Understanding the Role of Style in E-commerce Shopping

Hao Jiang[1]*, Aakash Sabharwal[2]*, Adam Henderson[2], Diane Hu[2], Liangjie Hong[2]

[1]Twitter, San Francisco, U.S.A
[2]Etsy, New York, U.S.A
{hjiang}@twitter.com,{asabharwal,ahenderson,dhu,lhong}@etsy.com

ABSTRACT

Aesthetic style is the crux of many purchasing decisions. When considering an item for purchase, buyers need to be aligned not only with the functional aspects (e.g. description, category, ratings) of an item's specification, but also its stylistic and aesthetic aspects (e.g. modern, classical, retro) as well. Style becomes increasingly important on e-commerce sites like Etsy, an online marketplace for handmade and vintage goods, where hundreds of thousands of items can differ by style and aesthetic alone. As such, it is important for industry recommender systems to properly model style when understanding shoppers' buying preference. In past work, because of its abstract nature, style is often approached in an unsupervised manner, represented by nameless latent factors or embeddings. As a result, there has been no previous work on predictive models nor analysis devoted to understanding how style, or even the presence of style, impacts a buyer's purchase decision.

In this paper, we discuss a novel process by which we leverage 43 named styles given by merchandising experts in order to bootstrap large-scale style prediction and analysis of how style impacts purchase decision. We train a supervised, style-aware deep neural network that is shown to predict item style with high accuracy, while generating style-aware embeddings that can be used in downstream recommendation tasks. We share in our analysis, based on over a year's worth of transaction data and show that these findings are crucial to understanding how to more explicitly leverage style signal in industry-scale recommender systems.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

supervised learning; embeddings; style; e-commerce; user behavior

ACM Reference Format:
Hao Jiang, Aakash Sabharwal, Adam Henderson, Diane Hu, Liangjie Hong. 2019. Understanding the Role of Style in E-commerce Shopping. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19), June 22–24, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330760

* These authors equally contributed to this work.
[1] This work was done while Hao Jiang was at Etsy.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330760

1 INTRODUCTION

A buyer's purchase preference is often shaped by a myriad of factors. These factors vary from attributes that have concrete values such as item specification, description, ratings, and cost, to more subjective characteristics such as aesthetics and style. At Etsy, an online marketplace for handmade and vintage goods with over 30 million diverse listings, the ability to capture style is particularly important as buyers often come to the site to find items that match their eclectic tastes. In many cases, style can be the differentiating factor between tens of thousands of items under the buyer's consideration. For example, a search on Etsy for "large canvas tote bag" under the "Bags & Purses" category for "under $25" and "ships to the U.S." still matches around 2,500 items. This example illustrates that even after satisfying a number of concrete item specifications (functionality, size, price, category, material, and shipping logistics), the buyer is still confronted with a large number of relevant items by which style is a major differentiating factor. As such, recommender systems should explicitly model style preference in order to holistically understand a buyer's purchase preference.

Unfortunately, style can be difficult to define. Many buyers may not have the vocabulary needed to describe a particular style that they like. Items may also be associated with a mix of different styles, or no strong style at all. This difficulty is reflected in the diverging ways in which style-based recommender systems have been approached in the past. Broadly speaking, there have been two approaches to represent style: (1) as a latent space, learned in an unsupervised manner, that tries to explain why certain groups of items are preferred by users over others [9, 11, 13, 16, 28], or (2) as a disparate collection of low-level attributes, such as color, material, texture, depth-of-field and other visual cues [1, 17, 22, 24, 32]. Both approaches, for the most part, end up rendering style in a nameless and un-interpretable manner, showing only how it can be used to improve overall recommendation accuracy, but imparts little insight into how distinctive styles, or even the presence of style, impacts the way buyers interact with them.

In contrast, merchandising experts view style in very concrete terms, and often refer to styles by name. For them, style is the vocabulary which one can use to evaluate whether or not an item is beautiful or aesthetically pleasing [35]. Though there exists a potentially infinite number of different styles, over time, specific stylistic names have emerged in the community to loosely describe aesthetically similar groupings for ease of communication. Some of these styles are rooted in historical art and cultural movements and time periods (e.g. Mid-century modern, Art Nouveau, Hollywood Regency); others have simply developed from modern-day culture

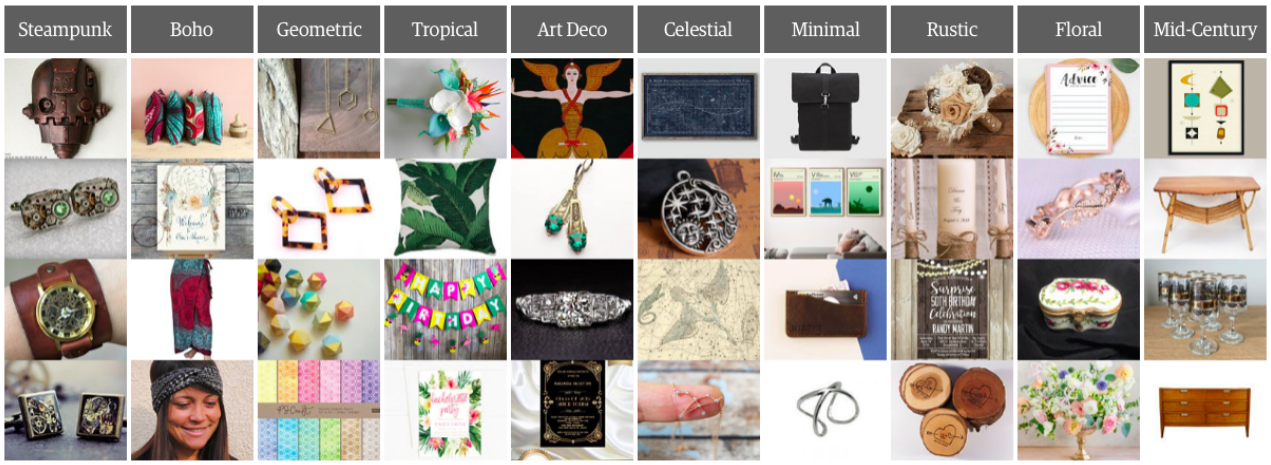


Figure 1: Test examples that scored highest for each style class, according to the style classifier described in Section 3.

to reflect the diverse niches of lifestyles and interests (e.g. Romantic Goth, Wanderlust, Cottage, Boho Chic, Woodland).

In this paper, we propose a novel method that blends the best of both worlds: We leverage domain expertise to extract latent style embeddings in a supervised manner that preserves the interpretability of the representation. More specifically, we obtain a list of 43 style classes from merchandising experts and collect a labeled "ground truth" dataset of items that belong to each of the style classes using implicit search log data. We then train a multi-modal, style-aware neural network using both text and image features that can: (1) generate low-dimensional embeddings that meaningfully capture style, and (2) classify items into one of the 43 style classes.

We evaluate our style model by deploying the learned embeddings as a candidate set in a production listing-to-listing recommender system and find that the new method produces recommendations that are more similar in style. We also show that the multi-modal style classifier achieves high accuracy over other competitive baselines on two labeled datasets, one derived from implicit feedback, and the other produced by merchandising experts. Using the style classifier, we obtain style predictions for all 50 Million active items on Etsy.com and use these predictions to understand how styles vary against taxonomies, seasons, and price ranges, as well as shopping behavior such as user favoriting, searching, and purchasing. Among our findings, we show that items with a strong Strength of Style (any style) are significantly correlated with more favorites, and more purchases. We also find that certain styles are purchased more or less frequently, depending on the season or occasion. To the best of our knowledge, this is the first system to (1) learn interpretable styles based on expert annotation that generalizes across all e-commerce related categories, and (2) perform a large-scale analysis of how style impacts shopping behavior.

2 RELATED WORK

The concept of visual style has been the subject of much discussion in past literature and touches on several different bodies of work, including visual and style-aware recommender systems, as well as models that focus on predicting styles or other related visual cues.

Before discussing visual style, we first describe general, visually-aware systems in which learned models detect or segment objects that are present in an image or scene [3, 7, 31]. Earlier work in this field made use of hand-crafted features and focused on sliding window classification [4, 5, 29], while more recent work leverages the power of CNNs and R-CNNs [6–8, 21, 37]. There have also been efforts to fuse image and text signals to achieve useful mid-level semantic representations and higher performance [2, 10, 18–20].

In recent years, many recommender systems have come to leverage image-recognition methods to improve recommender systems by finding visually similar items to aid in product discovery [14, 34, 36] or to understand user visual preferences to better personalize recommended items [13, 23]. This body of work is more concerned with identifying the identity of the object present, as opposed to understanding the style of the object. Of greater relevance to our proposed work are those visual recommender systems that have been developed for the online fashion industry, where explicit clothing items (e.g. blouse, pants, hat) are "parsed" [33], and visual attributes are extracted from each item in order to perform such tasks as judging clothing compatibility and fashionability [11, 12, 16, 27].

In general, some systems try to exploit the visual language of style, using a variety of low-level visual clues, such as image color, material, and texture to collectively model style [1, 17, 22, 24, 32]. Others represent style as latent spaces or embeddings that are learned by way of implicit user feedback, generalizing upon the idea that, for example, two t-shirts or two cups should have similar "style embeddings" if they are frequently favorited or purchased together by the same user, or same group of users [9, 13, 16, 23, 28].

While our proposed model is motivated by various multi-modal approaches for extracting image and text features, we optimize for style prediction instead of object detection. We also do so in a supervised manner by leveraging domain expertise such that the underlying system can give interpretable results, leading to more explainable recommendations. Most previous style-based recommender systems are also confined to a single domain, such as fashion, or artwork, rendering them sub-optimal for general e-commerce sites that span across many different product categories.

Our work specifically addresses styles that are diverse enough to capture user taste across all categories relevant to e-commerce.

3 STYLE MODEL

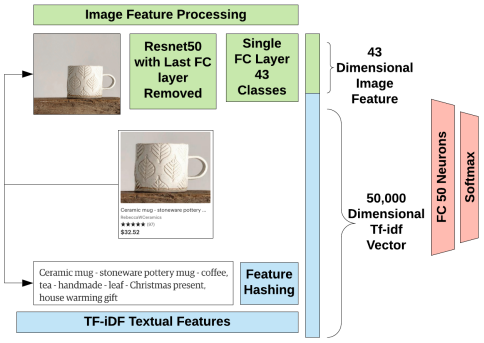


Figure 2: Neural Network Architecture to combine Image and Text (Tf-idf) features.

In order to explicitly understand how style impacts buyers' shopping behavior, we first develop a style recognition model that is capable of capturing the subjective notions of aesthetic and visual design of an item. In this section, we describe a supervised learning approach that learns a style representation for each item by collecting ground-truth labels from implicit feedback logs and extracting both text and image features. The outcome and application of this style representation are discussed next in Section 4.

3.1 Large-scale Style Data Collection

One of the central limitations of learning style-aware models in supervised fashion is the lack of labeled data for training. There are several reasons for why obtaining this type of labeled data is challenging: First, it can be difficult to define what "style" means, let alone decide what the different style classes should be. Even across existing style-based work, the nature of style classes vary: e.g. some focus on image-quality [6–8, 21**?**], others focus on a coherence of theme [11, 12, 16, 27]. Second, having decided on a consistent style definition and a set of style classes, significant time and domain expertise is required to provide a manually annotated dataset. Even if this is achieved, this method is not scalable for obtaining a dataset that is large enough to properly train a neural network.

At Etsy, we address these limitations by leveraging two unique aspects of the Etsy marketplace. First, we invite experts on the Etsy merchandising team to select a set of 42 styles that, from their experience, capture buyers' taste across all 15 top-level categories on Etsy. These styles vary from visual art styles (*mid-century*, *art nouveau*) to emotional themes (*fun and humor*, *inspirational*) to lifestyles (*boho*, *farmhouse*) and cultural trends (*memphis*, *scandi*, *hygge*). Second, we take advantage of the fact that a small population of Etsy buyers are, in fact, familiar with different stylistic names, and use them in their search queries (e.g. "mid-century modern nightstand").

From this, we are able to collect a large-scale dataset of "ground truth" examples for each of the 42 style classes. We do so by considering any search query that contains the text of one of the style classes. For each of these queries, we assign that style as the ground truth label to all items that the user interacted with from the search results for that query. For example, for a user who searched "art deco ring", the items the user clicked on, favorited, added-to-cart and purchased within the same search query session were assigned to the "art deco" class. In this way, we are trusting buyers who are savvy with style names on Etsy to implicitly provide style-based feedback for our training set. By considering only the items that were interacted with, we have more confidence that these items accurately embody the style in question. Based on this idea, we are able to collect a large labeled dataset of about 3 million instances just by considering query logs from the month of September, 2018. In addition, we generate a negative class, called **others**, by sampling the interacted listings from the queries that do no contain any of the 42 style keywords, hence leaving us with 43 style classes.

3.2 Supervised Learning of Style

Using the data collection approach discussed above, we train a deep neural network that learns a fused text and image representation that can best distinguish between the 43 different style classes for any item. We discuss the text and image features used, as well as the neural network architecture in turn below.

3.2.1 Text Features. Etsy sellers spend a lot of time optimizing the title and tags of their listings with keywords that best summarize different aspects of their listing. These keywords not only offer insight into the functional aspects of the listing, but also the stylistic aspects as well, often covering attributes such as: color, taxonomy, materials, time periods, patterns, etc. We compute a *TF-IDF* feature using a *Boolean Term Frequency* model for each listing that incorporates all of the text contained in the title and tags of a listing. We then use feature hashing [26, 30] to reduce the dimensionality of the *TF-IDF* feature space from $1.3M$ to $50k$ dimensions. This allows us to use more expensive, non-linear models such as neural networks, which perform more effectively with dense vectors.

We also note that our data collection method (Section 3.1) may suffer from bias when the text features themselves contain corresponding style keywords. While these features will, no doubt, be highly predictive of the style, we want the model to learn features that generalize well beyond the style keyword itself. To mitigate this, we filter out all terms that contain a style keyword before computing the hashed *TF-IDF* feature. We do this in order to ensure the robustness of our training setup and to ensure that our model generalizes to listings that don't explicitly have such style keywords as features, but still may fall into a style class.

3.2.2 Image Features. Images offer a rich source of information that is typically complementary to the textual features. Especially for style, the influence of visual aesthetic cannot be understated. On Etsy, each seller is given the opportunity to upload up to 10 images per listing. We proceed to extract features from these images using a Deep Convolutional Neural Network (CNN). Instead of using the raw features learned from a pre-trained network (most often optimized for object detection), we fine-tune a model based on our labeled dataset of 43 style classes in order to learn image-based mid-level representations that are style-aware.

We chose the Resnet50 architecture [8] as it offered a good balance between model complexity and accuracy versus infrastructure constraints such as GPU memory or time per epoch. The input to

the Resnet50 model is an image of size 224 x 224 pixels. Listing images often vary in size so various transformations including resizing and normalization are done before feeding them into the deep neural network. Before fine-tuning the Resnet model we replace the last fully connected layer (1000 neurons) with a new fully connected layer of 43 neurons. The output of this 43 neuron layer is passed through a softmax function to give a probability distribution over the 43 styles. Figure 2 shows our retraining setup.

Apart from re-training this last linear layer, we subsequently tried retraining the last n layers. The hypothesis here is that earlier layers are learning structural and foundational components of an image (that may be agnostic to something like detecting objects versus learning style), and the last few layers are fine-tuning the objective function towards a specific problem domain (e.g. the learning of style). (We would like to retrain the full network in future iterations.) Table 1 confirms this hypothesis as accuracy increases when retrained on the last two layers of Resnet50 instead of the last one. There are, however, diminishing returns in performance when $n > 2$.

Model	N-Last Layers Retrained	Accuracy
Alexnet	1	0.149
Resnet50	1	0.294
Resnet50	2	0.392
Resnet50	3	0.388

Table 1: Accuracy from retraining N-Last layers

3.2.3 Neural Network Architecture. To build the final style model, we use a two layer neural network (Figure 2) to combine the image and text features described above in non-linear fashion. The image features are from the final layer output of the retrained Resnet as detailed in Section 3.2.2. The text features are the *TF-IDF* values computed on the titles and tags of items, and then hashed to a vector of size $50,000$, as described in Section 3.2.1. The image and text vectors are then concatenated to obtain a $50,043$-dimensional vector that is used as input into the neural network model.

The neural network itself consists of a single fully connected hidden layer of 50 neurons. The final output layer is a fully connected layer with 43 neurons followed by softmax. We optimize for cross-entropy loss in our training. We note that the image features described in Section 3.2.2 are trained separately and used as input into this stage of the model. We trained our neural network model using Pytorch[25] with a batch size of 512 documents, initial learning rate of *0.1*, adam for stochastic optimization, a training dataset of size *2 million* instances and a test dataset size of *700,000* instances. 96 loader worker threads were used for loading the dataset on to the GPU. Each training epoch roughly took 15 mins on a single GPU core. Our early stopping criteria was that training loss not change for 3 consecutive epochs with a max of 20 epochs. In total our model had about *2.5 million* parameters.

4 EXPERIMENTS AND APPLICATIONS

In this section, we discuss two applications of the style model described in the previous section: (1) to predict the style of *all* items on Etsy, and (2) to extract mid-level style-aware representations that are useful for downstream recommender system tasks.

NO.	Structure	Features	Accuracy
1	MLR	TF-IDF	0.592
2	NN	TF-IDF	0.752
3	CNN	Image	0.535
4	MLR	TF-IDF + Image	0.710
5	NN	TF-IDF + Image	0.762

Table 2: Accuracy Results from Training different style classification models.

4.1 Style Prediction

As detailed in earlier sections, because of the time and domain expertise required, its impractical to get a style classification for all 50 million active listings on etsy. Examining the search queries with style keywords does not scale as only about 3% of the search queries contain a style keyword and correspondingly only a few items are interacted for those queries. Thus, we hope to leverage our style predictor to gather this style label for *all* our listings for downstream tasks.

We formulate our style prediction problem as a supervised, multi-class classification problem, in which the end goal is to classify each item into one of the 43 style classes. To perform this classification, we use the multi-modal neural network described in Section 3.2.3, and use the final 43-dimensional softmax output as our predicted probability class vector. We assign the class with the highest score as the final prediction. We can also use the 43-dimensional output to understand the mixture of different styles that are present in the item, as an item can often embody the characteristics of one or more style.

Table 2 shows the style prediction accuracy on our held-out test set (25% of the overall training set). We compare the accuracy against several ablation baselines to evaluate the effectiveness our proposed model, and describe each in turn below:

- **Text + MLR**: Multinomial Logistic Regression with text features only.
- **Text + NN**: 2-layer Neural Network with text features only.
- **Images + CNN**: Convolution Neural Network with a listings primary image only.
- **Text + Image + MLR**: Multi-nomial Logistic Regression with the concatenated text and image feature vectors.

We can see from the results in Table 2, neural network models are able to better predict the style of a listing and using image features in addition to text features helps in improving the final accuracy. We picked the neural network model that combines image and text features and gives an accuracy of about 76% as the final model for downstream tasks. The addition of images over text gives a pretty negligible benefit. We hypothesis that this is due to the images providing limited new information versus text. To address this, in future iterations, we would like to explore jointly training the image feature model (modified Resnet50 detailed in Section 3.2.2) and the final text + image feature style classifier.

4.2 Style Embeddings

In addition to style prediction, we describe an important, second application of our model: the ability to extract low-dimensional, style-aware representations that can capture style succinctly. To

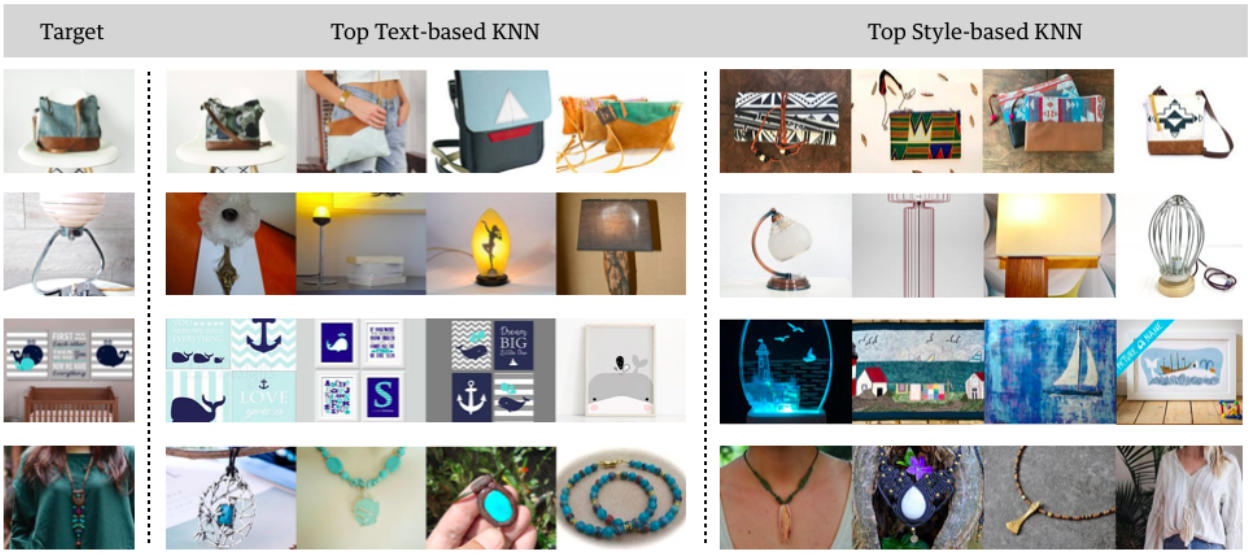


Figure 3: Nearest Neighbors for listings based on TF-IDF Vector & Style Embeddings

obtain this low-dimensional representation, we extract the penultimate layer of the neural network, giving us a 50-dimension vector, which we refer to as *style embeddings*. In this hidden layer output, the neural network tries to learn lower dimension embeddings that are most predictive of the final output label without having to label each listing as a particular style. This helps it represent the listing as a dense vector that best captures the correlations of different input text and image features with different style categories. This particular representation can be very useful in downstream tasks, such as a style-aware recommender system, as we discuss later in Section 4.2.2.

To evaluate the quality of these learned embeddings, we examine which characteristics of style have been captured and whether or not they align with our own intuition of style.

Style Similarity Within a Shop or Taxonomy. Sellers often adhere to a consistent style across all the items that they sell. As such, we expect that a pair of listings drawn from the same shop will have more similar style embeddings than if the pairs were drawn at random. In Table 3, we compare the cosine similarity between the style embeddings from different pairs of listings, across same/different shops, as well as same/different taxonomies. It is clear that style embeddings between pairs of listings from the same shop have a higher cosine similarity than listings from different shops, confirming our intuition of within-shop style coherence. In addition, we note that our style embeddings also capture some notion of taxonomy, as pairs of listings from the same taxonomy have higher similarity scores across the board.

Style Similarity among User Preferences. We perform the same type of evaluation as above, but with the hypothesis that groups of items favorited or purchased by the same user should also have similar style. In Table 4 we compare the cosine similarity of pairs of listings purchased by the same user vs a pair of listings purchased by different users, and we find a statistically significant difference in the mean of their distributions. The trend is the same

	Same Shop	Different Shop
Same Taxonomy	0.649	0.558
Different Taxonomy	0.601	0.521

* Any two values in the table differ statistically significantly.

Table 3: Cosine Similarity between pairs of listings from different (or same) shop and taxonomy

	Purchase	Favorite
Same User	0.877	0.871
Different User	0.856	0.851

* Any two values in the table differ statistically significantly.

Table 4: Cosine Similarity between pairs of listings from different (or same) user that were purchased or favorited

when we consider user favorites. This confirms our intuition that our style embeddings are able to capture the stylistically similar patterns in users' favoriting and purchasing behavior.

4.2.1 Quantifying Style Similarity. One application of these style embeddings is the ability to efficiently quantify how stylistically similar two listings are. We can do this by retrieving the nearest neighbors of a target listing based on its cosine similarity with other listings. We use Faiss[15], a library for efficient similarity search, to retrieve the nearest neighbors for each listing in the style embedding space.

In Figure 3 we show some examples of the top 4 nearest neighbors of a target listing based on both its text-based and style-based cosine similarity. Compared to the text-based nearest neighbors, the style-based nearest neighbors capture more stylistic similarity as opposed to content-based similarity. For example, in the first row the target listing is a geometric bag. Subsequently, all style-based items are

bags with different geometric patterns and shapes. In contrast, the text-based nearest neighbors are other satchels and bags that have title keywords in common with the target listing but do not necessarily capture the geometric style aspects of the bag.

4.2.2 Online Experiment: Style-based Recommender System. In order to evaluate the effectiveness of stylistic similarity, we use the style embeddings in a production listing-to-listing recommendations module that is served to live traffic. For this module, recommendations are obtained through a two-pass ranking system: First, the relevant candidate listings are generated using nearest neighbor search (called *candidate set selection*); second, candidate listings are re-ranked using a more expensive model to produce the final ordered list of recommendations.

For the control of the experiment, candidate set selection is powered by an embedding trained on sequences of user actions during a session using a skip gram model [38]. This base embedding provides a 100-dimension vector per listing which incorporates some notion of item similarity. For the treatment of the experiment, we incorporate our style embedding by appending a normalized style embedding to the base embedding, resulting in a 150-dimensional vector. The second-pass re-ranker remains unchanged.

For this experiment, online test result were very promising. The key metric of interest here was the average order value, which is one of the most important business metric at Etsy. Based on CLT with null hypothesis, control is better than treatment, the p-value from one-tailed z-test is 0.096, which means, under 0.1 significance level, we can conclude that style-embedding drives more revenue in our recommender system.

5 STYLE ANALYSIS

A particularly useful and novel application of the style prediction task described in Section 4.1 is the ability to infer style labels for all 50 million active listings on Etsy. We use the style-labeled listings to perform the first ever large-scale analysis on how different aesthetic styles impact e-commerce shopping behavior. For each of the analysis studies done below, we describe how this can better inform sellers on how to tailor their products. It also informs our recommender and search systems on how to better use predicted style to match user preferences in the future.

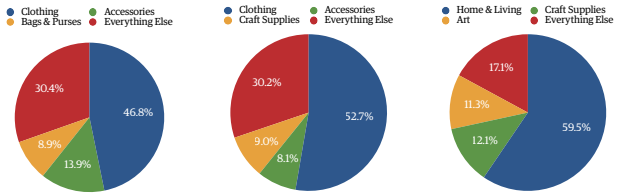
(a) Celestial (b) Floral (c) Geometric (d) Tropical

Figure 4: Four different styles from the same taxonomy

5.1 Style, Taxonomy, and Purchaseability

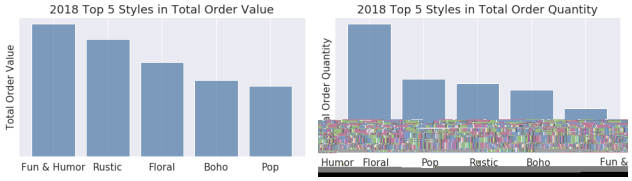
To begin, we first study some basic relationships between the style, taxonomy, and purchaseability of a listing. At Etsy, all listings are categorized into 15 top-level taxonomies, such as Jewelry, Clothing, Craft Supplies, etc. We first show that style and taxonomy are, in fact, highly dependent. We confirm this with a Chi-square test on the corresponding two-way contingency table, formulating the null hypothesis as style and taxonomy are independent. The resulting χ^2 statistic equals to 4291379 with degree of freedom equals 588. The corresponding p-value is 0.000, which confirms the dependency. We further investigate this dependency by analyzing taxonomy distributions for each style. Here, we measure the divergence between the taxonomy distribution for each specific style and the default taxonomy distribution over all listings using KL divergence. We show the 3 styles that deviate the most from the default taxonomy distribution (e.g. have the highest KL scores) in Figure 5 and see that a common characteristic is that all styles are heavily dominated by a single taxonomy class with proportion around 50%. This may inform sellers that it may make sense to focus on small subset of styles that are most dependent on the taxonomy that their products fall under.

We also show in Figure 6 the top 5 styles of 2018 by both total order value (revenue) and total quantity (purchase count). We see that items from *fun and humor* and *rustic* styles are particularly successful from a purchaseability standpoint.



(a) Sporty (b) 90s (c) Farmhouse

Figure 5: The taxonomy distribution for 3 styles that have the largest deviation from the taxonomy distribution of the general population of items



(a) Top 5 Styles in TOV (b) Top 5 styles in TQ

Figure 6: Top 5 styles in 2018 total order value and total order quantity

5.2 How Strength of Style Impacts Behavior

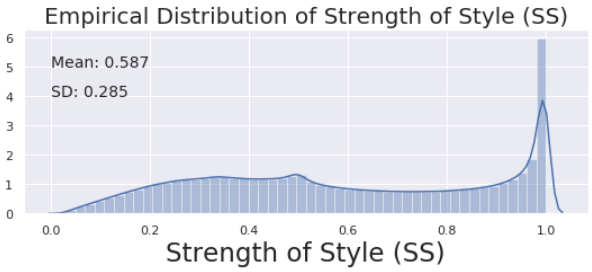
In this analysis, we describe a measure for *how strongly* a single style is exhibited in a particular listing and refer to this measure as *strength of style (SS)*. Given the 43-dimension predicted probability vector over styles, $\boldsymbol{p} = [p_1, p_2, \ldots, p_{43}]$, we hypothesize that this measure of strength of style (SS) determined by the statistical dispersion will be a strong signal for predicting user behavior. For an extreme case, a listing with one p_i equal to 1 and all others 0 is regarded as having the strongest strength of style while a listing with a uniform distribution over styles is the weakest one, meaning low dispersion transfers to strong strength of style and vice versa. There are several dispersion measures in statistics, like entropy and Gini

impurity (I_G) and we chose the latter to construct SS metric since compared with entropy, I_G is bounded, which makes comparison and understanding easier. We define

$$SS = 1 - I_G(\boldsymbol{p}) = 1 - \sum_{i=1}^{43} p_i(1-p_i) \in (0, 1]$$

Clearly, higher value of SS means stronger strength of style.

The empirical distribution of SS given more than 14-million active listings from Etsy is presented in Figure 7a. The peak around SS equals to 1 indicates that Etsy has relatively large proportion of listings with strong strength of style. The rest part of SS is roughly uniformly distributed and the drop around 0 to 0.2 shows that we have fewer listings with low strength of style listings.



(a) Empirical distribution of SS on Etsy active listings

(b) SS = 1.000

(c) SS = 0.492

(d) SS = 0.030

Figure 7: SS distribution and 3 listing examples with high, medium and low SS

We showed 3 listings as examples for high, medium and low strength of style in Figure 7. We can see that the wall art-decor in Figure 7b, which has the highest SS, has sailboat, anchor and whale; all of which are directly related to a *nautical* style. The avocado plush note holder in Figure 7c is very cute and funny, and holds a romantic note "you are my other half". Based on our prediction results, it has high probabilities on both *romantic* and *fun and humor*. Due to the mixture of the styles, it has SS equals to 0.49, which represents medium level strength of style. The black vest in Figure 7d shows no particular style and SS is a low value of 0.03. In the sections below, we examine the impact of a listing's strength of style on favoriting and purchasing behavior.

5.2.1 Strength of Style relating to Favoriting Behavior. Compared with clicks and impressions, favoriting an item is an interaction with higher intent since it directly discloses both users' preference and potential purchase motive. We try to understand whether users are more likely to favorite the listings with stronger strength of style. Rather than a pure prediction problem with strength of style as a feature and favorite count as the response, we focused on explaining the relationship between them statistically. Besides the strength of style, style and taxonomy indicators should be considered simultaneously since both of them may have heterogeneous impacts on user's favorite behavior. In addition, compared with marginally understanding the role of strength of style, the conditional effect by controlling both taxonomy and style is more convincing.

We sampled nearly 13-million listing favorite records that cover more than 2 million unique favorited listings from Etsy for analysis. As described above, the response of interest is the favorite count, C_{Fav}, with three kinds of features: strength of style (SS), style and taxonomy. In the regression equation 1 below, $\mathbf{T}$ indicates taxonomy set with cardinality 14, $\mathbf{S}$ means style set with cardinality 42 and $\mathbb{1}$ is the corresponding one-hot encoded dummy variable.

$$C_{\text{Fav}} = \beta_0 + \beta_{SS} \cdot SS + \sum_{i \in \mathbf{T}} \beta_i \cdot \mathbb{1}_i + \sum_{j \in \mathbf{S}} \beta_j \cdot \mathbb{1}_j + \varepsilon \quad (1)$$

Surprisingly, based on OLS estimation results, none of the taxonomy indicators has a statistically significant coefficient but 35 out of 42 style indicators are significant under 0.05 significance level. Moreover, SS is not only positive but also significant and F-test result validates the whole regression model. Detailed estimation and testing results are shown in Table 5.

The model tells us that in terms of favorite count, taxonomy does not have strong explanatory power with respects to favoriting behavior but style does. More importantly, the positive and significant coefficient of SS explains that users prefer to favorite listings with strong strength of style. A scatter plot for SS and C_{Fav} is shown in figure 8a, where the overall increasing trend between SS and favorite count is clearly showed.

Response	β_{SS}		F-test	
	coefficient	p-value	D.O.F	p-value
C_{Fav}	79.185	0.000	(57, 2247824)	0.000
C_{Pur}	4.039	0.000	(58, 7356954)	0.000

Table 5: Favorite count and purchase count regression models estimation and testing results

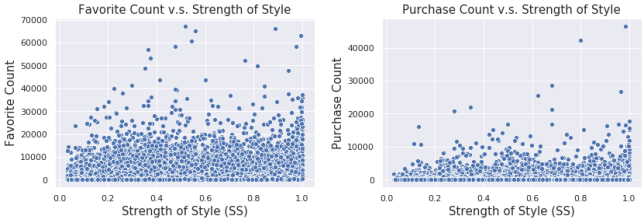
5.2.2 Strength of Style relating to Purchase Behavior. Compared with favoriting, purchasing is an interaction with even higher intent as it is expressing the user's final shopping decision. We perform an analysis similar to the prior section 5.2.1 for purchases, hoping to understand whether users are more interested in buying listings with high Strength of Style by taking purchase counts C_{Pur} as a response. The data we used for analysis here is a random sample of 2018 whole year transaction data which contains nearly 10-million transaction records that covers more than 7-million unique purchased listings. A slightly different part is that we also consider listing price, P, as a feature in regression model since favoriting comes with no cost but purchasing does. The regression model is shown below in equation 2.

$$C_{\text{Pur}} = \beta_0 + \beta_{SS} \cdot SS + \beta_P \cdot P + \sum_{i \in \mathbf{T}} \beta_i \cdot \mathbb{1}_i + \sum_{j \in \mathbf{S}} \beta_j \cdot \mathbb{1}_j + \varepsilon \quad (2)$$

Very different from the favorite model testing results, 12 out of 14 taxonomy indicators and 33 out of 42 style ones are significant under 0.05 significance level. Price is not only negative but also significant, which makes sense because of price advantage. Moreover, SS is not only positive but also significant as well and F-test result validates

the whole regression model. Detailed estimation and testing results are shown in Table 5.

The model tells us that in terms of purchase count, both taxonomy and style have strong explanatory power to purchase behavior. More importantly, the positive and significant coefficient of SS explains that users prefer to purchase listings with strong Strength of Style. A scatter plot for SS and C_{Pur} is shown in figure 8b, where the overall increasing trend between SS and purchase count is clearly showed. These two regression models inform Etsy sellers that they should create listings that have a single, strong style as opposed to no style or a blend of different styles.



(a) Favorite (b) Purchase

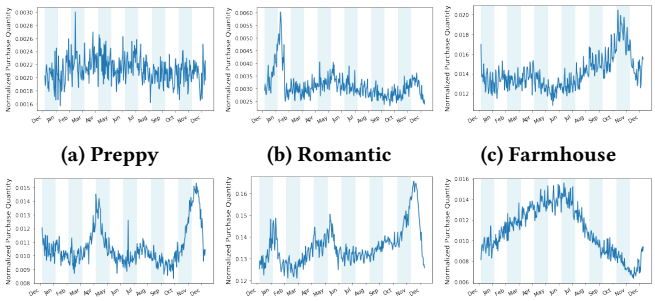
Figure 8: Scatter plots for both favorite count and purchase count versus SS, indicating an upward trend between SS and favoriting/purchase behavior

5.3 Seasonality Behind Shopping on Style

Studying cyclical patterns behind users' shopping behavior is a long-time researched topic in e-commerce as seasonality heavily impacts users' decision making. Therefore, incorporating seasonal signals into both recommender system and search ranking models can provide a better user experience by matching users' temporal shopping preference. With curiosity, we studied the purchase seasonality in terms of style and tried to understand the important insights behind it.

The dataset we used for seasonality study contains nearly 80-million transactions, which is a random sample from Etsy's transaction records in 2018. The quantity under interest here is normalized daily purchase quantity (NDPQ) per style, which is a ratio equalivent to daily purchase quantity per style divided by daily purchase quantity across all styles. For seasonality research, the normalization method we used is very necessary for Etsy data as it counteracts the overall trend for all styles, like shopping for holiday gifts that Etsy is very famous for.

For time series data, stationarity is an important property as it describes both fixed mean and fixed variance free of time index. In our study, stationary NDPQ represents stability of purchasing behavior, which implies there exist no specific shopping trend of that style across the whole year. Therefore, more attention should be paid to the styles with non-stationary NDPQs and to understanding the reason behind the non-stationarity. We applied augmented Dickey-Fuller (ADF) test for 43 styles' NDPQs and based on significance level equals 0.05, 16 styles, like preppy, Hollywood regency, etc, are stationary while 27 styles are not. One stationary style example (*preppy*) and 5 non-stationary ones are shown in Figure 9. Here, we see that the trends derived from the style prediction give fairly intuitive results and align with merchandising experts.



(a) Preppy (b) Romantic (c) Farmhouse

(d) Inspirational (e) Fun and Humor (f) Tropical

Figure 9: 6 styles NDPQs across 2018 (with light blue background for Jan, Mar, May, Jul, Sep and Nov)

For example, *romantic* has a clear peak around February for Valentine's Day, while *tropical* peaks in the warmer months. The style *farmhouse* is a warm and cozy style, often associated with colder months and the season of fall, thus peaking in November. Inspirational is a style that is popular during graduation (May/June), as well as the time leading up to a new Year (December) as gifts with motivational quotes are exchanged. Similarly, the style *fun and humor* is often applied to t-shirts and cards, and are common for Father's Day gifts as well as stocking stuffers during the holidays. Altogether, this analysis can help inform sellers as to how to tailor their products to certain styles during different times of the year.

5.4 Search Behavior on Style

We further analyzed the search log data to specifically show the potential use cases of style in order to improve search ranking models. We have the hypothesis that users are more interested in interacting with the listings belonging to a few preferred styles in search results. In order to verify our assumption, we measure the difference between averaged interacted listings style distribution ($\bar{\boldsymbol{p}}_{Inter}$) and averaged showed or impression level listings style distribution ($\bar{\boldsymbol{p}}_{Imp}$) for some queries. More specifically, under a specific query session, we collected all clicked, favorited, added-to-cart and purchased listings as interacted listings and averaged their style distributions to represent the user's preferred style distribution for this query, which is $\bar{\boldsymbol{p}}_{Inter}$ in equation 3. For the counterpart, we also measure the style distribution of shown search results for the same specific query, $\bar{\boldsymbol{p}}_{Imp}$ in equation 3. In order to capture how much the style distribution preference of the user deviates from that of the shown listings for each query, we quantify how different the two distributions are. We chose Jensen-Shannon divergence (JSD) as the metric since it is both symmetric and bounded with range from 0 to 1. The equation for JSD is showed in equation 4.

$$\bar{\boldsymbol{p}}_{Inter} = \frac{\sum_{i \in Inter} \boldsymbol{p}_i}{\#Inter}, \quad \bar{\boldsymbol{p}}_{Imp} = \frac{\sum_{j \in Imp} \boldsymbol{p}_j}{\#Imp} \tag{3}$$

$$JSD\left(\bar{\boldsymbol{p}}_{Inter} \,||\, \bar{\boldsymbol{p}}_{Imp}\right) = \frac{KL\left(\bar{\boldsymbol{p}}_{Inter} \,||\, \bar{\boldsymbol{p}}_{Avg}\right) + KL\left(\bar{\boldsymbol{p}}_{Imp} \,||\, \bar{\boldsymbol{p}}_{Avg}\right)}{2} \tag{4}$$

where $\bar{\boldsymbol{p}}_{Avg} = (\bar{\boldsymbol{p}}_{Inter} + \bar{\boldsymbol{p}}_{Imp})/2$ and $KL(\cdot \,||\, \cdot)$ is Kullback-Leibler divergence.

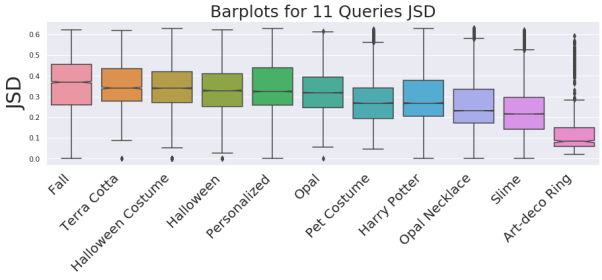


Figure 10: Barplots about JSD distributions for 10 top queries from Etsy with "art-deco ring" as baseline

We chose the top 10 queries from Etsy, as well as the query "art-deco ring", a popular query that explicitly contains style words as baseline for analysis. In each query session, there is one JSD score which measures the difference between user's preferred style distribution and the style distribution of all showed listings. Hence, we can aggregate by the same query to get a distribution of JSD for the 11 queries we considered, which is shown as 11 boxplots in Figure 10. From Figure 10, all 10 top queries JSD medians are obviously larger than the baseline, meaning that there exists a relatively large degree of mis-matched style distributions relative to user style preferences for our current top 10 queries. A potential solution is to better incorporate style into search ranking models or provide guided search options based on style.

6 DISCUSSION AND CONCLUSION

Etsy is a global community based marketplace where people come together to buy and sell unique items. The influence of aesthetic style on listings can not be understated yet until now we did not have the necessary tools and framework to study it. In this paper we proposed a style predictive model to better understand the influence of style on our listing inventory. We leverage 43 named styles given by merchandising experts in order to setup a supervised, style-aware deep neural network model for predicting a listings style. We detailed how we generate style-aware embeddings using the same model and showed one of its application in a downstream recommender task. With the help of these embeddings we performed the first ever large scale analysis to understand how aesthetic styles impact e-commerce purchase behavior. This gave us insights into how our recommender and search systems can better leverage predicted styles of a listing to match user preferences.

REFERENCES

[1] R. S. Arora and A. Elgammal. 2012. Towards automated classification of fine-art painting style: A comparative study. In *ICPR*.
[2] E. Bruni, N.-K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *JAIR* (2014).
[3] J. Dai, K. He, and J. Sun. 2016. Instance-aware semantic segmentation via multi-task network cascades. *CVPR* (2016).
[4] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. *CVPR* (2005).
[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part based models. *TPAMI* (2010).
[6] R. Girshick. 2015. Fast R-CNN. *ICCV* (2015).
[7] R. Girshick, andT. Darrell J. Donahue, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* (2014).
[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *CVPR* (2016).
[9] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: A Visually, Socially, and Temporally-aware Model for Artistic Recommendation. *Recsys* (2016).
[10] Felix Hill and Anna Korhonen. 2014. Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean. In *EMNLP*.
[11] W. Hsiao and K. Grauman. 2017. Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding from Fashion Images. *ICCV* (2017).
[12] Wei-Lin Hsiao and Kristen Grauman. 2018. Creating Capsule Wardrobes from Fashion Images. *CVPR* (2018).
[13] Diane J. Hu, Rob Hall, and Josh Attenberg. 2014. Style in the Long Tail: Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce. In *KDD*.
[14] Yushi Jing, David C. Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Visual Search at Pinterest. *KDD* (2015).
[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).
[16] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. *ICDM* (2017).
[17] Sergey Karayev, Aaron Hertzmann, Holger Winnemoeller, Aseem Agarwala, and Trevor Darrell. 2013. Recognizing Image Style. *CoRR* abs/1311.3715 (2013).
[18] Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. *EMNLP* (2014).
[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Multimodal Neural Language Models. *ICML* (2014).
[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). arXiv:1411.2539
[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (May 2017).
[22] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. 2015. Deep Multi-Patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *ICCV*.
[23] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR*.
[24] N. Murray, L. Marchesotti, and F. Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*.
[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.
[26] Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alexander J. Smola, and S. V. N. Vishwanathan. 2009. Hash Kernels for Structured Data. *Journal of Machine Learning Research* (2009).
[27] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*. 869–877.
[28] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge J. Belongie. 2015. Learning Visual Clothing Style with Heterogeneous Dyadic Co-occurrences. *ICCV* (2015).
[29] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. *CVPR* (2001).
[30] Kilian Q. Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford, and Alexander J. Smola. 2009. Feature Hashing for Large Scale Multitask Learning. *CoRR* (2009).
[31] J. Wu, Y. Yu, C. Huang, and K. Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. *CVPR* (2015).
[32] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. 2014. Architectural Style Classification Using Multinomial Latent Logistic Regression. *ECCV 2014* (2014), 600–615.
[33] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. 2012. Parsing clothing in fashion photographs. In *CVPR*. 3570–3577.
[34] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, M. Hadi Kiapour, and Robinson Piramuthu. 2017. Visual Search at eBay.. In *KDD*.
[35] Nick Zangwill. 2019. Aesthetic Judgement. In *The Stanford Encyclopedia of Philosophy*.
[36] Andrew Zhai, Dmitry Kislyuk, Yushi Jing, Michael Feng, Eric Tzeng, Jeff Donahue, Yue Li Du, and Trevor Darrell. 2017. Visual Discovery at Pinterest. *WWW* (2017).
[37] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. 2016. Is faster r-cnn doing well for pedestrian detection?. In *European conference on computer vision*. Springer, 443–457.
[38] Xiaoting Zhao, Raphael Louca, Diane Hu, and Liangjie Hong. 2018. Learning Item-Interaction Embeddings for User Recommendations. *CoRR* (2018).