2-2022

Collaborative curating for discovery and expansion of visual clusters

Duy Dung LE
Singapore Management University, ddle.2015@phdis.smu.edu.sg

Hady W. LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Collaborative Curating for Discovery and Expansion of Visual Clusters

Dung D. Le
dung.ld@vinuni.edu.vn
VinUniversity
Vietnam

Hady W. Lauw
hadywlauw@smu.edu.sg
Singapore Management University
Singapore

ABSTRACT

In many visually-oriented applications, users can select and group images that they find interesting into coherent clusters. For instance, we encounter these in the form of hashtags on Instagram, galleries on Flickr, or boards on Pinterest. The selection and coherence of such user-curated visual clusters arise from a user's preference for a certain type of content as well as her own perception of which images are similar and thus belong to a cluster. We seek to model such curation behaviors towards supporting users in their future activities such as expanding existing clusters or discovering new clusters altogether. This paper proposes a framework, namely Collaborative Curating that jointly models the interrelated modalities of preference expression and similarity perception. Extensive experiments on real-world datasets from a visual curating platform show that the proposed framework significantly outperforms baselines focusing on either clustering behaviors or preferences alone.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; **Content ranking**.

KEYWORDS

Collaborative Curating; Visual Curation; Visual Discovery

ACM Reference Format:
Dung D. Le and Hady W. Lauw. 2022. Collaborative Curating for Discovery and Expansion of Visual Clusters. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3488560.3498504

1 INTRODUCTION

The Web is probably the largest collective repository of visual imagery ever known. From Instagram stories to Facebook memories, we express ourselves using images in social media. The photos we take on our phones and cameras give fodder to image hosting services such as Flickr and Photobucket. Meanwhile, we fill our spare time seeking and providing visual inspirations through graphical designs, sketchings, infographics, etc., shared on Pinterest.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA

ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3498504

Given such vastness of the Web and the sheer volume of content within, to make sense of even just the sliver that we have encountered and are of personal interest to us, we would resort to creating our own "collections", to save and manage any visual items that catch our eyes. Beyond collecting them, we would further add structures, such as by tagging or categorizing them, for ease of continual access, browsing, and future discoveries.

Of particular interest to this work is a type of user behaviour, which we refer to as *curation*, as exemplified by a few paragons. On Pinterest, users can upload images from their devices or browser extensions and assign an image to a bulletin board. On Flickr, users can curate galleries of public photos from other users. Figure 1 illustrates example boards belonging to a Pinterest user and example galleries of a Flickr user, where images are apparently grouped together according to some semantic concepts (e.g., "teaching tips", "beach", "mimosa", "lotus") that are of the users' own volition. We call the resulting boards or groups of related images *user-curated visual clusters*. While related phenomena exist in other domains (e.g., Spotify song playlists, YouTube video playlists, or Yelp restaurant lists), in this work we maintain a cogent focus on visual clusters and reserve the exploration of other domains for future work.

Problem. We seek to build a learning model from users' curation behaviours. This has the potential to open up novel applications, particularly in supporting the users' future curation activities. As tangible instances, we identify two following potential tasks:

- *Cluster Expansion*: Could we suggest new images that a user would place into one of her existing visual clusters?
- *Cluster Discovery*: Could we suggest to the user new visual clusters that she may be interested to adopt?

We observe that user-curated visual clusters are manifestations of two salient modalities of user behaviour: **preference** - as the user actively seeks new images to be placed in her clusters (or new clusters altogether), she empathically indicates which images she would rather have over alternatives and **similarity perception** - as the user organizes her adopted images, she exercises her powers of discrimination, i.e., what it entails for two images to be related (in the same cluster) or not.

Traditionally, the two modalities mentioned above are studied separately (see Section 3). Preference modeling falls under the purview of collaborative filtering, which is oriented around item recommendations and with few exceptions is not concerned with clustering. On the other hand, similarity modeling is of interest in clustering, which focuses most of the time on an objective sense of similarity that does not usually depend on user preferences.

Approach. We postulate that jointly modeling user preferences and perceived similarities between images would be beneficial to modeling curation behaviors. However, modeling these modalities

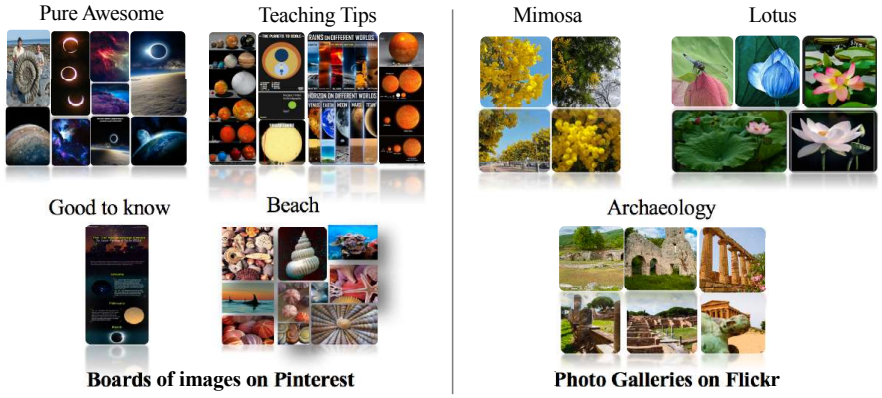

Figure 1: Examples of user-curated visual clusters on Pinterest and Flickr

even separately is already challenging. For one thing, there is the issue of *sparsity*. Since each user only clusters a few images, it is probably insufficient to learn both preferences and perceived similarities from individual user's historical feedback. For another, the *personalized* nature of preferences and user-perceived similarities means that content-based methods that do not take into account users' personal tastes might not be sufficient. To address these challenges, we adopt the collaborative learning paradigm, i.e., learning from the feedback of many users to better handle data sparsity and model user-specific preferences and similarity perceptions.

Contributions and Organization. Our contributions can be summarized as follows. *Firstly*, we introduce the problem of learning preference-infused user-perceived similarities from user-curated visual clusters for solving two personalized tasks: *cluster expansion* and *cluster discovery*. To our knowledge, this paper is the first to explore these problems for visual domain. *Secondly*, we propose the Collaborative Curating framework for modeling personalized curation behaviours in Section 2. In a nutshell, the model incorporates both preference and similarity learning objectives, arriving at latent representations for users, items, as well as clusters, for solving the two above-mentioned tasks. *Thirdly*, we conduct comprehensive experiments on real-world datasets of various categories. Section 4 shows that the proposed approach achieves significantly better performances as opposed to comparable baselines that model preferences and user-perceived similarities separately. To round this up, we review related literature in Section 3, concludes the paper and outlines future work in Section 5.

2 FRAMEWORK

We begin by introducing the notations and formally defining user-curated visual clusters. Thereafter, we describe the proposed Collaborative Curating, how we derive representations of the visual clusters, and how the model parameters would be learned.

User-Curated Visual Clusters. Let $\mathcal{I}$ be a universe of n images and $\mathcal{U}$ be a set of m users. For each user $u \in \mathcal{U}$, we observe his/her feedback in the form of clusters of visual items, formulated as $C_u = \{c_1, c_2, \ldots, c_{n_u}\}$, where $c_l \subset \mathcal{I}, \forall 1 \leq l \leq n_u$ is a subset of images in $\mathcal{I}$ and n_u is the number clusters created by user u. We denote the collection of all user clusters as C, i.e., $C = \bigcup_{u \in \mathcal{U}} C_u$.

2.1 Visual Cluster Representation

A particular image $i \in \mathcal{I}$ is associated with raw, high-dimensional input feature f_i. As per current practice, instead of working with raw features, we presume that from f_i we can derive an embedding vector $z_i \in \mathbf{R}^d$. We further intuit that a visual cluster c likely contains a set of images that cohere around a semantic concept, which manifests in the curation process. Such a "concept" could be represented by an embedding z_c in the same pace.

We now describe the derivation of the representation $z_i \in \mathbf{R}^d$ for a particular image $i \in \mathcal{I}$ and the representation z_c for a cluster c. This component comprises the following blocks (Figure 2):

- A base encoder *enc(.)* that extracts representation vector raw input f_i. One can choose from the rich variety of visual CNN encoders (whose weights are pre-trained or to be trained for the specific tasks at hand). In this study, we adopt a widely popular architecture: ResNet [10] with the depth of 50, pre-trained with ImageNet [6], to obtain the representation vector $h_i = \text{enc}(f_j)$ - the output after the average pooling layer with size 2048.
- A projection head $g(.)$ that maps the extracted vectors from $h_i = \text{enc}(f_j)$ to the embedding space R^d. Here, we use a simple MLP network with one fully connected layer, followed by a nonlinear activation function to obtain the image embedding $z_i = g(h_i) = \sigma_g(\theta_{\text{image}}.h_i) \in \mathbf{R}^d$, in which σ_g is the nonlinear activation function, which usually performs better than a linear function (as suggested in [4]) and θ_{image} is the embedding transformation matrix. In this study, we choose the $\sigma_g(.) = \tanh(.)$.

In short, given an image $i \in \mathcal{I}$ with the raw high-dimensional feature f_i, we derive the image embedding z_i as following:

$$z_i = \tanh(\theta_{\text{image}}.\text{enc}(f_i)) \tag{1}$$

Given a cluster c, we first extract the image vectors $\{z_i | i \in c\}$ of all images in c (described in Equation 1). To learn the representation for cluster c, we follow the architecture proposed in [30] and aggregate the vectors $\{z_i | i \in c\}$ by the *mean* operator $\oplus$ to get the representation z_c for cluster c:

$$z_c = \frac{1}{|c|} \sum_{i \in c} z_i, \tag{2}$$

in which $|c|$ is the number of images in c.

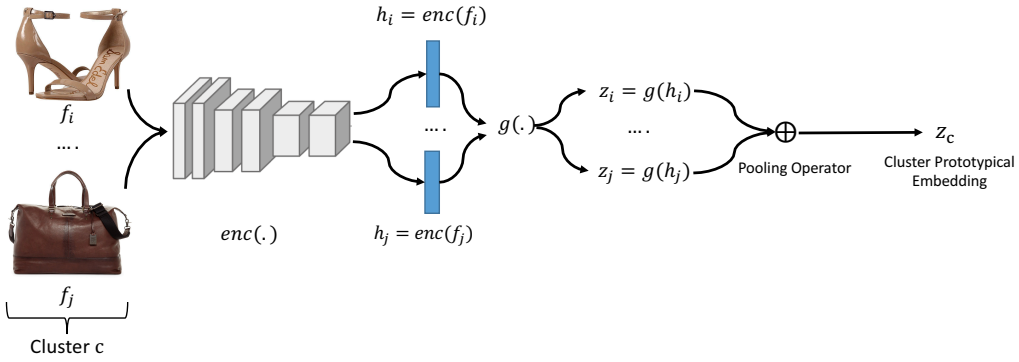

Figure 2: Image and cluster prototypical embedding pipeline

2.2 Collaborative Curating

We now model how user interacts with visual cluster c and an individual image i towards supporting *cluster discovery* and *expansion*.

Preference Modeling. When a user adopts a new cluster or a new item to be placed within one of the existing clusters, a key driver is her preference for the "concept" expressed by that cluster or item. Let us now associate each user u with a preference vector $x_u \in \mathcal{R}^d$. The collection of user vectors is denoted as $X \in \mathbf{R}^{d \times m}$.

We define the user preference score of a user u for an image $i \in \mathcal{I}$ as $\text{pref}_u(i) = x_u^T z_i$ and for a visual cluster $c \subseteq \mathcal{I}$ as $\text{pref}_u(c) = x_u^T z_c$. Note that besides the inner product kernel, alternative formulations of preference score function include l_2 distance [13], angular distance [17], or non-linear function [12, 24].

Let $\mathcal{I}_u^+$ be the collection of those images that u has adopted, i.e., $\mathcal{I}_u^+ = \bigcup_{c \in C_u} c$. As we do not presume any explicit rating over and above the act of adoption itself, we would only model the implicit feedback that u prefers those "positive" images in $\mathcal{I}_u^+$ to the complement $\mathcal{I} \backslash \mathcal{I}_u^+$. Moreover, since presumably u seeks to curate for "concepts" as captured by visual clusters rather than individual items, it would be more empathic to say that u would prefer an adopted visual cluster $c \in C_u$ to a "negative" image $i \in \mathcal{I} \backslash \mathcal{I}_u^+$. Therefore, we seek to minimize the following loss function:

$$\mathcal{L}_{\text{pref}}^u = -\left(\sum_{c \in C_u; i \in \mathcal{I} \backslash \mathcal{I}_u^+} \log \left(\text{SM} \left(\text{pref}_u(c) - \text{pref}_u(i) \right) \right) \right) \quad (3)$$

$\text{SM}(\cdot)$ is the sigmoid/logistic function $\frac{1}{1+\exp(-\cdot)}$. Though the above loss function is reminiscent of implicit feedback models for item preference [11, 23], there is a distinction in that we model preferences for clusters. In addition, the loss will be further coupled with other behavioral objectives to be discussed shortly.

User-Perceived Similarity Modeling. Whether a user would adopt a visual cluster would also likely be influenced by how well the individual items in the cluster collectively meet the user's semantic conception of the cluster. Moreover, when a user seeks to expand an existing cluster, she would be looking at those candidate images that will enhance or augment that conception. In both instances, similarity among images within a cluster matters, but only with respect to the user's perception of similarity.

Our intention is to model such user-perceived similarities so that it can generalize to new items not yet clustered by u. Towards this end, given a user u, we define the potential membership score of an image i to a cluster c as follows:

$$\text{mem}_u(c, i) \propto -\frac{(z_c - z_i)^T W_u (z_c - z_i)}{2} \quad (4)$$

where W_u is a $d \times d$ symmetric, positive semi-definite matrix that represents user u's similarity perception. As in [29], we restrict W_u to be diagonal with non-negative elements. In this paper, we set $W_u = \text{diag}(\text{ReLU}(w_u))$ is the diagonal matrix with $w_u \in \mathbf{R}^d$ whereby ReLU is the rectified linear activation function. Alternative choices for membership function can be found in [25, 29, 30].

For each cluster $c \in C_u$, we want its vector z_c to act as a *"representative"* for all images in c and to capture the user-perceived similarities that connect images in c together. Intuitively, for each cluster c, the similarity between z_c and z_i for $i \in c$ should be higher than that between z_c and z_k for $k \notin c$. To achieve this objective, we consider the following loss function for each cluster $c \in C_u$:

$$\mathcal{L}_c^u = -\sum_{i \in c} \left(\sum_{k \in \mathcal{I} \backslash c} \log \left(\text{SM} \left(\text{mem}_u(c, i) - \text{mem}_u(c, k) \right) \right) + \sum_{c^- \in C_u \backslash \{c\}} \log \left(\text{SM} \left(\text{mem}_u(c, i) - \text{mem}_u(c^-, i) \right) \right) \right) \quad (5)$$

in which $C_u \backslash \{c\}$ is the set of other clusters created by u, but different from c. By minimizing $\mathcal{L}_c^u$, (1) given one cluster c, the vector of an image i in c should be placed closer to the cluster embedding z_c as compared to the vectors of negative images i.e., $\{z_k | k \in \mathcal{I} \backslash c\}$ to z_c and (2) given one image $i \in c$, z_c should be placed closer to z_i as compared to z_{c^-}, the embedding of another cluster of user u to z_i.

Multi-Task Learning. The two modalities of preferences and user-perceived similarity modeling are mutually affecting one another, as manifested through shared parameters. As such, the underlying idea of the Collaborative Curating framework is to jointly model the two within a multi-task learning approach.

The overall objective function is formulated in Equation 6:

$$\mathcal{L} = \sum_{u \in \mathcal{U}} \mathcal{L}_{\text{pref}}^u + \sum_{u \in \mathcal{U}} \sum_{c \in C_u} \mathcal{L}_c^u + \lambda \left(\frac{||\theta_{\text{image}}||_F^2}{2} + \sum_{u \in \mathcal{U}} \frac{||x_u||^2}{2} + \sum_{u \in \mathcal{U}} \frac{||W_u - I_d||^2}{2} \right), \quad (6)$$

in which, $||.||_F$ is the Frobenius norm, λ is the regularization coefficient, θ_{image} is a parameter in deriving the representation of image z_i from its raw features (see Section 2.1). The regularization term $\frac{||W_u - I_d||^2}{2}$ (where I_d is the identity matrix of size $d \times d$), interpreted as an assumption that, a priori, each user perceives the similarity between images based on their visual content.

2.3 Parameter Learning

Algorithm 1 describes the parameter learning process of Collaborative Curating framework, by minimizing the objective function defined in Equation 6 using Adam optimization method. In each iteration, for each user u, we first sample K images from the set of unobserved images $\mathcal{I} \setminus \mathcal{I}_u^+$ to be used as negative examples I_u^- (Line 6). For each cluster $c \in C_u$, we compute the cluster prototypical representation z_c as described in Equation 2 . We sample K negative examples from $\mathcal{I} \setminus c$ to compute the clustering loss in Equation 5 (Line 11). The next step is to compute the pairwise ranking preference loss for user u as described in Equation 3 (Line 15). Lastly, we update the loss function with the regularization for Collaborative Curating's parameters (Line 20). The model parameters will be updated with Adam optimizer. Per iteration, the complexity of Collaborative Curating is $O\left(\sum_{u \in \mathcal{U}} \left(n_u K + \sum_{c \in C_u} \left(|c|K + |c|n_u\right)\right)\right)$.

Algorithm 1 Collaborative Curating

Require: Collection of clusters C, negative sample size K.

1: Initialize θ_{image}; x_u and W_u; $\forall u \in \mathcal{U}$
2: **while** *not converged* **do**
3: $\quad \mathcal{L} \leftarrow 0$
4: $\quad$ **for** user u in $\mathcal{U}$ **do**
5: $\quad\quad$ Sample K negative images I_u^- from $\mathcal{I} / I_u^+$
6: $\quad\quad$ **for** a cluster c in C_u **do**
7: $\quad\quad\quad$ Compute cluster vector z_c (Equation 2)
8: $\quad\quad\quad$ Sample K negative images from from $\mathcal{I} / \{c\}$
9: $\quad\quad\quad$ $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_c^u$, with $\mathcal{L}_c$ described in Equation 5.
10: $\quad\quad\quad$ **for** an image i in I_u^- **do**
11: $\quad\quad\quad\quad$ $\mathcal{L} \leftarrow \mathcal{L} - \log\left(\text{SM}\left(\text{pref}_u(c) - \text{pref}_u(i)\right)\right)$
12: $\quad$ $\mathcal{L} \leftarrow \mathcal{L} + \lambda . \sum_{u \in \mathcal{U}} \left(\frac{||x_u||^2}{2} + \frac{||W_u - I_d||^2}{2}\right) + \lambda . \frac{||\theta_{\text{image}}||_F^2}{2}$
13: $\quad$ Update $\theta_{\text{image}}; x_u, W_u \forall u \in \mathcal{U}$ to minimize $\mathcal{L}$
14: **return** $\theta_{\text{image}}; x_u, W_u \forall u \in \mathcal{U}$

3 RELATED WORK

Semi-Supervised Visual Clustering. Clustering deals with assigning data instances to groups, such that an instance would be more similar to members of the same group than to members of other groups [26]. Our problem is closer to the scenario where side information on the similarity between images is available, and is utilized as constraints to guide the parameter learning process. Such similarity observations may be produced based on human perception. The question is whether that perception is objective or subjective. In this work, we are interested in the latter, i.e., a user's personalized perception of similarity. Previous works such as [5, 9, 29] try to understand the clustering behaviors of the users. In particular, [29] proposes *personalized collaborative clustering or PCC* to model personalized similarity between pairs of items. However, these models only consider the clustering behaviors, but ignore the user preference.

Table 1: Data Summary

Category	#users	#clusters	#images	#adoptions
Science Nature	694	1912	22530	28114
Outdoors	854	2668	23980	28933
Products	923	2610	12759	13632
Technology	570	1623	22368	27238
Sports	1015	3396	91386	100227
Men Fashion	1076	3167	125619	171580

Conditional Similarity Learning. Another line of work focuses on similarity learning in the presence of multiple similarity concepts. The target is not a clustering structure per se, but a learned similarity function. Conditional Similarity Network or CSN [18, 25] assumes that the similarity between images is conditioned on certain semantic similarity concepts. It produces a disentangled representation for images and defines a mask vector for each similarity concept as a filter to map the disentangled representation to concept-specific representation space. There is no consideration of user preferences, which is an essential factor in Collaborative Curating. Our experiments in Section 4 validate this hypothesis.

Moreover, we are interested in *personalized* clustering and ranking of images. This is distinct from the problem in [19], whose goal is to objectively (i.e., non-personalized) cluster the objects into groups and rank these objects simultaneously.

Visual Recommendation. Collaborative filtering [16] learns personalized preference models from user-item interactions in the context of recommending individual items. Given the focus of this work, we pay particular attention to those where items are associated with visual content [11, 14, 28]. These venture beyond user-item interactions to also focus on extracting visual features of products and incorporate them into recommendation models. Our problem is distinct from this group in two important ways. For one, we are not modeling user preferences at the item level, but at the cluster level. For another, beyond user interactions with items or clusters, we expressly model the similarity between items (images) from the same clusters. Of these, we consider VBPR [11] most comparable, and include it as a baseline to test the significance of learning cluster representation and user preferences over clusters of images. Other variants tend to seek improvement in orthogonal means such as by complexifying the underlying CNN model [14].

Our problem is different from bundle recommendation [1–3]. For one, bundles are usually fixed-sized sets of items created by service providers or suppliers to be promoted to users. This setting is different from us, as our visual clusters are dynamically curated by users. For another, previous research on bundle recommendation deals with bundles in the form of baskets of online products [3], or playlists of songs [8, 22], not for visual collections.

4 EXPERIMENTS

4.1 Experimental Setup

Visual Curation Datasets. Since there is no existing benchmarking user-curation dataset for visual domain, we build the datasets

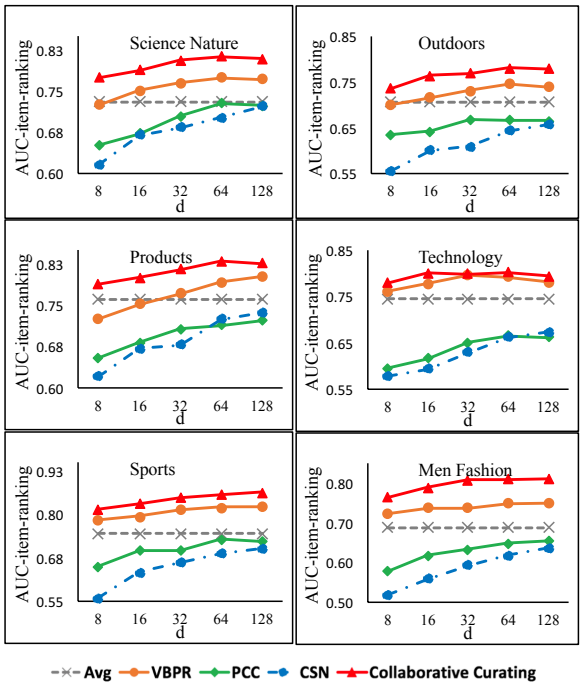

Figure 3: AUC Scores for image ranking (Step 1) as d varies

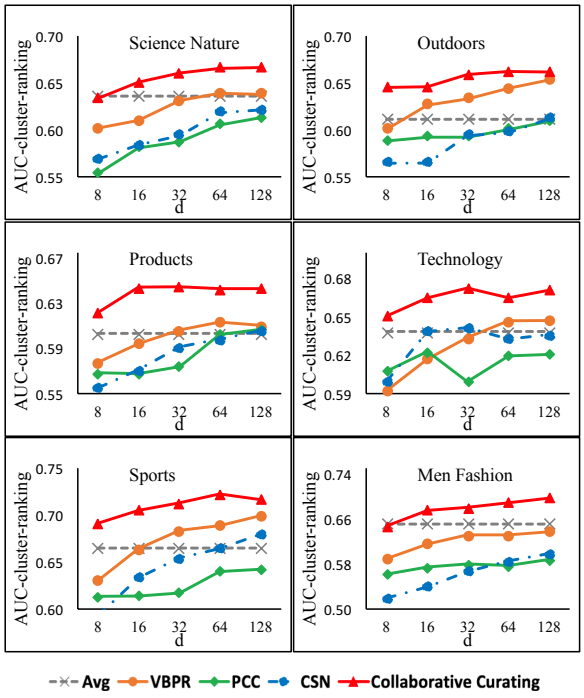

Figure 4: AUC scores for cluster ranking task as d varies

ourselves from data collected from Pinterest[1]. On Pinterest, a user can upload a new image or save an image from another user. The user then chooses a *board* to place the image. We consider each board as a visual cluster and the number of images varies significantly across the boards.

We evaluate our framework on multiple datasets derived from several categories, namely *Science Nature*, *Outdoors*, *Products*, *Technology*, *Sports*, and *Men Fashion*. For each category, we randomly sample a set of 100 seed users. Thereafter, for each seed user, we visit his/her profile and collect the corresponding boards and pins, and following/follower relationships. After collecting seed users' profiles, we pursue the same collection process with their following/follower users. For evaluation, only users with profiles completely collected are included. For each board, we can collect the name, description, and the images that belong to that board. Table 1 summarizes these categories. Adoptions are triplets in the form of (u, c, i), i.e., a user u placing an image i in a cluster c.

Visual Features. For each image i in the six categories, we extract visual feature $h_i = \text{enc}(f_j)$ using the ResNet [10] architecture with the depth of 50, pre-trained with ImageNet [6], where h_i is the output after the average pooling layer with size 2048. Without loss of generality, there could well be other approaches in the literature towards learning image representations for content-based recommendation [7, 15, 20, 27, 31, 32], and such orthogonal techniques could potentially be incorporated into the proposed framework.

Baselines. We consider the following baselines:

- For the first baseline, we simply represent a cluster c by the mean vector of all the visual features of images in c, i.e., h_c = mean($\{h_i | i \in c\}$). We use the cosine similarity between h_c and h_i to measure the membership score of an image i to a cluster c. We refer to this baseline as *Avg* model.
- The second baseline is VBPR [11], a personalized recommendation algorithm with visual signals. Comparing to VBPR will validate the effectiveness of modeling the user-perceived similarity for learning better representation for users and images[2]. The computational complexity of VBPR is $O\left(\sum_{u \in \mathcal{U}} |I_u^+| K\right)$.
- The third baseline Personalized Collaborative Clustering or PCC [29] learns the user-specific pairwise similarity score to fit clustering feedback. The fourth baseline Conditional Similarity Network or CSN [25] models similarities between items, conditioned on users. Comparing to PCC and CSN validates our modeling of user preferences for visual clusters. The computational complexity of PCC and CSN are $O\left(\sum_{u \in \mathcal{U}} \sum_{c \in C_u} \left(|c|^2 K\right)\right)$ and $O\left(\sum_{u \in \mathcal{U}} \sum_{c \in C_u} \left(|c|^2 K\right)\right)$ respectively.

Reproducibility. For project head g of Collaborative Curating, we use a simple nonlinear fully connected layer to map the ResNet features to the desired $d-$ dimensional embedding space. The learning rate and regularization coefficient are tuned for all models to achieve their best performances. Specifically, for all datasets, the learning rate for Collaborative Curating and VBPR is 0.005 and the regularization coefficient is 0.001, the number of negative examples is $K = 20$. For CSN and PCC, the best hyper-parameters setting are: learning rate 0.001, the regularization coefficient 0.001, the number of negative examples K = 20.

[1] www.pinterest.com

[2] The choice of VBPR is motivated by its similarity to our preference-infused component, which engenders greater comparability.

4.2 Personalized Cluster Expansion

The first evaluation task we are interested in this study is *cluster expansion*, i.e., to suggest the images that match user preferences and are also relevant to the similarity concepts expressed by the clusters. For this experiment, we only keep users who have created at least two clusters of images to ensure there is sufficient information to learn their similarity perceptions. To evaluate all models for this task, we use the *leave-one-out* strategy. Specifically, for each user u, we split each cluster c in C_u by selecting a random image to be used for validation and another for testing.

$$c = c^{\text{train}} \cup c^{\text{val}} \cup c^{\text{test}}, \forall c \in C_u; \tag{7}$$

This effectively creates the training clusters $C_u^{\text{train}} = \bigcup_{c \in C_u} c^{\text{train}}$, validation images $I_u^{\text{val}} = \bigcup_{c \in C_u} c^{\text{val}}$, and testing images $I_u^{\text{test}} = \bigcup_{c \in C_u} c^{\text{test}}$ for each user u.

A natural strategy to suggest images to existing clusters of user u is to perform the following sub-tasks:

(1) *Step 1: Image Ranking* Identify the top images that the user u might prefer.
(2) *Step 2: Cluster Selection* Assign each recommended image to the cluster that returns the highest membership score.

Since the observed test image of a cluster might not appear at the top of the image ranking list of the user (and thus would not participate in Step 2), a reasonable approach is to evaluate these sub-tasks separately so we can associate each with their respective known ground-truths.

For Step 1, we employ the widely used metric AUC (Area Under the ROC Curve), to evaluate the predicted personalized ranking on the hidden images I_u^{test} for each user u:

$$\text{AUC} - \text{item} - \text{ranking} = \frac{1}{|\mathcal{U}|} \sum_u \frac{1}{|E(u)|} \sum_{(i,j) \in E_u} \delta(\text{pref}_u(i) > \text{pref}_u(j)) \tag{8}$$

in which, $E(u) = \{(i,j) | i \in I_u^{\text{test}}; j \notin I_u^{+}\}$ and $\delta(x) = 1$ if $x > 0$ and 0 otherwise. Higher AUC indicates higher performance. For models that do not define an explicit preference score function such as CSN, PCC, and Avg, we rank the items based on their membership scores to all the user's training items instead. For Step 2, we also measure the AUC score to investigate how the algorithms select a cluster for a recommended image:

$$\text{AUC} - \text{cluster} - \text{selection} = \sum_{u \in \mathcal{U}} \sum_{i \in I_u^{\text{test}}} \sum_{c' \in C_u^{\text{train}} \setminus \{c_i\}} \frac{\delta(\text{mem}_u(c_i, i) > \text{mem}_u(c', i))}{|\mathcal{U}|.(n_u - 1).|I_u^{\text{test}}|}, \tag{9}$$

in which, c_i indicates the cluster image i belongs to.

For each split of a dataset, we select the best model via hyper-parameter tuning based on the validation set and report the performance on the test set. We conduct experiments on 5 different random splits and report the average results. Figures 3 and 5 respectively show the AUC scores of all models for the *image ranking* and *cluster selection* sub-tasks across different values of dimension $d = 8, 16, 32, 64, 128$. Across all categories, Collaborative Curating consistently shows better performances in both sub-tasks. The results are mostly statistically significant with $p < 0.01$ based on the paired sample t-test. For *image ranking* on *Technology*, Collaborative Curating is comparable to VBPR when $d = 32$ and for *cluster selection* on *Sports*, VBPR achieves competitive performance compared to Collaborative Curating. Jointly modeling preference and user-perceived similarity enhances the quality of image recommendation for each user and each cluster of the user.

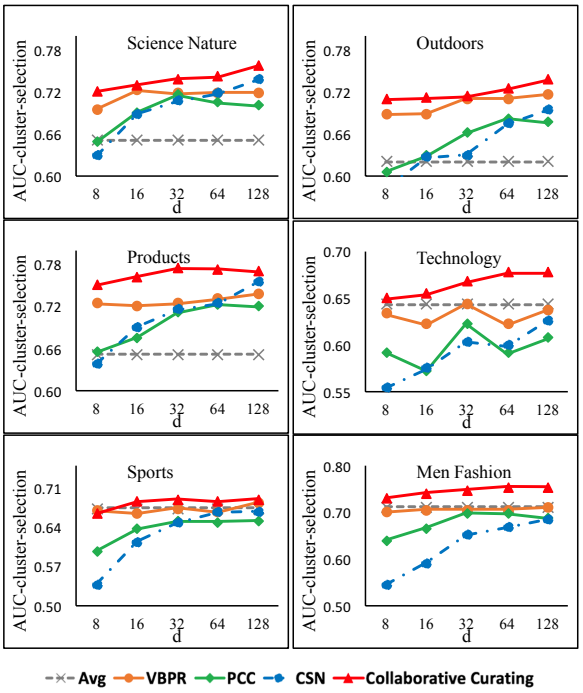

Figure 5: AUC scores for cluster selection (Step 2) as d varies

For *image ranking* (Figure 3), baselines CSN and PCC consistently show worse performances. This is attributed to how these models only consider the user-perceived similarity factor. Compared to these, Collaborative Curating shows significantly better results, implying that incorporating user preferences factor is critical for this task. For *cluster selection* (Figure 5), Collaborative Curating achieves substantial improvement over the baselines. The result shows that our framework can learn the user-perceived similarity perceptions and generalize to the new images effectively. For this task, Avg model performs fairly well on *Technology, Sports, Men Fashion* categories, while VBPR shows competitive results as well (except for *Science Nature* and *Products* categories, CSN outperforms VBPR when $d = 128$). This could be due to utilization of informative ResNet50 visual features to represent the images.

4.3 Personalized Cluster Discovery

The second evaluation task is to suggest a new cluster (created by another user) to a target user u. The motivation is to help the user u to discover new ideas from other users, who might share similar interests as u. For this experiment, we consider users with at least three clusters of images. Particularly, for each user u, we randomly hide one of their clusters for validation and another for testing, i.e.,

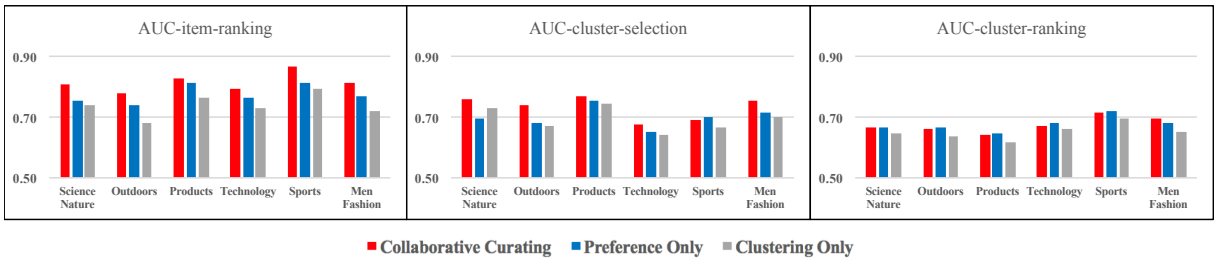

Figure 6: Ablation Analysis: Performances of Collaborative Curating and its variants $d = 128$

$C_u = C_u^{\text{observed}} \cup C_u^{\text{val}} \cup C_u^{\text{test}}$. Across all users, we denote all the testing clusters as C^{test}, i.e., $C^{\text{test}} = \bigcup_{u \in \mathcal{U}} C_u^{\text{test}}$.

For evaluation, for each user u, we first rank all the test clusters in C^{test} by the preference scores for clusters as defined in Eq. **??**. Here, we treat other test clusters from other users as negative clusters. We measure AUC to see how well the test cluster of a user u is ranked against u's negative clusters. The metric is given in Eq. 10.

$$\text{AUC} - \text{cluster} - \text{ranking} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{c \in C^{\text{test}} \setminus C_u^{\text{test}}} \frac{\delta\left(\text{pref}_u(C_u^{\text{test}}) - \text{pref}_u(c)\right)}{|C^{\text{test}} \setminus C_u^{\text{test}}|}. \quad (10)$$

Figure 4 shows the $\text{AUC}_{\text{cluster-ranking}}$ scores of all models on the test data (the average results of 5 random splitting). Across all datasets, Collaborative Curating achieves significant improvement over the baselines. The improvement is statistically significant with $p < 0.01$ based on the paired sample t-test (except for *Science Nature*, Avg baseline is slightly better than Collaborative Curating). Especially, Collaborative Curating shows better results than VBPR, supporting the effectiveness of modeling user preference for images at individual and cluster level.

Again, baselines CSN and PCC are the weaker models for this task, since both models are not optimized for learning user preferences. This again highlights the importance of incorporating user preference signals for suggesting new visual clusters to users. Avg model also achieves relatively good results for this task, especially on *Men Fashion* category, but still lower than ours. This shows that our adoption of the collaborative learning paradigm approach is more accurate compared to content-based solutions that do not take into account user personal preference and similarity perceptions.

4.4 Ablation Analysis

To investigate the contribution of our architecture components, we conduct an ablation analysis on several variants of Collaborative Curating. The first variant is *Preference Only*, which only models the user preference, but ignore the user perceived similarity factor. The second variant is *Clustering Only*, which only considers the user perceived similarity, but ignoring the user preference factor.

Figure 6 shows the AUC scores of all variants for the three tasks: *image ranking*, *cluster selection*, and *cluster ranking* on the six categories with $d = 128$. The full architecture Collaborative Curating (in red color) achieves highest AUC scores compared to other variants for *image ranking* and *cluster selection* tasks. This shows that jointly modeling user preferences and user-perceived similarities are beneficial. For *cluster ranking*, *Preference Only* achieves competitive results compared to the full architecture Collaborative Curating, and outperforms *Clustering Only*. This implies that modeling user preference for clusters is critical for *cluster discovery* task. Also, *Preference Only* variant achieves competitive performances for *cluster selection*, compared to *Clustering Only*, showing that user preference has useful signals towards learning the user-perceived similarity as well.

4.5 Case Studies

To better understand our framework, we illustrate several examples of recommendations generated by Collaborative Curating here.

Case Study #1: Cluster Expansion and Discovery. Figure 7 shows the visual clusters of a user from *Science Nature* category, and the users' recommendations for *cluster expansion* and *cluster discovery* tasks. The first row displays the existing clusters separated by the grey lines, with the corresponding names annotated by the corresponding user. The entire second row displays the top-20 recommended images returned by Collaborative Curating for the user, each of which is assigned to the cluster with highest membership score. Each column in the second row displays the recommended images (from the above top-20 images) to "expand" the corresponding cluster above. The third row is the top-1 cluster recommendation (with annotated name), created by another user. We observe that the top-20 images (second row) and the top-1 recommended cluster are significantly relevant to the images of the existing clusters. This shows that, Collaborative Curating can effectively learn the user preferences (both at the individual image and cluster level) for high quality recommendations. For each *existing* cluster, the recommended images are also semantically similar to the existing images in the cluster. This further illustrates the effectiveness of Collaborative Curating in learning the user-perceived similarities and generalizing to new images.

Case Study #2: Clusters Visualization. Figures 8 shows the 2-d t-SNE [21] visualization of all clusters from *Products* category with original $d = 128$. For each cluster, we choose the image whose embedding is closest to the cluster's embedding as the thumbnail.

From Figure 8, we observe that thumbnail images form several visual clusters that are semantically relevant, such as a cluster of *"shoes"* at the bottom-right corner, a cluster of *"interior furniture"* at the bottom left, a cluster of *"mugs"* at the middle left, or clusters of *"fashion items"*. This shows that Collaborative Curating is able to learn informative representation of the visual clusters, which is useful to help users to discover clusters created by other users.

Figure 7: Examples for cluster expansion and discovery recommendations (best viewed in color)

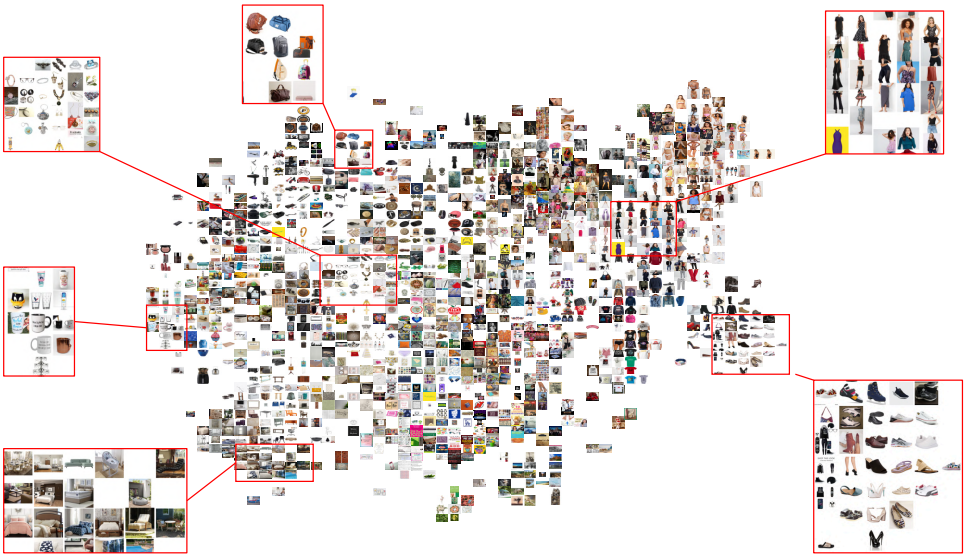

Figure 8: Visualization of clusters from *Products* category with $d = 128$.

5 CONCLUSION AND FUTURE WORK

We are interested in curation behaviors for visual content, whereby a user expresses preferences as well as perception of similarity. To model such behaviors towards supporting *cluster expansion* and *cluster discovery*, we propose Collaborative Curating that jointly models both modalities within a unified framework. Experiments on six categories show that the model outperforms baselines that consider preferences or user-perceived similarities separately.

There are several facets in Collaborative Curating that are worth further investigation such as the effect of different base encoder architectures on the quality of visual representations. It may also be interesting to study how the framework could be extended or modified to model curation behaviors in non-visual domains.

REFERENCES

[1] Jinze Bai, Chang Zhou, Junshuai Song, Xiaoru Qu, Weiting An, Zhao Li, and Jun Gao. 2019. Personalized bundle list recommendation. In *WWW*. 60–71.

[2] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle Recommendation with Graph Convolutional Networks. In *SIGIR*. 1673–1676.

[3] Liang Chen, Yang Liu, Xiangnan He, Lianli Gao, and Zibin Zheng. 2019. Matching User with Item Set: Collaborative Bundle Recommendation with Deep Attention Network.. In *IJCAI*. 2095–2101.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[5] Ting-Yu Cheng, Guiguan Lin, Kang-Jun Liu, Shan-Hung Brandon Wu, et al. 2016. Learning user perceived clusters with feature-level supervision. In *NIPS*. 532–540.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

[7] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *WWW*. 1775–1784.

[8] Anna Gatzioura, João Vinagre, Miquel Sànchez-Marrè, et al. 2019. A hybrid recommender system for improving automatic playlist continuation. *TKDE* (2019).

[9] Ryan G Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. 2011. Crowdclustering. In *NIPS*. 558–566.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[11] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[13] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *WWW*. 193–201.

[14] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *ICDM*. IEEE, 207–216.

[15] Dmitry Kislyuk, Yuchen Liu, David Liu, Eric Tzeng, and Yushi Jing. 2015. Human curation and convnets: Powering item-to-item recommendations on pinterest. *preprint arXiv:1511.04003* (2015).

[16] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[17] Dung D Le and Hady W Lauw. 2017. Indexable Bayesian personalized ranking for efficient top-k recommendation. In *CIKM*. 1389–1398.

[18] Jongpil Lee, Nicholas J Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam. 2020. Disentangled Multidimensional Metric Learning for Music Similarity. In *ICASSP*. IEEE, 6–10.

[19] Jiyi Li, Yukino Baba, and Hisashi Kashima. 2018. Simultaneous Clustering and Ranking from Pairwise Comparisons.. In *IJCAI*. 1554–1560.

[20] Yuchen Liu, Dmitry Chechik, and Junghoo Cho. 2016. Power of human curation in recommendation system. In *WWW*. 79–80.

[21] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, Nov (2008), 2579–2605.

[22] Piyush Papreja, Hemanth Venkateswara, and Sethuraman Panchanathan. 2019. Representation, Exploration and Recommendation of Playlists. In *ECML PKDD*. Springer, 543–550.

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and ST Lars. 2009. Bpr: Bayesian personalized ranking from implicit feedback. UAI'09. *Arlington, Virginia, United States* (2009), 452–461.

[24] Bo Song, Xin Yang, Yi Cao, and Congfu Xu. 2018. Neural collaborative ranking. In *CIKM*. 1353–1362.

[25] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. 2017. Conditional similarity networks. In *CVPR*. 830–838.

[26] Rui Xu and Don Wunsch. 2008. *Clustering*. Vol. 10. John Wiley & Sons.

[27] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *KDD*. 974–983.

[28] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *WWW*. 649–658.

[29] Yisong Yue, Chong Wang, Khalid El-Arini, and Carlos Guestrin. 2014. Personalized collaborative clustering. In *WWW*. 75–84.

[30] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep sets. In *NIPS*. 3391–3401.

[31] Andrew Zhai, Dmitry Kislyuk, Yushi Jing, Michael Feng, Eric Tzeng, Jeff Donahue, Yue Li Du, and Trevor Darrell. 2017. Visual discovery at pinterest. In *WWW*. 515–524.

[32] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. 2019. Learning a Unified Embedding for Visual Search at Pinterest. In *KDD*. 2412–2420.