Learning visual attributes from contextual explanations

by

Nils Ever Murrugarra Llerena

Msc, University of São Paulo, 2011

Submitted to the Graduate Faculty of
the School of Computing and Information in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Nils Ever Murrugarra Llerena

It was defended on

August 5, 2019

and approved by

Adriana Kovashka, PhD, Assistant Professor, Department of Computer Science

Rebecca Hwa, PhD, Associate Professor, Department of Computer Science

Milos Hauskrecht, PhD, Professor, Department of Computer Science

Daqing He, PhD, Professor, Department of Informatics and Networked Systems

Dissertation Director: Adriana Kovashka, PhD, Assistant Professor, Department of Computer Science

Learning visual attributes from contextual explanations

Nils Ever Murrugarra Llerena, PhD

University of Pittsburgh, 2019

In computer vision, attributes are mid-level concepts shared across categories. They provide a natural communication between humans and machines for image retrieval. They also provide detailed information about objects. Finally, attributes can describe properties of unfamiliar objects. These are some appealing properties of attributes, but learning attributes is a challenging task. Since attributes are less well-defined, capturing them with computational models poses a different set of challenges than capturing object categories does. There is a miscommunication of attributes between humans and machines, since machines may not understand what humans have in mind when referring to a particular attribute. Humans usually provide labels if an object or attribute is present or not without any explanation. However, attributes are more complex and may require explanations for a better understanding.

This Ph.D. thesis aims to tackle these challenges in learning automatic attribute predictive models. In particular, it focuses on *enhancing attribute predictive power with contextual explanations.* These explanations aim to enhance data quality with human knowledge, which can be expressed in the form of interactions and may be affected by our personality.

First, we emulate human learning skill to understand unfamiliar situations. Humans infer properties from what they already know (background knowledge). Hence, we study attribute learning in data-scarce and non-related domains emulating human understanding skills. We *discover* transferable knowledge to learn attributes from different domains.

Our previous project inspires us to *request* contextual explanations to improve attribute learning. Thus, we enhance attribute learning with context in the form of gaze, captioning, and sketches. Human gaze captures subconscious intuition and associates certain components to the meaning of an attribute. For example, gaze associates the tiptoe of a shoe to a pointy attribute. To complement this gaze representation, captioning follows conscious thinking with prior analysis. An annotator may analyze an image and may provide the following description: “This shoe is pointy because its sharp form at the tiptoe”. Finally, in image search, sketches provide a holistic view of an image query, which complement specific details encapsulated via attribute comparisons. To conclude, our methods with contextual explanations outperform many baselines via quantitative and qualitative evaluation.

Table of Contents

Preface

First of all, I would like to thank my advisor Professor Adriana Kovashka, who inspires and guides me along my Ph.D. journey. She helps me to unlock all my research potential via uncountable hours of discussion, insightful feedback for research and presentations, and a lot of passion and patience to pursue my projects. I feel grateful for believe in me, even in doubtful situations and provide me a spark of hope to be more perseverant.

I also like to thank my Ph.D. committee, Professors Rebecca Hwa, Milos Hauskrecht, and Daqing He. I appreciate their insights discussion, valuable feedback, and generous support to improve my research. I am grateful for their dedication and time in my thesis.

Besides my Ph.D. committee, I feel grateful to my labmates: Chris Thomas and Keren Ye. They always were willing to help with research and coding issues. They provide a friendly, intellectual and refreshing research environment. Also, I would like to thank my classmates at the computer science department, who promote pleasant times during my Ph.D. Specially, I would like to thank: Yuyu Zhou, Qin Guan, Gaurav Trivedi, Mao-Lin Li, Duncan Yung, Henrique Potter, Michael Cui, Angen Zheng, Jeongmin Lee, CharmGil Hong, Zhipen Luo, and Santiago Bock.

Besides Pitt community, I am grateful for internships at Education Testing Services (ETS) and ASEA Brown Bovery (ABB). I like to thank my mentors: Aoiffe Cahill, Mirrasoul Mousavi, and Mithun Acharya. They stretch my boundaries and extend my understanding of research in an industrial setting. I am grateful for all their insights in data collection strategies, which were crucial for this thesis.

Also, I would like to thank the Computer Science department at National University of Trujillo (UNT) and University of Sao Paulo (USP). First, I would like to thank my professors at UNT. Especially, I would like to thank Prof. Jose Saavedra, Violeta Chang, Ivan Sipiran and Jorge Guevara to spread their enthusiasm and curiosity in artificial intelligence. Also, I would like to thank Prof. Alneu de Andrade Lopes from USP to enhance my research skills and guide through my first serious research challenge: my master thesis.

In addition to my professors at National University of Trujillo, I would like to thank my friends from Computer Science Students' Society (SECC). They are always a source of inspiration in my life and one of the reasons to go abroad. I always remember our gatherings promoting computer science at our local University. I would like to thank specially to Marks Calderon, Clayder Gonzalez, David Wong, Fredy Carranza, Jorge Valverde, Pedro Shiguihara, Eduardo Rodriguez, Heider Sanchez, and Luis Lujan.

Also, I would like to thank my parents: Salomon and Genny; my brothers: Hans, Maisi and Jeffri; and my cousin Sandra for their love, sympathy, and support. They always inspire me to become a better person. I am grateful for their unconditional support in all projects in my life.

In addition to my family, I thank my friends in Pittsburgh. First, I thank Bazan family, who adopted me as an extra family member. Specially thanks to Cathy, Quique, Gaby, Claudia, Anton, Norma, Enrique, and Luis. Also, thanks to my friends: Lupita, Jorge, Ana Paula, Javier, and MariCarmen. All of them make my time at Pittsburgh pleasant and unforgettable. I really appreciate our board game gatherings.

Lastly, my whole appreciation goes to my girlfriend, Marleny. She always stays with me, even in hardest and unpredictable situations. She finds the correct words to support and encourage me to find a light of hope. She makes me happy when everything was going wrong. Without her support and love, I can not emotionally accomplish this Ph.D.

1.0 Introduction

In computer vision, attributes are mid-level concepts shared across categories. They provide a natural communication between humans and machines. For example, we can provide the query “I want a sky-blue and elegant shirt” to a search system. They also provide detailed information about objects. For example, let’s compare “cat” versus “a small domesticated animal with soft fur, and retractable claws”. The second statement provides much more detail about a cat. Finally, attributes can describe properties of unfamiliar objects. For example, if we know a horse and a cat, we can infer some properties of a zebra - even if we’ve never seen one. Zebra has four legs like a horse, and it has stripes as a cat.

These are some very appealing properties of attributes, but learning attributes is a challenging task. They are not well-defined: they can have different interpretations for different people, as opposed to an object, where the meaning is more standard. To see why, consider the following thought experiment. If a person is asked to draw a “boot”, the drawings of different people will likely not differ very much. But if a person is asked to draw what the attributes “formal” or “feminine” mean, drawings will vary. Similarly, drawings of a “forest” will likely all include a number of trees, but drawings of a “natural”, “open-area”, or “cluttered” scene will differ greatly among artists.

From the previous experiment, attributes are less well-defined than objects and may have different interpretations. There is a miscommunication of attributes between humans and machines, since machines may not understand what humans have in mind when referring to a particular attribute. Humans usually provide labels if an attribute is present or not without any explanation. However, attributes are more complex and may require explanations to understand them better.

This Ph.D. thesis aims to tackle these challenges in learning automatic attribute predictive models. Specifically, we investigate the following hypothesis:

Hypothesis. *Algorithms learning from contextual explanations will learn to predict and use attributes more accurately, compared to algorithms that don't use such explanations.*

Contextual explanations are based on the context in which attributes occur and may clarify their meaning to facilitate accurate learning. For example, to categorize face images as happy or sad, brain imaging may show brain regions associated with positive and negative feelings. Similarly, brain waves may show opposite waves. These two forms of contextual explanations provide subconscious thinking to take a decision. We complement subconscious analysis with conscious thinking on visual cues via a selection interface. In this scenario, annotators draw a polygon around the lips region on a face to denote happiness or sadness. Also, contextual explanations can be complementary. Visual cues may be complemented by physical interactions. For example, to identify furry animals, we can observe the texture of their fur and also touch them.

From the vast options to encode contextual explanations through sensors or visual interfaces, these contextual cues are also present in computer vision via saliency maps and gaze trackers. Saliency maps represent visual importance of a corresponding visual scene among its components (i.e. objects or parts) [62]. From our previous experiment, saliency maps can identify lips, teeth, and smiles as relevant parts highly correlated to identify happiness. Saliency maps also can encapsulate subconscious and conscious data. Saliency subconscious maps are acquired from a gaze tracker, while saliency conscious maps can be acquired from human interactions with a polygon drawing interface.

We discover and incorporate contextual explanations in recognition tasks. First, we emulate humans to understand unfamiliar situations. Humans try to infer properties from what they already know (background knowledge). In this setup, we represent unfamiliar situations via unrelated domains such as animal, scene, shoe, object, and texture. Also, we infer properties finding related attributes on unrelated domains. Given an attribute classifier, we aim to *discover* relevant components, that can be reused to learn other attributes. This finding inspires us to *request* contextual explanations via human rationale data to enhance attributes predictive power. We explore human rationales in the form of human gaze, text, and sketches. Human gaze captures subconscious intuition of the meaning of an attribute. For example, it identifies a tiptoe of a shoe as the most important component for pointy shoes. In contrast, text follows a conscious thinking with prior analysis. Following our pointy example, an annotator analyze an image and produce the following description: "This shoe is pointy because of its sharp form at the tiptoe". Finally, sketches encapsulate human rationale in a visual representation, which complements attributes representation. Sketches provide a holistic view of the query, in contrast to specific details encapsulated via attributes. Among these approaches, we incorporate contextual explanations for attribute learning via discovering relevant knowledge or requesting human intervention to enrich data, as shown in Figure 1.

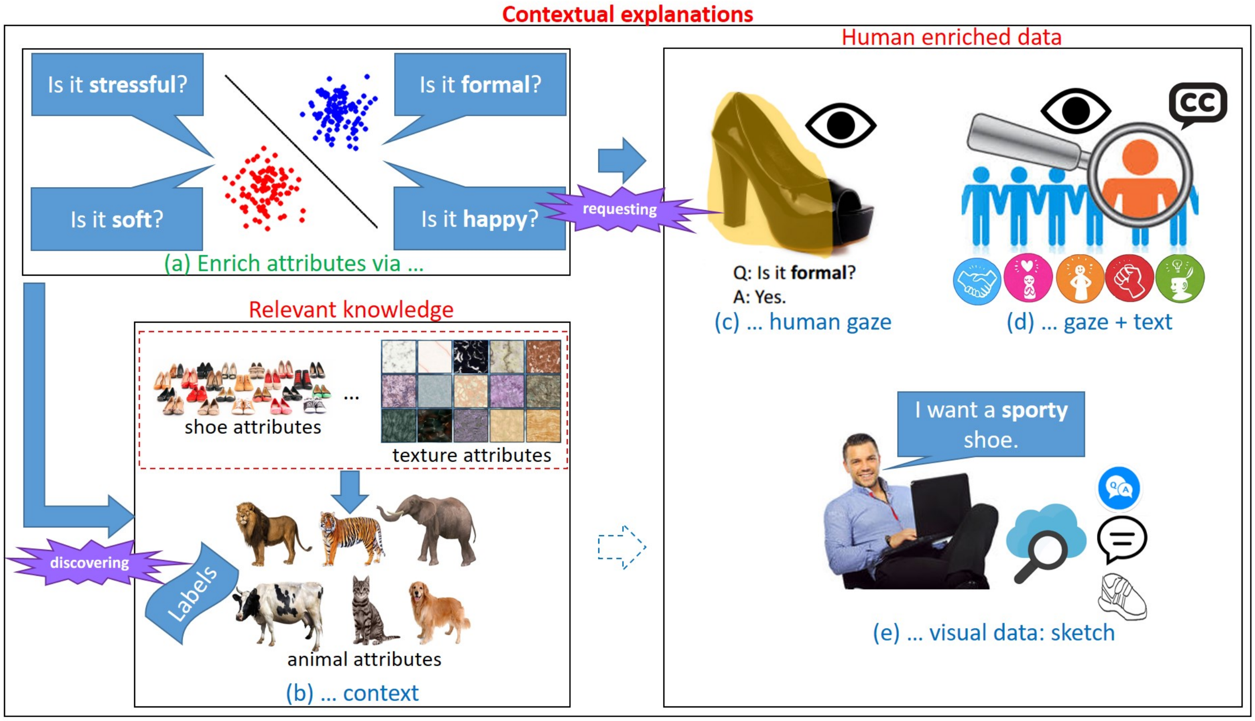


Figure 1: Overview of the work in this thesis: enrich attributes with contextual explanations. We start from discovering explanations in traditional data (bottom left) to request contextual explanations by human enriched data (right). In traditional data, we enrich attributes by *discovering* contextual information (b). Then, attributes are enriched by *requesting* human rationale data in the form of: human gaze (c), gaze and text (d) and sketches (e).

Following our intuitive approaches for contextual explanations, we describe different ways to work with attributes. In the first category, we investigate how to improve attribute learning by discovering relevant components that are shared among different attributes. In the second category, we aim to enrich attributes by requesting human contextual explanations. First, we improve attribute learning with gaze. Then, we improve the ability to classify personality-related attributes by contextualizing these through the ways in which people describe or look images. And finally, we complement attribute descriptive power with human-generated sketches to improve image retrieval. These sketches encapsulate a holistic view of the image query and provide contextual explanation in the form of visual cues. These holistic visual cues complement attribute-based textual representations.

The remainder of this chapter is organized as follows. In sections 1.1 and 1.2, we briefly introduce our approaches to enrich attributes with contextual explanations and discuss our solutions. In section 1.3, we show how all the projects in this thesis relate to and complement each other. In section 1.4, we describe our contributions. Finally, we outline the organization of this thesis in section 1.5.

1.1 ***Discovering*** contextual explanations for attribute learning

Attributes can be learned in isolation or in a multi-task scenario. These approaches require huge amounts of data to succeed because they require many objects with the attribute present or not to capture its real meaning [36]. Also, most recent successful approaches are based on deep learning [39, 128], and they require lots of data [15]. However, what can we do in the case of data scarcity? A usual solution is to perform transfer learning.

Transfer learning aims to transfer knowledge from a source domain with huge data to a target domain with scarce data. Source and target domain must be related in some sense. For example, a computer can adapt a spam detector from one mailbox to another [61]. Also, a computer can learn a race car detector from a traditional car detector [147].

In attribute learning, traditional transfer approaches perform adaptation between attributes from the same domain [17, 89, 48]. We define a domain as a set of semantically related categories. However, what could we do if in addition to the data scarcity, we do not have any data from semantic related categories? For example, let us imagine we have an entirely new domain of objects (e.g. deep sea animals) which is visually distinct from other objects we have previously encountered, and we have very sparse labeled data on that domain. Let us assume we have plentiful data from unrelated domains, e.g. materials, clothing, and natural scenes. Can we still use that unrelated data?

We examine how we can transfer knowledge from attribute classifiers on unrelated domains, as shown in Figure 2. For example, this transfer approach might mean we want to learn a model for the animal attribute "hooved" from the scene attribute "natural", the texture attribute "woolen", etc. We define semantic transfer as learning a target attribute using the remaining attributes in that same data set as source models. This is the approach used in prior work [17, 89, 48]. In contrast, in non-semantic transfer (our proposed approach), we use source attributes from *other* datasets. We show that allowing transfer from diverse datasets allows computers to learn more accurate models, but only when we intelligently select how to weigh the contribution of the source models. The intuition behind our approach is that the same visual patterns recur in different realms of the visual world, but language has evolved in such a way that they receive different names depending on which domain of objects they occur in.

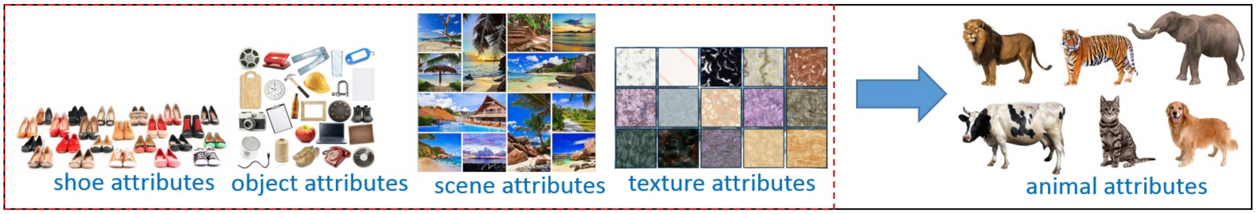


Figure 2: We study transfer of knowledge among disjointed domains. Can shoe, object, scene, and texture attributes be beneficial for learning *animal* attributes, despite the lack of semantic relation between the categories and attributes?

We propose an attention-guided transfer network. Briefly, our approach works as follows. First, the network receives training images for attributes in both the source and target domains. Second, it separately learns models for the attributes in each domain and then measures how related each target domain classifier is to the classifiers in the source domains via an attention mechanism. Finally, it uses these measures of similarity (relatedness) to compute a weighted combination of the source classifiers, which then becomes the new classifier for the target attribute. Importantly, we show that when the source attributes come from a diverse set of domains, the gain we obtain from this transfer of knowledge is greater than if only the attributes from the same domain had been used.

Note that our current solution aims to discover and select the most relevant and shareable knowledge, similar to humans. The discovered rationale aims to define an attribute as the combination of others. Discovering meaningful transferable knowledge motivates us to request contextual explanations to improve attribute learning. Hence, in the next section, we explore attribute learning closely involving humans, to improve our data quality. One of these approaches also focuses on select relevant knowledge in the form of localization via human gaze data.

1.2 ***Requesting*** contextual explanations for attribute learning via human interactions

We present approaches that combine human intervention and contextual explanations to improve attribute learning. The first one focuses on attribute learning as a core task, and the remaining two on attribute learning as a side task for cross-modality retrieval and image search. Also, the first project enriches attributes via human gaze, and the last two enhance attribute representation with (gaze, text) and (sketches, user simulations), respectively. Human gaze captures subconscious intuition of the meaning of an attribute. In contrast, text follows a conscious thinking with prior analysis. Finally, sketches encapsulate human rationale in a visual representation, which complements attributes interaction.

1.2.1 Learning attributes from human gaze

In terms of attribute learning as a core task, which is similar to object recognition, we can learn attributes with a traditional machine learning pipeline. However, attributes are less-well defined and there exists a disconnect between humans and machines in how they perceive attributes, as we described in our previous section. Thus, the best way to narrow the discrepancy is by learning from humans what attributes really mean.

We propose to learn attribute models using human gaze maps that show which part of an image contains the attribute, as shown in Figure 3. To obtain gaze maps for each attribute, we conduct human subject experiments where we ask viewers to examine images of faces, shoes, and scenes, and determine if a given attribute is present in the image or not. We use an inexpensive GazePoint eye tracking device which is simply placed in front of a monitor to track viewers' gaze and record the locations in the image that had some number of fixations. We aggregate the gaze collected from multiple people on training images, to obtain an averaged gaze map per attribute that we use to extract features from both train and test images. We also experiment with learning a saliency model that predicts which pixels will be fixated. To capture the potential ambiguity and visual variation within each attribute, we cluster the positive images per attribute and their corresponding gaze locations and obtain multiple gaze maps per attribute. We create one classifier per gaze map which only uses features from the region under nonzero gaze map values, for both training and testing.

The gaze maps that we learn from humans indicate the spatial support for an attribute

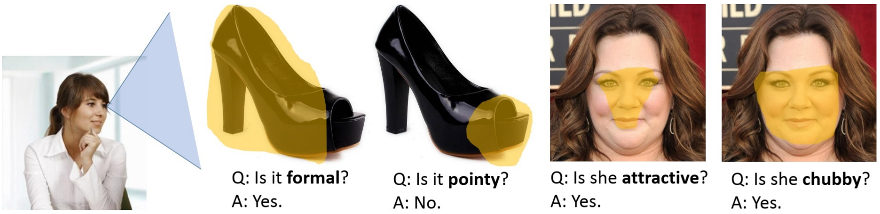


Figure 3: We learn the spatial support of attributes by asking humans to judge if an attribute is present in training images. We use this support to improve attribute prediction.

in an image and allow us to better understand what the attribute means. We use gaze maps to identify regions that should be used to train attribute models. We show this process achieves competitive attribute prediction accuracy compared to alternative ways to select relevant features. We also demonstrate additional applications showing how our method can be used to visualize attribute models, and how it can be employed to discover groups among users in terms of their understanding of attribute presence.

In this project, we study our sight sense in isolation as a contextual explanation. However, our perception through our senses is affected by our experience, personality, and bias. For example, "open-minded" people are more likely to combine visual elements and perceive them as a unified whole [5], disorganized people or ones with low self-confidence have a high tolerance of visual blur [164], and people who believe in paranormal events are more likely to perceive objects in images that only contain noise [109]. Hence, to learn gaze easily and more accurate, we learn gaze and personality jointly in our next project. Also, we enrich contextual explanations via text descriptions. Text encapsulates conscious thinking, complementary to gaze. These descriptions bridge the gap between gaze and personality and even enrich data by capturing the writing style of annotators.

1.2.2 Cross-modality personalization for retrieval

We extend our human gaze work for cross-modality personalized retrieval using gaze, captions, and personality questionnaires. Our goal is to find a shared embedding, where these paired data modalities are closed together. Hence, for example, we can retrieve the most probable caption given a gaze representation. Our gaze data collection does not use an eye-tracking device, which is not accessible for everybody and requires a meticulous calibration. In order to solve these issues, our project employs a *revealing mask* web interface. This interface does not require any calibration and is widely accessible from any web browser. Hence, data can be collected at a higher scale with crowdsourcing. This interface shows a blurred image, and users click on it to reveal certain parts. Collecting data with this interface is positively correlated with data from eye-trackers [70]. In addition to gaze human enriched data, we collect image captions (writing style) and personality questionnaires. This project uses attributes to represent personality, and as a side task to improve cross-modality personalized retrieval. Also, gaze and writing styles capture indirectly different interpretation of personality traits (i.e. attributes). We find that personality traits complement and enhance gaze and image captioning learning, which reaffirms the fact that personality influences our perception.

This variance in perception due to variance in personality is important to consider when predicting what meaning viewers will extract from imagery. It is especially important to model when predicting how humans will describe images that aim to impart opinions on the viewer in subtle ways. Prior work has examined the meaning that the average human extracts from images, by learning to predict what descriptive captions are appropriate for a given image. However, not all humans will describe the image in the same manner. Further, the way they describe it depends on how they look at it. We illustrate this in Fig. 4. When shown this car advertisement, an outgoing family man might first observe the children in front or behind the car, and interpret the message of the ad as emphasizing the safety features which are important for one's family. On the other hand, an artistic single woman might first fixate on the visual elegance of the car. As a result, viewers might describe the image content in a different order, or even omit elements that are not interesting to them.

We study the relationship between personality, gaze and captioning. For example, we predicted how users will caption an image, conditioned on how they looked images or conditioned on their personality. Similarly, other queries are performed for the remaining combinations of these data modalities. To do this, we learn a joint image-gaze-text-personality embedding space, in which we separately model content and style. We use these embeddings to retrieve content across modalities, in a pool of samples associated with different images and/or annotated by different users. For example, given how a person looked at an image, we learn to predict how that person might caption the image, in contrast to other users' captions on the same or different images.

We collect a cross-modality per-annotator dataset capturing gaze, captions, and personality. Using this data, we find that when retrieving samples for each user across modalities, it is important to model the similarity in the annotations that the user provided. In contrast, methods that only capture similarities in content but not personal style, produce weaker

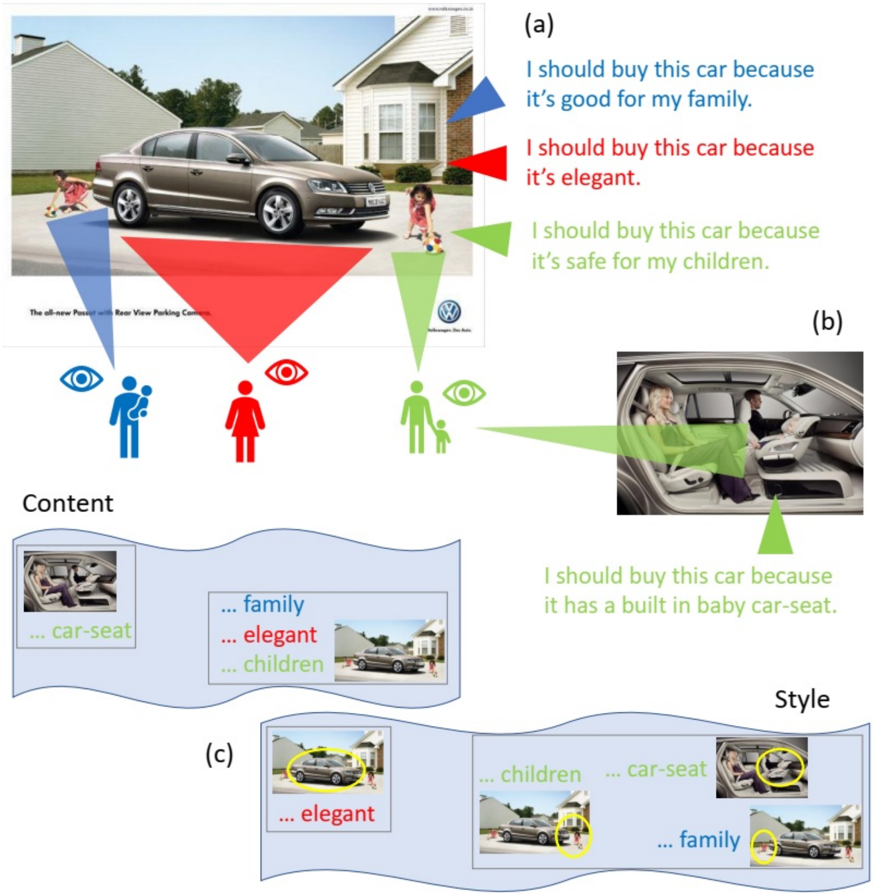

(a)
I should buy this car because it's good for my family.
I should buy this car because it's elegant.
I should buy this car because it's safe for my children.
The all-new Passat with Rear View Parking Camera.
(b)
I should buy this car because it has a built in baby car-seat.
Content
... car-seat
... family
... elegant
... children
Style
(c)
... elegant
... children
... car-seat
... family

Figure 4: People with different personalities might perceive and describe the same image differently. A social, family person might observe the children, and an artistic person might perceive the elegance of the vehicle, in this car advertisement (a). Further, we expect there is consistency between how the same person observes and describes different images (b). To link content across modalities, but preserve differences between how different users might observe and caption the image, we combine both content and style constraints (c). The former encourages samples provided on the same image to be close in a learned space, while the latter encourages samples provided by the same user to be close.

retrieval results. We also compare to a recent personality-aware method which considers single words in the form of tags, and we achieve a stronger result.

1.2.3 Image retrieval via reinforcement learning

Until now, two projects have focused on attribute learning as a core task, and binary predictions and one uses attributes as a side task for cross-modality retrieval. This last project also uses attributes as a side task and combine them with visual data in the form of sketches. Attributes encapsulate rationales via comparison, and sketches encapsulate a visual reasoning via drawings. Specifically, relative attributes are helpful for image search refining [79, 134, 75, 179, 51, 114]. Attributes provide an excellent channel for communication because humans naturally explain the world to each other with adjective-driven descriptions. For example, [75] show how a user can perform rich relevance feedback by specifying how the attributes of a results image should change to better match the user's target image. For example, the user might say "Show me people with longer hair than this one." Another approach has been to engage the user in question-answering with questions that the system estimated are most useful [72, 37]. Thus, in prior work, the initiative for what guidance to give to the system has been taken by either the user [74, 75, 79, 134, 179] or system [37, 146, 72] *but not both.* Previous interactions use attributes, which are useful when concepts can be expressed in language, but some visual concepts are not nameable, so we rely on visual cues in the form of sketches [32, 180, 181, 123] to express them. The system can then retrieve visually similar results. Thus, the user can use either language or visuals to search, but it is not clear which modality is more informative.

We propose a framework where *either* the user or system can drive the interaction, and the input modality can be *either* textual *or* visual, depending on what seems most beneficial at any point in time. Since it is the system that must rank the results, we propose to leave the choice of what is most informative to the system. In other words, the system can decide to let the user lead and *explore*, if it cannot *exploit* any relevant information in a certain iteration. The system can request that the user provides multimodal feedback, i.e. textual *or* visual feedback. To make all these decisions, we train a reinforcement learning agent (see

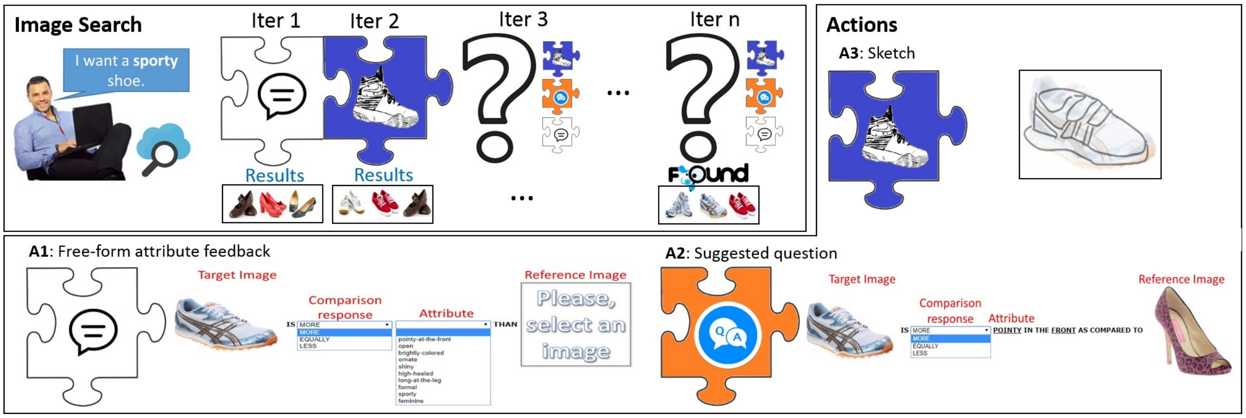


Figure 5: We learn how to intelligently combine different forms of user feedback for interactive image search, and find the user's desired content in fewer iterations. The *image search* section depicts our search agent that predicts an appropriate action at a certain iteration. For example, our agent selects free-form attribute feedback for iteration 1, and sketching for iteration 2. The *actions* section presents the three possible interactions (actions) of our agent.

Fig. 5).

In particular, the options that the reinforcement learning chooses between are (1) sketch feedback, (2) free-form attribute feedback, or (3) system-chosen attribute questions. At each iteration, the system adaptively chooses one of these interactions and asks the user to provide the corresponding type of feedback (e.g. it asks the user to choose an image and attribute to comment on).

Our agent optimizes both the informativeness and exploration capabilities allowing faster image retrieval. We find that our agent prefers human-initiated feedback in former iterations, and complements it with machine-based feedback (i.e. questions) in later iterations. We also outperform standard image retrieval approaches with simulated and real users.

Note that our solutions in this section employ contextual explanations in the form of human rationales. Our first solution improves attribute learning via human gaze maps with an eye-tracking device. Gaze captures subconscious intuition of the meaning of an attribute. Our second solution allows cross-modality retrieval from gaze, captions, and personality questionnaires. Personality is embedded via attributes, and attributes are considered as a side task to improve gaze and caption retrieval. Gaze data was collected in a large scale setup via crowdsourcing and a revealing mask web interface, in contrast to restrictive eye-tracker devices. Contextual explanations are represented via gaze and captions. Captions capture personality information via writing style. Finally, our third solution considers attribute learning as a side task to improve image search. In this proposed solution, we request contextual explanations in the form of sketches and generate human data in the form of user responses. Sketches encapsulate human rationales via drawings, which complement attribute comparison rationale. We create simulated users from previous relative attribute annotations. In addition, our reinforcement agent constantly creates new data as it learns.

1.3 Projects contextualization

All our projects are linked by attribute learning and presented in Table 1. Two of them are centered on attribute learning as a core task, and they employ supervised learning. Also, they center on the problem of binary attribute learning. Our remaining projects complement the current ones using attributes as a side task and employ metric learning and crowdsourcing. Specifically, our image retrieval project uses relative attribute learning and reinforcement learning.

In relation to machine learning paradigms, two of our projects incorporate transfer learning. One is used for attribute transfer learning, and the other to identify different attribute interpretations using gaze. The latter adapts a generic attribute classifier to group-specific attribute interpretations. Also, our *non-semantic attribute transfer learning* and *cross-modality personalization for retrieval* employ multi-task learning. One learns attributes jointly, and the other learns different embedding tasks at the same time. Also, our *image retrieval* and *cross-modality personalization* projects benefits from metric learning. The former uses it to retrieve images from sketches, and the latter to find common embeddings between gaze, captions, and personality. Finally, in relation to selection methods, our *non-semantic attribute transfer learning* aims to select source models using an attention mechanism. Similarly, our projects using *gaze* (P2 and P3) select image subregions.

In relation to data, three of our projects focus on enriched human data (i.e. gaze, writing style, sketches, and/or user simulations). Two projects focus on human sense data. We learn attributes from gaze (sight). Also, the last two projects use crowdsourcing, one to evaluate our system lively with real users, and the other for data collection.

Among these projects, we have explored different challenges such as:

1. How do we integrate human data properly to improve attribute learning?
   a. How to properly collect rationale data?
   b. How to properly represent rationale data?
2. How do we select relevant information in an effective and efficient way?
3. How do we combine different sources of knowledge effectively?
4. How to retrieve data effectively?
5. How to improve subjectivity-aware methods?

The first challenge was studied in three of our projects, and they were inspired by our project in *non-semantic attribute learning.* We ensure data quality via robust data collection interfaces. These interfaces are robust for device miscalibration and lack of participants concentration. We tackle these issues with validation images, which measures if our data is properly collected. Then, we find effective rationale data representations via spatial data

Table 1: Comparison table among all projects in this thesis.

	Core task	Binary attribute learning	Super-vised learning	Side task	Relative attribute learning	Reinforce-ment learning	Transfer learning	Multi-task learning	Selection methods	Human enriched data	Human sense data	Metric learning	Crowd-sourcing
P1: Non-semantic attribute learning	X	X	X				X	X	X				
P2: Learning attributes from gaze	X	X	X				X		X	X	X		
P3: Cross-modality personalization for retrieval				X				X	X	X	X	X	X
P4: Image retrieval with mixed initiative and multimodal feedback				X	X	X				X		X	X

and textual data, which are easy to incorporate in our learning framework. First, spatial data represent gaze saliency maps, which captures subconscious reasoning and encapsulates image query understanding via sketches. Second, textual data capture conscious reasoning in the form of a textual message.

Then, we tackle the second challenge using gaze and an attention mechanism. Gaze follows a human-engineered approach via eye-tracker outperforming gaze-based and data-driven selection methods. In contrast, attention follows a data-driven approach to denote a target attribute as the combination of attributes from unrelated domains.

For our next challenge, we combine effectively different data sources via transfer learning and reinforcement learning. Transfer learning uses an attention mechanism to combine unrelated attributes, while a reinforcement agent combines different actions via a reward function to perform accurate and faster image retrieval. Our transfer learning project encodes a target attribute as a combination of relevant attributes from unrelated domains. Similarly, our reinforcement agent predicts an image retrieval action at an iteration. These actions are combined iteratively among all iterations.

Then, we retrieve data effectively via metric learning and reinforcement learning. Metric learning directly learns a ranking function combining content and style constraints among three different data modalities: gaze, text, and personality. While a reinforcement agent learns a reward function to select an action in a certain iteration and refine an image retrieval ranking function combining textual and visual feedback.

Finally, we improve subjectivity-aware methods adding explainability in the form of gaze, and learning personalized perception combining gaze, writing style, and personality traits. Our former method employs matrix factorization on gaze data in contrast to simple attribute presence annotations. Our later method finds that personality affects gaze and writing style, and learning these three modalities jointly is beneficial.

1.4 Our contributions

The main contributions of this thesis are:

- Discovery of transferable rationale human knowledge components for attribute learning
  - We present a novel attention-guided transfer network to improve attribute learning in scarce and unrelated domains.
  - We show a study of transferability of attributes across semantic boundaries.
- Effective human intervention via contextual explanations in attribute learning for recognition and information retrieval
  - We present a novel method for learning attribute models, using inexpensive but rich data in the form of gaze.
  - We show two applications of how gaze can be used to visualize attribute models, and how it is useful to discover groups of users in terms of their interpretations of attributes.
  - We study the relationship between personality, gaze, and captions allowing cross-modality retrieval.
  - We find that learning about gaze, captions, and personality in the same framework is beneficial than learning them in isolation. Hence, these three modalities provide complementary sources of knowledge.
  - We present a faster mixed-initiative image search retrieval system combining attribute-based methods with sketch retrieval. Both the user and the system are active participants depending on who can provide high-quality search results.
  - We show a study of human-initiated and system-initiated actions in image retrieval.

1.5 Organization

This thesis is organized as follows. Chapter 2 reviews existing solutions for attribute learning, and how our projects solve some of its limitations. It also reviews related work for each of our projects. Chapter 3 shows our current solution to discover meaningful and shareable knowledge with traditional data in a scarcity of data of related attributes. We enrich this representation in chapter 4 using gaze as a meaningful representation. We also complement gaze representation with captions (writing style) and personality questionnaires in chapter 5. Then, chapter 6 investigates how to improve image retrieval using reinforcement learning. Finally, we conclude and present ideas to extend this thesis in Chapter 7.

2.0 Related work and background

In this chapter, we review challenges and different approaches to learn attributes. In section 2.1.7, we also show how we tackle some of these challenges, and how we contribute to them. We also review the most relevant topics for our projects in this thesis. In Section 2.2, we review related topics for our project in *non-semantic attribute transfer* that involves transfer learning, and how to use it for attributes. In section 2.3, our project of *learning attributes from human gaze* reviews attribute localization, learning from humans, and human gaze. In Section 2.4, our project *cross-modality personalization for retrieval* examines image captioning, relationship of gaze and captions, privileged information and style vs content approaches. Finally, in Section 2.5, our project of *image retrieval with mixed initiative and multimodal feedback* reviews image retrieval from attribute-based methods to sketch-based ones, active learning and reinforcement learning.

2.1 Attribute learning

Semantic visual attributes are properties of the visual world, akin to adjectives [81, 36, 10, 110]. Attributes bring recognition closer to human-like intelligence, since they allow generalization in the form of zero-shot learning, i.e. learning to recognize previously unseen categories using a textual attribute-based description and prediction models for these attributes learned on other categories [81, 36, 107, 59, 2]. Attributes have also been shown useful for actively learning object categories [108], scene recognition [110], and action recognition [88]. Attributes are also useful for interactively recognize fine-grained object categories [10, 158], and learn to retrieve images from precise human feedback [79, 75].

Previous works deal with attributes in different situations. Many of these situations have associated challenges, such as:

- Is the attribute vocabulary expressive enough? In order to learn attributes, we define a subset of them in a specific domain. All these attributes are expressive enough to describe all objects in our data. This is especially important in interactive systems [79, 75, 10, 158], where the user provide feedback with attributes. Attributes should allow an effective communication to the search user. Users should not be frustrated interacting with the system.

- How to learn attributes efficiently and confidently? Attributes can be grouped into many categories such as color, shape, texture, parts, and others. Which feature extractor should we use? Should one extractor be used per category? Should we focus on global or local descriptors? In relation to efficiency, should we learn all attributes together or should we learn separately? If we learn them together, some correlations can damage the true meaning of the attribute. For example, *made of metal* can be correlated with *has a wheel* attribute, and our attribute predictor can fail for a *wooden wheel.* These are some of the questions that we need to address depending on our problem setup.
- Attribute accuracy is even more important for attribute applications, where attributes are used as a supportive tool for a more complex task, such as image retrieval or fine-grained object recognition. If we can not trust our attribute models, the applications results are not reliable.
- Do people mostly agree in identifying attributes? Unlike objects, attributes are subjective and human-dependent. For example, if a group of people draws a boot. Most of the drawings are very similar. However, if we ask them to draw a formal shoe. These drawings will have much more variation (subjectivity).

In this chapter, we review attribute learning, and how they deal with these challenges. We assume a fixed attribute vocabulary $A = \{a_m\}, m \in \{1, ..., M\}$, where M is the number of attributes, and a_m is a function that determines if attribute m is present or not.

2.1.1 Multi-task learning approaches

Given that we have a set of attributes associated with an image, it is natural to learn them jointly. [128] employ multi-task learning to learn attributes for crowd scene understanding in videos. Their approach considers a deep network with an appearance and movement branches. Similarly, [39] also employs multi-task learning to recognize 3D shape attributes, however, the authors consider a $\varnothing$ label when they do not have an annotation. Also, they use an embedding loss to ensure that images of the same object are kept together in the feature space. Regularized hypergraphs [52] are also useful for joint learning. Hypergraphs represent instances and can capture correlations of multiple relations (i.e. attributes).

These approaches aim to find correlations among the presence of the attributes. For example, “made of metal” and “has wheels” can be highly correlated. However, these methods can fail to capture the real meaning of “has wheels”, and they do not recognize an object with a wooden wheel. [59] propose to decorrelate attributes using attribute grouping information (e.g. shape, color, texture, parts). They promote feature sharing among attributes from the same group, and feature competition across different groups.

2.1.2 Localization-based approaches

Other approaches claim that localization is a key step for attribute learning. [90] learn face attributes on the web employing a neural network for feature extraction, and linear SVMs for attribute predictions. Their architecture is composed of two localization components and an identity classifier. This classifier receives two images and determines if they belong to the same person. The localization components take care of localizing shoulders and face. Also, [66] learn facial attributes in conjunction with semantic segmentation, because many facial attributes describe local properties. Similarly, [7] employs poselets to localize body parts, and learn attributes.

In the relative attributes’ domain, [166] discover visual concepts that characterize an attribute, in a sequence of relative attribute comparisons. They generate visual chains among these comparisons, and select the most representative region using a ranking SVM. The drawbacks of this approach are that it is time-consuming, and that each step is done in isolation. Hence, optimal solutions are produced for individual steps, but an optimal solution is not produced for the whole problem. [136] tackle these problems using a siamese neural network with a localization and ranking network. The authors employ a localization network to transform the original image into a relevant subregion through translation and scaling operations. Then, image comparisons are learned through a ranking network.

These methods have the limitation that works properly for well-localized attributes. However, they do not provide any benefit for global attributes. For example, they work very well for parts-based attributes. However, they do not provide much benefit for texture and shape attributes, where the attribute lives on most of the whole image. Also, these methods do not consider different attributes interpretation, where attributes can be localized in different regions for each interpretation. In relation to faces, attractiveness is subjective. While some people only consider the eyes, others look for symmetries in the face. Also, these interpretations differ from localized features (i.e. eyes) to holistic ones (i.e. face).

2.1.3 Subjectivity-based approaches

Previous work assumes that there exists only one true annotation per attribute on an image. However, attributes are subjective and are interpreted differently by each user. [71] learn personalized attribute models to account for this issue. First, they learn a generic attribute classifier. Then, they adapt it to specific user annotations. In the same line of thought, [73] claim that user can be grouped in terms of how they interpret an attribute. In other words, users can be grouped in terms of how they respond to questions about the presence or absence of attributes, and how they use the attribute name. Then, a generic classifier is adapted for each group.

Previous approaches only consider one root model (generic classifier). However, this root model can not fit properly to all specific user needs. [80] learn an ensemble of multiple models. These root models are diverse, and are selected to best fit user personalized attributes.

All these approaches show a limited way of communication to learn user knowledge. They only require “what” attributes are in the image, and do not provide any explanation of “why” they are present. Thus, we should involve humans more closely in the learning process.

2.1.4 Category-based approaches

These approaches use information of categories (e.g objects) to improve attribute learning. Attributes are shared among different categories. [81, 82] combines category-specific features to learn attribute presence. For example, zebras, tigers, and bees are useful for the stripped attribute. [161] go a step further, and identify category-dependent and category-independent attribute relations. This knowledge is helpful for attribute and object learning. [52] model category and attribute data using a hyper-regularized graph. Hypergraphs represent instances and capture multiple relations (i.e. attributes). They aim to find a cut in the graph that minimizes the attribute prediction loss, and preserves the clustering in the data (i.e. categories).

Other approaches find an intermediate useful representation. [84] learn attribute models finding latent spaces. Their optimization objective is composed of an object category loss and a multi-task attribute loss. Also, [40] use category labels to create category-invariant features. These invariant features are natural for attributes due to their universality among different categories.

In a transfer learning setup, [17] create an attribute-category table, and infers attribute classifiers for unseen (attribute, category) pairs. They employ tensor completion techniques and category-specific attribute classifiers.

All these works are limited to provided attributes and categories. These data are usually provided by domain experts. However, are they expressive enough? do they cover most properties shared across categories? These questions are answered using a data-driven attribute vocabulary approach. For example, [3] aims to find automatically attribute vocabulary and their associations to categories from large-scale data. They aim to find numerous distinctive attributes shared across categories.

2.1.5 Context-based approaches

Previous section approaches are limited to attributes related to categories. We also can recover valuable knowledge from context [159, 44, 83, 160]. For example, most people wear *formal suits* in a *funeral*.

Contextual knowledge is useful for action recognition and attribute prediction [44]. The authors extend fast-RCNN [43] to find a secondary region, that encapsulates contextual data. Having a bigger bounding box as contextual data is restrictive and disorganized.

Thus, [83] use semantic organized context from human parts and the entire image. They employ the input image in conjunction with regions containing humans, human-parts regions, nearest human-parts, and the whole image in a neural network. In the same line of thought, [160] consider the context for pedestrian attribute recognition. Context is represented as a sequential set of sub images from top to bottom, and inter-person similarity, that consider visual similar images. These components are fed in a joint recurrent learning for attribute prediction. Also, [159] claim that location and weather are contextual information for facial attributes. They collect egocentric videos with location and weather labels.

Context is not restricted to knowledge in the same image. We can borrow meaningful knowledge from attributes in the same domain [17, 48, 89]. [17] use tensor factorization to transfer object-specific attribute classifiers to unseen object-attribute pairs. [48] learn a common feature space through maximum mean discrepancy and multiple kernels. [89] select features from the source and target domains, and transfer knowledge using Adaptive SVM [173] in a lower-dimensional space.

These approaches use complementary and related information in the form of context background and attributes from the same domain. However, they may not be applicable when this information is not available (i.e. images do not present a context background or there are not attributes in the same domain).

2.1.6 Applications

So far, we saw different ways to learn attributes. However, they are also helpful for more complicated tasks such as clothing style recognition [18], image captioning [162, 85], object retrieval [87], video annotation [104], and subjective tasks (i.e. aesthetics [27] and memorability [55, 69]). For example, [18] recognize clothing styles using attributes. They categorize clothing styles from famous people, and event-based clothing styles (e.g. weddings and basketball games). Also, [27] find the most aesthetically beautiful pictures from a search query or a photo album. They employ content, compositional and illumination attributes to recognize aesthetics in pictures. [55] help designers to create more effective memorable visual media. The authors identify the most relevant attributes in memorability. [69] extend the previous study for a large scale setting. They study if popularity, saliency, emotional and aesthetically attributes influences memorability.

Previous applications are appropriate for binary attributes. However, relative attributes can be used to provide useful feedback for image retrieval. [75, 72] ask feedback to the user via relative attribute comparisons. [75] receive feedback in the form “I am looking for a shoe that is more sporty and less pointy than this shoe”. In this setting, the user selects the reference image and the comparison attribute. On the other hand, [72] suggest these data and the user only need to answer with more, less or equal. Finally, [96] extend this approach incorporating confidence and diversity of attribute models to refine the retrieval.

2.1.7 Our work

In this work, we face some of the attribute learning challenges. Our projects focus on learning attributes more confidently using contextual explanations. Two projects enrich data explaining “why” an attribute is present, and three projects use contextual data in the form of non-semantic attributes, gaze, captions, and sketches. Also, one project explores different interpretations of an attribute using gaze. Finally, we deal with attribute learning as a side task for data retrieval tasks. One uses attributes as a complementary task for cross-modal retrieval and the other deals with attribute accuracy for image retrieval. We show the benefits of these projects as follows.

- First, we cope with the lack of explanation in subjectivity-based approaches via enriched data. We use gaze as a source of explanation and bring closer human-computer communication. We also incorporate gaze in [73], and show that gaze is more useful than plain attribute annotations. Also, this approach has time-efficient results comparable to data-driven approaches [166]. We also indirectly study writing styles in combination with gaze to capture different interpretation of personality traits via attributes.
- Second, we complement context-based approaches. Our non-semantic project shows that unrelated domains have valuable knowledge to improve attribute learning when there is no context background or semantically related attributes. Also, our cross-modality project learns together gaze, captions and personality attributes; which are contextual

data sources. Finally, we complement attribute-based image retrieval approaches with sketch-based ones via reinforcement learning. Sketches provide a visual context for attribute textual feedback.

- Third, in the applications domain, we developed two projects for data retrieval using attributes. The former performs cross-modality data retrieval using personality attributes as a side task via metric learning. Then, the later combines attribute-based [75, 72] and sketch-based [180] image retrieval approaches via reinforcement learning. These approaches complement each other, and they are beneficial in different retrieval stages.

Overall these projects, we focus on enhancing data representation for attribute learning with human knowledge and contextual explanations in the form of related/selected attributes, gaze, captions, and sketches.

2.2 Domain adaptation and transfer learning

In order to learn knowledge from non context-based domains, we review topics on transfer learning, and specifically how transfer learning is done for attributes. This related work is relevant for our project on *non-semantic attribute transfer learning.*

2.2.1 Transfer learning

Many researchers perform transfer learning via an invariant feature representation [40, 46], e.g. by ensuring a network cannot distinguish between two domains in the learned feature space [148, 41, 91], training a network that can reconstruct the target domain [42, 67, 9], through layer alignment [20] or shared layers that bridge different data modalities [14]. Other methods [173] perform transfer learning via parameter transfer where the source classifiers regularize the target one. [145] employ an adaptive least-squares SVM to transfer model parameters from source classifiers to a target domain.

2.2.2 Transfer learning for attributes.

We review some transfer learning methods for attributes in attribute learning for context-based approaches. A modern way of transfer learning is zero-shot learning, which aims to transfer knowledge for unseen categories.

Some recent zero-shot learning work [16, 165, 182] learns an underlying embedding space from the *seen* classes and some auxiliary information (e.g. text), and then queries this embedding with a sample belonging to a new *unseen* class, in order to make a prediction. For example, [165] use attributes and text as a class embedding. They also use a non-linear latent embedding to compute projections of image or text features, which are then merged through a Mahalanobis distance. A scoring function is learned which determines if the source domain (class descriptions) and the target domain (test image) belong to the same class.

Similarly, [16] find an intermediate representation for text and images with dictionary learning. [182] use a topic-modeling-based generative model as an intermediate representation. Usually, zero-shot learning is performed to make predictions about object categories, but it can analogously be used to predict a novel target attribute, from a set of known source attributes.

However, prior work only considers objects and attributes from the same domain. Our transfer learning project differs in that we study if transferability of unrelated attributes (from different domains) is more beneficial.

2.3 Localizing attributes, learning from humans and gaze

In order to understand and improve learning, we aim to improve the communication between humans and machines. Thus, we review topics on how to select relevant regions with humans, how to localize attributes, and human gaze. These topics are relevant to our project in *learning attributes from human gaze.*

2.3.1 Localizing attribute models

In the domain of *relative* attributes [107], which we do *not* study, [122] discover parts that improve relative attribute prediction accuracy. It is unclear whether the discovered parts capture the true meaning of attributes as humans perceive them, or simply exploit image features which are *correlated* [59] with the attribute of interest, but are *not* part of the human perception of the attribute.[1] In recent work, [166] propose to discover the spatial extent of relative attributes, as we discussed previously. While we model attributes as binary properties (in contrast to [166]), and use human insight to learn where an attribute lives, [166] is the most related work to ours so we compare to it in Section 4.2.

Other recent work applies deep neural networks to predict attributes [127, 128, 33, 159, 39]. While deep nets can improve the discriminative power of attribute models, they do not exploit human supervision on the meaning or spatial support of attributes. Thus, progress in deep nets is *orthogonal* to the objective of our study. We show that *even when deep features are employed*, using gaze maps to determine the spatial support of attributes improves performance.

2.3.2 Using humans to select relevant regions

[156] pair two humans in an image-based guessing game, where the goal is for the first person to reveal such image regions that allow the second person to most quickly guess the category of the image. The revealed regions are then assumed to be the most relevant for the category of interest. [25, 26] propose a single-player guessing game called "Bubbles," where the player must reveal as few circular regions of an image as possible, in order to match that image to one of two categories with several examples shown. There are three important differences between our work and [156, 25, 26]: (1) These approaches are used to learn objects, not attributes, and attributes have much more ambiguous spatial support; (2) They require that a human should *click* on a relevant image region, which means that the user is consciously aware of what the relevant regions are, whereas in our approach a human uses her potentially subconscious intuition about what makes an image "natural", "formal",

[1] This is also true for attention networks [132, 58] as they are data-driven, not based on human intuition.

or “chubby”; and (3) Clicking or drawing requires a bit more effort (looking is easier than moving one’s hand to use the mouse).

Our method can be seen as a form of annotator rationales [185, 28], which are annotations that humans provide to teach the classifier why a category is present. For example, the user can mark which regions of the face make a person “attractive”. However, providing gaze maps by looking is much faster than drawing rationales (see Section 4.1.2).

2.3.3 Gaze and saliency

[106] use human gaze to reduce the effort required in obtaining data for object detectors. They build bounding boxes from locations in a photo where a user fixates when judging which of two categories is portrayed in the image. [184] argue that using gaze can improve object detection—bounding box predictions that do not align with fixations can be pruned. They also use a gaze-based feature to classify detections into true and false positives, but only show small gains in detection accuracy.

In addition to gaze, saliency examines where a viewer will fixate in an image [57, 117, 101, 50, 63, 45, 62, 53]. We use [63]’s method to predict gaze maps for novel images. No prior work uses gaze to learn attribute models.

2.4 Cross-modality personalization for retrieval

In order to use attributes (personality) for gaze and caption retrieval, we review topics involving image captioning and gaze. We also focus on style and content approaches, as they are a key component of our approach. Finally, we review privileged information as we are learning different data modalities at the same time.

2.4.1 Image captioning

There is a large body of work [4, 118, 153, 178, 155, 68, 29] on automatic image captioning, or predicting a description for a given visual. Common approaches include learning a joint image-text embedding using triplet loss or by maximizing the correlation of the two modalities [34, 31]; training a recurrent network that predicts a sequence of words conditioned on the image and outputs at previous timesteps [155, 68, 29]; learning a template description and how to fill each position of the template with a word [93]; generative adversarial approaches [24]; etc.

Most captioning approaches assume all users would caption an image in the same way. In contrast, [22] learn individual differences in how an annotator describes an image, and [152] learn the types of hashtags a user might provide. However, none of these consider two manifestations or channels of personality as we do (i.e. gaze and captions). We show that having information from multiple modalities at training time allows us to better understand user differences.

2.4.2 Gaze

Saliency prediction work [57, 63, 62, 100] models what humans find fixate on in an image. Prior work has examined the relationship between sentiment and gaze [35] and the differences between viewers in how they look at an image [172], but none has examined the relationship between personalized *perception* and personalized *meaning*.

2.4.3 Relationship of captions and gaze

A few authors have examined the relationship between captions and attention. For example, [183, 142, 171, 92] predict captions conditioned on an attention map (learned from human gaze or discovered from a classification loss). However, these do not consider personalized captioning or gaze as we do.

2.4.4 Style vs content

In our work, we aim to separate similarities arising due to content (i.e. image and corresponding text should be close in our learned space) and similarities due to style (annotations produced by the same viewer should be close by). Prior work exists that separates content and style for different tasks. [186] separate content and style for handwritten Chinese characters, by training separate networks for each, and [143] use a model linear in both the content (character ID) and handwriting style. [46, 148, 41, 42, 91, 8] learn domain-invariant representations for object recognition, where objects are the "content" and modalities (e.g. paintings, sketches) are the "style". We have *multiple* content modalities, and multiple styles (one per user). Also relevant is [72] which train per-user attribute models, but this work only considers one modality.

2.4.5 Privileged information

Our approach utilizes a type of "privileged" feature information, which is available at training time only. Such information is useful to learn the structure of the space, and then utilize it at test time with only a subset of the input types. Prior work includes [150, 130, 131, 49, 97, 6]. For example, [130] use privileged information to learn which samples are easy to learn from, and [6] regularize the parameters of one network with another learned from privileged data. In contrast, we use privileged information for caption retrieval.

2.5 Image retrieval, active learning, and reinforcement learning

In order to combine different attribute-based image retrieval techniques with visual approaches, we review topics in image retrieval, active learning and reinforcement learning. Image retrieval focuses on topics about attribute-based search, sketch-based search, and interactive image retrieval. These topics are relevant to our project in *image retrieval with mixed initiative and multimodal feedback*.

2.5.1 Attribute-based search.

Prior work has explored the value of the fine-grained detail that attribute descriptions provide, by using attributes to initiate a search [134, 151] or provide iterative feedback on the results of a search system [75, 72, 96]. [74] browses the current search results, and can then provide a feedback statement of the form "The image I am looking for is more/less [attribute] than [this image in the results]." The choice of an attribute on which to comment is left to the user. This is helpful if the user is perceptive, or there are images which obviously differ from the user's desired content for particular attributes. On the other hand, browsing a set of images and choosing attributes is time-consuming for the user, as we find in experiments. [72] shows that given a limited budget of interactions that the user is willing to perform, more accurate search results can be achieved if the system asks the user questions of the form "Is the image you are looking for more/less/equally [attribute] than [this image]?" The chosen questions are those with high information gain. The disadvantage of [72] is that it limits the ability of the user to browse and explore the dataset space.

2.5.2 Sketch-based search.

While attribute-based feedback is appropriate when the user can concisely describe what content they wish to find using words, some searches involve concepts which are purely visual. In our setting, we assume the user does not have a photograph of what they wish to find, so cannot directly do a similarity-based search with a query image. However, the user does have a clear visual idea of what content they wish to find. Sketch-based search approaches allow the user to convey this visual idea to the system, via a sketch or drawing, which provides a complementary way of communication. The system can then extract features from this sketch and compare to the features of the images in a database [32, 133, 180, 123, 181].

We use a similar approach, but also propose to convert the sketch to an image using generative models. Other authors use generative learning to find a representation appropriate for cross-domain (sketch-to-image [105, 138, 137] or text-to-image [138]) search. We use sketch-based retrieval in a larger reinforcement learning framework that chooses which search interaction to propose (sketch, attribute-based feedback, or question-answering). Note that our focus is *not* in how we perform sketch-based retrieval, but rather *how to decide when* to request a sketch.

2.5.3 Interactive search.

Rather than ask the user to issue a query and return a single set of results, we engage the user in providing interactive relevance feedback and show results after each round. This is a popular idea [121, 187, 23, 38, 37] whose key benefit is that incorrect predictions by the system can be corrected. We also adopt interactive search, but combine the advantages of free-form feedback and exploration with the information-theoretic benefits of actively querying for feedback [37], via reinforcement learning.

2.5.4 Active learning.

In order to minimize the cost of data labeling, active learning approaches estimate the potential benefit of labeling any particular image, using cues such as entropy, uncertainty reduction, and model disagreement [146, 126, 47, 154, 64]. [163, 12, 139, 30] have explored mixed initiative between user and system as well as reinforcement learning, for improving active learning at training time, in contexts other than image search. In contrast, we use reinforcement learning to select interactions at test time (during an online search).

2.5.5 Reinforcement learning

[65, 95, 149] has recently gained popularity for a variety of computer vision tasks, e.g. object [11, 94] and action detection [176]. The most related work to ours is [177] which also uses reinforcement learning to choose the type of feedback method for requesting feedback from the user. This approach considers query vector modification, feature relevance estimation, and Bayesian inference, as three possible feedback mechanisms. Neither of these allows the user to *comparatively* describe how the results should change (via attributes); instead, each image property is defined as desirable/undesirable. [75] show such binary feedback is inferior to comparative attribute feedback. Further, unlike [177], we consider both visual and textual feedback among the mechanisms presented to our users.

3.0 Asking Friendly Strangers: Non-Semantic Attribute Transfer

Recent advances in computer vision rely on huge amount of data. This happens mainly of the unpredictable success of deep learning. This is not different for attribute learning approaches [128, 39, 159], where traditional data is represented as (image, labels) pairs. Labels represent binary attributes present in their respective image. However, what can we do if we have a limited amount of data? A common solution is transfer learning.

As we discussed before, transfer learning aims to transfer knowledge from a source task with many data to a related target task with limited data. This approach is similar to attributes, traditional attribute transfer learning aims to transfer knowledge between attributes from the same domain (Section 2.2.2). However, what can we do if we have data scarcity and no semantic related categories? In this work, we propose one solution to perform *non-semantic attribute transfer learning.*

This non-semantic approach aims to select valuable knowledge from unrelated data. Data is represented by a traditional feature matrix. Hence, we go from traditional to more complex data for learning attributes. We enrich data in each new chapter.

We test our method on 272 attributes from five datasets of objects, animals, scenes, shoes, and textures, and compare it with several baselines: learning using data from the target attribute only, transfer only from attributes in the same domain, uniform weighting of the source classifiers, learning an invariant representation through a confusion loss, and a fine-tuning approach. We also show qualitative results in the form of attention weights, which indicate what kind of information different target attributes borrowed.

While our target attributes come from well-defined and properly annotated datasets, our work demonstrates how non-semantic transfer can be used to learn attributes on novel domains where data is scarce. Our main contributions are an attention-guided transfer network, and a study of transferability of attributes across semantic boundaries. This project was published in [98].

The remainder of this chapter is organized as follows. In section 3.1, we describe our attention-guided transfer approach for non-semantic attribute transfer, including our network formulation, optimization and implementation details. In Section 3.2, we show that our method improves upon standard transfer learning approaches via quantitative experiments. We also show a transferability study across semantic categories. Finally, we summarize this chapter in Section 3.3.

3.1 Approach

We briefly give an overview of our approach, and how we formulate it on a neural network architecture. We also provide details about its optimization losses, and implementation details (e.g. model selection, frameworks).

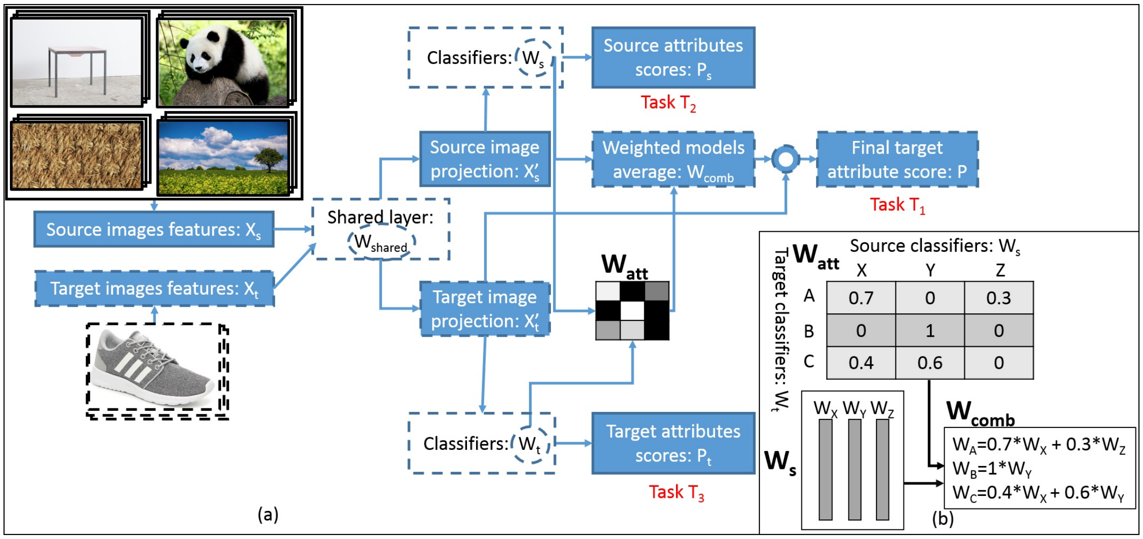
3.1.1 Overview

We first overview our multi-task attention network, illustrated in Fig. 6. Then, we give more details on its formulation, optimization procedure, and implementation.

An attention architecture allows us to select relevant information and discard irrelevant information. Attention has been used for tasks such as image segmentation [19], saliency detection [77], image captioning [178] and image question answering [169, 132, 174]. The latter use an attention mechanism to decide which regions in an image are relevant to a question input. In our problem scenario, we are not concerned with image regions, but we want to select source attribute models useful for predicting some particular target attribute.

We are interested in selecting relevant source models for our target attributes (e.g. "sporty"). For example, the network might determine attributes X and Z are useful for predicting target attribute A, but attribute Y is not (Fig. 6 (b)). The learned attention weights would reflect the predicted usefulness of the source attributes for the target task.

Our network contains source and target input branches, as depicted in Fig. 6. Similarly to [132, 174], we extract $fc7$ features from AlexNet for source and target images. These target (X_t) and source (X_s) visual features are embedded into a common space using a projection matrix W_{shared}, resulting in embedded features X_t' and X_s'. This common space


Classifiers: W_s
Source attributes scores: P_s
Task T_2
Source image projection: X'_s
Weighted models average: W_{comb}
Final target attribute score: P
Task T_1
Source images features: X_s
Shared layer: W_{shared}
Target images features: X_t
Target image projection: X'_t
W_{att}
Classifiers: W_t
Target attributes scores: P_t
Task T_3
(a)
W_{att}
Source classifiers: W_s
Target classifiers: W_t
W_s
W_X W_Y W_Z
W_{comb}
(b)

	X	Y	Z
A	0.7	0	0.3
B	0	1	0
C	0.4	0.6	0

$W_A=0.7*W_X + 0.3*W_Z$
$W_B=1*W_Y$
$W_C=0.4*W_X + 0.6*W_Y$

Figure 6: (a) Overview of our transfer attention network, using an example where the target attributes are from the shoes domain, and the source attributes are from the objects, scenes, animals and textures domains. Source and target images are projected through a shared layer. Then, target and source attribute models W_t and W_s are learned. An attention module selects how to weigh the available source classifiers, in order to produce a correct target attribute prediction. At test time, we only use the dashed-line modules. ⬚ denotes layers, and ◌ represents their parameters. (b) Example of how source models W_s are combined into the final target attribute classifiers W_{comb}, using as coefficients the attention weights W_{att}.

is required to find helpful features that bridge source and target attributes. Then we learn a set of weights (classifiers) W_t and W_s which we multiply by X'_t and X'_s, to obtain attribute presence/absence scores P_t and P_s for the target and source attributes, respectively.

In order to transfer knowledge between the target and source attribute classifiers, we calculate normalized similarities W_{att} between the classifiers W_t and W_s. We refer to W_{att} as the attention weights learned in our network. We then use W_{att} as coefficients to compute a linear combination of the source classifiers W_s. By doing so, we select the most relevant source classifiers related to our target attributes. We call this resulting combined classifier W_{comb}. Finally, we compute the product of W_{comb} with the target features X'_t, to produce the final attribute presence/absence scores for the target attributes.

At training time, our network requires source and target images to find helpful knowledge to our target task. However, once the relationship between source and target attributes is captured in W_{att}, we no longer need the source images. In Fig. 6, we denote modules that are used at test time with dashed boundaries. Layers are denoted with ⬚, and ◌ represents their parameters.

3.1.2 Network formulation

Our network receives target (X_t) and source (X_s) visual features. We process all source and target attributes jointly, i.e. we input training image features for all attributes at the same time. These are embedded in a new common feature space:

$$X'_t = X_t W_{shared} + 1b \qquad X'_s = X_s W_{shared} + 1b \tag{3.1}$$

where $X_t \in R^{NxD}$, $X_s \in R^{NxD}$ are the features, $W_{shared} \in R^{DxM}$ contains the shared embedding weights, $1 \in R^{Nx1}$ is a vector of ones, $b \in R^{1xM}$ is the bias term, N is the batch size, D is the number of input features, and M is the number of features of the embedding.

During backprop training, we learn target and source models W_t and W_s. Note that the target model is only used to compute its similarity to the source models, and will be replaced by a combination of source models in a later stage. We then compute P_t and P_s, which denote the probability of attribute presence/absence for the target and source attributes, respectively. These are only used so we can compute a loss during backprop (described below).

$$P_t = f(X'_t W_t) \qquad P_s = f(X'_s W_s) \tag{3.2}$$

where $W_t \in R^{MxK}$, $W_s \in R^{MxL}$ are learned model weights, f is a sigmoid function (used since we want to compute probabilities), L is the number of source attributes, and K the number of target attributes. We found it is useful to ensure unit-norm per column on W_t and W_s.

Attention weights W_{att} are calculated measuring the similarity between source classifiers W_s and target classifiers W_t. Then, a normalization procedure is applied.

$$O_{att_{i,j}} = \frac{W_{t_i}^T \cdot W_{s_j}^T}{||W_{t_i}^T||\ ||W_{s_j}^T||} \qquad W_{att_i} = \frac{[g(O_{att_{i,1}}), ..., g(O_{att_{i,L}})]}{\sum_{j=1}^{L} g(O_{att_{i,j}})} \tag{3.3}$$

where $W_{t_i}^T$ and $W_{s_j}^T$ are columns from W_t and W_s, $O_{att} \in R^{KxL}$, $W_{att} \in R^{KxL}$, g is a RELU function, $O_{att_{i,j}}$ is the similarity between target attribute i and source attribute j, and W_{att_i} are the attention weights for a single target attribute. We use cosine similarity in Eq. 3.3 to ensure distances are in the range [-1, 1].

When computing attention weights, we want to ensure we do not transfer information from classifiers that are inversely correlated with our target classifier of interest. Thus, we employ normalization over a RELU function (g in Eq. 3.3) and transfer information from classifiers positively correlated with the target classifier, but discard classifiers that are negatively correlated with it (negative similarities are mapped to a 0 weight).

Finally, a weighted combination of source models is created, and multiplied with the target image features X_t' to generate our final predictions for the target attributes:

$$W_{comb} = W_{att} W_s^T \qquad\qquad P = f(X_t' W_{comb}^T) \tag{3.4}$$

where $W_{comb} \in R^{KxM}$ is the weighted combination of sources, and f is a sigmoid function.

Note our model is simple to train as it only requires the learning of three sets of parameters, W_{shared}, W_s and W_t.

3.1.3 Optimization

Our network performs three tasks. The main task T_1 predicts target attributes using attention-guided transfer, and side tasks T_2 and T_3 predict source and target attributes, respectively. Each task T_i is associated with a loss L_i. Our optimization loss is defined as

$$L = \lambda_1 * L_1 + \lambda_2 * L_2 + \lambda_3 * L_3 \tag{3.5}$$

where $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$.[1] Since an image can posses more than one attribute, our predictions are multi-label and we employ binary cross-entropy loss for all L_i.

For task T_2, our source image branch contains attributes from different domains. Thus an image has annotations for attributes in its domain, but not for other domains. We solve this issue with a customized cross-entropy loss [39]. Suppose you have N samples and L

[1] The loss weights were selected similar to other transfer learning work [148] where the main task has a weight of 1, and side tasks have a weight of 0.1.

attributes. Each attribute is annotated with 0, 1 or $\varnothing$, where $\varnothing$ denotes no annotation. The customized loss is:

$$L(Y,P) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ Y_{i,j} \neq \varnothing}}^{L} Y_{i,j} log(P_{i,j}) + (1 - Y_{i,j}) log(1 - P_{i,j}) \tag{3.6}$$

where i is an image, j is an attribute label, $Y_{i,j} \in \{0, 1, \varnothing\}^{N,L}$ is the ground-truth attribute label matrix and $P_{i,j} \in [0,1]^{N,L}$ is the prediction probability for image i and attribute j. The constraint $Y_{i,j} \neq \varnothing$ means attribute annotations $\varnothing$ have no effect on the loss.

3.1.4 Implementation

We implemented the described network using the Theano [144] and Keras [21] frameworks and [136]'s attention network. First, we did parameter exploration using 70 random configurations of learning rate and L_2 regularizer weight. Each configuration ran for five epochs with the ADAM optimizer. Then the configuration with the highest accuracy on a validation set was selected and a network with this configuration ran for 150 epochs. In the end of each epoch, the network was evaluated on a validation set, and training was stopped when the validation accuracy began to decrease. Finally, note that we have fewer target images than source images, so the target images were sampled more times.

3.2 Experimental validation

We compare three types of source data for attribute transfer, i.e. three types of data that can be passed in the source branch of Fig. 6. This data can correspond to attributes from the same domain, from a disjoint domain, or from any domain. The first option corresponds to the standard manner of performing semantic (within-domain) attribute transfer [17, 48, 89]. The latter two options represent our *non-semantic transfer* approach.

To evaluate the benefit of transfer, we also compare to a method that learns target attributes from scratch with no source data, and two standard transfer learning approaches

[148, 103]. We do not directly compare to attribute transfer methods [17, 48, 89] as they do not use neural nets and the comparison would not be fair.

We evaluated our method and the baselines on five domains and 272 attributes. We observe that by transferring from disjoint domains or from any domain, i.e. by performing *non-semantic transfer* without the requirement for a semantic relationship between the source and target tasks, we achieve the best results. To better understand the transfer process, we also show attention weights and determine the most relevant source domains per target domain/attribute.

3.2.1 Datasets

We use five datasets: Animals with Attributes [81], aPascal/aYahoo Objects [36], SUN Scenes [111], Shoes [75], and Textures [13]. The number of attributes is 85, 64, 102, 10 and 11, respectively.

For each dataset, we split the data in 40% for training the source models, 10% for training the target models, 10% for selection of the optimal network parameters, and 40% to test the final trained network on the target data. The complexity of the experimental setup is to ensure fair testing. For transfer learning among different domains (Attention-DD and Attention-AD below), we can increase the size of our source data split to the full dataset, but for a fair comparison, we use the same split as for the Attention-SD setup.

Our splits mimic the scenario where we have plentiful data from the source attributes, but limited data for the attribute of interest.

3.2.2 Baselines

Let D_i represent a domain and its attributes, and $D = \bigcup_{i=1}^{5} D_i$ be the union of all domains. We compare seven methods. The first are two ways of performing non-semantic transfer:

- Attention-DD, which is our multitask attention network with D_i as our target domain and $D \backslash D_i$ as our source domains. We train five networks, one for each configuration of target/source.

- Attention-AD, which is our multitask attention network with D_i as our target domain and D as our source domains. We again train one network for each target domain. Some attributes on the source and target branches overlap, so we assign 0 values along the diagonal of W_{att} to avoid transfer between an attribute and itself.

We compare our methods against the following baselines:

- Attention-SD, which uses the same multitask attention network but applies it on attributes from only a single domain D_i, for both the source and target branches. We again train five networks, and assign values of 0 along the diagonal of W_{att}. Note that even though some form of transfer is already taking place between all target attributes due to the multi-task loss, the explicit transfer from the source domains is more effective because we have more training data for the sources than the targets.
- Target-Only, which uses the predictions P_t as the final predictions of the network, without any transfer from the source models.
- A replacement of the attention weights W_{att} with uniform weights, i.e. combining all source classifiers with the same importance for all targets. This results in baselines Attention-SDU, Attention-DDU and Attention-ADU.
- [148] which learns feature representations X'_s, X'_t invariant across domains, using domain classifier and confusion losses but no attention. This results in baselines Confusion-DD and Confusion-AD.
- Approaches Finetune-DD and Finetune-AD that fine-tune an AlexNet network using source data, then fine-tune those source networks again for the target domain. This method represents "standard" transfer learning for neural networks [103].

We found that Attention-SD is a weak baseline. Thus, we replace it by an ensemble of target-only with Attention-SD. This ensemble averages the probability outputs of these two models. We try a similar procedure for Attention-DD and Attention-AD, but it weakens their performance, so we use these methods in their original form.

3.2.3 Quantitative results

Tables 2 and 3 contain show average accuracy and F-measure, respectively. We show both per-domain and across-domains overall averages. We include F-measure because many attributes have imbalanced positive/negative data.

In both tables, we see that our methods Attention-DD and Attention-AD outperform or perform similarly to the baselines in terms of the overall average. While the strongest baselines Confusion-DD and Confusion-AD [148] perform similarly to our methods for accuracy, our methods have much stronger F-measure (Table 3). Accuracies in Table 2 seem misleadingly high because attribute annotations are imbalanced in terms of positives/negatives and a baseline that predicts all negatives will do well. Thus, the differences between the methods are larger than they seem.

Table 2: Method comparison using accuracy. Our Attention-DD and Attention-AD outperform or perform equal to the other methods on average. Best results are **bolded** per row.

	target -only	Attention -SDU	Attention -DDU	Attention -ADU	Attention -SD	Attention -DD (ours)	Attention -AD (ours)	Confusion -DD	Confusion -AD	Finetune -DD	Finetune -AD
avg animals	0.90	0.63	0.63	0.73	0.90	0.91	0.91	0.91	0.91	0.88	**0.92**
avg objects	0.92	0.89	0.89	0.89	0.92	**0.93**	**0.93**	**0.93**	**0.93**	0.91	0.92
avg scenes	**0.95**	0.93	0.93	0.93	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**
avg shoes	0.88	0.70	0.71	0.79	0.89	0.90	0.90	0.90	0.90	0.75	**0.92**
avg textures	0.91	0.87	0.91	0.91	0.95	**0.99**	**0.99**	**0.99**	**0.99**	0.91	0.91
avg overall	0.91	0.80	0.81	0.85	0.92	**0.94**	**0.94**	**0.94**	**0.94**	0.88	0.92

It is important to highlight the success of Attention-DD as it does not use any attributes from the target domain, as opposed to Attention-AD. In other words, transfer is more successful when we allow information to be transferred even from domains that are semantically unrelated to the target. In addition, note that the uniform weight baselines (Attention-SDU, Attention-DDU and Attention-ADU) are quite weak. This shows that only by selecting the source classifiers intelligently, we can perform transfer learning correctly. We see many 0 F-measure scores for Attention-SDU, Attention-DDU and Attention-ADU because they have a bias to predict negative labels.

Table 3: Method comparison using F-measure. Our approaches Attention-DD and Attention-AD outperform the other methods on average. Best results are **bolded** per row.

	target -only	Attention -SDU	Attention -DDU	Attention -ADU	Attention -SD	Attention -DD (ours)	Attention -AD (ours)	Confusion -DD	Confusion -AD	Finetune -DD	Finetune -AD
avg animals	0.81	0.00	0.00	0.27	0.82	0.82	**0.83**	0.82	0.82	0.69	0.79
avg objects	**0.50**	0.00	0.00	0.01	**0.50**	0.47	0.47	0.39	0.41	0.10	0.14
avg scenes	**0.28**	0.00	0.00	0.00	0.27	0.25	0.26	0.17	0.15	0.04	0.04
avg shoes	0.81	0.27	0.38	0.59	0.83	0.83	0.84	0.83	0.83	0.37	**0.87**
avg textures	0.68	0.09	0.00	0.00	0.78	**0.96**	**0.96**	0.95	0.95	0.06	0.09
avg overall	0.62	0.07	0.08	0.17	0.64	**0.67**	**0.67**	0.63	0.63	0.25	0.39

While Finetune-AD outperforms our methods for two domains in Table 2, it is weaker in terms of the overall average, and weaker in four out of five domains in Table 3.

Finally, the attention transfer methods with learned attention weights usually outperform Target-Only, which emphasizes the benefit of transfer learning. Our non-semantic transfer methods bring the largest gains.

We believe the success of our attention network is due to the combination of transfer learning via a common feature representation, and parameter transfer. The common feature representation is achieved via our shared layer, and the parameter transfer is performed via our attention-guided transfer. Finally, we believe that instance weighting also helps: this is accomplished by our choice to sample more target images than source images.

3.2.4 Qualitative results

In order to analyze the internal behavior of Attention-DD and Attention-AD, we extract and show the attention weights W_{att}. Hence, for each target classifier i, we extract the weights $W_{att_i} = (w_1, w_2, ..., w_L)$ for the source classifiers. This procedure also verifies if Attention-AD is primarily using transfer from attributes in the same domain, or attributes from disjoint domains with respect to the target. Due to the large number of attributes, we group attributes by their domain. Rows represent targets, and columns sources.

In Table 4 corresponding to Attention-DD, the attention weights over the source classifiers are distributed among animals, objects, and scenes. We believe that shoe attributes

Table 4: Attention weights summed per domain for our Attention-DD approach. Rows vs columns represent target vs source classifiers. The most relevant domains are **bolded** per row. – denotes Attention-DD does not transfer from attributes in the same domain.

tgt/src	animals	objects	scenes	shoes	textures
animals	-	0.29	**0.56**	0.06	0.09
objects	**0.48**	-	0.44	0.04	0.04
scenes	**0.59**	0.28	-	0.07	0.06
shoes	0.19	0.35	**0.38**	-	0.08
textures	0.33	0.19	**0.44**	0.04	-

Table 5: Attention weights summed per domain for our Attention-AD approach.

tgt/src	animals	objects	scenes	shoes	textures
animals	**0.43**	0.09	0.39	0.02	0.07
objects	0.26	0.21	**0.41**	0.04	0.08
scenes	0.36	0.19	**0.39**	0.02	0.04
shoes	0.10	0.30	**0.50**	0.00	0.10
textures	0.36	0.16	**0.39**	0.03	0.06

are not very helpful for other domains because shoe images only contain one object. Further, textures are likely not very helpful because they are a low-level representation mainly defined by edges. Interestingly, we observe that the most relevant domain for animals, shoes, and textures is scenes, and scenes is *not closely related to any of these domains*. Similarly, the most meaningful domain for objects and scenes is animals, another *semantically unrelated source domain*.

In Table 5, showing results when we perform transfer from *any* domain, we observe that shoes and textures attributes do not benefit almost at all from other attributes in the same domain. On the other hand, objects, scenes, animals do benefit from semantically related attributes, but the overall within-domain model similarity is lower than 50%, again reaffirming our choice to allow non-semantic transfer.

Finally, we illustrate what visual information is being transferred across domains. In

Table 6: Interesting selected source attributes from domains disjoint from the target domain.

domain	target attribute	some relevant source attributes from [domain]
textures	aluminium	muscular [animal], made of glass [object]
	linen	handlebars [object], railroad [scene]
	lettuce leaf	lives in forest [animal]
shoes	pointy	foliage [scene]
	bright-in-color	vegetation [scene], shrubbery [scene]
	long-on-the-leg	has leg [object]
object	has stem	dirty soil [scene], feed from fields [animal]
	vegetation	dirty soil [scene]
animal	tough-skinned	stressful [scene]
	fast	scary [scene]
	hunter	studying [scene]
scene	railroad	solitary [animal]
	shrubbery	tough-skinned [animal]

Table 6, we show relevant source attributes for several target attributes. The “aluminium” texture presents a “muscular” structure, and a color similar to “glass”. The “linen” texture has edges similar to “handlebars” and “railroads”. “Lettuce leaf” shows leaves’ textures, so “forest” animals (which might co-occur with leaves) are helpful. For shoes attributes, “foliage” is a set of “pointy” leaves, “vegetation” and “shrubbery” are “bright-in-color”, and “leg” is related to shoes that are “long-on-the-leg”. For object attributes, “vegetation” and objects with a “stem” grow on “dirty soil” and animals might “feed” on them. For animal attributes, “tough skin” gives us the feeling of a “stressful” situation, “fast” animals might “scare” people, and “hunter” animals “study” the best situation to catch their prey. Finally, “railroad” scenes might be “solitary” places, and “shrubbery” is rough like “tough-skinned” animals. In other words, while source attributes are selected from disjoint domains, it is possible to explain some selections, but note that many do not have an intuitive explanation. The latter is indeed what we expect when we perform non-semantic transfer.

3.3 Summary

We have explored the problem of attribute transfer learning using unrelated domains. We develop an approach that transfers knowledge in a common feature space, by performing parameter transfer from source models. Our attention mechanism intelligently weights source attribute models to improve performance on target attributes. We find that attributes from a different domain than the target attributes are quite beneficial for transfer learning, and improve accuracy more than transfer from semantically related attributes. We also outperform standard transfer learning approaches.

In this project, we *discover* contextual explanations by identifying human transferable knowledge. Specifically, we select models via an attention mechanism. In our next project, we extend this idea by *requesting* contextual explanations. We *request* gaze to select meaningful features.

One drawback of this project is that we do not study different attribute interpretations. Attributes are ambiguous, and they are understood in different ways by different people. In our next project, we solve this issue by capturing different attribute meanings using an eye-tracking device. Our main method consists of grouping similar gaze patterns, and learn specific classifiers. Also, we develop an application to group users in terms of their judgments for attribute presence.

Similarly to this current project, our application uses transfer learning. It adapts a generic attribute model to a group of users with the same understanding of attribute meaning. Hence, we complement the work in this chapter using transfer learning for attribute interpretation.

4.0 Learning Attributes from Human Gaze

This chapter focuses on attribute learning with contextual explanations. As discussed before, localization methods are mainly data-driven and applied in relative attributes (Section 2.3.1). In contrast, we specifically focus on how to involve humans more closely in the process of learning binary attributes via gaze. Gaze posses discriminative power to help learn different attribute interpretations, because it gives an explanation of "why" an attribute is present.

As we mentioned in the introduction, our gaze approach achieves competitive performance compared to other feature selection approaches. We first show success on shoes and faces datasets. Then, we adapt our method for more complicated datasets (i.e. scenes, that have more than one object). Finally, we show how gaze can be used to improve attribute visualization, and grouping users based on their judgments of attribute presence.

The main contribution of our work is a new method for learning attribute models, using inexpensive but rich data in the form of gaze. We show that our method successfully discovers the spatial support of attributes. Despite the close connection between attributes and human communication, gaze has never been used to learn attribute models before. This project was published in [100].

The remainder of this chapter is organized as follows. In Section 4.1, we describe our approach for learning attributes from human gaze, including how we collect gaze data, generate gaze maps, extract features from these maps and train attribute prediction models. In Section 4.2, we show that our method improves upon the standard method for learning attributes and alternative methods for selecting relevant regions, using a number of features, including ones extracted from a convolutional neural network. We also show several other applications of our method, including how gaze can be used to generate intuitive visualizations of attribute models, and to discover better groupings between users in terms of their interpretation of attributes [73]. Finally, we summarize this chapter in Section 4.3.

4.1 Approach

We first describe our datasets (Section 4.1.1) and how we collect gaze data from human subjects (Section 4.1.2). In Section 4.1.3, we discuss how we compute one or multiple gaze templates per attribute, and in Section 4.1.4, we describe how we use the templates to restrict the range of an image from which an attribute model is learned. Finally, in Section 4.1.5, we show how we predict an individual gaze template for each test image.

Like [166], our method is designed for images which contain a single object, specifically faces and shoes. See Section 4.2.2 for a preliminary adaptation of our work for scenes.

4.1.1 Datasets

We use two attribute datasets: the **Faces** dataset of [78] (also known as PubFig), and the **Shoes** dataset of [74]. All images are of the same square size (200x200 pixels for faces and 280x280 for shoes). The attributes we use are: for **Shoes**, “feminine”, “formal”, “open”, “pointy”, and “sporty”; and for **Faces**, “Asian”, “attractive”, “baby-faced”, “big-nosed”, “chubby”, “Indian”, “masculine”, and “youthful”. Like [166], we consider a subset of all attributes, in order to focus the analysis towards attributes whose spatial support does not seem obvious, i.e. it could not be predicted from the attribute name alone. This allows insight into the meaning of some particularly ambiguous attributes (e.g. “formal”, “feminine” and “attractive”). We also selected some attributes (“pointy” and “big-nosed”) where we had a fairly confident estimate of where gaze locations would be. This allows us to qualitatively evaluate the collected gaze maps via their alignment with the expected gaze locations. The annotation cost per attribute is small, about 1 minute per image-attribute pair (see below).

We select 60 images total per attribute. In order to get representative examples of each attribute, we sample: (a) 30 instances where the attribute is definitely present, (b) 18 instances where it is definitely not present, and (c) 12 instances where it may or may not be present. For **Faces**, we use the provided SVM decision values to select images in these three categories. For **Shoes**, we use the ordering of ten shoe categories from most to least having each attribute, which we map to individual images using their class labels.

4.1.2 Gaze data collection

We employ a \$495 GazePoint GP3 eye-tracker device[1] to collect gaze data from 14 participants. The 320x45x40mm eye-tracker is placed in front of a monitor, and the participants do *not* have to wear it, in contrast to older devices. Gaze data can also be collected via a webcam; see [170].

Our experiment begins with a screening phase in which we show ten images to each participant and ask him/her to look at a fixed region in the image that is marked by a red square, or to look at e.g. the nose or right eye for faces. If the fixated pixel locations lie within the marked region, the participant moves on to the data collection session. The latter consists of 200 images organized in four sub-sessions. In order to increase the participants' performance, we allow a five-minute break between sub-sessions. We ask the viewer whether a particular attribute is present in a particular image which we then show him/her. The participant has two seconds to look at the image and answer. His/her gaze locations and answers are recorded. We obtain 2.5 gaze maps on average, for each image-attribute question.

Of the 200 images, 20 are used for validation. If the gaze fixations on some validation image are not where they should be, we discard data from the annotator that follows that validation image and precedes the next one.

Each experiment took one hour, for a total of 14 hours of human time. Thus, obtaining the gaze maps for each of our 13 attributes took a short amount of time, *about one hour per attribute* or one minute per image-attribute pair. Our collected gaze data is available on our website[2]. Note that viewing an image is faster than drawing a rationale (45 seconds), so we save time and money compared to [28].

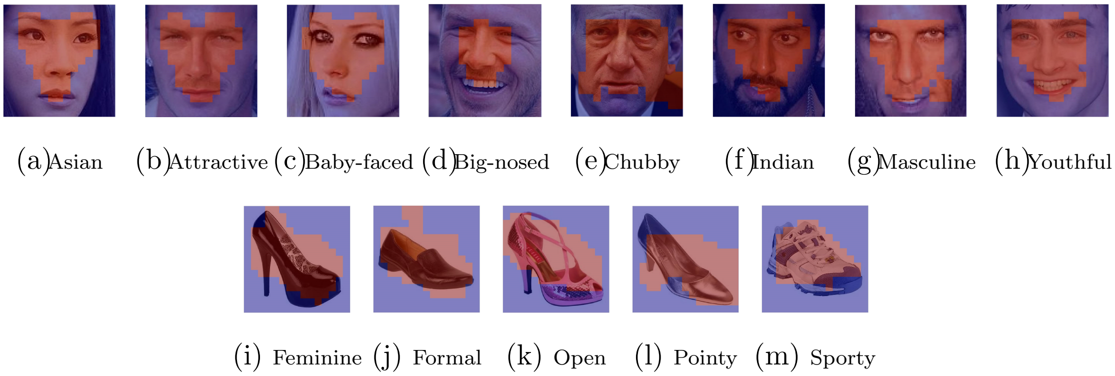
In contrast to our approach, some saliency work [62, 53] approximates gaze with mouse clicks, but as argued in relation to region selection methods (Section 2.3.2), clicks require conscious awareness of what makes an image "formal" or "baby-faced", which need not be true for attributes.

[1] http://www.gazept.com/product/gazepoint-gp3-eye-tracker/

[2] http://www.cs.pitt.edu/~nineil/gaze_proj/

4.1.3 Generating gaze map templates

The gaze data and labels are collected jointly but aggregated *separately* for each attribute. The format of a recorded gaze map is an array of coordinates (x, y) of the image being viewed. We convert this to a map with the same size as the image, with a value of 1 or 0 per pixel denoting whether the pixel was fixated or not. First, the gaze maps across all images that correspond to positive attribute labels are OR-ed (the maximum value is taken per pixel) and divided by the maximum value in the map. Thus we arrive at a gaze map gm_m for the attribute m with values in the range $[0, 1]$. Second, a binary template bt_m is created using a threshold of $t = 0.1$ on gm_m. All locations greater than t are marked as 1 in bt_m and the rest as 0. Third, we apply a 15x15 grid over the binary template to get a grid template gt_m. The process starts with a grid template filled with all 0 values. Then if a pixel with value 1 of bt_m falls inside some grid cell of gt_m, this cell is turned on (all pixels in that cell are replaced with 1). Some examples of the generated templates are shown in Fig. 7. Red regions are cells with value 1, while blue regions are cells with value 0.


(a)Asian (b)Attractive (c)Baby-faced (d)Big-nosed (e)Chubby (f)Indian (g)Masculine (h)Youthful

(i) Feminine (j) Formal (k) Open (l) Pointy (m) Sporty

Figure 7: Grid templates for the face (top row) and shoe attributes. Best viewed in color.

To get templates that capture the subtle variations of how an attribute might appear [73] and also separate different types of objects, a clustering is performed over the images labeled as positive by our human participants. For example, boots can be in one group and high-heels in another. We use K-means with $k = 5$.[3] After the clustering procedure, we

[3] We did not tune this parameter but also found the performance of our algorithm *not* to be sensitive to

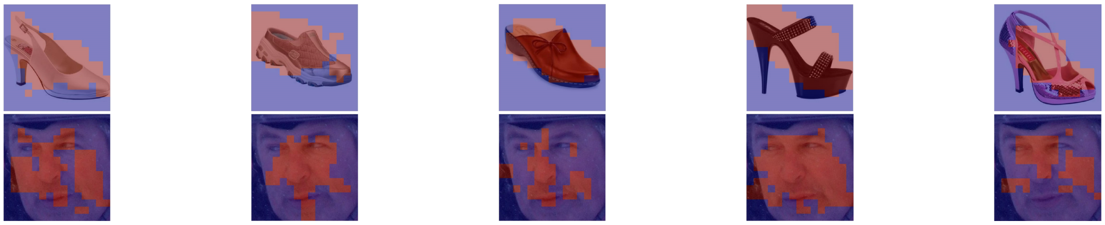


Figure 8: Grid templates for each positive cluster for the attributes "open" (top) and "chubby" (bottom). At the top, we show multiple templates capturing the nuances of "openness". At the bottom, we show how multiple templates for "chubby" look on the same image. Best viewed in color.

repeat the grid template generation, but now separately for each of the five clusters. Thus, we obtain five grid templates per attribute. Each attribute classifier can then specialize to a very concrete appearance, which might make learning a reliable model easier than learning an overall single-template model.

Examples of the five templates for the attribute "open" are shown in Fig. 8. We observe that each template captures a different meaning of "openness", e.g. open at the back (first, second and third image), front (fifth), or throughout (fourth). We also show multiple templates for the attribute "chubby" on the same image, for easier comparison. We quantitatively compare using one versus five grid templates in Tab. 7 and 9, and show additional qualitative results on Figures 9 and 10.

4.1.4 Learning attribute models using gaze templates

We consider two approaches: Single Template (ST) and Multiple Templates (MT). For Single Template, the parts of images involved in training and testing are multiplied by the grid template values, which results in image pixels under a 0 value being removed and keeping other pixels the same. We then extract both local and global features from the remaining part of the image, and train a classifier corresponding to the template using these features. At test time, we apply the template to each image, extract features

its choice. One can pick K using the silhouette coefficient [120] or a validation set.

from the 1-valued part, and apply the classifier. For Multiple Templates, we train five different classifiers (one per cluster), each corresponding to one grid template. We classify a new image as positive if at least one of the five classifiers predicts it contains the attribute.

Comparison to rationales. To test the effectiveness of our gaze template construction, we also tried implementing our gaze templates as rationales [185, 28]. In this work, the authors seek not only labels from their annotators (e.g. this person is attractive, and that person is not), but also ask annotators to mark with a polygon the region in the image that determined their positive/negative response. Our gaze templates resemble attributes since they indicate which region a human looked at to determine if an attribute is present. We implement gaze as a form of rationales as follows. If we have a positive image x_i and a template region within it r_i, we construct an artificial training example $x_i - r_i$ that excludes r_i, and then generate an additional constraint in the SVM formulation that enforces that x_i examples should receive a higher score than $(x_i - r_i)$ examples. This resulted in inferior results, thus confirming our choice of how to incorporate the gaze templates into attribute learning.

4.1.5 Learning attribute models with gaze prediction

So far we have used a single gaze template (or five templates) for each attribute, and applied it to all images. Rather than using a fixed template, one can also *learn* what a gaze map would look like for a novel test image. We construct a model following Judd's simple method [63], by inputting (1) our training gaze templates, from which 0/1 gaze labels are extracted per pixel, and (2) per-pixel image features (the same feature set as in [63] including color, intensity, orientation, etc; but excluding person and car detections). This saliency model learns an SVM which predicts whether each pixel will be fixated or not, using the per-pixel features. We learn a separate saliency model for each attribute.

For each attribute, as outlined in Alg. 1, we first learn a saliency model. Then we predict a real-valued saliency score for each pixel in each test image. Finally, we convert this real-valued saliency map to a binary template. To generate the latter, we consider thresholds u

Algorithm 1: Predicting a gaze template using saliency.

```
Data: Training grid templates $templates_{train,m}$ for attribute $m$; test image $i$
Result: Template for the test image $template_i$, to be used for feature extraction
Train a saliency model using $templates_{train,m}$;
Apply saliency model to $i$ to predict gaze map $gm_m^i$;
for $u \in \{0.1, 0.2, \ldots, 0.9\}$ do
    $r \leftarrow$ Threshold $gm_m^i$ at $u$;
    $score_u \leftarrow$ similarity of $r$ and $templates_{train,m}$
end
$fu \leftarrow$ Set the final threshold to $\arg\max_u(score_u)$;
$template_i \leftarrow$ Apply threshold $fu$ to gaze map $gm_m^i$
```

between 0.1 and 0.9. To score each u, we apply it to the predicted gaze template for our test image to obtain a binary test template. We compute the similarity between that test template and the training binary templates (Section 4.1.3), as the intersection over union of the 1-valued regions. Finally, we fix our choice of the threshold u to the one with the highest similarity score.

Once we have the binary grid template for the test image, we can extract features from it as in Section 4.1.4, only from the area predicted to have fixations on it. However, the size of the gaze template on test images is no longer guaranteed to be the same as the size of the template on training images, so we have a feature dimensionality mismatch. Thus, we opt for a bag-of-visual-words representation over dense SIFT features (from the part of the image under positive template values in the train/test images) and a vocabulary of 1000 visual words. Then, we build a new classifier using the templates on the training data as discussed above, and apply this model to the features extracted from our new *predicted* grid template. We call this approach Single Template Predicted (STP) or Multiple Templates Predicted (MTP), depending on whether a single or multiple templates were used per attribute at training time. The names denote that at test time, we use a predicted template.

4.2 Experimental validation

In this section, we present a comparison (Section 4.2.1) of our approach against six different baselines on the task of attribute prediction, five of which are alternative methods to select relevant regions in the image from which to extract features. We also include two additional applications: using gaze templates to visualize attribute models (Section 4.2.3), and discovering "schools of thought" among annotators which denote how they perceive attribute presence (Section 4.2.4). We primarily test our approach on the **Faces** and **Shoes** datasets, but in Section 4.2.2, we show an adaptation of our approach for scene attributes.

4.2.1 Attribute prediction

We build attribute prediction models using both standard vision features and features extracted from convolutional neural networks (CNNs). We use HOG+GIST concatenated, the *fc6* layer of CaffeNet [60], and dense SIFT bag-of-words extracted in stride of 10 pixels at a single scale of 8 pixels. Following [129], we use CaffeNet's *fc6* since *fc7* and *fc8* may be capturing full objects and not be very useful for learning attributes.

Our training data consists of the images chosen for the gaze data collection experiments (Section 4.1.1), for a total of 300 for shoes and 480 for faces. The training labels are those provided by our human subject annotators. We perform a majority vote over the labels in case the annotators who labeled an image disagree over its label. We might have more positive images for an attribute than we have negatives, so we set the SVM classifier penalty on the negative class to the ratio of positive images to negative images. We use a linear SVM, and employ a validation set to determine the best value of the SVM cost C in the range [0.1, 1, 10, 100], separately for each attribute.

The test data consists of 341 images from **Shoes** and 660 from **Faces**. The test labels are those that came with the dataset. We pool together positive and negative test data for different attributes, so we often have significantly more negatives than positives for any given attribute. Thus, we use the F-measure because it more precisely captures accuracy when the data distribution is imbalanced.

Our proposed techniques for computing the spatial support of an attribute and extracting features accordingly, MULTIPLE TEMPLATES and MULTIPLE TEMPLATES PREDICTED, as well as their simplified versions SINGLE TEMPLATE and SINGLE TEMPLATE PREDICTED, were compared with the following baselines:

- using the whole image for both training and testing (WHOLE IMAGE);
- DATA-DRIVEN, a baseline which selects features using an L1-regularizer over features extracted on a grid, then sets grid template cells on/off depending on whether at least one feature in that grid cell received a non-zero weight from the regularizer (note we do this only for localizable features);
- UNSUPERVISED SALIENCY, a baseline which predicts standard saliency using a state-of-the-art method [62][4] but without training on our attribute-specific gaze data, and the resulting saliency map is then used to compute a template mask;
- RANDOM, a baseline which generates a random template over a 15x15 grid, where the number of 1-valued cells is equal to the number of 1-valued cells in the corresponding SINGLE TEMPLATE template; and
- an ensemble of random template classifiers (RANDOM ENSEMBLE), which is the random counterpart to the ensemble used by MULTIPLE TEMPLATES.

Finally, we compare our method to the SPATIAL EXTENT (SE) method of Xiao and Lee [166] which discovers the spatial extent of *relative* attributes. While we do not study relative attributes, this is the work that is most relevant to our approach, thus prompting the comparison. [166] form "visual chains" from which they then build heatmaps showing which regions in an image are most responsible for attribute strength. We are only able to perform a comparison for attributes that have relative annotations on our datasets, which we take from [75, 107]. We use these heatmaps as saliency predictions, which in turn are used to mask the image and perform feature selection and attribute prediction (with the SVM cost C chosen on a validation set). We use dense SIFT and bag-of-words as for our SINGLE TEMPLATE PREDICTED.

[4] We used the authors' online demo to compute saliency on our images, as code was not available.

Table 7: F-measure using HOG+GIST features. WI = Whole Image, ST = Single Template, MT = Multiple Templates, DD = Data-driven, US = Unsupervised saliency, R = Random, RE = Random Ensemble. Bold indicates best performer excluding ties.

	WI	ST	MT (ours)	DD	US	R	RE
feminine	**0.80**	0.78	0.71	0.74	0.79	0.74	0.75
formal	0.78	**0.81**	0.80	0.79	0.77	0.77	0.77
open	0.52	0.53	**0.57**	0.45	0.55	0.51	0.51
pointy	0.17	0.17	**0.46**	0.00	0.10	0.14	0.10
sporty	0.74	0.70	**0.76**	0.72	0.71	0.72	0.72
avg	0.60	0.60	**0.66**	0.54	0.58	0.58	0.57
Asian	0.24	**0.33**	0.30	0.22	0.25	0.21	0.21
attractive	0.71	0.74	**0.81**	0.71	0.73	0.75	0.75
baby-faced	0.03	0.06	0.04	0.06	0.06	0.06	0.06
big-nosed	0.47	0.35	**0.52**	0.41	0.39	0.40	0.31
chubby	0.46	0.46	0.43	0.38	0.39	0.43	0.44
Indian	0.24	0.21	0.22	0.18	0.24	0.25	**0.27**
masculine	0.69	0.71	**0.77**	0.69	0.71	0.73	0.75
youthful	0.69	0.65	**0.7**	0.68	0.67	0.68	0.68
avg	0.44	0.44	**0.47**	0.42	0.43	0.44	0.43
total avg	0.52	0.52	**0.57**	0.48	0.51	0.51	0.50

In Tables 7 and 8, we show results for Single Template and Multiple Templates, for HOG+GIST and *fc6*, respectively. In all tables, "total avg" is the mean over the two per-attribute "avg" values above (for shoe and face attributes, respectively). Our MT performs better than the other approaches. In Tab. 7, MT improves the performance on shoes by 6 points or 10% (=0.66/0.60-1) relative to the second-best method, and on faces, it improves performance by 3 points or 7%. In Tab. 8, our method improves performance by 2% on shoes and 7% on faces.

Our MT approach captures the different meanings that an attribute can have and its possible locations. In contrast, ST imposes a fixed template and ignores possible shades of meaning and distinctions between the images viewed.

Table 8: F-measure using *fc6*. See legend in Tab. 7.

	WI	ST	MT (ours)	US	R	RE
feminine	**0.77**	0.73	0.66	0.70	0.69	0.74
formal	**0.63**	0.57	0.61	0.58	0.59	0.58
open	0.51	0.51	0.51	0.49	0.47	**0.53**
pointy	0.19	0.18	**0.38**	0.17	0.18	0.13
sporty	**0.82**	0.78	0.79	0.77	0.67	0.69
avg	0.58	0.55	**0.59**	0.54	0.52	0.53
Asian	0.25	**0.30**	0.22	0.26	0.21	0.24
attractive	0.72	0.73	**0.81**	0.77	0.71	0.73
baby-faced	0.08	**0.12**	0.09	0.10	0.09	0.09
big-nosed	0.46	0.44	**0.67**	0.44	0.40	0.31
chubby	**0.42**	0.37	0.41	0.35	0.34	0.32
Indian	**0.28**	0.13	0.27	0.22	0.16	0.13
masculine	0.7	0.67	0.71	0.66	0.69	**0.73**
youthful	0.65	0.60	**0.68**	0.58	0.61	0.64
avg	0.45	0.42	**0.48**	0.42	0.40	0.40
total avg	0.51	0.49	**0.54**	0.48	0.46	0.47

Also, we provide qualitative results comparing our ST and MT approaches in Figures 9 (for shoes) and 10 (for faces). For our MT approach, we select one meaningful template per image. Each subfigure contains two images: the left one shows the single template, and the right one shows a selected template from the MT method.

In Figure 9, we see that MT captured high-heel as a cue for the attribute “feminine”, while ST focus on the entire shoe. For the “formal” attribute, MT concentrates on the shoe center, while ST focuses on the entire shoe. For “pointy”, MT focuses on the front of the shoe, and for “open”, it concentrates on the center of the shoe, where the open attribute resides. Finally, for “sporty”, MT highlights shoelaces, which are a relevant part of any sporty shoe. In contrast, for these three attributes, ST could not determine a specific relevant part for the attribute.

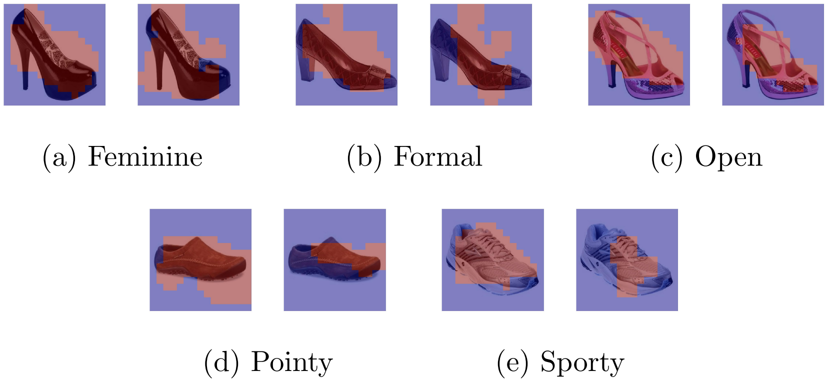
On our face data (Figure 10), MT focus on people’s eyes for the “Asian” attribute.

Table 9: F-measure using gaze maps predicted using the saliency method of [63]. STP = Single Template Predicted, MTP = Multiple Templates Predicted, SE = Spatial Extent. Other abbreviations are as before.

	WI	ST	MT	STP	MTP	DD	US	SE	R	RE
			(ours)		(ours)					
feminine	**0.83**	0.80	0.60	0.78	0.62	0.68	0.63	0.79	0.78	0.82
formal	0.75	0.75	**0.81**	0.76	0.76	0.55	0.66	0.78	0.75	0.74
open	0.53	0.58	0.57	0.53	0.56	0.30	0.43	**0.59**	0.50	0.57
pointy	0.16	0.30	0.53	0.10	0.48	0.55	0.00	**0.56**	0.23	0.20
sporty	0.74	0.81	**0.82**	0.80	0.77	0.54	0.66	0.72	0.70	0.72
avg	0.60	0.65	0.67	0.59	0.64	0.52	0.48	**0.69**	0.59	0.61
Asian	0.22	0.28	**0.32**	0.30	0.26	0.24	0.29	N/A	0.23	0.24
attractive	0.61	0.80	**0.84**	0.80	0.82	0.69	**0.84**	N/A	0.76	0.77
baby-faced	0.06	0.11	0.07	0.06	0.10	0.09	0.06	N/A	0.08	**0.22**
big-nosed	**0.64**	0.33	0.43	0.27	0.40	0.41	0.32	N/A	0.27	0.15
chubby	0.36	0.34	**0.40**	0.30	0.36	0.24	0.24	0.32	0.27	0.29
Indian	**0.25**	0.15	0.24	0.12	0.18	0.12	0.20	N/A	0.16	0.08
masculine	0.68	0.68	0.78	0.71	0.70	0.63	**0.80**	0.71	0.69	0.72
youthful	0.65	0.62	0.66	0.58	0.63	0.53	0.60	**0.69**	0.61	0.60
avg	0.43	0.41	**0.47**	0.39	0.43	0.37	0.42	N/A	0.38	0.38
total avg	0.52	0.53	**0.57**	0.49	0.53	0.45	0.45	N/A	0.49	0.50

Similarly, for “Indian”, it concentrates on the eyes and nose, while ST covers a wider area and picks the mouth also. For “chubby” and “big-nosed”, MT find a smaller relevant area concentrated on the cheeks and nose, respectively. For “baby-faced”, MT determines that the eyes, cheeks and nose are relevant; the template is better localized than the one found by ST. Finally, for the “attractive”, “masculine” and “youthful” attributes, MT finds the same face components as ST, however MT templates are a bit better localized and covers a smaller area.

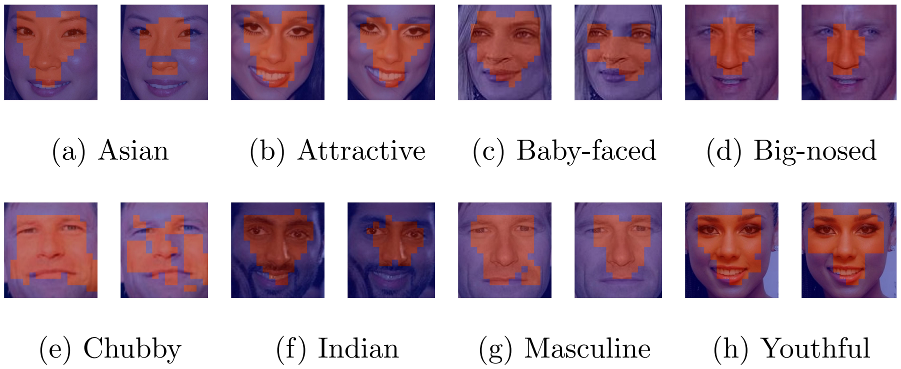
In Tab. 9, we examine the performance of Single Template Predicted and Multiple Templates Predicted. We observe that *predicting* the gaze map, as opposed to using a fixed map, only helps to improve the performance of the proposed feature selection



(a) Feminine (b) Formal (c) Open

(d) Pointy (e) Sporty

Figure 9: A comparison of the single and multiple template methods, for shoe attributes. Left = ST, right = MT.



(a) Asian (b) Attractive (c) Baby-faced (d) Big-nosed

(e) Chubby (f) Indian (g) Masculine (h) Youthful

Figure 10: A comparison of the single and multiple template methods, for face attributes. Left = ST, right = MT.

approach on a few attributes ("formal", "Asian" and "masculine" for STP vs ST, and "feminine" and "baby-faced" for MTP vs MT). This may be because for our face and shoe data, the object of interest is fairly well-centered (although faces can be rotated to some degree). We show some unthresholded predicted gaze maps in Fig. 11. Note how our raw gaze maps correctly detect cheeks as salient for "chubbiness", and shoe toes and heels as salient for "pointiness".

As before, our best results are achieved by using multiple templates. The MT method outperforms the standard way of learning attributes, namely WI, by 10% on average.

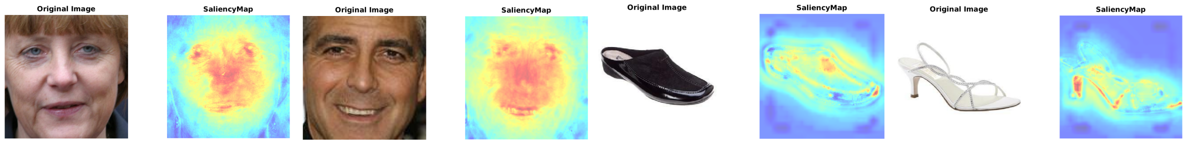


Figure 11: Representative predicted templates for “chubby” and “pointy”. Red = most, blue = least salient.

In terms of region selection baselines, the Random and Random Ensemble baselines perform somewhat worse than Whole Image. The Single Template method performs similarly to Whole Image (slightly better or worse, depending on the feature type). In contrast, our Multiple Templates perform much better. This indicates that capturing the meaning of an attribute does indeed lie in determining where the attribute lives, by also accounting for different participants’ interpretations. The Data-driven baseline performs weaker than the random baselines and our method, indicating the need for rich human supervision. The Unsupervised saliency baseline outperforms our method in a few cases (e.g. “feminine”), but overall performs similarly to Random Ensemble and weaker than our multiple template methods. Thus, attribute information is required to learn accurate gaze templates.

The results of [166] (Spatial Extent) are better than MT for four of the eight attributes available to test for SE, but the average over the eight attributes is almost the same (ours is slightly higher). However, for each attribute, SE required *38 hours* to run on average, on 2.6GHz Xeon processor with 256GB RAM. In contrast, our method only requires the time to capture the gaze maps, i.e. about *one hour*. In Fig. 13 (a), we compare MT with different configurations of SE that take a different amount of time to compute. (The results in Tab. 9 used the original most expensive setting.) Overall our method has similar or better performance than the different runs of SE, but it requires much less time.

4.2.2 Adaptation for scene attributes

Similar to [166], the method most relevant to our work, we have so far only attempted our method on faces and shoes. Given our encouraging performance, we also tested it on ten

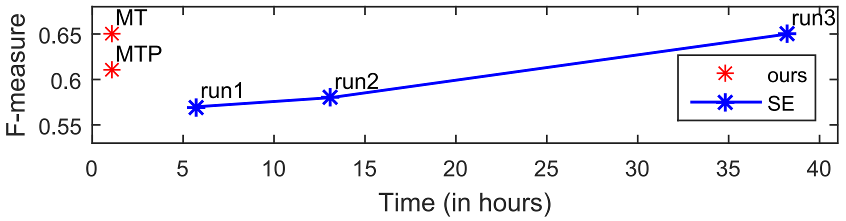


Figure 12: Time comparison of our MT and MTP with SE. On the y-axis is the average F-measure over the attributes tested. Run1, run2, and run3 use different parameter configurations for SE (each one requiring more processing time). Our MT is *more accurate* than the cheaper SE versions and *as accurate* as the most expensive one.

scene attributes [110] (see Tab. 10 for the list), using 60 images per attribute for training and 700 for testing.

A direct application of our MT and MTP performed weaker or similar to WI, likely because scene images contain more than one object. Thus, we adapted our method for this dataset, using five seconds of gaze data. The intuition for our adapted method is as follows: For the attributes "natural" and "sailing", people might look at e.g. trees and water, respectively. Thus, we can use *objects* as cues for where people will look. Such an approach computes location-invariant masks that depend on *what* is portrayed, not *where* it is portrayed.

Our approach consists of three steps: learning an object detector, modeling attributes via objects, and predicting attributes on test images. We fine-tuned the VGG16 network [135] with object annotations from SUN [167] on images not contained in our gaze experiments or test set. We trained three CNNs grouping the objects with similar bounding box size. To learn attributes, we first ran the object detector on our training images. For a given attribute, we counted how many objects intersect with its gaze fixations. Next, we normalized these values and compiled a list of the five most frequently fixated, hence *most relevant* categories for each attribute. At test time, if at least one of these is present, we predict the attribute is present as well.

This simple approach achieves an average F-measure of 0.37, compared to 0.33, 0.34 and 0.45 for WI with HOG+GIST, dense SIFT, and *fc6*, respectively. It outperforms *fc6* on the attributes “driving” and “open area”. A more elaborate approach which extracts *fc6* features on a grid and masks out cells of the grid based on overlap with relevant objects, achieves 0.40.

The objects selected per attribute are shown in Tab. 10. We observe that for “natural”, the fixated objects are trees, grass, sky, and mountains; for “driving”, one of the objects is road, for “swimming” water, and for “climbing” mountains and buildings. This result confirms our intuition that scene attributes can be recognized by detecting relevant objects associated with the attributes through gaze. In our future work, we will formulate this intuition such that it allows us to outperform whole-image *fc6* features on more attributes.

Table 10: The top five objects most frequently fixated per scene attribute.

Attribute	Relevant objects	Attribute	Relevant objects
climbing	mountain, sky, tree, trees, building	open area	sky, trees, grass, road, tree
cold	tree, building, mountain, sky, trees	soothing	trees, sky, wall, floor, tree
competing	wall, floor, grass, trees, tree	sunny	sky, tree, building, grass, trees
driving	sky, road, tree, trees, building	swimming	tree, trees, water, sky, building
natural	trees, tree, grass, sky, mountain	vegetation	tree, trees, sky, grass, road

4.2.3 Visualizing attribute models

We conclude with two applications of our method. First, our gaze templates can be employed to visualize attribute classifiers. We use Vondrick et al.’s Hoggles [157], a method used for object model visualization, and apply it to attribute visualization, on (1) models learned from the whole image, and (2) models learned from the regions chosen by our templates. We show examples in Fig. 13. Using the templates produces more meaningful visualizations than using the whole image. For example, for the attribute “baby-faced”, our visualization shows a smooth face-like image that highlights the form of the nose and the cheeks, and for “big-nosed”, we see a focus on the nose.

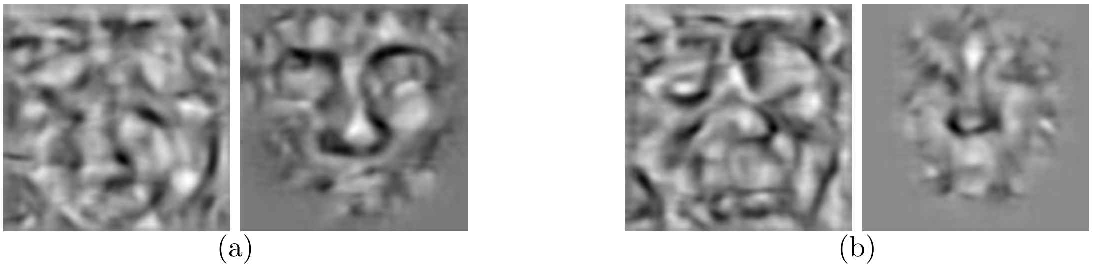


Figure 13: Model visualizations for (a) the attribute “baby-faced”, using whole image features (left) and our template masks (right), and (b) the attribute “big-nosed”.

4.2.4 Using gaze to find schools of thought

Kovashka and Grauman [73] show there exist “schools of thought” (groupings) of users in terms of their judgments about attribute presence. They discover these groupings and use them to build accurate attribute sub-models, each of which captures an attribute variation (e.g. open at the toe as opposed to at the heel). The goal is to disambiguate attributes and create clean attribute models. First, they build a “generic” model (by pooling labels from many annotators). They discover schools using the users’ labels, by clustering in a latent space representation for each user, computed using matrix factorization on the annotators’ sparse labels. Then they use domain adaptation techniques to adapt this “generic” model towards sparse labeled data from each school. At test time, they apply the user’s group’s model to predict the labels on a sample from that user. We follow the same approach, but employ gaze to discover the schools.

We factorize an (annotator, image) table where the entry for annotator i and image j is the cluster membership of image j, computed by clustering images using their gaze maps on positive and negative annotations separately. Thus, for each user, we capture what type of gaze maps they provide, using the intuition that how a user perceives an attribute affects where he/she looks. On our data, the original method of [73] achieves 0.37, and our gaze-based discovery achieves 0.40. Our method is particularly useful for the attributes “big-nosed” (0.41 vs 0.29 for [73]), “masculine” (0.40 vs 0.35), “feminine” (0.43 vs 0.36), “open” (0.58 vs 0.52), and “pointy” (0.43 vs 0.36), most of which are fairly subjective.We present our full results on Table 11. This indicates using gaze is very informative for disambiguating attributes, the original goal of [73].

Table 11: Quantitative comparison of the original schools of thought approach and our gaze-based approach.

	original approach	gaze-based approach
feminine	0.36	0.43
formal	0.40	0.44
open	0.52	0.58
pointy	0.36	0.43
sporty	0.41	0.43
asian	0.43	0.34
attractive	0.13	0.19
baby	0.49	0.52
big-nosed	0.29	0.41
chubby	0.38	0.35
indian	0.46	0.43
masculine	0.35	0.40
youthful	0.29	0.26
avg	0.37	0.40

4.3 Summary

We showed an approach for learning more accurate attribute prediction models by using supervision from humans in the form of gaze locations. These locations indicate where in the image space a given attribute "lives". We demonstrate that on a set of face and shoe attributes, our method improves performance compared to six baselines including alternative methods for selecting relevant image regions. This indicates that human gaze is an effective cue for learning attribute models. We also show applications of gaze for attribute visualization and finding users who perceive an attribute in a similar fashion.

From our transfer learning project, we expand our contextual explanations via related/selected attributes to contextual explanations in the form of gaze. Gaze provides supportive data to explain "why" an attribute is present.

Gaze also can be seen as a selection method, specifically a feature selection approach. This complement our previous project, where selection procedures aim to select relevant source models via an attention mechanism. Here, gaze works as a feature selection and aims to select relevant regions for easy attribute learning.

In conjunction with our first project, this project focus on attribute learning. They consider attribute learning as their core task. However, attributes can be useful for other tasks. We will complement these two works with cross-modality personalization and image retrieval, where attribute learning is used as a side task.

Cross-modality personalization uses metric learning and image retrieval employs reinforcement learning. Both machine learning paradigms complement our previous projects, which are only based on supervised learning. Also, image retrieval uses relative attributes, which complement binary attribute classification.

One drawback of this project is that eye-trackers require calibration and a controlled environment. Thus, they are not suitable for uncontrolled large scale experiments such as crowdsourcing. In our next project, we solve this issue via a *revealing mask* web interface on a blurred image. This procedure captures data, which is highly correlated with acquired data via an eye-tracker, however, it does not require any special equipment.

Finally, we continue using contextual explanations. Cross-modality personalization still uses gaze, and caption annotations, which capture writing style; and image retrieval focuses on visual sketches and attribute comparisons to feed our reinforcement learning agent.

5.0 Cross-modality personalization for retrieval

This chapter focuses on attributes as a side task to improve cross-modality personalized retrieval. Human enriched data is represented in the form of gaze and writing style (captions). Hence, in addition to modeling gaze and captions, we also explicitly model the personality of the users providing these samples via attributes. We incorporate constraints that encourage samples on the same image to be close in a learned space; we refer to this as *content modeling*. We also model *style*: we encourage samples provided by the same user to be close in a separate embedding space, regardless of the image on which they were provided. To leverage the complementary information that content and style constraints provide, we combine the embeddings from both networks.

Our content/style approach achieves better performance than existing approaches for cross-modal retrieval. We consider two strong baselines: one uses metric learning with hard negative mining, and the other employs matrix factorization to find latent factors in order to couple different data modalities.

The main contribution of our work is a novel method that separately considers style and content, and combines them to achieve effective personality-aware retrieval across three modalities. We also examine the latent interdependency of these three modalities: learning all three jointly can be beneficial, even if only two are used at test time. In order to evaluate our method, we collect two datasets of caption-gaze samples for (139, 79) unique users, and over (2700, 1350) annotations on (543, 363) unique images, with worker identity preserved. These dataset can be used by other researchers investigating personalized perception.

The remainder of this chapter is organized as follows. In section 5.1, we describe how we collect our dataset using Amazon Mechanical Turk, and our approach to combine base, content, and style networks via metric learning constraints. Then, in Section 5.2, we describe our setup, evaluation metrics, and comparison with two baselines. We also perform an experiment of modeling all embedding tasks jointly. Finally, we summarize this chapter in Section 5.3.

5.1 Approach

Since no prior dataset exists that considers personalized annotations in multiple modalities, we first collect such a dataset (Sec. 5.1.1). We next describe the retrieval scenarios we consider (Sec. 5.1.2). We then describe the cues we use to learn a space in which we can perform cross-modal personalized retrieval, using standard content (Sec. 5.1.3) and personality-aware style (Sec. 5.1.4), in combination with a base network (Sec. 5.1.5, 5.1.6). We finally describe how we learn a joint space for all modalities (Sec. 5.1.7) and conclude with implementation details (Sec. 5.1.8).

5.1.1 Dataset

We collected two datasets. First, we collected an *ads* dataset of 2700 annotations total, over 543 unique images (of which three were used for annotation quality validation), 3 modalities, and from 139 unique viewers (180 separate tasks, but some users completed more than one task). We used the dataset of [54] which contains 64,832 advertisements. In particular, we constructed 60 sets with 15 randomly sampled images each, from topics alcohol, travel, beauty, and animal rights. We showed each set to three viewers/annotators. Second, we complement our *ads* dataset with a subset of images from *COCO* dataset [86]. We selected cluttered images with many objects. Our *COCO* data contains 1350 annotations total, over 363 unique images, 3 modalities, and 79 unique viewers. For each image in the set, annotators were asked to provide the following annotations.

- We simulated gaze capture, using the BubbleView interface [70] shown to return data strongly correlated with gaze data. BubbleView shows a blurred version of an image and asks the viewer to click on parts of the image, revealing clear circle-shaped regions. This interface allows us to crowdsource the collection. We recorded both the locations and order of clicks.
- We also asked annotators to describe the meaning of the advertisement in the form “I should [action that the ad prompts] because [reasoning that the ad provides].” e.g. “I should buy this perfume because it will make me attractive.” In the case of *COCO* data,

we ask annotators to provide a caption to the image.

- Finally, we ask them to complete a ten-question personality questionnaire where they provide multiple-choice answers. The survey was developed in [116] and it is provided in Table 12. It measures five dimensions of personality: neuroticism, extraversion, openness, agreeableness, and conscientiousness. Each question queries for a response in the range from “disagree strongly” to “agree strongly”. Neuroticism is closely related to people tendencies for anxiety, hostility, depression and low self-esteem, while extraversion for positive, energetic and encouraging tendencies. Openness encompasses personality traits such as curiosity, artistry, flexibility, and wisdom, while agreeableness is related to kindness, generosity, empathy, altruism and trusting others. Finally, conscientiousness measures people traits such as efficiency, reliableness, and rationality.

We used Amazon Mechanical Turk to collect our data. To ensure quality, we restricted access to our task to annotators with 98% approval on completed tasks, over at least 1000 submitted tasks. As a form of quality control, we incorporate validation ad images. These validation images have objects in a small portion of the image and a plain background. We check the intersection of the acquired gaze map with the object region. If there is no intersection, the whole set of annotations are discarded, the work is rejected and resubmitted for new annotations.

Samples from different users. In Fig. 14, we show text and gaze samples that different users provided on the same image. We show three columns, and each column shows the results of the same two users; thus we show results from six users total. The top responses are from one user, and the bottom responses are from another user.

In the first column, we observe that the first user (in blue) uses more adjective words, while the second (in red) uses more verbs. For example, in the second row, the first annotator describes the drink as being “chilled and refreshing” while the second describes the ad in a more active way, i.e. the bottle “gives you” a certain pour. From their answers to the personality questions, the second viewer is more extroverted, which aligns with energetic feelings and using verbs.

In the second column, the first user (in green) says “I deserve”, “I am in the mood

Table 12: Personality survey [116] as shown to Amazon Mechanical Turkers. Each question starts with “I see myself as someone who...”

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
... is reserved	○	○	○	○	○
... is generally trusting	○	○	○	○	○
... tends to be lazy	○	○	○	○	○
... is relaxed, handles stress well	○	○	○	○	○
... has few artistic interests	○	○	○	○	○
... is outgoing, sociable	○	○	○	○	○
... tends to find fault with others	○	○	○	○	○
... does a thorough job	○	○	○	○	○
... gets nervous easily	○	○	○	○	○
... has an active imagination	○	○	○	○	○

for”, “I enjoy”, i.e. the responses come from an ego-centric perspective. The second viewer (in purple) focuses more on the state of the world and properties of products, i.e. a more analytical perception. We observe a correlation between the personality inferred from text, and the gaze maps provided. For example, the “self-centered” viewer in green has a lazier approach to examining the image, while the more analytical one is more thorough. From their personality responses, the second viewer exhibits more neuroticism (low self-esteem) than the first. Self-esteem appears related to egocentrism.

In the third column, the first viewer (in black) emphasizes his or her relationship with

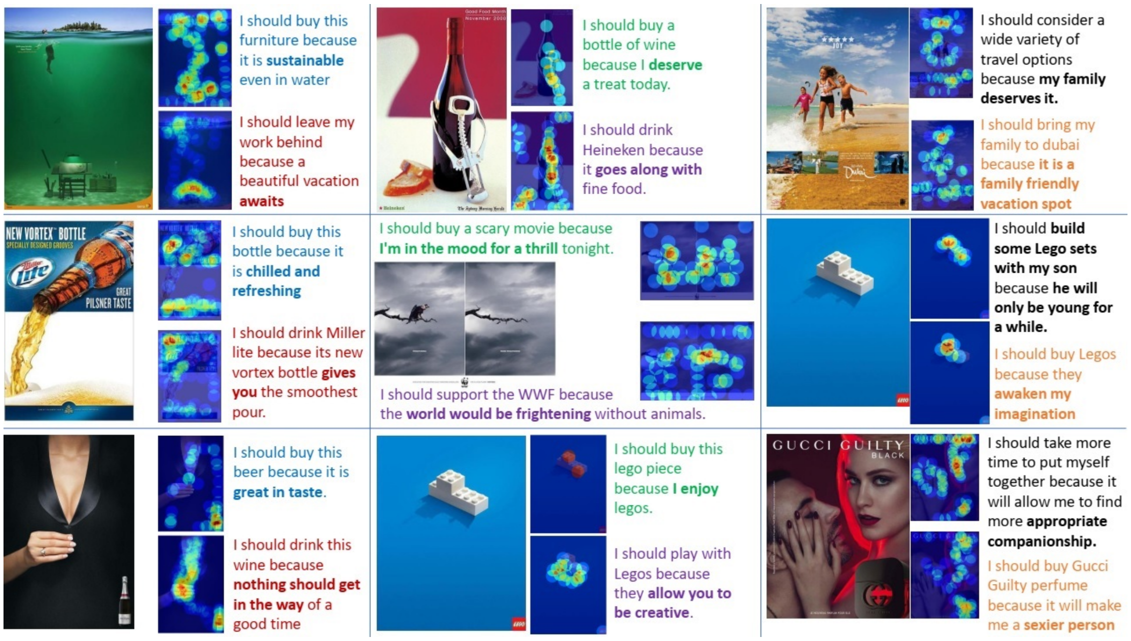

I should buy this furniture because it is sustainable even in water
I should leave my work behind because a beautiful vacation awaits
I should buy a bottle of wine because I deserve a treat today.
I should drink Heineken because it goes along with fine food.
I should consider a wide variety of travel options because my family deserves it.
I should bring my family to dubai because it is a family friendly vacation spot
NEW VORTEX BOTTLE
Miller Lite
GREAT PILSNER TASTE
I should buy this bottle because it is chilled and refreshing
I should drink Miller lite because its new vortex bottle gives you the smoothest pour.
I should buy a scary movie because I'm in the mood for a thrill tonight.
I should support the WWF because the world would be frightening without animals.
I should build some Lego sets with my son because he will only be young for a while.
I should buy Legos because they awaken my imagination
I should buy this beer because it is great in taste.
I should drink this wine because nothing should get in the way of a good time
I should buy this lego piece because I enjoy legos.
I should play with Legos because they allow you to be creative.
GUCCI GUILTY BLACK
I should take more time to put myself together because it will allow me to find more appropriate companionship.
I should buy Gucci Guilty perfume because it will make me a sexier person

Figure 14: Text and gaze samples for different users on our *ads* data. Each column represents a different set of images shown to two users per column. Gaze data is simulated via BubbleView interface [70], which produces data strongly correlated with gaze patterns.

others (e.g. family, child, companion). The second viewer (in orange) focuses more on themselves (e.g. "awaken my imagination", "make me sexier"). Similarly, in the third image, the first viewer pays close attention to the face of the man. In contrast, the more self-centered viewer only looks at the woman (the "protagonist" of the ad). From their personality responses, this first viewer is more agreeable than the second one. Agreeableness is closely related to generosity, empathy, and sympathy, which relates to making a connection with others.

Representation We represent the collected data in the following way. For images, we extract Inception-v4 CNN features [141]. We then mask the image convolution feature with the BubbleView saliency map, by resizing the saliency map to the convolution feature size and multiplying them together. Finally, average pooling is performed to obtain a 1536-dimensional feature vector. We represent textual descriptions as a 200-dimensional Glove embedding [112]. For personality, we used a 10-dimensional feature vector obtained from the personality questionnaire in [116]. Below, we describe how we learn projections of these representations that place them in the same feature space.

5.1.2 Tasks and embeddings

We consider three modalities: gaze, text, and personality. We consider six retrieval tasks: gaze to personality (g2p), text captions to personality (t2p), personality to gaze (p2g), text to gaze (t2g), gaze to text (g2t), and personality to text (p2t). In all of these, we wish to retrieve an annotation that a given user provided, upon receiving another sample from that same user on the same image, but in a different modality (e.g. retrieve the text the annotator wrote to describe the image, conditioned on how the user looked at that image).

We learn a joint embedding of images, gaze, captions, and personality. We separately account for similarities between data from different modalities on the same image, and data on different images from the same user. Our key hypothesis is that bridging modalities through a content loss that ensures samples on the same image, regardless of modality, project closeby, is insufficient for this task. In addition, we need to model the type of captions/gaze/personality that a user demonstrates, by also bridging samples from the same user, *regardless of the image* on which they were provided.

We ensure these similarities through triplet constraints. First, we project each modality to a shared 200-dimensional feature vector via a fully connected layer. For text, we use an embedding layer and calculate the average 200-word embeddings of the words. Then, for every pair of modalities, x (input) and y (output), we generate the content and style constraints described below. Our approach's key intuition is summarized in Fig. 15.

5.1.3 Content Network

We use the following constraints to learn a joint embedding that couples the representations across modalities, for data samples that correspond to the same image. Let us denote a textual description of image i provided by user a as t_i^a, and a gaze map for the same image from the same user by g_i^a. The image that was shown to obtain this text/gaze is denoted

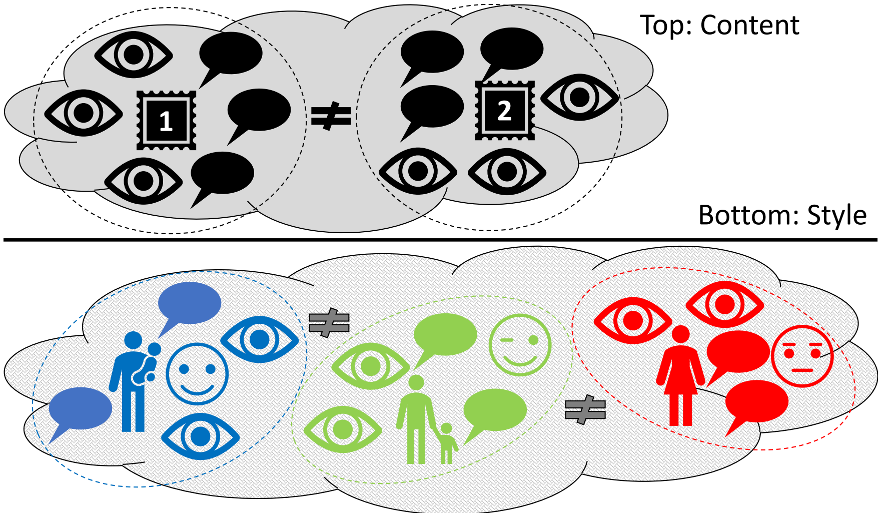


Figure 15: Standard approaches use a content-type loss for cross-modal retrieval, which ensures that samples provided for the same image map are placed in a similar position in the learned space. Here these samples are gaze-masked images and captions. In contrast, we argue that a style-based loss is also necessary. In particular, we wish to ensure that samples that a particular user provided, regardless of the image on which they were provided, cluster together.

by v_i^a. For compactness, we show constraints in a more general form, using x to denote one modality embedding and y to denote a different modality embedding. The original image is only used as an anchor modality; it is not part of our $\{x, y\}$ modality pairs, and is denoted separately.

The embeddings for the following pairs should be similar (where $*$ denotes *any user*[1], and i and j denote distinct images): $\{x_i^*, y_i^*\}$; $\{x_i^*, x_i^*\}$; $\{y_i^*, y_i^*\}$; $\{v_i^*, x_i^*\}$; and $\{v_i^*, y_i^*\}$.

For example, if x refers to text and y refers to gaze, text and gaze samples provided on the same image should be similar; text samples from different users provided on the same image should be similar (and same for gaze samples); and the text and gaze samples' representations should be similar to the original image representation. The last two constraints are necessary because each image is observed by three users, and each provides a potentially different gaze

[1] *Any user* is used because samples come from the same or diff. users, and user differences don't matter for content.

map or caption. We primarily model visual content through the gaze-masked image, which we refer to as the gaze map. However, we would like to ensure the maps for the same image have similar representation.

The following representations should be dissimilar: $\{x_i^*, y_j^*\}$; $\{x_i^*, x_j^*\}$; $\{y_i^*, y_j^*\}$; $\{v_i^*, x_j^*\}$; and $\{v_i^*, y_j^*\}$. These are the same as before, but the subscript in the second sample in each pair is j, referring to a *different* image than the anchor. We generate triplet constraints from these, using all data in the current batch.

For content, we consider the following pairs of modalities as $\{x, y\}$: $\{t, g\}$, and $\{g, t\}$. We train a single network using Eq. 5.1 to bridge the text and gaze modalities. It does not, however, make sense to consider the following: $\{g, p\}$, since the same personality matches multiple images, yet multiple different users (with different personalities) annotated the same images; nor $\{t, p\}$, $\{p, g\}$, $\{p, t\}$.

We would like to ensure that the distances between samples across modalities minimize the following loss:

$$\begin{aligned} L_c(x, y, v; \theta) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \big[\|x_i^* - y_i^*\|_2^2 - \|x_i^* - y_j^*\|_2^2 + \alpha \big]_+ \\ & + \sum_{j \in N} \big[\|y_i^* - x_i^*\|_2^2 - \|y_i^* - x_j^*\|_2^2 + \alpha \big]_+ \\ & + \sum_{j \in N} \big[\|x_i^* - x_i^*\|_2^2 - \|x_i^* - x_j^*\|_2^2 + \alpha \big]_+ \\ & + \sum_{j \in N} \big[\|y_i^* - y_i^*\|_2^2 - \|y_i^* - y_j^*\|_2^2 + \alpha \big]_+ \\ & + \sum_{j \in N} \big[\|v_i^* - x_i^*\|_2^2 - \|v_i^* - x_j^*\|_2^2 + \alpha \big]_+ \\ & + \sum_{j \in N} \big[\|v_i^* - y_i^*\|_2^2 - \|v_i^* - y_j^*\|_2^2 + \alpha \big]_+ \Big] \end{aligned} \tag{5.1}$$

where K is batch size; N is the set of negative samples in the batch; and α is the triplet margin.

5.1.4 Style Network

The style network captures the similarities between different samples that the same user provided. Thus, the embeddings for the following should be similar, where $*$ denotes *any* *image*, and a and b are distinct users: $\{x^a_*, x^a_*\}$[2]; $\{y^a_*, y^a_*\}$; and $\{x^a_*, y^a_*\}$. Thus, annotations provided by the same user (in the same or different modalities) should be similar, regardless of the image. The following should be dissimilar: $\{x^a_*, x^b_*\}$; $\{y^a_*, y^b_*\}$; and $\{x^a_*, y^b_*\}$.

We consider the following three symmetric pairs of input-output modalities $\{x, y\}$: $\{t, g\}$, $\{g, p\}$, $\{t, p\}$, We train separate networks, each bridging the corresponding two modalities. Note that when the input modality is p, there can be fifteen positives (or more if an annotator completed more than one task) for text/gaze.

Thus, we seek to minimize the following expression:

$$\begin{aligned} L_s(x, y; \theta) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \left[\|x^a_* - y^a_*\|_2^2 - \|x^a_* - y^b_*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|y^a_* - x^a_*\|_2^2 - \|y^a_* - x^b_*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|x^a_* - x^a_*\|_2^2 - \|x^a_* - x^b_*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|y^a_* - y^a_*\|_2^2 - \|y^a_* - y^b_*\|_2^2 + \alpha \right]_+ \Big] \end{aligned} \tag{5.2}$$

5.1.5 Base network

We ensure these similarities through the triplet constraint losses described above, which are added on top of a base network. As our base network, we use VSE++ on Ads, which is an adaptation of VSE++ [34] on the dataset of [54], implemented in [175]. This network also employs content-type constraints. It employs the following loss:

$$\begin{aligned} L_b(x, y; \theta) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \left[\|x^a_i - y^a_i\|_2^2 - \|x^a_i - y^a_j\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|y^a_i - x^a_i\|_2^2 - \|y^a_i - x^a_j\|_2^2 + \alpha \right]_+ \Big] \end{aligned} \tag{5.3}$$

In other words, two samples (in different modalities) from the same user on the same image should be close by, while samples from the same user on different images should be

[2] We do not consider personality because it does not change among different images annotated by the same user.

further. However, each user only provided a single sample from each modality on a given image, so we cannot constrain samples on the same image to be close.

Note that we also experimented with ADVISE from [175] as our base network, but it performed worse. ADVISE models image features, while we use a gaze-masked image. In particular, we masked the last convolution layer of Inception-v4 with our BubbleView gaze map. This procedure may hide some relevant information. Also, ADVISE extracts regions of interest (ROI) from the image and finds an embedding space for the image and ROIs. However, in our approach, we do not employ the full image, instead, we use some salient locations, which could hamper the generated embedding space.

5.1.6 Combining base, content and style

We also compute a combined embedding. We assign weights on each embedding; β_b for base, β_c for content, and β_s for style. The embedding for each modality becomes:

$$x = \beta_b * x^b + \beta_c * x^c + \beta_s * x^s \tag{5.4}$$

where x^b denotes the embedding obtained from Eq. 5.3, x^c from Eq. 5.1, and x^s from Eq. 5.2. We optimize the weights on a validation set, separately for each task, using values in the range [0, 1] with step 0.25. In the case of text-to-personality and gaze-to-personality (and vice versa), we use a subset of content constraints, only to ensure gaze/text samples on the same image are similar, and those samples are similar to the corresponding image representation.

5.1.7 Joint embedding and privileged information

In the above description, we create separate networks for each pair of modalities. However, we can also embed all constraints for all pairs into the same space. This means that even if our goal is to retrieve text given personality, and we do not plan to retrieve e.g. text with gaze as input, knowing about the relationship between text and gaze provides additional useful information for the main task. This can be seen as an approach that exploits privileged information, i.e. information that is only available at training time (since at test time, we do not receive gaze as input). Thus, we combine all constraints into the same network, i.e. we add the terms from Eqs. 5.1, 5.2, 5.3 for any pair of modalities, into the same loss, and train a single network. We show in Sec. 5.2.4 that a joint embedding and privileged information improve our system's accuracy.

5.1.8 Implementation details

We implemented the networks using TensorFlow [1]. We use the Adagrad optimizer, a learning rate of 2, an L2 regularizer of 1e-6, 10,000 steps and $\alpha = 0.2$. Every thirty seconds, the network was evaluated on a validation set, and the network with the highest accuracy was selected for testing. For the base network, we found semi-hard negative mining [125] worked best. We selected the smallest negative example that satisfies $d(a,p) < d(a,n)$, where a denotes an anchor, p its positive annotation, n a negative example and d denotes a distance measure. If the condition was not satisfied, a hard negative with the largest $d(a,n)$ was selected.

5.2 Experimental validation

We first verify the contribution of both the content and style components of our method. We compare the combined network against the base, content and style networks separately, and to [152]. We next show the relationship between all three modalities, using a single network for all tasks.

5.2.1 Setup and metrics

We use a test setup where one image is considered a positive; for example, if the input is a gaze sample, the one desired retrieval result is the caption the same user provided on the same image. The negatives are samples provided on any image but from different users. In other words, given a sample x_i^a (caption, gaze, personality) from user a on image i, retrieve sample y_i^a from the same user on the same image, in the presence of 14 other samples: two negatives y_i^b, i.e. on the same image but from other users, and 12 negative y_j^b, where i and

j are distinct images. We split the data over users in 70% for training, 10% for validation and 20% for testing. We run our experiments in five different shuffle splits.

We show three evaluation metrics: top-1 accuracy (is the top-retrieved result the correct one), top-3 accuracy (are any of the top-3 results the correct one), and rank (what is the rank of the correct result among the 15 ranked samples, where lower is better). We use top-1 accuracy to select the best network snapshot per task and per method, because retrieving the correct result at the very top of the 15 samples is the most challenging task.

5.2.2 Methods compared

Our method is the one described in Sec. 5.1.6. It is composed of three constituents, each described in Sec. 5.1.5, 5.1.3 and 5.1.4. We compare all three components below, and their combination, and refer to these as Base, Content, Style, and Combined. The Base result captures the performance of VSE++ [34], which is a state of the art cross-modality embedding method but does not consider personality. We also compare to Veit [152], which is a method that considers personality and predicts hashtags that a particular user would provide on a given image.

Table 13: Summary table for *ads* dataset using top-1, top-3 accuracy and rank metrics for the task-specific setup. We show the average rank (lower is better) for each method across the three metrics. The best performer per task is in **bold**.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	**1.33**	1.67	5.00	3.33	3.67
t2p	4.00	2.00	5.00	2.67	**1.33**
p2g	2.67	**1.67**	5.00	3.67	2.00
t2g	4.00	2.67	2.33	5.00	**1.00**
g2t	3.33	3.67	2.00	5.00	**1.00**
p2t	4.00	3.00	5.00	2.00	**1.00**
avg	3.22	2.44	4.06	3.61	**1.67**

Table 14: Summary table for *coco* dataset using top-1, top-3 accuracy and rank metrics for the task-specific setup. We show the average rank (lower is better) for each method across the three metrics. The best performer per task is in **bold**.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	2.67	**2.00**	5.00	3.00	**2.00**
t2p	3.67	3.00	5.00	2.00	**1.33**
p2g	2.33	3.67	5.00	**1.67**	2.33
t2g	4.00	3.00	2.00	5.00	**1.00**
g2t	4.00	3.00	1.67	5.00	**1.33**
p2t	3.67	2.33	5.00	3.00	**1.00**
avg	3.39	2.83	3.94	3.28	**1.50**

5.2.3 Benefit of combining content and style

We separately evaluate all methods according to each metric described above, and summarize the results. For each task and each metric, we rank each method from best to worst (with rank 1 being best). We then average the ranks across the three metrics, and show the result in Tables 13 and 14. We present the top-3 accuracy, rank and top-1 accuracy results in Tables 15, 16, 17, 18, 19 and 20. As discussed in Sec. 5.1.3, the Content method only makes sense in the case of retrieving gaze from captions, and vice versa, so it produces no result for the other tasks. Here we model all tasks separately i.e. the first/third, second/sixth, and fourth/fifth rows in each table correspond to the same network.

From the comprised Tables 13 and 14, we see our approach outperforms in nine out of twelve tasks, and ranks second in two of the remaining ones. In contrast, Veit and Style are the best for two tasks, and Base for other two. We also observe that Veit is not among the top baselines. A possible reason could be the difficulty to find latent variables due to matrix factorization. Also, it requires more parameters than the other methods, which makes the optimization function harder. Veit has a spatial complexity in the number of parameters: $(d1 + d2) * m + m^2$ (due to two FC layers and matrix factorization; $d1, d2$ are

Table 15: Top-3 accuracy for task-specific setup (higher is better) in *ads* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	0.2107	**0.2111**	N/A	0.206	0.2051
t2p	0.2625	**0.2894**	N/A	0.2806	0.2861
p2g	0.1671	**0.1754**	N/A	0.1643	0.1704
t2g	0.3783	0.4023	0.4384	0.2704	**0.4426**
g2t	0.3801	0.3745	0.4366	0.3074	**0.4463**
p2t	0.2556	0.2718	N/A	0.2741	**0.2768**

the modality input dims and m is the embedding dim) vs $(d1 + d2) * m$ (other approaches).

From the detailed tables, our best result is for the rank measure (Tables 16 and 19), where our approach outperforms all other baselines in four out of the six tasks for both *ads* and *coco* datasets. In this setup, our weakest result is for g2p/p2g, where Veit outperforms our approach. We believe Veit find a latent link between these modalities, which allow easy retrieval in constrast to our methods, which does not use any matrix factorization. Our best competitors for top-3 and top-1 accuracy are Base and Style (Tables 15, 17, 18 and 20). However, overall from our comprised measures Tables 13 and 14, our method performs strongest in the context of all metrics and all tasks. In contrast, other methods have inconsistent performance, i.e. they do well on some metrics but not others.

5.2.4 Joint modeling of all tasks

We next show that all three modalities are inter-dependent. Even if the task is to retrieve a caption based on gaze, i.e. personality is neither input nor output, it helps to model personality jointly with text and gaze. For this experiment and the following ones, we use our *ads* data, because it is the most challenguing task.

In Table 22, we show the top-3 accuracy result using our joint modeling of all modalities. We exclude content because it doesn't apply to all modality pairs. We see that **our combined method is the strongest in three out of six tasks**. This is consistent with

Table 16: Rank for task-specific setup (lower is better) in *ads* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	**7.9912**	8.0199	N/A	8.0361	8.0718
t2p	7.3523	7.1445	N/A	7.0819	**7.0495**
p2g	**7.9241**	7.9949	N/A	8.0625	8.0259
t2g	5.6254	5.4213	5.1315	6.5926	**5.0393**
g2t	5.7305	5.7551	5.2292	6.6616	**5.1417**
p2t	7.4148	7.2403	N/A	7.1894	**7.1653**

Table 17: Top-1 accuracy for task-specific setup (higher is better) in *ads* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	0.0838	**0.0829**	N/A	0.0769	0.0792
t2p	0.1213	0.1463	N/A	0.144	**0.15**
p2g	0.0398	0.0472	N/A	0.0431	**0.0495**
t2g	0.1088	0.119	0.1139	0.0764	**0.1241**
g2t	0.138	0.1514	0.1616	0.1157	**0.1648**
p2t	0.1121	0.1148	N/A	0.1218	**0.1264**

the summary result considering top-3 (see Table 22), top-1 accuracy (see Table 24) and rank (see Table 23) in Table 21. We observe that our joint method outperforms the baselines in three of the tasks and occupies the second position for the remaining three.

Most related modalities. We observe that in terms of top-3 accuracy for the combined method, the easiest task (and hence the most related two modalities) are g2t/t2g, followed by p2t/t2p, then by g2p/p2g, which is the hardest. However, for top-1 accuracy (see Table 24), the easiest and second-easiest tasks are swapped, but the hardest is the same as for top-3 (see Table 22). Thus, text and gaze, and personality and text, are most tightly coupled, while the connection between gaze and personality is weaker. This finding is also confirmed

Table 18: Top-3 accuracy for task-specific setup (higher is better) in *coco* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	0.2121	**0.2222**	N/A	0.2194	**0.2222**
t2p	0.2954	0.2926	N/A	**0.3102**	0.3074
p2g	0.1685	0.1556	N/A	**0.1759**	0.1639
t2g	0.4852	0.5371	0.6139	0.3269	**0.625**
g2t	0.4639	0.5204	0.5972	0.3657	**0.6065**
p2t	0.2722	0.2769	N/A	0.2787	**0.2833**

Table 19: Rank for task-specific setup (lower is better) in *coco* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	**7.8537**	8.1509	N/A	8.0333	8.0917
t2p	7.0389	6.9482	N/A	7.0324	**6.8713**
p2g	**7.7972**	8.0685	N/A	8.0509	8.0407
t2g	4.7426	4.2713	3.7815	6.112	**3.6555**
g2t	4.8593	4.4833	3.8861	6.2833	**3.7352**
p2t	7.1241	6.9482	N/A	7.0306	**6.8695**

Table 20: Top-1 accuracy for task-specific setup (higher is better) in *coco* dataset.

	Veit [152]	Base [34]	Content	Style	**Ours**
g2p	0.0982	**0.1074**	N/A	0.0972	0.1037
t2p	0.1361	0.15	N/A	0.1537	**0.1639**
p2g	0.0389	0.0454	N/A	**0.0481**	0.0463
t2g	0.1371	0.1593	0.1713	0.0805	**0.1945**
g2t	0.1954	0.2037	**0.2472**	0.1195	0.2463
p2t	0.1167	0.1185	N/A	0.1157	**0.1259**

by our identity classifier (Sec. 5.2.5).

Joint vs task-specific. For easier comparison of task-specific and joint modeling, in Table 25, we show the benefit of modeling all modalities jointly compared to per-task, for the style constraints only. The Joint method is trained with all three modalities at training time, and the Task-specific one is trained just the corresponding two modalities. Both methods receive the same inputs at test time.

We see that the largest improvement (10%) between joint and task-specific is for the personality-to-gaze task, which is the most challenging task. We also see a large gain (4-8%) between joint and per-task when the input/output pair is text-to-personality and vice versa, which we saw above is the second most challenging set of tasks. This makes sense because joint modeling is a double-edged sword. On one hand, leaning the structure of the space from multiple modalities helps; e.g. knowing about the captions a user provides helps us learn what types of users there are at training time, so even if at test time we do not have their captions, we can better predict gaze or personality than if we didn't know about captions at training time. On the other hand, task-specific networks are more focused, thus easier to learn the task. Thus, we expect that using a third modality at training time will only help

Table 21: Summary table showing rank of each method for the joint setup in *ads* data (lower is better). Content doesn't apply; see text.

	Veit [152]	Base [34]	Style	**Ours**
g2p	3.33	3.33	**1.33**	2.00
t2p	4.00	3.00	1.67	**1.33**
p2g	4.00	3.00	**1.00**	2.00
t2g	3.00	2.00	4.00	**1.00**
g2t	2.67	2.33	4.00	**1.00**
p2t	3.67	3.33	**1.33**	1.67
avg	3.44	2.83	2.22	**1.50**

Table 22: Top-3 accuracy for joint setup in *ads* data (higher is better). Content doesn't apply because it does not consider personality.

	Veit [152]	Base [34]	Style	**Ours**
g2p	0.2056	0.2042	**0.2083**	0.2051
t2p	0.2611	0.2852	0.3019	**0.3134**
p2g	0.1532	0.1787	**0.1815**	0.1792
t2g	0.3625	0.3843	0.2671	**0.4079**
g2t	0.382	0.3847	0.294	**0.412**
p2t	0.2528	0.2569	**0.2847**	0.281

Table 23: Rank for joint setup (lower is better) for *ads* data.

	Veit [152]	Base [34]	Style	**Ours**
g2p	8.0843	8.0296	8.0236	**8.0111**
t2p	7.3778	7.2398	**6.8875**	6.9185
p2g	8.0676	8.012	**7.9732**	8.0093
t2g	5.6593	5.5903	6.7218	**5.3875**
g2t	5.6782	5.7245	6.7796	**5.4935**
p2t	7.394	7.3977	**7.0398**	7.0935

when that third modality provides a latent link between the input and output modalities. The weakest performance of joint modeling is on the text-to-gaze task, since gaze and text are already tightly coupled. They are more closely linked by the meaning of each image.

5.2.5 In-depth look

In this subsection, we provide in-depth intuitions to the task and the performance of our methods. We first quantitatively show how different the samples provided by different users are; see Fig. 14 for a qualitative version. We next show the selected combination weights for

Table 24: Top-1 accuracy for joint setup (higher is better) for *ads* data.

	Veit [152]	Base [34]	Style	**Ours**
g2p	0.0694	0.0754	**0.0778**	0.0768
t2p	0.1167	0.1426	0.1523	**0.1593**
p2g	0.0366	0.0403	**0.0472**	0.0435
t2g	0.0912	0.1102	0.0699	**0.1241**
g2t	0.1366	0.1449	0.1009	**0.1593**
p2t	0.1065	0.1116	0.1264	**0.131**

our studied tasks.

Identity classifier. If the samples from different users are very unique, it will be easy to distinguish between users. To examine how unique samples are, we train an identity classifier where the features are the samples, and the labels are the IDs of the users who provided the samples. We follow a five-fold stratified cross-validation procedure with a linear and RBF support vector machine. We select parameters for nine configurations of gamma and cost for RBF SVM and three configurations of cost for linear SVM.

In the text domain, we employed averaged 200-dimensional Glove embeddings of words in the caption. In the gaze domain, we calculated the percentage of image explored and the max/min distance among all revealed "bubbles." These features produced the best performance for the identity classifier. In the text space, we achieve 6.77% accuracy (while chance is about 0.7%). In the gaze space, we achieve a lower performance of 3.89%; and combining these two spaces, we achieve 9.11%. Thus, users provide reasonably different samples in all modalities, but there is more overlap in the space of gaze samples.

If we use the same features as for retrieval, for text we achieve comparable performance of 6.77%, a lower 0.71% for gaze, and 8.99% for their combination. We opt not to use percentage of exploration and bubbles distances in our retrieval task for gaze, because they won't capture any image content, hence it would be harder to find relations with text.

Table 25: Results of style network only; top-3 accuracy.

	ads	
	per_task	joint
g2p	0.206	**0.2083**
t2p	0.2806	**0.3019**
p2g	0.1643	**0.1815**
t2g	**0.2704**	0.2671
g2t	**0.3074**	0.294
p2t	0.2741	**0.2847**

Content/style/base weights. Our combined approach works by combining the base, content, and style embeddings, with appropriate weights. These weights are chosen on the validation set and applied on the test set. We perform five different shuffle splits, so we obtain five sets of weights for each task. In Table 26, we show the average weight assigned to style, base and content. For the most content-dependent task, gaze to text and vice versa, Content is most important. Then, for text to personality and viceversa, Style is the most important. Ads have subjectivity, thus it requires to capture more style of the different annotators. Finally, for gaze to personality and viceversa, which is the hardest task, ads give the same importance for Style and Base networks.

Table 26: Averaged weights selected for each network on five different shuffled splits.

Tasks	ads		
	style	base	cont.
g2t/t2g	0.2	0.25	**0.7**
t2p/p2t	**0.7**	0.55	N/A
g2p/p2g	0.55	0.55	N/A

5.3 Summary

We described an approach for retrieving samples capturing different perceptions of the same image input, across modalities. To understand how different viewers perceive and describe images, we use two types of constraints. One bridges samples across modalities, using images as anchors in the learned space. The other set of constraints employs viewers as anchors, i.e. samples that came from the same user should be similar, regardless of the viewed image. We combine both sets of constraints and show that the combination usually outperforms the individual constraints. Further, it usually outperforms two baseline approaches. Importantly, learning about gaze, captions, and personality in the same framework improves performance over learning networks for each separate input-output pair of modalities.

In this work, we still use contextual explanations via gaze as in our previous project. We collect gaze at a large scale with crowdsourcing. We employ a *revealing mask* web interface, which can be accessed by any web browser, as opposed to an eye-tracker device, that is not accessible for everybody. We also complement gaze representation with writing style via image captions, and personality via attributes. Gaze is a form to analyze an image, and this analysis is captured in image captions. Both procedures are influenced by our personality. We also observe that gaze capture can be an unconscious analysis, and image captioning provides more conscious and thoughtful thinking. Both data representations are complementary and provide human enriched data, which is associated with personality traits.

This project differs from the two previous in that it uses attributes as a side task. We use attributes to represent personality and improve cross-modality retrieval for gaze and captions. This project also complements the use of attributes for applications as in our image retrieval project. Also, these two projects use metric learning. One to retrieve different data modalities (gaze, captions and/or personality questionnaires) and the other to retrieve images given a sketch. Image retrieval is the focus of our last project, and it also complements traditional retrieval approaches (based on metric learning) with reinforcement learning.

6.0 Image retrieval with mixed initiative and multimodal feedback

This chapter focuses on attribute learning and its application to image retrieval. We propose a *mixed-initiative* framework where both the user and system can be active participants, depending on whose initiative will be more beneficial for obtaining high-quality search results. We develop a reinforcement learning approach which dynamically decides which of three interaction opportunities to give to the user: drawing a sketch, providing free-form attribute feedback, or answering attribute-based questions. By allowing these three options, our system optimizes both the informativeness and exploration capabilities allowing faster image retrieval.

Our reinforcement learning agent achieves competitive performance with standard image retrieval approaches for simulated and real users. We find that our agent learned to prioritize human-initiated feedback early on and complement it with machine-initiated feedback in later iterations. This project was published in [99].

The remainder of this chapter is organized as follows. In Section 6.1, we describe our approach for image retrieval with reinforcement learning, including our system setup and our agent state, actions, reward, and learning. In Section 6.2, we show that our method improves upon standard image retrieval approaches via quantitative experiments for simulated and real users. We also show a study on human-initiated and machine-initiated actions. Finally, we summarize this chapter in Section 6.3.

6.1 Approach

We develop an approach for interactive image retrieval, where the user can provide guidance to the system via two text-based and one sketch-based modalities, described below. The search scenario we envision is the following: The user has a clear idea of the exact target image they wish to find, but does not have that image in hand. Our system's goal is to determine which type of interaction to suggest to the user at any point in time.

6.1.1 Search setup and interactions

Interactions. The user can initiate a search with random images from the database, or ones that match a simple keyword query. Then the user can perform a combination of the following three types of feedback. First, the user can browse the returned images, and relate them to her desired target via attribute comparisons, e.g. "The person I am looking for is *younger* than this person," where "this person" is an image chosen from the returned results. Second, the system can ask the user a question, e.g. "Is the person you are looking for more or less chubby than this person?" Third, the user can draw a sketch to visually convey to the system their desired content. These search interactions are based on prior work [75, 74, 72, 32, 123, 180], and we learn how to combine them.

System interface. Our system is illustrated in Fig. 16, and it has three components: i) a target image, ii) user feedback using attributes or a sketch, and iii) current top images. User feedback is received in each iteration, and updates the top images.

Relevance models. After one of the three interactions is used and feedback from the user is received, the system must rank all database images by estimating their relevance using the feedback the user provided. For free-form attribute feedback and suggested question interactions, following [75], the relevance of a database image is proportional to the likelihood that it satisfies each attribute constraint, e.g. it is more shiny than a reference image.

For sketch interaction, we "convert" the sketch to a photograph (i.e. we add color) using a conditional GAN [56]. An alternative is to directly learn a space whether sketches and images are aligned, and perform retrieval in this space; we show an experiment using this approach as well. Then, CNN features are extracted and we train a one-class SVM [124] whose output probabilities for each image are used to rank the images. The final relevance of an image is a product (multiplication) of all attribute-based and sketch-based relevance estimates.

6.1.2 Reinforcement learning representation

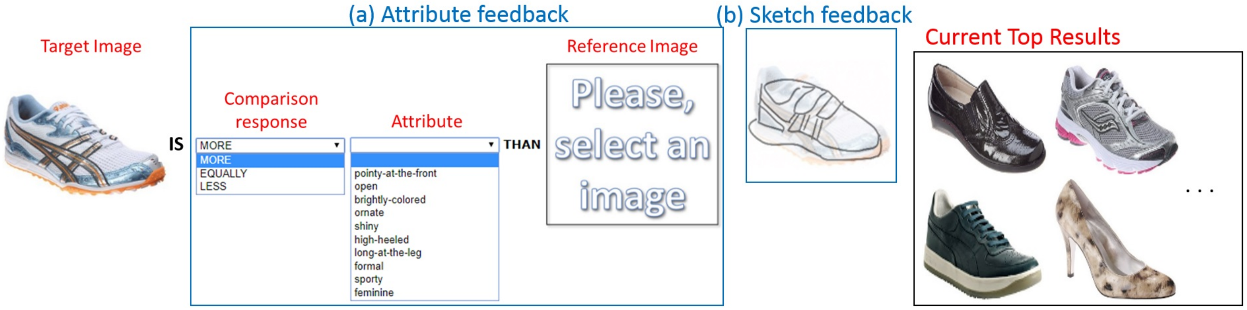


Figure 16: Image retrieval system setup. The system's goal is to find the target image. Users refine the image retrieval using an (attribute, reference image, comparison response) triplet or a sketch. User interactions are used to update the current top image results.

We formulate the selection over search interactions as a Markov Decision Process composed of actions, states, and rewards, defined below.

Actions. We train a reinforcement learning algorithm to select one of three interactions for a given iteration. In order to train it, we require user selections of image-attribute pairs (the free-form feedback proposed in [74]), responses to attribute-based questions proposed in [72] (the more/less/equally value of a comparison between the target and reference image along a certain attribute dimension), and sketches (used for search in [180, 32, 123]). User selections are simulated by selecting an (image, attribute) pair that reduces the part of the multi-attribute space that needs to be searched in order to find the target image. In particular, our simulated users are given a subset of the attribute vocabulary[1], and a set of reference images. They are also given information about how many images in the database satisfy a given image-attribute constraint, e.g. how many images are "less chubby than [this person]," according to the system's model of "chubbiness." The simulated user then chooses the image-attribute pair that results in the smallest number of images satisfying the

[1] Since our simulated users receive system-level information as described next, allowing them to use the full vocabulary results in unrealistic alignment between the user's mental model and the system's predictions.

constraint. This simplifies search as only a few images remain relevant after each feedback constraint is given.

In terms of question responses, we also simulate users' feedback, similarly to [72], by adding Gaussian noise to the attribute model predictions, and choosing the more/less/equally response based on the difference in the attribute values predicted for the target image and the reference image which the system chose. The original method of [72] requires entropy computation, which is computationally expensive if it needs to be repeated many times, as we require for reinforcement learning. Hence, we use an ablation presented in [72] which performs similarly but is much faster. It uses the per-attribute binary search trees of [72] but alternates between attribute pivots in a round-robin fashion.

Sketches are simulated using edge maps [168] generated from the target image, similarly to [56]. We also show experiments using real human-drawn sketches. We then convert them to photographs using a GAN [56], and rank database images by their similarity to the photo, using the probabilities from a one-class SVM [124].

State. Let h_{+prox} and h_{-prox} be positive and negative proxy sets for the target image, defined as the five neighbors closest to the target (excluding the target itself), and five neighbors furthest from the target. We represent our state as (h_{top_ims}, h_{+prox}, h_{-prox}, $h_{actions}$), where h_{top_ims} is the history of top images (i.e. those ranked at the top in previous iterations), and $h_{actions}$ are the actions taken in previous iterations. Images are represented by features extracted from AlexNet [76], and actions by a 3-dimensional binary vector, where all values are zero, except the one corresponding to the taken action. We use a history size of 3.

Rewards. We would like that in each iteration, our top images become more and more similar to the target image (which is unknown to the system). We can measure this using two cues: distance to positive proxy images, and distance to negative proxy images. We encourage a decrement of the first distance, and an increment of the later distance. We do this using a reward function $r(s, s')$ which is evaluated when an action is performed and causes a transition from state s to state s'. Each state has associated top images (*top_ims*) and proxies (*+prox* and *−prox*). We calculate the Euclidean distance d between (1) the average features of all top images and (2) the average features of the positive/negative proxy images. Then the function r is defined as:

$$r(s,s') = sign[d(top\_ims,+prox) - d(top\_ims',+prox)] + sign[d(top\_ims',-prox) - d(top\_ims,-prox)] \quad (6.1)$$

In other words, we want the distance of the top images to positive proxies to decrease, and distance to negative proxies to increase. One might think that using positive proxies is enough, however we prefer a more fine-grained representation. Both sets of proxies are helpful, especially at the beginning when the search space is large and could be misleading. For example, imagine a two-dimensional search space where $+prox = (4,1)$, $-prox = (1,4)$, $top\_ims = (3,3)$ and $top\_ims' = (2,2)$. Thus, $r(s,s') = sign(2-1.4) + sign(2-2.8) = 1-1 = 0$. We observe that decreasing the distance to $+prox$ does not necessarily enforce an increment on the distance to $-prox$, so we need to explicitly encourage this.

We also want to encourage that the sketch action is used only once. Hence, we assign a penalty of -1 if the sketch interaction is requested more than once.

6.1.3 Learning

The goal of our agent is to update the search results by selecting actions. There are many possible states, so using a transition matrix with all states and actions is not recommended. Also, our reward function is data-dependent (i.e. we use image ranking to calculate it). Q-learning [140], which receives a state and predicts the best action, is a good fit for our task. Our Q-learning agent aims to maximize the future discounted reward $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ at each timestep t, where $r_{t'}$ is the reward at time t', T is the time when the search episode ends and γ is the discount factor. We maximize R_t learning a policy to select an action by $\pi(s) = argmax_a Q(s,a)$ at state s.

We approximate the Q function with a neural network, which is based on [11] and is depicted in Fig. 17. Our top images and proxies data uses the same convolution architecture composed of a convolutional layer with 8 filters of size 3x3 and a max-pooling layer. The outputs of the top images and proxies branches are concatenated with the history of actions, and projected using 3 fully-connected layers to generate action scores. We employ RELU

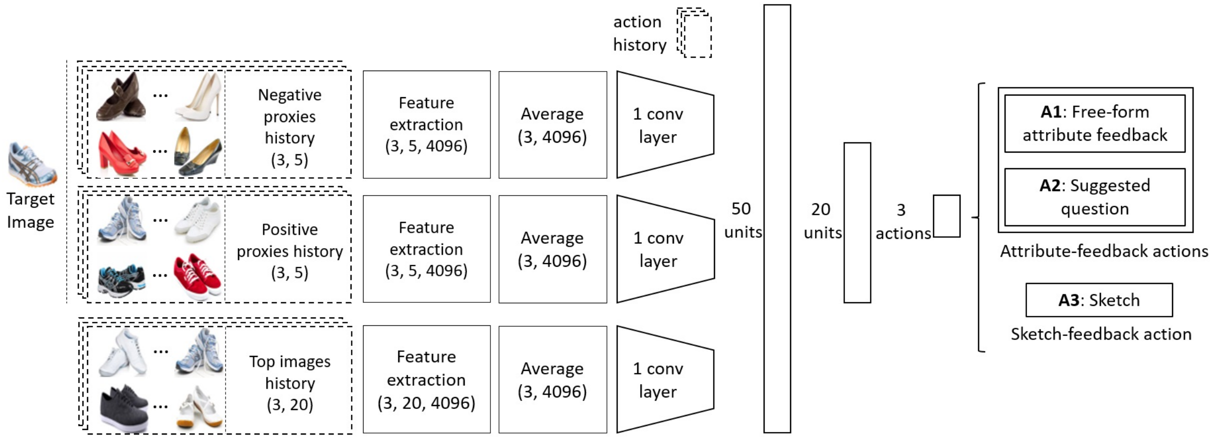

action history
Target Image
Negative proxies history (3, 5)
Positive proxies history (3, 5)
Top images history (3, 20)
Feature extraction (3, 5, 4096)
Feature extraction (3, 5, 4096)
Feature extraction (3, 20, 4096)
Average (3, 4096)
Average (3, 4096)
Average (3, 4096)
1 conv layer
1 conv layer
1 conv layer
50 units
20 units
3 actions
A1: Free-form attribute feedback
A2: Suggested question
Attribute-feedback actions
A3: Sketch
Sketch-feedback action

Figure 17: Architecture of our proposed Q-network. It receives histories of top-ranked images, positive and negative proxy images, and taken actions. It predicts the best action given a specific state. Inputs are denoted with dotted lines. Please see text for further explanation.

activation for the convolutional and fully-connected layers. We employ convolutional layers in the top result image and proxies branches, because they capture information about image features and ordering. Our Q-network learning requires data in the form of $[s,\, s',\, a,\, r]$, which denotes current state, next state, action and reward; and aims to maximize the following loss, where V represents the true future discounted reward using r and s'.

$$L = \frac{1}{2} * [V - Q(s,a)]^2 \qquad\qquad V = r + \gamma * max_{a'} Q(s',a') \tag{6.2}$$

Our approach also considers replay-memory to collect many data instances as it is running. Each instance follows our previous format $[s,\, s',\, a,\, r]$. This information enriches our training data, and in each iteration, a random subset of this data is used for training. This procedure also removes short-term correlation between subsequent states, and makes our algorithm more robust and stable.

At initial stages of learning, random actions are beneficial so the agent can *explore* [140] and get information about the problem. Later this information is *exploited* to select actions. We generate random actions with probability decreasing from 1 to 0.1 as training progresses.

Implementation. We implemented the described network using the Theano [144], Keras [21] and DEER[2] frameworks. We use the RMSProp optimizer, a discount factor of 0.9, a learning rate of 1e-5, and 30 epochs. At the end of each epoch, the network was evaluated on a validation set, and the network that successfully completed more searches (i.e. found the target image in at most 10 iterations) over a validation set was selected for testing.[3]

6.2 Experimental validation

Datasets. We use three datasets which have frequently been used for image search: Pubfig [78] with 11 attributes (e.g. smiling, rounded-face, masculine) and 769 images (after deduplication); Scenes [102, 107] with 6 attributes (open, in perspective, etc.), and 2668 images; and Shoes [75] with 10 attributes (formal, high-heeled) and 12,807 images. We extracted fc6 deep features for Pubfig and Shoes; and fc7 features for Scenes as in [96]. To speed up the interaction of our reinforcement learning agent and the image retrieval system, we reduce the number of images to 1000 by clustering in the predicted attribute strengths space.

Evaluation protocol. For each dataset, we split the data in 70% for training, 10% for validation and 20% for testing. Our reinforcement learning approach uses the train and validation splits to learn to predict actions. To compare the methods more precisely, we tell the user which image to search for (*target image*). In each iteration, the user provides a comparison of the target and pivot/reference image, or a sketch of the target. We report percentile rank of the target, defined as the fraction of database images ranked lower than the target (in the range [0, 1], higher is better).

Baselines. We compare our reinforcement learning agent (RL) with three baselines:

- Whittle Search [75] (*WS*): In each iteration, users select a (reference image, attribute)

[2] https://github.com/VinF/deer/

[3] For our Scenes dataset, the best model is acquired using percentile rank.

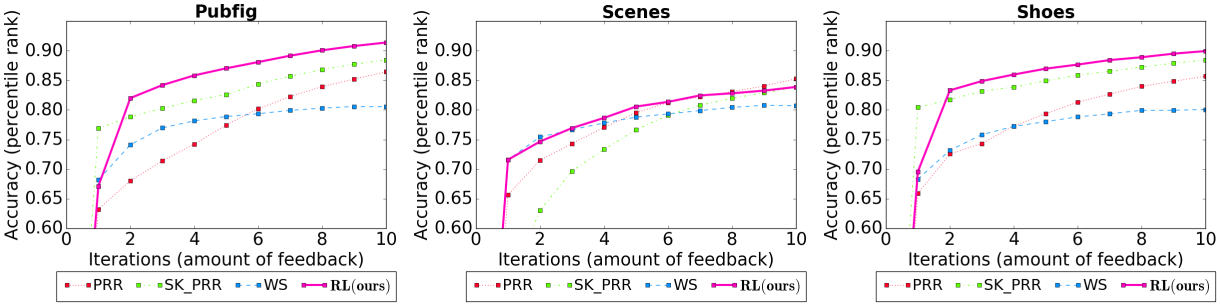


Figure 18: Percentile rank plots for Pubfig, Scenes, and Shoes. Our mixed-initiative RL agent outperforms the other baselines on Pubfig and Shoes, and performs competitively for Scenes.

and compare target and reference for the chosen attribute dimension ("more / less / equally"). The relevance of database images which satisfy this feedback increases.

- Pivot round robin [72] (*PRR*): In contrast to *WS*, *PRR* provides an (image, attribute) pair, and users only need to provide a more/less/equally response.
- Sketch retrieval [180] + pivot round robin [72] (*SK_PRR*): In the first iteration, we ask the user for a sketch of the target image, then attribute questions follow.

6.2.1 Simulated experiments

We simulate ten users as described in Sec. 6.1.2. Fig. 18 shows percentile rank curves for our proposed method and the three baselines. For the Pubfig and Shoes datasets, our reinforcement agent outperforms the baselines with a large margin. However, for Scenes, the improvement is reduced. Hence, we also inspect AUC for the percentile rank curves in Table 27. We observe that our approach outperforms all baselines for all datasets.

We observe that *WS* achieves high accuracy at the very first iterations and outperforms the *PRR* method. This follows the intuition that with *WS*, which allows *exploration*, the user can provide more meaningful feedback that reduces the search space, in contrast to earlier stages of the *PRR* method. However, in later iterations, *PRR* improves accuracy because it follows a binary-search strategy iterating over all attributes. Hence, *PRR* ensures diversity of feedback, in contrast to *WS* which can be repetitive. *SK_PRR* outperforms *WS* and *PRR*

Table 27: AUC for percentile rank curves from Fig. 18. Best scores are highlighted per dataset.

	PRR [72]	WS [75]	SK_PRR [180, 72]	**RL (ours)**
Pubfig	0.729	0.737	0.789	**0.810**
Scenes	0.741	0.741	0.699	**0.754**
Shoes	0.745	0.731	0.806	**0.810**
avg	0.738	0.736	0.764	**0.791**

in two of the three datasets. Incorporating sketch feedback enhances the informativeness of attribute-based feedback, except for Scenes. A possible explanation is that scenes are more complex than faces and shoes, as they contain more than one object. This prevents our GAN from being able to generate good photo versions of our scene edge maps (see Fig. 21).

6.2.2 Live experiments

In order to run a user study, we develop a web interface that implements our three baselines, and our approach. Our approach queries the next action using a REST API[4], that connects to our web interface. For this experiment, we replace sketch-to-photo coloring with sketch retrieval [180] directly comparing features of the sketches to images, as an alternative to get diverse and realistic images. This helps avoid GPU memory problems due to multiple queries for the GAN conversion. We only conduct an experiment for the Shoes dataset because we did not find any appropriate sketch annotations for training, for Faces[5] and Scenes. The result for simulated users (Fig. 19 left) in this setting is similar to our previous findings: our approach outperforms all baselines.

We recruit workers on Amazon Mechanical Turk and university students to search for 100 images. Each participant searches for one image, which is the same for the four methods. We request Turkers with location in the US, HIT approval rate greater or equal to 98%, and at least 1000 approved HITs. We remove blank and careless sketches (i.e. just straight lines),

[4] https://blog.keras.io/building-a-simple-keras-deep-learning-rest-api.html

[5] Fine-grained sketches are available but most real users cannot provide such high-quality sketches.

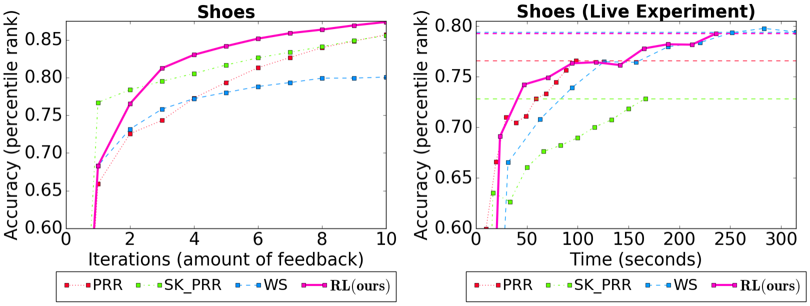


Figure 19: Percentile rank plots for Shoes dataset with simulated (left) and live users (right). Both experiments use sketch retrieval. Live user experiment results are plotted over time.

which results in 88 searches. The results are shown in Fig. 19 (right). Because different interactions require very different amount of user time (PRR: 9s, SK_PRR: 16s, WS: 31s, and RL: 23s), we plot time on the x-axis, multiplying each iteration by the number of seconds it requires. We show horizontal lines with the final (highest) percentile rank a method achieves. Our RL method and WS achieve similar peak performance (79.2% for RL and 79.4% for WS) while PRR only achieves 76.6% at the end of 10 iterations. However, our method achieves higher performance early on; the curve for RL is higher than that for WS until about 230s of user time spent, then performance is similar. Thus, our approach achieves higher performance in a smaller amount of time, compared to the strongest baseline WS.

We examine provided sketches from our live users in Fig. 20. We observe that many of them do a good job. For example, in (row 1, column 4), the sketch has finer details such as the flower ornaments of the flat shoe. Similarly, for (row 3, column 1), the boot was drawn with laces in its top as in its middle. Finally, a sneaker sketch (row 3, column 2) contains shoelaces and details at its bottom part.

6.2.3 Qualitative Results

In order to understand the success of our approach, we visualize some of the generated colored pictures (Fig. 21), and we also show the predicted actions on our test split (Fig. 22).

For our sketch-to-photo generated images, we observe that the most realistic ones cor-

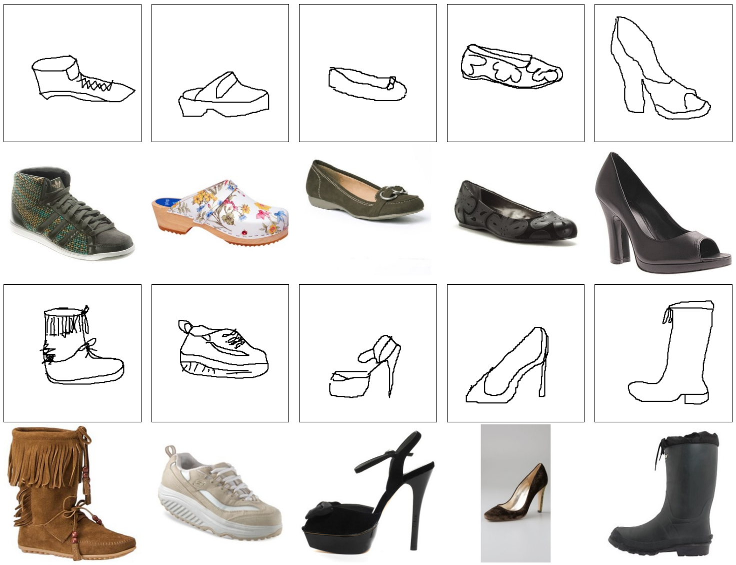


Figure 20: Sketches provided by annotators from Amazon Mechanical Turk and university students for our live experiment. Rows 1 and 3 are user sketches, and rows 2 and 4 are their correspondent target images.

respond to Pubfig, then Shoes, and finally Scenes. This order also corresponds with the performance of our method in terms of percentile rank, where Pubfig and Shoes achieve the best performance. Scenes did not benefit from the generated images as much because they are not realistic and present poor quality. However, our GAN intuitively associates brown color to coast (panels 1 and 2 in row 6, from Fig. 21). Similarly, it learns green color for forest (panels 5 and 6 in row 6, from Fig. 21). Even apart from the generations' quality, edge maps from Scenes do not provide as much detail as edge maps for Faces and Shoes. For example, only the exterior surface of buildings was present in the edge map (see last two panels in row 5). High-level edge maps also can remove crucial objects in the scene, that can not be colored. For example, some trees were removed in (row 5, column 6), which hampers coloring.

We also want to understand our mixed-initiative RL agent, so we count its predicted actions per iteration in Fig. 22. Note that the action at each iteration is chosen by our agent. Partial not available history information is filled with 0s. For Pubfig and Shoes, we observe that *SK* (sketch) and *WS* actions are mainly performed in iterations 1 and 2,

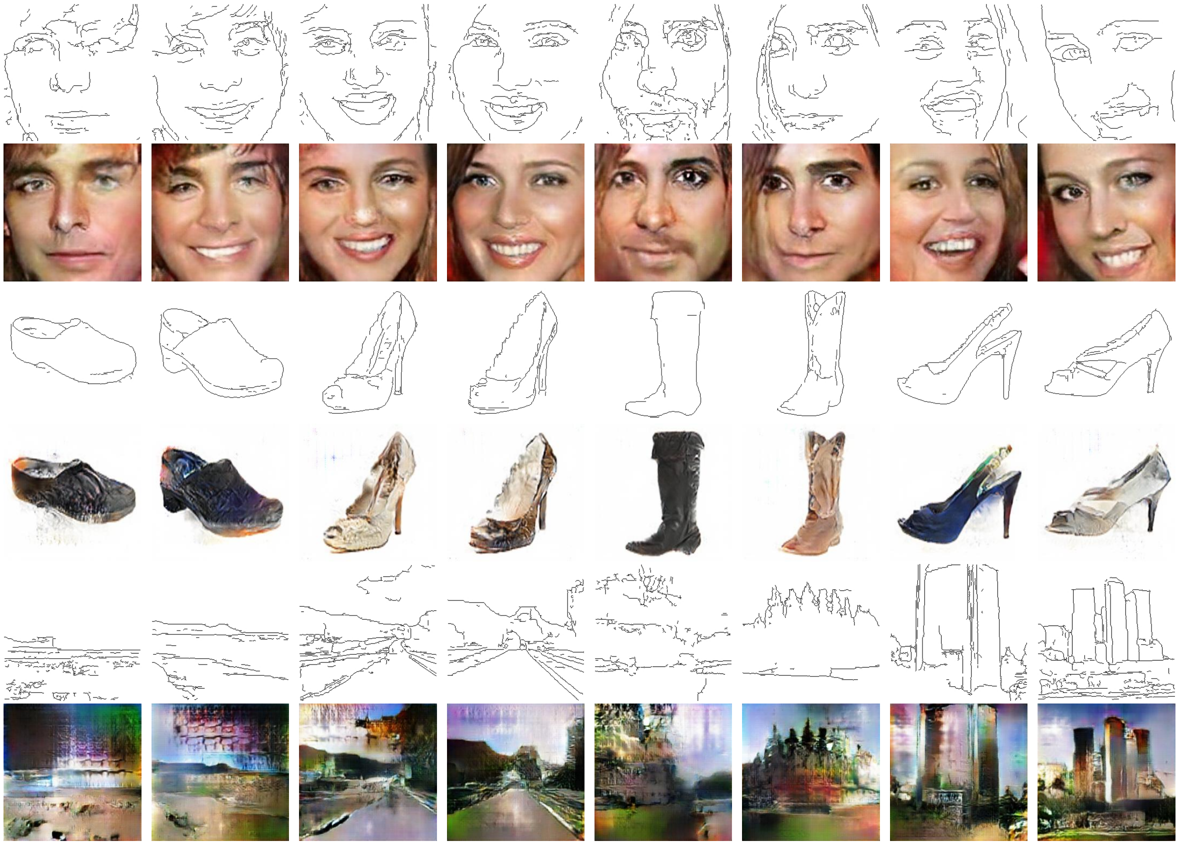


Figure 21: Sample sketch-to-photo colored images for Pubfig (rows 1-2), Shoes (rows 3-4), and Scenes (rows 5-6). Each pair of images denotes the same class category. For each dataset, the first row shows the edge maps, and the second row shows the colored picture.

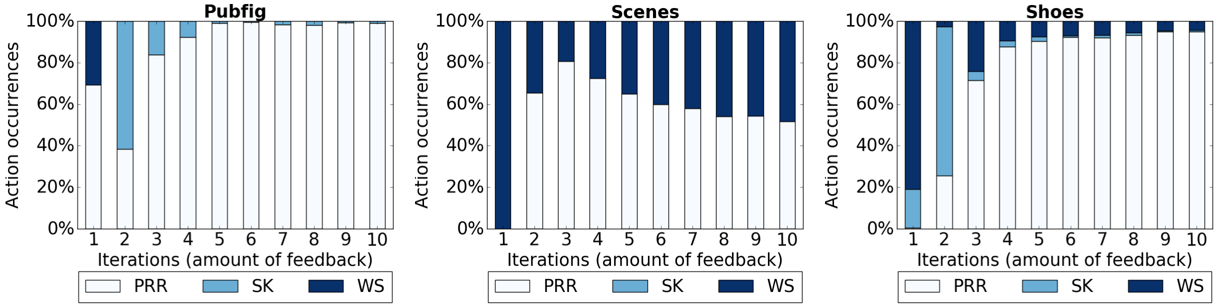


Figure 22: Percentage of actions predicted by our approach in the test set.

because these are the *exploration*-like actions. Then, after iteration 3, the *PRR* is the most common one. Once the most beneficial human knowledge is acquired, having a computer suggest feedback (in the form of questions) helps reduce the search space the fastest. Hence, our agent learned to prioritize human-initiated feedback early on, and complement it with machine-initiated feedback in later iterations. For Scenes, our method prioritizes WS early on and PRR later, and ignores SK because it does not provide much benefit.

6.3 Summary

We explored the problem of selecting interactions in a mixed-initiative image retrieval system. Our approach selects the most appropriate interaction per iteration using reinforcement learning. We find that our model prefers human-initiated feedback in former iterations, and complements it with machine-based feedback requests (e.g. questions) in later iterations. We outperform standard image retrieval approaches with simulated and real users.

This project complements the previous ones because it uses attributes as a side task for image retrieval. It is closely related to our transfer learning project because both projects aim to combine intelligently different source of data. For transfer learning, we are interested in combining source models. And for this project, we are interested in combining image retrieval systems. Also, we employ different techniques. The former uses an attention mechanism, and the latter employs reinforcement learning. It is also closely related to our cross-modality retrieval project because both methods aim to retrieve data. One retrieves images, and the other retrieves gaze, captions and/or personality.

Also, both projects use metric learning and crowdsourcing. Metric learning is used for image retrieval given a sketch, and to find a common embedding for gaze, captions, and personality. Crowdsourcing is used for evaluation with real users, and for data collection.

Finally, in this project, we use contextual explanations in the form of sketches. Sketches are a form of human-enriched data because we have to reason the most salient features of the target image to draw an sketch. They also are a holistic view of an image query that complement attribute comparisons.

7.0 Conclusions

In computer vision, attributes are mid-level concepts shared across categories. They are useful for efficient communication between humans and machines, the description of objects in fine-grained details, and description of unfamiliar objects.

These are very attractive properties for attributes, however, they present many challenges. In this thesis, we address many of them and contribute to boosting its performance and applicability. Specifically, we demonstrate how to use contextual explanations to enhance attributes predictive power. Our contributions are categorized in learning and applications categories.

Related to attribute learning, we contribute to improve subjectivity-based and contextual-based attribute classifiers. For subjective-based classifiers, our cross-modality project learns personalized perception through gaze, writing style, and personality traits, and our human gaze project uses gaze to incorporate explanations and capture different attribute interpretations via clustering and matrix factorization. For contextual-based classifiers, our non-semantic project shows that human-relevant knowledge can be extracted for unrelated domains when there is a lack of contextual information or semantically related attributes. Also, our cross-modality project learns together gaze, caption and personality; which are contextual data sources. Finally, we complement attribute-based image retrieval approaches with sketch-based ones via reinforcement learning. Sketches provide a visual context for attribute textual feedback. Notice that the last two contributions are also attribute applications for data and image retrieval.

Overall these projects, we focus on enhancing data representation for attribute learning with human knowledge and contextual explanations. We enrich attribute representation from *discovering* human rationale shareable knowledge to *providing* human contextual explanations. Our contextual explanations are composed of gaze, writing style, and visual data in the form of sketches. All these representations encapsulate different human rationales. For example, human gaze captures subconscious intuition of the meaning of an attribute. In contrast, text follows a conscious thinking with gaze prior rationale. Finally, sketches encapsulate human rationale in a visual representation. All these data embed personality and can capture different interpretations of knowledge.

The remainder of this chapter is organized as follows. In Section 7.1, we describe the main contributions for data enhancing in this thesis. Then, we summarize some limitations and promising future ideas in Sections 7.3 and 7.4.

7.1 Main contributions

The main contributions of this thesis are:

- Discovery of transferable rationale knowledge to improve attribute learning
  - We develop a novel attention-guided transfer network for attributes for non-semantic related domains and in a data scarcity scenario.
  - We show a study of transferability of attributes from unrelated domains.
- Effective use of human contextual explanations in attribute learning for recognition and data retrieval
  - We develop a new approach for learning attributes using explainable rich data in the form of gaze.
  - We develop two applications: one to visualize attribute models using gaze templates, and another to discover groups of users according to different attributes interpretation.
  - We find that learning gaze, captions, and personality together is beneficial. Thus, these three data modalities have complementary transferable knowledge.
  - We develop a quick mixed-initiative image retrieval system combining attribute-based methods with sketch-based retrieval.
  - We find that human-initiated and system-initiated actions are complementary and beneficial for image retrieval.

Finally, we couple our contributions under a general framework. This framework integrates human contextual explanations on machine learning tasks and is depicted in Figure 23.

It has four components: data acquisition, rationale encoding, attribute learning, and multi-modal learning. We acquire data in the form of labels, gaze, text, and data simulation for user attribute comparison responses and sketches. We develop two data collections interfaces, which are robust to device miscalibration and data quality. Then, our main component is rationale encoding among four data modalities, depicted in Figure 1. First, our *non-semantic project* encodes rationales as background knowledge on unrelated domains. Second, our *gaze project* masks images using human gaze saliency maps. Third, our *cross-modality project* masks images with human-gaze masks and learns jointly with image captions, which are complementary reasoning modalities. Finally, our *reinforcement learning project* encodes visual reasoning in the form of sketch drawings, and combine with attribute comparisons. Finally, our first two projects learn attributes using attention or SVM classifiers. In contrast, our last two projects follow multi-modal learning. Our *cross-modality project* learns a new space where paired data (gaze, image captions and personality traits) are close by. Also, our *reinforcement learning project* combines attributes and sketch retrieval interactions for accurate and faster image retrieval.

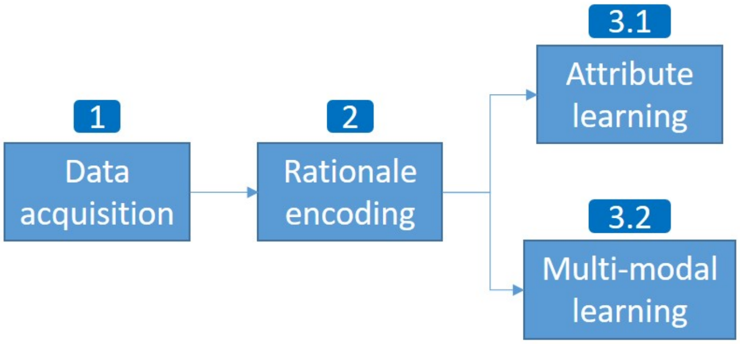


Figure 23: Human rationale framework. First, we develop interfaces to collect our enriched data. Second, we find appropriate encodings to represent rationales. And finally, we learn attributes or multi-modal data representations.

In addition to our contributions, all previous projects generate new knowledge and contribute to the scientific community with the following conference publications:

- N. Murrugarra-Llerena and A. Kovashka. *Learning attributes from human gaze.* IEEE Winter Conference on Applications of Computer Vision (WACV), 2017.
- N. Murrugarra-Llerena and A. Kovashka. *Asking friendly strangers: non-semantic attribute transfer.* Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018.
- N. Murrugarra-Llerena and A. Kovashka. *Image retrieval with mixed initiative and multimodal feedback.* British Machine Vision Conference (BMVC), September 2018.
- N. Murrugarra-Llerena and A. Kovashka. *Cross-Modality Personalization for Retrieval.* Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

Also, our projects were published in workshops in the extended abstract format:

- N. Murrugarra-Llerena and A. Kovashka. *Image retrieval with mixed initiative and multimodal feedback.* LatinX in AI research workshop. Thirty-second Conference on Neural Information Processing Systems (NeurIPS), 2018.
- N. Murrugarra-Llerena and A. Kovashka. *Asking friendly strangers: non-semantic attribute transfer.* LatinX in AI research workshop. Thirty-six International Conference on Machine Learning (ICML), 2019.

In addition to these publications, we are working on extending our image retrieval project for submission to the *International Journal of Computer Vision (IJCV)*. Also, our cross-modality project was accepted to the doctoral consortium at CVPR. Finally, we are grateful for four travel grants to attend AAAI, NeurIPS, ICML, and CVPR.

7.2 Implications

This thesis contributes to the research community with methodologies and findings, which may be useful to other researchers and designers of data collection interfaces.

7.2.1 For researchers

Researchers can benefit from methodologies and findings from this thesis. They can benefit from methods resembling human problem-solving skills, understanding reinforcement learning via action occurrences, and a methodology to conduct experiments.

First, our approaches analysis resembles traditional human problem-skills. We identify the key component of each approach and visualize it. Each visualization provides an explanation as described below.

- In our *non-semantic project*, the key component is an attention mechanism. We visualize attention weights across different domains. We observe which domains are more important to a studied attribute. For example, a "tough skin" (animal domain) attribute gives us the feeling of a "stressful" situation (scene domain). The attention mechanism emulates human skills to understand unfamiliar situations. Humans try to infer properties from what they already know. Similarly, our attention mechanism transfer knowledge from familiar domains to unfamiliar ones.
- In our *gaze project*, the key component is our human-generated template. Then, we visualize template-based attribute models, and we observe that they resemble human intuitions of attributes. A baby-faced attribute classifier highlights the cheeks and nose of a person, similarly, as a human does.
- In our *cross-modality project*, the key components are base, content and style networks. We compare these networks in isolation and find the best weight configuration to combine them. This representation resembles a problem-solving skill, where we ask opinions, suggestions, and solutions from our close friends. Then, we find the best solutions and combine them. This paradigm is similar to combining base, content and style network and determine their importance by a validation set.
- In our *reinforcement learning project*, the key component is action prediction. Thus, we plot action occurrences among iterations. We observe that our agent prioritizes human-initiated actions and complement it with machine-based ones for fast and accurate image retrieval. This observation follows the exploration-exploitation paradigm. When a problem is not clear, humans first explore the problem and then exploit the acquired knowledge to propose a solution. In this case, the exploration phase is composed of human-initiated actions, and exploitation comprises system-initiated actions. Sketches and user-defined attribute comparisons provide a clear representation of the image query.

Once, we clearly understand the image query, the system exploits knowledge and provides relevant questions to eliminate the search space.

We believe that by resembling human problem-solving skills, researchers can develop novel methods to solve any problem. As a first step, researchers can use our key components to their research problems and then, they can find other methods to encapsulate different human problem-solving skills.

Second, our study of action occurrences for reinforcement learning is generalizable and can be applied to any reinforcement learning agent. Hence, researchers can identify early and later type of actions. Also, they can draw conclusions, research questions and reveal the rationale of their agent. In our case, our agent prioritizes human-initiated actions and complement with machine-initiated ones. This behavior resembles the exploration-exploitation paradigm as we stated before.

Finally, from an experimental setup view, we recommend researchers to document properly all their experiments. There could exist a huge amount of prototypes. Each prototype has its parameter configurations, network architectures, etc. Hence, for documentation purposes, we develop a simple tool that for each experiment, it saves an identifier, a description and all parameters of the current prototype. This tool provides a modular interface, where each experiment is in its folder, and we can generate comparative plots combining different baselines and prototypes. Also, this tool promotes experiment reusability without the need of re-running. It only requires access to a previously generated evaluation metrics.

7.2.2 For designers of data collection interfaces

The findings in this thesis provide implications on how to design data collection interfaces to acquire valid and high-quality data.

First, to acquire valid data, we find useful to run some preliminary studies. These preliminary studies help us to identify some issues such as device miscalibration, annotators loss of concentration, and strange software bugs. For device miscalibration, we find that annotators loss concentration and miscalibrate eye-trackers device after a long period of usability. Hence, we split our experiment session in small ones and add some validation images to motivate annotators to pay more attention during data collection. Also, for strange software bugs, our web data collection interface reinitializes components and erase intermediate information when the website is resized. Due to preliminary user studies, we identify this issue and fix it. Finally, in our *reinforcement learning project*, for our sketch action, some users do not draw any object and provide an empty drawing. Thus, in a second round of experiments, we add some validation code to tackle this issue.

Second, even if the current thesis does not necessarily focus on acquiring high-quality data, it suggests some guidelines to be considered. We believe user engagement is a key component for high-quality data. For example, acquiring high-quality sketches is challenging because annotators have different artistic skills. Thus, some annotators can provide naturally high-quality data, while others no. Hence, we envision a tool to provide some guidelines to improve drawing quality or to generate automatically an enhanced drawing from the user-provided drawing. In summary, the goal of the tool is to engage users to provide more accurate and meaningful feedback.

7.3 Limitations

Limitations in this work are organized in general and specific settings. For general setting, first, most of our methods use spatial information without considering time information. For example, gaze data can assign more importance to former gaze fixations than later ones. Similarly, words at the beginning of the sentence are more important than words at ending positions.

Second, rationale encodings such as visual cues (sketch) and writing style (text) are related to analytical, creative and artistic personality traits. However, many annotators can have deficient skills in these scenarios. For example in sketch drawings, some annotators with minimal artistic skills can provide simple drawings that can not be informative. We should provide some guidelines to improve these data modalities. We also can provide an interface to improve data quality, however, we should preserve each annotator unique rationale.

Third, our cues are mainly visual. However, physical interactions with objects can provide complementary information and identify easily attribute presence (e.g. heaviness, furriness, softness). For example, a furry couch can be identified by visual and physical features. Visual features can be acquired from a texture descriptor, and physical features can be acquired from pressure or muscle sensors.

Fourth, our rationales encode reasoning indirectly via gaze, writing style, or sketches. However, there are more direct ways to capture reasoning via brain waves or brain imaging, which can be more related to personality traits.

Also, our current approaches do not consider contextual cues such as browsing history, object properties, and events. For example, furriness is different from a dog and a couch. Similarly, a formal shoe has a different meaning for a wedding or at work.

In a more specific setting, first, our gaze rationales capture image subconscious reasoning affected by background human knowledge. We can provide more conscious reasoning by drawing a polygon around a distinctive region associated with the presence or absence of a category. However, polygon drawing is much slower than gaze capturing. Second, we represent personality with coarse granularity. However, it could require a fine granularity to differentiate a bigger quantity of personalities. Fine granularity can be encapsulated using personality questionnaires with more questions. These fine-grained specialized questionnaires capture additional personality traits in contrast to our current questionnaire. Third, annotators can lie in our questionnaire showing a person that they are not. We can overcome this situation with indirect or redundant questions [115]. Indirect questions can ask you for an action in a certain situation and capture a personality trait.

7.4 Future work

This thesis may lead to new future work, which should be explored and studied. Here, we comprise a set of promising ideas and organize them in short-term, medium-term and long-term future work.

For short-term future work, we can tackle some of our limitations. First, we can include temporal data for gaze, text captioning and sketch drawing. Second, we can provide tutoring systems to improve data quality acquisition. For example in sketch drawing, annotators have different drawing skills depending on their artistic skills. Thus, a tutoring system can provide more realistic sketches improving deficient drawings. Third, we can experiment with more direct rationale modalities via brain waves and brain imaging. Similarly, we might use conscious reasoning representations via polygon drawings. Finally, we can improve our personality representation, we can provide more fine-grained questionnaires and use indirect questions to improve data quality.

For medium-term future work, we can provide reasoning modalities (e.g. physical interactions), which are complementary to visual cues. This future work requires data collection, physical objects, and physical sensors such as touch, weight, muscles, and others. Similar to [113], we can simulate physical interactions with a robot arm. Also, some projects follow a general understanding of attributes or sketches, in contrast to individual attribute interpretations. We can tackle this issue with "school of thoughts" to find groups among users in terms of their understanding of attribute presence. These "school of thoughts" can capture similar visual perception and sketching style of users. This procedure can also group similar user for gaze, writing style, and personality traits. Initial experiments can use matrix factorization approaches to identify latent features to group similar users.

Finally, for our long-term future work, we can combine different human sense data, and train data-driven approaches to identify rationales via region selection of our learned models. For the former case, we only explore visual attribute via our sight sense. However, there are other attributes that can be perceived by our other senses: taste for sweet, sour, bitter and salty flavors; smell for floral, lemon, bleach, chocolate, and rotting meat; hearing for load, quiet and peaceful attributes; and touch for heavy and soft properties. Some of them can be complementary to our sight sense, and others can be captured by an isolated sense. For the latter case, our methods incorporate rationales as human enriched data, however, we can also ask programs to identify region rationales for each query image. We can follow [119]'s approach, where an explanation of an image classifier is depicted by image regions, which provide explanations. For example, for a dog classifier, its explanation is a region that encloses a dog. Similar, we can highlight the most relevant regions of our enriched data and our input images to add interpretability in our current setup. Thus, region rationales on enriched data will complement traditional region rationales on images. In this way, we combine human-engineered (enriched data) and data-driven approaches ([119]) for interpretable machine learning.

Bibliography

[1] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar andPaul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[3] Ziad Al-Halah and Rainer Stiefelhagen. Automatic discovery, association estimation and learning of semantic attributes for a thousand categories. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018.

[5] Anna Antinori, Olivia L Carter, and Luke D Smillie. Seeing it both ways: Openness to experience and binocular rivalry suppression. *Journal of Research in Personality*, 68:15–22, 2017.

[6] Guido Borghi, Stefano Pini, Filippo Grazioli, Roberto Vezzani, and Rita Cucchiara. Face verification from depth using privileged information. In *The 29th British Machine Vision Conference (BMVC)*, 2018.

[7] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1543–1550. IEEE, 2011.

[8] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR (IEEE)*, 2017.

[9] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[10] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference of Computer Vision (ECCV)*. Springer, 2010.

[11] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2488–2496, 2015.

[12] Maya Cakmak and Andrea L Thomaz. Mixed-initiative active learning. In *International Conference on Machine Learning (ICML)*. PMLR, 2017.

[13] Barbara Caputo, Eric Hayman, and P Mallikarjuna. Class-specific material categorisation. In *International Conference on Computer Vision (ICCV)*. IEEE, 2005.

[14] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[15] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[17] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[18] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. *Computer Vision–ECCV 2012*, pages 609–623, 2012.

[19] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[20] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[21] Francois Chollet. Keras, 2015.

[22] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] Ingemar J Cox, Matthew L Miller, Stephen M Omohundro, and Peter N Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *International Conference on Pattern Recognition (ICPR)*. IEEE, 1996.

[24] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2017.

[25] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[26] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *TPAMI*, 38(4):666–676, 2016.

[27] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[28] Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. In *ICCV*, 2011.

[29] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[30] Sandra Ebert, Mario Fritz, and Bernt Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[31] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *Transactions on visualization and computer graphics (TVCG)*, 2011.

[33] Victor Escorcia, Juan Carlos Niebles, and Bernard Ghanem. On the relationship between visual attributes and convolutional networks. In *CVPR*, 2015.

[34] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. *BMVC*, 2018.

[35] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[37] Marin Ferecatu and Donald Geman. A statistical framework for image category search from a mental picture. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2009.

[38] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 29–38. ACM, 2008.

[39] David F. Fouhey, Abhinav Gupta, and Andrew Zisserman. 3D shape attributes. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[40] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[41] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference of Machine Learning (ICML)*. IEEE, 2015.

[42] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[43] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[44] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[45] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *TPAMI*, 34(10):1915–1926, 2012.

[46] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[47] Yuhong Guo and Russell Greiner. Optimistic active-learning using mutual information. In *International Joint Conference on Artificial Intelligence (IJCAI)*. Morgan Kaufmann Publishers Inc., 2007.

[48] Yahong Han, Yi Yang, Zhigang Ma, Haoquan Shen, Nicu Sebe, and Xiaofang Zhou. Image attribute adaptation. *IEEE Transactions on Multimedia*, 2014.

[49] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.

[50] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *ICCV*, 2007.

[51] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[52] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang. Learning hypergraph-regularized attribute predictors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[53] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015.

[54] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110. IEEE, 2017.

[55] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2429–2437. Curran Associates, Inc., 2011.

[56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[57] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, (11):1254–1259, 1998.

[58] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[59] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[60] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[61] Jing Jiang and Chengxiang Zhai. Instance weighting for domain adaptation in nlp. In *In Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[62] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *ICCV*, 2015.

[63] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[64] Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[65] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research (JAIR)*, 1996.

[66] Mahdi M. Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[67] Meina Kan, Shiguang Shan, and Xilin Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[68] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[69] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[70] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):36, 2017.

[71] Adriana Kovashka and Kristen Grauman. Attribute Adaptation for Personalized Image Search. In *ICCV*, 2013.

[72] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 297–304, 2013.

[73] Adriana Kovashka and Kristen Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision (IJCV)*, 2015.

[74] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[75] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision (IJCV)*, 2015.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2012.

[77] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[78] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *International Conference of Computer Vision (ICCV)*. IEEE, 2009.

[79] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Describable visual attributes for face verification and image search. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011.

[80] Shrenik Lad, Bernardino Romera Paredes, Julien Valentin, Philip Torr, and Devi Parikh. Knowing who to listen to: Prioritizing experts from a diverse ensemble for attribute personalization. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4463–4467. IEEE, 2016.

[81] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[82] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[83] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016.

[84] Kongming Liang, Hong Chang, Shiguang Shan, and Xilin Chen. A unified multiplicative framework for attribute learning. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[85] Xiaodan Liang, Lisa Lee, and Eric P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[86] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV (Springer)*, 2014.

[87] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[88] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[89] Siqi Liu and Adriana Kovashka. Adapting attributes by selecting features similar across domains. In *Applications of Computer Vision (WACV)*. IEEE, 2016.

[90] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[91] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[92] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[93] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[94] Stefan Mathe, Aleksis Pirinen, and Cristian Sminchisescu. Reinforcement learning for visual object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2894–2902, 2016.

[95] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[96] Bhavin Modi and Adriana Kovashka. Confidence and diversity for active selection of feedback in image retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[97] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Information bottleneck learning using privileged information for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2016.

[98] Nils Murrugarra-Llerena and Adriana Kovashka. Asking friendly strangers: Non-semantic attribute transfer. In *Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018.*

[99] Nils Murrugarra-Llerena and Adriana Kovashka. Image retrieval with mixed initiative and multimodal feedback. In *British Machine Vision Conference, BMVC 2018.*

[100] Nils Murrugarra-Llerena and Adriana Kovashka. Learning attributes from human gaze. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2017.*

[101] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, 2006.

[102] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 2001.

[103] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[104] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[105] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. British Machine Vision Conference (BMVC), 2017.

[106] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *ECCV*. Springer, 2014.

[107] Devi Parikh and Kristen Grauman. Relative attributes. In *International Conference of Computer Vision (ICCV)*. IEEE, 2011.

[108] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *European Conference of Computer Vision (ECCV)*. Springer, 2012.

[109] Timea R Partos, Simon J Cropper, and David Rawlings. You don't see what i see: Individual differences in the perception of meaning from visual stimuli. *PloS one*, 11(3):e0150615, 2016.

[110] Genevieve Patterson and James Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*, 2012.

[111] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision (IJCV)*, 2014.

[112] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014.

[113] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016.

[114] Nikita Prabhu and R. Venkatesh Babu. Attribute-graph: A graph based approach to image ranking. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[115] Beatrice Rammstedt, Daniel Danner, Christopher Soto, and Oliver P. John. Validation of the short and extra-short forms of the big five inventory-2 (bfi-2) and their german adaptations. *European Journal of Psychological Assessment*, pages 1–13, 08 2018.

[116] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[117] Rajesh PN Rao, Gregory J Zelinsky, Mary M Hayhoe, and Dana H Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463, 2002.

[118] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017.

[119] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[120] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[121] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 1998.

[122] Ramachandruni N Sandeep, Yashaswi Verma, and CV Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014.

[123] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.

[124] Bernhard Scholkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computing (NC)*, 2001.

[125] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[126] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Workshop on Computational Learning Theory (WCLT)*. ACM, 1992.

[127] Sukrit Shankar, Vikas K Garg, and Roberto Cipolla. Deep-carving: Discovering visual attributes by carving deep neural nets. In *CVPR*, 2015.

[128] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[129] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2014.

[130] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 825–832, 2013.

[131] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to transfer privileged information. *arXiv preprint arXiv:1410.0389*, 2014.

[132] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[133] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Data-driven visual similarity for cross-domain image matching. *Transactions on Graphics (TOG)*, 2011.

[134] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[135] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[136] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[137] Jifei Song, Yu Qian, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Deep spatialsemantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the International Conference in Computer Vision (ICCV)*, 2017.

[138] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[139] Jina Suh, Xiaojin Zhu, and Saleema Amershi. The label complexity of mixed-initiative classifier training. In *International Conference on Machine Learning (ICML)*. IEEE, 2016.

[140] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. 2011.

[141] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017.

[142] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[143] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[144] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, 2016.

[145] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[146] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)*, 2001.

[147] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[148] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[149] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Conference on Artificial Intelligence (AAAI)*. AAAI, 2016.

[150] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049):2, 2015.

[151] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Winter Conference on Computer Vision (WACV)*. IEEE, 2009.

[152] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[153] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1170–1178, 2017.

[154] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision (IJCV)*, 2014.

[155] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[156] Luis Von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI*, 2006.

[157] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013.

[158] Catherine Wah and Serge Belongie. Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[159] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[160] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[161] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*, 2013.

[162] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[163] Steven A Wolfman, Tessa Lau, Pedro Domingos, and Daniel S Weld. Mixed initiative interfaces for learning tasks: Smartedit talks back. In *International Conference on Intelligent User Interfaces (IUI)*. ACM, 2001.

[164] Russell L Woods, C Randall Colvin, Fuensanta A Vera-Diaz, and Eli Peli. A relationship between tolerance of blur and personality. *Investigative ophthalmology & visual science*, 51(11):6077–6082, 2010.

[165] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[166] Fanyi Xiao and Yong Jae Lee. Discovering the spatial extent of relative attributes. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[167] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[168] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[169] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[170] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[171] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[172] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: personalized saliency prediction with multi-task cnn. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3887–3893, 2017.

[173] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *International Conference on Multimedia (ICM)*. ACM, 2007.

[174] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[175] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[176] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[177] Peng-Yeng Yin, Bir Bhanu, Kuang-Cheng Chang, and Anlei Dong. Integrating relevance feedback techniques for image retrieval using reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1536–1551, 2005.

[178] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[179] Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[180] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[181] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision (IJCV)*, 2017.

[182] Xiaodong Yu and Yiannis Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *European Conference on Computer vision (ECCV)*. Springer, 2010.

[183] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[184] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, 2013.

[185] Omar Zaidan, Jason Eisner, and Christine D Piatko. Using" annotator rationales" to improve machine learning for text categorization. In *HLT-NAACL*, pages 260–267. Citeseer, 2007.

[186] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[187] Xiang Sean Zhou and Thomas S Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems (MS)*, 2003.