Intents and Preferences Prediction Based on Implicit Human Cues

A dissertation submitted towards the degree
Doctor of Engineering
(Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
Hosnieh Sattar, M.Sc.

Saarbrücken
May 2019

Day of Colloquium	2nd of July, 2019
Dean of the Faculty	Univ.-Prof. Dr. Sebastian Hack Saarland University, Germany

Examination Committee

Chair	Prof. Dr. Antonio Krüger
Reviewer, Advisor	Dr. Mario Fritz
Reviewer	Prof. Dr. Bernt Schiele
Reviewer	Prof. Dr. Yusuke Sugano
Academic Assistant	Dr. Paul Swoboda

ABSTRACT

Visual search is an important task, and it is part of daily human life. Thus, it has been a long-standing goal in Computer Vision to develop methods aiming at analysing human search intent and preferences. As the target of the search only exists in mind of the person, search intent prediction remains challenging for machine perception. In this thesis, we focus on advancing techniques for search target and preference prediction from implicit human cues.

First, we propose a search target inference algorithm from human fixation data recorded during visual search. In contrast to previous work that has focused on individual instances as a search target in a closed world, we propose the first approach to predict the search target in *open-world settings* by learning the compatibility between observed fixations and potential search targets.

Second, we further broaden the scope of search target prediction to categorical classes, such as object categories and attributes. However, state of the art models for categorical recognition, in general, require large amounts of training data, which is prohibitive for gaze data. To address this challenge, we propose a novel *Gaze Pooling Layer* that integrates gaze information into CNN-based architectures as an attention mechanism – incorporating both spatial and temporal aspects of human gaze behaviour.

Third, we go one step further and investigate the feasibility of combining our gaze embedding approach, with the power of generative image models to visually decode, i.e. create a visual representation of, the search target.

Forth, for the first time, we studied the effect of body shape on people preferences of outfits. We propose a novel and robust multi-photo approach to estimate the body shapes of each user and build a conditional model of clothing categories given body-shape. We demonstrate that in real-world data, clothing categories and body-shapes are correlated. We show that our approach estimates a realistic looking body shape that captures a user's weight group and body shape type, even from a single image of a clothed person. However, an accurate depiction of the naked body is considered highly private and therefore, might not be consented by most people. First, we studied the perception of such technology via a user study. Then, in the last part of this thesis, we ask if the automatic extraction of such information can be effectively evaded.

In summary, this thesis addresses several different tasks that aims to enable the vision system to analyse human search intent and preferences in real-world scenarios. In particular, the thesis proposes several novel ideas and models in visual search target prediction from human fixation data, for the first time studied the correlation between shape and clothing categories opening a new direction in clothing recommendation systems, and introduces a new topic in privacy and computer vision, aimed at preventing automatic 3D shape extraction from images.

ZUSAMMENFASSUNG

Visuelle Suche ist eine wichtige Aufgabe und ein Teil unseres täglichen Lebens. Deswegen ist es seit langem ein Ziel des maschinellen Sehens, Methoden zu entwickeln, die Analyse menschlicher Suchvorhaben und Präferenzen zur Aufgabe haben. Da das Ziel der Suche nur in der Vorstellung der Person existiert, bleibt die Vorhersage von Suchvorhaben herausfordernd für die maschinelle Wahrnehmung. In dieser Arbeit fokussieren wir uns auf fortgeschrittene Techniken zur Vorhersage von Suchzielen und Präferenzen anhand impliziter menschlicher Hinweise.

Erstens schlagen wir einen Suchziel-Inferenz-Algorithmus vor, der auf Grundlage menschlicher Fixierungsdaten, die während der visuellen Suche aufgenommen wurden, arbeitet. Im Gegensatz zu vorherigen Arbeiten die auf individuelle Beispiele als Suchziel-Objekte einer geschlossenen Welt fokussiert waren, schlagen wir den ersten Ansatz vor, der die Suchziele in einer offenen Welt vorhersagt, indem die Kompatibilität zwischen beobachteten Fixierungen und potentiellen Suchzielen gelernt wird.

Zweitens erweitern wir den Anwendungsbereich der Suchzielvorhersage auf kategorische Klassen wie zum Beispiel Objektkategorien und Attribute. Führende Modelle der Kategorienerkennung benötigen jedoch im Allgemeinen große Trainingsdatenmengen von menschlichen Blicken, welche bislang schwierig zu beschaffen sind. Um diese Herausforderung anzugehen schlagen wir eine neuartige Blick-Zusammenfassungsschicht vor, die Blickinformationen in CNN-basierte Architekturen als eine Art Aufmerksamkeitsmechanismus integriert. Dies bezieht sowohl räumliche als auch zeitliche Aspekte des menschlichen Blickverhaltens mit ein.

Drittens gehen wir noch einen Schritt weiter und untersuchen die Durchführbarkeit der Kombination unseres Blickeinbettungsansatzes mit der Stärke von generativen Bildmodellen um eine visuelle Repräsentation des Suchzieles zu dekodieren.

Zum vierten sind wir die Ersten, die den Effekt von Körperform auf die Vorlieben für verschiedene Kleidungen untersuchen. Wir präsentieren einen neuartigen, robusten Ansatz zur Schätzung der Körperform individueller Nutzer, der auf mehreren Eingabebilder basiert, und erstellen ein Modell von Kleidungskategorien bedingt auf Körperform. Wir zeigen, dass Kleidungskategorien und Körperform in Echtweltdaten korreliert sind. Weiterhin zeigen wir, dass unser Ansatz realistisch aussehende Körperformen robust schätzen kann, welche die Gewichtsgruppen und Körperformtypen der Nutzer abbilden, und das selbst dann, wenn nur ein einzelnes Eingabebild verfügbar ist.

Gleichzeitig ist eine naturgetreue Darstellung des nackten Körpers eine äußerst sensible Angelegenheit, die sicherlich nicht bei allen Nutzern auf Zustimmung stoßen dürfte. Daher untersuchen wir mithilfe einer Nutzerbefragung auch die öffentliche Einstellung einer solchen Technologie gegenüber. Im letzten Teil der Arbeit beleuchten wir darüber hinaus auch die Frage, ob die automatisierte Auswertung

solch privater Informationen effektiv umgangen werden kann.

Zusammenfassend betrachtet die vorliegende Dissertation verschiedene Fragestellungen, die es zum Ziel haben, Bildverarbeitungssysteme für die Analyse von menschlichen Suchabsichten und Präferenzen in Echtweltszenarien bereitzustellen. Insbesondere werden verschiedene neue Ideen und Modelle für die Zielvorhersache in der visuellen Suche durch menschliche Fixationsdaten beleuchtet. Ebenso wird erstmalig die Korrelation zwischen Körpermaßen und Kleidungskategorien untersucht, was neue Möglichkeiten für Bekleidungsempfehlungssysteme eröffnet. Darüber hinaus wird mit der Verhinderung der automatischen Gewinnung von Körpermaßen aus Bilddaten eine neue Herausforderung für weitere Forschungsvorhaben in den Bereichen Datenschutz und Bildverarbeitung definiert.

ACKNOWLEDGEMENTS

I want to thank my advisor Dr. Mario Fritz for the great support during my PhD study. Mario was a constant source of knowledge, inspiration and encouragement from the beginning to the end of my PhD. I want to thank people who I had chances to collaborate with: Prof. Andreas Bulling, Dr. Gerard Pons-Moll, Dr. Katharina Krombholz. Their experiences in research and also fruitful discussions helped me to improve the quality of my works. I also want to thank my thesis reviewer Bernt Schiele and Yusuke Sugano for the effort and time invested on reviewing my thesis. I want to give my special thanks to Prof. Bernt Schiele for building a friendly research environment for the whole D2 group. I am forever thankful to all my colleagues for their friendship and support, and for creating a fantastic working environment. I would like to especially thank Theimo Alldieck for his valuable insights on shape estimation models and for sharing his code with me. I would like to thank my parents, who provide me with the possibility to study and supported me in my life under any circumstances.

CONTENTS

1 INTRODUCTION

Man-Computer symbiosis was first introduced by (Licklider, 1960). Symbiosis is defined as "the living together in more or less intimate association or close union of two dissimilar organisms"[1]. In his work, he metaphorically compared this symbiosis system with fig tree which is pollinated only by the insect Blastophaga grossorum. As fig tree is dependent on the larva of an insect to be able to reproduce and the insect needs the tree to eat. Hence the insect and the tree constitute a productive and thriving partnership. Licklider envisioned the type of human-machine system in which human and machine become fluidly interdependent, sharing complementary abilities toward solving a shared goal that neither could achieve alone:

"The hope is that in not too many years, human brains and computing machines will be coupled together very tightly and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."

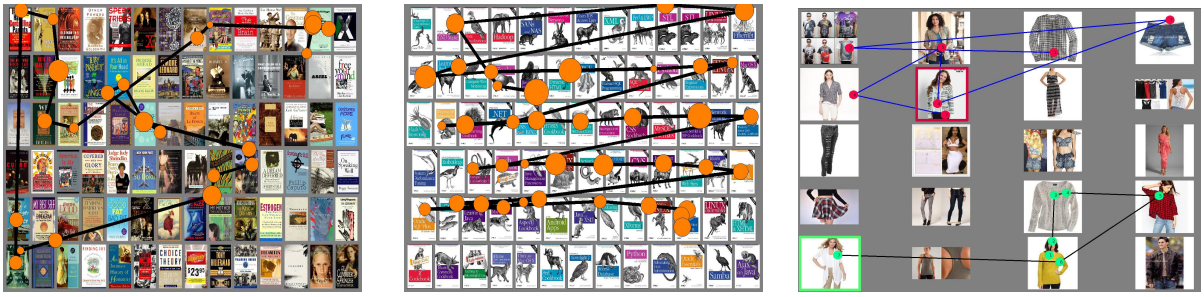
At the time a promising existing example of this symbiosis was the Semi-Automatic Ground Environment (SAGE), a system of large computers with associated networking equipment and human operators. Over 50 years later, we live in a world of human-machine symbiosis due to advances in physiological computing, biometrics, sensing technologies, and machine learning.

Today, in a **symbiotic relationship** machines develop solutions and make decisions based on the user's data. A **symbiotic system** helps users with information collected implicitly (Janlert and Stolterman, 2017; Verbeek, 2015; Gamberini and Spagnolli, 2016) using human bio-signal (Negri *et al.*, 2015) or from traces left by users of networked devices. Hence, users not necessarily need to be aware of what is happening. The symbiotic device, is able to learn and model the user in order to make decisions that simplify the users' task, clarifies possible issues or refines a service for them (e.g., recommendations or intent prediction), and can evolve in close relation with the environment by analysing the consequences of process jointly started with human (Jacucci *et al.*, 2015). Since the user does not need to provide the necessary information's to the system explicitly (e.g. rating system or clicking on the filters), could enjoy from an improved service. Hence, a symbiotic system is highly **'user-centered'**, which fits autonomously to user's need without the user need to ask. However, such a system raises serious security, ethics, legal and psychological risks to the user if designed incautiously (Spagnolli *et al.*, 2016).

Understanding the user intent and preferences is one of the main building blocks of symbiotic systems. In the last few years, several attempts have been made to understand human implicit intention and interest using techniques such

[1] Webster's New International Dictionaries

Figure 1.1: Gaze pattern of two observers searching for a same search target. Although both observer are looking for a same target, they have a very different search pattern.



(a) High variation in structure, context and colors

(b) Low variation in structure, context and colors

(c) Two observer searching for "Cardigan"

Figure 1.2: Gaze pattern of observers depends on the complexity of the scene and the search target. Furthermore, each individual could have different set of attributes for a share target category.

as Electroencephalography (EEG) (Zhang *et al.*, 2017c; Park *et al.*, 2014), Functional magnetic resonance imaging (fMRI) (Naselaris *et al.*, 2009; Nestor *et al.*, 2016), etc. Although EEG and fMRI could reveal a significant amount of information about users state of mind, tools used to acquire these types of information are expensive and not available for everyday use.

Eye movement is considered as a great source of information about a person's state of mind. Thanks to recent advances in eye tracking technologies, gaze data becomes practical and affordable to collect and use in many application scenarios of future interfaces (Sato *et al.*, 2016; Zhang *et al.*, 2015a). Nowadays, there exist many brands of eye trackers providing mobile or stationary tracking devices with high accuracy and precision.

In his seminal work from 1967, Yarbus showed that visual behaviour is closely linked to task when looking at a visual scene (Yarbus *et al.*, 1967). This work is an essential demonstration of task influence on fixation patterns and sparked a large number of follow-up works in a range of disciplines, including human vision, neuroscience, artificial intelligence, and computer vision. A common goal in these

human and computer vision works is to analyse visual behaviour, i.e. typically fixations and saccades, to make predictions about user behaviour or more recently predicting the target of visual search (Zelinsky *et al.*, 2013; Borji *et al.*, 2014). Visual search is a complicated daily life task of a person. During the search, human tends to fixate on objects which share visual features similar to those of the search target (Wolfe and Horowitz, 2017). However, each observer could have their own individual set of target features to find a shared search target. Figure 1.1 presents gaze data of two different users searching for the same target. Each observer had their approach (scan-path) to find the target. As one can see, although they both found the search target, they have very different fixation patterns.

Another common issue relates to the complexity of the scene and the target. As the complexity of the search target increases, only sub-patterns will guide observers' attention. In Figure 1.2, we present two different search set. In the first scene (Figure 1.2(a)) the sub-images of each collage has a higher variation in colour, shape and texture. Hence, fixated patches have more discriminative features. Whereas in the other collage (Figure 1.2(b)) the sub-images shares higher similarities in texture and colour. As one can see, the observer skims through the images to find the target rather than fixating on patches with discriminative features.

Moreover, users might search for a common category with a different set of attributes. Example of these scenario is presented in Figure 1.2 (c). Both observers were searching for "carding". However, one user found a cardigan with properties such as short and grey (red box) whereas the cardigan found by other user is white and long (green box).

Thus, to build a model for search target prediction, we need to know the answer to these several fundamental questions such as: what are the most critical algorithmic principles for this task? How could the complexity of the search target affect the prediction power of the model? How much gaze data is needed for such a model? To what level such a model are robust to noise in the input gaze data? Moreover, the inter and intra user dependencies of such models. In this thesis, we focus on addressing these challenges. First, we introduced a compatibility measure to learn features similarities between the search target and gaze data during the visual search on a limited set of targets. We further present a gaze embedding model that could generalised search target prediction over a broad set of targets.

Body shape of a person could affect a person's interest and decisions. A person shape could influence their interest in the type of clothing items they would buy, the food they would eat, or amounts of sports they would do. Hence, we used 3D shape data to understand how body shape correlates with people's clothing preferences. The main challenge here is to build a model that could obtain an accurate 3D shape estimate from an unconstrained 2D image of a fully clothed person. Getting such a realistic estimate from a 2D image is very challenging. Figure 1.3 presents several of these challenges. People could appear in different poses, wearing many different types of garments, standing at a different distance to the camera, and photos are taken from different camera viewpoints. We address these challenges by introducing a robust multi-photo approach that can obtain a realistic 3D shape even from 2D

Figure 1.3: Web photos of a person wearing different clothing items. The person in the image appears in different poses and viewpoint. Furthermore, the person is standing at different distances to the camera. These variations cause difficulties in having an accurate 3D shape estimates from 2D images.

web images of people. However, the automated extraction of 3D shape data from regular, readily available images might equally raise concerns about users' privacy. Hence, we present an approach to prevent multi-stage shape estimation techniques to acquire such a realistic estimation.

1.1 THESIS CONTRIBUTIONS

The core contribution of the thesis is to build a system which could do prediction using implicit human cues. We used gaze data to predict the intent of the visual search and shape to infer the clothing preferences of a user. In the following, we detail the challenges involved in these tasks, as well as the contributions this thesis makes to address them.

1.1.1 Gaze Based Inference

The first part of this thesis is dedicated to predicting the target of visual search from gaze data. Predicting the target of a visual search task is particularly interesting, as the corresponding internal representation, the mental image of the search target, is difficult if not impossible to assess using other modalities. While there exist several attempts (Zelinsky *et al.*, 2013; Borji *et al.*, 2014) in predicting the search target of users, these works are limited to a close set of targets where all potential search targets are part of the training set, and fixations for all of these targets were observed. This excludes searches for broader classes of objects that share the same semantic category or specific object attributes. Such searches commonly occur if the user does not have a concrete target instance in mind but is only looking for an object from a specific category or with particular attributes.

To address these limitations, we first introduced a *open-world setting* in which we no longer assume that we have fixation data to train for these targets. We further

broaden the scope of search target prediction to categorical classes, such as object categories or attributes by the introduction of *Gaze Pooling Layer*. This Pooling Layer utilised readily trained CNN architectures and combines them with gaze data. The gaze information is used as an attention mechanism that acts selectively on the visual features to predict users' search target. These design choices make our approach compatible and practical with current deep learning architectures.

We even went one step further and investigate the feasibility of combining gaze pooling layer with generative image models to visually decode, i.e. create a visual representation of, the search target. Such visual decoding is challenging for two reasons: 1) the search target only resides in the user's mind as a subjective visual pattern, and can most often not even be described verbally by the person, and 2) it is, as of yet, unclear if gaze fixations contain sufficient information for this task at all. We show, for the first time, that visual representations of search targets can indeed be decoded only from human gaze fixations.

The main contributions of this part are:

- We present an annotated dataset of human fixations on synthesised collages of natural images during the visual search that lends itself to studying our new open-world setting. Compared to previous works, our dataset is more challenging because of its larger number of distractors, higher similarities between search image and distractors, and a more significant number of potential search targets.
- We introduce a novel problem formulation and method for learning the compatibility between observed fixations and potential search targets.
- Using this dataset, we report on a series of experiments on predicting users' search target from fixations by moving from closed-world to open-world settings.
- We present an annotated dataset of human fixations on synthesised collages of 10 clothing category and ten attributes during the visual search that allows us to study our new *Gaze Pooling Layer*.
- We propose an approach for predicting categories and attributes of search targets that utilise readily trained CNN architectures and combines them with gaze data in a novel *Gaze Pooling Layer*.
- Through extensive experiments, we show that our method achieves accurate search target prediction for ten category and ten attribute tasks on our new gaze data set.
- We evaluate different parameter settings and design choices of our Gaze Pooling Layer, visualise internal representations and perform a robustness study w.r.t. noise in the eye tracking data.
- We present the first proof of concept that visual representations of search targets can be decoded from human gaze data.

- We present a practical approach, as it respects the difficulties in collecting large human gaze datasets. Encoder and decoders are trained from large image corpora, and a semantic layer facilitates transfer between the two representation in between.

- We show the importance of localised gaze information for improved search target reconstruction.

1.1.2 Shape Based Inference

In the second part of this thesis, we used 3D shape data to understand how body shape correlates with people's clothing preferences. Many e-commerce companies, such as Amazon or Zalando, makes it possible for their users to buy clothing online. However, based on a recent study, [2] around 50% of bought items were returned by users. One major reason for return is "It doesn't fit" (52%). Fit goes beyond the mere size — certain items look good on certain body shapes, and others do not. Consequently, understanding how body shape correlates with people's clothing preferences could avoid such confusions and reduce the number of returns. To study the correlation between clothing garments and body shape, we collected a new dataset (*Fashion Takes Shape*), which includes images of female users with clothing category annotations. Despite the progress in body shape estimation from pictures, it turns out to be challenging to infer body shape from such diverse, real-world photos. Hence, we propose a novel and robust *multi-photo approach* to estimate body shapes of each user and build a conditional model of clothing categories given body-shape. Our body shape estimate could provide us with a realistic depiction of the users' naked body. However, such representation is considered highly private and therefore might not be consented by most people. Hence, we ask if the automatic extraction of such information can be effectively evaded. While adversarial perturbations are sufficient for manipulation, the output of a range of machine learning models - in particular, end-to-end deep learning approaches - state of the art shape estimation methods are composed of multiple stages. We perform the first investigation of different strategies that can be used to effectively manipulate the automatic shape estimation while preserving the overall appearance of the original image.

The main contributions of this part are:

- Introducing Fashion Takes Shape data set.

- A Robust multi-photo shape estimation techniques with an increase in accuracy of 3D shape estimates.

- The first work which studied the relationship between clothing categories and body shape.

[2] https://www.ibi.de/files/Competence%20Center/Ebusiness/PM-Retourenmanagement-im-Online-Handel.pdf

- A motivational user study to get a rough estimate of users' concerns w.r.t. privacy and body shape estimation in different application contexts.
- An evasion attack to mitigate privacy violations via body shape estimation.
- Analysis of synthetic attacks on human skeleton joints.
- Evaluation of practicability and effectiveness of the real attack on skeleton joints.
- New localised attack on human skeleton joints feature maps that show smaller norm at the same effectiveness.
- Evaluation of overall effectiveness of different attacks strategies on shape estimation. We show the first successful attacks that offer an increase in privacy with negligible loss in visual quality.

1.2 OUTLINE OF THE THESIS

In this section, we summarise each chapter of the thesis. We also indicate the respective publications and connections with other previous works.

Chapter 2: Related Works. In this chapter, we review the related works on visual search target prediction, outfit recommendation, 3D shape estimation, and privacy in computer vision. We analyse the relations of previous and subsequent works to the research presented in this thesis.

Chapter 3: Search Targets Prediction From Fixations in Open-World. In this chapter, we go beyond state of the art by studying search target prediction in an open-world setting in which we no longer assume that we have fixation data to train for the search targets. We then present a new problem formulation for search target prediction in the open-world setting that is based on learning compatibilities between fixations and potential search targets. We present a dataset containing fixation data of 18 users searching for natural images from three image categories within synthesised image collages of about 80 images. The content of this chapter corresponds to the CVPR 2015 publication "Prediction of Search Targets From Fixations in Open-World Settings (Sattar *et al.*, 2015). Hosnieh Sattar was the lead author.

Chapter 4: Search Intent Prediction Using Deep Gaze Pooling. In this chapter, we propose the first approach to predict the categories and attributes of search targets based on gaze data. We propose a novel *Gaze Pooling Layer* that integrates gaze information into CNN-based architectures as an attention mechanism – incorporating both spatial and temporal aspects of human gaze behaviour. We present an annotated dataset of human fixations on synthesised collages of 10 clothing category and ten attributes during the visual search that allows us to

study our new *Gaze Pooling Layer*. The content of this chapter corresponds to the ICCVW 2017 publication "Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling" (Sattar *et al.*, 2017). Hosnieh Sattar was the lead author.

Chapter 5: Visual Search Target Decoding From Human Eye Fixations. In this chapter, we go one step further and investigate the feasibility of combining the *Gaze Pooling Layer*, which encodes human gaze information using deep convolutional neural networks with the power of generative image models to decode visually, i.e. create a visual representation of, the search target. The content of this chapter corresponds to the arXiv 2017 publication "Visual Decoding of Targets During Visual Search From Human Eye Fixations" (Sattar *et al.*, 2019a). Hosnieh Sattar was the lead author.

Chapter 6: Understanding Clothing Preference Based on Body Shape. In this chapter, we study the correlation between clothing garments and body shape on our new Fashion Takes Shape dataset. We introduced a new multi-photo approach for 3D shape estimation and build a conditional model of clothing categories given body-shape. The content of this chapter corresponds to the WACV 2019 publication "Fashion is Taking Shape: Understanding Clothing Preference Based on Body Shape From Online Sources" (Sattar *et al.*, 2019c).

Chapter 7: Shape Evasion: Preventing Body Shape Inference. In this chapter, we introduced an evasion method that can be used to effectively manipulate the automatic shape estimation while preserving the overall appearance of the original image. We performed a user study to understand users' concerns w.r.t. privacy depiction of the 3D body shape in public and different application contexts. The content of this chapter corresponds to the arXiv 2019 publication "Shape Evasion: Preventing Body Shape Inference of Multi-Stage Approaches" (Sattar *et al.*, 2019b)

Chapter 8: Conclusions and Future Prospects. In this chapter, we summarise this thesis' contributions and discuss current limitations, as well as possible directions to overcome the limitations. Besides, we provide an outlook on future work and discuss future directions.

RELATED WORKS 2

VISUAL search is part of daily human life. It is goal oriented and involves active scanning of the environment and directing attention to objects that might be the target. Understanding how human attention is guided during visual search and prediction of the user intent is a hot topic of research among cognitive and computer vision scientist. One of the fundamental studies has done by (Treisman and Gelade, 1980) on the feature integration theory of visual attention. This theory suggested that whenever there exists more than one separable feature to characterise or distinguish the present object, attention is directed serially to each stimulus. Many followed up studies attempt to testify this theory and were concerned with mainly four important questions (Müller and Krummenacher, 2006):

1. The bottom-up and top-down mechanisms in guiding the visual search.
2. The role of implicit and explicit memory mechanism.
3. How these mechanisms are implemented in the brain.
4. Simulation of visual search processes in neurocomputational models.

Wolfe and Horowitz (2017), state that five factors guide attention during visual search, such as bottom-up salience, top-down feature guidance, scene structure and meaning, the previous history of search and the relative value of target and distractors. Role of attention and human visual behaviour during search has been extensively studied over the past 40 years. For a detailed review we would refer the reader to Müller and Krummenacher (2006) and Wolfe and Horowitz (2017). Visual search in computer vision created the field of content-based image retrieval (CBIR). This term was first used by Gudivada and Raghavan (1995), to describe a system that automatically retrieves images from a database based on the presented colour and shape features. Since then, many CBIR systems have been developed. However, retrieving images based on their pixel content is still an open research problem (Lew *et al.*, 2006). To predict the search intent of the users, researchers in both areas start using explicit or implicit human cues, such as reaction time (RT) to detect a target, gaze data, pupil size, fixation duration, saliency maps. Meanwhile, CBIR systems get equipped by users data such as click-through data and mouse movement to detect user intent.

In the remaining of this chapter, we revisit related works focusing on the directions explored in this thesis. In Section 2.1, we give a brief introduction to the implicit and explicit way of communication and type of implicit data. We then review previous works on search target and task prediction using gaze (subsection 2.2.3), body shape estimation methods and clothing preference prediction in section 2.3.

In the last section (section 2.4), we cover related works on privacy aspects of implicit data and application of privacy in computer vision.

2.1 EXPLICIT VS. IMPLICIT COMMUNICATION

Human communicate with each other in explicit and implicit ways. However, only a small fraction of this communication is in the form of explicit (McDonald and Flanagan, 2004). The human ability to decode cognitive states of another person makes human effective at communication and collaboration (Frith and Frith, 2006). Such decoding is not done via explicit information instead from plenty of implicit data such as gaze behaviour, gesture, speech tune, and so on.

Recent communication techniques between human and computer rely majorly on explicit inputs. In explicit communication, users need to tell a computer what they expect from the computer to do for them. This type of communication is asymmetrical since the computer does not have access to users mental state. The communication is done via typing a command in a terminal, direct manipulation in a graphical user interface with a mouse, through gesture or speech. In this case, the actions are certain and discontinue (Schmidt, 2000). Explicit communications (traditional), assumes that human is explicit, unambiguous and fully attentive while controlling the information and command flow (Pantic *et al.*, 2007).

To have a symmetrical human-computer communication, the device needs to have access to the user's mental status. In such a system, the computer can understand the internal state of the human user and use this information to improve the interaction and adapt to the user need (Schmidt, 2000; Wagner *et al.*, 2013). The implicit data comes for free and in huge amount as the primary goal of the action is not to interact with the computerised system. However, the system understands it as input. Multi-modal wearable devices are used to acquire implicit data such as Psychophysiological measurement and Gaze data. Computer vision techniques could be used to detect facial expression, gesture, shape and pose. Mouse movement, search history, purchase history, user profiles on social media, time spending on different web pages are another source of implicit human data.

Implicit Vs. Explicit Data In Recommendation Systems. Recommendation and information retrieval systems are two examples that move from explicit only towards adding implicit data of users into their model. A recommendation system provides personalised suggestions to the user via collecting information on the user preferences for a set of items(e.g. movies, songs, books, travel destination, clothing item) (Bobadilla *et al.*, 2013). Explicit rating is used to determine the interest of the users. In explicit rating, users tell what they think about a piece of information by rating it into one or more ordinal or qualitative scales. These ratings are well understood and reasonably precise (Watson and Sasse, 1998). However, an explicit rating is known to have several issues. First, stopping a user to enter an explicit rating will alter the regular pattern of browsing (Claypool *et al.*, 2001). The other major problem is that

the collaborative filtering techniques used in recommendation systems usually need many rating for each item to have accurate prediction (i.e. sparsity problem) (Sarwar *et al.*, 1998). The user only provides rating if they see a benefit in doing it (Grudin, 1995; Sparling and Sen, 2011). Since the user may continue using the system without entering any rating (Herlocker *et al.*, 2004). Research on GroupLens system (Sarwar *et al.*, 1998), also found that users tend to read more articles and do fewer rating. Recent recommendation systems use implicit data to build their rating system, such as the number of play counts of a song, or clicks on webpages (Oard *et al.*, 1998; Joachims *et al.*, 2005), geographical location of person (Bouros *et al.*, 2015), mouse movement and scrolling elapsed time (Claypool *et al.*, 2001; Kim and Chan, 2005), and most recently gaze data (Song and Moon, 2019; Castagnos *et al.*, 2010; Silva *et al.*, 2018). Using implicit rating removes the burden from user examining and rating items. Also, every interaction with the system can contribute to the rating system (Claypool *et al.*, 2001).

Implicit Vs. Explicit Data In Information Retrieval Systems. An information retrieval system which relies on explicit user input assumes that users know exactly what they are searching for or they have sufficient knowledge domain to formulate their queries (Tripathi *et al.*, 2019). However, most often, the user has difficulties expressing what they are searching for (Chowdhury *et al.*, 2011). This can be seen specifically in the exploratory search, where users are trying to discover new information and interesting items and may struggle without additional support (Athukorala *et al.*, 2016a,b; Marchionini, 2006; Medlar and Glowacka, 2018). There are several attempts to close such a semantic gap such as using free-hand human sketches as queries to perform instance-level retrieval of images (Yu *et al.*, 2016). Ferecatu and Geman (2009) used mouse clicks and A. Kovashka (2012) used a set of attributes and required users to operate on a large attribute vocabulary to describe their mental images. In Yu *et al.* (2016) the feedback was provided by sketching the target to convey concepts such as texture, colour, material, and style, which is a non-trivial step for most users. Providing these explicit user inputs if not impossible (such as sketches) is cumbersome and time-consuming and not compatible with the user-centered era of communication in which computer is expected to understand users with few or no effort from user side.

Implicit data are shown to be useful in information retrieval systems as well. Human gaze behaviour reflects cognitive processes of the mind, such as intentions (Brigham *et al.*, 2001; Kleinke, 1986; Land and Furneaux, 1997), and is influenced by the user's task (Yarbus *et al.*, 1967). Hence, several works started to use gaze behaviour to understand users decision making behaviour (Joachims *et al.*, 2005; Granka *et al.*, 2004) to improve their retrieval algorithm. Miller and Agne (2005) used eye tracking to detects users attention on words while reading a text to retrieve relevant text, McNamara *et al.* (2019) provided content based information based on user attention in virtual reality. Sun *et al.* (2019) used eye tracking data as relevance feedback to retrieve face images.

Implicit Data Types. Implicit data can come in a huge variety. However, to use implicit data, we need to know what is communicated, how information is passed on, in which context the information passed on and which reaction should be taken to satisfy users need (Pantic *et al.*, 2007). Users facial activity (Ekman, 1977), vocal expression and non-linguistic expression (Juslin and Scherer, 2005; Russell *et al.*, 2003), language and choice of words (Furnas *et al.*, 1987) are several sources of implicit data which is mainly used to model human emotion in order to build anticipatory ambient interfaces. Atomic facial signals and action units (AU) were used to recognise basic emotion (Ekman, 1977), interest, disagreement and puzzlement (Cunningham *et al.*, 2004), suicidal depression or pain (Ekman, 1997), and social signals like emblems, regulators, and illustrators (Ekman and Friesen, 1969). Auditory features like pitch, intensity, and speech rate (Pierre-Yves, 2003; Pantic and Rothkrantz, 2003) was used for discrete emotion recognition.

Information such as users purchase history, click through data, mouse movement, visited web pages, and amount of time spend on web pages are another source of implicit data used to model user intent in web search.

Psychophysiological measurement such as electroencephalogram (EEG), electrodermal activity (EDA), heart rate (electrocardiogram (ECG)), respiration rates, bold oxygenation level dependent(BOLD) response is known to be an indication of user states and intent. Among these measurements, EEG contains a higher amount of information about the user's intent. However, this signal usually suffers from low signal to noise ration (Spapé *et al.*, 2015; Luck, 2014), it contains artefacts (such as eye blink or body part movement) (Muthukumaraswamy, 2013; Whitham *et al.*, 2007), needs specific experimental design, and is difficult to analyse (Lemm *et al.*, 2011).

Human gaze behaviour is shown to be a reliable source of information about human intentions (Brigham *et al.*, 2001; Kleinke, 1986; Land and Furneaux, 1997), and task (Yarbus *et al.*, 1967) as well. Thanks to advances in eye tracking technology, acquisition of gaze data become very easy and accessible for daily use. New eye-tracking devices provide gaze data with high accuracy and precision in stationary and mobile settings. Hence, a growing body of computer vision and human-machine interaction literature starts to employ gaze.

2.2 VISUAL SEARCH TARGET PREDICTION USING GAZE

Several researchers recently aimed to reproduce Yarbus's findings and to extend them by automatically predicting the observers' tasks. Green et al. reproduced the original experiments, but although they were able to predict the observers' identity and the observed images from the scan paths, they did not succeed in predicting the task itself (Greene *et al.*, 2012). Borji et al., Kanan et al., and Haji-Abolhassani et al. conducted follow-up experiments using more sophisticated features and machine learning techniques (Borji and Itti, 2014; Haji-Abolhassani and Clark, 2014; Kanan *et al.*, 2014). All three works showed that the observers' tasks could be successfully predicted from gaze information alone.

Other works investigated means to recognise more general aspects of user behaviour. Bulling et al. investigated the recognition of everyday office activities from visual behaviours, such as reading, taking hand-written notes, or browsing the web (Bulling *et al.*, 2011). Based on long-term eye movement recordings, they later showed that high-level contextual cues, such as social interactions or being mentally active, could also be inferred from visual behaviour (Bulling *et al.*, 2013). They further showed that cognitive processes, such as visual memory recall or cognitive load, could be inferred from gaze information (Bulling and Roggen, 2011; Tessendorf *et al.*, 2011) as well – of which the former finding was recently confirmed by Henderson et al. (Henderson *et al.*, 2013).

Subjects' gaze patterns during categorical search such as the number of fixations made prior to search judgements as well as the percentage of first eye movements landing on the search target were predicted by Zelinsky et al (Neider and Zelinsky, 2006; Zhang *et al.*, 2005; Chen and Zelinsky, 2006). They later showed how to predict the categorical search targets themselves from eye fixations Zelinsky *et al.* (2013). Borji et al. focused on predicting search targets from fixations (Borji *et al.*, 2014). In three experiments, participants had to find a binary pattern and 3-level luminance patterns out of a set of other patterns, as well as one of 15 objects in 11 synthetic natural scenes. They showed that binary patterns with higher similarity to the search target were viewed more often by participants. Additionally, they found that when the complexity of the search target increased, participants were guided more by sub-patterns rather than the whole pattern.

The works of Zelinsky et al. (Zelinsky *et al.*, 2013) and Borji et al. (Borji *et al.*, 2014) are most related to ours. However, both works only considered simplified visual stimuli or synthesised natural scenes in a closed-world setting. In that setting, all potential search targets were part of the training set, and fixations for all of these targets were observed. In Chapter 3, we introduced the open-world setting in which we no longer assume that we have fixation data to train for these targets, and we presented a new problem formulation for this open-world search target recognition. Although we could predict the search target in an open world setting, this method is still limited to predicting a specific instance of a search query. In Chapter 4, we aim to infer the general properties of a search target represented by the object's category and attributes. In this scenario, the search task is guided by the mental model that the user has for an object class rather than a specific instance of an object (Ferecatu and Geman, 2009; Wilson *et al.*, 2008). This presents additional challenges as mental models might differ substantially among subjects. In contrast to previous works which required gaze data for training, our approach can be pre-trained on visual data alone, and then combined with gaze data at test time.

2.2.1 Image generation and multi-modal learning

Due to recent advances in representation learning and convolution neural networks, image generation becomes possible. Recently, generative adversarial networks (GANs) (Goodfellow *et al.*, 2014a; Reed *et al.*, 2016; Denton *et al.*, 2015; Isola *et al.*,

2017; Pathak *et al.*, 2016) were used to generate realistic and novel images. GANs consists of two parts: a generator and a discriminator. The discriminator is designed to discriminate between generated images and training data. However, training GANS is a challenging task due to the min-max objective. A stochastic variational inference and learning algorithm was introduced by Kingma and Welling (2013). A lower bound estimator is achieved via re-parameterization of the variational lower bound. Consequently, the standard stochastic gradient method can be used to optimize the estimator. However, the posterior distribution of latent variables is usually unknown. Yang et al. introduced a general-optimization based approach that uses image generation models and latent priors for posterior inference (Yan *et al.*, 2016). They generated images conditioned on visual attributes. In Chapter 5, we employ their idea of conditional generative models in the context of generating the search intends from gaze data.

2.2.2 Visual Experience Reconstruction using fMRI

Recent developments in functional magnetic resonance imaging (fMRI) make it possible for neuroscientists to generate links between brain activity and the visual world. Nishimoto et al. reconstructed natural movies from brain activity (Nishimoto *et al.*, 2011). They proposed a motion energy encoding methods to decode the fast visual information and BOLD signals in occipitotemporal visual cortex and fit the model separately to individual voxels.

In another work, Cowen et al. proposed to reconstruct human faces from evoked brain activity using multi-variant regression and PCA (Cowen *et al.*, 2014). In their experiment, they asked participants to look at an image and then tried to reconstruct this specific image from fMRI data. All of the above tasks tried to reconstruct the *seen* images. In contrast, in Chapter 5, we demonstrate that using gaze data, we can decode the *visual search target* of user's which only resides in the user's mind. Also, we are not using fMRI but gaze data, which is arguably more practical and affordable to collect and use.

2.2.3 Gaze-Supported Computer Vision

The content of Chapter 3, and Chapter 4, and Chapter 5, is related to an increasing body of computer vision literature that employs gaze as a means to provide supervision or indicate Salient regions in the image in a variety of recognition tasks. Visual fixations have been used in Li *et al.* (2015); Xu *et al.* (2014) to indicate object locations in the context of saliency predictions, and in Karthikeyan *et al.* (2013); Papadopoulos *et al.* (2014); Shcherbatyi *et al.* (2015) as a form of weak supervision for the training of object detectors. Gaze information has been used to analyze pose estimation tasks in Marinoiu *et al.* (2013); Subramanian *et al.* (2011) as well as for action detection (Mathe and Sminchisescu, 2014). Gaze data has also been employed for active segmentation (Mishra *et al.*, 2009), localizing important objects in egocentric videos Damen *et al.* (2014); Toyama *et al.* (2012), image captioning and scene understanding (Sugano and

Bulling, 2016), as well as zero-shot image classification (Karessli *et al.*, 2017).

2.2.4 User Feedback for Image Search and Retrieval.

To close the semantic gap between a user's envisioned search target and the images retrieved by search engines, Ferecatu and Geman (2009) proposed a framework to discover the semantic category of user's mental image in unstructured data via explicit user input. A. Kovashka (2012) introduced a novel explicit feedback method to assess the mental models of users. Most recently, Yu *et al.* (2016) proposed to use free-hand human sketches as queries to perform instance-level retrieval of images. They considered these sketches to be manifestations of users' mental model of the target. The common theme in these approaches is that they require explicit user input as part of their search refinement loop. Mouse clicks were used as input in Ferecatu and Geman (2009). A. Kovashka (2012) used a set of attributes and required users to operate on a large attribute vocabulary to describe their mental images. In Yu *et al.* (2016) the feedback was provided by sketching the target to convey concepts such as texture, colour, material, and style, which is a non-trivial step for most users.

Several previous works investigated the use of gaze information as an implicit measure of relevance in image retrieval tasks. For example, Oyekoya and Stentiford compared similarity measures based on a visual saliency model as well as real human gaze patterns, indicating better performance for gaze (Oyekoya and Stentiford, 2004). In later works, the same and other authors showed that gaze information yielded significantly better performance than random selection or using saliency information (Oyekoya and Stentiford, 2007; Schulze *et al.*, 2013). Coddington presented a similar system but used two separate screens for the task (Coddington *et al.*, 2012) while Kozma et al. focused on implicit cues obtained from gaze in real-time interfaces (Kozma *et al.*, 2009). To make implicit relevance feedback richer, Klami proposed to infer which parts of the image the user found most relevant from gaze (Klami, 2010).

In contrast, in our work (Chapter 3 and Chapter 4), we do not rely on a feedback loop as in A. Kovashka (2012), or a relevance measure as in (Klami, 2010; Oyekoya and Stentiford, 2007; Schulze *et al.*, 2013; Coddington *et al.*, 2012; Kozma *et al.*, 2009) or explicit user input or some form of initial description of a target as in A. Kovashka (2012); Ferecatu and Geman (2009); Yu *et al.* (2016). In Chapter 4, we show only using fixation information that can be acquired implicitly during the search task itself, allows us to predict categories as well as attributes of search targets in a single search session.

2.3 PREFERENCE PREDICTION

In this section, we cover previous works related to the second part of this thesis. The amount of information available on the web is growing daily. Users of web services are often facing a countless number of products, restaurants, movie or clothing. A

recommender systems can predict the preferences of their users and suggest users proactively items that they might like (Ricci *et al.*, 2015; Adomavicius and Tuzhilin, 2005). They play a vital role in information access and facilitate the decision-making process and boosting business (Jannach *et al.*, 2010). Such a system has been widely used by companies such as Netflix or YouTube. 80 % of watched movies on Netflix (Gomez-Uribe and Hunt, 2016), and 60% of video clicks on YouTube (Davidson *et al.*, 2010) came from recommendation systems. One of the areas that highly takes advantage of a recommendation system is fashion. In the following, we cover related works in the field of computer vision and fashion.

2.3.1 Fashion Understanding in Computer Vision.

Recently, fashion image understanding has gained a lot of attention in computer vision community, due to large range of its human-centric applications such as clothing recommendation (Kiapour *et al.*, 2014; Simo-Serra *et al.*, 2015; Liu *et al.*, 2016a; Han *et al.*, 2017b; Hsiao and Grauman, 2018), retrieval (Wang and Zhang, 2011; Yamaguchi *et al.*, 2013; Kiapour *et al.*, 2015; Liu *et al.*, 2012; Ak *et al.*, 2018), recognition (Chen *et al.*, 2012; Liu *et al.*, 2016a; Hsiao and Grauman, 2017; Han *et al.*, 2017a; Al-Halah *et al.*, 2017), parsing (Yamaguchi *et al.*, 2012; Yang *et al.*, 2014) and fashion landmark detection (Liu *et al.*, 2016a,b; Wang *et al.*, 2018).

Whereas earlier work in this domain used handcrafted features (e.g. SIFT, HOG) to represent clothing (Chen *et al.*, 2012; Kiapour *et al.*, 2014; Wang and Zhang, 2011; Liu *et al.*, 2012), newer approaches use deep learning (Bo Zhao, 2017) which outperforms prior work by a large margin. This is thanks to availability of large-scale fashion datasets (Simo-Serra *et al.*, 2015; Liu *et al.*, 2016a,b; Han *et al.*, 2017a; Rostamzadeh *et al.*, 2018) and blogs. Recent works in clothing recommendation leverage metadata from fashion blogs. In particular, Han *et al.* (2017a) used the fashion collages to suggest outfits using multimodal user input. Zhang *et al.* (2017d) studied the correlation between clothing and geographical locations and introduced a method to recommend location-oriented clothing automatically. In another work, Simo-Serra *et al.* (2015) used user votes of fashion outfits to obtain a measure of fashionability. Vasileva *et al.* (2018); Cucurull *et al.* (2019) and Han *et al.* (2017b), predicted whether a set of fashion items goes well together, this task is called compatibility prediction.

Although the relationship between location, users' vote and fashion compatibilities is well investigated, there is no work which studies the relationship between human body shape and clothing. In Chapter 6, we introduce an automatic method to estimate the 3D shape and a model that relates it to clothing preferences. We also introduce a new dataset to promote further research in this direction.

2.3.2 3D Body Shape Estimation.

Recovery of 3D human shape from a 2D image is a very challenging task which has been facilitated by the availability of 3D generative body models learned from

thousands of scans of people (Anguelov *et al.*, 2005; Pons-Moll *et al.*, 2015; Loper *et al.*, 2015). Such models capture anthropometric constraints of the population and therefore reduce ambiguities. Several works (Sigal *et al.*, 2008; Guan *et al.*, 2009; Hasler *et al.*, 2010; Zhou *et al.*, 2010; Chen *et al.*, 2010; Bogo *et al.*, 2016; Huang *et al.*, 2017) leverage these generative models to estimate 3D shape from single images using shading cues, silhouettes and appearance.

Recent model-based approaches leverage deep learning based 2D detections (Cao *et al.*, 2017) – by either fitting a model to them at test time (Bogo *et al.*, 2016; Alldieck *et al.*, 2018) or by using them to supervise bottom-up 3D shape predictors (Pavlakos *et al.*, 2018; Kanazawa *et al.*, 2018; Tung *et al.*, 2017; Tan *et al.*, 2017). Similar to Bogo *et al.* (2016), we fit the SMPL model to 2D joint detections, but, to obtain better shape estimates, we include a silhouette term in the objective like Alldieck *et al.* (2018); Huang *et al.* (2017). In contrast to previous work, we leverage multiple web photos of the same person in different poses. In particular, we jointly optimize a single coherent static shape and pose for each of the images. This makes our multi-photo shape estimation approach robust to challenging poses and shape occluded by clothing. Other works have exploited temporal information in a sequence to estimate shape under clothing (Bălan and Black, 2008; Zhang *et al.*, 2017a; Yang *et al.*, 2016) in constrained settings – in contrast, we leverage web photos without known camera parameters. Furthermore, we can not assume pose coherency over time (Huang *et al.*, 2017) since our input are photos with various poses. None of the previous works leverages multiple unconstrained pictures of a person to estimate body shape. This approach is introduced in Chapter 6.

2.3.3 Virtual try-on:

Another popular application of computer vision and computer graphics to fashion is virtual try-on, which boils down to a clothing re-targeting problem. Pons-Moll *et al.* (2017) jointly capture body shape and 3D clothing geometry – which can be re-targeted to new bodies and poses. Other works bypass 3D inference; using simple proxies for the body, Han *et al.* (2017c) retarget clothing to people directly in image space. Using deep learning and leveraging SMPL (Loper *et al.*, 2015), Lassner *et al.* (2017a) predicts images of people in clothing, and Zanfir *et al.* (2018) transfer appearance between subjects. These works leverage a body model to re-target clothing but do not study the correlation between body shape and clothing categories.

2.4 PRIVACY ASPECTS OF IMPLICIT DATA COLLECTION

Implicit data are easier to collect compared to explicit data. Explicit data are collected by asking explicitly to disclose data while implicit data collection is known as behind the curtain approach (Huang and Lin, 2005). Implicit data is collected without direct involvement or consent of individuals. Implicit data are very important for

organisations as it provides them with information about their customer's style of life, shopping habits as well as psychological characteristics (Reed, 1999; Robertshaw and Marr, 2006). Such insights enable organisations to expand their customer base through customised advertising and more personalised products (Deighton and Sorrell, 1996), causing personalisation privacy trade-off (Lee *et al.*, 2011)

Hence, ethical issues arise because of the potential non-voluntary disclosure of implicit data that disregards the fundamental right of privacy of the individuals (Spagnolli *et al.*, 2016) — thus introducing a new topic of research in several fields such as ethics, law, information security, computer vision, human-computer interactions and psychology.

2.4.1 Privacy and Computer Vision

Recent developments in computer vision techniques increases concerns about extraction of private information from visual data such as age (Bauckhage *et al.*, 2010), social relationships (Wang *et al.*, 2010), face detection (Sun *et al.*, 2017; Viola and Jones, 2001), landmark detection (Zheng *et al.*, 2009), occupation recognition (Shao *et al.*, 2013), and license plates (Zhou *et al.*, 2012; Zhang *et al.*, 2006; Chang *et al.*, 2004).

Hence several studies on keeping the private content in visual data began only recently, where hardware or software was used to protect private users' data. Orekondy *et al.* (2017) propose a visual privacy advisor that could predict users specific privacy preferences of image content. In the follow-up work, Orekondy *et al.* (2018) introduced a model for the automatic redaction of different private information in images. To evaluate their model, they collected a dataset of private images "in the wild", which was annotated with pixel and instance level labels across a broad range of privacy classes. Wilber *et al.* (2016), and Harvey (2012), studied different techniques that could prevent faces from being detected and recognised. Oh *et al.* (2016) studied privacy implications of visual data shared on social media by analysing how well people are recognisable.

As wearable devices that continuously capture video becomes widespread, it increases the risk that sensitive object/region will appear accidentally in videos, or a person will get captured unwillingly. Hence, several works proposed a framework to protect private data in videos. Privacy-preserving video capture frameworks was introduced in (Aditya *et al.*, 2016; Pittaluga and Koppal, 2015; Neustaedter *et al.*, 2006; Raval *et al.*, 2014). privacy-Aware eye tracking system was introduced in Steil *et al.* (2018a,b). Hoyle *et al.* (2015) explored people privacy preferences on collected images of lifelogging cameras. Korayem *et al.* (2016) proposed a model to detect computer screens in photo lifelogs.

None of the previous work in this domain studied the users shape privacy preferences. Hence, in Chapter 7, we present a new challenge in computer vision aimed at preventing automatic 3D shape extraction from images.

2.4.2 Adversarial Image Perturbation

Adversarial examples for deep neural networks were first reported in Szegedy *et al.* (2014); Goodfellow *et al.* (2014b) demonstrating that deep neural networks are being vulnerable to small adversarial perturbations. This phenomenon was analyzed in several studies (Arnab *et al.*, 2018; Xu *et al.*, 2017; Fawzi *et al.*, 2018; Shaham *et al.*, 2015; Fawzi *et al.*, 2016), and different approaches have been proposed to improve the robustness of neural networks (Papernot *et al.*, 2016; Cissé *et al.*, 2017). Fast Gradient Sign Method (FGSM) and several variations of it were introduced in (Goodfellow *et al.*, 2014b; Moosavi-Dezfooli *et al.*, 2016) for generating adversarial examples that are indistinguishable–to the human eye–from the original image, but can fool the networks. However, these techniques do not apply to state of the art body shape estimation as those are based on multi-stage processing. Typically, shape inference consists in fitting a body model to detected skeleton key points. Consequently, we perturb the 2D joints to produce an error in the shape fitting step. Cisse et al. (Cisse *et al.*, 2017), proposed a surrogate loss function call Houdini, which could be used to fool 2D pose estimations deep models by generating false poses or transferring one pose to another. As we do not know the persons' shape beforehand, it is ambiguous how to generate new pose that could lead to a success full shape evasion. Hence in Chapter 7, we studied the effect of each joint by selectively attacking them and report the importance of each joint in the final accuracy of the shape estimation.

Part I

GAZE BASED INFERENCE

Predicting the target of visual search from eye fixation (gaze) data is a challenging problem with many applications in human-computer interaction.

In this Chapter 3, we go beyond state of the art by studying search target prediction in an open-world setting in which we no longer assume that we have fixation data to train for the search targets. We then present a new problem formulation for search target prediction in the open-world setting that is based on learning compatibilities between fixations and potential.

In Chapter 4, we propose the first approach to predict the categories and attributes of search targets based on gaze data. We propose a novel *Gaze Pooling Layer* that integrates gaze information into CNN-based architectures as an attention mechanism – incorporating both spatial and temporal aspects of human gaze behaviour.

In Chapter 5, we go one step further and investigate the feasibility of combining recent advances in encoding human gaze information using deep convolutional neural networks with the power of generative image models to visually decode, i.e. create a visual representation of, the search target.

3 PREDICTION OF SEARCH TARGETS FROM FIXATIONS IN OPEN-WORLD SETTINGS

Previous work on predicting the target of visual search from human fixations only considered closed-world settings in which training labels are available and predictions are performed for a known set of potential targets. In this work we go beyond the state of the art by studying search target prediction in an open-world setting in which we no longer assume that we have fixation data to train for the search targets. We present a dataset containing fixation data of 18 users searching for natural images from three image categories within synthesised image collages of about 80 images. In a closed-world baseline experiment we show that we can predict the correct target image out of a candidate set of five images. We then present a new problem formulation for search target prediction in the open-world setting that is based on learning compatibilities between fixations and potential targets.

3.1 INTRODUCTION

In his seminal work from 1967, Yarbus showed that visual behaviour is closely linked to task when looking at a visual scene (Yarbus *et al.*, 1967). This work is an important demonstration of task influence on fixation patterns and sparked a large number of follow-up works in a range of disciplines, including human vision, neuroscience, artificial intelligence, and computer vision. A common goal in these human and computer vision works is to analyse visual behaviour, i.e. typically fixations and saccades, in order to make predictions about user behaviour. For example, previous work has used visual behaviour analysis as a means to predict the users' tasks (Borji and Itti, 2014; Borji *et al.*, 2012; DeAngelus and Pelz, 2009; Haji-Abolhassani and Clark, 2014; Kanan *et al.*, 2014; Zelinsky *et al.*, 2013; Zhang *et al.*, 2005), visual

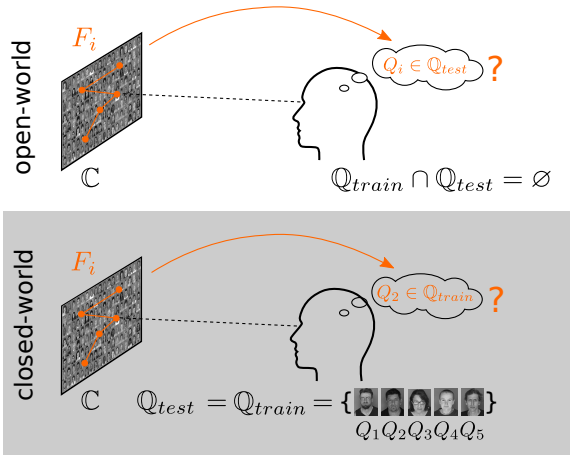


Figure 3.1: Experiments conducted in this work. In the *closed-world* experiment we aim to predict which target image (here Q_2) out of a candidate set of five images $\mathbb{Q}_{train} = \mathbb{Q}_{test}$ the user is searching for by analysing fixations F_i on an image collage $\mathbb{C}$. In the *open-world* experiments we aim to predict Q_i on the whole $\mathbb{Q}_{test}$.

activities (Bulling *et al.*, 2012, 2011, 2013; Land, 2006; Peters and Itti, 2008), cognitive processes such as memory recall or high cognitive load (Bulling and Roggen, 2011; Tessendorf *et al.*, 2011), abstract thought processes (Coen-Cagli *et al.*, 2009; Mast and Kosslyn, 2002), the type of a visual stimulus (Brandt and Stark, 1997; Cerf *et al.*, 2008; King, 2002), interest for interactive image retrieval (Coddington *et al.*, 2012; Guo *et al.*, 2002; Hussain *et al.*, 2014; Kozma *et al.*, 2009; Papadopoulos *et al.*, 2014; Stefanou and Wilson; Zhang *et al.*, 2010), which number a person has in mind (Loetscher *et al.*, 2010), or – most recently – to predict the search target during visual search (Borji *et al.*, 2014; Haji-Abolhassani and Clark, 2013; Rajashekar *et al.*, 2006; Zelinsky *et al.*, 2013).

Predicting the target of a visual search task is particularly interesting, as the corresponding internal representation, the mental image of the search target, is difficult if not impossible to assess using other modalities. While (Zelinsky *et al.*, 2013) and (Borji *et al.*, 2014) underlined the significant potential of using gaze information to predict visual search targets, they both considered a closed-world setting. In this setting, all potential search targets are part of the training set, and fixations for all of these targets were observed.

In contrast, in this chapter we study an open-world setting in which we no longer assume that we have fixation data to train for these targets. Search target prediction in this setting has significant practical relevance for a range of applications, such as image and media retrieval. This setting is challenging because we have to develop a learning mechanism that can predict over an unknown set of targets. We study this problem on a new dataset that contains fixation data of 18 users searching for five target images from three categories (faces as well as two different sets of book covers) in collages synthesised from about 80 images. The dataset is publicly available online.

The contributions of this work are threefold. First, we present an annotated dataset of human fixations on synthesised collages of natural images during visual search that lends itself to studying our new open-world setting. Compared to previous works, our dataset is more challenging because of its larger number of distractors, higher similarities between search image and distractors, and a larger number of potential search targets. Second, we introduce a novel problem formulation and method for learning the compatibility between observed fixations and potential search targets. Third, using this dataset, we report on a series of experiments on predicting users' search target from fixations by moving from closed-world to open-world settings.

3.2 DATA COLLECTION AND COLLAGE SYNTHESIS

Given the lack of an appropriate dataset, we designed a human study to collect fixation data during visual search. In contrast to previous works that used 3×3 squared patterns at two or three luminance levels, or synthesised images of natural scenes (Borji *et al.*, 2014), our goal was to collect fixations on collages of natural

(a) O'Reilly book covers (b) Amazon book covers (c) Mugshots (blurred for privacy reason)

Figure 3.2: Sample image collages used for data collection. Participants were asked to find different targets within random permutations of these collages.

images. We therefore opted for a task that involved searching for a single image (the target) within a synthesised collage of images (the search set). Each of the collages are the random permutation of a finite set of images. To explore the impact of the similarity in appearance between target and search set on both fixation behaviour and automatic inference, we have created three different search tasks covering a range of similarities.

In prior work, colour was found to be a particularly important cue for guiding search to targets and target-similar objects (Hwang *et al.*, 2009; Motter and Belky, 1998). Therfore we have selected for the first task 78 coloured O'Reilly book covers to compose the collages. These covers show a woodcut of an animal at the top and the title of the book in a characteristic font underneath (see Figure 4.4 top). Given that overall cover appearance was very similar, this task allows us to analyse fixation behaviour when colour is the most discriminative feature.

For the second task we use a set of 84 book covers from Amazon. In contrast to the first task, appearance of these covers is more diverse (see Figure 4.4 middle). This makes it possible to analyse fixation behaviour when both structure and colour information could be used by participants to find the target.

Finally, for the third task, we use a set of 78 mugshots from a public database of suspects. In contrast to the other tasks, we transformed the mugshots to grey-scale so that they did not contain any colour information (see Figure 4.4 bottom). In this case, allows abalysis of fixation behaviour when colour information was not available at all. We found faces to be particularly interesting given the relevance of searching for faces in many practical applications.

We place images on a grid in order to form collages that we show to the participants. Each collage is a random permutation of the available set of images on the grid. The search targets are subset of images in the collages. We opted for an independent measures design to reduce fatigue (the current recording already took 30 minutes of concentrated search to complete) and learning effects that both may have influenced fixation behaviour.

3.2.1 Participants, Apparatus, and Procedure

We recorded fixation data of 18 participants (nine male) with different nationalities and aged between 18 and 30 years. The eyesight of nine participants was impaired but corrected with contact lenses or glasses. To record gaze data we used a stationary Tobii TX300 eye tracker that provides binocular gaze data at a sampling frequency of 300Hz. Parameters for fixation detection were left at their defaults: fixation duration was set to 60ms while the maximum time between fixations was set to 75ms. The stimuli were shown on a 30 inch screen with a resolution of 2560x1600 pixels.

Participants were randomly assigned to search for targets for one of the three stimulus types. We first calibrated the eye tracker using a standard 9-point calibration, followed by a validation of eye tracker accuracy. After calibration, participants were shown the first out of five search targets. Participants had a maximum of 10 seconds to memorise the image and 20 seconds to subsequently find the image in the collage. Collages were displayed full screen and consisted of a fixed set of randomly ordered images on a grid. The target image always appeared only once in the collage at a random location.

To determine more easily which images participants fixated on, all images were placed on a grey background and had a margin to neighbouring images of on average 18 pixels. As soon as participants found the target image they pressed a key. Afterwards they were asked whether they had found the target and how difficult the search had been. This procedure was repeated twenty times for five different targets, resulting in a total of 100 search tasks. To minimise lingering on search target, participants were put under time pressure and had to find the target and press a confirmation button as quickly as possible. This resulted in lingering of 2.45% for Amazon (O'Reilly: 1.2%, mugshots: 0.35%).

3.3 METHOD

In this chapter we are interested in search tasks in which the fixation patterns are modulated by the search target. Previous work focused on predicting a fixed set of targets for which fixation data was provided at training time. We call this the *closed-world setting*. In contrast, our method enables prediction of new search targets, i.e. those for which no fixation is available for training. We refer to this as the *open-world setting*. In the following, we first provide a problem formulation for the previously investigated *closed-world setting*. Afterwards we present a new problem formulation for search target prediction in an *open-world setting* (see Figure 3.1).

3.3.1 Search Target Prediction

Given a query image (search target) $Q \in \mathbb{Q}$ and a stimulus collage $C \in \mathbb{C}$, during a search task participants $P \in \mathbb{P}$ perform fixations $F(C, Q, P) = \{(x_i, y_i, a_i), i = 1, \ldots, N\}$, where each fixation is a triplet of positions x_i, y_i in screen coordinates and

appearance a_i at the fixated location. To recognise search targets we aim to find a mapping from fixations to query images:

$$F(C, Q) \mapsto Q \in \mathbb{Q} \tag{3.1}$$

We use a bag of visual world featurisation ϕ of the fixations. We interpret fixations as key points around which we extract local image patches. These are clustered into a visual vocabulary V and accumulated in a count histogram. This leads to a fixed-length vector representation of dimension $|V|$ commonly known as a bag of words. Therefore, our recognition problem can more specifically be expressed as:

$$\phi(F(C, Q, P), V) \mapsto Q \in \mathbb{Q} \tag{3.2}$$

3.3.2 Closed-World Setting

We now formulate the previously investigated case of the closed-world setting where all test queries (search targets) $Q \in \mathbb{Q}_{test}$ are part of our training set $\mathbb{Q}_{test} = \mathbb{Q}_{train}$ and, in particular, we assume that we observe fixations $\mathbb{F}_{train} = \{F(C, Q, P) | \forall Q \in \mathbb{Q}_{train}\}$. The task is to predict the search target while the query and/or participant changes (see Figure 3.1).

$$\phi(F(C, Q, P), V) \mapsto Q \in \mathbb{Q}_{train} \tag{3.3}$$

We use a one-vs-all multi-class SVM classifier $\mathcal{H}_i$ and the query image with the largest margin:

$$Q_i = _{i=1,\ldots,|\mathbb{Q}_{test}|} \mathcal{H}_i(\phi(F_{test}, V)) \tag{3.4}$$

3.3.3 Open-World Setting

In contrast, in our new open-world setting, we no longer assume that we have fixation data to train for these targets. Therefore $\mathbb{Q}_{test} \cap \mathbb{Q}_{train} = \emptyset$. The main challenge that arises from this setting is to develop a learning mechanism that can predict over a set of classes that is unknown at training time (see Figure 3.1).

Search Target Prediction. To circumvent the problem of training for a fixed number of search targets, we propose to encode the search target into the feature vector, rather than considering it a class that is to be recognised. This leads to a formulation where we learn compatibilities between observed fixations and query images:

$$(F(C, Q_i, P), Q_j) \mapsto Y \in \{0, 1\} \tag{3.5}$$

Training is performed by generating data points of all pairs of Q_i and Q_j in $\mathbb{Q}_{train}$ and assigning a compatibility label Y accordingly:

$$Y = \begin{cases} 1 & if \text{i=j} \\ 0 & \end{cases} \tag{3.6}$$

The intuition behind this approach is that the compatibility predictor learns about similarities in fixations and search targets that can also be applied to new fixations and search targets.

Similar to the closed-world setting, we propose a featurisation of the fixations and query images. Although we can use the same fixation representation as before, we do not have fixations for the query images. Therefore, we introduce a sampling strategy S which still allows us to generate a bag-of-words representation for a given query. In this work we propose to use sampling from the saliency map as a sampling strategy. We stack the representation of the fixation and the query images. This leads to the following learning problem:

$$\begin{pmatrix} \phi(F(C, Q_i, P), V) \\ \phi(S(Q_j)) \end{pmatrix} \mapsto Y \in \{0, 1\} \tag{3.7}$$

We learn a model for the problem by training a single binary SVM $\mathcal{B}$ classifier according to the labelling as described above. At test time we find the query image describing the search target by

$$Q = _{Q_j \in \mathbb{Q}_{test}} \mathcal{B} \begin{pmatrix} \phi(F_{test}, V) \\ \phi(S(Q_j)) \end{pmatrix} \tag{3.8}$$

Note that while we do not require fixation data for the query images that we want to predict at test time, we still search over a finite set of query images $\mathbb{Q}_{test}$.

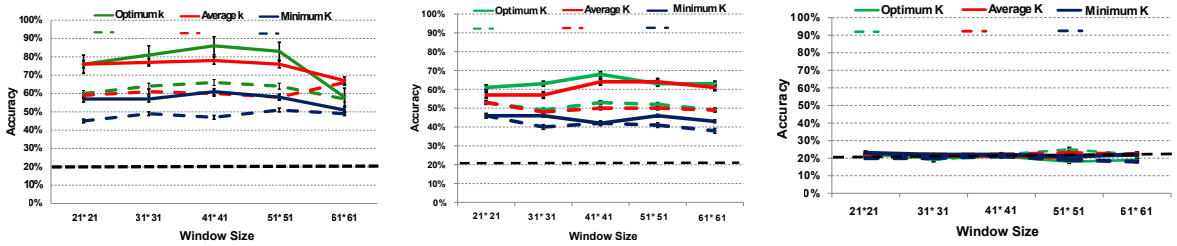
3.4 EXPERIMENTS

Our dataset contains fixation data from six participants for each search task. To analyse the first and second search task (O'Reilly and Amazon book covers) we used RGB values extracted from a patch (window) of size $m \times m$ around each fixation as input to the bag-of-words model. For the third search task (mugshots) we calculated a histogram of local binary patterns from each fixation patch. To compensate for inaccuracies of the eye tracker we extracted eight additional points with non-overlapping patches around each fixation (see Figure 3.3). Additionally, whenever an image patch around a fixation had overlap with two images in the collage, pixel values in the area of the overlap were set to 128.

3.4.1 Closed-World Evaluation

In our closed-world evaluation we distinguish between within-participant and cross-participant predictions. In the "within participant" condition we predict the search target for each participant individually using their own training data. In contrast, for the "cross participant" condition, we predict the search target across participants. The "cross participant" condition is more challenging as the algorithm has to generalise across users. Chance level is defined based on the number of search targets or classes our algorithm is going to predict. Participants were asked to search for five different targets in each experiment (chance level $1/5 = 20\%$).

Figure 3.3: Proposed approach of sampling eight additional image patches around each fixation location to compensate for eye tracker inaccuracy. The size of orange dots corresponds to the fixation's duration.



(a) O'Reilly (b) Amazon (c) Mugshots

Figure 3.4: Closed-world evaluation results showing mean and standard deviation of cross-participant prediction accuracy for Amazon book covers, O'Reilly book covers), and mugshots. Results are shown with (straight lines) and without (dashed lines) using the proposed sampling approach around fixation locations. The chance level is indicated with the dashed line.

Within-Participant Prediction. Participants looked for each search target 20 times. To train our classifier we used the data from 10 trials and the remaining 10 trials were used for testing. We fixed the patch (window) size to 41×41 and optimised k (vocabulary size) for each participant. Figure 3.5 summarises the within-participant prediction accuracies for the three search tasks. Accuracies were well above chance for all participants for the Amazon book covers (average accuracy 75%) and the O'Reilly book covers (average accuracy 69%). Accuracies were lower for mugshots but still above chance level (average accuracy 30%, chance level 20%).

Cross-Participant Prediction. We investigated whether search targets could be predicted within and across participants. In the accross-participants case, we trained one-vs-all multi-class SVM classifier using 3-fold cross-validation. We trained our model with data from three participants to map the observer-fixated patch to the target image. The resulting classifier was then tested on data from the remaining three participants. Prior to our experiments, we ran a control experiment where we uniformly sampled from 75% of the salient part of the collages. We trained the classifier with these randomly sampled fixations and confirmed that performance was around the chance level of 20% and therefore any improvement can indeed be

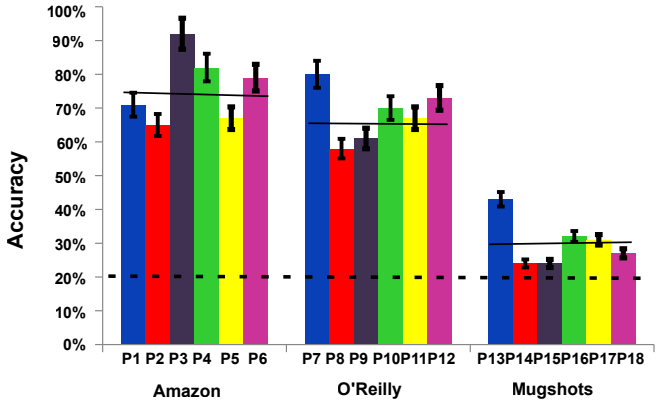


Figure 3.5: Closed-world evaluation results showing mean and standard deviation of within-participant prediction accuracy for Amazon book covers, O'Reilly book covers, and mugshots. Mean performance is indicated with black lines, and the chance level is indicated with the dashed line.

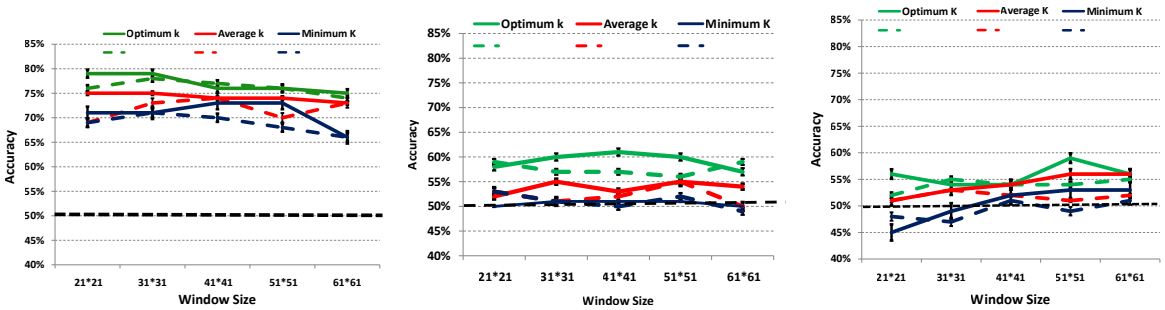
attributed to information contained in the fixation patterns.

Figure 3.4 summarises the cross-participant prediction accuracies for Amazon book covers, O'Reilly book covers, and mugshots for different window sizes and size of vocabulary k, as well as results with (straight lines) and without (dashed lines) using the proposed sampling approach around fixation locations. The optimum k represents the upper bound and corresponds to always choosing the value of k that optimises accuracy, while the minimum k correspondingly represents the lower bound. Average k refers to the practically most realistic setting in which we fix $k = 60$. Performance for Amazon book covers was best, followed by O'Reilly book covers and mugshots. Accuracies were between $61\% \pm 2\%$ and $78\% \pm 2\%$ for average k for Amazon and O'Reilly book covers but only around chance level for mugshots.

3.4.2 Open-World Evaluation

In the open-world evaluation the challenge is to predict the search target based on the similarity between fixations $F(C, Q)$ and query image $S(Q)$. In absence of fixations for query images Q we uniformly sample from the GBVS saliency map (Harel *et al.*, 2006). We chose the number of samples on the same order as the number of fixations on the collages. For the within-participant evaluation we used the data from three search targets of each participant to train a binary SVM with RBF kernel. The data from the remaining two search targets was used at test time. The average performance of all participants in each group was for Amazon: 70.33%, O'Reilly: 59.66%, mugshots: 50.83%.

Because the task is more challenging in the cross-participant evaluation, we report results for this task in more detail. As described perviously, we train a binary SVM with RBF kernel from data of three participants to learn the similarity between the observer-fixated patch when looking for three of the search targets and the corresponding target images. Our positive class contains data coming from the



(a) O'Reilly (b) Amazon (c) Mugshots

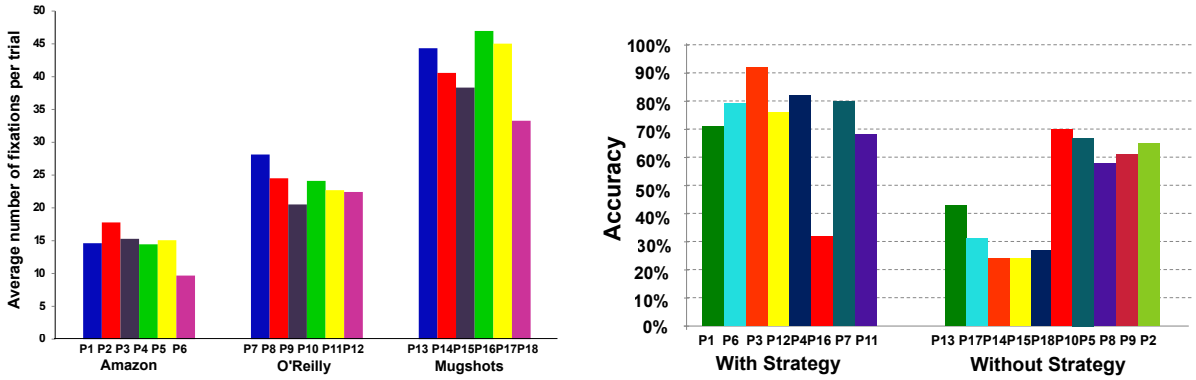
Figure 3.6: Open-World evaluation results showing mean and standard deviation of cross-participant prediction accuracy for Amazon book covers (top), O'Reilly book covers (middle), and mugshots (bottom). Results are shown with (straight lines) and without (dashed lines) using the proposed sampling approach around fixation locations. The chance level is indicated with the dashed line.

concatenation of $\Phi(F(C, Q_i, P), V)$ and $\Phi(S(Q_j))$ when $i = j$. At test time, we then test on the data of remaining three participants looking for two other search targets that did not appear in the training set and the corresponding search targets. The chance level is $1/2 = 50\%$ as we have a target vs, non-target decision.

Figure 3.6 summarises the cross-participant prediction accuracies for Amazon book covers, O'Reilly book covers, and mugshots for different window sizes and size of vocabulary k, as well as results with (straight lines) and without (dashed lines) using the proposed sampling approach around fixation locations. With average k the model achieves an accuracy of 75% for Amazon book covers, which is significantly higher than chance at 50%. For O'Reilly book covers accuracy reaches 55% and for mugshots we reach 56%. Similar to our closed-world setting, accuracy is generally better when using the proposed sampling approach.

3.5 DISCUSSION

In this chapter, we studied the problem of predicting the search target during visual search from human fixations. Figure 3.5 shows that we can predict the search target significantly above chance level for the within-participant case for the Amazon and O'Reilly book cover search tasks, with accuracies ranging from 50% to 78%. Figure 3.4 shows similar results for the cross-participant case. These findings are in line with previous works on search target prediction in closed-world settings (Zelinsky *et al.*, 2013; Borji *et al.*, 2014). Our findings extend these previous works in that we study synthesised collages of natural images and in that our method has to handle a larger number of distractors, higher similarities between search image and distractors, and a larger number of potential search targets. Instead of a large number of features, we rely only on colour information as well as local binary pattern features. We extended these evaluations with a novel open-world evaluation setting in which we no longer assume that we have fixation data to train for these targets. To learn under such



(a) Fixation count

(b) Scan-path variation

Figure 3.7: (a) average number of fixations per trial performed by each participant during the different search tasks. (b) difference in accuracies of participants who have a strategic search pattern vs participants that mainly skim the collage to find the search image.

a regime we proposed a new formulation where we learn compatibilities between observed fixations and query images. As can be seen from Figure 3.6, despite the much more challenging setting, using this formulation we can still predict the search target significantly above chance level for the Amazon book cover search task, and just about chance level for the other two search tasks for selected values of k. These results are meaningful as they underline the significant information content available in human fixation patterns during visual search, even in a challenging open-world setting. The proposed method of sampling eight additional image patches around each fixation to compensate for eye tracker inaccuracies proved to be necessary and effective for both evaluation settings and increased performance in the closed-world setting by up to 20%, and by up to 5% in the open-world setting. These results also support our initial hypothesis that the search task, i.e. in particular the similarity in appearance between target and search set and thus the difficulty, has a significant impact on both fixation behaviour and prediction performance. Figures 3.4 and 3.6 show that we achieved the best performance for the Amazon book covers, for which appearance is very diverse and participants can rely on both structure and colour information. The O'Reilly book covers, for which the cover structure was similar and colour was the most discriminative feature, achieved the second best performance. In contrast, the worst performance was achieved for the greyscale mugshots that had highly similar structure and did not contain any colour information. These findings are in line with previous works in human vision that found that colour is a particularly important cue for guiding search to targets and target-similar objects (Motter and Belky, 1998; Hwang *et al.*, 2009). Analysing the visual strategies that participants used provides additional interesting (yet anecdotal) insights. As the difficulty of the search task increased, participants tended to start skimming the whole collage rather than doing targeted search for specific visual features (see Figure 3.7(b) for an example). This tendency was the strongest for the most difficult search task, the mugshots, for which the vast majority of participants

Targeted search behaviour Skimming behaviour

Figure 3.8: Sample scan-paths of P8: Targeted search behaviour with a low number of fixations, and skimming behaviour with a high number of fixations. Size of the orange dots corresponds to fixation duration.

assumed a skimming behaviour. Additionally, as can be seen from Figure 3.7(b), our system achieved higher accuracy in search target prediction for participants who followed a specific search strategy than for those who skimmed most of the time. Well-performing participants also required fewer fixations to find the target (see Figure 3.7(a)). Both findings are in line with previous works that describe eye movement control, i.e. the planning of where to fixate next, as an information maximization problem (Butko and Movellan, 2010; Renninger *et al.*, 2004). While participants unconsciously maximised the information gain by fixating appropriately during search, in some sense, they also maximised the information available for our learning method, resulting in higher prediction accuracy.

3.6 CONCLUSION

In this chapter we demonstrated how to predict the search target during visual search from human fixations in an open-world setting. This setting is fundamentally different from settings investigated in prior work, as we no longer assume that we have fixation data to train for these targets. To address this challenge, we presented a new approach that is based on learning compatibilities between fixations and potential targets. We showed that this formulation is effective for search target prediction from human fixations. These findings open up several promising research directions and application areas, in particular gaze-supported image and media retrieval as well as human-computer interaction. Adding visual behaviour features and temporal information to improve performance is a promising extension that we are planning to explore in future work.

4 PREDICTING THE CATEGORY AND ATTRIBUTES OF VISUAL SEARCH TARGETS USING DEEP GAZE POOLING

PREDICTING the target of visual search from eye fixation (gaze) data is a challenging problem with many applications in human-computer interaction. In contrast to previous work that has focused on individual instances as a search target, we propose the first approach to predict categories and attributes of search targets based on gaze data. However, state of the art models for categorical recognition, in general, require large amounts of training data, which is prohibitive for gaze data. To address this challenge, we propose a novel *Gaze Pooling Layer* that integrates gaze information into CNN-based architectures as an attention mechanism – incorporating both spatial and temporal aspects of human gaze behavior. We show that our approach is effective even when the *gaze pooling layer* is added to an already trained CNN, thus eliminating the need for expensive joint data collection of visual and gaze data. We propose an experimental setup and data set and demonstrate the effectiveness of our method for search target prediction based on gaze behavior. We further study how to integrate temporal and spatial gaze information most effectively, and indicate directions for future research in the gaze-based prediction of mental states.

4.1 INTRODUCTION

As eye tracking technology is beginning to mature, there is an increasing interest in exploring the type of information that can be extracted from human gaze data. Within the wider scope of eye-based activity recognition Bulling *et al.* (2011); Steil and Bulling (2015), search target prediction Borji *et al.* (2014); Sattar *et al.* (2015); Zelinsky *et al.* (2013) has recently received particular attention as it aims to recognize users' search intents without the need for them to verbally communicate these intends. Previous work on search target prediction from gaze data (e.g. Borji *et al.* (2014) and Chapter 3) is limited to specific target instances that

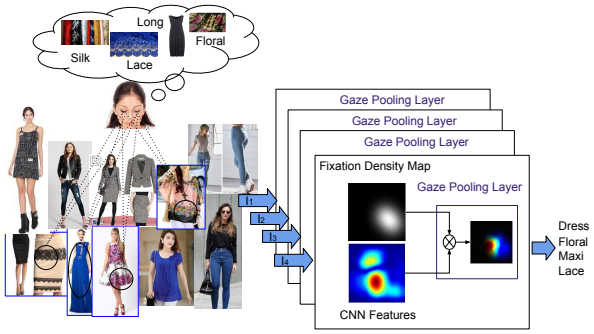


Figure 4.1: We propose a method to predict the target of visual search in terms of categories and attributes from users' gaze. We propose a *Gaze Pooling Layer* that leverages gaze data as an attention mechanism in a trained CNN architecture.

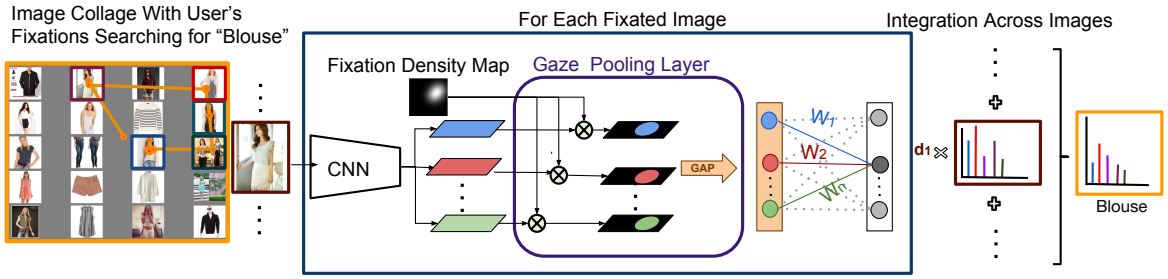


Figure 4.2: Overview of our approach. Given a search task (e.g. "Find a blouse"), participants fixate on multiple images in an image collage. Each fixated image is encoded into multiple spatial features using a pre-trained CNN. The proposed Gaze Pooling Layer combines visual features and fixation density maps in a feature-weighting scheme. The output is a prediction of the category or attributes of the search target. To obtain one final prediction over image collages, we integrate the class posteriors across all fixated images using average pooling.

users searched for, e.g. a particular object. This excludes searches for broader classes of objects that share the same semantic category or certain object attributes. Such searches commonly occur if the user does not have a concrete target instance in mind but is only looking for an object from a certain category or with certain characteristic attributes.

To address these limitations, we broaden the scope of search target prediction to categorical classes, such as object categories or attributes. One key difficulty towards achieving this goal is acquiring sufficient training data. We have to recall that object categorization only in the past decade has seen a breakthrough in performance by combining deep learning techniques with large training corpora. Collecting such large corpora is prohibitive for human gaze data, which poses a severe challenge to achieve our goal.

Therefore, we propose an approach for predicting categories and attributes of search targets that utilize readily trained CNN architectures and combines them with gaze data in a novel *Gaze Pooling Layer* (see Figure 6.1). The gaze information is used as an attention mechanism that acts selectively on the visual features to predict users' search target. These design choices make our approach compatible and practical with current deep learning architectures.

Through extensive experiments, we show that our method achieves accurate search target prediction for 10 category and 10 attribute tasks on a new gaze data set that is based on the DeepFashion data set Liu *et al.* (2016a). Furthermore, we evaluate different parameter settings and design choices of our approach, visualize internal representations and perform a robustness study w.r.t. noise in the eye tracking data. All code and data will be made publicly available upon acceptance.

4.2 PREDICTION OF SEARCH TARGETS USING GAZE

In this chapter, we are interested in predicting the category and attributes of search targets from gaze data. We address this task by introducing the Gaze Pooling Layer (GPL) that combines CNN architectures with gaze data in a weighting mechanism. Figure 5.4 gives an overview of our approach. In the following, we describe the four major components of our method in detail: The image encoder, human gaze encoding, the Gaze Pooling Layer, and search target prediction. Finally, we also discuss different integration schemes across multiple images that allow us to utilize gaze information obtained from collages. As a mean of inspecting the internal representation of our Gaze Pooling Layer, we propose Attended Class Activation Maps (ACAM).

4.2.1 Image Encoder

We build on the recent success of deep learning and use a convolutional neural network (CNN) to encode image information K. Simonyan and Zisserman (2013); Krizhevsky *et al.* (2012). Given a raw image I, a CNN is used to extract image feature map $F(I)$.

$$F(I) = \mathbf{CNN}(I) \tag{4.1}$$

The end-to-end training properties of these networks allow us to obtain domain-specific features. In our case, the network will be trained with data and labels relevant to the fashion domain. As we are interested in combining spatial gaze features with the image features, we use features $F(I)$ of the last convolutional layer that still has a spatial resolution. This results in a task-dependent representation with spatial resolution. In addition, to gain a higher spatial resolution we used the same architecture as describe in Zhou *et al.* (2016). We use their VGGnet-based model where layers after `conv5-3` are removed to gain a resolution of 14×14.

4.2.2 Human Gaze Encoding

Given a target category or attributes, participant $P \in \mathbb{P}$ look at image I and performs fixations $G(I,P) = (x_i, y_i), i = 1, ..., N$ in screen coordinates. We aggregate these fixations into fixation density maps $FDM(G)$ that capture the spatial density of fixations over the full image. Therefore, we represent the fixation density map $FDM(g)$ for a single fixation $g \in G(I, P)$ by a Gaussian:

$$FDM(g) = \mathcal{N}(g, \sigma_{fix}), \tag{4.2}$$

centered at the coordinates of the fixation, with a fixed standard deviation σ_{fix} – the only parameter of our representation. The fixation density map for all fixations $FDM(G)$ is obtained by coordinate-wise summation:

$$FDM(G) = \sum_{g \in G} FDM(g) \tag{4.3}$$

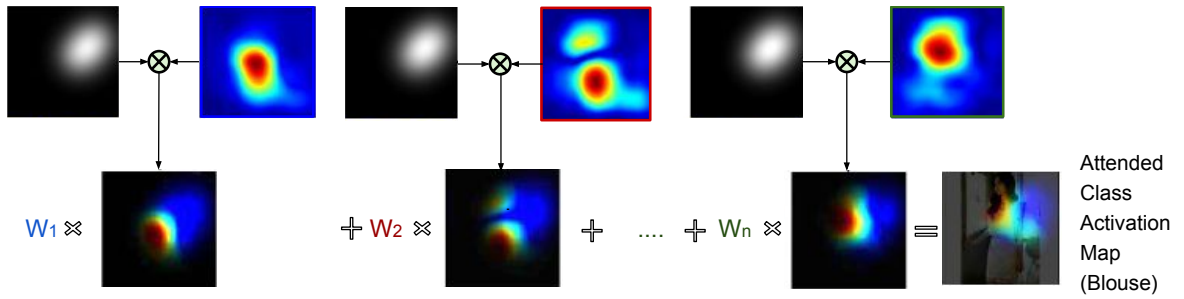


Figure 4.3: The proposed Gaze Pooling Layer combines fixation density maps with CNN feature maps via a spatial re-weighting (top row). Attended class activation maps are shown in the bottom row, which the predicted class scores are mapped back to the previous convolutional layer. The attended class activation maps highlight the class-specific discriminative image regions.

This corresponds to an average pooling integration. We also propose a max pooling version as follows:

$$FDM(G) = \max_{g \in G} FDM(g) \tag{4.4}$$

4.2.3 Gaze Pooling Layer

We combine the visual features $F(I)$ with fixation density map $FDM(G)$ in a Gaze Pooling Layer. The integration is performed by element-wise multiplication between both to obtain a gaze-weighted feature map (GWFM)

$$GWFM(I, G) = F(I) \otimes FDM(G). \tag{4.5}$$

In spirit of Zhou *et al.* (2016), we then perform Global Average Pooling (GAP) on each feature channel separately in order to yield a vector-valued feature representation.

$$GAP_{GWFM}(I, G) = \sum_{x,y} GWFM(I, G) \tag{4.6}$$

We finish our pipeline by classification with a fully connected layer and a soft-max layer.

$$p(C|I, G) = softmax(W\ GAP_{GWFM}(I, G) + b), \tag{4.7}$$

where W are the learned weights and b is the bias and C are the considered classes. The classes represent either categories or attributes depending on the experiment and we decide for the class with the highest class posterior (see Figure 5.4).

4.2.4 Integration Across Images

In our study, a stimulus is a collage with a set of images $I_i \in \mathbb{I}$. During the search task, participants fixate on multiple images in the collage, which generates fixations

$G_i \in \mathbb{G}$ for each image I_i. Hence, we need a mechanism to aggregate information across images. To do this, we propose a weighted average scheme of the computed posteriors per image:

$$p(C|\mathbb{I},\mathbb{G}) = \sum_i \frac{d_j}{\sum_j d_j} p(C|I_i, G_i). \tag{4.8}$$

We consider for the weights d_i the total fixations duration on image I_i as well as fixed d_i (see Figure 5.4). The latter corresponds to plain averaging.

4.2.5 Attended Class Activation Mapping

In order to inspect the internal representation of our Gaze Pooling Layer, we propose the attended class activation map visualization. It highlights discriminative image regions for a hypothesized search target based on CNN features combined with the weights from the gaze data. In this vein, it shares similarities to the CAM of Zhou *et al.* (2016) but incorporates the gaze information as attention scheme. The key idea is to delay the average pooling, which allows us to show spatial maps as also illustrated in Figure 5.4. In more detail, our network consists of several convolutional layers which the features of last convolutional layer is weighted by our fixation density map (GWFM). We do global average pooling over the GWFM and use those features for a fully connected layer to get the user attended categories or attributes. Given that our features maps are weighted by gaze data of users, it represents their attended classes. We can identify the importance of the image region for attended categories by projecting back the weights of the output layer onto a gaze-weighted convolutional feature map, which we call Attended Class Activation Map (ACAM):

$$ACAM_c(x,y) = \sum_k w_k^c \; GWFM_k(I,G) \tag{4.9}$$

where w_k^c indicates the importance $GWFM_k(I,G)$ of unit k for class c. The procedure for generating the class activation map are shown in Figure 4.3.

4.2.6 Implementations Details

In order to obtain the CNN features maps, we follow Zhou *et al.* (2016) and build on the recent VGGnet-GAP model. For our categorization experiments, we fine-tune on a 10 class classification problem on the DeepFashion data set Liu *et al.* (2016a). For attribute prediction, we fine-tune a model with 10 times 2-way classification in the final layer. We perform a validation of the VGGnet image classification performance model in the same setting as Liu *et al.* (2016a) and obtained comparable results ($\pm$5%) for category and attribute classification. To ensure that the images and collages are not informative of the category or attribute search tasks, we have performed a sanity check by using only the CNN prediction on the images of our collages. The resulting performance is at chance level, which validates our setup as search task information cannot be derived from the images or collages and therefore can only come from the gaze data.

(a) Attributes (c) Categories

Figure 4.4: Sample image collages used for data collection: Attributes, Categories. Participants were asked to find different clothing attributes and categories within these collages.

4.3 DATA COLLECTION

No existing data set provides image and gaze data that is suitable for our search target prediction task. We, therefore, collected our own gaze data set based on the DeepFashion data set Liu *et al.* (2016a). DeepFashion is a clothes data set consisting of 289,222 images annotated with 46 different categories and 1,000 attributes. We used the top 10 categories and attributes in our data collection. The training set of DeepFashion was used to train our CNN image model for clothes category and attribute prediction; the validation set was used to train participants for each category and attribute (see below). Finally, the test set was used to build up image collages for which we recorded human gaze data of participants while searching for specific categories and attributes. In the following, we describe our data collection in more detail.

4.3.1 Participants and Apparatus

We collected data from 14 participants (six females), aged between 18 and 30 years and with different nationalities. All of them had normal or corrected-to-normal vision. For gaze data collection we used a stationary Tobii TX300 eye tracker that provides binocular gaze data at a sampling frequency of 300Hz. We calibrated the eye tracker using a standard 9-point calibration, followed by a validation of eye tracker accuracy. For gaze data processing we used the Tobii software with the parameters for fixation detection left at their defaults (fixation duration: 60ms, the maximum time between fixations: 75ms). Image collages were shown on a 30-inch screen with a resolution of 2560x1600 pixels.

4.3.2 Procedure

We first trained participants by showing them exemplar images of all categories and attributes in a game like session to familiarise themselves with the categories and attributes. We did not collect any gaze data at this stage. For each category and attribute, we then generated 10 image collages, each containing 20 images. Each target category or attribute appeared twice in each collage at a random location (see Figure 4.4 for an example). Participants were then asked to search for ten different categories and attributes on these image collages (see Figure 4.4) while their gaze was being tracked. We stress again that we did not show participants a specific target instance of a category or attribute that they should search for. Instead, we only instructed them to find a matching image from a certain category, i.e "dress", or with a certain attribute, i.e "floral". Consequently, search session guided by the mental image of participants from the specific category or attributes. Participants had a maximum of 10 seconds to find the asked target category or attribute in the collage that was shown full-screen. As soon as participants found a matching target, they were asked to press a key. Afterwords, they were asked whether they had found a matching target and how difficult the search had been. This procedure was repeated ten times for ten different categories or attributes, resulting in a total of 100 search tasks.

4.4 EXPERIMENTS

To evaluate our method for search target prediction of categories and attributes, we performed a series of experiments. We first evaluated the effectiveness of our Gaze Pooling Layer, the effect of using a local vs global representation, and of using a weighting by fixation duration. We then evaluated the gaze encoding that encompasses the pooling scheme of the individual fixation as well as the σ_{fix} parameter to represent a single fixation. Finally, we evaluated the robustness of our method to noise in the eye tracking data, which sheds light on different possible deployment scenarios and hardware that our approach is amendable to. Additionally, we provide visualization of the internal representations in the Gaze Pooling Layer. Across the results, we present Top-N accuracies denoting correct predictions if the correct answer is among the top N predictions.

4.4.1 Evaluation of the Gaze Pooling Layer

Fixation information enters our method in two places: The fixation density maps in the Gaze Pooling Layer(subsection 4.2.3) as well as the weighted average across images in the form of fixation duration (see subsection 4.2.4 and Figure 5.4). In order to evaluate the effectiveness of our Gaze Pooling Layer, we evaluate two

Global vs.		——— Category ———			Attribute
Local	◔	Top1	Top2	Top3	Accuracy
Global		31%±5	48% ±8	62% ±8	20%±1
Local		49%±7	68%±6	78%±6	26 %±1
Global	✓	52%±6	68%±6	78%±6	25%±1
Local	✓	**57%±8**	**74%±7**	**84%±4**	**34%±1**

Table 4.1: Evaluation of global vs. local gaze pooling with and without weighting based on the fixation duration ◔.

conditions: "*local*" makes full use of the gaze data and generates fixation density maps using the fixation location as described in our method section. "*global*" also generates a fixation density map, but does not use the fixation location information and therefore generates for each fixation a uniform weight across the whole fixated image. In addition, we evaluate two more conditions, where we either used the fixation duration (◔) as a weight to the average class posterior of each fixated image (see subsection 4.2.4) or ignore the duration.

Table 6.2 shows the result of all 4 combinations of these conditions, with the first column denoting if local or global information was used and the second column ◔ whether fixation duration was used. Absolute performance of our best model using local information and fixation duration were 57%, 74%, and 84% on top1-3 accuracy respectively for the categorization task and 34% accuracy for attributes. The results show a consistent improvement (16 to 18 pp for categories, 6 pp for attributes) across all measures and tasks going from a global to a local representation (first to the second row). Adding the weighting by fixation duration yields another consistent improvement for both local and global approach (another 6 to 5 pp for categories). Our best method improves overall by 22 to 26 pp on the categorization task and 14 pp on the attributes. The global method without fixation duration (first row) is in a spirit similar to Chapter 3– although the specific application differs. All further experiments will consider our best model (last row) as the reference and justify the parameter choices (average pooling, $sigma_{fix}$) by varying each parameter one by one.

4.4.2 Evaluation of the Gaze Encoding

We then evaluated the gaze encoding that takes individual fixations as input and produces a fixation density map. We first evaluated the representation of a single fixation that depends on the parameterσ_{fix}, followed by the pooling scheme that combines multiple fixations into fixation density maps.

$\sigma_{fix} \rightarrow$	1	1.2	1.4	1.6	1.8	2
Top1	55%	54%	56%	56%	57%	57%
Top2	74%	74%	74%	74%	74%	75%
Top3	83%	84%	84%	85%	85%	84%

Table 4.2: Evaluation of different gaze encoding schemes using different per-fixation σ_{fix}.

Fixation	——— Category ———			Attribute
Pooling	Top1	Top2	Top3	Accuracy
Max	54%±8	73%±9	83%±6	32%±1
Average	**57%±8**	**74%±7**	**84%±4**	**34%±1**

Table 4.3: Evaluation of different fixation pooling strategies using average or max pooling.

Effects of Fixation Representation Parameter f_σ. The parameter σ_{fix} controls the spatial extend of a single fixation in the fixation density maps as described in subsection 4.2.2. We determined an appropriate setting of this parameter to be $\sigma_{fix} = 1.6$ in a pilot study to roughly match the eye tracker accuracy and analyzed here the influence on the overall performance by varying this parameter in a sensible range (given eye tracker accuracy and coarseness of feature map) from 1 to 2 as shown in Table 4.2. As can be seen from the Table, our method is largely insensitive to the investigated range of reasonable choices of this parameter and our choice of 1.6 is on average a valid choice within that range.

Fixation Pooling Strategies. We evaluated two options for how to integrate single fixations into an fixation density map: Either using average or max pooling. The results are shown in Table 4.3. As the Table shows, while both options perform well, average pooling consistently improves over the max pooling option.

4.4.3 Noise Robustness Analysis

While our gaze data is recorded with a highly-accurate stationary eye tracker, there are different modalities and types of eye trackers available. One key characteristic in which they differ is the error at which they can record gaze data – typically measured in degrees of visual angle. While our controlled setup provides us with an

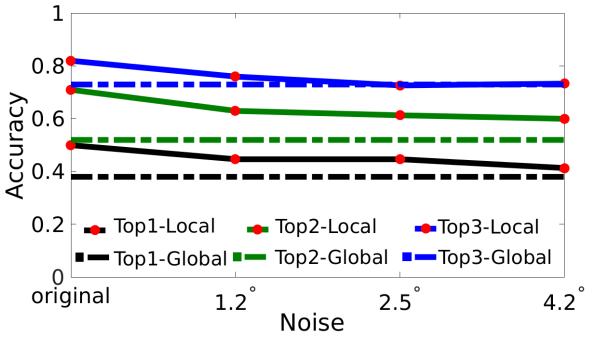


Figure 4.5: Accuracy for different amounts of noise added to the eye tracking data. Our method is robust to this error which suggests that it can also be used with head-mounted eye trackers or learning-based methods that leverage RGB cameras integrated into phones, laptops, or public displays.

accuracy of about 0.7 degrees of error, state-of-the-art eye trackers based on webcams, tablets or integrated into glasses can have up to 4 degrees depending also on the deployment scenario Zhang *et al.* (2015b). Therefore, we finally investigated the robustness of our approach w.r.t. different levels of (simulated) noise in the eye tracker. To this end, we sampled noise from a normal distribution with $\sigma = 1, 3, 5$. This corresponds roughly to $60, 120$ and 200 pixels and to $1.2, 2.5$ and 4.2 degrees of visual angles and hence covers a realistic range of errors. The results of this evaluation are shown in Figure 4.5. As can be seen, our method is quite robust to noise with only a drop of 5 to 10pp for Top3 to Top1 accuracy, respectively – even at the highest noise level. In particular, all the results are consistently above the performance of the corresponding global methods shown as dashed lines in the plot.

4.4.4 Visualization and Analysis of Gaze Pooling Layer on Single Images

We provide further insights into the working of our Gaze Pooling Layer by showing visual examples of the attended class activation maps, associated fixation density map and search target prediction results. While the quantitative evaluation was conducted on full collages, this is impracticable for inspection. Therefore, we show in the following visualizations and analysis on single images.

Predictions. Figure 4.7(a) shows results for the categorization task and Figure 4.7 (b) for the attribute task. Each of these figures shows the output of the "global" method that uses uniform fixation density map as well as the "local" method that makes full use of the gaze data. We observe that for the "local" method a relevant part of the images is fixated on which in turn leads to correct prediction of the intended search task.

Search target prediction over image collage. In Figure 4.9 presents fixation data of one participant searching for category "Blouse" and attribute "Lace". The posterior

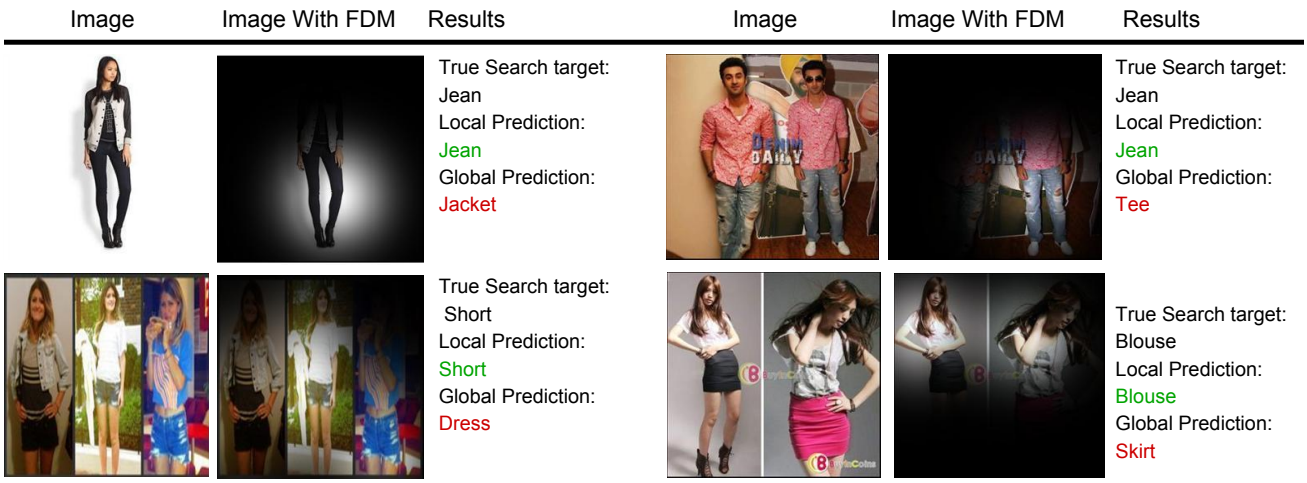
(a)Attributes (b) Categories

Figure 4.6: Example fixation data of 2 participants (red and green dots) with search target attribute='Floral' on top and category='Cardigan' below.

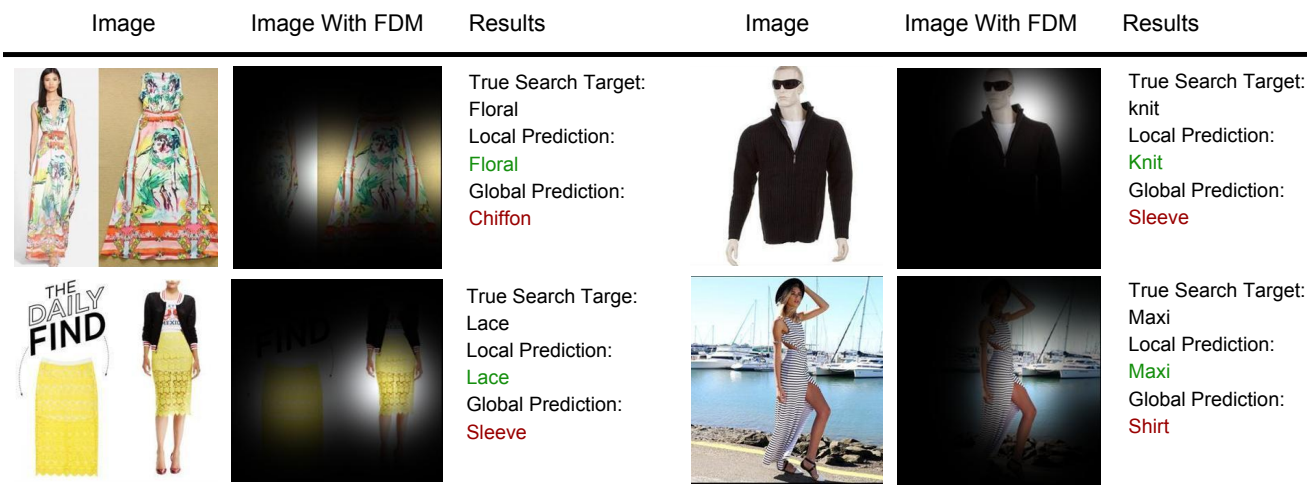
of all fixated images are average to get one final prediction over all fixated images. For each fixated image we show the attended class activation map (ACAM) and the result of global and local method with and without fixation duration.

Attended Class Activation Map (ACAM) Visualization. Figure 4.8 shows the attended class activation map (ACAM) of top 3 predictions, for "local" as well as "global" approach. The "global" method exploits that this image was fixed on - but does not exploit the location information of the fixations. Therefore it reduces in the case of a single image to a standard CAM. E.g the lower part of the image is activated for "skirt" and the upper part is activated for "Tee". One can see that highlighted regions vary across predicted class. The first row shows the ACAM for the "local" method. It can be seen how the local weighting due to the fixation is selective to the relevant features of the search target, e.g. eliminating the "skirt" responses and retaining the "blouse" responses.

Figure 4.10 and Figure 4.11, shows the attended class activation map(ACAM) of top 1 predictions of the fixated images, for "local" and "global" approach. The left column represent the image and the task of the user, the second column shows fixation density maps of the user searching for the given task and the last two columns are ACAM for the local and global method.



(a) Categories



(b) Attributes

Figure 4.7: Example responses of local and global method. Green means correct and red means wrong target prediction.

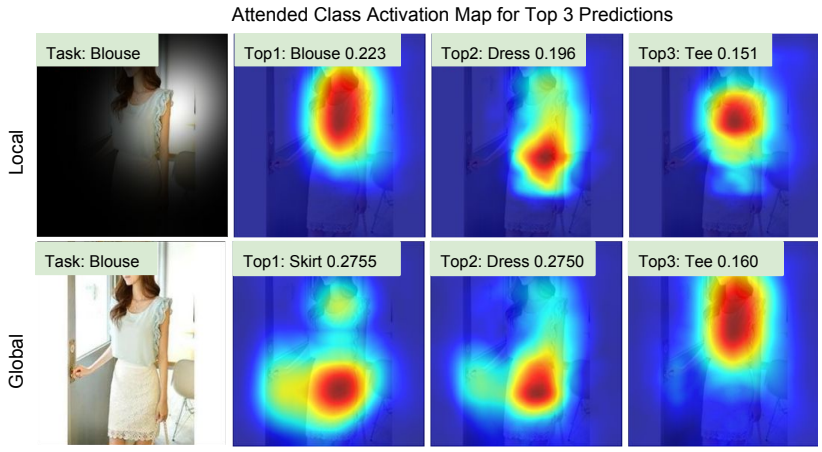


Figure 4.8: Attended class activation maps of top 3 predictions in local and global method for a given image. Participants were searching for target category “Blouse". The maps shows the discriminative image regions used for for this search task.

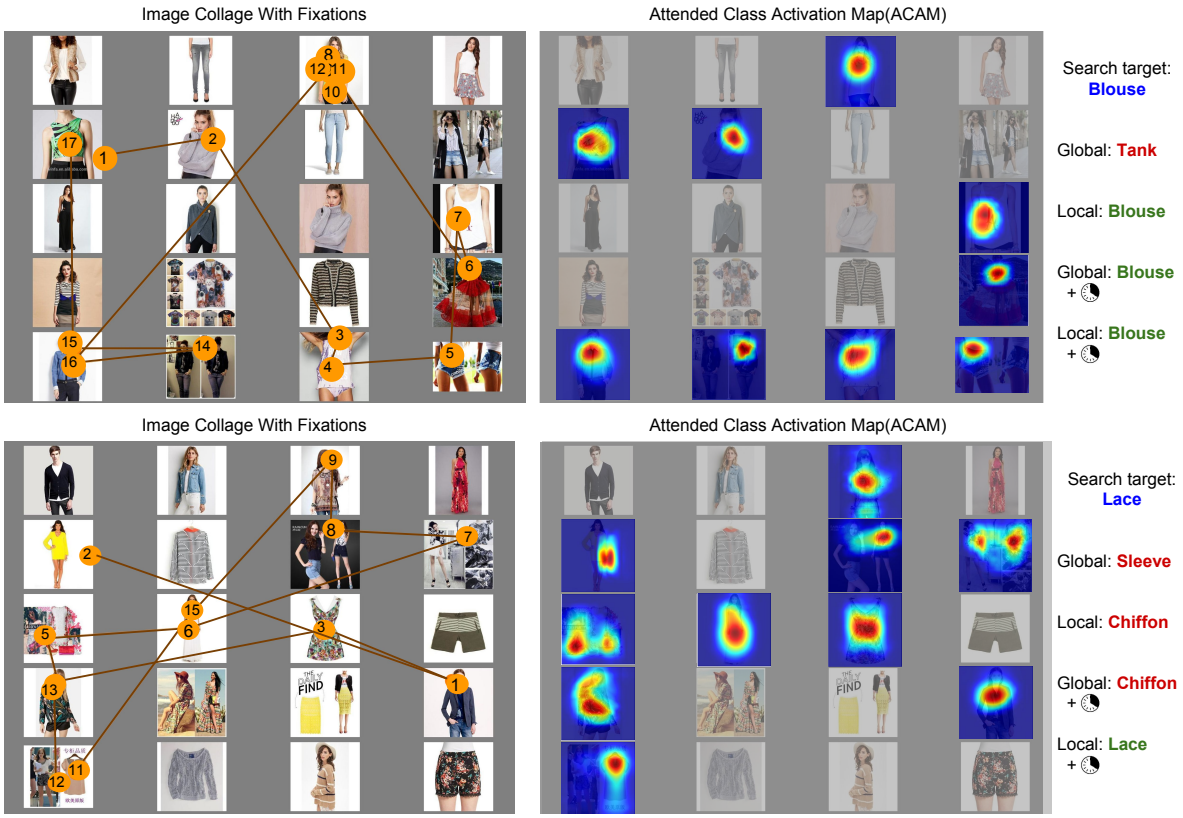


Figure 4.9: Image collage with fixations of a participant searching for “Blouse” and “Lace”. The right image show the ACAM of each fixated image in the collage. The last column represent top 1 prediction for global and local method without and with fixation durations.

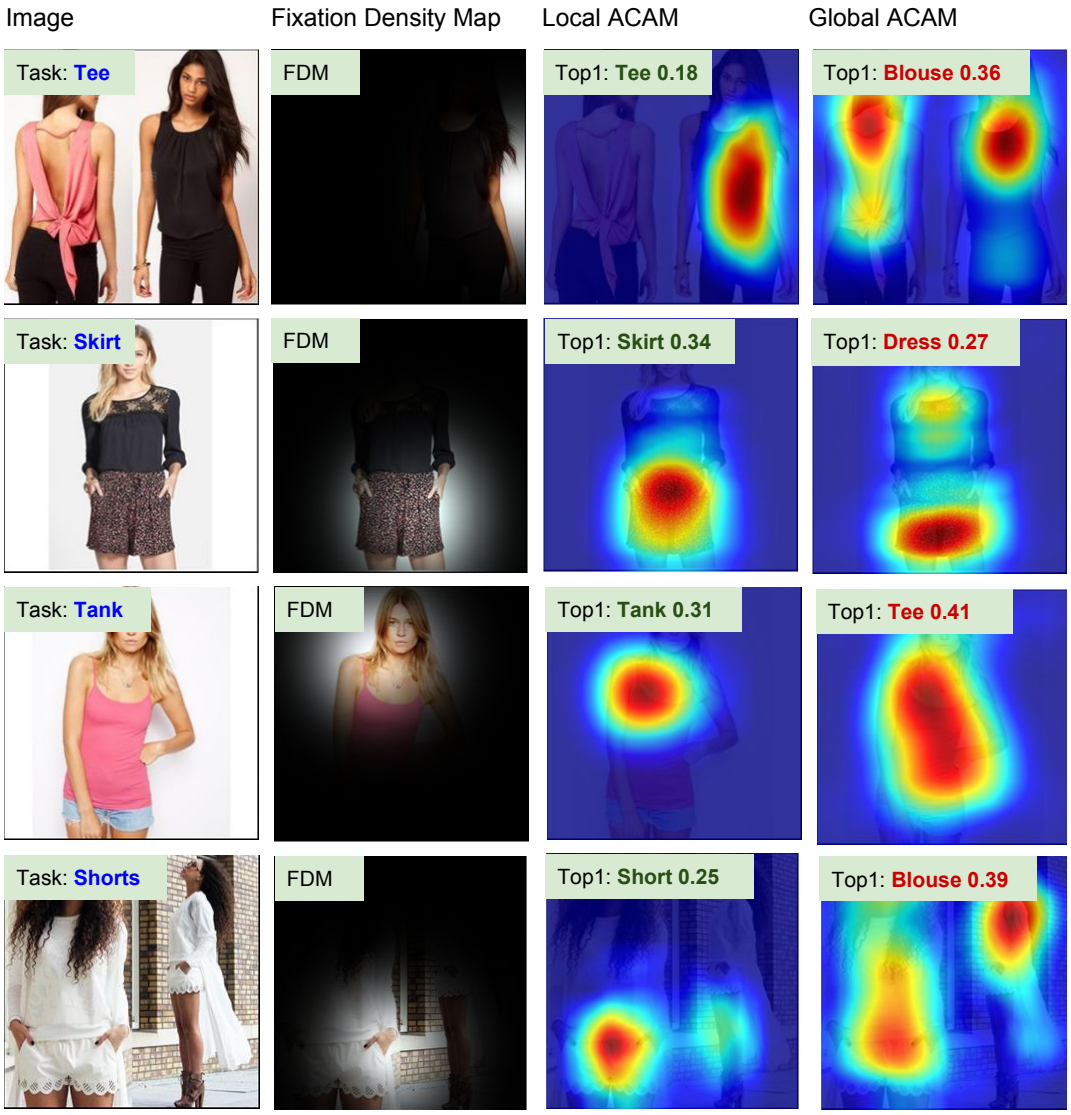


Figure 4.10: Attended class activation maps of top1 prediction in local and global method for a single fixated image. Participants were searching for the given category. The maps shows the discriminative image regions used for this search task

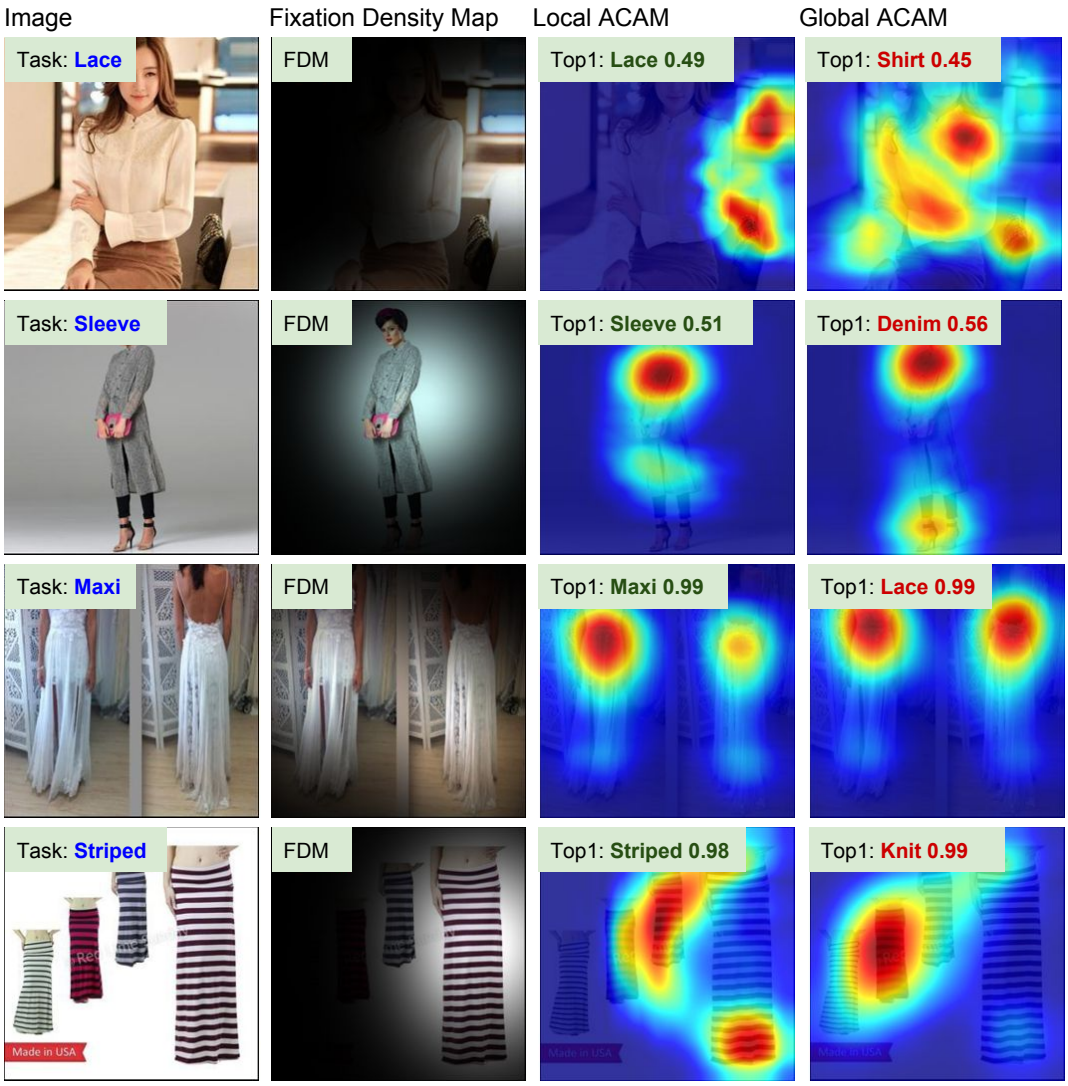


Figure 4.11: Attended class activation maps of top1 prediction in local and global method for a single fixated image. Participants were searching for the given attribute. The maps shows the discriminative image regions used for this search task

4.5 DISCUSSION

In this chapter, we studied the problem of predicting categories and attributes of search targets from gaze data. Table 6.2 shows strong performance for both tasks. Our Gaze Pooling Layer represents a modular and effective integration of visual and gaze features that are compatible with modern deep learning architectures. Therefore, we would like to highlight three features that are of particular practical importance.

Parameter Free Integration Scheme. First, our proposed integration scheme is basically parameter-free. We introduce a single parameter σ_{fix} but the gaze encoding is only input to the integration scheme and, in addition, the method turns out to be not sensitive to the choice (see experiments in subsection 4.2.2).

Training from Visual Data. Second, fixing the fixation density maps to uniform maps yields a deep architecture similar to a GAP network that is well-suited for various classification tasks. While this no longer addresses the task of predicting categories and attributes intended by the human in the loop, it allows us to train the remaining architecture for the task at hand and on visual data, which is typically easier to obtain in larger quantities than gaze data. This type of training results in a domain-specific image encoding as well as a task-specific classifier.

Training Free Gaze Deployment. Gaze data is time-consuming to acquire – which makes it rather incompatible with today's data hungry deep learning models. In our model, however, the fixations density maps computed from the gaze data can be understood as spatially localized feature importance that is used to weight feature importance in the spatial image feature maps Figure 4.8. Our results demonstrate that strong performance can be obtained with this re-weighting scheme without the need to re-train with gaze data. As a result, our approach can be deployed without any gaze-specific training. This result is surprising, in particular as the visual model on its own is completely uninformative without gaze data on the task of search target prediction (as we have validated in subsection 4.4.1. We believe this simplicity of deployment is a key feature that makes the use of gaze information in deep learning practical.

Biases in Mental Model of Attributes and Categories Among Users. In order to illustrate the challenges our Gaze Pooling Layer has to deal with in terms of the variations in the observed gaze data, we show example fixation data in Figure 4.6. In each image, fixation data of two participants (red and green dots) is overlaid over a presented collage. Although both participants had the same search target (top: attribute 'Floral'; bottom: category 'Cardigan'), we observe a drastically different fixation behaviour. One possible explanation is that the mental models of the same target category or attribute can vary widely depending on personal biases Ferecatu and Geman (2009). Despite these strong variations in the gaze information, our

Gaze Pooling Layer allows predicting the correct answer in all 4 cases. The key to this success is aggregating relevant local visual feature across all images in the collage, that in turn represent one consistent search target in terms of categories and attributes.

4.6 CONCLUSION

In this chapter, we proposed the first method to predict the category and attributes of visual search targets from human gaze data. To this end, we proposed a novel Gaze Pooling Layer that allows us to seamlessly integrate semantic and localized fixation information into deep image representations. Our model does not require gaze information at training time, which makes it practical and easy to deploy. We believe that the ease of preparation and compatibility of our Gaze Pooling Layer with existing models will stimulate further research on gaze-supported computer vision, particularly methods using deep learning.

5 VISUAL DECODING OF TARGETS DURING VISUAL SEARCH FROM HUMAN EYE FIXATIONS

What does human gaze reveal about a users' intents and to what extent can these intents be inferred or even visualized? In Chapter 3 and Chapter 4, gaze was proposed as an implicit source of information to predict the object class and attributes of the search target.

In this chapter, we go one step further and investigate the feasibility of combining our gaze embedding approach introduced in Chapter 4, with the power of generative image models to visually decode, i.e. create a visual representation of the search target. Such visual decoding is challenging for two reasons: 1) the search target only resides in the user's mind as a subjective visual pattern, and can most often not even be described verbally by the person, and 2) it is, as of yet, unclear if fixation data contain sufficient information for this task at all. We show, for the first time, that visual representations of search targets can indeed be decoded only using human fixation data. We propose first to encode fixations into a semantic representation and then decode this representation into an image. We evaluate our method on our recent gaze dataset form section 4.3 and validate the model's predictions using two human studies. Our results show that 62% (Chance level = 10%) of the time users were able to select the categories of the decoded image right.

5.1 INTRODUCTION

Predicting the visual search target of users is essential for a range of applications, particularly human-computer interaction, given that it reveals information about users' search intents. During visual search, gaze conveys rich information about the target a user has in mind.

Pioneer works tried to predict the search target in the limited set of known targets from gaze data (Zelinsky *et al.*, 2013; Borji *et al.*, 2014). In the Chapter 3, we move smoothly from a limited set of targets to an open-world setting. In Chapter 4, we introduced a gaze pooling layer that can recognise categories and attributes of the target. We also removed

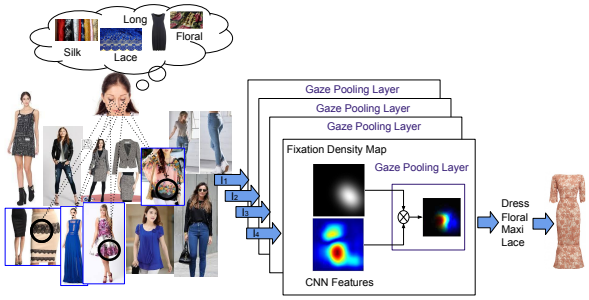


Figure 5.1: The *Gaze Pooling Layer* allows us to encode the target of visual search into a semantic vector, in terms of categories and attributes. The representation is used as a condition, in a Conditional Variational Auto-Endocer to generate images of the search target.

the need of gaze to train a classifier to identify the users' search target. This enables the usage of gaze in conjuncture with state of the art deep learning recognition system to recognise users intent. Yet, we only addressed a classification scenario – hence a finite set of discrete classes.

On the other hand, cognitive neuroscientists have shown the first success of visualising images based on fMRI data (Cowen *et al.*, 2014; Nishimoto *et al.*, 2011). While we also want to access aspects of the mental state, our task is fundamentally different in two points: (1) While they aim at decoding a specific image that was shown to a person, our goal is to decode visual search target. (2) They are using fMRI data, we are using gaze data. On a related note, we believe that gaze data is particularly interesting to investigate, as it is practical and affordable to collect and can be use in many application scenarios of future interfaces (Sato *et al.*, 2016; Zhang *et al.*, 2015a).

As we are targeting decoding of categorical search targets, generating such a visualisation is challenging due to substantial intra-class variations. However, recent advances in deep learning have led to a new generation of generative models for images. Recently, Yan *et al.* (2016) generated images of objects from high-level descriptions. They transfer the high-level text information to a set of attributes (e.g. hair colour: Brown, gender: female). These attributes are used later on to build an attribute-conditioned generative model.

Hence, we approach decoding of search targets from gaze data by bringing together recent success in gaze encoding and categorisation with state of the art category conditioned generative image models.

The main contributions of this chapter are:

- First proof of concept that visual representations of search targets can be decoded from human gaze data.
- We present a practical approach, as it respects the difficulties in collecting large human gaze datasets. Encoder and decoders are trained from large image corpora, and a semantic layer facilitates transfer between the two representation in between.
- We show the importance of localised gaze information for improved search target reconstruction.

5.2 SEARCH TARGET DECODING MODEL

In this chapter, we address decoding of users visual search target from the gaze. To achieve our goal, we build on our recent success in gaze encoding and categorisation as well as state of the art category conditioned generative image models. The gaze encoding is used to encode the raw gaze data into a semantic categorical space. The generative image model is conditioned on the encoded gaze data to decode visual search target of users as illustrated in Figure 5.2.

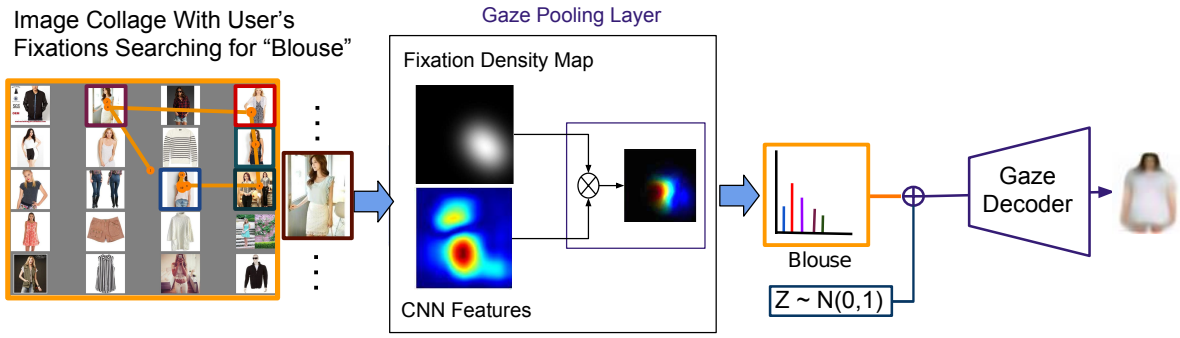


Figure 5.2: Overview of our approach. The user is searching for a category "Blouse", the gaze data is recorded during the search task. We encode the gaze information into a semantic representation $p(C|I,F)$. The representation is used as a condition over the learned latent space to decode the gaze into visualisations of the categorical search target.

We assume that a participant $P \in \mathbb{P}$ is searching for a target category $C \in \mathbb{C}$ in an image collage I. During the search task the participant, perform fixations $F(I,C,P) = (x_i, y_i, t_i), i = 1, ..., N$, where x_i, y_i are position of fixations in screen coordinates and t_i is fixation duration. Given the fixation data for a target category $P(c|F(I,C,P))$, we want to sample the search target ST as:

$$P(ST|F(I,C,P)) = \sum_c P(ST|c)P(c|F(I,C,P)), \tag{5.1}$$

where $P(ST|c)$ is a decoding from the semantic space to the visual search target. In the following, we explain the gaze encoder and search target decoder.

5.2.1 Semantic Gaze Encoder

As in Chapter 4, we use Fixation Density Map (FDM) to represent fixations:

$$FDM(F) = \sum_{f \in F} FDM(f) \tag{5.2}$$

where each fixation f is represented by a Gaussian function of fixed variance at the location of the fixated point FDM(f). The FDM are then combined with visual features F(I) using the Gaze Pooling Layer. The integration is done via element-wise multiplication of FDM and F(I):

$$G(I,F) = \sum_{x,y} F(I) \times FDM(F) \tag{5.3}$$

As in Chapter 4, we then perform Global Average Pooling (GAP) on each feature

channel separately in order to yield a vector-valued feature representation.

$$GAP_G(I,F) = \sum_{x,y} G(I,F) \tag{5.4}$$

To get a final class prediction the resulted feature maps G are averaged and fed into a fully connected and *softmax* layer:

$$p(C|I,F) = softmax(W\ GAP_G(I,F) + b), \tag{5.5}$$

W are learned weights, b is the bias. To decode the search target, and c is the considered as class. Previously (Chapter 4), we used $p(C|I,F)$ to predict the search target of users. Hence, we hypothesise that this encoding conveys rich information on the user's visual search target.

5.2.2 Visual Search Target Decoder

To sample visual search targets of users, we employ a generative image model that is conditioned on the class posteriors predicted via gaze pooling layer and a latent random variable z. Yan *et al.* (2016) introduced a Conditional Variational Auto-Encoder (CVAE) which generates images conditioned on a list of attributes. Here, we build on this recent success and train a model to generate images of different clothing categories. Given the category vector $c \in \mathbb{R}^{d_c}$ and the latent variable $z \in \mathbb{R}^{d_z}$, our goal is to build a generative model $p_\theta(I|c,z)$, which generates image $I \in \mathbb{R}^{d_I}$. The generated image is conditioned on the categorical information and the latent variable. In conditional variational auto-encoder, the auxiliary distribution $q_\phi(z|I,c)$ is introduced to approximate the true posterior $p_\theta(z|I,c)$. The goal of the learning process is to find the best parameter θ which maximises the lower bound of the log-likelihood $log\ p_\theta(I|c)$. Hence the conditional log-likelihood is

$$log\ p_\theta(I|c) = KL(q_\phi(z|I,c)||p_\theta(z|I,c)) + \mathcal{L}_{CVAE}(I,c,;\theta,\phi), \tag{5.6}$$

where the variational lower bound

$$\mathcal{L}_{CVAE}(I,c,;\theta,\phi) = -KL(q_\phi(z|I,c)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|I,c)}[log\ p_\theta(I|c,z)] \tag{5.7}$$

is maximised for learning the model parameter. We assume that the prior $p_\theta(z)$ follows a isotropic multivariate Gaussian distribution. The conditional distributions $p_\theta(I|c,z)$ and $q_\phi(z|I,c)$ are multi-variate Gaussian distribution with mean and variance of $\mathcal{N}\left(\mu_\theta(I,c), diag(\sigma^2_\theta(z,c))\right)$ and $\mathcal{N}\left(\mu_\phi(I,c), diag(\sigma^2_\phi(I,c))\right)$. The recognition model here is $q_\phi(z|,I,c)$ and the generation model is the conditional distribution $p_\theta(I|c,z)$.

During training the first term $KL(q_\phi(z|I,c)||p_\theta(z))$ acts as a regularisation term that minimises the gap between the prior $p_\theta(z)$ and the proposal distribution $q_\phi(z|I,c)$.

To generate gaze conditioned image, we replace the recognition network with a VGGNet-16-GAP network, as explained in subsection 4.2.6 during the test time. The Z is sampled from isotropic Gaussian distribution.

5.3 EXPERIMENTS

we used the dataset from section 4.3, to decode visual target of our users. We evaluate our results via two user study. One measures the success in reconstructing meaningful visual representations of visual search targets and the second highlight the importance of a gaze encoder that respects localised information.

5.3.1 Implementation details

We evaluate our model on the gaze data set from section 4.3. The data set contains gaze data of 14 participants searching for ten different clothing categories and attributes. In this data set, a stimulus is a collage $C_k, k \in [1,\ 10]$ with a set of images $I_i \in \mathbb{I}$. As described in Chapter 4, during the search task, participants fiIate on multiple images in the collage , which generates fixations $F_i \in \mathbb{F}$ for each image I_i. Hence to aggregate information across images we averaged the computed posteriors per fixated image:

$$p(C_k|\mathbb{I}, \mathbb{F}) = \sum_i \frac{d_j}{\sum_j d_j} p(C_k|I_i, F_i). \tag{5.8}$$

where d_i is the total fixations duration on image I_i.

Furthermore, participant were searching for a category c in 10 consecutive collages. Hence to achieve a unique description of the participant search target for each category, we average the resulted posterior from Equation 5.8, over all collages that belongs to target category c:

$$p(C|\mathbb{I}, \mathbb{F}) = \sum_{k=1}^{10} p(C_k|\mathbb{I}, \mathbb{F}) \tag{5.9}$$

The CAVE, have two convolutional neural networks for recognition and generation. The encoder network consists of 5 convolution layers, followed by 2 fully-connected layers (convolution layers have 64, 128, 256, 256 and 1024 channels with filter size of 5×5, 5×5, 3×3, 3×3 and 4×4, respectively; the two fully-connected layers have 1024 and 192 neurons). The category stream is merged with image stream at the end of the recognition network. The decoder network consists of 2 fully-connected layers, followed by 5 convolution layers with 2-by-2 upsampling (fully-connected layers have 256 and 8×8×256 neurons; the convolution layers have 256, 256, 128, 64 and 3 channels with filter size of 3 × 3, 5× 5, 5 × 5,5×5 and 5×5. Furthermore, we train a classifier (VGGNet-16-GAP) as explained in subsection 4.2.6 to get class posterior from gaze data $p(C|\mathbb{I}, \mathbb{F})$. This posterior is used in the category stream of the encoder to generate search target of our users.

Both networks are trained over the top 10 categories of Deep-fashion. We use the same train, test and validation split as proposed in the deep-fashion dataset. The CVAE is trained to generate images of clothing, conditioned on top ten categories of deep fashion data set. We used Adam for stochastic optimization. For the CAVE network we used a learning of 0.0003 and a mini-batch of size 32.

Coping with Noise in Gaze Encoder Prediction. Despite the aggregation as described in section 5.3.1, inferred quantities from gaze are typically still noisy. To cope with this challenge, we try different pruning strategies to suppress weak activations in the semantic representation c.

Specifically, we tried four scenarios to decode visual search target from the gaze. In the first case, we used plain posterior as a conditioned vector. In the remaining cases, we used only the top 1 to top 3 highest activated classes in the posterior as a condition vector for encoder. All other probabilities are set to zero, and the posterior is renormalised afterwards.

5.3.2 Qualitative Results of Search Target Decoding

Figure 5.3, Figure 5.4,and Figure 5.5 show qualitative results of our approach. Noise is inherent in the gaze encoding, and we, therefore, consider four pruning strategies to suppress minor activation. In the first case, we used plain posterior as a conditioned vector. In the remaining cases, we used only the top 1 to top 3 highest activated classes in the posterior as a condition vector for encoder. All other probabilities are set to zero, and the posterior is renormalised afterwards.

Using directly the posterior of the CNN with gaze pooling layer causes images that contain several categories rather than the intended visual search target. As shown in Figure 5.3, the image reconstructed from unaltered posteriors (first row) is mostly blurry and does not seem to contain one specific category (e.g. Z1 looks like a Blouse, Z6 is a dress and Z9 is a skirt). Images from top2 and top1 appear to be more focused on the intended category. Using top2 posteriors generates images which are a mixture between the two posteriors. For example the top2, is a composition between dress and skirt, whereas the top1 only contains skirt.

Images from top1 are sharper and mostly contain one category. However, if the predicted category is wrong, we can not decode the intended category (last row of Figure 5.4). As we observed in Chapter 4, we achieved stronger recognition performance for top 2 and top 3 compared to top1. Hence, using the top2 and top3 is more likely to contain information about the target.

As one can see in Figure 5.4, top1 decoded a dress, although the intended category was a tank. The intended category is recovered in the top2 and top3 decodings. As top2 results are giving images with preferred search target, for further analysis, we choose the decoded images from top2. Figure 5.5, shows ten samples for each category using top2 posteriors based on fixations from one user. Our approach can generate images of different categories. The model performs better for several classes, as Jean, Shorts, Skirt, Dress, Tank, Sweater and Tee. Images from the cardigan

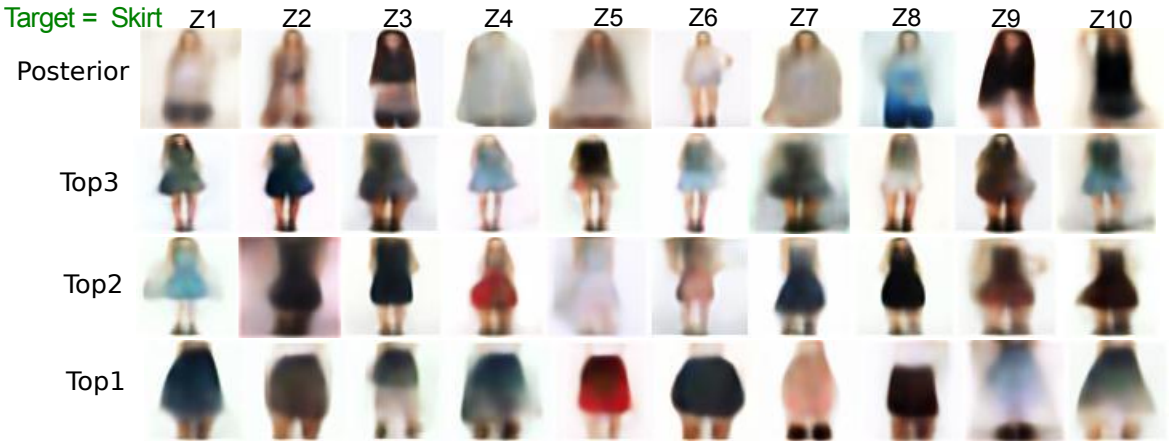


Figure 5.3: Using all posteriors gives images that contain several categories. Using the only top3 to top1 posterior gives images which contain the intended categories. As we move from posteriors to top1, the decoded image is more localised and contains fewer classes. Top3 images have a full body part, as we move to top 1, can see only lower body part that contains a skirt.

and jacket are more similar to each other, although there are still differences in the appearance. In particular, images depicting the search target cardigan appear more elongated compared to those for jackets.

5.3.3 User Study: Search Target Recognition

To assess the accuracy of the search target reconstruction, we run a user study. In this user study, we show our participant 10 generate samples from a category along with a list of all possible categories in our dataset. The participant should pick one category that represents the generated image the best. The generated images are based on the fixation data. The average accuracy of 19 users across categories was 62%, and the detailed confusion matrix is shown in Table 5.2. There are confusion between cardigan and jacket, also a skirt and blouse with a dress. All search targets were recognised significantly above chance (10%). Users were most confident for jeans, shorts and dress.

5.3.4 User Study: Local Vs Global Gaze Encoding

In this user study, we test the importance of local gaze information for the decoding of visual search targets. Our full model is denoted as "local" , as it uses the full gaze information – in particular the fixation location on the image. We compare it to the "global" model which uses gaze information – but only to the extent that we know an image was fixated without knowing the exact location. This also connects to the

Figure 5.4: Top3 and top2 were able to capture the right category, the decoded images contain the target "Tank". However, due to wrong prediction for top1 resulted decoding looks like a "Dress".

	P1	P2	P3	P4	P5	P6	P7
Local	**80%**	**70%**	**60%**	**70%**	**70%**	**60%**	**50%**
Global	20%	30%	40%	30%	30%	40%	**50%**

Table 5.1: All of the users, preferred the decoding using local information over global method. This indicates the importance of local information on decoding the users' intents.

analysis performed on the recognition task for the gaze pooling layer in Chapter 4. We ask how much of a difference these approaches make in terms of search target reconstruction.

In this user study, each participant saw two rows of search target reconstructions. One row was generated by the local, the other by the global method (Figure 5.6. The users were instructed to select the row, which matches the best, the given search target category. The users selected the local encoding method in 65% of the cases. The chance level for this experiment is 50% as for each image; the participates do a binary task. Consequently, the gain is $65\%/50\% = 13\%$. The gain indicates how much performance of users differs from a random selection. Also, we performed Chi-Square Goodness of Fit Test, to investigate the significance of our result. The null hypothesis was that both local and global decoding are equal. The χ^2 value is 6.914 and P-Value is 0.009. Hence, the result is significant at $p \leq 0.05$ and therefore, local information is key for improved search target reconstruction. Detailed results are shown in Table 5.1.

Figure 5.5: Each row is the decoded search target of a user for the given category using only top2 posteriors. Each column is for different samples of z from a normal distribution. As one can see the decoded search targets are distinctive from one another and they represent their corresponding categories properly.

	Blouse	T-Shirt	Jean	Shorts	Skirt	Cardigan	Dress	Jacket	Sweater	Tanks
Blouse	**42%**	5%	0%	0%	5%	0%	47%	0%	0%	0%
T-Shirt	21%	**74%**	5%	0%	0%	0%	0%	0%	0%	0%
Jean	0%	0%	**95%**	5%	0%	0%	0%	0%	0%	0%
Shorts	0%	0%	0%	**95%**	5%	0%	0%	0%	0%	0%
Skirt	0%	0%	0%	0%	**42%**	0%	58%	0%	0%	0%
Cardigan	0%	0%	0%	0%	0%	**37%**	0%	58%	5%	0%
Dress	0%	16%	0%	0%	5%	5%	**74%**	0%	0%	0%
Jacket	0%	0%	0%	0%	0%	47%	0%	**47%**	0%	5%
Sweater	16%	0%	0%	0%	0%	10%	0%	16%	**58%**	0%
Tanks	16%	0%	0%	0%	5%	5%	21%	0%	0%	**53%**

Table 5.2: Confusion Matrix of "Search Target Recognition". One can see in all of the cases, users were able to recognize the right categories above chance level 10%. (Bold number on diagonal corresponds to classification accuracy per class). However, Classes "Blouse" and "Skirt" are confused with "Dress" (in red). "Jacket" and "Cardigan" (in blue) where the other classes which users tend to be more confuse about them.

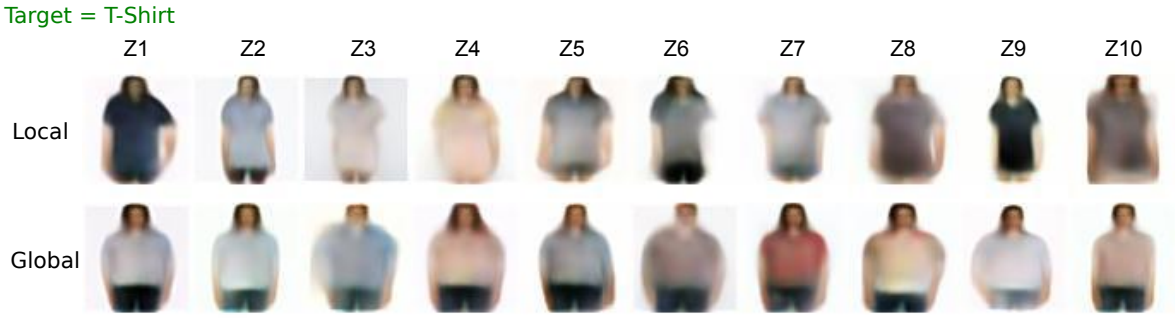


Figure 5.6: Example image used in our second user study. For each category, users need to select between local and global decoded target. The local method encodes the gaze data using gaze-pooling layer which benefits from user intended local image regions.

5.4 CONCLUSION

In this chapter, we introduce the first approach to decode visual search target of users from their gaze data. This task is very challenging as the target only resides in user mind. For this aim, we used recent advances in generative image models and search target prediction using gaze data. Two user studies establish proof of concept, showing that the decoded target lead to human recognisable visual representations

as well as highlighting the importance of localised gaze information. We like to emphasise that due to the training setup, the method remains highly practical and applicable, as no large scale gaze data had to be collected or used. Key is rather the utilisation of a semantic layer that connects the gaze encoder with the conditional generative image model.

Part II

SHAPE BASED INFERENCE

Body shape of a person could affect a person's interest and decisions. A person shape could influence their interest in the type of clothing items they would buy, the food they would eat, or amounts of sports they would do. Hence, in addition to gaze data, we also used 3D shape data to understand how body shape correlates with people's clothing preferences.

In Chapter 6, we study the correlation between clothing garments and body shape on our new Fashion Takes Shape dataset. We introduced a new multi-photo approach for 3D shape estimation and build a conditional model of clothing categories given body-shape.

In Chapter 7, we introduced an evasion method that can be used to effectively manipulate the automatic shape estimation while preserving the overall appearance of the original image. We performed a user study to understand users' concerns w.r.t. privacy depiction of the 3D body shape in public and different application contexts.

6 FASHION IS TAKING SHAPE: UNDERSTANDING CLOTHING PREFERENCE BASED ON BODY SHAPE FROM ONLINE SOURCES

To study the correlation between clothing garments and body shape, we collected a new dataset (Fashion Takes Shape), which includes images of female users with clothing category annotations. Despite the progress in body shape estimation from images, it turns out to be challenging to infer body shape from such diverse, real-world photos. Hence, we propose a novel and robust multi-photo approach to estimate body shapes of each user and build a conditional model of clothing categories given body-shape. We demonstrate that in real-world data, clothing categories and body-shapes are correlated and show that our multi-photo approach leads to a better predictive model for clothing categories compared to models based on single-view shape estimates or manually annotated body types. We see our method as the first step towards the large-scale understanding of clothing preferences from body shape.

6.1 INTRODUCTION

Fashion is a $2.4 trillion industry[3] which plays a crucial role in the global economy. Many e-commerce companies such as Amazon or Zalando makes it possible for their users to buy clothing online. However, based on a recent study, [4] around 50% of bought items were returned by users. One major reason of return is "It doesn't fit" (52%). Fit goes beyond the mere size — certain items look good on certain body shapes and others do not. In contrast to in store shopping where one can try on clothing, in online shopping users are limited to a coarse set of numeric size ranges (e.g. 36, 38 and so on) to predict the fitness of the clothing item. Also, they only see the clothing worn by a professional model, which does not account for the diverse body shape of people. A clothing item that looks very good on a professional model body could look very different on another person's body. Consequently, understanding how body shape correlates with people's clothing preferences could avoid such confusions and reduce the number of returns.

Due to importance of the fashion industry, the application of computer vision in fashion is rising rapidly. Especially, clothing recommendation Kiapour *et al.* (2014); Simo-Serra *et al.* (2015); Liu *et al.* (2016a); Han *et al.* (2017b) is one of the

[3] https://www.mckinsey.com/industries/retail/our-insights/the-state-of-fashion

[4] https://www.ibi.de/files/Competence%20Center/Ebusiness/PM-Retourenmanagement-im-Online-Handel.pdf

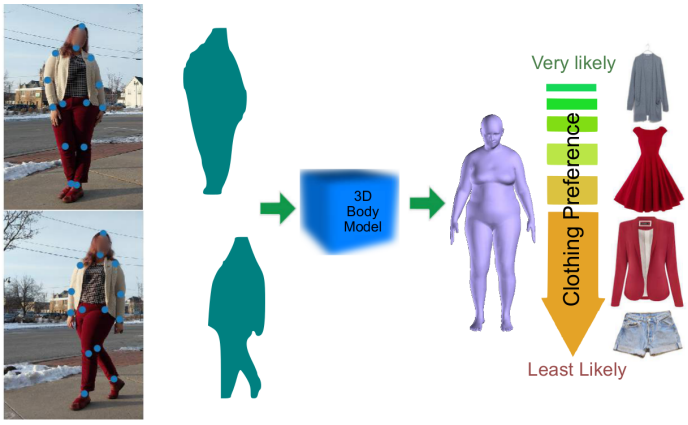


Figure 6.1: Our multi-photo approach uses 2D body joint and silhouette to estimate 3D body shape of the person in the photo. Our shape conditioned model of clothing categories uses the estimated shape to predict the best fitting clothing categories.

hot topics in this field along with cloth parsing Yamaguchi *et al.* (2012); Yang *et al.* (2014); Gong *et al.* (2017), recognition Chen *et al.* (2012); Liu *et al.* (2016a); Hsiao and Grauman (2017); Han *et al.* (2017a); Al-Halah *et al.* (2017) and retrieval Wang and Zhang (2011); Yamaguchi *et al.* (2013); Kiapour *et al.* (2015); Liu *et al.* (2012). Research in the domain of clothing recommendation studies the relation between clothing and categories, location, travel destination and weather. However, there is no study on the correlation between human body shape and clothing. This is probably due to the fact that there exists no dataset with clothing category annotations together with detailed shape annotations.

Therefore, our main idea is to leverage fashion photos of users including clothing category meta-data and, for every user, automatically estimating their body shape. Using this data we learn a conditional model of clothing given the inferred body shape.

Despite recent progress, the visual inference of body shape in unconstrained images remains a very challenging problem in computer vision. People appear in different poses, wearing many different types of garments, and photos are taken from different camera viewpoints and focal lengths.

Our key observation is that users typically post several photos of themselves, while viewpoint and body pose varies across photos, the body shape *does not* (within a posted gallery of images). Hence, we propose a method that leverages multiple photos of the same person to estimate their body shape. Concretely, we first estimate body shape by fitting the SMPL body model Loper *et al.* (2015) to each of the photos separately, and demonstrate that exhaustively searching for depth improves performance. Then, we reject photos that produce outlier shapes and optimize for a single shape that is consistent with each of the inlier photos. This results in a robust *multi-photo* method to estimate body shape from unconstrained photos on the internet.

We use image from the web [5] and collected a dataset (*Fashion Takes Shape*) which includes more than 18000 images with meta-data including clothing category, and a *manual shape annotation* indicating whether the person's shape is above average or average. The data comprises 181 different users. Using our multi-photo method, we estimated the shape of each user. This allowed us to study the relationship between clothing categories and body shape. In particular, we compute the conditional distribution of clothing category conditioned on body shape parameters.

To validate our conditional model, we compute the likelihood of the data (clothing categories worn by the user) under the model and compare it against multiple baselines, including a marginal model of clothing categories, a conditional model built using the manual shape annotations, and a conditional model using a state of the art single view shape estimation method Bogo *et al.* (2016).

Experiments demonstrate that our conditional model with multi-photo shape estimates always produces better data-likelihood scores than the baselines. Notably, our model using automatic multi-photo shape estimation even outperforms a model using a coarse manual shape annotations. This shows that we extract more fine-grained shape information than manual annotations. This is remarkable, considering the unavoidable errors that sometimes occur in automatic shape estimation from monocular images.

We see our method as the first step towards the large-scale understanding of clothing preferences from body shape. To stimulate further research in this direction, we will make the newly collected Fashion Takes Shape Dataset (FTS), and code available to the community. FTS includes clothing meta-data, 2D joint detection, semantic segmentation and our 3D shape-pose estimates.

6.2 ROBUST HUMAN BODY SHAPE ESTIMATION FROM PHOTO-COLLECTIONS

Our goal is to relate clothing preferences to body shape automatically inferred from photo-collections. Here, we build on the SMPL Loper *et al.* (2015) statistical body model that we fit to images. However, unconstrained online images make the problem very hard due to varying pose, clothing, illumination and depth ambiguities.

To address these challenges, we propose a robust multi-photo estimation method. In contrast to controlled multi-view settings where the person is captured simultaneously by multiple cameras, we devise a method to estimate shape leveraging multiple photos of the same person in different poses and camera viewpoints.

From a collection of photos, our method starts by fitting SMPL (Sec. 6.2.1) to each of the images. This part is similar to Bogo *et al.* (2016); Alldieck *et al.* (2018) and not part of our contribution and we describe it in Sec. 6.2.2 for completeness. We demonstrate (Sec. 6.2.3) that keeping the height of the person fixed and initializing optimization at multiple depths significantly improves results and reduces scale ambiguities. Then, we reject photos that result in outlier shape estimates. Using

[5] http://www.chictopia.com

the inlier photos, our multi-photo method (Sec. 6.2.4) jointly optimizes for multiple cameras, multiple poses, and a *single coherent shape*.

6.2.1 Body Model

SMPL Loper *et al.* (2015) is a state of the art generative body model, which parameterizes the surface of the human body with shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the PCA coefficients of a shape space learned from thousands of registered 3D scans. The shape parameters encode changes in height, weight and body proportions. The body pose $\boldsymbol{\theta} \in \mathbb{R}^{3P}$, is defined by a skeleton rig with $P = 24$ joints. The joints $J(\boldsymbol{\beta})$ are a function of shape parameters. The SMPL function $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ outputs the $N = 6890$ vertices of the human mesh transformed by pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$.

In order to "pose" the 3D joints, SMPL applies global rigid transformations $\mathbf{R}_{\boldsymbol{\theta}}$ on each joint i as $R_{\boldsymbol{\theta}}(J_i(\boldsymbol{\beta}))$.

6.2.2 Single View Fitting

We fit the SMPL model to 2D body joint detection $\mathbf{J}_{est}$ obtained using model of Insafutdinov *et al.* (2016), and a foreground mask $\mathbf{S}$ computed using method of Pinheiro *et al.* (2016). Concretely, we minimize an objective function with respect to pose, shape and camera translation $\mathbf{K} = [X, Y, Z]$:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_P(\boldsymbol{\beta}, \boldsymbol{\theta}) + E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{J}_{\text{est}}) + E_h(\boldsymbol{\beta}) + E_S(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{S}), \tag{6.1}$$

where $E_P(\boldsymbol{\beta}, \boldsymbol{\theta})$ are the four prior terms as described in Bogo *et al.* (2016), and the other terms are described next.

Joint-based data term:. We minimize the re-projection error between SMPL 3D joints and detection:

$$E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{J}_{\text{est}}) = \sum_{\text{joint}j} \omega_j \rho(\Pi_{\mathbf{K}}(R_\theta(J(\boldsymbol{\beta})_j)) - \mathbf{J}_{\text{est},j}) \tag{6.2}$$

where $\Pi_{\mathbf{K}}$ is the projection from 3D to 2D of the camera with parameters $\mathbf{K}$. ω_j are the confidence scores from CNN detection and ρ a Geman-McClure penalty function which is robust to noise.

Height term:. Previous work Bogo *et al.* (2016); Lassner *et al.* (2017b) jointly optimizes for depth (distance from the person to camera) and body shape. However, the overall size of the body and distance to the camera are ambiguous; a small person closer to the camera can produce a silhouette in the image just as big as a bigger person farther from the camera. Hence, we aim at estimating body shape up to a scale factor. To that end, we add an additional term that constrains the height of the person to remain very close to the mean height T_H of the SMPL template

$$E_h(\boldsymbol{\beta}) = ||M_h(0, \boldsymbol{\beta}) - T_H||_2^2,$$

where height $M_h(0, \boldsymbol{\beta})$, is computed on the optimized SMPL model before applying pose. This step is especially crucial for multi-photo optimization as it allows us to analyze shapes at the same scale.

Silhouette term:. To capture shape better, we minimize the miss-match between the model silhouette $\mathbf{I_s}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{K})$, and the distance transform of the CNN-segmented mask $\mathbf{S}$ Pinheiro *et al.* (2016):

$$E_S(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{S}) = G(\lambda_1 \mathbf{I_s}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{K})\mathbf{S} + \lambda_2(\mathbf{1} - \mathbf{I_s}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{K}))\overline{\mathbf{S}}), \tag{6.3}$$

where G is a Gaussian pyramid with 4 different levels, $\mathbf{K}$ is the camera parameter, and $\overline{\mathbf{S}}$ is the distance transform of the inverse segmentation, and λ is a weight balancing the terms.

6.2.3 Camera Optimization

Camera translation and body orientation are unknown. However, we assume that rough estimation of the focal length is known. We set the focal length as two times the width of the image. We initialize the depth Z via the ratio of similar triangles, defined by the shoulder to ankle length of the 3D model and the 2D joint detection. To refine the estimated depth, we minimize the re-projection error E_J of only torso, knee, and ankle joints (6 joints) with respect to camera translation and body root orientation. At this stage, $\boldsymbol{\beta}$ is held fixed to the template shape. We empirically found that a good depth initialization is crucial for good performance. Hence, we minimize the objective in 7.1 at 5 different depth initializations – we sample in the range of [-1,+1] meters from the initial depth estimate. We keep the shape estimate from the initialization that leads to a lower minimum after convergence. After obtaining the initial pose and shape parameter we refine the body shape model adding silhouette information.

6.2.4 Robust Multi-Photo Optimization

The accuracy of the single view method heavily depends on the image view-point, the pose, the segmentation and 2D joint detection quality. Therefore, we propose to jointly optimize one shape to fit several photos at once. Before optimizing, we reject photos that are likely to be outliers in order to make the optimization more robust. First, we compute the median shape from all the single view estimates and keep only the views whose shape is closest to the median. Using these inlier views we jointly optimize V poses $\boldsymbol{\theta}^i$, and a single shape $\boldsymbol{\beta}$. We minimize the re-projection error in all the photos at once:

$$E_{mp}(\boldsymbol{\beta}, \boldsymbol{\theta}^{\forall i}; \mathbf{K}^{\forall i}, \mathbf{J}^{\forall i}_{\text{est}}, \mathbf{S}^{\forall i}) = \sum_{i=1}^{K} E(\boldsymbol{\beta}, \boldsymbol{\theta}^i, ; \mathbf{K}^i, \mathbf{J}^i_{\text{est}}, \mathbf{S}^i) \tag{6.4}$$

where $E()$ is the single view objective function, Eq. 7.1 and K are number of views we kept after outlier rejection. Our multi-photo method leads to more accurate shape

estimates as we show in the experiments.

6.3 EVALUATION

To evaluate our method, we proposed two datasets: synthetic and real images. We used the synthetic dataset to perform an ablation study of our multi-photo body model. Using unconstrained real-world fashion images, we evaluate our clothing category model conditioned on the multi-photo shape estimates.

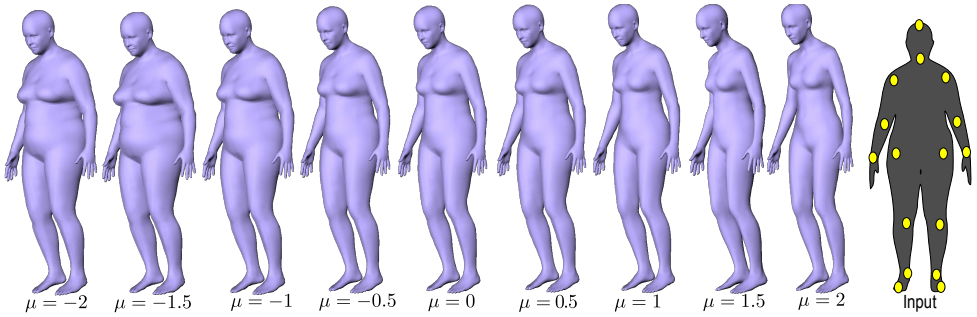
6.3.1 Synthetic Bodies

SMPL is a generative 3D body model which is parametrized with pose and shape. We observe that while the first shape parameter produces body shape variations due to scale, the second parameter produces shapes of varied weight and form. Hence, we generate 9 different bodies by sampling the second shape parameter from $\mathcal{N}(\mu, 1)$ and $\mu \in [-2 : 0.5 : 2]$. In Figure 6.2, we show the 9 different body shapes and a representative rendered body silhouette with 2D joints which we use as input for prediction. We also generate 9 different views for each subject to evaluate our multi-photo shape estimation in a controlled setting (Figure 6.2). In all experiments, we report the mean Euclidean error between the estimated shape parameters $\boldsymbol{\beta}$ and the ground truth shapes.

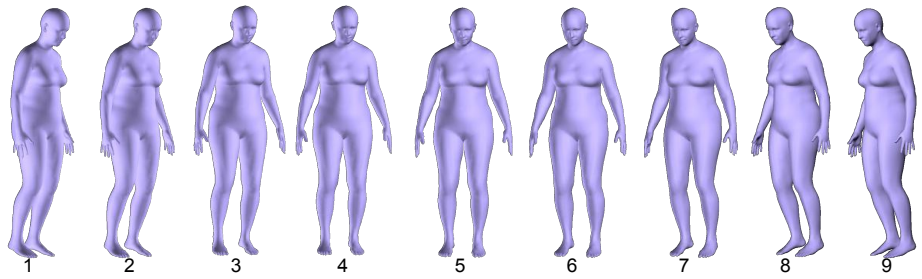
Single-View Shape Estimation:. In this controlled, synthetic setting, we have tested our model in several conditions. We summarize the results of our single view method in the first column of Table 6.1 (mean shape error over all 9 subjects and 9 views), and plot the error w.r.t. viewpoints in Figure 6.3. Please note in Table 6.1 that SMPLifyBogo *et al.* (2016) can only use one photo, and therefore columns corresponding to multiple photos are marked as "na" (not available).

Overall, we see in Table 6.1 a reduction of shape estimation error from 1.05 by SMPLify Bogo *et al.* (2016) to 0.91 of our method by adding joint estimation (J), silhouette features (S) and depth selection (DS). The depth selection (DS) strategy yield the strongest improvement. We see an additional decrease to 0.84 by considering multiple photos ($k = 5$) despite having to estimate camera and body pose for each additional photo.

People with similar joint length could have different body mass. As SMPLify only uses 2D body joint as input, it is not able to estimate the body shape with high accuracy. Silhouette is used to capture a better body shape. However adding silhouette with a wrong depth data, decreases the accuracy of shape estimation drastically (Ours using 2D joint and silhouette (Ours+J+S), red curve in Figure 6.3 from 0.91 to 1.20. Hence, to study the impact of depth accuracy in shape estimation, we provide ground-truth depth (in Figure 6.3, green curve: *Ours(D)*) to our method. We observe that using ground truth depth information improves the error of Our+J+S from 1.20 to 0.86.



(a) Variation in body shapes



(b) Variation in views

Figure 6.2: Using SMPL 3D body model we generate 9 subject each with 9 views. We study the effectiveness of our method on this dataset. As shown in (a) the input to our system is only image silhouette and a set of 2D body joints.

This argues for the importance of our introduced depth search procedure. Indeed, we find that our model with depth selection ("Ours+J+S+DS") yields a reduced error of 0.91 *without using any ground truth information.*

Multi-Photo Shape Estimation:. Table 6.1 present also our result for multi-photo case. Real-world images exhibit noisy silhouettes and 2D joints, body occlusion, variation in camera parameters, viewpoints and poses. Consequently, we need a robust system that can use all the information to obtain an optimized shape. Since every single photo may not give us a very good shape estimation, we jointly optimize all photos together. However, for certain views, estimating the pose and depth is very difficult. Consequently, adding those views leads to worse performance. Hence, before optimization, we retain only the K views with shape estimates closest to the median shape estimate of all views. This effectively rejects outlier views. The results are summarized in Table 6.1. K is the number of photos we kept out of the total of 9 to perform optimization. Using only 2D joint data, we optimized the shape in multi-photo setting (Ours+J in Table 6.1). In the second step we add silhouette term to our method in multi-photo optimization (Ours+J+S). Both of these experiments shows decrease in accuracy of the the estimated shape compared to SMPLify. However, for our full method which uses up to $K = 5$ views, we observe a consistent decrease in error; beyond 5 the error increases, which supports the

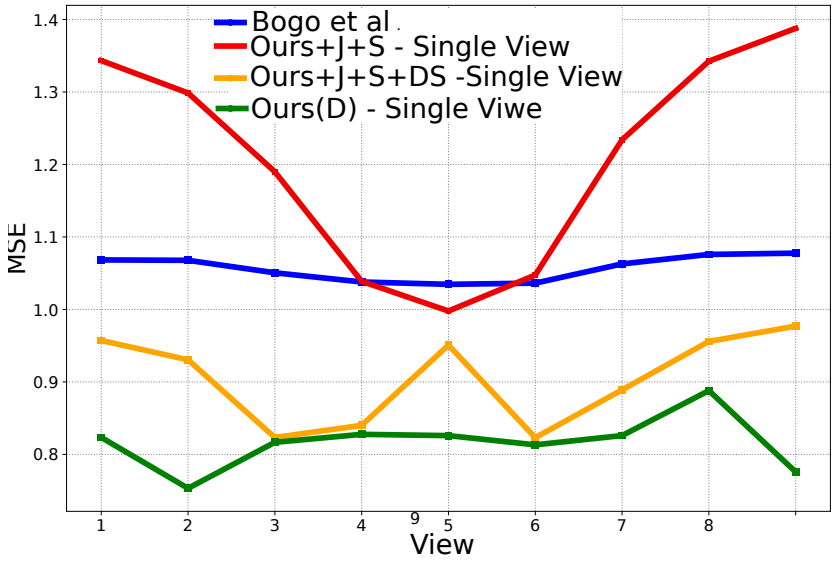


Figure 6.3: The plot shows the mean euclidean norm between estimated and the ground truth shape among all subjects for each view on synthetic data.

	single-view	k=2	k=3	k=4	k=5	k=6	k=7
SMPLify Bogo *et al.* (2016)	1.05	na	na	na	na	na	na
Ours+J	1.05	1.81	1.80	1.80	1.80	1.80	1.82
Ours+J+S	1.20	1.80	1.80	1.80	1.80	1.80	1.77
Ours+J+S+DS	0.91	0.91	0.88	0.87	0.84	0.85	0.88
Ours(D)	0.86	0.84	0.79	0.77	0.80	0.83	0.92

Table 6.1: We present the results for Multi-Photo optimization on the synthetic data. The error is the L2 distance between ground-truth and estimated shape parameter. Ours(D) has ground truth camera translation(depth) data.

effectiveness of the proposed integration and outlier detection scheme. We improve over our single view estimate reducing the error *further* from 0.91 to 0.84 (for $K = 5$) – and even approach the oracle performance (0.86) where ground truth depth is given.

6.3.2 Fashion Takes Shape Dataset

Not every clothing item matches every body shape. Hence our goal is to study the link between body shapes and clothing categories. In order to study these correlations, we collected data from 181 female users of "Chictopia"[6] (online fashion

[6]http://www.chictopia.com

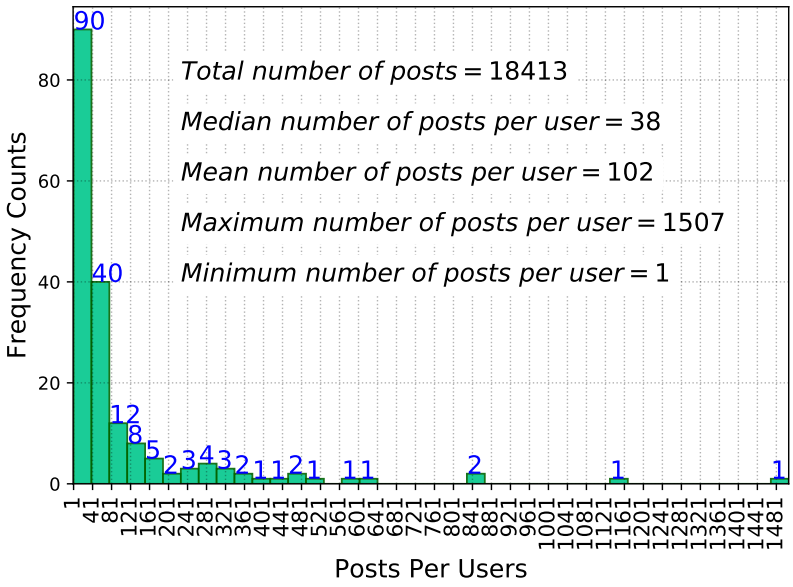


Figure 6.4: Histogram of posts in our dataset. A total number of posts from all our users are 18413. Each post has 1 or more images of a person with clothing.

blog).

We look for two sets of users: in the first set, we collected data of users with average and below the average size, which we call group G_a; the second set contains data of above average (plus size) users referred to as G_p. In total, we have 141 users in group G_a and 40 in group G_p which constitute a diverse sample of real-world body shapes. Figure 6.4, shows the summary of our dataset. The total number of posts from all users is 18413 – each post can contains one or more images (usually between 2 to 4). The minimum number of posts per user is 1 and the maximum 1507. In average, we have 102 posts per user, and a median of 38 posts per user. Furthermore, each post contains data about clothing garments, other users opinions (Votes, likes) and comments. Figure 6.5 shows a post uploaded by a user.

6.3.3 Shape Representation

In order to build a model conditioned on shape, we first need a representation of the users' shapes. Physical body measurements can be considered as an option which is not possible when we only have access to images of the person. Hence, we use our multi-photo method to obtain a shape $\beta \in \mathbb{R}^{10}$ estimate of the person from multiple photos.

Since we do not have ground truth shape for the unconstrained photos, we have trained a binary Support Vector Machine (SVM) on the estimated shape parameters β for classification of the body type into G_a and G_p. The intuition is that if our shape estimations are correct, above average and below average shapes should be separable.

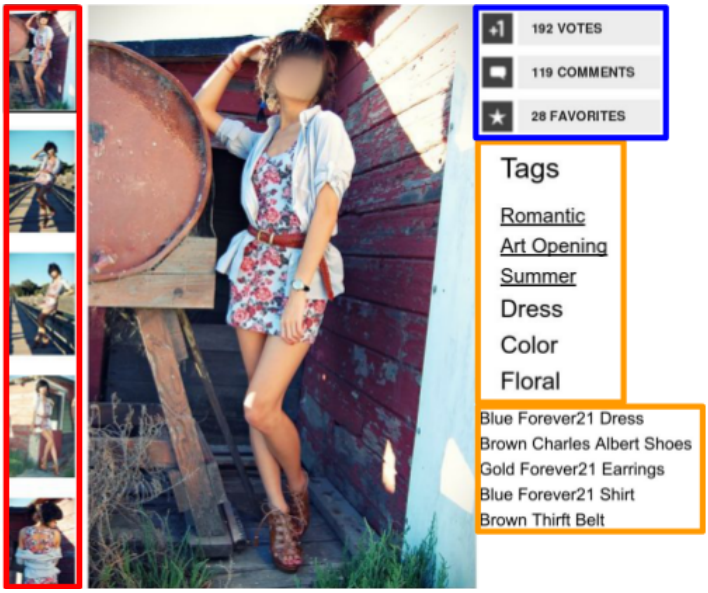


Figure 6.5: Chictopia's users upload images of themselves wearing different garments. Each post has 1 or more image of the person(red box). In addition to images, meta data such as "Tags"(orange box) and users opinion(blue box) are available.

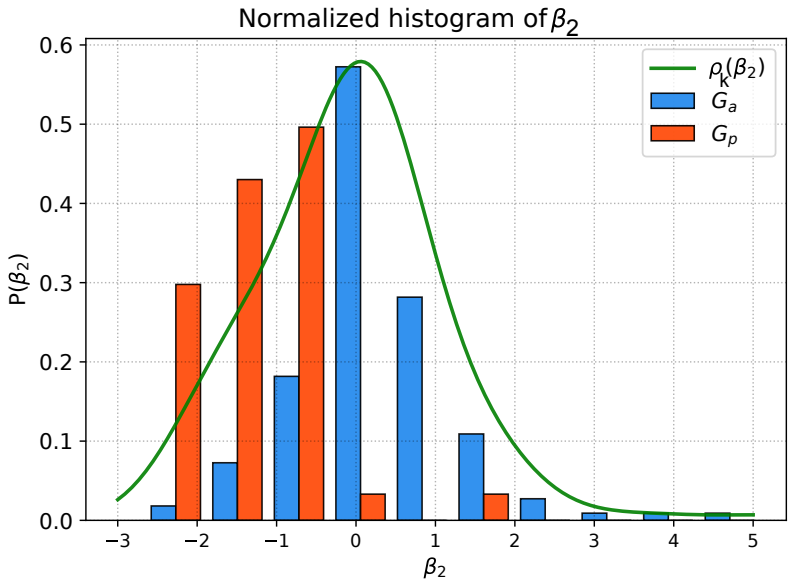


Figure 6.6: Shape distribution of our dataset. The thinner the person the higher values of β_2 they have. While the group G_p has lower values (negative).

Indeed, the SVM obtains an accuracy (on a hold-out test set) of 87.17% showing that the shape parameter is at the very least informative of the two aforementioned body classes. Looking at the SVM weights, we recognize that the second entry of the β

vector has the most contribution to the classifier. Actually, classifying the data by simple thresholding of the second dimension of the β vector results in an even higher accuracy of 88.79%. Hence, for following studies, we have used directly the second dimension of β. We illustrate the normalized histogram of this variable in Figure 6.6. The normalized histogram suggests that users in group G_p have negative values whereas group G_a have positive values. For later use, we estimate a probability density function (pdf) for this variable β_2 with a kernel density estimator – using a Gaussian kernel:

$$\rho_K(\beta_2) = \frac{1}{N}\sum_{i=1}^{N} K((\beta_2 - \beta_2^i)/h) \tag{6.5}$$

The green line in Figure 6.6 illustrate the estimated pdf $\rho_K(\beta_2)$ of all users.

6.3.4 Correlation Between Shape and Clothing Categories

The type of clothing garment people wear is very closely correlated with their shape. For example, "Leggings" might look good on one body shape but may not look very good on other shapes. Hence, we introduce 3 models to study the correlation of the shape and clothing categories. Our basic model (**Model 1**) uses only data statistics with no information about users shape. In the second approach (**Model 2**), clothing is conditioned on binary shape categories G_a and G_p – which in fact requires manual labels. The final approach (**Model 3**) is facilitated by our automatic shape parameter estimation β_2.

We evaluate the quality of the model via the negative log likelihood of held-out data. A good model minimizes the negative log-likelihood. Hence a better model should have smaller value in negative log likelihood. The negative log likelihood is defined as:

$$LL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{Z} \log P_M^{c_j^i} \tag{6.6}$$

Where N is the number of users, Z is the number of clothing categories and $c_j^i \in C_i$ is a vector of user's clothing categories. M represents the model. In the following, we present the details of these three approaches. The log likelihood of each approach is reported in the Table 6.2.

Model 1: Prediction Using Probability of Clothing Categories:. We established a basic model using the probability of the clothing categories $P_M = p(c)$. The clothing categories tag of the dataset is parsed for fourteen of the most common clothing's categories. Category "Dress" has the highest amount of images (Figure 6.7) whereas "Tee and Tank" were not tagged very often.

Model 2: Prediction for Given G_a and G_p:. This is based on the conditional probability of the clothing category given the annotated body type $P_M = p(c|G)$ where $G \in G_a, G_p$. From this measure we find out that several clothing categories

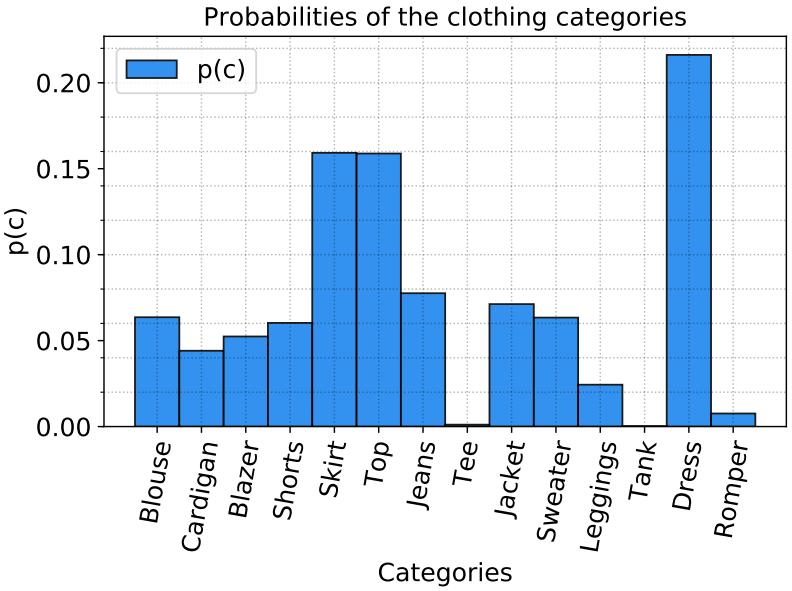


Figure 6.7: Probabilities of 14 different clothing garments of our dataset.

are more likely for certain group (Figure 6.8). As an example, while "Cardigan" and "Jacket" have higher probabilities for the G_p group, users in G_a were more likely to wear "Short" and "Skirt".

Model 3: Prediction for a Given Shape β_2:. As shown in the second model, body types and clothing garments are correlated. However, categorizing people only into two or more categories is not desirable. First, it requires tedious and time consuming manual annotation of body type. Second, the definition of the shape categories is very personal and fuzzy.

The estimated shape parameters of our model provides us with a continuous fine-grained representation of human body shape. Hence, we no further need to classify people in arbitrary shape groups. Using the shape parameter β_2 and statistics of our data, we are able to measure the conditional probability of shapes for a given clothing category $p(\beta_2|c)$. This probability is measured for wearing and not wearing a certain category. The result is shown in Figure 6.9 where for each category the Blue line represents wearing the category. Similarly to the previous model, one can see users with negative values of β_2 wearing "Cardigan", where the probabilities of wearing "Short" and "skirt" is skewed towards positive values of β_2. Furthermore, using the Bayes rule we can predict clothing condition on the body shape $P_M = p(c|\beta_2)$ as:

$$p(c|\beta_2) = \frac{p(\beta_2|c)p(c)}{P(\beta_2)} \tag{6.7}$$

The green line in Figure 6.9 illustrate the $p(c|\beta_2)$.

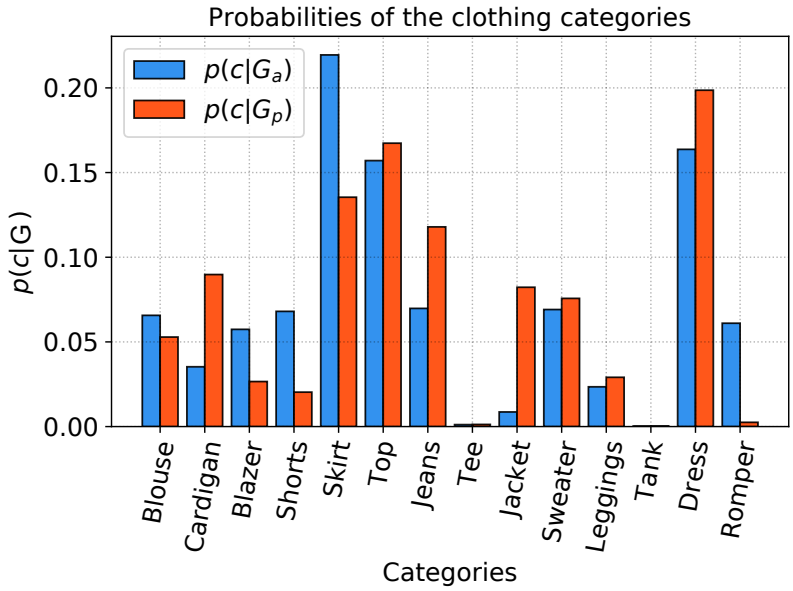


Figure 6.8: Given the Body type (G_a and G_p), we measured the probability of each clothing category in our dataset.

Negative Log likelihood. We quantify the quality of our prediction models by the negative log likelihood of held out data. As we are using negative of log likelihood, the model with smallest values is the best. The results, of each model on our dataset, is summarized in the Table 6.2. Also, we used the estimated shape parameters of Bogo et al Bogo *et al.* (2016) and ours for comparison. In addition to our method which optimized multi-photo, we only used the median shape among photos of a user as a baseline as well. We also include the likelihood under the prior as a reference. For better analysis, we split the users into 4 groups: The first group contains users which there is only a single photo ($I = 1$) of them in each post. In the second group, users always have 2 images($I = 2$) and third group contains users with 3 or more images of a clothing($I \geq 3$). Finally, we also show results with taking into account all users. Table 6.2, shows that our method obtained the smallest negative log-likelihood on the full dataset – in particular outperforming the model that conditions on the two discrete labeled shape classes, shape based on prior work SMPLifyBogo *et al.* (2016), as well as a naive multi-photo integration based on a median estimate. While the median estimate is comparable if only two views are available, we see significant gains for multiple viewpoint – that also show on the full dataset.

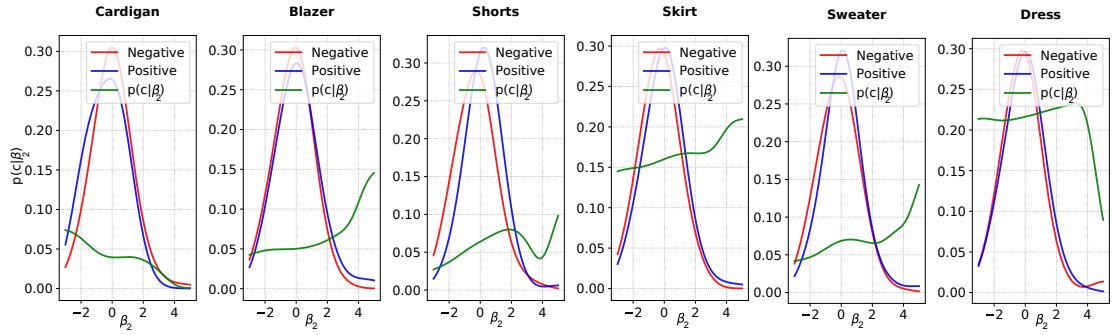


Figure 6.9: Probabilities of $p(\beta_2|c)$ for wearing (blue curves) and not wearing (red curves) of a clothing category on our dataset. Using Bayes rule we can estimate the probability of clothing given the shape $p(c|\beta_2)$(green curve). Negative values of β_2 corresponds to above average while average and below average users have positive values for β_2.

	Model 1 p(c)	Model 2 p(c|G_a/G_p)	$p(c\|\beta_2)^1$	Medianβ_2	Model 3 $p(c\|\beta_2)^2$
I = 1	12.81	12.80	13.63	-	**12.16**
I = 2	13.31	13.47	13.34	**13.09**	13.11
I $\geq$ 3	19.06	19.11	18.8	18.59	**17.85**
All	20.13	20.39	20.48	20.12	**19.81**

Table 6.2: We measured the negative Log-Likelihood of our different models on held out data. Numbers are comparable within the rows. Smaller is better. $p(c|\beta_2)^1$ uses estimated shape from SMPLifyBogo *et al.* (2016) and $p(c|\beta_2)^2$ uses our estimated shape.

6.4 QUALITATIVE RESULTS ON SHAPE ESTIMATION

In Figure 6.10 we present example results obtained with our method and compare it to the result obtained with SMPLify Bogo *et al.* (2016). SMPLify fits the body model based on 2D positions of body joints that often do not provide enough information regarding body girth. This leads to shape estimates that are rather close to the average body shape for above-average body sizes (rows 1 and 2 in Figure 6.10). SMLPify also occasionally fails to select the correct depth that results in body shape that is too tall and has bent knees (red box). The single-view variant of our approach improves over the result of SIMPLify for the first example in Figure 6.10. However it still fails to estimate the fine-grained pose details such as orientation of the feet (blue box). In the second example in Figure 6.10 the body segmentation includes

a handbag resulting in a shape estimate with exaggerated girth by our single-view approach (yellow box). These mistakes are corrected by our multi-photo approach that is able to improve feet orientation in the first example (blue box) and body shape in the second example (yellow box).

Figure 6.11 presents estimated shapes per each initialized depth along with the multi-photo shape estimates and its comparison with SMPLify and Kanazawa *et al.* Hence, we minimize the objective in 7.1 at 5 different depth initializations – we sample in the range of [-1,+1] meters from the initial depth estimate (column 4). We keep the shape estimate from the initialization that leads to a lower minimum after convergence. After obtaining the initial pose and shape parameter, we refine the body shape model with adding silhouette information. The effect of refinement with silhouette term is especially visible in the second row of Figure 6.11, where SMPLify failed to recover the shape of the person (person is unnaturally thin). Please note that the comparison is not completely fair because the method of Kanazawa *et al.* (2018) uses a gender-neutral SMPL model. However, the issues with scale ambiguity are an issue regardless of whether the models are gender specific or not.

Figure 6.12 shows several users from group G_a and G_p. For each user the shape estimates of our method and of SMPlifyBogo *et al.* (2016) are shown as well. Our method was able to recover more diverse shapes, whereas SMPLify tends to produce shapes that are closer to the average body shape regardless of the true shape of the person in the image. The good performance of our method is due to the fact that our model leverages the silhouette of the person in addition to 2D joints as input. Furthermore, our scale selection and multi-photo optimization refine the estimated shapes further. For a better comparison, all of the estimated shapes in Figure 6.12 are rendered with the same camera parameters.

6.5 CONCLUSION

In this chapter we aimed to understand the connection between body shapes and clothing preferences by collecting and analyzing a large database of fashion photographs with annotated clothing categories. Our results demonstrate that clothing preferences and body shapes are correlated and that we can build predictive models for clothing categories based on the output of automatic shape estimation. To obtain estimates of 3D shape we proposed a new approach that incorporates evidence from multiple photographs and body segmentation and is more accurate than popular recent SMPLify approach Bogo *et al.* (2016). We are making our data and code available for research purposes.

Figure 6.10: Shape estimation results on real data. Note that the shape estimates obtained with SMPLify Bogo *et al.* (2016) are rather close to the average body shape whereas our multi-photo approach is able to recover shape details more accurately both for above-average (rows 1 and 2) and average (rows 3 and 4) body types.

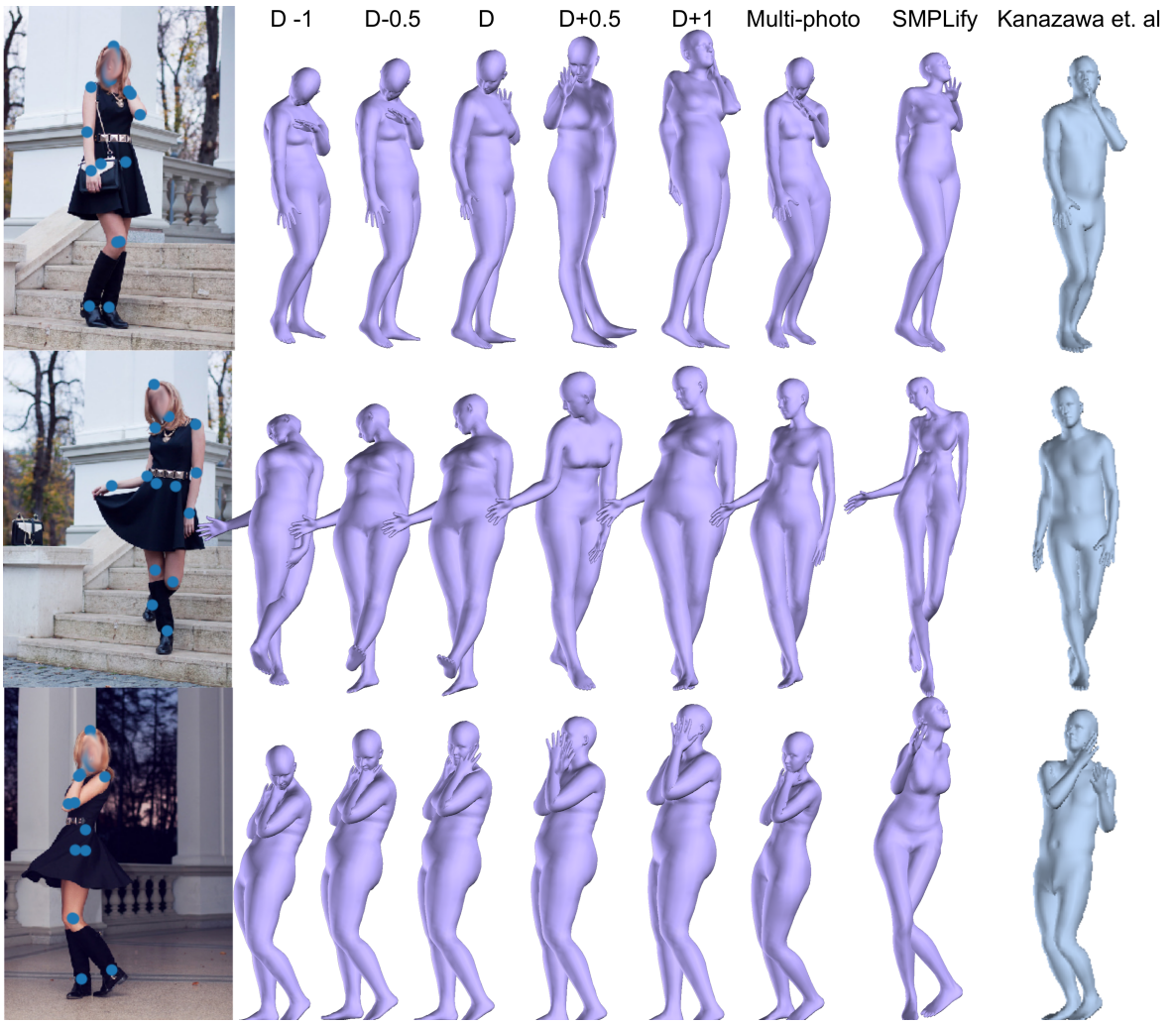


Figure 6.11: Effect of depth selection D on the shape and pose estimation. Initializing the model with the correct camera parameter is crucial and results in a better shape estimates. Our multi-photo approach chooses the shape which is estimated with the initialization that leads to a lower minimum after convergence. We compare our method with Bogo *et al.* (2016); Kanazawa *et al.* (2018).

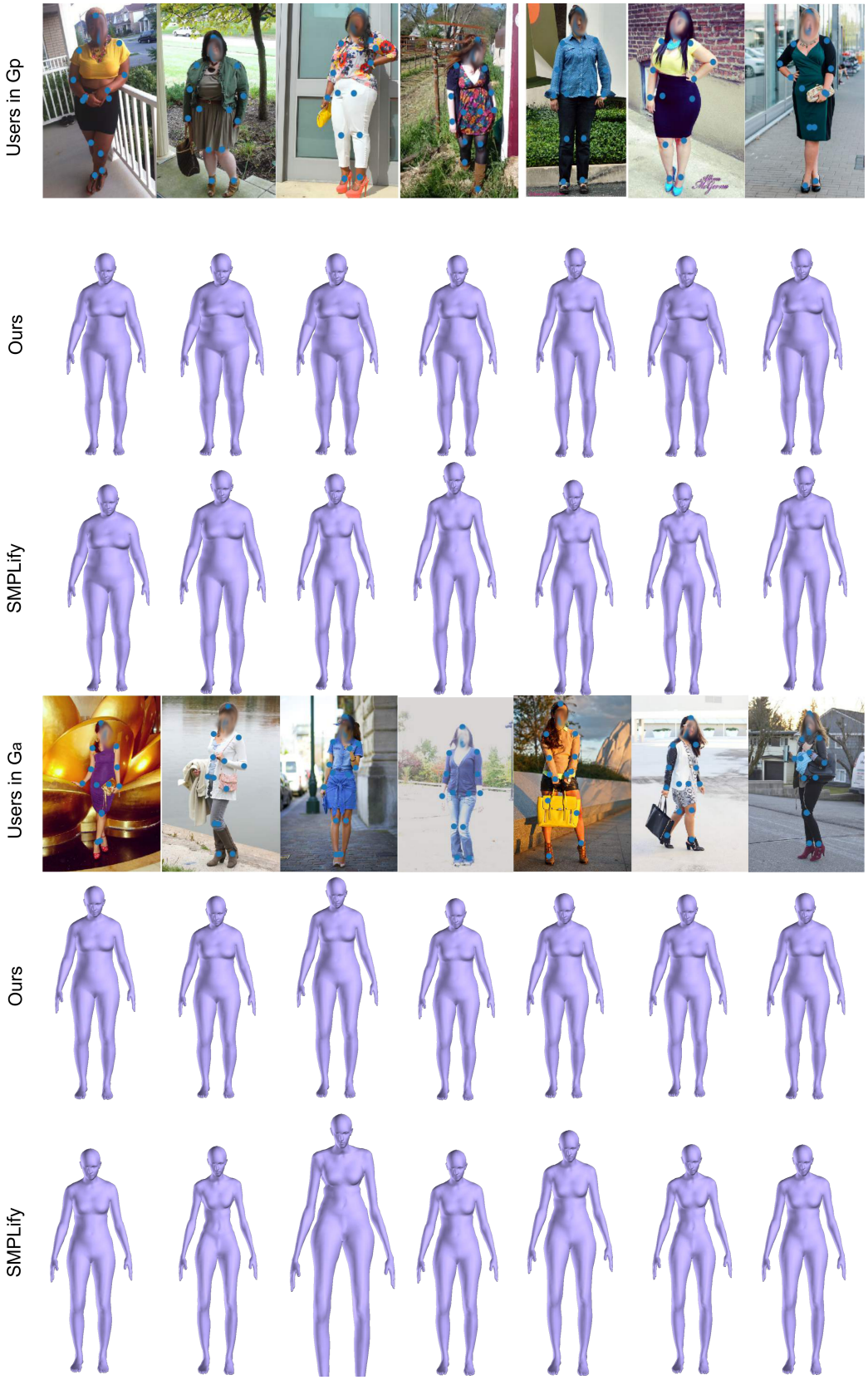


Figure 6.12: Example images of users in our dataset for average G_a and above average G_p group. For each user, the shape estimates of our method and Bogo *et al.* (2016) is shown. The shapes estimates of our method are more diverse compared to Bogo *et al.* (2016) and closer to the person's shape.

7 PREVENTING BODY SHAPE INFERENCE OF MULTI-STAGE APPROACHES

Modern approaches to pose and body shape estimation have recently achieved strong performance even under challenging real-world conditions. Even from a single image of a clothed person, realistic looking body shape can be inferred that captures a users weight group and body shape type well. This opens up a whole spectrum of applications – in particular in fashion – where virtual try-on and recommendation systems can make use of these new and automatized cues. However, a realistic depiction of the undressed body is regarded highly private and therefore might not be consented by most people. Hence, we ask if the automatic extraction of such information can be effectively evaded. While adversarial perturbations are effective for manipulating the output of machine learning models – in particular, end-to-end deep learning approaches – state of the art shape estimation methods are composed of multiple stages. We perform the first investigation of different strategies that can be used to effectively manipulate the automatic shape estimation while preserving the overall appearance of the original image.

7.1 INTRODUCTION

Since the early attempts to recognize human pose in images (Wren *et al.*, 1997; Gavrila, 1999), we have seen a transition to real-world applications where methods operate on challenging real-world conditions in uncontrolled pose and lighting. We have seen more recently progress towards extracting richer representations beyond the pose. Most notably, a full body shape that is represented by a 3D representation or a low dimensional manifold (SMPL) (Loper *et al.*, 2015). It has been shown that such representations can be obtained from fully clothed persons – even in challenging conditions from a single image (Bogo *et al.*, 2016) as well as from web images of a person (Chapter 6 (Sattar *et al.*, 2019c).

On the one hand, this gives rise to various applications – most importantly in the fashion domain. The more accurate judgment of fit could minimize clothing returns, and avatars and virtual try-on may enable new shopping experiences. Therefore, it is unsurprising that such technology already sees gradual adaption in businesses[7], as well as start-ups [8] [9].

On the other hand, the automated extraction of such highly personal information from regular, readily available images might equally raise concerns about privacy.

[7] https://www.cnet.com/news/amazon-buys-body-labs-a-3d-body-scanning-tech-startup/
[8] https://bodylabs.io/en/
[9] https://www.fision-technologies.com/

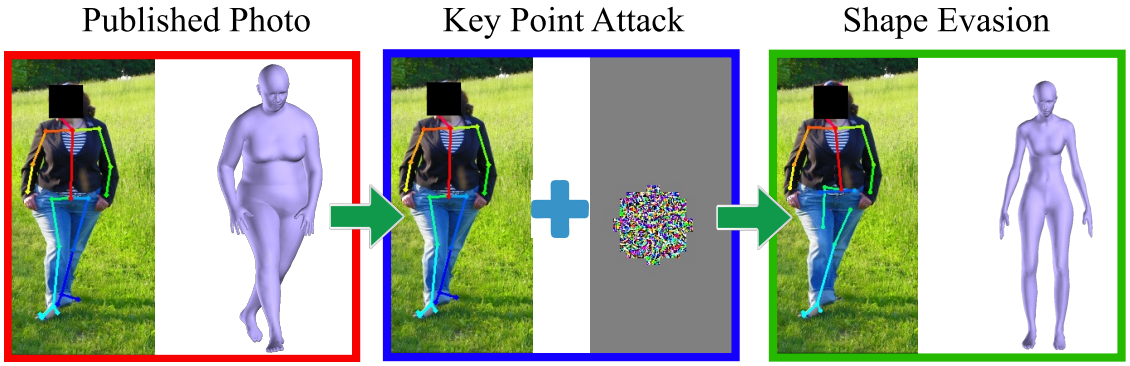


Figure 7.1: A realistic depiction of the naked body is considered highly private and therefore might not be consented by most people. We prevent automatic extraction of such information by small manipulations of the input image that keep the overall aesthetic of the image.

Images contain a rich source of implicit information that we are gradually learning to leverage with the advance of image processing techniques. Only recently, the first organized attempts were made to categorize private information in images (Orekondy *et al.*, 2017) to raise awareness and to activate automatic protection mechanisms.

To control and control private information in images, a range of redaction and sanitization techniques have been introduced (Orekondy *et al.*, 2018; Sun *et al.*, 2018; Tretschk *et al.*, 2018). For example, evasion attacks have been used to disable classification routines to avoid extraction of information. Such techniques use adversarial perturbations to throw off a target classifier. It has been shown that such techniques can generalize to related classifiers (Oh *et al.*, 2017), or can be designed under unknown/black-box models (Oh *et al.*, 2018; Chen *et al.*, 2017; Brendel and Bethge, 2017; Su *et al.*, 2017; Ilyas *et al.*, 2018).

Unfortunately, such techniques are not directly applicable to state-of-the-art shape estimation techniques (Bogo *et al.*, 2016; Alldieck *et al.*, 2018; Lassner *et al.*, 2017b; Alldieck *et al.*, 2019; Habermann *et al.*, 2019), as they are based on multi-stage processing. Typically, deep learning is used to extract person *keypoints*, and a model-fitting/optimization stage leads to the final keypoint estimation of pose and shape. As a consequence, there is no end-to-end architecture that would allow the computation of an image gradient needed for adversarial perturbations.

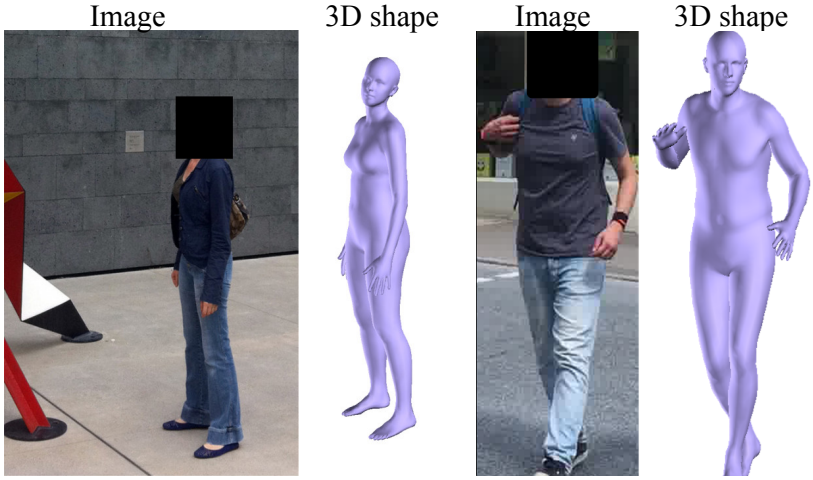
Today, we are missing successful evasion attacks on shape extraction methods. In this chapter, we investigate to what extent shape extraction can be avoided by small manipulations of the input image. We follow the literature on adversarial perturbations and require our changes in the input image to be of a small Euclidean norm. After analyzing a range of synthetic attack strategies that operate at the keypoints level, we experimentally evaluate their effectiveness to throw off multi-stage methods that include a model fitting stage. These attacks turn out to be highly effective while leaving the images visually unchanged. In summary, our contributions are:

- An orientative user study of concerns w.r.t. privacy and body shape estimation in different application contexts.
- An evasion attack to prevent privacy violations via body shape estimation.
- Analysis of synthetic attacks on 2D keypoint detections.
- Evaluation of practicability and effectiveness of the real attack on keypoints.
- A new localized attack on keypoint feature maps that require smaller noise norm for the same effectiveness.
- Evaluation of overall effectiveness of different attacks strategies on shape estimation. We show the first successful attacks that offer an increase in privacy with negligible loss in visual quality.

7.2 UNDERSTANDING USERS SHAPE PRIVACY PREFERENCES

Modern body shape methods (Sattar *et al.*, 2019c; Bogo *et al.*, 2016; Omran *et al.*, 2018; Kanazawa *et al.*, 2018) infer a realistic looking 3D body shape from a single photo of a person. The estimated 3D body captures user weight group and body shape type. However, such a realistic depiction of the naked body is considered highly private and therefore might not be consented by most people. We performed a user study to explore the users' personal privacy preferences related to their body shape data. Our goal was to study the degree to which various users are sensitive to sharing their shape data such as height, different body part measurement, and their 3D body shape in different contexts. This study was approved by our university's ethical review board and is described next.

User Study. We split the survey into three parts. In the first part of the survey, our goal was to understand users image sharing preferences and the user's knowledge of what type of information could be extracted from a single image.



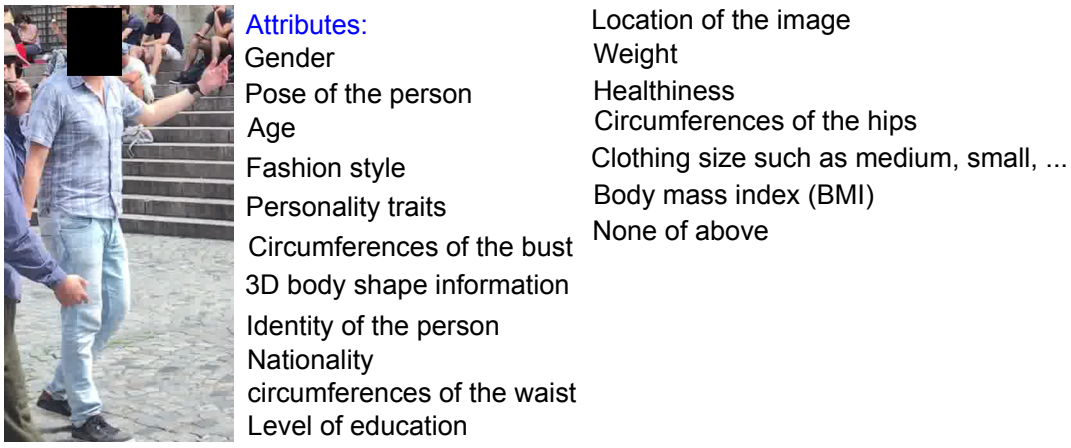


Figure 7.2: In Question 2 participant was shown this image and was asked to select attributes that could be extracted from this list. Furthermore, we asked our participant to indicate which of the listed attributes could be extracted from online purchase history.

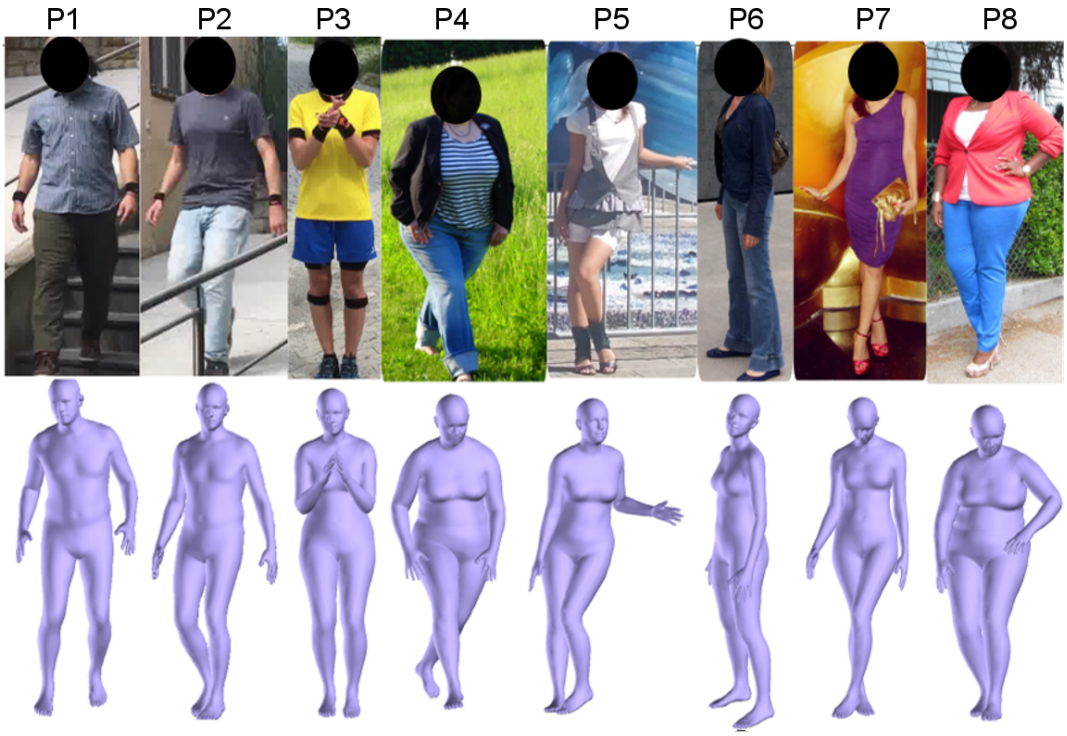


Figure 7.3: Participants were asked to judge the closeness of the depicted 3D shape to the actual body of the person in the images.

Question 1: Users are shown section 7.2 without the 3D shape data. Participants are asked how comfortable they are sharing such images publicly, considering they are the subject in these images. Responses are collected on a scale of 1 to 5, where: (1) Extremely comfortable, (2) Slightly comfortable, (3) Somewhat comfortable, (4) Not comfortable, and (5) Extremely uncomfortable.

Question 2.1: Participant were shown Figure 7.2, and were asked which of the attributes could be extracted from this image?

Question 2.2: In addition, we also ask our participant to indicate what type of information could be extracted from online purchase history? (Pose was removed from the list as it was not applicable to this question.)

In the second part, users were introduced to 3D shape models by showing them images of 8 people along with their 3D body shape, as shown in Figure 7.3. The purpose of part 2 was to understand the user's perceived closeness of extracted 3D shapes to the original images, and their level of comfort with them.

Question3: Participants were asked to rate how close the estimated 3D shape is to the person in the image. Responses are collected on a scale of 1 to 5, where: (1) Untrue of the person in the image, (2) somewhat untrue of the person in the image, (3) Neutral, (4) Somewhat true of the person in the image, and (5) True of the person in the image.

Question 4: Participants were shown section 7.2 asked to indicate how comfortable they are sharing such a photograph along with 3D shape data publicly, considering they are the subject in these images. We collected responses on a scale of 1 to 5, similar to Question 1.

In the third part of this survey, we explore users preferences on what type of body shape information they would share for applications such as (a) Health insurance, (b) Body scanners at airport, (c) Online shopping platforms, (d) Dating platforms, and (e) Shape tracking applications (for sport, fitness, ...).

Question 5: Users were asked their level of comfort on a scale of 1 to 5 as in *Question 1* for the applications mentioned above.

Participants. We collect responses of 90 unique users in this survey. Participants were not paid to take part in this survey. Out of the 90 respondents, 43.3% were female, 55.6% were male, and we have 1.1% as queer. The dominant age range of our participants (63.3%) was in 21-39, followed by 30-39 (23.3%). Participants have a wide range of education level, where 46.7% has a master degree, 21.1% has a bachelor degree. Further details on participants demographic data are presented in the Table 7.1

Analysis. The results of *Question 1* and *Question 4* are shown in Figure 7.6a. We see that majority of the users do not feel comfortable or they feel extremely uncomfortable(36%, 30%) sharing their 3D data publicly compared to sharing only their images (29%, 14%). More detailed results for each gender is shown in Figure 7.7

In *Question 2* top three selected attributes was: Gender (98.9%), pose (87.8%), and age (85.6%). Whereas shape related attributes such as body mass index (BMI)

Gender	Female	Male	Other
	39	50	1
Age			
18 - 20	2	2	0
21 - 29	22	34	1
30 - 39	10	11	0
40 - 49	3	2	0
50 - 59	1	0	0
60 or older	1	1	0
Highest Completed Education			
High school degree or equivalent (e.g., GED)	3	4	0
Some college but no degree	1	0	0
Bachelor's degree	7	11	1
Master's degree	20	22	0
Graduate degree	1	3	0
Professional degree	2	0	0
Doctorate degree	5	10	0

Table 7.1: **Participants demographic. Total N = 90**. Majority of our participant were in age range 21-29 with Masters degree.

(47.8%), weight (63.3%), and 3D body shape (66.7%) were not in the top selected attributes. Indicating that the majority of the participants did not consider such information could be extracted from an image and lack awareness of shape extraction techniques. The full ranking of all attributes are shown in Figure 7.4.

In *Question3*, users were asked to judge the quality of the presented 3D models. Around 43% of the participants believe the presented shape is Somewhat true of the person in the image, and 31% thinks the 3D mesh is true to the person in the picture. This indicates that recent approaches can infer perceptually faithful 3D body shapes under clothing from a single image. The detailed responses are shown in Figure 7.5

Figure 7.6b presents the results from *Question 5*. Participants show a high level of discomfort in sharing their 3D shape data for multiple applications. In all investigated applications except for fitness, the majority of the users responded with "discomfort of some degree".

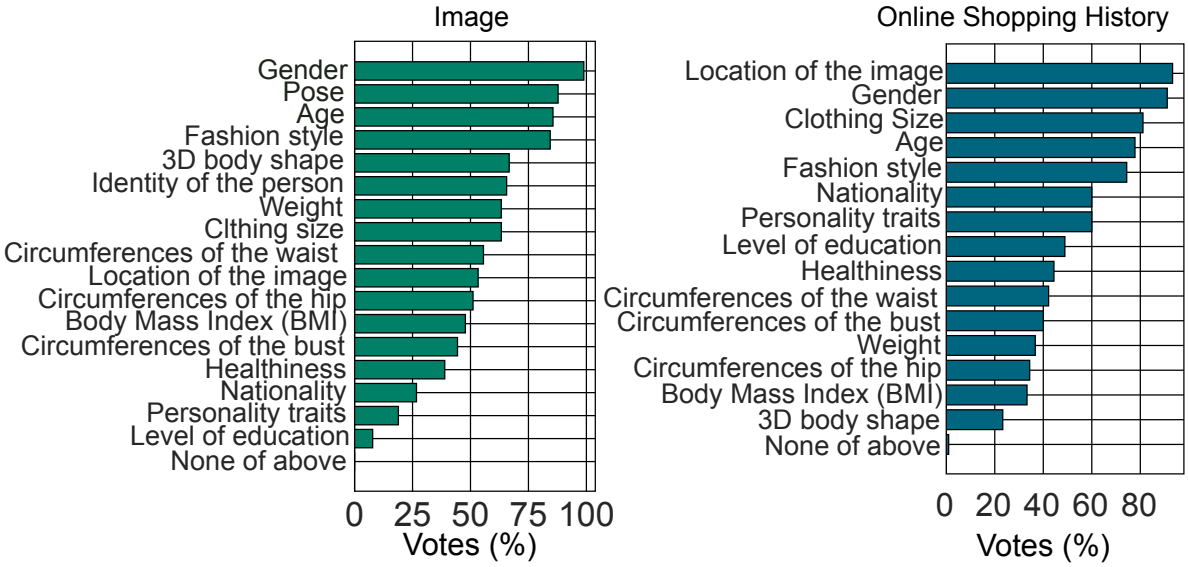


Figure 7.4: List of attributes selected by our participants in ***Question 2.1,*** and ***Question 2.2***

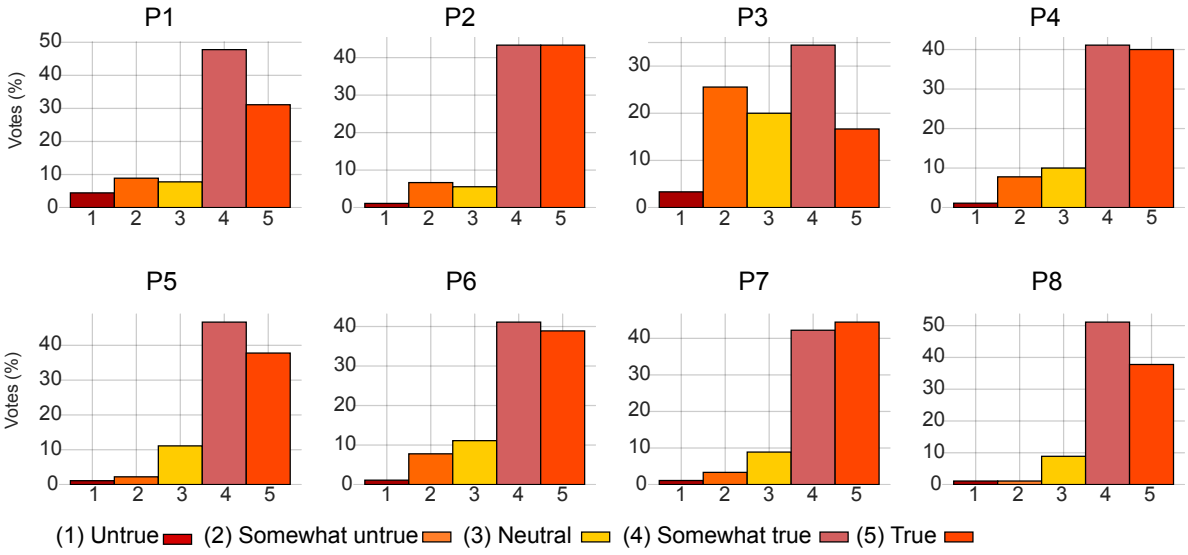


Figure 7.5: Participant were asked to rate how well the depicted 3D shape reflects the person in the image. Somewhat true and true was selected by majority of participants.

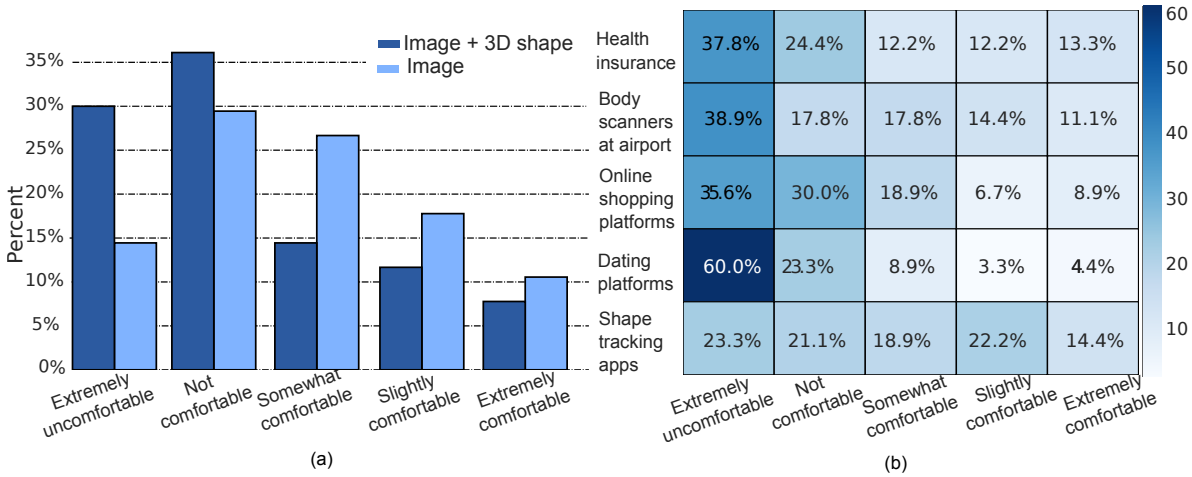


Figure 7.6: (a) Comfort level of the participant for sharing an image of a person publicly, considering they are the subject in these images with and without 3D mesh data. (b) Comfort level of the participant for sharing their 3D mesh data with multiple applications. Values are reported in percentage of a time an answer is chosen.

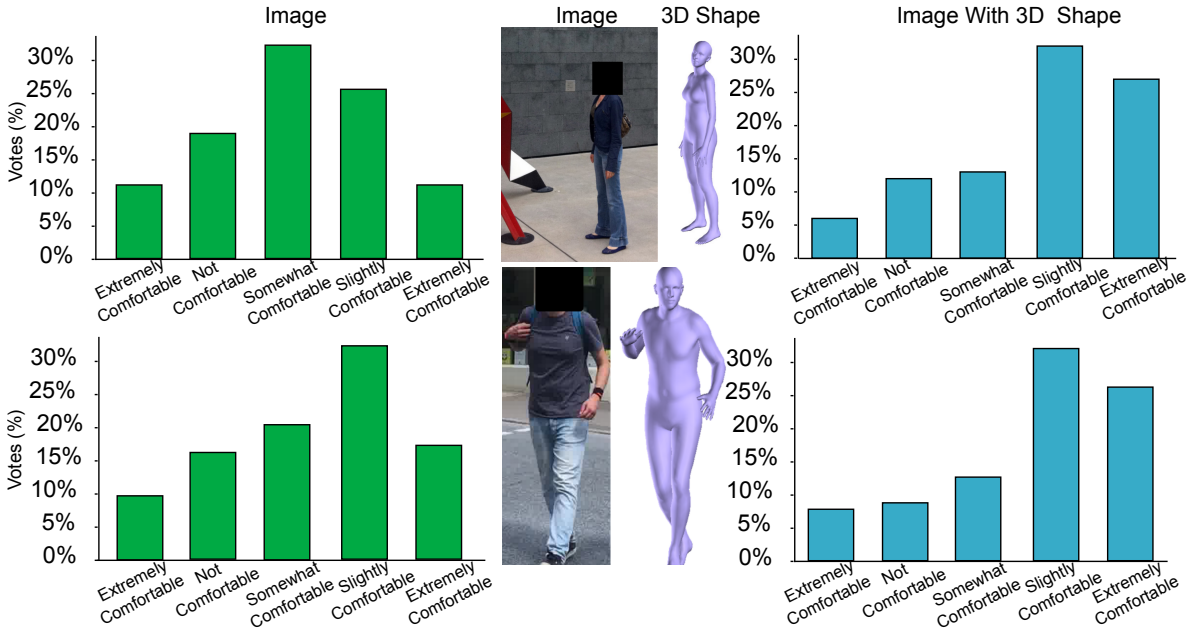


Figure 7.7: Participants were shown the image with and without 3D shape data and were asked to indicate their comfort level for sharing this data publicly. One can see that the majority of participant have a high level of discomfort in sharing their 3D shape data publicly. This can be seen especially for the female subject were the comfort distribution towed towards not comfortable when having 3D shape data along.

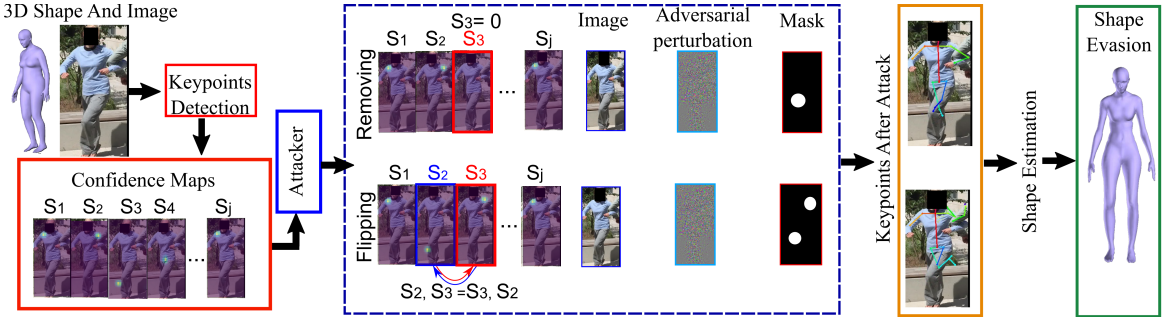


Figure 7.8: The summary of our framework. We assume that we have full access to the parameter of the network. The attacker breaks the detections by removing or flipping of a keypoint. Hence the final estimated shape does not depict the person in the image.

7.3 SHAPE EVASION FRAMEWORK

Model-based shape estimation methods from a 2D image are based on a two-stage approach. First, a neural network is used to detect a set of 2D body keypoints, and then a 3D body model fits the detected keypoints. Since this approach is not end-to-end, it does not allows direct computation of the image gradient needed for adversarial perturbation. To this end, we approach the shape evasion by attacking the keypoints detection network. In section 7.3.1, we give a brief introduction on model-based shape estimation method. In section 7.3.2, we introduced a local attack that allows targeted attacks on keypoints. Figure 7.8 shows an overview of our approach.

7.3.1 Model Based Shape Estimation

A Skinned Multi-Person Linear Model (SMPL) (Loper *et al.*, 2015) is a state of the art generative body model. The SMPL function $M(\boldsymbol{\beta}, \boldsymbol{\theta})$, uses shape $\boldsymbol{\beta}$ and poses $\boldsymbol{\theta}$ to parametrize the surface of the human body that is represented using $N = 6890$ vertices. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ encode changes in height, weight and body proportions. The body pose $\boldsymbol{\theta} \in \mathbb{R}^{3P}$, is defined by a skeleton rig with $P = 24$ keypoints. The 3D skeleton keypoints are predicted from body shape via $J(\boldsymbol{\beta})$. We can use a global rigid transform R_θ To pose the SMPL keypoint. Hence, $R_\theta(J(\boldsymbol{\beta})_i)$ denotes a posed 3D keypoint i. In order to estimated 3D body shape from a 2D image I, several works (Sattar *et al.*, 2019c; Bogo *et al.*, 2016; Lassner *et al.*, 2017b), minimizes an objective function composed of a keypoint-based data term, pose priors, and a shape prior.

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_{P_\theta}(\boldsymbol{\theta}) + E_{P_\beta}(\boldsymbol{\beta}) + E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{J}_{\text{est}}) \tag{7.1}$$

where $E_{P_\theta}(\boldsymbol{\theta})$, and $E_{P_\beta}(\boldsymbol{\beta})$ are the pose and shape prior terms as described in Bogo *et al.* (2016). The $E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{J}_{\text{est}})$ is the keypoint-based data term which penalizes the

weighted 2D distance between estimated 2D keypoints, $\mathbf{J}_{\text{est}}$, and the projected SMPL body keypoint $R_\theta(J(\boldsymbol{\beta}))$:

$$E_J(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}; \mathbf{J}_{\text{est}}) = \sum_{\text{keypoint}i} w_i \rho(\Pi_{\mathbf{K}}(R_\theta(J(\boldsymbol{\beta})_i)) - \mathbf{J}_{\text{est},i}) \tag{7.2}$$

where $\Pi_{\mathbf{K}}$ is the projection from 3D to 2D of the camera with parameters $\mathbf{K}$ and ρ a Geman-McClure penalty function which is robust to noise in the 2D keypoints detections. w_i indicates the confidence of each keypoints estimate, provided by 2D detection method. For cases such as occluded or missing keypoints, w is very low, and hence the data term will be driven by pose prior term. Furthermore, the prior term avoids impossible poses. Shape evasion can be achieved by introducing error in 2D keypoints detection $\mathbf{J}_{est}$. We use Adversarial perturbation to fool the pose detection method by either removing a keypoint or filliping two keypoints with each other.

7.3.2 Adversarial Image Generation

The state of the art 2D pose detection methods uses a neural network f parametrized by ϕ, to predict a set of 2D locations of anatomical keypoints $\mathbf{J}_{est}$ for each person in the image. The network produces a set of 2D confidence maps $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, ..., \mathbf{S}_P\}$, where $\mathbf{S}_i \in \mathbb{R}^{w \times h}$, $i \in 1, 2, 3, ..., P$, is a confidence map for the keypoints i and P is total number of Keypoints. Assuming that a single person is in the image, then each confidence map contains a single peak if the corresponding part is visible. The final set of 2D keypoints $\mathbf{J}_{est}$ are achieved by performing non-maximum suppression per each confidence map. These confidence maps are shown in Figure 7.8.

To attack a keypoint we used adversarial perturbation. Adding adversarial perturbation $\mathbf{a}$ to an image I will causes a neural network to change its predictionSzegedy *et al.* (2014).The adversarial perturbation $\mathbf{a}$ is defined as the solution to the optimization problem

$$\arg\min_{\mathbf{a}} \ \|\mathbf{a}\|_2 + L(f(I + \mathbf{a}; \phi), \mathbf{S}^*). \tag{7.3}$$

L is the loss function between the network output and desired confidence maps $\mathbf{S}^*$.

Removing and Flipping of Keypoints:. The $\mathbf{S}^*$ is defined for removing and flipping of keypoints. to remove a keypoint, we put its confidence map to zero. For example if we are attacking the first keypoint we have: $\mathbf{S}^* = \{\mathbf{S}_1 = 0, \mathbf{S}_2, \mathbf{S}_3, ..., \mathbf{S}_P\}$. To flip two key points we exchanged the values of two confidence map as $\mathbf{S}^i, \mathbf{S}^j = \mathbf{S}^j, \mathbf{S}^i$. In case $i, j = 2, 3$ we have $\mathbf{S}^* = \{\mathbf{S}_1, \mathbf{S}_3, \mathbf{S}_2, ..., \mathbf{S}_P\}$. An example of removing and flipping of the keypoint is shown in Figure 7.8.

Fast Gradient Sign Method (FGSM) Goodfellow *et al.* (2014b):. FGSM is a first order optimization schemes used in practice for Equation 7.3, which approximately minimizes the ℓ_∞ norm of perturbations bounded by the parameter ϵ. The adversarial

examples are produced by increasing the loss of the network on the input I as

$$I^{adv} = I + \epsilon \, sign(\nabla_I L(f(I;\phi), \mathbf{S}^*)). \tag{7.4}$$

We call this type of attack global as the perturbation is applied to the whole image. This perturbation results in poses with several missing keypoints or poses outside of natural human pose manifold. While this will often make the subsequent shape optimization step fail (Eq. eq:single), the approach has two limitations: i) this attack requires a large perturbation and ii) the attack is very easy to identify by the defender.

Masked Fast Gradient Sign Method (MFGSM):. To overcome the limitations of the global approach, we introduced Masked FGSM. This allows for localized perturbation for more targeted attacks. This method will generate poses, which are close to ground truth pose, yet have a missing keypoint that will cause shape evasion. We will refer to this scheme as "local" in the rest of this chapter. To attack a specific keypoint we solve the following optimization problem in a iterative manner as:

$$\mathbf{I}_0^{adv} = I$$

$$I_{t+1}^{adv} = clip(I_t^{adv} - \alpha \cdot sign(\nabla_{I_t^{adv}} L(f(I_t^{adv};\phi), \mathbf{S}^*) \odot M), \epsilon)$$

(7.5)where mask $\mathbf{M} \in \mathbb{R}^{w \times h}$ is used to attack a keypoint $\mathbf{J}_{est,i} \in \mathbb{R}^2$ selectively. $\mathbf{M}$ is defined as:

$$\mathbf{M} = \begin{cases} 1 & if(x - \mathbf{J}_{est,i})^2 = r^2 \\ 0 & \end{cases} \tag{7.6}$$

r controls the spread of the attack and $\mathbf{x} \in \mathbb{R}^2$ are the pixel coordinates. To ensure the max norm constraint of perturbation $\mathbf{a}$ being no greater than ϵ is preserved, the $clip(z, \epsilon)$ is used, which keeps the values of z in the range$[z - \epsilon, z + \epsilon]$.

7.4 EXPERIMENTS

The overall goal of the experimental section is to provide an understanding and the first practical approach to evade body shape estimation and hence protect the privacy of the user. We approach this by systematically studying the effect of attacking keypoint detections on the overall shape estimation pipeline. First, we study synthetic attacks based on direct manipulation of keypoints locations, where we can observe the effects on body shape estimation in an idealized scenario. This study is complemented by real image-based attacks which make keypoint estimation fail. Together, we evaluate our approach that provides the first and effective defence against body shape estimation on real-world data.

Dataset.. We used 3D Poses in the Wild Dataset (3DPW) (von Marcard *et al.*, 2018), which includes 60 sequences with 7 actors. To achieve ground truth shape parameter β, actors were scanned and SMPL was non-rigidly fit to them to obtain their 3D

Attack	Right ankle	Right knee	Right hip	Left hip	Left knee	Left ankle	Right wrist
Real	**1.32**	**1.4**	1.39	1.37	**1.38**	**1.32**	**1.36**
Synthetic	1.17	1.18	**1.79**	**1.94**	1.18	1.18	1.18
Attack	Right elbow	Right shoulder	Left shoulder	Left elbow	Left wrist	Head top	Average
Real	**1.41**	1.40	1.35	**1.28**	**1.37**	**1.35**	**1.37**
Synthetic	1.17	**1.43**	**1.49**	1.15	1.16	1.19	1.32

Table 7.2: Shape estimation error on 3DPW with Procrustes analysis with respect to the ground truth shape. Error in cm. The goal of each attack is to induce a bigger error in the estimated shape. Hence, higher errors are an indication of a successful attack. The average is calculated over all keypoints for each of the real and synthetic attacks.

models similar to (Pons-Moll *et al.*, 2017; Zhang *et al.*, 2017b). To the best of our knowledge, 3DPW is the only in wild image dataset which provides the ground truth shape data as well, which makes this dataset most suitable for our evaluation. For our evaluation, for each actor, we randomly selected multiple frames from different sequences. All reported results are averaged across subjects and sampled sequence frames.

Model.. We used OpenPose (Cao *et al.*, 2017) for keypoint detection as it is the most widely used. OpenPose consists of a two-branch multi-stage CNN, which process images at multi-scales. Each stage in the first branch predicts the confidence map S, and each stage in the second branch predicts the Part Affinity Fields (PAFs). For the shape estimation, we used the public code of Smplify (Bogo *et al.*, 2016), which infers a 3D mesh by fitting the SMPL body model to a set of 2D keypoints. To improve the 3D accuracy, we refined the estimations using silhouette as described in (Sattar *et al.*, 2019c). We used MFGSM (Eq. eq:iterative with $\alpha = 1$) in an iterative manner. We evaluated attacks when setting the ℓ_∞ norm of the perturbations to $\epsilon = 0.035$ since we observed that higher values lead to noticeable artifacts in the image. We stop the iterations if we reach an Euclidean distance (between the original and perturbed images) of 0.02 in image space for local, and 0.04 for global attacks.

7.4.1 Synthetic Modification of The keypoints

First, we studied the importance of each keypoint on the overall body shape estimation by removing one keypoint at a time–which simulate miss-detections. The error on shape estimation caused by this attack is reported in the second row of Table 7.2. We observe that removing“Hips”, and “Shoulder” keypoints results in the highest increase of error of 34%, and 25.86% whereas "Elbows" and "Wrists" result in an increase of only 1%.

We also studied the effect of flipping keypoints. The results of this experiment are shown in Figure 7.9. Flipping the “Head” with the left or right“Hip” caused

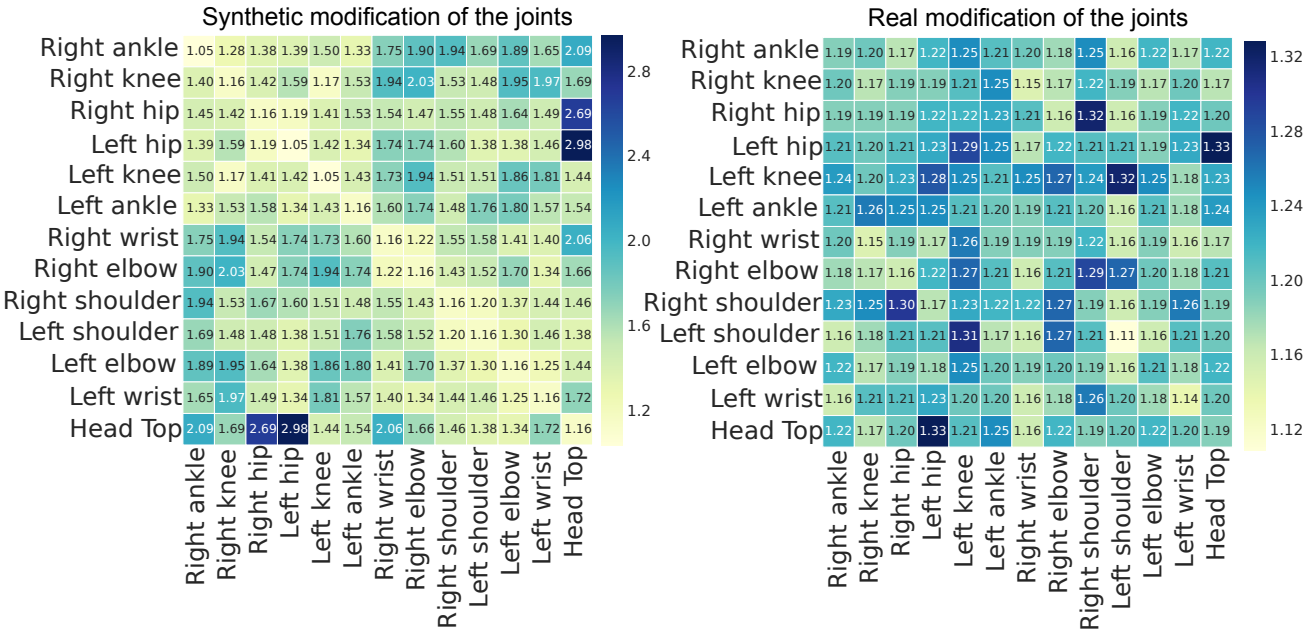


Figure 7.9: Shape estimation error on 3DPW with Procrustes analysis. Error in cm for synthetic and real flipping of the keypoints.

an increase in error of 143.96%. Flipping the " Elbow" and "Knee" was the second most effective attack causing 67% increase of error in average. The least effective attack was by flipping the left and right knee (2.58%). The average error introduced by removing or flipping of each keypoint is illustrated in Figure 7.10 – higher error is larger in size and darker in colour. We can see that, overall "Hip", "Shoulder", and "Head" keypoints play a crucial role in the quality of the final estimated 3D mesh, and are the most powerful attacks.

7.4.2 Attacking keypoint Detection by Adversarial Image Perturbation

To apply modifications to the keypoints, we used our proposed local Mask Iterative Fast Gradient Sign Method (MIFGSM). Figure 7.11 shows the keypoint confidence map values when removing and adding a keypoint using local and global attacks with respect to the amount of perturbation added to the image. We can see that the activation's per each keypoint decreases after each iteration. Interestingly, the rate of decrease is slower for global attacks for the same amount of perturbation (0.015 Mean Squared Error (MSE) between perturbed and original image). Global attacks require much higher amount of perturbations (0.035 MSE) to be successful, causing visible artifact in the image. We observed similar behavior when adding a "fake" keypoint detections (required to flip two keypoints). Similarly, the rate of increase in activation was slower for global compared to local for the same amount of perturbation (0.015 MSE, blue bar in the plot). From Figure 7.11 we can also see that shoulders and head are more resistant to the removal. Furthermore, the

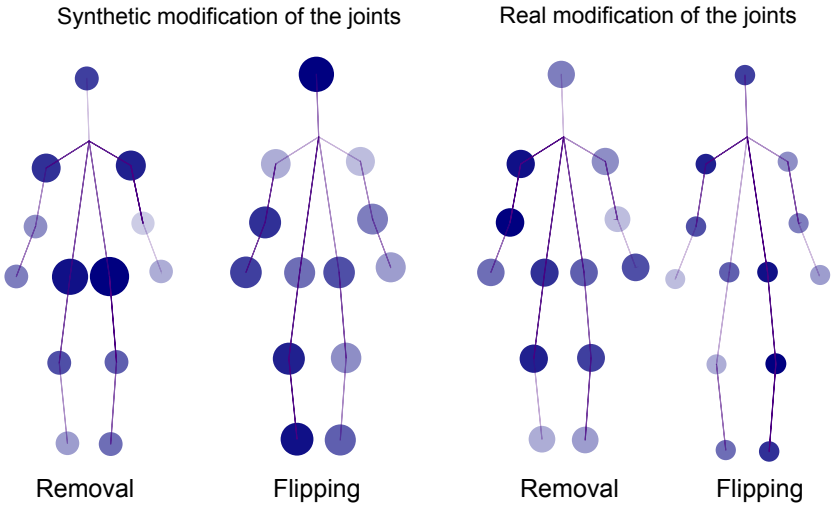


Figure 7.10: The overall shape estimation error induced by synthetic and real (local) attacks. The darker and bigger circles shows higher error.

attack was the most successful in the creation of wrists. Since local attacks are more effective, we consider only the local attack method for further analysis.

7.4.3 Shape Evasion

In this section, we evaluate the effectiveness of the whole approach for evading shape estimation and therefore, protecting the users' privacy. We used our proposed local method to remove and flip keypoints instead of the synthetic modification of the keypoints as described in section 7.4.1. Hence, we call this attack as a real modification of keypoints.

The error on shape estimation caused by removing of the keypoints using our local method are reported in the first row of Table 7.2, we refer to it as real. We see that attacks on "Right Elbow" and "Right knee" causes 21% increase of error in shape estimation. The least amount of error 10% and 13% was produced by removing "Left Elbow" and "Ankles" respectively. However, "Hip" and "Shoulder" gained higher error in average for left and right keypoint by 18%. On average, the real attack for removing keypoints caused an even higher error than the synthetic mode (18% to 13%), showing the effectiveness of this approach in shape evasion and hence protecting the users' privacy.

The result for flipping the keypoints is shown in Figure 7.9 (Real modification of the keypoint). The highest increase in error was (14%) caused by flipping the "Head" with "Left Hip", the second most effective attack was for flipping the "Shoulder" and "Knee" keypoints (12% in average over left and right keypoints). The least effective attack was on "wrist" with increase of 2% error on average.

Real flipping of keypoints achieves an error of (3%) compared to real removing attacks (18%), which shows they are slightly less effective. In addition, similar to global attacks, filliping of keypoints causes more changes in the keypoints, making

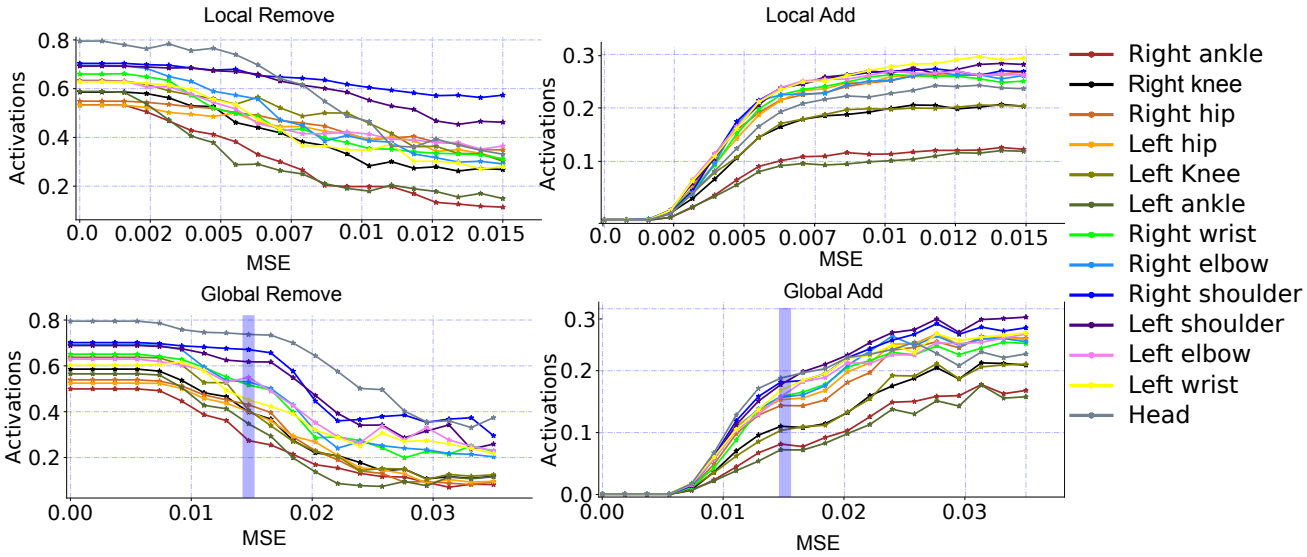


Figure 7.11: Comparison of local and global attacks for removing and adding a keypoint. The local attack has a higher rate of decrease or increase of activation compared to the global method for the same amount of perturbation. The blue bar on the global plots shows where the local methods end.

the detection of these attacks easier.

7.4.4 Qualitative Results

In Figure 7.12, we present example results obtained for each type of attack. The global attack causes pose estimation to hallucinate multiple people in the image, destroying the body signal of the person in the picture. As the predicted poses in the global attack are not in human body manifold, the optimization step in SMPL will fail to fit these keypoints resulting in average shape estimates. In the local attack, we were able to apply small changes in the keypoints. Hence, these small changes make the shape optimization stage predict shapes that are not average and also not close to the person in the image. Overall, shape evasion was most successful when removing the keypoints than flipping them, and when using the local attacks.

7.5 DISCUSSION

As our study of privacy on automatically extracted body shapes and method for evading shape estimation is the first of its kind, it serves as a starting point – but naturally needs further investigations to extend on both lines of research that we have touched on. The following presents a selection of open research questions.

Targeted vs untargeted shape evasion.. While our method for influencing the key point detection is targeted, the overall method to shape evasion remains untargeted.

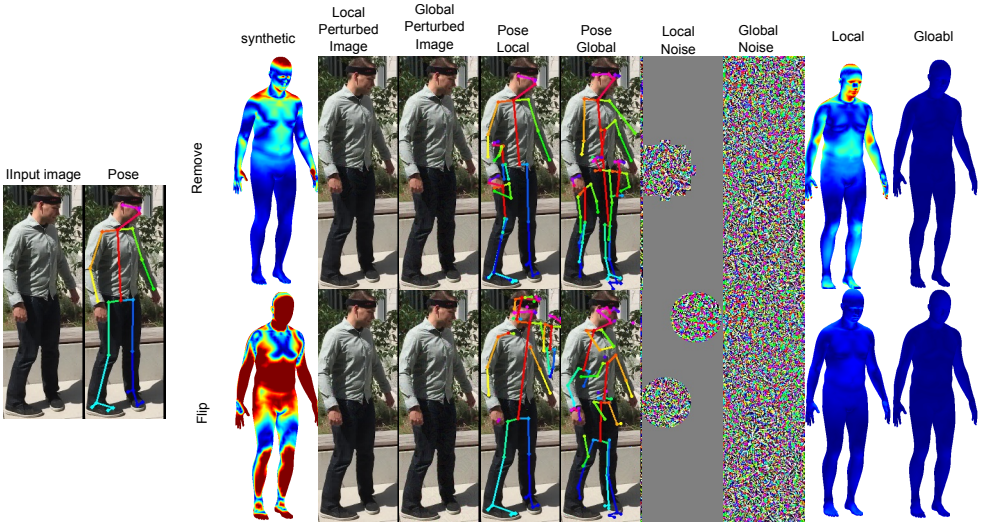
Depending on the application scenario, a consistent change or particular randomization of the change in shape might be desired, which is not addressed by our work.

Effects of adversarial training.. It is well known that adversarial training against particular image perturbations can lead to some robustness against such attacks (Szegedy *et al.*, 2014; Oh *et al.*, 2017) and in turn, the attack can again be made to some extent robust against such defenses. Preventing this cat-mouse-game is subject of on-going research and – while very important – we consider outside of the scope of our first demonstration of shape evasion methods.

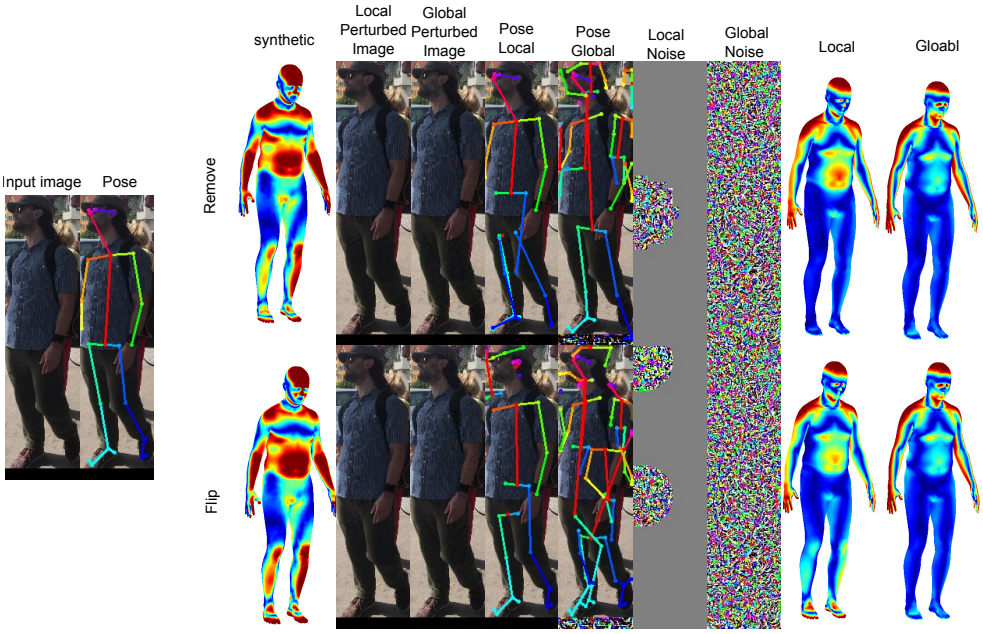
Scope of the user study.. While our user study encompasses important aspects of privacy of body shape information, clearly a more detailed understanding can be helpful in order to inform the design evasion techniques and privacy-preserving methodologies that comply with the users' expectations on handling personal data. As our study shows that such privacy preferences are personal as well as application domain specific, there seem ample opportunities to leverage the emerging methods of high-quality body shape estimation in compliance with user privacy.

7.6 CONCLUSION

Methods for body shape estimation from images of clothed people are getting more and more accurate. Hence we have asked the timely question to what extent this raises privacy concerns and if there are ways to evade shape estimation from images. In order to better understand the privacy concerns, we conduct a user study that sheds light on the privacy implication as well as the sensitivity of shape data in different application scenarios. Overall, we observe a high sensitivity which is also dependent on the use case of the data. Based on this understanding, we follow up with a defense mechanism that can hamper or even prevent body shape estimation from real-world images. Today's state of the art body shape estimation approaches are frequently optimization based and therefore don't lend themselves to gradient-based adversarial perturbation. We tackle this problem by a two-stage approach that first analysis the effect of individual keypoints on the shape estimate and then proposes adversarial image perturbations in order to influence the keypoints. In particular, our novel localized perturbation techniques constitute an effective technique to evade body shape estimation at negligible changes to the original image.



Person with body shape close to SMPL template (0.04 cm)



Person with a higher distance to SMPL template (2 cm)

Figure 7.12: The left side shows the original image with the estimated pose, and the right the output when modified with local and global adversarial perturbations with corresponding error heatmaps with respect to ground truth shapes (red means $>$ 2cm). Here we applied local and global attack for removing the "Right Hip", and flipping the " Right Hip" and 'Head Top". The global attack causes the pose estimation to hallucinate multiple people in the image, while or local attack only changes the selected keypoints. The predicted shape in case of global attack is always close to the average template of SMPL causing a lower error for people with an average shape.

8 CONCLUSIONS AND FUTURE PROSPECTS

User intent and preferences prediction is one of the main building blocks of a symbiotic system. Such a 'user-centred' system needs to be able to predict users' need and assist them in achieving their goal. There exist several approaches in web search (Joachims *et al.*, 2005; Granka *et al.*, 2004; Miller and Agne, 2005; McNamara *et al.*, 2019; Sun *et al.*, 2019), which aimed to predict the user intent to improve the search experience and provide users with personalised recommendations. All these applications are limited to web-based interfaces and cannot be deployed in a real-world environment such as a real shopping mall. In this thesis, we focus on predicting the user intent and preferences using shape and gaze. Our gaze-based visual search prediction is not limited to the web-based application and can be deployed in real-world scenarios using wearable eye trackers. Our method does not need any input from the user, such as a sketch or keywords. Furthermore, using data from online fashion blog users, we show for the first time that clothing and shape are correlated. These results indicate that even though the size of the garment may be correct, yet clothing item may not be an excellent fit to specific body shape. In this chapter, we further discuss the contributions of this thesis (Section 8.1) and review open problems as well as potential prospects (Section 8.1).

8.1 DISCUSSION OF CONTRIBUTIONS

The overall goal of this thesis is to enable intent and preference prediction using implicit data. We tackled two specific research topics: *gaze based inferences* and *shape based inferences*. In the following, we will discuss the main findings and insights of this thesis concerning the individual chapters.

8.1.1 Search Target Prediction Using Gaze

In Chapter 3, we demonstrated how to predict the search target during the visual search from human fixations in an open-world setting. This setting is fundamentally different from works investigated in prior work, as we no longer assume that we have fixation data to train for these targets. To address this challenge, we presented a new approach that is based on learning compatibilities between fixations and potential targets. We showed that this formulation is effective for search target prediction from human fixations.

In Chapter 4, we proposed the first method to predict the category and attributes

of visual search targets from human gaze data. To this end, we introduced a novel Gaze Pooling Layer that allows us to seamlessly integrate semantic and localised fixation information into deep image representations. Our model does not require gaze information at training time, which makes it practical and easy to deploy.

In Chapter 5, we introduce the first approach to decode visual search target of users from their gaze data. This task is very challenging as the target only resides in user mind. For this aim, we used recent advances in generative image models and our introduced gaze embedding technique as described in Chapter 4. We introduced a gaze conditioned auto-encoder to generate mental images of the user during visual search. To evaluate our method, we performed two user studies. The first user study shows the decoded target lead to human recognisable visual representations. The second user study highlights the importance of localised gaze information. We like to emphasise that due to the training setup, the method remains highly practical and applicable, as no large scale gaze data had to be collected or used. Key is instead the utilisation of a semantic layer that connects the gaze encoder with the conditional generative image model.

We believe that the ease of preparation and compatibility of our Gaze Pooling Layer with existing models will stimulate further research on gaze-supported computer vision, particularly methods using deep learning.

8.1.2 Shape Based Clothing Preference Prediction

In Chapter 6, we aimed to understand the connection between body shapes and clothing preferences by collecting and analysing an extensive database of fashion photographs with annotated clothing categories. Our results demonstrate that clothing preferences and body shapes are correlated and that we can build predictive models for clothing categories based on the output of automatic shape estimation. To obtain estimates of 3D shape, we proposed a new approach that incorporates evidence from multiple photographs and body segmentation and is more accurate than popular recent SMPLify approach (Bogo *et al.*, 2016). However, as methods for body shape estimation from images of clothed people are getting more accurate, we have asked the timely question to what extent this raises privacy concerns and if there are ways to evade shape estimation from images. In Chapter 7, we conduct a user study that sheds light on the privacy implication as well as the sensitivity of shape data in different application scenarios. Overall, we observe a high sensitivity, which is also dependent on the use case of the data. Based on this understanding, we follow up with a defence mechanism that can hamper or even prevent body shape estimation from real-world images. Today's state of the art body shape estimation approaches are frequently optimisation based and therefore don't lend themselves to gradient-based adversarial perturbation. We tackle this problem by a two-stage approach that first analysis the effect of individual body keypoints on the shape estimate and then proposes adversarial image perturbations to influence the key points. In particular, our novel localised perturbation techniques constitute an effective technique to evade body shape estimation at negligible changes to the

original image.

These findings will open up new research directions and applications areas in outfit recommendations, and raises awareness about the privacy aspects of the shape.

8.2 FUTURE PROSPECTS

In this section, we discuss several challenges that remain for enabling search target and preference prediction in real-world settings, and possible solutions to address them.

Search Target Prediction in Real World Shopping. Although our recent search target prediction, cover variations in gaze behaviour of users and can generalises over users, our experiments have been done in a laboratory environment using natural image collages. Thus, the implementation of our approach in the real world scenario is still missing. For example, to predict the search target of a user in real-world shopping scenario, we need to handle objects occlusions by either other objects or people, object clutter, motion blur in videos, and depth ambiguity on fixation locations. Another challenge in the real world is the lack of annotated shopping mall data set along with gaze data. Furthermore, during shopping, people tend to have exploratory search behaviour rather than targeted search, which makes search target prediction more challenging. Hence, understanding user gaze behaviour in real-world shopping scenario, and building a model which could predict users interest and search target is still a challenging task and need further work. As a first step, a data collection in a small mock-up shop environment could simplify several challenges mentioned above. The other possibility is using a virtual environment setting to build a virtual shop.

Search Target Prediction in VR. Virtual reality offers a lab environment with a high immersion and close alignment with reality. In VR, we can have full control over the environment, which allows for a more in-depth study of user behaviour. To find out where a person is looking in a real world, we need to have 3D gaze vector. One can calculate the 3D gaze vector by using a 3D eye model to detect pupil (Khan *et al.*, 2019; Nagamatsu *et al.*, 2009). In this case depth can be calculated from the crossing point of the gaze from both eyes. However, this calculation is imprecise and provides acceptable results in case of a perfect calibration (Jansen *et al.*, 2009). In contrast, in a VR environment, as we know the exact distance between the eyes and objects, depth can be calculated precisely using a 3D eye model for pupil detection (Clay *et al.*, 2019). Hence, analysing user behaviour during shopping and building model to predict user intent is more feasible in VR Speicher *et al.* (2017). This direction is auspicious as several companies such as Amazon[10], are investing in

[10]https://venturebeat.com/2018/07/12/watch-amazons-vr-kiosks-transform-the-future-of-shopping/

VR and AR shopping [11]. For example, the environment could be updated based on the predicted user intent.

Combining Additional Implicit Cues For Preference Prediction. Recent clothing outfit recommendations only use clothing features for recommendations. There exist only very few works that take into account location (Zhang *et al.*, 2017d) or location and skin tone (Garude *et al.*, 2019) to recommend outfit. However, outfit recommendation models should be more human-centered than clothing-center. For example, during shopping, a person usually tries on different outfits and chooses the one that fits them the best. The decision is often makes based on if the colour fits their skin tone or hair colour, or if the clothing item is an excellent match to their body shape. In Chapter 6, we studied the correlation between a person body and the clothing item. Hence, a system that takes into account more body features is still missing. Such a model need to take into account shape, skin colour, hair colour, maybe even eye shape. However, due to lack of any data set that covers all this metadata, this task is challenging. One possible way is to use automatic detection methods (Oh *et al.*, 2015; Schroff *et al.*, 2015), to extract information about skin and hair colour from our Fashion Take Shape data set. Recent recommendations based on graph-auto encoder for compatibility predictions show state of the art results in this area (Kipf and Welling, 2016; Ying *et al.*, 2018; Cucurull *et al.*, 2019). Hence, one can use graph autoencoders to learn compatibility between clothing items and body cues.

Search Target and Preference Prediction in Real World. Combination of search target prediction with human-centered outfit recommendation is another interesting yet not studied application. There exist several similar attempts such as shop the look (Kiapour *et al.*, 2015) or street to shop (Liu *et al.*, 2012; Shankar *et al.*, 2017) recommendation systems[12]. In which a user upload a photo of a person and the recommendation systems would find the exact match and recommend a shop to buy the item. In this scenario, the person needs to make a photo of the things they like and upload it to the recommendation model. However, a system that could use eye tracking to detect people interest in the wild (like walking in the street) and give a recommendation based on a mixture of body cues and the interesting item would be more practical as it removes the need of the user to take photos.

[11] https://insights.samsung.com/2018/07/31/vr-in-retail-the-future-of-shopping-is-virtual-and-augmented/

[12] https://theblog.adobe.com/see-it-search-it-shop-it-how-ai-is-powering-visual-search/

PUBLICATIONS

[6] *Deep Gaze Pooling: Inferring and Reconstructing Search Intent From Human Fixations*
Hosnieh Sattar, Andreas Bulling, and Mario Fritz.
In IEEE Trans. on NeuroComputing, 2019.

[5] *Shape Evasion: Preventing Body Shape Inference of Multi-Stage Approaches.*
Hosnieh Sattar, Katharina Krombholz, Gerard Pons-Moll, and Mario Fritz.
In Technical Report, 2019, (arXiv: 1905.0450)

[4] *Fashion is Taking Shape: Understanding Clothing Preference Based on Body Shape From Online Sources*
Hosnieh Sattar, Gerard-Pons Moll, and Mario Fritz.
in Proc. IEE Conf Winter Conference on Applications of Computer Vision (**WACV**), 2019.

[3] *Visual Decoding of Targets During Visual Search From Human Eye Fixations*
Hosnieh Sattar, Mario Fritz, and Andreas Bulling.
In Technical Report, 2017, (arXiv: 1706.05993).

[2] *Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling*
Hosnieh Sattar, Andreas Bulling, and Mario Fritz.
In Proc. IEEE Int. Conf. on Computer Vision Workshops (**ICCVW**), 2017.

[1] *Prediction of Search Targets from Fixations in Open-world Settings.*
Hosnieh Sattar, Mario Fritz, and Andreas Bulling.
In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2015.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

K. G. A. Kovashka, D. Parikh (2012). WhittleSearch: Interactive Image Search with Relative Attribute Feedback, *International Journal of Computer Vision*, vol. 115(2). Cited on pages 11 and 15.

P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. T. Wu (2016). I-pic: A platform for privacy-compliant image capture. Cited on page 18.

G. Adomavicius and A. Tuzhilin (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge & Data Engineering*, (6), pp. 734–749. Cited on page 16.

K. Ak, A. Kassim, J. Hwee Lim, and J. Yew Tham (2018). Learning Attribute Representations With Localization for Flexible Fashion Search, in *CVPR 2018*. Cited on page 16.

Z. Al-Halah, R. Stiefelhagen, and K. Grauman (2017). Fashion Forward: Forecasting Visual Style in Fashion, in *ICCV 2017*. Cited on pages 16 and 68.

T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll (2019). Learning to Reconstruct People in Clothing from a Single RGB Camera, in *CVPR 2019*. Cited on page 86.

T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll (2018). Video Based Reconstruction of 3D People Models, in *CVPR Spotlight 2018*. Cited on pages 17, 69, and 86.

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis (2005). SCAPE: shape completion and animation of people, in *ACM Transactions on Graphics 2005*. Cited on page 17.

A. Arnab, O. Miksik, and P. H. S. Torr (2018). On the Robustness of Semantic Segmentation Models to Adversarial Attacks, in *CVPR 2018*. Cited on page 19.

K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, and J. Vreeken (2016a). Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks, *Journal of the Association for Information Science and Technology*, vol. 67(11), pp. 2635–2651. Cited on page 11.

K. Athukorala, A. Medlar, A. Oulasvirta, G. Jacucci, and D. Glowacka (2016b). Beyond relevance: Adapting exploration/exploitation in information retrieval, in *Proceedings of the 21st International Conference on Intelligent User Interfaces 2016*. Cited on page 11.

A. O. Bălan and M. J. Black (2008). The naked truth: Estimating body shape under clothing, in *ECCV 2008*. Cited on page 17.

C. Bauckhage, A. Jahanbekam, and C. Thurau (2010). Age Recognition in the Wild, in *ICPR 2010*. Cited on page 18.

X. W. S. Y. Bo Zhao, Jiashi Feng (2017). Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search, in *CVPR 2017*. Cited on page 16.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez (2013). Recommender systems survey, *Knowledge-based systems*, vol. 46, pp. 109–132. Cited on page 10.

F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black (2016). Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image, in *ECCV 2016*. Cited on pages 17, 69, 70, 72, 74, 79, 80, 81, 82, 83, 84, 85, 86, 87, 93, 96, 104, 112, and 115.

A. Borji and L. Itti (2014). Defending Yarbus: Eye movements reveal observers' task, *Journal of vision*, vol. 14(3), p. 29. Cited on pages 12 and 23.

A. Borji, A. Lennartz, and M. Pomplun (2014). What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations, *Neurocomputing*. Cited on pages 3, 4, 13, 24, 31, 35, and 53.

A. Borji, D. N. Sihite, and L. Itti (2012). Probabilistic learning of task-specific visual attention, in *CVPR 2012*. Cited on page 23.

P. Bouros, N. Lathia, M. Renz, F. Ricci, and D. Sacharidis (2015). LocalRec'15: Workshop on Location-Aware Recommendations, in *Proceedings of the 9th ACM Conference on Recommender Systems 2015*. Cited on page 11.

S. A. Brandt and L. W. Stark (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene, *Journal of Cognitive Neuroscience*, vol. 9(1). Cited on page 24.

W. Brendel and M. Bethge (2017). Comment on "Biologically inspired protection of deep networks from adversarial attacks", *arxiv*, vol. 1704.01547. Cited on page 86.

F. J. Brigham, E. Zaimi, J. J. Matkins, J. Shields, J. McDonnouugh, and J. J. Jakubecy (2001). *The eyes may have it:Reconsidering eye-movement research in human cognition*, vol. 15, Advances in Learning and Behavioral Disabilities. Cited on pages 11 and 12.

A. Bulling and D. Roggen (2011). Recognition of Visual Memory Recall Processes Using Eye Movement Analysis, in *Proc. UbiComp 2011*. Cited on pages 13 and 24.

A. Bulling, J. A. Ward, and H. Gellersen (2012). Multimodal Recognition of Reading Activity in ransit Using Body-Worn Sensors, *ACM Transactions on Applied Perception*, vol. 9(1). Cited on page 24.

A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster (2011). Eye Movement Analysis for Activity Recognition Using Electrooculography, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33(4), pp. 741–753. Cited on pages 13, 24, and 35.

A. Bulling, C. Weichel, and H. Gellersen (2013). EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour, in *CHI 2013*. Cited on pages 13 and 24.

N. J. Butko and J. R. Movellan (2010). Infomax control of eye movements, *IEEE Transactions on Autonomous Mental Development*, vol. 2(2), pp. 91–107. Cited on page 33.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in *CVPR 2017*. Cited on pages 17 and 96.

S. Castagnos, N. Jones, and P. Pu (2010). Eye-tracking product recommenders' usage, in *Proceedings of the fourth ACM conference on Recommender systems 2010*. Cited on page 11.

M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch (2008). Decoding what people see from where they look: Predicting visual stimuli from scanpaths, in *International Workshop on Attention in Cognitive Systems 2008*. Cited on page 24.

S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen (2004). Automatic license plate recognition, *IEEE Trans. Intelligent Transportation Systems*, vol. 5, pp. 42–53. Cited on page 18.

H. Chen, A. Gallagher, and B. Girod (2012). Describing clothing by semantic attributes, in *ECCV 2012*. Cited on pages 16 and 68.

P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security 2017*. Cited on page 86.

X. Chen and G. J. Zelinsky (2006). Real-world visual search is dominated by top-down guidance, *Vision Research*, vol. 46(24), pp. 4118 – 4133. Cited on page 13.

Y. Chen, T.-K. Kim, and R. Cipolla (2010). Inferring 3D Shapes and Deformations from Single Views, in *ECCV 2010*. Cited on page 17.

S. Chowdhury, F. Gibb, and M. Landoni (2011). Uncertainty in information seeking and retrieval: A study in an academic environment, *Information Processing & Management*, vol. 47(2), pp. 157–175. Cited on page 11.

M. Cissé, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier (2017). Parseval Networks: Improving Robustness to Adversarial Examples, in *ICML 2017*. Cited on page 19.

M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet (2017). Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples, in *NIPS 2017*. Cited on page 19.

V. Clay, P. König, and S. König (2019). Eye tracking in virtual reality, *Journal of Eye Movement Research*, vol. 12(1). Cited on page 105.

M. Claypool, P. Le, M. Wased, and D. Brown (2001). Implicit interest indicators, in *Proceedings of the 6th international conference on Intelligent user interfaces 2001*. Cited on pages 10 and 11.

J. Coddington, J. Xu, S. Sridharan, M. Rege, and R. Bailey (2012). Gaze-based image retrieval system using dual eye-trackers, in *2012 IEEE International Conference on Emerging Signal Processing Applications 2012*. Cited on pages 15 and 24.

R. Coen-Cagli, P. Coraggio, P. Napoletano, O. Schwartz, M. Ferraro, and G. Boccignone (2009). Visuomotor characterization of eye movements in a drawing task, *Vision Research*, vol. 49(8), pp. 810–818. Cited on page 24.

A. S. Cowen, M. M. Chun, and B. A. Kuhl (2014). Neural portraits of perception: Reconstructing face images from evoked brain activity, *NeuroImage*, vol. 94, pp. 12 – 22. Cited on pages 14 and 54.

G. Cucurull, P. Taslakian, and D. Vazquez (2019). Context-Aware Visual Compatibility Prediction, *arXiv preprint arXiv:1902.03646*. Cited on pages 16 and 106.

D. W. Cunningham, M. Kleiner, H. H. Bülthoff, and C. Wallraven (2004). The components of conversational facial expressions, in *Proceedings of the 1st Symposium on Applied perception in graphics and visualization 2004*. Cited on page 12.

D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas (2014). You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video, in *BMVC 2014*. Cited on page 14.

J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, *et al.* (2010). The YouTube video recommendation system, in *Proceedings of the fourth ACM conference on Recommender systems 2010*. Cited on page 16.

M. DeAngelus and J. B. Pelz (2009). Top-down control of eye movements: Yarbus revisited, *Visual Cognition*, vol. 17(6-7), pp. 790–811. Cited on page 23.

J. Deighton and M. Sorrell (1996). The future of interactive marketing., *Harvard business review*, vol. 74(6), pp. 151–160. Cited on page 18.

E. L. Denton, S. Chintala, A. Szlam, and R. Fergus (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks, in *NIPS 2015*. Cited on page 13.

P. Ekman (1977). Facial action coding system. Cited on page 12.

P. Ekman and W. V. Friesen (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding, *semiotica*, vol. 1(1), pp. 49–98. Cited on page 12.

R. Ekman (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, Oxford University Press, USA. Cited on page 12.

A. Fawzi, O. Fawzi, and P. Frossard (2018). Analysis of classifiers' robustness to adversarial perturbations, *Machine Learning*, vol. 107(3), pp. 481–508. Cited on page 19.

A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard (2016). Robustness of classifiers: from adversarial to random noise, in *NIPS 2016*. Cited on page 19.

M. Ferecatu and D. Geman (2009). A Statistical Framework for Image Category Search from a Mental Picture, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(6), pp. 1087–1101. Cited on pages 11, 13, 15, and 50.

C. D. Frith and U. Frith (2006). The neural basis of mentalizing, *Neuron*, vol. 50(4), pp. 531–534. Cited on page 10.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais (1987). The vocabulary problem in human-system communication, *Communications of the ACM*, vol. 30(11), pp. 964–971. Cited on page 12.

L. Gamberini and A. Spagnolli (2016). Towards a definition of symbiotic relations between humans and machines, in *International Workshop on Symbiotic Interaction 2016*. Cited on page 1.

D. Garude, A. Khopkar, M. Dhake, S. Laghane, and T. Maktum (2019). Skin-Tone and Occasion Oriented Outfit Recommendation System, *Available at SSRN 3368058*. Cited on page 106.

D. Gavrila (1999). The Visual Analysis of Human Movement: A Survey, *Computer Vision and Image Understanding*, vol. 73, pp. 82–98. Cited on page 85.

C. A. Gomez-Uribe and N. Hunt (2016). The netflix recommender system: Algorithms, business value, and innovation, *ACM Transactions on Management Information Systems (TMIS)*, vol. 6(4), p. 13. Cited on page 16.

K. Gong, X. Liang, X. Shen, and L. Lin (2017). Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing, in *CVPR 2017*. Cited on page 68.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014a). Generative Adversarial Nets, in *NIPS 2014*. Cited on page 13.

I. J. Goodfellow, J. Shlens, and C. Szegedy (2014b). Explaining and harnessing adversarial examples, *arxiv*, vol. abs/1412.6572. Cited on pages 19 and 94.

L. A. Granka, T. Joachims, and G. Gay (2004). Eye-tracking analysis of user behavior in WWW search, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval 2004*. Cited on pages 11 and 103.

M. R. Greene, T. Liu, and J. M. Wolfe (2012). Reconsidering Yarbus: A failure to predict observers? task from eye movement patterns, *Vision Research*, vol. 62. Cited on page 12.

J. Grudin (1995). Groupware and social dynamics: Eight challenges for developers, in *Readings in Human–Computer Interaction 1995*, pp. 762–774, Elsevier. Cited on page 11.

P. Guan, A. Weiss, A. O. Bălan, and M. J. Black (2009). Estimating human shape and pose from a single image, in *ICCV 2009*. Cited on page 17.

V. N. Gudivada and V. V. Raghavan (1995). Content based image retrieval systems, *Computer*, vol. 28(9), pp. 18–22. Cited on page 9.

G.-D. Guo, A. K. Jain, W.-Y. Ma, and H.-J. Zhang (2002). Learning similarity measure for natural image retrieval with relevance feedback, *IEEE Transactions on Neural Networks*, vol. 13(4). Cited on page 24.

M. Habermann, W. Xu, , M. Zollhoefer, G. Pons-Moll, and C. Theobalt (2019). LiveCap: Real-time Human Performance Capture from Monocular Video, *ACM Transactions on Graphics, (Proc. SIGGRAPH)*. Cited on page 86.

A. Haji-Abolhassani and J. J. Clark (2013). A computational model for task inference in visual search, *Journal of Vision*, vol. 13(3), p. 29. Cited on page 24.

A. Haji-Abolhassani and J. J. Clark (2014). An Inverse Yarbus Process: Predicting Observer's Task from Eye Movement Patterns, *Vision research*. Cited on pages 12 and 23.

X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis (2017a). Automatic Spatially-aware Fashion Concept Discovery, in *ICCV 2017*. Cited on pages 16 and 68.

X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis (2017b). Learning fashion compatibility with bidirectional lstms, in *ACM on Multimedia Conference 2017*. Cited on pages 16 and 67.

X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis (2017c). VITON: An Image-based Virtual Try-on Network, *arXiv preprint arXiv:1711.08447*. Cited on page 17.

J. Harel, C. Koch, and P. Perona (2006). Graph-based visual saliency, in *Proc. NIPS 2006*. Cited on page 30.

A. Harvey (2012). CV Dazzle: Camouflage from computer vision, *Technical report*. Cited on page 18.

N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel (2010). Multilinear pose and body shape estismation of dressed subjects from image sets, in *CVPR 2010*. Cited on page 17.

J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk (2013). Predicting cognitive state from eye movements, *PloS one*, vol. 8(5), p. e64937. Cited on page 13.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl (2004). Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems (TOIS)*, vol. 22(1), pp. 5–53. Cited on page 11.

R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia (2015). Sensitive Lifelogs: A Privacy Analysis of Photos from Wearable Cameras, in *CHI 2015*. Cited on page 18.

W.-L. Hsiao and K. Grauman (2017). Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding From Fashion Images, in *ICCV 2017*. Cited on pages 16 and 68.

W.-L. Hsiao and K. Grauman (2018). Creating Capsule Wardrobes From Fashion Images, in *CVPR 2018*. Cited on page 16.

E. Y. Huang and C.-Y. Lin (2005). Customer-oriented financial service personalization, *Industrial Management & Data Systems*, vol. 105(1), pp. 26–44. Cited on page 17.

Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black (2017). Towards Accurate Marker-less Human Shape and Pose Estimation over Time, in *3DV 2017*. Cited on page 17.

Z. Hussain, A. Klami, J. Kujala, A. P. Leung, K. Pasupa, P. Auer, S. Kaski, J. Laaksonen, and J. Shawe-Taylor (2014). Pinview: Implicit Feedback in Content-Based Image Retrieval, arXiv 1410.0471. Cited on page 24.

A. D. Hwang, E. C. Higgins, and M. Pomplun (2009). A model of top-down attentional control during visual search in complex scenes, *Journal of Vision*, vol. 9(5). Cited on pages 25 and 32.

A. Ilyas, L. Engstrom, and A. Madry (2018). Prior convictions: Black-box adversarial attacks with bandits and priors, *arXiv preprint arXiv:1807.07978*. Cited on page 86.

E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schieke (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model, in *ECCV 2016*. Cited on page 70.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros (2017). Image-to-image translation with conditional adversarial networks, in *CVPR 2017*. Cited on page 13.

G. Jacucci, A. Spagnolli, J. Freeman, and L. Gamberini (2015). Symbiotic interaction: a critical definition and comparison to other human-computer paradigms, in *International Workshop on Symbiotic Interaction 2015*. Cited on page 1.

L.-E. Janlert and E. Stolterman (2017). The meaning of interactivity—some proposals for definitions and measures, *Human–Computer Interaction*, vol. 32(3), pp. 103–138. Cited on page 1.

D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich (2010). *Recommender systems: an introduction*, Cambridge University Press. Cited on page 16.

L. Jansen, S. Onat, and P. König (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes, *Journal of Vision*, vol. 9(1), pp. 29–29. Cited on page 105.

T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay (2005). Accurately interpreting clickthrough data as implicit feedback, in *Sigir 2005*. Cited on pages 11 and 103.

P. N. Juslin and K. R. Scherer (2005). Vocal expression of affect, *The new handbook of methods in nonverbal behavior research*, pp. 65–135. Cited on page 12.

A. V. K. Simonyan, Karen and A. Zisserman (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv:1312.6034*. Cited on page 37.

C. Kanan, N. A. Ray, D. N. F. Bseiso, J. H. wen Hsiao, and G. W. Cottrell (2014). Predicting an observer task using multi-fixation pattern analysis, in *ETRA 2014*. Cited on pages 12 and 23.

A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik (2018). End-to-End Recovery of Human Shape and Pose, in *CVPR 2018*. Cited on pages 17, 81, 83, 87, and 112.

N. Karessli, Z. Akata, B. Schiele, and A. Bulling (2017). Gaze Embeddings for Zero-Shot Image Classification, in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 15.

S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath (2013). From where and how to what we see, in *ICCV 2013*. Cited on page 14.

M. S. Khan, R. Malik, A. Siddique, and A. Nawaz (2019). A new 3D eyeball tracking system to enhance the usability of page scrolling, *Optik*, vol. 185, pp. 1270–1276. Cited on page 105.

M. H. Kiapour, Y. M. Hadi, A. C. Berg, and T. L. Berg (2014). Hipster Wars: Discovering Elements of Fashion Styles, in *ECCV 2014*. Cited on pages 16 and 67.

M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg (2015). Where to Buy It: Matching Street Clothing Photos in Online Shops., in *ICCV 2015*. Cited on pages 16, 68, and 106.

H.-r. Kim and P. K. Chan (2005). Implicit indicators for interesting web pages, in *Web Information System and Technologies 2005*. Cited on page 11.

L. King (2002). The relationship between scene and eye movements, in *HICSS 2002*. Cited on page 24.

D. P. Kingma and M. Welling (2013). Auto-Encoding Variational Bayes, in *ICLR 2013*. Cited on page 14.

T. N. Kipf and M. Welling (2016). Variational graph auto-encoders, *arXiv preprint arXiv:1611.07308*. Cited on page 106.

A. Klami (2010). Inferring task-relevant image regions from gaze data, in *MLSP 2010*. Cited on page 15.

C. L. Kleinke (1986). Gaze and Eye Contact: A Research Review, *Psychological Bulletin*, vol. 100(1), pp. 78–100. Cited on pages 11 and 12.

M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia (2016). Enhancing Lifelogging Privacy by Detecting Screens, in *CHI 2016*. Cited on page 18.

L. Kozma, A. Klami, and S. Kaski (2009). GaZIR: Gaze-based Zooming Interface for Image Retrieval, in *ICMI-MLMI 2009*. Cited on pages 15 and 24.

A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks, in *NIPS 2012*. Cited on page 37.

M. F. Land (2006). Eye movements and the control of actions in everyday life, *Progress in Retinal and Eye Research*, vol. 25(3). Cited on page 24.

M. F. Land and S. Furneaux (1997). The knowledge base of the oculomotor system, *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 352(1358). Cited on pages 11 and 12.

C. Lassner, G. Pons-Moll, and P. V. Gehler (2017a). A Generative Model of People in Clothing, in *ICCV 2017*. Cited on page 17.

C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler (2017b). Unite the people: Closing the loop between 3d and 2d human representations, in *CVPR 2017*. Cited on pages 70, 86, and 93.

D.-J. Lee, J.-H. Ahn, and Y. Bang (2011). Managing consumer privacy concerns in personalization: a strategic analysis of privacy protection, *Mis Quarterly*, pp. 423–444. Cited on page 18.

S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller (2011). Introduction to machine learning for brain imaging, *Neuroimage*, vol. 56(2), pp. 387–399. Cited on page 12.

M. S. Lew, N. Sebe, C. Djeraba, and R. Jain (2006). Content-based multimedia information retrieval: State of the art and challenges, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2(1), pp. 1–19. Cited on page 9.

Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille (2015). The secrets of salient object segmentation, in *CVPR 2015*. Cited on page 14.

J. C. R. Licklider (1960). Man-computer symbiosis, *IRE transactions on human factors in electronics*, (1), pp. 4–11. Cited on page 1.

S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set, in *CVPR 2012*. Cited on pages 16, 68, and 106.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang (2016a). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. Cited on pages 16, 36, 39, 40, 67, and 68.

Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang (2016b). Fashion Landmark Detection in the Wild, in *ECCV 2016*. Cited on page 16.

T. Loetscher, C. J. Bockisch, M. E. Nicholls, and P. Brugger (2010). Eye position predicts what number you have in mind, *Current Biology*, vol. 20(6). Cited on page 24.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black (2015). SMPL: A Skinned Multi-Person Linear Model, *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34(6), pp. 248:1–248:16. Cited on pages 17, 68, 69, 70, 85, and 93.

S. J. Luck (2014). *An introduction to the event-related potential technique*, MIT press. Cited on page 12.

G. Marchionini (2006). Exploratory search: from finding to understanding, *Communications of the ACM*, vol. 49(4), pp. 41–46. Cited on page 11.

E. Marinoiu, D. Papava, and C. Sminchisescu (2013). Pictorial Human Spaces. How Well do Humans Perceive a 3D Articulated Pose?, in *ICCV 2013*. Cited on page 14.

F. W. Mast and S. M. Kosslyn (2002). Eye movements during visual mental imagery, *Trends in Cognitive Sciences*, vol. 6(7), pp. 271–272. Cited on page 24.

S. Mathe and C. Sminchisescu (2014). Multiple Instance Reinforcement Learning for Efficient Weakly-Supervised Detection in Images, *arXiv:1412.0100*. Cited on page 14.

S. McDonald and S. Flanagan (2004). Social perception deficits after traumatic brain injury: interaction between emotion recognition, mentalizing ability, and social communication., *Neuropsychology*, vol. 18(3), p. 572. Cited on page 10.

A. McNamara, K. Boyd, D. Oh, R. Sharpe, and A. Suther (2019). Using Eye Tracking to Improve Information Retrieval in Virtual Reality, in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) 2019*. Cited on pages 11 and 103.

A. Medlar and D. Glowacka (2018). How Consistent is Relevance Feedback in Exploratory Search?, in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management 2018*. Cited on page 11.

T. Miller and S. Agne (2005). Attention-based information retrieval using eye tracker data, in *International Conference On Knowledge Capture: Proceedings of the 3 rd international conference on Knowledge capture 2005*. Cited on pages 11 and 103.

A. Mishra, Y. Aloimonos, and C. L. Fah (2009). Active segmentation with fixation, in *ICCV 2009*. Cited on page 14.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks, in *CVPR 2016*. Cited on page 19.

B. C. Motter and E. J. Belky (1998). The guidance of eye movements during active visual search, *Vision Research*, vol. 38(12). Cited on pages 25 and 32.

H. J. Müller and J. Krummenacher (2006). Visual search and selective attention, *Visual Cognition*, vol. 14(4-8), pp. 389–410. Cited on page 9.

S. Muthukumaraswamy (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations, *Frontiers in human neuroscience*, vol. 7, p. 138. Cited on page 12.

T. Nagamatsu, J. Kamahara, and N. Tanaka (2009). Calibration-free gaze tracking using a binocular 3D eye model, in *CHI'09 Extended Abstracts on Human Factors in Computing Systems 2009*. Cited on page 105.

T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity, *Neuron*, vol. 63(6), pp. 902 – 915. Cited on page 2.

P. Negri, L. Gamberini, and S. Cutini (2015). A review of the research on subliminal techniques for implicit interaction in symbiotic systems, in *International Workshop on Symbiotic Interaction 2015*. Cited on page 1.

M. B. Neider and G. J. Zelinsky (2006). Searching for camouflaged targets: Effects of target-background similarity on visual search, *Vision Research*, vol. 46(14), pp. 2217 – 2235. Cited on page 13.

A. Nestor, D. C. Plaut, and M. Behrmann (2016). Feature-based face representations and image reconstruction from behavioral and neural data, *Proceedings of the National Academy of Sciences*, vol. 113(2), pp. 416–421. Cited on page 2.

C. Neustaedter, S. Greenberg, and M. Boyle (2006). Blur filtration fails to preserve privacy for home-based video conferencing. Cited on page 18.

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies, *Current Biology*, vol. 21(19), pp. 1641 – 1646. Cited on pages 14 and 54.

D. W. Oard, J. Kim, *et al.* (1998). Implicit feedback for recommender systems, in *Proceedings of the AAAI workshop on recommender systems 1998*. Cited on page 11.

S. J. Oh, M. Augustin, B. Schiele, and M. Fritz (2018). Towards Reverse-Engineering Black-Box Neural Networks, in *ICLR 2018*. Cited on page 86.

S. J. Oh, R. Benenson, M. Fritz, and B. Schiele (2015). Person Recognition in Personal Photo Collections, in *ICCV 2015*. Cited on page 106.

S. J. Oh, R. Benenson, M. Fritz, and B. Schiele (2016). Faceless Person Recognition: Privacy Implications in Social Media, in *ECCV 2016*. Cited on page 18.

S. J. Oh, M. Fritz, and B. Schiele (2017). Adversarial Image Perturbation for Privacy Protection – A Game Theory Perspective, in *ICCV 2017*. Cited on pages 86 and 100.

M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele (2018). Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation, in *3DV 2018*. Cited on page 87.

T. Orekondy, M. Fritz, and B. Schiele (2018). Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images, in *CVPR 2018*. Cited on pages 18 and 86.

T. Orekondy, B. Schiele, and M. Fritz (2017). Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images, in *ICCV 2017*. Cited on pages 18 and 86.

O. Oyekoya and F. Stentiford (2004). Eye tracking as a new interface for image retrieval, *BT Technology Journal*, vol. 22(3), pp. 161–169. Cited on page 15.

O. Oyekoya and F. Stentiford (2007). Perceptual image retrieval using eye movements, *International Journal of Computer Mathematics*, vol. 84(9), pp. 1379–1391. Cited on page 15.

M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang (2007). Human computing and machine understanding of human behavior: A survey, in *Artifical Intelligence for Human Computing 2007*, pp. 47–71, Springer. Cited on pages 10 and 12.

M. Pantic and L. J. Rothkrantz (2003). Toward an affect-sensitive multimodal human-computer interaction, *Proceedings of the IEEE*, vol. 91(9), pp. 1370–1390. Cited on page 12.

G. Papadopoulos, K. Apostolakis, and P. Daras (2014). Gaze-Based Relevance Feedback for Realizing Region-Based Image Retrieval, *ACM-MM*, vol. 16(2). Cited on pages 14 and 24.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami (2016). Distillation as a defense to adversarial perturbations against deep neural networks, in *Security and Privacy (SP), 2016 IEEE Symposium on 2016*. Cited on page 19.

U. Park, R. Mallipeddi, and M. Lee (2014). Human implicit intent discrimination using EEG and eye movement, in *International Conference on Neural Information Processing 2014*. Cited on page 2.

D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros (2016). Context Encoders: Feature Learning by Inpainting, in *CVPR 2016*. Cited on page 14.

G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis (2018). Learning to Estimate 3D Human Pose and Shape From a Single Color Image, in *CVPR 2018*. Cited on page 17.

R. J. Peters and L. Itti (2008). Congruence between model and human attention reveals unique signatures of critical visual events, in *NIPS 2008*. Cited on page 24.

O. Pierre-Yves (2003). The production and recognition of emotions in speech: features and algorithms, *International Journal of Human-Computer Studies*, vol. 59(1-2), pp. 157–183. Cited on page 12.

P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár (2016). Learning to refine object segments, in *ECCV 2016*. Cited on pages 70 and 71.

F. Pittaluga and S. J. Koppal (2015). Privacy preserving optics for miniature vision sensors, in *CVPR 2015*. Cited on page 18.

G. Pons-Moll, S. Pujades, S. Hu, and M. Black (2017). ClothCap: Seamless 4D Clothing Capture and Retargeting, *ACM Transactions on Graphics*, vol. 36(4). Cited on pages 17 and 96.

G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black (2015). Dyna: a model of dynamic human shape in motion, *ACM Transactions on Graphics*, vol. 34, p. 120. Cited on page 17.

U. Rajashekar, A. C. Bovik, and L. K. Cormack (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis, *Journal of Vision*, vol. 6(4). Cited on page 24.

N. Raval, A. Srivastava, K. Lebeck, L. Cox, and A. Machanavajjhala (2014). Markit: Privacy markers for protecting visual secrets, in *UbiComp 2014*. Cited on page 18.

D. Reed (1999). Consumer lifestyle data, *Precision Marketing*, vol. 8, pp. 7–9. Cited on page 18.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee (2016). Generative Adversarial Text-to-Image Synthesis, in *ICML 2016*. Cited on page 13.

L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik (2004). An information maximization model of eye movements, in *Proc. NIPS 2004*. Cited on page 33.

F. Ricci, L. Rokach, and B. Shapira (2015). Recommender systems: introduction and challenges, in *Recommender systems handbook 2015*, pp. 1–34, Springer. Cited on page 16.

G. S. Robertshaw and N. E. Marr (2006). The implications of incomplete and spurious personal information disclosures for direct marketing practice, *Journal of Database Marketing & Customer Strategy Management*, vol. 13(3), pp. 186–197. Cited on page 18.

N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal (2018). Fashion-Gen: The Generative Fashion Dataset and Challenge, in *arXiv preprint arXiv:1806.08317 2018*. Cited on page 16.

J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols (2003). Facial and vocal expressions of emotion, *Annual review of psychology*, vol. 54(1), pp. 329–349. Cited on page 12.

B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl (1998). Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system, in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 1998*. Cited on page 11.

Y. Sato, Y. Sugano, A. Sugimoto, Y. Kuno, and H. Koike (2016). Sensing and Controlling Human Gaze in Daily Living Space for Human-Harmonized Information

Environments, in *Human-Harmonized Information Technology, Volume 1 2016*, pp. 199–237, Springer Japan. Cited on pages 2 and 54.

H. Sattar, A. Bulling, and M. Fritz (2017). Predicting the category and attributes of visual search targets using deep gaze pooling, in *ICCVW 2017*. Cited on page 8.

H. Sattar, M. Fritz, and A. Bulling (2019a). Visual decoding of targets during visual search from human eye fixations, *arXiv preprint arXiv:1706.05993*. Cited on page 8.

H. Sattar, K. Krombholz, G. Pons-Moll, and M. Fritz (2019b). Shape Evasion: Preventing Body Shape Inference of Multi-Stage Approaches, *arXiv preprint arXiv:1905.11503*. Cited on page 8.

H. Sattar, S. Müller, M. Fritz, and A. Bulling (2015). Prediction of Search Targets From Fixations in Open-World Settings, in *CVPR 2015*. Cited on pages 7 and 35.

H. Sattar, G. Pons-Moll, and M. Fritz (2019c). Fashion is Taking Shape: Understanding Clothing Preference Based on Body Shape From Online Sources, in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV) 2019*. Cited on pages 8, 85, 87, 93, and 96.

A. Schmidt (2000). Implicit human computer interaction through context, *Personal technologies*, vol. 4(2-3), pp. 191–199. Cited on page 10.

F. Schroff, D. Kalenichenko, and J. Philbin (2015). Facenet: A unified embedding for face recognition and clustering, in *CVPR 2015*. Cited on page 106.

C. Schulze, R. Frister, and F. Shafait (2013). Eye-tracker based part-image selection for image retrieval., in *ICIP 2013*. Cited on page 15.

U. Shaham, Y. Yamada, and S. Negahban (2015). Understanding adversarial training: Increasing local stability of neural nets through robust optimization, *arxiv*, vol. abs/1511.05432. Cited on page 19.

D. Shankar, S. Narumanchi, H. Ananya, P. Kompalli, and K. Chaudhury (2017). Deep learning based large scale visual recommendation and search for e-commerce, *arXiv preprint arXiv:1703.02344*. Cited on page 106.

M. Shao, L. Li, and Y. Fu (2013). What do you do? Occupation recognition in a photo via social context, in *CVPR 2013*. Cited on page 18.

I. Shcherbatyi, A. Bulling, and M. Fritz (2015). GazeDPM: Early Integration of Gaze Information in Deformable Part Models, arxiv:1505.05753 [cs.cv]. Cited on page 14.

L. Sigal, A. Balan, and M. J. Black (2008). Combined discriminative and generative articulated pose and non-rigid shape estimation, in *NIPS 2008*. Cited on page 17.

N. Silva, T. Schreck, E. Veas, V. Sabol, E. Eggeling, and D. W. Fellner (2018). Leveraging eye-gaze and time-series features to predict user interests and build a recommendation model for visual analysis, in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications 2018*. Cited on page 11.

E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun (2015). Neuroaesthetics in Fashion: Modeling the Perception of Fashionability, in *CVPR 2015*. Cited on pages 16 and 67.

H. Song and N. Moon (2019). Eye-tracking and social behavior preference-based recommendation system, *The Journal of Supercomputing*, vol. 75(4), pp. 1990–2006. Cited on page 11.

A. Spagnolli, M. Conti, G. Guerra, J. Freeman, D. Kirsh, and A. van Wynsberghe (2016). Adapting the system to users based on implicit data: ethical risks and possible solutions, in *International Workshop on Symbiotic Interaction 2016*. Cited on pages 1 and 18.

M. M. Spapé, M. Filetti, M. J. Eugster, G. Jacucci, and N. Ravaja (2015). Human computer interaction meets psychophysiology: a critical perspective, in *International Workshop on Symbiotic Interaction 2015*. Cited on page 12.

E. I. Sparling and S. Sen (2011). Rating: how difficult is it?, in *Proceedings of the fifth ACM conference on Recommender systems 2011*. Cited on page 11.

M. Speicher, S. Cucerca, and A. Krüger (2017). Vrshop: A mobile interactive virtual reality shopping environment combining the benefits of on-and offline shopping, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1(3), p. 102. Cited on page 105.

G. Stefanou and S. P. Wilson (). Mental Image Category Search: a Bayesian Approach. Cited on page 24.

J. Steil and A. Bulling (2015). Discovery of everyday human activities from long-term visual behaviour using topic models, in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015*. Cited on page 35.

J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling (2018a). Privacy-Aware Eye Tracking Using Differential Privacy, *arXiv preprint arXiv:1812.08000*. Cited on page 18.

J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling (2018b). PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features, *arXiv preprint arXiv:1801.04457*. Cited on page 18.

J. Su, D. V. Vargas, and K. Sakurai (2017). One pixel attack for fooling deep neural networks, *arxiv*, vol. abs/1710.08864. Cited on page 86.

R. Subramanian, V. Yanulevskaya, and N. Sebe (2011). Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements, in *ACM-MM 2011*. Cited on page 14.

Y. Sugano and A. Bulling (2016). Seeing with Humans: Gaze-Assisted Neural Image Captioning, arxiv:1608.05203. Cited on page 14.

M. Sun, J. Wang, and Z. Chi (2019). Eye-tracking based relevance feedback for iterative face image retrieval, in *Tenth International Conference on Graphics and Image Processing (ICGIP 2018) 2019*. Cited on pages 11 and 103.

Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele (2018). A Hybrid Model for Identity Obfuscation by Face Replacement, in *ECCV 2018*. Cited on page 86.

X. Sun, P. Wu, and S. C. H. Hoi (2017). Face Detection using Deep Learning: An Improved Faster RCNN Approach, *CoRR*, vol. abs/1701.08289. Cited on page 18.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus (2014). Intriguing properties of neural networks, in *ICLR 2014*. Cited on pages 19, 94, and 100.

V. Tan, I. Budvytis, and R. Cipolla (2017). Indirect deep structured learning for 3D human body shape and pose prediction, in *BMVC 2017*. Cited on page 17.

B. Tessendorf, A. Bulling, D. Roggen, T. Stiefmeier, M. Feilner, P. Derleth, and G. Tröster (2011). Recognition of Hearing Needs From Body and Eye Movements to Improve Hearing Instruments, in *Proc. Pervasive 2011*. Cited on pages 13 and 24.

T. Toyama, T. Kieninger, F. Shafait, and A. Dengel (2012). Gaze guided object recognition using a head-mounted eye tracker, in *ETRA 2012*. Cited on page 14.

A. M. Treisman and G. Gelade (1980). A feature-integration theory of attention, *Cognitive psychology*, vol. 12(1), pp. 97–136. Cited on page 9.

E. Tretschk, S. J. Oh, and M. Fritz (2018). Sequential Attacks on Agents for Long-Term Adversarial Goals, in *ACM Computer Science in Cars Symposium – Future Challenges in Artificial Intelligence Security for Autonomous Vehicles (CSCS) 2018*. Cited on page 86.

D. Tripathi, A. Medlar, and D. Glowacka (2019). How Relevance Feedback is Framed Affects User Experience, but not Behaviour, in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval 2019*. Cited on page 11.

H. Tung, H. Wei, E. Yumer, and K. Fragkiadaki (2017). Self-supervised Learning of Motion Capture, in *NIPS 2017*. Cited on page 17.

M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth (2018). Learning type-aware embeddings for fashion compatibility, in *Proceedings of the European Conference on Computer Vision (ECCV) 2018*. Cited on page 16.

P.-P. Verbeek (2015). Beyond interaction: A short introduction to mediation theory, *Interactions*, vol. 22(3), pp. 26–31. Cited on page 1.

P. A. Viola and M. J. Jones (2001). Robust Real-Time Face Detection, *International Journal of Computer Vision*, vol. 57, pp. 137–154. Cited on page 18.

T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll (2018). Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera, in *ECCV 2018*. Cited on page 95.

J. Wagner, F. Lingenfelser, E. André, D. Mazzei, A. Tognetti, A. Lanatà, D. De Rossi, A. Betella, R. Zucca, P. Omedas, *et al.* (2013). A sensing architecture for empathetic data systems, in *Proceedings of the 4th Augmented Human International Conference 2013*. Cited on page 10.

G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth (2010). Seeing People in Social Context: Recognizing People and Social Relationships, in *ECCV 2010*. Cited on page 18.

W. Wang, Y. Xu, J. Shen, and S.-C. Zhu (2018). Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification, in *CVPR 2018*. Cited on page 16.

X. Wang and T. Zhang (2011). Clothes Search in Consumer Photos via Color Matching and Attribute Learning, in *ACM International Conference on Multimedia 2011*. Cited on pages 16 and 68.

A. Watson and M. A. Sasse (1998). Measuring perceived quality of speech and video in multimedia conferencing applications, in *Proceedings of the 6th ACM International Conference on Multimedia 1998*. Cited on page 10.

E. M. Whitham, K. J. Pope, S. P. Fitzgibbon, T. Lewis, C. R. Clark, S. Loveless, M. Broberg, A. Wallace, D. DeLosAngeles, P. Lillie, *et al.* (2007). Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG, *Clinical Neurophysiology*, vol. 118(8), pp. 1877–1888. Cited on page 12.

M. J. Wilber, V. Shmatikov, and S. Belongie (2016). Can we still avoid automatic face detection?, in *WACV 2016*. Cited on page 18.

S. P. Wilson, J. Fauqueur, and N. Boujemaa (2008). *Mental Search in Image Databases: Implicit Versus Explicit Content Query*, Springer Berlin Heidelberg. Cited on page 13.

J. M. Wolfe and T. S. Horowitz (2017). Five factors that guide attention in visual search, *Nature Human Behaviour*, vol. 1(3), p. 0058. Cited on pages 3 and 9.

C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland (1997). Pfinder: real-time tracking of the human body, *TPAMI*, vol. 19(7), pp. 780–785. Cited on page 85.

J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao (2014). Predicting human gaze beyond pixels, *Journal of vision*, vol. 14(1). Cited on page 14.

X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darell, and D. Song (2017). Can you fool AI with adversarial examples on a visual Turing test?, *CoRR*, vol. abs/1709.08693. Cited on page 19.

K. Yamaguchi, M. H. Kiapour, and T. L. Berg (2013). Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items, in *ICCV 2013*. Cited on pages 16 and 68.

K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg (2012). Parsing clothing in fashion photographs, in *CVPR 2012*. Cited on pages 16 and 68.

X. Yan, J. Yang, K. Sohn, and H. Lee (2016). Attribute2Image: Conditional Image Generation from Visual Attributes, in *ECCV 2016*. Cited on pages 14, 54, and 56.

J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer (2016). Estimation of Human Body Shape in Motion with Wide Clothing, in *ECCV 2016*. Cited on page 17.

W. Yang, P. Luo, and L. Lin (2014). Clothing Co-Parsing by Joint Image Segmentation and Labeling, in *CVPR 2014*. Cited on pages 16 and 68.

A. L. Yarbus, B. Haigh, and L. A. Rigss (1967). *Eye movements and vision*, vol. 2, Plenum press New York. Cited on pages 2, 11, 12, and 23.

R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec (2018). Graph convolutional neural networks for web-scale recommender systems, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018*. Cited on page 106.

Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M.Hospedales, and C. C. Loy (2016). Sketch Me That Shoe, in *CVPR 2016*. Cited on pages 11 and 15.

M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu (2018). Human appearance transfer, in *CVPR 2018*. Cited on page 17.

G. J. Zelinsky, Y. Peng, and D. Samaras (2013). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets, *Journal of Vision*, vol. 13(14). Cited on pages 3, 4, 13, 23, 24, 31, 35, and 53.

C. Zhang, S. Pujades, M. Black, and G. Pons-Moll (2017a). Detailed, accurate, human shape estimation from clothed 3D scan sequences, in *CVPR 2017*. Cited on page 17.

C. Zhang, S. Pujades, M. Black, and G. Pons-Moll (2017b). Detailed, accurate, human shape estimation from clothed 3D scan sequences, in *CVPR 2017*. Cited on page 96.

D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, and R. Boots (2017c). EEG-based intention recognition from spatio-temporal representations via cascade and parallel convolutional recurrent neural networks, *arXiv preprint arXiv:1708.06578*. Cited on page 2.

H. Zhang, W. Jia, X. He, and Q. Wu (2006). Learning-Based License Plate Detection Using Global and Local Features, in *ICPR 2006*. Cited on page 18.

W. Zhang, H. Yang, D. Samaras, and G. J. Zelinsky (2005). A Computational Model of Eye Movements During Object Class Detection, in *NIPS 2005*. Cited on pages 13 and 23.

X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian (2017d). Trip Outfits Advisor: Location-Oriented Clothing Recommendation, *IEEE Transactions on Multimedia*. Cited on pages 16 and 106.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2015a). Appearance-based gaze estimation in the wild, in *CVPR 2015*. Cited on pages 2 and 54.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2015b). Appearance-Based Gaze Estimation in the Wild, in *CVPR 2015*. Cited on page 44.

Y. Zhang, H. Fu, Z. Liang, Z. Chi, and D. Feng (2010). Eye Movement As an Interaction Mechanism for Relevance Feedback in a Content-based Image Retrieval System, in *ETRA 2010*. Cited on page 24.

Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven (2009). Tour the world: Building a web-scale landmark recognition engine, in *CVPR 2009*. Cited on page 18.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2016). Learning Deep Features for Discriminative Localization, in *CVPR 2016*. Cited on pages 37, 38, and 39.

S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han (2010). Parametric reshaping of human bodies in images, in *ACM Transactions on Graphics 2010*. Cited on page 17.

W. Zhou, H. Li, Y. Lu, and Q. Tian (2012). Principal Visual Word Discovery for Automatic License Plate Detection, *IEEE Trans. Image Processing*, vol. 21, pp. 4269–4279. Cited on page 18.