Sample, Nudge and Rank: Exploiting Interpretable GAN Controls for Exploratory Search

Yang Liu
University of Helsinki
Helsinki, Finland
yang.liu@helsinki.fi

Alan Medlar
University of Helsinki
Helsinki, Finland
alan.j.medlar@helsinki.fi

Dorota Głowacka
University of Helsinki
Helsinki, Finland
dorota.glowacka@helsinki.fi

ABSTRACT

Exploratory search is characterized by open-ended search tasks and uncertainty with respect to the clarity of users' information needs. In the context of image retrieval, generative adversarial networks (GANs) present numerous opportunities for satisfying the information needs of users engaged in exploratory search compared to a collection of images. In this article, we present a novel approach for performing exploratory search on a GAN's image space using interpretable GAN controls that can be summarized as *sample*, *nudge*, and *rank*. At each search iteration, we *sample* images from the GAN's latent space. We implement faceted search by *nudging* the sampled images towards regions of the latent space containing the attributes associated with selected facets. Lastly, we *rank* the nudged images using reinforcement learning with relevance feedback. We present a comprehensive evaluation of the proposed approach, incorporating results from simulations and a user study. In simulation, we show that our approach efficiently adapts to user preferences, while preserving a high-level of image diversity. In the user study (N=30), a majority of participants (23/30) preferred our system to the baseline. Concordant with simulation results, users reported both higher perceived search efficiency and image diversity compared to the baseline. Indeed, due to the baseline system's dependence on a warm-start procedure, users of our system examined significantly fewer images while achieving task outcomes of similar subjective quality.

CCS CONCEPTS

• **Information systems** → **Search interfaces**; **Users and interactive retrieval**.

KEYWORDS

exploratory search, image retrieval, GANs, contextual bandits, Thompson sampling

ACM Reference Format:
Yang Liu, Alan Medlar, and Dorota Głowacka. 2024. Sample, Nudge and Rank: Exploiting Interpretable GAN Controls for Exploratory Search. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3640543.3645156

IUI '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645156

1 INTRODUCTION

Exploratory search is a broadly defined search process that involves significant cognitive processing and interpretation [62]. Unlike known item search, where users search for discrete, well-defined facts or specific documents, exploratory search tasks are open-ended and have ill-defined success criteria [2, 40]. Indeed, exploratory search is characterized by uncertainty with respect to the clarity of users' information needs, the scope of available documents and the nature of appropriate search outcomes [62, 63]. Exploratory search is, therefore, considered challenging [2, 16, 41] and requires search systems to provide additional support to help users achieve their search goals [3, 42].

In the context of image retrieval, generative adversarial networks (GANs) present numerous opportunities for satisfying the complex information needs associated with exploratory search [32]. First, GANs are generative models that map points in a continuous latent space to photorealistic images of, for example, human faces [25, 26]. This representation allows GANs to smoothly interpolate between images and, therefore, provides a significantly expanded search space for users to explore. Second, GANs generate images with uniform attributes, such as image dimensions, resolution and similar subject orientations [23], making it easier for users to compare and provide feedback on different images. Third, recent advances in interpretable GAN controls allow for fine-grain control over the image generation process [64], making it possible, for example, to manipulate an image by changing a person's hair color, while keeping other facial attributes constant. This potentially allows users to truly satisfy their search goals, without needing to compromise on a "close enough" image that is merely adequate – as is often the case with traditional interactive image retrieval systems based on collections of discrete images [11]. To our knowledge, however, there are only two studies that have investigated using GANs for interactive image retrieval [32, 58]. Ukkonen et al. used an augmented version of Rocchio's algorithm to perform a greedy search of the latent space [58]. Their system, however, did not provide any support for users performing exploratory search. Kropotov et al. investigated how Gaussian Process bandits could be used to support exploratory search of GANs [32]. Unfortunately, this approach is computationally intensive and was only evaluated in simulation.

In this article, we present a novel approach for exploratory search of GAN image space called ***Sample***, ***Nudge***, and ***Rank*** (SNR). At each search iteration, we ***sample*** images from the GAN's latent space. We implement faceted search [67] using supervised interpretable GAN controls [53]. In our system, facets work by ***nudging*** the sampled images towards regions of the image space that contain the specific attribute associated with selected facets. Next, the nudged images are ***ranked*** using reinforcement learning with

relevance feedback [34]. We use unsupervised interpretable GAN controls to extract features [15] and Thompson sampling, a linear contextual bandit algorithm, to balance exploration (showing more diverse images) and exploitation (showing images predicted to be relevant) [7]. To achieve high performance, all computations are performed in the latent space, only performing the costly image generation procedure for images that are shown to the user. The main contributions of this paper are as follows:

- We present a novel approach for exploratory search of GAN image spaces called Sample, Nudge and Rank. More broadly, our approach can be viewed as a general framework for integrating generative models into interactive systems, and facilitating users' exploration of complex latent spaces.
- We demonstrate how to implement faceted search and relevance feedback - two fundamental interaction mechanisms in exploratory search - using supervised and unsupervised interpretable GAN controls, respectively.
- We validate the effectiveness of our approach in simulation and show that our approach efficiently accommodates user preferences while maintaining image diversity.
- We present a comprehensive user study (N=30), where participants where situated as the casting director of a future Harry Potter movie. A majority of study participants (23/30) preferred using our system to complete the search tasks compared to the baseline.

2 RELATED WORK

In this section, we review related research on interactive image retrieval, GAN architectures and manipulation, and their use in information retrieval.

2.1 Interactive Image Retrieval

Interactive image retrieval describes the iterative process of searching for images through a dialog between users and search systems [57]. While early image retrieval systems used textual search queries [20], users had difficulties specifying their search goals in a way that would match their query terms with textual tags associated with specific images [39, 57]. The shortcomings of these systems lead to increased interest in content-based image retrieval (CBIR), where image features are derived from the images themselves [56, 72]. CBIR was initially based on hand-crafted features (e.g. [38, 56]), but more recent systems use deep learning techniques for image representation learning [31, 55] or rely on cross-modal interactions as in image-text retrieval [48, 59, 71]. In the case of GANs, the representation space is not designed with interactive image retrieval in mind and overcoming this issue is the focus of this article.

User interface design is an important topic in CBIR, with substantial effort going into developing methods for specifying visual queries based on, for example, images [60, 72], sketching [6], color maps [60] and concept maps [65]. Despite such advances, CBIR can still fail to capture users' underlying search intents due to the "semantic gap" that comes from trying to describe high-level semantic concepts with low-level visual features [66, 72]. Alternative approaches based on relevance feedback have been developed to allow searchers to flag relevant images to iteratively refine the scope of their search without needing to explicitly describe their search goals [34, 66]. Relevance feedback has been combined with reinforcement learning to trade off exploration and exploitation in CBIR [29, 34] and in exploratory search more generally [18, 19]. Recent approaches have started to move beyond relevance feedback to using relative attributes [30] and natural language [14] to critique search results, however, such techniques may be unsuitable for search tasks that involve browsing and exploration. In this paper, we use both faceted search and relevance feedback to bridge the semantic gap for users performing interactive image retrieval for exploratory search tasks.

2.2 GAN Architectures and Manipulation

Generative adversarial networks (GANs) are deep generative models that learn a mapping function from a latent space ($\mathcal{Z}$-space) to image space [13]. GANs use adversarial training to produce two networks: a generator, that learns to generate synthetic data with the same distributional properties as the training data, and a discriminator, that is trained to distinguish between real and synthetic data [13]. In this section, we briefly introduce relevant research about GAN architectures and manipulation.

2.2.1 GAN Architectures. Research into GANs initially focused on developing better architectures and training schemes to improve the quality of generated images [64]. For example, DCGAN uses convolutions in both the generator and discriminator to improve image quality [68]. BigGAN adopted the self-attention module from SAGAN [69] and introduced the "truncation trick" to trade off image fidelity and diversity [5]. PCGAN progressively grows its architecture during training, facilitating image synthesis at higher resolutions [23]. More recently, StyleGAN [24–26] introduced an additional intermediate latent space, $\mathcal{W}$-space, to enable better control over image generation by disentangling features.

2.2.2 GAN Manipulation. Recent studies have investigated how GAN image synthesis can be manipulated using vector arithmetic in latent space, i.e. by adding a vector corresponding to a given attribute [1, 12, 21, 53]. These vectors are often referred to as interpretable directions or controls, and have been identified using both supervised and unsupervised approaches [64]. For example, Shen et al. trained a classifier and used the normal vector of the separating hyperplane as an interpretable direction [53]. Whereas Härkönen et al. identified interpretable directions using principal component analysis (PCA) on an intermediate representation of the GAN [15]. In our work, we used a combination of these two approaches to implement faceted search and relevance feedback. Other approaches for GAN manipulation include GANalyze [12], StyleFlow [1] and SeFa [54]. We refer readers to a recent survey by Xia et al. [64] for more details.

2.3 GANs in IR

Previous studies of applying GANs in IR can be roughly divided into two categories: (i) using GANs as ranking functions [61, 70] and (ii) searching through GAN space [32, 58]. We are only aware of two studies focusing on interactive image retrieval from GANs. Ukkonen et al. proposed an interactive image retrieval system based on

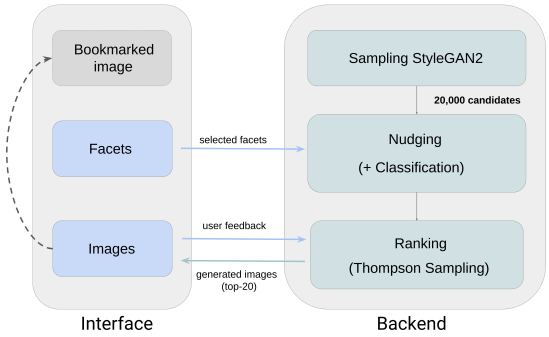


Figure 1: System overview: At each iteration, the system randomly samples images from StyleGAN2, nudges and filters those images, and ranks them using Thompson sampling. Searchers can provide positive and negative relevance feedback to guide the search process. Searchers can bookmark the current best image seen during the search session.

Rocchio's algorithm to generate images according to users' information needs [58]. This was the first method to use relevance feedback to interactively search a GAN's latent space, but provided no support for exploratory search and only evaluated system performance based on target image descriptions. Kropotov et al. investigated the use of Gaussian Process (GP) bandits for exploratory search over a GAN's image space [32]. While GP bandits outperformed Ukkonen et al. in simulation, it was shown to be too computationally intensive to be used interactively and was, therefore, not evaluated with users. In this article, we demonstrate that it is possible to support exploratory search of GANs, while maintaining interactivity and satisfying users.

3 APPROACH

We implemented the Sample, Nudge and Rank approach in a practical exploratory search system [36]. The system has two components: (1) an interface that contains standard features for exploratory search including faceted search, relevance feedback and bookmarking, and (2) a backend that uses a GAN to generate images and then ranks those images using Thompson sampling. In our implementation, we used StyleGAN2 [26] to generate images.

3.1 System Overview

Figure 1 shows an overview of the system. Each search session spans multiple iterations. At each iteration, searchers are shown 20 images of human faces generated by StyleGAN2. The initial set of images shown to users in the first iteration is randomly sampled, but in subsequent iterations images are ranked on the basis of user feedback. Searchers can provide feedback using two interaction mechanisms: faceted search and relevance feedback. Faceted search is used to quickly filter out images that do not have specific facial features, whereas relevance feedback allows searchers to indicate which images are positive and negative examples of their search goals. Searchers proceed to the next iteration by clicking the "Next" button, which sends the currently selected facets and images that received relevance feedback to the search engine backend.

The backend generates images for the next iteration using the following procedure: (i) sample random images from StyleGAN2, (ii) "nudge" the random images towards the facial attributes associated with the selected facets (we explain what it means to nudge an image in Section 3.4.2), (iii) filter out images using the classifiers associated with the selected facets, (iv) update the posterior distribution from the Thompson sampling model using relevance feedback, (v) sample model weights from the posterior distribution and rank the remaining images on the basis of the estimated probability of relevance and, (vi) return the top-20 images to the search interface. The search session ends when the user clicks the "End" button. Each of these steps is described in detail below.

3.2 Interface

The search interface is shown in Figure 2. Searchers interact with facets, relevance feedback and are able to bookmark the best image that currently satisfies their information needs.

3.2.1 Faceted Search. Facets are used to filter search results to include only images with specific facial features (Figure 2, part A). We currently support the following categories of facet: sex (male, female), hair color (blonde, brown and black), hair style (bangs, wavy and straight) and a miscellaneous category (glasses, children). Facets are mutually exclusive in each category. We provide these particular facets as each one is implemented using a classifier and we are limited by the available training data (described in Section 3.4.1).

3.2.2 Relevance Feedback. Searchers can provide positive relevance feedback to images that satisfy their information needs and negative feedback to images that do not (Figure 2, part B). The interface is toggled between positive and negative relevance feedback modes by clicking the "Good" and "Bad" buttons, respectively (Figure 2, part C). Images that receive positive feedback are highlighted with a green border, whereas images receiving negative feedback are bordered in red. Searchers can click on images that received relevance feedback during the current search iteration to put them back into a neutral state of having received neither positive nor negative relevance feedback. Both positive and negative relevance feedback is explicit: when searchers click the "Next" button, only the images that received feedback are passed on to the backend to generate images for the next iteration.

3.2.3 Bookmarks. Searchers can bookmark the best image currently seen during their search session. Bookmarked images are highlighted with a yellow border in the right margin of the interface (Figure 2, part D). Our current implementation of the interface only supports a single bookmark to force study participants to be explicit about their search progress during experiments. In a non-experimental version of our search system, users would be allowed to bookmark any number of images.

3.3 Sampling StyleGAN2

We used StyleGAN2 [26][1] to generate images of faces using a pretrained model[2] based on the Flickr-Faces-HQ (FFHQ) data set [25].

[1] https://github.com/NVlabs/stylegan2-ada-pytorch
[2] https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada/pretrained/ffhq.pkl

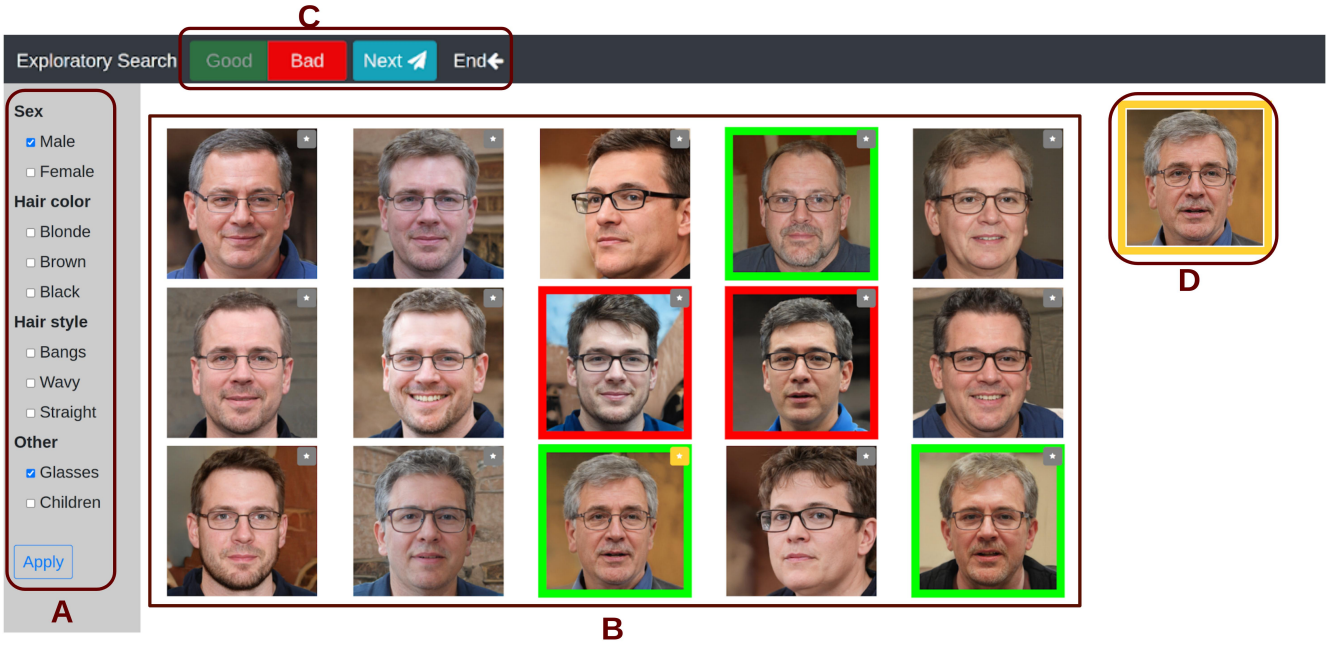


Figure 2: Search engine interface. *Part A:* search facets are shown in the left margin; *Part B:* 20 images are shown at each iteration (image is cropped). Images with green/red boxes have received positive/negative relevance feedback; *Part C:* "Good" and "Bad" buttons toggle between positive and negative relevance feedback modes. "Next" button submits feedback and initiates the next iteration. "End" button ends the current search session; *Part D:* the bookmarked image is shown in the right margin.

3.3.1 Image Generation. StyleGAN2 consists of a mapping network between $\mathcal{Z}$-space and $\mathcal{W}$-space, and a synthesis network that generates synthetic images on the basis of vectors from $\mathcal{W}$-space. To generate an image, we sample a latent vector $\boldsymbol{z}$ from a standard multivariate normal distribution (i.e. with zero mean and identity covariance), map $\boldsymbol{z}$ to an intermediate latent vector $\boldsymbol{w}$ using the mapping network and feed $\boldsymbol{w}$ into the synthesis network to produce an image. StyleGAN2 uses the truncation trick (constraining $\boldsymbol{w}$ via $\boldsymbol{w'} = \bar{\boldsymbol{w}} + \psi(\boldsymbol{w} - \bar{\boldsymbol{w}})$) to improve image quality [5]. We set the truncation parameter to $\psi = 0.5$ as it provides high quality images without loosing significant variation.

3.3.2 PCA Transformation. In addition to $\mathcal{W}$-space, we created an auxiliary latent space based on PCA (from this point referred to as PCA-space), following the method introduced by Härkönen et al. [15]. In brief, we sampled 750,000 random latent vectors in $\mathcal{Z}$-space, used StyleGAN2's mapping network to generate $\mathcal{W}$-space vectors and trained a PCA model with 512 dimensions. This approach produces unsupervised semantic features that, being based on PCA, are guaranteed to be orthogonal and, therefore, independent of one another. These features allow us to use linear models to implement search facets and ranking, which are computationally cheap and allow us to maintain interactivity.

3.3.3 Sampling. In the first iteration of a search session, we randomly sample 20 vectors from $\mathcal{Z}$-space and generate images using StyleGAN2. In all subsequent iterations, we randomly sample 20,000 vectors from $\mathcal{Z}$-space, transform them into $\mathcal{W}$-space and then transform the $\mathcal{W}$-space vectors into PCA-space. The PCA vectors are passed on to the faceted search module.

3.4 Nudging with Faceted Search

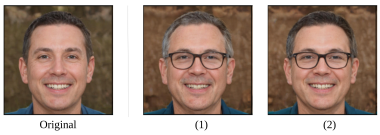
Despite randomly sampling 20,000 vectors at each search iteration, the particular combination of user-selected facets may only be present in a small proportion of generated images. We, therefore, used supervised GAN controls [53] to implement search facets that we use to "nudge" vectors towards regions of the latent space more likely to contain the selected facial features. Each facet was implemented using a linear classifier in PCA-space. We use a similar multi-label classifier to filter out the resulting nudged images that do not contain those attributes.

3.4.1 Facet Classifier Training. We created a classifier for each search facet using the following procedure:

- We fine-tuned a ResNet50 classifier [17] using the CelebA data set [37]. CelebA contains 202,599 face images associated with 40 binary attributes, e.g. female, glasses, etc.[3]
- We randomly generated 270,000 images using StyleGAN2 and labelled them with facial attributes using the ResNet50 classifier.
- We created linear classifiers for a subset of attributes from the CelebA data set. We used logistic regression with L1 regularization to predict whether a given binary attribute

[3] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

(a) Nudging examples



(b) Decoupling example

Figure 3: (a) Nudging a randomly sampled image to "male", "black hair", "brown hair" and "blonde hair", respectively. (b) Nudging a randomly sampled image to "glasses": (1) without decoupling, "glasses" is entangled with "old"; (2) eliminating the effect of "old" with decoupling.

is present in an image based on its associated PCA-space vector.

We perform as much processing in PCA-space as possible to avoid the computational cost of generating images and predicting attributes for them. From the 40 attributes in the CelebA data set, we selected 9 that could be reliably used as facets. We created an additional attribute, "Children", as we found it could be reliably classified with only a small amount of manually annotated data. Each classifier was trained with a balanced data set and achieved over 90% accuracy on their test sets with the exception of "wavy hair" and "straight hair", which obtained accuracies of 87% and 80%, respectively. The unused CelebA facial attributes were either less reliable in terms of accuracy or did not make good GAN controls as they failed to isolate a single facial attribute (determined by manual inspection). We speculate that these failures were due to biases in the CelebA data set (resulting in entangled attributes) or in StyleGAN2 (either due to its own training data or because of the truncation trick).

3.4.2 Nudging. Given a binary facial attribute, e.g. glasses vs. no glasses, there exists a separation hyperplane with a normal vector $\boldsymbol{h}$ in PCA-space. We transform or "nudge" a PCA vector, $\boldsymbol{v}$, towards images with this attribute via: $\boldsymbol{v'} = \boldsymbol{v} + \alpha\boldsymbol{h}$, where $\boldsymbol{h}$ is the vector of coefficients from a facet's logistic regression model and α is a control parameter from 1-5 that was set via manual inspection to ensure it worked on a variety of generated faces [53]. Figure 3a shows a female face that was nudged towards the attributes "male", "black hair", "brown hair" and "blonde hair", using this procedure.

Unfortunately, some facial attributes are entangled, e.g. adding glasses to a face may inadvertently age it as well (because older people are more likely to wear glasses than younger people). We decouple the normal vectors for the two attributes, $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$, by forcing them to be orthogonal: $\boldsymbol{h'_1} = \boldsymbol{h_1} - \cos(\boldsymbol{h_1}, \boldsymbol{h_2}) \cdot \boldsymbol{h_2}$ [53]. In our system, decoupling is applied to weaken the influence of age on both gender and glasses wearing. Figure 3b shows the impact of decoupling on nudging a male face towards the "glasses" attribute.

3.4.3 Filtering. As a final check to ensure the images passed on for ranking contain all facial attributes selected by the searcher, we use a multi-label classifier for the 10 attributes used in the interface to filter PCA vectors. This step is necessary as nudging does not always succeed in adding all attributes to each sampled image. While we could have filtered images using the single-label classifiers that were used to obtain normal vectors for nudging, we used a single classifier to improve efficiency.

3.5 Ranking with Thompson Sampling

After sampling and nudging, we obtain a large set of candidate images from which we need to select the images most likely to be considered relevant by the user. Searchers provide relevance feedback at each search iteration with the goal of retrieving similar images (exploitation), but also want to explore the wider search space of images (exploration). We use Thompson sampling, a Bayesian approach to the contextual bandit problem [7], to balance exploration and exploitation.

3.5.1 Problem Setting. We have a matrix $\boldsymbol{V}$, where each row $\boldsymbol{v_i}$ is the PCA vector representation of images for which the searcher has provided relevance feedback. Let $\boldsymbol{r} = (r_1, r_2, \dots r_t)^\top$ be the column vector of relevance scores up to time t, where $r_i = 1$ for positive relevance feedback and 0 otherwise. Following [7], we assume the probability of $\boldsymbol{v_i}$ being relevant is:

$$P(r_i = 1|\boldsymbol{v_i}, \boldsymbol{\theta}) = (1 + exp(-\boldsymbol{v_i}^\top \boldsymbol{\theta}))^{-1}, \quad (1)$$

where $\boldsymbol{\theta}$ is a weight vector to be learnt.

3.5.2 Thompson Sampling. While $\boldsymbol{\theta}$ in Equation 1 could be estimated with, for example, logistic regression, the resulting weight vector can be problematic as users have only explored a small proportion of the image space. Instead, we assume $\boldsymbol{\theta}$ follows a probability distribution and select the top-k images that maximize:

$$\int \mathbb{I}(E[r|\boldsymbol{v}, \boldsymbol{\theta}] = \max_{\boldsymbol{v'}} E[r|\boldsymbol{v'}, \boldsymbol{\theta}]) P(\boldsymbol{\theta}|V) d\boldsymbol{\theta}, \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function and $P(\boldsymbol{\theta}|V)$ is the posterior distribution of θ, where:

$$P(\boldsymbol{\theta}|V) \propto P(\boldsymbol{\theta}) \prod_{\boldsymbol{v} \in V_P} P(r = 1|\boldsymbol{v}, \boldsymbol{\theta}) \prod_{\boldsymbol{v'} \in V_N} [1 - P(r' = 1|\boldsymbol{v'}, \boldsymbol{\theta})], \quad (3)$$

where $P(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, and V_P and V_N are sets of PCA vectors that received positive and negative relevance feedback, respectively.

In Thompson sampling, maximizing Equation 2 is achieved by sampling from the posterior distribution, i.e. we sample the parameters $\boldsymbol{\theta}^*$ from $P(\boldsymbol{\theta}|V)$, and then calculate the probability, $P(r = 1|\boldsymbol{v}, \boldsymbol{\theta^*})$, for each image. This weight sampling procedure addresses the exploration-exploitation trade-off as $\boldsymbol{\theta}$ is drawn according to its probability of being optimal.

Sampling directly from the posterior distribution is challenging, however, so we approximate the posterior using a multivariate normal distribution with a diagonal covariance matrix, $P(\boldsymbol{\theta}|V) \sim \mathcal{N}(\bar{\boldsymbol{\theta}}, \Sigma)$, where $\Sigma = \alpha\boldsymbol{\sigma}^{-1} \odot I$ (I is the identity matrix, α is the exploration parameter set to 1.0 and $\boldsymbol{\sigma}$ is the inverse diagonal vector of Σ). In the prior distribution, $\bar{\boldsymbol{\theta}} = \mathbf{0}$ and $\Sigma = \lambda I$, where λ is a regularization parameter set to 0.5. At each iteration, we estimate the mean $\bar{\boldsymbol{\theta}}$ by the posterior mode ($\bar{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|V)$) and the

Figure 4: Randomly generated images from StyleGAN2 with the first 30 principal components held fixed produces almost identical faces. The main differences between images appear to be the background and minor differences in the orientation of the subject's head and hair.

covariance Σ with the inverse of the negative Hessian of the logarithm posterior at the mode ($\Sigma^{-1} = -\nabla^2 \log[P(\tilde{\theta}|V)]$), according to Laplace approximation [50].

3.5.3 Ranking. To limit the amount of computation needed to be performed interactively, we truncate each PCA vector, $\boldsymbol{v_i}$, performing Thompson sampling on only the first 30 principal components. This reduces the explained variance to 0.718 , but we found that randomly generated faces with these components kept fixed were almost indistinguishable from one another, with only minor differences in the background and orientation of the subject's head and hair (see Figure 4).

After ranking all images by probability, the top-20 PCA vectors are converted back to $\mathcal{W}$-space by inverse PCA, transformed into images using StyleGAN2's synthesis network and returned to the interface to be shown to the user for the next search iteration.

3.6 Evaluation Methodology

Evaluating an exploratory search system is challenging due to the interactive nature of search systems and the subjectivity of task success. Evaluation, therefore, should involve users performing an appropriate simulated work task [4, 63] to investigate the system's impact on search behavior, task performance and perceived usability [27]. However, running user experiments to understand the usefulness of each component of a given system would be impractical as it would require a prohibitively large number of study participants. Hence, we present both a simulation study and a user study to evaluate different aspects of our approach. More specifically, we performed a simulation study to verify whether the core system components, i.e. nudging and ranking, achieved the expected functionality and to understand the broad characteristics of the system. Subsequently, we conducted a user study that situates participants in an open-ended search task designed to elicit exploratory search behavior to assess system performance with real users. In both simulation and in the user study, our system was compared with a baseline method for interactive GAN search.

3.6.1 Baseline. We adopted the Rocchio algorithm-based approach proposed by Ukkonen et al. [58] as the baseline. Unlike our approach that operates in PCA-space, the baseline uses latent vectors in $\mathcal{Z}$-space [58]. In general, the Rocchio algorithm models user preferences with a centroid latent vector $\boldsymbol{c}$ and at each iteration samples m vectors close to it (from a multivariate normal distribution with $\boldsymbol{c}$ as the mean and a covariance matrix parameterized by σ). We use $m = 20$ to match our approach setting. The centroid is iteratively updated with $\boldsymbol{c_i} = (1-\alpha)\boldsymbol{c_{i-1}} + \alpha \boldsymbol{v_{pos}}$, where $\boldsymbol{v_{pos}}$ is the mean latent vector of images that received positive relevance feedback. Following [58], σ is set to 0.2 and α is set to 0.7. The initial centroid is obtained by averaging the latent vectors of a set of seed images selected by the user.

4 SIMULATION STUDY

We conducted a series of simulations to examine the performance of the main system components: nudging and ranking. Our goal is to answer the following three questions:

(1) Does nudging ensure the presence of selected facial features?
(2) Does ranking negatively impact the effectiveness of nudging?
(3) Does Thompson sampling converge rapidly to accurately reflect users' preferences?

4.1 Nudging Evaluation

We investigated the effectiveness of nudging, in isolation and in combination with ranking.

4.1.1 Nudging Effectiveness. We assessed the effectiveness of nudging by examining whether randomly sampled images gained the selected facial attributes associated with each search facet. We randomly generated 100,000 images using StyleGAN2. As the generated images are not labeled with attributes, we employed the fine-tuned ResNet50 classifier (see Section 3.4.1) to assign attributes to the images. We analyzed all attributes used to create search facets, with the exception of "Children" as the ResNet50 classifier is unable to predict this label (see Section 3.4.1). We calculated the proportion of images that featured each facial attribute, considering the case of nudging with and without filtering (see details in Section 3.4.3).

The results depicted in Figure 5 indicate that randomly sampled images (i.e. without nudging) are biased against certain attributes, such as "blonde hair" (~2%), "bangs" (~8%) and "glasses" (~10%). After nudging, however, the percentage of images containing each attribute is significantly improved, with improvements ranging from 55% (for "female") to 3750% (for "blonde hair"). The incorporation of filtering yields further improvements, with four out of nine attributes ("black hair", "male", "female", "glasses") reaching at least 97%. The lowest percentage obtained is approximately 62%, which is higher than the highest percentage achieved without nudging (~60%).

4.1.2 Nudging Effectiveness after Ranking. Although we can manipulate the attributes that a generated image features to a large extent, it is unclear how well nudging performs when integrated into the system. One concern is that iterative re-ranking could degrade the performance of nudging. For example, a user who has selected the search facet "blonde hair" may reasonably give positive relevance feedback to images of people with brown hair if they feature other attributes that the user considers relevant to their search goals. As nudging fails for 18% of images in the case of blonde hair (see Figure 5), in the worst-case scenario this inconsistent combination

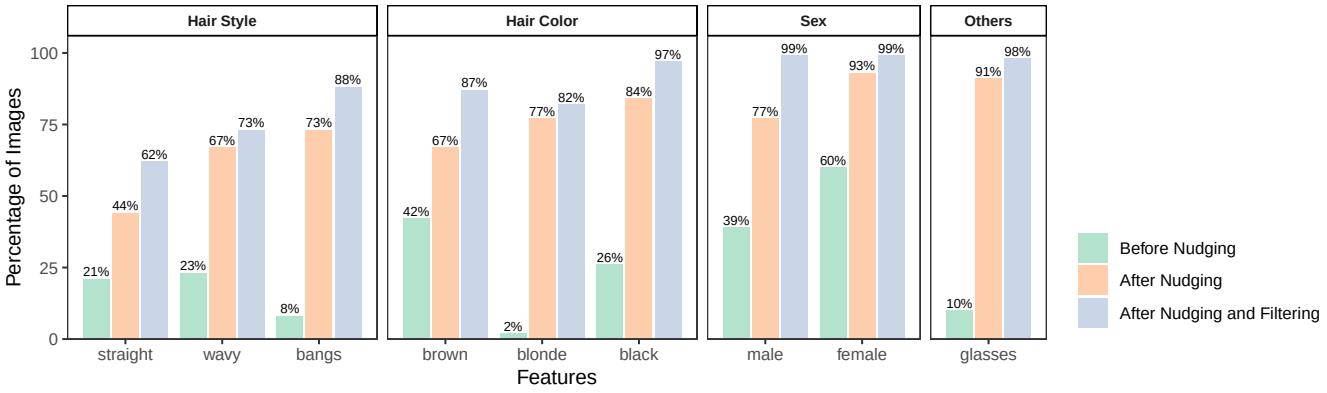


Figure 5: Percentage of images featuring each facial attribute over 100,000 generated images. Both nudging and filtering increase the percentage of images with each attribute.

of search facets and relevance feedback could result in the system not displaying any images to the user with blonde hair.

We conducted a simulation where we generated random relevance feedback to examine the effectiveness of nudging during iterative search. As in the previous experiment, we examined the same 9 out of 10 facial attributes. For each attribute, we simulated 100 search sessions with the target attribute's search facet selected. Each search session spanned 30 iterations. At each iteration, we randomly selected one positive and one negative example to update the ranking function (for real users, we assume that relevance feedback is more likely to reinforce the target attribute if that attribute is important to their search goals). We compared the number of images featuring the target attribute in the first and last iterations, i.e. before and after 30 iterations of re-ranking.

Figure 6 illustrates no significant difference in the distribution of the number of images featuring each attribute in the first and the last iterations ($P > 0.05$ for all attributes, Mann-Whitney U test). This suggests that ranking does not affect the performance of nudging despite receiving potentially inconsistent feedback and irrespective of the attributes' baseline nudging effectiveness.

4.2 Ranking Evaluation

We validated the use of Thompson sampling in our system following a similar simulation approach to Chapelle and Li [7]. For each simulation run, we simulated user feedback using an arbitrary weight vector $\boldsymbol{w}^*$. Hence, the probability of an image being considered relevant can be determined by $p = (1 + exp(-\boldsymbol{v_i}^T \boldsymbol{w}^*))^{-1}$, where $\boldsymbol{v_i}$ is the PCA vector representation of the image. We then generated relevance feedback for the image by sampling from a Bernoulli distribution with the probability p. At each iteration, we synthesized relevance feedback for all 20 images. In order to make the simulation more realistic, we set a parameter n, which limits the number of positive and negative examples provided per iteration. More precisely, we randomly sample n positive and n negative examples without replacement, which is then used to update the ranking model as described in Section 3.5.2. We tested $n \in \{3, 5, 10, 20\}$, where $n = 20$ represents giving relevance feedback every image. The simulation results indicate a consistent pattern across all values of n, so we have chosen to present the results for $n = 3$, because it most closely aligns with a real-world scenario, as demonstrated in our user study (see Section 6.5). We compared our system with the baseline approach described in Section 3.6.1. We simulated 100 search sessions, where each session spanned 1000 iterations using both approaches. We examine the simulation results from three perspectives: convergence, effectiveness, and diversity, as shown in Figure 7 (only the results for the first 500 iterations are presented for clarity).

We note that these simulations are not intended to represent actual search conditions or user search behaviour: the use of search facets dramatically reduces the search space and users' subjective assessment of task outcomes will likely end the search session much sooner.

4.2.1 Convergence. We analyzed the convergence of Thompson sampling by observing the trend of regret across iterations. Similar to [7], the regret at iteration t is defined as the difference between the highest attainable reward and the actual reward obtained. More precisely, the regret is $R_t(a) = \max_a X_t(a) - X_t(a)$, where $X_t(a)$ is the reward of selecting a list of images a to display and is defined as the proportion of images considered relevant by the simulated user model. Following Chapelle and Li [7], however, we instead used the expectation of the proportion of images to avoid unnecessary variance. The Rocchio algorithm was not included in this assessment because it is not a learning algorithm.

The regret as a function of iteration (t) for Thompson sampling is plotted in Figure 7a. As shown, the regret declines dramatically within the first 10 iterations, and eventually converges to ~0.2 after 50 iterations. This implies that Thompson sampling can quickly adapt to users' preferences, and, therefore, is appropriate for an interactive search system.

4.2.2 Effectiveness. To make a comparison between our approach and the baseline in terms of effectiveness, we show the average number of images that would obtain positive relevance feedback at each iteration in Figure 7b. Both approaches converge rapidly after ~10 iterations. However, the baseline achieves near-optimal

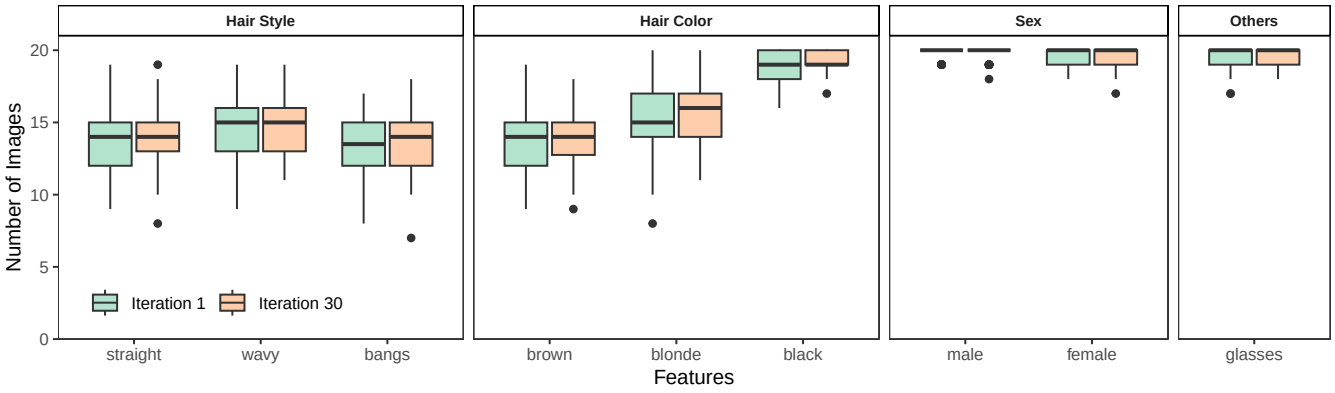


Figure 6: Distribution of the number of images featuring each attribute at the 1st iteration (green) and the 30th iteration (orange) of 100 simulated search sessions.

performance (i.e. almost all 20 images displayed are relevant images), whereas Thompson sampling converges to the slightly lower number of ∼16.5 relevant images.

4.2.3 Diversity. As our goal is to facilitate exploratory search, an essential aspect to consider is the diversity of images displayed across iterations. We define the diversity at each iteration as the average pairwise face distance. The face distance is calculated using the Python Face Recognition library [9] that uses an approach similar to FaceNet [52]. To put the face distance into context, we randomly sampled 10,000 $\mathcal{Z}$-space vectors and used StyleGAN2 to generate the corresponding images twice. As StyleGAN2 injects some noise with each image generation, the images are not identical, but highly similar (data not shown). The face distance between these pairs of almost identical images ranged from 0.043 to 0.279.

As shown in Figure 7c, Thompson sampling maintains a consistently high level of image diversity throughout, whereas diversity drops drastically with the Rocchio algorithm and converges to a value lower than 0.25. Moreover, the average pairwise face distance drops below the threshold after ∼225 iterations for the baseline (see the dashed line in Figure 7c). This reveals how the near-optimal performance of the Rocchio algorithm demonstrated above (Section 4.2.2) is actually achieved by sacrificing the diversity of displayed images. More precisely, despite all 20 images displayed being considered relevant after reaching the convergence point (see Figure 7b), most of them are highly similar or even almost identical. This observation is in line with the feedback from real users in our user study (see details in Section 6.6).

5 USER STUDY METHODOLOGY

We conducted a user study to evaluate the proposed approach. The aim of the study was threefold. First, we wanted to understand how well the system supported users performing exploratory search tasks. Second, we wanted to know how our system compared to an existing system in terms of usability and user satisfaction. Finally, we were interested in understanding how users perceived the system and what could be changed to improve user experience.

We used a within-subject study design, where each participant used both our system and a baseline system to complete two different exploratory search tasks. Each experiment was conducted through Zoom: the system was run on the experimenter's local computer and we gave participants the ability to interact with each system remotely. Each participant was compensated with a book voucher after completing the study.

5.1 Baseline System

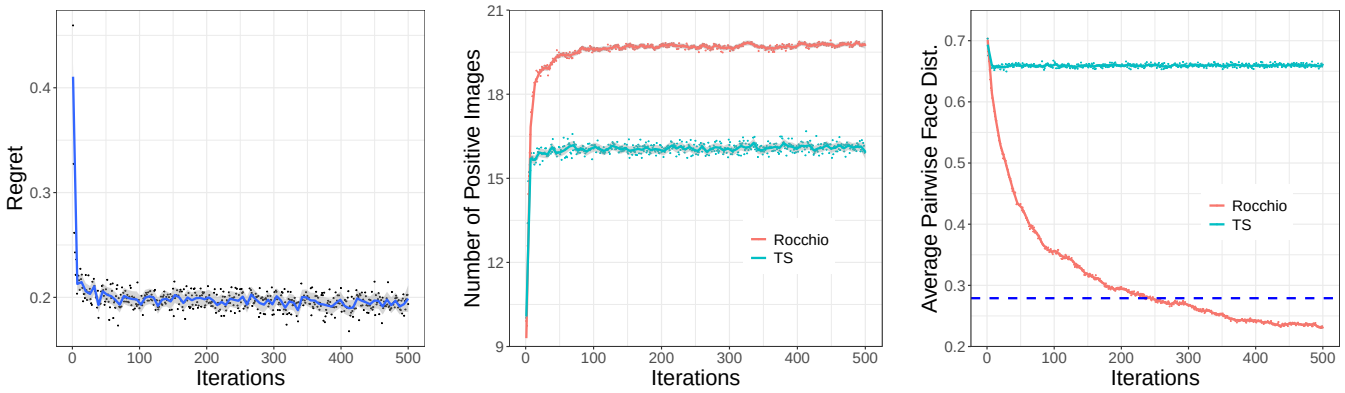
As in the simulation study, we compared our system with the Rocchio-based approach proposed by Ukkonen et al. [58] (see Section 3.6.1 for details). We made the interface for the baseline as similar to our system as possible, but without the search facets in the left margin nor the ability to switch between positive and negative relevance feedback (see Figure 8). We perform warm start in the baseline system by asking users to select at least one seed image from 100 randomly generated images before the search process starts (as in the "near task" in the original study [58]).

5.2 Participants

We recruited 30 study participants (16 female, 13 male, 1 rather not say) aged 23-53 (mean age 28.6). The participants ranged from Master's students to postdoctoral researchers and had backgrounds in various disciplines, including game design, forestry, cell biology, computer science and business analytics. According to the background questionnaire, almost all participants stated that they had used image search engines at least once (frequently (7), sometimes (8), rarely (11), never (4)), while about 70% of participants had experience of using image search engines without textual search queries (sometimes (8), rarely (14), never (8)).

5.3 Task

In the search tasks, participants were situated in the role of a casting director and asked to identify the faces of actors who they felt looked suitable to play the roles of Harry Potter and Hermione Granger in an upcoming Harry Potter movie. They were given a brief summary of the movie's plot:



(a) Convergence (b) Effectiveness (c) Diversity

Figure 7: Simulation results of Thompson sampling and the Rocchio Algorithm based on three performance measures: *convergence*, *effectiveness*, and *diversity*. Convergence only shows Thompson sampling. Diversity additionally shows the maximum face distance obtained from 10,000 images pairs derived from the same $\mathcal{Z}$-space vector.

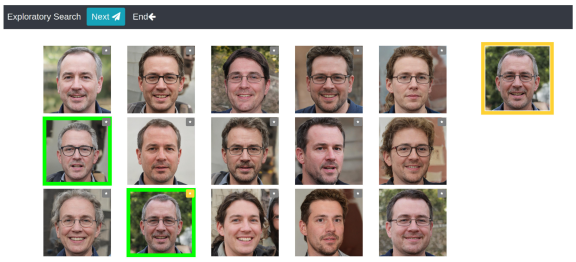


Figure 8: Search engine interface of the baseline system. Images with green boxes have received positive relevance feedback. The bookmarked image is shown in the right margin.

> *"The movie takes place when Harry Potter and Hermione Granger are around 30 years old. Harry was framed for a crime he did not commit and was imprisoned in Azkaban (a prison for wizards). At the start of the movie, Harry escapes from Azkaban. His time in prison has been tough. Harry is angry and wants revenge. Hermione is now a teacher of the dark arts at Hogwarts, but is unhappy and disillusioned with the world of magic."*

The tasks were designed to elicit exploratory search behavior: they are open-ended, and do not include any visual criteria describing success outcomes. Instead, searchers needed to combine their understanding of the characters with the circumstances of the movie [63]. We ensured that all participants had watched at least one Harry Potter movie in order to perform the search tasks.

5.4 Procedure

Prior to the study, we provided participants with a 5-minute instructional video to teach them how to use both systems. During the experiment, we asked each participant to perform two image retrieval tasks - one with each system. The tasks were the same for all participants, i.e. to find the face of a person who they feel could play the role of Harry Potter or Hermione Granger, given the description of the movie provided. We counter-balanced the sequence of systems and tasks to avoid order effects. We limited each task to 10 minutes to avoid possible biases caused by excessively long search sessions. After each task, we asked participants to (i) rate how satisfied they are with their bookmarked image on a 5 point Likert scale (from "very dissatisfied" to "very satisfied"), (ii) complete the standard System Usability Scale (SUS) questionnaire with 10 questions [22], and (iii) complete a modified ResQue questionnaire with 11 questions [46]. While the SUS questionnaire can help us evaluate system usability, the ResQue questionnaire was chosen to measure users' satisfaction with the system.

While participants performed search tasks, we logged all their interactions with each system including images displayed, facets selected, relevance feedback, and which images were bookmarked. After participants completed both tasks, we asked them to answer a post-experiment questionnaire and conducted a semi-structured interview to better understand their perceptions of system functionality and the differences between systems. This procedure lasted approximately 30-40 minutes.

6 RESULTS

We present a quantitative analysis of task performance, system usability, and user satisfaction, preferences and search behavior, and summarize the outcomes of the semi-structured interviews with study participants.

6.1 Task Performance

We assessed task performance using the ratings study participants gave to their final bookmarked images. There was no significant difference in ratings for images found using our system (mean rating = 4.33) versus the baseline (4.2) ($P = 0.317$, Wilcoxon signed-rank test). Furthermore, in both systems, 27/30 participants stated they were either satisfied or very satisfied with their bookmarked images. Anecdotally, 3 more study participants were "very satisfied"

(a) Task I: Harry Potter

(b) Task II: Hermione Granger

Figure 9: Final bookmarked images found by 30 study participants for two exploratory search tasks. For each task, *top row*: search results obtained with the Rocchio system; *Bottom row*: search results obtained with our system.

with the final image obtained from our system than the baseline (13 vs 10), while the only instance (out of 60 search tasks) of a study participant being "dissatisfied" was from the baseline system. The final bookmarked images for two search tasks are shown in Figure 9, which shows a high diversity of user preferences.

6.2 System Usability

Our system obtained a SUS score of 68.4, indicating an above-average level of usability [33], whereas the baseline obtained a score of 73.9. While there was no statistically significant difference between SUS scores ($P = 0.056$, Wilcoxon signed-rank test), it seems likely that the difference would have been significant given a slightly larger sample size.

Table 1 breaks down the results of each question in the SUS questionnaire, highlighting statistically significant differences in the responses to questions 2, 4, and 10. These questions were all related to system complexity, which is understandable as our system contained more user interface elements (i.e. search facets and negative relevance feedback) and aimed to show more diverse images to users to encourage exploration. We note, however, that despite these differences being significant, the average responses for our system were between 2.2 and 2.4, which is between the "disagree" (2) and "neutral" (3) responses. These scores suggest that our system was not considered complex in an absolute sense, just slightly more complex than the baseline.

6.3 User Satisfaction

In our modified ResQue questionnaire, we found no significant differences in responses between our system and the baseline (see Table 2). For both systems, users stated that they were recommended good images (Q1, Q4, and Q10), that the system was easy to start using and they felt confident providing relevance feedback (Q5 and Q7), and they were generally satisfied (Q11).

6.4 User Preferences

While ResQue was administered after each search task, our post-experiment questionnaire contained similar questions, but asked users to explicitly compare the two systems (see Table 3). Our system showed users more diverse images than the baseline (Q2, $P = 0.016$, binomial test), which, as a result, included more images that were irrelevant to the search task (Q4, $P = 0.043$, binomial test). Furthermore, despite the potential complexity issues highlighted by the SUS results, our interface was preferred to the baseline by 22/30 study participants (Q5, $P = 0.016$, binomial test). Finally, for performing the exploratory search task of casting roles in a new Harry Potter movie, 23/30 study participants preferred our system over the baseline (Q1, $P = 0.005$, binomial test).

The post-experiment questionnaire also contained questions related to our interface's system components (Table 3, Q6-9). Study participants were divided between whether it was easier to stick with only giving positive relevance feedback (Q6) and whether it was easy to give negative feedback (Q7). However, 90% of participants agreed that it was easy to use search facets (Q8, $P = 8.4\times10-6$, binomial test) and that the meanings of the facet labels were clear (Q9, $P = 8.4 \times 10^{-6}$, binomial test).

6.5 User Behavior

On average, users spent 62.9 seconds longer completing the search task with our system compared to the baseline (407.5 vs 344.6 seconds, $P = 0.046$, Wilcoxon signed-rank test). There was no significant difference in the number of search iterations before finding the most satisfying search result, i.e. the final bookmarked image (3.6 iterations for our system compared to 4.3 with the baseline, $P = 0.115$, Wilcoxon signed-rank test). However, due to the baseline system's dependence on warm-start, users of our system examined significantly fewer images in total than with the baseline system (106.7 vs 188.7, $P = 2.50 \times 10^{-5}$, Wilcoxon signed-rank test).

When using our system, users provided more relevance feedback compared to the baseline (28.3 vs 15.7, $P = 0.014$, Wilcoxon signed-rank test). Furthermore, we found that users tended to give more

Table 1: Average SUS responses for our system (Ours) and the baseline (Base.), P-values from Wilcoxon signed-rank test. Better scores are bolded. Statistically significant differences ($P < 0.05$) are highlighted with *.

Base.	Ours	P-value	Question
3.7	3.6	0.518	1. I think that I would like to use this system frequently.
1.9	2.3	0.041*	2. I found the system unnecessarily complex.
4.0	3.8	0.261	3. I thought the system was easy to use.
2.0	2.4	0.018*	4. I think that I would need the support of a technical person to be able to use this system.
3.7	3.7	0.953	5. I found the various functions in this system were well integrated.
1.9	2.1	0.153	6. I thought there was too much inconsistency in this system.
4.1	4.0	0.592	7. I would imagine that most people would learn to use this system very quickly.
2.0	2.2	0.216	8. I found the system very cumbersome to use.
3.7	3.5	0.152	9. I felt very confident using the system.
1.8	2.2	0.032*	10. I needed to learn a lot of things before I could get going with this system.

Table 2: Average ResQue scores obtained by our system (Ours) and the baseline (Base.), P-value from Wilcoxon signed-rank test. Better scores are bolded. No differences were statistically significant.

Base.	Ours	P-value	Question
3.8	3.8	0.930	1. The images recommended to me matched what I was searching for.
3.7	3.5	0.182	2. The system helped me discover new images.
3.1	**3.2**	0.639	3. The images recommended to me are diverse.
3.8	**3.9**	0.439	4. The system helped me find the ideal images.
4.3	4.1	0.251	5. I became familiar with the system very quickly.
3.6	3.3	0.084	6. I found it easy to notice if the search results were not correct anymore.
3.9	3.8	0.593	7. I felt confident to give feedback.
3.7	3.6	0.429	8. Using the system to find what I like is easy.
3.1	3.0	0.565	9. I found it easy to re-find images I had been recommended before.
4.0	4.0	1.000	10. The system gave me good suggestions.
3.9	**4.0**	0.480	11. Overall, I am satisfied with the system.

negative than positive relevance feedback in our system (3.79 negative vs 2.56 positive per iteration, $P = 0.049$, Wilcoxon signed-rank test), whereas there was no significant difference in the amount of positive relevance feedback given between our system and the baseline (2.56 vs 3.24, $P = 0.116$, Wilcoxon signed-rank test).

Finally, we investigated how participants used search facets in our system. A majority of users (24/30, $P = 0.0014$, binomial test) only selected facets during the first iteration (i.e. before providing relevance feedback) and did not adjust them in subsequent iterations. We observed that these users took significantly fewer iterations (5.2 vs 8.2, $P = 0.038$, Wilcoxon rank-sum test) compared to users who did not use facets (4/30) or who adjusted facets after the first search iteration (2/30).

6.6 Qualitative Feedback

The comments made by study participants during semi-structured interviews helped us further understand why a majority of users preferred our system over the baseline for exploratory search tasks. The most often mentioned reasons included (i) facets helped to narrow down the search space and made image selection easier (12 users), (ii) our system showed more diverse images (9), (iii) negative relevance feedback was perceived to increase the efficiency of searching (8), and (iv) our system provided better recommendations and/or helped to find more satisfying images (8). While on the other hand, several users complained that (i) not enough search facets were provided (3), (ii) negative relevance feedback was unnecessary (4), and (iii) our system returned many irrelevant images (4).

For the baseline, although some users appreciated its simplicity (8) and better convergency (8), many users complained that the system returned very similar faces that were difficult to distinguish (14), e.g.: *"I kind of felt to have face blindness"* [Participant 26], *"the [baseline] converges too fast and faces look very similar to each other"* [P29], *"at the last several iterations, all images looked like images of the same person"* [P9].

While there was no significant difference in task performance according to users' ratings (see Section 6.1), five users stated that our system helped them find more satisfying results, though only three of these participants gave different ratings. Furthermore, five different users stated that the baseline may not have helped them find what they were looking for: *"the [baseline] lost the images I was looking for. The more I use it, the more I feel that the results deviate from my goal"* [P2], *"the [baseline] only captured features of the image I selected during initialization, but that was a bad one"* [P23].

Table 3: Post-experiment questions with proportions (Prop.) and corresponding P-values from a binomial test (the null hypothesis being a proportion of 0.5 for random responses). Questions 1-5 were binary responses and questions 6-9 were rated on a 5-point Likert scale. Prop. was the proportion of users that selected our system over the baseline in question 1-5 and agreeing or strongly agreeing with the statement in questions 6-9. Statistically significant results ($P < 0.05$) are marked with *.

Prop.	P-value	Question
0.767*	0.005	1. Which system did you prefer to use for finding an actor if you are a casting director?
0.733*	0.016	2. Which system in your opinion provided more diverse images?
0.500	1.000	3. Which system in your opinion provided more ideal images?.
0.700*	0.043	4. Which system in your opinion provided more irrelevant images?
0.733*	0.016	5. Which interface did you prefer?
0.500	1.000	6. I found it easier to give only positive feedback.
0.667	0.099	7. I found it easy to give negative feedback.
0.900*	8.4e-6	8. I found it easy to perform the search with facets.
0.900*	8.4e-6	9. The facet labels are clear.

7 DISCUSSION AND CONCLUSIONS

In this paper, we presented a novel approach for exploratory search using interpretable GAN controls. Unlike traditional interactive image retrieval systems that operate over a collection of discrete images, our approach seeks to capture users' information needs using the GAN's latent space to generate images, which can be summarized as *sample*, *nudge*, and *rank*. We used interpretable GAN controls to implement faceted search, to help users quickly narrow down the search space, and relevance feedback, that used Thompson sampling to help users explore the search space by balancing exploration and exploitation.

7.1 Summary of Results

We conducted a series of simulation studies to validate the efficacy of nudging and ranking. Our results showed that nudging significantly increased the occurrence of selected attributes, with improvements from 55% to 3750% (Figure 5), highlighting the effectiveness of nudging in limiting the search space. Additionally, the number of images featuring selected attributes was relatively consistent throughout a search session and remained unaffected by the iterative re-ranking process, even with inconsistent relevance feedback (Figure 6). In simulated, Thompson sampling-based ranking accommodated user preferences quickly without sacrificing image diversity, whereas the baseline converged rapidly, but produced results with very limited diversity (Figure 7).

Our user study showed that a majority of participants (23/30) preferred our system to a state-of-the-art baseline [58] for performing exploratory search tasks using GANs (Table 3, Q1). While our analysis of SUS responses highlighted our system's higher perceived complexity than the baseline (Table 1, Q2, Q4 and Q10), the additional interface components (search facets, negative relevance feedback) were praised in participant interviews for improving the perceived efficiency of searching (Section 6.6 and Table 3, Q8). Indeed, participants that used facets to initialize their search took significantly fewer iterations to complete the tasks (Section 6.5). Our results are also in line with previous studies claiming that negative relevance feedback improves search performance and user satisfaction [43, 45] (Section 6.6 and Table 3, Q6). Despite several users preferring the convergence properties of the baseline (i.e. lower image diversity), it was highlighted as a pain point for many users (Section 6.6).

7.2 Implications

This article focused on exploring a new category of search task: exploratory search in a continuous latent image space. In image search, the explosive growth of visual data on the internet has raised concerns within the community due to the amount of effort required to annotate, organize, and represent such tremendous numbers of images [10]. However, such challenges can be ameliorated by instead training a large generative model and searching with interpretable controls as in our approach. Furthermore, despite our work focusing on GANs, we argue that the essence of our approach is helping searchers' to navigate a continuous latent space, which is not limited to GANs. Indeed, recent studies have demonstrated that the latent space of diffusion models (a type of image generative model) also exhibit continuity similar to GANs and can be manipulated via text prompts [28, 49, 51]. This seems to suggest that our approach can also be applied to diffusion models. Moreover, the high diversity image generation capabilities of diffusion models have the potential to extend our approach beyond searching within a single domain.

Our work also sheds light on the role of user modeling in navigating complex latent spaces. Many recent studies have been devoted to building applications that can generate images based on textual descriptions given by the user (similar to the case of searching for images given a search query in traditional image retrieval). More recent studies have suggested employing prompt engineering as an interactive text-based retrieval mechanism, facilitating users in identifying desirable prompts (i.e. textual descriptions) to better describe their information needs [8, 44]. However, these systems can fail in situations where users have hard-to-describe or subjective information needs [35]. In contrast, our approach can overcome this limitation by modeling user preferences with faceted search and relevance judgments.

There are several narrower implications related to how we implemented faceted search and relevance feedback that could impact our ability to expand the capabilities of our current system. While our search facets performed well (see Figure 3) and were useful to

searchers, expanding the range of facets to cover even just the facial attributes in the CelebA data set would be exceptionally challenging as the majority of features in real images are highly entangled, making it difficult to produce supervised controls that change only a single isolated attribute. One possible solution would be to use StyleGAN itself to create synthetic training data, but this would require considerable manual annotation. For GANs other than faces, it seems likely that manual (and potentially expert) annotation would be the only way to develop faceted search. Furthermore, it is unclear how entangled features impact the efficacy of relevance feedback. Our implementation of relevance feedback uses unsupervised GAN controls to generate orthogonal features, but many of these features are composed of complex combinations of attributes and it is unclear how this affects search effectiveness. Finally, while we used interpretable controls to hide the underlying GAN from the user, this could be supplemented by interaction mechanisms that exploit the properties of the GAN more effectively. For example, the ability to interpolate between images could be used to foreshadow the impact of feedback on future search results. Similar ideas were investigated in the FutureView system [18], but in the context of traditional interactive image retrieval.

7.3 Limitations

The study we conducted has several limitations. First, we only presented results from a single domain: human faces, that could have biased our results because humans are highly adept at facial recognition (to the extent that we see faces in everyday objects, a phenomena known as *pareidolia*), but are less capable in other domains. This implies that participants searching a GAN of, for example, cat or car images may be less able to discern the differences between similar images and, therefore, need a higher level of exploration to search effectively. Domain could also introduce experimental confounders, such as domain knowledge, as highly unlikely images could impact user experience, e.g. images of cats with a male facial structure and a tri-color coat.

Second, our study was limited to a single GAN architecture, StyleGAN2, which could have biased our system towards particular design decisions or parameter settings. As StyleGAN2 is designed to disentangle features, some aspects of our system, such as search facet decoupling (see Figure 3b) and ranking, may not work as well with other models. However, as both supervised and unsupervised interpretable GAN controls have been developed for other GANs (e.g. [15]), we believe that our broad approach is generally applicable.

7.4 Practical and Social Impact

The approach proposed in this article is intended to be deployed in any scenario where there is a need for images of faces that fit highly subjective criteria. These scenarios are generally in the creative industries, such as casting actors for a movie role or as an ideation tool for characters when writing a novel. For example, our system could be used to perform exploratory search in co-creation activities to identify images with the right aesthetic for a given movie role. The process of casting could then focus on finding a similar looking actor, instead of being influenced by the limited diversity of the pool of applicants. Of course, biases in the GAN training data could make some searches difficult or even impossible. We know, for example, that the CelebA data set contains relatively few images of children and is limited in terms of ethnic diversity. Furthermore, as interactive information retrieval studies generally focus on search tasks that can succeed [27], we did not investigate how performance degrades in sparser areas of the latent space. Practical applications would necessitate the creation of more diverse underlying models that have been thoroughly tested in their application domain to avoid perpetuating existing biases.

7.5 Future Work

In future work, we want to investigate more scalable approaches for implementing faceted search for GANs. Our current method for creating facets relies on supervised learning and, therefore, requires labelled data for a range of attributes. Instead, we plan to use pre-trained multi-modal models, such as CLIP [47], to associate latent vectors from a GAN with textual information directly. By doing this, we can not only expand the categories of facets in the system, but could also use this textual information to further facilitate exploratory search.

ACKNOWLEDGMENTS

This work has been supported by Helsinki Institute for Information Technology HIIT.

REFERENCES

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)* 40, 3 (2021), 1–21.

[2] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[3] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. 2016. Beyond relevance: adapting exploration/exploitation in information retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 359–369.

[4] Pia Borlund. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation* 56, 1 (2000), 71–90.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv* abs/1809.11096 (2019).

[6] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. 2010. Mindfinder: interactive sketch-based image search on millions of images. In *Proceedings of the 18th ACM international conference on Multimedia*. 1605–1608.

[7] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems* 24 (2011).

[8] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2023. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 172–186.

[9] Adam Geitgey. 2017. *Face Recognition*. Retrieved January 25, 2023 from https://github.com/ageitgey/face_recognition

[10] Dorota Głowacka and Sayantan Hore. 2014. Balancing exploration–exploitation in image retrieval. *UMAP 2014 Extended Proceedings* (2014).

[11] Dorota Glowacka and John Shawe-Taylor. 2010. Content-based image retrieval with multinomial relevance feedback. In *Proceedings of 2nd Asian Conference on Machine Learning*. JMLR Workshop and Conference Proceedings, 111–125.

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*. 5744–5753.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[14] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems* 31 (2018).

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems* 33 (2020), 9841–9850.

[16] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 53–62.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Sayantan Hore, Dorota Glowacka, Ilkka Kosunen, Kumaripaba Athukorala, and Giulio Jacucci. 2015. FutureView: Enhancing Exploratory Image Search. In *IntRS@RecSys*.

[19] Sayantan Hore, Lasse Tyrvainen, Joel Pyykko, and Dorota Glowacka. 2015. A reinforcement learning approach to query-less image retrieval. In *International Workshop on Symbiotic Interaction*. Springer, 121–126.

[20] Mark J Huiskes and Michael S Lew. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. 39–43.

[21] Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

[22] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. 1996. *Usability evaluation in industry*. CRC Press.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.

[25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[27] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2426–2435.

[29] Ksenia Konyushkova and Dorota Glowacka. 2013. Content-based image retrieval with hierarchical Gaussian Process bandits with self-organizing maps. In *21st European Symposium on Artificial Neural Networks, ESANN 2013, Bruges, Belgium, April 24-26, 2013*.

[30] Adriana Kovashka and Kristen Grauman. 2017. Attributes for image retrieval. In *Visual Attributes*. Springer, 89–117.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[32] Ivan Kropotov, Alan Medlar, and Dorota Glowacka. 2021. *Exploratory Search of GANs with Contextual Bandits*. Association for Computing Machinery, New York, NY, USA, 3157–3161. https://doi.org/10.1145/3459637.3482103

[33] James R Lewis. 2018. Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction* 34, 12 (2018), 1148–1156.

[34] Jing Li and Nigel M Allinson. 2013. Relevance feedback in content-based image retrieval: a survey. In *Handbook on neural information processing*. Springer, 433–469.

[35] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[36] Yang Liu, Alan Medlar, and Dorota Glowacka. 2022. ROGUE: A System for Exploratory Search of GANs. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3278–3282. https://doi.org/10.1145/3477495.3531675

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[38] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[39] Shiyang Lu, Tao Mei, Jingdong Wang, Jian Zhang, Zhiyong Wang, and Shipeng Li. 2014. Browse-to-Search: Interactive Exploratory Search with Visual Entities. *ACM Trans. Inf. Syst.* 32 (2014), 18:1–18:27.

[40] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[41] Alan Medlar and Dorota Glowacka. 2018. How Consistent is Relevance Feedback in Exploratory Search?. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1615–1618.

[42] Alan Medlar, Jing Li, and Dorota Głowacka. 2021. Query Suggestions as Summarization in Exploratory Search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 119–128.

[43] Henning Muller, Wolfgang Muller, Stéphane Marchand-Maillet, Thierry Pun, and David McG Squire. 2000. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Vol. 1. IEEE, 1043–1046.

[44] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. 2023. Prompting AI art: An investigation into the creative skill of prompt engineering. *arXiv preprint arXiv:2303.13534* (2023).

[45] Jaakko Peltonen, Jonathan Strahl, and Patrik Floréen. 2017. Negative Relevance Feedback for Exploratory Search with Visual Interactive Intent Modeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 149–159. https://doi.org/10.1145/3025171.3025222

[46] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[48] Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. Where Does the Performance Improvement Come From? - A Reproducibility Concern about Image-Text Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2727–2737. https://doi.org/10.1145/3477495.3531715

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[50] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2017. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038* (2017).

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=08Yk-n5l2Al

[52] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[53] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.

[54] Yujun Shen and Bolei Zhou. 2021. Closed-Form Factorization of Latent Semantics in GANs. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 1532–1540.

[55] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). http://arxiv.org/abs/1409.1556

[56] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1349–1380.

[57] Bart Thomee and Michael S Lew. 2012. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval* 1, 2 (2012), 71–86.

[58] Antti Ukkonen, Pyry Joona, and Tuukka Ruotsalo. 2020. Generating Images Instead of Retrieving Them: Relevance Feedback on Generative Adversarial Networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1329–1338.

[59] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6439–6448.

[60] Jingdong Wang and Xian-Sheng Hua. 2011. Interactive image search by color map. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 1 (2011), 1–23.

[61] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 515–524. https://doi.org/10.1145/3077136.3080786

[62] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.

[63] Barbara M Wildemuth and Luanne Freund. 2012. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the symposium on human-computer interaction and information retrieval*. 1–10.

[64] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2023. GAN Inversion: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3121–3138. https://doi.org/10.1109/TPAMI.2022.3181070

[65] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. 2010. Image search by concept map. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 275–282.

[66] Heng Xu, Jun-yi Wang, and Lei Mao. 2017. Relevance feedback for Content-based Image Retrieval using deep learning. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 629–633.

[67] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 401–408. https://doi.org/10.1145/642611.642681

[68] Yang Yu, Zhiqiang Gong, Ping Zhong, and Jiaxin Shan. 2017. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International conference on image and graphics*. Springer, 97–108.

[69] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks. In *ICML*.

[70] Weinan Zhang. 2018. Generative adversarial nets for information retrieval: Fundamentals and advances. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1375–1378.

[71] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1012–1021. https://doi.org/10.1145/3477495.3532047

[72] Wengang Zhou, Houqiang Li, and Qi Tian. 2017. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064* (2017).