72.7**	**91.5**
DeiT-S [238]	22M	224^2	79.8	95.0
+ BFV2	22M	224^2	**80.4**	**95.2**
DeiT-B [238]	86M	224^2	81.7	95.5
+ BFV2	86M	224^2	**82.2**	**95.8**

3.4.7 Object Detection

In this subsection, we evaluate BatchFormerV2 for common object detection using popular transformer-based object detectors such as DETR [23], Conditional-DETR [181], SMCA [68], and Deformable-DETR [314].

We perform all experiments on the most popular common object detection benchmark dataset, COCO 2017 [164], which contains 118k training images and 5k validation images. During training, the backbone network is initialized from the weights pretrained on ImageNet-1K [50]. We run experiments on eight NVIDIA V100 GPUs (16GB) for DETR, and eight NVIDIA A100 GPUs (40GB) for Deformable DETR. If not otherwise stated, the batch-size for DETR, Conditional-DETR and SMCA is 16 and the default batch size for Deformable-DETR is 24. We insert the BatchFormerV2 module in the first transformer encoder layer in all experiments. The number of heads in BatchFormerv2 is 4. For fair comparison, all other hyperparameters follow the default configurations described in DETR [23], Conditional-DETR [181], SMCA [68] and Deformable-DETR [314].

Results on COCO. In Table 3.12, BatchFormerV2 significantly improves the corresponding baseline methods. For example, without bells and whistles, BatchFormerV2 improves DETR by 2.1% and Deformable-DETR by 1.7% when using a ResNet-50 backbone. We observe consistent improvement on Conditional-DETR and SMCA. Moreover, we find that BatchFormerV2 mainly improves the object detection performance on small and medium objects. For example, BatchFormerV2 increases Deformable-DETR in APS by 1.9% and APM by 1.5%, respectively. For DETR, BatchFormerV2 increases AP_S and AP_M by 1.6% and 2.8%, respectively. We think via building Transformer along the pixel of the feature map in the batch dimension, BatchFormerV2 utilizes features from other images to facilitates object detection in current image. For small objects which is usually challenging to detect, BatchFormerV2 is able to incorporate objects from other images to detect (See visualization results in this section). Therefore, BatchFormerV2 significantly improves corresponding baselines. To better understand BatchFormerV2 for object detection, we also perform ablation studies on some key factors that may have influences on model performance as follows.

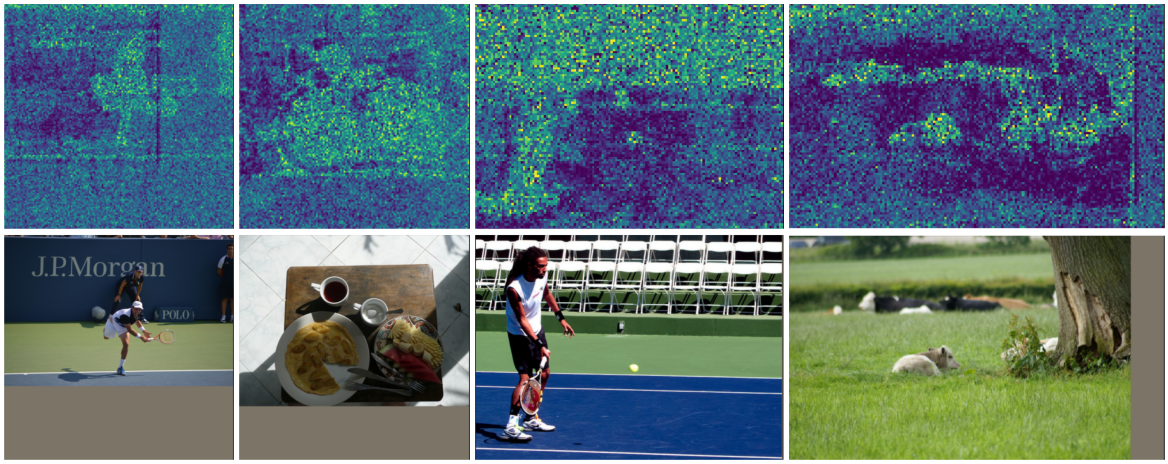

FIGURE 3.10. Visualization of the attention maps. We apply BatchFormerV2 in the first transformer layer in Deformable-DETR.

Ablation Study on Batch Size. Considering that BatchFormerV2 aims to learn sample relationships among each mini-batch during training, we evaluate the influence of different mini-batch size on BatchFormerV2 as follows. As shown in Table 3.14, 1) when increasing the batch-size from 16 to 24, the performance can be further improved with a small margin; and 2) when further increasing the batch-size to 32, the performance is comparable, i.e., no additional improvements. Here, we maintain other hyper-parameters when increasing the batch size. We consider that it may require to tune other hyperparameters after increasing the batch size to achieve further improvements.

TABLE 3.14. Ablation study on batch size. Specifically, we use Deformable-DETR as our baseline and insert BatchFormerV2 module in the last transformer layer.

Batch Size	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
16	50	44.7	63.5	48.9	27.3	48.1	59.1
24	50	**45.1**	**64.1**	**49.3**	**28.5**	**48.4**	59.4
32	50	44.9	63.8	48.8	27.7	48.3	**60.0**

Ablation Study on Insert Position. We perform ablation studies on different insert positions of BatchFormerV2 and show the influences on model performance. Specifically, we use Deformable-DETR as the baseline, which contains six transformer layers. As shown in Table 3.15, we find that: 1) the insert positions do have an important effect on the performance;

and 2) the number of BatchFormerV2 modules does not have significant influence on the final performance, i.e., more BatchFormerV2 modules cannot further improve the object detection performance; and 3) applying BatchFormerV2 in early layers seems to be more effective for dense prediction tasks.

TABLE 3.15. Ablation study on insert position. Specifically, "L1-3" indicates that we apply BatchFormerV2 modules from the first layer to the third layer.

	L1-2	L1-3	L3-6	L4-6	L5-6	L1	L2	L3	L4	L5	L6
AP	45.2	44.9	45.4	44.9	45.0	**45.5**	45.3	45.2	45.2	45.2	45.1

Ablation Study on Shared Modules. We evaluate whether sharing all BatchFormerV2 modules among different layers will benefit dense prediction tasks. As shown in Table 3.16, we find that, without sharing the modules, it could bring 0.5% improvement compared to the shared scheme. It may suggest that the dense sample relationships are varying among different layers/levels, which also explains that, in different visual recognition tasks, BatchFormerV2 may need to be added into different layers.

TABLE 3.16. Ablation study on shared modules. Specifically, here we apply BatchFormerV2 from the third to the sixth layers. All models are trained with 50 epochs.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Deformable-DETR	43.8	62.6	47.7	26.4	47.1	58.0
+ BFV2 (shared)	44.9	63.6	49.1	27.7	47.9	59.6
+ BFV2	**45.4**	**64.3**	**49.5**	**28.6**	**48.5**	**59.9**

Visualization. BatchFormerV2 is applied into transformer encoder layers along each spatial position, and it enables the information propagation among samples in a mini-batch via attention mechanism. Here, we visualize the attention of each slot on other slots along with the same position among the batch dimension. Specifically, we show the feature map in the first transformer encoder layer of Deformable-DETR. Figure 3.10 shows that BatchFormerV2 mainly focuses on the objects (e.g., person, chairs), while paying less attention on the background (e.g., play ground, grass). This actually demonstrates that BatchFormerV2 improves the object localization via the attention among different samples. This is also

consistent to the panoptic segmentation results in Section 3.4.8, where the improvements on the *things* categories are significant than those on the *stuff* categories. In addition, we observe that self-attention highlights all the regions of the objects with clear boundaries.

FIGURE 3.11. A visual comparison between DETR [23] with and without BatchFormerV2. Specifically, the first row is the original image, the second row is the result without BatchFormerV2, and the last row is the result with BatchFormerV2.

TABLE 3.17. Panoptic segmentation with DETR [23] on the COCO val dataset.

Method	PQ	SQ	RQ	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}	SQ^{st}	RQ^{st}	AP
DETR [23]	43.4	79.3	53.8	48.2	79.8	59.5	36.3	78.5	45.3	31.1
+ BFV2	**45.1**	**80.3**	**55.3**	**50.5**	**81.1**	**61.5**	**37.1**	**79.1**	**46.0**	**33.4**

3.4.8 Panoptic Segmentation

In this subsection, we evaluate BatchFormerV2 for panoptic segmentation, i.e., a combination of instance and semantic segmentation, on the MS-COCO dataset. We use the panoptic annotation in [137], which contains additional 53 *stuff* categories in addition to 80 *things* categories from the original MS-COCO dataset. We use DETR [23] as our baseline, i.e., we utilize a mask head to generate panoptic segmentation results for both *stuff* and *things* classes in a unified way [137]. Following [23], we first train the model with BatchFormerV2 for

object detection to predict bounding boxes around *stuff* and *things* classes 300 epochs. We then finetune the new mask head for extra 25 epochs.

Results. We report the panoptic quality (PQ) and the breakdown performances on things (PQ^{th}) and stuff (PQ^{st}) in Table 3.17. Specifically, we observe that BatchFormerV2 significantly improves AP by 2.3% and PQ by 1.7%. We also notice the improvement on PQ^{th} is much larger than PQ^{st}. That is, BatchFormerV2 improves PQ^{th} by 2.3%, while the improvement on PQ^{st} is only 0.8%. This result is consistent with the results of object detection: by enabling the information propagation, BatchFormerV2 mainly facilitates object detection and instance segmentation. Furthermore, following [23], we actually freeze the bounding box branch and transformer layers (include BatchFormerV2) when finetuning the mask head, we find that the performance of panoptic segmentation is also significantly improved. A possible explanation is that BatchFormerV2 improves the optimization of the backbone and the transformer encoder for better object detail modeling for bounding box detection and subsequently facilitates the segmentation performance when finetuning the mask head.

Visualization. To better understand how BatchFormerV2 helps dense prediction, we provide visualization results in Figure 3.11. Specifically, BatchFormerV2 mainly improves the segmentation details and the small object segmentation. For example, the segmentation boundaries of the door have been significantly improved, while the baseline model mistakenly segments the door as the wall in Figure 3.11. Meanwhile, the legs of the desk are more clear with the help of BatchFormerV2. In the second column, the segmentation details of airplane become better with BatchFormerV2, e.g., the tail and front wheels. In the last column, the baseline model ignores the segmentation of the grass (i.e., small stuffs), while it can correctly segment the grass with BatchFormerV2.

3.4.9 HOI Detection

This chapter proposes a novel BatchFormer module for sample relationship exploration. Chapter 2 demonstrates the compositional learning framework for Human-Object Interaction Exploration, which actually explicitly leverages the relationships among different HOI images,

i.e., different interactions might share similar objects or verbs. Different from explicit or supervised compositional learning according to bounding boxes and the corresponding verb or object labels in Human-Object Interaction, BatchFormer presents an implicit way to explore the sample relationship. Table 3.18 presents the results when we apply BatchFormer to HOI detection approaches. BatchFormer is also able to effectively improve HOI detection based on DETR-based methods, particularly for rare category.

TABLE 3.18. Evaluation of BatchFormerV2 on HICO-DET

Method	Full	Rare	NonRare
Baseline	28.16	19.48	30.76
+ BFV2	**28.46**	**21.55**	30.52

3.5 Summary

In this chapter, we propose to enable deep neural networks themselves with the abilities to explore the sample relationships and enable the implicit representation transfer from each training mini-batch. Specifically, we consider each image in a mini-batch as one node of a sequence, and employ a transformer encoder network among all images to learn sample relationships among them. The proposed BatchFormerV1 enables the gradient propagation of each label to all images in the mini-batch, which can be intuitively seen as virtual sample augmentation, and thus benefits robust representation learning. To further explore multi-scale sample relationships for dense prediction tasks, we generalize the proposed module as BatchFormerV2. Meanwhile, we introduce a two-stream training pipeline, where two streams share all other layers/blocks except the proposed modules. By doing this, the proposed module can be a plug-and-play module and easily integrated into different vision transformers without any extra inference cost. To evaluate the proposed module. We perform extensive experiments on over ten datasets, which show that the proposed method achieves significant improvements in 1) different data scarcity settings, including long-tailed recognition, zero-shot learning, domain generalization, and contrastive learning; and 2) different visual recognition tasks, including image classification, object detection, and panoptic segmentation.

CHAPTER 4

3D Human-Object Interaction Animation

After exploring the hierarchical visual relationships from existing relationships and sample relationships in the 2D perspective, the thesis further investigates the 3D visual relationships based on the 3D human-object interaction and learns to transfer the poses for novel objects. Neural rendering of animatable 3D human avatars has been intensively explored by implicit neural representations, while the rich human-object interactions (HOIs) are crucial for numerous human-centric scene capturing/understanding applications such as AR/VR and robotics. In this thesis, we address the challenge of HOI animation in a compositional manner, i.e., animating novel HOIs including the novel interaction, human and/or object via a sequence of novel driving poses. Specifically, we first adopt the neural human-object deformation to model and render HOI dynamics based on the neural representations. We then devise a new compositional conditional neural radiance field (or CC-NeRF), which decomposes the interdependence between the human and object latent codes to enable compositionally controlling the animation of novel HOIs. Extensive experiments show that the proposed method generalizes well to novel HOI animation.

4.1 Motivations and Contributions

Rendering 3D human-object animation is of great importance for human-centric generation with a wide range of real-world applications such as telepresence, video games, films, AR/VR and robotics. However, reconstructing and rendering human avatars with the interactive objects remains poorly investigated. Since traditional 3D reconstruction methods highly depend on dense cameras or depth sensors [215, 54, 55], implicit neural representations for graphical

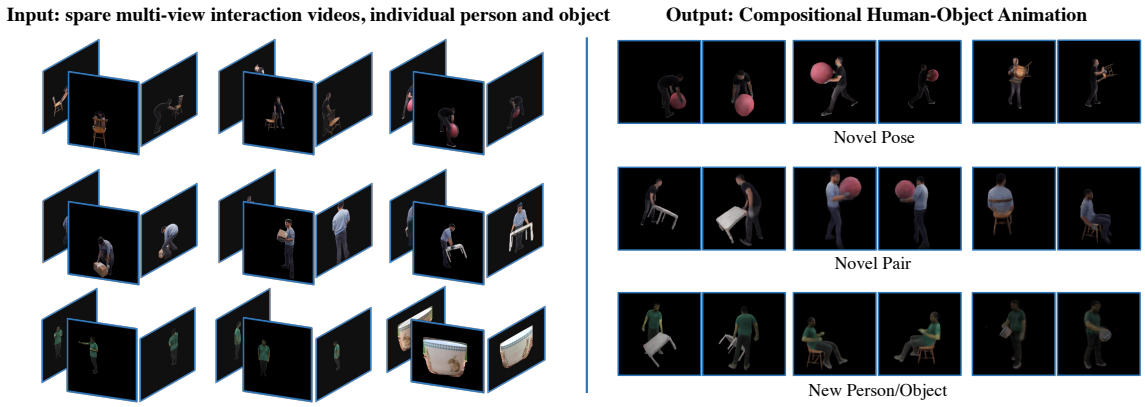

FIGURE 4.1. An illustration of compositional human-object neural animation. Given a set of sparse multi-view RGB HOI short videos with less than 50 frames, we render the neural animation of novel HOIs with novel pose, human, and object. Specifically, most faces in the training dataset are partly blurred.

objects present appealingly realistic results without the requirement of complex hardware and thus receive increasing attention from the community [182, 183, 12, 188]. Specifically, [183] introduces an implicit representation, i.e., neural radiance fields (NeRF), which represents static or rigid 3D objects/scenes as color and density fields and is capable of efficiently learning 3D geometry from images with differentiable volume rendering techniques.

To explore dynamic non-rigid scenes and objects, vanilla NeRF has been recently extended to handle deforming scenes [196, 240, 200] and motion modeling [159, 270, 207]. On the one hand, several methods propose to represent multiple frames of human body with implicit neural representations under the control of skeleton [199, 165, 230, 143, 279, 190, 303, 151, 255], achieving considerable performance in free-viewpoint human avatar rendering and demonstrating good generalization to novel human poses. However, the above-mentioned methods usually focus on the animation of either individual human body or object, leaving the interactions between human and object poorly investigated. On the other hand, several methods propose to explore the interactions between human and the surrounding objects or environments [233, 88, 82, 299, 46, 294, 280, 169]. Nevertheless, they mainly aim to reconstruct human/object shape and appearance rather than rendering animatable HOIs.

To jointly capture dynamic human body and objects, we thus generalize deformable neural radiance fields for human-object interaction. Specifically, the objects are regarded as disconnected joints in correspondence to the human body, and we then construct "pseudo bones" based on the object joint together with human body bones to model and control the dynamics of human-object interactions. By doing this, we extend the idea of animatable volumetric avatars to human-object interactions with non-linear pose-dependent deformation fields. Therefore, with the carefully designed canonical human-object pose, the generalized deformation fields and coordinate-based volumetric rendering, we can reconstruct and animate existing HOIs. Notably, recent animation approaches usually leverage meshes or poses to control the rendering for human or animals [199, 230, 190, 303, 151, 255, 229]. Considering that mesh or prior model is not always available for novel objects, and we thus adopt 3D human pose and 6-DoF object pose to simplify the control of HOI animation, which requires only the person skeleton and object 6-DoF pose and multi-view images. We introduce the details of neural human-object deformation in Section 4.3.3.

Considering the compositional nature of HOI, which is composed of a person, a verb/action, and an object, the animation of HOIs is thus not only related to novel poses/actions but also novel human and object. Specifically, due to the combinatorial explosion, it is impractical to collect all possible interactions between human and object, involving various kinds of objects and human with different appearances and shapes. Therefore, we proposed to compositionally animate human-object interaction by introducing a new compositional conditional neural radiance fields (CC-NeRF). Specifically, we first utilize the conditional latent variables to control different people and objects [217, 167, 188], and then decompose the human and object latent codes via the compositional invariant learning. By doing this, we thus enable the controllable animation for novel human-object interactions. We introduce the details of compositional animation in Section 4.3.4.

In this thesis, we present a novel approach, named as compositional 3D human-object neural animation or CHONA, to implicitly reconstruct HOIs from sparse multi-view videos via coordinate-based neural representations, and compositionally animate HOIs under novel

poses/interactions, novel person and novel object. An illustration of compositional HOI animation is shown in Figure 4.1. Our contributions can be summarized as:

- We introduce an HOI animation framework by exploring neural HOI deformations.
- We devise a compositional conditional NeRF for compositional HOI animation, which enables generalizing to novel human and object.
- We perform comprehensive experiments to demonstrate that the proposed method not only improves the animation performance but also the compositional generalization.

4.2 Related Work

4.2.1 3D Human-Object Modeling

The interaction with objects is common in the people's daily life [120, 78]. Early work mainly investigate synthesizing human pose and object [120], human body reconstruction [62], object recognition [266], or human 3D pose estimation [2, 133, 38, 160] under the interaction with objects or environments. Recently, Human-Scene Synthesis [295, 296, 89, 298, 251, 90, 253] has attracted extensive interests from the community due to the potential applications in the meta universe. Those methods usually depend on the prior human model (e.g., SMPL) and only synthesize the human motion in the scenes, while compositional Human-Object neural animation aims to animate both human and object in a compositional manner. Recently, increasing approaches [16, 233, 232, 84, 125, 46, 280, 294, 115, 262, 276] focus on 3D Interactions between Human and its surrounding objects. Zhang [294] present to reconstruct the spatial arrangements of Human-Object Interaction. [280, 46] reconstruct the meshes of human-object interactions, while recent work [125] introduces the neural representations to human-object interaction and significantly advances the novel view synthesis performance. Particularly, a real HOI dataset, BEHAVE [16], consisting of 8 subjects and diverse objects, is introduced with spare views of HD videos and the poses of human and object. We mainly conduct our experiments based on BEHAVE. Concurrent work [85, 262, 276, 115] focus on reconstruction or 3D tracking, significantly ignoring interaction animations. Besides,

though current compositional approaches on human-centric interactions have studied the recognition [128], detection [111], object affordance [108], 2D generation [187], and 3D human-scene synthesis [298], the compositional 3D animation remains unsolved.

4.2.2 Animatable Avatars

3D Avatars [174, 200, 255, 151, 37] have been through a significant progress. Early work usually leverages SMPL [174] model to reconstruct or synthesize human body, however the body is usually naked. Recently, neural fields have dominated 3D shape representations and novel view synthesis. Peng *et al*. [200] present to implicitly reconstruct human body from spare videos with neural radiance fields with a carefully designed skinning deformation. Next, [199, 165, 143, 190, 230, 158, 303, 229] significantly facilitate the performance in novel view rendering for novel poses. More recently, [151, 255] demonstrate appealing avatar generation under out of distribution poses. Meanwhile, [158] presents to reconstruct high-fidelity human avatars from the monocular RGB video observation and the avatar prior. Specifically, [229, 151] requires only 3D skeletons and multiple multi-view frames to construct an animatable Avatar, while [255] requires a pre-trained body SDF model to control the animation. Considering the fewer requirements on body models (e.g., SMPL) of pose-dependent animation, we follow [229, 151] to compositionally control the animation with poses.

4.2.3 Neural 3D Representations

Neural representations [195, 182, 39] have revolutionized the 3D surface representation, and achieved continuous, high resolution outputs of arbitrary shape. Recently, NeRF [183] represents 3D points in the scene with density and color, and renders the scene with volumetric rendering techniques, achieving photorealistic novel rendering. Next, [12] extends NeRF to represent the scene at a continuously-valued scale with conical frustum. GRAF [217] represents the neural radiance conditioned on shape/appearance latent codes. GIRAFFE [188] further presents controllable image synthesis with Compositional Generative Neural Feature

Fields. However, those approaches [217, 188] mainly learn the representations from static scenes, in which the objects are not frequently occluded. Differently, Compositional Human-Object animation requires to control the synthesis from spare multi-view HOI videos, which includes massive occlusion for human body and objects. Meanwhile, [217, 188, 281] mainly control the image synthesis for the static 3D scenes with rigid objects via linear transformation, while our approach is able to deform the interaction in a non-linear way.

4.3 Methodology

In this section, we introduce the proposed compositional 3D human-object animation approach. Specifically, we first provide an overview of HOI animation and the popular neural radiance fields (NeRF). We then introduce the neural human-object deformation and the compositional conditional radiance fields in detail.

4.3.1 Overview

Given sparse multi-view inputs, including interaction videos, single person videos, and objects, Compositional 3D Human-Object Animation enables to not only render the interaction under novel interaction pose, but also animate the interaction between a novel person and novel objects. Specifically, we build a pseudo bone for the object, and then treat the pseudo bone equally with body bones. Next, following the popular body skinning deformation techniques [199, 303, 151, 255], we devise a neural Human-Object deformation method to construct animatable human-object interactions as illustrated in Figure 4.2. Moreover, in order to control the interaction animation with novel people or objects, we devise compositional conditional radiance fields with two conditional latent codes, representing human and object respectively, to control the human-object identity. Specifically, we devise a compositional invariant learning strategy to decompose the interdependence between human and object latent codes, and thus enable to compositionally control the animation for novel objects or people.

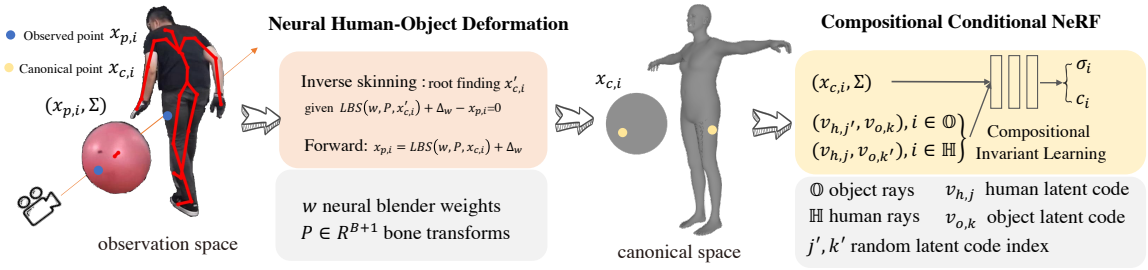

FIGURE 4.2. Overview of the proposed approach. The proposed compositional human-object neural animation approach leverages the neural Human-Object deformation module to deform the canonical points to posed points, and identify the corresponding canonical points of observed points via inverse skinning. Next, we obtain the density and color of the ray points conditioned on human and object latent codes, and accumulate the samples to render the pixel color. In addition, a compositional invariant learning strategy is introduced to decompose the interdependence between the two latent codes, and facilitate compositional human-object animation.

4.3.2 Neural Radiance Fields

NeRF [183] leads to significant progress in a wide range of 3D vision topics. It implicitly represents the geometry and appearance of the scene with a multi-layer perceptron neural network and volumetric rendering techniques. For a ray $\boldsymbol{r}$ and the viewing direction (θ, ϕ), NeRF first queries the emitted color c and density σ at the 3D point $\boldsymbol{x} = (x, y, z)$ in the ray $\boldsymbol{r}$, then uses volumetric rendering to get the pixel color $C(\boldsymbol{r})$ via accumulating the view-dependent colors along the ray $\boldsymbol{r}$ as follow,

$$C(\boldsymbol{r}) = \sum_{i}^{N} T_i(1 - exp(\sigma_i \delta_i))\boldsymbol{c}_i, \tag{4.1}$$

where $T_i = exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$, δ_i indicates the distances between the sample points along the ray. Then, it optimizes the network via calculating the distance loss between $C(\boldsymbol{r})$ and ground truth pixel color. Recently, Mip-NeRF [12] extends NeRF via taking each point in the ray as a cone, the samples $\boldsymbol{x}$ along the ray as conical frusta modeled as multivariate Gaussians (μ, Σ). Mip-NeRF accumulates the pixel color in a similar way to NeRF [183].

4.3.3 Neural Human-Object Deformation

To represent the human-object interaction via neural volumetric representation, we devise a canonical Human-Object representation as follows. For each point $\boldsymbol{x}_p$ in the observed/posed space, we have a corresponding point $\boldsymbol{x}_c$ in the canonical space, which can be deformed into $\boldsymbol{x}_p$ via neural skinning deformation. The canonical representation includes a Lambertian neural radiance field $F_{\Theta_r} : (\boldsymbol{x}_c, \boldsymbol{\Sigma}) \rightarrow (\boldsymbol{c}, \boldsymbol{\delta})$, where r denotes the pixel ray, $\boldsymbol{c} = (r, g, b)$ indicates the material color, $\boldsymbol{\delta}$ represents the material density respectively. Besides, we follow Mip-NeRF [12] to accumulate the samples $(\boldsymbol{x}_p, \boldsymbol{\Sigma})$ (a multivariate Gaussian) to render the pixel color at each ray. As illustrated in Figure 4.2, the canonical human-object space includes a T-pose body and an object placed in front of the body. We denote the transformation of each body bone as $\boldsymbol{T}_i, 0 \leq i < B$, where B is the number of body bones and $\boldsymbol{T}_i \in R^{4\times4}$. We represent the 6 DoF transformation $(\boldsymbol{t}, \boldsymbol{r})$ of the object from canonical space to the observed space as $\boldsymbol{T}_B$. As a result, we have $\boldsymbol{P} = \{\boldsymbol{T}_0, \boldsymbol{T}_1, ..., \boldsymbol{T}_B\} \in R^{(B+1)\times4\times4}$ representing the transformation of a human-object interaction.

Forward Skinning. Following the popular animatable avatar methods [37, 256, 255, 151], we revise the traditional linear blend skinning (LBS) [3, 174, 193] into neural skinning to deform a canonical Human-Object pose according to rigid bone transformations. We treat background as an additional bone. Thus, we have $B + 2$ bones for the human-object interaction. Similar to [199, 255, 151], a MLP function $F_{\Theta_s} : \boldsymbol{x_c} \rightarrow \boldsymbol{w}$ is used to project a canonical point $\boldsymbol{x_c}$ into corresponding weights $\boldsymbol{w}$. Given the skinning weights $\boldsymbol{w} = (w_0, w_1, ..., w_{B-1}, w_B, w_{bg}) \in R^{B+2}$ and a pose $\boldsymbol{P} = \{\boldsymbol{T}_0, \boldsymbol{T}_1, ..., \boldsymbol{T}_B\}$, we use forward LBS to define the deformation of a sample $\boldsymbol{x}_c$ in the canonical space to $\boldsymbol{x}_p$ in the view space:

$$\begin{aligned} \boldsymbol{x}_p &= LBS(F_{\Theta_s}(\boldsymbol{x}_c), \boldsymbol{P}, \boldsymbol{x}_c) + F_{\Theta_\nabla}(\boldsymbol{x}_c, \boldsymbol{P}) \\ &= [\sum_{j=0}^{B+1} F_{\Theta_s,j}(\boldsymbol{x}_c) \cdot \boldsymbol{T}_j + w_{bg} \cdot \boldsymbol{I}] \cdot \boldsymbol{x}_c + F_{\Theta_\Delta}(\boldsymbol{x}_c, \boldsymbol{P}), \end{aligned} \tag{4.2}$$

where $\boldsymbol{I} \in R^{4\times4}$ denotes identity matrix, $F_{\Theta_\Delta} : (\boldsymbol{x}_c, \boldsymbol{P}) \rightarrow \Delta_w \in R^3$ is for modeling the non-linear deformations [151]. Similar to [267, 151], rather than the surface points as

traditional LBS, we skin all points in the 3D space with $\boldsymbol{w}_{bg} \cdot \boldsymbol{I}$, which stops deforming the points in the background and empty space.

Inverse Skinning. It requires to transformer the observed points into canonical space for rendering the model. Similar to [37, 255, 151], we leverage the root finding strategy [37] to deform $\boldsymbol{x}_p$ to $\boldsymbol{x}_c^{'}$ subject to,

$$f(\boldsymbol{x}_c^{'}) = LBS(F_{\Theta_s}(\boldsymbol{x}_c^{'}), \boldsymbol{P}, \boldsymbol{x}_c^{'}) + F_{\Theta_\Delta}(\boldsymbol{x}_c^{'}, \boldsymbol{P}) - \boldsymbol{x}_p = 0, \tag{4.3}$$

where $\boldsymbol{x}_c^{'}$ denotes the potential canonical point of $\boldsymbol{x}_p$. Then we solve it numerically via Newton's method similar to [255, 151]. Following [151], we simply use the $K = 5$ nearest bones of the point in observed space to initialize the Newton's method, such that the computational burden for the inverse skinning can be significantly reduced. Then, we get K corresponding points $\{\boldsymbol{x}_{c,0}^{'}, \boldsymbol{x}_{c,1}^{'}, ..., \boldsymbol{x}_{c,K-1}^{'}\}$ for the observed point $\boldsymbol{x}_p$. The same as [151, 37], we analytically compute the gradients of the network parameters for the inverse skinning. During volumetric rendering, similar to previous works [37, 151], we choose the density and color for the observed point $\boldsymbol{x}_v$ as follow,

$$\boldsymbol{c}_v = \boldsymbol{c}_{c,m}^{'}, \boldsymbol{\sigma}_v = \boldsymbol{\sigma}_{c,m}^{'}, \tag{4.4}$$

where $m = argmax_i(\boldsymbol{\sigma}_{c,i}^{'}), 0 \leq i < K$. we then use $(\boldsymbol{c}_v, \boldsymbol{\sigma}_v)$ for volumetric rendering.

4.3.4 Compositional Conditional Radiance Fields

Though the proposed human-object neural deformation enables the animation for a given human-object interaction, it fails to animate novel combinations, i.e.an interaction involves a novel person or a novel object. Due to the combinatorial explosion of Human-Object interactions, we can not collect the multi-view videos for all possible interactions, which significantly limits the potential applications of the proposed human-object neural deformation. Therefore, we devise Compositional Conditional Radiance Fields to enable compositionally animating the interactions from novel combinations, and even novel person and objects. To decouple the controlling of human and object, we use two latent codes for the Conditional

Radiance Fields. Denote $\boldsymbol{v}_h \in R^{N_h \times C}$ and $\boldsymbol{v}_o \in R^{N_o \times C}$, where N_h and N_o are the number of person and object categories, as the latent codes of human and object respectively, we have the conditional radiance fields as follows,

$$F_{\Theta_{\boldsymbol{r}}} : (\boldsymbol{x}_c, \boldsymbol{\Sigma}, \boldsymbol{v}_{h,j}, \boldsymbol{v}_{o,k}) \rightarrow (\boldsymbol{c}, \boldsymbol{\delta}), \tag{4.5}$$

where $0 \leq j < N_h$ and $0 \leq k < N_o$ denote the corresponding person and object for observed interaction. With the conditional radiance fields, we can control the rendering for different human-object pairs with $\boldsymbol{v}_h$ and $\boldsymbol{v}_o$. However, the two latent codes $\boldsymbol{v}_h$ and $\boldsymbol{v}_o$ are entangled together, i.e., the rendering is controlled jointly by two latent codes. Therefore, the conditional radiance field in Eq. (4.5) fails to animate the interactions of novel human or objects.

Compositional Invariant Learning. To decouple the interdependence between the human and object latent codes, we further introduce a compositional invariant learning strategy for conditional radiance fields, named as Compositional Conditional Radiance Fields, to ease the spurious correlation between human and object latent codes, and thus enable compositional neural animation. Specifically, for the pixel in human body, we expect $(\boldsymbol{c}, \boldsymbol{\delta})$ only dependent on $\boldsymbol{v}_h$, regardless of the value of $\boldsymbol{v}_o$. Thus, when the ray $\boldsymbol{r}$ is located in the human body, we randomly set the value for the object latent codes, and vice versa for rays in the object. Then, the color and density of the Compositional Conditional Radiance Fields for the points in ray $\boldsymbol{r}$ are presented as follows,

$$\boldsymbol{c}, \boldsymbol{\delta} = \begin{cases} F_{\Theta_{\boldsymbol{r}}}(\boldsymbol{x}_c, \boldsymbol{\Sigma}, \boldsymbol{v}_{h,j'}, \boldsymbol{v}_{o,k}) & \boldsymbol{r} \in \mathbb{O} \\ F_{\Theta_{\boldsymbol{r}}}(\boldsymbol{x}_c, \boldsymbol{\Sigma}, \boldsymbol{v}_{h,j}, \boldsymbol{v}_{o,k'}) & \boldsymbol{r} \in \mathbb{H} \end{cases} \tag{4.6}$$

where $\mathbb{H} \cap \mathbb{O} = \emptyset$, $\mathbb{H}$ and $\mathbb{O}$ represent the rays set of human and object respectively. $j^{'}$ and $k^{'}$ are random latent human and object codes respectively. Via randomly setting the latent codes, we can decouple the interdependence of the Human-Object pairs in the training set. Therefore, we can control the animation via human or object latent codes individually.

4.4 Implementation Details

In this thesis, we follow the linear blender skinning deformation [199, 255, 303, 151] to devise an additional pseudo bone for human-object interaction and two latent codes for compositional animation. We localize the human and object rays according to the provided segmentation in [16]. We model the shading effect similar to [151], and sample 64 points along a ray. Due to the camera distance difference between different datasets, we sample 2048 rays for BEHAVE images, 1024 rays for ZJU-mocap images, and 512 rays for CO3D images in each mini-batch. Meanwhile, we utilize two additional losses for skinning weights and non-linear deformations in [151] for optimization. The overall loss function thus is $\mathcal{L} = \mathcal{L}_{img} + \lambda\mathcal{L}_w + \beta\mathcal{L}_\Delta$, where $\mathcal{L}_{img}$ indicates image loss similar to [183], $\mathcal{L}_w$ represents the loss to encourage the onehot skinning weights $\boldsymbol{w}$, $\mathcal{L}_\Delta$ is to encourage the non-linear deformation term $F_{\Theta_\Delta}(\boldsymbol{x}_c, \boldsymbol{P})$ close to zero. Both $\mathcal{L}_w$ and $\mathcal{L}_\Delta$ are MSE losses. In our experiments, we set λ to 1.0 and β to 0.1. The Adam optimizer [134] is adopted for the training with an initial learning rate 5e-4 and an exponentially decay strategy to 5e-6.

4.5 Experiments

In this section, we perform experiments and show quantity results on HOI reconstruction, quality results on HOI animation, to evaluate the proposed method.

Datasets We leverage three datasets, i.e., BEHAVE [16], ZJU-mocap [200] and CO3D [211] for the compositional human-object neural animation. BEHAVE [16] is a 4D datasets with 8 subjects performing a wide range of interactions with 20 common objects from 4 camera views. BEHAVE provides estimated body poses and object poses for each interaction, while each interaction has less than 50 frames. BEHAVE includes many blurred faces and frames, which we provide the analysis in Appendix. ZJU-mocap [200] consists of 10 sequences captured with 23 calibrated cameras. We select one subject (386) and four cameras for evaluating compositional HOI animation on novel persons. CO3D is a large 3D object dataset

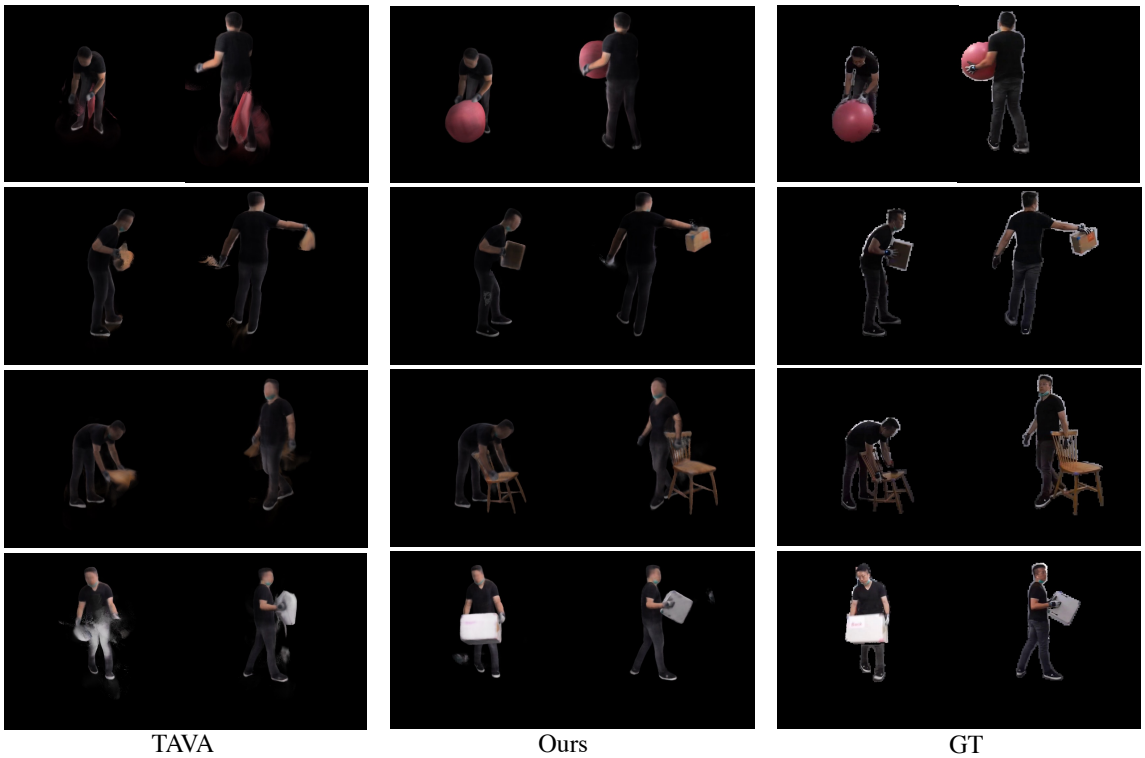

FIGURE 4.3. Visualized Comparisons between the proposed method and baseline method (TAVA [151]). We demonstrate the results of “yogaball”, “boxsmall”, “chairwood”, “boxlarge” with two distinct views.

with multiple sequences. We select "bowl" for evaluation on novel objects. More objects are illustrated in Appendix.

Metrics We mainly adopt the popular metrics in animatable avatars, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), for evaluating Human-Object animation. Meanwhile, for novel object and human, we mainly provide qualitative results.

4.5.1 Novel Pose Animation

In order to evaluate the novel pose animation, we select the first subject (S01), five kinds of different boxes, and seven classes of objects with distinct interactions from the BEHAVE dataset to construct a benchmark for novel pose animation. The objects with distinct interactions, including “backpack”, “chairwood”, “chairblack”, “suitcase”, “tablesmall”, “tablesquare” and “yogaball” are utilized to evaluate novel actions animation. The boxes consist of five scales,

i.e., “boxtiny”,“boxsmall”,“boxmedium”,“boxlarge” and “boxlong”, and we leverage it to demonstrate the effectiveness of the proposed method on different scales of objects. We randomly split the training and validation set for boxes, while we randomly choose two or one interaction in other classes for training and the remaining one for validation. The details can be found in Appendix. Previous work mainly reconstructs the Human-Object, while ignoring the HOI animation from sparse videos. We thus utilize the template-free animatable avatar method, TAVA [151], as our baseline method to demonstrate the deficiency without the modeling of objects for Human-Object Interaction. For each interaction in BEHAVE [16], there are only less than 50 frames. Therefore, we use one V100 GPU to run our experiments with 100,000 iterations.

Quantitative comparisons Table C.3 demonstrates that the proposed method consistently improves the baseline method. We notice the larger the box is, the better the performance of the proposed method is. It is because the human body dominates the statistics of PSNR and SSIM when the object is small. For novel action evaluation, Table 4.2 illustrates the proposed method considerably improves the baseline among different objects.

Qualitative comparisons The person usually occupies the main region of the interaction in the image. As a result, the proposed method does not improve the numbers significantly compared to the baseline for small objects if the person animation achieves good results as illustrated in Table C.3. Therefore, we further present the qualitative comparisons. Figure 4.3 shows our method can effectively animate the human-object interactions with the control of poses. Without the modeling of objects, the baseline method achieves poor performance on object rendering though it can still render the human body correctly. Meanwhile, we find the small object, e.g., “boxsmall”, only occupies a small region in the Human-Object Interaction images. Therefore, for those interactions, the PNSR and SSIM can not demonstrate the performance well.

TABLE 4.1. Human-Object Animation for the boxes, i.e., different sizes of objects.

Method	boxlarge		boxlong		boxmedium		boxsmall		boxtiny	
	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
TAVA [151]	22.6	0.949	26.8	0.966	25.9	0.967	26.8	0.970	27.5	0.973
CHONA (ours)	**27.2**	**0.971**	**28.1**	**0.974**	**28.5**	**0.976**	**28.0**	**0.974**	**28.3**	**0.976**

TABLE 4.2. Human-Object Animation under novel interactions.

Method	backpack		chairblack		chairwood		suitcase	
	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
TAVA [151]	27.9	0.960	28.3	0.959	26.0	0.960	28.9	0.964
CHONA (ours)	**28.4**	**0.971**	**29.1**	**0.971**	**27.3**	**0.969**	**29.4**	**0.974**

TABLE 4.3. Human-Object Animation under novel interactions.

Method	tablesmall		tablesquare		yogaball	
	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
TAVA [151]	25.7	0.965	22.8	0.943	24.6	0.950
CHONA (ours)	**27.9**	**0.974**	**27.9**	**0.966**	**28.2**	**0.974**

4.5.2 Compositional Animation

In this section, we provide experiments to evaluate the proposed method on the compositional Human-Object Animation. We first construct a compositional benchmark for compositional animation with two subjects (S01, S02) and nine objects, totally 18 combinations, from BEHAVE to construct a sub-dataset. Human-Object Interaction is composed of person, action, and object. There are at most three actions in BEHAVE. Therefore, the subset includes 37 combinations of $< person, action, object >$. Given a person, there are different novel compositions as follows,

- Novel Action, i.e., the combination of the object and the person exists in the training set, but the action is novel. This is similar to novel pose animation.
- Novel Object, i.e., there are no combinations of the person and the object in the training set, but the combination of the action and the object exists in the training set.
- Novel Action and Object, i.e., the combination of the action and the object does not exist in the training set.

TABLE 4.4. Compositional Human-Object Animation. The subscripts of $_a$, $_o$ and $_{ao}$ indicate the results of novel action set, novel object set, novel action-object set respectively. The baseline is CHONA without compositional invariant learning.

Method	PSNR$_a$	SSIM$_a$	PSNR$_o$	SSIM$_o$	PSNR$_{ao}$	SSIM$_{ao}$
Baseline	28.2	0.967	26.5	0.961	27.4	0.964
CC-NeRF	28.1	0.966	**27.0**	**0.966**	**28.0**	**0.968**

Then, we split the dataset into a training set and three validation sets, i.e., novel action validation, novel object validation and novel action-object validation. For similar objects, we treat the actions with the same name equally. For example, the action "sit" between "chairwood" and "chairblack" is treated equally. Then, we randomly select 13 (around 1/3) combinations as the training set, and split the remaining combinations into three novel categories according to the description above. There are 614 frames in the training set, we thus use two V100 GPUs to run our experiments with 300,000 iterations.

Quantitative Comparisons The proposed Compositional Conditional Radiance Fields effectively improves the rendering for both human and object as illustrated in Table 4.4. We notice CC-NeRF achieves similar performance to the network without compositional invariant learning on novel action/pose animation. However, for the novel object split and novel action-object split, The proposed method effectively decomposes the control of different people and objects, and thus illustrates better performance on the compositional animation.

Qualitative Comparisons Figure 4.4 demonstrates that the model without compositional invariant learning strategy fails to render the novel human-object interactions. We notice the head (the mask is missing) in the baseline is dissimilar to the ground truth, but more similar to another subject. The objects of the baseline become red due to the entangling of the two latent codes. Those cases indicate the proposed Compositional Conditional NeRF effectively decomposes the latent codes and achieves compositional animation.

FIGURE 4.4. Visualized Comparisons between Compositional Conditional Radiance Fields and baseline method (w/o compositional invariant learning). The first row is the baseline, the second row is the proposed method, while the last row is the Ground Truth. The first three columns indicate the novel object categories, and the last three columns show the novel action-object categories.

4.5.3 Novel Person and Object

To demonstrate the effectiveness of the proposed method on novel object and novel person without interactions, we leverage the person (386) in ZJU-mocap and object bowl in CO3D to jointly train with BEHAVE. Here, for the dataset in BEHAVE, we directly adopt the subset in Section 4.5.2. Given that we do not have the ground truth for this experiment, we show the results qualitatively in Figure 4.5. We observe the Compositional Conditional NeRF significantly improves the animation for novel object and person as illustrated in Figure 4.5. We find the baseline method will incur wired color on the object (e.g., the suitcase) or render the color of the novel person into the object (e.g., the "tablesquare" is green). Besides, for the novel object, we observe the baseline without compositional invariant learning fails to render the human body. More visualized demonstration with additional objects and persons is provided in Appendix.

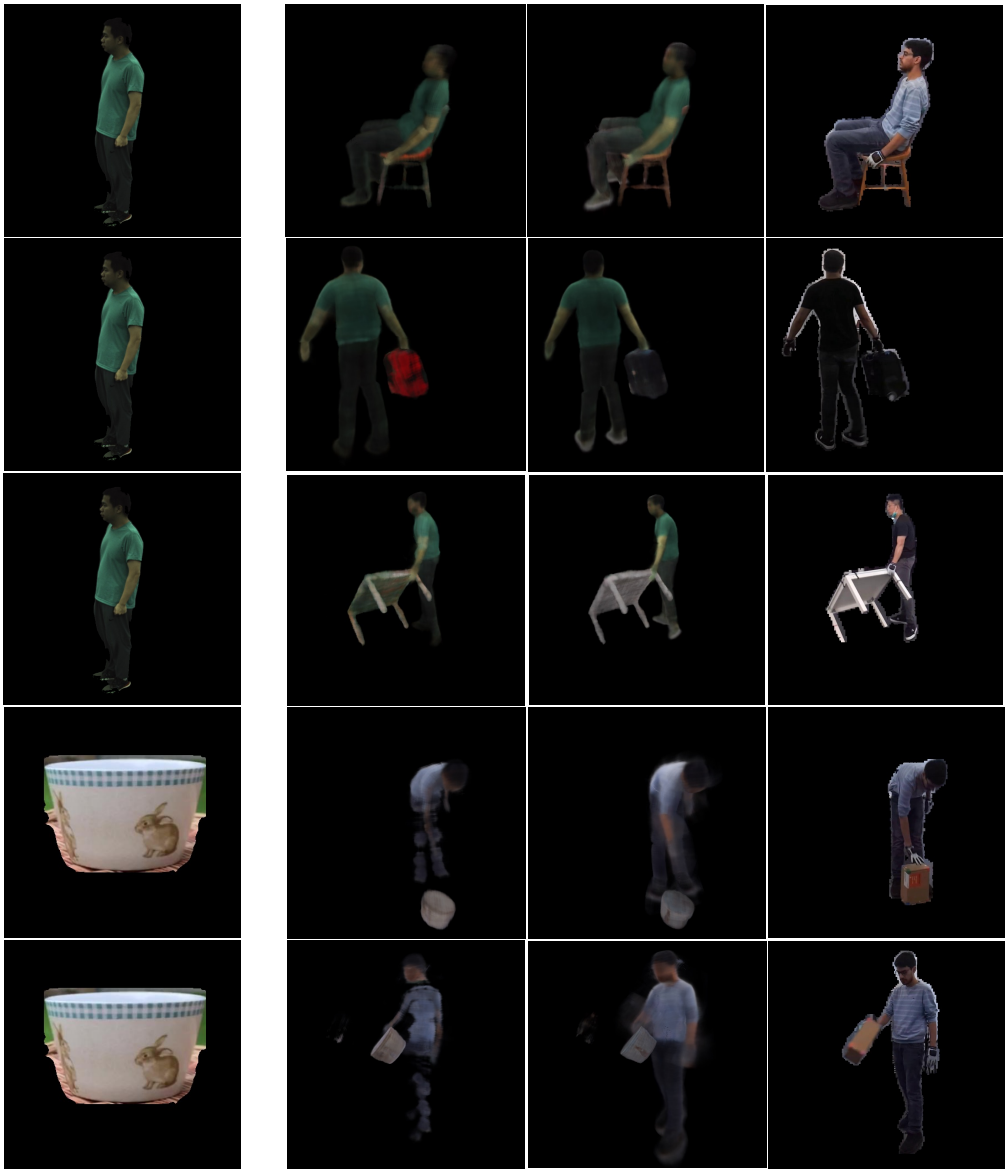

FIGURE 4.5. Visualized Comparisons between Compositional Conditional Radiance Fields and baseline (without compositional invariant learning) on novel object and person animation. The first column is novel/object, the second column is the baseline method, the third column is the proposed method, and the last column is the given pose. The first three rows indicate the novel person animation, while the last two rows show the novel object animation.

4.6 Summary

In this thesis, we address the challenge of compositional human-object animation via neural Human-Object skinning deformations and compositional conditional radiance fields. Specifically, we construct a pseudo bone for the object, and devise a human-object skinning deforming approach to model the interactions between human and object. Moreover, to enable compositional Human-Object animation, we further present compositional conditional neural radiance fields, which decompose the human and object latent codes via compositional invariant learning, to compositionally control the animation for novel human-object combinations, and even novel person and objects. Comprehensive experiments demonstrate the proposed method significantly improves the animation performance, as well as the compositional generalization.

Though we achieve considerable performance with the proposed methods, there still are some challenges in the human-object neural animations, e.g., the interactions with non-rigid objects. We think we can extend the deformation module by designing multiple pseudo bones for the non-rigid object, which we leave to future work. There is also another challenge that is how to understand the interaction region (i.e., affordance region of the object). As human interacts with similar objects in a similar way, we think we can make use of the similarity of affordances among similar objects for the affordance localization to novel objects in the future.

CHAPTER 5

Conclusion

---

hierarchical relationship understanding is one of the most challenges in visual perception and the visual scenes are usually layout in a hierarchical and compositional way. Compositionality and Hierarchy are of importance for the visual scene understanding. However, the compositionality and hierarchy, especially the knowledge transfer among different concepts, are poorly investigated by previous approaches. The thesis focuses on learning transferable representations for the hierarchical relationship exploration from the hierarchical and compositional perspective. In this thesis, we first propose a visual compositional learning framework to facilitate the compositional generalization for HOI detection. Then, we introduce a transfer learning approach to transfer the verb/affordance representations to novel objects for reasoning object categories. Next, we devise a self-compositional learning framework with a pseudo-label strategy to reason novel possible HOI categories. Except for the existing visual relationship reasoning, the thesis further devises a simple yet effective module, Batch Transformer or BatchFormer, to implicitly explore the sample relationships in the penultimate layer for robust representation learning. We further extend the idea of BatchFormer into vision transformer networks, and achieve consistent improvement among different DETR-based methods. Lastly, the thesis presents a novel 3D compositional human-object animation to explore the 3D geometry and animation for the hierarchical visual relationship understanding with a neural human-object deformation and compositional invariant learning strategy. Overall, the content of this thesis can be categorized as follows,

Chapter 3 comprehensively explores the compositional learning approach for Human-Object Interaction understanding. This chapter first presents a new visual compositional learning framework, which first decomposes the verb and object representation according to bounding

boxes, and then stitches the verb and object representations among pair-wise images to generate composite HOI features for end-to-end optimization. The visual compositional learning framework effectively improves Long-tailed HOI detection and compositional zero-shot HOI detection. To address the open long-tailed HOI detection, this chapter further introduces an object representation fabricator to balance the distribution, named as fabricated compositional learning, to significantly improve the few-shot and compositional zero-shot HOI detection. Next, a transfer learning framework is introduced to transfer the object affordance representation to novel objects, and enables the HOI model for human-novel-object interaction detection and object affordance recognition. Lastly, this chapter presents a self-training strategy to build pseudo labels from the online concept discovery to facilitate compositional learning, and thus significantly improve concept discovery and object affordance recognition.

Chapter 4 mainly investigates the sample relationship exploration for representation learning. While previous approaches implicitly explore the sample relationships from a perspective of either the input or the loss function, we introduce a batch transformer, BatchFormerV1, to equip the deep neural networks themselves with the ability to explore the sample relationships in a learnable way. With shared classifier strategy, the module can consistently improve those data-scarcity tasks without incurring any computational budget during inference. BatchFormerV2 module is introduced to explore the sample relationships for pixel-/patch-level dense representations. Meanwhile, BatchFormerV2 module is shared among the spatial positions in the network, and a two-stream pipeline is introduced for achieving batch-invariant learning for dense prediction tasks. Extensive experiments demonstrate the effectiveness for data-scarcity tasks and visual recognition tasks ranging from image classification to object detection and panoptic segmentation.

Chapter 5 focuses on the 3D Compositioanl human-object animation (CHONA) based on the neural radiance fields. The chapter first introduces a neural human-object deformation for achieving pose-driven human-object interaction animation. Considering the limitation of collecting enough interaction poses, a compositional conditional NeRF or CC-NeRF is introduced to enable the transfer of poses among different person and objects, and thus compositionally control the animation for human-object interaction under novel poses, novel person and novel object. Extensive experiments demonstrate the proposed method for 3D

Human-Object Animation under novel poses, novel person, novel objects and even non-interactive person and static objects.

5.1 Future outlook

Hierarchical relationship understanding is significant for the intelligent agent to achieve the ability of human-level perception and reasoning. There are still extensive challenging directions that are worth exploring in the future and I list a few directions here.

- **3D Compositions:** With the gradual slowdown of 2D computer vision, the community has show great interest in 3D vision. Compared to 2D semantic understanding, 3D vision consists of more compositionality and hierarchy challenges. First of all, the scene in the 3D vision is composed of multiple objects that are spatially distributed in the scene and can be redistributed into new scenes. Compared to 2D vision, in which we can usually collect enough images, 3D vision might suffer from the data scarcity tasks more seriously. Exploring compositionality and hierarchy is beneficial for understanding the 3D world. Meanwhile, the deformation reasoning is also interesting when we combine different parts or objects in the new scene.
- **Hierarchical Neural Architectures:** While current deep neural networks are internally built in a hierarchical way, there is still valuable potential to redesign the basic architectures from the perspective of hierarchy and composition. Current deep neural networks highly rely on the large scale of data, and usually achieve poor performances in those data scarcity tasks, regardless of rare class samples and rare domain samples. However, the class correlations and sample correlations are crucial for addressing the data scarcity challenges. Though the deep neural network is hierarchically built, the current architectures do not consider the hierarchical relations for the network design. Therefore, it is valuable to design a new architecture that is internally useful for data scarcity tasks and simultaneously achieves good interpretability from the perspective of hierarchy.

- **Compositional Theory:** Though compositionality and hierarchy are ubiquitous in the visual world, the community lacks the corresponding theory for generalization. Currently, there are several benchmarks for compositional generalization, e.g., compositional zero-shot learning. Those approaches mainly focus on proposing novel approaches to improve compositional generalization. However, due to the correlation (e.g. shared elements) between the training set and test set in compositional generalization settings, there might be possible to infer the generalization boundary according to the side of the training set.

Bibliography

[1] Amit Alfassy et al. 'LaSO: Label-Set Operations networks for multi-label few-shot learning'. In: *CVPR*. 2019, pp. 6548–6557.

[2] Mykhaylo Andriluka and Leonid Sigal. 'Human Context: Modeling human-human interactions for monocular 3D pose estimation'. In: *International Conference on Articulated Motion and Deformable Objects*. Springer. 2012, pp. 260–272.

[3] Dragomir Anguelov et al. 'Scape: shape completion and animation of people'. In: *ACM SIGGRAPH 2005 Papers*. 2005, pp. 408–416.

[4] Martin Arjovsky et al. 'Invariant risk minimization'. In: *arXiv preprint arXiv:1907.02893* (2019).

[5] Yuval Atzmon et al. 'A causal view of compositional zero-shot recognition'. In: *NeurIPS*. 2020.

[6] Samaneh Azadi et al. 'Compositional gan: Learning conditional image composition'. In: *IJCV* (2020).

[7] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 'Neural Machine Translation by Jointly Learning to Align and Translate'. In: *arXiv preprint arXiv:1409.0473* (2014).

[8] Yogesh Balaji, Swami Sankaranarayanan and Rama Chellappa. 'Metareg: Towards domain generalization using meta-regularization'. In: *NIPS*. Vol. 31. 2018, pp. 998–1008.

[9] Ankan Bansal et al. 'Detecting Human-Object Interactions via Functional Generalization'. In: *AAAI*. 2020.

[10] Ankan Bansal et al. 'Detecting Human-Object Interactions via Functional Generalization'. In: *AAAI* (2020).

[11] Hangbo Bao, Li Dong and Furu Wei. 'Beit: Bert Pre-training of Image Transformers'. In: *arXiv preprint arXiv:2106.08254* (2021).

[12] Jonathan T. Barron et al. 'Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields'. en. In: *CVPR*. IEEE, June 2022, pp. 5460–5469. ISBN: 978-1-66546-946-3. (Visited on 15/10/2022).

[13] Peter W Battaglia et al. 'Relational inductive biases, deep learning, and graph networks'. In: *arXiv preprint arXiv:1806.01261* (2018).

[14] Sara Beery, Grant Van Horn and Pietro Perona. 'Recognition in terra incognita'. In: *ECCV*. 2018, pp. 456–473.

[15] Yoshua Bengio, Aaron Courville and Pascal Vincent. 'Representation learning: A review and new perspectives'. In: *IEEE PAMI* 35.8 (2013), pp. 1798–1828.

[16] Bharat Lal Bhatnagar et al. 'Behave: Dataset and method for tracking human object interactions'. In: *CVPR*. 2022, pp. 15935–15946.

[17] Irving Biederman. 'Recognition-by-components: a theory of human image understanding.' In: *Psychological review* 94.2 (1987), p. 115.

[18] Gilles Blanchard, Gyemin Lee and Clayton Scott. 'Generalizing from several related classification tasks to a new unlabeled sample'. In: *NIPS* 24 (2011), pp. 2178–2186.

[19] Christopher P Burgess et al. 'Monet: Unsupervised scene decomposition and representation'. In: *arXiv preprint arXiv:1901.11390* (2019).

[20] Jonathon Byrd and Zachary Lipton. 'What is the effect of importance weighting in deep learning?' In: *ICML*. PMLR. 2019, pp. 872–881.

[21] Jiarui Cai, Yizhou Wang and Jenq-Neng Hwang. 'ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot'. In: *ICCV*. 2021.

[22] Kaidi Cao et al. 'Learning imbalanced datasets with label-distribution-aware margin loss'. In: *NeurIPS* (2019).

[23] Nicolas Carion et al. 'End-to-end object detection with transformers'. In: *ECCV*. Springer. 2020, pp. 213–229.

[24] Joao Carreira and Andrew Zisserman. 'Quo vadis, action recognition? a new model and the kinetics dataset'. In: *CVPR*. 2017, pp. 6299–6308.

[25] Junbum Cha et al. ‘SWAD: Domain Generalization by Seeking Flat Minima’. In: *NeurIPS*. 2021.

[26] Yu-Wei Chao et al. ‘Hico: A benchmark for recognizing human-object interactions in images’. In: *ICCV*. 2015, pp. 1017–1025.

[27] Yu-Wei Chao et al. ‘Hico: A benchmark for recognizing human-object interactions in images’. In: *ICCV*. 2015, pp. 1017–1025.

[28] Yu-Wei Chao et al. ‘Learning to detect human-object interactions’. In: *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE. 2018, pp. 381–389.

[29] Wei-Lun Chao et al. ‘An empirical study and analysis of generalized zero-shot learning for object recognition in the wild’. In: *ECCV*. Springer. 2016, pp. 52–68.

[30] Prithvijit Chattopadhyay, Yogesh Balaji and Judy Hoffman. ‘Learning to balance specificity and invariance for in and out of domain generalization’. In: *ECCV*. Springer. 2020, pp. 301–318.

[31] Nitesh V Chawla et al. ‘SMOTE: synthetic minority over-sampling technique’. In: *JAIR* 16 (2002), pp. 321–357.

[32] Chun-Fu Chen, Rameswar Panda and Quanfu Fan. ‘RegionViT: Regional-to-Local Attention for Vision Transformers’. In: *ICLR*. 2022.

[33] Mingfei Chen et al. ‘Reformulating hoi detection as adaptive set prediction’. In: *CVPR*. 2021, pp. 9004–9013.

[34] Xinlei Chen, Saining Xie and Kaiming He. ‘An Empirical Study of Training Self-supervised Vision Transformers’. In: *ICCV*. 2021, pp. 9640–9649.

[35] Xinlei Chen et al. ‘Improved Baselines with Momentum Contrastive Learning’. In: *arXiv preprint arXiv:2003.04297* (2020).

[36] Xinyun Chen et al. ‘Compositional generalization via neural-symbolic stack machines’. In: *NeurIPS* 33 (2020), pp. 1690–1701.

[37] Xu Chen et al. ‘SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes’. In: *ICCV*. 2021, pp. 11594–11604.

[38] Yixin Chen et al. 'Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense'. In: *ICCV*. 2019, pp. 8648–8657.

[39] Zhiqin Chen and Hao Zhang. 'Learning implicit fields for generative shape modeling'. In: *CVPR*. 2019, pp. 5939–5948.

[40] Xinlei Chen*, Saining Xie* and Kaiming He. 'An Empirical Study of Training Self-Supervised Vision Transformers'. In: *CVPR*. 2021.

[41] Bowen Cheng, Alex Schwing and Alexander Kirillov. 'Per-Pixel Classification is Not All You Need for Semantic Segmentation'. In: *NeurIPS* 34 (2021).

[42] Noam Chomsky. *Aspects of the Theory of Syntax*. Vol. 11. MIT press, 2014.

[43] Xiangxiang Chu et al. 'Twins: Revisiting the Design of Spatial Attention in Vision Transformers'. In: *NeurIPS*. 2021.

[44] Jiequan Cui et al. 'Parametric Contrastive Learning'. In: *ICCV*. 2021.

[45] Yin Cui et al. 'Class-balanced loss based on effective number of samples'. In: *CVPR*. 2019, pp. 9268–9277.

[46] Rishabh Dabral et al. 'Gravity-Aware Monocular 3D Human-Object Reconstruction'. en. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 12345–12354. ISBN: 978-1-66542-812-5. (Visited on 16/10/2022).

[47] Rishabh Dabral et al. 'Gravity-aware monocular 3d human-object reconstruction'. In: *ICCV*. 2021, pp. 12365–12374.

[48] Dima Damen et al. 'Scaling egocentric vision: The epic-kitchens dataset'. In: *ECCV*. 2018, pp. 720–736.

[49] Francesco De Comité et al. 'Positive and unlabeled examples help learning'. In: *International Conference on Algorithmic Learning Theory*. Springer. 1999, pp. 219–230.

[50] Jia Deng et al. 'Imagenet: A large-scale hierarchical image database'. In: *CVPR*. Ieee. 2009, pp. 248–255.

[51] Shengheng Deng et al. '3d affordancenet: A benchmark for visual object affordance understanding'. In: *CVPR*. 2021, pp. 1778–1787.

[52] Jacob Devlin et al. ‘Bert: Pre-training of deep bidirectional transformers for language understanding’. In: *arXiv preprint arXiv:1810.04805* (2018).

[53] Alexey Dosovitskiy et al. ‘An image is worth 16x16 words: Transformers for image recognition at scale’. In: *ICLR*. 2020.

[54] Mingsong Dou et al. ‘Fusion4d: Real-time performance capture of challenging scenes’. In: *ACM Transactions on Graphics* 35.4 (2016), pp. 1–13.

[55] Mingsong Dou et al. ‘Motion2fusion: real-time volumetric performance capture’. en. In: *ACM Transactions on Graphics* 36.6 (Nov. 2017), pp. 1–16. ISSN: 0730-0301, 1557-7368. (Visited on 15/10/2022).

[56] Charles Elkan and Keith Noto. ‘Learning classifiers from only positive and unlabeled data’. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. 2008, pp. 213–220.

[57] Mark Everingham et al. ‘The pascal visual object classes (voc) challenge’. In: *IJCV* 88.2 (2010), pp. 303–338.

[58] Chen Fang, Ye Xu and Daniel N Rockmore. ‘Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias’. In: *ICCV*. 2013, pp. 1657–1664.

[59] Hao-Shu Fang et al. ‘Pairwise body-part attention for recognizing human-object interactions’. In: *ECCV*. 2018, pp. 51–67.

[60] Kuan Fang et al. ‘Demo2Vec: Reasoning Object Affordances from Online Videos’. In: *CVPR*. 2018.

[61] Li Fei-Fei, Rob Fergus and Pietro Perona. ‘One-shot learning of object categories’. In: *IEEE TPAMI* 28.4 (2006), pp. 594–611.

[62] Mihai Fieraru et al. ‘Three-dimensional reconstruction of human interactions’. In: *CVPR*. 2020, pp. 7214–7223.

[63] David F. Fouhey et al. ‘People Watching: Human Actions as a Cue for Single View Geometry’. In: *IJCV* 110 (2014), pp. 259–274.

[64] John Freeman. ‘The modelling of spatial relations’. In: *Computer Graphics and Image Processing* 4.2 (1975), pp. 156–171. ISSN: 0146-664X. DOI: https://

doi.org/10.1016/S0146-664X(75)80007-4. URL: https://www.sciencedirect.com/science/article/pii/S0146664X75800074.

[65] Andrea Frome et al. 'Devise: A deep visual-semantic embedding model'. In: *NIPS*. 2013, pp. 2121–2129.

[66] Chen Gao, Yuliang Zou and Jia-Bin Huang. 'ican: Instance-centric attention network for human-object interaction detection'. In: *arXiv preprint arXiv:1808.10437* (2018).

[67] Chen Gao et al. 'DRG: Dual Relation Graph for Human-Object Interaction Detection'. In: *ECCV*. 2020.

[68] Peng Gao et al. 'Fast Convergence of DETR with Spatially Modulated Co-Attention'. In: *ICCV*. 2021, pp. 3621–3630.

[69] Marta Garnelo and Murray Shanahan. 'Reconciling deep learning with symbolic artificial intelligence: representing objects and relations'. In: *Current Opinion in Behavioral Sciences* 29 (2019), pp. 17–23.

[70] Spandana Gella, Frank Keller and Mirella Lapata. 'Disambiguating visual verbs'. In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2017), pp. 311–322.

[71] Golnaz Ghiasi et al. 'Simple copy-paste is a strong data augmentation method for instance segmentation'. In: *CVPR*. 2021, pp. 2918–2928.

[72] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[73] James J. Gibson. 'The Ecological Approach to Visual Perception'. In: (1979).

[74] Justin Gilmer et al. 'Neural message passing for quantum chemistry'. In: *ICML*. PMLR. 2017, pp. 1263–1272.

[75] Georgia Gkioxari et al. 'Detecting and recognizing human-object interactions'. In: *CVPR*. 2018, pp. 8359–8367.

[76] Ian Goodfellow et al. 'Generative adversarial nets'. In: *NIPS*. 2014, pp. 2672–2680.

[77] Chunhui Gu et al. 'Ava: A video dataset of spatio-temporally localized atomic visual actions'. In: *CVPR*. 2018, pp. 6047–6056.

[78] Abhinav Gupta, Aniruddha Kembhavi and Larry S Davis. 'Observing human-object interactions: Using spatial and functional compatibility for recognition'. In: *IEEE*

Transactions on Pattern Analysis and Machine Intelligence 31.10 (2009), pp. 1775–1789.

[79] Agrim Gupta, Piotr Dollar and Ross Girshick. 'LVIS: A dataset for large vocabulary instance segmentation'. In: *CVPR*. 2019, pp. 5356–5364.

[80] Saurabh Gupta and Jitendra Malik. 'Visual semantic role labeling'. In: *arXiv preprint arXiv:1505.04474* (2015).

[81] Tanmay Gupta, Alexander Schwing and Derek Hoiem. 'No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques'. In: *ICCV*. 2019, pp. 9677–9685.

[82] Vladimir Guzov et al. 'Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors'. en. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 4316–4327. ISBN: 978-1-66544-509-2. (Visited on 16/10/2022).

[83] Hui Han, Wen-Yuan Wang and Bing-Huan Mao. 'Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning'. In: *ICIC*. Springer. 2005, pp. 878–887.

[84] Sanjay Haresh et al. *Articulated 3D Human-Object Interactions from RGB Videos: An Empirical Analysis of Approaches and Challenges*. en. arXiv:2209.05612 [cs]. Sept. 2022. (Visited on 14/10/2022).

[85] Sanjay Haresh et al. 'Articulated 3D Human-Object Interactions from RGB Videos: An Empirical Analysis of Approaches and Challenges'. In: *arXiv preprint arXiv:2209.05612* (2022).

[86] Bharath Hariharan and Ross Girshick. 'Low-shot visual recognition by shrinking and hallucinating features'. In: *ICCV*. 2017, pp. 3018–3027.

[87] Mahmudul Hassan and Anuja Dharmaratne. 'Attribute based affordance detection from human-object interaction images'. In: *Image and Video Technology*. Springer. 2015, pp. 220–232.

[88] Mohamed Hassan et al. 'Populating 3D Scenes by Learning Human-Scene Interaction'. en. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR). IEEE, June 2021, pp. 14703–14713. ISBN: 978-1-66544-509-2. (Visited on 16/10/2022).

[89] Mohamed Hassan et al. 'Populating 3D scenes by learning human-scene interaction'. In: *CVPR*. 2021, pp. 14708–14718.

[90] Mohamed Hassan et al. 'Stochastic scene-aware motion prediction'. In: *ICCV*. 2021, pp. 11374–11384.

[91] Mohammed Hassanin, Salman Khan and Murat Tahtali. 'Visual affordance and function understanding: A survey'. In: *ACM Computing Surveys (CSUR)* 54.3 (2021), pp. 1–35.

[92] Munawar Hayat et al. 'Gaussian affinity for max-margin class imbalanced learning'. In: *ICCV*. 2019, pp. 6469–6479.

[93] Haibo He and Edwardo A Garcia. 'Learning from imbalanced data'. In: *TKDE* 21.9 (2009), pp. 1263–1284.

[94] Ju He, Adam Kortylewski and Alan Yuille. 'COMPAS: Representation Learning with Compositional Part Sharing for Few-Shot Classification'. In: *arXiv preprint arXiv:2101.11878* (2021).

[95] Kaiming He et al. 'Deep Residual Learning for Image Recognition'. In: *CVPR*. 2016.

[96] Kaiming He et al. 'Masked Autoencoders are Scalable Vision Learners'. In: *arXiv preprint arXiv:2111.06377* (2021).

[97] Kaiming He et al. 'Momentum Contrast for Unsupervised Visual Representation Learning'. In: *CVPR*. 2020.

[98] Kaiming He et al. 'Momentum Contrast for Unsupervised Visual Representation Learning'. In: *CVPR*. 2020, pp. 9729–9738.

[99] Yin-Yin He, Jianxin Wu and Xiu-Shen Wei. 'Distilling Virtual Examples for Long-tailed Recognition'. In: *ICCV*. 2021.

[100] Irina Higgins et al. 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.' In: *ICLR*. Vol. 2. 5. 2017, p. 6.

[101] Irina Higgins et al. 'Scan: Learning hierarchical compositional visual concepts'. In: *ICLR*. 2018.

[102] Geoffrey Hinton, Oriol Vinyals and Jeff Dean. 'Distilling the knowledge in a neural network'. In: *arXiv preprint arXiv:1503.02531* (2015).

[103] Donald D Hoffman and Whitman A Richards. 'Parts of recognition'. In: *Cognition* 18.1-3 (1984), pp. 65–96.

[104] Richang Hong et al. 'Learning visual semantic relationships for efficient visual retrieval'. In: *IEEE Transactions on Big Data* 1.4 (2015), pp. 152–161.

[105] Ruibing Hou et al. 'Cross attention network for few-shot classification'. In: *NeurIPS* (2019).

[106] Zhi Hou, Baosheng Yu and Dacheng Tao. 'BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning'. In: *CVPR*. 2022.

[107] Zhi Hou et al. 'Affordance Transfer Learning for Human-Object Interaction Detection'. In: *CVPR*. 2021.

[108] Zhi Hou et al. 'Affordance transfer learning for human-object interaction detection'. In: *CVPR*. 2021, pp. 495–504.

[109] Zhi Hou et al. 'BatchFormerV2: Exploring Sample Relationships for Dense Representation Learning'. In: *arXiv preprint arXiv:2204.01254* (2022).

[110] Zhi Hou et al. 'Detecting human-object interaction via fabricated compositional learning'. In: *CVPR*. 2021, pp. 14646–14655.

[111] Zhi Hou et al. 'Visual compositional learning for human-object interaction detection'. In: *ECCV*. Springer. 2020, pp. 584–600.

[112] Jie Hu, Li Shen and Gang Sun. 'Squeeze-and-Excitation Networks'. In: *CVPR*. 2018.

[113] Jie Hu, Li Shen and Gang Sun. 'Squeeze-and-Excitation Networks'. In: *CVPR*. 2018, pp. 7132–7141.

[114] Chen Huang et al. 'Learning deep representation for imbalanced classification'. In: *CVPR*. 2016, pp. 5375–5384.

[115] Yinghao Huang et al. 'InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction'. In: *DAGM German Conference on Pattern Recognition*. Springer. 2022, pp. 281–299.

[116] Zeyi Huang et al. 'Self-challenging improves cross-domain generalization'. In: *ECCV*. Springer. 2020, pp. 124–140.

[117] Dat Huynh and Ehsan Elhamifar. 'Interaction Compass: Multi-Label Zero-Shot Learning of Human-Object Interactions via Spatial Relations'. In: *ICCV*. 2021, pp. 8472–8483.

[118] Sergey Ioffe and Christian Szegedy. 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *ICML*. PMLR. 2015, pp. 448–456.

[119] Phillip Isola, Joseph J Lim and Edward H Adelson. 'Discovering states and transformations in image collections'. In: *CVPR*. 2015, pp. 1383–1391.

[120] Sumit Jain and C. Karen Liu. 'Interactive synthesis of human-object interaction'. en. In: *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09*. ACM Press, 2009, p. 47. ISBN: 978-1-60558-610-6. (Visited on 17/10/2022).

[121] Muhammad Abdullah Jamal et al. 'Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective'. In: *CVPR*. 2020, pp. 7610–7619.

[122] Nathalie Japkowicz and Shaju Stephen. 'The class imbalance problem: A systematic study'. In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[123] Jingwei Ji, Rishi Desai and Juan Carlos Niebles. 'Detecting Human-Object Relationships in Videos'. In: *ICCV*. 2021, pp. 8106–8116.

[124] Junguang Jiang et al. *Transfer-Learning-library*. https://github.com/thuml/Transfer-Learning-Library. 2020.

[125] Yuheng Jiang et al. 'NeuralHOFusion: Neural Volumetric Rendering under Human-object Interactions'. en. In: *CVPR*. IEEE, June 2022, pp. 6145–6155. ISBN: 978-1-66546-946-3. (Visited on 14/10/2022).

[126] Bingyi Kang et al. 'Decoupling representation and classifier for long-tailed recognition'. In: *ICLR* (2020).

[127] Shyamgopal Karthik, Massimiliano Mancini and Zeynep Akata. 'KG-SP: Knowledge Guided Simple Primitives for Open World Compositional Zero-Shot Learning'. In: *CVPR*. 2022, pp. 9336–9345.

[128] Keizo Kato, Yin Li and Abhinav Gupta. 'Compositional learning for human object interaction'. In: *ECCV*. 2018, pp. 234–251.

[129] Rohit Keshari, Richa Singh and Mayank Vatsa. 'Generalized Zero-Shot Learning Via Over-Complete Distribution'. In: *CVPR*. 2020.

[130] Daniel Keysers et al. 'Measuring compositional generalization: A comprehensive method on realistic data'. In: *arXiv preprint arXiv:1912.09713* (2019).

[131] Bumsoo Kim et al. 'HOTR: End-to-End Human-Object Interaction Detection with Transformers'. In: *CVPR*. 2021, pp. 74–83.

[132] Najoung Kim and Tal Linzen. 'COGS: A compositional generalization challenge based on semantic interpretation'. In: *arXiv preprint arXiv:2010.05465* (2020).

[133] Vladimir G Kim et al. 'Shape2pose: Human-centric shape analysis'. In: *ACM Transactions on Graphics* 33.4 (2014), pp. 1–12.

[134] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980* (2014).

[135] Diederik P Kingma and Max Welling. 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114* (2013).

[136] Diederik P Kingma et al. 'Semi-supervised Learning with Deep Generative Models'. In: *NIPS*. 2014.

[137] Alexander Kirillov et al. 'Panoptic Segmentation'. In: *CVPR*. 2019, pp. 9404–9413.

[138] Hedvig Kjellström, Javier Romero and Danica Kragić. 'Visual object-action recognition: Inferring object affordances from human demonstration'. In: *Computer Vision and Image Understanding* 115.1 (2011), pp. 81–90.

[139] Jannik Kossen et al. 'Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning'. In: *NeurIPS* 34 (2021).

[140] Alex Krizhevsky, Geoffrey Hinton et al. 'Learning multiple layers of features from tiny images'. In: (2009).

[141] David Krueger et al. 'Out-of-Distribution Generalization via Risk Extrapolation (REx)'. In: *ICML*. 2021.

[142] Vinay Kumar Verma et al. 'Generalized zero-shot learning via synthesized examples'. In: *CVPR*. 2018, pp. 4281–4289.

[143] Youngjoong Kwon et al. ‘Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering’. In: *NeurIPS*. Vol. 34. Curran Associates, Inc., 2021, pp. 24741–24752. (Visited on 15/10/2022).

[144] Brenden Lake and Marco Baroni. ‘Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks’. In: *ICML*. PMLR. 2018, pp. 2873–2882.

[145] Brenden M Lake et al. ‘Building machines that learn and think like people’. In: *Behavioral and brain sciences* 40 (2017).

[146] Christoph H Lampert, Hannes Nickisch and Stefan Harmeling. ‘Learning to detect unseen object classes by between-class attribute transfer’. In: *CVPR*. IEEE. 2009, pp. 951–958.

[147] Yann LeCun, Yoshua Bengio et al. ‘Convolutional Networks for Images, Speech, and Time Series’. In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[148] Dong-Hyun Lee et al. ‘Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks’. In: *Workshop on challenges in representation learning, ICML*. 2013.

[149] Da Li et al. ‘Deeper, broader and artier domain generalization’. In: *ICCV*. 2017, pp. 5542–5550.

[150] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV].

[151] Ruilong Li et al. ‘TAVA: Template-free animatable volumetric actors’. In: 2022.

[152] Shuang Li et al. ‘MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition’. In: *CVPR*. 2021, pp. 5212–5221.

[153] Xiangyu Li et al. ‘Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning’. In: *CVPR*. 2022, pp. 9326–9335.

[154] Yong-Lu Li et al. ‘HOI Analysis: Integrating and Decomposing Human-Object Interaction’. In: vol. 33. 2020.

[155] Yong-Lu Li et al. ‘PaStaNet: Toward Human Activity Knowledge Engine’. In: *CVPR*. 2020, pp. 382–391.

[156] Yong-Lu Li et al. ‘Symmetry and group in attribute-object compositions’. In: *CVPR*. 2020, pp. 11316–11325.

[157] Yong-Lu Li et al. ‘Transferable interactiveness prior for human-object interaction detection’. In: *CVPR*. 2019.

[158] Zhe Li et al. ‘AvatarCap: Animatable Avatar Conditioned Monocular Human Volumetric Capture’. In: *ECCV*. 2022.

[159] Zhengqi Li et al. ‘Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes’. en. In: *CVPR*. IEEE, June 2021, pp. 6494–6504. ISBN: 978-1-66544-509-2. (Visited on 15/10/2022).

[160] Zongmian Li et al. ‘Estimating 3d motion and forces of person-object interactions from monocular video’. In: *CVPR*. 2019, pp. 8640–8649.

[161] Yue Liao et al. ‘PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection’. In: *CVPR*. 2020.

[162] Kevin Lin, Lijuan Wang and Zicheng Liu. ‘Mesh Graphormer’. In: *ICCV*. 2021.

[163] Tsung-Yi Lin et al. ‘Focal loss for dense object detection’. In: *CVPR*. 2017, pp. 2980–2988.

[164] Tsung-Yi Lin et al. ‘Microsoft Coco: Common Objects in Context’. In: *ECCV*. Springer. 2014, pp. 740–755.

[165] Lingjie Liu et al. ‘Neural actor: Neural free-view synthesis of human actors with pose control’. In: *ACM Transactions on Graphics* 40.6 (2021), pp. 1–16.

[166] Qian Liu et al. ‘Compositional generalization by learning analytical expressions’. In: *NeurIPS* 33 (2020), pp. 11416–11427.

[167] Steven Liu et al. ‘Editing conditional radiance fields’. In: *ICCV*. 2021, pp. 5773–5783.

[168] Yanbin Liu et al. ‘Learning to propagate labels: Transductive propagation network for few-shot learning’. In: *ICLR*. 2018.

[169] Yunze Liu et al. ‘HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction’. In: *CVPR*. 2022, pp. 21013–21022.

[170] Ze Liu et al. ‘Swin Transformer: Hierarchical Vision Transformer using Shifted Windows’. In: *ICCV*. 2021.

[171] Ziwei Liu et al. ‘Large-Scale Long-Tailed Recognition in an Open World’. In: *CVPR*. 2019.

[172] Francesco Locatello et al. ‘Challenging common assumptions in the unsupervised learning of disentangled representations’. In: *ICML*. 2019.

[173] Jonathan Long, Evan Shelhamer and Trevor Darrell. ‘Fully Convolutional Networks for Semantic Segmentation’. In: *CVPR*. 2015, pp. 3431–3440.

[174] Matthew Loper et al. ‘SMPL: A skinned multi-person linear model’. In: *ACM transactions on graphics* 34.6 (2015), pp. 1–16.

[175] Cewu Lu et al. ‘Visual relationship detection with language priors’. In: *European conference on computer vision*. Springer. 2016, pp. 852–869.

[176] Xiaojian Ma et al. ‘Relvit: Concept-guided vision transformer for visual relational reasoning’. In: *arXiv preprint arXiv:2204.11167* (2022).

[177] Laurens van der Maaten and Geoffrey Hinton. ‘Visualizing data using t-SNE’. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[178] Massimiliano Mancini et al. ‘Open World Compositional Zero-Shot Learning’. In: *CVPR*. 2021, pp. 5222–5230.

[179] David M. Mark and Max J. Egenhofer. ‘Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing’. In: *Cartography and Geographic Information Systems* 21.4 (1994), pp. 195–212.

[180] Joanna Materzynska et al. ‘Something-else: Compositional action recognition with spatial-temporal interaction networks’. In: *CVPR*. 2020, pp. 1049–1059.

[181] Depu Meng et al. ‘Conditional DETR for Fast Training Convergence’. In: *ICCV*. 2021.

[182] Lars Mescheder et al. ‘Occupancy networks: Learning 3d reconstruction in function space’. In: *CVPR*. 2019, pp. 4460–4470.

[183] Ben Mildenhall et al. ‘Nerf: Representing scenes as neural radiance fields for view synthesis’. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[184] Arnab Kumar Mondal, Vineet Jain and Kaleem Siddiqi. ‘Mini-batch graphs for robust image classification’. In: *arXiv preprint arXiv:2105.03237* (2021).

[185] MF Naeem et al. ‘Learning Graph Embeddings for Compositional Zero-shot Learning’. In: *CVPR*. IEEE. 2021.

[186] Tushar Nagarajan and Kristen Grauman. ‘Learning affordance landscapes for interaction exploration in 3d environments’. In: *NeurIPS* 33 (2020), pp. 2005–2015.

[187] Megha Nawhal et al. ‘Generating videos of zero-shot compositions of actions and objects’. In: *ECCV*. Springer. 2020, pp. 382–401.

[188] Michael Niemeyer and Andreas Geiger. ‘Giraffe: Representing scenes as compositional generative neural feature fields’. In: *CVPR*. 2021, pp. 11453–11464.

[189] Mitja Nikolaus et al. ‘Compositional generalization in image captioning’. In: *arXiv preprint arXiv:1909.04402* (2019).

[190] Atsuhiro Noguchi et al. ‘Neural articulated radiance field’. In: *ICCV*. 2021, pp. 5762–5772.

[191] Donald A. Norman. *The Design of Everyday Things*. USA: Basic Books, Inc., 2002. ISBN: 9780465067107.

[192] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[193] Ahmed AA Osman, Timo Bolkart and Michael J Black. ‘Star: Sparse trained articulated human body regressor’. In: *ECCV*. Springer. 2020, pp. 598–613.

[194] Mark M Palatucci et al. ‘Zero-shot learning with semantic output codes’. In: (2009).

[195] Jeong Joon Park et al. ‘Deepsdf: Learning continuous signed distance functions for shape representation’. In: *CVPR*. 2019, pp. 165–174.

[196] Keunhong Park et al. ‘Nerfies: Deformable Neural Radiance Fields’. en. In: *ICCV*. IEEE, Oct. 2021, pp. 5845–5854. ISBN: 978-1-66542-812-5. (Visited on 15/10/2022).

[197] Adam Paszke et al. ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’. In: *NeurIPS*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[198] Adam Paszke et al. ‘Pytorch: An Imperative Style, High-Performance Deep Learning Library’. In: *NeurIPS*. 2019.

[199] Sida Peng et al. ‘Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies’. en. In: *ICCV*. IEEE, Oct. 2021, pp. 14294–14303. ISBN: 978-1-66542-812-5. (Visited on 15/10/2022).

[200] Sida Peng et al. ‘Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans’. en. In: *CVPR*. IEEE, June 2021, pp. 9050–9059. ISBN: 978-1-66544-509-2. (Visited on 15/10/2022).

[201] Xingchao Peng et al. ‘Domain agnostic learning with disentangled representations’. In: *ICML*. 2019, pp. 5102–5112.

[202] Xingchao Peng et al. *VisDA: The Visual Domain Adaptation Challenge*. 2017. eprint: arXiv:1710.06924.

[203] Jeffrey Pennington, Richard Socher and Christopher D Manning. ‘Glove: Global vectors for word representation’. In: *EMNLP*. 2014, pp. 1532–1543.

[204] Julia Peyre et al. ‘Detecting Unseen Visual Relations Using Analogies’. In: *ICCV*. 2019.

[205] Julia Peyre et al. ‘Detecting Unseen Visual Relations Using Analogies’. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[206] Vihari Piratla, Praneeth Netrapalli and Sunita Sarawagi. ‘Efficient domain generalization via common-specific low-rank decomposition’. In: *ICML*. PMLR. 2020, pp. 7728–7738.

[207] Albert Pumarola et al. ‘D-NeRF: Neural Radiance Fields for Dynamic Scenes’. en. In: *CVPR*. IEEE, June 2021, pp. 10313–10322. ISBN: 978-1-66544-509-2. (Visited on 15/10/2022).

[208] Senthil Purushwalkam et al. ‘Task-driven modular networks for zero-shot compositional learning’. In: *ICCV*. 2019, pp. 3593–3602.

[209] Alec Radford et al. ‘Learning Transferable Visual Models from Natural Language Supervision’. In: *ICML*. PMLR. 2021, pp. 8748–8763.

[210] Aditya Ramesh et al. ‘Zero-shot Text-to-Image Generation’. In: *ICML*. PMLR. 2021, pp. 8821–8831.

[211] Jeremy Reizenstein et al. ‘Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction’. In: *ICCV*. 2021.

[212] Jiawei Ren et al. ‘Balanced meta-softmax for long-tailed visual recognition’. In: *NeurIPS*. 2020.

[213] Shaoqing Ren et al. ‘Faster r-cnn: Towards real-time object detection with region proposal networks’. In: *NeurIPS*. 2015, pp. 91–99.

[214] Tim Salimans et al. ‘Improved techniques for training gans’. In: *NIPS*. 2016, pp. 2234–2242.

[215] Johannes L Schonberger and Jan-Michael Frahm. ‘Structure-from-motion revisited’. In: *CVPR*. 2016, pp. 4104–4113.

[216] Edgar Schonfeld et al. ‘Generalized zero-and few-shot learning via aligned variational autoencoders’. In: *CVPR*. 2019, pp. 8247–8255.

[217] Katja Schwarz et al. ‘Graf: Generative radiance fields for 3d-aware image synthesis’. In: *NeurIPS* 33 (2020), pp. 20154–20166.

[218] Clayton Scott and Gilles Blanchard. ‘Novelty detection: Unlabeled data definitely help’. In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 464–471.

[219] Henry Scudder. ‘Probability of error of some adaptive pattern-recognition machines’. In: *IEEE Transactions on Information Theory* 11.3 (1965), pp. 363–371.

[220] Ramprasaath R Selvaraju et al. ‘Grad-cam: Visual explanations from deep networks via gradient-based localization’. In: *ICCV*. 2017, pp. 618–626.

[221] Ramprasaath R. Selvaraju et al. ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization’. In: *ICCV*. 2017.

[222] Shiv Shankar et al. ‘Generalizing across domains via cross-gradient training’. In: *ICLR*. 2018.

[223] Shuai Shao et al. ‘Objects365: A Large-Scale, High-Quality Dataset for Object Detection’. In: *ICCV*. 2019.

[224] Liyue Shen et al. ‘Scaling human-object interaction recognition through zero-shot learning’. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 1568–1576.

[225] Jake Snell, Kevin Swersky and Richard Zemel. ‘Prototypical networks for few-shot learning’. In: *NIPS*. 2017, pp. 4077–4087.

[226] Elizabeth S Spelke. ‘Principles of object perception’. In: *Cognitive science* 14.1 (1990), pp. 29–56.

[227] Jost Tobias Springenberg. 'Unsupervised and semi-supervised learning with categorical generative adversarial networks'. In: *arXiv preprint arXiv:1511.06390* (2015).

[228] Robin Strudel et al. 'Segmenter: Transformer for Semantic Segmentation'. In: 2021.

[229] Shih-Yang Su, Timur Bagautdinov and Helge Rhodin. 'DANBO: Disentangled Articulated Neural Body Representations via Graph Neural Networks'. In: *ECCV*. 2022.

[230] Shih-Yang Su et al. 'A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose'. In: *NeurIPS*. Vol. 34. Curran Associates, Inc., 2021, pp. 12278–12291. (Visited on 15/10/2022).

[231] Baochen Sun and Kate Saenko. 'Deep coral: Correlation alignment for deep domain adaptation'. In: *ECCV*. 2016.

[232] Guoxing Sun et al. 'Neural Free-Viewpoint Performance Rendering under Complex Human-object Interactions'. en. In: *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, Oct. 2021, pp. 4651–4660. ISBN: 978-1-4503-8651-7. (Visited on 14/10/2022).

[233] Omid Taheri et al. 'GRAB: A Dataset of Whole-Body Human Grasping of Objects'. In: *ECCV*. 2020.

[234] Masato Tamura, Hiroki Ohashi and Tomoaki Yoshinaga. 'QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information'. In: *CVPR*. 2021.

[235] Masato Tamura, Hiroki Ohashi and Tomoaki Yoshinaga. 'QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information'. In: *CVPR*. 2021, pp. 10410–10419.

[236] Jingru Tan et al. 'Equalization loss for long-tailed object recognition'. In: *CVPR*. 2020, pp. 11662–11671.

[237] Pavel Tokmakov, Yu-Xiong Wang and Martial Hebert. 'Learning compositional representations for few-shot recognition'. In: *CVPR*. 2019, pp. 6372–6381.

[238] Hugo Touvron et al. 'Training Data-efficient image transformers & distillation through attention'. In: *ICML*. Vol. 139. 2021, pp. 10347–10357.

[239] Hugo Touvron et al. 'Training data-efficient image transformers & distillation through attention'. In: *ICML*. 2021, pp. 10347–10357.

[240] Edgar Tretschk et al. 'Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video'. en. In: *ICCV*. IEEE, Oct. 2021, pp. 12939–12950. ISBN: 978-1-66542-812-5. (Visited on 15/10/2022).

[241] Grant Van Horn et al. 'The inaturalist species classification and detection dataset'. In: *CVPR*. 2018, pp. 8769–8778.

[242] Ashish Vaswani et al. 'Attention is all you need'. In: *NeurIPS*. 2017, pp. 5998–6008.

[243] Petar Velickovic et al. 'Graph Attention Networks'. In: 2018.

[244] Petar Veličković et al. 'Graph attention networks'. In: *ICML* (2018).

[245] Hemanth Venkateswara et al. 'Deep hashing network for unsupervised domain adaptation'. In: *CVPR*. 2017, pp. 5018–5027.

[246] Oriol Vinyals et al. 'Matching networks for one shot learning'. In: *NIPS*. 2016, pp. 3630–3638.

[247] Catherine Wah et al. 'The caltech-ucsd birds-200-2011 dataset'. In: (2011).

[248] Bo Wan et al. 'Pose-aware Multi-level Feature Network for Human Object Interaction Detection'. In: *ICCV*. 2019, pp. 9469–9478.

[249] Daixin Wang, Peng Cui and Wenwu Zhu. 'Structural deep network embedding'. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. 2016, pp. 1225–1234.

[250] Jianfeng Wang et al. 'RSG: A Simple but Effective Module for Learning Imbalanced Datasets'. In: *CVPR*. 2021, pp. 3784–3793.

[251] Jiashun Wang et al. 'Synthesizing long-term 3d human motion and interaction in 3d scenes'. In: *CVPR*. 2021, pp. 9401–9411.

[252] Jindong Wang et al. 'Generalizing to Unseen Domains: A Survey on Domain Generalization'. In: *arXiv preprint arXiv:2103.03097* (2021).

[253] Jingbo Wang et al. 'Towards Diverse and Natural Scene-aware 3D Human Motion Synthesis'. In: *CVPR*. 2022, pp. 20460–20469.

[254] Peng Wang et al. 'Contrastive Learning based Hybrid Networks for Long-Tailed Image Classification'. In: *CVPR*. 2021, pp. 943–952.

[255] Shaofei Wang et al. 'ARAH: Animatable Volume Rendering of Articulated Human SDFs'. In: *ECCV*. 2022.

[256] Shaofei Wang et al. ‘MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images’. In: *NeurIPS*. Vol. 34. Curran Associates, Inc., 2021, pp. 2810–2822. (Visited on 15/10/2022).

[257] Suchen Wang et al. ‘Discovering human interactions with large-vocabulary objects via query and multi-scale detection’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13475–13484.

[258] Suchen Wang et al. ‘Discovering Human Interactions with Novel Objects via Zero-Shot Learning’. In: *CVPR*. 2020, pp. 11652–11661.

[259] Tiancai Wang et al. ‘Learning Human-Object Interaction Detection using Interaction Points’. In: *CVPR*. 2020.

[260] Wei Wang et al. ‘A survey of zero-shot learning: Settings, methods, and applications’. In: *ACM TIST* 10.2 (2019), pp. 1–37.

[261] Wenhai Wang et al. ‘Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions’. In: *ICCV*. 2021, pp. 568–578.

[262] Xi Wang et al. ‘Reconstructing Action-Conditioned Human-Object Interactions Using Commonsense Knowledge Priors’. In: *arXiv preprint arXiv:2209.02485* (2022).

[263] Yu-Xiong Wang, Deva Ramanan and Martial Hebert. ‘Learning to model the tail’. In: *NIPS*. 2017, pp. 7032–7042.

[264] Yu-Xiong Wang et al. ‘Low-shot learning from imaginary data’. In: *CVPR*. 2018, pp. 7278–7286.

[265] Xudong Wang et al. ‘Long-tailed recognition by routing diverse distribution-aware experts’. In: *ICLR*. 2021.

[266] Ping Wei et al. ‘Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization’. In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1165–1179.

[267] Chung-Yi Weng et al. ‘Humannerf: Free-viewpoint rendering of moving people from monocular video’. In: *CVPR*. 2022, pp. 16210–16220.

[268] Ross Wightman. *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[269] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[270] Wenqi Xian et al. 'Space-time Neural Irradiance Fields for Free-Viewpoint Video'. en. In: *CVPR*. IEEE, June 2021, pp. 9416–9426. ISBN: 978-1-66544-509-2. (Visited on 15/10/2022).

[271] Yongqin Xian et al. 'Feature generating networks for zero-shot learning'. In: *CVPR*. 2018, pp. 5542–5551.

[272] Yongqin Xian et al. 'Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly'. In: *PAMI* 41.9 (2018), pp. 2251–2265.

[273] Liuyu Xiang, Guiguang Ding and Jungong Han. 'Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification'. In: *ECCV*. Springer. 2020, pp. 247–263.

[274] Enze Xie et al. 'SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers'. In: *NeurIPS* 34 (2021).

[275] Qizhe Xie et al. 'Self-training with noisy student improves imagenet classification'. In: *CVPR*. 2020, pp. 10687–10698.

[276] Xianghui Xie, Bharat Lal Bhatnagar and Gerard Pons-Moll. 'CHORE: Contact, Human and Object REconstruction from a single RGB image'. In: *arXiv preprint arXiv:2204.02445* (2022).

[277] Bingjie Xu et al. 'Learning to Detect Human-Object Interactions With Knowledge'. In: *CVPR*. 2019.

[278] Danfei Xu et al. 'Scene graph generation by iterative message passing'. In: *CVPR*. 2017, pp. 5410–5419.

[279] Hongyi Xu, Thiemo Alldieck and Cristian Sminchisescu. 'H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion'. In: *NeurIPS*. Vol. 34. Curran Associates, Inc., 2021, pp. 14955–14966. (Visited on 15/10/2022).

[280] Xiang Xu et al. 'D3d-hoi: Dynamic 3d human-object interactions from videos'. In: *arXiv preprint arXiv:2108.08420* (2021).

[281] Bangbang Yang et al. ‘Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 13779–13788.

[282] Xiangli Yang et al. ‘A Survey on Deep Semi-supervised Learning’. In: *arXiv preprint arXiv:2103.00550* (2021).

[283] Bangpeng Yao and Li Fei-Fei. ‘Modeling mutual context of object and human pose in human-object interaction activities’. In: *CVPR*. IEEE. 2010, pp. 17–24.

[284] Bangpeng Yao, Jiayuan Ma and Fei Fei Li. ‘Discovering Object Functionality’. In: *ICCV*. 2013.

[285] Aron Yu and Kristen Grauman. ‘Fine-grained visual comparisons with local learning’. In: *CVPR*. 2014, pp. 192–199.

[286] Li Yuan et al. ‘Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet’. In: *ICCV*. 2021, pp. 558–567.

[287] Sangdoo Yun et al. ‘Cutmix: Regularization Strategy to Train Strong Classifiers with Localizable Features’. In: *ICCV*. 2019, pp. 6023–6032.

[288] Wei Zhai et al. ‘One-shot object affordance detection in the wild’. In: *arXiv preprint arXiv:2108.03658* (2021).

[289] Yibing Zhan et al. ‘On Exploring Undetermined Relationships for Visual Relationship Detection’. In: *CVPR*. 2019, pp. 5128–5137.

[290] Aixi Zhang et al. ‘Mining the Benefits of Two-stage and One-stage HOI Detection’. In: *NeurIPS*. Vol. 34. 2021.

[291] Chiyuan Zhang et al. ‘Understanding deep learning requires rethinking generalization’. In: *ICLR*. 2017.

[292] Hongyi Zhang et al. ‘Mixup: Beyond Empirical Risk Minimization’. In: *arXiv preprint arXiv:1710.09412* (2017).

[293] Hongyi Zhang et al. ‘mixup: Beyond empirical risk minimization’. In: *ICLR* (2018).

[294] Jason Y. Zhang et al. ‘Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild’. In: *ECCV*. 2020.

[295] Siwei Zhang et al. ‘PLACE: Proximity learning of articulation and contact in 3D environments’. In: *2020 International Conference on 3D Vision (3DV)*. IEEE. 2020, pp. 642–651.

[296] Yan Zhang et al. ‘Generating 3d people in scenes without people’. In: *CVPR*. 2020, pp. 6194–6204.

[297] Yuchen Zhang et al. ‘Bridging theory and algorithm for domain adaptation’. In: *ICML*. 2019.

[298] Kaifeng Zhao et al. ‘Compositional Human-Scene Interaction Synthesis with Semantic Control’. In: *arXiv preprint arXiv:2207.12824* (2022).

[299] Kaifeng Zhao et al. *Compositional Human-Scene Interaction Synthesis with Semantic Control*. en. arXiv:2207.12824 [cs]. July 2022. (Visited on 15/10/2022).

[300] Yuyang Zhao et al. ‘Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification’. In: *CVPR*. 2021, pp. 6277–6286.

[301] Qi Zheng, Chaoyue Wang and Dacheng Tao. ‘Syntax-aware action targeting for video captioning’. In: *CVPR*. 2020, pp. 13096–13105.

[302] Sixiao Zheng et al. ‘Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers’. In: *CVPR* (2021).

[303] Zerong Zheng et al. ‘Structured Local Radiance Fields for Human Avatar Modeling’. en. In: *CVPR*. IEEE, June 2022, pp. 15872–15882. ISBN: 978-1-66546-946-3. (Visited on 16/10/2022).

[304] Xubin Zhong et al. ‘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*. 2021, pp. 13234–13243.

[305] Xubin Zhong et al. ‘Polysemy deciphering network for human-object interaction detection’. In: *European Conference on Computer Vision*. Springer. 2020, pp. 69–85.

[306] Zhisheng Zhong et al. ‘Improving Calibration for Long-Tailed Recognition’. In: *CVPR*. 2021, pp. 16489–16498.

[307] Bolei Zhou et al. ‘Places: A 10 million image database for scene recognition’. In: *PAMI* 40.6 (2017), pp. 1452–1464.

[308] Boyan Zhou et al. ‘Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition’. In: *CVPR*. 2020, pp. 9719–9728.

[309] Kaiyang Zhou et al. ‘Domain generalization with mixstyle’. In: *ICLR* (2021).

[310] Kaiyang Zhou et al. ‘Domain generalization: A survey’. In: *arXiv preprint arXiv:2103.02503* (2021).

[311] Kaiyang Zhou et al. ‘Learning to generate novel domains for domain generalization’. In: *ECCV*. Springer. 2020, pp. 561–578.

[312] Penghao Zhou and Mingmin Chi. ‘Relation Parsing Neural Network for Human-Object Interaction Detection’. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[313] Linchao Zhu and Yi Yang. ‘Inflated episodic memory with region self-attention for long-tailed visual recognition’. In: *CVPR*. 2020, pp. 4344–4353.

[314] Xizhou Zhu et al. ‘Deformable DETR: Deformable Transformers for End-to-End Object Detection’. In: *arXiv preprint arXiv:2010.04159* (2020).

[315] Yizhe Zhu et al. ‘A generative adversarial approach for zero-shot learning from noisy texts’. In: *CVPR*. 2018, pp. 1004–1013.

[316] Yuke Zhu, Alireza Fathi and Li Fei-Fei. ‘Reasoning about object affordances in a knowledge base representation’. In: *ECCV*. Springer. 2014, pp. 408–424.

[317] Christian Zimmermann et al. ‘Freihand: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images’. In: *ICCV*. 2019, pp. 813–822.

[318] Cheng Zou et al. ‘End-to-end human object interaction detection with hoi transformer’. In: *CVPR*. 2021, pp. 11825–11834.

APPENDIX A

Appendix of Chapter 2

A1 Visual Compositional Learning

A1.1 Hyper-Parameters

In our proposed framework, there are two hyper-parameters λ_1 and λ_2. We evaluate the performance when we set different values for the two hyper-parameters.

From Table A.1, we can find when we increase the value of λ_1. We can witness a considerable increase in the Full category. If we choose the value more than 2.0 for λ_1, the performace slightly decreases. From Table A.2, if we set 0.5 or 0.1, the performance is similar. But, when λ_2 is more than 1.0 or less than 0.1, the performance drops quickly.

Like [66, 157], we first detect the objects in the image and then use the object detection results to infer the HOI categories during test. we use the same score threshold (0.8 for human and 0.3 for object) as [157] in resnet50 coco detector. We use 0.3 for human and 0.1 for object in resnet101 detector that is finetuned on HICO-DET dataset since the object detection result is largely better.

TABLE A.1. The results of setting different values for λ_1 when λ_2 is 0.5 in HICO-DET.

λ_1	1.0	1.5	2.0	2.5	3
Full	18.96	18.95	19.43	19.29	19.34

TABLE A.2. The results of setting different values for λ_2 when λ_1 is 2.0 in HICO-DET.

λ_2	0.05	0.1	0.5	1.0	1.5
Full	19.18	19.30	19.43	19.10	18.90

TABLE A.3. The results of the number of interactions in minibatch in HICO-DET.

the number of interactions	VCL	Full	Rare	NonRare
1	-	18.41	14.17	19.68
1	✓	18.85	14.98	20.01
5	-	18.43	14.14	19.71
5	✓	**19.43**	**16.55**	**20.29**

A1.2 The effect of the number of interactions in minibatch

In order to compose enough interactions for Visual Compositional Learning, we increase the number of interactions in each minibatch while reducing the number of augmentations for each interaction and the number of negative interactions to keep the batch size unchanged in our experiment. We evaluate the effect in this section. We set the maximum number of interactions 5 in our experiment. Noticeably, most training images in HICO-DET only contain one interaction.

From Table A.3, we can find the baseline model of different iteractions has similar results with 18.43 mAP and 18.47 mAP respectively. However, we witness a better improvement (1.0 mAP vs 0.44 mAP) if we increase the interaction classes in the minibatch. It shows that increasing the number of interactions is considerably beneficial for Visual Compositional Learning.

A1.3 The two branches in zero-shot HOI detection

From Table A.4, we can find the performance of verb-object branch in Seen category and Full category is similar to that of spatial-human branch, while verb-object branch is 3.52% and **4.90%** better than spatial branch in selecting rare first and selecting non-rare first respectively

TABLE A.4. Two branches ablation study of the proposed Visual Compositional Learning framework in zero-shot HOI detection on HICO-DET test set during inference.

Method	Unseen	Seen	Full
Verb-Object branch (rare first)	**7.85**	15.48	13.95
Spatial-Human branch (rare first)	4.33	15.92	13.60
Two branches (rare first)	7.55	**18.84**	**16.58**
Verb-Object branch (non-rare first)	**10.61**	10.95	10.88
Spatial-Human branch (non-rare first)	5.71	11.82	10.60
Two branches (non-rare first)	9.13	**13.67**	**12.76**

TABLE A.5. Illustrations of VCL with language priors.

Strategy	Full (mAP %)	Rare (mAP %)	NonRare (mAP %)
VCL	19.43	16.55	20.29
VCL + Language prior	19.56	16.27	20.55

in the Unseen category. **Particularly, after we fuse the result of the two branches, the Unseen category witnesses a considerable decrease in the two selecting strategies.**

A1.4 Verb Polysemy Problem

There is a verb polysemy problem in HOI detection, that is the verb "play" has different meanings between "play guita" and "play football". But, HICO restricts itself to a single sense of a verb (with the exceptions of a couple of verbs) [27, 70], which means that the verb polysemy problem is not serious. Previous HOI approaches [224, 277, 205] usually regard the verb from different HOIs as same, and successfully achieve good performance. We also conduct a simple experiment to validate this problem. We use the language priors to choose the suitable composited HOIs according to the object similarity of word embedding in Table A.5. We can find the improvement of language priors is very limited.

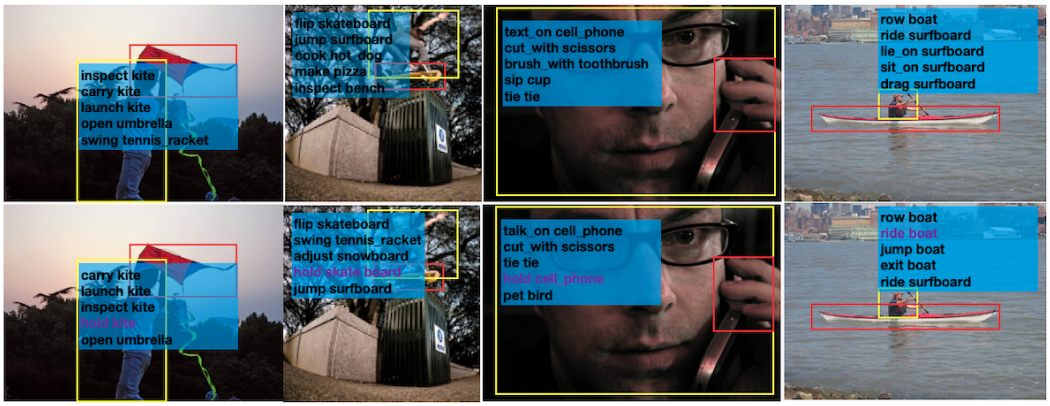

FIGURE A.1. Some HOI detections detected by the proposed Compositional Learning and the model without Compositional Learning in zero-shot HOI detection (selecting nonrare first). The first row is the results of our baseline model without VCL. The second row is the results of the proposed composition learning. The unseen interactions are marked with purple. We illustrate top 5 score results for the human object pair.

A2 Fabricated Compositional Learning

A2.1 Visual Illustration of zero-shot HOI detection

Similar to Figure 4 in the paper, we qualitatively show that our proposed Visual Compositional Learning framework can detect those unseen interactions efficiently in Figure A.1 while the baseline model without Visual Compositional Learning is misdetected on HICO-DET. It shows our proposed Visual Compositional Learning framework is significantly beneficial for Unseen categories.

A2.2 Unseen labels on HICO-DET dataset

In zero-shot detection in HICO-DET, we select randomly unseen labels for zero-shot detection. In detail, we first sorted the labels according to the number of instances of categories. Then we select the HOIs out for unseen data according to the sorted label list and meanwhile make sure that all types of objects and verbs exist in seen data. we provide the unseen label id in two zero-shot learning settings.

rare first ids: 509, 279, 280, 402, 504, 286, 499, 498, 289, 485, 303, 311, 325, 439, 351, 358, 66, 427, 379, 418, 70, 416, 389, 90, 395, 76, 397, 84, 135, 262, 401, 592, 560, 586, 548, 593, 526, 181, 257, 539, 535, 260, 596, 345, 189, 205, 206, 429, 179, 350, 405, 522, 449, 261, 255, 546, 547, 44, 22, 334, 599, 239, 315, 317, 229, 158, 195, 238, 364, 222, 281, 149, 399, 83, 127, 254, 398, 403, 555, 552, 520, 531, 440, 436, 482, 274, 8, 188, 216, 597, 77, 407, 556, 469, 474, 107, 390, 410, 27, 381, 463, 99, 184, 100, 292, 517, 80, 333, 62, 354, 104, 55, 50, 198, 168, 391, 192, 595, 136, 581

non-rare first ids: 38, 41, 20, 18, 245, 11, 19, 154, 459, 42, 155, 139, 60, 461, 577, 153, 582, 89, 141, 576, 75, 212, 472, 61, 457, 146, 208, 94, 471, 131, 248, 544, 515, 566, 370, 481, 226, 250, 470, 323, 169, 480, 479, 230, 385, 73, 159, 190, 377, 176, 249, 371, 284, 48, 583, 53, 162, 140, 185, 106, 294, 56, 320, 152, 374, 338, 29, 594, 346, 456, 589, 45, 23, 67, 478, 223, 493, 228, 240, 215, 91, 115, 337, 559, 7, 218, 518, 297, 191, 266, 304, 6, 572, 529, 312, 9, 308, 417, 197, 193, 163, 455, 25, 54, 575, 446, 387, 483, 534, 340, 508, 110, 329, 246, 173, 506, 383, 93, 516, 64

A2.3 Additional Details

A2.3.1 More examples of Open Long-tailed HOI Detection

Figure A.2 provides more clear illustration of open long-tailed HOI detection. Open long-tailed HOI detection aims to detect head, tail and unseen classes in one integrated way from long-tailed HOI examples.

A2.3.2 Factorized model

We implement the factorized model under our framework. In details, we replace the HOI branch in Figure 3 in the paper with verb and object stream. The two streams predict the verb and object respectively. During inference, we merge the score of verb and object to obtain HOI score as follows,

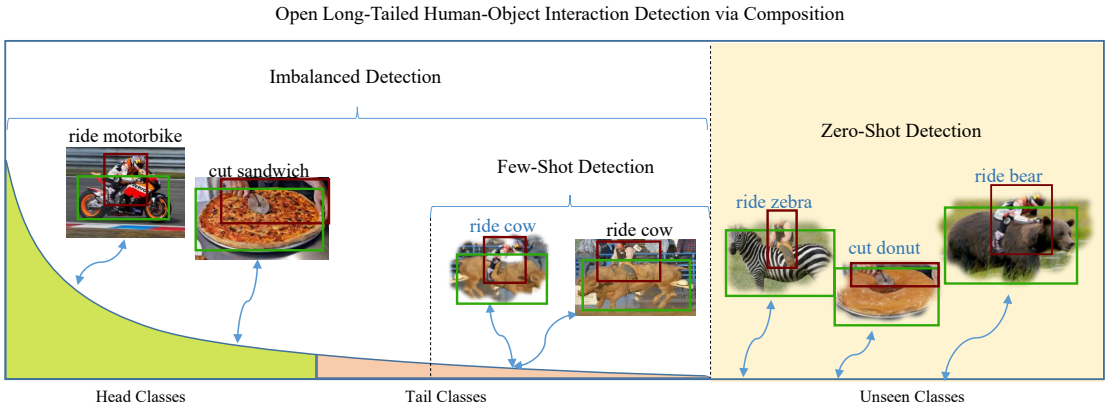

FIGURE A.2. Open long-tailed HOI detection addresses the problem of imbalanced learning and zero-shot learning in a unified way. We propose to compose new HOIs for open long-tailed HOI detection. Specifically, the blurred HOIs, e.g., “ride bear", are composite, while the black HOIs are real.

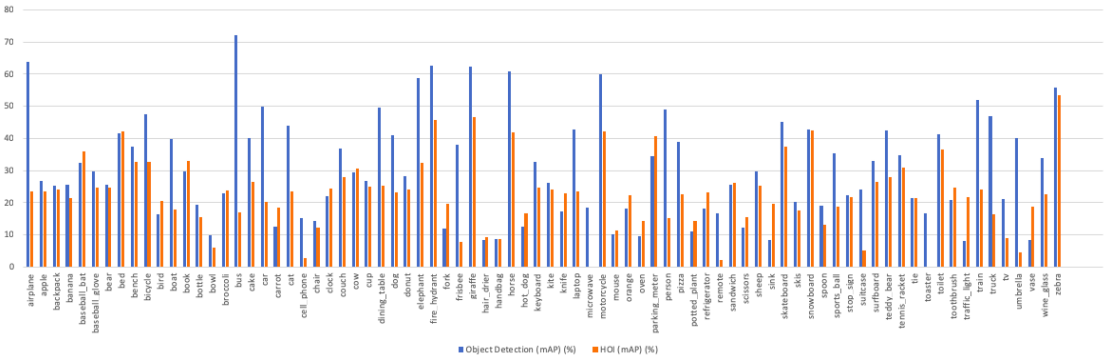

FIGURE A.3. Illustration of Object detection result and HOI detection result in HICO-DET dataset. Blue is Object result. Yellow is HOI result. We average HOI detection AP according to the object categories for a direct comparison.

$$\mathbf{S}_{hoi} = (\mathbf{S}_o\mathbf{A}_o) + (\mathbf{S}_v\mathbf{A}_v), \tag{A.1}$$

where $\mathbf{A}_v$ ($\mathbf{A}_o$) is the co-occurrence matrix between verbs (objects) and HOIs, $\mathbf{S}_o$ is the score from object stream and $\mathbf{S}_v$ is the score from verb stream.

A2.3.3 The Effect of Objects on HOI Detection

In the nature, different types of objects form a long-tail distribution. Then, all those actions that people perform on those objects are inevitably long-tailed. As a result, those HOIs that

we observed are long-tailed. This motivates us to fabricate balanced objects for composing HOI samples with visual verbs. We have demonstrated the long-tailed distribution of objects in Figure 2 in the paper and the effect of different object detector on HOI detection in Table 6 in paper. We further illustrate HOI detection has roughly similar performance to object detection among most object categories in Figure A.3, which also illustrates the importance of object detector for HOI detection at the same time. Meanwhile, it is necessary to balance the the distribution of objects.

A2.4 Additional Quantitative analysis

A2.4.1 Object Identity

In Table A.6, we compare three kinds of object identity. The object variables are identified after we fine-tune the fabricator in the first step. Meanwhile, in the end-to-end optimization, the object variables can maintain object semantic information. We find word embedding and object variables achieve similar performance (24.78% vs 24.68%), while the performance of one-hot representation is a bit worse. Particularly, the HOI model is initialized with a pretrained object detector model. Thus, one-step optimization can also optimize the Fabricator according to the pre-trained backbone.

A2.4.2 V-COCO

We evaluate V-COCO based on the state-of-the-art method PMF [248]. Thus, our baseline is PMFNet. For V-COCO, we do not use auxiliary verb loss since there are only two kinds of objects (stru, obj) on V-COCO. We set λ_1 as 1 and λ_2 as 0.25. We find although the data on V-COCO is balanced, FCL still improves the baseline in Table A.8. In fact, we only change a few codes based on PMFNet. Our code is provided in the supplementary material.

A2.4.3 Visual Relation Detection

We also present the efficiency of FCL in Predicate Detection on Visual Relation Detection [175] in Table A.7. Here, we combine subject, predicate and fabricated object to generate

TABLE A.6. Illustration of the effect of different object identity in the proposed fabricator on HICO-DET dataset[28].

Method	Full	Rare	NonRare
object variables	**24.78**	**20.05**	**26.19**
word embedding	24.68	20.03	26.07
one-hot	24.38	19.49	25.84

Method	Zero-Shot	All
MFURLN [289]	-	58.2
MFURLN [289]*	25.26	57.87
Ours	**27.31**	**58.31**

TABLE A.7. Illustration of Predicate Detection in Visual Relation Detection. Zero-shot means the relation (subject, predicate, object) do not exist in the training data.

Method	AP_{role}
PMFNet (reproduced)	51.85
FCL	**52.35**

TABLE A.8. Illustration of Fabricated Compositional Learning on V-COCO based on PMF [248]

novel relation samples [289]. Table A.7 illustrates an important improvement on zero-shot predicate detection compared to the state-of-the-art approach with FCL.

A2.4.4 Semantic Verb Regularization

We also experiment with semantic verb regularization similar to [277] with Graph Convolutional Network and verb word embeddings graph. In details, we use the cosine distance loss to regularize the visual verb representation to be similar to the corresponding word embedding. Here, similar to [277], we equally treat same category of verbs among different HOIs as same. Table A.9 illustrates FCL is orthogonal to semantic regularization. Meanwhile, auxiliary verb loss achieve similar performance compared to semantic verb regularization [277]. When we incorporate both semantic regularization and auxiliary verb loss, the improvement is limited. This means verb regularization loss in the paper and semantic verb regularization have similar effect on the model.

FCL	S	V	Full	Rare	NonRare	Unseen
-	✓	-	18.22	15.69	20.74	12.98
✓	✓	-	19.39	17.99	21.21	14.83
✓	-	✓	19.61	18.69	21.13	15.86
✓	✓	✓	19.62	18.38	21.61	14.73

TABLE A.9. Illustration of semantic regularization modules based on the ablated setting in paper. FCL Means proposed Compostional Learning. S means semantic regularize loss. V means auxiliary verb loss (verb regularization loss in paper).

TABLE A.10. Illustration of auxiliary object loss on HICO-DET dataset[28]. Here, auxiliary object loss aims to regularize visual objects

Method	Full	Rare	NonRare
w/o object loss	**24.78**	**20.05**	**26.19**
auxiliary object loss	24.54	19.93	25.92

A2.4.5 Object Feature Regularization

visual object feature regularization. Object features are usually more discriminative. Meanwhile, we initialize our backbone with the faster-rcnn pre-trained in COCO dataset, which largely helps us to obtain discriminative object features. Thus, it is unnecessary to use auxiliary object loss to regularize object features (See Table A.10). Meanwhile, we find the object features is more discriminative from the t-SNE graph in Figure A.5.

A2.4.6 The Effect of Union Box on FCL

We extract verb representation from the union box of human and object. In Table A.11, we illustrate with human box verb, FCL still effectively improves the baseline. This shows the proposed method is orthogonal to the verb representation. Noticeably, although the union box contains the object, the HOI model mainly learns the verb representation via compositional learning, and largely ignores the identity information of the object. Thus, the object in the union box do not have much effect on Fabricator. By comparing human box and union box for verb representation in Table 2 in paper and Table A.11, we find verb representation from union box largely improves the performance since it provides more context information for verb representation.

TABLE A.11. Illustration of the box for verb representation on HICO-DET dataset[28].

Method	Full	Rare	NonRare
baseline(human box)	22.91	16.66	24.77
FCL (human box)	23.83	18.62	25.39

TABLE A.12. The result while filtering out the composite HOIs according to the similarity between the fake objects and original objects. Neighbors (K) means top K neighbors according to similarity. This experiment is based on ablated setting in Table 3 in paper. When the number of neighbors is 80, it means we do not filter out composite HOIs according to similarity.

#Neighbors (K)	1	5	10	20	40	80
FCL (Full)	18.70	19.15	19.19	19.48	19.60	19.61

TABLE A.13. Comparison between step-wise optimization and one step optimization in unseen object HOI detection.

Method	Full	Rare	NonRare	Unseen
one step	19.87	15.01	22.51	**15.54**
step-wise	**20.13**	**16.71**	**22.82**	13.85

A2.4.7 Verb Analysis

The same verb might has different meanings in different HOIs. However, the verb in HOI dataset (e.g. HICO-DET) mainly represents action. Thus, the verb in HOI dataset is usually not ambiguous. Meanwhile, the deep convolutional network (e.g.Resnet) is able to fit some ambiguous and even random data [291]. Therefore, we can use factorized method [277] for HOI detection and the ambiguous verbs do not affect the compositional learning on HICO-DET [111], even if there are still some ambiguous verbs (e.g. hold) who can be related to multiple objects.

Besides, we further demonstrates the improvement of FCL among different categories of verbs in Figure A.4. We find the ambiguity does not affect the performance of those verbs in fact. For example, although the verb "hold" is related to 61 kinds of objects in HICO-DET, the correpsonding HOIs of "hold" still achieves considerable improvement.

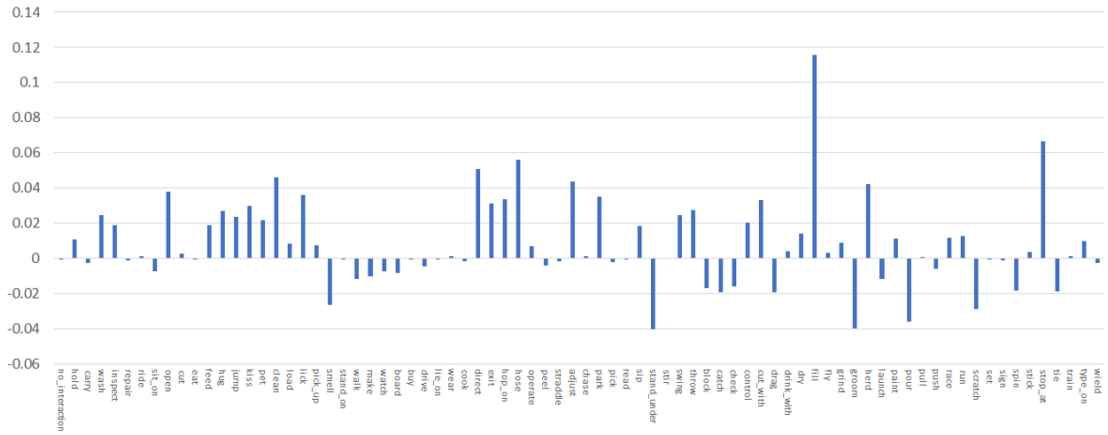

FIGURE A.4. The improvement among the classes of verbs on HICO-DET. The verbs are sorted by the number of HOIs that the particular verb is related. The clear figure is in the directory of Compressed package.

Inspired by that people interact similar objects in a similar manner. we also design an approach to select composite HOIs according to the similarity between different object of objects, i.e.we only keep those composite HOIs whose object is in the top K neighbors of the verb's original object. The original object of the verb is the visual object paired with the verb in the HOI annotation. This helps us to filter out those ambiguous composite HOIs. Specifically, we calculate the similarity between different classes of objects by its word embedding [203]. Then we can obtain the top K neighbors for each class of objects. Table A.12 shows with more similar objects, the performance steadily improves. Particularly, there are only one verb relating to more than 40 HOIs, and 4 verbs with more than 20 HOIs in HICO-DET. When $K = 1$, we only keep composite HOIs whose objects have the same label to the original object.

A2.4.8 Complementarity and Orthogonality to previous methods

Complementary to previous zero-shot method. We incorporate VCL [111] (the released code and model) to evaluate the complementarity of FCL to previous zero-shot approach. Table A.14 shows FCL is complementary to the compositional approach [111] between pair-wise images by fusing FCL and VCL (See Table 2 in Paper).

TABLE A.14. Illustration of the fusion of FCL and VCL on HICO-DET dataset[28] under ground truth object result.

Method	Detector	Full	Rare	NonRare
VCL	GT	43.09	32.56	46.24
FCL	GT	44.11	**36.62**	46.35
FCL + VCL	GT	**45.25**	36.27	47.94

TABLE A.15. Illustration of FCL without re-weighting on long-tailed HOI detection.

FCL	Full	Rare	NonRare
-	20.79	13.19	23.06
✓	21.20	15.48	22.90

TABLE A.16. Illustration of proposed modules on long-tailed HOI detection. FCL Means proposed Fabricated Compostional Learning. V means verb regularization loss.

FCL	V	Full	Rare	NonRare
-	-	23.35	17.08	25.22
✓	-	23.86	18.16	25.56
-	✓	23.94	17.48	25.87
✓	✓	**24.78**	**20.05**	**26.19**

Orthogonal to spatial pattern. Table A.18 illustrates that the spatial pattern strategy [66, 157, 312, 248] largely improves the performance, and the proposed compositional learning is orthogonal to spatial pattern.

Orthogonal to re-weighting. In our baseline, we utilize the re-weighting strategy that is used in [157, 111] to compare directly with [111]. We demonstrate FCL is orthogonal to re-weighting in Table A.15. Without the useful re-weighting strategy, FCL still achieves similar improvement than baseline.

A2.4.9 Complementary Analysis of fabricator

In this section, we conduct analysis of fabricator on HOI detection without unseen data (the full long-tailed HOI detection). We witness the similar trend compared to the ablation study in the paper.

Method	Full	Rare	NonRare
FCL	**24.78**	**20.05**	**26.19**
FCL w/o noise	24.22	19.23	25.72
FCL w/o verb	24.29	18.98	25.87
verb fabricator	23.93	17.10	25.97

TABLE A.17. Ablation study of fabricator. Verb fabricator means we fabricate verb features.

FCL	SP	ZS	Full	Rare	NonRare	Unseen
-	-	-	21.07	14.11	23.15	-
✓	-	-	21.68	16.92	23.11	-
✓	✓	-	**24.78**	**20.05**	**26.19**	-
-	-	✓	15.29	14.45	17.85	8.27
✓	-	✓	16.82	16.57	18.17	12.94
✓	✓	✓	**19.61**	**18.69**	**21.13**	**15.86**

TABLE A.18. Illustration of spatial pattern. SP means we use spatial pattern. ZS means zero-shot setting.

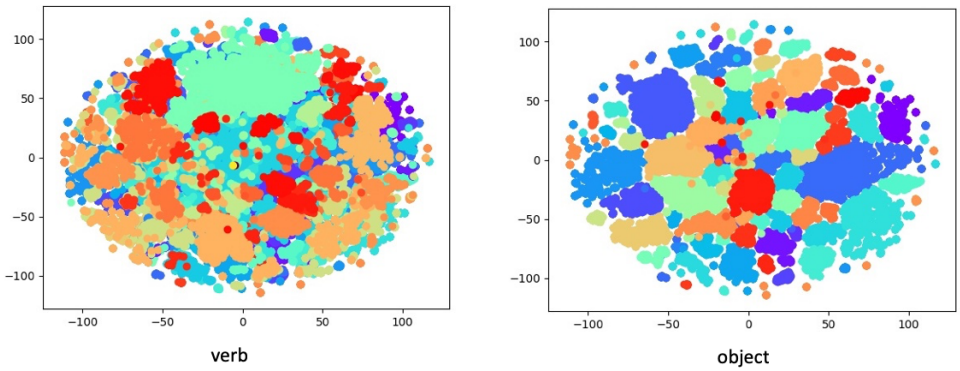

FIGURE A.5. The comparison between verb features and object features.

Verb and Noise for fabricating objects. Table A.17 demonstrates the efficiency of verb and noise. Particularly, the performance in the full HOI detection drops larger than that in zero-shot study in the paper. We think it is because the improvement on unseen category is large, while there are no unseen category in the full HOI detection.

Verb Fabricator. Table A.17 illustrates if we fabricate verb features to augment HOI samples, the performance apparently decreases to 23.93% in long-tailed HOI detection. This again illustrates that the verb feature is more complex and it is difficult to generate efficient verb features to facilitate HOI detection.

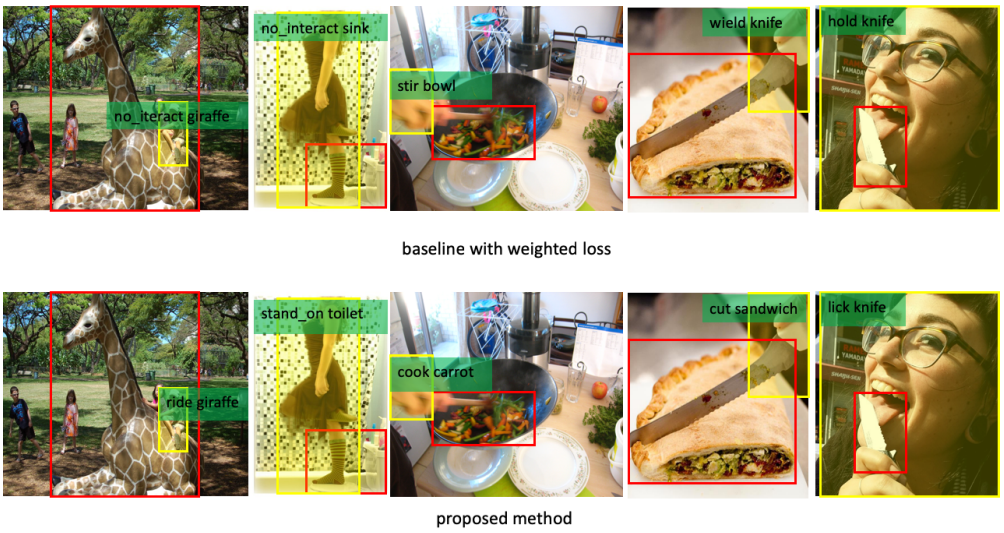

FIGURE A.6. Visual Comparison between FCL and our baseline. The two models use same detector.

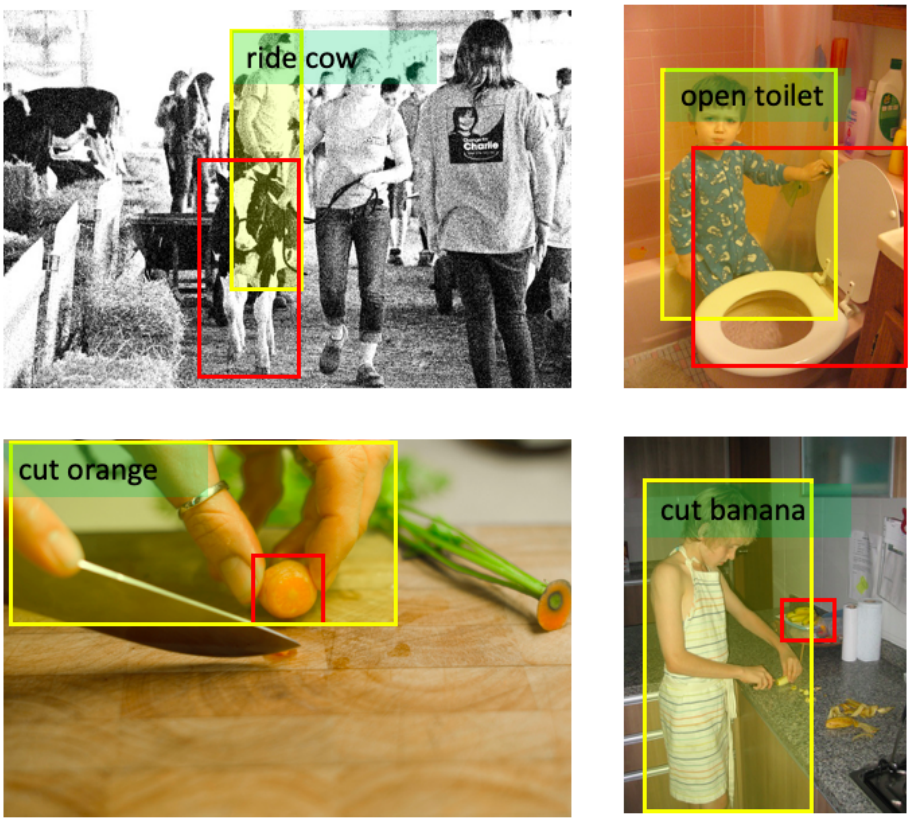

FIGURE A.7. Illustration of failure cases.

TABLE A.19. Illustration of ablated study on λ_3 in HICO-DET based on open long-tailed HOI detection (corresponding to Table 3 in paper).

λ_3	0.1	0.3	0.5
FCL	19.30	19.61	19.10

A2.5 Additional Ablation Study

Step-wise optimization. We also provide the comparison between step-wise optimization and one-step optimization in unseen object HOI detection in Table A.13.

Hyper-Parameters. We follow the hyper-parameters in [111] for λ_1 and λ_2. For λ_3, we provide the ablated experiment in Table A.25 based on 0.5 because we think L_{reg} is less important than L_{CL}.

A2.6 Qualitative Analysis

A2.6.1 Primitive Features

Figure A.5 illustrates verb features are apparently more difficult to distinguish. The verb representation is abstract and complicated. By contrast, object representations extracted from modern object detector are more discriminative.

A2.6.2 Qualitative Comparison

In Figure A.6, we compare our baseline with our proposed method. Apparently, our proposed method efficiently detects rare categories, while the corresponding baseline can not. In fact, all the HOIs detected by our method in Figure A.6 have less than five samples in training set which is much less than the rare setting (less than 10 samples).

A2.6.3 Failure cases analysis

We provide some false positive results on Rare category in Figure A.7. All failure cases can be separated into four groups: blurry image, wrong verb, wrong object, wrong match. If the

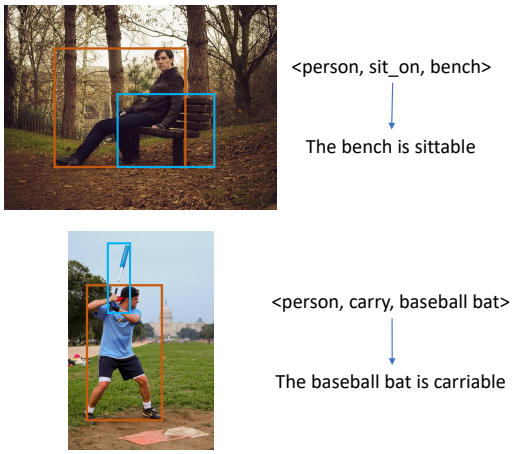

FIGURE A.8. The examples about HOI and Affordance.

image is blurry or has partial occlusion, it is hard to detection the interaction right. Besides, verb is usually hard to classify. Meanwhile, small objects also cause that the network detect object wrongly (e.g.the carrot in Figure A.7). Lastly, even though the network can recognize action and object correctly, it also possibly mismatches the interaction. For example, in Figure A.7, the women do not interact with the banana on the corner of the table.

A3 Affordance Transfer Learning

A3.1 More Examples of HOI and Object Affordance

Images labeled with HOI annotations simultaneously show the affordance of the objects. Therefore, we can not only learn to detect HOIs, but also recognize the affordance of the objects as illustrated in Figure A.8. By combining the affordance representation with various kinds of its corresponding objects, we then enable the model to recognize the affordance of novel objects.

A3.2 Non-COCO classes in Object365

For evaluating ATL on affordance recognition of unseen classes, we manually select 12 non-coco classes from object365: glove, microphone, american football, strawberry, flashlight,

TABLE A.20. Affordances of Non-COCO classes in Object365 based on HOI-COCO.

name	verbs/affordances
glove	carry, throw, hold
microphone	talk_on_phone, carry, throw, look, hold
american football	kick, carry, throw, look, hit, hold
strawberry	cut, eat, carry, throw, hold
flashlight	carry, throw, hold
tape	carry, throw, hold
baozi	eat, carry, look, hold
durian	eat, carry, hold
boots	carry, hold
ship	ride, sit, lay, look
flower	look, hold
basketball	throw, hold

TABLE A.21. Affordances of Non-COCO classes in Object365 based on HICO-DET.

name	verbs/affordances
glove	buy, carry, hold, lift, pick_up, wear
microphone	carry, hold, lift, pick_up
american football	block, carry, catch, hold, kick, lift, pick_up, throw
strawberry	buy, eat, hold, lift, move
flashlight	buy, hold, lift, pick_up
tape	buy, hold, lift, pick_up
baozi	buy, eat, hold, lift, pick_up
durian	buy, hold, lift, pick_up
boots	buy, hold, lift, pick_up, wear
ship	adjust, board
flower	buy, hold, hose, lift, pick_up
basketball	block, hold, kick, lift, pick_up, throw

tape, baozi, durian, boots, ship, flower, basketball. The actions that we can act on those objects (i.e.affordance) are list on Table A.20 and Table A.21.

A3.3 Detailed Analysis for the Motivation

Actually, after we generate the composite HOI features, we have features for both known and unknown concepts. We merely know the HOI features of the known concepts are existing,

while we do not know whether the HOI features of unknown concepts are reasonable or not. This actually fall into a typical semi-supervised learning, in which part of samples are labeled (known). Therefore, inspired by the popular semi-supervised learning method, we propose to design a self-training strategy with pseudo labels.

SCL largely improves concept discovery. At first, during training, SCL involves both HOI instances from known or unknown concepts (via pseudo-labeling). Another important thing is that SCL uses both positive and negative unknown concepts, which prevents the model from only fitting the verb patterns. For example, the classifier may predict a reasonable concept for the verb “eat" regardless of the object representation, if there are no negative unknown concepts, e.g., “eat TV". Lastly, as shown in Figure A.9, SCL also reduces the risk of overfitting known concepts compared with ATL. e.g., we observe high confidence for the novel concept "squeeze banana"(sort in 2027) in SCL, while the confidence of "squeeze banana" is merely 0.0017 (sort in 7554) in ATL.

A3.4 Annotation

In order to evaluate the proposed method, we manually annotate the novel concepts for both HICO and V-COCO dataset. Specifically, we annotate the concepts that people can infer from existing concepts. The final set of concepts are provided in the supplemental material.

Statistically, there are about 1.3% and 1.9% mislabeled pairs on HICO-DET and V-COCO, respectively. Meanwhile, there are about 1.7% and 1.1% unlabeled pairs (including ambiguous verbs) on the remaining categories of HICO-DET and V-COCO.

To evaluate the effect of annotation quality of concept annotation on HOI concept discovery, we illustrate the result of different models with different annotations. We compare two versions of annotations, both of which are provided in supplemental materials. Specifically, the file “label_hoi_concept.csv" is the worse version, while “label_hoi_concept_new.csv" is the refined version. Table A.22 shows SCL even achieves better performance when evaluate SCL with better annotation, while the performance of baseline is not improved. This experiments together with Table 1 in the main paper show the quality of current annotation is enough for the evaluation of the proposed method.

TABLE A.22. The performance of the proposed method for HOI concept discovery under different annotations. Better Annotation indicates we remove some wrongly labeled concepts in annotation. We report all performance using the average precision (AP) (%). UC means unknown concepts and KC means known concepts. SCL means self-compositional learning. SCL− means online concept discovery without self-training.

Method	Better Annotation	UC	KC
SCL−		22.36	83.04
SCL		33.26	93.06
SCL−	✓	22.25	83.04
SCL	✓	**33.58**	**92.65**

TABLE A.23. The illustration of discovered concepts.

Method	Concepts with high confidence	Concepts with low confidence
SCL−	type_on sink,inspect refrigerator,feed suitcase, inspect chair,carry stop_sign	zip zebra, sign dog, chase broccoli, set parking_meter, tag teddy_bear
SCL	ride bear, board truck, carry bowl, wash fire_hydrant, hop_on motorcycle	zip zebra, flush parking_meter, stop_at hair_drier, stop_at microwave

A3.5 Qualitative illustration

We also illustrate the discover concepts in this Section. Here, we choose the concepts after removing the known concepts from the prediction list because the confidence of known concepts in the prediction of SCL is usually very higher. We choose 5 concepts with high confidence and 5 concepts with low confidence to illustrate. Table A.23 shows the discovered concepts in SCL are usually more reasonable.

A3.6 Ablation Studies

A3.6.1 Modules

We conduct ablation studies on three modules: verb auxiliary loss [110], union verb [111], and spatial branch [66]. Union verb indicates that we extract verb representation from the union box of human and object. When we remove the union verb representation, we directly extract verb representation from the human bounding box; In our experiment, we remove the spatial branch. Here, we demonstrate we achieve better performance without the spatial branch.

TABLE A.24. Ablation studies of different modules on HICO-DET. UC means unknown concepts and KC means known concepts. Verb aux loss means Verb auxiliary loss (i.e., binary cross entropy loss). Results are reported by average precision (%).

Spatial branch	Verb aux loss	Union Verb	UC	KC
✓	✓	✓	32.56	94.39
-	✓	✓	**33.26**	93.06
✓	-	✓	29.56	93.36
✓	✓	-	28.30	94.27

Spatial branch. We remove the spatial branch in [66], which is very effective for HOI detection. We find that the spatial branch degrades the performance of HOI concept discovery: the performance of HOI concept discovery increases from 32.56% to 33.26% without spatial branch, as shown in Table A.24. We thus remove spatial branch.

Verb auxiliary loss. We follow [110] to utilize a verb auxiliary loss to regularize verb representations. As shown in Table A.24, the model without using a verb auxiliary loss drops by nearly 3% on unseen concepts, which demonstrates the importance of verb auxiliary loss for HOI concept discovery.

Union verb. Table A.24 demonstrates that extracting verb representation from union box is of great importance for HOI concept discovery. When we extract verb representation from human bounding box, the result of HOI concept discovery apparently drops from 32.56% to 28.30%.

Though verb auxiliary loss and union verb representation are very helpful for concept discovery, the performance without the two strategies still outperform our baseline, i.e., online concept discovery without self-training.

A3.6.2 Convergence Analysis

To some extent, the self-training approach makes use of all composite HOIs, and thus significantly enriches the training data. As a result, the self-training strategy usually requires more iterations to converge to a better result. Figure A.9 illustrates the comparison of convergence between online concept discovery and self-training. For online concept discovery,

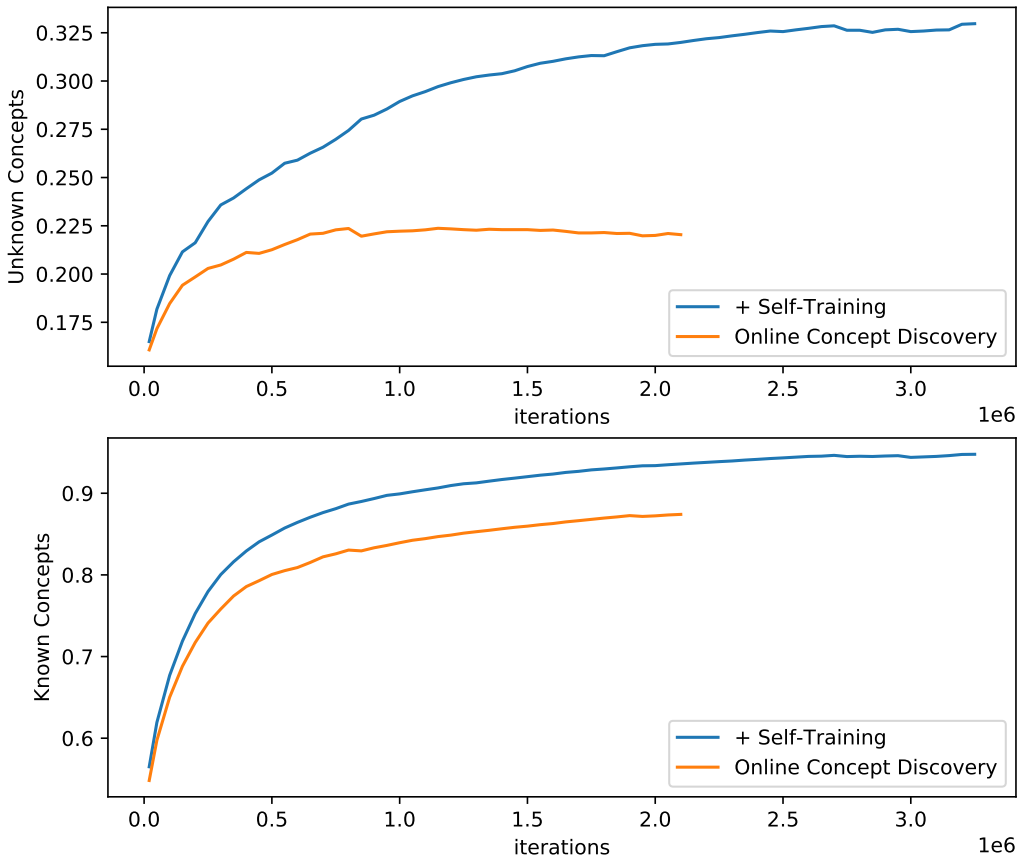

FIGURE A.9. Illustration of the convergence with self-training strategy.

we observe that the model begins to overfit the known concepts after 2,000,000 iterations, and we thus have an early stop during the optimization. We notice that the result on unknown concepts of self-training increases to 32.%, while the baseline (i.e., online concept discovery) begins to overfit after 800,000 iterations. This might be because the self-training utilizes all composite HOIs including many impossible combinations (i.e., negative samples for HOI concept discovery).

TABLE A.25. Ablation studies of hyper-parameters on V-COCO. UC means unknown concepts and KC means known concepts. Results are reported by average precision (%).

λ_3	0.5	0.5	0.5	0.25	1.	2.	4.
T	1	2	0.5	1.	1.	1.	1.
UC (%)	29.52	28.60	29.69	28.06	29.94	**31.33**	29.78
KC (%)	97.57	96.76	97.57	95.32	97.87	97.81	97.94

A3.6.3 Hyper-parameters

In the main paper, we have several hyper-parameters (i.e.λ_1, λ_2, λ_3, T, where $\lambda_1 = 2.$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ and $T = 1.$). For λ_1 and λ_2, we follow the settings in [111]. For λ_3 and T, we perform ablation studies on V-COCO as shown in Table A.25. We notice that both T and λ_3 have an important effect on the HOI concept discovery. As shown in Table A.25, the performance increases from 29.52% to **31.33%** on unseen concepts when we set $\lambda_3 = 2.$, which is much better than the results reported in the main paper. This also illustrates that $\mathcal{L}_d$ is more important than $\mathcal{L}_{CL}$ for HOI concept discovery.

In our experiment, we apply the temperature T to predictions. As shown in Table A.25, we find that when T decreases to 0.5, the performance also slightly increases from 29.52% to 29.69%. Thus, we further conduct ablation experiments on T in Table A.26. Specifically, to quickly evaluate the effect of T, we remove spatial branch and run all experiments with 1,000,000 iterations. Noticeably, when we set $T = 0.25$, the performance on concept discovery further increases from 30.36% to **33.66%**, which indicates a smaller temperature helps HOI concept discovery. In our experiments, we also find this result further increases to over 35.% when $T = 0.5$ after convergence, which is much better than the result (33.26%) of $T = 1$. This might be because smaller temperature is less sensitive to noise data, since composite HOIs can be regard as noise data.

A3.6.4 Normalization for Pseudo-labels

In our experiment, we normalize the confidence matrix for pseudo-labels. Table A.27 illustrates the normalization approach has a slight effect on the concept discovery performance.

TABLE A.26. Ablation studies of hyper-parameter T on HICO-DET. Here, we run all experiments with only 1,000,000 iterations and remove the spatial branch to evaluate T. UC means unknown concepts and KC means known concepts. Results are reported by average precision (%).

T	2	1	0.5	0.25	0.125
UC (%)	27.15	30.36	33.54	**33.66**	33.25
KC (%)	85.53	88.72	91.71	93.62	**94.32**

TABLE A.27. Illustration of normalized pseudo labels on HICO-DET and V-COCO. Experiments results are reported by average precision (%). Here, the SCL model uses spatial branch.

Method	HICO-DET		V-COCO	
	UC (%)	KC (%)	UC (%)	KC (%)
SCL	**32.56**	**94.39**	**29.52**	**97.57**
w/o normalization	32.30	94.2	29.32	97.93

A4 Self-Compositional Learning

A4.1 HOI Detection with Unknown Concepts

A4.1.1 Additional Comparisons

Table A.28 demonstrates SCL consistent improves the baseline (i.e., SCL without Self-Training). Here, we use the same concepts for a fair comparison. Thus, the recall is the same. Meanwhile, Table A.28 also shows Self-Training effectively improves the HOI detection. when we select all concepts to evaluate HOI detection, it is common zero-shot HOI detection, i.e., all unseen classes are known. Particularly, for application, one can directly detect unknown concepts with concept discovery from the model itself, e.g., Qpic [234]. Here, we mainly demonstrate different methods with the same concept confidence for a fair comparison.

A4.1.2 Novel Objects

In the main paper, we illustrate the result on two compositional zero-shot settings. Here, we further illustrate the effectiveness of HOI concept discovery for novel object HOI detection. Novel object HOI detection requires to detect HOI with novel objects, i.e., the object of an

TABLE A.28. Illustration of HOI detection with unknown concepts and zero-shot HOI detection with SCL. K is the number of selected unknown concepts. HOI detection results are reported by mean average precision (mAP)(%). We also report the recall of the unseen categories in the top-K novel concepts. K = all indicates the results of selecting all concepts, i.e., common zero-shot. $*$ means we train Qpic [234](ResNet-50) with the released code in zero-shot setting and use the discovered concepts of SCL to evaluate HOI detection with unknown concepts.

Method	K	Rare First				Non-rare First			
		Unknown	Known	Full	Recall (%)	Unknown	Known	Full	Recall (%)
Baseline	0	1.68	22.10	18.52	0.00	5.86	16.30	14.21	0.00
Baseline	120	3.06	22.10	18.29	10.83	6.16	16.30	14.27	21.67
Baseline	240	3.28	22.10	18.34	13.33	6.90	16.30	14.42	25.00
Baseline	360	3.86	22.10	18.45	15.83	7.29	16.30	14.50	30.83
Baseline	all	9.62	22.10	19.61	100.00	12.82	16.30	15.60	100.00
SCL	0	1.68	22.72	18.52	0.00	5.86	16.70	14.53	0.00
SCL	120	2.26	22.72	18.71	10.83	7.05	16.70	14.77	21.67
SCL	240	3.66	22.72	18.91	15.00	7.17	16.70	14.80	25.00
SCL	360	4.09	22.72	19.00	15.83	7.91	16.70	14.94	30.83
SCL	all	9.64	22.72	19.78	100.00	13.30	16.70	16.02	100.00

unseen HOI is never seen in the HOI training set. We follow [107] to select 100 categories as unknown concepts. The remaining categories do not include the objects of unseen categories. Here we use a unique object detector to detect objects. To enable the novel object HOI detection and novel object HOI concept discovery, we follow [107] to incorporate external objects (e.g.COCO [164]) to compose novel object HOI samples. Specifically, we only choose the novel types of objects from COCO [164] as objects images in the framework [107] for novel object HOI detection with unknown concepts.

Table A.29 demonstrates concept discovery largely improves the performance on unseen category from 3.92% to **11.41%** (relatively by 191%) with top 100 unknown concepts. We meanwhile find the recall increases to 41.00% with only the top 100 unknown concepts. Nevertheless, when we select all unknown concepts, the performance on unseen category is 17.19%. This shows we should improve the performance of concept discovery.

TABLE A.29. Illustration of the effectiveness of HOI concept discovery for HOI detection with unknown concepts (novel objects). K is the number of selected unknown concepts. HOI detection results are reported by mean average precision (mAP)(%). Recall is evaluated for the unseen categories under the top-k novel concepts. The last row indicates the results of selecting all concepts.

K	Unseen	Seen	Full	Recall (%)
0	3.92	19.45	16.86	0.00
100	11.41	19.45	18.11	41.00
200	12.40	19.45	18.28	48.00
300	13.52	19.45	18.46	52.00
400	13.52	19.45	18.46	52.00
500	13.91	19.45	18.53	56.00
600	13.91	19.45	18.53	56.00
all	17.19	19.45	19.07	100.00

TABLE A.30. Additional Comparison on HOI concept discovery. We report all performance using the average precision (AP) (%). UC means unknown concepts and KC means known concepts. SCL means self-compositional learning. SCL− means online concept discovery without self-training. SCL (COCO) means we train the network via composing between verbs from HICO and objects from COCO 2014 training set.

Method	HICO-DET		V-COCO	
	UC (%)	KC (%)	UC (%)	KC (%)
Random	12.52	6.56	12.53	13.54
language embedding	16.08	29.64	-	-
Re-Training	26.09	50.32	-	-
SCL− (COCO)	17.01	55.50	26.04	81.47
SCL (COCO)	31.92	86.43	27.90	90.04
SCL−	22.36	83.04	26.64	95.59
SCL	**33.26**	**93.06**	**29.52**	**97.57**

TABLE A.31. Illustration of the effectiveness of self-training on HOI detection based on ground truth box. Results are reported by mean average precision (%).

Method	Full	Rare	NonRare
SCL	**42.92**	**36.60**	**44.81**
w/o Self-Training	42.66	35.81	44.70

TABLE A.32. Illustration of the effectiveness of self-training for Qpic (ResNet-50). Results are reported by mean average precision (%). $*$ means we use the released code to reproduce the results for a fair comparison. S1 means Scenario 1, while S2 means Scenario 2.

Method	HICO-DET			V-COCO	
	Full	Rare	NonRare	S1	S2
GGNet [304]	23.47	16.48	25.60	-	54.7
ATL [107]	23.81	17.43	25.72	-	-
HOTR [131]	25.10	17.34	27.42	55.2	**64.4**
AS-Net[33]	28.87	24.25	30.25	-	53.9
Qpic [234]	29.07	21.85	31.23	58.8	61.0
Qpic* [234]	29.19	23.01	31.04	61.29	62.10
Qpic + SCL	**29.75**	**24.78**	**31.23**	**61.55**	**62.38**

A4.2 HOI Detection

One-Stage Method. We also evaluate SCL on Qpic [234], i.e., the state-of-the-art HOI detection method based on Transformer, for HOI detection. Code is provided in https://github.com/zhihou7/SCL. We first obtain concept confidence similar as Section 3.3.2 in the main paper. Denote $\hat{\mathbf{Y}}_v \in R^{N \times N_v}$ as verb predictions, $\hat{\mathbf{Y}}_o \in R^{N \times N_o}$ as verb predictions, we obtain concept predictions $\hat{\mathbf{Y}}_h$ as follows,

$$\hat{\mathbf{Y}}_h = \hat{\mathbf{Y}}_v \otimes \hat{\mathbf{Y}}_o. \tag{A.2}$$

Then, we update M according to Equation 2 and Equation 3 in the main paper. After training, we evaluate HOI concept discovery with M.

For self-training on Qpic [234], we use M to update the verb label $\mathbf{Y}_v \in R^{N \times N_v}$ for annotated HOIs. Here, we do not have composite HOIs because Qpic has entangled verb and object predictions, and we update verb labels with $\mathbf{M}$. Specifically, given an HOI with a verb labeled as $y_v \in R_v^N$ and an object labeled as $y_o \in R_o^N$, where $0 \leq y_o < N_o$ denotes the index of object category, we update y_v as follows,

$$\widetilde{y}_v = max(y_v + \mathbf{M}(:, y_o), 1) \tag{A.3}$$

where max means we clip the value to 1 if the value is larger than 1. Then, we obtain pseudo verb label $\widetilde{y}_v$ to optimize the samples of the HOI similar as Equation 7 (here, we only have annotated HOI samples). We think the running concept confidence $\mathbf{M}$ have **implicitly counted the distribution of verb and object in the dataset**. Meanwhile, the denominator in Equation 2 can also normalize the confidence according to the frequency, and thus ease the long-tailed issue. Thus, with the pseudo labels constructed from $\mathbf{M}$, we can re-balance the distribution of the dataset, which is a bit similar to re-weighting strategy [20, 45]. However, SCL does not require to set the weights for each class manually.

Table A.33 demonstrates SCL greatly improves Qpic on Unseen category on rare first zero-shot detection, while SCL significantly facilitates rare category on non-rare first zero-shot detection. In Full HOI detection on HICO-DET, Table A.32 shows SCL largely facilitates HOI detection on rare category. Particularly, the seen category in rare first setting includes 120 rare classes, while the seen category in non-rare first setting only includes 18 classes (all rare classes are in unseen category in non-rare first setting). Thus, SCL actually improves HOI detection for rare category. We think the concept confidence matrix internally learns the distribution of verb and objects and in the dataset. e.g., given an object, $\mathbf{M}$ illustrates the corresponding verb distribution.

TABLE A.33. Zero-Shot HOI detection based on Qpic. Results are reported by mean average precision (%). Here, we split the classes of HOI into four categories in zero-shot setting, i.e., Seen are categorized into rare and non-rare.

Method	Unseen	Rare	NonRare	Full
Qpic [234] (non-rare first)	21.03	19.12	25.59	23.19
Qpic+SCL (non-rare first)	**21.73**	**22.43**	**26.03**	**24.34**
Qpic [234] (rare first)	15.24	16.72	30.98	27.40
Qpic+SCL (rare first)	**19.07**	16.19	30.89	**28.08**

Two-Stage method. Considering the HOI concept discovery is mainly based on two-stage HOI detection approaches [111], it is direct and simple to evaluate the performance of self-training on HOI detection. Table A.31 demonstrates the HOI detection results on ground truth boxes. Noticeably, we directly predict the verb category, rather than HOI category. Thus, the baseline of HOI detection (i.e.visual compositional learning [111]) is a bit worse. We can find self-training also slightly improves the performance, especially on rare category.

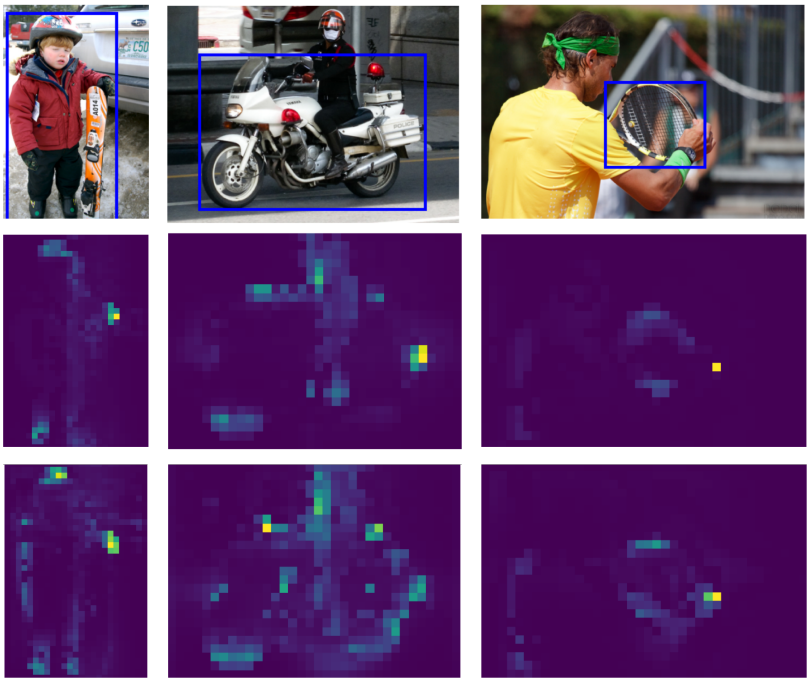

FIGURE A.10. Visualized Illustration of SCL+Qpic and Qpic [234].

A4.3 Visualization

In this section, we provide more visualized illustrations.

More Grad-CAM Visualizations Figure A.10 demonstrates the visualization of Qpic and Qpic+SCL: the second row is Qpic and the third row is Qpic+SCL, where we observe a similar trend to the Gram-CAM illustration in main paper.

Concept Visualization. We illustrate the visualized comparisons of concept discovery in Figure A.11. According to the ground truth and known concepts, we find some verb (affordance) classes can be applied to most of objects (the row is highlighted in the ground truth figure). This observation is reasonable because some kinds of actions can be applied to most of objects in visual world, e.g., hold. As shown in Figure A.11, there are many false positive predictions in the results of affordance prediction, and affordance prediction tends to overfit the known concepts, especially those with frequently appeared verbs. Methods of online HOI concept discovery on V-COCO have fewer false positive predictions compared to affordance prediction. However, the two methods tend to predict concepts composed of

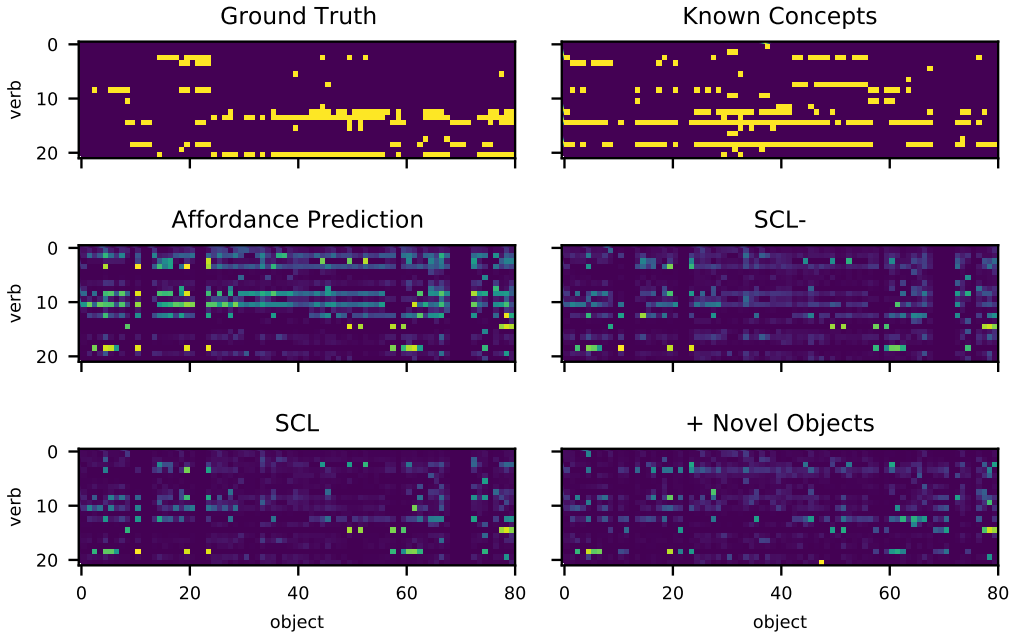

FIGURE A.11. Visualized Comparison of different methods on V-COCO dataset. The column is the object classes and the row represents the verb classes. Known Concepts are the concepts that we have known. SCL− means online concept discovery without self-training. For better illustration, we filter out known concepts in proposed methods. "+ Novel Objects" means self-training with novel object images.

frequent verbs in known concepts due to the verb and object imbalance issues in HOI dataset [110]. Particularly, the false positive predictions are largely eased with self-training (e.g., the top right region). In addition, the blank columns in Figure A.11 are because there are only 69 objects in V-COCO training set, and we can ease it via training network with additional object images [107] as illustrated in the last figure of Figure A.11. See more visualized results on HICO-DET and V-COCO in the supplemental material. Particularly, we further notice there are dependencies between verb classes (See verb dependency analysis).

A4.4 Additional Concept Discovery Approaches

We provide More comparisons in this Section. For a fair comparison with ATL [107] (i.e., affordance prediction), we use the same number of verbs (21 verbs) on V-COCO. The code

includes how to convert V-COCO to 21 verbs, i.e.merge "_instr" and "_obj" and remove actions without object (e.g., stand, smile, run).

Language embedding baseline. In the main paper, we illustrate a random baseline. Here we further illustrate the results with language embedding [203]. Different from extracting verb/object features from real HOI images, we use the corresponding language embedding representations of verb/object as input, i.e.discovering concepts from language embedding. Table A.30 shows the performance is just a bit better than random result, and is much worse than online concept discovery. Similar to the main paper, when we evaluate the unknown concepts, we mask out the known concepts to avoid the disturbance from known concepts.

Re-Training. We first train the HOI model via visual compositional learning [111], and then predict the concept confidence. Next, we use the predicted concept confidence to provide pseudo labels for the composite HOIs. Table A.30 shows the performance of Re-Training is worse than SCL.

With COCO dataset. Table A.30 also demonstrates the baseline (SCL$-$) with COCO datasets has poor performance on concept discovery. We think it is because the domain shift between COCO dataset and HICO-DET dataset. However, SCL still achieves significant improvement on concept discovery.

Qpic+SCL. The details are provided in Section D.

A4.5 Object Affordance Recognition

SCL requires more iterations to converge, and achieves better performance on object affordance recognition. Table A.34 shows the performance of the model without self-training does not improve with more training iterations.

TABLE A.34. Comparison of object affordance recognition with HOI network (trained on HICO-DET) among different datasets. Val2017 is the validation 2017 of COCO [164]. Here we illustrate the result of SCL− under different training iterations.

Method	Type	Val2017	Obj365	HICO	Novel
SCL− (0.8M iters)	U	43.61	41.14	47.56	14.46
SCL− (1.5M iters)	U	44.07	39.05	50.27	10.19

APPENDIX B

Appendix of Chapter 3

B1 Additional Experiments

In this section, to better demonstrate the effectiveness of BatchFormerV1, we provide more experimental results.

B1.1 Long-Tailed Recognition

We notice BatchFormer mainly improves PaCo on Many category on CIFAR-100-LT (imbalance ratio 100). We thus conduct additional ablation study on BatchFormer for PaCo. Here, we remove the PaCo loss with Balanced loss [44] to build the baseline. we observe consistent results in the Table B.1. Noticeably, the baseline without PaCo loss is even better than the one in the main paper.

B1.2 Generalized Zero-Shot Learning

We also evaluate BatchFormerV1 on generalized zero-shot learning task. Specifically, we report the accuracy of “seen", “unseen", and the harmonic mean of them (unseen and seen).

TABLE B.1. Illustration of BatchFormer without PaCo loss [44] on CIFAR-LT-100.

Method	100				200			
	All	Many	Med	Few	All	Many	Med	Few
Baseline [52]	52.0	68.1	53.2	31.6	47.31	67.8	52.6	27.3
+ BFV1	**52.6**	**68.7**	53.2	**33.1**	**48.13**	**68.9**	**53.1**	**28.2**

TABLE B.2. Illustration of BatchFormerV1 for MoCo on VOC2007 [57]. Here, for fair comparison, we use the official code of MoCo-v2 and MoCo-v3 to run all experiments in the same setting.

Methods	AP	AP50	AP75
MoCo-v2* [35]	56.4	82.1	63.1
+ BFV1	**56.7**	**82.0**	**63.6**
MoCo-v3 [35]	46.6	78.2	48.9
+ BFV1	**48.0**	**78.8**	**51.1**

We perform experiments on one of the most popular datasets for generalized zero-shot learning, CUB [247], which includes 11,788 images from 200 bird species. We build a baseline with the released code of [254] and achieve better results than [254]. As shown in Table **??**, the proposed BatchFormerV1 achieves a new state-of-the-art on Unseen and Harmonic mean.

B1.3 Self-Supervised Learning

Object detection on VOC2007. We also evaluate Object Detection of MoCo on VOC2007 [57] in Table B.2. Similar as MoCo-v2 [35], we use the pre-trained model to fine-tune Faster-RCNN on VOC2007 based on Detectron2 [269]. We find MoCo-v3 achieves worse result on VOC2007. However, BatchFormerV1 consistently improves the object detection on VOC2007. Here, we train MoCo-v2 for 200 epochs, and MoCo-v3 for 100 epochs. Specifically, we think that the number of training epochs (only 100 epochs) of MoCo-v3 might limit the performance on VOC2007.

B1.4 Image Recognition

Table B.3 demonstrates BatchFormerV1 for image classification. We find BatchFormerV1 achieves comparable performance among ResNet50 and Swin Transformer [170]. This shows BatchFormerV1 does not degrade the performance when the distribution of data is balanced. Table B.3 also shows our updated experiment on ViT [53, 239]. We find that BatchFormerV1 effectively improves the baseline. Here, we apply BatchFormerV1 before the last normalization layer since there is no average pooling.

TABLE B.3. Illustration of BatchFormerV1 for image recognition.

Method	Epochs	Top-1	Top-5
ResNet50 [268]	200	78.9	
+ BFV1	200	78.9	-
Swin-T [170]	300	81.3	95.5
+ BFV1	300	81.3	95.6
DeiT-S [239]	300	79.8	95.0
+ BFV1	300	**80.3**	95.1

TABLE B.4. Illustration of BatchFormerV1 for domain generalization on PACS [149]. Here, the baseline is from [124]. SWAD [25] is reproduced based on the official code.

Method	art_paint	cartoon	sketches	photo	Avg.
Baseline	81.3±0.7	76.1±0.6	75.5±2.6	**95.4** ±0.2	82.0
+ BFV1	**82.4**±1.5	**76.4**±1.2	**75.7**±1.0	95.1±0.4	**82.4**
CORAL [231]	79.2±1.7	**75.5** ±1.1	71.4±3.1	94.7±0.3	80.2
+ BFV1	**80.6**±0.9	74.7±1.9	**73.1**±0.3	**95.1**±0.3	**80.9**
IRM [4]	**81.0**±0.6	**71.4**±4.1	68.1±7.1	95.0±0.6	78.9
+ BFV1	78.9±3.1	71.0±7.1	**71.5**±2.8	**96.0**±0.3	**79.4**
V-REx [141]	80.8±1.8	75.3±1.4	73.3±0.9	95.9±0.0	81.3
+ BFV1	**82.0**±0.3	**76.3**±0.7	**75.2**±1.7	95.3±0.1	**82.2**
MixStyle [309]	81.7±0.1	**76.8**±0.0	80.8±0.0	93.1±0.0	83.1
+ BFV1	**84.8** ±0.4	75.3±0.0	**81.1** ±0.4	**93.6**±0.0	**83.7**
SWAD* [25]	83.1±1.5	75.9±0.9	77.1±2.4	95.6±0.6	82.9
+ BFV1	**84.3**±0.8	**76.9** ±1.2	**78.2**±1.8	95.7±0.6	**83.9**
ResNet50					
V-REx [141]	83.8±4.8	**81.0**±0.0	**97.7**±0.4	77.7±3.1	85.0
+ BFV1	**87.3** ±5.0	80.2±4.6	97.1±1.7	**77.9**±4.4	**85.6**
IRM [4]	88.2±0.6	79.8±1.0	97.6±0.5	77.6±0.7	85.8
+ BFV1	**89.0**±0.98	**80.1** ±1.0	**98.0**±0.4	**79.8**±0.4	**86.8**
SWAD [25]	89.4±0.7	83.7±1.2	**97.7**±0.6	82.5±0.8	88.1
+ BFV1	**90.2**±0.5	**84.0**±1.0	97.3±0.3	**83.0**±0.6	**88.6**

B1.5 Domain Generalization

We provide more experimental results based on [124] in Table B.4. Experiments on Office-Home, VLCS, TerraIncognita are provided in Table B.5, Table B.6 and Table B.7 respectively. If not otherwise stated, the default backbone is ResNet-18.

TABLE B.5. Illustration of BatchFormerV1 for domain generalization on OfficeHome using [25] as the baseline.

Method	Art	Clipart	Product	RealWorld	Avg.
SWAD* [25]	54.5±0.8	49.4±0.1	70.9±0.1	72.7±0.2	62.1
+ BFV1	**57.8**±0.1	**51.0**±0.1	**73.4**±0.2	**75.1**±0.1	**64.3**
ResNet-50					
IRM [4]	66.8±0.2	54.9±0.8	77.5±0.7	80.5±0.4	69.9
+ BFV1	**67.7** ±0.2	**55.5**±0.8	**78.4**±0.5	**81.0**±0.3	**70.6**
SWAD* [25]	65.9±0.8	58.0±0.1	78.5±0.5	80.2±0.7	70.6
+ BFV1	**66.7**±0.3	57.9±0.3	**79.2**±0.4	**80.6**±0.7	**71.1**

TABLE B.6. Illustration of BatchFormerV1 for domain generalization on VLCS using [25] as the baseline. The backbone is ResNet-18.

Method	Caltech101	LabelMe	SUN09	SUN09	Avg.
SWAD* [25]	97.2±1.4	61.4±0.1	71.2±1.7	75.5±0.8	76.3
+ BFV1	97.2±0.8	61.3±1.1	**71.7**±1.0	**77.4**±0.4	**76.9**

TABLE B.7. Illustration of BatchFormerV1 for domain generalization on TerraIncognita using a recent work [25] as the baseline. The backbone is ResNet-18.

Method	Art	Clipart	Product	RealWorld	Avg.
SWAD* [25]	47.6±3.0	33.8±4.5	53.6±1.8	33.3±0.6	42.1
+ BFV1	**49.8**±1.8	**40.3**±2.0	**55.2**±1.2	**34.0**±1.1	**44.8**

TABLE B.8. Illustration of BatchFormerV1 for domain adaption on VisDA2017 [202]. The backbone is ResNet-101. Experiments are based on [124].

Method	Synthetic $->$ Real
MDD [297]	76.8±1.5
+ BFV1	**77.8** ±2.0

B1.6 Domain Adaption

We also demonstrate BatchFormerV1 on Domain Adaption on VisDA2017 [202]. Table B.8 shows BatchFormerV1 effectively improves the corresponding baseline, i.e., MDD [297].

Method	Insert Position	All	Many	Med	Few
Deit-S	-	32.8	52.5	24.3	7.0
+ BFV2	1-12	**35.5**	**55.4**	**27.2**	**8.6**
+ BFV2	8-12	34.7	54.7	26.3	7.2
+ BFV2	4-12	35.5	55.3	26.8	8.4
+ BFV2 (non-shared)	1-12	35.2	55.3	26.7	8.3

TABLE B.9. Illustration of Deit-S on ImageNet-LT. By default, we share all the modules among different layers on this experiments. BatchFormerV2 (non-shared) indicates we do not share the modules among different layers. We observe sharing BatchFormerV2 on image classification achieves a bit better performance.

B2 Additional Experiments

In this section, we provide more experimental results for BatchFormerV2.

B2.1 Long-Tailed Recognition

In Table B.9, we show the model performances with BatchFormerV2 on ImageNet-LT. Here, all experiments are based on DeiT-S [238] and we do not use any re-balance strategies. We find that BatchFormerV2 can significantly improve model performance comparing with the baseline.

B2.2 3D Hand Reconstruction

In addition to object detection and panoptic segmentation, we further provide results on another important pixel-level task, i.e., 3D Reconstruction. Specifically, we use the popular 3D hand reconstruction benchmark, i.e., FreiHAND [317] and evaluate the proposed BatchFormerV2 module for 3D hand mesh reconstruction using a recent state-of-the-art method [162], MeshGraphormer. Here, we report the performance on FreiHand dataset under single-scale inference for a quick evaluation. As shown in Table B.10, the proposed BatchFormerV2 module clearly improves the baseline by over 1.% on both two metrics, PA-MPVPE and PA-MPJPE.

TABLE B.10. BatchFormerV2 for 3D Hand Mesh Reconstruction. $*$ indicates we train the network with the released official code of Mesh-Graphormer [162].

Method	PA-MPVPE ↓	PA-MPJPE ↓	F@5 mm ↑	F@15 mm ↑
Lin *et al*. [162]*	62.8	64.3	74.7	98.3
+ BFV2	**61.3**	**62.6**	**75.4**	**98.5**

TABLE B.11. MAE with BatchFormerV2. * indicates we use the released code to MAE for 800 epochs. We illustrate the result of Linear Probe.

Method	Epochs	ViT-Base	ViT-Large
MAE [96]*	800	65.6	73.5
+ BFV2	800	**66.1**	73.9

B2.3 Masked AutoEncoder

Here, we also utilize a simple experiment to evaluate BatchFormerV2 on recent self-supervised learning framework, i.e., Masked Auto Encoder [96](MAE). We insert BatchFormerV2 into all layers in the decoder in MAE [96]. We use the ViT-Base model to evaluate BatchFormerV2. Here, due to the computation limitation, we train the network 800 epochs with the released code of [96], and verify the model via linear probe. All other hyper-parameters are following [96]. Table B.11 demonstrates BatchFormerV2 is also beneficial for MAE. Without bells and whistles, BatchFormerV2 improves the baseline by 0.5%.

B2.4 Without Two-Stream Strategy

We conduct experiments about the two-stream training strategy. As shown in Table B.12, the performance significantly drops if we use a single stream with BatchFormerV2, since the distribution between with and without BatchFormerV2 changes in each layer. Therefore, a single-stream network can not enable the inference without BatchFormerV2 modules.

TABLE B.12. Ablation study on two-stream training strategy. Here, TS indicates the two-stream strategy. The backbone is ResNet-50.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
w/ TS	**45.5**	**64.3**	**49.8**	**28.3**	**48.6**	**59.4**
w/o TS	12.3	33.9	6.3	5.3	21.1	14.6

B2.5 Mini-batch Inference

In our experiment, we remove BatchFormerV2 for inference, since we can not always assume a mini-batch of testing data. In Table B.13, we also show the inference results of BatchFormerV2 with a mini-batch testing data. Here, we apply BatchFormerV2 in the first layer, and use the model to evaluate via mini-batch inference. We find that “inference without BatchFormerV2" achieves similar performance comparing with “inference with BatchFormerV2". Therefore, we consider that the two-stream strategy enables the semantically invariant learning, and thus make it possible to remove BatchFormerV2 during inference.

TABLE B.13. Ablation study on mini-batch inference with BatchFormerV2. “BFV2 (BI)” indicates mini-batch inference. The backbone is ResNet-50.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
BFV2	45.6	64.5	49.8	28.3	48.8	59.7
BFV2 (BI)	45.6	64.4	49.8	28.3	48.7	59.7

B2.6 Inference with Feature Fusion

We observe that the inference strategy with and without BatchFormerV2 achieve similar performance in previous section. Taking a further step, we design a feature fusion to evaluate the feature around the current feature space. Specifically, let x denote the feature without BatchFormerV2 (i.e., the feature we use for inference in main paper), $\hat{x}$ denote the feature with BatchFormerV2 during inference, we then update x as follows,

$$x = \lambda x + (1 - \lambda)\hat{x}, \tag{B.1}$$

where λ indicates the weight of feature without BatchFormerV2. Interestingly, Table B.14 shows that the models with different λ achieve similar performances. Here, we fix the order of images during inference. We think this experiment shows all the features between the features with and without BatchFormerV2 in the feature space are valid for prediction, i.e., the two-stream training strategy also augment the feature space. As illustrated in Figure B.4, these two features are actually different.

TABLE B.14. Feature fusion between with and without BatchFormerV2 during inference. λ indicates the weights for the feature without BatchFormerV2.

Method	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
Fusion	45.6	45.6	45.6	45.6	45.6

TABLE B.15. Illustration the effect of mixup [293] and cutmix [287] on BatchFormerV2. Experiments are conducted on Tiny-ImageNet. We follow the same experimental setups described in DeiT [238]. "w/o mixup" indicates that we remove both mixup and cutmix.

Model	#Params	Input	Top-1	Top-5
DeiT-S [238]	22M	224^2	81.8	94.1
+ BFV2	22M	224^2	**82.9**	**94.3**
DeiT-S [238] (w/o mixup)	22M	224^2	75.80	89.6
+ BFV2 (w/o mixup)	22M	224^2	**78.2**	**95.6**
DeiT-S [238] (w/o cutmix)	22M	224^2	79.8	92.5
+ BFV2 (w/o cutmix)	22M	224^2	**81.1**	**92.3**

B2.7 Classification Without Mixup

We notice that there are frequent training crashes when applying BatchFormerV2 with multiple layers on large datasets. We also evaluate BatchFormerV2 without mixup on Tiny-ImageNet. Except for BatchFormerV2 modules, all configurations follow [238]. We run the experiments on four NVIDIA V100 (16GB) GPUs. Table B.15 demonstrates that without using cutmix [287] or mixup [293], BatchFormerV2 significantly improves the baseline with a larger margin.

B2.8 Shared BatchFormerV2 Modules

We can also share BatchFormerV2 modules among different layers on image classification. The motivation behind of this setting is that we further encourage different layers to discover the same batch attention pattern. Here, we illustrates the effect of sharing modules among different layers on ImageNet-LT. As shown in Table B.9, it achieves a bit better performance on small datasets if we share the modules among different layers. This is different from the observation on object detection, possibly because that sharing modules plays a role of regularization which benefits the learning on small datasets. Meanwhile, it is also challenging

to optimize the DeiT model with BatchFormerV2 modules if we do not share the modules among different layers. In this thesis, we mainly focus on a general BatchFormerV2 module which can be well generalized for different levels of tasks. We leave the further exploration of sharing strategy, and crash collapse on ImageNet when inserting BatchFormerV2 into multiple layers to future work.

B3 Visualization

In this section, we provide more visualization results for BatchFormerV1 and BatchFormerV2.

B3.1 Visualization Results on ImageNet-LT.

We provide more comparisons in Figure B.1, Figure B.2, and Figure B.3, where the top 100 classes on ImageNet are chosen for demonstration.

B3.2 Visualization of Features

Table B.13 shows BatchFormerV2 without mini-batch inference achieves similar performance to that with mini-batch inference. To further analyze this phenomenon, we visualize the feature maps between with and without BatchFormerV2. As shown in Figure B.4, we find that there are significant differences (i.e., different distribution) between the above-mentioned two feature maps during inference. We think the two feature maps represent similar semantics though the distribution is diverse, i.e., representing similar semantics for the the same prediction modules.

B3.3 Visualization of Panoptic Segmentation

We further provide more panoptic segmentation examples in Figure B.5. We find that BatchFormerV2 usually helps object segmentation and improves the segmentation boundaries of the stuffs.

B3.4 Visualization of Attention

Visualization of the multi-head self-attention provides rich semantic interpretations. Here we provide more observations from the visualization of attentions in Figure B.6. First, we observe that the images with objects usually have higher attentions to other images, i.e., the objects are usually highlighted as illustrated in Figure B.6. Second, and more importantly, *the attention of background in current image is suppressed if the corresponding positions in other images have objects*. For example, the region (grass) under the zebra in row 2 in Figure B.6 is suppressed because there is a person in the first image. There is a region suppressed like a person in “row 4, column 3" in Figure B.6 because there is a person in second column. However, if the region has objects, the region will not be suppressed. For example, the airplane is highlighted in “row 4, column 1" though the corresponding region is object in “row 4, column 2".

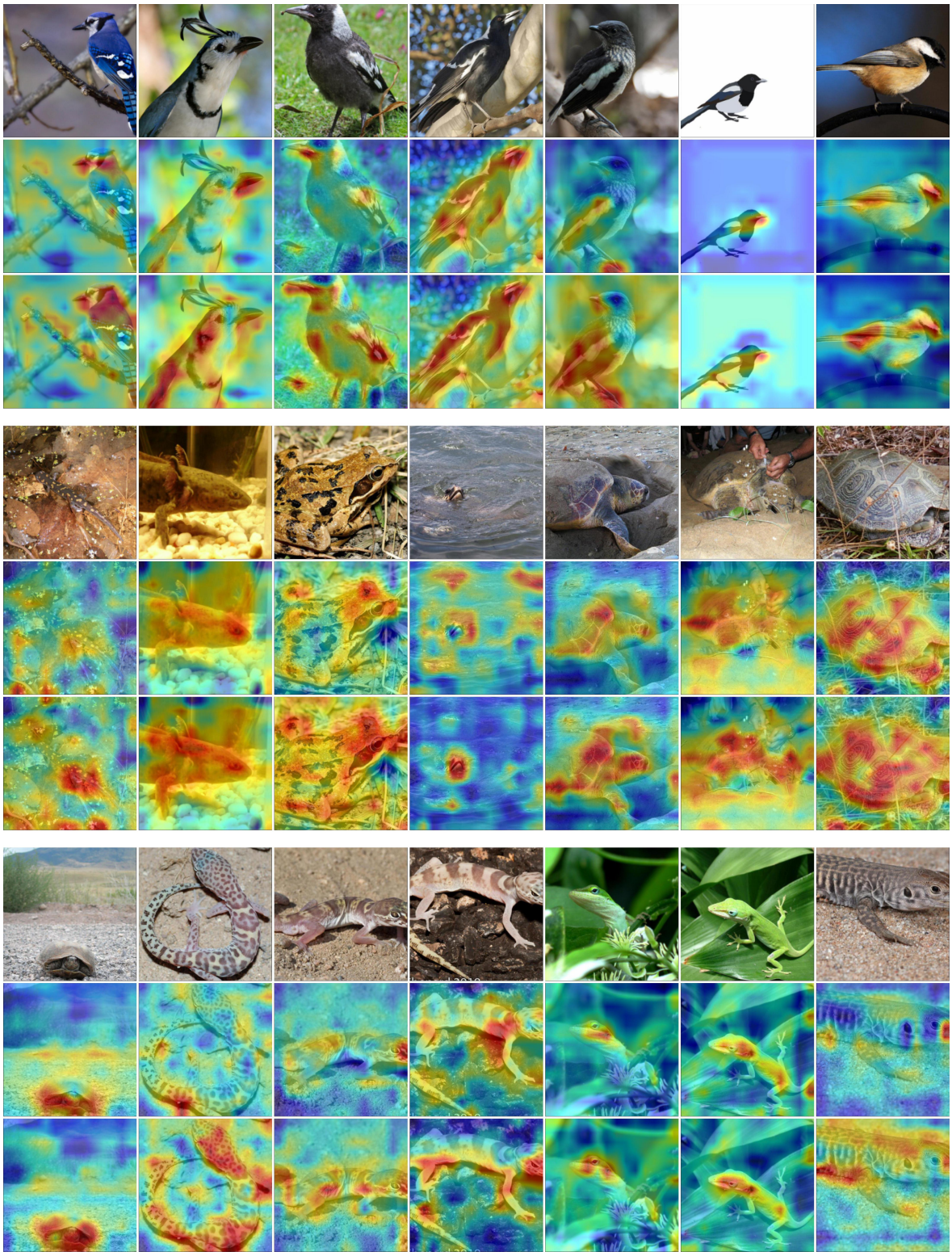

FIGURE B.1. Additional visualization results of BatchFormerV1 on low-shot test images based on [212].

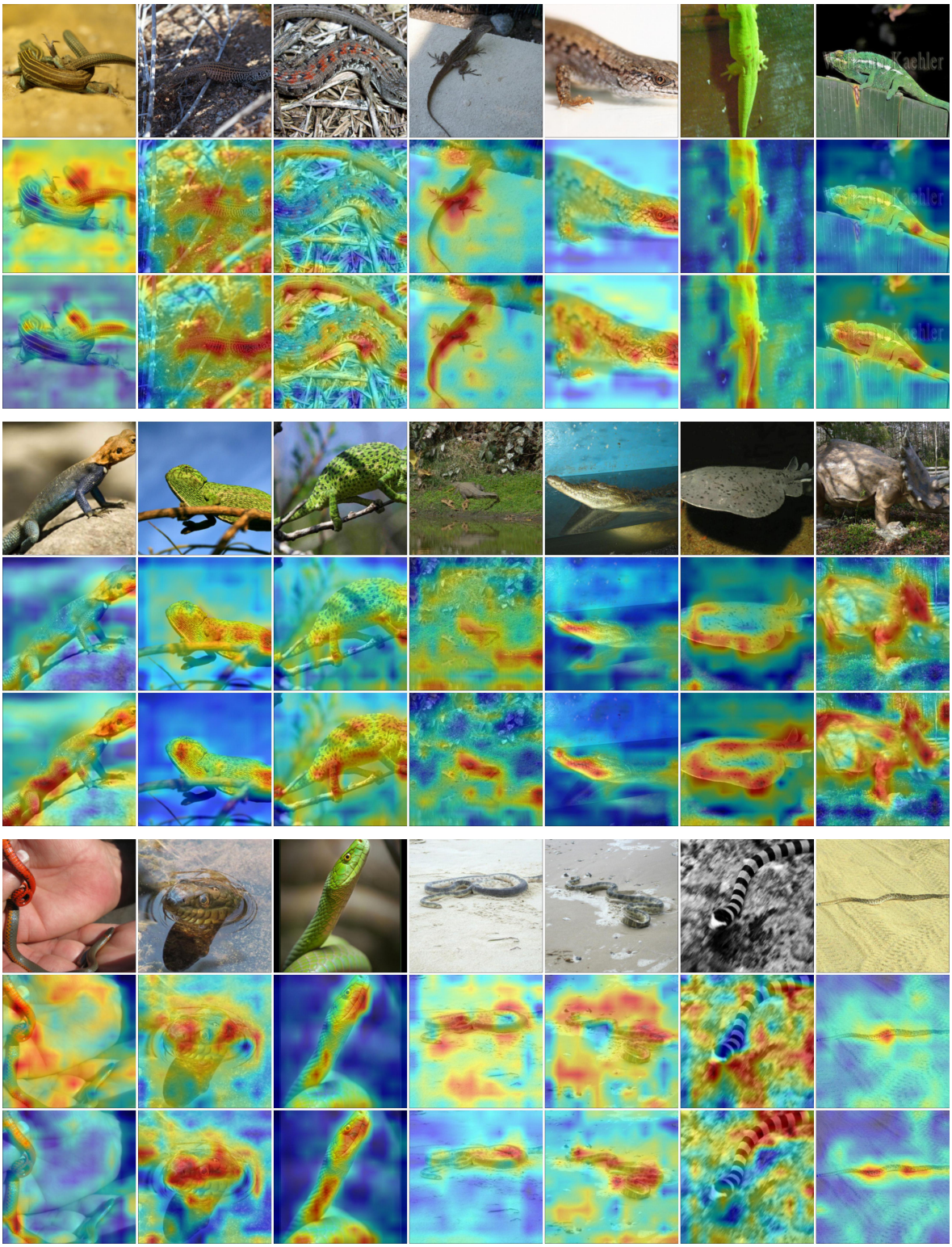

FIGURE B.2. Additional visualization results of BatchFormerV1 on low-shot test images based on [212].

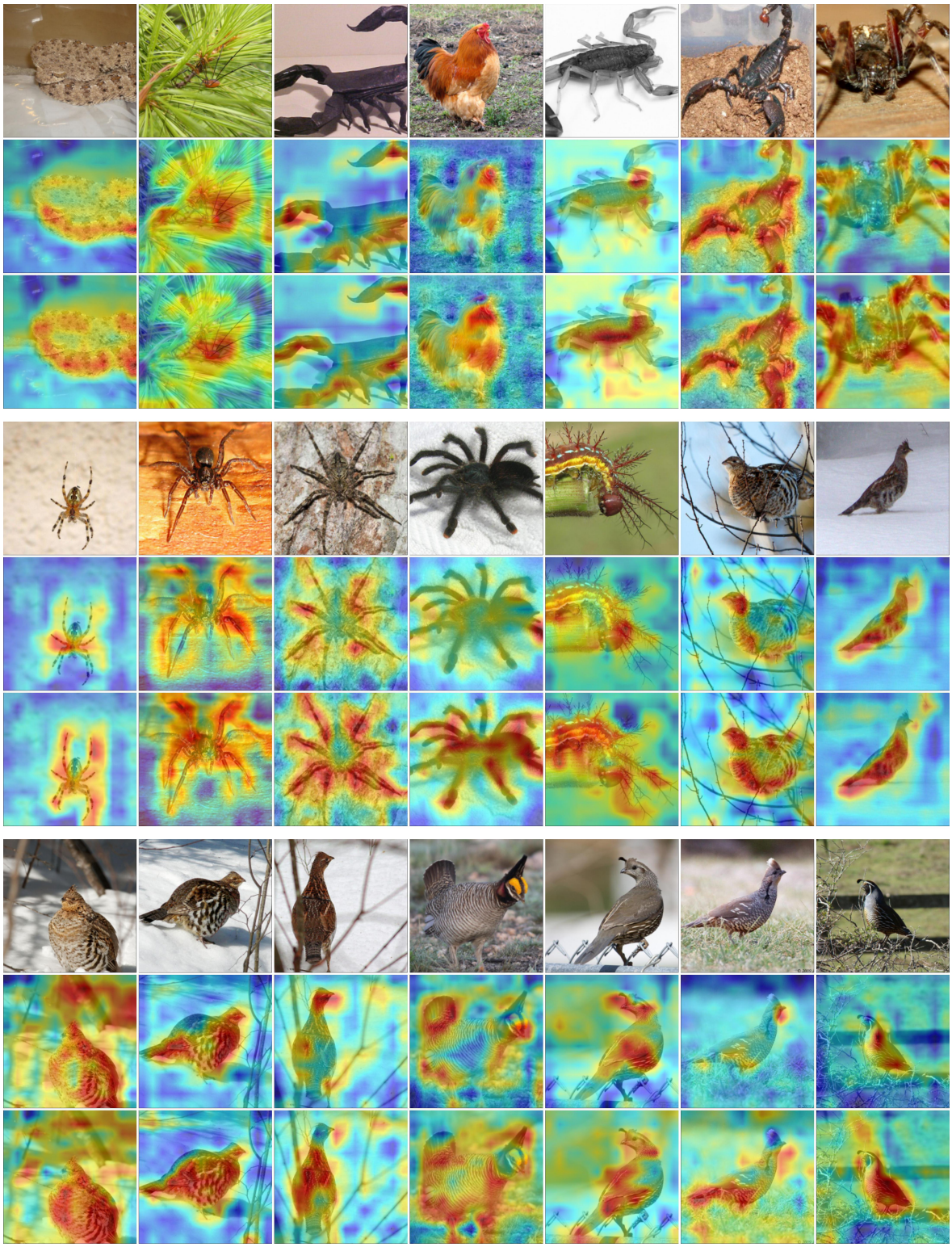

FIGURE B.3. Additional visualization results of BatchFormerV1 on low-shot test images based on [212].

FIGURE B.4. Visualization of the difference between the representations with and without BatchFormerV2 during inference. Here, we choose the largest feature map and use the model that we trained with BatchFormerV2 which is inserted into the first Transformer Encoder layer. The first row is image, the second row is the feature without BatchFormerV2, and the last row indicates the feature with BatchFormerV2 (mini-batch inference).

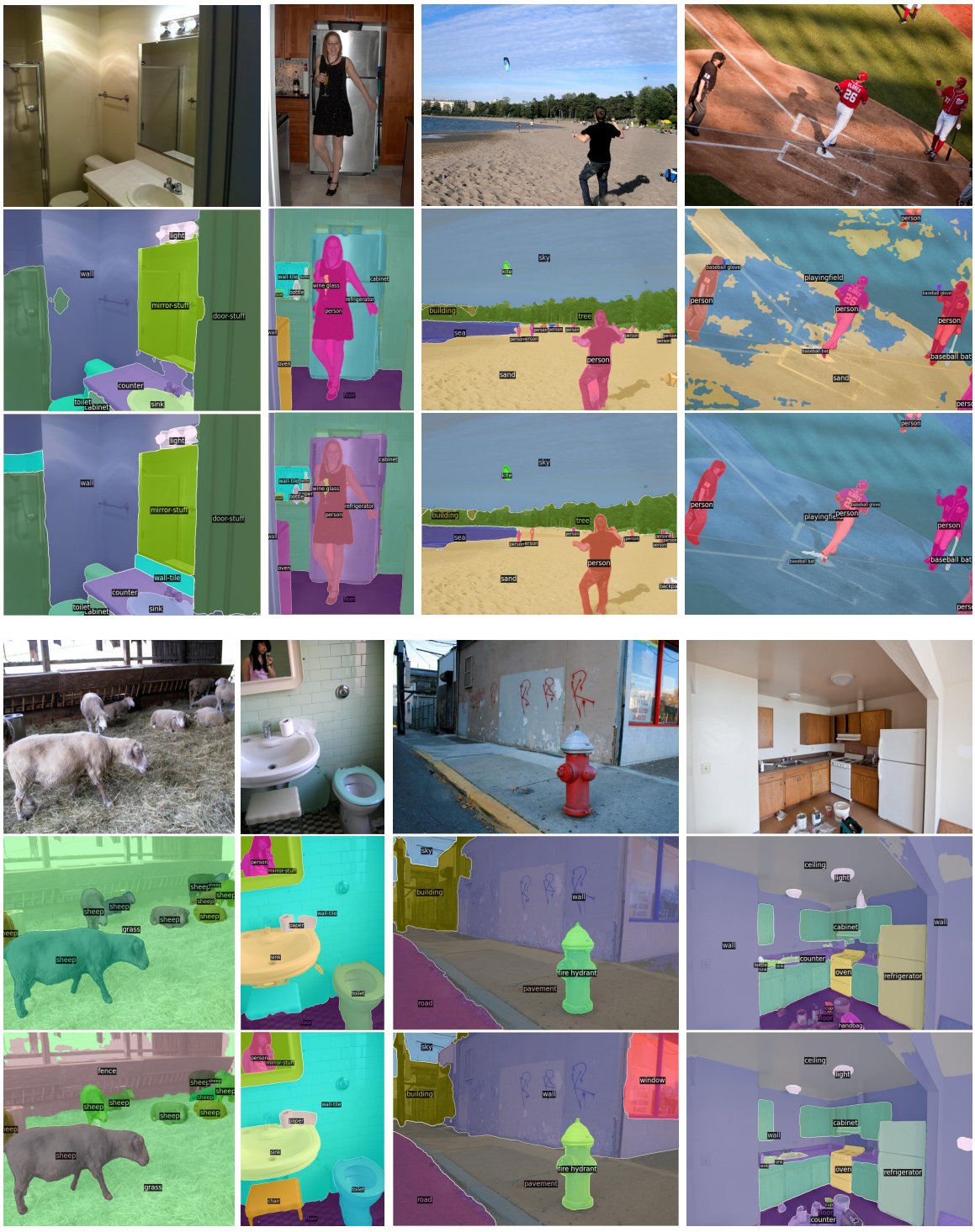

FIGURE B.5. Visualization of additional panoptic segmentation examples. The first row is original image, the second row is DETR and the thrid row is DETR with BatchFormerV2.

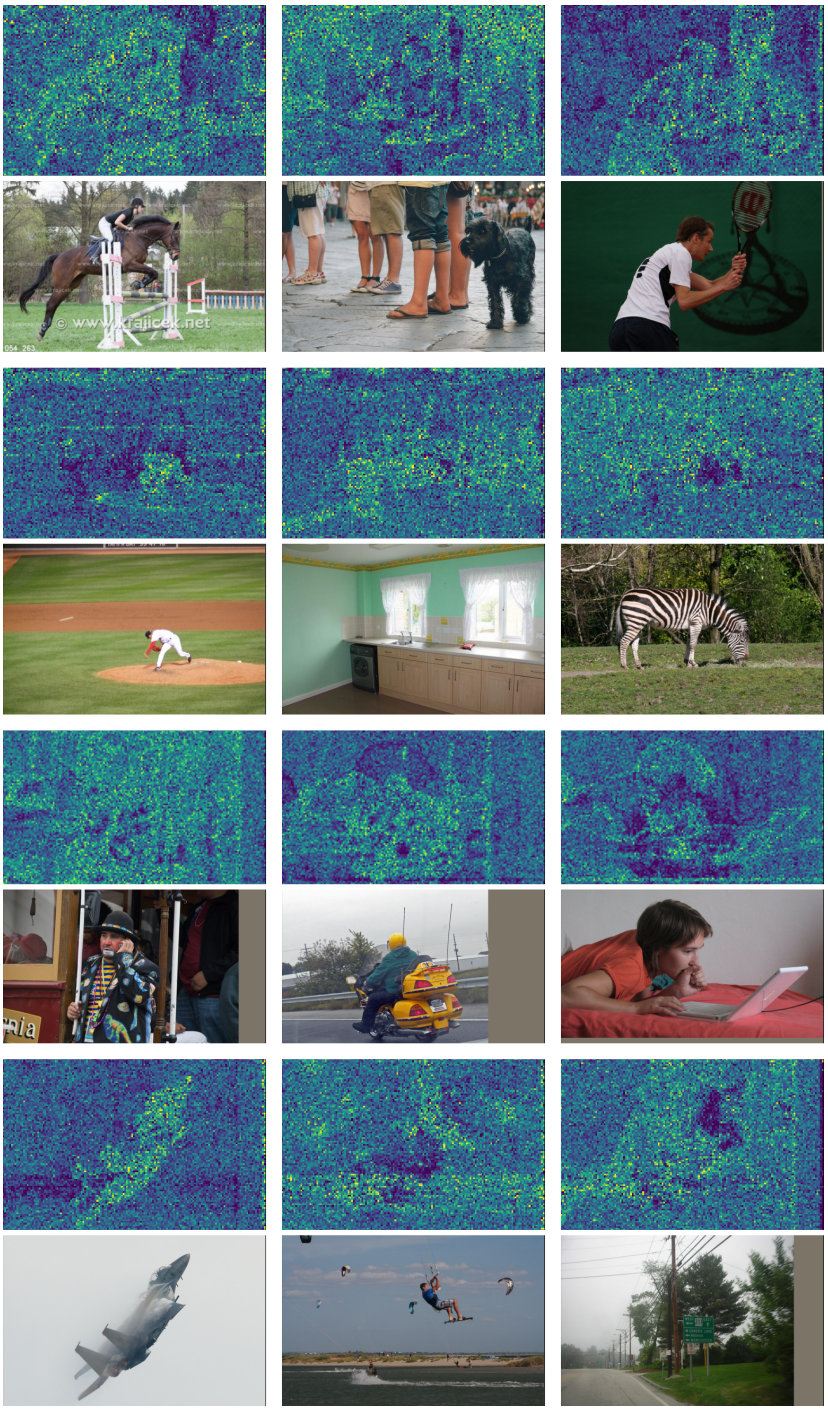

FIGURE B.6. Visualization of self-attention in the same mini-batch. Each row represents a mini-batch during inference. The model and settings are the same as those in Figure 5 in main paper

APPENDIX C

Appendix of Chapter 4

C1 More Implementation Details

To construct a "pseudo bone" for the object, and make it more convenient to implement and calculate the distance between the points in the ray and the object bone, we build the "pseudo bone" as a segment which starts from the center point of the object and ends at a point very close to the center point (≤ 0.01). Then, we can equally treat the "pseudo bone" as body bones in the code implementation. In our experiment, the dimension of the latent code is 64.

Neural Human-Object Deformation Similar to [267, 151], rather than the surface points as traditional LBS, we skin all points in the 3D space with $\boldsymbol{w}_{bg} \cdot \boldsymbol{I}$, which stops deforming the points in the background and empty space in Equation (2) in main paper. After Equation (3), the same as [151, 37], we analytically compute the gradients of the network parameters for the inverse skinning. During volumetric rendering, similar to previous works [37, 151], we choose the density and color for the observed point $\boldsymbol{x}_v$ as follow,

$$\boldsymbol{c}_v = \boldsymbol{c}^{'}_{c,m}, \boldsymbol{\sigma}_v = \boldsymbol{\sigma}^{'}_{c,m}, \tag{C.1}$$

where $m = argmax_i(\boldsymbol{\sigma}^{'}_{c,i}), 0 \leq i < K$. we then use $(\boldsymbol{c}_v, \boldsymbol{\sigma}_v)$ for volumetric rendering.

Training loss The overall loss function is

$$\mathcal{L} = \mathcal{L}_{img} + \lambda \mathcal{L}_w + \beta \mathcal{L}_{\Delta} \tag{C.2}$$

where $\mathcal{L}_{img}$ indicates image loss similar to [183], $\mathcal{L}_w$ represents the loss to encourage the onehot skinning weights $\boldsymbol{w}$, $\mathcal{L}_{\Delta}$ is to encourage the non-linear deformation term $F_{\Theta_{\Delta}}(\boldsymbol{x}_c, \boldsymbol{P})$

close to zero. Both $\mathcal{L}_w$ and $\mathcal{L}_\Delta$ are MSE losses. The learning rate and training iteration are provided in the experimental sections. For ARAH, we follow the default hyper-parameters.

C2 Baseline Method Details Analysis

In our experiments, we leverage two state-of-the-art Animatable Avatar methods (i.e., TAVA [151] and ARAH [255]) as our baselines. We find the both the two methods fail to animate the object. Meanwhile, we also apply the neural human-object deformation method in ARAH, which we named it as ARAH$*$. In ARAH$*$, we treat the object as a unique bone, and use the object meshes to guide the optimization of SDF model. we observe the model fails to reconstruct the geometry of the complex object (e.g., chairs). However, we notice the implicit model (SDF) of simple objects (e.g., yogaball, suitcase, boxes) can be successfully reconstructed and animated.

Besides, we devise a baseline method which first localize the object position, and then leverage the implicit object code embedding to model the rotation of the object. According to the visualization in Figure 3 in main paper, we notice the model can reconstruct the simple object, e.g., yogaball. For the complex objects, e.g., chairwood, it fails to reconstruct the object.

Lastly, we notice the "chairblack", "suitcase" and "backpack" are usually black, which is similar to the background. Therefore, the numbers do not always indicate the results well.

C3 Benchmark Construction

In our experiment, we randomly choose one action as validation set for each subject-object pair to evaluate the performance on out-of-distribution poses.

Here we provide the interactions splits of our experiments in Table C.1. For boxes, we randomly split the frames into training set and validation set because there is only a single interaction for each box. For compositional animation, we select "yogaball", "chairblack", "chairwood","tablesquare", "tablesmall", "suitcase", "boxmedium", "boxlarge", "boxsmall"

TABLE C.1. Dataset splits for novel pose animation.

Objects	training set	validation set
backpack	Sub01_backpack_hug,Sub01_backpack_back	Sub01_backpack_hand
chairwood	Sub01_chairwood_hand,Sub01_chairwood_sit	Sub01_chairwood_lift
chairblack	Sub01_chairblack_lift,Sub01_chairblack_hand	Sub01_chairblack_sit
suitcase	Sub01_suitcase_lift	Sub01_suitcase
tablesmall	Sub01_tablesmall_lift,Sub01_tablesmall_move	Sub01_tablesmall_lean
tablesquare	Sub01_tablesquare_hand,Sub01_tablesquare_lift	Sub01_tablesquare_sit
yogaball	Sub01_yogaball	Sub01_yogaball_play

TABLE C.2. Dataset splits for compositional animation.

training set	novel action validation
Sub01_chairwood_hand, Sub01_chairwood_lift, Sub01_tablesmall_lean, Sub01_tablesmall_lift, Sub01_yogaball_play, Sub02_boxmedium_hand, Sub02_boxsmall_hand, Sub02_chairblack_hand, Sub02_chairblack_lift, Sub02_suitcase_ground, Sub02_tablesquare_sit, Sub02_tablesquare_lift, Sub01_boxlarge_hand	Sub01_yogaball, Sub02_suitcase_lift, Sub01_chairwood_sit, Sub02_chairblack_sit, Sub01_tablesmall_move, Sub02_tablesquare_move, Sub01_suitcase
novel object validation	novel action object validation
Sub01_chairblack_sit, Sub02_chairwood_sit, Sub01_suitcase_lift, Sub02_yogaball_sit, Sub02_tablesmall_move, Sub01_tablesquare_hand	Sub01_chairblack_hand, Sub01_chairblack_lift, Sub02_chairwood_hand, Sub02_yogaball_play, Sub02_tablesmall_lean, Sub02_tablesmall_lift, Sub01_tablesquare_sit, Sub01_tablesquare_lift, Sub02_boxlarge_hand, Sub01_boxmedium_hand, Sub01_boxsmall_hand

from BEHAVE [16] to construct the benchmark. Table C.2 present the splits of compositional animation.

TABLE C.3. Human-Object Animation for the "boxlong". This is a complementary table to Table 2 in main paper.

Method	boxlarge		boxlong		boxmedium		boxsmall		boxtiny	
	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
TAVA [151]	22.6	0.949	26.8	0.966	25.9	0.967	26.8	0.970	27.5	0.973
CHONA	**27.2**	**0.971**	**28.1**	**0.974**	**28.5**	**0.976**	**28.0**	**0.974**	**28.3**	**0.976**

TABLE C.4. Human-Object Animation for the boxes, i.e., different sizes of objects.

Method	boxlarge		boxmedium		boxsmall		boxtiny	
	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
TAVA [151]	22.6	0.949	25.9	0.967	26.8	0.970	27.5	0.973
ARAH [255]	23.3	0.963	26.3	0.972	27.0	0.974	27.7	0.977
ARAH$*$	**27.8**	**0.975**	**28.5**	**0.978**	**28.7**	**0.978**	27.9	0.978
Baseline	26.2	0.968	27.9	0.973	28.2	0.974	28.3	0.975
CHONA	27.2	0.971	28.5	**0.976**	**28.0**	0.974	**28.3**	0.976

TABLE C.5. Comparison on the novel non-interactive person (Sub02). Here, we use the shared body shape. We evaluate it on the object "chairblack". Therefore, the baseline is also good. CIL indicates compositional invariant learning.

Method	PNSR_{ind}	SSIM_{ind}	PNSR_{ood}	SSIM_{ood}
w/o CIL	25.4	0.951	26.5	0.959
CC-NeRF	**25.6**	**0.951**	**26.7**	**0.958**

C4 Challenges Analysis on BEHAVE

Occlusions BEHAVE [16] is a real-world 3D HOI dataset with *only four camera views and extensive occlusions*, which poses a significant challenge for the detailed reconstruction of Human and Object. Meanwhile, each interaction has less than 50 frames, which is challenging for the model to implicitly reconstruct the human body and object.

Blurry faces and frames To protect privacy, BEHAVE [16] uses the mask or fuzzy technique to blur most of the faces as illustrated in Figure C.1. This makes it very difficult to reconstruct the face of Subject01. Meanwhile, there are also blurry frames in BEHAVE. This further poses a significant challenge for a detailed reconstruction of HOI as illustrated in Figure C.1.

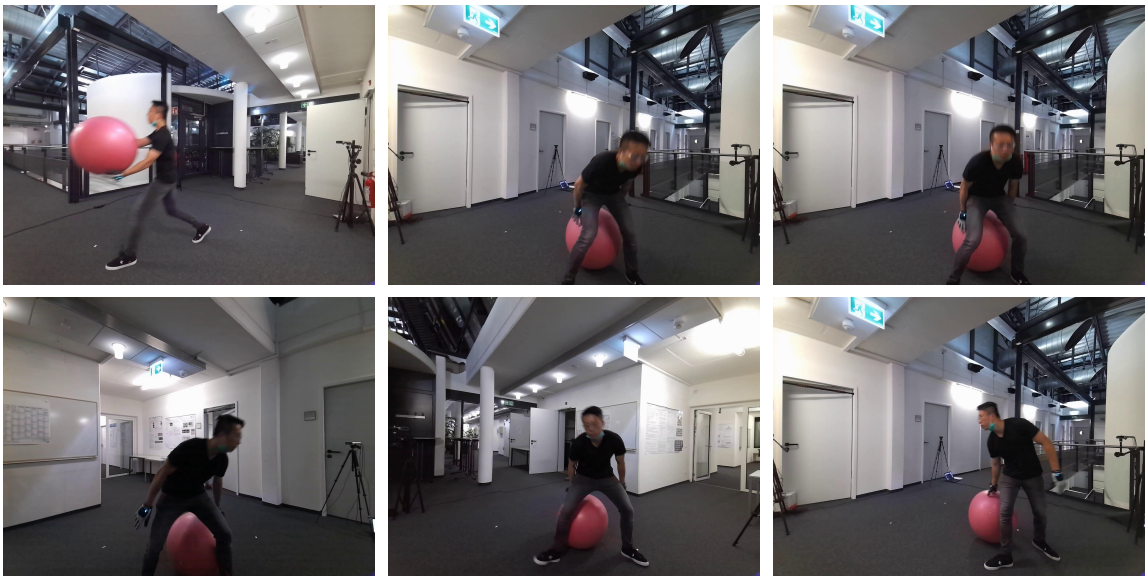

FIGURE C.1. Illustration of the blurry faces and frames.

Inaccurate Segmentation Besides, the segmentation mask in BEHAVE [16] is not much accurate due to the occlusion and complex background as illustrated in Figure C.2. One can find more inaccurate segmentation in the ground truth of the video comparison. In our experiment, we find the proposed method is able to marginally implicitly reconstruct the object and human body. However, the segmentation problem also degrades the accuracy of reconstruction.

C5 Demonstration of different pose quality

Without ground truth poses, we can also transfer the interaction poses (actions) among similar objects as shown in the novel person or novel object animation. Besides, Table. C.6 also shows the effectiveness when using noised and predicted object poses for animation.

C6 Comparison on different boxes

Table C.3 shows the effect of object size on evaluation metrics. Figure C.3 illustrates TAVA completely fails to reconstruct the object. Figure C.4 demonstrates neural human-object

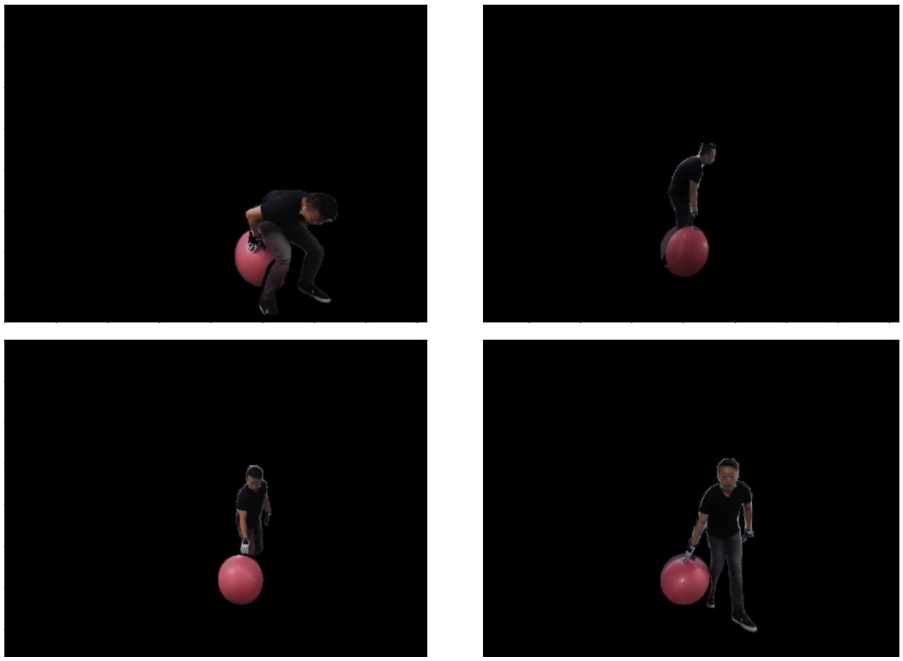

FIGURE C.2. Illustration of inaccurate masks. The boundary between the yogaball and human is not correct. The wrong boundary even causes the shape of yogaball changes.

TABLE C.6. The left table is for evaluation on different object poses on a subset of Mediumbox. "Predicted" is we predict the object poses with the method in BEHAVE[4]. "Noised" indicates we add Gaussian noise to the ground truth. The right part is for quantitative evaluation on a novel static object ("chairwood").

	GT	Predicted	Noised	Baseline
PNSR	27.7	27.2	27.0	26.5
SSIM	0.975	0.973	0.973	0.966

deformation can also reconstruct the boxes based on ARAH. Here, we leverage the box meshes to guide the SDF model optimization. Nevertheless, we still found it is challenging to reconstruct the complex object (e.g. chairs) in the main paper. Meanwhile, we can not make sure we have the meshes for the object. Therefore, we mainly evaluate the compositional human-object animation based on the template-free methods.

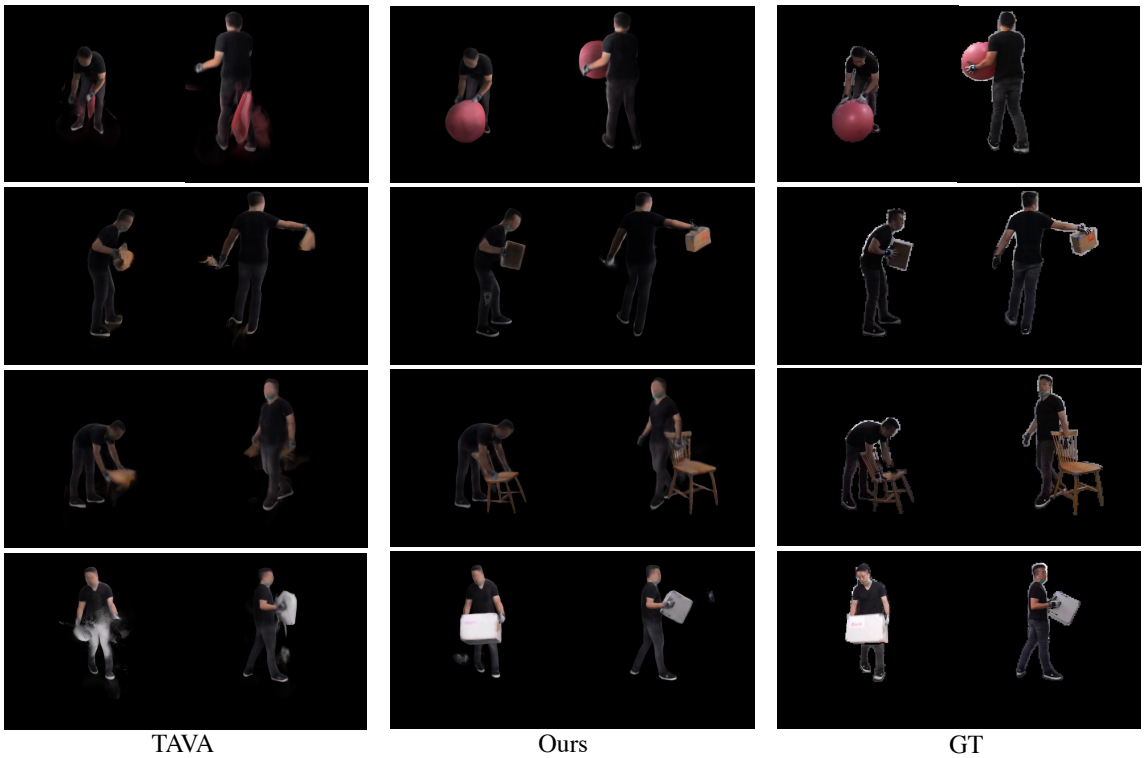

FIGURE C.3. Visualized Comparisons between the proposed method and baseline method (TAVA [151]). We demonstrate the results of "yogaball", "boxsmall", "chairwood", "boxlarge".

Besides, Table C.4 and Figure C.4 show we can achieve better performance when we utlize the SDF-based neural animation methods. But the limitation of those methods is that we can not always have the prior models for the object, even for the novel objects.

C7 Additional Experiments on Novel Non-interactive Person and static Objects

In this section, we first provide experiments on the person (ZJU subject 387) and the object (chair in CO3D) in Figure C.5. *Because the object segmentation in CO3D [211] is usually not accurate, we select two objects with relatively better segmentation to demonstrate the proposed method on novel static objects*. Figure C.5 demonstrates the baseline achieves worse rendering results on human body, while CC-NeRF with compositional invariant learning

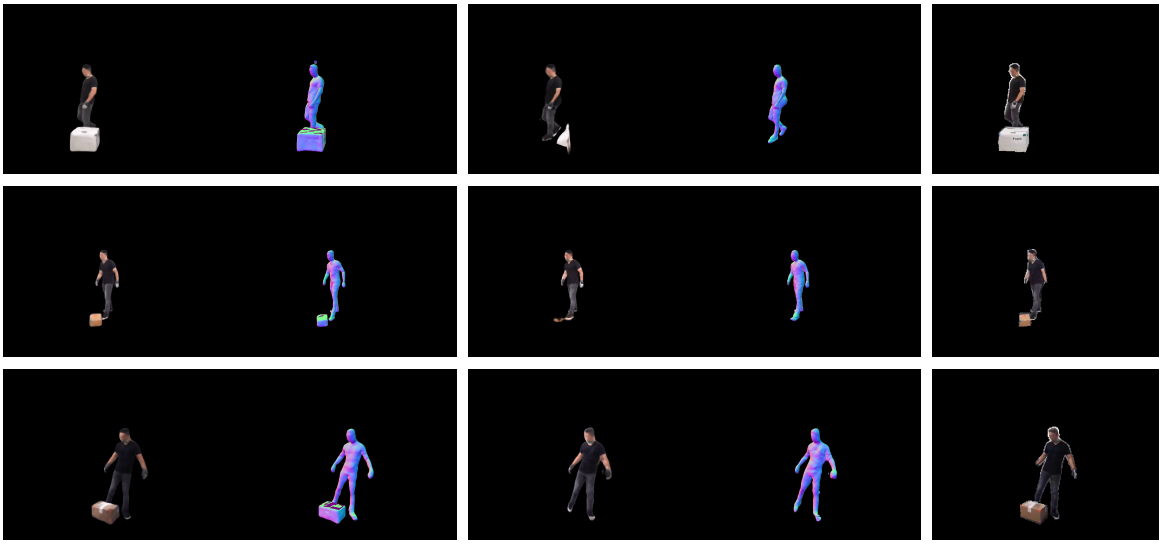

FIGURE C.4. Visualized Illustration between ARAH$*$ (with the proposed neural human-object deformation) and ARAH [255]. We demonstrate the results of "boxlarge", "boxsmall", "boxmedium".

significantly improves the baseline. Particularly, the objects in CO3D [211] are not fully scanned, i.e., the objects have some views (e.g., the bottom) that are unseen. As a result, the rendering result is poor when the object is transformed into a novel view during the interaction.

Figure C.6 also demonstrates CC-NeRF with compositional invariant learning achieves better rendering performance on human body. We observe the face of the baseline tends to be similar to Subject01 or Subject02 in BEHAVE, while the proposed method achieves better controllable rendering with the latent codes.

Additional Quantitative comparison. Table C.5 shows that the proposed method is also able to improve the animation on novel non-interactive person. Here, we first choose a single frame from subject02 without the object. We then train the network together with the "chairwood" HOI videos from subject01. We evaluate the animation on the interaction between subject02 and "chairwood". For the actions in validation that are similar to the training set, we treat it as the similar distribution generalization ($_{ind}$), while we treat the validation samples that have different actions as out-of-distribution evaluation ($_{ood}$). In our experiments, "chairwood_hand" and "chairwood_lift" are training actions and we choose "chairwood_sit" as ood validation.

For the experiments on novel object animation in the main paper (Table 4), we choose a single frame of “chairwood” and the two actions (“hand” and ”lift”) from “chairblack” in subject01 as training set. We use the interactions between subject01 and “chairwood” as validation set. Figure C.7 illustrates that the model without compositional invariant learning fails to render the person. Compared to the experiments in Figure 5 in the main paper, we here do not have a lot of objects and person in the training set. As a result, it is more difficult to render the person for the novel object. Therefore, the baseline method totally fails to render the person. However, the proposed method still effectively renders the person. Besides, this experiment also demonstrates Compositional Invariant Learning is more beneficial for decoupling the object and person latent code compared to the baseline when the number of HOI classes is limited in the training.

C8 Potential Applications

The proposed compositional neural animation approach is also able to reconstruct the human body or object separately from the interaction scenes as illustrated in Figure C.8. It is expensive to obtain dense cameras to scan all the 3D surfaces of the object. With the proposed method, we think we can reconstruct the full object by moving and rotating the objects under a single camera. Figure C.8 shows we can reconstruct the object with massive occlusions and render the objects individually from a few camera views.

Besides, Human-Object animation is important for Human-centric generation. We can generate the interaction videos according to the poses. Furthermore, we can also use the language-to-motion model (e.g. MotionDiffusion [**tevet2022human**]) to generate poses given the language description.

C9 Animation from Monocular Videos

We further present compositional 3D Human-Object Animation from a single monocular video (single view). We notice the proposed method is able to achieve remarkable performance

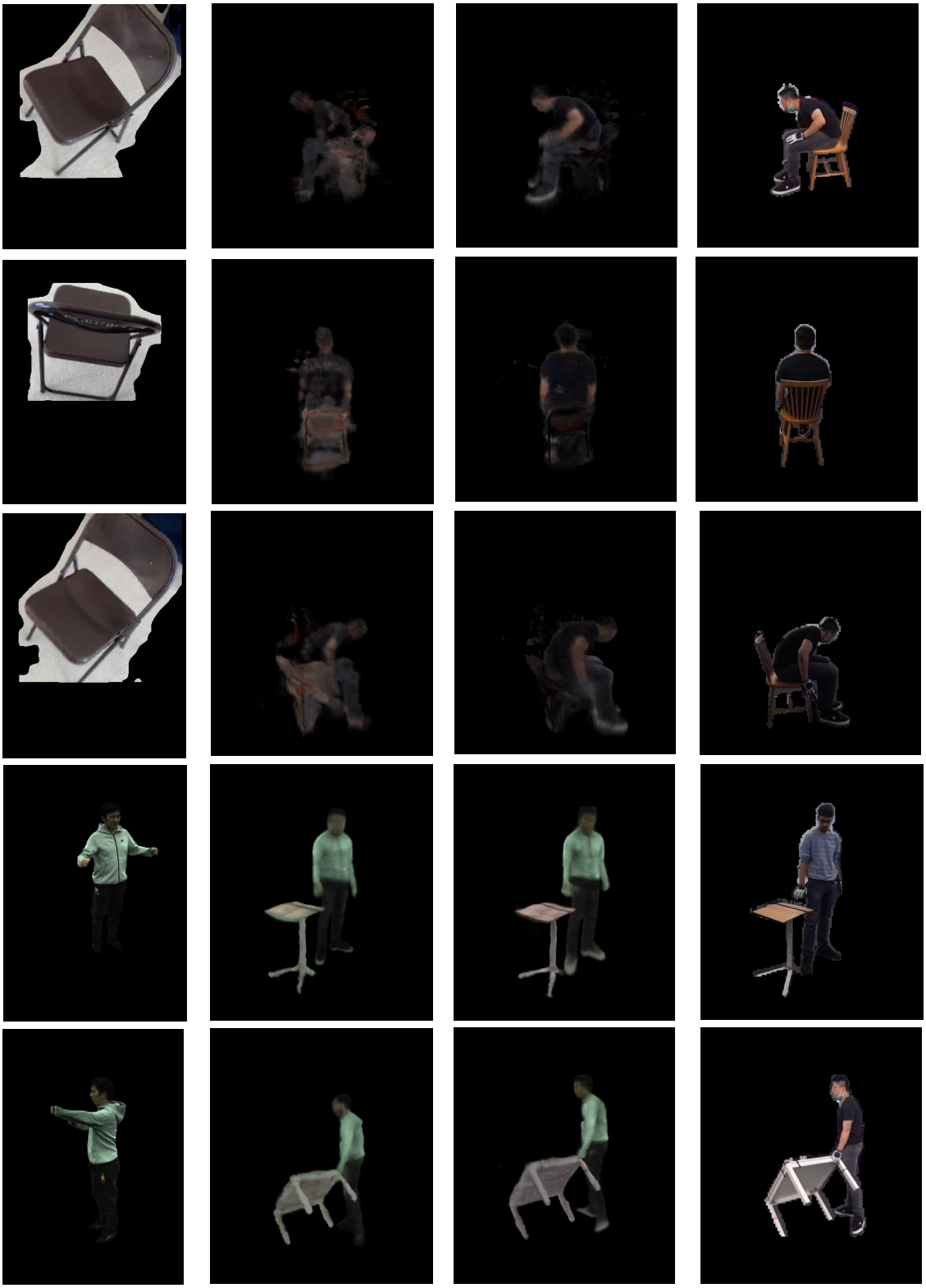

FIGURE C.5. Illustration of Compositional 3D Human-Object Neural Animation on Novel static object and non-interactive person. The first column is the guided person/object, the second column is the baseline, the third column is CC-NeRF with compositional invariant learning, and the last column is guided poses.

FIGURE C.6. Illustration of Compositional 3D Human-Object Neural Animation on non-interactive person (ZJU386). The first column is baseline without compositional invariant learning, the second column is CC-NeRF, and the last is guided poses.

as illustrated in Figure C.9. Compared with the experiments from four views, we notice the details of a single view are worse, (e.g., the legs in the left images and the medium boxes in the right images).

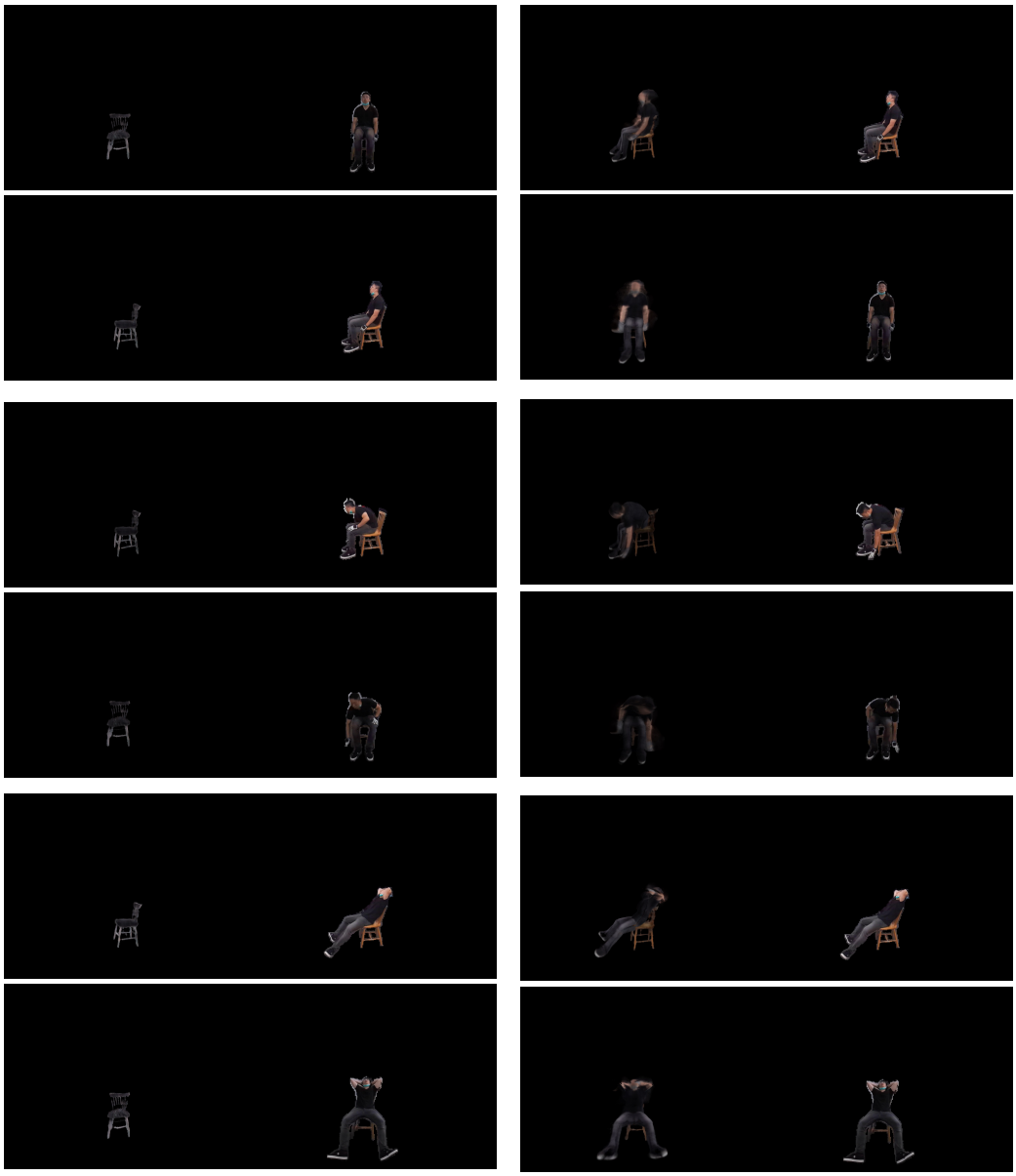

FIGURE C.7. Illustration of Compositional Human-Object Animation on novel static object. Here, we have only two objects ("chairblack" and "chairwood") in the training set. The first column is the model without compositional invariant learning. The second column is the model with compositional invariant learning.

FIGURE C.8. Illustration of Object Reconstruction and Rendering. Here we directly disable the human rendering via changing the person latent code.

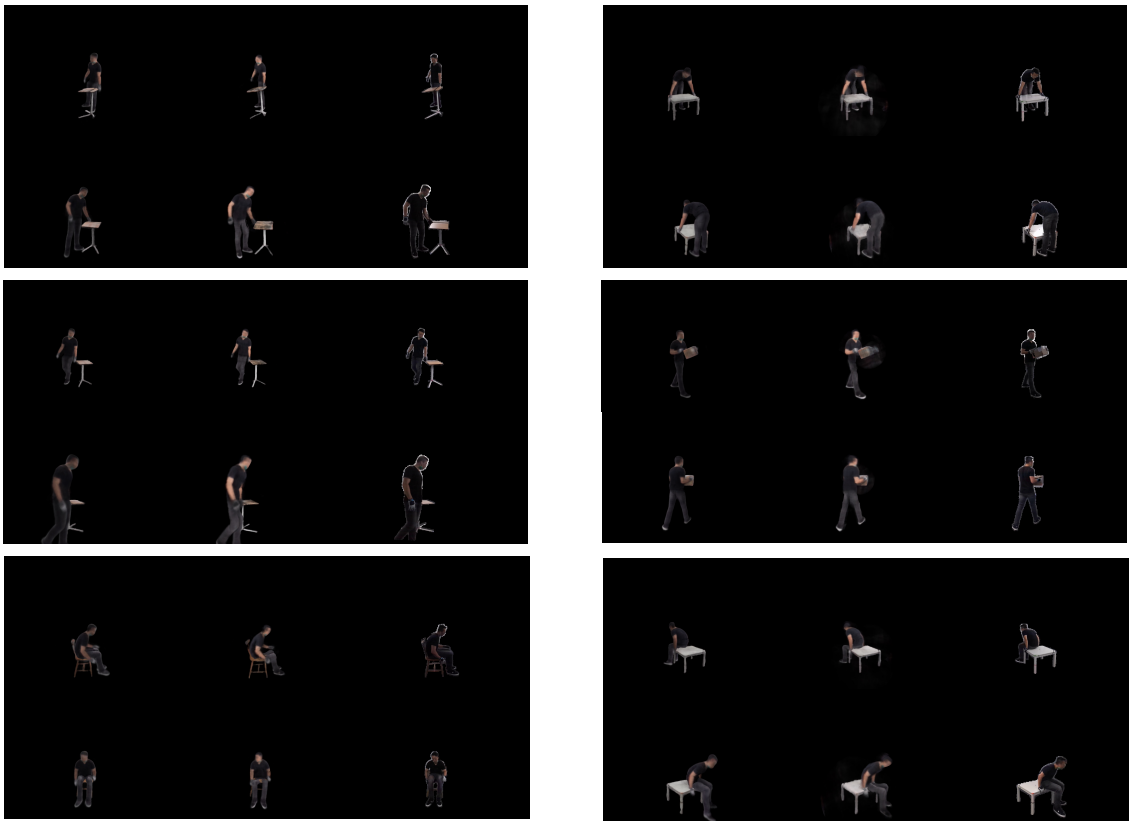

FIGURE C.9. Illustration of Compositional 3D Human-Object Neural Animation from a single view video. The left images indicate the novel action validation, while the right images present the novel object validation. The first column is CHONA from four views, the second column is CHONA from a monocular video, and the third column is ground truth. For each interaction, we choose two views (not the training view) for demonstration.