Automatically Discovering Visual Product Characteristics

Ankit Sisodia* Alex Burnap* Vineet Kumar*

Abstract

Marketing models typically focus on how structured product characteristics impact consumer preferences. However, visual characteristics of products present in unstructured image data play an important role in impacting preferences for many categories. We seek to automatically discover and quantify visual characteristics (attributes) from image data using a disentanglement-based approach. While the deep learning literature has shown that supervision is required to obtain unique disentangled representations, ground truth visual characteristics are typically unknown. We develop a method that does not require such supervision, and instead uses readily available structured product characteristics as supervisory signals to enable disentanglement. Our method does not need prior knowledge of characteristics, yet we are able to discover semantically interpretable and statistically independent characteristics. Moreover, the method quantifies the levels of each discovered product characteristic, necessary for managerial tasks such as demand modeling and conjoint analysis. We apply this method to automatically discover visual product characteristics of watches, and discover 6 semantically interpretable visual characteristics providing a disentangled representation. Our results find the supervisory signal 'brand' best promotes disentanglement relative to an unsupervised approach. We lastly demonstrate how consumers preferences may be assessed over these discovered visual characteristics using a choice-based conjoint analysis.

Keywords: discovery of product characteristics, deep learning, disentanglement

*Yale School of Management

INTRODUCTION

Product characteristics form the basis of consumer choice and willingness to pay (WTP) for almost all products and services. Hedonic demand theory initially posited that market demand for a product is the aggregation of demand over its underlying product characteristics (Lancaster 1966). Unsurprisingly, characteristics form the foundation for quantitative models used in marketing and economics, for tasks ranging from quantifying consumer preferences (Guadagni and Little 1983), pricing products and services (Mahajan, Green, and Goldberg 1982) as well as modeling competitive markets (Berry, Levinsohn, and Pakes 1995). For these tasks, the relevant set of characteristics for the model must be defined in advance; for example, for a car this may include horsepower, fuel efficiency, and towing capacity. While this task is straightforward for structured product characteristics, it is not obvious for visual characteristics, which are often fully captured only with unstructured data (e.g. text, images, audio, video).

Unstructured data has received growing interest in marketing, due to both the increasing quantity and variety of such data, alongside the simultaneous development of machine learning methods capable of its analysis. However to date, to discover unstructured characteristics for use in marketing models, researchers have *pre-specified characteristics they are looking for* and labeled them using human judgment. For example, recent work has used Airbnb images labeled by humans to train a machine learning model to predict image quality, and then demonstrate that consumer preferences and demand are impacted by such quality (Zhang et al. 2021a). Similarly, deep learning methods have been used to extract researcher-defined characteristics from images (Zhang and Luo 2018; Troncoso and Luo 2020; Zhang et al. 2021b; Anand and Kadiyali 2020; Li, Shi, and Wang 2019) or directly used unstructured data to predict an outcome such as return rate (Dzyabura et al. 2019; Yang, Zhang, and Zhang 2021) or brand characteristics (Liu, Dzyabura, and Mizik 2020).

In contrast, our research has a different goal. In many product categories, such as fashion goods or automobiles, visual characteristics are critical drivers of consumer preference and purchase, yet enunciating *why* a product looks appealing is challenging for researchers and consumers alike (Berlyne 1973)—our method seeks to identify specifically such characteristics.

Research Goal: Our research aims to develop a method to *automatically discover and quantify* multiple (visual) characteristics that are independent and obtained directly from unstructured product image data, with aid from other structured product data.

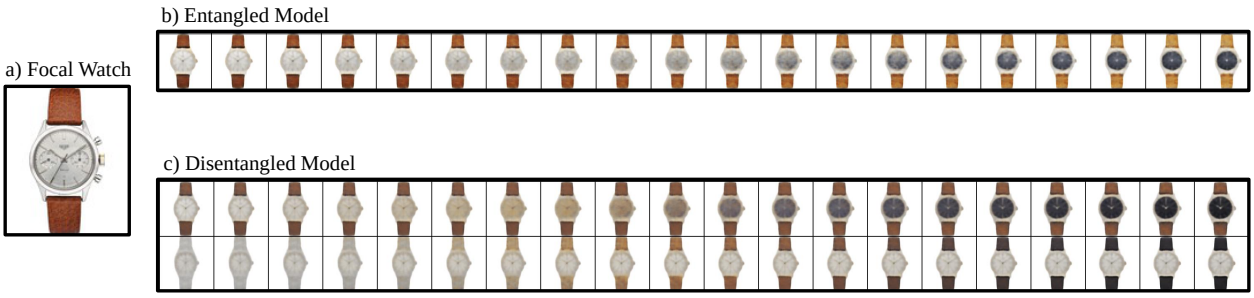
Quantification of these visual characteristics is essential to many marketing applications, e.g. relative product positioning, identifying the set of products that are most similar to a focal product on a specific visual dimension (or some set of visual dimensions). Another important application using such visual characteristics involves individual preferences. We demonstrate using a visual conjoint analysis to enumerate consumer preferences over visual characteristics of products, by independently varying each characteristic, while holding the other characteristics constant.

Current practice for discovery of visual characteristics relies primarily on human judgment. The (human) modeler first defines visual characteristics using a combination of intuition and expertise, based on available data and qualitative methods based on consumer input like Voice of the Customer (Griffin and Hauser 1993). In product categories such as automobiles and furniture, firms routinely spend upwards of $100,000 on specialized focus groups that aim to both identify important visual characteristics and their impact on consumer preferences (Burnap, Hauser, and Timoshenko 2019). Likewise with digital products, best practices include conducting design ethnography and usability testing when launching new product features or changes to human-computer interface (Norman 2004).

However, there are several reasons why it would be helpful to have an automated (as opposed to human) approach to discovering and quantifying visual characteristics. First, while humans can discover what the characteristics are, it is a significant and time-consuming process for humans to enumerate or quantify. [1]. Even with humans we would need to create a quantification scale for each specified characteristic, which is also not trivial. Second, given the large scope of the manual task, if multiple individuals are used, then we would need to have a principled approach to aggregating individual judgments, especially when the differ significantly from one another. Third, even if

[1] With K characteristics for each of N products we would need a total of (N x K) human evaluations, which is not quite scalable especially across categories. In our application with watches we have $N = 6,187$ and $K = 6$, resulting in $37,122$ manual human judgments regarding the levels of characteristics.

Figure 1: Example of Entanglement and Disentanglement in Visual Characteristics



Notes: **a**: Focal watch **b**: Entangled model outputs a characteristic that changes both the dial color and strap color as its the level is changed. **c**: Disentangled model outputs two independent characteristics for dial color and strap color.

humans are able to identify and quantify the levels of visual characteristics for existing watches, this manual approach would *not* be able to generate counterfactual watches, which are required for visual conjoint analysis. might be useful in visualization (e.g. if an existing watch model were to have a different strap color). Our VAE-based generative modeling approach can obtain such counterfactual images as part of the learning process. Finally, although humans might be able to identify some set of visual characteristics, it is possible that an algorithmic approach could obtain a different set, with some potentially new characteristics.

Our methodology is built on *disentangled* representation learning, an emerging area of deep learning that aims at identifying independent yet semantically-meaningful factors of variation within data (Bengio, Courville, and Vincent 2013). Entanglement (in contrast) implies that a change in the level across one discovered characteristic impacts *multiple* semantically-interpretable characteristics. In contrast, with a disentangled representation, a change in one characteristic (e.g. size or color) would result in a change to only one characteristic. In our application, we disentangle (or separate out) visual characteristics that are present in unstructured image data of watches. Figure 1 illustrates the difference between disentangled and entangled representations.

Advantages of Our Approach: Our approach provides practical advantages to both academics and practitioners. First, our disentanglement-based approach is designed to work with unstructured *big image data* that would be practically obtainable in real managerial contexts (e.g., product images). It does not require labeled data on visual characteristics, and is designed to leverage typically

available structured characteristics. Second, the researcher does not need to define the (unknown) visual characteristics in advance, and does not even need to specify the number of such characteristics that must be discovered. Our method is also flexible with regard to image quality, and works with low resolution images (like 128x128 pixels). Third, our method discovers not just the product characteristics, but quantifies their level for each product, and is capable of generating counterfactual products that vary only along a specified characteristic. This is critical since real data is unlikely to feature variation along only one characteristic, since products differ from one another along many (visual) dimensions. This ability to independently vary visual characteristics makes the method suitable for use with a variety of downstream tasks like demand estimation or conjoint analysis, among others. Fourth, the method typically discovers semantically-interpretable characteristics. Although no method can guarantee that automatically discovered characteristics are interpretable by humans, those discovered by our method are typically semantically-interpretable, a claim we validate with human subjects. Finally, our approach can be applied in a scalable manner across product categories as long as image data from the relevant category is available.

Methodology: Our primary goal is to infer disentangled representations from image data, rather than generating realistic and high quality images. We therefore develop our disentanglement approach using a VAE, which includes an encoder neural net and decoder neural net, both of which are parametrized by highly nonlinear deep neural networks. The encoder neural net takes high-dimensional unstructured data (e.g., images) as input and outputs a latent low-dimensional vector for each discovered characteristic, whereas the decoder neural net takes as input the low-dimensional vector and attempts to reconstruct the original data as output. The idea of representation learning is that the "true" dimension of images in the data belonging to a category is much lower than the dimensionality possible in raw data. [2]

Deep autoencoders have found recent application in business and marketing; for example, Dew, Ansari, and Toubia (2021) used VAEs to study logo design and Malik, Singh, and Srinivasan (2019)

[2]For instance, images are high-dimensional data since even a modest-sized image of 1,000×1,000 pixels exists in a 1,000,000-dimensional space. But suppose we know that each of the images represents a black circle on a white background; each circle can then be completely represented by the location of its center (x, y) and its radius r, thus essentially making the data 3-dimensional.

used conditional adversarial autoencoders to study the impact of beauty premium in human faces on career outcomes, and (Cheng, Lee, and Tambe 2022). However, these studies did not focus on automatically discovering disentangled representations that are both statistically independent and semantically interpretable.

To promote disentanglement, we penalize the *total correlation* of the discovered characteristics such that they are statistically independent (Chen et al. 2018; Hoffman and Johnson 2016; Kim and Mnih 2018). With successful disentanglement, changes along any one of the discovered characteristics leads the reconstructed image to change visually only along that one characteristic.

A critical challenge in *any* disentanglement approach is that without a supervisory signal, the discovered set of characteristics is not unique. Instead, there may be (infinitely) many representations that are equally probable (Locatello et al. 2019).[3] However, since our goal is to discover precisely these visual product characteristics, it is not consistent with having such ground truth information available. Even partial labeling, which has been shown to be effective (Locatello et al. 2020) requires human intervention, which also contrasts with our goal of automatically disentangling characteristics from data typically found in marketing applications.

Our method instead leverages supervision on readily available, complete, and precise structured data that are collected in marketing data sets. Specifically, the encoder neural net is additionally connected to a supervised neural net, thereby connecting the discovered visual characteristics to structured characteristics (e.g., brand identity). This enables us to obtain supervision without additional labeling, and thus overcoming the theoretical limitation of unsupervised algorithms.

Application and Results: We apply the proposed approach in the visual domain, with the goal of automatically discovering visual characteristics of watches; this application is appropriate for several reasons. First, watches represent a product category where visual and design aspects captured in the images are likely to play an important role in consumer valuation and choice behavior (Kotler and Rath 1984). Second, as typical with marketing data, we have a set of structured data

[3] Specifically, any bijective function of the characteristics of a discovered representation would be representationally equivalent (Khemakhem et al. 2020).

appropriately matched up with the images. Finally, for our validation exercise, human respondents are typically familiar with this product category, providing a useful benchmark to understand how our method's results align with human interpretation.

Our method automatically discovers six visual characteristics of the watches. These discovered characteristics correspond to 'size of the dial', 'dial color', 'strap color', 'dial shape', 'size of the knob', and 'rim color'. Figure 2b gives an example of these discovered characteristics for one randomly selected watch. This example allows the reader to visually evaluate disentanglement performance, defined as both independence across characteristics (i.e., how each characteristic changes independently of each other) and semantic interpretability (i.e., how well can humans understand) (Higgins et al. 2017; Burgess et al. 2017; Higgins et al. 2021). For example, as the characteristic level for 'dial color' increases, the 'dial color' increases from light to dark but other visual characteristics remain the same.

We also validate the discovered visual characteristics by using a survey with 600 human respondents to independently identify whether these characteristics have semantic meaning. For a visual characteristic to have semantic meaning, a strong majority of respondents need to reliably identify the characteristic, and the agreement among respondents needs to be high. We use a panel of selected human respondents, and find that a majority of respondents agree with one another that the discovered visual characteristic has identical semantic meaning, with the greatest agreement was for strap color (94%) and lowest agreement was for knob size (42%). For the latter, the visual quality of the images might have contributed to the low degree of agreement.

To validate the quantification of characteristics, we also examine whether the quantified the level of the characteristic has semantic meaning to humans. One way to test this aspect is to show respondents 2 pairs of watches (randomly drawn from the disentangled representations), where each pair varies on a specific visual characteristic (e.g. dial color). We then ask respondents which pair is visually "more similar" and measure the divergence between human responses and the algorithm's quantification of characteristics. The idea is that if humans and algorithms view the semantic meaning of the quantification of characteristics in a similar manner. We find that human

respondents and the algorithm match 85% of the time along this similarity metric, reflecting that the algorithm's quantification of the levels of visual characteristics is semantically meaningful to humans.

We next study the issue of supervised vs unsupervised disentanglement. We evaluate how well various supervised and unsupervised modeling specifications affect disentanglement performance of the six discovered characteristics. Our supervised disentanglement model specifications include supervisory signals related to product (e.g., brand, circa, material), place (e.g., auction location), and price (e.g., willingness to pay), while our unsupervised disentanglement method uses no supervisory signals at all. For model selection of supervised disentanglement models, we choose the hyperparameter settings that lead to the lowest supervised loss on a validation dataset (Locatello et al. 2020). For the unsupervised approach, we use unsupervised disentanglement ranking (UDR), a metric for evaluating disentanglement performance when the ground-truth product characteristics are unknown as typical in real-world datasets like ours (Duan et al. 2020). In our case, we use UDR for unsupervised disentanglement model selection (on a validation set), as well as for model evaluation (on a test set) for both the supervised and unsupervised disentanglement models.

Our results are in parts both expected and unexpected. In our comparison of supervisory signals for disentanglement, we find that 'brand' helps but 'price' hurts disentanglement performance relative to an unsupervised approach. This is surprising as 'price' is one of, if not the most significant economic primitives affecting product design. Since 'price' is often assumed as function of product characteristics (and consumer preferences over those characteristics), we expected using it as a supervisory signal would improve disentanglement—instead, it resulted in worse disentanglement than no supervision at all. This gives evidence that visual characteristics associated with a given watch brand captures more visual variation than across brands when varying price.

Lastly, to examine whether the discovered visual characteristics impact consumer preferences, we conducted a choice-based conjoint survey with N=384 respondents. Participants in the conjoint survey were presented with pairs of images of different watches obtained from our disentangled representations. The watches varied on each of the visual characteristics that were identified by our

algorithm, and participants were shown 2 distinct designs and then asked to choose which of the visual designs they preferred or choose a None option. For the purposes of this study, participants were asked to only focus on visual characteristics, and not prices or other structured characteristics (which were not provided to them). We find that consumers indeed have preferences over visual characteristics discovered by our proposed approach. Consumers tend to prefer visual contrast between the color of a watch's dial and the rim that surrounds it, while tending to prefer visual similarity between the color of the dial and the watch strap.

Overall, we find that our approach discovers and quantifies visual characteristics from unstructured data in a manner that is meaningful to humans (semantically interpretable), and that these individual consumers have clear preferences over these characteristics.

Contributions: We provide an automated approach to discover product characteristics from unstructured data typically found in marketing. Such characteristics could be used in competitive analysis, product positioning and pricing decisions by firms. In addition to the direct level of such discovery, and understanding how they impact consumers, they can have an indirect impact on structured characteristics. When inferring economic valuation of structured characteristics, we could have omitted variables (from unstructured data), leading to biased inference due to unobserved correlation between unstructured (visual) and structured characteristics.

From a methodological perspective, our paper contributes on the issue of supervised versus unsupervised disentanglement in representation learning. First, we show that characteristics can be discovered without access to ground truth from visual or unstructured data. We show how structured data typically available in marketing applications may be used as supervisory signals for obtaining better disentanglement. This aspect is useful as the machine learning literature typically assumes the presence of ground truth supervisory signals, which are seldom available in typical business applications. Second, and equally important, we demonstrate that just the idea of using any supervised signal might not work and may indeed backfire. The machine learning literature has focused on using a supervised approach due to known theoretical challenges with recovering a unique latent representation via unsupervised disentanglement. However, our research points out

that, in practice, supervised learning may not be a panacea and that the choice of supervisory signal is critically important. In fact, many supervisory signals actually lead to *worse* disentanglement than using no supervision at all (i.e., unsupervised disentanglement).

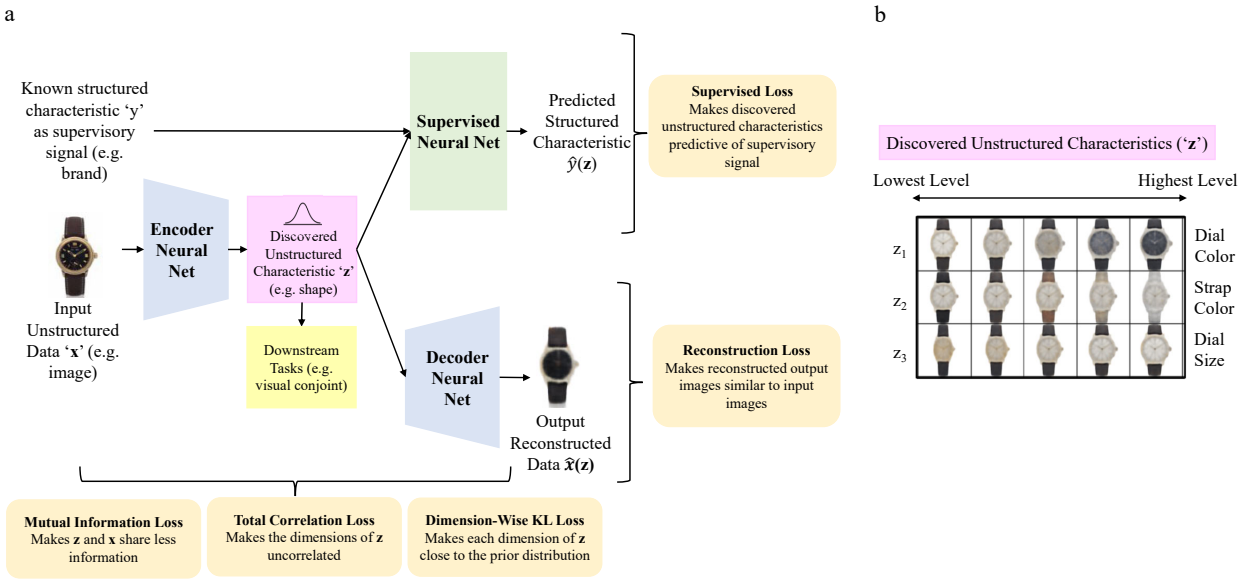
Limitations: Our approach has specific limitations worth noting and addressing in future research. First, it requires structured data to be matched to corresponding unstructured data. In our application the watch images are matched to corresponding structured characteristics, but other applications may not have such structured data that correspond to image data. Second, although the algorithm does not require human intervention, the data is typically preprocessed to ensure centering, similar size, background color, and orientation. Third, no algorithm can *guarantee* semantic interpretability for newly discovered features, because that is a uniquely human ability (Locatello et al. 2019; Higgins et al. 2021). However, we validate that in practice we observe that our proposed method performs well in a realistic and practical setting.

METHODOLOGY

Our proposed approach builds on recent advances in disentangled representation learning, a stream of machine learning focused on learning lower-dimensional re-representations of high-dimensional data. Most disentanglement methods are built on deep generative models, most notable variational autoencoders (VAE) and generative adversarial networks (GAN), which we describe more comprehensively in Appendix A. Our model is a VAE (Kingma and Welling 2014) extended for supervised learning and disentanglement.

Our method is illustrated in Figure 2. We *encode* unstructured data (e.g. text or images) to discover unstructured (visual) characteristics that are independent, low-dimensional and semantically interpretable (e.g., shape) and then *decode* the discovered unstructured (visual) characteristics to reconstructed unstructured data as well as *predict* a supervised signal (e.g., typical marketing structured data such as brand) from the discovered unstructured (visual) characteristics. The model minimizes the weighted sum of five different type of losses — reconstruction loss, mutual information loss, total correlation loss, dimension-wise Kullbeck-Leibler (KL) loss and supervised loss.

Figure 2: Schematic Illustration of Proposed Approach



Notes: **a**: The encoder neural net maps an input image into low-dimensional unstructured (visual) data characteristics, which are then used by both the decoder neural net to reconstruct the original image and by the supervised neural net to predict a supervisory signal corresponding to the image. **b**: Varying the levels of discovered characteristics to visualise the semantic meaning encoded by single disentangled visual characteristic of a trained model. In each row the level of a single visual characteristic is varied while the other characteristics are fixed. The resulting effect on the reconstruction is visualised. We show three discovered visual characteristics here for illustration purposes.

Model: Supervised Variational Autoencoder with Disentanglement Losses

We first describe a VAE and subsequently describe how it is extended with disentanglement constraints and supervision using structured data. We denote the observed dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ where the i-th observation is a high-dimensional product image $\mathbf{x}_i$ and its corresponding supervised signal y_i. The VAE assumes a two-step data generating process. The first step samples the (visual) discovered characteristics denoted by $\mathbf{z}_i \in \mathbb{R}^J$, where J is the maximum number of characteristics to be discovered. In the second step, the product image $\mathbf{x}_i$ is reconstructed from the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z}) = f(\mathbf{x}; \mathbf{z}, \theta)$, where $f(\mathbf{x}; \mathbf{z}, \theta)$ is a multivariate Gaussian distribution whose probabilities are formed by nonlinear transformation of the characteristics, $\mathbf{z}$, using a neural network with parameters θ. Likewise, the signal y_i is predicted from the conditional distribution $p_w(y|\mathbf{z}) = f(y; \mathbf{z}, \mathbf{w})$, where $f(y; \mathbf{z}, \mathbf{w})$ is a function formed by non-linear transformation, with parameters $\mathbf{w}$, of unstructured (visual) characteristics $\mathbf{z}$.

We refer to $p_\theta(\mathbf{x}|\mathbf{z})$ as the decoder neural net, $q_\phi(\mathbf{z}|\mathbf{x})$ as the encoder neural net, and $p_\mathbf{w}(y|\mathbf{z})$ as the supervised neural net. As in variational Bayesian inference (Blei, Kucukelbir, and McAuliffe 2017) the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is intractable, so we follow the original VAE assumption that the true posterior can be approximated using a variational family of Gaussians with diagonal covariance $\log q_\phi(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and the s.d. of the approximate posterior (Kingma and Welling 2014). We simultaneously train the encoder neural net, the decoder neural net, and the supervised neural net, by minimizing a variational bound to the negative log-likelihood. In practice, this results in a loss minimization problem to find point estimates of the neural network parameters, $(\theta, \phi, \mathbf{w})$, while inferring a full distribution over the discovered characteristics, $\mathbf{z}_i \in \mathbb{R}^J$. The parameter space of the deep neural networks in our intended applications are often in the range of hundreds of thousands to hundreds of millions depending on architectural decisions (e.g., our architecture in Appendix B has 1,216,390 parameters).

The overall loss is composed of several loss terms corresponding to a VAE extended with supervision and disentanglement terms. We detail these losses starting with the loss of the original VAE in Equation (1), and refer readers to Kingma and Welling (2014) for its detailed derivation.

$$\underbrace{L(\theta, \phi, \mathbf{w}; \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction Loss}} + \underbrace{KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\text{Regularizer Term}}$$

To learn disentangled representations, a recent model denoted β-VAE (Higgins et al. 2017) extends Equation 1 by imposing a heavier penalty on the regularizer term using an adjustable hyperparameer $\beta > 1$.[4] Intuitively, β-VAE uses the hyperparameter β to sacrifice reconstruction

[4] Higgins et al. (2017) derive the β-VAE loss function as a constrained optimization problem. Specifically, the goal is to maximize the reconstruction accuracy subject to the inferred visual characteristics being matched to a prior isotropic unit Gaussian distribution. This can be seen in Equation 1 where ϵ specifies the strength of the applied constraint.

$$\max_{\theta,\phi} \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \text{ subject to } KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right] < \epsilon$$

We can re-write Equation 1 as a Lagrangian under the KKT conditions (Kuhn and Tucker 2014; Karush 1939), where the KKT multiplier β is a regularization coefficient. This explicit coefficient β is used as a hyperparameter (set by the researcher) to promote disentanglement, and results in the β-VAE formulation in Equation 1.

$$\mathcal{L}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta(KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right])$$

accuracy in order to learn more disentangled representations. We adopt this decomposition and further extend it by decomposing the regularizer term in Equation (1) into three terms (Chen et al. 2018; Hoffman and Johnson 2016; Kim and Mnih 2018). These three terms enable us to directly and separately control disentanglement constraints of the model as follows in Equation (1).

$$\underbrace{KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\substack{\text{Regularizer Term}\\ \text{of Total Loss}}} = \underbrace{I_q(\mathbf{z},\mathbf{x})}_{\substack{\text{Mutual Information}\\ \text{Loss}}} + \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}} + \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}}$$

We finally add a supervised loss term to enforce the discovered characteristics to help predict the supervised signal y in Equation (1). This enables us to study whether using typical structured data (e.g., 'brand') in a supervised approach helps improve disentanglement, and moreover, compare supervised disentanglement versus unsupervised disentanglement.

$$\begin{aligned}\underbrace{L(\theta,\phi,\mathbf{w});\mathbf{x},\mathbf{z})}_{\text{Total Loss}} = {} & \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction}\\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z},\mathbf{x})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} \\ & + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}} + \gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{y}(\mathbf{z}),y)}_{\substack{\text{Supervised}\\ \text{Loss}}}\end{aligned} \tag{1}$$

Our model's total loss is comprised of five loss terms weighted using scaling hyperparameters, $(\alpha,\beta,\gamma,\delta)$. Adjusting these hyperparameters critically affects disentanglement performance by adjusting the relative weight of each of the five loss terms, for which we detail the intuition below:[5]

Reconstruction Loss: Penalizing the reconstruction loss encourages the reconstructed output $\hat{x}(\mathbf{z})$ to be as close as possible to the input data x. This ensures that the discovered characteristics possess the necessary information to be able to reconstruct the product image.

[5]Note that adjusting these hyperparameters also leads to different models as special cases. In the original VAE, $\alpha=\beta=\gamma=1$ and $\delta=0$. In the β-VAE, $\alpha=\beta=\gamma>1$ and $\delta=0$, meaning that a heavier penalty is imposed on all three terms of the decomposed regulariser term in Equation 1. Finally, in β-TCVAE, $\alpha=\gamma=1$, $\beta>1$ and $\delta=0$ and thus there is a heavier penalty only on the total correlation loss term. In our proposed supervised approach, we impose $\alpha=\gamma=1$ and find levels of the hyperparameter set $\Omega=\{\beta,\delta\}$. We compare it with an unsupervised approach in which we impose $\alpha=\gamma=1$, $\delta=0$ and find the levels of the hyperparameter set $\Omega=\{\beta\}$.

Mutual Information Loss: $I_q(\mathbf{z}, \mathbf{x}) = \mathbf{E}_{q(x,z)} \log\left(\frac{q(x,z)}{q(x)q(z)}\right)$ is the mutual information between the discovered unstructured (visual) characteristic $\mathbf{z}$ and the product image $\mathbf{x}$. From an information-theoretic perspective (Achille and Soatto 2018), penalizing this term reduces the minimum amount of information about $\mathbf{x}$ stored in $\mathbf{z}$ that is sufficient to reconstruct the data by ensuring $\mathbf{z}$ does not store nuisance information. A low α would result in $\mathbf{z}$ storing nuisance information, whereas a high α results in loss of sufficient information needed for reconstruction. We set $\alpha = 1$ to encourage the visual characteristics to store the minimum amount of information about the raw data sufficient to reconstruct the raw data while not compromising on the reconstruction accuracy.

Total Correlation Loss: The total correlation, $KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]$, represents a measure of dependence of multiple random variables in information theory (Watanabe 1960). If the latent variables $\mathbf{z}$ are independent, then the KL divergence is zero. More generally, a high penalty for the total correlation term forces the model to find statistically independent visual characteristics. A high β results in a more disentangled representation but with potentially worse reconstruction quality. We follow the β-TCVAE approach (Chen et al. 2018) and find the level of the hyperparameter β in order to learn disentangled representations for both supervised and unsupervised approaches.

Dimension-Wise KL Loss: The dimension-wise KL loss term, $\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]$, penalizes the objective to push $q(z_j)$ to the prior $p(z_j)$ encouraging the probabilistic structure imposed by the parametric assumptions of the prior (e.g., Gaussian). This term promotes continuity in the latent space, which allows generation from a smooth and compact region of latent space. We set $\gamma = 1$ to encourage individual visual characteristics to not deviate much from the prior while also not overly compromising reconstruction accuracy.

Supervised Loss: Penalizing the supervised loss $P(\hat{y}(\mathbf{z}), y)$, where $\hat{y}(\mathbf{z}) \sim p_{\mathbf{w}}(y|\mathbf{z})$ prioritizes the discovered visual characteristics $\mathbf{z}$ to obtain high accuracy in predicting y. We find the level of the hyperparameter δ for the supervised disentanglement approach by model selection and set $\delta = 0$ for the unsupervised disentanglement approach. When the signal is discrete (e.g. brand), we use cross-entropy loss for the multiclass classification prediction task, and for a continuous signal (e.g. price), we use mean squared loss for the regression prediction task.

Supervised Disentanglement vs Unsupervised Disentanglement

A key issue we examine in this work is whether structured data variables typically found in marketing contexts (e.g., brand) can be used as supervisory signals to improve disentanglement, and thus our ability to discover unstructured (visual) characteristics. (Locatello et al. 2019) showed that this is challenging in theory, as there is no guarantee for finding a unique disentangled representation using an unsupervised approach.[6] Locatello et al. (2020) further showed that this challenge could be resolved by using *supervision* with ground truth characteristics, in which lower supervised loss is correlated with a high score on disentanglement performance metrics. However, their approach (and several related in machine learning) are not aligned with our goal of characteristic discovery for several reasons. First, needing ground truth labels of the characteristics conflicts with our goals as these labels are that of the characteristics we are trying to discover in the first place. Second, if the approach requires humans to (partially) label characteristics, then the approach is not fully automated. Third, disentanglement performance metrics used in machine learning are generally only usable for synthetic datasets with access to ground truth characteristics (Higgins et al. 2021).

Our work instead takes an empirical viewpoint that is practical to marketing contexts. We do not assume access to ground truth characteristics, and consequently, we measure disentanglement performance using two evaluation methods: (1) supervised classification accuracy on held-out data, and (2) Unsupervised Disentanglement Ranking (UDR). UDR is a metric proposed by Duan et al. (2020) to work in real-world data where ground truth is not available, and is based on heuristics of good disentanglement rather than theoretical guarantees. This metric posits that for a particular dataset and a particular VAE-based disentangled representation learning model, the unstructured (visual) characteristics learned using different random seeds should be similar, whereas every entangled representation is different in its own way. This is because while the model defines all the

[6] One drawback of using a supervised disentanglement approach is that it assumes a single canonical factorisation of generative factors (Duan et al. 2020). For example, color can be represented by alternative representations like RGB, HSV, HSL, CIELAB or YUV.

So, if a disentangled representation learns color aligned with HSV, then it will perform poorly if the supervised metric assumes that color should be represented by RGB.

hyperparameter levels, the random seed levels only determine the initial levels of the parameters for the neural net and any sampling within the algorithm (e.g., dataset splitting or batch-level data sampling during training). Specifically, UDR expects two disentangled representations learned from the same model on the same dataset with two different random seeds to be similar up to permutation and sign inverse. Appendix C has details on how UDR is calculated as well as comparisons with related disentanglement metrics.

We investigate how well each of the (six) supervisory signals lead to better disentanglement than the one learned by the purely unsupervised approach. We select informative visual characteristics and ignore uninformative visual characteristics by calculating the $KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]$ for each visual characteristic and then select characteristics with KL divergence above a threshold (Duan et al. 2020). Variation across an uninformative characteristic would produce little to zero visual change in the image.[7] Both the supervised and unsupervised disentanglement approaches require model training (i.e., how model parameters are estimated), model selection (i.e., how model hyperparameters are chosen), and model evaluation (i.e., how good is a selected model's resulting disentanglement). However supervised and unsupervised approaches require different model training and selection steps, while having the same evaluation step (so the comparison between unsupervised and supervised approaches is apples-to-apples). We therefore describe these steps separately for the supervised and unsupervised approaches in Appendix C.

EMPIRICAL APPLICATION

We consider an application of our proposed approach in the visual domain. Examples of visually important aspects of products that impact consumer demand include a product's design, packaging and even promotion materials. Understanding their impact on consumer demand is of considerable interest (Kang et al. 2019; Burnap, Hauser, and Timoshenko 2019; Liu et al. 2017).

[7] Rolinek, Zietlow, and Martius (2019) showed that during training, models based on VAEs enter a *polarised regime*. By polarised regime, we mean that many unstructured (visual) characteristics are switched off by being reduced to the prior $q_\phi(z_j) = p(z_j)$. This is due to the choice of a diagonal posterior. Entering this polarized regime allows the models to disentangle. These switched off characteristics are referred to as uninformative characteristics. Duan et al. (2020) showed that models with fewer uninformative characteristics do not disentangle well and their unstructured (visual) characteristics are hard to semantically interpret.

Existing methods either ignore visual characteristics completely, or collapse all visual (and other) unobservable characteristics to form an unobserved product characteristic (Cho, Hasija, and Sosa 2015).

Data

Our data includes watches auctioned at Christie's auction house, spanning the years 2010 to 2020. We choose this data for two main reasons. First, visual characteristics of watches are important considerations for consumers in this market. Second, the auction mechanism leads to a truthful revelation of the buyer's willingness to pay (WTP) for the watches. This allows us to use price as a supervisory signal as it corresponds to an economic primitive, i.e. demand, for each watch in our dataset.

For each auctioned watch in the dataset, we have its image, structured product characteristics, and the hammer price paid at the auction (i.e., the willingness to pay). Structured characteristics include the brand of the watch, model of the watch, year of manufacture or *circa*, type of movement associated with the watch, dimensions of the watch and materials used in the watch. Figure 3 shows a sample of watch images in our dataset. The hammer price corresponding to a consumer's willingness to pay (in $1000s) are in constant 2000 dollars, adjusted for inflation using the Consumer Price Index.

A total of 199 unique brands are present in the data. Audemar's Piguet, Cartier, Patel Philippe and Rolex are the four brands with the largest share of observations, while the remaining brands are coded as Others. Circa is coded as Pre-1950, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s and 2010s. Movement of a watch is classified as either mechanical, automatic or quartz. Dimensions of the watch refers to the watch diameter in case of a circular dial or the length of the longest edge in case of a rectangular dial (in millimeters). Material is coded as gold, steel, a combination of gold and steel or other materials. Table 1 provides summary statistics of the auctioned watches.

Figure 3: Sample of Watches Auctioned at Christie's

Table 1: Summary Statistics of Structured characteristics of Auctioned Watches

Statistic	Mean	SD	Min	Max
Brand (Audemar's Piguet)	0.06	0.24	0	1
Brand (Cartier)	0.07	0.25	0	1
Brand (Patek Philippe)	0.20	0.40	0	1
Brand (Rolex)	0.18	0.38	0	1
Brand (Others)	0.49	0.50	0	1
Circa (Pre-1950s)	0.05	0.21	0	1
Circa (1950s)	0.05	0.22	0	1
Circa (1960s)	0.07	0.26	0	1
Circa (1970s)	0.10	0.30	0	1
Circa (1980s)	0.08	0.26	0	1
Circa (1990s)	0.19	0.39	0	1
Circa (2000s)	0.33	0.47	0	1
Circa (2010s)	0.14	0.35	0	1
Movement (Automatic)	0.54	0.50	0	1
Movement (Mechanical)	0.36	0.48	0	1
Movement (Quartz)	0.11	0.31	0	1
Watch Dimensions (in mm)	36.21	6.83	9	62
Material (Gold)	0.60	0.49	0	1
Material (Gold and Steel)	0.05	0.22	0	1
Material (Steel)	0.28	0.45	0	1
Material (Others)	0.07	0.25	0	1
Hammer Price (in $000s)	23.25	55.18	1.00	950.20

Notes: The unit of analysis for each auction is a single watch.

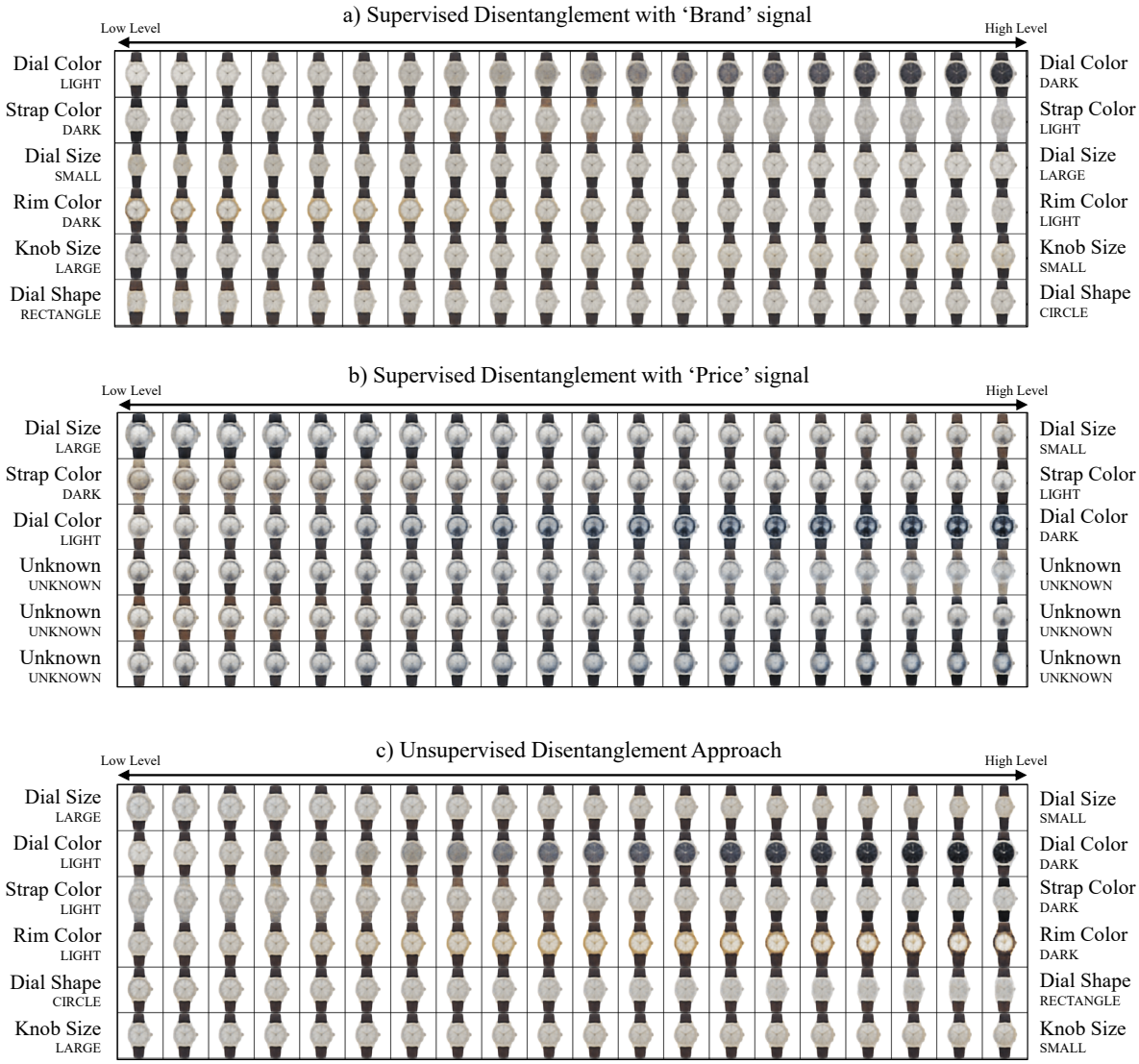
Model: Training, Selection and Evaluation

The dataset is segmented into training, validation and test data. To avoid data leakage, each watch model was present only in one of the above subsets. The model training process requires us to specify the dimension of the latent space, and the number of random seeds used. The process for supervised and unsupervised both involve training across multiple seeds, but differ in the model selection step, which obtains the hyperparameters for subsequent use (see Table A.4 in Appendix C for obtained hyperparameters corresponding to each disentanglement approach). Finally, the model evaluation compares the set of supervised models (with each structured characteristic serving as a supervisory signal) and the unsupervised model to evaluate which of these discover visual characteristics that are (a) predictive of structured characteristics, and (b) obtain the highest level on the UDR metric. This process is illustrated in Figure A.2 and detailed in Appendix C. The architecture of the model is further specified in Appendix C.

Results: Discovered Visual Characteristics

Figure 4 gives example output of discovered visual characteristics corresponding to supervisory signals 'brand' and 'price' as well as the 'unsupervised approach'. In each row of the figure, we show how the watch image changes based on changes in levels of one visual characteristic, while keeping all the other characteristics fixed. We only show six visual characteristics as rest of the characteristics are found to be uninformative. By uninformative, we mean that traversing along those dimensions leads to no visual changes. For a quantitative analysis detailing this aspect, see Appendix C. From ex-post human inspection (by researchers), we observe that both 'brand (supervised approach)' and 'unsupervised approach' are able to discover six distinct unstructured (visual) characteristics that are independent as well as semantically interpretable. These are 'dial color', 'strap color', 'dial size', 'rim color', 'knob size' and 'dial shape'. However, 'price (supervised approach)' is only able to discover 'dial size', 'strap color' and 'dial color' but is not able to discover 'rim color', 'dial shape' and 'knob size'. We refer readers to Appendix D to see the visual characteristics discovered by other supervisory signals.

Figure 4: Discovered Visual characteristics from Supervised and Unsupervised Approaches



Notes: Latent traversals along a *focal watch* used to visualise the semantic meaning encoded by single visual characteristic learnt by a trained model. In each row, the quantitative level of a single characteristic is varied keeping the other characteristics fixed. The resulting reconstruction is visualized. **a**: Discovered visual characteristics learned by supervising the characteristics to predict the brand simultaneously. **b**: Discovered visual characteristics learned by supervising the characteristics to predict the price simultaneously. **c**: Discovered visual characteristics learned with no supervision.

We conduct two surveys to validate our method. We choose respondents based in the US who are fluent in English. For both surveys, we employ an attention check in which we ask respondents to write the name of the object in the survey while we show them an image of giraffe. In the first survey, we address the interpretability of the discovered characteristics from unstructured data. We do so by generating counterfactual images that vary along only one visual characteristic. For example, each watch image (see Figure A.7 from Appendix F) is generated by fixing all except one focal visual characteristic and only changing the level of the focal visual characteristic. We ask 600 respondents (100 for each visual characteristic) to select which part of the watch (if any) is changing as move from left to right. Then, we ask respondents to specify how that part of the watch changing. Now that we have got a semantic meaning of the visual characteristic, we address the quantification of this characteristic through a second survey. Here, we generate several pairs of watch images that are different only along one visual characteristic. In this survey (see Figure A.8 from Appendix F), we ask 300 respondents (50 for each visual characteristic) to select respondents to select the pair whose watches are closer to each other. We then verify whether the responses matched with our algorithm's results. Table 2 shows us the percentage of respondents who agreed with the algorithm's output. We find that the agreement ranged from a high of 94% for strap color to a low of 42% for knob size. We note that "*Nothing is visually changing*" was the second most chosen option when the watch was changing along the knob size. We expect that this could be due to the low generative quality of the counterfactual images by the VAE.

When we measure whether the algorithms's quantification of characteristic levels (as measured by similarity between pairs of images) was consistent with human raters' judgment, we find that a strong majority agree with the algorithm's quantification scale for dial size, dial color, strap color, dial shape and knob size respectively. See Table 2 for details.

We next compare the set of disentanglement approaches using two different methods for quantitative evaluation. First, we evaluate the visual characteristics discovered from different supervisory signals by their average performance on downstream tasks. Specifically, we characterize how well discovered visual characteristics from each of the seven disentanglement models (six super-

vised disentanglement models and single unsupervised disentanglement model) are able to predict structured product characteristics - namely brand, year of manufacture or *circa*, type of movement associated with the watch and the materials used in the watch. Using the trained models, we classify the watches in the test set. We then calculate the average accuracy for each supervisory signal. Second, we evaluate the models based on the UDR metric, where a higher UDR is preferable since it corresponds to a more disentangled representation.

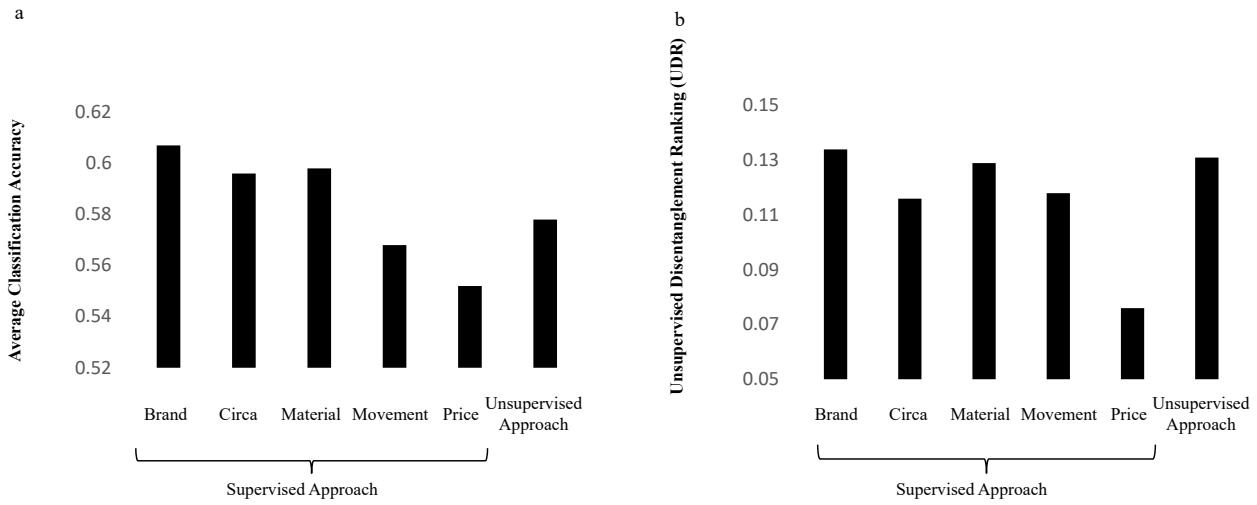
Table 2: Summary Statistics of Discovered Visual characteristics (from 'Brand' Signal)

Visual characteristic	Interpretability Survey	Quantification Survey
Dial Size	72%	83%
Dial Color	87%	92%
Strap Color	94%	92%
Rim Color	87%	88%
Dial Shape	72%	68%
Knob Size	42%	85%

The results of the two quantitative evaluations of disentanglement performance are detailed in Figure 5. We plot the average accuracy for each supervisory signal in panel (a). We find that the visual characteristics learned from the supervisory signal 'brand' has the highest accuracy and thus is most useful for downstream tasks. On the other hand, visual characteristics learned from 'price' has the lowest accuracy and thus are the least useful for downstream tasks. We also plot the UDR for each supervisory signal in panel (b). We find that the visual characteristics learned from the supervisory signal 'brand' also has the highest UDR. Similarly, visual characteristics learned from 'price' has the lowest UDRs. Thus, this UDR metric matches with evaluating visual characteristics on multiple downstream tasks. Interestingly, while the visual characteristics learned from the unsupervised approach has the second highest UDR, its downstream performance is not quite as good (it was fourth highest).

From both these evaluation approaches, we find that supervision with a typical dependent variable in marketing such as 'price' does not necessarily lead to to the best disentanglement, whereas 'brand' serves as the best signal. Perhaps even more surprising, unsupervised disentanglement *in*

Figure 5: Disentanglement Performance of Supervised & Unsupervised Approaches



Notes: Using each set of discovered visual characteristics corresponding to the unsupervised approach as well as six different supervised approaches, **a**: we classify watches according to their structured product characteristics and then calculate average classification accuracy; **b**: we calculate the UDR metric.

practice leads to better disentanglement than price and other supervisory signals. In other words, while supervised signals are necessary for guaranteed disentanglement *in theory*, in practice we find that several supervisory signals lead to worse performance than unsupervised disentanglement. Further, conditional on using supervised methods, deep learning literature (Locatello et al. 2020) has assumed ground truth on the visual characteristics as the supervisory signals. We use other structured characteristics as the supervisory signal because obtaining ground truth on real-world datasets is not feasible. We show that supervising on structured characteristics helps in discovering disentangled visual characteristics. Thus, supervision can help even in the absence of ground truth on visual characteristics.

Rolinek, Zietlow, and Martius (2019) had shown that the loss function used in unsupervised approaches does not in itself encourage disentanglement. Indeed Locatello et al. (2019) showed that any rotationally invariant prior makes disentangled representations learnt in an unsupervised setting unidentifiable when optimizing the loss function for unsupervised approaches. However, Rolinek, Zietlow, and Martius (2019) showed that the interactions between the reconstruction objective and the enhanced pressure to match a diagonal prior created by the modified objectives of

the disentangling VAEs force the decoder neural net to pursue orthogonal representations. During training, models based on VAEs enter a polarised regime. By polarised regime, we mean that many unstructured (visual) characteristics are switched off by being reduced to the prior $q_\phi(z_j) = p(z_j)$. Entering this polarized regime and ensuring sufficient dimensionality of the latent space are critical in allowing the models to disentangle.

It is also useful to understand why using brand as a supervisory signal helps in disentanglement while a signal such as price does not. We might expect this is because, in our case, watches have less pronounced variation in visual aesthetics by the price at which they are auctioned. At the same time, brand appears to be the best supervisory signal according to both the evaluation approaches. We conjecture that this is because watches of different brands have different visual aesthetics (or "signatures"). Further, existing marketing research has shown that brands have different personalities (Aaker 1997) that can be expressed through their product-related characteristics, product category associations, brand name, symbol or logo, advertising style, price, distribution channel and user imagery (Batra, Lehmann, and Singh 1993; Liu, Dzyabura, and Mizik 2020). This allows the brand variable to serve as a good supervisory signal in our setting. We also find that other structured product characteristics such as circa, material and movement have reasonable performance. We hope our findings provide guidance to future researchers and managers using this method to automatically discover visual characteristics for their data. We provide statistics for the discovered characteristics from supervision using brand in Appendix E of the Supplement.

MANAGERIAL APPLICATION: VISUAL CONJOINT ANALYSIS

We next assess whether or not consumer preferences are indeed impacted by discovered visual product characteristics, and in doing so, show how these characteristics can be used in a common managerial application. To this end, we conducted a choice-based conjoint analysis that varied images of watches along each of the six discovered visual characteristics. Our application is in line with recent work aiming to measure the effect of product images on consumer preferences (Dotson et al. 2019; Zhang et al. 2021a), with major difference being the method of how the product images

are obtained, with our interest in having variation along specified visual characteristics.

A total of 401 respondents were initially sourced from the survey panel Prolific. After a consent page and introduction page, each respondent was presented with 15 choice-based conjoint questions. Each question consisted of asking the respondent to choose between two images of watches or a "None" option based on their individual preferences. Respondents were not given any other information about the watches (e.g., brand) and were instructed to only choose based on visually comparing the two watches. Of the 15 choice-based conjoint questions, 3 were repeated (to assess respondent consistency), with the remaining 12 randomly generated. Appendix F and Figure A.9 show snapshots of the conjoint survey. To improve statistical efficiency, each of the random pairs of watches were generated prior to fielding the conjoint survey by balancing pairwise overlap between each visual characteristics using Sawtooth Software's Lighthouse Studio. After the choice-based conjoint questions, respondents were directed to an instruction manipulation check (IMC) question to assess attentiveness by asking respondents to choose the last answer in a multiple choice question. Of the 401 respondents, 17 respondents did not pass the IMC and were filtered out leaving a final total of 384 respondents.

We then obtained individual-level part-worths corresponding to each respondent's preferences over the six visual characteristics. Moreover, following recent work in visual conjoint analysis, we also include in the choice model specification all 15 pair-wise interactions of the six characteristics (Sylcott, Michalek, and Cagan 2015). We follow industry-standard estimation using a hierarchical Bayesian model (Lenk et al. 1996). Specifically, we assume that individual-level part-worths are drawn from a multivariate Gaussian common to all respondents, and use Markov chain Monte Carlo (MCMC) to obtain the posterior distribution using 10,000 drawn samples after initially discarding the first 10,000 samples to allow posterior mixing and convergence.

Table 3 details the aggregate-level part-worth estimates. Respondents indeed have preferences over the six visual characteristics as well as their interactions. While at an aggregate level, most main effects had large posterior standard deviation due to consumer heterogeneity (e.g., some consumers prefer brown watch straps over black ones, and vis-a-versa), certain interactions between

visual characteristics were more homogeneous (Sylcott, Michalek, and Cagan 2015). In particular, the interactions of 'Dial Color x Rim Color', 'Dial Color x Strap Color' both had estimated part-worths that were greater than one posterior standard deviation. The corresponds to respondents having visual preferences for contrasting 'Dial Color' and 'Rim Color', such as a black dial with silver rim, while preferring similar colors for the 'Dial Color' and 'Strap Color', such as silver dial with silver strap.

Table 3: Estimated consumer preference part-worths via choice-based conjoint analysis

Visual characteristic	Average Part-Worth	Posterior Std. Deviation
Dial Size	$\mathbf{-0.760}$	$\mathbf{0.394}$
Dial Color	0.011	1.151
Strap Color	-0.216	0.403
Rim Color	0.102	0.997
Dial Shape	-0.015	0.178
Knob Size	0.242	0.526
Dial Size * Dial Color	0.102	0.218
Dial Size * Strap Color	0.012	0.189
Dial Size * Rim Color	0.022	0.205
Dial Size * Dial Shape	0.073	0.179
Dial Size * Knob Size	-0.063	0.271
Dial Color * Strap Color	$\mathbf{0.404}$	$\mathbf{0.210}$
Dial Color * Rim Color	$\mathbf{-0.312}$	$\mathbf{0.205}$
Dial Color * Dial Shape	-0.053	0.159
Dial Color * Knob Size	0.255	0.380
Strap Color * Rim Color	0.053	0.175
Strap Color * Dial Shape	0.060	0.156
Strap Color * Knob Size	-0.090	0.168
Rim Color * Dial Shape	0.060	0.151
Rim Color * Knob Size	0.001	0.284
Dial Shape * Knob Size	-0.054	0.167
None	0.827	1.934

Notes: Bold indicates average aggregate effect greater than one posterior standard deviation. See works such as (Lenk et al. 1996; Rao et al. 2014) for more details on hierarchical Bayesian conjoint and related metrics.

DISCUSSION AND CONCLUSION

Visual characteristics are known to be important across several product categories, but have been challenging to discover in a way that is automated, scalable and human-interpretable. We have proposed a methodology to automatically identify visual characteristics from unstructured image data, which builds upon the disentanglement literature in machine learning.

Our approach leverages typically available marketing variables like brand to act as a supervisory signal to obtain better disentanglement performance, in terms of both semantic interpretability and accuracy. From a methodological perspective, our approach helps in cases where such visual characteristics are not labeled, but we have structured product data. Our approach complements the research in machine learning that has focused on cases where ground truth about visual characteristics is available.

Prior theoretical research in computer science showed that multiple visual characteristic representations are possible in the absence of a supervisory signal, leading to uncertainty about the true representation (Locatello et al. 2019). This uncertainty led to the machine learning literature focusing mostly on supervised approaches for disentanglement, with the assumption that ground truth on visual characteristics was available as a supervisory signal. This assumption in turn is challenging when the research objective is precisely the discovery of these visual characteristics. Our research sidesteps this need for ground truth on visual characteristics by using structured characteristics for supervision.

Applying our method to data on images of watches combined with structured product characteristics, we discover 6 visual characteristics, along with a quantification of their levels for each product. We find that all structured characteristics are not equally helpful for achieving disentanglement, with brand proving to be a valuable supervisory signal. We validate that our method discovers visual characteristics that are more semantically interpretable, whereas most automated approaches do not focus on interpretability. We have demonstrated the application of our disentanglement method in an application involving visual conjoint analysis: examining how such

unstructured (visual) characteristics impact consumer preferences.

Our approach can also be extended in several ways. First, it can be used for other downstream marketing tasks like product positioning maps in visual space (in addition to structured characteristics space), and also other models of demand incorporating the discovered visual characteristics. Second, from a methodological perspective it would be helpful to obtain more insight into the specific conditions under which certain combinations of signals might produce better disentanglement. Finally, our method can be extended to other modalities of data, e.g., text, voice, or video.

REFERENCES

Aaker, Jennifer L (1997), "Dimensions of brand personality," *Journal of Marketing Rresearch*, 34 (3), 347–356.

Achille, Alessandro and Stefano Soatto (2018), "Emergence of Invariance and Disentanglement in Deep Representations," *Journal of Machine Learning Research*, 19 (1), 1947–1980.

Anand, Piyush and Vrinda Kadiyali (2020), "Smoke and Mirrors: Impact of E-Cigarette Taxes on Underage Social Media Posting," *Working Paper*.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou "Wasserstein Generative Adversarial Networks," "International Conference on Machine Learning," pages 214–223 (2017).

Batra, Rajeev, Donald Lehmann, and Dipinder Singh (1993), "The Brand Personality Component of Brand Goodwill: Some Antecedents and Consequences.," *Brand Equity & Advertising: Advertising's Role in Building Strong Brands*, pages 83–96.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013), "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, 35 (8), 1798–1828.

Berlyne, Daniel E (1973), "Aesthetics and psychobiology," *Journal of Aesthetics and Art Criticism*, 31 (4).

Berry, Steven, James Levinsohn, and Ariel Pakes (1995), "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, pages 841–890.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112 (518), 859–877.

Burgess, C., I. Higgins, A. Pal, Loic Matthey, Nick Watters, G. Desjardins, and Alexander Lerchner "Understanding disentangling in β-VAE," "Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems," (2017).

Burnap, Alex, John R. Hauser, and Artem Timoshenko (2019), "Design and Evaluation of Product Aesthetics: A Human-Machine Hybrid Approach," *Available at SSRN 3421771*.

Chen, Ricky T. Q., Xuechen Li, Roger B Grosse, and David K Duvenaud "Isolating Sources of Disentanglement in Variational Autoencoders," "Advances in Neural Information Processing Systems," pages 2615–2625 (2018).

Chen, Xi, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," "Advances in Neural Information Processing Systems," pages 2180–2188 (2016).

Cheng, Zhaoqi, Dokyun Lee, and Prasanna Tambe (2022), "InnoVAE: Generative AI for Understanding Patents and Innovation," *Available at SSRN*.

Cheung, Brian, Jesse A. Livezey, Arjun K. Bansal, and Bruno A. Olshausen "Discovering Hidden Factors of Variation in Deep Networks," "Workshop at International Conference on Learning Representations," (2015).

Cho, Hallie, Sameer Hasija, and Manuel Sosa (2015), "How Important is Design for the Automobile Value Chain?," *Available at SSRN 2683913*.

Dew, Ryan, Asim Ansari, and Olivier Toubia (2021), "Letting Logos Speak: Leveraging Multiview Representation Learning for Data-Driven Branding and Logo Design," *Marketing Science*.

Dotson, Jeffrey P, Mark A Beltramo, Elea McDonnell Feit, and Randall C Smith (2019), "Modeling the Effect of Images on Product Choices," *Available at SSRN 2282570*.

Duan, Sunny, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, and Irina Higgins "Unsupervised Model Selection for Variational Disentangled Representation Learning," "International Conference on Learning Representations," (2020).

Dzyabura, Daria, Siham El Kihal, John R Hauser, and Marat Ibragimov (2019), "Leveraging the power of images in managing product return rates," *Available at SSRN 3209307*.

Eastwood, Cian and Christopher KI Williams "A framework for the quantitative evaluation of disentangled representations," "International Conference on Learning Representations," (2018).

Gabbay, Aviv, Niv Cohen, and Yedid Hoshen (2021), "An image is worth more than a thousand words: Towards disentanglement in the wild," *Advances in Neural Information Processing Systems*, 34.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020), "Generative adversarial networks," *Communications of the ACM*, 63 (11), 139–144.

Green, Paul E, Abba M Krieger, and Yoram Wind (2001), "Thirty years of conjoint analysis: Reflections and prospects," *Interfaces*, 31 (3_supplement), S56–S73.

Griffin, Abbie and John R Hauser (1993), "The voice of the customer," *Marketing Science*, 12 (1), 1–27.

Guadagni, Peter M and John DC Little (1983), "A logit model of brand choice calibrated on scanner data," *Marketing Science*, 2 (3), 203–238.

Higgins, Irina, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick (2021), "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons," *Nature Communications*, 12 (1), 1–14.

Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," "International Conference on Learning Representations," (2017).

Hoffman, Matthew D and Matthew J Johnson "Elbo surgery: yet another way to carve up the variational evidence lower bound," "Workshop in Advances in Approximate Bayesian Inference, Neural Information Processing Systems," (2016).

Hyvärinen, Aapo and Petteri Pajunen (1999), "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, 12 (3), 429–439.

Kang, Namwoo, Yi Ren, Fred Feinberg, and Panos Papalambros (2019), "Form + Function: Optimizing Aesthetic Product Design via Adaptive, Geometrized Preference Elicitation," *arXiv preprint arXiv:1912.05047*.

Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila "Training Generative Adversarial Networks with Limited Data," "Advances in Neural Information Processing Systems," Vol. 33., pages 12104–12114 (2020).

Karras, Tero, Samuli Laine, and Timo Aila "A style-based generator architecture for generative adversarial networks," "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition," pages 4401–4410 (2019).

Karush, William (1939), "Minima of functions of several variables with inequalities as side constraints," *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*.

Khemakhem, Ilyes, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen "Variational autoencoders and nonlinear ica: A unifying framework," "International Conference on Artificial Intelligence and Statistics," pages 2207–2217, PMLR (2020).

Kim, Hyunjik and Andriy Mnih "Disentangling by Factorising," "International Conference on Machine Learning," pages 2649–2658 (2018).

Kingma, Diederik P, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling "Semi-supervised Learning with Deep Generative Models," "Advances in Neural Information Processing Systems," pages 3581–3589 (2014).

Kingma, Diederik P and Max Welling "Auto-Encoding Variational Bayes," "International Conference on Learning Representations," (2014).

Klys, Jack, Jake Snell, and Richard S. Zemel "Learning Latent Subspaces in Variational Autoencoders," "Advances in Neural Information Processing Systems," pages 6444–6454 (2018).

Kotler, Philip and G Alexander Rath (1984), "Design: A powerful but neglected strategic tool," *Journal of Business Strategy*, 5 (2), 16–21.

Kuhn, Harold W and Albert W Tucker "Nonlinear programming," "Traces and emergence of nonlinear programming," pages 247–258 (2014).

Kulkarni, Tejas D., William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum "Deep convolutional inverse graphics network," "Advances in Neural Information Processing Systems," pages 2539–2547 (2015).

Kumar, Abhishek, Prasanna Sattigeri, and Avinash Balakrishnan "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations," "International Conference on Learning Representations," (2017).

Lancaster, Kelvin J (1966), "A new approach to consumer theory," *Journal of Political Economy*, 74 (2), 132–157.

Lee, Wonkwang, Donggyun Kim, Seunghoon Hong, and Honglak Lee "High-fidelity synthesis with disentangled representation," "European Conference on Computer Vision," pages 157–174, Springer (2020).

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, and Martin R Young (1996), "Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs," *Marketing Science*, 15 (2), 173–191.

Li, Xi, Mengze Shi, and Xin Shane Wang (2019), "Video mining: Measuring visual information using automatic methods," *International Journal of Research in Marketing*, 36 (2), 216–231.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020), "Visual listening in: Extracting brand image portrayed on social media," *Marketing Science*, 39 (4), 669–686.

Liu, Yan, Krista J Li, Haipeng Chen, and Subramanian Balachander (2017), "The effects of products' aesthetic design on demand and marketing-mix effectiveness: The role of segment prototypicality and brand consistency," *Journal of Marketing*, 81 (1), 83–102.

Locatello, Francesco, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," "International Conference on Machine Learning," pages 4114–4124 (2019).

Locatello, Francesco, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen "Weakly-supervised disentanglement without compromises," "International Conference on Machine Learning," pages 6348–6359, PMLR (2020).

Locatello, Francesco, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem "Disentangling Factors of Variations Using Few Labels," "International Conference on Learning Representations," (2020).

Mahajan, Vijay, Paul E Green, and Stephen M Goldberg (1982), "A conjoint model for measuring self-and cross-price/demand relationships," *Journal of Marketing Research*, 19 (3), 334–342.

Malik, Nikhil, Param Vir Singh, and Kannan Srinivasan (2019), "A Dynamic Analysis of Beauty Premium," *Available at SSRN 3208162*.

Mathieu, Michael, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun "Disentangling factors of variation in deep representations using adversarial training," "Advances in Neural Information Processing Systems," pages 5040–5048 (2016).

Nie, Weili, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar "Semi-supervised StyleGAN for disentanglement learning," "International Conference on Machine Learning," pages 7360–7369, PMLR (2020).

Norman, Donald A (2004), *Emotional design: Why we love (or hate) everyday things* Civitas Books.

Rao, Vithala R et al. (2014), *Applied conjoint analysis* 2014, Springer.

Ridgeway, Karl and Michael C. Mozer "Learning Deep Disentangled Embeddings With the F-Statistic Loss," "Advances in Neural Information Processing Systems," pages 185–194 (2018).

Rolinek, Michal, Dominik Zietlow, and Georg Martius "Variational autoencoders pursue pca directions (by accident)," "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition," pages 12406–12415 (2019).

Roweis, Sam and Zoubin Ghahramani (1999), "A unifying review of linear Gaussian models," *Neural Computation*, 11 (2), 305–345 00715.

Siddharth, N., Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H.S. Torr "Learning Disentangled Representations with Semi-Supervised Deep Generative Models," "Advances in Neural Information Processing Systems," pages 5925–5935 (2017).

Sylcott, Brian, Jeremy J Michalek, and Jonathan Cagan (2015), "Exploring the role of interaction effects in visual conjoint analysis," *Journal of Mechanical Design*, 137 (9), 094503.

Troncoso, Isamar and Lan Luo (2020), "Look the Part? The Role of Profile Pictures in Online Labor Markets," *Available at SSRN 3709554*.

Voynov, Andrey and Artem Babenko "Unsupervised discovery of interpretable directions in the gan latent space," "International Conference on Machine Learning," pages 9786–9796, PMLR (2020).

Watanabe, Satosi (1960), "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, 4 (1), 66–82.

Yang, Jeremy, Juanjuan Zhang, and Yuhan Zhang (2021), "First law of motion: Influencer video advertising on tiktok," *Available at SSRN 3815124*.

Zhang, Mengxia and Lan Luo (2018), "Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp," *Available at SSRN 3108288*.

Zhang, Shunyuan, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan (2021a), "What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features," *Management Science*.

Zhang, Shunyuan, Nitin Mehta, Param Vir Singh, and Kannan Srinivasan (2021b), "Frontiers: Can an Artificial Intelligence Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb," *Marketing Science*, 40 (5), 813–820.

APPENDIX

A. Literature

Our work is related to two broad streams of literature. First, it is related to marketing methods for discovering characteristics and quantifying their levels. Second, our work relates to a stream of literature in machine learning known as representation learning and more specifically to disentangled representations.

Discovery of Product Characteristics Conventional methods in marketing science require a defined list of product characteristics over which consumers form preferences as inputs. Examples of methods that need a set of characteristics as inputs range from conjoint analysis and factor analysis, to reduced-form regression models and structural models. Along with the product characteristics themselves, these methods also need their characteristic levels (i.e., the values or levels of characteristics for a given product).

Researchers have long relied on widely-adopted qualitative methods such as focus groups, in-depth consumer interviews, and internal firm expertise to define the set of characteristics and their levels (Green, Krieger, and Wind 2001). While these qualitative methods are market research staples for good reason (Griffin and Hauser 1993), they require extensive human input on both the researcher and consumer side. Moreover, there are no guarantees on discovering characteristics that might be non-obvious to researchers ex-ante or hard to enunciate by consumers. Our approach complements existing market research methods to discover additional independent product characteristics automatically from unstructured data. In contrast to qualitative methods, we also discover characteristics that can be semantically interpreted by humans even if they might be non-obvious ex-ante since we correlate them with ex-post consumer decisions based on historical observed data.

Disentangled Representation Learning Our work builds on a stream of literature in machine learning known as *disentangled* representation learning, which aims to separate distinct informative factors of variation in the data (Bengio, Courville, and Vincent 2013). For example, a model to extract disentangled representations trained on a dataset of 3D objects might learn independent

factors of variation corresponding to object identity, position, scale, lighting and color.

We seek *good* disentangled representations that are both independent as well as semantically interpretable by humans. Promoting statistical independence is relatively straightforward by penalizing statistical moments, whether certain moments (e.g., minimizing correlation (Kumar, Sattigeri, and Balakrishnan 2017)) or all (e.g., penalizing mutual information (Chen et al. 2018)). From this viewpoint, while recent deep learning methods are generally aimed at learning disentangled representations from high-dimensional unstructured data (e.g., images, text, video), they may also be viewed as nonlinear extensions of classic marketing methods such as factor analysis and principle component analysis, in which the learned representations are statistically independent; albeit lower-dimensional and obtained using linear projections (Roweis and Ghahramani 1999).

Generative Modeling using GANs and VAEs The two broad classes of generative models are based on variational autoencoders (VAEs) (Kingma and Welling 2014) and generative adversarial networks (GAN) [8] (Goodfellow et al. 2020). Most state-of-the-art disentangled *representation learning* methods are based on VAEs. VAEs are comprised of two models – the encoder neural net and the decoder neural net. The encoder neural net compresses high-dimensional input data to a lower-dimensional latent vector (latent characteristics), followed by inputting the latent vector to the decoder neural net which outputs a reconstruction of the original input data. VAEs balance having both a low reconstruction error between the input and output data (e.g., images, text), as well as a KL-divergence of the latent space distribution (latent characteristics) from a researcher-defined prior distribution (e.g., Gaussian). The KL-divergence term acts as a regularizer on the latent space, such that it has desired structure (smoothness, compactness). VAEs are parametrized in both the encoder neural net and decoder neural net using neural networks whose parameters are learned jointly.

Several methods based on GANs have also been used for disentanglement. InfoGAN was one of the first scalable unsupervised methods for learning disentangled representations (Chen et al. 2016). While GANs are typically less suited relative to VAEs for representation learning,

[8] In a GAN, two neural networks compete with each other in a zero-sum game to become more accurate.

as GANs traditionally do not infer a representation[9], InfoGAN explicitly constrains a small subset of the 'noise' variables to have high mutual information with generated data. Several VAE-based methods have proven to be superior (Kim and Mnih 2018; Chen et al. 2018) than InfoGAN. Recent methods based on StyleGAN (Karras, Laine, and Aila 2019) such as Info-StyleGAN (Nie et al. 2020) are able to perform disentanglement at a much higher resolution (1024×1024) unlike the VAE-based methods. However, unlike InfoGAN, Info-StyleGAN suffers from the need for human labels or pretrained models, which can be expensive to obtain (Voynov and Babenko 2020).

We choose a VAE-based approach over a GAN-based approach for several reasons. First, our goal is to propose an easy-to-train method that can be used by researchers as well as practitioners (Lee et al. 2020). Second, our goal of discovering unique (visual) characteristics that are semantically meaningful and independent of each other requires high disentanglement performance, but reconstruction accuracy is not our primary goal (Lee et al. 2020). GANs suffer from lower disentanglement performance because they focus on localized concepts but not global concepts of the image (Gabbay, Cohen, and Hoshen 2021). On the other hand, discovered characteristics from VAEs are much more globally distributed as compared with GANs. This allows the VAE-based methods to discover few important and semantically interpretable unstructured (visual) characteristics that can represent the input raw data. Third, one of the benefits of our approach is that we are able to not just discover disentangle characteristics, but infer the levels of these characteristics for all dataums in the data. This enables use in downstream marketing tasks that require characteristic levels, for example, visual conjoint analysis to understand consumer preferences. GANs do not conventionally infer a representation of the data, and hence do not have this benefit. Finally, VAEs often require less data to train in comparison with GANs (Karras, Laine, and Aila 2019). Thus, even though GANs can provide much better reconstruction and work better for small and detailed objects (Locatello et al. 2020), we choose a VAE-based approach because of its suitability to our research question. Table A.2 summarizes the recent disentanglement methods and Table A.3 summarizes metrics to measure disentanglement.

[9]Moreover, GANs tend to suffer from training instability. Common failure modes are vanishing gradients, mode collapse, and failure to converge.

Table A.1: Comparison between VAE and GAN based methods

#	Topic	VAE	GAN	Source
1	Disentanglement Performance	High	Low	(Lee et al. 2020)
2	Quality of generated image	Low	High	(Lee et al. 2020)
3	Training instability	Low	High	(Lee et al. 2020)
4	Local v Global Concepts	Global	Local	(Gabbay, Cohen, and Hoshen 2021)
5	Data requirement	Low	High	(Karras et al. 2020)
6	Ability to work on small or detailed objects	No	Yes	(Locatello et al. 2020)

Notes: **1,2,3** According to Lee et al. (2020): "VAE-based approaches are effective in learning useful disentangled representations in various tasks, but their generation quality is generally worse than the state-of-the-arts, which limits its applicability to the task of realistic synthesis. On the other hand, GAN based approaches can achieve the high-quality synthesis with a more expressive decoder and without explicit likelihood estimation. However, they tend to learn comparably more entangled representations than the VAE counterparts and are notoriously difficult to train, even with recent techniques to stabilize the training." **4:** According to Gabbay, Cohen, and Hoshen (2021): "Such methods that rely on a pretrained unconditional StyleGAN generator are mostly successful in manipulating highly-localized visual concepts (e.g. hair color), while the control of global concepts (e.g. age) seems to be coupled with the face identity." **5:** According to Karras et al. (2020): "Acquiring, processing, and distributing the 10^5 — 10^6 images required to train a modern high-quality, high-resolution GAN is a costly undertaking. The key problem with small datasets is that the discriminator overfits to the training examples; its feedback to the generator becomes meaningless and training starts to diverge." **6** According to Locatello et al. (2020): "It is however interesting to notice how the GAN based methods perform especially well on the data sets SmallNORB and MPI3D where VAE based approaches struggle with reconstruction as the objects are either too detailed or too small."

Table A.2: Recent Disentanglement Literature

Method	Authors	Architecture	Supervised
InfoGAN	Chen et al. 2016	GAN	Unsupervised
InfoWGAN-GP	Arjovsky, Chintala, and Bottou 2017	GAN	Unsupervised
β-VAE	Higgins et al. 2017	VAE	Unsupervised
AnnealedVAE	Burgess et al. 2017	VAE	Unsupervised
FactorVAE	Kim and Mnih 2018	VAE	Unsupervised
β-TCVAE	Chen et al. 2018	VAE	Unsupervised
DIP-VAE-I and DIP-VAE-II	Kumar, Sattigeri, and Balakrishnan 2017	VAE	Unsupervised
XCov	Cheung et al. 2015	Autoencoder	Semi-Supervised
VAE-GAN	Mathieu et al. 2016	VAE-GAN	Semi-Supervised
VAE	Kingma et al. 2014	VAE	Semi-Supervised
DC-IGN	Kulkarni et al. 2015	VAE	Semi-Supervised
Conditional Subspace VAE	Klys, Snell, and Zemel 2018	VAE	Semi-Supervised
Graphical Model Structures in VAE	Siddharth et al. 2017	VAE	Semi-Supervised
Info-StyleGAN	Nie et al. 2020	GAN	Semi-Supervised
β-VAE, FactorVAE, β-TCVAE etc,	Locatello et al. 2020	VAE	Supervised

Table A.3: Various Disentanglement Metrics

Metric	Authors	Ground Truth Required?
β-VAE	Higgins et al. 2017	Yes
FactorVAE	Kim and Mnih 2018	Yes
Mutual Information Gap (MIG)	Chen et al. 2018	Yes
Modularity	Ridgeway and Mozer 2018	Yes
DCI Disentanglement	Eastwood and Williams 2018	Yes
SAP score	Kumar, Sattigeri, and Balakrishnan 2017	Yes
Unsupervised Disentanglement Ranking (UDR)	Duan et al. 2020	No

Several extensions of the VAE explicitly enforce disentanglement. β-VAE (Higgins et al. 2017) introduces an adjustable hyperparameter β that balances reconstruction accuracy with constraints to ensure that the latent space produces disentangled characteristics. AnnealedVAE (Burgess et al. 2017) modifies the training regime of β-VAE so that the tradeoff in learning disentangled representations and reconstruction accuracy is reduced. Finally, both FactorVAE (Kim and Mnih 2018) and β-TCVAE (Chen et al. 2018) propose methods to show improved disentanglement than β-VAE for the same reconstruction quality. This is because unlike β-VAE, both these methods downweight penalties on the mutual information between the data and the recovered disentangled representations that helps in the reconstruction. Most importantly, they upweight penalities that encourage the learned disentangled representations to be more factorized and hence more independent.

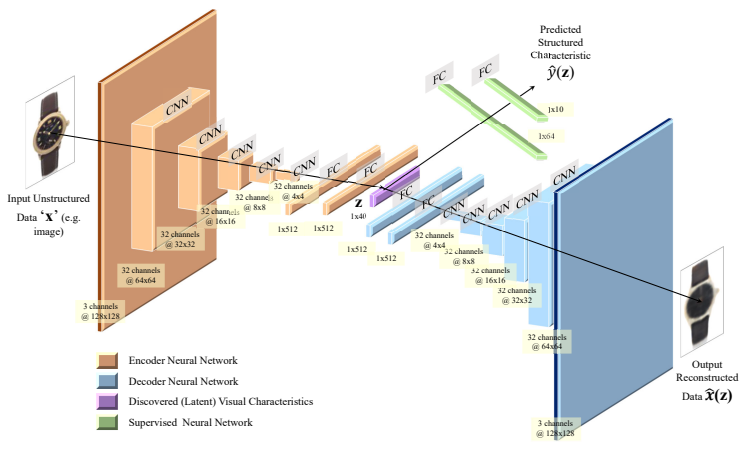
Disentanglement with and without Supervision A key challenge in *any* disentangled representation learning approach is whether a unique set of discoverable characteristics exists and whether it can be learned. This is especially relevant to real-world unstructured data that exists in the field of marketing where ground truth cannot be known. Contextualizing results from independent component analysis (Hyvärinen and Pajunen 1999) to disentanglement learning, Locatello et al. (2019) showed that there is no theoretical guarantee for learning independent characteristics using an unsupervised disentanglement approach. To address this concern, Locatello et al. (2020) showed that a small number of labelled examples with even potentially imprecise and incomplete labels is sufficient to perform model selection to learn disentangled representations.

Our work is instead motivated by marketing applications which typically do not have even partial ground truth of unstructured (visual) characteristics. For the supervised disentanglement models in this work, we use structured data typically found in marketing as supervisory signals. Specifically, we add a supervised objective to the β-TCVAE objective so that the disentangled representations are also helpful in predicting the supervisory signal in addition to ensuring lower reconstruction loss, independence between disentangled representations and an organized latent space.

B. Model Architecture

Before we build the model architecture, we decide the number of latent codes J or the maximum number of characteristics that our model aims to find. On the one hand, a small J might combine multiple visual characteristics into one which results in entanglement. On the other hand, when J is large, the model discovers redundant or irrelevant characteristics or it might even break up a true characteristic across multiple dimensions. We choose $J = 20$ to balance these considerations, based on our empirical setting. Next, we describe the model architecture.s‘ Figure A.1 shows the detailed model architecture. We modify the architecture used in Burgess et al. (2017) in order to use images of 128×128 pixels as well as to incorporate a supervised neural net. Since we

Figure A.1: Model Architecture



Notes: The encoder neural net for the VAEs consisted of 5 convolutional layers, each with 32 channels, 4×4 kernels, and a stride of 2. This was followed by 2 fully connected layers, each of 512 units. The latent distribution consisted of one fully connected layer of 40 units parameterizing the mean and log standard deviation of 20 Gaussian random variables. The decoder neural net architecture was the transpose of the encoder neural net but with the output parameterizing Bernoulli distributions over the pixels. Leaky ReLU activations were used throughout. We used the Adam optimizer with the learning rate 5e-4 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set batch size equal to 64. We train for 100 epochs.

provide an application of our proposed method in the visual domain, we use Convolutional Neural Net (CNNs) to construct the encoder neural net. In the encoder neural net, we stack a sequence of CNN layers in order to learn high-level concepts for images. Finally, we introduce 2 fully-connected (FC) layers to first flatten the output of the sequence of CNN layers and then reduce the number of dimensions in order to learn a maximum of J visual characteristics. The decoder neural net is the transpose of the encoder neural net, and is designed to reconstruct the image from the J-dimensional latent visual characteristics. Finally, we connect fully connected layers to the discovered visual characteristics to create the supervised neural net in order to predict the structured characteristics.

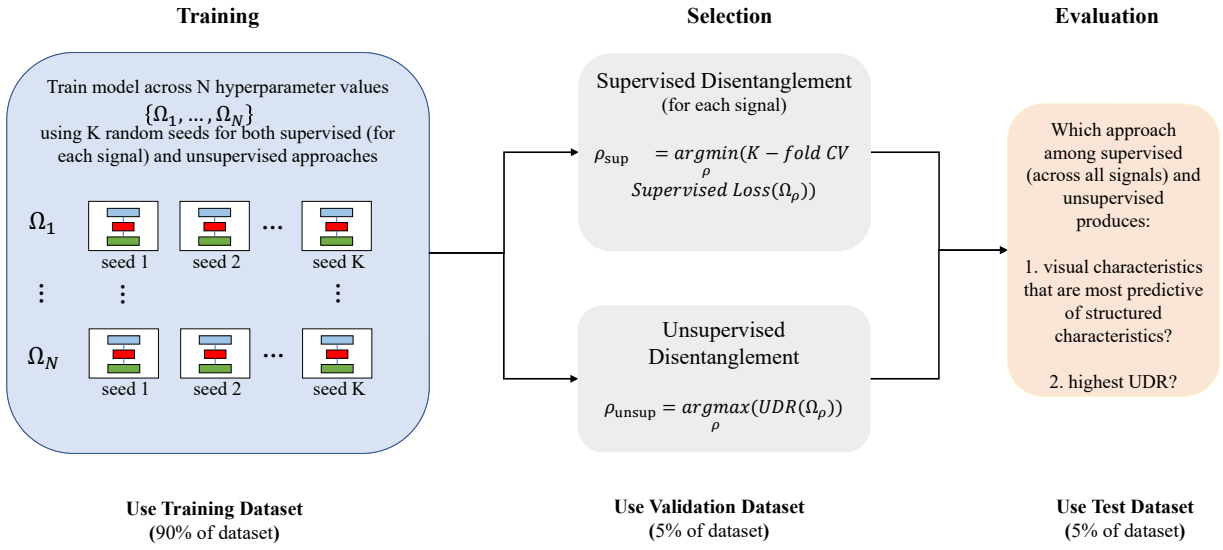
In order to train this model architecture, we need to tune the learning rate, batch size and number of training steps or epochs. A very low learning rate can lead the model to get stuck on a local minima or converge very slowly and a very high learning rate can lead the model to overshoot the minima. A low batch size increases the time required to train the model till convergence while a large batch size significantly degrades the quality of the model so that it is not generalizable beyond the training dataset. Training for low number of epochs may result in the model not converging while training for a very high number of epochs may result in the model overfitting on the train dataset.

C. Modeling Details: Model Training, Model Selection, and Model Evaluation

We detail below the process for training the models given the initial weights and hyperparameters, followed by model selection using cross-validation or UDR metrics, and finally model evaluation.
Model Training and Model Selection We divide the dataset into a training dataset for training (or learning) disentangled representations, a validation dataset for model selection using the Unsupervised Disentanglement Ranking (UDR) and a test dataset to compare various supervised approaches and the unsupervised approach in the ratio 90:5:5. Figure A.2 provides a schematic diagram for the model training and selection for the supervised and the unsupervised approaches. The training process takes in the unstructured data (watch images) as input, and chooses a structured

watch characteristics (e.g., brand) as the supervisory signal to the model (only for the supervised approach). We fix the hyperparameters corresponding to α, γ, batch size, latent space dimensions J, type of optimizer, optimizer parameters such as learning rate, type of decoder neural net and number of training steps (Locatello et al. 2020).

Figure A.2: Model Training, Model Selection, & Model Evaluation



Notes: We train N different hyperparameter (Ω) levels for both supervised and unsupervised approaches. For supervised approaches, we choose the hyperparameter level that minimize the supervised loss $P(\hat{y}(\mathbf{z}), y)$ on the validation dataset. For the unsupervised approach, we choose the hyperparameter level that maximise the UDR. We evaluate different sets of visual characteristics learned by various approaches by their predictive ability of structured product characteristics and by the UDR metric.

Supervised Approach: We sweep over levels of hyperparameters corresponding to β (weight on the total correlation loss term) and δ (weight on the prediction loss term). For each β and δ level, we calculate a 10-fold cross-validation supervised loss. We select the hyperparameter setting corresponding to the lowest cross-validated supervised loss. Table A.4 lists the hyperparameters obtained for all the supervised disentanglement approaches. Finally, we retrain the model on the entire training dataset with the chosen β and δ. We then use the trained model to extract the discovered unstructured (visual) characteristics on the test dataset.

Unsupervised Approach: We sweep over hyperparameters corresponding to β (weight on the total correlation loss term). In the unsupervised approach $\delta = 0$ by definition. We use Unsupervised Disentanglement Ranking (UDR), a metric proposed by Duan et al. (2020), for the purpose of model selection. This UDR metric allows for an automated way to select a model and does not require access to the ground truth data generative process, unlike other metrics such as β-VAE metric (Higgins et al. 2017), the FactorVAE metric (Kim and Mnih 2018), Mutual Information Gap (MIG) (Chen et al. 2018) and DCI Disentanglement scores (Eastwood and Williams 2018). We select the hyperparameter setting corresponding to the highest UDR. Appendix C has details on how UDR is calculated. Table A.4 lists the hyperparameters obtained for the unsupervised approach. Similar to the supervised approach, we use the chosen trained model to extract the discovered unstructured (visual) characteristics on the test dataset as well.

Table A.4: Hyperparameters Obtained by Model Selection Criteria

Disentanglement Approach	Signal	β	δ
Supervised	Brand	18	50
Supervised	Circa	4	35
Supervised	Material	6	25
Supervised	Movement	4	20
Supervised	Price	1	16
Unsupervised	–	18	0

Model Evaluation One of the contributions of our paper is identify the class of marketing signals which help in disentangling factors of variation. We also compare the use of such supervisory signals with an unsupervised approach. We evaluate the model along two dimensions: (a) performance in predicting the set of structured characteristics, and (b) unsupervised disentanglement ranking (UDR).

Performance on Predicting Structured characteristics: A good disentangled representation should be developed to help in a variety of downstream tasks (Bengio, Courville, and Vincent 2013). Based on this logic, we compare various supervisory signals by their ability to discover

characteristics that can be used to classify the watches according to different structured characteristics. In the downstream classification models, we train the discovered disentangled characteristics to predict a particular structured characteristics on the training dataset. Next, we predict the structured product characteristics from the test set using the trained classification model. For each set of discovered disentangled characteristics corresponding to a particular supervisory signal, we calculate the average accuracy across different classification tasks. Finally, we select the supervisory signal that provides the highest average accuracy.

Unsupervised Disentanglement Ranking: UDR is a metric developed in the deep learning literature and provides a useful benchmark to compare the approaches. There are two advantage of this metric. First, it does not require ground truth labels for the latent space (or visual characteristics), which would necessarily be human sourced. Second, it allows for a principled way to compare both unsupervised and supervised approaches. We calculate UDR for the all the unstructured (visual) characteristics discovered using supervisory signals similar to the unsupervised approach, and select the supervisory signal that provides the highest UDR.

The key idea behind Unsupervised Disentanglement Ranking (UDR) (Duan et al. 2020) is that two visual characteristics z_i and z_j would be scored highly similar if they axis align with each other up to *permutation*, *sign inverse* and *subsetting*. By permutation, we mean that the same ground truth factor c_k may be encoded by different visual characteristics within the two models $z_{i,a}$ and $z_{j,b}$ where $a \neq b$. By sign inverse, we mean that the two models may learn to encode the levels of the generative factor in the opposite order to each other, $z_{i,a} = -z_{j,b}$. By subsetting, we mean that one model may learn a subset of the factors that the other model has learnt if the relevant disentangling hyperparameters encourage a different number of latents to be switched off in the two models.

For each trained model, we perform $\kappa = 45$ pairwise comparisons with all other models trained with the same β level but with different seed levels and calculated the UDR_{ij}, where i and j index the two models. Each UDR_{ij} score is calculated by computing the similarity matrix R_{ij}, where each entry is the Spearman correlation between the responses of individual latent units of the two

models. The absolute value of the similarity matrix is then taken $|R_{ij}|$ and the final score UDR_{ij} for each pair of models is calculated according to the Equation (2). However, since this approach is an unsupervised method, it does not have theoretical guarantees to disentangle as shown by Locatello et al. (2019).

$$UDR_{ij} = \frac{1}{d_a + d_b}\left[\Sigma_b \frac{r_a^2 I_{KL}(b)}{\Sigma_a R(a,b)} + \Sigma_a \frac{r_b^2 I_{KL}(a)}{\Sigma_b R(a,b)}\right] \tag{2}$$

where a and b index the latent units of models i and j, respectively, $r_a = max_a R(a,b)$ and $r_b = max_b R(a,b)$. I_{KL} indicates an *informative* visual characteristics within a model and d is the number of such characteristics: $d_a = \Sigma_a I_{KL}(a)$ and $d_b = \Sigma_b I_{KL}(b)$. The final score for model i (UDR_i)is calculated by taking the median of UDR_{ij} across all j.

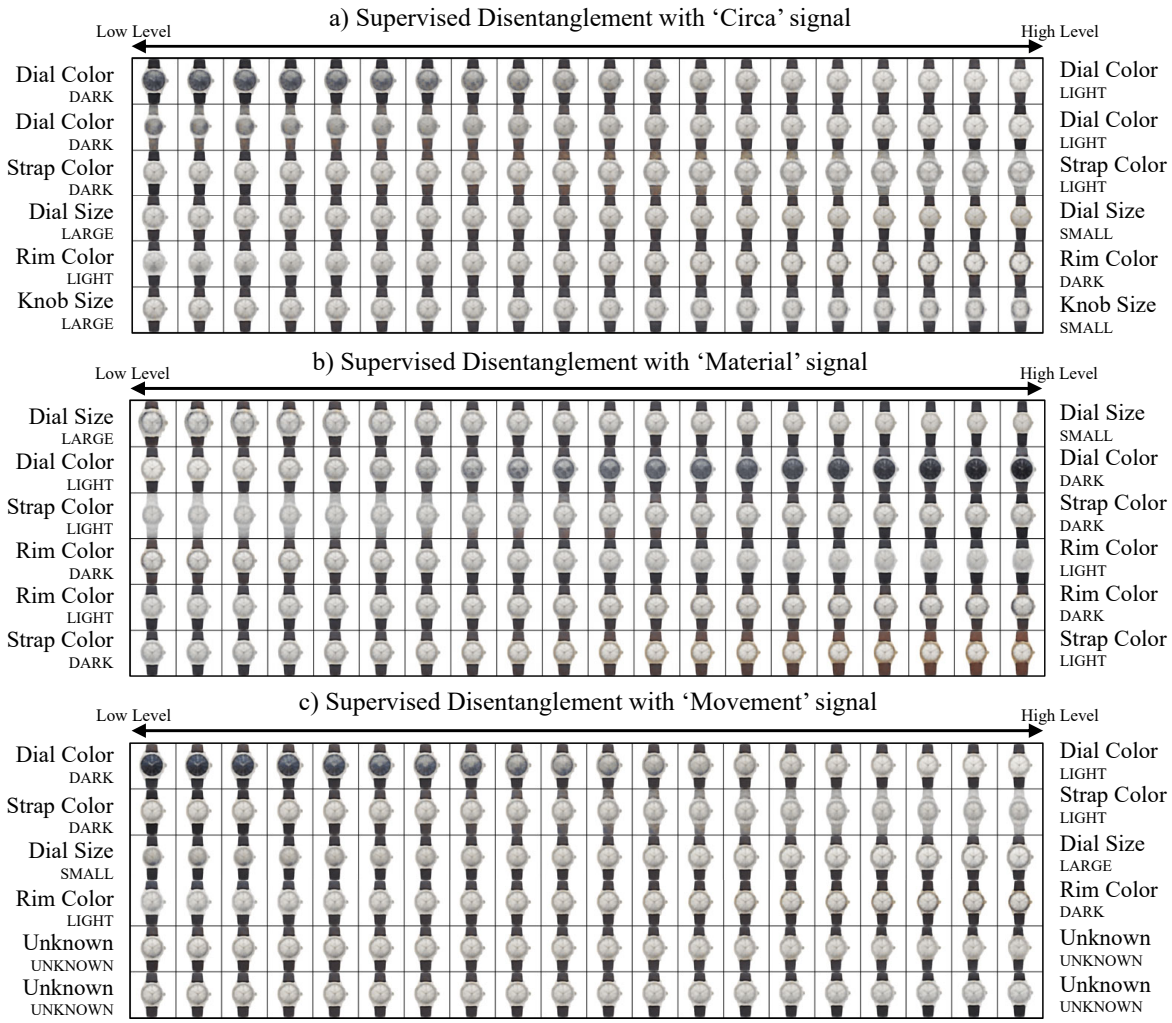
D. Results: Discovered Visual characteristics

Figure A.3 shows the discovered visual characteristics. It contains the characteristics learnt by the below supervisory signals (Circa or Decade of Manufacture, Watch Material and Type of Movement. The visual characteristics learnt by the unsupervised approach as well as supervising on brand and price are in Figure 4.

E. Quantitative Analysis of Individual-Level Discovered characteristics

Table A.5 has the summary statistics of the visual characteristic levels learned by using the supervisory signal 'brand'. Figure A.4 shows the histogram of these discovered visual characteristics. We see that the distribution of 'dial color' and 'strap color' do not seem to follow a standard normal distribution. This is because the method does not enforce any conventional parameterization on the distribution of the visual characteristics of our data. The histogram also shows that the algorithm is able to find continuous as well as discrete visual characteristics. While 'dial size', 'dial color', 'rim color' and 'dial shape' can be interpreted as continuous visual characteristics with a distribution close to gaussian, 'dial color' and 'strap color' seem to be discrete visual characteristics. A watch's 'dial color' or 'strap color' could come from one of two gaussian distributions.

Figure A.3: Discovered Visual characteristics from other Supervised Approaches



Notes: Latent traversals along a *focal watch* used to visualise the semantic meaning encoded by single visual characteristic learnt by a trained model. In each row, the value (or level) of a single visual characteristic is varied keeping the other characteristics fixed. The resulting reconstruction is visualized. **a**: Discovered visual characteristics learned by supervising the characteristics to predict the circa simultaneously. **b**: Discovered visual characteristics learned by supervising the characteristics to predict the material simultaneously. **c**: Discovered visual characteristics learned by supervising the characteristics to predict the movement simultaneously.

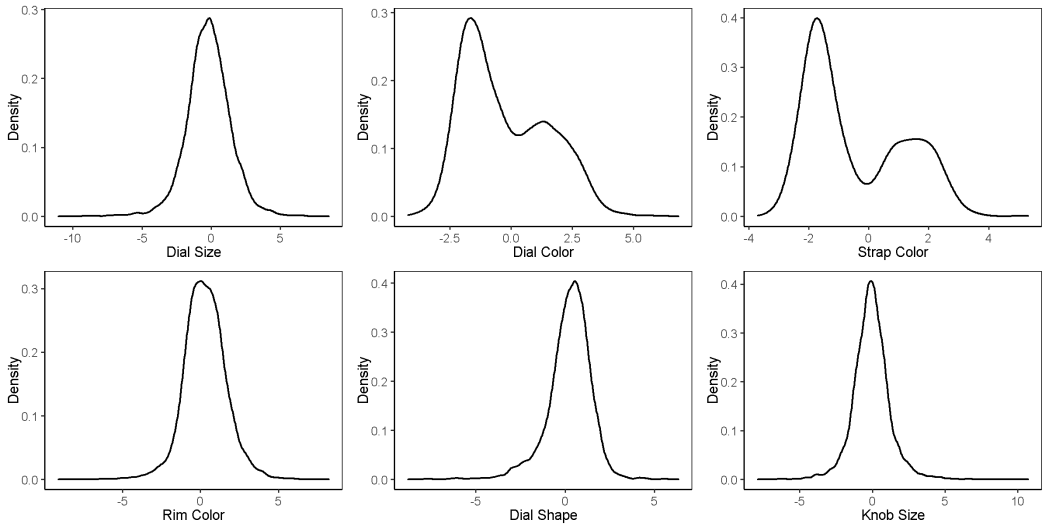
Table A.5: Summary Statistics of Discovered Visual characteristics (from 'Brand' Signal)

Visual characteristic	Mean	SD	Min	Max
Dial Size	−0.20	1.62	−11.04	8.49
Dial Color	−0.25	1.75	−4.20	6.83
Strap Color	−0.46	1.66	−3.70	5.32
Rim Color	0.30	1.35	−9.10	8.23
Dial Shape	0.23	1.22	−8.80	6.34
Knob Size	−0.08	1.26	−7.89	10.72

Why does Brand serve as a good supervisory signal?

We next provide evidence to why brand served as a good signal. We motivated the use of brand as a supervisory signal by the observation that watches from different brands would be different visually. This is because each brand would use visual aesthetics to differentiate themselves. If this is so, then the visual characteristics would have a different distribution for each brand. Figure A.6 plots the brand-wise density graph on the six discovered visual characteristics. From Figure A.6a, we see that while Audemar's Piguet and Cartier have a smaller proportion of medium-sized watches and instead have a larger proportion of smaller-sized and large-sized watches. On the other hand, Patek Philippe and Rolex have a larger proportion of medium-sized watches. From Figure A.6b, we see that Cartier and Patek Philippe has a very high proportion of light colored dials. On the other hand, Audemar's Piguet and Rolex have a similar proportion of light colored and dark colored dials. From Figure A.6c, we see that Cartier and Patek Philippe has a very high proportion of dark colored strap; Rolex has a high proportion of light colored strap; Audemar's Piguet has a similar proportion of light colored and dark colored strap. From Figure A.6d, Figure A.6e, and Figure A.6f, we see that all brands are similar in terms of rim color, dial shape, and knob size. Next, we provide evidence to why price did not serve as a good signal. From Figure A.5, we see that low priced as well as high priced watches are similar across different visual characteristics. This provides explanation as to why brand serves as good supervisory signal while price does not.

Figure A.4: Histogram of Discovered Visual characteristics (from 'Brand' Signal)



Notes: The distribution of the visual characteristics corresponding to dial size, rim color, dial shape and knob size is close to a standard normal distribution. However, the distribution of dial color and strap color is not similar to any standard distribution.

Figure A.5: Density Graph of Visual characteristics of Low Priced and High Priced Watches

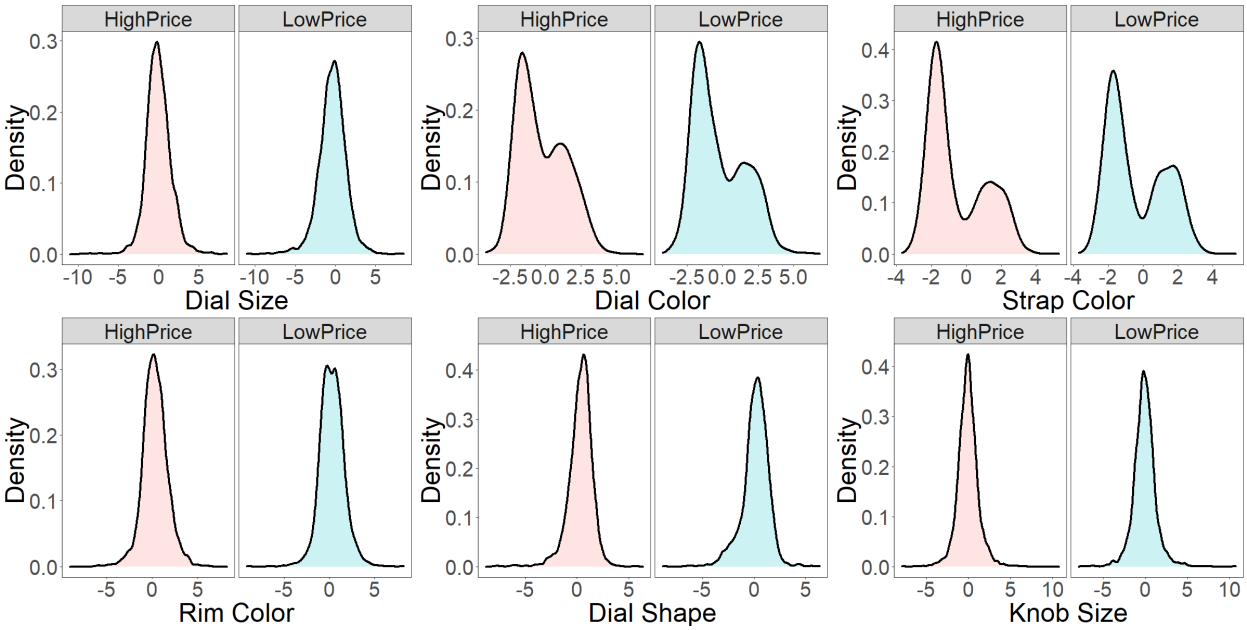
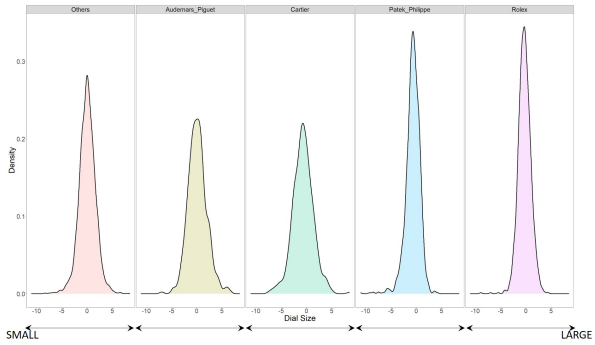
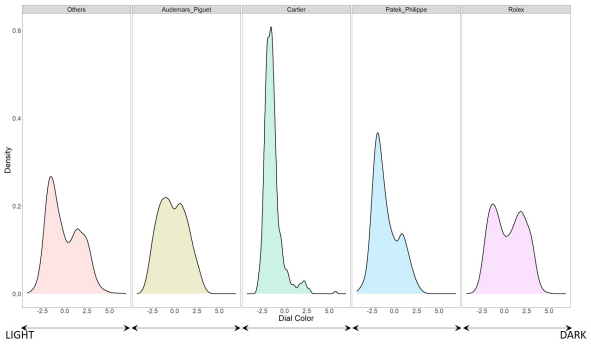


Figure A.6: Brand-Wise Density Graph of Visual characteristics

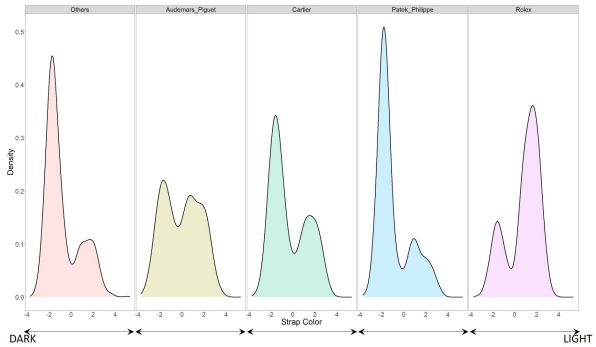
(a) Dial Size



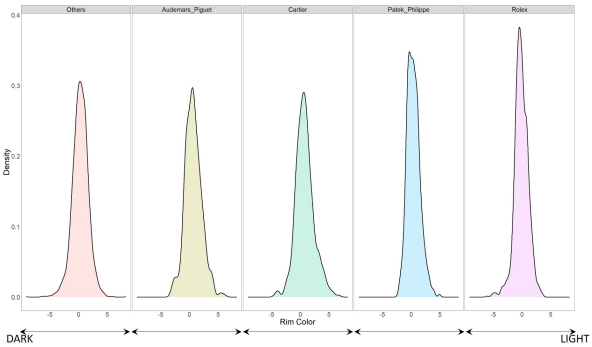
(b) Dial Color



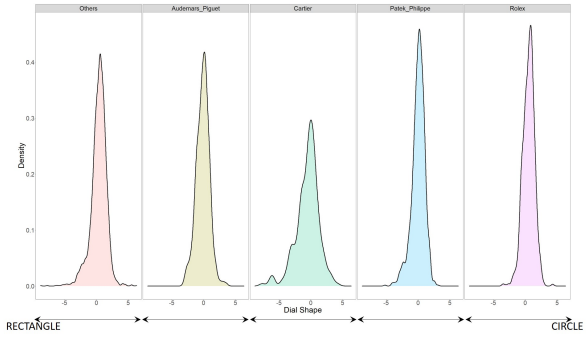
(c) Strap Color



(d) Rim Color



(e) Dial Shape



(f) Knob Size

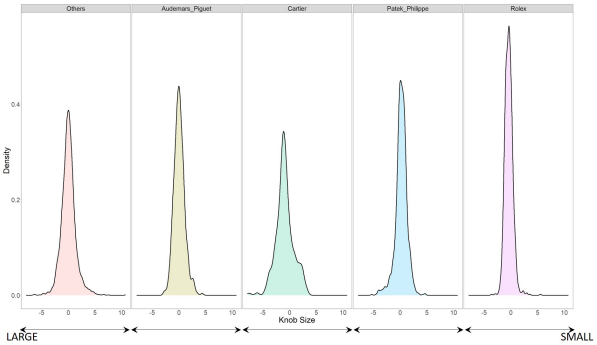


Figure A.7: Survey Question to Validate Interpretability

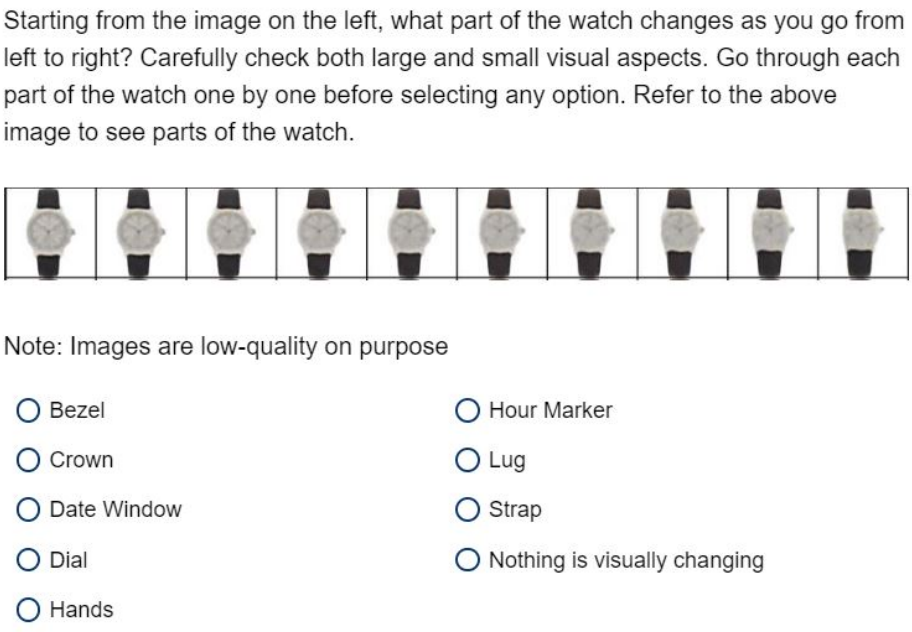


Figure A.8: Survey Question to Validate Quantification

F. Survey Screenshots

Figure A.7 and Figure A.8 are screenshots of the interpretability and the quantification survey respectively. Figure A.9 are screenshots of the conjoint survey.

Note: The terminology in the surveys specifies "Bezel" which is the same as "Rim" and "Crown" which is also known as "Knob."

Figure A.9: Conjoint Survey Screenshots