An interaction of horse riding in a rather uncommon context. The image is generated by computer, using OpenAI’s recent DALL-E 2.

Does DALL-E know horses would require an air tank within the context of space as well? The model can definitely benefit from a contextual understanding of human-object interactions, which is presented in this thesis.

Contextual Understanding of Human-Object Interactions

Mert Kılıçkaya

Contextual Understanding of Human-Object Interactions

Mert Kılıçkaya

Contextual Understanding of Human-Object Interactions

Mert Kılıçkaya

This book was typeset by the author using $\LaTeX\,2_\varepsilon$.

Printing: Ridderprint — `www.ridderprint.nl`

ISBN 978-94-6416-757-3

Contextual Understanding of Human-Object Interactions

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. Peter-Paul Verbeek
ten overstaan van een door het College voor promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 28 November 2022, te 14:00 uur

door

Mert Kılıçkaya

geboren te Ankara

Promotiecommissie

Promotor:	prof. dr. ir. A. W. M. Smeulders	Universiteit van Amsterdam
Co-promotor:	prof. dr. C. G. M. Snoek	Universiteit van Amsterdam
Overige leden:	prof. dr. A. del Bimbo	Universiteit van Florence
	prof. dr. J. A. Good	Universiteit van Amsterdam
	prof. dr. C. Monz	Universiteit van Amsterdam
	dr. P. S. M. Mettes	Universiteit van Amsterdam
	dr. E. Gavves	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has been carried out within the graduate school ASCI, dissertation number XXX, at the QUVA Lab of the University of Amsterdam.

Annem Solmaz'a, Babam İlyas'a, Kardeşim Yiğit'e, Ablam İlknur'a, Yeğenim Eylül'e ve Eniştem Kamil'e ithafen...

CONTENTS

1 INTRODUCTION 1

1.1 Origins 6

2 WHERE IS THE INTERACTION CONTEXT? AN EMPIRICAL STUDY 7

2.1 Introduction 7

2.2 Related Work 8

2.3 Method 10

2.4 Empirical method 11

2.4.1 System 11

2.4.2 Implementation 12

2.5 Experiments 13

2.6 Evaluation 13

2.6.1 Exp 1: How do the 6 basic observable regions discriminate? . . 13

2.6.2 Exp 2: Per class and per image what region yields the best discrimination? 14

2.6.3 Exp 3: Are Details Important Within The Best Latent Context? 17

2.7 Conclusion 19

3 SELF-SELECTIVE CONTEXT FOR INTERACTION RECOGNITION 21

3.1 Introduction 21

3.2 Related Work 22

3.2.1 Human-object interaction recognition 22

3.2.2 Combining multiple contexts in action recognition 23

3.3 Method 23

3.3.1 Overview 23

3.3.2 Self-Selection Context 24

3.3.3 Gating Module 24

3.3.4 Human-object Features $\mathbf{X}_{ho}$ 25

3.3.5 Contextual Features $\mathbf{X}_c$ 26

3.4 Experiments 27

3.4.1 Contextualized Interactions Dataset 27

3.4.2 Other Datasets 28

3.4.3 Baseline models 28

3.4.4 Evaluation 28

3.5 Evaluation 29

3.5.1 Self-Selection of the Single Context 29

3.5.2 Self-Selection of the Multiple Contexts 29

3.5.3 Further Analysis of Self-Selection 31

3.5.4 Self-Selection with the State-of-the-art 33

3.6 Conclusion 34

4 STRUCTURED VISUAL SEARCH VIA COMPOSITION-AWARE LEARNING 37

4.1 Introduction 37
4.2 Related Work 38
4.3 Method 40
4.3.1 Composition-aware Transformation 41
4.3.2 Composition-aware Loss 42
4.3.3 Composition-Equivariant Backbone 43
4.4 Experiments 44
4.4.1 Datasets 44
4.4.2 Evaluation Metrics 44
4.4.3 Performance Comparison 45
4.5 Evaluation 46
4.5.1 Ablation of Composition-aware Learning 46
4.5.2 Feature and Data Efficiency 47
4.5.3 Comparison with the State-of-the-Art 48
4.6 Conclusion 49
5 HUMAN-OBJECT INTERACTION DETECTION WITHOUT ALIGNMENT SUPERVISION 51
5.1 Introduction 51
5.2 Related Work 52
5.3 Method 53
5.3.1 HO-I Align Layer 55
5.3.2 HO-I Classification Layer 56
5.3.3 HO-I Feature Extraction Layer 56
5.3.4 HO-I Loss Layer 57
5.4 Experiments 57
5.5 Evaluation 58
5.5.1 Comparison to The State-of-The-Art 58
5.5.2 Further Analysis 59
5.6 Conclusion 60
6 SUMMARY AND DISCUSSION 63
6.1 Summary 63
6.2 Discussion 65
6.2.1 Future Work 65
Bibliography 73
Acknowledgments 77

1

INTRODUCTION

Context allows us to make sense of the world. While this thesis is being edited in January, many cities like Istanbul are in complete lock-down in Turkey. The lock-down is not due to the pandemic, but the ongoing blizzard. There is, however, a cold, distant city named Kars, where the snow does not come as a surprise. The readers of the novel *The Snow* (2002) from Nobel Laureate author Orhan Pamuk will remember the main character ”KA”, and his struggle in reaching Kars from the nearby city Erzurum. KA turns his struggle into an advantage, however, as the ongoing blizzard slows him down, simplifies the city view and gives sufficient time to think about his journalistic investigation on the women. Such realization is rendered possible with *the temporal context* of his small journey, as it takes time in the winter.

Famous piano composer Fazıl Say translates Turkish poetry to the musical language. To make sense of his piece *Black Earth* (2003), one needs to refer to *the lingual context* of Aşık Veysel's poem with the same title. Aşık Veysel refers to the soil with Black Earth, as the one and the only unconditional lover, from where we emerge and eventually will go back to. In his famous painting *The Tortoise Trainer* (1906), Osman Hamdi depicts himself as the restless trainer of multiple tortoises around him. To understand this piece, one needs to consider *the visual context* between the trainer and the tortoises, as Hamdi's head and gaze, and the relative location of the tortoises play a crucial role.

Inspired by the temporal and lingual context, this thesis focuses on the visual context. The absence of the context challenges human in recognizing the visual objects [114, 115, 128, 143]. The visual context determines our expectations about the scene. We expect the objects to be in a typical place, in a typical location surrounded by common objects. Consider the human-cow pairs in Figure 1. On the left, we depict an abstract human and cow, whereas on the right the same objects within their usual context. What are the differences? First, the cow and the rider are in a rural scene on yellow grass, which is comfortable for the target animal. Second, the rider is on top of the cow, which is one of the many expected configurations between a human and the cow object. Third, more than half of the pixels of the human rider are occluded behind the object, whereas they are fully visible on the left. Fourth, both the human and the cow are dressed accordingly with their immediate interaction, as is visible from the rodeo clothing of the human. Fifth, their shadows are visible from the grass, as the interaction takes place in a typical day time. Lastly, the human and the cow are surrounded by semantically similar objects and stuff: The by-stander, the car in the background and the mountains in the far-away region. All of these highlight the abundance of the visual context information we leverage on a daily basis to perform simple tasks of recognition.

Visual context is anything secondary to the target object appearance. Given an object, the immediate pixels around the vicinity of the object is the local context. The type of the

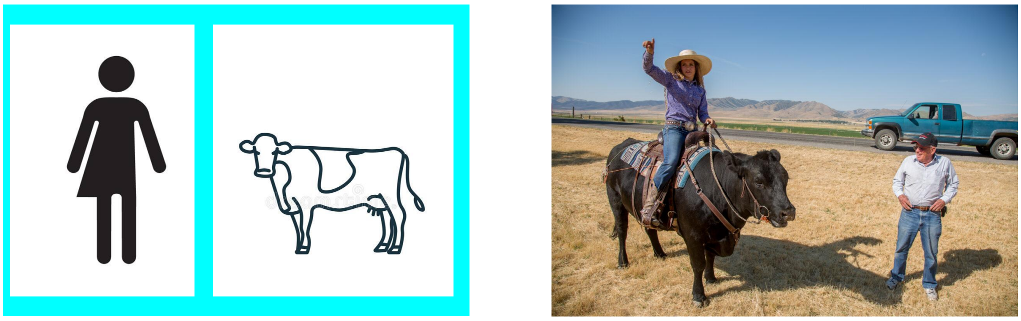


Figure 1: On the left, we depict an abstract human and cow, whereas on the right the same objects within their usual context. What are the differences? First, the cow and the rider are in a rural scene on yellow grass, which is comfortable for the target animal. Second, the rider is on top of the cow, which is one of the many expected configurations between a human and the cow object. Third, more than half of the pixels of the human rider are occluded behind the object, whereas they are fully visible on the left. Fourth, both the human and the cow are dressed accordingly with their immediate interaction, as is visible from the rodeo clothing of the human. Fifth, their shadows are visible from the grass, as the interaction takes place in a typical day time. Lastly, the human and the cow are surrounded by semantically similar objects and stuff: The by-stander, the car in the background and the mountains in the far-away region.

object location is the scene context. The spatial configuration between the target object and all the other objects is the compositional context. The type of the other objects is co-occurrence context. The clothing of the objects is attire context. The time of the day is the atmospheric context.

Interactions, just like context, are fundamental to the universe. The planets interact with each other via gravitational waves to keep the solar system in balance. The atmosphere interacts with the sun and the soil to regulate the temperature. Humans interact with each other to create societal systems, even in times of social distancing. Human cells interact with each other to generate immunity. Words interact with one another to make up sentences, paragraphs and consequently novels. Where we are inspired by the interactions in astronomical, biological, social and lingual entities, in this thesis we focus on visual interactions.

Visual interactions exhibit themselves at different scales in an image. Given an image, two pixels can interact with each other to generate edges, blobs or color changes to define an object region. Such narrow-range interactions give rise to prominent techniques for fundamental techniques of image filtering, such as smoothing or denoising [66, 146]. Object parts interact with each other to generate the holistic object view. Consider the interaction between the legs, the head and the main body of the cow in Figure 1. Cow legs are bent, the head points to the forward, and the body connects these two to afford the interaction oh holding and supporting. Thirdly, objects can interact with each other, such as the rider and the cow. This thesis studies visual interaction between human subjects and arbitrary objects.

In this thesis, we represent a Human-Object interaction as a `<verb, noun>` pair, such as ride-cow or eat-donut. Human interactors could be of any role, such as a rider, a player or holder. Object interactees could be of any type, such as an elephant, a cow, or another human in case of social interaction.

Many times, various types of interactions co-occur with one another. For example, to ride a cow, one needs to sit on, hold and straddle the cow. To hit a ball, one needs to hold and swing the baseball bat. In this thesis, we employ these preferences of co-occurrence to improve understanding of visual interactions.

Many verb-noun combinations contain only few examples in a typical dataset. Therefore, Human-Object interactions exhibit a long-tailed distribution, where the majority of interactions dominate the tail categories [70]. Consider, we type the keyword ”riding” in a search engine. The result would yield lots of exemplars from riding-bicycle, or riding-horse (well-represented). However, exemplars for riding-cow, or riding-elephant could only be very few (under-represented) if any. This challenges modern learning algorithms, as they begin to associate the target verb (riding) with the most popular nouns (*i.e.* bicycle, horse) during training, therefore limiting the accuracy [70].

Human-Object interaction research has pre-dominantly focused on video sequences [16, 37, 125, 137, 138], as interactions are generally more obvious from a video sequence. What moves together in a video will generally be subjected to a form of interaction (*i.e.* a bicycle and the rider). In this thesis, we prefer to focus on visual interaction understanding from a single image. Single image interactions are abundant, as the majority of images on the Web exhibits humans manipulating objects. Single image interaction is easily picked up by humans but in the state of the art a considerable challenge to machines, and by the absence of co-motion, arguably much harder than video. In the absence of the crucial temporal information within a video, we resort to a contextualized understanding of Human-Object interactions from an image.

Human-object interactions are contextual. Consider Figure 2. On the left, we compare a cow rider with a by-stander within the same image. Even though they are lit by the same illumination source, captured from a similar viewpoint, they exhibit drastic appearance changes. All body joints of the by-stander are easily visible and non-occluded. The by-stander stands in a canonical human pose. However, nearly the half the body parts (*i.e.* legs) of the cow rider is occluded behind the cow. Her body is bend, and her legs are straddling the object of interaction. In this thesis, we call such changes in contextual appearance with respect to the target interaction ”the locality of visual interactions”.

In the middle, we compare the spatial configuration of the cow and the rider, and the cow and the by-stander. Observe how the interaction of riding leads to drastic changes within the spatial relationship of the human and the object. Humans configure their full body as well as the body-parts to afford the target interaction. In this thesis, we call such changes in contextual appearance with respect to the interaction ”the compositionality of the visual interactions”.

On the right, we compare the bounding box pairs of the cow rider and the three other objects within the same image: The cow, the human, and the truck within the background. Even before arriving at an interaction decision, the computer needs to understand the interactivity: Who is interacting with whom? In a typical scene, there are hundreds of potential human-object pairs to consider before arriving at an interactivity decision. However, amongst them, only few are in an interaction (*i.e.* rider and cow), whereas the rest is background (*i.e.* rider and car). In this thesis, we call such phenomenon observed in human-object interactions: ”The sparsity context”.

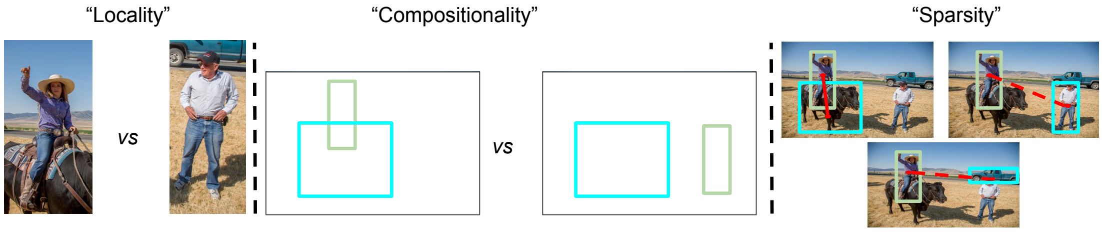


Figure 2: Three different contextual information leveraged within this thesis. (Left) The locality context concerns the transformation of human body appearance w.r.t. the interaction. Observe how cow-riding occludes half the human body in comparison to the by-stander within the same image. (Middle) The composionality context concerns the transformation of the human spatial configuration. (Right) The sparsity context concerns the interactivity between a human interactor and an object.

Observing that human-object interactions exhibit rich source of contextual information, this thesis studies the following central research question:

RQ: Can we understand the role of context in single image human-object interactions?

To understand the role of context, we first need to determine the sources of context in visual interactions from a single image. To determine the source of context, we propose to identify the visual extent of human-object interactions. Previous research on visual extent focuses on the temporal extent of video actions [59, 122, 123]. The authors identify the temporal borders of a video action, by inferring the start and end frames of the target activity. In our work, we focus on single images, therefore lacking the crucial time information. Therefore, we propose to identify the spatial borders of a human-object interaction in a 2D image plane. Uijlings *et al.* studies the spatial extent of visual object recognition [133]. The authors conclude that the surround of visual object region carries the most discriminative information. Our work not only discriminates between different objects, but also between different states of the same object from a single image, which brings us to the following research question:

RQ1: How can we identify contextual cues in visual interactions?

We propose to understand the visual extent of human-object interactions through the lens of Convolutional Networks [83]. We take a hierarchical approach, where we start from a full image frame and approach to the vicinity of human-object region. We gradually limit the amount of information visible to a pre-trained interaction classifier and record discriminative regions that yield good classification accuracy. Our first conclusion is that the locality of the visual interactions play an important role. The surround of human-object regions carries the highest information for classification.

Furthermore, we observe that the visual extent of interactions is ambiguous, as there is no fixed amount of context that performs best across all categories. This motivates us to build models that can dynamically select, from a pool of local contextual representations, discriminative context(s) to recognize the interaction. To that end, in the next chapter, we ask the following research question:

RQ2: How does local context help in interaction recognition?

To understand the role of local context in interaction recognition, we propose a locality-aware context which can represent the locality of the surrounding scene, the locality of human body pose configuration as well as its deformation. Since not all contextual features are equally important, we propose a self-selective c ontext, a neural network that can select contextual features conditioned on the joint appearance of human-objects as well as the context. We note that the importance of contextual information is relative to a given human-object, as small-scale human-object regions have higher contextual dependencies. From the output of self-selection, we conclude that indeed interactions have strong contextual preference for interaction recognition.

Dynamic selection of local context yields a dramatic improvement in recognition. However, it neglects an important, fundamental contextual source in visual interactions: The composition. Humans configure their body according to the spatial affordance of the target interaction. To ride a cow, humans need to be on top, whereas to walk a cow, humans need to be next to the cow. Spatial arrangement not only represents itself in the full human body, but also in the human body part configurations, as we arrange our fingers in accordance with the object to grasp. The unique compositional arrangement between humans and objects provides a visual, much more descriptive way to search for visual interactions from large image databases in comparison to traditional, text-based image search. To leverage the compositional relationship between humans and objects, the next chapter asks the following research question:

RQ3: How does compositional context help in interaction search?

This chapter proposes a method to enable interaction search via spatial context of human-objects. Human users draw on a 2D canvas, the location as well as the category of the objects of their interest, which we then use to search over large image databases. Resulting images not only should satisfy the arrangement of human-object locations, but also their semantics. To achieve this, we propose composition-aware learning, a technique that leverages the symmetrical changes between input (query) and output (visual feature) spaces. We observe that imposing such a constraint not only leads to computational efficiency, but also to sample-space efficiency.

While searching for images via compositional queries, a limiting factor in retrieving relevant images are the distractors. In a typical, unconstrained scene, there are potentially tens of humans and objects passing by, if not hundreds, whereas only few of them are in a form of interaction. This is the result of object co-occurrence in natural scenes. Where a photographer focuses on a cyclist on a road, there could potentially be many pedestrians or cars passing by. This challenges the computer to identify the interactivity between humans and objects, even before arriving at the interaction decision. To determine the interactivity, we take advantage of the sparsity context, that is, given an exhaustive list of all human and object regions within an image, only a sparse subset will exhibit interaction patterns, in terms of composition and appearance. By forcing the learner to focus on these sparse subsets, we can naturally learn to identify the real human-object interactors, neglecting the rest. To leverage the sparsity of human-object interactivity, the next chapter asks the following research question:

RQ4: How does sparsity context help in interaction detection?

This chapter proposes a method to detect human-object interactors from an image. In doing so, we only rely on weak, image-level supervision, as opposed to popular, strong instance-level supervision. In the absence of strong supervision, we force the network to find a sparse subset of human-object interactors that can classify the interaction within the image. Enforcing such a sparsity constraint guides the detector to select correct targets for supervision, while eliminating the distractor noise.

1.1 ORIGINS

This thesis is based on the following publications:

- **Chapter 2** is based on ”Where is the Context of Interaction? An Empirical Study”. Under submission to CVPR Workshop on Visual Learning from Limited Labels, 2022, by Mert Kilickaya, Efstratios Gavves and Arnold Smeulders [73].

  Contribution of authors

  Mert Kilickaya: All aspects,
  Efstratios Gavves: Insight,
  Arnold Smeulders: Supervision and insight.

- **Chapter 3** is based on ”Self-Selective Context for Interaction Recognition”. In: *International Conference on Pattern Recognition (ICPR)*, 2021, by Mert Kilickaya, Noureldien Hussein, Efstratios Gavves and Arnold Smeulders [69].

  Contribution of authors

  Mert Kilickaya: All aspects,
  Noureldien Hussein: Writing and experiments,
  Efstratios Gavves: Insight,
  Arnold Smeulders: Supervision and insight.

- **Chapter 4** is based on ”Structured Visual Search via Composition-aware Learning”. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, by Mert Kilickaya and Arnold Smeulders [72].

  Contribution of authors

  Mert Kilickaya: All aspects,
  Arnold Smeulders: Supervision and insight.

- **Chapter 5** is based on ”Human-Object Interaction Detection without Alignment Supervision”. In: *British Machine Vision Conference*, 2021, by Mert Kilickaya and Arnold Smeulders [71].

  Contribution of authors

  Mert Kilickaya: All aspects,
  Arnold Smeulders: Supervision and insight.

2

WHERE IS THE INTERACTION CONTEXT? AN EMPIRICAL STUDY

2.1 INTRODUCTION

In a story, many sentences relate the interaction of a subject with an object. Whether it is talking to a friend, walking the dog or cooking potatoes, interactions of subject and object are the basic ingredient of the story.

Also in pictures, interaction steals the focus of attention [13]. When the picture contains a prominent subject and a salient object, the attention is drawn to their interaction. Even the absence of interaction is noted immediately. The subject acts upon an object for a purpose; to feed, to amuse or to achieve: eating an apple, riding a bicycle, or stuccoing a wall. Recognition of interaction holds the key to the purpose in an image.

Interaction recognition is not limited to the subject and the corresponding object. Their joint appearance in an image does not guarantee interaction. Contact between subject and object usually matters but not necessarily. Also context matters in most interactions. These elements of composition may aid in the recognition. Riding a horse is limited to a specific contact of the subject with the horse and usually restricted to the number of preferred surroundings, see Figure 3. Where interaction is important, its recognition does not yet match the performance of person or object recognition since [77]. In spite of good progress, the performance on HICO [18], the biggest benchmark of interactions, has been limited in recent years.

Current research does not subscribe to the same definition of interaction. Especially the definition of the localization varies significantly: some researchers focus solely on the subject-object appearance [30, 34, 42, 94], whereas others follow a hybrid strategy where the goal is to combine contextual information like the surround scene [95] or surround objects [46] with subject and object information. Given this, in this paper, we ask: *Where is the interaction?* We identify two main problems in answering this question. First, do we need context for classification of interactions? Second, to what extent subject-object regions contribute to the recognition?

We study the *"where"* of an interaction by analyzing the spatial extent of human-object interaction from a single image. Our work operates in three steps. We first discern six main regions illustrated in Figure 3: the subject box, the object box, the box of the subject-object intersection which we refer to as "contact", the subject-object union, the context region which is the complement of the union to the whole image, and finally the whole image itself. Then in the first step, we study which of these six regions is best in learning to recognize interactions. Given this, we then ask whether a latent context region exists within the image that yields better interaction classification. Lastly, we

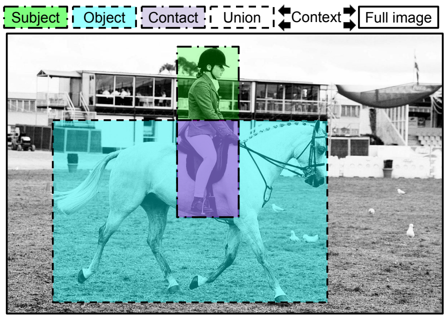


Figure 3: It is easy to pinpoint the participants of the interaction: the subject (equestrian), the object (the horse), their intersection (contact), their union, or region between the union and the full image (context). It is hard, however, to pinpoint the interaction itself. Where is the visual evidence for riding a horse?

analyze the impact of the small, specific details within the latent context for interaction recognition.

The results of this paper demonstrate that *i)* in general context carries important information for interaction recognition, therefore one needs the full image to learn to recognize interactions. However, *ii)* full context is not needed in recognizing the interactions, as we will show that there exists a latent context within the union region and the full image that occupies 75% of the image, and yields a significant increase of up to 20 mean Average Precision (mAP) points once localized by an oracle procedure. Lastly, *iii)* we observe that within this latent context visual details have a major impact: blocking small regional details within the latent context can reduce the recognition rates by up to 30 mAP points.

We make the following contributions: *i)* we study the visual extent of human-object interactions from single image and determine the best source to learn about interactions, *ii)* we identify by oracle the best possible amount of context for interaction recognition and its properties among different interaction classes, *iii)* we determine the importance of visual details within this context. In conclusion, we show that localization per image is indeed critical for interaction recognition, and leaves substantial room for improvement. In addition, we find that localization can not be the only solution for interaction recognition. We pay a special attention to the results of *" no interaction"*, since to perform *" no interaction"* is a real test whether the interaction is recognized or just the coincidence of a subject with an object.

2.2 RELATED WORK

Spatial extent of an object. The spatial extent of an object from a single image studies what parts carry the information for object recognition [82, 133]. It has been studied in the context of Bag-of-Words representations by [133], where it is determined among 20 PASCAL VOC categories [33]. The interior and the close surround of the object bounding box are found to be most discriminating. Later, this work has been extended in [82] to Fisher [120] and Convolutional Neural Networks (CNN) [77, 83] representations. The context in these modern representations is found to be less important for recognition, thanks to the better localization capabilities. Inspired by these studies, we focus on the visual extent of interactions.

Action recognition. Action recognition is an extensively studied problem in computer vision, especially from videos [16,125,137], and also from single images [8,94,119,147]. This work focuses on single images. Single image action recognition mostly focuses on identifying only verbs from an image, such as running, cooking, eating or sitting [8,94]. Studies in [119,147] deviate from verb-only and also include phrases, like riding a bicycle. Visual-Phrases [119] consider non-human subjects, including spatial-only relationships, providing a limited set of interactions (32 in total). Stanford-40 [147] has 40 verb-only and verb+noun image actions. Although 6 of the verb+noun image actions are common with the HICO dataset [18] we consider in this study, the majority of the classes are verb-only. We, therefore, focus on [18] which features a great range of 600 possible interactions between 80 different object nouns and 117 different verbs.

Interaction recognition from a single image. The goal of the interaction recognition is to describe the relationship between a subject and an object. In [18] the relationship is defined as a set of `<verb,noun>` pairs, such as `<ride,horse>` or `<eat,cake>`, between a human subject and an arbitrary object from the list of MSCOCO objects [89]. In this paper, we follow the definition as well as the dataset and annotations provided in [17,18]. Looking closer to interactions, we identify two important sources of information from the references, namely the context of the interaction, and the union of the interaction.

First, we consider the context. The importance of context for interaction recognition has been well recognized in [46,51,60,95,97,110]. The various forms of contextual information fed into the interaction classifier include the pose context [51,110], the human-object spatial context [51], context of the surround object [46,60], or context of the surround scene around the subject and the object [60,95,97]. [95] obtained a significant improvement over the base model of [46] by considering the surround across the subject for interaction recognition. To this end we limit our focus to the surround context and study whether we need context for interaction recognition.

Second, we consider the union of subject body and object body [30,34,42,94]. In their paper on Poselets [94], the authors suggest that for interaction recognition full body is not needed, some discriminative parts would suffice. They localize small, discriminating body parts of subject-object via discriminative clustering of gradient histograms [28]. In an extension of this work by Phraselets in [30] the joint appearance of distinct body parts is considered: hands holding a handlebar while legs push a pedal to identify cycling. An important contribution in this work is the explicit consideration of the intersection of subject and object, namely the contact region. The focus on body parts by themselves has been extended to CNN by [42], and later by [34]. Attentional-Pooling [42] localizes body parts via an attention mechanism for interaction recognition. Pairwise-Attention [34] focuses on finding two-parts of the body. Body-part localization for interaction is fundamentally difficult because it is unknown what to localize before knowing the interaction. And, as we will demonstrate, context is too important for interaction recognition to leave out. In this work, we quantify the relative importance of the subject-object union for interaction recognition in comparison to context, and determine which region(s) carry the most discriminative information for interaction recognition.

Localizing visual evidence via occlusion sensitivity. In this work, our base classifier is a vanilla CNN [55] to study the visual extent of human-object interactions. Whereas different methods exist [7], we rely on occlusion sensitivity [109, 117, 149] to find important regions within the image that support the classifier decision. Occlusion sensitivity is first proposed in [149] where the authors slide a square box of a certain size over the image, zeroing-out regions one-by-one, to measure the influence of each region on the performance of a CNN.

In our work we re-purpose occlusion sensitivity as follows. We first progressively limit the amount of the context pixels seen by the classifier to determine the amount of context that yields the best performance. Then, we further remove sub-region(s) within this limited context one-by-one to determine which parts actually contribute to the accuracy. Similar to [117, 149], we use bounding boxes [134] and superpixels [5] as the form of occlusion.

2.3 METHOD

An interaction tuple `<subject, interaction verb, object>` has a direct relation to two bounding boxes in the image: the subject bounding box and the object bounding box.

Based on the locations between the two boxes, we identify the following 6 basic regions, illustrated in Figure 3: the *subject* box; the *object* box; the *contact* box; the *union* of the subject and the object boxes; the *context*, and finally *full* image.

In an image, one or more of the 6 regions may be void. Specifically, when there is no physical contact between the subject and object (*i.e.* subject inspecting object from a distance), the contact region will be empty. We refer to the 6 regions as the *observable* visual extent. The actual interaction region is unknown, and referred to as the *latent visual context* of the interaction. The goal of this work is to reveal the latent visual extent of interactions.

This brings research question 1: *Which of the six basic observable box regions can better discriminate between the interaction classes?*

The experiments on RQ1 will give a general impression on the localization of an interaction. In the second part, we investigate whether there is a dependency between the context and the interactions on a class basis and image basis.

In the second experiment, the question is: *Per class and per image what latent region yields the best performance?*

Third, while a good latent interaction context may lead to accurate interaction classification, can it be further improved by looking at the details? Hence, the third research question is: *In recognizing the interaction are local details particularly important?* In the next section, we concretely describe how each research question is implemented.

2.4 EMPIRICAL METHOD

2.4.1 *System*

Data. For training and testing we rely on the HICO [18] dataset. HICO contains a total of 50*k* single images of human subjects interacting with objects. The dataset is splitted into 40*k* training images and 10*k* testing images. The interaction categories are not mutually exclusive. A subject can conduct multiple simultaneous interactions. An image may be annotated as "sitting on a bicycle" and simultaneously as "riding a bicycle", as long as both are valid and contained in the set of interaction verbs. In total, there are 600 different pairs of tuple interactions of the form `<subject, interaction verb, object noun>`, built from 117 verbs and 80 nouns. In HICO, the distribution of examples is heavily imbalanced. While 167 interaction tuples in the learning set have more than 100 examples, 155 interaction tuples in the learning set contain less than 10 examples. We consider this property of the dataset a realistic asset as the free choice of verbs and nouns from a large pool will often contain rare combinations. And, combinations which are rare will often hold interesting phrases. As a consequence, we have adapted the evaluation metric from [18] that Average Precision is weighed per image rather than per class.
Annotations. HICO [18] is recently augmented with bounding box annotations by [17]. For HICO interactions, annotators are being asked to draw a bounding box around the subject and object for each interaction pair in the image. Apart from multiple interactions, also group interactions, where multiple subjects interact with the same object are allowed. For example, a group of people having dinner together at a table. We use 20*k* subject-object bounding box annotations, single or multiple, from 10*k* test images. Similarly, we gather 70*k* subject-object bounding box annotations from 40*k* training images.
Model for interaction classification. In all the experiments, we use a ResNet-50 [55] model pre-trained on ImageNet [29] and fine-tuned to discriminate between the 600 `<subject, interaction verb, object>` tuples in HICO, see Figure 4. Following [95], the model is optimized via ADAM [76] with a batch size of 4 for 60*k* iterations. We use a learning rate of 0.001 decaying to half after 30*k* iterations. The model is implemented in Tensorflow [4].

To verify the model independence, we have repeated the experiments also with a VGG-architecture [124]. The results were consistent with the Resnet-50 results. So in this paper, we only provide results with Resnet-50 [55], and refer the reader to Supplementary material for VGG-based replications of the experiments.

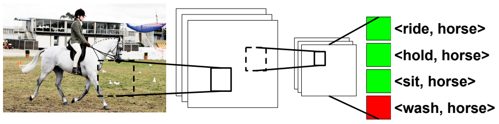


Figure 4: We rely on a ResNet-50 [55] CNN, pre-trained on Imagenet [29] and fine-tuned on HICO [18] to discriminate among 600 *interaction categories. Full image model is shown.*

Evaluation measure. We evaluate the various models in our experiments by computing mean Average Precision over images. Due to the long tail distribution of interaction tuples, many classes in the learning phase have fewer than 10 elements. We obtain the score by adding the score per image and then averaging over all images. We note that the result will be rather different from evaluating per class as was done in [18], which gives an unbalanced view of the many classes which are rarely occupied. We find image-based average a better fit for this paper since most of the analysis focuses on conducting experiments on a per image basis.

2.4.2 *Implementation*

Research Question 1. For research question 1, we train classifiers f, each trained on one of 6 different basic observable regions according to Figure 3, where we zero-out the rest of the image apart from the region in consideration. We compare the recognition accuracy on the interaction classes. Note that the 6 regions tend to have different sizes. The contact region is by definition (much) smaller than the other areas. The image regions are not re-sized to avoid any distortions due to aspect ratio change, especially with small regions.

Research Question 2. To answer research question 2 we need an oracle that provides the best latent region of interaction. As the number of possible regions in an image is practically infinite, we opt for the following approach. Given an image I of size $[H, W, 3]$, initial mask is given as the *subject-object-union* region mask, that is m_0 $m_0 \equiv r^{h-o-u} = b^h \cup b^o$, where b^h is the subject box in that image and b^o the object box. At every new step t the mask is shrunk or expanded by the mathematical morphological operations of erosion and dilation respectively, until the empty set or the complete set is reached. In total, we obtain about 350 masks per image. The classifier then receives as input the pixels within the current mask, $f(m_t \odot I)$ where $\odot$ is the element-wise multiplication. The best performing mask m^* for the true class, as determined by the oracle, is returned as the latent context of interaction.

Research Question 3. For the third research question we focus on local visual details. Specifically, we examine the positive or negative impact of local visual details contained in the latent context m^* resulting from question 2. To answer this question we devise a removal procedure.

Specifically, we divide the latent context into sub-regions $d_j, j = 1, ..., J$, removing one of them at a time. This results in masks $m' = m^* - d_j, j = 1, ..., J$ which we pass through the classifier as $f(m' \odot I)$. Then, we evaluate the Average Precision obtained by removing each sub-region. As for the choice of sub-regions, we opt for selective search boxes [134] or superpixels [5,36].

2.5 EXPERIMENTS

2.6 EVALUATION

2.6.1 *Exp 1: How do the 6 basic observable regions discriminate?*

Context matters

In the first experiment, we start from the 6 basic observable regions, to determine the best suited region for interaction recognition. We present the results in Table 1.

	Subject	Object	Contact	Union	Context	Full image
AP	.42	.36	.27	.42	.30	.55

Table 1: Discrimination of 6 different regions for recognition.

Among these 6 basic observable regions, the full image returns the best accuracy over all images in the dataset. This shows that the context of an interaction captured by the surroundings is indeed relevant. The context may help discriminating between similar interactions, with different context preference. A `<ride, horse>` interaction happens in a "jockey club" context, whereas a `<ride, bicycle>` is unexpected.

The results on the full image demonstrate the capacity of CNN's to perform soft detection implicitly [47, 153]. As an interaction does not have a clearly-defined, observable visual extent, it is preferable to give the full image as input to the classifier and let the model decide which image region is important.

The second highest accuracy - at a substantial margin - is obtained when using the subject box or the union box. The subject box is easy to detect with the modern object detectors [116], and per image it apparently contains sufficient context of the interaction to arrive at a substantial classification score. For illustration in Figure 3, the dressing suffices to understand that the subject is riding a horse. Conversely, when looking at the object box in Figure 3, the foot also implies someone riding that horse, but on average the subject box is preferred. In conclusion, context matters as the full image leads to the highest accuracy followed by the subject box at a wide margin.

Half of the full image suffices

As from Table 1 it was concluded that context is important for interaction, we investigate to what degree the rest of the image is needed for recognition. By shrinking or expanding the subject-object union box as described in Section 2.4.2, we obtain tighter and larger versions of the interaction regions. In Figure 5, we plot the recognition accuracy as a function of the extent of the interaction region used in the classification. Using an increasingly larger portion of the image, right of the black line in the Figure, eventually yields 55%, which was expected from Table 1. The blue line for the classifier, trained on the subject - object union box alone, shows a substantially lower performance, declining when more context is included. The red line for the classifier, trained on the context

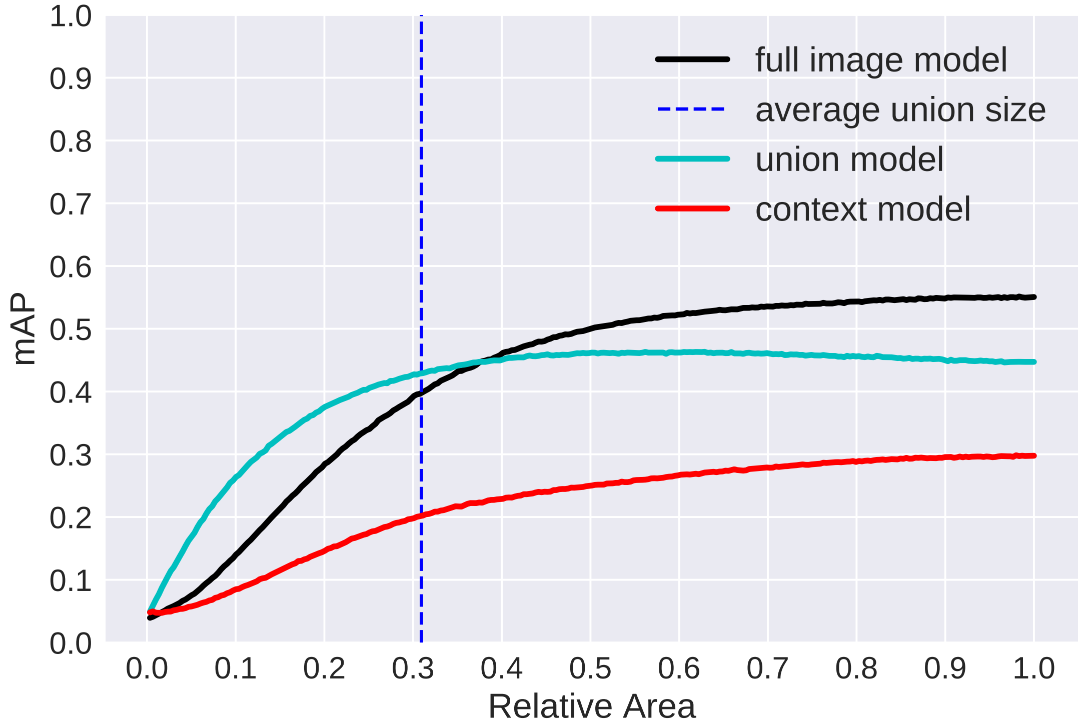


Figure 5: Performance of the three different models averaged over all dataset. Half of the image suffices in recognition for the full image model.

alone, is increasing when more of the image is included but never comes close to either the full image or union box performances.

It is concluded that half of the image suffices for good performance when a good part of the context is included as well.

As the full image classifier is the best and the most complete one, from now on we analyze the results of that classifier.

2.6.2 *Exp 2: Per class and per image what region yields the best discrimination?*

Given the importance of context, next we examine whether it is class specific. To this end, we shrink and expand the object and subject regions as before, and classify the region with the full image classifier. We present the per class accuracy as a function of the region extent in Figure 6 for selected example classes and include the same plots for all classes in the Supplementary material. We observe three types of distributions, that are positive, indifferent and negative to context.

1 out of 10 classes are positive to context

The top row in Figure 6 shows a gradual increase in performance when more of the context is taken into account. In 1 out of 10 interaction classes, the interacted object typically co-occurs with many identical objects, like picking an apple from an apple tree,

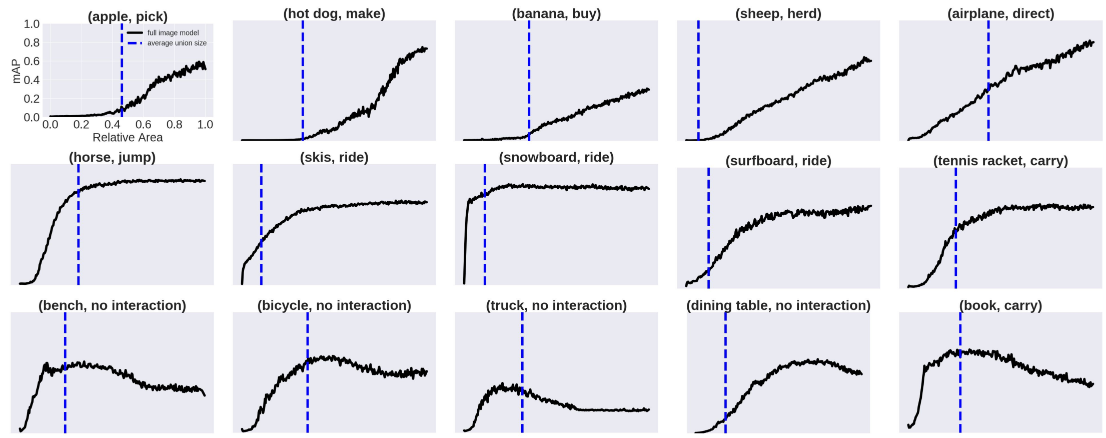


Figure 6: Performance of full image model averaged per class. 15 *interactions are shown. We note three different trends in performance with context, namely positive to context (top row), indifferent to context (middle row) and negative to context (bottom row). Axes are the same as top-left. The blue dashed line indicates the average union size.*

preparing hot dog in a restaurant kitchen, herding a group of sheep, or directing a group of airplane.

8 out of 10 classes are indifferent to context

The middle row in Figure 6 shows the dominant type of distribution, for 8 out of 10 classes indifferent to context. Classes featuring this distribution are sports interactions leading to unique subject-object deformations, such as “horse jumping” or “riding skis”. In conclusion, the vast majority of cases, knowledge of the class does not help in determining the proper context for interaction classification.

1 out of 10 classes are negative to context

The third row in Figure 6, however, illustrates the 1 out of 10 classes where the context has a negative effect on the recognition of the interaction. This is especially the case for the important class of “no interaction”. Obviously, the more context is provided to the classifier, the higher the chances are that an arbitrary interaction will be picked up. Therefore, the important class of ”no-interaction” profits from proper framing in the context.

Per image context is important

If the context is not important per class, is it important per image? When accumulating the average accuracy to the proper interaction class or classes per image, given an oracle definition of subject and object regions fed into the full image classifier, we arrive at an overall performance of 74% mAP, by using 75% of all image pixels. This is the upper bound what can be achieved when the best latent context is perfectly known.

Next, we analyze which images and interaction categories benefit from the oracle context in Table 2. We identify two mutually non-exclusive groups of interaction images

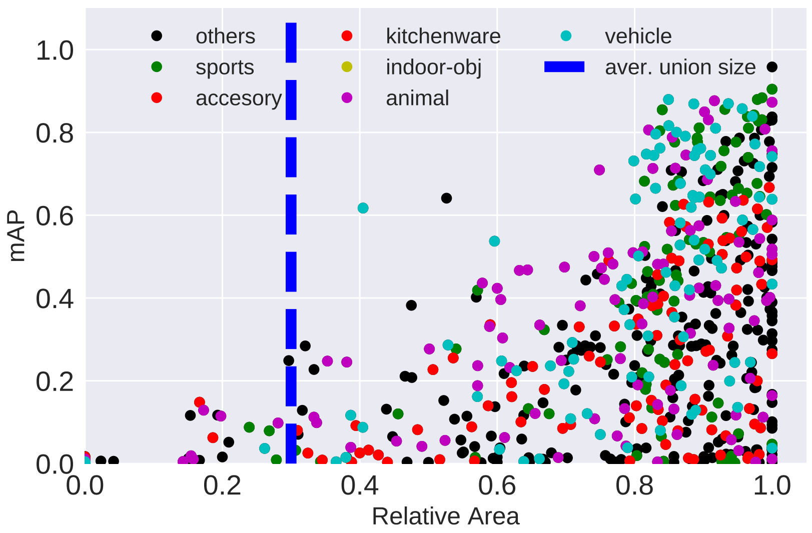


Figure 7: The area of the best context per class (x-axis) against the classification accuracy (y-axis). Colours indicate a grouping of interactions by object super-category [89]. The Figure shows that the classification accuracy does not depend on the type or the size.

	Size		Population	
	Small	Large	Under-pop.	Well-pop.
Dataset ratio	.29	.71	.11	.89
Full image	.32	.56	.34	.58
Best latent context	.46	.75	.47	.78

Table 2: The statistics of the improvement. Best latent context is helpful in cases when the union region is small or large, or when the interaction class is rare or frequent.

relevant to this question, in terms of *i)* population in the dataset (well-populated if containing more than 10 training instances, under-populated otherwise), *ii)* or pixel size of the union (small if occupying less than 10% of the image, large otherwise). We compare the results with the case of the full image. We also note the ratio of each group in the test set above.

First, we observe that the best latent context is useful not only for the well-populated but also for the under-populated classes. That being said, well-populated classes benefit more, as the classifier is able to generalize better. There is still a 31% gap observed between these two groupings. Moreover, it is clear that small-size and large-size interactions benefit equally from the best latent context.

Second, even with the upper bound performance obtained by the best latent context there is still 26% gap till reaching the 100% accuracy. We observe that many of the still unrecognized classes either belong to "no interaction" or to the group of under-populated classes.

	Slic [5]				Felzenswalb [36]				Selective Search [134]			
	Small	Mid.	Large	All	Small	Mid.	Large	All	Small	Mid.	Large	All
Max. increase	.78	.67	.73	.79	.78	.69	.67	.78	.78	.75	.67	.79
Max. decrease	.42	.51	.40	.29	.48	.47	.04	.04	.33	.25	.04	.04

Table 3: Visual details matter. We group sub-region(s) according to their pixel size as either small (< 5%) , mid-size (> 5% & < 10%) or large (> 10%) of the whole image. Removing even a small region yields an increase or a decrease in the accuracy.

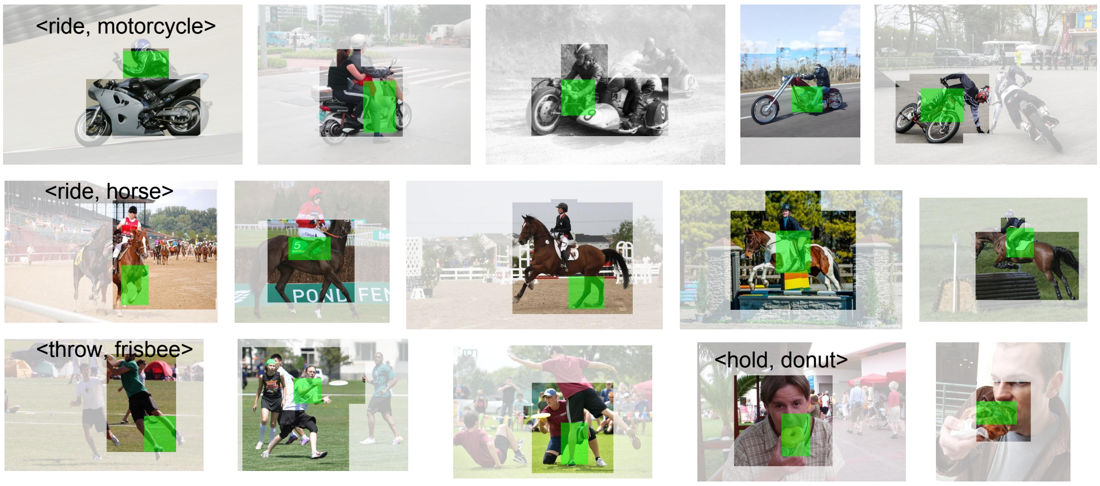


Figure 8: Details matter for interaction recognition. Green box is Selective search box [134] that leads to the maximum decrease when removed from the best latent context. See how body parts such as legs of the subject (frisbee players) or object (horse) are used in recognition.

We provide qualitative examples of best latent context obtained by the oracle results in Figure 9. These are example images where the preferred best latent context is rather small, as the best accuracy is obtained around the union region. See how the best latent context is helpful especially in cases when the union region is small (*e.g.*, group activities like races), and also in cases when the background includes distracting objects (*e.g.*, for hand-object interactions in the second row).

We confirm and conclude that the best latent context can only be discovered on a per image basis. What is more, factors like the class frequency of the interaction as well the size of the interaction are important for deciding the best latent context.

2.6.3 *Exp 3: Are Details Important Within The Best Latent Context?*

Visual details matter

In the previous experiments, we showed that there is a best latent context, with an oracle procedure provided to us using morphological operations. We also found that the best latent context makes sense only on a per image basis. It is unclear, however, whether the best latent context is determined in the entirety of the region, or specific local details

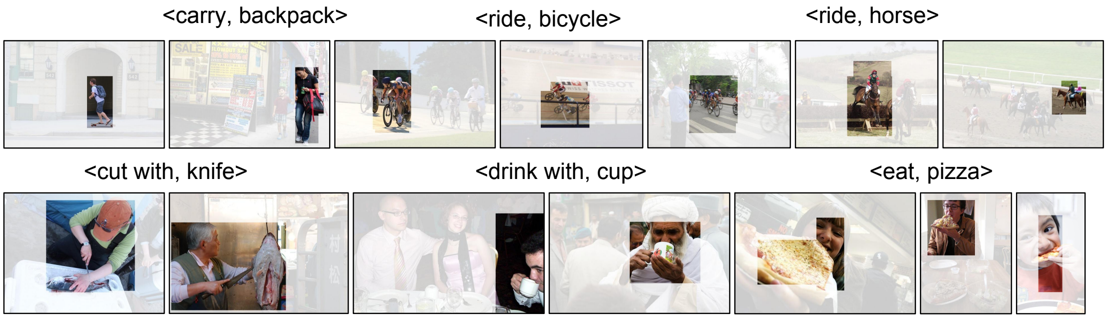


Figure 9: Results of the best latent context per image on some test cases. Latent context is highlighted whereas rest of the image is semi-transparent. Even though context is shown to be highly beneficial for recognition, these examples prefer the vicinity of the union. When i) union region occupies a small region (`<carry, backpack>`, or group activities like `<ride,bicycle>` or `<ride,horse>`) or ii) when hand is active in interaction (`<cut with, knife>`, `<drink with, cup>` or `<eat, pizza>`) latent context focuses around the union.

influence the best latent context considerably. To test this in this part we remove selective search [134] or superpixel sub-regions [5, 36] according to the procedure described in Section 2.4.2. We present results in Table 3. We provide both the maximum increase or the maximum decrease attained once a sub-region is removed. We group sub-region(s) according to their pixel size as either small (< 5%) , mid-size (> 5% & < 10%) or large (> 10%) of all pixels in the full image. The reference accuracy is the best latent context performance of 0.74% mAP.

We observe that even removing a small region leads to a slight improvement of 4% from the reference mAP of 0.74%. These visual details correspond to negative evidence for the target interaction categories within the best latent context region: once this negative evidence is removed, we obtain better recognition. While removing negative evidence has a small but noticeable effect to the final recognition accuracy, the impact of removing positive evidence local details is much larger. When focusing on the small sub-regions only, we observe a drop in accuracy in the range of 20% – 40% depending on the type of sub-region. We observe similar drops also for middle and large sub-regions. This indicates that specific, local visual details serve as crucial positive evidence for recognizing interactions. Once these are removed, the recognition is seriously impaired. We visualize some of these sub-regions in Figure 8. See how the classifier relies on the legs, a small portion of the torso or the hands to recognize the interaction.

We conclude that the selection of the visual details within the best latent context helps further interaction recognition.

Visual details prefer contact region

While the previous experiment showed that small visual details have a noticeable effect on the best latent context, it is not clear whether these visual details have a location preference. Thus, we examine whether the visual details leading to an increase or decrease in the recognition of interactions usually lie within one of the 6 observable regions defined in Figure 3. We take the average over all sub-region(s) that overlaps with one of the 6 observable regions that we consider in this work. The results are presented in Table 4.

	Subject	Object	Contact	Union	Context	Full image
Size	.15	.17	.04	.28	.47	1.00
AP	.31	.30	.32	.33	.11	.74

Table 4: Visual detail location preference. Contact region, despite being 4% *of the image, is highly discriminative among the observable regions.*

We note that contact region, despite only occupying 4% of the image, is better among the observable regions. This is also visible from Figure 8, where the contact region between the subject and object during riding or holding interactions is highly discriminative. Lastly, even though the context is a necessary part of the recognition, as shown in previous experiments, the union region carries more information about the interaction.

We conclude that visual details prefer different parts of the image, with a special focus on the contact region.

2.7 CONCLUSION

In this paper we study the visual extent of human-object interactions. We show that there is a potential gain in considering an expanded union box up to the half of the image, as can be seen from Figure 5. This simple finding can benefit many interaction recognition models, especially the ones that focus mostly on body-part localization [30, 34, 42, 94]. This also shows that the full image is not needed for most of the interaction classes.

Indeed, only 10% of classes keep increasing in performance until the full image, evident from Figure 6. These are interactions where the interacted object co-occurs within the image like `<pick, apple>` `<herd, sheep>`. In the rest of the cases context is either negative (10% of the classes) or indifferent (80% of the classes). The context is negative when there is ”no-interaction” between subject-object. 80% of cases are indifferent to context, where picking half of the image suffices. This three-way categorization of interaction classes indicates that interactions should not be treated equally. We show that there is a potential gain by finding the best latent context per-image in an oracle fashion, an improvement even for hard cases like small-size or under-populated interactions, as can be seen from Table 2. These results are motivating for [95] which considers the full context for all images for recognition to arrive at a higher accuracy than the base model of [46]. It shows that there is even a bigger potential gain for this model by determining the amount of context depending on the interaction class or more importantly depending on the image. Lastly, we show that interaction recognition is sensitive to small details that, once they are invisible, the recognition is altered. These regions generally lie within the contact region, where the subject-object touches each other.

We conclude that the localization pays off well for interaction recognition, leaving a room for substantial improvement.

3

SELF-SELECTIVE CONTEXT FOR INTERACTION RECOGNITION

3.1 INTRODUCTION

The goal of this paper is to recognize Human-object interactions from a single image. Human-object interaction recognition is an important problem with applications in areas such as robotics [31, 81, 148], image captioning [56] or visual question answering [95]. The task is to identify the relationship between a human and an object in terms of a `<verb, noun>` pair, such as `<ride, bicycle>`. Since the type of interaction is highly correlated with the scene (*i.e.* riding bicycle on the city street), researchers tackle this problem via integrating scene into the deep Convolutional Neural Networks.

Initially, Gkioxari *et al.* [45] augments human appearance with the global scene in a late-fusion manner by combining human and scene classifiers. However, late-fusion cannot model the correlation between the human and the scene. To that end, Mallya and Lazebnik [95] combines human appearance with the global scene early in the network layers, leading to a significant increase in the performance. Later, Fang *et al.* [34] improves this model by further augmenting early-fused global scene appearance with the scene objects.

We identify the following problems with this approach. First, incorporating scene early in the network layers increases network parameters, limiting the efficiency. Second, changes in the scene appearance (*i.e.* a clutter object) yields noisy filter responses, limiting the accuracy. Third, Human-object interactions offer a multitude of contexts beyond the global scene, see Figure 24, which is yet to be explored.

In this work, we propose Self-Selective Context (SSC) to circumvent the aforementioned problems. SSC learns to select the discriminative context(s) conditioned on the input image. It considers the joint appearance of human-object and context to decide which context feature(s) are discriminative. To take advantage of the multitude of contextual features offered by human-object interactions, we also devise contextual features that model the locality of the interactions. Our experiments on three large-scale benchmarks reveal that indeed the proposed contextual features are discriminative, and SSC can further boost the performance by selecting the discriminative feature in a scalable fashion.

Our contributions are as follows:

1. We propose novel contextual features to represent the locality of humans, objects, scene, and human-objects.

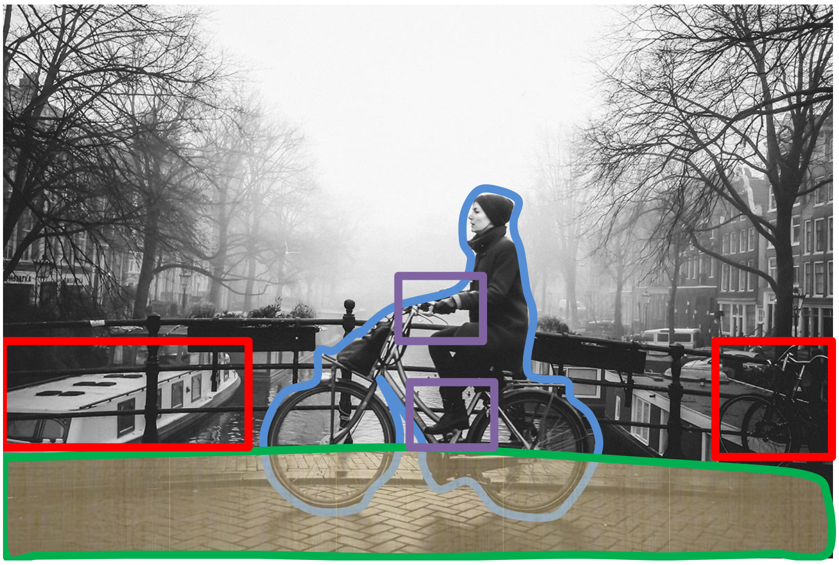


Figure 10: Human-object interactions come with many contexts that can help in recognition. In the example above, utilizing the body-parts, the deformation, and the surround scene can ease the recognition of `<ride, bicycle>`*. However, background objects like the boats can mislead the recogniton. In this paper, we first develop novel contextual features, as well as a context selection scheme Self-Selective Context to rely only on the most discriminative contexts.*

2. We propose Self-Selective Context to selectively utilize the discriminative context depending on the input image.

3. On three benchmarks for interaction recognition, SSC, combined with our novel contextual features, improves baseline models while being three times more parameter-efficient.

3.2 RELATED WORK

3.2.1 *Human-object interaction recognition*

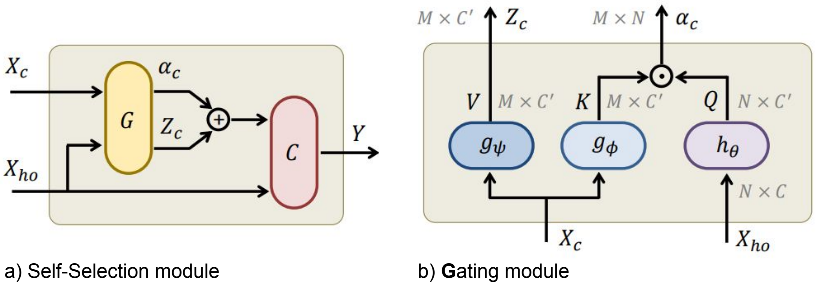
Recently there has been good progress on interaction detection [17] which requires bounding box annotations for each interaction in the image. In our work, we instead focus on interaction recognition, which is an image-level classification task since many images exhibit humans manipulating objects.

Human-object interaction recognition is defined as single image multi-label classification task in [18]. The authors collect a large-scale dataset named HICO with multiple image-level annotations for concurrent interactions, such as `<ride, bicycle>` and `<hold, bicycle>`. In our paper, we resort to the definition of HICO for interaction recognition. HICO allows researchers to train deep CNNs where they combine human and global scene context [34, 45, 95] to classify the interaction.

HICO dataset is collected by `<verb, noun>` queries-only, therefore the contextual diversity is limited (*i.e.* most interactions occur in their canonical contexts). This prevents seeing the models generalization abilities across different environments of the same interaction. To tackle this, in our paper we collect a new dataset we name **C**ontextualized **Int**eractions (CINT) which exhibits interactions within diverse contexts. In addition, we develop novel context features to leverage the locality of the human-object interactions. The locality is more robust to changes in the visual context, as we demonstrate through our experiments on HICO [18], V-COCO [52], and CINT.

3.2.2 *Combining multiple contexts in action recognition*

Multiple cues are helpful in action recognition, where the dominant approaches are fusion [67, 107, 126]. Such approaches can handle single context, however, they fall short in case of multiple contexts. Early fusion scales quadratically with the number of fused contexts. Late fusion does not leverage the correlation of human-object feature and context feature. To that end, in our work, as inspired by the Self-attention [135], we develop Self-Selective Context. SSC scales sub-linearly with the number of fused contexts. And it leverages the correlation of human-object and context.



a) Self-Selection module

b) **G**ating module

Figure 11: Overview of our method. On the left, the Self-Selection module. It takes as an input the features $\mathbf{X}_{ho}$ of N human-object pairs in a certain image, and M context features $\mathbf{X}_c$ corresponding to the image. Then, it modulates the context features $\mathbf{X}_c$ using a novel Gating module $G(\cdot)$. The final image-level features are then feed-forwarded to the classifier $C(\cdot)$ to predict the human-object interactions in the image. On the right, the Gating module $G(\cdot)$, inspired by the Self-attention [135]. The main purpose of $G(\cdot)$ is to embed the heterogeneous context features $\mathbf{X}_c$ into a compact representation $\mathbf{Z}_c$. $G(\cdot)$ predicts the vectors α used for Self-Selection of the embedded context features $\mathbf{Z}_c$.

3.3 METHOD

3.3.1 *Overview*

The goal of this paper is to map an input image to the correct interaction class $I \rightarrow Y$. If the image contains one pair of human and object, we can represent this pair as the feature x_{ho} using off-the-shelf CNN [124]. Then, using a classifier $C(\cdot)$, one can predict the probability scores of interaction classes Y. However, this paper argues that recognizing the interaction based on only the the human-object feature x_{ho} is sub-optimal. It is preferred to complement x_{ho} with more prior representations. Several sources of contexts are the perfect choices for such prior representations. As such, this paper contributes to the following. 1) We define new sources of contexts, along with their feature representations $\mathbf{X}_c$, see section 3.3.5. 2) We propose a new method, Self-Selection Context (SSC) that learns how to complement the human-object feature x_{ho} with the corresponding context features $\mathbf{X}_c$, see section 3.3.2.

3.3.2 *Self-Selection Context*

Given an input image I comprising a set of N human-object pairs. Each pair possibly describes the interaction in such an image. These N pairs are represented as features $\mathbf{X}_{ho} = \{x^j_{ho}\}^N_{j=1}$ using off-the-shelf human-object extractor $f_{ho}(\cdot)$. In addition, we are given a set of M sources of contexts, corresponding to the input image I. These contexts can be represented as features $\mathbf{X}_c = \{x^i_c\}^M_{i=1}$. Each context feature x^i_c is obtained from a different off-the-shelf context extractor $f^i_c(\cdot)$. All of our feature extractors $[f_{ho}(\cdot), f^i_c(\cdot)]$ build upon the same CNN [124] for a fair comparison. Hereafter, the *global layer* is used to refer to the last fully convolutional layer `conv5_3` of the CNN.

The goal of SSC, see Figure 11b, is to complement each human-object feature x^j_{ho} with the corresponding context features $\mathbf{X}_c$. That is why the core of the SSC is a novel Gating module $G(\cdot)$, see Figure 11a. $G(\cdot)$ serves two purposes. First, as the context features are heterogeneous, comes from different extractors, and have different feature dimensions, $G(\cdot)$ embeds $\mathbf{X}_c$ into a compact features $\mathbf{Z}_c$ with a unified feature dimension. Second, it predicts the gating vectors $\mathbf{ff} = \{\alpha^j\}^N_{j=1}$, where $\alpha^j \in \mathbb{R}^M$ is the gating vector corresponding to the j-th human-object feature and all the M context features $\mathbf{X}_c$. After the Gating module, SSC complement each human-object features x^j_{ho} with multitude of corresponding context features $\mathbf{Z}_{ho}$ in an adaptive manner:

$$x^j = \texttt{cat}\left(x^j_{ho}\ ,\ \sum_{i=1}^{M} \alpha^{ij} * z^i_c\right), \tag{3.1}$$

where $\texttt{cat}(\cdot)$ is the concatenation operation along the feature dimension. The output interaction feature x^j is feed-forwarded to a Multi-Layer Perception (MLP) $C(\cdot)$ for interaction classification.

3.3.3 *Gating Module*

The main goal of the Gating module is to select (*i.e.* gate) the most relevant context sources for each human-object pair $G(\cdot)$. This gating helps in recognizing human interaction by incorporating the prior knowledge in such contexts. Tthe gating mechanism $G(\cdot)$ is conditioned on the human-object x^i_{ho} and it contexts X_c, *jointly*.

The Gating module $G(\cdot)$ takes as an input the set of N human-object features $\mathbf{X}_{ho} = \{x^j_{ho}\}^N_{j=1}$. Also, $G(\cdot)$ takes the set of M context features $\mathbf{X}_c = \{x^i_c\}^M_{i=1}$. These features $\mathbf{X}_c$ are heterogeneous, each x^i_c is obtained from a different context extractor $f^i_c(\cdot)$ with a different space dimension. Thus, the first step is to embed each x^i_c into a common dimension using linear mapping $g^i_\psi(\cdot)$. The outcome is the embedded context features $\mathbf{Z}_c = \{z^i_c\}^M_{i=1}$. Then, we need to measure the correlation between the human-object features $\mathbf{X}_{ho}$ and their corresponding context features $\mathbf{X}_c$. But since both $\mathbf{X}_{ho}$ and $\mathbf{X}_c$

have different space dimensions, we use two linear embeddings $g_\theta(\cdot)$ and $g_\phi(\cdot)$ to map $\mathbf{X}_{ho}$ and $\mathbf{X}_c$, respectively, into two spaces with common dimension C',

$$\mathbf{Q} = g_\theta(\mathbf{X}_{ho}) \tag{3.2}$$

$$\mathbf{K} = g_\phi(\mathbf{X}_c). \tag{3.3}$$

The outcome is the features $\mathbf{Q} \in \mathbb{R}^{N\times C'}$ and $\mathbf{K} \in \mathbb{R}^{M\times C'}$, respectively. Then, we measure the pairwise correlation between the N human-object pairs $\mathbf{Q}$ and M contexts $\mathbf{K}$ using an inner product

$$\alpha = \texttt{softmax}\left(\mathbf{Q} \odot \mathbf{K}^\top\right), \tag{3.4}$$

where $\alpha \in \mathbb{R}^{N\times M}, \mathbf{ff} = \{\alpha^j\}_{j=1}^N$ are the gating vectors. Each gating vector $\alpha^j \in \mathbb{R}^M$ represents how the j-th human-object pair correlated with all the M contexts. Notice that α is activated with `softmax` along with the M contexts as a way of normalization. The next step is to use the gating vector α^j to pool the embedded context features $\mathbf{Z_c}$ into the final context feature $x'_c \in \mathbb{R}^{C'}$. x'_c is calculated as the inner product $\odot$ between α and $\mathbf{Z}_c$

$$x'_c = \alpha^\top \odot \mathbf{Z}_c. \tag{3.5}$$

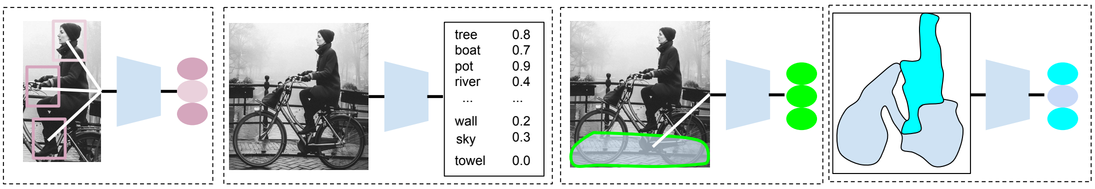
To obtain the final interaction feature x^j, both the human-object feature x^j_{ho} and its corresponding pooled context feature x'_c are concatenated along the feature dimension as shown in Equation 3.1. This feature x^j is feed-forwarded to the classifier $C(\cdot)$ to obtain the probability scores $Y = C(x^j)$ of classifying the interaction represented by input human-object pair x^j_{ho}.

Classifying interaction. So far we have described how to combine human-object feature x^j_{ho} with context features $\mathbf{X}_c$. An image potentially has multiple human-objects ($N = 12$ in our case), and it is not known which human-object pair is conducting the interaction (among all possible pairs). To that end, we resort to Multiple Instance Learning (MIL) as is the common practice [46] to obtain the final image-level classifier response. Specifically, we first obtain the classifier predictions for each human-object pairings $Y \in \mathbb{R}^{N\times S}$ where S denotes the number of interaction categories, then we apply max-pooling over the human-object dimension to obtain the final image-level interaction response $Y' = \texttt{pool}\left(Y\right), Y' \in \mathbb{R}^S$.

Worth mentioning that the classifier $C(\cdot)$ is a Multi-Layer Perceptron (MLP) with two hidden layers. Each hidden layer is followed by `BatchNorm` and `ReLU` non-linearity. The output layer uses `sigmoid` non-linearity, to handle multiple labels per-image. Up till now, we have described our method, SSC. In the following, we first describe the human-object extractor $f_{ho}(\cdot)$. Then, we complement with our contextual extractors $f^i_c(\cdot)$.

3.3.4 *Human-object Features* $\mathbf{X}_{ho}$

Our human-object features are obtained from CNN [124] that takes as input an image and returns the global appearance features for N human-object pairs in the image. To achieve this, we first detect possible human-object locations using an off-the-shelf object



a) body-part context b) stuff context c) surround context d) deformation context

Figure 12: The context features proposed in this paper. a) *Body-part context models the appearance of the human joints,* b) *Stuff-context models the occurrence of stuff-like regions in the image,* c) *Surround context models the appearance of local segments around humans,* d) *Deformation context models the shape of the human-object posture.*

detector [116]. We select the top-3 detected humans and the top-4 detected objects, based on the detector confidence. The permutations of 3 humans and 4 objects yield $3 \times 4 = 12$ distinct human-object pairs. We compute the union region for each pair, and then apply Region-of-Interest (ROI) pooling [43] over this region to obtain the final features per human-object. ROI pooling is applied over the global layer of the network.

3.3.5 *Contextual Features* $\mathbf{X}_c$

For an image, we extract M different context features, $\mathbf{X}_c$, see Figure 12, where $M = 4$ where each feature x_c^i is obtained from a different extractor $f_c^i(\cdot)$, which we detail below.
Body-part context feature. This context, see Figure 12a, models the local body-parts of the human interactor. Local body-parts carry distinctive information about an interaction, especially for grasping. The context is implemented as follows.

First, the context requires an initial set of human body-part regions. Hence, given each detected human, we run an off-the-shelf human keypoint detector [35]. This yields 17 distinct body-part regions such as the knee, the hand or the head. We draw a regular bounding box around each keypoint region. Given these bounding boxes, we apply ROI pooling over each region for each detected human from the global layer of the CNN.

However, not all body-parts contribute equally to the interaction. To select the most discriminative body-part, we then feedforward each ROI-pooled part feature to an attentional sub-network $f_{att}(\cdot)$ that yields a scalar value per-part indicating their relevance. Based on the obtained scores, we only select top-k regions for further processing ($k = 3$). Nonetheless, this indexing operation is non-differentiable. To that end, we employ a strategy called straight-through estimator [11] to by-pass the gradient computation for the non-selected body-parts. In practice, the gradients of the non-selected body-parts are set to 0. Finally, the selected body-part features are concatenated and further compressed with two fully connected layers, leading to a $1k$ dimensional feature summarizing distinctive human parts.
Stuff context feature. This context models the existence of local stuff-like regions, such as trees, wall, river, see Figure 12b. The existence of such regions can give hint about the interactions, *e.g.* river for boat riding.

We implement the stuff-context as the class probabilities of a stuff-classifier. Stuff-classifier is a linear classifier on top of the global layer response of the CNN using the annotations from [14]. The final response is a 91−dimensional feature vector summarizing the existence of local stuff-like regions in the image.

Surround context feature. This context models the local surround around human-object, see Figure 12c. To represent the local surround, we make use of the semantic segments like the road, the sky or the sideways, obtained from an off-the-shelf model [155] over the input image.

Specifically, we first create binary masks from the input image of size $m \in \mathbb{R}^{H \times W \times K}$, where $[H, W]$ denotes the height and width of the global layer feature map and K is the number of distinct segments. For each image, we choose the top K segments with the largest scales ($K = 5$). For each mask, the values of only the respective segment are set to 1 and all else is set to 0. Masks are then applied to the global activations (conv5_3) of the CNN, which is then average-pooled over the spatial dimension to obtain the features of size $\mathbb{R}^{K \times D}$. Finally, we aggregate over the semantic segment dimension K to obtain $\mathbb{R}^{1 \times D}$ feature response ($d = 512$) via max pooling.

Deformation context feature. This context models the mutual deformation of human-objects, see Figure 12d. Our goal is to encode two important cues of interactions: 1) Shape of the human-object deformation, 2) Spatial relation of the human-object location simultaneously.

To that end, given an image, we first obtain object segmentation predictions [54] of size $\mathbb{R}^{H \times W \times K}$ where $[H, W]$ are the heights and width of the image, and K is the number of distinct objects ($K = 80$). We resize this feature map such that the longest side is 64 pixels, using bi-linear interpolation, and process it with a three-layer CNN along the channel dimensions $80 \rightarrow 128 \rightarrow 256 \rightarrow 512$. The first layer is a 1×1 convolution followed by a 3×3 kernel. Keeping the first convolution 1×1 is crucial – The input mask is sparse (*i.e.* most locations are 0), which is hard to process with a dense filter of size 3×3 from the start. To that end, we first generate a denser feature map using 1×1 filters to process with subsequent layers. Finally, we pool the response over the spatial dimension to obtain 512-dimensional deformation context feature.

Implementation details. All the models are implemented in PyTorch [106], trained and optimized with SGD. The CNN [124] is pre-trained on ImageNet [29]. Then it is fine-tuned on the HICO dataset for *epochs* with a learning rate 0.001 that is decayed by a factor 0.1 after 15 epochs.

3.4 EXPERIMENTS

3.4.1 *Contextualized Interactions Dataset*

Existing datasets [18, 52] are limited in their context repertoire since they are collected with `<verb, noun>` queries only. This prevents observing the contribution of the context in diverse environments. To that end, we collect a new challenging dataset we name **C**ontextualized **Int**eractions (CINT).

To create CINT, we first queried Google Images [48] via triplets of `<verb, noun, context>` queries, where `<verb, noun>` pairs are from HICO and 40 context queries are derived from scene datasets [80, 144, 154]. Context queries are the time of the day (*e.g.* day, night, 7/40), state of the scene (*e.g.* sunrise, snowy, dark, 12/40) or the location of the scene (*e.g.* railroad, beach, street, 21/40). After omitting the queries with no results, we ended up with around 2*k* distinct `<verb, noun, context>` queries of which we use in the annotation procedure.

We initially downloaded 250 images per-query and then removed images where the human, the object and the context are not visible. We have 16*k* images from 200 distinct interactions.

3.4.2 *Other Datasets*

HICO dataset [18]. For training we rely on HICO. It contains 40*k* training and 10*k* testing images from 600 distinct `<verb, noun>` pairs, for 117 verbs and 80 nouns. HICO exhibits a long-tailed distribution: 155 classes have less than 10 examples. A special case is `no-interaction`, where the target object and a human is visible whereas not interacting.

V-COCO dataset [52]. V-COCO was initially designed for interaction detection [44]. To demonstrate the generality of our approach, we re-purpose the dataset for interaction recognition as follows. We align class namings with HICO, for example by splitting `<skateboarding>` into `<ride, skateboard>`). We omit actions like smiling that do not correspond to any object. Finally, we follow the best practice of [18] and aggregate different interaction instances within the same image over the image label. As a result, V-COCO has 4*k* test images with 226 distinct interactions. A big portion of the dataset belongs to interactions with rare categories, making the inference challenging.

3.4.3 *Baseline models*

Interaction recognition baselines. To prove the generality of our approach with different human-object representations, we plug in our SSC to three different human-object feature extractors, namely: 1) VGG-16 [124] that is pre-trained on ImageNet and fine-tuned on HICO, 2) ContextFusion [95], 3) Global stream of PairAtt [34]. All the features are extracted from the penultimate layer of `fc7` using the code from the respective authors.

Fusion baseline. We compare SSC to a fusion baseline, where we concatenate human-object features with different context features.

3.4.4 *Evaluation*

For evaluation, we use the instance-based Mean Average Precision (mAP). That is, we evaluate the prediction performance per-image, which is then averaged over the respective dataset. This metric allows us to observe the effect of fusing or Self-Selecting different forms of context features on a per-image basis. Additionally, it allows us to analyze the effect of the context over the characteristics of interactions such as the size, the population, or the existence of interaction.

3.5 EVALUATION

3.5.1 *Self-Selection of the Single Context*

In the first experiment, we validate the discriminative ability of the proposed context features on HICO. We select and combine single context features with the human-object feature, to see their individual contributions. Results are in Table 5.

Table 5: Self-Selection of The Single Context.

Context Feature	mAP(%)	Improvement $\Delta \uparrow$
Human-Object (HO)-only	62.50	-
HO + Body-part context	68.55	6.05
HO + Stuff context	68.20	5.70
HO + Surround context	68.44	5.94
HO + Deformation context	68.30	5.80

Each of the 4 context features, by itself, are complementary to the human-object feature. Each improves the performance of human-object features considerably.

Body-part context helps the most with 6.05 mAP since many interactions are localized on the fine-grained body-parts such as the hand-object interactions like cutting, eating or cooking. Also, the surround context helps with 5.94 mAP, indicating that the immediate surround of the human-objects is distinctive, such as the road for transport vehicles. Deformation context helps with 5.80 mAP since it complements the human-object features with the mutual position and the deformation information, which is crucial in dynamic interactions like jumping or throwing. Lastly, stuff context helps with 5.70 mAP which confirms that high-level representation of surrounding object-like regions can help distinguish the interaction.

To conclude, we observe that context features are distinctive for recognition and complement human-object features. Also, each different context feature specializes in different interactions which call for an effective combination.

3.5.2 *Self-Selection of the Multiple Contexts*

This experiment validates the complementary power of the proposed context features via Self-Selection. We select and combine all 4 contextual features with human-objects. Results are in Table 6.

It is observed that both fusion and SSC improves upon human-object-only feature, confirming the complementary power of the proposed contexts. We also see that SSC performs better than fusion across all three datasets. The difference is more pronounced on CINT, which highlights the need for selecting the discriminative context in diverse environments.

Table 6: Self-Selection of The Multiple Contexts.

	Dataset		
Method	HICO	V-COCO	CINT
HO-only	62.50	52.27	45.24
HO + Fusion	69.59	54.74	49.74
HO + SSC (Ours)	**70.78**	**55.00**	**54.36**

SSC does so by using three times fewer parameters, as can be seen from Figure 13. Figure 13 plots the recognition mAP as a function of the number of cumulative contexts added at each step (from 1 to 4). We initially add stuff context and then add 1 more context at each step. As can be seen, Self-Selective Context uses 3 times fewer parameters than the fusion counterpart (4.9 Million vs. 13.6 Million) while yielding better results. This is expected since many human-object interactions have limited examples in the training set, hence making the learning difficult.

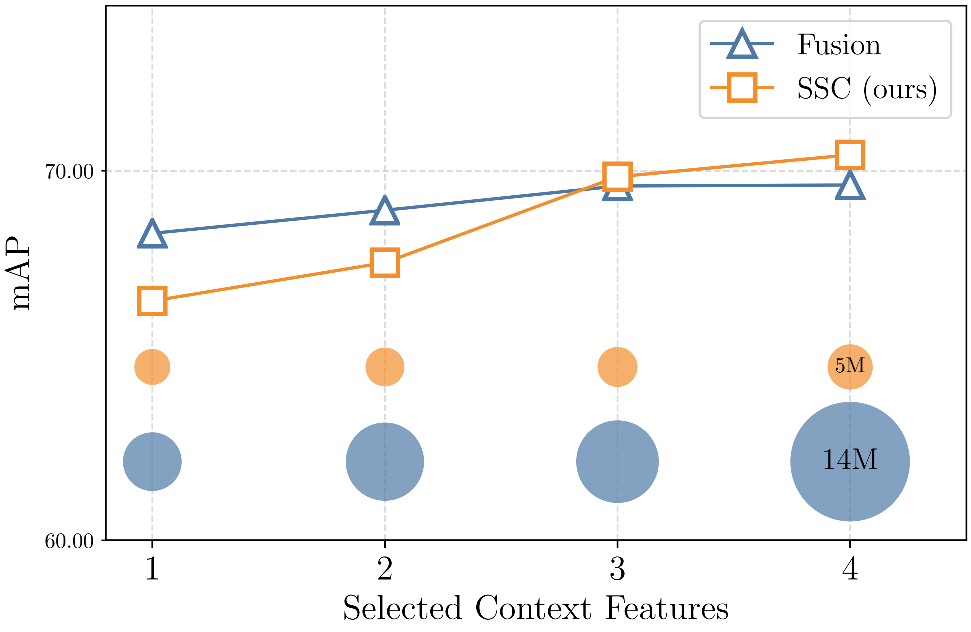


Figure 13: Parameter efficiency of Self-Selection vs. Fusion. Amount of parameters for each step for the respective technique is presented in log-scale using circles.

To conclude, we observe that the context features are complementary to each other, and, Self-Selective context provides a parameter efficient and accurate combination of contexts.

3.5.3 *Further Analysis of Self-Selection*

In this Section, we present further analysis of Self-Selection, on the source of the improvement, the ablation of the joint conditioning, and the distribution of Self-Selection.
Source of the improvement. To shed some light into where the gain comes from, we marginalize the improvement of SSC over human-object features in Table 7.

Table 7: The Source of the Improvement.

Method	Pixel Area		Population		Existence	
	Small	Large	Rare	Frequent	No	Yes
HO-only	60.09	63.98	39.31	61.68	36.96	64.26
HO + **SSC**	69.45	72.70	51.38	70.79	47.67	73.24
Δ ↑	**9.36**	8.72	**12.07**	9.11	**10.71**	8.98

1) SSC helps slightly better for small human-object interactions. During this experiment, an interaction is deemed to be small if the human-objects occupy less than 20% of the whole image, and large otherwise. This indicates that when the visual details of the human-objects are limited due to size, the context becomes more important.

2) SSC helps considerably better for rare human-object interactions. During this experiment, an interaction is deemed to be rare if it has less than 10 examples in the HICO training set, frequent otherwise. This indicates that rare human-object interactions (*i.e.* `<ride, giraffe>`) exhibit distinctive local contextual appearance that, once modeled with SSC, becomes easier to recognize.

3) SSC helps considerably better for the case of no-interaction. During this experiment, we aggregate the performance over no-interaction and interaction categories separately. This indicates that SSC encodes the distinctive signals of the interaction, which is once used, helps the model to discriminative interaction from no-interaction.

Contribution of the joint conditioning. For an ablation, study we remove the joint conditioning from SSC. In this way, our model learns context relevance values α by considering the *context only* (as opposed to human-object and context together). Results can be seen from Table 8.

Table 8: Contribution of Joint Conditioning.

Condition	Dataset		
	HICO	V-COCO	CINT
context-*only*	67.77	49.58	49.26
human-object & context	**70.78**	**55.00**	**54.36**

As can be seen, using context-*only* to select the context leads to a drop over all three datasets, confirming the importance of Self-Selection jointly based on the human-object and context.

Distribution of Self-Selection. In this experiment we visualize the distribution of Self-Selection. We aggregate Self-Selection ratios over distinct nouns and verbs in HICO in Figure 14.

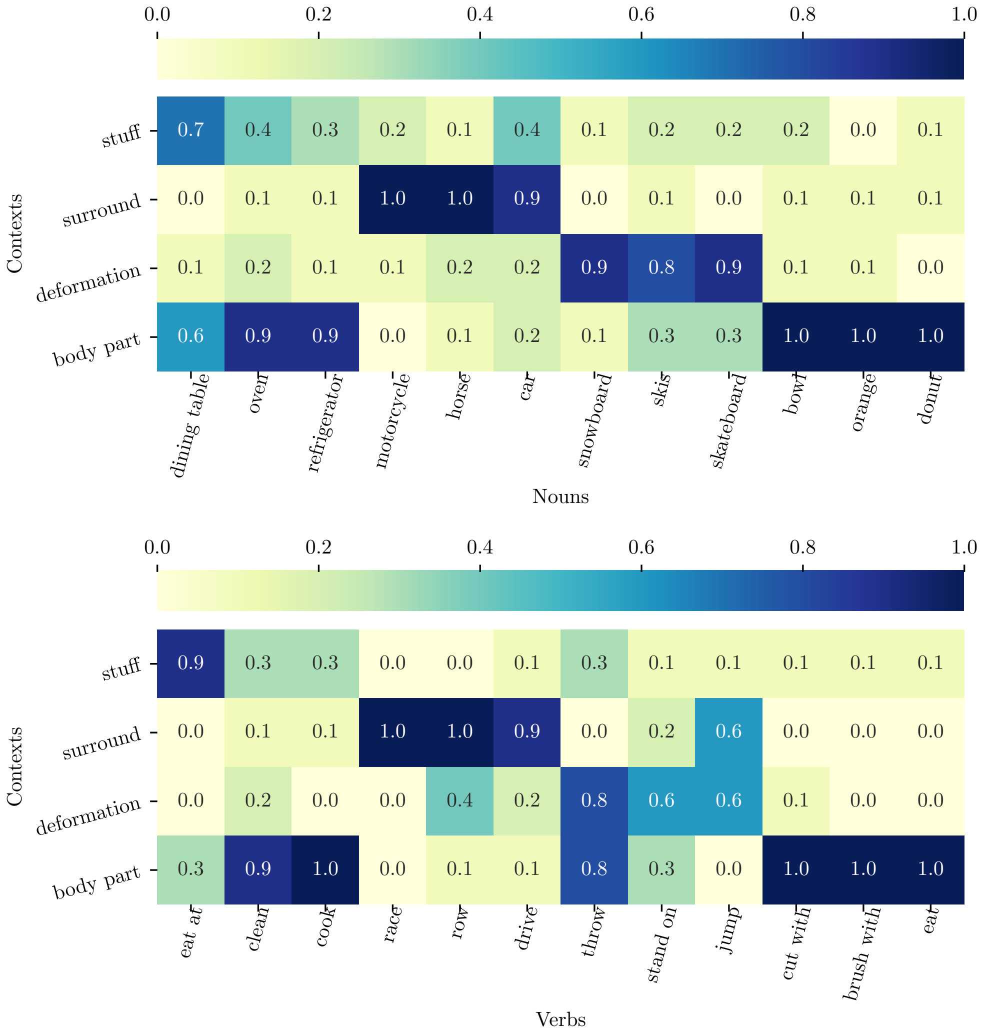


Figure 14: Distribution of Self-Selection over nouns and verbs.

Stuff context is preferred in house activities of cooking and cleaning, where the co-occurring objects are distinctive. The surround context is preferred by transportation interactions that use horses, cars, or motorcycles for racing or driving. Deformation context is preferred by sport objects like skateboard or skis for jumping or standing, where the interaction leads to unique postures in the human-object body. Lastly, hand-object interactions like cutting, brushing, or eating prefer body-part context, where the hand leads to distinctive occlusion patterns over the object region. We re-assure that the contribution of the Self-Selection is based on the interaction type.

3.5.4 *Self-Selection with the State-of-the-art*

So far, we experimented with our proposed human-object features. This experiment plugs in our Self-Selective module to existing models, namely VGG-16 [124], Context-Fusion [95], and global stream of PairAtt [34]. Results from the three datasets can be seen from Table 9.

Table 9: Combining SSC with the State-of-the-art.

Method	Dataset		
	HICO	V-COCO	CINT
VGG-16 [124]	56.10	46.91	44.83
VGG-16 [124] + **SSC**	**67.59**	**51.17**	**49.56**
ContFus [95]	63.47	51.36	46.72
ContFus [95] + **SSC**	**65.79**	**52.24**	**51.92**
PairAtt [34]	65.10	53.62	48.99
PairAtt [34] + **SSC**	**68.29**	**54.24**	**51.22**
Human-Object	62.50	51.60	47.85
Human-Object + **SSC**	**70.78**	**55.00**	**54.36**

As can be seen in all cases, incorporating Self-Selective context improves upon the respective model alone. This indicates that Self-Selective context carries complementary information for State-of-the-art models, even though these features incorporate some contexts like global surround or object co-occurrence intrinsically. An important result is that human-object features coupled with Self-Selective context still outperforms all other models, despite its simplicity. This indicates that modeling human-object and context separately and adaptively is essential for recognizing human-object interactions.

Qualitative analysis. We present success and failure cases in Figure 23. We compare the performance of PairAtt with PairAtt + **SSC** using images from CINT dataset.

In the top, we provide three examples where SSC improves upon PairAtt. We can see that SSC helps when the scene is not strongly correlated with the target interaction, such as `<sit on, chair>` on city street, or `<ride, ski>` in the night. In such particular cases, SSC can suppress the contribution of the irrelevant context feature, therefore leading to accurate classification performance.

In the bottom, we provide three examples SSC decreases the performance. We can see that when the amount of visual context is limited, as in `<carry, backpack>` on left bottom example, or when the context is too noisy, such as the background humans in the middle bottom, SSC is challenged in identifying the discriminative context. This leaves for improvement for such cases.

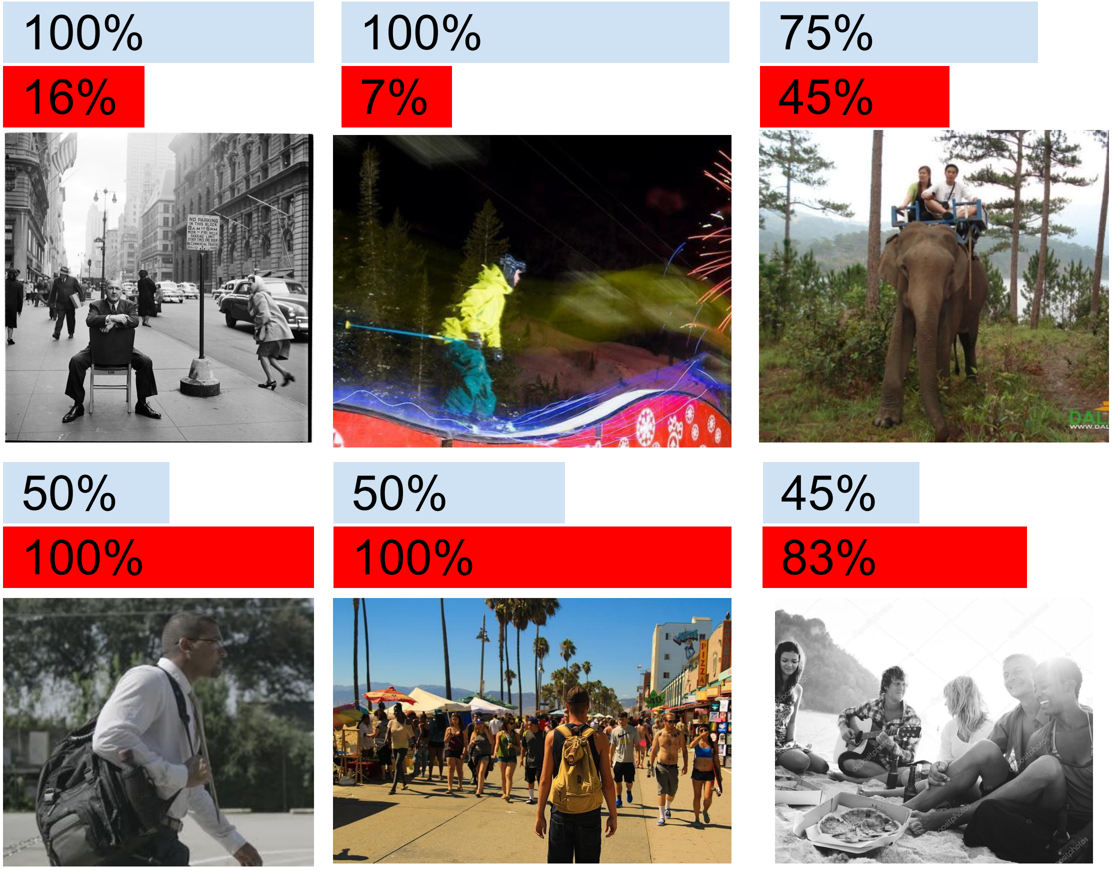


Figure 15: Qualitative examples from CINT dataset for PairAtt [34] and PairAtt+SSC(ours). We provide mAP % on top for both. SSC helps when the context is unexpected (top), however may decrease the result if the context is not visible or too noisy (bottom).

To conclude, Self-Selective context carries information for the State-of-the-art models, as shown in Table 9. The difference is considerable for all three models, on HICO, V-COCO, and CINT. Also, Self-Selective context helps the most when the context is radically different as shown in Figure 23.

3.6 CONCLUSION

In this work, we addressed the task of recognizing human-object interactions from a single image. We treated human-object interaction recognition as a task of context selection. We first devised context features to model the locality of the human, the object, the surround scene, and the human-objects. Then, we proposed a new model, namely scalable Self-Selective Context (SSC), along with a novel gating module. The gating mechanism considers the correlation between the human-objects and the corresponding contexts. And the gating module succeeds in selecting the context(s) that are most relevant to the interactions in each image. Our experiments reveal that, indeed, the proposed context features are discriminative and they complement the appearance features of human-objects. In addition, our method, SSC, improves State-of-the-art in interaction recognition on three challenging benchmarks: HICO, V-COCO and CINT.

4

STRUCTURED VISUAL SEARCH VIA COMPOSITION-AWARE LEARNING

4.1 INTRODUCTION

Visual image search is a core problem in computer vision, with many applications, such as organizing photo albums [118], online shopping [65], or even in robotics [12, 105]. Two popular means of searching for images are either text-to-image [20, 85] or image-to-image [111, 152]. While simple, text-based search could be limited in representing the *intent* of the users, especially for the spatial interactions of objects. Image-based search can represent the spatial interactions, however, an exemplar query may not be available at hand. Due to these limitations, in our work, we focus on a structured visual search problem of compositional visual search.

The composition is one of the key elements in photography [108]. It is the spatial arrangement of the objects within the image plane. Therefore, composition offers a natural way to interact with large image databases. For example, a big stock image company already offers tools for its users to find images from their databases by composing a query [2]. The users compose an abstract, 2D image query where they arrange the location and the category of the objects of interest, see Figure 24.

Compositional visual search is initially tackled as a learning problem [145], recently using deep Convolutional Neural Networks (CNN) [93]. Mai *et al.* treats the problem as a visual feature synthesis task where they learn to map a given 2D query canvas to a 3 dimensional feature representation using binary metric learning which is then used for querying the database [93]. We identify the following limitations with this approach: *i)* The method requires a large-dimensional feature ($7 \times 7 \times 2048 \approx 100k$) to account for the positional and categorical information of the input objects, limiting the memory efficiency especially while searching across large databases. *ii)* The method requires a large-scale dataset ($\approx 70k$ images) for training, limiting the sample efficiency. *iii)* The method only considers binary relations between images, limiting the compositional-awareness. To overcome these limitations, in our work, we introduce composition-aware learning.

Compositional queries exhibit continuous-valued similarities between each other. Objects within the queries transform in two major ways: 1) Their positions change (translational transformation), 2) Their categories change (semantic transformation), see Figure 24. Our composition-aware learning approach takes advantage of such transformations using the principle of equivariance, see Figure 17. Our formulation imposes the transformations within the input (query) space to have a symmetrical effect within the output (feature) space. To that end, we develop novel representations of the

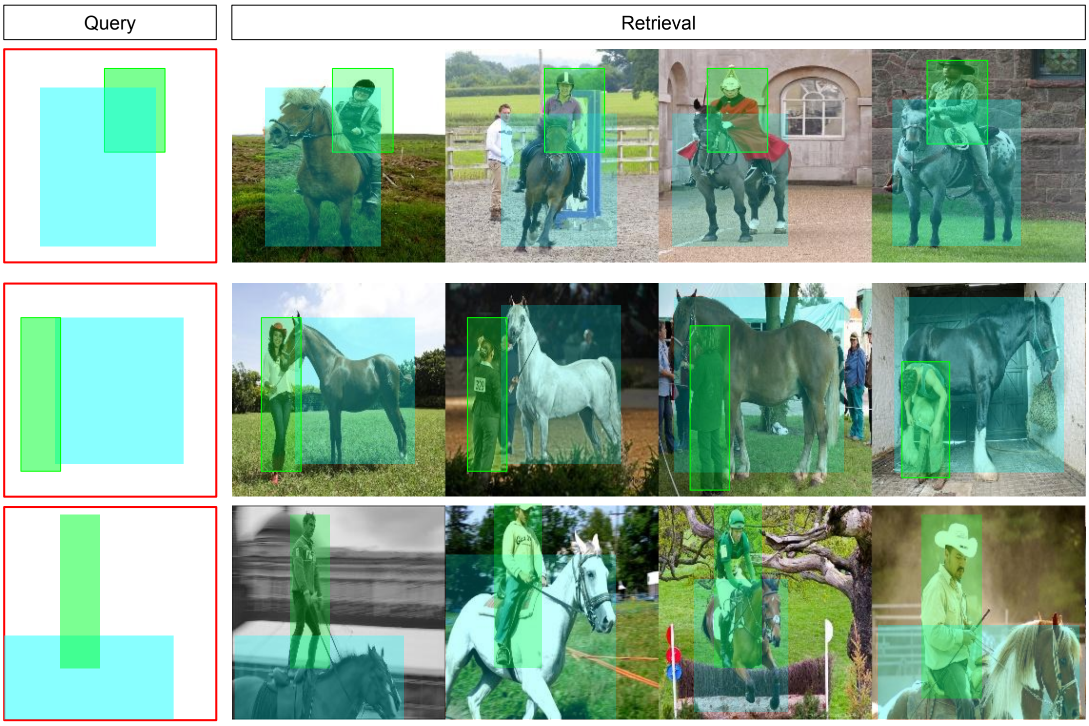


Figure 16: The compositional visual search takes a 2D canvas (left) as a query and then returns the relevant images that satisfy the object category and location constraints. Retrieval set (right) is in descending order by their mean Intersection-over-Union with the query canvas. Observe how small changes in the composition of the horse and the person lead to drastic transformations within the images. In this work, our goal is to learn these transformations for efficient compositional search.

input and the output transformations, as well as a novel loss function to learn these transformations within a continuous range.

Our contributions are three-fold:

I. We introduce the concept of composition-aware learning for structured image search.

II. We illustrate that our approach is efficient both in feature-space and data-space.

III. We benchmark our approach on two large-scale datasets of MS-COCO [89] and HICO-DET [17] against competitive techniques, showing considerable improvement.

4.2 RELATED WORK

Compositional Visual Search. Visual search mostly focused on text-to-image [19, 20, 85, 98, 121, 139] or image-to-image [9, 49, 50, 64, 84, 86, 111–113, 131, 152] search. Text-to-image is limited in representing the user intent, and a visual query may not be available for image-to-image search. Recent variants also combine the compositional query either with text [38] or image [91]. In this paper, we focus on compositional visual

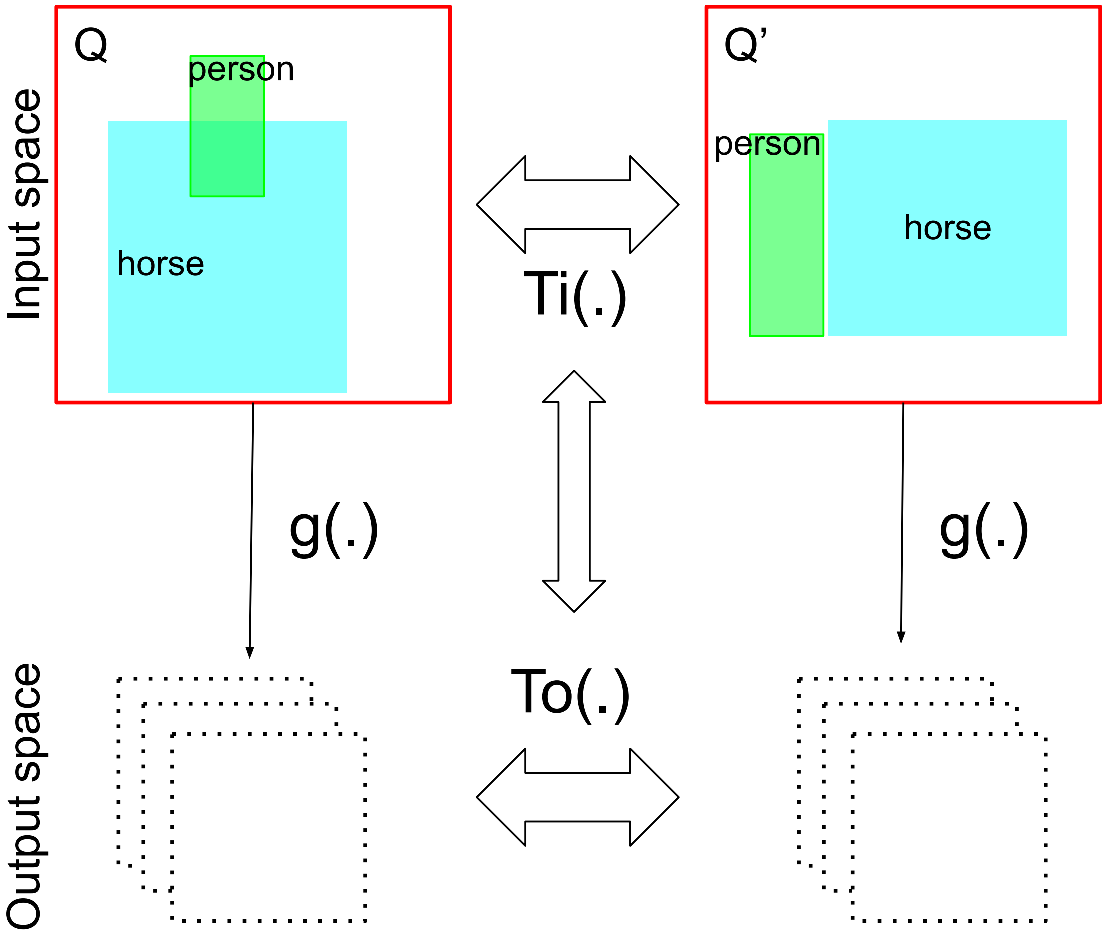


Figure 17: At the core of our technique is the principle of equivariance, which enforces a symmetrical change within the input and output spaces. We achieve this via mapping a query Q and its transformed version $Q' = T_i(Q)$ to a feature space where the transformation holds $g(Q') = T_o(g(Q))$.

search [93, 102, 145]. A user composes an abstract, 2D query representing the objects, their categories, and relative locations which is then used to search over a potentially large database. A successful example is VisSynt [93] where the authors treat the task as a visual feature synthesis problem using a triplet loss function. Such formulation is limited in the following ways: 1) VisSynt is high dimensional in feature-space ($100k$ dimensional), limiting memory efficiency, 2) VisSynt requires a large training set ($70k$ examples), limiting data efficiency, 3) VisSynt does not consider the compositional transformation between queries due to binary nature of the triplet loss [57], limiting the generalization capability of the method. In our work, inspired by the equivariance principle, we propose composition-aware learning to overcome these limitations and test our efficiency and accuracy on two well-established benchmarks of MS-COCO [89] and HICO-DET [17].

Learning Equivariant Transformations. Equivariance is the principle of the symmetry: Any change within the input space leads to a symmetrical change within the output space. Such formulation is highly beneficial, especially for model and data efficiency [32]. In computer vision, equivariance is used to represent transformations such as object rotation [26, 140, 141], object translation [68, 96, 142, 151] or discrete motions [62, 63]. Our composition-aware learning approach is inspired by these works, as we align the continuous transformation between the input (query) and output (feature) spaces, see Figure 17.

Continuous Metric Learning. Continuous metric learning takes into account the continuous transformations between the image instances [75, 79, 100], since such relationships can not be modeled with conventional metric learning techniques [25, 57]. Recently, Kim *et al.* [75] proposed LogRatio, a loss function that matches the relative ratio of the input similarities with the output feature similarities. It yields significant gain over competing methods for pose and image caption search. Since compositional visual search is a continuous-valued problem, we bring LogRatio as a strong baseline to this problem. LogRatio intrinsically assumes a dense set of relevant images given an anchor point for an accurate estimation. However, compositional visual search follows Zipf distribution [101], where, given a query, only a few images are relevant, limiting LogRatio performance.

4.3 METHOD

Our method consists of three building blocks:

1. Composition-aware transformation that computes the transformations in the input and output space,
2. Composition-aware loss function that updates the network parameters according to the divergence of input-output transformations,
3. Composition-equivariant CNN, used as the backbone to learn the transformation.

Method Overview. An overview of our method is provided in Figure 18. Our method takes as an input a 2D compositional query $q \in \mathbb{R}^{H \times W}$, where H, W are the height and width of the query canvas. This query contains a set of objects, along with their categories and positions (in the form of bounding boxes). The goal of our method is, given a target dataset of images, we want to retrieve the top-k images that are most relevant to the query q – *i.e.* relevant to both the objects and their positions. Each image I can initially be represented as feature $x \in \mathbb{R}^{H' \times W' \times C'}$ using the last convolutional layer of an off-the-shelf, ImageNet pre-trained deep CNN, *e.g.* ResNet-50 [55]. Such feature x preserves the spatial information as well as the object category information within the image I. Furthermore, we assume access to a tuple (c, x, I), where $c \in \mathbb{R}^{H \times W \times C}$ is a compositional map constructed using the object categories and bounding boxes of the query q. In addition, let $q' = T(q)$ be the transformed version of the query q, and (c', x', I') are the corresponding composition map, CNN feature and the image. The transformation T can correspond to a translation of object location(s), or a change in object categories in q. Our method trains a 3-layer CNN $g_\Theta(\cdot)$ with the parameters Θ, by minimizing the following objective function:

$$\min_{\Theta}(L_{comp}(T_i(c, c'), T_o(g_\Theta(x), g_\Theta(x')))), \tag{4.1}$$

where T_i measures the input transformation between compositional maps c and c', and T_o measures the transformation between output feature maps $g(x)$ and $g(x\prime)$, and

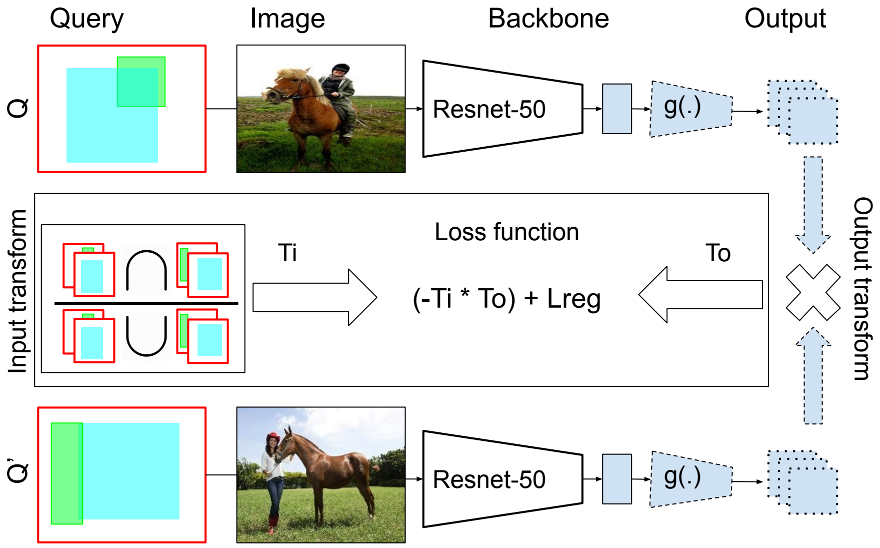


Figure 18: Our composition-aware learning approach. Our approach is trained with pairs of queries (Q, Q') *with identical backbones. 1) Given a pair of queries, we sample the corresponding images and feed them through a frozen ResNet-*50 *up to layer-*4 *of size* $7 \times 7 \times 2048$. *2) Then, we process these activations with our light-weight* 3*-layer CNN* $g(\cdot)$ *to map the channel dimension to a smaller size (*i.e. $2048 \rightarrow 256$*) while preserving the spatial dimension of* 7*. 3) In the mean-while, we compute the input (*T_i*) and the output (*T_o*) transformations, which are then forced to have similar values using the loss function.*

L_{comp} is the composition-aware loss function measuring the discrepancy between these transformations. In the following, we first describe the compositional map c, and the input and the output transformations T_i and T_o. Then, we describe composition-aware loss function L_{comp}. Finally, we describe our CNN architecture $g_{\Theta}(\cdot)$ that learns the mapping. We drop Θ from now for the sake of clarity.

4.3.1 *Composition-aware Transformation*

The goal of the composition-aware transformation is to quantify the amount of transformation between the input compositions (c, c') and output feature maps $(g(x), g(x'))$ in the range $[0, 1]$. For this, first, we construct compositional maps from the input user queries, then we measure the input transformation using these maps, and finally we describe the output transformation.

Constructing compositional map c**.** First, given a user query q that reflects the category and the position of the objects, we create a one-hot binary feature map c of size $\mathbb{R}^{H,W,C}$ where $[H, W]$ are the spatial dimension of the composition map ($H = W = 32$), and C is the number of object categories (*i.e.* 80 for MS-COCO [89]). In this map, only the corresponding object locations and the categories are set to $1s$ and otherwise $0s$. This simple map encodes both the positional and categorical information of the input composition, which we will then use to measure the transformation within the input space. We apply the same procedure to the transformed query q' which yields c'. Now given the pair of compositional maps (c, c'), we can quantify the input transformation.

Input transformation T_i. Then, our goal is to measure the similarity between these two compositions as:

$$T_i(c, c') = \frac{\sum_{xyz}(c_{xyz} \cdot c'_{xyz})}{\sum_{xyz} 1(c_{xyz} + c'_{xyz})}, \tag{4.2}$$

where 1 is an indicator function that is 1 for only non-zero pixels. This simple expression captures the proportion of the intersection of the same-category object locations in the numerator and the union of the same-category object locations in the denominator. T_i output is in the range $[0, 1]$, and will return 1 if the two compositions c and c' are identical in terms of object location and the categories, and 0 if no objects share the same location. T_i will smoothly change with the translation of the input objects in the compositions. Given the input transformation, we now need to compute the output transformation which will then be correlated with the changes within the input space.

Output transformation T_o. Output transformation is computed as the dot product between the output features as follows:

$$T_o(g(x), g(x')) = g(x) \times (g(x'))^\top, \tag{4.3}$$

where $(g(x'))^\top$ is the transpose of the output feature $g(x')$. We choose the dot product due to its simplicity and convenience in a visual search setting. T_o can take arbitrary values in the range $[-\infty, \infty]$. In the following, we describe how to bound these values and measure the discepancy between the input-output transformations T_i and T_o.

4.3.2 *Composition-aware Loss*

Given the input-output transformations, we can now compute their discrepancy to update the parameters Θ of the network $g(\cdot)$. A *naive* way to implement this would be to minimize the Euclidean distance between the input-output transformations as:

$$\min_{\Theta} ||\mathbf{T_i} - \text{œ}(\mathbf{T_o})||, \tag{4.4}$$

where $\sigma(\cdot)$ is the exponential non-linearity $\frac{1}{1+\exp(\cdot)}$ to bound T_o in range $[0, 1]$. However, such a function generates unbounded gradients therefore leading to instabilities during training [87], and reducing the performance, as we show through our experiments. Instead, cross entropy is a stable and widely used function that is used to update the network weights. However, cross entropy can only consider binary labels as $(0, 1)$ whereas in our case the transformation values vary within $[0, 1]$. To that end, we derive a new loss function inspired by the cross entropy that can still consider in-between values.

Consider that our goal is to maximize the correlation between input-output transformations as:

$$\max_{\Theta}(T_i \cdot \sigma(T_o^\top)). \tag{4.5}$$

We can also equivalently minimize the negative of this expression due to convenience:

$$\min_{\Theta}(-T_i \cdot \sigma(T_o^\top)). \tag{4.6}$$

The divergence of T_o and T_i at the beginning of the training leads to instabilities during the training. To overcome this, we include additional regularization via the following two terms as:

$$\min_{\Theta}(T_o - T_i \cdot T_o^\top + \log(1 + \exp(-T_o))), \tag{4.7}$$

where the two terms T_o and $\log(1 + \exp(-T_o))$ penalize for larger values of T_o in the beginning of the training, leading to lesser divergence from T_i. To further avoid over-flow, the final form of the regularizer terms are:

$$\min_{\Theta}(\max(T_o, 0.) - T_i \cdot T_o^\top + \log(1 + \exp(-\|T_o\|))). \tag{4.8}$$

This is the final expression for L_{comp} which we use throughout the training of our network $g(\cdot)$.

4.3.3 *Composition-Equivariant Backbone*

Our model $g(\cdot)$ is a lightweight 3-layers CNN that maps the bottleneck representation x obtained from the pre-trained network ResNet-50 of dimension $\mathbb{R}^{7\times7\times2048}$ to a smaller channel dimension of the same spatial size, *i.e.* $\mathbb{R}^{h\times w\times C}$, such as $7\times7\times256$ unless otherwise stated. Our intermediate convolutions are 2048 → 1024 → 512 → 256. The first two convolutions use 3×3 kernels whereas the last layer uses 1×1. We use stride$=1$ and apply zero-padding to preserve the spatial dimensions which are crucial for our task. We use *LeakyReLU* with slope parameter $s = 0.2$, batch-norm and dropout with $p = 0.5$ in between layers. We do not apply any batch-norm, dropout, or *LeakyReLU* at the output layer as this leads to inferior results.

Since our goal is to preserve positional and categorical information, a network with standard layers may not be a proper fit. Convolution and pooling operations in standard networks are shown to be lacking translation (shift) equivariance, contrary to wide belief [151]. To that end, we use the anti-aliasing trick suggested by [151] to preserve shift equivariance throughout our network. Specifically, before computing each convolution, we apply a Gaussian blur on top of the feature map. This simple operation helps to keep translation information within the network layers.

4.4 EXPERIMENTS

4.4.1 *Datasets*

Constructing Queries. To evaluate our method objectively, without relying on user queries and studies, we rely on large-scale benchmarks with bounding box annotations. We evaluate our method on MS-COCO [89] and HICO-Det [17]. The training is only conducted on MS-COCO. Given an image, we select at most 6 objects based on their area as is the best practice in [93].
MS-COCO. MS-COCO is a large-scale object detection benchmark. It exhibits 80 object categories such as animals (*i.e.* dog, cat, zebra, horse) or house-hold objects. The dataset contains 120k training and 5k validation images. We split the training set to two mutually exclusive random sets of 50k training and 70k gallery images. The number of objects in each image differs in the range $[1,6]$.
HICO-DET. HICO-DET is a large-scale Human-object interaction detection benchmark [17, 69]. HICO-DET builds upon 80 MS-COCO object categories, and collects interactions for 117 different verbs, such as ride, hold, eat or jump, for 600 unique `<verb, noun>` combinations. Interactions exhibit fine-grained spatial configurations which makes it a challenging test for the compositional search. The dataset includes 37k training and 10k testing images. The training images are used as the gallery set and the testing set is used as the query set. A unique property of the dataset is that 150 interactions have less than 10 examples in the training set, which means a query can only match very few images within the gallery set, leading to a challenging visual search setup [70]. HICO-DET is only used for evaluation.

4.4.2 *Evaluation Metrics*

We evaluate the performance of the proposed model with three metrics. Standard mean Average Precision metric as is used in VisSynt [93]. Also, we borrow continuous Normalized Discounted Cumulative Gain (cNDCG) and mean Relevance (mREL) metrics used in continuous metric learning literature [75, 79, 100] All metric values are based on the mean Intersection-over-Union (mIOU) scores between a query and all gallery images described below. For all three metrics, higher indicates better performance.

Mean Intersection-over-Union

To measure the relevance between a query and a retrieved image, we resort to mean Intersection-over-Union as is the best practice [93]. Concretely, to measure the relevance between a Query q and a retrieved image r

$$mIOU(q,r) = \frac{1}{B_q} \sum_{b_i \in B_Q} \max_{b_j \in B_I} 1(k(b_i) = k(b_j)) \frac{b_i \cap b_j}{b_i \cup b_j}, \tag{4.9}$$

where B_Q and B_I represents all the available objects in the query Q and retrieved image I respectively, 1 is an indicator function that checks whether objects i and j are from the same class k, which is then multiplied with the intersection-over-union between these two regions. This way, the metric measures both the spatial and semantic localization of the query object.

Metrics

mAP. Based on the relevance score, we use mean Average Precision to measure the retrieval performance. We first use a heuristic relevance threshold ≥ 0.30 as recommended in [93], to convert continuous relevance values to discrete labels. Then, we measure the mAP values @$\{1, 10, 50\}$.

mAP metric does not respect the continuous nature of the compositional visual search since it binarizes continuous relevance values with a heuristic threshold. To that end, we resort to two additional metrics, continuous adaptation of NDCG and mean Relevance values which are used to evaluate continuous-valued metric learning techniques in [75, 79, 100].

cNDCG. We make use of the continuous adaptation of the Normalized Discounted Cumulative Gain as follows:

$$cNDCG(q) = \frac{1}{Z_k} \sum_{i=1}^{K} \frac{2^{r_i}}{\log_2(i+1)}, \tag{4.10}$$

that takes into account both the rank and the scores of the retrieved images and the ground truth relevance scores. In our experiments we report cNDCG@$\{1, 50, 100\}$.

mREL. mREL measures the mean of the relevance scores of the retrieved images per query, which is then averaged over all queries. In our experiments, we report mREL@$\{1, 5, 20\}$. We also note the **oracle** performance where we assume access to the ground truth mIOU values to illustrate the upper bound in the performance.

4.4.3 *Performance Comparison*

ResNet-50 **[55].** We use the activations from layer-4 of ResNet-50 to retrieve images. In this work, we build upon this feature since it captures the object semantics and positions within the feature map of size $\mathbb{R}^{7\times7\times2048}$. We also experimented with the earlier layers, however we found that layer-4 performs the best. The network is pre-trained on ImageNet [29].

Textual. We assume access to the ground truth object labels for a query and retrieve images that contain the same set of objects. This acts as a textual query baseline and is blind to the spatial information.

VisSynt [93]. This baseline uses a triplet loss formulation coupled with a classification loss to perform a compositional visual search. We use the same backbone architecture $g(\cdot)$ and the same target feature ResNet-50 to train this baseline for a fair comparison.

LogRatio [75]. This method is the state-of-the-art technique in continuous metric learning, originally evaluated on human pose and image caption retrieval. In this work, we bring this technique as a strong baseline since the visual composition space also exhibits continuous relationships. We use the authors code [1] and the recommended setup.

[1] https://github.com/tjddus9597/Beyond-Binary-Supervision-CVPR19

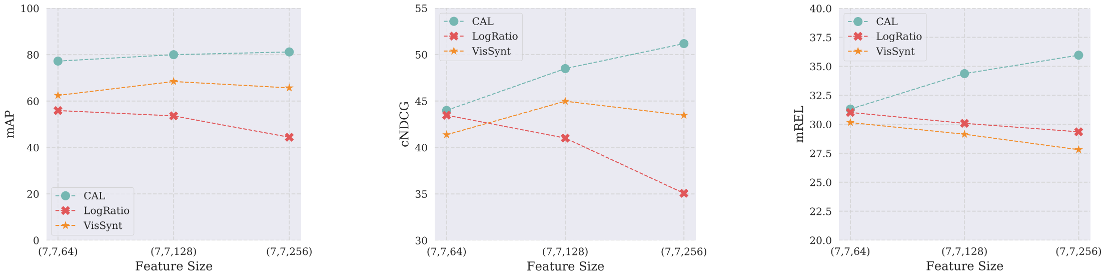


Figure 19: Feature efficiency. Our model performs better even when the feature-space is compact.

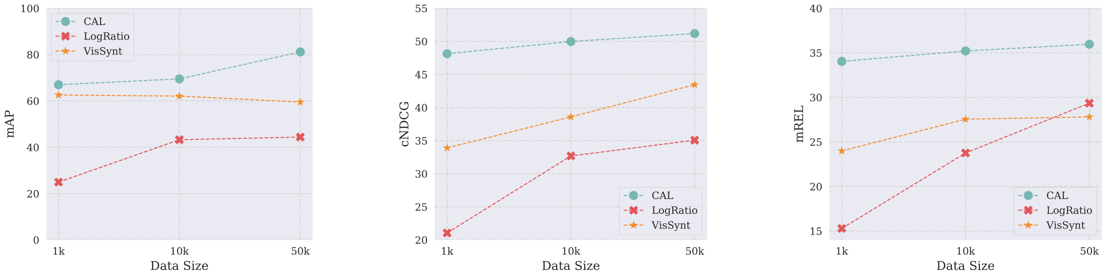


Figure 20: Data efficiency. Our model outperforms VisSynt and LogRatio within small data regime.

We convert mIOU scores to distance values as $1 - mIOU$ since the method minimizes the distances.

Implementation details. We use PyTorch [106] to implement our method. We use the same backbone ($g(\cdot)$) and the input feature (ResNet-50) for all the baselines. All the models are trained for 20 epochs using SGD with momentum ($= 0.9$). We use an initial learning rate of 10^{-2} which is decayed exponentially with 0.004 at every epoch. We use weight decay ($wd = 0.005$) for regularization. In practice, we compute input-output transformations between all examples within the batch to get the best out of each batch. We set the batch size to 36, and given each query in the batch, we sample 1 highly relevant and 1 less relevant examples for each query, which leads to an effective batch size of $36 \times 3 = 108$.

4.5 EVALUATION

In this Section, we present our experiments. For Experiments $1-2$, we use all three metrics $@k = 1$. For the third experiment of the State-of-the-Art comparison, we provide performance at different k values.

4.5.1 *Ablation of Composition-aware Learning*

Euclidean vs. Composition-aware loss. In our first ablation study, we compare the Euclidean loss described in Equation 4.4 with our composition-aware loss. The results are presented in Table 10.

	mAP	cNDCG	mREL
Euclidean	66.87	39.73	28.49
CAL (ours)	81.17	51.18	35.96

Table 10: Euclidean vs. Composition-aware loss.

It is observed that Composition-aware loss outperforms Euclidean alternative by a large-margin, confirming the effectiveness of the proposed loss function.

	mAP	cNDCG	mREL
Lingual	65.14	27.77	19.56
Visual (ours)	81.17	51.18	35.96

Table 11: Lingual vs. Visual input transformation.

Lingual vs. Visual transformation. In our second ablation study, we test the domain of the input transformation (Eq 4.2). In our work, we proposed a visual-based input transformation whereas VisSynt [93] utilizes a lingual-based input transformation using semantic Word2vec embeddings [99]. As can be seen from Table 11, vision-based transformation outperforms the lingual counterpart, since it can better encode the relationships within the visual world.

4.5.2 *Feature and Data Efficiency*

In this experiment, we test the efficiency. Specifically, we first test the feature-space efficiency to see how the performance changes with varying sizes of the query embedding. Second, we test the data-space efficiency by sub-sampling the training data.

Feature-space efficiency. We change the feature embedding size by varying the number of channels as $64, 128, 256$ by keeping the spatial dimension of 7×7. We compare our approach to VisSynt [93] and LogRatio [75]. The results can be seen from Figure 19.

As can be seen, our approach performs the best for all metrics and across all feature sizes. This indicates that composition-aware learning is effective even when the feature size is compact (*i.e.* $7 \times 7 \times 64$). Another observation is that the performance of *CAL* increases with the increased feature size, whereas the performance of the two other techniques is lower. This indicates that *CAL* can leverage bigger feature sizes while other objectives tend to over-fit.

It is concluded that *CAL* is a feature-efficient approach for compositional visual search.

Data-space Efficiency. In this experiment we vary the number of training data as $1k, 10k, 50k$. The results can be seen from Figure 20.

Our method performs the best regardless of the training size. The gap in the performance is even more significant when the training set size is highly limited (*i.e.* $1k$ only), confirming the data efficiency of the proposed approach.

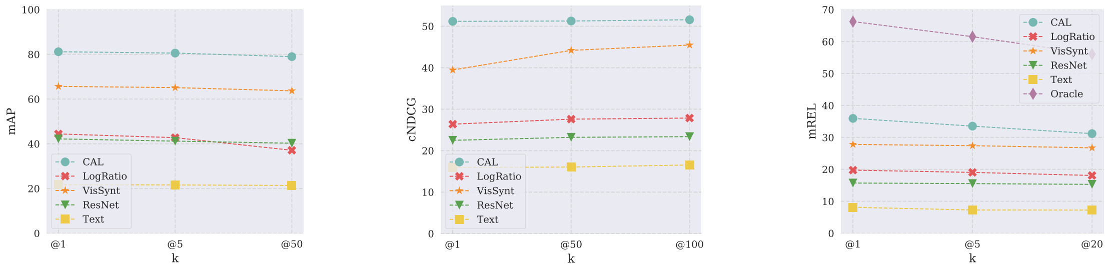


Figure 21: Benchmarking on MS-COCO [89]. Our method outperforms existing techniques for all three metrics.

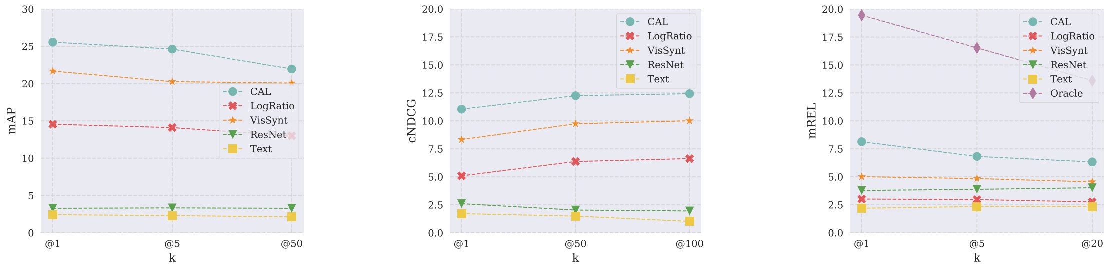


Figure 22: Benchmarking on HICO-DET [17]. Our method transfers better to HICO-DET dataset for object-interaction search.

It is concluded that *CAL* can learn more from fewer examples by leveraging the continuous-valued transformations and the regularized loss function.

4.5.3 *Comparison with the State-of-the-Art*

In the last experiment, we compare our approach to competing techniques on MS-COCO in Figure 21 and HICO-DET in Figure 22 datasets.

As can be seen, our method outperforms the compared baselines in both datasets, and in 3 metrics. This confirms the effectiveness of composition-aware learning for object (MS-COCO) and object-interaction (HICO-DET) search. The results in HICO-DET are much lower compared to MS-COCO since 1) HICO-DET has a higher number of query images (10k vs. 5k), 2) Many queries have only a few relevant images within the gallery set (as can be seen from the oracle performance of only 0.19 mREL in Figure 22), 3) No training is conducted on HICO-DET, revealing the transfer-learning abilities of the evaluated techniques.

Qualitative analysis. Lastly, we showcase a few qualitative examples in Figure 23. First, as a sanity check, we illustrate single object queries (stop signs). As can be seen, our method successfully retrieves images relevant to the query category and the position. Then, we illustrate some object-interaction examples, such as human-on-bench, or human-with-tennis racket, or human-on-skateboard. Our model can still generalize to such examples, meaning that compositional learning benefits the case of the object interaction. We illustrate a failure case in the last row, where our model retrieves a mix of snowboard-skateboard objects given the query of a skateboard. This indicates that our model performance can be improved by incorporating scene context, which we leave as future work.

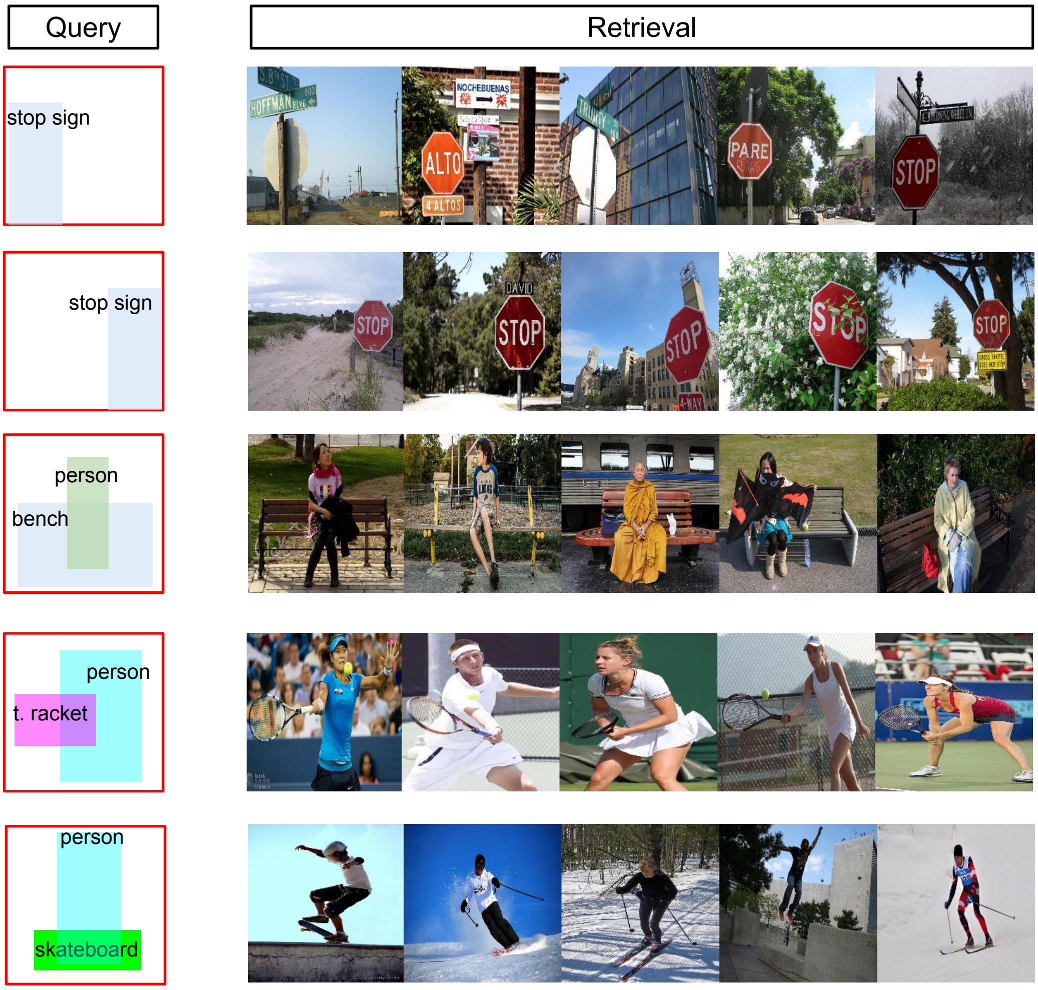


Figure 23: Qualitative examples. First two rows show a single-object query, and last three rows show multi-object queries. As can be seen, our approach considers the object category, location and interaction into account while retrieving examples.

4.6 CONCLUSION

In this work, we tackled a structured visual search problem called compositional visual search. Our approach is based on the observation that the visual compositions are continuous-valued transformations of each other, carrying rich information. Such transformations mainly consists of the positional and categorical changes within the queries. To leverage this information, we proposed composition-aware learning, which consists of the representation of the input-output transformations as well as a new loss function to learn these transformations. Our experiments reveal that defining the transformations within the visual domain is more useful than the lingual counterpart. Also, a regularized loss function is necessary to learn such transformations. Leveraging transformations with this loss function leads to an increase in the feature and data efficiency, and outperforms existing techniques on MS-COCO and HICO-DET. We hope that our work will inspire further research to incorporate structure for the structured visual search problems.

5

HUMAN-OBJECT INTERACTION DETECTION WITHOUT ALIGNMENT SUPERVISION

5.1 INTRODUCTION

This paper strives for Human-object Interaction (HO-I) detection from an image. HO-I detection receives an astounding attention from the community recently [17, 21, 39, 40, 44, 53, 58, 69, 72, 74, 88, 90, 129], thanks to the large-scale benchmark of HICO-DET [17]. The goal is to identify the tuples of `<human, object, verb, noun>` from the input, where human-object is an interacting bounding box pair, and verb-noun is the interaction type, such as ride-horse.

To tackle this problem, researchers leverage strong HO-I alignment supervision, see Figure 24-(a). Annotators first draw a bounding box around all humans and objects, then align humans with the object-of-interaction (*e.g.*, rider and horse). Finally, they align the interaction category with each human-object pairs.

However, collecting such annotation is costly [1]. Annotation costs time, since in a typical image there are tens of potential human-object interactors, if not hundreds. One can instead rely on image-level HO-I annotations, see Figure 24-(b). Image-level annotations enumerate existing HO-I within the image, without specifying where they occur. Image-level annotations are much faster and cheaper to collect.

There are few attempts to perform HO-I detection via image-level supervision weaklyhoi1,weaklyhoi2. Initially, Zhang *et al.* weaklyhoi1 proposes a two-stream architecture based on Region-FCN rfcn, focusing on the regional appearance of subject-objects and spatial relations. Later, Kumaraswamy *et al.* weaklyhoi2 adapted this technique for HO-I detection, and improve it via an additional stream of human pose. These techniques yield remarkable results on HICO-DET benchmark hicodet in the absence of alignment supervision. However, they are limited in three major ways: *i)* These methods isolate human-objects from their context via Region-of-Interest (RoI) pooling fastrcnn,faster-rcnn, however, contextual information is crucial in understanding the interaction, *ii)* The authors propose multiple streams of context to circumvent the missing contextual information, which increases model complexity. Increased model complexity results in low performance on especially rarely represented HO-I (*i.e.* `<ride, cow>`) as we will show. *iii)* Hand-crafted context (*i.e.* body-pose configuration using key-points) may not be sufficient to account for the complexity of HO-I detection problem.

To that end, in this paper, we propose Align-Former, a visual-transformer-based architecture based on detr. Align-Former is a single-stream HO-I detector that is trained

[1] Try-it-yourself! HICO-DET-Annotator

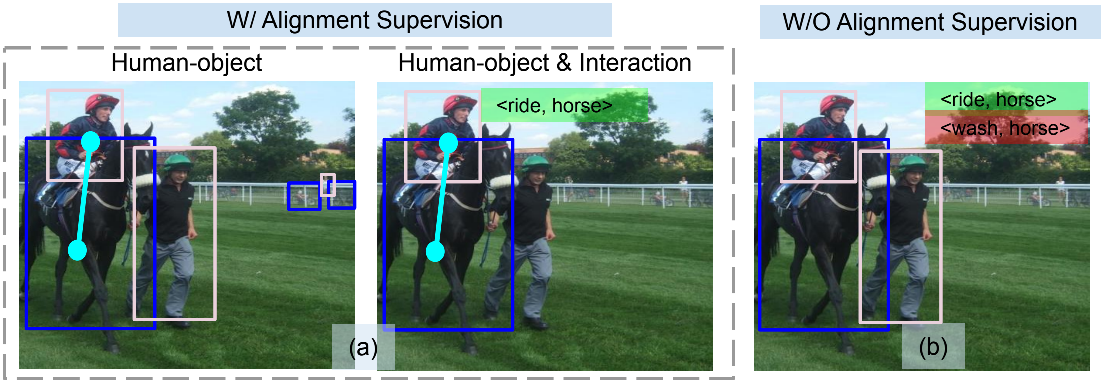


Figure 24: Alignment (left) vs. Image-level HO-I supervision (right). a) Alignment supervision annotates each human-objects, aligns humans to their interacting objects, then aligns human-objects to their type of interaction. b) Image-level supervision only lists existing interactions without pointing where they happen. Our goal is to detect HO-I without costly alignment supervision, by only using image-level labels.

end-to-end using image-level supervision only. Align-Former is equipped with a novel HO-I Align layer that learns to align a few candidate target HO-I with predictions, allowing detector supervision. The decision of alignment is based on geometric and visual priors that are crucial in HO-I detection.

This paper makes the following contributions:

I. We propose Align-Former, an end-to-end HO-I detector that is supervised via image-level annotation.

II. We equip Align-Former with a novel HO-I align layer, that learns to match few HO-I predictions with HO-I target(s), therefore allowing detector supervision.

III. We evaluate Align-Former on HICO-DET [17] and V-COCO [52], and show that Align-Former outperforms competing baselines with the same level of supervision (by **4.71** mAP) on the large-scale benchmark of HICO-DET [17], especially within the low-data regime of rare categories (by **6.17** mAP).

5.2 RELATED WORK

Alignment-Supervised HO-I Detection. In HO-I detection, the goal is to find quadruplets of `<human,object,verb, noun>` where human-object are bounding boxes and verb-noun are interaction pairs like `<ride, horse>`. Initially, HICO-DET authors collect more than 150k instance annotations to match humans to their interacted object, as well as to their interaction categories. Then, there has been a surge in detecting HO-I, initially via two-stage techniques [17, 40, 44, 53, 58, 90], and later by one-stage architectures [21, 39, 74, 88, 129] leveraging costly strong alignment supervision, see Figure 24-(a).

In this work, our goal is to train HO-I detectors without alignment supervision, by only relying on image-level HO-I annotations.

HO-I Detection via Image-level Supervision. Few works attempt to train HO-I detectors by only image-level supervision [78, 150]. Initially, Zhang *et al.* [150] proposes a

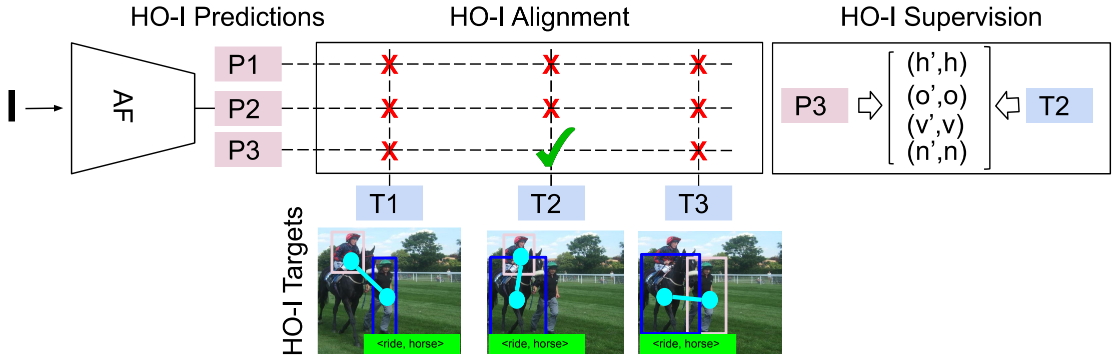


*Figure 25: To perform HO-I detection via image-level supervision: i) Align-Former maps the input image I to HO-I predictions P . ii) We also prepare a set of HO-I targets by exhaustively matching human-object detections and list of interactions. iii) Finally, we find the least costly prediction-target pair(s) (*i.e. (T_2, P_3)*) which will be used for detector supervision.*

two-stream architecture based on Region-FCN [27] to model the subject-object region appearance and spatial relations. Later, Kumaraswamy *et al.* [78] extends this approach via additional pose-stream. These methods operate on the isolated appearance of human-objects, neglecting the crucial context. Consequently, they supplement Region-FCN with additional streams, increasing the model size, decreasing the performance.

To circumvent this, in this work, we propose a single-stream HO-I detector based on visual-transformer [15]. Our network naturally encodes the surrounding context of human-objects thanks to self-attention [135] and learns to align few candidate HO-I targets with HO-I predictions to perform detector supervision, see Figure 25.

Discrete Variable Sampling in Computer Vision. In this work, we treat HO-I target alignment as a hard-valued, discrete variable sampling: Amongst all possible target-prediction pair(s), which subset(s) should be selected for detector supervision? Such decision is non-differentiable therefore ill-suited in convolutional network training. To that end, we resort to a continuous relaxation procedure named Gumbel-Softmax trick, which allows end-to-end training via discrete variables [61, 92]. Gumbel-Softmax has successfully been used to sample convolutional layers [136], filters [22] or channels [10].

In this work, we adapt Gumbel-Softmax to select the target HO-I for detector supervision.

5.3 METHOD

Method Overview. An overview of our technique is presented in Figure 25-26. The goal of our network $g_\theta(\cdot)$ is to produce HO-I prediction tuples given an image I as $I \xrightarrow{g_\theta(\cdot)} t'$. Here, HO-I prediction is of size P and represented via $t' = (h', o', v', n')$, where $(h' \in \mathbb{R}^{P\times 4}, o' \in \mathbb{R}^{P\times 4})$ are human-object bounding box predictions, and $(v' \in \mathbb{R}^{P\times V}, n' \in \mathbb{R}^{P\times N})$ are verb-noun class predictions for V verbs and N nouns.

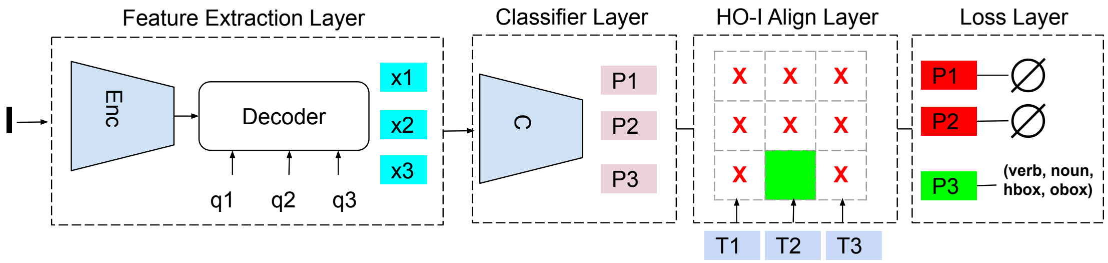


Figure 26: Align-Former consists of four main layers. ***Feature Extraction Layer*** *is an Encoder-Decoder-based visual-transformer that extracts a set of human-object features x_i using the positional queries q_i. Then,* ***Classifier Layer*** *generates HO-I predictions P in the form of human-object bounding boxes and verb-noun classes.* ***HO-I Align Layer*** *compares HO-I predictions P with potential HO-I targets T to find few-matching pair(s) that are used for HO-I detector supervision using* ***Loss Layer****.*

Then, assume we have access to a set of HO-I targets of size T with the same structure $t = (h \in \mathbb{R}^{T\times4}, o \in \mathbb{R}^{T\times4}, v \in \mathbb{R}^{T\times V}, n \in \mathbb{R}^{T\times N})$. To supervise Align-Former, we propose to minimize the following objective:

$$\min_{\theta}(A \times t, t') \tag{5.1}$$

where we omit θ from now on for clarity. A is a binary matrix of size $P \times T$ where only few entries are non-zero. A is applied separately on all tuple members, as $A \times t = (A \times h, A \times o, A \times v, A \times n)$. Here, $A(i, j) = 1$ means prediction i matches (*i.e.* aligns) with target j to use in supervision. Similarly, $A(i, j) = 0$ indicates target i should not be used in detector supervision. To identify which target-prediction pairs should be used in detector supervision, we rely on geometric and visual priors detailed later.

Finally, replacing t' with $g(I) = C(Dec(Enc(CNN(I)), Q))$ yields:

$$\min(A \times t, C(Dec(Enc(CNN(I)), Q))) \tag{5.2}$$

which is detailed in four Sections:

- **HO-I Align Layer (section 5.3.1)** generates the alignment matrix A that pairs few HO-I prediction(s) with HO-I target(s),
- **Classification Layer (section 5.3.2)** generates human-object bounding boxes and verb-noun classification via $C(x)$ using human-object features x,
- **Feature Extraction Layer (section 5.3.3)** generates features via $x = Dec(Enc(CNN(I)), Q)$ via positional queries Q using Encoder-Decoder architecture,
- **HO-I Loss Layer (section 5.3.4)** computes the human-object box and verb-noun classification losses to supervise the detector with the generated HO-I targets t.

5.3.1 *HO-I Align Layer*

HO-I align layer consists of two sub-layers, *i)* Prior layer that judges the compatibility between all HO-I targets and predictions, *ii)* Discretization layer that binarizes the likelihood values to obtain the final hard-alignment.

Discretization Layer

Assume we are given a scoring function $S \in \mathbb{R}^{P \times T}$ where $S(i, j)$ encodes how compatible HO-I prediction t'_i and HO-I target t_j matches. Our goal is to discretize this matrix to obtain the final hard-valued alignment decision.

To perform this, we discretize S such that only few members will be non-zeros. Specifically, given raw values of S, we apply the following operation:

$$A = \sigma(S + G) \geq \delta \tag{5.3}$$

where $\delta = 0.5$ is the hard-threshold value, G is the Gumbel noise [61, 92] added to the matrix S for regularization, and $\sigma(\cdot)$ is the sigmoid activation to bound S between $[0, 1]$. Note that Gumbel-noise is crucial to avoid any degenerate solutions like all 1s.

This operation yields the binary alignment matrix $A \in \{0, 1\}$ where only a few entries are non-zero.

Prior Layer

To compute the compatibility between HO-I targets & predictions, we resort to a convex combination of geometric and visual priors as $S = \alpha_g * GP + \alpha_v * VP$. Our intuition is that for an HO-I target to be a good candidate for detector supervision, it needs to be compatible both in terms of human-object bounding boxes (geometric) and verb-noun classes (visual).

Geometric Prior $GP(\cdot)$ computes the bounding box compatibility of human-objects via L_1 distance as:

$$GP = \exp(-\frac{\sum_{ij} ||h'_i - h_j|| + ||o'_i - o_j||}{\tau}) \tag{5.4}$$

where the exponential function $\exp(\cdot)$ converts the distance values to similarity where $\tau = 1$.

Visual Prior $VP(\cdot)$ computes how well a given target-prediction pair matches in terms of HO-I classes. Remember that our HO-I targets enumerate existing HO-I from the image in terms of verb-noun pairs. Therefore, $VP(\cdot)$ is calculated as:

$$VP = v' * v^T + n' * n^T \tag{5.5}$$

where verb-predictions are of size $v' \in \mathbb{R}^{P \times V}$ and verb-targets are of size $v \in \mathbb{R}^{T \times V}$ for V distinct verbs. Similarly, noun-predictions are of size $n' \in \mathbb{R}^{P \times N}$ and noun-targets $n' \in \mathbb{R}^{T \times N}$ for N distinct nouns.

5.3.2 *HO-I Classification Layer*

Classifier layer is responsible for generating HO-I predictions t' consisting of human-object bounding box predictions (h', o') as well as verb-noun category predictions (v', n').
Human-Object Bounding Box Classifiers are two multi-layer perceptrons $g^h(\cdot)$ and $g^o(\cdot)$ that maps human-object features x to coordinates as $(h', o') = (\sigma(g^h(x)), \sigma(g^o(x)))$.
Verb-Noun Classifiers are also two multi-layer perceptrons as $g^v(\cdot)$ and $g^n(\cdot)$ that learns to map human-object features x to corresponding verb-nouns as $(v', n') = (\sigma(g^v(x)), (g^n(x)))$.

5.3.3 *HO-I Feature Extraction Layer*

Our backbone needs to encode: *i)* Object-object relations, *ii)* Relative object positions that are critical to perform HO-I alignment and detection. To that end, we implement the feature extractor as a visual-transformer based on DETR [15]. The feature extractor yields human-object features $x \in \mathbb{R}^{P \times D}$, and consists of three sub-layers: Backbone, Encoder and Decoder, which are detailed below.
Backbone ($x = CNN(I)$). Backbone is a deep CNN [55] that extracts global feature maps from the input image I of size $x \in \mathbb{R}^{H \times W \times C}$ where $[H, W]$ are the height-width of the feature map, and C is the number of channels.
Encoder ($x = Enc(x)$). Encoder further processes the global feature map from the backbone to increase positional and contextual information. We first reduce the number of channels from the backbone to a much smaller size via 1×1 convolutions of $C \times D$. Then, the resulting feature map $\mathbb{R}^{H \times W \times D}$ is collapsed in the spatial dimension as $\mathbb{R}^{D \times HW}$ where each pixel becomes a "token" represented by D dimensional features. Finally, this feature undergoes a few self-attention operations via few multi-layer perceptrons, residual operations, and dropout. At each step, pixel positions are added to the feature map to retain position information.
Decoder ($x = Dec(x, Q)$). The Decoder is a combination of self-attention and cross-attention layers, which yields the final human-object features. The Decoder takes as input the Encoder output $x \in \mathbb{R}^{D \times HW}$ as well as fixed positional query embeddings $Q \in \mathbb{R}^{P \times D}$. Decoder alternates between the cross-attention between the feature map x and Q, as well as self-attention across queries. Cross-attention extracts features from the global feature maps, whereas self-attention represents object-object relations necessary for HO-I detection. Decoder is implemented as multi-layer perceptrons. Final output is $x \in \mathbb{R}^{P \times D}$ that encodes positional and appearance-based representations of potential human-object pairs within the image.

5.3.4 *HO-I Loss Layer*

Our loss function ensures that the predicted human-object bounding boxes as well as the verb-noun predictions are in line with the aligned HO-I targets.

The loss function $\mathcal{L}$ is a composite of bounding box, classification, and sparsity losses as $\mathcal{L} = \mathcal{L}_{box} + \mathcal{L}_{class} + \mathcal{L}_{sparse}$. Here, $\mathcal{L}_{box}$ computes the L_1 distances between human-object predictions and (aligned) targets as $\mathcal{L}_{box} = \mathcal{L}_{human} + \mathcal{L}_{object}$. And, $\mathcal{L}_{class} = \mathcal{L}_{verb} + \mathcal{L}_{noun}$ are implemented via classical cross-entropy. As there can be multiple verbs for each instance, we use sigmoid activation before computing the verb loss.

Sparsity Loss. Finally, sparsity loss minimizes $\mathcal{L}_{sparse} = \frac{1}{P \times T} \sum_{ij} A_{ij}$ where $\frac{1}{P \times T}$ is a constant normalizing factor to bound the loss. This ensures the sum over all entries within the alignment matrix A is minimized, leading to only few pairs of HO-I predictions and targets to be aligned for further supervision.

Implementation. We set the number of predictions as $|P| = 100$. Our network is implemented using PyTorch [106]. Feature size D from the last layer of the Decoder is set to $D = 256$. Both human-object bounding box classifiers and verb and noun predictors are 2-layer perceptrons with ReLU activation in between.Initial learning rate is set to 10^{-6} for the ResNet backbone and 10^{-5} for the rest of the parameters. We use weight-decay to regularize the network with 10^{-4}. We train the network for 150 epochs with an effective batch size of 16 over 8 GPU Titan cards. We decay the learning rate linearly with $10^{-}1$ after epoch 100.

5.4 EXPERIMENTS

Datasets. We experiment on two large-scale standard datasets, namely HICO-DET [17] and V-COCO [52]. *i) HICO-DET* contains 38k images for training and 9.6k images for testing. Images contain 117 distinct verbs and 80 distinct nouns together, making 600 `<verb, noun>` pairs. For each noun, there exists a case of ”no-interaction”, where at least a single human and the target object is visible, even though not interacting. We only use HO-I alignment annotations for testing, and not training, since our goal is to evaluate HO-I detection via image-level supervision. *ii) V-COCO* builds upon MS-COCO [89] where the authors annotate subset of images with human-object alignments and their (inter-)action. The type of interactions is riding, reading and smiling. The dataset exhibits 2.5k images for training, 2.8k images for validation, and 4.9k images for testing.

Metric. We use the mean Average Precision (mAP) metric for evaluation as is the standard [17,52]. A human-object interaction is true positive only if both humans and objects have an Intersection-over-Union with a ground-truth HO-I pair above > 0.50 *and* they are assigned to the correct interaction categories.

Evaluation. *i) HICO-DET:* We use the evaluation code presented in the server [3]. We compute the mean over all three splits of full, rare, and non-rare in HICO-DET. We provide comparison on three standard splits. *Full*: All 600 categories, *Rare*: 138 categories with less than or equal to 10 training instances, *Non-Rare*: 462 categories with more than 10 training instances. *ii) V-COCO:* We use the evaluation code presented in authors' code [1]. We evaluate using three different standard scenarios. *Agent*: We

report the human interactor detection performance, *Scenario-1*: We report the detection of humans and objects together, *Scenario-2*: We report the detection of humans and objects where the object predictions for object-less interactions (*i.e.* smiling) is ignored.
Baselines. We compare Align-Former to *i) Weakly-supervised HO-I detectors*: PPR-FCN [150] and MX-HOI [78] that performs HO-I detection without alignment supervision. *ii) Strongly-supervised variants*: To measure the upper bound performance as a reference, we also report MX-HOI and Align-Former performance via strong alignment supervision.

5.5 EVALUATION

5.5.1 *Comparison to The State-of-The-Art*

Method	Backbone	Alignment-Supervised?	Full	Rare	Non-Rare
PPR-FCN [150]	ResNet-101	✗	15.14	10.65	16.48
MX-HOI [78]	ResNet-101	✗	16.14	12.06	17.50
Align-Former (ours)	ResNet-50	✗	19.26	14.00	20.83
Align-Former (ours)	ResNet-101	✗	**20.85**	**18.23**	**21.64**
MX-HOI [78]	ResNet-101	✓	17.82	12.91	19.17
Align-Former (ours)	ResNet-50	✓	25.10	17.34	27.42
Align-Former (ours)	ResNet-101	✓	27.22	20.15	29.57

Table 12: Human-Object Interaction Detection mAP on HICO-DET [17]. Our method outperforms existing techniques over all splits of full, rare, and non-rare.

HICO-DET Results are presented at Table 12. Overall, Align-Former outperforms the other two techniques by 3.12 mAP via ResNet-50 and 4.71 mAP via ResNet-101 on all categories. This confirms that HO-I detection benefits from the end-to-end alignment of the targets and the predictions. Our improvement is even more pronounced on the rare split via 6.17 mAP using ResNet-101, exhibiting the sample efficiency of our technique.

Method	Backbone	HICO-DET Pre-Trained?	Alignment-Supervised?	Agent	Scenario 1	Scenario 2
Align-Former	ResNet-50	✗	✗	24.63	13.90	14.15
Align-Former	ResNet-50	✓	✗	27.95	15.52	16.06
Align-Former	ResNet-101	✗	✗	20.00	10.44	10.79
Align-Former	ResNet-101	✓	✗	**30.02**	**15.82**	**16.34**
Align-Former	ResNet-50	✗	✓	66.78	50.20	56.42
Align-Former	ResNet-101	✗	✓	68.00	55.40	62.15

Table 13: Human-Object Interaction Detection mAP on V-COCO [52]. Even though the performance is limited when trained from scratch on V-COCO, HICO-DET pre-training yields a considerable improvement on V-COCO.

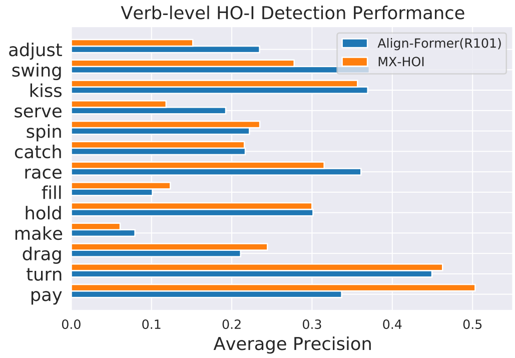


Figure 27: Verb-level Performance on HICO-DET [17]

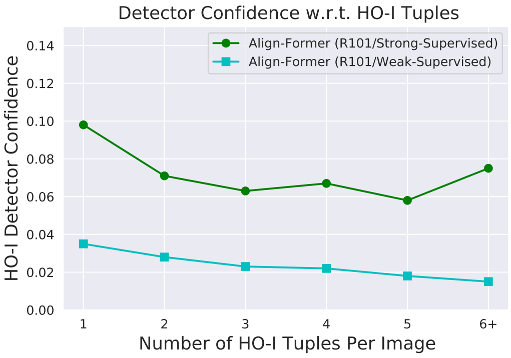


Figure 28: HO-I detector confidence w.r.t. number of HO-I tuples in an image on HICO-DET [17].

V-COCO Results are presented at Table 13. We only compare to our own baselines [2]. We evaluate two different settings. *i) Training on V-COCO from scratch*: Since the number of training images are quite limited (only $2k$ examples), training on V-COCO without alignment supervision yields limited accuracy on all three settings. *ii) Transfer learning from HICO-Det*: where we fine-tune a HICO-DET pre-trained model on V-COCO. In all cases, pre-training on HICO-DET helps significantly. As one of the major goal of annotation-free training is the ability to pre-train on large-scale benchmarks, we see this as a promising direction in HO-I detection with cheap image-level supervision.

We confirm that our model yields competitive performance on HICO-DET against competing benchmarks on all full, rare and non-rare splits, and showcases promising first results without alignment supervision on V-COCO, especially via transfer learning.

5.5.2 *Further Analysis*

In this section, we provide analysis to better understand the contribution of Align-Former.
Verb-level Performance Comparison. We visualize verb-level performance difference between weakly supervised Align-Former and MX-HOI in Figure 27. We observe that Align-Former outperforms for pose and part-driven interactions like adjust, swing or kiss, while underperforming for scene-driven interactions like pay or turn. This indicates end-to-end learning of pose-based representations is more valuable than hand-crafted pose representations as in MX-HOI. For more results, refer to our Supp. material.
W/ *vs.* W/O Alignment Supervision. To better understand the gap between strongly *vs.* weakly supervised HO-I detection, we provide results of MX-HOI with strong supervision on HICO-DET in Table 12 as well as strongly supervised Align-Former in both datasets (Table 12- 13). Our method is flexible as it can be easily trained with strong and weak supervision with no change in architecture, whereas MX-HOI ensembles two CNNs (a weak [150] and strong [53] CNN) to do so.

We have three main findings. *i)* Weakly-supervised Align-Former outperforms strongly supervised MX-HOI on HICO-DET (Table 12), which indicates our method compensates

[2] Neither of the existing baselines (PPR-FCN and MX-HOI) evaluates on V-COCO. Additionally, strongly supervised stream of MX-HOI (No-Frills HO-I [53].) also is not evaluated on V-COCO

for the lack of supervision with its representational power. *ii)* Strongly supervised Align-Former outperforms weakly supervised Align-Former on both datasets (Table 12- 13). This shows Align-Former better leverages the supervision when is used, and there is a room for improvement in weakly-supervised techniques. *iii)* In Figure 28, we plot the confidence of strongly *vs.* weakly supervised Align-Former as a function of number of HO-I tuples in an image on HICO-DET. As can be seen, strongly-supervised variant retains its performance whereas weakly-supervised degrades in confidence, which may help explain the performance gap between the two variants of Align-Former.

ResNet-101 vs. ResNet-50. We implement Align-Former with ResNet-50 and 101. Even though we do not observe significant difference at the verb- or object- level, the difference is at the interaction-level. Our findings are: *i)* ResNet-101 outperforms $ResNet-50$ on both datasets across all settings, *ii)* Surprisingly, ResNet-101 outperforms especially on the rare split of HICO-DET, and exhibits better transferability to V-COCO, despite higher number of parameters.

Qualitative Inspection. *i) Attention Analysis*: To understand where Align-Former is looking at to perform HO-I alignment and detection, we present the attention matrix for a set of queries from the last layer of the Decoder in Figure 29-(a). We observe that Align-Former attends on body-parts when the visual information is sufficient, and full-body when the human-object has small scale. *ii) Qualitative Results*: Finally, we visualize high-confident detection examples in Figure 29-(b). We observe that Align-Former can detect both dynamic interactions like `<kick, sports ball>` or static interactions like `<eat, sandwich>`. However, our method fails when humans can not be paired with their object of interaction, as is visualized in the bottom row.

5.6 CONCLUSION

This paper addressed HO-I detection from images. We proposed Align-Former, a visual-transformer based CNN that can learn to detect HO-I without alignment supervision, via image-level supervision. We equip Align-Former with HO-I align, a novel layer that learns to select correct detection targets based on geometric and visual priors. We show that Align-Former outperforms existing techniques for HO-I detection on HICO-DET especially on rare HO-I, and yields promising results on V-COCO, confirming the efficacy of our method. We hope our work inspires future research on reducing supervision in HO-I detection.

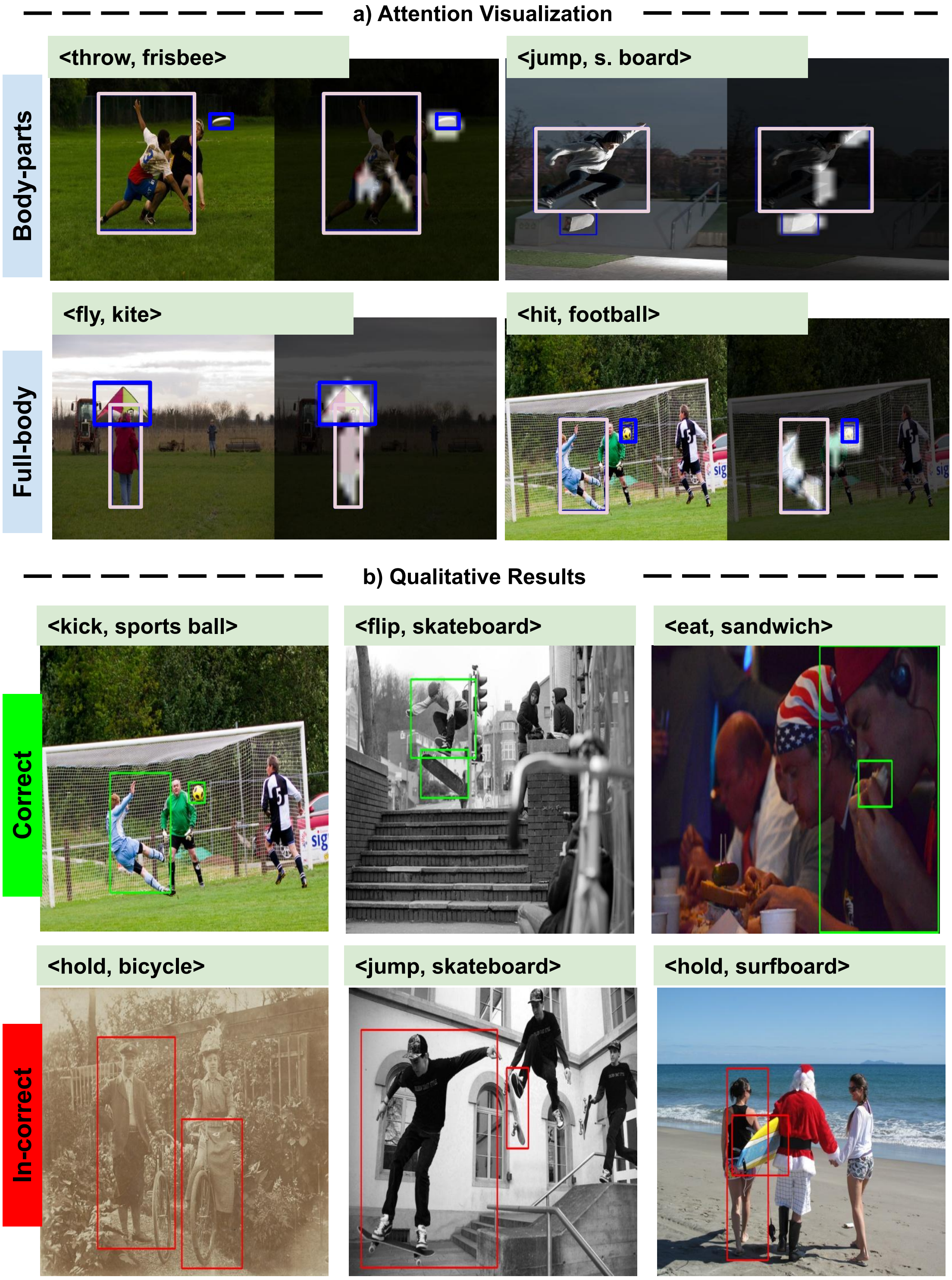


Figure 29: a) Attention analysis of Align-Former reveals the focus on body-part and full-body. b) Qualitative analysis of Align-Former reveals it can detect both dynamic and static interactions.

6

SUMMARY AND DISCUSSION

6.1 SUMMARY

This thesis proposes a contextual understanding of human-object interactions. We alleviate the challenge of human-object interaction understanding by incorporating multiple sources of context into deep learners. We first investigate the sources of contextual information in interactions, by studying the visual extent of human-object interactions in Chapter 2. We observe that the locality and compositionality of the human-object interactions play a significant role in interaction understanding. Then, we develop multiple sources of local contextual information in Chapter 3, which we show to help drastically in interaction recognition. Then, we incorporate the compositional context between the human interactor and the object interactee to perform interaction search within large image databases in Chapter 4. With the observation that given an image, only few human-object pairs are in an interaction, we incorporate sparsity context to perform interaction detection in Chapter 5.

Chapter 2: Where is the Interaction Context? An Empirical Study

Chapter 2 studies the visual extent of human-object interactions. Visual human-object interactions are hard to pinpoint in an image. Where objects and subjects have clear boundaries, their interaction does not. In this work, we try to pinpoint the human-object interactions from a single image by studying their visual extent. Where is the visual evidence for the interactions in an image? We start from observable regions like the subject and the object to determine which region is effective in learning to recognize interactions. Then, we devise an oracle strategy to determine the region that yields the best recognition performance. This provides an upper bound for interaction recognition in our setting. Finally, we explore the importance of visual details within this limited region. Our findings show that: *i)* interactions can benefit from even simple inclusion of the context into the recognition, *ii)* finding the best context per image helps even greater and, *iii)* small details around the intersection of subject-object is important in recognition.

Chapter 3: Self-Selective Context for Interaction Recognition

This chapter studies the local context of human-object interactions. Human-object interaction recognition aims for identifying the relationship between a human subject and an object. Researchers incorporate global scene context into the early layers of deep

Convolutional Neural Networks as a solution. They report a significant increase in the performance since generally interactions are correlated with the scene (*i.e.* riding bicycle on the city street). However, this approach leads to the following problems. It increases the network size in the early layers, therefore not efficient. It leads to noisy filter responses when the scene is irrelevant, therefore not accurate. It only leverages scene context whereas human-object interactions offer a multitude of contexts, therefore incomplete. To circumvent these issues, in this work, we propose Self-Selective Context (SSC). SSC operates on the joint appearance of human-objects and context to bring the most discriminative context(s) into play for recognition. We devise novel contextual features that model the locality of human-object interactions and show that SSC can seamlessly integrate with the State-of-the-art interaction recognition models. Our experiments show that SSC leads to an important increase in interaction recognition performance, while using much fewer parameters.

Chapter 4: Structured Visual Search via Composition-Aware Learning

This chapter studies visual search using structured queries. The structure is in the form of a 2D composition that encodes the position and the category of the objects. The transformation of the position and the category of the objects leads to a continuous-valued relationship between visual compositions, which carries highly beneficial information, although not leveraged by previous techniques. To that end, in this work, our goal is to leverage these continuous relationships by using the notion of symmetry in equivariance. Our model output is trained to change symmetrically with respect to the input transformations, leading to a sensitive feature space. Doing so leads to a highly efficient search technique, as our approach learns from fewer data using a smaller feature space. Experiments on two large-scale benchmarks of MS-COCO [89] and HICO-DET [17] demonstrates that our approach leads to a considerable gain in the performance against competing techniques.

Chapter 5: Human-Object Interaction Detection without Alignment Supervision

The goal of this chapter is Human-object Interaction (HO-I) detection. HO-I detection aims to find interacting human-objects regions and classify their interaction from an image. Researchers obtain significant improvement in recent years by relying on strong HO-I alignment supervision from [17]. HO-I alignment supervision pairs humans with their interacted objects, and then aligns human-object pair(s) with their interaction categories. Since collecting such annotation is expensive, in this paper, we propose to detect HO-I without alignment supervision. We instead rely on image-level supervision that only enumerates existing interactions within the image without pointing where they happen. Our paper makes three contributions: *i)* We propose Align-Former, a visual-transformer based CNN that can detect HO-I with only image-level supervision. *ii)* Align-Former is equipped with HO-I align layer, that can learn to select appropriate targets to allow detector supervision. *iii)* We evaluate Align-Former on HICO-DET [17] and V-COCO [52], and show that Align-Former outperforms existing image-level supervised HO-I detectors by a large margin (**4.71**% mAP improvement from 16.14% → 20.85% on HICO-DET [17]).

6.2 DISCUSSION

In this thesis, we asked ourselves: *Can we understand the role of context in single image human-object interactions?* We identified that the context determines the appearance of an interaction, as illustrated by Figure 1, as without context there is no interaction. To that end, firstly, we establish that the role of context is to determine *the existence of an interaction*. Identifying if there is an interaction is a critical first step, as there could be multiple humans and objects within an image although not interacting. Secondly, the role of context is to identify *the where of an interaction*. Given an input image where we establish that exhibits an interaction, not all regions are equally viable. Also, amongst many human and object candidates, only few are in an interaction. Lastly, the role of context is to identify *the what of an interaction*. Within the image, after locating the interaction regions, the final step is to determine the type of the interaction. With the help of the context, the computer can distinguish amongst a multitude of interactions within the visual world. Here, one critical set of category is rare interactions, with only few exemplars within the training data. In this thesis, we establish that the context provides a stable signal to be able to distinguish rare interactions from the others, for recognizing, searching and detecting visual interactions.

6.2.1 *Future Work*

While we tackled visual interaction understanding from different angles of recognition, search, detection, and generalization, in this part, we discuss the potential ways to further improve our understanding of visual interactions.

We humans interact with the visual world to understand the properties of objects, such as material properties or affordances [41]. With interactions, we are able to manipulate novel objects to meet our daily human needs. In this work, we consider an important part of that scenario: a passive, static dataset of well curated single images that have been observed multiple times during training. While useful from Vision-for-Web point of view, we realize such approach has limitations for Vision-for-Action in real life.

In real life, humans are presented with sequential, multiple views of an object, collected through movements and tactile sensors. To that end, we see active vision [6, 104] for interaction as a plausible new direction of exploration. We believe that to fully leverage the signals present in visual interactions, the computers need to actively acquire visual data via exploration [23]. As potential interaction is anywhere around us, and when there are N significant objects in the scene, this active interaction vision is of N**2 complexity, the new task is quite formidable from the start. We believe, all aspects of human-object interaction in this thesis may contribute to active interaction vision. To make further progress, we identify three possible directions for future research.

Generalization from Single to Multiple Modalities. This thesis has focused on the single modality of visual images. However, interactions can be recorded via multiple modalities. One modality is tactile, which can be used to represent grasping interactions [127]. Another modality is lingual, which can be represented either via textual information surrounding the input, or as audio [103]. In instances where the visual modality does not suffice, one can learn to rely on or one may seek to combine the information from multiple modalities. To contribute to this, the self-selective context presented in Chapter 2 may be re-purposed to determine the relative importance of modalities.

Generalization from Images to Video. This thesis has focused on single images to understand interactions. To benefit from active exploration and to learn from interactions, the agents need to be able to process temporal information within a video. A trivial technique to adapt the presented models for videos would be to run them on each video frame. This can be implemented in a straight-forward and robust manner, as our models make no assumption in the processing of single images. While having the advantage of being simple, this ignores the crucial temporal information which is helpful in distinguishing between interactions. Adapting models from image to video is challenging due to domain discrepancies, such as image quality, or translation bias [130] including the habitual difference of framing between photographs and single video frames. And finally, for temporal signals like swinging, opening, closing, the additional task of finding the beginning and the end of the interaction looms. Such interactions may lead to confusion when treated as a set of separate frames. To alleviate this, one simple solution is to replace 2D convolutions in our models with 3D counterparts [132].

Generalization from Within-Context to Out-of-Context. This thesis has mostly focused on within-context examples, where we assume the interactions take place in their usual context. Riding a bicycle takes place on the street, whereas eating a donut takes place in a donut shop. Such an assumption is unrealistic in real-life active learning, since humans can perform similar interactions in a variety of places. To circumvent this, for achieving context-free detection and recognition, a potential remedy is to train the models to be invariant against the scene information while optimizing for interaction understanding performance [24]. This way, the models are forced to choose representations that are stable across scenes.

But even then, even when properly detecting and categorizing out-of-context common interactions like shaking hands or eating, the context brings an important interpretation of the action. To shake hands on top of a mountain is in its meaning quite different from shaking hands at the doorstep. To eat in a fast food restaurant has a different meaning than to eat while riding a bicycle. So even when there is an interaction which is quickly classified as free of context, the context plays an important role in determining the interpretation.

To conclude, we believe there is still quite some topics to delve deeper in human-object interaction understanding.

BIBLIOGRAPHY

[1] Vcoco evaluation server. In *Link*, 2015.

[2] Shutterstock compositional visual search. In *https://www.shutterstock.com/blog/composition-aware-search-tool*, 2017.

[3] Hico-det evaluation server. In *Link*, 2018.

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016.

[5] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012.

[6] M. Al Haj, C. Fernández, Z. Xiong, I. Huerta, J. Gonzàlez, and X. Roca. Beyond the static camera: Issues and trends in active vision. In *Visual Analysis of Humans*. 2011.

[7] M. Ancona, E. Ceolini, C. Oztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*, 2018.

[8] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[9] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *ICCV*, 2015.

[10] B. E. Bejnordi, T. Blankevoort, and M. Welling. Batch-shaping for learning conditional channel gated networks. *arXiv preprint*, 2019.

[11] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint*, 2013.

[12] N. Bore, R. Ambrus, P. Jensfelt, and J. Folkesson. Efficient retrieval of arbitrary objects from long-term robot observations. *RAS*, 2017.

[13] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *ECCV*, 2016.

[14] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[16] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[17] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[18] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[19] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019.

[20] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han. Cross-modal image-text retrieval with semantic consistency. In *ACM MM*, 2019.

[21] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021.

[22] Z. Chen, Y. Li, S. Bengio, and S. Si. You look twice: Gaternet for dynamic filter selection in cnns. In *CVPR*, 2019.

[23] R. Cheng, A. Agarwal, and K. Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *CORL*, 2018.

[24] J. Choi, C. Gao, J. C. Messou, and J.-B. Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *NeurIPS*, 2019.

[25] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[26] T. Cohen and M. Welling. Group equivariant convolutional networks. In *ICLR*, 2016.

[27] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *NeurIPS*, 2016.

[28] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[30] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[31] A. Edsinger and C. C. Kemp. Human-robot interaction for cooperative manipulation: Handing objects to one another. In *RO-MAN*, 2007.

[32] C. Esteves, A. Makadia, and K. Daniilidis. Spin weighted spherical cnns. *arXiv preprint arXiv:2006.10731*, 2020.

[33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[34] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018.

[35] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[36] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[37] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, 2008.

[38] R. Furuta, N. Inoue, and T. Yamasaki. Efficient and interactive spatial-semantic image retrieval. *MTA*, 2019.

[39] C. Gao, J. Xu, Y. Zou, and J.-B. Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020.

[40] C. Gao, Y. Zou, and J.-B. Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018.

[41] E. J. Gibson and A. S. Walker. Development of knowledge of visual-tactual affordances of substance. *Child development*, 1984.

[42] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. In *NeurIPS*, 2017.

[43] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[44] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[45] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *ICCV*, 2015.

[46] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *ICCV*, 2015.

[47] A. Gonzalez-Garcia, D. Modolo, and V. Ferrari. Do semantic parts emerge in convolutional neural networks? In *IJCV*, 2015.

[48] Google. Google image search engine toolkit. `https://developers.google.com/image-search/v1/devguide`, 2015.

[49] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016.

[50] Y. Gu, C. Li, and Y.-G. Jiang. Towards optimal cnn descriptors for large-scale image retrieval. In *ACM MM*, 2019.

[51] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009.

[52] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint*, 2015.

[53] T. Gupta, A. Schwing, and D. Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. *arXiv preprint*, 2018.

[54] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[55] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[56] S. Herdade, A. Kappeler, K. Boakye, and J. Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.

[57] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[58] Z. Hou, X. Peng, Y. Qiao, and D. Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020.

[59] D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. C. Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *CVPR*, 2018.

[60] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[61] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint*, 2016.

[62] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In *ICCV*, 2015.

[63] D. Jayaraman and K. Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *CVPR*, 2016.

[64] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto. Class-weighted convolutional features for visual instance search. *arXiv preprint arXiv:1707.02581*, 2017.

[65] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel. Visual search at pinterest. In *ACM SIGKDD*, 2015.

[66] L. Karacan, E. Erdem, and A. Erdem. Structure-preserving image smoothing via region covariances. *TOG*, 2013.

[67] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[68] O. S. Kayhan and J. C. v. Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *CVPR*, 2020.

[69] M. Kilickaya, N. Hussein, E. Gavves, and A. W. Smeulders. Self-selective context for interaction recognition. *arXiv preprint arXiv:2010.08750*, 2020.

[70] M. Kilickaya and A. W. Smeulders. Diagnosing rarity in human-object interaction detection. In *CVPRW*, pages 904–905, 2020.

[71] M. Kilickaya and A. W. Smeulders. Human-object interaction detection without alignment supervision. 2021.

[72] M. Kilickaya and A. W. Smeulders. Structured visual search via composition-aware learning. In *WACV*, 2021.

[73] S. A. W. Kilickaya, Mert and E. Gavves. Subjects and objects, where is the interaction? 2019.

[74] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020.

[75] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak. Deep metric learning beyond binary supervision. In *CVPR*, 2019.

[76] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[78] S. K. Kumaraswamy, M. Shi, and E. Kijak. Detecting human-object interaction with mixed supervision. In *WACV*, 2021.

[79] S. Kwak, M. Cho, and I. Laptev. Thin-slicing for pose: Learning to understand pose without explicit pose estimation. In *CVPR*, 2016.

[80] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *TOG*, 2014.

[81] S. Lallée, E. Yoshida, A. Mallet, F. Nori, L. Natale, G. Metta, F. Warneken, and P. F. Dominey. Human-robot cooperation based on interaction learning. In *From motor learning to interaction learning in robots*. 2010.

[82] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *CVPR*, 2016.

[83] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[84] H. Lejsek, B. . Jónsson, L. Amsaleg, and F. H. Ásmundsson. Dynamicity and durability in scalable visual instance search. *arXiv preprint*, 2018.

[85] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019.

[86] Y. Li, Z. Miao, J. Wang, and Y. Zhang. Nonlinear embedding neural codes for visual instance retrieval. *Neurocomputing*, 2018.

[87] S. Liao, E. Gavves, and C. G. Snoek. Spherical regression: Learning viewpoints, surface normals and 3d rotations on n-spheres. In *CVPR*, 2019.

[88] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

[89] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[90] Y. Liu, Q. Chen, and A. Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020.

[91] J. Ma, S. Pang, B. Yang, J. Zhu, and Y. Li. Spatial-content image search in complex scenes. In *WACV*, 2020.

[92] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint*, 2016.

[93] L. Mai, H. Jin, Z. Lin, C. Fang, J. Brandt, and F. Liu. Spatial-semantic image search by visual feature synthesis. In *CVPR*, 2017.

[94] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[95] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016.

[96] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia. Rotation equivariant vector field networks. In *ICCV*, 2017.

[97] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[98] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[99] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint*, 2013.

[100] G. Mori, C. Pantofaru, N. Kothari, T. Leung, G. Toderici, A. Toshev, and W. Yang. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint*, 2015.

[101] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 2005.

[102] D. Nunes, L. A. Ferreira, P. E. Santos, and A. Pease. Representation and retrieval of images by means of spatial relations between objects. In *AAAI*, 2019.

[103] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[104] N. P. Papanikolopoulos. *Controlled active vision*. Carnegie Mellon University, 1992.

[105] H. W. Park and A. M. Howard. Retrieving experience: Interactive instance-based learning methods for building robot companions. In *ICRA*, 2015.

[106] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[107] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CVIU*, 2016.

[108] B. Peterson. *Learning to see creatively: Design, color, and composition in photography*. Amphoto Books, 2015.

[109] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

[110] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *TPAMI*, 2012.

[111] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[112] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *PAMI*, 2018.

[113] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual instance retrieval with deep convolutional networks. *TMTA*, 2016.

[114] F. Rémy, L. Saint-Aubert, N. Bacon-Macé, N. Vayssière, E. Barbeau, and M. Fabre-Thorpe. Object recognition in congruent and incongruent natural scenes: A life-span study. *Vision research*, 2013.

[115] F. Rémy, N. Vayssière, D. Pins, M. Boucart, and M. Fabre-Thorpe. Incongruent object/context relationships in visual scenes: Where are they processed in the brain? *Brain and cognition*, 2014.

[116] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[117] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, 2016.

[118] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *SIGCHI*, 2003.

[119] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[120] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.

[121] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[122] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010.

[123] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, 2017.

[124] K. Simonyan and A. Ziserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[125] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[126] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[127] E. Smith, D. Meger, L. Pineda, R. Calandra, J. Malik, A. Romero Soriano, and M. Drozdzal. Active 3d shape reconstruction from vision and touch. *NeurIPS*, 2021.

[128] E. Spaak, M. V. Peelen, and F. de Lange. Scene context impairs perception of semantically congruent objects. *BioRxiv*, 2020.

[129] M. Tamura, H. Ohashi, and T. Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[130] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NeurIPS*, 2012.

[131] R. Tao, A. W. Smeulders, and S.-F. Chang. Attributes and categories for generic instance search from one example. In *CVPR*, 2015.

[132] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[133] J. R. Uijlings, A. W. Smeulders, and R. J. Scha. What is the spatial extent of an object? In *CVPR*, 2009.

[134] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[136] A. Veit and S. Belongie. Convolutional networks with adaptive inference graphs. In *ECCV*, 2018.

[137] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018.

[138] X. Wang and A. Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[139] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019.

[140] M. Weiler and G. Cesa. General e (2)-equivariant steerable cnns. In *NeurIPS*, 2019.

[141] M. Weiler, F. A. Hamprecht, and M. Storath. Learning steerable filters for rotation equivariant cnns. In *CVPR*, 2018.

[142] D. Worrall and G. Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–584, 2018.

[143] M. F. Wurm and R. I. Schubotz. What's she doing in the kitchen? context helps when actions are hard to recognize. *Psychonomic bulletin & review*, 2017.

[144] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[145] H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *SIGIR*, 2010.

[146] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via l 0 gradient minimization. In *SIGGRAPH Asia*, 2011.

[147] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[148] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint*, 2018.

[149] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[150] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017.

[151] R. Zhang. Making convolutional networks shift-invariant again. *ICML*, 2019.

[152] L. Zheng, Y. Yang, and Q. Tian. Sift meets cnn: A decade survey of instance retrieval. *PAMI*, 2017.

[153] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015.

[154] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2018.

[155] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.

SAMENVATTING

Dit proefschrift stelt een contextueel begrip van mens-object interacties voor. Wij verlichten de uitdaging van het begrijpen van mens-object interacties door meerdere bronnen van context op te nemen in deep learners. We onderzoeken eerst de bronnen van contextuele informatie in interacties, door de visuele omvang van mens-object interacties in hoofdstuk 2 te bestuderen. We stellen vast dat de lokaliteit en de compositionaliteit van de mens-object interacties een belangrijke rol spelen bij het begrijpen van interacties. Vervolgens ontwikkelen we meerdere bronnen van lokale contextuele informatie in hoofdstuk 3, waarvan we laten zien dat ze drastisch helpen bij interactieherkenning. Vervolgens integreren we de compositorische context tussen de menselijke interactor en de object interactee om interactie te zoeken in grote beelddatabases in hoofdstuk 4. Met de observatie dat, gegeven een beeld, slechts enkele mens-object paren in een interactie zijn, integreren we sparsity context om interactie detectie uit te voeren in Hoofdstuk 5.

Hoofdstuk 2: Waar is de interactiecontext? Een empirische studie

Hoofdstuk 2 bestudeert de visuele omvang van interacties tussen mensen en objecten. Visuele interacties tussen mens en object zijn moeilijk aan te wijzen in een beeld. Waar objecten en subjecten duidelijke grenzen hebben, is hun interactie dat niet. In dit werk proberen we de interacties tussen mens en voorwerp in een enkel beeld vast te stellen door hun visuele omvang te bestuderen. Waar is het visuele bewijs voor de interacties in een beeld? Wij gaan uit van waarneembare gebieden zoals het onderwerp en het voorwerp om te bepalen welk gebied effectief is bij het leren herkennen van interacties. Vervolgens bedenken we een orakelstrategie om de regio te bepalen die de beste herkenningsprestatie oplevert. Dit levert een bovengrens op voor interactieherkenning in onze setting. Tenslotte onderzoeken we het belang van visuele details binnen dit beperkte gebied. Onze bevindingen tonen aan dat: *i)* interacties kunnen profiteren van zelfs eenvoudige opname van de context in de herkenning, *ii)* het vinden van de beste context per beeld nog meer helpt en, *iii)* kleine details rond het snijpunt van subject-object belangrijk zijn in de herkenning.

Hoofdstuk 3: Zelf-selectieve context voor interactieherkenning

Dit hoofdstuk bestudeert de lokale context van mens-object interacties. Mens-object interactie herkenning is gericht op het identificeren van de relatie tussen een menselijk subject en een object. Onderzoekers nemen als oplossing globale scènecontext op in de eerste lagen van diepe Convolutionele Neurale Netwerken. Zij melden een aanzienlijke verbetering van de prestaties omdat interacties in het algemeen gecorreleerd zijn met de scène (een fietsende man in een stadsstraat). Deze aanpak leidt echter tot de volgende problemen. Het verhoogt de netwerkgrootte in de eerste lagen, en is daarom niet efficiënt. Het leidt tot lawaaierige filterresponsen wanneer de scène niet relevant is, dus niet accuraat. Het maakt alleen gebruik van de context van de scène, terwijl interacties tussen mens en object een veelheid aan contexten bieden en dus onvolledig zijn. Om deze problemen te omzeilen, stellen wij in dit werk Self-Selective Context (SSC) voor. SSC werkt op de gezamenlijke verschijning van mens-object en context om de meest discriminerende context(en) in te zetten voor herkenning. Wij bedenken nieuwe contextuele kenmerken die de lokaliteit van mens-object interacties modelleren en tonen aan dat SSC naadloos kan integreren

met de state-of-the-art interactie herkenningsmodellen. Onze experimenten tonen aan dat SSC leidt tot een belangrijke verbetering van de interactieherkenning, terwijl er veel minder parameters nodig zijn.

Hoofdstuk 4: Gestructureerd visueel zoeken via compositiebewust leren

Dit hoofdstuk bestudeert visueel zoeken met behulp van gestructureerde zoekopdrachten. De structuur heeft de vorm van een 2D samenstelling die de positie en de categorie van de objecten codeert. De transformatie van de positie en de categorie van de objecten leidt tot een doorlopende relatie tussen visuele composities, die zeer nuttige informatie bevat, hoewel die door eerdere technieken niet werd benut. Daarom is ons doel in dit werk deze continue relaties te benutten door gebruik te maken van het begrip symmetrie in equivariantie. Onze modeluitvoer wordt getraind om symmetrisch te veranderen ten opzichte van de invoertransformaties, wat leidt tot een gevoelige kenmerkruimte. Dit leidt tot een zeer efficiënte zoektechniek, aangezien onze aanpak leert van minder gegevens en een kleinere kenmerkruimte gebruikt. Experimenten op twee grootschalige benchmarks van MS-COCO [89] en HICO-DET [17] tonen aan dat onze aanpak leidt tot een aanzienlijke prestatiewinst ten opzichte van concurrerende technieken.

Hoofdstuk 5: Mens-voorwerp interactie detectie zonder uitlijningstoezicht

Het doel van dit hoofdstuk is de detectie van mens-object interactie (HO-I). HO-I-detectie heeft tot doel interacterende mens-objectgebieden te vinden en hun interactie uit een beeld te classificeren. Onderzoekers hebben de laatste jaren aanzienlijke verbeteringen bereikt door te vertrouwen op een sterke HO-I alignment supervision van [17]. HO-I alignment supervision koppelt mensen aan de objecten waarmee ze interageren, en stemt vervolgens mens-objectparen af op hun interactiecategorieën. Omdat het verzamelen van dergelijke annotatie duur is, stellen we in dit artikel voor om HO-I te detecteren zonder alignment supervision. In plaats daarvan vertrouwen we op toezicht op beeldniveau dat alleen bestaande interacties binnen het beeld opsomt zonder aan te geven waar ze plaatsvinden. Ons artikel levert drie bijdragen: *i)* We stellen Align-Former voor, een op visuele transformatie gebaseerde CNN die HO-I kan detecteren met alleen toezicht op beeldniveau. *ii)* Align-Former is uitgerust met een HO-I align laag, die kan leren om geschikte doelen te selecteren om detector supervisie mogelijk te maken. We evalueren Align-Former op HICO-DET [17] en V-COCO [52], en laten zien dat Align-Former het veel beter doet dan bestaande HO-I-detectoren met beeldtoezicht ($4,71\%$ mAP-verbetering ten opzichte van $16,14\%$ mAP-verbetering op HICO-DET [17]).

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my supervisor, Prof. Arnold Smeulders. Peter Higgs, famous physicist who transformed the world of physics with his discovery, Higgs Boson, claims he would not be productive enough in today's fast-paced, result oriented academic system. In a world looking forward to the next State-of-the-Art results, you've guided me to the real objective of a scientist: The intuition. You patiently waited for me to build my intuition on one of the most interesting and challenging problem of vision: The visual interactions. Thank you very much for sticking with me throughout the PhD. I will always remember our philosophical discussions on how the visual world of humans and computer work, with big joy, thank you!

Second, I would like to thank my co-promotor, Prof. Cees Snoek. Cees, you are the prototypical Dutch, with your result orientedness and concise communication. I had neither of them before arriving to the Netherlands, so tried to leverage our short face time to improve on both sides. Thank you Cees!

Third, I would like to express my gratitude to my committee members, Prof. Judith Good, Prof. Christof Monz, Prof. Alberto del Bimbo, Dr. Pascal Mettes and Dr. Efstratios ”Stratis” Gavves. Thank you very much for taking your valuable time, and your enthusiastic feedback! Special thanks to Prof. Christof for the invaluable discussions from the language perspective throughout the PhD.

Dear Amsterdam, I would like to thank you for being an amazing city. Your art, your canals, your parks, your diversity, your enthusiasm and ability to surprise anybody, is just wonderful. Thanks to you, I have witnessed all sorts of interactions one can have with a bicycle, which was a great inspiration to me. Talking about Amsterdam, I would like to thank my friends outside the academic circle, with whom I fully enjoyed what Amsterdam can offer. I would like to especially thank Canan and Jason for taking me to amazing movies and exhibitions, or just enjoying a nice coffee with me at Kriterion.

Dearest Virginie Mes, I do not know how to put my gratitude to you into English words. You are simply the best of the best. I really enjoyed our hallway jokes. You were one of my biggest emotional support during the pandemic. Thank you, Virginie!

I would like to thank Dr. Dennis Koelma. Dennis, you have supported me thoroughly in all technical aspects of a PhD. Thanks for your patience with my never ending questions!

I would like to thank my dear friend, Dr. William Thong. William, without your support, this PhD would not be possible. You are one of the greatest listener I have ever met, your jokes are fantastic, and you are one of the best person to take a coffee walk or discuss movies with. I am so lucky to have you besides me whenever I needed. Thank you, William!

My dearest QUVAns, I learned a lot from you. Dear Dr. Nour Hussein, you took me down back to Earth, and taught me how to properly train a neural network. You always took your time to listen and exchange ideas. My PhD took a great turn after my chance to collaborate with you, thank you very much. Dear Dr. Changyong OH, you taught me how to translate my abstract ideas into a mathematical formulation that can be implemented. Dear Maurice Weiler, you taught me the importance of Equivariance, which was a great inspiration for the third chapter of this thesis. I also would like to thank Dr. Berkay Kıcanaoğlu, Dr. Anıl Başlamışlı, Hoàng-Ân Lê and Dr. Peter O'Connor for their wonderful chat, funny jokes and everything.

I also would like to thank my current labmates. I enjoy discovering and working on something completely different with you: Ceren and Onur Yıldırım, Fangqin Zhou, Israel Jurado, Dr. Joaquin Vanschoren and Dr. Faysal Boughorbel.

I also would like to thank my M.Sc. supervisor, Prof. Nazlı İkizler-Cinbiş, for inspiring me to work on human-object interactions, and training me well to continue with my PhD. Thank you, Prof. Nazlı!

Degerli annem Solmaz, babam İlyas, ikiz kardesim Yiğit, ablam Ilknur ve yegenim Eylül. Değerli ailem, sizler benim en büyük destekçimsiniz. Sizlerin varlığı içimi ısıtıyor, beni güvenli hissettiriyor, ve aklıma gelen her türlü zorluga girişme ve dayanma gücü veriyor. Beni, küçücük bir çocukken izledigim Nobel ödülü törenlerinden esinlenip bilim insani olmak istediğimi söylediğim günden beri, koşulsuz ve dopdolu bir sevgiyle desteklediniz. Şimdi sizinle birlikte bu cocukluk hayaline ulaşmış olmanın haklı coşkusu içerisindeyim. Sizin gibi muhteşem bir ailenin içerisine doğmus olmak bana hayatta sunulan en büyük şans. Umarım hayatım boyunca bu şansımı en güzel şekilde değerlendiririm. Hepinize çok ama çok teşekkur ederim. Eylül'ümün yanaklarından da kocaman öperim!

Mert Kılıçkaya
November 2022