A Deep Learning Approach to 2D/3D Object Affordance Understanding

by

Spyridon Thermos

Submitted to the Department of Electrical and Computer Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

at the

UNIVERSITY OF THESSALY

May 2020

Author . . . .
Department of Electrical and Computer Engineering
May, 2020

Certified by . . . .
Professor Gerasimos Potamianos, Chair
Thesis Supervisor

Certified by . . . .
Dr. Petros Daras
Thesis Supervisor

Certified by . . . .
Professor Antonios Argyriou
Thesis Supervisor

“It is the people who no one imagines anything of, who do the things that no one can imagine"

Alan Turing

A Deep Learning Approach to 2D/3D Object Affordance Understanding

by

Spyridon Thermos

Submitted to the Department of Electrical and Computer Engineering
on May, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Abstract

A long-standing challenge in the computer vision field is to recognize the perceived objects and leverage their rich visual information. In fact, objects constitute key elements for a wide variety of real-world applications; from scene understanding and industry automation to security and robotics. Significant steps have been made towards 2D/3D object detection and recognition over the last few years that were complemented by the rapid advancements in processing units technology. However, robust object understanding remains an open challenge since recent works focus mostly on the appearance attributes of the objects, such as shape and texture, and omit any information about their functionalities.

In this dissertation we develop models and techniques that allow us to understand and exploit these functionalities, also known as object "affordances", *i.e.* the set of actions that humans can perform while interacting with the object. In particular, first we investigate the impact of object affordances to RGB-D object recognition through the "function from motion" perspective, where the affordance information is extracted by observing human-object interactions. Motivated by the research findings of cognitive neuroscience, we are the first to apply the so-called "sensorimotor" learning theory in computer vision, using end-to-end deep neural networks to fuse the object appearance (sensory) and affordance (motor) information and improving object recognition in RGB-D videos. Second, we develop an encoder-decoder model that is able to localize and segment the object parts that support specific human-object interactions. Rather than relying on object-specific information, such as bounding boxes and class labels, our model is able to learn to focus in the interaction spot through processing spatio-temporal information and predict affordance segmentation masks both in RGB-D videos and static images. Lastly, we introduce SOR3D, the first large-scale RGB-D dataset that consists of human-object interaction sequences and facilitates affordance-related research. The corpus is publicly available and includes various object and action related annotations, ranging from video-level object and action class labels to frame-level affordance heatmaps and segmentation masks.

Extensive experiments on the introduced SOR3D dataset demonstrate the efficacy

of the proposed models in the aforementioned tasks. From the presented results, we observe that: a) the utilization of object affordance information leads to improved object recognition, and b) object affordance localization and segmentation in videos and static images can be achieved without the need for extra object-related information, such as object class and location.

Greek Abstract

Αναγνώριση Χαρακτηριστικών 2Δ/3Δ Αντικειμένων με Καινοτόμες Τεχνολογίες Βαθιάς Μάθησης και Αλληλεπίδρασης

Η δυνατότητα να αναγνωρίζουμε τα αντικείμενα που μας περιβάλουν και να αξιοποιούμε την πλούσια οπτική πληροφορία που τα χαρακτηρίζει, αποτελεί μια σημαντική πρόκληση για τον τομέα της όρασης υπολογιστών. Τα αντικείμενα αποτελούν στοιχεία κλειδιά για ένα ευρύ πεδίο εφαρμογών που εκτείνεται από την κατανόηση χαρακτηριστικών σκηνής και τον αυτοματισμό, μέχρι την ασφάλεια και τη ρομποτική. Τα τελευταία χρόνια έχουν γίνει σημαντικά βήματα προς τον εντοπισμό και την αναγνώριση 2Δ/3Δ αντικειμένων χρησιμοποιώντας τεχνικές βαθιάς μάθησης, που συνοδεύτηκαν από τη σημαντική βελτίωση στον τομέα της υπολογιστικής ισχύος. Ωστόσο, η εύρεση αποτελεσματικών αλγορίθμων για την κατανόηση των χαρακτηριστικών ενός αντικειμένου παραμένει μια ανοιχτή πρόκληση, μιας και οι υπάρχουσες ερευνητικές εργασίες επικεντρώνονται κυρίως στα χαρακτηριστικά εμφάνισης των αντικειμένων, όπως το σχήμα και το χρώμα, αγνοώντας τη λειτουργικότητά τους.

Στην παρούσα διατριβή αναπτύσσονται μοντέλα και τεχνικές για την κατανόηση της λειτουργικότητας των αντικειμένων, η οποία καθορίζει τους τρόπους με τους οποίους μπορούν να χρησιμοποιηθούν τα αντικείμενα αυτά από τον άνθρωπο. Αρχικά, εξετάζεται η επίδραση της λειτουργικότητας των αντικειμένων ως πρόσθετο χαρακτηριστικό για την αναγνώρισή τους. Το χαρακτηριστικό αυτό εξάγεται παρατηρώντας ακολουθίες αλληλεπίδρασης ανθρώπου-αντικειμένου. Μάλιστα, αξιοποιώντας πρόσφατα αποτελέσματα από έρευνες στον τομέα των νευροεπιστημών, εφαρμόζεται για πρώτη φορά η λεγόμενη «αισθητικοκινητική» μάθηση στο πεδίο της όρασης υπολογιστών, χρησιμοποιώντας μοντέλα βαθιάς μάθησης ώστε να συνδυαστούν χαρακτηριστικά εμφάνισης και λειτουργικότητας (μέσω κίνησης) με σκοπό τη βελτίωση της αναγνώρισης 2Δ/3Δ αντικειμένων σε βίντεο και εικόνες. Στη συνέχεια, παρουσιάζεται ένα μοντέλο κωδικοποίησης-αποκωδικοποίησης πληροφορίας για τον εντοπισμό και το διαχωρισμό (σε επίπεδο εικονοστοιχείου) του μέρους του αντικειμένου που υποστηρίζει συγκεκριμένες χρήσεις. Η παραπάνω διαδικασία είναι εφαρμόσιμη και σε δεδομένα βίντεο και εικόνας. Μάλιστα, το συγκεκριμένο μοντέλο έχει τη δυνατότητα να επικεντρώνεται στο σημείο της επαφής του ανθρώπου με το αντικείμενο κατά τη διάρκεια της αλληλεπίδρασης, χωρίς την ανάγκη χρησιμοποίησης πρόσθετης πληροφορίας όπως είναι η κλάση ή η ακριβής τοποθεσία του αντικειμένου. Τέλος, παρουσιάζεται η πρώτη εκτενής βάση δεδομένων που μπορεί να χρησιμοποιηθεί για την εκπαίδευση και την αξιολόγηση μοντέλων που επεξεργάζονται χαρακτηριστικά λειτουργικότητας αντικειμένων. Η συγκεκριμένη βάση δεδομένων είναι διαθέσιμη για δημόσια χρήση και αποτελείται από δεδομένα RGB-D βίντεο (περιέχοντας δηλαδή σε κάθε πλαίσιο εικόνας και δεδομένα χρωματικού πεδίου και δεδομένα βάθους) τα οποία απεικονίζουν αλληλεπιδράσεις ανθρώπων με αντικείμενα. Ακόμα, περιέχει επισημειώσεις για τα παραπάνω δεδομένα σε μορφή κλάσεων για τα αντικείμενα και τις αλληλεπιδράσεις, σε επίπεδο βίντεο, εικόνας, αλλά και εικονοστοιχείου.

Η αποτελεσματικότητα των μοντέλων που σχεδιάστηκαν για τους παραπάνω σκοπούς αποδεικνύεται μέσω εκτενών πειραμάτων που αξιοποιούν δεδομένα από την παραπάνω βάση. Συγκρίνοντας τα παραπάνω αποτελέσματα με αντίστοιχα της βιβλιογραφίας εξάγονται δύο συμπεράσματα. Πρώτον, είναι σαφής η βελτίωση στην αναγνώριση αντικειμένων όταν αξιοποιείται η λειτουργικότητά τους ως πρόσθετο χαρακτηριστικό, και δεύτερον είναι δυνατός ο ακριβής εντοπισμός και διαχωρισμός του μέρους του αντικειμένου που υποστηρίζει μια συγκεκριμένη λειτουργικότητα σε δεδομένα βίντεο και εικόνας, και μάλιστα χωρίς να είναι απαραίτητη η ύπαρξη πρόσθετης πληροφορίας για το αντικείμενο.

Thesis Supervisor: Professor Gerasimos Potamianos, Chair

Thesis Supervisor: Dr. Petros Daras

Thesis Supervisor: Professor Antonios Argyriou

Acknowledgments

There are many people I would like to thank for contributing to the 4.5 years of my experience as a PhD student.

First I must thank my advisor Gerasimos Potamianos who, through countless emails and many meetings over these years, helped me develop into a competent researcher. Gerasimos' foresight and zeal for perfection are unique. I recall several times that I produced a terrible paper draft and he patiently repeated the same thing over and over until I understand how to communicate my work appropriately. I am grateful to Gerasimos for teaching me not just about some aspects of research, but for teaching me how to think.

I would also like to thank Petros Daras who co-advised my PhD. I have been very fortunate to have his constant support and experience in computer vision. When I pitched an idea Petros was there to distill its essence and provide valuable insights for transforming it to a research paper.

I would like to note the importance of working in a lab next to remarkably inspiring people. The Visual Computing Lab of CERTH has researchers who I have developed a lot of respect for, and who contributed to my PhD journey in many ways. I would like to especially thank Athanasios Psaltis for the countless discussions and pair-programming hours since we began our exploration in the deep learning universe. I must also thank Anargyros Chatzitofis and Nikolaos Zioulis for their unique perspectives and insightful comments, feedback, and advice. I would also like to thank Georgios Floros, Georgios Zoumpourlis, Vasileios Magoulianitis, Konstantinos Gkountakos, Konstantinos Apostolakis, and Alexandros Doumanoglou who all shaped my approach to research and inspired me to aim higher.

Finally, a big thank you to my wife Iordana, my parents, and my sister for the constant support during the undergrad, masters and PhD long long years. It is their faith in me that drives my ambition and my desire to make them proud.

Contents

List of Figures

List of Tables

List of Abbreviations

ACOF	**A**ccumulated **C**olorized **O**ptical **F**low
ACOFM	**A**ccumulated **C**olorized **O**ptical **F**low **M**agnitude
CDM	**C**olorized **D**epth**M**ap
cLSTM	**c**onvolutional **L**ong-**S**hort **T**erm **M**emory
CNN	**C**onvolutional **N**eural **N**etwork
COF	**C**olorized **O**ptical **F**low
COFM	**C**olorized **O**ptical **F**low **M**agnitude
CPU	**C**entral **P**rocessing **U**nit
GPU	**G**raphics **P**rocessing **U**nit
HHA	**H**orizontal disparity **H**eight **A**ngle
IoU	**I**ntersection **o**ver **U**nion
KLD	**K**ullback-**L**eibler **D**ivergence
LSTM	**L**ong-**S**hort **T**erm **M**emory
MLP	**M**ulti-**L**ayer **P**erceptron
RNN	**R**ecurrent **N**eural **N**etwork
SOR3D	**S**ensorimotor **O**bject **R**ecognition **3D**
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

1.1 Object Affordance

Objects constitute a significant part of the perceived environment. Their presence is usually all we need to define the rest of the scene. In fact, it is our brain that identifies objects and connects them with specific environments. That is, a “dish” and a “pan” indicate a kitchen environment, while a “board” and a “chalk” indicate a classroom one. Thus, due to the objects high correlation with the scene, computer vision researchers have dedicated tremendous efforts in detecting and recognizing them in the 2D and 3D space [3, 27, 40, 41, 83, 115]. To perform these tasks, the majority of the existing methods focus on the appearance attributes of the objects, such as shape and color [13,32,41,53–55]. However, the aforementioned characteristics vary dramatically in real-world scenarios, where object deformation, occlusions, and illumination variation occur.

Besides appearance, objects can also be defined by their functionality, the so-called object “affordance”. According to Gibson [22], “the affordances of the environment are what it offers the animal”, implying the complementarity between the animals and the environment. Based on this theory, Minsky [59] argues on the significance of understanding items according to what they can be used for, *i.e.* what they afford. These theoretical foundations have resulted to the so-called function-based object understanding, which can be viewed as an approach applicable to environments in

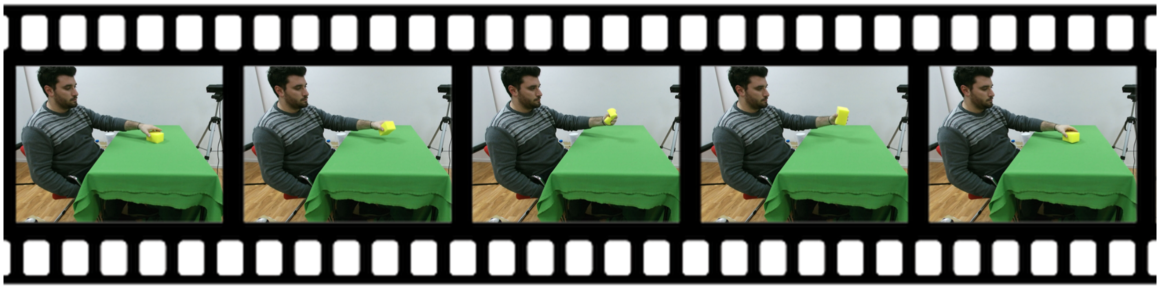


Figure 1-1: Example of the "function from motion" perspective. Here we observe a "squeeze sponge" video from the SOR3D dataset (see Chapter 4), where the human-object interaction indicates that the "sponge" is "squeezable".

which objects are designed or used for specific purposes [86]. Moreover, the work in [98] describes three possible ways for extracting affordance information for an object: a) "Function from shape", where the object shape provides some indication of its function; b) "Function from motion", where an observer attempts to understand the object function by perceiving a task being performed with it; and c) "Function from manipulation", where function information is extracted by manipulating the object.

In this thesis we adopt the "function from motion" perspective, focusing on understanding the affordance information through observing human-object interaction videos such as the one depicted in Fig. 1-1. In particular, we define affordance understanding as: a) the exploitation of the information "revealed" during the hand-object interaction for improving object recognition, and b) the localization and segmentation of the object parts that support specific interactions.

1.2 Recognition, Reasoning, and Segmentation

If we manage to understand and exploit object affordances, we can confidently answer two crucial questions: a) what is the identity of the perceived object, and b) how can we use it?

Regarding affordance-based object recognition, the existing types of information, *i.e.* the object appearance and affordance attributes, have motivated the investigation of the so-called “sensorimotor” learning approach [31, 44]. This approach is based on research findings in cognitive neuroscience that describe the human object perception as the fusion of sensory (object appearance) with motor (object affordance) information [107]. In this thesis, we adopt the sensorimotor learning paradigm and design a neuro-biologically inspired two-stream model for RGB-D object recognition. We implement both streams as state-of-the-art deep neural networks that process and fuse appearance and affordance information in multiple ways.

To answer the object functionality question, recent studies focus on optimizing the affordance localization and segmentation tasks. However, existing approaches in the localization context rely mostly on predicting saliency-based heatmaps [8, 34, 50, 71] in static images, without associating the heatmaps with specific affordance classes. Similarly, existing affordance segmentation methods [12, 62, 65, 66, 90] predict pixel-level affordance labels on objects that are detected in static images. However, these image-based approaches of affordance localization and segmentation are limited by a significant knowledge gap. That is, affordance information is spatio-temporal by its nature, and the temporal domain is fully omitted when learning from static representations. In this thesis, we design a deep encoder-decoder model that is able to process human-object interactions and predict object affordances based on spatio-temporal information. Using the aforementioned model, we jointly investigate the affordance reasoning, *i.e.* recognition and localization of the object affordance, and segmentation tasks.

1.3 Contributions and Outline

In this dissertation we design two learning frameworks for understanding the object affordance information. Both frameworks consist of deep neural network architectures that are trained and evaluated using human-object interaction videos. A brief description of each chapter follows that includes a brief overview of the Thesis contributions that can be found in Chapters 4, 5, and 6.

In **Chapter 2** we present existing works in the literature that exploit affordance information. First we report the relevant works for recognizing objects through embodiment, *i.e.* active manipulation of the object, and by observing human-object interactions. Then we present the works related to the object affordance localization and pixel-level segmentation.

In **Chapter 3** we describe the structure and advantages of widely-known convolutional and recurrent neural networks. These networks are partially or fully utilized in the model architectures proposed in this thesis for the investigated tasks. Additionally, we provide details about the deep learning frameworks used to implement the deep model architectures presented in this thesis.

In **Chapter 4** we provide details of our developed Sensorimotor Object Recognition 3D corpus (SOR3D), the first ever large-scale RGB-D dataset in the literature that consists of multiple object types and complex affordances. The content of this chapter is based primarily on [100], where the dataset is introduced, as well as on [99] where the SOR3D-AFF subset of SOR3D is presented in detail.

In **Chapter 5** we present our novel, neuro-biologically inspired two-stream model for the RGB-D object recognition task. Both streams are realized as deep neural networks that process and fuse appearance and affordance information in multiple ways. We develop three model variants to efficiently encode the spatio-temporal nature of the hand-object interaction, and investigate an attention mechanism that relies on the appearance stream confidence. The content of this chapter is based on the spatial and spatio-temporal models presented in [100] and [102], as well as on [101] where the attention mechanism is proposed.

In **Chapter 6** we propose a deep encoder-decoder model that learns to encode spatio-temporal information from human-object interaction videos and predict affordance heatmaps and pixel-level labels with minimal supervision. First, we focus on affordance reasoning as an independent task, using a model variant with only one decoder that predicts affordance heatmaps. Then, we employ a second decoder and train them jointly for both tasks, using the predicted affordance heatmap to guide the segmentation decoding process. Note that our model is able to infer affordance heatmaps and pixel-level labels both in videos and static images. The content of this chapter is based on [99] and [103].

Finally, in **Chapter 7** we identify the remaining challenges and discuss possible future directions, while in the **Appendix** we list the publications related to this thesis.

Chapter 2

Related Work

This chapter presents existing works in the literature that investigate the utilization of affordance information in object recognition, as well as affordance reasoning and segmentation. In particular, Section 2.1 discusses works on appearance-based object recognition using traditional computer vision methods or adopting the deep learning paradigm, but also recent studies that exploit object affordances to boost recognition. Subsequently, Section 2.2 presents works that investigate object affordance reasoning in static images and videos, while Section 2.3 discusses the state-of-the-art in pixel-level affordance segmentation.

2.1 Affordance-based Object Recognition

Object recognition is a fundamental problem in computer vision. Focusing solely on object appearance attributes, relevant approaches can be divided into two main categories: methods that represent the object with hand-crafted features and ones that learn deep object representations exploiting the deep learning paradigm. Characteristic works of the first category are reported in the survey of [3]. Regarding deep learning-based methods, numerous works appear in the literature. For example, among others, [54] proposes a recurrent Convolutional Neural Network (CNN) framework to classify objects based on their appearance and the context of the scene; [97] presents a multi-view CNN with a view-pooling layer to categorize 3D objects; [80]

proposes a volumetric CNN for object point-cloud processing and classification; [117] utilizes polynomial kernels and bilinear pooling in a CNN to aggregate local convolutional features in a 3D object representation; and [17] proposes a group-view CNN that models the hierarchical correlations among multiple 2D views of a 3D object, leading to a powerful 3D descriptor.

Besides appearance-based learning, there are extensive studies related to functional object recognition exploiting object affordances. In particular, affordance-oriented object recognition is investigated in the literature from two viewpoints: a) embodiment and b) observation. The former indicates the scenario where there is direct interaction of the perceiver with the object, while the latter denotes the scenario where the perceiver observes others interacting with the object.

2.1.1 Inferring Object Affordances from Embodiment

Regarding the embodiment scenario, object affordances that are inferred from agent-object interaction have been recently leveraged in object recognition. In particular, [91] concentrates on robotic grasping of novel objects using a set of 2D object views labeled with grasping points. Additionally, [31] proposes a Gaussian process to model object-related sensorimotor "contigencies" [69] and categorizes objects by "pushing" them and observing their displacement. Further, [56] employs the iCub and Meka robots to categorize objects by combining visual and proprioceptive knowledge with motion behavior observed during interaction. Focusing on more composite actions, [18] utilizes robotic "push", "pull", and "poke" actions to further explore object representations, while [60] presents a scenario where a robot with basic motor skills categorizes objects by observing human-object interactions and subsequently selects its own action that will have the same effects on the object. Additionally, [38] proposes a system for active visual recognition through agent-object interaction, where, given an initial view of the object, the system predicts how the choice of motion alters the environment and integrates the result of the object manipulation at each time-step to classify the object. The system is trained end-to-end using reinforcement learning.

Besides recognition, object affordances provide valuable feedback for numerous tasks in the field of cognitive vision and developmental robotics, such as scene understanding [9,110], action anticipation [47,48,85,121], and action prediction [20,21, 36,46,68,114,118]. However, further elaboration on this aspect of affordance-based learning lies outside the scope of this Thesis.

2.1.2 Observation-based Sensorimotor Learning

Learning to recognize objects by observing others interacting with them is a challenging machine learning task. However, recent works on observation-based sensorimotor object recognition mostly rely on simple fusion schemes (*e.g.* using simple Bayesian models or the product rule), hard assumptions (*e.g.* naive Gaussian prior distributions), and simplified experimental settings (*e.g.* few object types and simple affordances).

For example, [44] utilizes histograms of oriented gradients to model object appearance, while the global velocity, orientation, and joint angles of the hand are used to encode the affordance information. A binary Support Vector Machine (SVM) is trained for each stream, while the predicted object-hand pairs of 3 consecutive frames are utilized by factorial conditional random fields for the final object class prediction. This method is evaluated using a dataset of 6 objects and 3 affordances. Further, [45] proposes a framework where GIST-features of object appearance and affordance are used to form sensorimotor representations. Then, probabilistic reasoning comprised of a Bayesian network with information gain strategy is used for object classification, exploiting these representations. The method is evaluated on a dataset that consists of 8 object classes and a single affordance. Additionally, [6] encodes the object appearance as frequency histograms of 200 bins, while 22 motor features provided by a motion-capture glove sensor are used as affordance representation. The appearance and affordance features are fused using positively weighted linear combination of Mercer kernels and are used to train a one-versus-all SVM for object classification. The algorithm is evaluated on a dataset of 7 objects and 5 affordances. Finally, [122] proposes a framework aiming at understanding the affordance and the functional basis (*e.g.* the part of the hammer that touches a surface when hammering) of tool objects through observing a human, using them for task-oriented object recognition. Object appearance, action sequence, and physical quantities produced by the interaction are modeled using graphs and a ranking-SVM classifier is then trained to recognize the objects. The framework is evaluated on a dataset consisting of 10 objects and 3 affordances.

2.2 Reasoning About Object Affordances

Affordance reasoning is realized as the combination of the affordance localization, *i.e.* prediction of a heatmap on the object part that supports an interaction, and affordance recognition tasks. Early studies focus solely on the localization part, proposing saliency-based methods to predict affordance heatmaps in static images [8, 71]. More recent approaches adopt the "learning from observation" perspective by processing human-object interaction videos, and reason about object affordances by associating each predicted heatmap with the corresponding affordance class. In particular, [15] presents "Demo2Vec" that learns spatio-temporal embeddings from product demonstrations and predicts keypoints on the object affordance part. Further, [63] proposes a model that infers spatial hotspot maps on static images using gradient-weighted attention maps for pre-defined actions.

2.3 Segmenting Object Affordances

Object affordance segmentation, *i.e.* the pixel-wise identification of the object part that enables a specific interaction, is a challenging task that has been mostly treated as a static semantic segmentation problem, usually coupled with object detection. For example, [62] uses hierarchical matching pursuit, as well as normal and curvature features derived from RGB-D data, to learn pixel-wise labeling of affordances for common household objects, while [65] proposes an encoder-decoder architecture to predict pixel-wise affordances based on depthmaps. Further, [12] expands the architecture of [65] by adding a region proposal network [84] to predict the bounding box of the target object and also investigate the joint learning of detecting and segmenting the object affordance part. All aforementioned works rely on strong supervision, as each object affordance part must be fully annotated at pixel-level. On the other hand, [90] proposes a weakly-supervised setting using CNNs and keypoints annotation to predict reasonable but not precise pixel-level affordances, which are then refined using the GrabCut algorithm [88].

Chapter 3

Deep Learning Background

This chapter provides the technical background for the deep neural networks that are used in the context of affordance understanding. In particular, Section 3.1 describes two widely-known CNNs, Section 3.2 presents the vanilla RNN and a more efficient variant, while Section 3.3 provides details about the deep learning frameworks we used to implement the models presented in this thesis.

3.1 Convolutional Neural Networks

CNNs have dominated most computer vision neural network architectures specifically designed for handling data with some spatial topology, such as images, videos, and 3D voxels. They apply a hierarchy of operations, such as convolutions and non-linear activations, to embed the input into a feature space. This space is learned by optimizing an objective function that varies depending on the task. Some of the most popular CNN models in the literature are: a) the AlexNet [49], b) the VGGNet [95], and c) the ResNet [28]. In this thesis, we utilize a VGGNet variant as base network for all investigated tasks, which consists of 16 layers and is denoted as VGG16.

The main characteristics of the VGG16 model, depicted in Fig. 3-1, are the depth and simplicity. This model is widely used in the context of several computer vision tasks, *e.g.* semantic segmentation [92] and action recognition [16], as it reinforces the notion that CNNs should be deep networks to really take advantage of the hierarchical

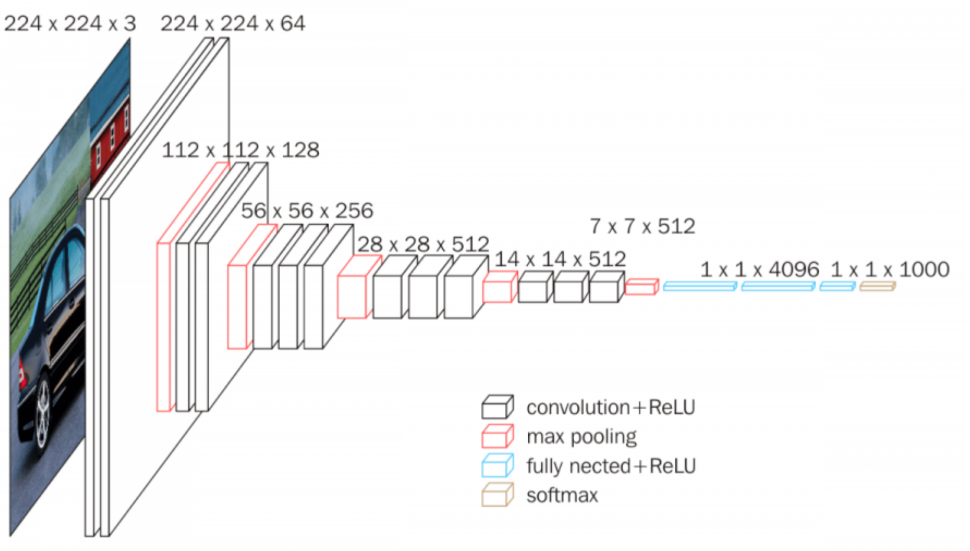


Figure 3-1: VGG16-CNN for image recognition (figure from [67]). The input image is being processed through a hierarchy of local convolutions, non-linear activation functions, and pooling. A classifier is trained to predict whether the input image belongs to one of 1000 classes of the ImageNet dataset [10].

representation of visual data. In particular, it consists of: a) thirteen convolutional (CONV) layers, b) five pooling (POOL) layers that perform 2 × 2 max-pooling with stride set to 1, and c) three fully connected (FC) layers on top of the network. Every CONV layer is followed by a Rectified Linear Unit (RL) non-linearity to filter its activations. In total the VGG16 architecture has approximately 138M parameters. The main advantage of VGG16 compared to similar CNN architectures is that all used CONV layers perform 3×3 convolutions. This enables the stacking of multiple CONV layers between consecutive POOL layers, as shown in Fig. 3-1, that creates an effective receptive field of 5 × 5 or 7 × 7 resolution (for stacking two or three CONV layers, respectively). The presented CONV stacking has another benefit. It enables the utilization of three RL activations instead of one, which leads to more discriminative learned representations. On the other hand, the main disadvantage of the VGG16 model is that it requires more memory compared to more recent architectures, such as ResNet [28], due to the high number of parameters (138M). However, most of these parameters are in the FC layers that are used only for the object recognition task, as VGG16 is used for convolutional feature extraction, *i.e.* up to the last CONV layer,

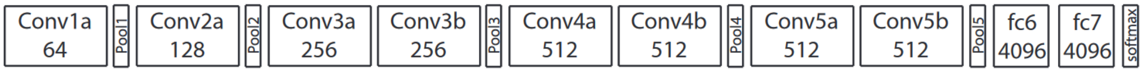


Figure 3-2: A schematic representation of the C3D model [105]. It can be seen that C3D follows the VGG16 structure, but utilizes 3D kernels in all convolutional and pooling layers.

in the context of affordance reasoning and segmentation.

However, the VGG16 model cannot encode any temporal dependencies and correlations as it relies solely on 2D convolutions. On the contrary, 3D convolutions are applied to both the spatial and the temporal domain, thus are advantageous in modeling spatio-temporal and 3D characteristics. This is the reason that they are used in several computer vision applications that process either 3D input data (*e.g.* voxelized representations) for tasks such as 3D object recognition [32,41,115] and 3D shape retrieval [51,89], or spatio-temporal data for activity [42] and action recognition [105]. Since this thesis focuses on understanding affordances by observing human-object interactions, we use 3D convolutions for encoding spatio-temporal information.

A widely used deep neural network that utilizes 3D convolutions for spatio-temporal information processing is the C3D model [105]. As depicted in Fig. 3-2, it consists of 8 3D convolutional (3DCONV), 5 POOL, and 2 FC layers, mimicking the VGG16 overall structure. This model processes groups of video frames, stacked along the RGB-channel axis to form 3D representations. Note that a very important module of C3D is the 3D POOL operators, as they further reduce the size of the input data, while preserving the encoded motion patterns and removing irrelevant information.

3.2 Recurrent Neural Networks

An RNN is a connectivity pattern that forms a directed graph along a temporal sequence of vectors $\{x_1, \ldots, x_T\}$. The ability to process variable length sequences is based on the internal state h_t of the model, also known as internal memory. The model uses a recurrence formula of the form $h_t = f_\theta(h_{t-1}, x_t)$, where f is a linear combination of the previous hidden state, *i.e.* h_{t-1}, with the current input vector x_t, followed by a non-linear activation function. Note that at each time step the model utilizes the same parameters θ, enabling the processing of sequences with an arbitrary number of vectors. A typical form of an RNN network is the following:

$$h_t = \tanh\left(W\begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}\right), \tag{3.1}$$

where the hidden vector h_t can be interpreted as a running summary of all vectors x until the time step t, and f is realized as the parameters W followed by a hyperbolic tangent *tanh* non-linearity. Note that in (3.1) the bias terms are omitted for brevity. However, the vanilla RNN has a major disadvantage when processing long sequences. In particular, during the backpropagation through time, the gradients either decrease or increase exponentially, depending on the used activation functions. The exponential increase problem, known as "exploding gradients", has been circumvented using a heuristic approach of clipping the gradients at some maximum value [74]. Nevertheless, the RNNs still suffer from the exponential decrease problem, also termed "vanishing gradients".

To address the aforementioned limitations of the RNN, the Long-Short Term Memory (LSTM) model is introduced in [30]. Its recurrence formula allows the input and hidden state vectors to interact in a more computationally complex manner that includes multiplicative interactions and propagates the gradients back in time more efficiently. The main contribution of the LSTM is its memory vector c_t, which can be used to read from, write to, or reset at each time step, using explicit gating mechanisms. In more detail, the update of an LSTM cell at each time step is described by the following equation:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W\begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}, \qquad \begin{aligned} c_t &= f \odot c_{t-1} + i \odot g \\ h_t &= o \odot \tanh(c_t) \end{aligned} \quad , \tag{3.2}$$

where the sigmoid function, denoted as σ, and the hyperbolic tangent are applied in

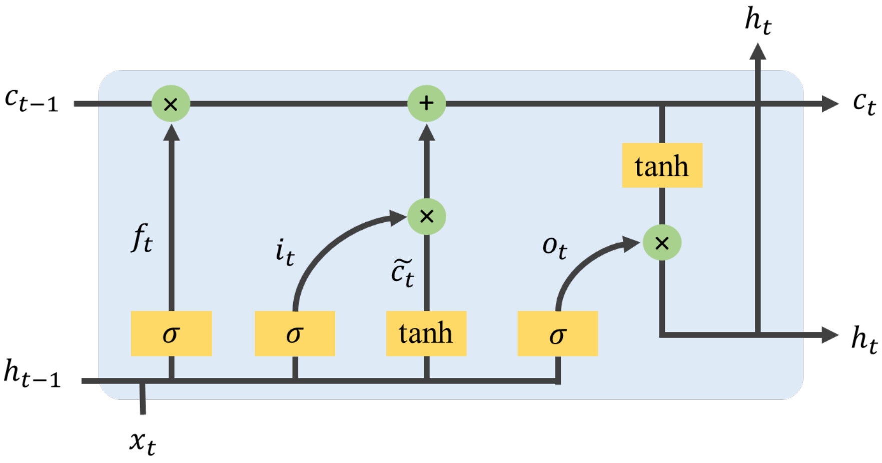


Figure 3-3: A schematic representation of a Long-Short Term Memory cell [30].

an element-wise manner. Note, that the i, f, and o vectors are realized as binary gates that control whether each memory cell is updated, reset to zero, or its local state is propagated to the hidden vector, respectively. This binary nature originates from the utilized sigmoids that are used as activation functions and make the model differentiable. Further, the vector g has values in the $[-1, 1]$ range due to the *tanh* activation function and is used to additively modify the memory cell c content. This additive form is a significant part of the LSTM as the sum operation distributes the gradients equally during back-propagation, allowing the gradients of c to flow backwards through time for long time periods.

3.3 Torch and PyTorch Frameworks

All models investigated in the context of this thesis are implemented using the Torch7[1] and PyTorch [75] frameworks. Both frameworks can be used to build arbitrary graphs of neural networks and parallelize them over CPUs and/or GPUs in an efficient manner. Torch7 supports LuaJIT, *i.e.* a user-friendly scripting language that is used for scientific programming, while its packages are implemented in C and CUDA. On the other hand, PyTorch supports the Python programming language and is a highly

[1] `http://torch.ch/`

efficient library for building complex and production ready deep learning architectures. Since Python is a highly flexible language, PyTorch has replaced Torch7 in most deep learning communities, such as in computer vision, natural language processing, and medical imaging. One of its key functionalities is that it applies a graph meta-programming based approach, where the executive code for defining layers, composing models, loading data, and running the optimizer are expressed by general purpose programming. This design ensures that any new neural network architecture can be easily implemented and generalized. Other advantages of this framework are its interoperability and extensibility, which enable the bi-directional exchange of data with several external libraries.

Chapter 4

The SOR3D Dataset

In this chapter, the large-scale dataset designed to advance research in affordance understanding is detailed. The corpus, denoted as SOR3D, consists of multiple object types and complex affordances, and focuses on the task of affordance-based object recognition. Note, that it constitutes the broadest and most challenging dataset in the affordance-based object recognition literature. A subset of SOR3D, denoted as SOR3D-AFF, provides extra annotations targeting the affordance reasoning and segmentation tasks. SOR3D, as well as the aforementioned subset, facilitate the development and efficient evaluation of affordance-based object recognition, affordance reasoning, and affordance segmentation approaches. The dataset is publicly available at `http://sor3d.vcl.iti.gr/`.

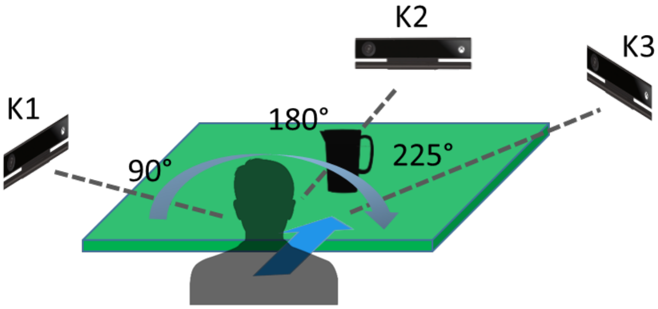


Figure 4-1: Schematic representation of the SOR3D capturing setup. The three Kinect sensors (K1-K3) are placed from left to right at 90°, 180°, and 225° with respect to the subject orientation.

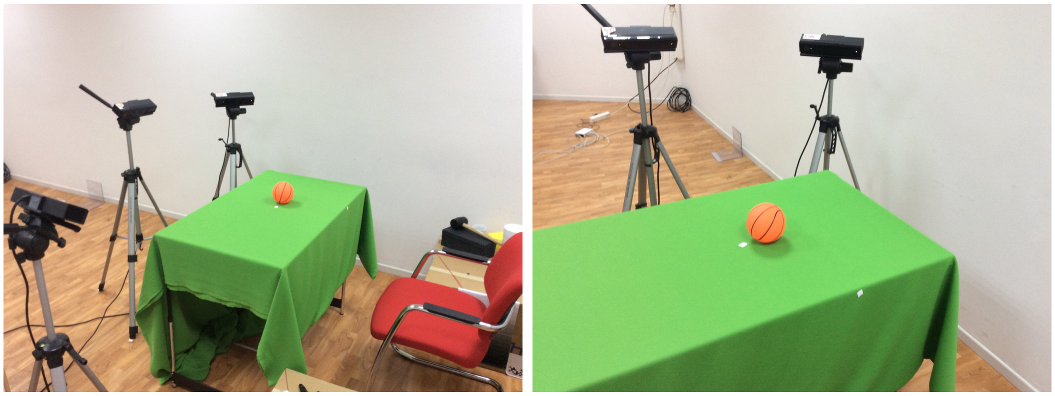


Figure 4-2: SOR3D capturing snapshots. The two markers on the tablecloth indicate the starting position of the object (at the middle of the table) and the hand (at the edge of the table).

4.1 Capturing Setup

The SOR3D dataset recording setup involved three synchronized Microsoft Kinect II sensors [58] in order to acquire RGB (1920 × 1080 resolution) and depth (512 × 424 resolution) streams from three different viewpoints, as depicted in Fig. 4-1, all at 30 Hz frame rate and an approximate 1.5 meters "head-to-device" distance. A monitor was utilized for displaying the "prototype" instance before the execution of every human-object interaction. Additionally, all involved subjects were provided with a ring-shaped remote mouse, held by the other hand than that interacting with the objects. This allowed the participants to indicate by themselves the start and end of each session (*i.e.* performing real-time annotation). Before the execution of any interaction, all objects were placed at a specific position on a desk, indicated by a marker on the table cloth, while a similar marker was placed at the edge of the table as a starting point for the hand that would participate in the hand-object interaction. The two markers are depicted in the two recording snapshots of Fig. 4-2.

The dataset was recorded under controlled environmental conditions, *i.e.* with negligible illumination variations (no external light source was present during the experiments) and a homogeneous static background (all human-object interactions were performed on top of a desk covered with a green tablecloth). Snapshots of the captured video streams from each viewpoint are depicted in Fig. 4-3.

Table 4.1: Supported object and affordance types in the SOR3D corpus. Considered object-affordance combinations are marked with √.

Object types	**Affordances**												
	Grasp	Lift	Push	Rotate	Open	Hammer	Cut	Pour	Squeeze	Unlock	Paint	Write	Type
Ball	√	√	√										
Book	√	√	√		√	√							
Bottle	√	√	√					√					
Box	√	√	√	√	√								
Brush	√	√									√		
Can	√	√	√										
Cup	√	√	√	√									
Hammer	√	√				√							
Key	√	√					√			√			
Knife	√	√					√						
Pen	√	√										√	
Pitcher	√	√	√	√				√					
Smartphone	√	√	√										√
Sponge	√	√	√	√					√				

4.2 Classes and Data Splits

Regarding the nature of the supported human-object interactions, a set of 14 object types was considered (each type having two individual instantiations, *e.g.* small and big ball). The appearance characteristics of the selected object types varied significantly, ranging from distinct shapes (like "box" or "ball") to more challenging ones (like "knife"). Taking into account the selected objects, a respective set of 13 affordance types was defined, covering typical manipulations of the defined objects. Concerning the complexity of the supported affordances, relatively simple (*e.g.* "grasp"), complex (*e.g.* leading to object deformations, like affordance "squeeze"), and continuous-nature ones (*e.g.* affordance "write") were included. In contrast, other experimental settings in the literature have mostly considered simpler and less time evolving affordances, like "grasp" and "push". In Table 4.1, all supported types of objects and affordances, as well as all combinations that have been considered in the dataset, are provided. As listed, a total of 54 object-affordance combinations (*i.e.* human-object interactions) are supported. All participants were asked to execute all object-affordance combinations indicated in Table 4.1 at least once. The experimental protocol resulted in a total of 20,830 videos, considering the data captured from each Kinect as a different human-object interaction instance. The length of every recording varied between 4

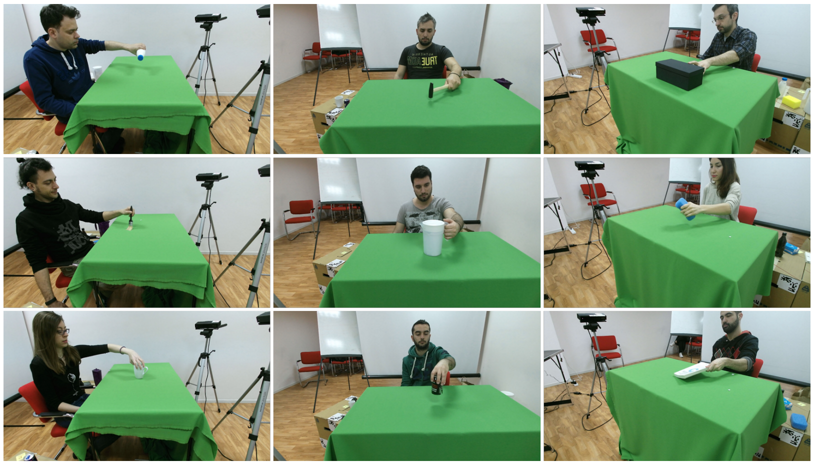


Figure 4-3: Examples of human-object interactions captured by the 3 Kinect sensors employed in the SOR3D corpus recording setup.

and 8 seconds. The dataset was split into training, validation, and test sets (25%, 25% and 50%) that correspond to approximately 5k, 5k and 10k hand–object interaction videos, respectively.

4.3 SOR3D-AFF Subset

In order to facilitate research in affordance reasoning and segmentation, the SOR3D-AFF dataset was created, providing heatmap and pixel-level segmentation mask annotations for object affordances. In particular, it consists of 1201 RGB-D videos, being a subset of SOR3D corpus. The data is split into 902 videos for training. 60 for validation, and 239 for testing. SOR3D-AFF supports 9 affordances types, namely "grasp", "cut", "lift", "push", "rotate", "hammer", "squeeze", "paint", and "type", of 10 common household objects, such as "pitcher" and "knife". Note that we choose to omit some object categories from SOR3D, as well as their corresponding affordances, as the affordance-related annotation for these objects is problematic (*e.g.* "pen" is fully occluded during interaction and has very noisy depthmap, "box" has a removable layer so the object shape in the last frame is not the same with the one during the interaction). Note that only videos captured from the K1 and K3 viewpoints (see Fig. 4-1) are used, as from the K2 viewpoint the interaction spot was not always visi-

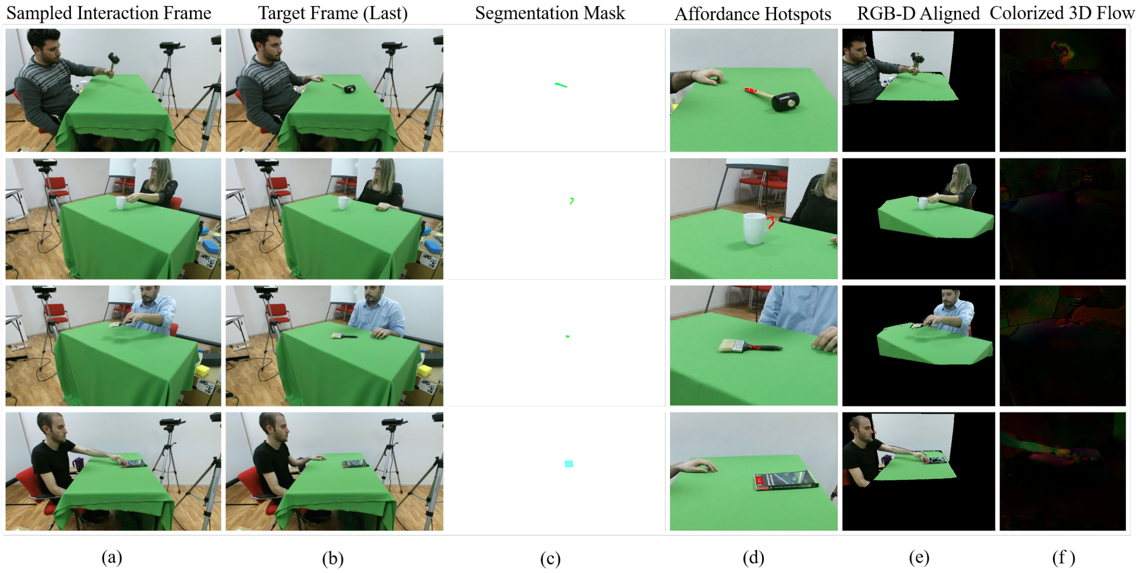


Figure 4-4: Four indicative SOR3D-AFF samples. From left to right: a) a frame sampled from the interaction sequence (1920 × 1080 pixel resolution), b) the last frame of the sequence that is used as target frame, c) the segmentation mask of the target frame (1920 × 1080 pixel resolution), d) hotspot annotations on the target frame (1920 × 1080 pixel resolution), e) the sampled interaction frame after RGB-D alignment (the color image is mapped to the depthmap resolution, *i.e.* 512 × 424), and f) the colorized 3D optical flow representation of the interaction frame after center-cropping to 300 × 300 pixel resolution.

ble. For each video, the following annotations are provided only for the last frame: a) pixel-level affordance segmentation mask, b) affordance heatmap based on Gaussian blurring of marked pixels that indicate the human-object interaction hotspot, c) object bounding box, and d) object label. The action label of each video is also provided, which is complementary to the corresponding affordance, *i.e.* “grasp”, “squeeze”. Some indicative annotated samples are depicted in Fig. 4-4.

4.4 Relevant Datasets in the Literature

Table 4.2 lists datasets that consist of tool-objects captured indoors and include affordance information. From this Table, we can observe that the datasets that include hand-object interaction sessions, apart from SOR3D, consist of small numbers of samples and are not publicly available (note that the TTU dataset [122] provides

Table 4.2: Datasets that consist of "tool-objects" (indoor scenes) and have affordance information available in the form of instance or pixel-level annotations. The SOR3D dataset is reported in the last row for comparison.

Dataset	Interaction	Format	Objects	Affordances	Subjects	Samples	Public availability
[44]	yes	RGB	6	3	4	28	no
[6]	yes	RGB	7	5	20	130	no
[45]	no	RGB	8	1	n/a	n/a	no
TTU [122]	yes	RGB-D	10	3	1	452	no
ADE-Affordance [96]	no	RGB	8	4	n/a	10,360	yes
IIT-AFF [66]	no	RGB-D	10	9	n/a	8,835	yes
COQE [61]	no	RGB	10	1	n/a	5000	yes
UMD [62]	no	RGB-D	17	7	n/a	10,000	yes
OPRA [15]	yes	RGB	n/a	7	n/a	20,612	yes
SOR3D	yes	RGB-D	14	13	105	20,800	yes

only the colored point clouds of the objects and not the hand-object sequences). The OPRA corpus [15] is an exception, however this dataset focuses on affordance reasoning by providing affordance heatmaps and action labels. On the other hand, ADE-Affordance [96], IIT-AFF [66], COQE [61], and the one from [45] include only static objects with no interaction, while the affordance information is represented as pixel-wise annotation of the object part that enables a specific affordance (*e.g.* the handle of a cup is annotated as "graspable") followed by the corresponding bounding box. These datasets are mostly used for affordance detection, reasoning, and segmentation. Part of the information in Table 4.2 is also reported in the recent survey on visual affordances by [26].

The datasets that are used for training and/or evaluation purposes in this thesis, apart from SOR3D/SOR3D-AFF, are further detailed below:

IIT-AFF. The dataset consists of a combination of images from ImageNet [10] and a collection from two RGB-D sensors at various resolutions. All images depict cluttered scenes that include multiple objects. The 9 supported affordance classes are: "contain", "cut", "display", "engine", "grasp", "hit", "pound", "support", and "w-grasp". For each image, there are pixel-level affordance labels and object bounding boxes.

UMD. The dataset provides pixel-level affordance labels for 105 kitchen, workshop, and garden tools. The tools were collected from 17 different object categories covering 7 affordance classes, namely "grasp", "cut", "scoop", "contain", "pound", "support”, “wrap-grasp”. For each dataset sample, a color image and the corresponding aligned depthmap are available, both in 640 × 480 pixel resolution.

OPRA. The dataset consists of 20,612 RGB video clips of various resolutions, depicting product reviews for appliances, such as pans and washing machines, split into 16,976 clips for training and 3,798 clips for validation. Each video contains the demonstration of an appliance feature, *e.g.* scoop food from the pan, and is paired with a static image that depicts the same object without any background or occlusion (target image). The supported affordance classes (7 in total) are the following: “hold”, “touch”, “rotate”, “push”, “pull”, “pick up”, “put down”. Note that each target image is annotated with an affordance heatmap, which is the result of Gaussian blurring applied on 10 marked pixels that indicate the location of the human-object interaction.

Chapter 5

Sensorimotor Object Recognition

5.1 Introduction

In this chapter we investigate the contribution of the affordance information to object recognition. As discussed in Chapter 1, objects constitute key elements for scene understanding, action identification, interaction prediction, and other computer vision tasks. Thus, the process of recognizing them in the context of an image or video has been an challenging research topic over the last decades.

There is accumulated evidence that humans, at an early stage of their lives, perceive objects by combining their visual attributes with the feedback from interacting with them. This process is known as "sensorimotor learning" [11, 19, 76], due to the parallel processing of the "sensory" and "motor" information in the human brain. Indeed, it is well established by cognitive scientists that there are two main streams that process the aforementioned information [106, 107]: the ventral stream that runs in the inferotemporal cortex is involved in the recognition of objects, while the dorsal one that projects to the posterior parietal cortex is involved in the understanding of 3D space and action planning. Research findings indicate that the two streams process information both independently and in parallel, utilizing feedback loops and sharing information through neural connections that exist in multiple stages [4, 7]. These identified interconnections enable the human brain to fuse sensory and motor information, so as to achieve robust cognition.

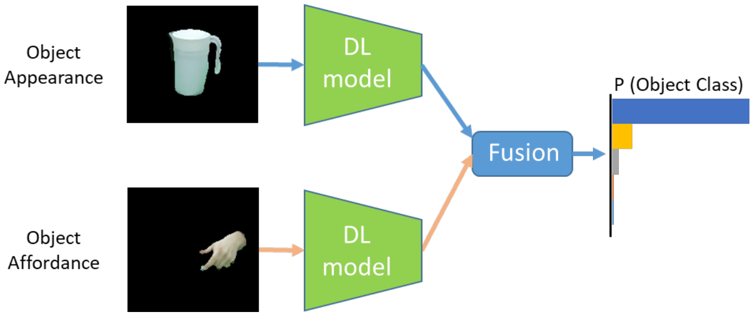


Figure 5-1: Schematic of the deep learning based architecture of the proposed sensorimotor 3D object recognition framework. Following fusion of the object appearance and affordance processing streams, the object class is predicted.

Motivated by the above facts, we investigate sensorimotor learning for RGB-D object recognition in the context of "function from motion". Further, inspired by the complex neural network of the human brain, we adopt the deep learning paradigm [52] to form two parallel information streams that process object appearance (sensory) and affordance (motor) information. These streams exploit deep learning architectures, primarily CNNs and RNNs, and are fused in multiple ways, in order to mimic the complex information exchange between the brain processing pathways. A schematic of our approach is depicted in Fig 5-1.

The main contribution of this chapter is therefore the deep sensorimotor learning approach for the RGB-D object recognition task. Specifically, three variants of the proposed two-stream sensorimotor modeling approach are considered that utilize different deep neural networks to encode the spatial-only or spatio-temporal correlations of suitable appearance and affordance input representations.

Additionally, inspired by the aforementioned neuro-scientific findings for the human brain complex information exchange at different levels of granularity, fusion at one or multiple layers of each model variant is extensively investigated. Regarding spatio-temporal information processing, the incorporation of an attention mechanism is also proposed, which forces the model to selectively attend to the affordance information, when the appearance one is not discriminative enough, as indicated by appropriate stream confidence measures.

Further, an auxiliary loss function is introduced, based solely on affordance predictions. The new loss is combined with the object prediction one, and the result is used to optimize both streams during training. In order to compute the auxiliary loss, a classifier is added after the last affordance stream layer, but later removed during inference.

Finally, an extensive quantitative evaluation of the proposed models is presented, using the challenging SOR3D corpus (see Chapter 4) that includes a significantly increased number of affordances compared to existing works in the literature (see Section 4.4 and Table 4.2). Besides comparison of the two-stream models with the appearance-only baseline, the best performing one is further benchmarked against traditional probabilistic fusion approaches. The evaluation is concluded with a cross-view analysis, providing valuable insights about how view-dependent is the affordance information and how each viewpoint affects model performance.

5.2 Visual Front-end and Single-stream Models of Appearance and Affordance

In this section, the preprocessing framework, as well as the appearance and affordance input representations are detailed. Additionally, three single-stream models capable of encoding either spatial-only or spatio-temporal information are presented.

5.2.1 Input Streams and Preprocessing

The SOR3D data preprocessing begins with the RGB and depthmap frame alignment, based on the Kinect intrinsic parameters. Then, the region that includes the hand-object interaction is defined and a centered rectangular region (300×300 pixels) is cropped. Subsequently, using a simple thresholding method in the HSV color space [108], the background is removed, and the skin color pixels (*i.e.* those corresponding to the hand region) are separated from the object ones. As a result, the hand and object RGB and depthmap frames are provided separately, as depicted in Fig. 5-2. Note that the hand-object separation leads to the two types of information

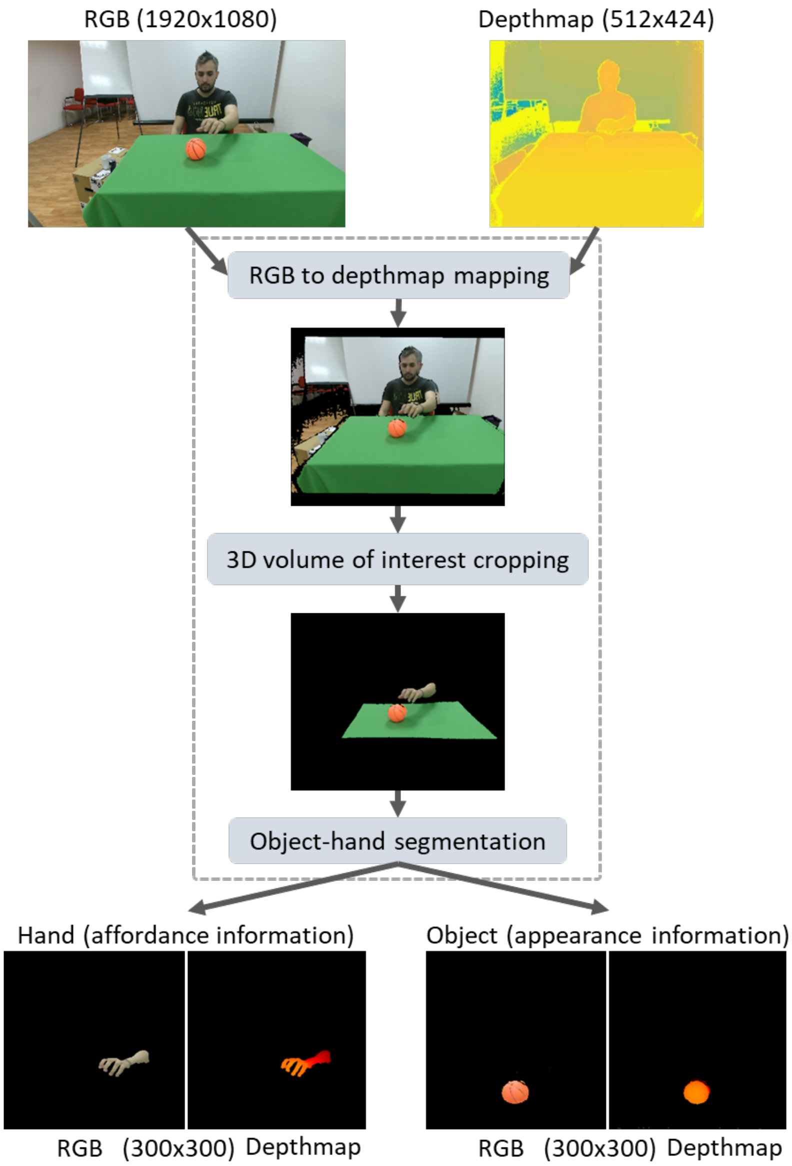


Figure 5-2: Preprocessing overview. The captured RGB and depth raw data (top) are initially aligned, the 3D volume of interest is cropped (middle), and the hand and object RGB and depth representations are separated (bottom).

utilized in this thesis, namely the object appearance that is related to the object shape, color, and texture, and the object affordance that is related to the hand movement. This process aims to remove information that is not relevant to the interaction (*e.g.* background, tablecloth, etc.), in order to investigate the true added value of the affordance information. Note also that as this database is collected in a controlled lab environment (*i.e.* illumination, green tablecloth, no long sleeves), traditional segmentation approaches are very accurate, thus a more sophisticated semantic segmentation algorithm would offer no substantial gains.

Due to the low object intra-class variance, we choose to ignore the RGB information and instead encode the depth information using two different approaches. In the first one, we adopt the depth encoding algorithm introduced by [24], which relies on computing three depth-based features, namely the Horizontal disparity, the Height above the ground, and the Angle between the surface normals and the gravity direction of the captured scene (HHA). The three computed features are stacked to form a 3-channel representation that has the same width and height as the original depthmap. The second depth encoding approach is depthmap "colorization". Motivated by [14], depth colorization is performed by normalizing all depth values in the interval $[0, 255]$ and then mapping each pixel distance to color values ranging from red (near) to yellow (far), transforming the one-channel depthmap to a three-channel color image. Note that the aforementioned approaches also enable the exploitation of transfer learning by using deep learning models pre-trained on large-scale image datasets [72, 104, 116].

Besides depth encoding, we further process the original depthmaps and RGB frames in order to compute the 3D optical flow of the interaction (relating to object affordance). Due to the development of affordable RGB-D sensors, several 3D flow computation methods have been proposed in the literature [25, 33, 81]. Here, we utilize the primal-dual algorithm proposed in [35] due to its efficiency. In detail, the 3D motion vectors between two pairs of RGB-D images, as well as their magnitude are first computed. The 3D flow and its magnitude are then colorized by normalizing each axis values in the interval $[0, 255]$, transforming the 3D motion vectors into a three-channel image. We further choose to encode the 3D flow sequence into a single motion map, by accumulating the flow over the entire sequence, as such representations can be very informative [109].

To summarize, the information streams that are utilized as input to the proposed models are: a) HHA encoding, b) colorized depthmaps (CDM), c) colorized 3D op-

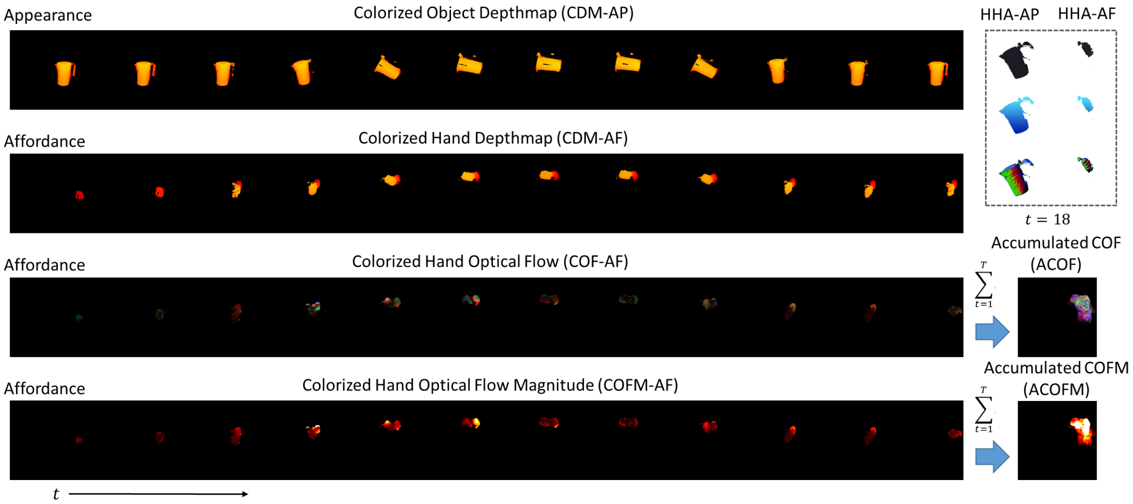

Appearance
Colorized Object Depthmap (CDM-AP)
HHA-AP
HHA-AF
Affordance
Colorized Hand Depthmap (CDM-AF)
t = 18
Accumulated COF (ACOF)
Affordance
Colorized Hand Optical Flow (COF-AF)
$\sum_{t=1}^{T}$
Accumulated COFM (ACOFM)
Affordance
Colorized Hand Optical Flow Magnitude (COFM-AF)
$\sum_{t=1}^{T}$
t

Figure 5-3: Example video session "pour from pitcher" from the SOR3D corpus, sampled every 4 frames. The object appearance is depicted as colorized depthmaps and HHA encoding (only for an example frame, top-down: disparity, height, normals channels are shown), while the affordance information is depicted as colorized depthmaps, HHA encoding (same example frame and top-down presentation as HHA-AP), 3D optical flow, and 3D optical flow magnitude, as well as the accumulation of the latter two over the sequence of T frames.

tical flow (COF) along with the accumulated colorized 3D optical flow (ACOF), and d) colorized 3D optical flow magnitude (COFM), coupled with the corresponding accumulated one (ACOFM). Fig. 5-3 depicts an example of two appearance and six affordance input representations of a "pour from pitcher" session. For the sake of clarity, the information stream that processes the object appearance is denoted as the "appearance stream", while the one that processes the hand-object interaction is denoted as the "affordance stream". Additionally, the notation {AP, AF} is used to state that a specific input is received from the appearance or the affordance stream (*e.g.* CDM-AP denotes that the appearance stream receives colorized depthmaps as input).

5.2.2 Single-stream Models

For the appearance and affordance information processing, three single-stream models are proposed, as detailed next.

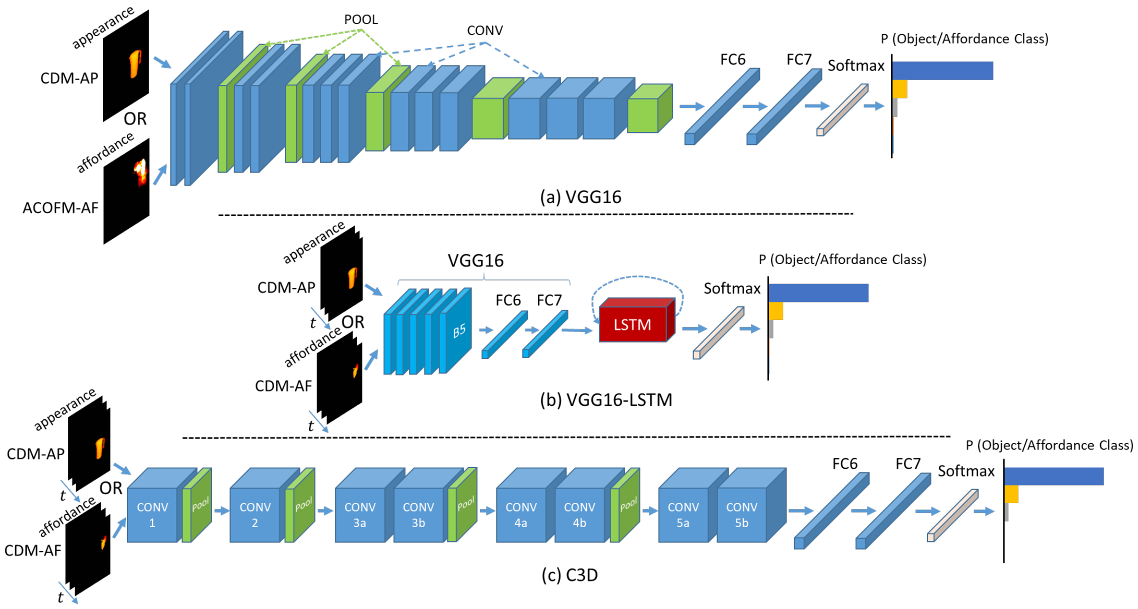


Figure 5-4: Detailed architecture of the adopted single-stream models: a) VGG16 that is capable of encoding spatial information; b) VGG16-LSTM that utilizes a VGG16 and an LSTM to encode spatio-temporal information; and c) C3D that exploits 3D convolutions to encode spatio-temporal information. The CDM-AP or HHA-AP can be used as input appearance representations, while various affordance input representations of Fig. 5-3 can be used, as evaluated in Table 5.1 of Section 5.4.1.

The first model, depicted in Fig. 5-4(a), is the VGG16 model, which encodes the spatial-only information of an input image. Note that this model can efficiently learn complex spatial feature representations and has been widely used for visual recognition purposes. The second model is capable of encoding both spatial and temporal information, by processing sequences of 2D frames. As shown in Fig. 5-4(b), the model consists of a VGG16 model followed by an LSTM one. Finally, as a third model we use the C3D network, depicted in Fig. 5-4(c), which is capable of encoding spatio-temporal information.

The aforementioned models are separately trained for object and affordance recognition, using the 14 object and 13 affordance classes as ground truth, respectively. Note that all models use a Softmax layer for class prediction. Additionally, HHA encoding and CDM are used as input representations of object appearance, whereas all six affordance input representations reported in Section 5.2.1 are utilized to investigate their impact on RGB-D object recognition.

Regarding the VGG16 model, which predicts classes for individual images, the video-level object prediction is obtained by averaging all frame-level predictions. However, this process is not effective for affordance recognition. Intuitively, object affordances are explicitly described by hand motion, which is time-evolving. Thus, using the VGG16 model to predict the affordance class of a sequence would be inconsistent. This intuition is confirmed in Section 5.4.1 (see also Table 5.1), hence for most of the Thesis we utilize only the ACOF and ACOFM representations as input to the affordance VGG16, as they summarize the entire motion of the sequence accumulated within a single frame.

5.3 Two-stream Models Fusing Appearance and Affordance

Motivated by the two-stream hypothesis of the human brain sensorimotor learning process, the aforementioned single-stream models are fused in multiple ways in order to achieve robust object recognition. Three sensorimotor models are presented, where the appearance and affordance information exchange between the two streams is extensively investigated.

5.3.1 Spatial-only Model (SP)

The two-stream spatial-only model (SP), depicted in Fig. 5-5, utilizes two VGG16 networks, one for appearance and the other for affordance information processing. The appearance VGG16 receives HHA-AP or CDM-AP input, while the affordance one processes either ACOF-AF or ACOFM-AF representations, similarly to the single-stream affordance VGG16 described in Section 5.2.2. Three fusion schemes of the two streams are investigated: a) late fusion at the FC layer level (SP$_{L-FC}$); b) late fusion at the CONV layer level (SP$_{L-CONV}$); and c) multi-layer fusion (SP$_{ML}$) that combines the aforementioned approaches.

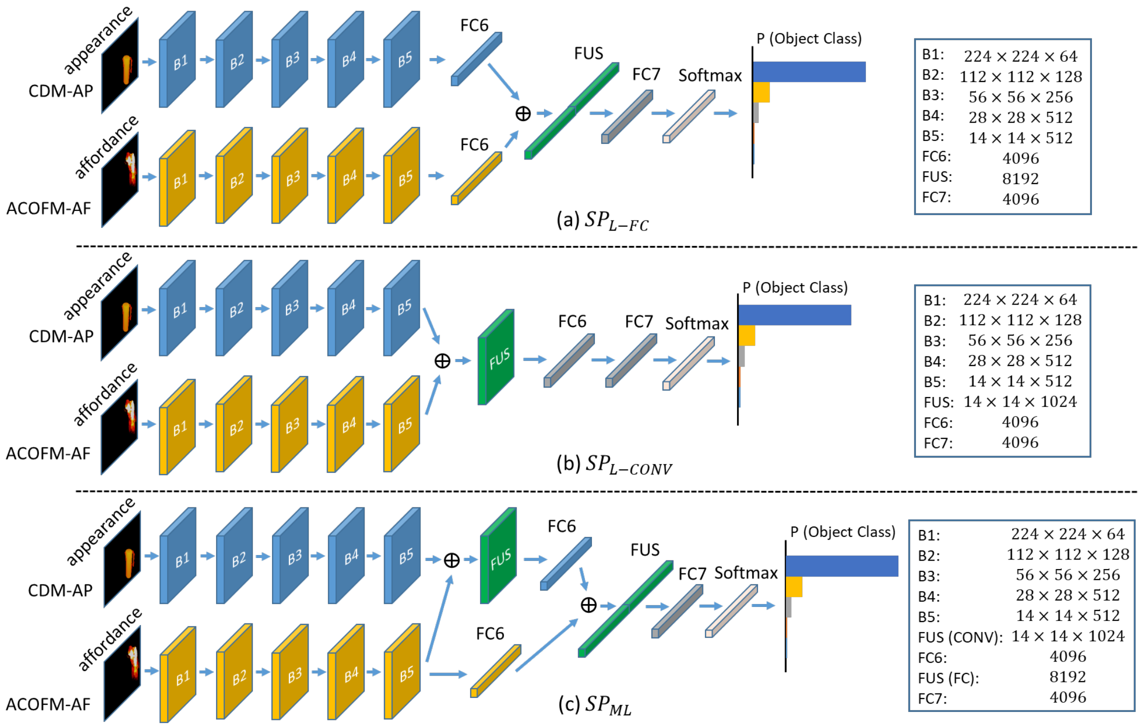


Figure 5-5: Detailed architecture of the SP model for: a) late fusion at the FC layer; b) late fusion at the CONV layer; and c) multi-layer fusion. Each block (B1-B5) corresponds to a CONV-RL-POOL sequence of VGG16, while FUS indicates feature fusion (concatenation). At the right side of each fusion scheme, the dimensionality of the activation matrix for each CONV block is reported as "*height* × *width* × *channels*" and for each FC layer as the number of neurons. The CDM-AP or HHA-AP representations can be used as input to the appearance stream, whereas the ACOF-AF and ACOFM-AF as input to the affordance stream.

Late fusion at the FC layer level (Fig. 5-5(a)) is realized by concatenating the activations of FC6 (*i.e.* the sixth VGG16 layer, which is a FC one) of each stream, after the RL non-linearity. After fusion, a single stream of a 4096-dimensional (dim) FC layer and a Softmax layer is formed.

Regarding late fusion at the CONV layer level (Fig. 5-5(b)), the activation maps after RL5 (non-linearity of CONV5) are concatenated along the channel dimension. In more detail, if $X_s^{h \times w \times d}$ represents each activation matrix, where $s \in \{\text{AP}, \text{AF}, \text{FUS}\}$ and h, w, d correspond to the height, width, and number of channels, then $X_{AP}^{14 \times 14 \times 512}$, $X_{AF}^{14 \times 14 \times 512}$ are the inputs and $X_{FUS}^{14 \times 14 \times 1024}$ is the output of the fusion. The latter is further convolved with 512 filters of 1×1 size and downsampled using a max-pooling layer (2×2 size), thus resulting in a $X_{FUS}^{7\times7\times512}$ activation matrix. Similarly to the FC layer late fusion, a single processing stream is formed that consists of 2 FC layers (4096-dim) and a Softmax layer.

Finally, in order to allow more complex information exchange at different levels of granularity between the two streams, a multi-layer fusion scheme is also investigated (Fig. 5-5(c)). In particular, the two streams are initially fused after the last CONV layer (RL5) and then fused again after FC6 (RL6). The appearance FC6 layer receives as input the fused activations, while the affordance FC6 receives the activations from POOL5 (POOL layer after CONV5) of the affordance stream only. Subsequently, the activations after RL6 of both streams are concatenated forming a 8192-dim feature, followed by a 4096-dim FC layer and a Softmax layer. Note that, in the multi-layer fusion case only, the weights of the affordance B1-B5 layers are not updated from the gradients computed at the CONV fusion level during backpropagation. In that way, the affordance stream contributes to the appearance one in multiple levels, without being particularly affected by the appearance information.

For the video-level object class prediction, the object probabilities for each frame of the sequence are averaged. Note that the affordance input representation remains unaltered, as it includes the aggregated information of the entire sequence.

5.3.2 Spatio-temporal 2D Model (ST2D)

Another approach for modeling the dynamic nature of the affordance information is realized with the proposed two-stream Spatio-Temporal 2D model (ST2D). As shown in Fig. 5-6, we adopt the VGG16-LSTM structure to model the spatio-temporal nature of the hand-object interaction. Two fusion approaches are considered, namely intermediate (ST2D$_{IM}$) and late fusion (ST2D$_{L}$).

Regarding the ST2D$_{IM}$ model, the 4096-dim spatial feature vectors extracted by each VGG16 model (*i.e.* the activations after the RL7 layer) are concatenated and then processed by the LSTM at every time instant, namely at every frame of the input sequence. The LSTM encodes the temporal correlations of the interaction, while its

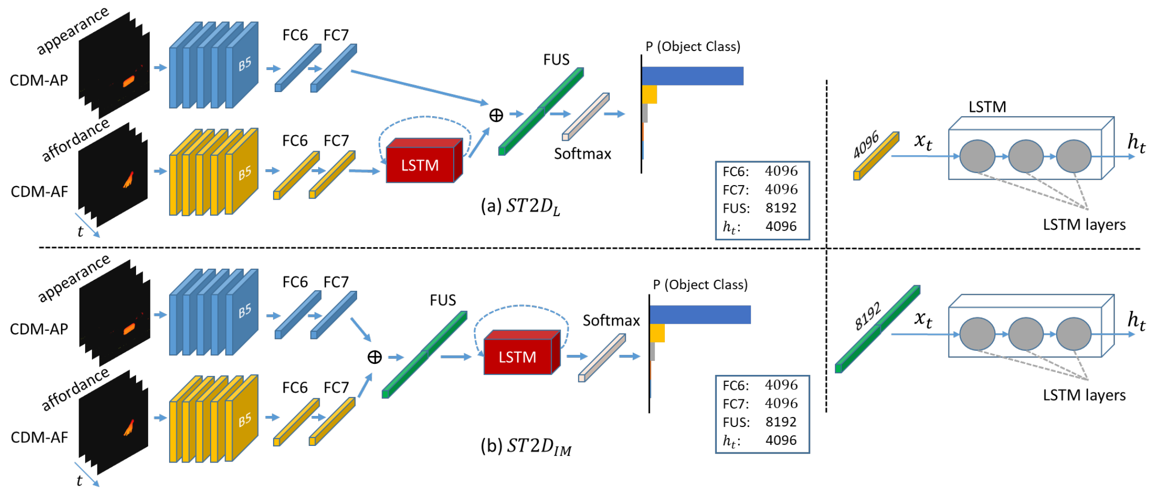


Figure 5-6: Detailed architecture of the ST2D model for: a) late fusion and b) intermediate fusion. Blocks B1-B5, FC6, and FC7 correspond to the VGG16 network, while FUS indicates feature fusion. At the right side, the internal structure of the LSTM, as well as the input x and output h vector dimensionality for each fusion scheme are depicted. The appearance stream processes HHA-AP or CDM-AP inputs, while the HHA-AF, CDM-AF, COF-AF, and COFM-AF representations can be used as input to the affordance stream.

internal state vector $[h(t)]$ (4096-dim) is further processed by a Softmax layer for the object class prediction.

On the other hand, ST2D$_L$ adopts the VGG16-LSTM structure only for the affordance information processing, since the appearance VGG16-LSTM performs significantly worse than VGG16 as a single-stream object classifier (see Section 5.4.1 and Table 5.1). Thus, for this model, the RL7 activations of the appearance stream (4096-dim) are concatenated with the internal state vector $[h(t)]$ (4096-dim) of the affordance VGG16-LSTM at every time instant. The outcome of the concatenation is further processed by a Softmax layer.

Regarding the final prediction, two approaches are investigated for both fusion schemes. These approaches aggregate the frame-level prediction to yield a video-level classification decision. Given a series of frame-level posteriors $p_{t,c}$, where $t = 1, \ldots, T$ is the frame number and $c = 1, \ldots, C$ the object class, the video-level classification decision $\hat{c}$ is given either by:

$$\hat{c}_{avg} = \arg\max_{c} \frac{1}{T} \sum_{t=1}^{T} p_{t,c} \quad , \tag{5.1}$$

employing the averaging approach, or by:

$$\hat{c}_{w} = \arg\max_{c} \frac{1}{T} \sum_{t=1}^{T} t \, p_{t,c} \quad , \tag{5.2}$$

using the weighting approach, respectively. Clearly, (5.1) indicates that all frame-level predictions contribute equally to the video-level one. However, the LSTM weights are updated after every processed frame, thus the affordance features prior to fusion should be more discriminative at the end of the sequence. Thus, we utilize (5.2) to force the model to focus more on the frame-level predictions over the last video frames.

5.3.3 ST2D Model with Attention

The proposed attention mechanism is based on the appearance stream confidence. As depicted in Fig. 5-7 (green box), a Softmax layer is added after the last FC layer of the appearance CNN, which predicts the label of the object for each frame. The new layer is followed by a module that measures the appearance-based classifier confidence for the entire frame sequence. The output of the latter is used to selectively attend to the affordance features extracted by the affordance CNN-LSTM stream, prior to the fusion Multi-Layer Perceptron (MLP).

In order to measure the appearance classifier confidence, we investigate three different metrics. Let $c_{t,n}, n = 1, \ldots, N$ be the ranked $N-$best object class predictions of the appearance CNN classifier, $\mathcal{C}$ the number of the object classes, and $p_{t,n} = Pr(c_{t,n}|x_t)$ the probability distribution after the Softmax given the appearance feature vector x_t at frame t. As the first metric, the entropy $\mathcal{I}_{t,E}$ is computed for the probability distribution as:

$$\mathcal{I}_{t,E} = -\sum_{n=1}^{\mathcal{C}} p_{t,n} \log{(p_{t,n})}. \tag{5.3}$$

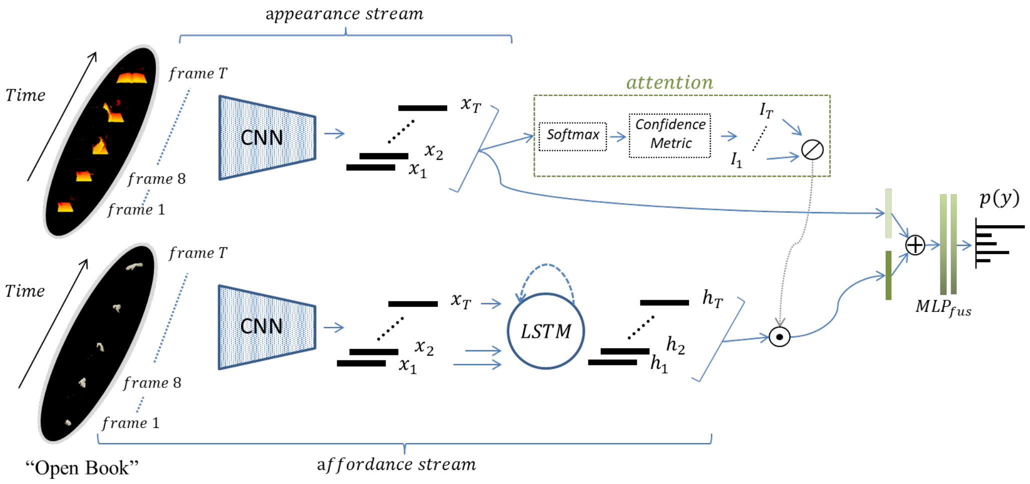


Figure 5-7: Detailed architecture of the proposed spatio-temporal late fusion model. The green box includes the attention mechanism modules attached to the final FC layer of the appearance CNN (top), which selectively attends to the affordance CNN-LSTM output (bottom). The feature fusion MLP follows (right-most), while ⊘, ⊙, and ⊕ represent normalization, frame-level multiplication, and concatenation.

Clearly, $\mathcal{I}_{t,E}$ values that are close to zero indicate strong confidence, while larger values indicate difficulty in discrimination. The second investigated metric is the average $N-$best log-likelihood difference, computed as:

$$\mathcal{I}_{t,A} = \frac{1}{N-1}\sum_{n=2}^{N}(\log{(p_{t,1})} - \log{(p_{t,n})}), \tag{5.4}$$

where $N \geq 2$. In contrast to the entropy metric, larger values of $\mathcal{I}_{t,A}$ indicate high-confidence predictions. The last metric measures the log-likelihood dispersion among the $N-$best class predictions, and is given by:

$$\mathcal{I}_{t,D} = \frac{2}{N(N-1)}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}(\log{(p_{t,n})} - \log{(p_{t,m})}), \tag{5.5}$$

where $N \geq 2$. Similarly to (2), larger $\mathcal{I}_{t,D}$ values indicate high classification confidence. It must be noted that the presented metrics have been also used in the context of audio-visual speech recognition [1, 78]. Following the appearance classifier confidence measurement, the $\mathcal{I}_t$ values of all frames are normalized to $[0, 1]$ by:

$$w_t = \frac{\mathcal{I}_t - \mathcal{I}_{min}}{\mathcal{I}_{max} - \mathcal{I}_{min}}, \tag{5.6}$$

where $\mathcal{I}_{min}$, $\mathcal{I}_{max}$ are calculated over the entire frame sequence, and $w \in [0, 1]$ is the video confidence vector. The last step of the mechanism is given by:

$$\hat{H} = \begin{cases} w \odot H & \text{if (5.3)} \\ (1 - w) \odot H & \text{if (5.4) or (5.5)} \end{cases}$$

where $\odot$ indicates the frame-level multiplication of confidence values with the LSTM output matrix $H^{T \times M}$. Notice that by multiplying w_t with the corresponding h_t, the mechanism alters the impact of the affordance information on the final prediction, since $\hat{H}^{T \times M}$ is fused with the appearance features as:

$$\hat{p}_t = \text{softmax}(\phi(\text{concat}(x_t, \hat{h}_t))), \tag{5.7}$$

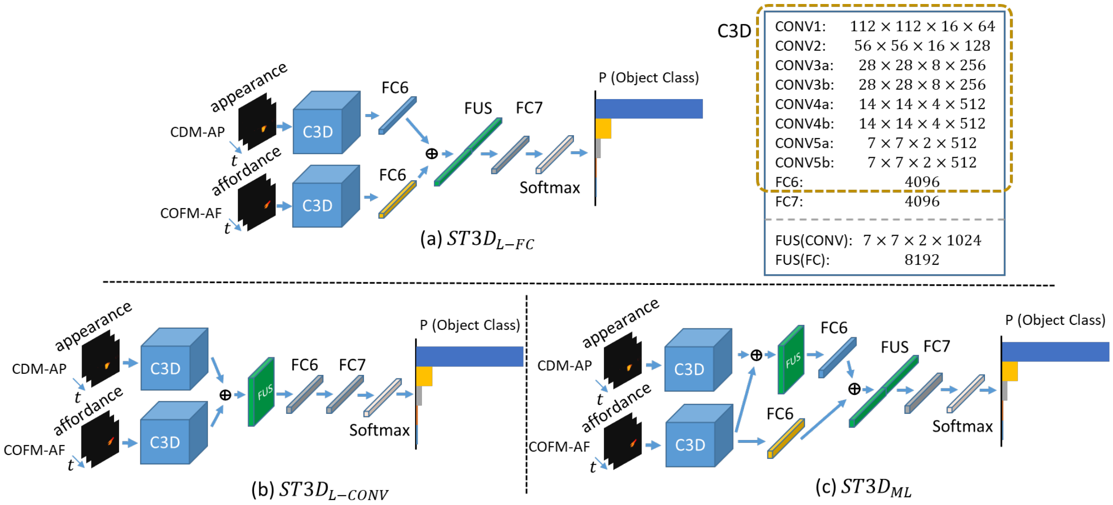
where $x_t \in X^{T \times F}$ denotes the appearance feature vector, ϕ is the fusion MLP followed by a Softmax function, and $\hat{p}_t$ is the probability distribution of the attention-based ST model (Fig. 5-7) for the $t-$th frame.

Regarding the final prediction, two approaches are investigated. Both aggregate the frame-level prediction of the attention-based ST model to yield a video-level decision for the object label. Given a series of frame-level predictions $\hat{p}_{1,c}, \ldots, \hat{p}_{t,c}, \ldots, \hat{p}_{T,c}$ from (5.7), the video-level classification decision y is given either by:

$$y_{avg} = \arg\max_c \frac{1}{T} \sum_{t=1}^{T} \hat{p}_{t,c}, \tag{5.8}$$

as the averaging approach, or by:

$$y_w = \arg\max_c \frac{1}{T} \sum_{t=1}^{T} t \, \hat{p}_{t,c}, \tag{5.9}$$



C3D

CONV1:	$112 \times 112 \times 16 \times 64$
CONV2:	$56 \times 56 \times 16 \times 128$
CONV3a:	$28 \times 28 \times 8 \times 256$
CONV3b:	$28 \times 28 \times 8 \times 256$
CONV4a:	$14 \times 14 \times 4 \times 512$
CONV4b:	$14 \times 14 \times 4 \times 512$
CONV5a:	$7 \times 7 \times 2 \times 512$
CONV5b:	$7 \times 7 \times 2 \times 512$
FC6:	4096
FC7:	4096
FUS(CONV):	$7 \times 7 \times 2 \times 1024$
FUS(FC):	8192

Figure 5-8: Detailed architecture of the ST3D model for: a) late fusion at the FC layer; b) late fusion at the CONV layer; and c) multi-layer fusion. At the upper right side the dimensionality of the activation matrix for each CONV block is reported as "*height* × *width* × *channels*" and for each FC layer as the number of neurons. The appearance stream processes HHA-AP or CDM-AP inputs, while the HHA-AF, CDM-AF, COF-AF, and COFM-AF representations can be used as input to the affordance stream.

as the weighting approach, respectively. Clearly, the latter forces the model to focus more on the frame-level predictions over the last frames of the video, while the former treats all frame-level predictions equally.

5.3.4 Spatio-temporal 3D Model (ST3D)

An alternative approach for modeling the spatio-temporal nature of time-evolving interactions is by incorporating the 3D CNN structure in a two-stream model. The two-stream Spatio-Temporal 3D model (ST3D) consists of two C3D ones, one for appearance and the other for affordance information processing. Note that we choose to process the appearance information using a C3D instead of a VGG16 model, as we observed that despite its slightly inferior performance as single-stream classifier for object recognition (see Section 5.4.1 and Table 5.1), it performs better when it is combined with the affordance C3D. Additionally, since its structure is very similar to VGG16, we investigate the same three fusion schemes as for the SP model (*i.e.* ST3D$_{L-FC}$, ST3D$_{L-CONV}$, ST3D$_{ML}$). The aforementioned fusion schemes are depicted in Fig. 5-8.

Note that, unlike ST2D, the C3D models used as appearance and affordance streams can selectively attend to both appearance and motion information. To support this hypothesis, [105] use the deconvolution method proposed in [119] to visualize the patterns learned by the C3D weights over video samples. They report that based on observations, the C3D starts by focusing on appearance in the first few frames and tracks the salient motion in the subsequent ones. Thus, unlike ST2D, no extra attention mechanism is incorporated to the model.

5.3.5 Auxiliary Loss Function

The training objective for the proposed two-stream models is to minimize the cross-entropy loss between the predicted object class and the ground truth. This loss is used to compute the gradients and update the weights of both streams. However, besides incorporating affordance information to improve object class prediction, further optimization of the models weights using an auxiliary loss function based solely on the affordance stream performance can be beneficial. In order to compute the auxiliary loss, the affordance features prior to fusion are used to train a Softmax classifier. The training objective of the new classifier is to minimize the cross-entropy loss between the predicted affordance class and the affordance ground truth. The two loss functions can be combined and used to optimize both streams. This aggregated loss is computed as:

$$\mathcal{L}_{agg} = -\frac{1}{K}\sum_{k=1}^{K}\Big(y_{o,k}\log(p_{o,k}) + y_{a,k}\log(p_{a,k})\Big) \ , \tag{5.10}$$

where K is the total number of training samples, $y_{o,k}$ and $p_{o,k}$ are the object ground truth and predicted probability, and $y_{a,k}$ and $p_{a,k}$ are the affordance ground truth and predicted probability of sample k. The auxiliary loss can be applied to the affordance stream of any two-stream model, except for the ST2D$_{IM}$ where the two streams are fused before the LSTM. Fig. 5-9 depicts an example of the auxiliary loss applied to

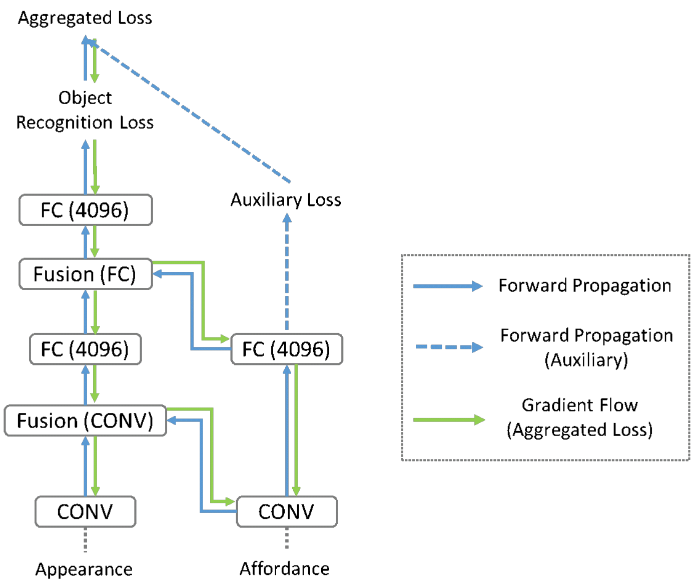


Figure 5-9: Example of the auxiliary loss application on the SP$_{ML}$ model. The auxiliary loss is computed based on the affordance classifier and then combined with the object recognition loss.

the SP$_{ML}$ model.

5.4 Experimental Results

The presented single-stream and fusion models were evaluated using the SOR3D dataset for the task of object recognition. The data captured from the three viewpoints (see also Fig. 4-1) were accumulated into a unified (*i.e.* all-viewpoint) dataset, which was then split into training, validation, and test sets (25%, 25% and 50%) that correspond to approximately 5k, 5k and 10k hand-object interaction videos, respectively, as also discussed in Section 4.2. For all 300×300 pixel extracted video frames, a 224×224 patch was randomly cropped and used as input to the models. All models were trained with the negative log-likelihood criterion, whereas for backpropagation, Stochastic Gradient Descent with 0.9 momentum was used. The standalone VGG16 network was pre-trained on ImageNet [10], while the VGG16-LSTM and the C3D were pre-trained on Sports-1M [42]. Subsequently, all models were fine-tuned on the SOR3D dataset with learning rate set to 5×10^{-3}, decreased by a factor of 5×10^{-1}

Table 5.1: Recognition accuracy of the three single-stream models of Section 5.2.2 on the test set of the SOR3D database for various appearance and affordance input stream representations. Object recognition accuracy (%) is reported in the upper part of the table (appearance stream) and affordance recognition accuracy (%) in the lower part (affordance stream).

Input Stream	VGG16	VGG16-LSTM	C3D
HHA-AP	84.98	73.96	84.45
CDM-AP	85.12	74.33	84.67
HHA-AF	56.89	67.44	79.12
CDM-AF	57.28	69.27	81.44
COF-AF	58.32	68.02	82.68
COFM-AF	58.49	68.85	83.19
ACOF-AF	80.84	n/a	n/a
ACOFM-AF	81.92	n/a	n/a

when the validation accuracy curve plateaued. For fusion models training, $L2$ regularization [64] was incorporated in order to prevent over-fitting. All experiments are conducted on 2 Nvidia Titan X GPUs.

5.4.1 Single-stream Model Evaluation

The first set of experiments deals with the evaluation of the single-stream models presented in Section 5.2.2. The results are reported in Table 5.1 in terms of overall object and affordance recognition accuracy. For each video sequence, a set of 20 uniformly selected video frames was provided to each single-stream model, while due to computational and memory restrictions the input sequence length for the C3D model was set to 8 frames. For the latter, a sliding window of 8 frames was applied to each sequence and the window-level predictions were averaged to provide a video-level one. The aforementioned setup was used during both training and testing. The frame-level predictions of the VGG16 were also averaged to provide a single prediction for each video.

Regarding object recognition, VGG16 yielded the best overall accuracy compared to the VGG16-LSTM and C3D models for both CDM-AP and HHA-AP input representations. From the aforementioned representations, CDM-AP performed slightly better than HHA-AP (*i.e.* 85.12% over 84.98%), mainly due to the nature of the captured data, *i.e.* height and disparity are more informative in outdoor scenes, or indoor ones that consist of large objects (*e.g.* furniture). Based on the reported results, the VGG16 model that processes CDM-AP input representation was considered as the appearance-only baseline for the rest of the experiments. Further, due to the CDM-AP superiority over HHA-AP, the former was considered as appearance input representation for all two-stream models evaluation.

In order to truly understand the impact of the affordance information on object recognition, we firstly evaluated the affordance encoding efficiency of each single-stream model. For this experiment, only the affordance information was utilized providing the target labels, with the experimental framework remaining unaltered. However, the last layer of each network was restructured to predict probabilities based on the 13 affordance classes. From the results reported in Table 5.1, we can conclude that when processing individual frames (*i.e.* HHA-AF, CDM-AF, COF-AF, and COFM-AF) the VGG16 model cannot implicitly capture the temporal information of the affordance. Additionally, we observe that the VGG16-LSTM model cannot efficiently encode the temporal correlations of the sequence, mainly due to the short and fine-grained interaction. On the other hand, the C3D model yields satisfactory results for all affordance input representations, while the VGG16 one performs considerably well when using accumulated 3D flow as input.

5.4.2 Two-stream Model Evaluation

In this section, the fusion models evaluation is detailed. It should be noted that for all fusion model experiments the appearance stream receives CDM input (CDM-AP), as discussed in Section 5.2.1. Thus, the appearance input is not reported in Tables 5.2-5.6 for simplicity.

Table 5.2: Object recognition results (in accuracy, %) on the SOR3D test set, using different SP-based fusion and training schemes and affordance inputs (in conjunction with CDM-AP input).

Input Stream (Regularization)	SP_{L-FC}	SP_{L-CONV}	SP_{ML}
ACOF-AF	87.03	87.93	89.10
ACOFM-AF	87.40	88.24	89.43
ACOFM-AF (aux. loss)	88.37	89.63	90.79
ACOFM-AF ($L2$)	87.92	88.55	89.95
ACOFM-AF (aux. loss, $L2$)	88.54	89.81	91.12

5.4.3 SP Model Evaluation

Table 5.2 shows the performance of the SP model, in terms of object recognition accuracy. From the presented results, it can be seen that using the ACOFM input representation is advantageous compared to the ACOF one. Thus for the rest of the SP model evaluation, the former representation is utilized. Further, the late fusion of CONV features (*i.e.* fusion after RL5) appears to perform better compared to the late fusion at the FC layer level. Note that at the FC layers the spatial information is lost, thus fusing CONV layer activations leads to more discriminative post-fusion features. Interestingly, SP_{ML} outperforms the aforementioned late fusion schemes. Using this fusion approach, the model learns both mid-level and high-level feature representations, without loosing the spatial correspondence due to the feature-flattening at the FC layers.

Additionally, significant performance improvement can be observed when the auxiliary loss (see Section 5.3.5) is incorporated. This result reflects the importance of the affordance modeling optimization in parallel with the overall object recognition task. Further, regularization with the $L2$ norm leads to higher accuracy. In fact, the SP_{ML} model trained using the auxiliary loss and the $L2$ norm outperforms the appearance-only VGG16 by an absolute 6%.

Fig. 5-10(b) visualizes the confusion matrix of the best performing SP_{ML} model (ACOFM-AF, aux. loss, $L2$) on the SOR3D test set. It can be observed that

Table 5.3: Object recognition results using the ST2D$_{IM}$ and ST2D$_L$ models in conjunction with the averaging (AVG) and weighting (W) video-level prediction approaches for various affordance input representations and CDM-AP input.

Input Stream	ST2D$_{IM}$-AVG	ST2D$_{IM}$-W	ST2D$_L$-AVG	ST2D$_L$-W
HHA-AF	79.33	80.17	86.12	86.53
CDM-AF	79.65	80.43	86.50	86.87
COF-AF	78.98	79.94	86.30	86.64
COFM-AF	79.08	80.04	86.38	86.72

this fusion scheme boosts recognition performance of all supported objects over the appearance-only VGG16 with CDM-AP input (see Fig. 5-10(a)), demonstrating the additional discriminative power of affordance information.

5.4.4 ST2D Model Evaluation

Experimental results of the ST2D-based fusion evaluation are reported in Tables 5.3-5.5. In all cases, as in Section 5.4.1, a set of 20 uniformly selected frames was provided as input to the respective networks.

Table 5.3 reports the comparative evaluation of the averaging and weighting video-level prediction for the ST2D$_L$ and ST2D$_{IM}$ models. It can be observed that, for both fusion schemes and all affordance input representations, the weighting approach leads to better overall accuracy than the averaging one. Thus, for the rest of the ST2D experiments reported in Tables 5.4 and 5.5, the weighting video-level prediction is adopted.

Table 5.4 shows that the ST2D$_{IM}$ performs worse than the appearance-only VGG16 (*i.e.* 80.43% over 85.12%). Thus, we conclude that the LSTM cannot efficiently encode the time-evolving object manipulation, using a sequence of fused representations as input. Note that the latter is the result of fusing the two information streams at the FC layer-level, where the spatial correspondence is lost; thus, the LSTM has difficulty learning temporal correlations for both appearance and affordance. In contrast, the ST2D$_L$ fusion scheme outperforms the appearance-only VGG16 for all affordance

Table 5.4: Object recognition results using different ST2D-based fusion and training schemes and affordance inputs (in conjunction with CDM-AP input).

Input Stream (Regularization)	ST2D$_{IM}$-W	ST2D$_L$-W	ST2D$_L$-W (attention)
HHA-AF	80.17	86.53	89.14
CDM-AF	80.43	86.87	89.84
COF-AF	79.94	86.64	89.91
COFM-AF	80.04	86.72	90.02
COFM-AF (aux. loss)	n/a	86.86	90.18
COFM-AF ($L2$)	80.42	86.78	90.09
COFM-AF (aux. loss, $L2$)	n/a	86.95	90.31

input representations. In detail, the ST2D$_L$ scheme with HHA-AF input yields an absolute improvement of 1.41% compared to the appearance-only VGG16, which is further improved by CDM-AF, COF-AF, and COFM-AF to 1.75%, 1.52%, and 1.6% boosts, respectively.

The performance of ST2D$_L$ is further improved when the attention mechanism is incorporated. Based on Table 5.5, the $N-$best log-likelihood dispersion metric ($N = 3$) is selected as it yields the best overall accuracy. The inclusion of the attention mechanism leads to a performance boost for all affordance input representations (see right-most column of Table 5.4). Note also that the attention-based model using COFM-AF slightly outperforms the ones that use HHA-AF and CDM-AF as input representations. One plausible reason is that the 3D optical flow of the hand movement, prior to and after the interaction, may not contain significant affordance information, thus its impact to the final prediction should be small for the corresponding frames. The application of the auxiliary loss to the attention-based ST2D$_L$ model with COFM-AF input, in combination with $L2$ regularization, yields a 90.31% object recognition accuracy.

The confusion matrix of the attention-based ST2D$_L$ (COFM-AF, aux. loss, $L2$) model on the SOR3D test set is given in Fig. 5-10(c). It can be seen that it confuses objects that are very small or thin and their manipulation is very similar (*e.g.* small-

Table 5.5: Object recognition results of the ST2D$_L - W$ model with CDM-AP and CDM-AF inputs using attention in conjunction with the following confidence estimation metrics: a) the entropy, b) the average $N-$best log-likelihood difference ($N = 3$), and c) the $N-$best log-likelihood dispersion ($N = 3$).

Confidence Metric	Test Acc. (%)
Entropy	88.91
$N-$best difference	89.27
$N-$best dispersion	89.84

Table 5.6: Object recognition results using different ST3D-based fusion schemes and affordance inputs. CDM-AP is used as appearance input representation.

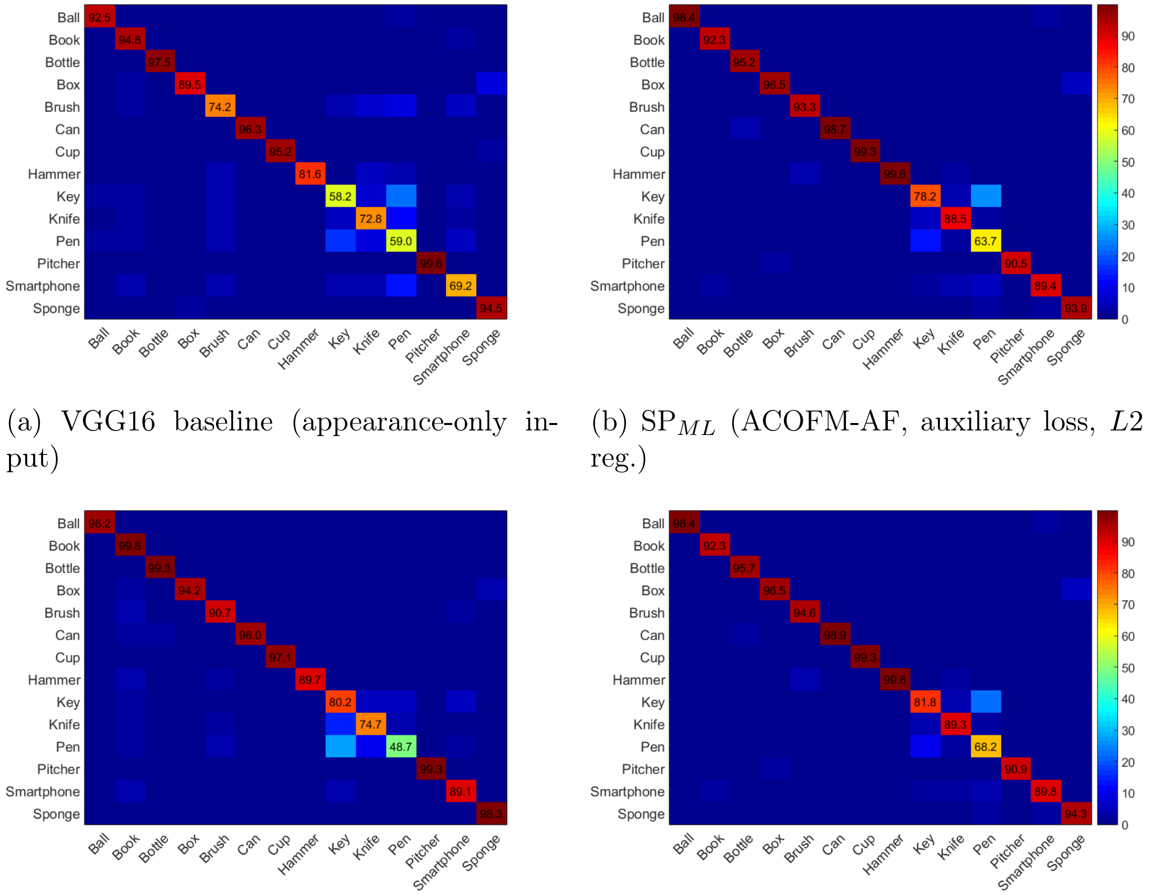
Input Stream (Regularization)	ST3D$_{L-FC}$	ST3D$_{L-CONV}$	ST3D$_{ML}$
HHA-AF	87.12	87.92	88.76
CDM-AF	87.97	88.32	89.23
COF-AF	88.06	88.65	89.79
COFM-AF	88.49	89.14	90.47
COFM-AF (aux. loss)	89.12	90.02	91.44
COFM-AF ($L2$)	88.86	89.58	90.70
COFM-AF (aux. loss, $L2$)	89.67	90.88	**91.98**

size ones, like “Key”, “Pen”, etc.).

5.4.5 ST3D Model Evaluation

For the ST3D model evaluation, we used sequences of 20 uniformly selected frames as input for each stream combined with an 8-frame sliding window, similarly to the single-stream C3D experiment (see Section 5.4.1). Note that in contrast to the LSTM learning process (see Section 5.3.2), the window-level predictions of the C3D model are independent from each other, thus for the final prediction the averaging approach was used.

Table 5.6 reports the overall accuracy of the ST3D models. Similarly to the SP



(a) VGG16 baseline (appearance-only input)

(b) SP$_{ML}$ (ACOFM-AF, auxiliary loss, $L2$ reg.)

(c) ST2D$_{L}$ (COFM-AF, auxiliary loss, $L2$ reg.) with attention

(d) ST3D$_{ML}$ (COFM-AF, auxiliary loss, $L2$ reg.)

Figure 5-10: Object recognition confusion matrices of the appearance-only VGG16 and the best performing fusion scheme of each two-stream model. Training parameters, such as affordance input and regularization of each model, are reported inside the parentheses. In all cases, CDM-AP is used as the appearance stream representation.

evaluation, ST3D$_{ML}$ outperforms ST3D$_{L}$ for all affordance inputs, due to the information sharing at different levels of granularity. Additionally, training both schemes with the auxiliary loss and $L2$ regularization leads to additional performance improvement. From the reported results, it can be observed that using 3D flow information instead of colorized depthmaps is advantageous. The latter is in accordance with the results presented in Table 5.1 for affordance recognition. Furthermore, it must be noted that ST3D$_{ML}$ with COFM-AF input, which is the best performing approach presented in this chapter, outperforms the appearance-only VGG16 by 6.86% (*i.e.* 91.98% over

Table 5.7: Comparative evaluation of the ST3D$_{ML}$ (CDM-AP, COFM-AF, aux. loss, $L2$) model, three probabilistic fusion methods, and the appearance-only VGG16 baseline. In all cases, object recognition accuracy (%) is reported.

Model	Fusion Layer	Test Acc. (%)
Appearance-only VGG16 baseline	no fusion	85.12
Product Rule	Softmax	77.91
SVM [44]	RL7	84.77
Bayes [31]	RL7	80.63
ST3D$_{ML}$	RL5, RL6	**91.98**

85.12%), which corresponds to an approximately 46% relative error reduction.

From a practical perspective, the ST3D model handles both the lack of temporal information modeling of the SP model and the difficulty of the ST2D one to learn the spatio-temporal correlations of fine-grained interactions. Additionally, it can better exploit 3D optical flow information, which explicitly describes the motion between sequential frames, thus making the recognition easier as the network does not need to estimate motion implicitly.

Finally, the confusion matrix of the ST3D$_{ML}$ model (COFM-AF, aux. loss, $L2$) on the SOR3D test set is depicted in Fig. 5-10(d). Notice that this model boosts recognition performance of all objects, while further improving it for the most challenging ones (*e.g.* “Key”, “Knife”, and “Pen”), by modeling the affordance information more efficiently.

5.4.6 Comparison with Probabilistic Fusion

The best performing fusion model, namely ST3D$_{ML}$ that utilizes COFM-AF as input and is trained with auxiliary loss and $L2$ regularization, is also comparatively evaluated against typical probabilistic fusion approaches of the literature. To perform a fair comparison, two C3D models are trained following the process presented in Section 5.4.1, using CDM-AP and COFM-AF as input representations. The product rule for fusing the appearance and the affordance C3D output probabilities is adopted

Table 5.8: Cross-view object recognition results using the appearance-only VGG16 and ST3D$_{ML}$ (COFM-AF, aux. loss, $L2$) models. The last row reports the results using the original SOR3D training and test sets.

Training Set	Test Set	VGG16	ST3D$_{ML}$
K_1	K_2, K_3	51.74	55.13
K_2	K_1, K_3	53.28	57.80
K_3	K_1, K_2	49.42	53.96
K_2, K_3	K_1	62.43	69.74
K_1, K_3	K_2	66.14	72.86
K_1, K_2	K_3	78.65	85.33
K_1, K_2, K_3	K_1, K_2, K_3	85.12	91.98

as the first probabilistic approach. Additionally, after removing both Softmax classifiers, the concatenated FC7 activations of the appearance and affordance C3D models are used to train a one-versus-all SVM classifier with RBF kernel [6, 44], as well as a naive Bayes classifier [31]. From the results presented in Table 5.7, it can be observed that the evaluated probabilistic fusion approaches fail to increase object recognition accuracy compared to the appearance-only VGG16 baseline. On the contrary, the proposed ST3D$_{ML}$ model exhibits a significant performance increase.

5.4.7 Cross-view Analysis

In this section, we perform a cross-view analysis on the SOR3D data, in order to evaluate the contribution of each viewpoint to the performance of the appearance-only VGG16 and the ST3D$_{ML}$ model. For this analysis, the three viewpoints of the SOR3D capturing setup, depicted in Fig. 4-1, are denoted as K_1, K_2, and K_3. For each $K_i, i \in V = \{1, 2, 3\}$, the evaluated model is initially trained using the K_i data and tested on the $K_{V-\{i\}}$ set, and then trained with $K_{V-\{i\}}$ data and tested on the K_i one. It must be noted that no viewpoint fusion is considered for any of the experiments. The appearance-only VGG16 employs the CDM-AP input representation, while the ST3D$_{ML}$ utilizes CDM-AP and COFM-AF inputs, auxiliary loss, and $L2$

regularization.

Intuitively, the affordance information should be significantly more viewpoint-dependent, since the starting point of the hand movement is different from each viewpoint and the actual interaction may not always be visible (*e.g.* the handle of the cup might be from the opposite side of the RGB-D sensor). From the results presented in Table 5.8, it can be observed that both models perform worse when trained on one or two viewpoints and tested on the rest. Additionally, it can be seen that, contrary to the aforementioned intuition, the starting point of the hand does not significantly affect the models performance, and the affordance information is discriminant even if some parts of the interaction are not entirely visible. We can further conclude that K_1 and K_2 are the most critical viewpoints for both appearance and affordance exploitation, as their absence from the training set leads to inferior overall classification accuracy.

5.5 Conclusion

In this chapter, we investigate the application of sensorimotor learning in RGB-D object recognition, following the observation learning scenario. Three deep learning-based models that fuse appearance and affordance information by adopting multiple fusion schemes are presented. Further, six alternative representations are used as input to the affordance stream in order to maximize the information gain by incorporating affordance information. An attention mechanism based on appearance stream confidence is developed, and an auxiliary loss for fusion model optimization based on the affordance stream performance is also introduced. The 3D convolution based two-stream model with multi-layer fusion is experimentally shown to significantly improve the appearance-only baseline and outperform the rest of the proposed models, as well as alternative probabilistic fusion methods of the literature. A cross-view analysis conclude the study, providing intuition concerning viewpoint contribution to model performance and viewpoint-dependency of the affordance information.

Chapter 6

Affordance Reasoning and Segmentation

6.1 Introduction

In this chapter we investigate the joint reasoning, *i.e.* the combination of affordance localization and recognition, and pixel-wise segmentation of the object affordances. Recent work in affordance reasoning [15, 63] has demonstrated the advantages of learning spatio-temporal feature embeddings instead of predicting salient hotspots on static images. The proposed methods learn to predict the object part that affords specific actions by processing hand-object interaction videos. Interestingly, these methods do not depend on object details. Inspired by the these findings, we go further and argue that affordance reasoning can be exploited to improve affordance segmentation in a joint-learning scenario, which constitutes a critical step towards truly robust affordance understanding.

In particular, we argue that it is possible to localize and segment the object affordance using hand-object interaction information, without the need for strong object-related supervision (*i.e.* object labels and bounding boxes) or any intermediate object detection step. Following this intuition, visualized in Fig. 6-1, we adopt the "funtion from motion" perspective (discussed in Chapter 1) and propose an end-to-end encoder-decoder model to exploit the spatio-temporal information for improving

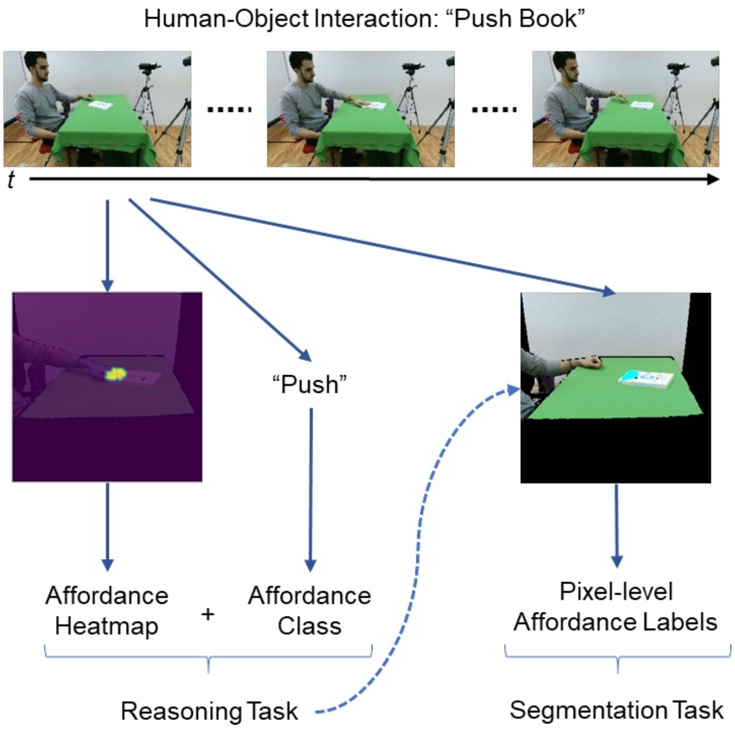


Figure 6-1: Overview of the proposed approach. Our approach aims to exploit the spatio-temporal information of the human-object interaction to localize and classify the affordances of the object, and further utilize this information to improve their pixel-level segmentation. In particular, given a human-object interaction (top), our model encodes the provided spatio-temporal information and predicts the heatmap for the interaction spot and the corresponding affordance label (left branch). The predicted heatmap is then used as auxiliary information to improve the pixel-level segmentation prediction of the model (right branch). Note that our approach does not rely on object-related annotations during training (*e.g.* object label and bounding box), and it is able to infer affordance heatmaps and pixel-level segmentation maps both in videos and static images.

affordance segmentation. The decoding part of this model consists of two decoders that can be trained jointly for the tasks of affordance reasoning and segmentation. To demonstrate the advantages of having a specific decoder for localization, we first investigate the reasoning task independently using only one decoder and comparing this model variant with state-of-the-art models in this task (see Fig. 6-2(a)). Then, as shown in Fig. 6-2(b), we add the second decoder targeting semantic segmentation, and train the model jointly without any object-related supervision. Besides the decoding part, this model exploits an attention module designed for the implicit localization

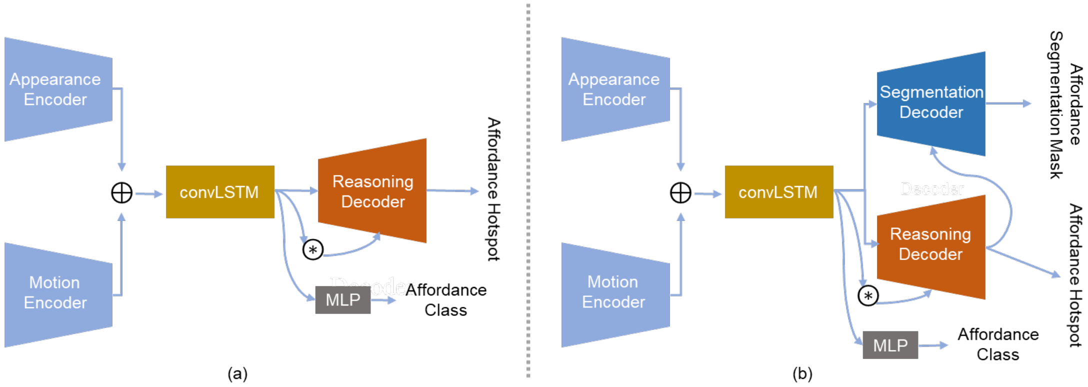


Figure 6-2: The proposed model encodes the appearance and motion information of human-object interaction sequences and produces spatio-temporal embeddings. The embeddings are then used as input to the soft-attention mechanism ($*$) that focuses on the human-object interaction spot, and to the affordance recognition branch that consists of an MLP. There are two different decoder variants that can process the output of the soft-attention mechanism and the predicted affordance class: a) the reasoning-only decoder that is designed for the affordance reasoning task, which predicts affordance heatmaps and associates them with the predicted affordance classes (see Fig. 6-3 for more details), and b) the two-decoder model that is trained jointly for affordance reasoning and segmentation. The joint model predicts affordance heatmaps that are further used to improve performance of the segmentation decoder (see Fig. 6-5 for more details).

of the hand-object interaction. This soft-attention mechanism fuses the frame-level spatial information with the video-level temporal one, forcing the network to focus on the object part that participates in the interaction.

6.2 Model Architecture and Learning Approach

In this section we present the proposed encoder-decoder model, which is depicted in Fig. 6-2(b). Its architecture, inspired by the U-Net model [87], consists of two encoders, a convolutional LSTM (cLSTM) [93], a soft-attention mechanism, an MLP for affordance recognition, and two decoders. We adopt an architecture that includes skip connections between the encoder and the decoder layers, as they help to recover the full spatial resolution and improve the gradient flow [28, 87, 92]. Further, note that the utilized cLSTM enables the robust video representation learning as it is able to capture the temporal dependencies of the human-object interaction. Prior to the model components description, we present the supported input representations.

6.2.1 Input Representations

Localizing and segmenting object parts based on human-object interaction can benefit from both appearance and motion information. In our approach, we use color and depth information to accurately represent appearance, while for motion, we use 3D optical flow information that can efficiently encode the temporal dynamics of the hand movement [73]. We choose to combine RGB and depth information by concatenating the color image and the depthmap along the channel dimension forming a $4 \times H \times W$ input, where $H = W = 300$ represent the height and the width of the input image/depthmap. Regarding 3D optical flow, we use the same algorithm as in Section 5.2.1, which computes the 3D motion vectors between two pairs of RGB-D images, and colorize them by normalizing each axis values within $[0, 255]$, thus transforming them into a three-channel image of size $3 \times H \times W$.

6.2.2 Appearance and Motion Feature Encoders

In order to exploit the appearance and motion features of the human-object interaction, we encode the RGB-D and the 3D flow information using two decoders, as shown in Fig. 6-3, where the affordance reasoning model variant is depicted. Both encoders follow the typical structure of a VGG16 model, while the encoded feature maps are concatenated at the bottleneck of the model. In particular, let $X_{RGBD}^{d \times h \times w}$ be the output feature of the RGB-D encoder, and $X_{3DOF}^{d \times h \times w}$ be the corresponding 3D flow one, where $d = 512$, $h = w = 37$ are the number of channels, height, and width of both features. Then, the two convolutional features are concatenated along the channel dimension and are convolved with d kernels of 1×1 size, producing the $X_{CAT}^{d \times h \times w}$ activation map.

6.2.3 Bottleneck and Affordance Recognition Branch

The bottleneck of the model, also visible in Fig. 6-3, consists of a residual block and two cLSTM layers. The residual block follows the RL-CONV-RL-CONV structure,

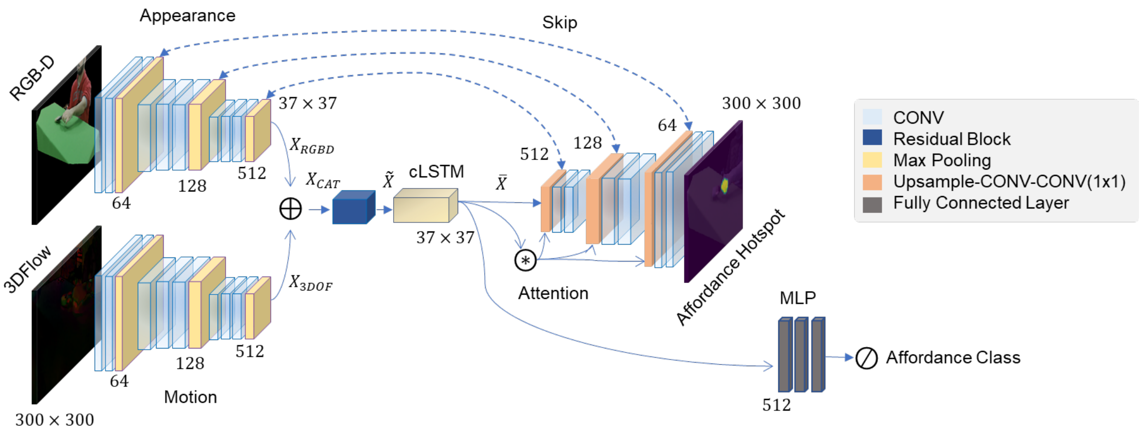


Figure 6-3: Detailed architecture of a variant of the proposed encoder-decoder model that is used for the affordance reasoning task (for the full model see Fig.6-2(b)). From left to right: a) the model receives RGB-D and 3D flow information using two convolutional encoders and fuses the encoded features, b) the latent space consists of one residual block and a convolutional LSTM module (cLSTM), followed by a soft-attention mechanism ($*$), denoted with the asterisk, c) the decoder receives the output of the attention module to predict the affordance hotspot, and d) a fully connected network (MLP) receives the convolutional LSTM output to predict the affordance class using a softmax operator ($\oslash$). The skip connections help to recover the full spatial resolution and improve the gradient flow. The numbers under/over layers indicate number of channels (*e.g.* 64, 128, 512), while $H \times W$ numbers indicate spatial resolution (*e.g.* 37×37).

adopting the pre-activation method and the identity mapping proposed in [29] for performance improvement. Both the residual block and the cLSTMs use CONV layers with 3×3 kernel size and stride equal to 1. The activation maps after the residual block and the last cLSTM layer have the same dimensionality, and they are denoted as $\tilde{X}^{d \times h \times w}$ and $\bar{X}^{d \times h \times w}$, respectively. The cLSTM layers are followed by a soft-attention mechanism that is detailed in Section 6.2.4.

Besides the attention module, the output of the cLSTM is further processed by three 512-dimensional MLP layers. The MLP is followed by a softmax classifier, which is used for affordance recognition. We use affordance recognition to further regularize the model parameters during training, and to associate the predicted affordance label to the corresponding affordance heatmap of the reasoning decoder (detailed in Section 6.2.5). Note that we choose to place the affordance recognition branch after the cLSTM module, inspired by the 2D/3D action recognition literature, where both

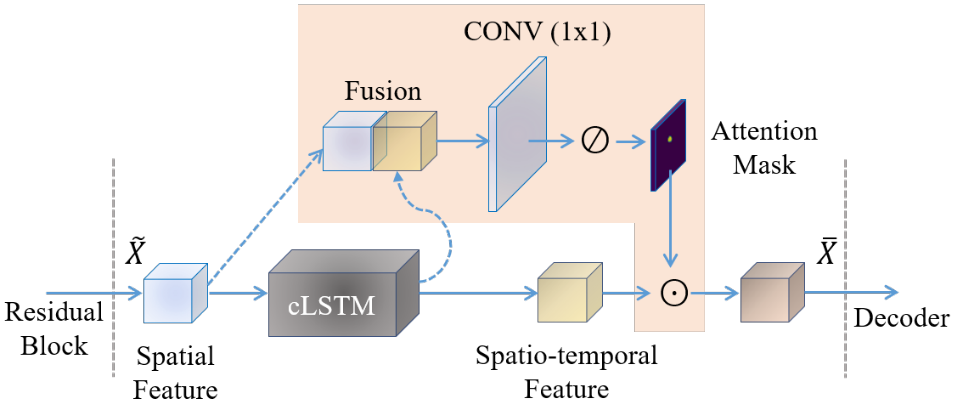


Figure 6-4: Detailed architecture of the proposed spatio-temporal soft-attention mechanism. The spatial feature of the last residual block of the network bottleneck is concatenated with the output of the convolutional LSTM. The result is processed by a CONV layer with 1×1 kernel size and a softmax operator ($\oslash$) to get the attention mask. The mask is then multiplied with the spatio-temporal feature in an element-wise manner ($\odot$).

appearance and motion features are encoded in the context of various CNN-LSTM model architectures [79, 94].

6.2.4 Soft-attention Mechanism

Using detection mechanisms to localize the object before predicting its affordance requires extra knowledge about its class label and bounding box, while also adding significant complexity to the model architecture. Since the affordance part of the object is "exposed" during the interaction with the human, we design a soft-attention mechanism that forces the model to focus on that specific part using both spatial and temporal information [120].

The aforementioned mechanism is object-agnostic and its structure is summarized in three steps, as depicted in Fig. 6-4. First, we concatenate the $\tilde{X}$ and $\bar{X}$ activation maps at the channel dimension and convolve the produced feature using a kernel of 1×1 size. Second, we use the softmax function to normalize the activation values to the $[0, 1]$ space, forming the "excitation" mask $M^{1 \times h \times w}$. Finally, M is multiplied with each channel of $\bar{X}$ in an element-wise manner, and then it is upsampled using nearest neighbor interpolation and re-applied to the activation maps of the reasoning

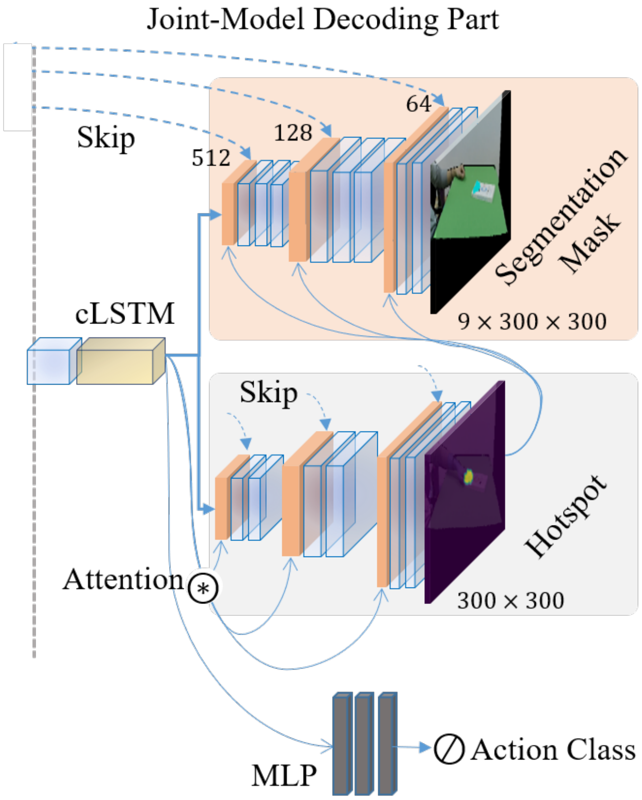


Figure 6-5: Detailed representation of the joint model decoding part. Both decoders receive the spatio-temporal feature of the convolutional LSTM as input. The soft attention mechanism ($*$) is used to guide the reasoning decoder at different levels of granularity. The predicted interaction hotspot operates as a second attention mechanism, masking the segmentation decoder activations to achieve a more accurate pixel-wise affordance class prediction.

decoder after each upsampling layer, as shown in Figs. 6-3 and 6-5.

6.2.5 Reasoning and Segmentation Decoders

We use different decoders for the tasks of affordance reasoning and segmentation. The two decoders share similar structure, however the segmenation one is deeper, as more detailed spatial information is required for semantic segmentation at the pixel level compared to the coarser heatmap prediction.

First we use solely the reasoning decoder for the affordance reasoning task, proposing a model variant that is depicted in Fig. 6-3. This decoder is a combination of 6 CONV, 6 RL, and 3 upsampling layers, and predicts an $H \times W$ heatmap. After each upsampling layer, a CONV layer follows, while its output feature is concatenated with the corresponding one from the appearance and motion encoders through the corresponding skip connections. A CONV layer with 1×1 kernel size follows, forcing intra-channel correlation learning. Note that each channel of produced activation map is multiplied with the attention mask M in a element-wise manner (as described in the previous section). The last CONV layer results in a $1 \times H \times W$ feature, where a softmax function is applied to get the final heatmap. We use RL as activation after each CONV layer, and nearest neighbor interpolation for upsampling.

As reported, the segmentation decoder shares similar structure with the reasoning one, using 14 CONV layers to preserve the spatial information details. The main difference is that instead of using the output of the soft-attention mechanism to mask the activations after each upsampling module, the predicted affordance heatmap is exploited. In particular, the affordance heatmap is multiplied in an element-wise manner with each channel of the activation map after each upsampling layer. Note that the heatmap is downsampled to two different spatial resolutions, namely 75×75 and 150×150, to match the height and width of the activation map after each upsampling layer. The decoder results in a $C \times H \times W$ dimensional feature, where each $c = 1, \ldots, C$ corresponds to a predicted affordance map, and C is the number of affordance classes. The structure of the segmentation decoder is depicted in Fig. 6-5(top).

6.2.6 Joint-task Learning

We argue that the affordance reasoning and segmentation tasks are complementary to each other as: a) their predictions are based on the same spatio-temporal embedding that is designed to focus on the human-object interaction hotspot, and b) the segmentation task can benefit from the localization of this hotspot, as the affordance heatmap and segmentation mask should overlap.

To take advantage of this complementarity, we train our model jointly for the two tasks, by minimizing the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{heat} + \lambda_3 \mathcal{L}_{aff}, \tag{6.1}$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ are hyper-parameters that add to 1. We compute $\mathcal{L}_{heat}$ as the Kullback–Leibler Divergence (KLD) between the predicted and the ground-truth heatmaps as follows:

$$\mathcal{L}_{heat} = \frac{1}{N} \sum_k \hat{D}_k \log \frac{\hat{D}_k}{D_k}, \tag{6.2}$$

where $N = H \times W$, while $\hat{D}$ and D are the probability distributions of the predicted and the ground-truth heatmaps, normalized over the total number of pixels. $\mathcal{L}_{seg}$ is the per-pixel cross-entropy of the predicted and ground-truth affordance labels, defined as:

$$\mathcal{L}_{seg} = -\frac{1}{H \times W} \sum_{c,i,j} U_{c,i,j} \log(\hat{U}_{c,i,j}), \tag{6.3}$$

where $\hat{U}$, U are the predicted and the ground-truth affordance maps, normalized over the total number of pixels. Similarly, we define $\mathcal{L}_{aff}$ as:

$$\mathcal{L}_{aff} = -\sum_c a_c \log(\hat{a}_c), \tag{6.4}$$

where $\hat{a}$, a are the predicted and ground-truth affordance labels, respectively.

6.3 Experimental Framework and Results

In this section, we present the quantitative and qualitative evaluation of the proposed model for the affordance reasoning and segmentation tasks.

6.3.1 Datasets

We use our own SOR3D-AFF dataset, discussed in Section 4.3, to evaluate our model in both tasks. Besides SOR3D-AFF, we also use the OPRA video dataset to evaluate the reasoning model variant, as this dataset does not contain pixel-level affordance annotations. Additionally, we use the UMD and IIT-AFF datasets to qualitatively evaluate our segmentation model on unseen objects. These two datasets consist of static images coupled with segmentation annotations, however there are no human-object interactions and hotspot annotations included. Note that OPRA, UMD, and IIT-AFF are presented in detail in Section 4.4.

6.3.2 Reasoning Evaluation

Implementation Details

The videos of SOR3D-AFF and OPRA are subsampled to 10 FPS, while their frames are center-cropped to 300×300 pixel resolution. We pre-train both encoders on separate datasets; the RGB-D encoder followed by the cLSTM is trained for 50 epochs on the UTKinect action recognition dataset [113], while for the colorized 3D flow encoder we utilize the weights of a VGG16 model pre-trained on ImageNet. To enable fair comparison with the RGB-only baseline models, we further compute the 2D flow of each video, and pre-train the respective encoder using only color information. For the decoder and the MLP layer weights initialization, we employ the Xavier method [23]. The model is fine-tuned in an end-to-end fashion for 80 epochs, using batch size equal to 6, Adam optimization [43], and learning rate set to 2×10^{-5}. Since the model is trained using small batch size, we choose to use group normalization [112] between each CONV and RL layers. Further, in (6.1) we set $\lambda_1 = 0$, since there is no segmentation for this task, $\lambda_2 = 0.3$ and $\lambda_3 = 0.7$ for the first 50 epochs, as affordance recognition is a critical step towards affordance hotspot prediction and should guide the total loss. For the last 30 epochs, both hyperparameters are set to 0.5. All experiments are conducted on 2 Nvidia Titan X GPUs.

Alternative Models Considered

We evaluate our reasoning model against the following state-of-the-art methods:

- **SalGAN [71]:** SalGAN estimates the most salient regions in an image by predicting heatmaps. It is trained in a supervised manner using saliency annotations. We use the original implementation of [71] and pre-train the model using the SALICON dataset [39].

- **Demo2Vec [15]:** Demo2Vec predicts affordance heatmaps on target object images based on demonstration videos. It is trained in a supervised setting using heatmap annotations. We re-implement the model to support both SOR3D-AFF and OPRA input resolution.

- **Img2Heatmap:** Adopting the term and goal from [63], we define Img2Heatmap as a static Demo2Vec variant. The model architecture is identical with the Demo2Vec one, however it is trained without the video context, *i.e.* using only static images.

- **Grounded Human-Object Interactions (GHOI) [63]:** GHOI also predicts affordance heatmaps based on human-object interaction videos. However, the model is trained in a weakly supervised manner using only affordance class annotations. We use the original implementation of [63].

Note that we consider SalGAN as a weakly supervised model for affordance prediction, as it is trained using saliency heatmaps that do not correspond to specific affordance classes. Similarly, GHOI is weakly supervised in the content of affordance heatmap prediction as it is trained using affordance class labels. On the other hand, both the original Demo2Vec and its static variant are strongly supervised using affordance heatmap annotations. Note that in the context of affordance reasoning, our model is also strongly supervised using heatmap annotations.

Evaluation Metrics

To quantitatively evaluate our reasoning model against the aforementioned baselines we use two metrics: a) KLD as described in (6.2), and b) SIM, which is a popular metric in the saliency research community [5] and measures the similarity between two heatmaps $\hat{D}$ and D that have values in the $[0, 1]$ range and add up to 1. In

Table 6.1: Comparative evaluation of the proposed affordance reasoning model against the state-of-the-art on the SOR3D-AFF and OPRA datasets. The performance of the RGB-D based model variant is reported only for SOR3D-AFF. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

	SOR3D-AFF		**OPRA**	
Method	KLD (↓)	SIM (↑)	KLD (↓)	SIM (↑)
SalGAN [71]	2.452	0.289	2.121	0.308
GHOI [63]	1.992	0.319	1.425	0.363
Img2Heatmap	2.026	0.312	1.481	0.352
Demo2Vec [15]	1.961	0.322	1.198	0.483
Ours (RGB-only)	1.818	0.332	**1.189**	**0.488**
Ours (RGB-D)	**1.439**	**0.412**	n/a	n/a

particular, SIM is defined as:

$$\mathrm{SIM}(\hat{D}, D) = \sum_i \min(\hat{D}_i, D_i), \qquad (6.5)$$

summing over all pixels i. In our case $\hat{D}, D$ correspond to the predicted and ground-truth heatmaps, respectively.

Quantitative Evaluation Results

Table 6.1 reports the performance of the aforementioned models on the SOR3D-AFF and OPRA datasets. Since SOR3D-AFF consists of RGB-D data, we evaluate two model variants, one using only color as input representation (RGB-only) coupled with 2D optical flow, while the second utilizes both color and depth information (RGB-D) along with 3D optical flow. On the other hand, for the experiments on the OPRA dataset, where there is no depth information, we evaluate our color-based model only. Note that for the SOR3D-AFF experiments, we use the last frame of each video as "target" frame.

From the reported results, we observe that the strongly supervised models achieve better overall performance than the weakly supervised ones on both datasets. Regarding the SOR3D-AFF experiments, our RGB-D model achieves the best results in

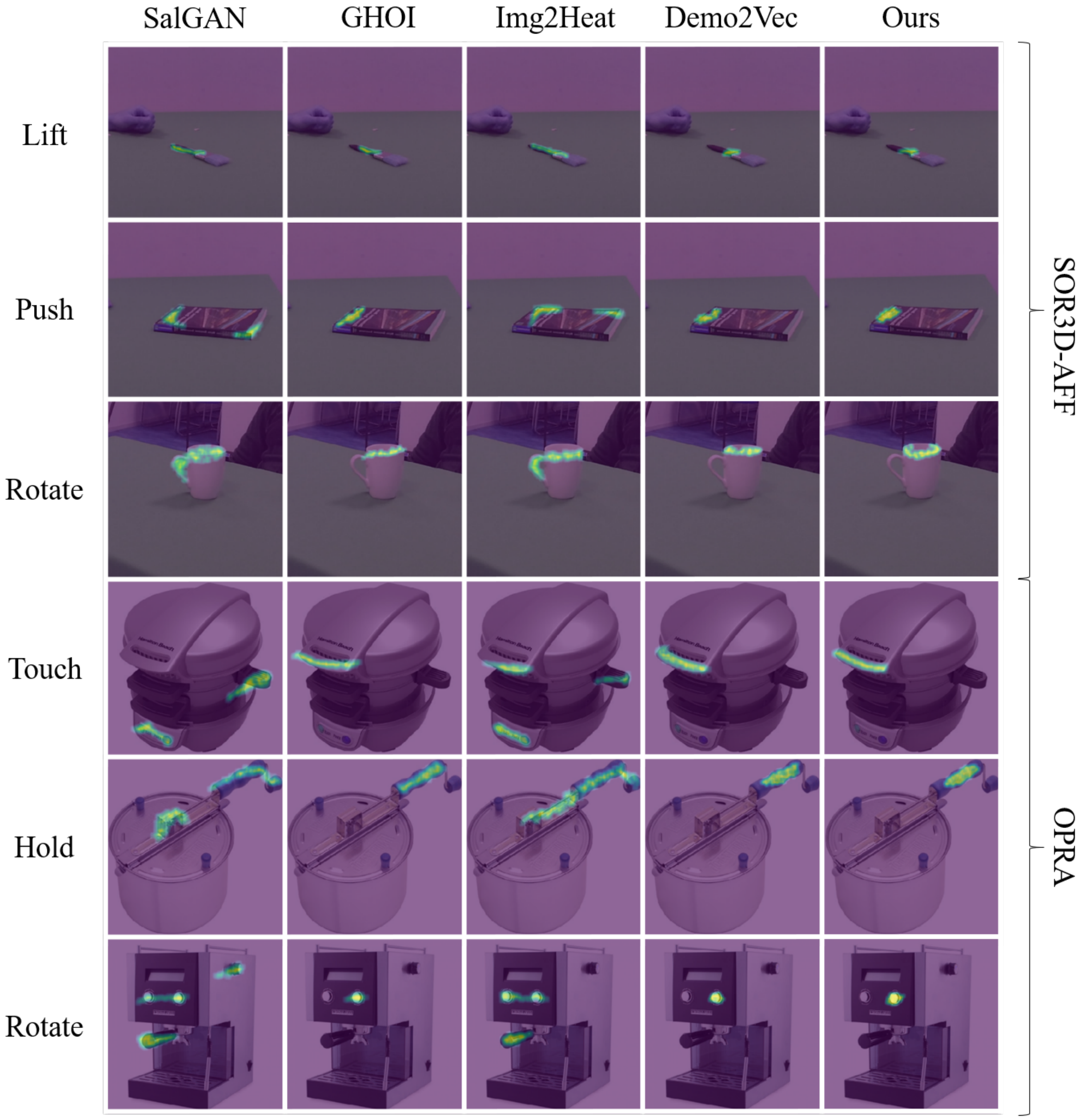


Figure 6-6: Affordance heatmap prediction on target images from the SOR3D-AFF (upper three rows) and the OPRA [15] (lower three rows) datasets. For the former, the target frame corresponds to the last frame of the video, while the latter utilizes an image that depicts the object of the video, but without background or occlusions. Each heatmap is associated with the corresponding affordance class predicted from each model (left), while the model name is shown at the top of the image. The heatmaps are overlaid on the target images for better visualization, while the object classes are shown for clarity.

both metrics, *i.e.* 1.439 KLD and 0.412 SIM, while our RGB-only model marginally outperforms Demo2Vec. We believe that this performance difference between our model and Demo2Vec is a result of our reasoning decoder design, namely the use of

Table 6.2: Comparative evaluation of a number of variations in the reasoning model architecture on the SOR3D-AFF test set. The soft-attention mechanism is denoted as α. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Model Parameters	KLD (↓)	SIM (↑)
RGB	2.209	0.231
RGB + α	2.031	0.294
RGB + α + 2D flow	1.818	0.332
RGB-D	2.189	0.292
RGB-D + α	1.914	0.368
RGB-D + α + 3D flow	1.439	0.412

the "upsampling, CONV layers" instead of using transposed convolution layers. In this way, our decoder is able to preserve more fine-grained spatial information up to the heatmap prediction [111]. Based on the RGB-D model results, we observe that the utilization of depth information in both appearance and motion, *i.e.* 3D flow, leads to more discriminative features. Regarding the OPRA experiments, we observe similar performance, with our RGB-only model marginally outperforming Demo2Vec. Fig. 6-6 shows some indicative hotspot predictions using samples from both datasets. From the visualized samples, we observe that our model is able to predict accurate hotspots, associated with the affordance label of the interaction (*e.g.* highlight the handle of the pan for affordance "hold"). In contrast, the static version of Demo2Vec (Img2Heat) and the saliency model, which are trained on static images, highlight the most salient regions of the objects regardless the affordance class. Finally, by observing the performance of the GHOI model, we conclude that the temporal information of the human-object interaction is more critical than the strong supervision for the affordance reasoning task.

Ablation Study

In order to demonstrate the contribution of each individual component to the proposed model architecture, we perform an ablation study using the following variations: a) single-stream RGB-only encoder, b) single-stream RGB-only encoder and soft-attention mechanism, and c) two-stream encoder for RGB-only and 2D optical flow, and soft-attention mechanism. Note, that the same model variations are investigated for the RGB-D and 3D optical flow information.

Table 6.2 reports the results of the investigated model variants on the SOR3D-AFF test set. Evidently, RGB-D information leads to better overall performance, while the integration of the soft-attention mechanism improves both RGB-only and RGB-D based models in terms of KLD and SIM. Finally, we observe that the use of a second stream that processes optical flow information significantly boosts each model performance, mostly due to its contribution to the affordance recognition part of the network.

6.3.3 Segmentation Evaluation

Implementation Details

Similarly to the reasoning setup, all data from the SOR3D-AFF, UMD, and IIT-AFF datasets are resized to 300×300 pixel resolution, while each video from SOR3D-AFF is subsampled to 10 FPS. Focusing on the joint model, we pre-train both encoders as reported in the reasoning setup, and fine-tune the model for 200 epochs. We set batch size equal to 4 and use group normalization to normalize the activations after each CONV layer, while we set the learning rate to 2×10^{-5}. The model is optimized using the Adam algorithm. Following (6.1), we set $\lambda_1 = 0.3$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.6$, and train our models on 2 Nvidia Titan X GPUs.

Alternative Models Considered

We compare our model with AffordanceNet [12], a convolutional encoder-decoder model that utilizes a region proposal network [84] in order to restrict affordance segmentation to the detected object bounding box in the static image. We re-implement AffordanceNet in order to be able to receive 300×300 inputs, and train it for 50 epochs with the batch size set to 8 and learning rate equal to 2×10^{-5}.

Evaluation Metrics

We use two different metrics to assess model performance: a) the Intersection over Union (IoU), and b) the F_β-score. IoU, originally proposed in [70], quantifies the overlap between $\hat{U}$ and U, namely the predicted affordance and the corresponding ground-truth set of pixels, and is defined as:

$$\mathrm{IoU}(\hat{U}, U) = \frac{|\hat{U} \cap U|}{|\hat{U}| + |U| - |\hat{U} \cap U|}, \tag{6.6}$$

where $| * |$ denotes the cardinality of set $*$. Additionally, F-score provides helpful insight about the model robustness based on false positive and negative predictions. Since some affordances are associated with more objects, we choose to evaluate the performance of the model using a weighted version of the F-score metric, which is proposed in [57] and is denoted as:

$$\mathrm{F}_\beta^w = (1 + \beta^2)\frac{\mathrm{P}^w \mathrm{R}^w}{\beta^2 \mathrm{P}^w + \mathrm{R}^w}, \text{with } \beta = 1, \tag{6.7}$$

where $\mathrm{P}^w, \mathrm{R}^w$ are the weighted versions of the standard precision and recall metrics, respectively. Since "grasp" and "lift" are the most dominant affordance labels, we set their weight to 0.2, while next dominant label "push" is set to 0.1. We want the weights of the labels to sum to 1, thus the weight of each remaining label is set to 0.083.

Quantitative Evaluation Results

Table 6.3 reports the overall performance of the joint model on the SOR3D-AFF test set. Since there is no alternative model in the literature for inferring pixel-level affordance labels based on videos, only our model results are reported (top row). For the static, *i.e.* image-only based, affordance prediction, our model is compared to the AffordanceNet and achieves competitive results (bottom rows). Recall also that the goal of our work is to perform affordance segmentation using minimal supervision, *i.e.* using affordance-related ground-truth only at the last frame of the video while

Table 6.3: Overall object affordance segmentation results on the SOR3D-AFF test set based on video (top) and static image inference (bottom). ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Model	IoU (↑)	F_β (↑)	F_β^w (↑)
Ours	0.731	0.820	0.821
AffordanceNet [12]	0.561	0.618	0.621
Ours	0.559	0.617	0.622
Ours (extra supervision)	**0.575**	**0.625**	**0.638**

omitting any object-related annotations (*i.e.* the object class and bounding box). The results also support our argument that a model can be trained using interaction sequences and infer affordance labels for both videos and static images. However, since SOR3D-AFF provides object bounding boxes and object classes, we also investigate training our model using 2 extra losses to exploit the object-related annotations. We use the $L2$-norm to measure the bounding box error, defined as:

$$\mathcal{L}_{bbox} = \|\hat{\rho} - \rho\|_2, \tag{6.8}$$

where $\hat{\rho}, \rho$ are vectors with the predicted and ground-truth bounding box coordinates (top left corner, width, and height). Further, we utilize (6.4) with $C = 10$ (object classes) to measure the object recognition loss, which we define as $\mathcal{L}_{obj}$. The total loss for this experiment becomes:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{heat} + \lambda_3 \mathcal{L}_{aff} + \lambda_4 \mathcal{L}_{bbox} + \lambda_5 \mathcal{L}_{obj}, \tag{6.9}$$

where we set $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, $\lambda_3 = 0.3$, $\lambda_4 = 0.1$, and $\lambda_5 = 0.3$. As observed in the last row of Table 6.3, the model trained with extra supervision outperforms AffordanceNet, achieving 2.50%, 1.13%, and 2.74% relative improvement in IoU, F_β, and F_β^w, respectively.

We also present results per affordance category in Table 6.4, based on both video (top) and static image (bottom) inference. From the reported results, we can observe

Table 6.4: Category-specific object affordance segmentation results of our model on the SOR3D-AFF test set based on video (top) and static image inference (bottom).

Metric	cut	grasp	hammer	lift	paint	push	rotate	squeeze	type
IoU	0.477	0.878	0.761	0.920	0.713	0.777	0.754	0.671	0.633
F_β	0.612	0.929	0.859	0.952	0.790	0.892	0.847	0.769	0.731
F_β^w	0.613	0.931	0.860	0.952	0.791	0.894	0.859	0.772	0.734
IoU	0.381	0.673	0.594	0.709	0.532	0.634	0.579	0.533	0.408
F_β	0.447	0.707	0.652	0.761	0.590	0.678	0.641	0.592	0.479
F_β^w	0.468	0.716	0.661	0.772	0.592	0.681	0.643	0.594	0.481

the superiority of the dominant affordances, *i.e.* these associated with most of the objects, such as “grasp” and “lift”, as well as the adequate performance of complex affordances that change the visual representation of the object, such as “squeeze”. Note, that affordance label weighting leads to a slightly better overall performance in terms of F-score, which is expected given the very confident predictions for the dominant affordances.

Qualitative Evaluation Results

Besides quantitative evaluation, we use samples from the image-only UMD and IIT-AFF datasets to qualitatively evaluate our model performance. As depicted in Fig. 6-7, our model is able to confidently predict learned affordances on UMD corpus objects that are similar to that of SOR3D-AFF (*e.g.* “bottle”, “hammer”, and “knife”), while also inferring reasonable affordance labels of the dominant affordances on samples from the challenging, due to the cluttered scenes, IIT-AFF (*e.g.* “rotate” or “grasp” pixel-level predictions on the unseen object class “cable”). Note that we center-crop a 300 × 300 portion of each image and visualize the affordance pixel-level predictions with values greater than 0.75.

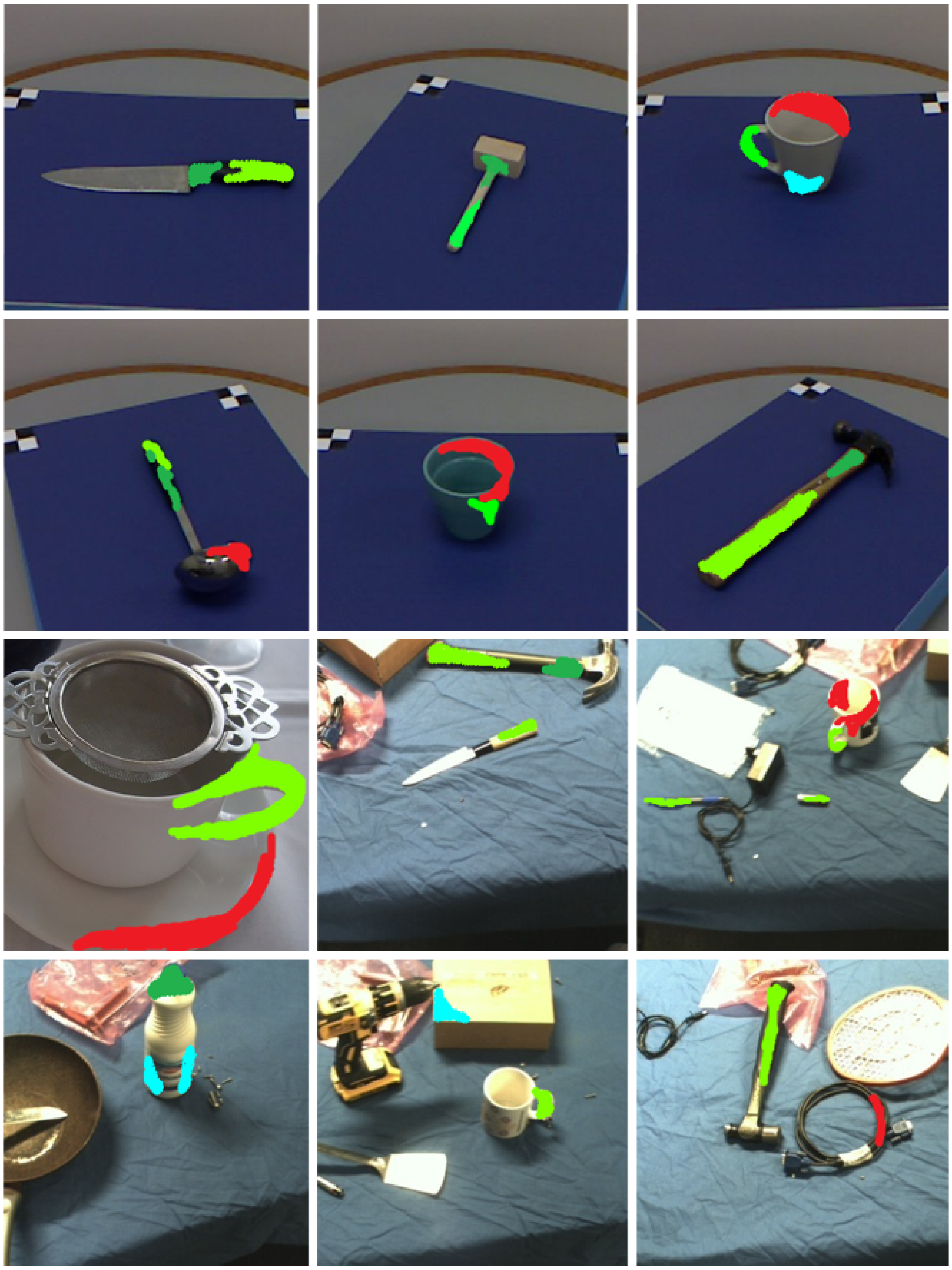


Figure 6-7: Pixel-level affordance class predictions on unseen objects from UMD [62] (upper two rows) and IIT-AFF [66] (lower two rows) datasets. The predictions are color-coded based on SOR3D-AFF annotation: “grasp” in light green, “lift” in green, “rotate” in red, “push” in cyan and are shown when the classifier confidence exceeds 0.75.

Ablation Study

Similarly to Section 6.3.2, we evaluate different segmentation model variants in order to demonstrate the contribution of each individual component. Note that for this

Table 6.5: Comparative evaluation of a number of variations in the encoder-decoder architecture on the SOR3D-AFF test set based on video (top) and static image inference (bottom). The 2^{nd} column reports the decoder used for segmentation: a) "Seg" denotes the segmentation-only decoder (*i.e.* without using the reasoning decoder), and b) "Joint" denotes the model with the two decoders proposed in this Thesis. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Model Parameters	Decoder	IoU (↑)	F_β (↑)
RGB	Seg	0.619	0.733
RGB + α		0.652	0.769
RGB + α + 2D flow		0.663	0.784
RGB-D		0.640	0.771
RGB-D + α		0.703	0.797
RGB-D + α + 3D flow		0.718	0.801
RGB-D + α + 3D flow	Joint	0.731	0.820
RGB + α + 2D flow	Seg	0.471	0.529
RGB-D + α + 3D flow		0.537	0.582
RGB-D + α + 3D flow	Joint	0.559	0.617

ablation we utilize a model variant that uses only the segmentation decoder (similar to the reasoning-only decoder variant depicted in Fig. 6-2(a)), which is reported as "Seg" in Table 6.5, while the jointly trained model is reported as "Joint". We follow the same logic and present from the single-stream RGB-only encoder up to the two-stream RGB-D one coupled with attention.

The segmentation results of the aforementioned variations on the full sequences of the SOR3D-AFF test set are reported in Table 6.5. As in the reasoning task, we observe that the depth information, the optical flow, as well as the attention mechanism lead to superior performance in both metrics. Besides the contribution of the aforementioned components, we observe a significant performance boost when training jointly the reasoning and segmentation decoders for the RGB-D model. In fact, the joint model yields 1.81% IoU and 2.37% F_β relative improvement when tested on videos, while achieving 4.10% IoU and 6.01% F_β relative improvement when tested on static images.

6.4 Conclusion

In this chapter, the affordance reasoning and segmentation tasks are jointly investigated, following the “function from motion” scenario by processing videos that include human-object interactions. In particular, an end-to-end deep encoder-decoder model is proposed, which encodes color, depth, and motion information from human-object interaction videos, and predicts affordance hotspots and segmentation maps. The model uses a spatio-temporal soft-attention mechanism that enforces implicit localization of the interaction hotspot, which leads to performance improvement in both tasks. The reported results on the SOR3D-AFF corpus show that the proposed model predicts more accurate affordance heatmaps compared to alternative state-of-the-art methods in the literature, while regarding the segmentation task, it outperforms AffordanceNet when predicting affordance masks from either videos or static images. Finally, the model generalization ability is demonstrated through qualitative evaluation using unseen objects from two image-only datasets.

Chapter 7

Conclusions and Future Directions

This final chapter summarizes the thesis contributions and the significance of our work. Additionally, based on the presented research findings, this chapter discusses possible future research directions.

7.1 Conclusions

In this thesis we investigate the notion of the object affordance, namely the set of interactions supported by the object, in the context of computer vision. In particular, we focus on the so-called "function from motion" approach, where the affordance information is visible when observing human-object interactions. To truly understand the nature of the affordance information, we first use it as an auxiliary object attribute to improve object recognition, while then we use it to localize and segment the part of the object that supports the corresponding interaction. The aforementioned tasks are investigated by adopting the deep learning paradigm. That is, we use state-of-the-art deep neural networks to design each learning framework.

Regarding the affordance-based object recognition direction, we show that it is possible to apply the sensorimotor learning approach on a computer vision problem by designing a two-stream deep neural network. Additionally, we demonstrate that the affordance information can be encoded in several representations, which are then fused with the object appearance and lead to improved object recognition performance.

On the other hand, we highlight the ability to localize and segment the affordance part of an object in both videos and images. We show that by exploiting the spatio-temporal nature of the affordance information, it is possible to reason about the object affordances by localizing the actual object part that participates in the interaction. Further, we confirm that the reasoning task can contribute to the pixel-level segmentation of the affordance part of the object, by designing a deep neural network to learn the two tasks jointly.

7.2 Future Directions

The ability to answer the “what object?” and “how to use it?” questions by exploiting affordance information can impact many practical computer vision and robotics applications. There are two distinct research directions that can be investigated: a) the affordance information contribution to more complex indoor scenes, focusing on both human-object and object-object interactions through observation, and b) the sensorimotor learning approach in an active learning scenario that involves a robot and an indoor environment.

The first direction can be adopted to design algorithms with application in real-world scenarios that involve a passive perception system, such as smart homes [2, 77]. These algorithms will adopt the observation-based learning in cluttered scenes, focusing not only on the prediction of the object functionality with respect to the human-object interaction, but also with respect to the interaction among multiple objects for a specific purpose. Note that this scenario enables the exploitation of other modalities, such as speech in an audio-visual learning setup.

Regarding the active learning scenario, teaching a home robot to recognize objects and understand how to use them is vital. One possible but rather impractical way would be to collect data from the robot perspective in different indoor environments and train different models for each task at a time. However, an active sensorimotor learning scenario, where the robot combines the visual with the motor information and is continuously rewarded or punished seems more promising [37,82]. In this case, the sensorimotor framework should be re-designed to be able to improve based on the feedback of each learning step. Since visual recognition still heavily relies on detailed annotated data, this reinforcement learning scenario seems necessary for the transition from passive perception to active visual intelligence.

Appendix A

Publications

This thesis has led to the following publications:

- S. Thermos, G. Potamianos, P. Daras. A deep learning approach to joint object affordance reasoning and segmentation in RGB-D videos. *Submitted to the IEEE Transactions on Circuits and Systems for Video Technology*, May, 2020.

- S. Thermos, G.T. Papadopoulos, P. Daras, G. Potamianos. Deep sensorimotor learning for RGB-D object recognition. *Computer Vision and Image Understanding*, vol. 190 (4), Jan. 2020.

- S. Thermos, P. Daras, G. Potamianos. A deep learning approach to object affordance segmentation. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2358-2362, Barcelona, Spain, 2020.

- S. Thermos, G.T. Papadopoulos, P. Daras, G. Potamianos. Attention-enhanced sensorimotor object recognition. *In Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 336-340, Athens, Greece, 2018.

- S. Thermos, G.T. Papadopoulos, P. Daras, G. Potamianos. Deep affordance-grounded sensorimotor object recognition. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 49-57, Honolulu, Hawaii, USA, 2017.

Bibliography

[1] A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. In D. G. Stork and M. E. Hennecke (eds.). *Speechreading by Humans and Machines: Models, Systems, and Applications*, pages 461–471. Springer, Berlin Heidelberg, 1996.

[2] M. R. Alam, M. B. I. Reaz, and M. Mohd Ali. A review of smart homes – Past, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(6):1190–1203, 2012.

[3] A. Andreopoulos and J. K. Tsotsos. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, 117(8):827–891, 2013.

[4] M. L. Brandi, A. Wohlschläger, C. Sorg, and J. Hermsdörfer. The neural correlates of planning and executing actual tool use. *The Journal of Neuroscience*, 34(39):13183–13194, 2014.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

[6] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo. Using object affordances to improve object recognition. *IEEE Transactions on Autonomous Mental Development*, 3(3):207–215, 2011.

[7] L. L. Cloutman. Interaction between dorsal and ventral processing streams: Where, when and how? *Brain & Language*, 127(2):251–263, 2013.

[8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 3488–3493, 2016.

[9] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. In *Proc. European Conference on Computer Vision (ECCV)*, pages 284–298, 2012.

[10] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[11] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[12] T. Do, A. Nguyen, and I. Reid. AffordanceNet: An end-to-end deep learning approach for object affordance detection. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5, 2018.

[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proc. International Conference on Machine Learning (ICML)*, pages 647–655, 2014.

[14] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, 2015.

[15] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim. Demo2Vec: Reasoning object affordances from online videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2139–2147, 2018.

[16] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016.

[17] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao. GVCNN: Group-view convolutional neural networks for 3d shape recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–272, 2018.

[18] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action - initial steps towards artificial cognition. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, volume 3, pages 3140–3145, 2003.

[19] J. H. Flavell. Cognitive development: Past, present, and future. *Developmental Psychology*, 28(6):998–1005, 1992.

[20] A. Ghadirzadeh, J. Bütepage, D. Kragic, and M. Björkman. Self-learning and adaptation in a sensorimotor framework. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 551–558, 2016.

[21] A. Giagkos, D. Lewkowicz, P. Shaw, S. Kumar, M. Lee, and Q. Shen. Perception of localized features during robotic sensorimotor development. *IEEE Transactions on Cognitive and Developmental Systems*, 9(2):127–140, 2017.

[22] J. J. Gibson. The theory of affordances. In R. Shaw and J. Bransford (eds.). *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum, Hillsdale NJ, 1977.

[23] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[24] S. Gupta, R. B. Girshick, P. A. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 345–360, 2014.

[25] S. Hadfield and R. Bowden. Scene particles: Unregularized particle-based scene flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):564–576, 2014.

[26] M. Hassanin, S. Khan, and M. Tahtali. Visual affordance and function understanding: A survey. *CoRR*, abs/1807.06775, 2018.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[29] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 630–645, 2016.

[30] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[31] V. Högman, M. Björkman, A. Maki, and D. Kragic. A sensorimotor learning framework for object categorization. *IEEE Transactions on Cognitive and Developmental Systems*, 8(1):15–25, 2016.

[32] C. Hong, J. Yu, J. You, X. Chen, and D. Tao. Multi-view ensemble manifold regularization for 3D object recognition. *Information Sciences*, 320:395–405, 2015.

[33] M. Hornácek, A. Fitzgibbon, and C. Rother. SphereFlow: 6 DoF scene flow from RGB-D pairs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3526–3533, 2014.

[34] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proc. European Conference on Computer Vision (ECCV)*, pages 754–769, 2018.

[35] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers. A primal-dual framework for real-time dense RGB-D scene flow. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 98–104, 2015.

[36] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine. End-to-end learning of semantic grasping. In S. Levine, V. Vanhoucke, and K. Goldberg, editors, *Proc. Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research (PMLR)*, pages 119–132. 2017.

[37] E. Jang, S. Vijaynarasimhan, P. Pastor, J. Ibarz, and S. Levine. End-to-end learning of semantic grasping. In *Proc. Conference on Robot Learning (CoRL)*, 2017.

[38] D. Jayaraman and K. Grauman. End-to-end policy learning for active visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1601–1614, 2019.

[39] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, 2015.

[40] Z. Jie, W. F. Lu, S. Sakhavi, Y. Wei, E. H. F. Tay, and S. Yan. Object proposal generation with fully convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):62–75, 2018.

[41] A. Kanezaki, Y. Matsushita, and Y. Nishida. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5010–5019, 2018.

[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014.

[43] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[44] H. Kjellström, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011.

[45] T. Kluth, D. Nakath, T. Reineking, C. Zetzsche, and K. Schill. Affordance-based object recognition using interactions obtained from a utility maximization principle. In *Proc. European Conference on Computer Vision Workshops (ECCVW)*, pages 406–412, 2014.

[46] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. *International Journal of Robotics Research*, 32(8):951–970, 2013.

[47] H. S. Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. In *Proc. European Conference on Computer Vision (ECCV)*, pages 831–847, 2014.

[48] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[50] M. Kümmerer, T. S. Wallis, and M. Bethge. Deepgaze II: Reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563, 2016.

[51] F. Langenfeld, A. Axenopoulos, A. Chatzitofis, D. Cracium, P. Daras, B. Du, A. Giachetti, Y. Lai, H. Li, Y. Li, M. Masoumi, Y. Peng, P. L. Rosin, J. Sirugue, L. Sun, S. Thermos, M. Toews, Y. Wei, Y. Wu, Y. Zhai, T. Zhao, Y. Zheng, and M. Montes. SHREC'18 track: Protein shape retrieval. In *Proc. Eurographics Workshop on 3D Object Retrieval (3DOR)*, 2018.

[52] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[53] K. Lee, K. Lee, K. Min, Y. Zhang, J. Shin, and H. Lee. Hierarchical novelty detection for visual object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1034–1042, 2018.

[54] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3367–3375, 2015.

[55] Y. Liu, H. Zha, and H. Qin. Shape topics: A compact representation and new algorithms for 3D partial shape retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2025–2032, 2006.

[56] N. Lyubova, S. Ivaldi, and D. Filliat. From passive to interactive object learning and recognition through self-identification on a humanoid robot. *Autonomous Robots*, 40(1):33–57, 2016.

[57] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2014.

[58] Microsoft. Meet Kinect for Windows. `https://developer.microsoft.com/en-us/windows/kinect`. [Online].

[59] M. Minsky. Society of mind: A response to four reviews. *Artificial Intelligence*, 48(3):371–396, 1991.

[60] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: from sensory motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.

[61] R. Mottaghi, C. Schenck, D. Fox, and A. Farhadi. See the glass half full: Reasoning about liquid containers, their volume and content. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1889–1898, 2017.

[62] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381, 2015.

[63] T. Nagarajan, C. Feichtenhofer, and K. Grauman. Grounded human-object interaction hotspots from video. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 8688–8697, 2019.

[64] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proc. International Conference on Machine Learning (ICML)*, pages 78–85, 2004.

[65] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Detecting object affordances with convolutional neural networks. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770, 2016.

[66] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915, 2017.

[67] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[68] J. Oberlin and S. Tellex. Autonomously acquiring instance-based object models from experience. In A. Bicchi and W. Burgard, editors, *Robotics Research*, volume 2, pages 73–90. Springer International Publishing, 2018.

[69] J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–1031, 2001.

[70] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[71] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. *CoRR*, abs/1701.0108, 2017.

[72] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[73] G. T. Papadopoulos and P. Daras. Human action recognition using 3D reconstruction data. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(8):1807–1823, 2018.

[74] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proc. International Conference on Machine Learning*, page 1310–1318, 2013.

[75] 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. In *In Proc. Advances in Neural Information Processing Systems (NeurIPS) Autodiff Workshop*. 2017.

[76] J. Piaget and T. Brown. *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development.* University of Chicago Press, 1985.

[77] M. P. Poland, C.D. Nugent, H. Wang, and L. Chen. Smart home research: Projects and issues. *International Journal of Ambient Computing and Intelligence*, 1(4):32–45, 2009.

[78] G. Potamianos and C. Neti. Stream confidence estimation for audio-visual speech recognition. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 746–749, 2000.

[79] A. Psaltis, G. T. Papadopoulos, and P. Daras. Deep 3D flow features for human action recognition. In *Proc. IEEE International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2018.

[80] C. R. Qi, H. Su, M. Niessner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view CNNs for object classification on 3D data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2016.

[81] J. Quiroga, T. Brox, F. Devernay, and J. L. Crowley. Dense semi-rigid scene flow estimation from RGBD images. In *Proc. European Conference on Computer Vision (ECCV)*, pages 567–582, 2014.

[82] S. Reddy, A. D. Dragan, and S. Levine. SQIL: Imitation learning via reinforcement learning with sparse rewards. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[83] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[84] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[85] N. Rhinehart and K. M. Kitani. Learning action maps of large environments via first-person vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–588, 2016.

[86] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding*, 62(2):164–176, 1995.

[87] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[88] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004.

[89] M. Savva, F. Yu, H. Su, A. Kanezaki, T. Furuya, R. Ohbuchi, Z. Zhou, R. Yu, S. Bai, X. Bai, M. Aono, A. Tatsuma, S. Thermos, A. Axenopoulos, G. T. Papadopoulos, P. Daras, X. Deng, Z. Lian, B. Li, H. Johan, Y. Lu, and S. Mk. SHREC'17 track: Large-scale 3D shape retrieval from ShapeNet Core55. In *Proc. Eurographics Workshop on 3D Object Retrieval (3DOR)*, 2017.

[90] J. Sawatzky, A. Srikantha, and J. Gall. Weakly supervised affordance detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2017.

[91] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *International Journal of Robotics Research*, 27(2):157–173, 2008.

[92] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[93] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 802–810, 2015.

[94] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 568–576, 2014.

[95] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[96] H. O. Song, M. Fritz, D. Goehring, and T. Darrell. Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13(2):798–809, 2016.

[97] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, 2015.

[98] M. Sutton, L. Stark, and K. Bowyer. Function from visual analysis and physical interaction: a methodology for recognition of generic classes of objects. *Image and Vision Computing*, 16(11):745–763, 1998.

[99] S. Thermos, P. Daras, and G. Potamianos. A deep learning approach to object affordance segmentation. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2358–2362, 2020.

[100] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos. Deep affordance-grounded sensorimotor object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–57, 2017.

[101] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos. Attention-enhanced sensorimotor object recognition. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 336–340, 2018.

[102] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos. Deep sensorimotor learning for RGB-D object recognition. *Computer Vision and Image Understanding*, 190(4), 2020.

[103] S. Thermos, G. Potamianos, and P. Daras. A deep learning approach to joint object affordance reasoning and segmentation in RGB-D videos. *(Submitted to:) IEEE Transactions on Circuits and Systems for Video Technology.*

[104] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3081–3088, 2010.

[105] D. Tran, L. Bourdev R., Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[106] L. G. Ungerleider and J. V. Haxby. ‘What’ and ‘where’ in the human brain. *Current Opinion in Neurobiology*, 4:157–165, 1994.

[107] V. van Polanen and M. Davare. Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, 79:186–191, 2015.

[108] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. GRAPHICON*, pages 85–92, 2003.

[109] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 416–425, 2017.

[110] X. Wang, R. Girdhar, and A. Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3366–3375, 2017.

[111] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings. The devil is in the decoder: Classification, regression and GANs. *International Journal of Computer Vision (IJCV)*, 127:1694–1706, 2019.

[112] Y. Wu and K. He. Group normalization. In *Proc. European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[113] L. Xia, C. C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proc. IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–27, 2012.

[114] T. Xiao, Q. Fan, D. Gutfreund, M. Monfort, A. Oliva, and B. Zhou. Reasoning about human-object interactions through dual attention networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3919–3928, 2019.

[115] Z. Yang and L. Wang. Learning relationships for multi-view 3D object recognition. In *Proc. International Conference on Computer Vision (ICCV)*, pages 7505–7514, 2019.

[116] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 3320–3328, 2014.

[117] T. Yu, J. Meng, and J. Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 186–194, 2018.

[118] M. Zambelli and Y. Demiris. Online multimodal ensemble learning using self-learned sensorimotor representations. *IEEE Transactions on Cognitive and Developmental Systems*, 9(2):113–126, 2017.

[119] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 818–833, 2014.

[120] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[121] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian. Cascaded interactional targeting network for egocentric video analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1904–1913, 2016.

[122] Y. Zhu, Y. Zhao, and S. C. Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2855–2864, 2015.