



# Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations

Hosnieh Sattar<sup>a,\*</sup>, Mario Fritz<sup>b</sup>, Andreas Bulling<sup>c</sup>

<sup>a</sup>Max-Planck-Institut für Informatik, Saarland Informatics Campus, Campus E1 4, Saarbrücken 66123, Germany

<sup>b</sup>CISPA Helmholtz Center for Information Security, Saarland Informatics Campus, Stuhlsatzenhaus 5, Saarbrücken 66123, Germany

<sup>c</sup>Institute for Visualisation and Interactive Systems, University of Stuttgart, Pfaffenwaldring 5a, Stuttgart 70569, Germany

## ARTICLE INFO

### Article history:

Received 14 February 2019

Revised 12 December 2019

Accepted 4 January 2020

Available online 13 January 2020

Communicated by Dr. H. Yu

### Keywords:

Gaze pooling

Visual search

Deep learning

Mental image

Visual search target prediction

Visual search target reconstruction

## ABSTRACT

Predicting the target of visual search from human eye fixations (gaze) is a difficult problem with many applications, e.g. in human-computer interaction. While previous work has focused on predicting specific search target instances, we propose the first approach to predict categories and attributes of search intents from gaze data and to visually reconstruct plausible targets. However, state-of-the-art models for categorical recognition, in general, require large amounts of training data, which is prohibitive for gaze data. To address this challenge, we further propose a novel Gaze Pooling Layer that combines gaze information with visual representations from Deep Learning approaches. Our scheme incorporates both spatial and temporal aspects of human gaze behavior as well as the appearance of the fixated locations. We propose an experimental setup and novel dataset and demonstrate the effectiveness of our method for gaze-based search target prediction and reconstruction. We highlight several practical advantages of our approach, such as compatibility with existing architectures, no need for gaze training data, and robustness to noise from common gaze sources.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

As eye tracking technology is continuing to mature, there is an increasing interest in exploring the type of information that can be extracted from human gaze data. Within the wider scope of eye-based activity recognition [1,2], search target prediction [3–5] has recently received particular attention as it aims to recognize users' search intents without the need for them to verbally communicate these intents.

Previous work on search target prediction from gaze data (e.g. [3,4]) is limited to specific target instances that users searched for, e.g. a particular object. This excludes searches for broader classes of objects that share the same semantic category or specific object attributes. However, such searches commonly occur if the user does not have a concrete target instance in mind but is only looking for an object from a certain category or with certain characteristic attributes.

To address these limitations, we broaden the scope of search target prediction to categorical classes, such as object categories

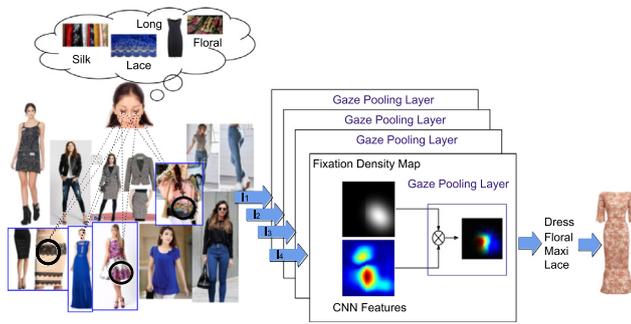
or attributes. One key difficulty in achieving this goal is acquiring sufficient training data. We have to recall that object categorization only in the past decade has seen a breakthrough in performance by combining deep learning techniques with large training corpora. Collecting such large corpora is prohibitive for human gaze data, which poses a severe challenge to achieve our goal.

Therefore, we propose an approach for predicting categories and attributes of search targets that utilize readily trained CNN architectures and combines them with gaze data in a novel *Gaze Pooling Layer* (see Fig. 1). The gaze information is used as an attention mechanism that acts selectively on the visual features to predict users' search targets. These design choices make our approach compatible and practical with current deep learning architectures. Through extensive experiments, we show that our method achieves accurate search target prediction for 10 categories and 10 attribute tasks on a new gaze data set that is based on the DeepFashion data set [6]. Furthermore, we evaluate different parameter settings and design choices of our approach, visualize internal representations, and perform a robustness study w.r.t. noise in the eye-tracking data.

In addition to search target prediction, we want to understand to what level of detail such targets and intents can be visually reconstructed. Cognitive neuroscientists have shown the first success of visualizing mental images based on fMRI data [7,8]. While we

\* Corresponding author.

E-mail addresses: [sattar@mpi-inf.mpg.de](mailto:sattar@mpi-inf.mpg.de) (H. Sattar), [fritz@cispa.saarland](mailto:fritz@cispa.saarland) (M. Fritz), [andreas.bulling@vis.uni-stuttgart.de](mailto:andreas.bulling@vis.uni-stuttgart.de) (A. Bulling).



**Fig. 1.** We propose a *Gaze Pooling Layer* that leverages gaze data as an attention mechanism in a trained CNN architecture. Our methods using this new layer predict the target of visual search in terms of categories and attributes from users' gaze as well as decode gaze data into a visualization of the search target.

also want to assess aspects of the mental state, our task is fundamentally different from prior work in two aspects: (1) While they aim to decode a specific image that was shown to a person, our goal is to decode a visual search target. (2) While they are using fMRI data, we are using gaze data. On a related note, we believe that gaze data is particularly interesting to investigate, as it is practical and affordable to collect and use in many application scenarios of future interfaces [9,10].

As we are targeting to decode images of categorical search targets, generating visualization is difficult due to strong intra-class variations. However, recent advances in deep learning have led to a new generation of generative models for images. Recently, [11] generates images of objects from high-level descriptions. They transfer the high-level text information to a set of attributes (e.g. hair color: brown, gender: female). These attributes are used later on to build an attribute-conditioned generative model.

Hence, we approach the decoding of search targets from gaze data by bringing together our gaze encoding method and categorization with state-of-the-art category-conditioned generative image models. We show the first proof of concept that visual representations of search targets can be inferred from human gaze data. We present a practical approach, as it respects the difficulties in collecting large human gaze datasets. Encoder and decoders are trained from large image corpora and transfer between the two representations is facilitated by a semantic layer in between. We show the importance of localized gaze information for improved search target reconstruction.

## 2. Related work

Inferring the target of visual search is a task studied both in computer vision [3–5,12–14] and human perception [15–18]. Existing approaches vary in the granularity of the predictions, either focusing on predicting specific object instances [3,4,19] or operating at the coarser level and predicting target categories [5,13]. The type of user feedback varies as well. While [3–5] solely use implicit information obtained from human gaze, [12–14] require the user to provide explicit relevance feedback. In the following, we summarize previous works on gaze-supported computer vision, user feedback for image search and retrieval, methods for search target prediction, Decoding visual experience using from EEG or fMRI signals, as well as, works in computer vision area which used autoencoders to decoded images from feature space.

### 2.1. Gaze-supported computer vision

Our approach is related to an increasing body of computer vision literature that employs gaze as a means to provide supervision or indicate salient regions in the image in a variety of recog-

niton tasks. Visual fixations have been used in [20,21] to indicate object locations in the context of saliency predictions, and in [22–24] as a form of weak supervision for the training of object detectors. Gaze information has been used to analyze pose estimation tasks in [25,26] as well as for action detection [27]. Gaze data has also been employed for active segmentation [28], localizing important objects in egocentric videos [29–31], image captioning and scene understanding [32,33], as well as zero-shot image classification [34].

### 2.2. User feedback for image search and retrieval

To close the semantic gap between the user's envisioned search target and the images retrieved by search engines, Ferecatu and Geman [13] proposed a framework to discover the semantic category of the user's mental image in unstructured data via explicit user input. Kovashka et al. [12] introduced a novel explicit feedback method to assess the mental models of users. Most recently Yu et al. [14] proposed to use free-hand human sketches as queries to perform instance-level retrieval of images. They considered these sketches to be manifestations of users' mental model of the target. The common theme in these approaches is that they require explicit user input as part of their search refinement loop. Mouse clicks were used as input in [13]. A. Kovashka [12] used a set of attributes and required users to operate on a large attribute vocabulary to describe their mental images. In [14] the feedback was provided by sketching the target to convey concepts such as texture, color, material, and style, which is a non-trivial step for most users. In contrast, in our work, we do not rely on a feedback loop as in [12] or explicit user input or some form of the initial description of a target as in [12–14]. We instead use fixation information that can be acquired implicitly during the search task itself, and demonstrate that such information allows us to predict categories as well as attributes of search targets in a single search session.

### 2.3. Visual search target prediction

Human gaze behavior reflects cognitive processes of the mind, such as intentions [35–37], and is influenced by the user's task [38]. In the context of visual search, previous work typically focused on predicting targets corresponding to specific object instances [3–5]. For example, users were required to search for a specific book [4], a specific binary patterns [3] among other distracting objects. Zelinsky et al. predicted search targets from subjects' gaze patterns during a categorical search task [5]. In their experiments, participants were asked to find two categorical search targets among four visually similar distractors.

In contrast, in this work, we aim to infer the general properties of a search target represented by the object's category and attributes. In this scenario, the search task is guided by the mental model that the user has of the object class rather than a specific instance of an object [13,39]. This presents additional challenges as mental models might differ substantially among subjects. Furthermore, [3–5] required gaze data for training, whereas our approach can be pre-trained on visual data alone, and then combined with gaze data at test time.

### 2.4. Image generation and multi-modal learning

Due to recent advances in representation learning and convolution neural networks, image generation becomes possible. Recently, generative adversarial networks (GANs) [31,40–46,46–50] were used to generate realistic and novel images. GANs consists of two parts: a generator and discriminator. The discriminator is designed to discriminate between generated images and

training data. However, training GANS is a challenging task due to the min-max objective. A stochastic variational inference and learning algorithm was introduced by Kingma and Welling [51]. A lower bound estimator is achieved via the re-parameterization of the variational lower bound. Consequently, the standard stochastic gradient method can be used to optimize the estimator. However, the posterior distribution of latent variables is usually unknown. Yang et al. introduced a general-optimization based approach that uses image generation models and latent priors for posterior inference [11]. They generated images conditioned on visual attributes. In our work, we employ their idea of conditional generative models in the context of inferring search intends from gaze data. We are the first to address the reconstruction of the search target from fixation data, which is more difficult than prediction as it addresses a continuous output space.

### 2.5. Visual experience reconstruction using fMRI

Recent developments in functional magnetic resonance imaging (fMRI) make it possible for neuroscientist to generate links between brain activity and the visual world. In a more advanced setting, Nishimoto et al. reconstructed natural movies from brain activity [8]. They proposed a motion energy encoding method to decode the fast visual information and BOLD signals in the occipitotemporal visual cortex and fit the model separately to individual voxels. In another work, Cowen et al. proposed to reconstruct human faces from evoked brain activity using multi-variant regression and PCA [7]. In their experiment, they asked participants to look at an image and then tried to reconstruct this specific image from fMRI data. All of the above tasks tried to reconstruct the *seen* images. In contrast, our approach decodes the *visual search target* of user's which only resides in the user's mind. Also, we are not using fMRI but gaze data which is arguably more practical and affordable to collect and use.

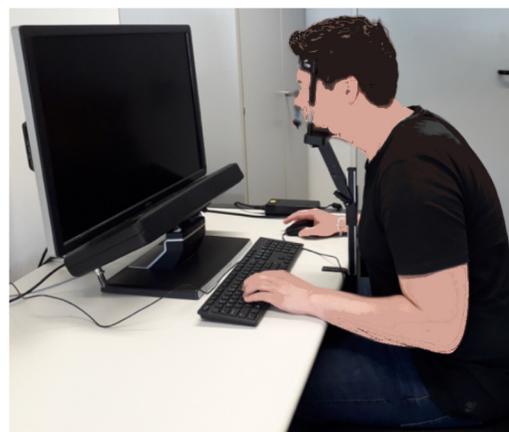
## 3. Data collection

No existing dataset provides image and gaze data that is suitable for our search target prediction task. We, therefore, collected our own gaze data set based on the DeepFashion data set [6].<sup>1</sup> DeepFashion is a clothes data set consisting of 289,222 images annotated with 46 different categories and 1000 attributes. We used the top 10 categories<sup>2</sup> and attributes<sup>3</sup> in our data collection. The training set of DeepFashion was used to train our CNN image model for clothes category and attribute prediction; the validation set was used to train participants for each category and attribute (see below). Finally, the test set was used to build up image collages for which we recorded human gaze data of participants while searching for specific categories and attributes. In the following, we describe our data collection in more detail.

### 3.1. Participants and apparatus

We collected data from 14 participants (six females), aged between 18 and 30 years and with different nationalities. All of them had a normal or corrected-to-normal vision. For gaze data collection we used a stationary Tobii TX300 eye tracker that provides binocular gaze data at a sampling frequency of 300Hz. A chinrest was used to stabilize the head position of the participants. an illustration of a sample test environment can be seen in Fig. 2.

We calibrated the eye tracker using a standard 9-point calibration, followed by a validation of eye tracker accuracy. For gaze



**Fig. 2.** an illustration of a sample test environment. The eye tracker is placed on a table with the test subject behind the chin rest. The subject is instructed to not use any electronic devices that could cause distraction. The room is kept quiet during data collection. No one enters the room or exits the room during the data collection. Only the main author and the subject were in the room

data processing we used the Tobii software with the parameters for fixation detection left at their defaults (fixation duration: 60ms, the maximum time between fixations: 75ms). Image collages were shown on a 30-inch screen with a resolution of 2560x1600 pixels.

### 3.2. Procedure

We first trained participants by showing them exemplary images of all categories and attributes in a game like sessions to familiarize themselves with the categories and attributes. We did not collect any gaze data at this stage. For each category and attribute, we then generated 10 image collages, each containing 20 images. Each target category or attribute appeared twice in each collage at a random location (see Fig. 3 for an example). Participants were then asked to search for ten different categories and attributes on these image collages (see Fig. 3) while their gaze was being tracked. We stress again that we did not show participants a specific target instance of a category or attribute that they should search for. Instead, we only instructed them to find a matching image from a certain category, i.e “dress”, or with a certain attribute, i.e “floral”. Consequently, the search session guided by the mental image of participants from the specific category or attributes. Participants had a maximum of 10 s to find the asked target category or attribute in the collage that was shown full-screen. As soon as participants found a matching target, they were asked to press a key. Afterward, they were asked whether they had found a matching target and how difficult the search had been. This procedure was repeated ten times for ten different categories or attributes, resulting in a total of 100 search tasks.

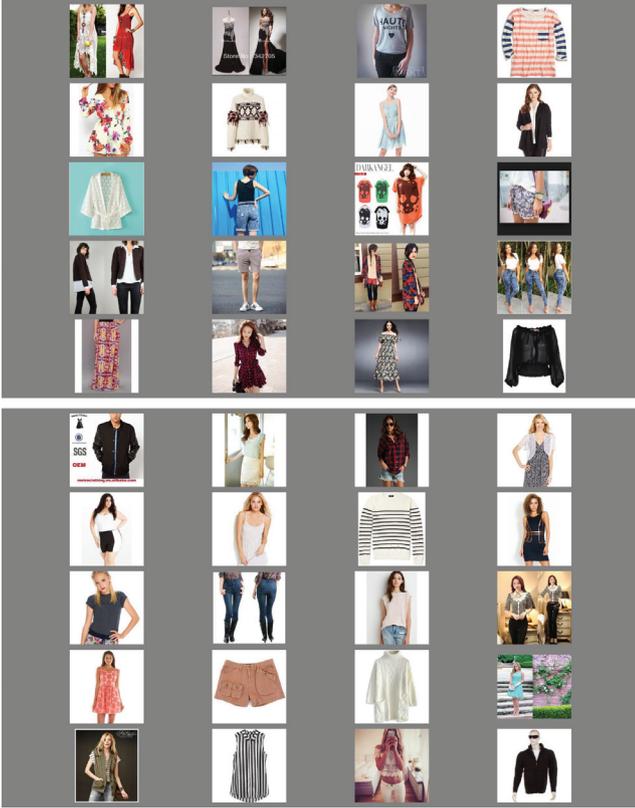
## 4. Integration of gaze into deep learning architectures

In this work, we propose a modular and effective integration scheme which facilitates current deep architectures with gaze data. We address this task by introducing the Gaze Pooling Layer (GPL) that combines CNN architectures with gaze data in a weighting mechanism. We further show the application of our Gaze Pooling Layer in two different frameworks, for prediction and decoding of the visual search targets of users from their gaze data during visual search. The proposed layer is parameter-free and does not need any gaze data to be trained on. In the following, we describe the major components of our method in detail: the Gaze Pooling Layer, search target prediction and search target decoder. Finally, we also discuss different integration schemes across multiple images that

<sup>1</sup> Data set is Available at [GazePooling](#)

<sup>2</sup> Categories: dress, tee, blouse, shorts, tank, skirt, cardigan, sweater, jacket, jean.

<sup>3</sup> Attributes: print, floral, lace, knit, sleeve, maxi, shirt, denim, striped, chiffon.



**Fig. 3.** Sample image collages used for data collection: Attributes (top), Categories (bottom). Participants were asked to find different clothing attributes and categories within these collages.

allow us to utilize gaze information obtained from collages. As a mean of inspecting the internal representation of our Gaze Pooling Layer, we propose Attended Class Activation Maps (ACAM).

#### 4.1. Image encoder

We build on the recent success of deep learning and use a convolutional neural network (CNN) to encode image information [52,53]. Given a raw image  $I$ , a CNN is used to extract image feature map  $F(I)$ .

$$F(I) = \text{CNN}(I) \quad (1)$$

The end-to-end training properties of these networks allows us to obtain domain-specific features. In our case, the network will be trained with data and labels relevant to the fashion domain. As we are interested in combining spatial gaze features with the image features, we use features  $F(I)$  of the last convolutional layer that still has a spatial resolution. This results in a task-dependent representation with spatial resolution. Also, to gain a higher spatial resolution we used the same architecture as describe in [54]. We use their VGGnet-based model where layers after conv5-3 are removed to gain a resolution of  $14 \times 14$ .

#### 4.2. Human gaze encoding

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

$$FDM(g) = \mathcal{N}(g, \sigma_{\text{fix}}), \quad (2)$$

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

$$FDM(G) = \sum_{g \in G} FDM(g) \quad (3)$$

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

$$FDM(G) = \max_{g \in G} FDM(g) \quad (4)$$

#### 4.3. Deep Gaze Pooling Layer

During visual search, users pay attention more to image regions with local similarities to the target in mind. To extract these local regions from the rest of the image features, we introduced the Gaze Pooling Layer. Via this layer, we can ignore regions in which users did not pay attention to, and only encode image features in the interest of users. Hence, these features are not generic image features, rather represents the target of visual search in mind of the user.

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

$$GWFM(I, G) = F(I) \otimes FDM(G). \quad (5)$$

In spirit of [54], we then perform Global Average Pooling (GAP) on each feature channel separately in order to yield a vector-valued feature representation.

$$GAP_{GWFM}(I, G) = \sum_{x,y} GWFM(I, G) \quad (6)$$

#### 4.4. Visual search target inference

The resulted features from the Gaze Pooling Layer encode information about the visual search target of the users. These features can be used to predict the search target of users. For predicting the visual search targets, we add a fully connected and a soft-max layer after the Gaze Pooling Layer.

$$p(C|I, G) = \text{softmax}(W \text{ GAP}_{GWFM}(I, G) + b), \quad (7)$$

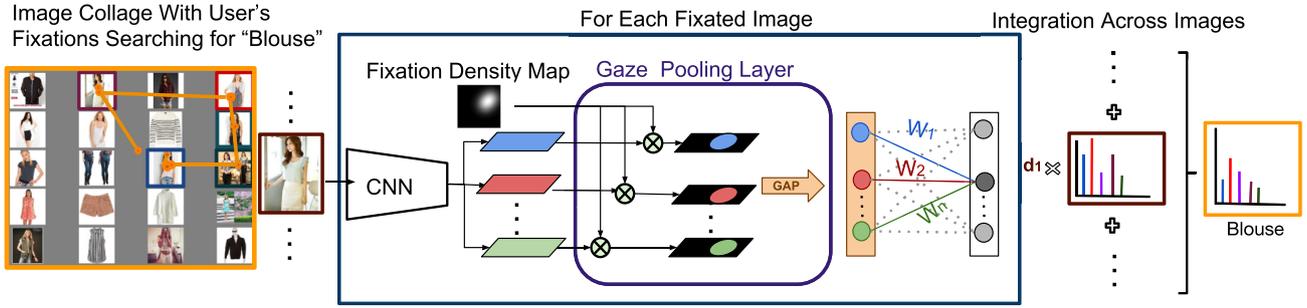
where  $W$  are the learned weights and  $b$  is the bias and  $C$  are the considered classes. The classes represent either categories or attributes depending on the experiment and we decide for the class with the highest class posterior.

In our study, a stimulus is a collage with a set of images  $I_i \in \mathbb{I}$ . During the search task, participants fixate on multiple images in the collage, which generates fixations  $G_i \in \mathbb{G}$  for each image  $I_i$ . Hence, we need a mechanism to aggregate information across images. To do this, we propose a weighted average scheme of the computed posteriors per image:

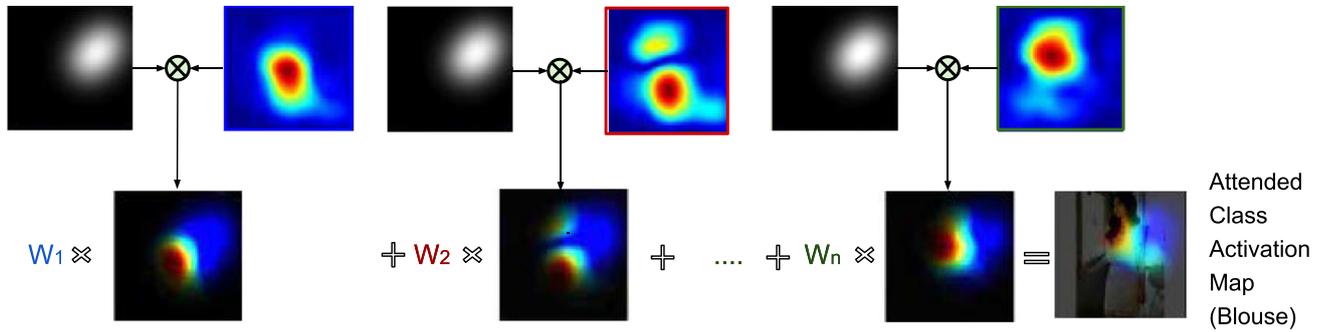
$$p(C|\mathbb{I}, \mathbb{G}) = \sum_i d_i \times p(C|I_i, G_i). \quad (8)$$

We consider for the weights  $d_i$  the total fixation duration of all fixations on image  $I_i$  as well as fixed  $d_i$  which is assumed to be one. The fixation duration  $d_i$  is normalized by total fixation duration of all the fixations on collage  $C$ .

In order to obtain the CNN features maps, we follow [54] and build on the recent VGGnet-GAP model. For our categorization experiments, we fine-tune on a 10 class classification problem on the DeepFashion data set [6]. For attribute prediction, we fine-tune a model with 10 times 2-way classification in the final layer. We used Caffe to train our models using a NVIDIA Tesla 2 × K40m



**Fig. 4.** Overview of our approach. Given a search task (e.g. “Find a blouse”), participants fixate on multiple images in an image collage. Each fixated image is encoded into multiple spatial features using a pre-trained CNN. The proposed Gaze Pooling Layer combines visual features and fixation density maps in a feature-weighting scheme. The output is a prediction of the category or attributes of the search target. To obtain one final prediction over image collages, we integrate the class posteriors across all fixated images using average pooling.



**Fig. 5.** The proposed Gaze Pooling Layer combines fixation density maps with CNN feature maps via a spatial re-weighting (top row). Attended class activation maps are shown in the bottom row, in which the predicted class scores are mapped back to the previous convolutional layer. The attended class activation maps highlight the class-specific discriminative image regions.

12GB GPU card. We perform a validation of the VGGnet image classification performance model in the same setting as [6] and obtained comparable results ( $\pm 5\%$ ) for category and attribute classification. To ensure that the images and collages are not informative of the category or attribute search tasks, we have performed a sanity check by using only the CNN prediction on the images of our collages. The resulting performance is at chance level, which validates our setup as search task information cannot be derived from the images or collages and therefore can only come from the gaze data.

#### 4.5. Visual search target decoding

In order to generate visual search targets of users, we employ a Conditional Variational Auto-Encoder (CVAE) [11]. The conditional variational autoencoder is trained gaze free, for the generation of images of different clothing categories using the deep fashion data set. However, at test time, the decoder is using the gaze weighted feature maps  $GAP_{GWF}(I, G)$ , to decode visual search target of users as illustrated in Fig. 4. Given the posteriors  $p(c|I, G)$ , our goal is to sample the visual search target  $ST$  of the target category  $c$  from

$$P(ST|I, G) = \sum_c P(ST|c)P(c|I, G), \quad (9)$$

In the following, we explain the encoder and search target decoder.

#### 4.6. Category-conditioned image generation model

In contrast to traditional Variational Auto Encoder(VAE), where we have no control on the data generation process, Conditional Variational Auto-Encoder (CVAE) can generate specific data.

Given the category vector  $c \in \mathbb{R}^{d_c}$  and the latent variable  $z \in \mathbb{R}^{d_z}$ , CVAE is a generative model  $p_\theta(I|c, z)$ , which generates image  $I \in \mathbb{R}^{d_I}$ . The generated image  $I$ , is conditioned on the categori-

cal information  $c$  and the randomly sampled latent variable  $z$  from prior distribution  $p(z)$ . In Conditional Variational Auto-Encoder, the auxiliary distribution  $q_\phi(z|I, c)$  is introduced to approximate the true posterior  $p_\theta(z|I, c)$ . The goal of learning process is to find the best parameter  $\theta$  which maximizes the lower bound of the log-likelihood  $\log p_\theta(I|c)$ . Hence the conditional log-likelihood is

$$\log p_\theta(I|c) = KL(q_\phi(z|I, c)||p_\theta(z|I, c)) + \mathcal{L}_{CVAE}(I, c, ; \theta, \phi), \quad (10)$$

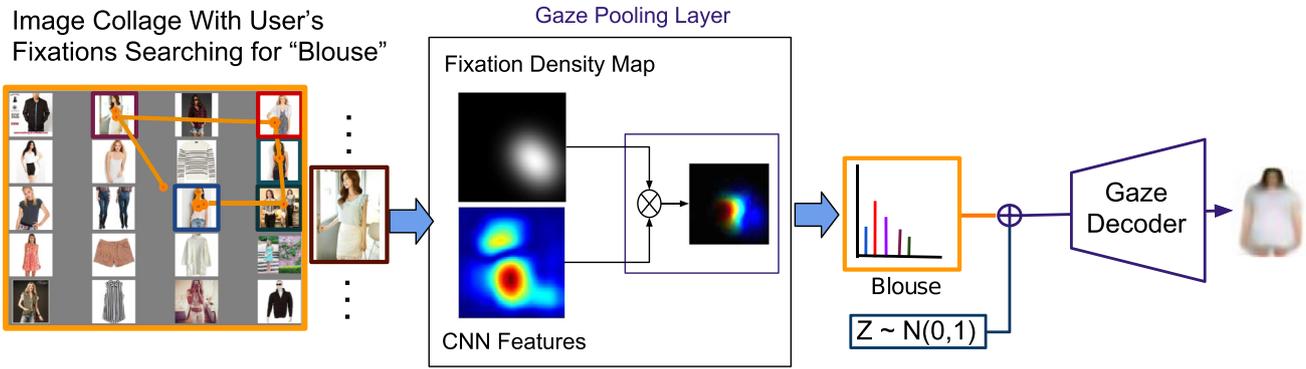
where the variational lower bound

$$\begin{aligned} \mathcal{L}_{CVAE}(I, c, ; \theta, \phi) \\ = -KL(q_\phi(z|I, c)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|c, I)}[\log p_\theta(I|c, z)] \end{aligned} \quad (11)$$

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

The recognition model here is  $q_\phi(z|I, c)$  and the generation model is the conditional distribution  $p_\theta(I|c, z)$ . During training the first term  $KL(q_\phi(z|I, c)||p_\theta(z))$  acts as a regularisation term that minimises the gap between the prior  $p_\theta(z)$  and the proposal distribution  $q_\phi(z|I, c)$ . To generate gaze conditioned image, we replace the recognition network with a VGGNet-16-GAP network, as explained in Section 4.4 during the test time. The  $Z$  is sampled from isotropic Gaussian distribution.

The CAVE, have two convolutional neural networks for recognition and generation. The encoder network consists of 5 convolution layers, followed by 2 fully-connected layers (convolution layers have 64, 128, 256, 256 and 1024 channels with filter size of  $5 \times 5$ ,  $5 \times 5$ ,  $3 \times 3$ ,  $3 \times 3$  and  $4 \times 4$ , respectively; the two fully-connected layers have 1024 and 192 neurons). The category stream is merged with the image stream at the end of the recognition network. The decoder network consists of 2 fully-connected layers,



**Fig. 6.** This image gives an overview of search target decoding. The user is searching for a category “Jean”, the gaze data is recorded during the search task. We encode the gaze information into a semantic representation  $p(C|I, F)$ . The representation is used as condition over the learned latent space to decode the gaze into visualizations of the categorical search target.

followed by 5 convolution layers with 2-by-2 upsampling (fully-connected layers have 256 and  $8 \times 8 \times 256$  neurons; the convolution layers have 256, 256, 128, 64 and 3 channels with filter size of  $3 \times 3$ ,  $5 \times 5$ ,  $5 \times 5$ ,  $5 \times 5$  and  $5 \times 5$ ). Furthermore, we train a classifier (VGGNet-16-GAP) to get class posterior from gaze data  $p(C|I, F)$ . This posterior is used in the category stream of the encoder to generate the search target of our users.

Both networks are trained over the top 10 categories of Deep-fashion. We use the same train, test and validation split as proposed in the deep-fashion dataset. The CVAE is trained to generate images of clothing, conditioned on the top ten categories of the deep fashion data set. We used Adam for stochastic optimization. For the CAVE network, we used a learning rate of 0.0003 and a mini-batch of size 32. Torch is used to train the CVAE using an NVIDIA Tesla 2  $\times$  K40m 12GB GPU card.

#### 4.7. Attended class activation maps

To inspect the internal representation of our Gaze Pooling Layer, we propose the attended class activation map visualization. It highlights discriminative image regions for a hypothesized search target based on CNN features combined with the weights from the gaze data. In this vein, it shares similarities to the CAM of [54] but incorporates the gaze information as an attention scheme. The key idea is to delay the average pooling, which allows us to show spatial maps as also illustrated in In more detail, our network consists of several convolutional layers which the features of the last convolutional layer is weighted by our fixation density map (GWFM). We do global average pooling over the GWFM and use those features for a fully connected layer to get the user attended categories or attributes. Given that our features maps are weighted by gaze data of users, it represents their attended classes. We can identify the importance of the image region for attended categories by projecting back the weights of the output layer onto a gaze-weighted convolutional feature map, which we call Attended Class Activation Map (ACAM):

$$ACAM_c(x, y) = \sum_k w_k^c GWFM_k(I, G) \quad (12)$$

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

## 5. Evaluation

To evaluate our method for search target prediction of categories and attributes, we performed a series of experiments. We first evaluated the effectiveness of our Gaze Pooling Layer, the

**Table 1**

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

Global vs.	---- Category ----			Attribute
	Top1	Top2	Top3	
Local	31% $\pm$ 5	48% $\pm$ 8	62% $\pm$ 8	20% $\pm$ 1
Local	49% $\pm$ 7	68% $\pm$ 6	78% $\pm$ 6	26% $\pm$ 1
Global	52% $\pm$ 6	68% $\pm$ 6	78% $\pm$ 6	25% $\pm$ 1
Local	57% $\pm$ 8	74% $\pm$ 7	84% $\pm$ 4	34% $\pm$ 1

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

#### 5.1. Evaluation of the Gaze Pooling Layer

Fixation information enters our method in two places: The fixation density maps in the Gaze Pooling Layer (Section 4.3) as well as the weighted average across images in the form of fixation duration (see Eq. (8) and Fig. 6). To evaluate the effectiveness of our Gaze Pooling Layer, we evaluate two conditions: “local” makes full use of the gaze data and generates fixation density maps using the fixation location as described in our method section. “global” also generates a fixation density map, but does not use the fixation location information and therefore generates for each fixation a uniform weight across the whole fixated image. Besides, we evaluate two more conditions, where we either used the fixation duration as a weight to the average class posterior of each fixated image (see Eq. (8)) or ignore the duration.

Table 1 shows the result of all 4 combinations of these conditions, with the first column denoting if local or global information was used and the second column whether fixation duration was used. Absolute performance of our best model using local information and fixation duration was 57%, 74%, and 84% on top1-3 accuracy respectively, for the categorization task and 34% accuracy for attributes. The results show a consistent improvement (16 to 18 pp for categories, 6 pp for attributes) across all measures and tasks going from a global to a local representation (first to the second row). Adding the weighting by fixation duration yields another

**Table 2**  
Evaluation of different fixation pooling strategies using average or max pooling.

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

**Table 3**  
Evaluation of different gaze encoding schemes using different per-fixation  $\sigma_{\text{fix}}$ .

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

consistent improvement for both local and global approach (another 6 to 5 pp for categories). Our best method improves overall by 22 to 26 pp on the categorization task and 14 pp on the attributes. The global method without fixation duration (first row) is in a spirit similar to [4] – although the specific application differs. All further experiments will consider our best model (last row) as the reference and justify the parameter choices (average pooling,  $\sigma_{\text{fix}}$ ) by varying each parameter one by one.

## 5.2. Evaluation of the gaze encoding

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

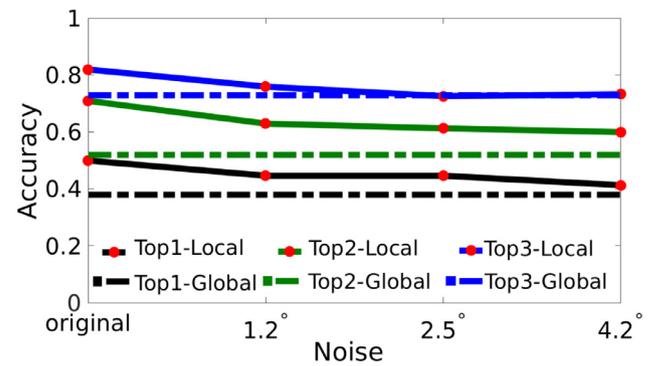
*Effects of fixation representation parameter  $f_{\sigma}$ .* The parameter  $\sigma_{\text{fix}}$  controls the spatial extend of a single fixation in the fixation density maps as described in Section 4. We determined an appropriate setting of this parameter to be  $\sigma_{\text{fix}} = 1.6$  in a pilot study to roughly match the eye tracker accuracy and analyzed here the influence on the overall performance by varying this parameter in a sensible range (given eye tracker accuracy and coarseness of feature map) from 1 to 2 as shown in Table 3. As can be seen from the Table, our method is largely insensitive to the investigated range of reasonable choices of this parameter and our choice of 1.6 is on average a valid choice within that range.

*Fixation pooling strategies.* We evaluated two options for how to integrate single fixations into a fixation density map: Either using average or max pooling. The results are shown in Table 2. As the Table shows, while both options perform well, average pooling consistently improves over the max pooling option.

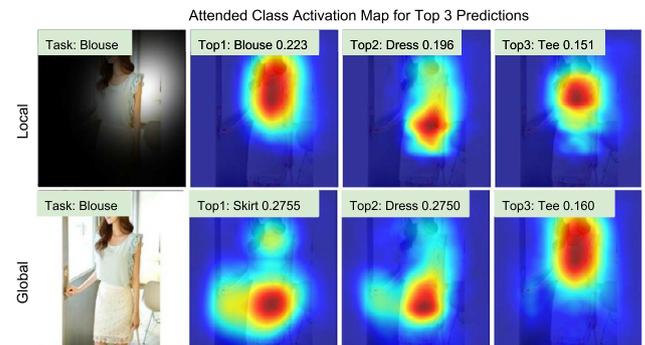
## 5.3. Noise robustness analysis

Several factors such as size and resolution of displays, the visual angle at which the stimuli are presented to the viewers could increase gaze estimation error. However, it remains difficult to study those systematically and exhaustively. Therefore, we decided to study noise as it gives some indication of robustness across a range of causes related to the sensing process.

While our gaze data is recorded with a highly-accurate stationary eye tracker, there are different modalities and types of eye trackers available. One key characteristic in which they differ is the error at which they can record gaze data – typically measured in degrees of visual angle. While our controlled setup provides us with an accuracy of about 0.7 degrees of error, state-of-the-art eye trackers based on webcams, tablets or integrated into



**Fig. 7.** Accuracy for different amounts of noise added to the eye tracking data. Our method is robust to this error which suggests that it can also be used with head-mounted eye trackers or learning-based methods that leverage RGB cameras integrated into phones, laptops, or public displays.



**Fig. 8.** Attended class activation maps of top 3 predictions in local and global method for a given image. Participants were searching for target category “Blouse”. The maps shows the discriminative image regions used for for this search task.

glasses can have up to 4 degrees depending also on the deployment scenario [55]. Therefore, we finally investigated the robustness of our approach w.r.t. different levels of (simulated) noise in the eye tracker. To this end, we sampled noise from a normal distribution with  $\sigma = 1, 3, 5$ . This corresponds roughly to 60, 120 and 200 pixels and 1.2, 2.5 and 4.2 degrees of visual angles and hence covers a realistic range of errors. The results of this evaluation are shown in Fig. 7. As can be seen, our method is quite robust to noise with only a drop of 5 to 10pp for Top3 to Top1 accuracy, respectively – even at the highest noise level. In particular, all the results are consistently above the performance of the corresponding global methods shown as dashed lines in the plot.

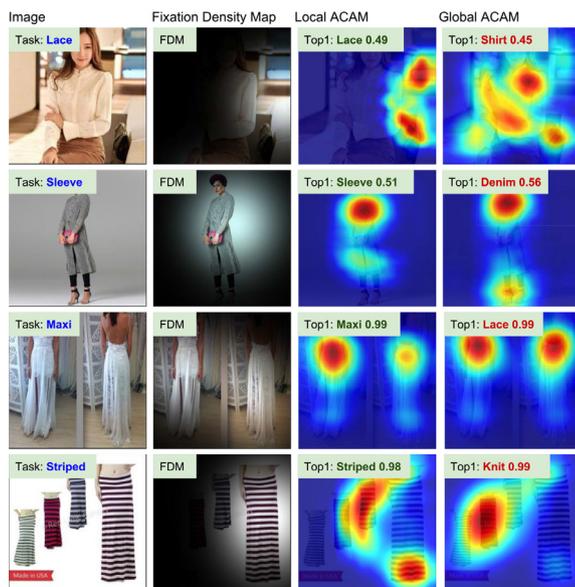
## 5.4. Visualization and analysis of Gaze Pooling Layer on single images

We provide further insights into the working of our Gaze Pooling Layer by showing visual examples of the attended class activation maps, associated fixation density map and search target prediction results. While the quantitative evaluation was conducted on full collages, this is impracticable for inspection. Therefore, we show in the following visualizations and analysis of single images. *Predictions.* Fig. 12 shows results for the categorization task and Fig. 13 for the attribute task. Each of these figures shows the output of the “global” method that uses uniform fixation density map as well as the “local” method that makes full use of the gaze data. We observe that for the “local” method a relevant part of the images is fixated on which in turn leads to correct prediction of the intended search task.

*Attended Class Activation Map (ACAM) Visualization.* Fig. 8 shows the attended class activation map (ACAM) of top 3 predictions, for “local” as well as “global” approach. The “global” method exploits



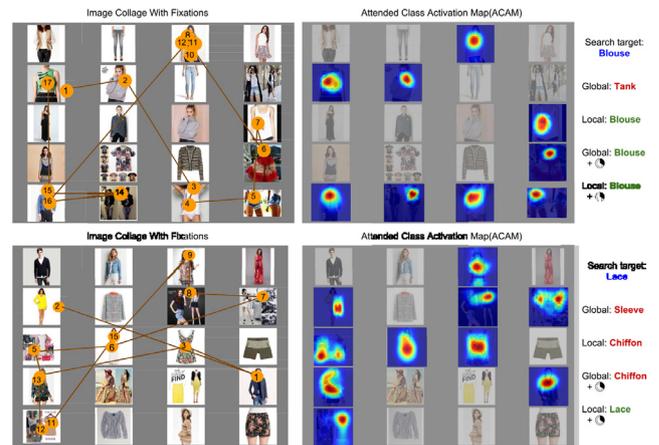
**Fig. 9.** Attended class activation maps of top1 prediction in local and global method for a single fixed image. Participants were searching for the given category. The maps show the discriminative image regions used for this search task



**Fig. 10.** Attended class activation maps of top1 prediction in the local and global method for a single fixed image. Participants were searching for the given attribute. The maps show the discriminative image regions used for this search task

that this image was fixed on - but does not exploit the location information of the fixations. Therefore it reduces in the case of a single image to a standard CAM. E.g. the lower part of the image is activated for “skirt”, and the upper part is activated for “Tee”. One can see that highlighted regions vary across the predicted class. The first row shows the ACAM for the “local” method. It can be seen how the local weighting due to the fixation is selective to the relevant features of the search target, e.g. eliminating the “skirt” responses and retaining the “blouse” responses.

Figs. 9 and 10, shows the attended class activation map (ACAM) of top 1 predictions of the fixed images, for “local” and “global” approach. The left column represents the image and the task of the user, the second column shows fixation density maps of the user



**Fig. 11.** Image collage with fixations of a participant searching for “Blouse” and “Lace”. The right image shows the ACAM of each fixed image in the collage. The last column represents the top 1 prediction for a global and local method without and with fixation durations.

searching for the given task and the last two columns are ACAM for the local and global method.

*Search target prediction over image collages.* In Fig. 11 presents fixation data of one participant searching for category “Blouse” and attribute “Lace”. The posterior of all fixated images are average to get one final prediction overall fixated images. For each fixed image we show the attended class activation map (ACAM) and the result of the global and local method with and without fixation duration.

*Performance over time.* Fig. 16, illustrate the effect of information accumulation over time. The number of fixations varies across participants. Hence, we measure the accuracy of our method using different number of fixations. As shown in Fig. 16, after 8 fixations the prediction accuracy of our model does not improve so much.<sup>4</sup>

### 5.5. Evaluation of the search target decoding

We evaluate our approach on our gaze dataset. Besides qualitative results, we show two user studies. One measures the success in reconstructing meaningful visual representations of visual search targets and the second highlights the importance of a gaze encoder that respects localized information.

*Qualitative results of search target decoding.* Figs. 14, 15 and 17 show qualitative results of our approach. Noise is inherent in gaze encoding. To cope with this challenge, we try different pruning strategies to suppress weak activation in the semantic representation  $c$ . Specifically, we tried four scenarios to decode the visual search target from the gaze. In the first case, we used plain posterior as a conditioned vector. In the remaining cases, we used only the top 1 to top 3 highest activated classes in the posterior as a condition vector for CVAE. All other probabilities are set to zero, and the posterior is re-normalized afterward.

Using directly the posterior of the CNN with the gaze pooling layer causes images that contain several categories rather than the intended visual search target. As shown in Fig. 14, the image reconstructed from unaltered posteriors (first row) is more blurry and does not seem to contain one specific category (e.g. Z1 looks like a Blouse, Z6 is a dress and Z9 is a skirt).

Images from top2 and top1 appear to be more focused on the intended category. Using top2 posteriors generates images which are a mixture between the two posteriors. One can see more details in top2, which is a composition between dress and skirt,

<sup>4</sup> Further illustration of prediction over time could be seen in this [video](#)

Image	Image With FDM	Results	Image	Image With FDM	Results
		True Search target: Jean Local Prediction: Jean Global Prediction: Jacket			True Search target: Jean Local Prediction: Jean Global Prediction: Tee
		True Search target: Short Local Prediction: Short Global Prediction: Dress			True Search target: Blouse Local Prediction: Blouse Global Prediction: Skirt

Fig. 12. Example category responses of the local and global method. Green means correct and red means wrong target prediction (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Image	Image With FDM	Results	Image	Image With FDM	Results
		True Search Target: Floral Local Prediction: Floral Global Prediction: Chiffon			True Search Target: knit Local Prediction: Knit Global Prediction: Sleeve
		True Search Target: Lace Local Prediction: Lace Global Prediction: Sleeve			True Search Target: Maxi Local Prediction: Maxi Global Prediction: Shirt

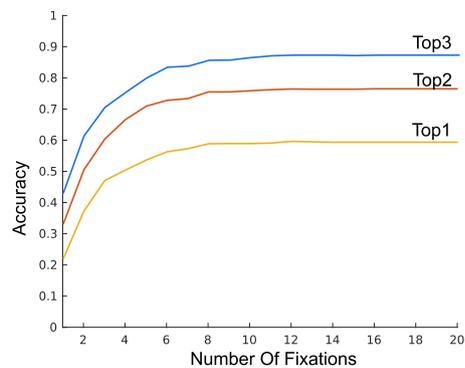
Fig. 13. Example attribute responses of the local and global methods. Green means correct and red means wrong target prediction (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).



Fig. 14. Using all posteriors gives images that contain several categories. Using the only top3 to top1 posterior gives images that contain the intended categories. As we move from posteriors to top1, the decoded image is more localized and contains fewer classes. Top3 images have a full-body part, as we move to top 1, can see only lower body part that contains a skirt.



**Fig. 15.** Top3 and top2 were able to capture the right category; the decoded images contain the target “Tank”. However, due to the wrong prediction for top1 resulted decoding looks like a “Dress”.



**Fig. 16.** Performance of our best model(local+fixation duration) using increasing number of fixations. The top1-3 category accuracy is reported.

whereas the top1 only contains a skirt. Images from top1 are sharper and mostly contain one category. However, if the predicted category is wrong, we can not decode the intended category(last row of Fig. 15). Also Table 2 showed strong recognition performance for top 2 and top 3 classification performance. Hence, using the top2 and top3 is likely to contain information about the target. This is reflected in better reconstructions for the top2 and top3 strategies.

As one can see in Fig. 15, top1 decoded a dress, although the intended category was a tank. The intended category is recovered in the top2 and top3 decodings.

As top2 results are giving images with preferred search targets, for further analysis, we choose decoded images from top2. Fig. 17 shows ten samples for each category using top2 posteriors based on fixations from one user. Our approach can generate images of different categories. The model performs better for several classes, as Jean, Shorts, Skirt, Dress, Tank, Sweater and Tee. Images from the cardigan and jacket are more similar to each other, although there are still differences in the appearance. In particular, images depicting the search target cardigan appear more elongated compared to those for jackets. *User study I: search target recognition.* To assess the accuracy of the search target reconstruction, we run a user study. For each participant and category in the study, we show ten samples from our model. The reconstructions are based on human fixations from our dataset. The users are asked to pick one category among ten categories for a given image. The average accuracy of 19 users across categories was 62%, and the detailed confusion matrix is shown in Table 5. There are confusion between cardigan and jacket, also a skirt and blouse with a dress.

**Table 4**

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

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

All search targets were recognized significantly above chance (10%). Users were most confident about jeans, shorts, and dresses.

*User study II: local vs global gaze encoding.* In this user study, we evaluated the importance of local gaze information for the decoding of visual search targets. Our full model is denoted as “local” here, as it uses the complete gaze information – in particular the fixation location on the image. We compare it to a “global” model which uses gaze information – but only to the extent that we know an image was fixated without knowing the exact location. This also connects to the analysis performed on the recognition task for the gaze pooling layer. We ask how much of a difference these approaches make in terms of search target reconstruction. In this user study, each participant saw two rows of search target reconstructions. One row was generated by the local, the other by the global method (Fig. 18). The users were instructed to select the row, which matches the best, the given search target category. The users selected the local encoding method in 65% of the cases. The chance level for this experiment is 50% as for each image the participants do a binary task. Consequently, the gain is  $65\%/50\% = 13\%$ . The gain indicates how much performance of users differs from a random selection. Also, we performed the Chi-Square Goodness of Fit Test to investigate the significance of our result. The null hypothesis was that both local and global decoding are equal. The  $\chi^2$  value is 6.914 and P-Value is 0.009. Hence, the result is significant at  $p \leq 0.05$  and therefore, local information is key for improved search target reconstruction. Detailed results are shown in Table 4.

## 6. Discussion

In this work, we studied the problem of predicting and decoding categories and attributes of search targets from gaze data. Table 1 shows strong performance for both tasks. Our Gaze Pooling Layer represents a modular and effective integration of visual and gaze feature that is compatible with modern deep learning architectures. Therefore, we would like to highlight three features that are of particular practical importance.



Fig. 17. Each row is the decoded search target of a user for the given category using only top2 posteriors. Each column is for different samples of  $z$  from a normal distribution. As one can see the decoded search targets are distinctive from one another and they represent their corresponding categories properly.

Target = T-Shirt

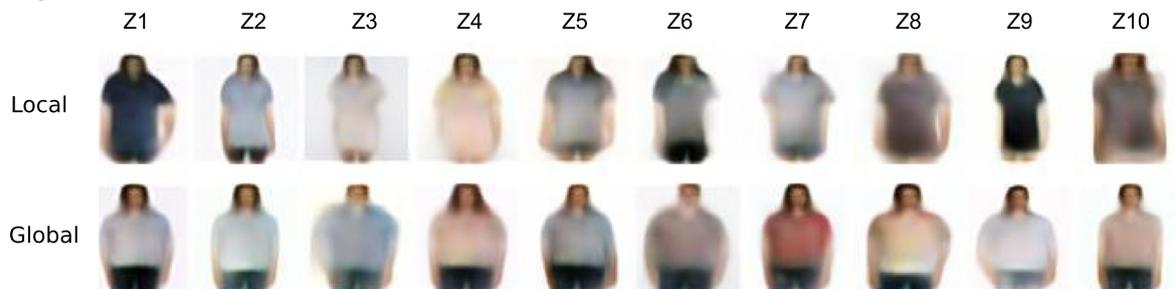


Fig. 18. Example image used in our second user study. For each category, users need to select between local and global decoded target. The local method encodes the gaze data using gaze-pooling layer which benefits from user intended local image regions.

**Table 5**

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

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

### 6.1. Parameter free integration scheme

First, our proposed integration scheme is parameter-free. We introduce a single parameter  $\sigma_{\text{fix}}$  but the gaze encoding is only input to the integration scheme and, also, the method turns out to be not sensitive to the choice (see experiments in Section 4).

### 6.2. Training from visual data

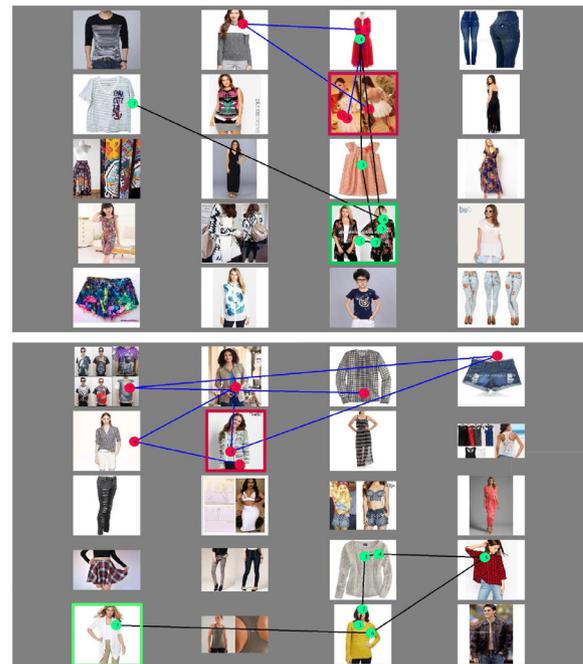
Second, fixing the fixation density maps to uniform maps yields a deep architecture similar to a GAP network that is well-suited for various classification tasks. While this no longer addresses the task of predicting categories and attributes intended by the human in the loop, it allows us to train the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

### 6.3. Training free gaze deployment

Gaze data is time-consuming to acquire – which makes it rather incompatible with today’s data-hungry deep learning models. In our model, however, the fixations density maps computed from the gaze data can be understood as spatially localized feature importance that is used to weight feature importance in the spatial image feature maps Fig. 8. Our results demonstrate that strong performance can be obtained with this re-weighting scheme without the need to re-train with gaze data. As a result, our approach can be deployed without any gaze-specific training. This result is surprising, in particular as the visual model on its own is entirely uninformative without gaze data on the task of search target prediction and decoding (as we have validated in Section 5.1. We believe this simplicity of deployment is a critical feature that makes the use of gaze information in deep learning practical.

### 6.4. Biases in mental model of search targets among users

To illustrate the challenges our Gaze Pooling Layer has to deal with in terms of the variations in the observed gaze data, we show example fixation data in Fig. 19. In each image, fixation data of two participants (red and green dots) is overlaid over a presented collage. Although both participants had the same search target (top: attribute ‘Floral’; bottom: category ‘Cardigan’), we observe a drastically different fixation behavior. One possible explanation is that the mental models of the same target category or attribute



**Fig. 19.** Example fixation data of 2 participants (red and green dots) with search target attribute=‘Floral’ on top and category=‘Cardigan’ below (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

can vary widely depending on personal biases [13]. Despite these strong variations in the gaze information, our Gaze Pooling Layer allows us to predict the correct answer in all 4 cases. The key to this success is aggregating relevant local visual feature across all images in the collage, that in turn represent one consistent search target in terms of categories and attributes.

### 6.5. Privacy risks and mitigation strategies

While previous work has already illustrated the richness of information contained in gaze data, our work contributes to the ease of integration, connecting gaze data to powerful deep learning architectures as well as showing for the first time reconstruction of search targets. We are aware that these techniques will most likely lead to even more powerful techniques for extraction of seman-

tic information from gaze data as well as a new quality of the extracted information, as visual information can now be reconstructed. Hence, we see the necessity to not only raise awareness of the opportunities, but equally to the risks pertaining to those new methods. Recent work has already picked up on this challenge by protecting privacy related to gaze based inference and wearable cameras, e.g. using hardware solutions [56] or software/algorithmic solutions such as differential privacy [57].

## 7. Conclusion

We introduce the first approach to predict and decode the visual search target of users from their gaze data. This task is very difficult as the target only resides in the user mind. Our approach that addresses both tasks is facilitated by a novel Gaze Pooling Layer that integrates gaze and visual data in a modular way. We validate our approach for search target prediction in a quantitative experiment as well as two user studies for search target decoding, showing that the decoded target leads to human recognizable visual representations as well as highlighting the importance of localized gaze information. We like to emphasize that due to the training setup, the method remains highly practical and applicable, as no large scale gaze data had to be collected or used. The key is rather the utilization of a semantic layer that connects the gaze encoder with the image features for prediction and conditional generative image model. We believe that our modular approach of integrating gaze data into standard deep learning architecture will further stimulate and facilitate research in this interesting research direction.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## CRediT authorship contribution statement

**Hosnieh Sattar:** Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization, Investigation. **Mario Fritz:** Conceptualization, Supervision, Validation, Writing - review & editing. **Andreas Bulling:** Conceptualization, Supervision, Validation, Writing - review & editing.

## Acknowledgments

A. Bulling was supported by the [European Research Council](#) (ERC; grant agreement 801708).

## References

- [1] A. Bulling, J.A. Ward, H. Gellersen, G. Tröster, Eye Movement Analysis for Activity Recognition Using Electrooculography, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (4) (2011) 741–753, doi:10.1109/TPAMI.2010.86.
- [2] J. Steil, A. Bulling, Discovery of everyday human activities from long-term visual behaviour using topic models, in: *Proceedings of the UbiComp*, 2015.
- [3] A. Borji, A. Lennartz, M. Pomplun, What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations, *Neurocomputing* (2014).
- [4] H. Sattar, S. Müller, M. Fritz, A. Bulling, Prediction of search targets from fixations in open-world settings, in: *Proceedings of the CVPR*, 2015.
- [5] G.J. Zelinsky, Y. Peng, D. Samaras, Eye can read your mind: Decoding gaze fixations to reveal categorical search targets, *J. Vis.* 13 (14) (2013).
- [6] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, *Proceedings of the CVPR*, 2016.
- [7] A.S. Cowen, M.M. Chun, B.A. Kuhl, Neural portraits of perception: Reconstructing face images from evoked brain activity, *NeuroImage* 94 (2014), doi:10.1016/j.neuroimage.2014.03.018.
- [8] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J.L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies, *Current Biol.* 21 (19) (2011), doi:10.1016/j.cub.2011.08.031.
- [9] Y. Sato, Y. Sugano, A. Sugimoto, Y. Kuno, H. Koike, Sensing and controlling human gaze in daily living space for human-harmonized information environments, in: *Human-Harmonized Information Technology, Volume 1*, Springer Japan, 2016, pp. 199–237.
- [10] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Mpiigaze: Real-world dataset and deep appearance-based gaze estimation, *IEEE Trans. Pattern Anal. Mach. Intell.* (TPAMI) (2018), doi:10.1109/TPAMI.2017.2778103.
- [11] X. Yan, J. Yang, K. Sohn, H. Lee, Attribute2image: Conditional image generation from visual attributes, in: *Proceedings of the ECCV*, 2016.
- [12] K.G.A. Kovashka, D. Parikh, Whittlesearch: Interactive image search with relative attribute feedback, *Int. J. Comput. Vis.* 115 (2) (2012).
- [13] M. Ferecatu, D. Geman, A statistical framework for image category search from a mental picture, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (6) (2009) 1087–1101, doi:10.1109/TPAMI.2008.259.
- [14] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, C.C. Loy, Sketch me that shoe, in: *Proceedings of the CVPR*, 2016.
- [15] M.P. Eckstein, Visual search: a retrospective, *J. Vis.* 11 (5) (2011) 14, doi:10.1167/11.5.14.
- [16] X. Chen, G.J. Zelinsky, Real-world visual search is dominated by top-down guidance, *Vision Res.* 46 (24) (2006) 4118–4133.
- [17] D.T. Levin, Y. Takarae, A.G. Miner, F. Keil, Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision, *Percept. Psychophys.* 63 (4) (2001) 676–697.
- [18] M.B. Neider, G.J. Zelinsky, Searching for camouflaged targets: Effects of target-background similarity on visual search, *Vision Res.* 46 (14) (2006) 2217–2235.
- [19] H. Sattar, A. Bulling, M. Fritz, Predicting the category and attributes of visual search targets using deep gaze pooling, in: *Proceedings of the ICCVW*, 2017, doi:10.1109/ICCVW.2017.322.
- [20] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: *Proceedings of the CVPR*, 2015.
- [21] J. Xu, M. Jiang, S. Wang, M.S. Kankanhalli, Q. Zhao, Predicting human gaze beyond pixels, *J. Vis.* 14 (1) (2014).
- [22] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Eckstein, B. Manjunath, From where and how to what we see, in: *Proceedings of the ICCV*, 2013, doi:10.1109/ICCV.2013.83.
- [23] G. Papadopoulos, K. Apostolakis, P. Daras, Gaze-based relevance feedback for realizing region-based image retrieval, *IEEE Trans. Multimed.* 16 (2) (2014) 440–454, doi:10.1109/TMM.2013.2291535.
- [24] I. Shcherbatyi, A. Bulling, M. Fritz, GazeDPM: Early Integration of Gaze Information in Deformable Part Models, arXiv:1505.05753, 2015.
- [25] E. Marinoiu, D. Papava, C. Sminchisescu, Pictorial human spaces. how well do humans perceive a 3D articulated pose? in: *Proceedings of the ICCV*, 2013.
- [26] R. Subramanian, V. Yanulevska, N. Sebe, Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements, in: *Proceedings of the ACM-MM*, 2011.
- [27] S. Mathe, C. Sminchisescu, Multiple instance reinforcement learning for efficient weakly-supervised detection in images, arXiv:1412.0100, 2014.
- [28] A. Mishra, Y. Aloimonos, C.L. Fah, Active segmentation with fixation, in: *Proceedings of the ICCV*, 2009.
- [29] T. Toyama, T. Kieninger, F. Shafait, A. Dengel, Gaze guided object recognition using a head-mounted eye tracker, in: *Proceedings of the ETRA*, 2012.
- [30] J. Steil, M.X. Huang, A. Bulling, Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets, in: *Proceedings of the International Symposium on Eye Tracking Research and Applications (ETRA)*, 2018, doi:10.1145/3204493.3204538.
- [31] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [32] Y. Sugano, A. Bulling, Seeing with Humans: Gaze-Assisted Neural Image Captioning, arXiv:1608.05203, 2016.
- [33] A.B. Vasudevan, D. Dai, L.V. Gool, Object referring in videos with language and human gaze, *CoRR abs/1801.01582* (2018).
- [34] N. Kareszli, Z. Akata, B. Schiele, A. Bulling, Gaze embeddings for zero-shot image classification, in: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [35] F.J. Brigham, E. Zaimi, J.J. Matkins, J. Shields, J. McDonnough, J.J. Jakubecy, The eyes may have it: Reconsidering eye-movement research in human cognition, pp. 39–59. 10.1016/S0735-004X(01)80005-7
- [36] C.L. Kleinke, Gaze and Eye Contact: A Research Review, *Psychol. Bull.* 100 (1) (1986) 78–100.
- [37] M.F. Land, S. Furneaux, The knowledge base of the oculomotor system, *Philosoph. Trans. R. Soc. Lond. B Biol. Sci.* 352 (1358) (1997) 1231–1239, doi:10.1098/rstb.1997.0105.
- [38] A.L. Yarbus, B. Haight, L.A. Riggs, *Eye Movements and Vision*, 2, Plenum press New York, 1967.
- [39] S.P. Wilson, J. Fauqueur, N. Boujema, Mental Search in Image Databases: Implicit Versus Explicit Content Query, Springer Berlin Heidelberg, 2008.
- [40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *Proceedings of the NIPS*, 2014.
- [41] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text-to-image synthesis, in: *Proceedings of the ICML*, 2016.
- [42] E.L. Denton, S. Chintala, A. Szlam, R. Fergus, Deep generative image models using a Laplacian pyramid of adversarial networks, in: *Proceedings of the NIPS*, 2015.

- [43] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, arXiv (2016).
- [44] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A. Efros, Context encoders: Feature learning by inpainting, in: *Proceedings of the CVPR, 2016*.
- [45] J. Kossaifi, L. Tran, Y. Panagakis, M. Pantic, Gagan: Geometry-aware generative adversarial networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [46] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, M. Fritz, Disentangled person image generation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [47] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, Y.-Y. Chuang, Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [49] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social gan: Socially acceptable trajectories with generative adversarial networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [50] K. Regmi, A. Borji, Cross-view image synthesis using conditional gans, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.
- [51] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: *Proceedings of the ICLR, 2013*.
- [52] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv:1312.6034 (2013).
- [53] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Proceedings of the NIPS, 2012*.
- [54] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: *Proceedings of the CVPR, 2016*.
- [55] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: *Proceedings of the CVPR, 2015*, doi:10.1109/CVPR.2015.7299081.
- [56] J. Steil, M. Koelle, W. Heuten, S. Boll, A. Bulling, Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features, in: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA), 2019*.
- [57] J. Steil, I. Hagedstedt, M.X. Huang, A. Bulling, Privacy-aware eye tracking using differential privacy, in: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA), 2019*.



**Hosnieh Sattar** is a Postdoctoral Research Associate at Roche Pharma AG (Switzerland, Basel). Before she was a PhD student at Max-Planck-Institut für Informatik, Saarland Informatics Campus. She received her MSc. in Visual Computing from Saarland University, Germany, in 2014. Her research interests include human-computer interaction, machine learning, and computer vision.



**Mario Fritz** is faculty member at the CISPA Helmholtz Center for Information Security and a Professor at the University of Saarland. Before, he was a senior researcher and research group head at the Max Planck Institute for Informatics and PostDoc at International Computer Science Institute and UC Berkeley. His current work focuses on trustworthy information processing and the intersection of Artificial Intelligence and Machine Learning with Security and Privacy.



**Andreas Bulling** is Full Professor of Human-Computer Interaction and Cognitive Systems at the University of Stuttgart, Germany. Before, he was a Feodor Lynen and Marie Curie Research Fellow at the University of Cambridge, UK, Senior Researcher at the Max Planck Institute for Informatics, and an Independent Research Group Leader at Saarland University, Germany. His research interests include computer vision, machine learning, and human-computer interaction.