Automatic Discovery and Generation of Visual Design Characteristics: Application to Visual Conjoint

Ankit Sisodia, Alex Burnap and Vineet Kumar[*]

Spring 2023

Abstract

Visual design characteristics of products play an important role in consumer preferences for many categories. However, characterization of quantification of visual design is a challenging problem. We provide a method 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 in real world applications. Our method does not require such supervision, and instead uses readily available structured product characteristics as supervisory signals to enable disentanglement. No prior knowledge on design characteristics is required, yet we are able to discover human interpretable and statistically independent characteristics. We apply this method to automatically discover visual product characteristics of watches, and discover 6 human interpretable visual characteristics providing a disentangled representation. We conduct visual conjoint analysis to obtain consumer preferences over visual characteristics. Our generative method is also able to create novel visual designs that correspond to ideal points of different consumer segments.

Keywords: discovery of product characteristics, deep learning, disentanglement

[*] Ankit Sisodia is a Ph.D. student at Yale School of Management. email: ankit.sisodia@yale.edu. Alex Burnap is an Assistant Professor of Marketing at Yale School of Management. email: alex.burnap@yale.edu. Vineet Kumar is an Associate Professor of Marketing at Yale School of Management. email: vineet.kumar@yale.edu. We thank participants at the AIM Conference (2022) at USC Marshall and INFORMS Marketing Science Conference (2022). We also thank participants in seminars at National University of Singapore, Nanyang Technological University, Purdue University, Santa Clara University, Washington University at St. Louis, and Yale University. We thank Raghuram Iyengar, Artem Timoshenko, Olivier Toubia for their comments.

INTRODUCTION

Market demand for products is impacted by the underlying product characteristics (Lancaster 1966). For example, in the automobile market, structured characteristics like horsepower and fuel efficiency impact consumer choices. Similarly, visual design is also a significant driver of consumer purchase in automobiles and other product categories like apparel and home furnishings (Simonson and Schmitt 1997; Bloch 1995).

However, clearly articulating *why* a product looks appealing and what aspects contribute to such appeal is challenging for consumers, practitioners, and researchers alike (Berlyne 1973). Methods for modeling the visual characteristics of products require significant product knowledge, expertise and judgment. The expert must *manually* define which visual characteristics represent a product's visual form. Even after defining visual characteristics, the question remains of how to quantify these characteristics. To our knowledge, there is no extant research in marketing that automatically characterizes and quantifies different aspects of visual product design in a human interpretable manner.

Research Goal: Our research instead aims to *automatically discover* (extract) and *quantify* multiple independent and interpretable visual characteristics directly from unstructured product image data, with the aid of structured product data. Our method also *generates* novel visual designs across the span of interpretable visual characteristics. Both the discovery of interpretable visual characteristics and the generation of novel visual designs can then be used in a variety of marketing applications.[1]

We demonstrate how to use these quantified visual characteristics in an application of *visual conjoint analysis*. We obtain consumer preferences over these visual characteristics, along with demographic and psychographic variables to use in segmentation. We then generate novel visual designs targeted to the "ideal points" of distinct customer segments.

[1] Our focus here is not on discovering *outlier* characteristics that are particularly *surprising* to humans, especially experts. Rather it is to identify aspects apparent from visual product images.

Our method of obtaining interpretable visual characteristics is valuable to researchers interested in understanding consumer preferences, demand responses, and firms' strategic choices in the visual domain. Discovery and quantification of these characteristics is a first step in enabling analysis of these issues. The framework here also has value to practitioners like product managers, who can use the ability to generate visual designs to evaluate prototypes, or seek to differentiate products in terms of visual design.

Approaches to Quantifying Visual Design: Obtaining quantified interpretable characteristics *manually* from humans leads to multiple challenges. First, quantifying the levels of these characteristics for each product in the market is costly in terms of resouces, time and effort and not very scalable.[2] 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 they differ significantly from one another. Moreover, even if humans could identify and quantify the levels of visual characteristics for existing products, this manual approach would *not* be able to generate counterfactual visual design, which are required for visual conjoint analysis.

On the other hand, existing methods like Principal Components Analysis (PCA) and Multidimensional Scaling (MDS) are scalable and *automatically* obtain a representation from image data. However, these methods offer no interpretability of characteristics, since they are focused only on obtaining orthogonal representations that capture the most variation in the data. While PCA and MDS have been widely-used in marketing to reduce data dimensionality for managerial interpretation, these methods are also not well suited to capturing complex nonlinear relationships in unstructured data like images (Linting et al. 2007). See the Web Appendix for a comparison of select methods for obtaining a low-dimensional representation. Developing a methodology that is *both automatic and obtains human interpretable characteristics* is what makes our approach unique.

[2] With K characteristics for each of N products we would need a total of $(N \times K)$ 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 for just one product category.

Methodological Basis: We build upon the disentanglement stream of literature in representation learning, an area of deep learning, with our primary goal of inferring interpretable representations from image data. According to Locatello et al. (2019), "the key idea behind this [disentanglement learning] model is that the high-dimensional data x can be explained by the substantially lower dimensional and semantically meaningful [to humans] latent variable z."

Disentanglement learning builds upon variational autoencoders (VAE), which includes an encoder neural net and decoder neural net, both of which are parameterized 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 of distributions for each discovered characteristic. 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 (e.g. a set of images of various watches) is much lower than the dimensionality of the raw images.[3]

Disentanglement learning using only images with unsupervised learning has theoretical limitations (Locatello et al. 2019). To remedy this issue, recent research recommends using supervised learning with "ground truth" visual characteristics for each data point (i.e., product image) as a supervisory signal (or label) (Locatello et al. 2020).[4] However, in our case, and in many practical marketing and business applications, these "ground truth" visual characteristics *are unknown and exactly what we would like to learn*. Thus, we cannot use standard methods suggested in machine learning.

Our methodology aims to overcome this issue by building upon deep learning models of disentanglement. We show that supervised disentanglement, with structured product characteristics as

[3] 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.

[4] Specifically, the prediction problem is to predict the ground truth visual characteristics using the discovered characteristics in the latent representation. For real-world data, researchers first decide a set of visual characteristics to obtain annotations for and then, ask human coders to quantify the "ground truth" labels corresponding to the chosen set of visual characteristics. For example, in a dataset of celebrity faces, human annotations were created for a wide variety of visual characteristics including eyeglasses, shape of face, wavy hair, mustache etc (Liu et al. 2015). Similarly, in a dataset of 3D chairs, human annotations like object pose and scale were created (Aubry et al. 2014). Broadly, this manual approach requires obtaining annotations from multiple human coders and reconciling these noisy measures to create "ground truth" labels.

signals (labels), readily-available in typical marketing datasets, can both address known theoretical limitations and improve disentanglement performance.

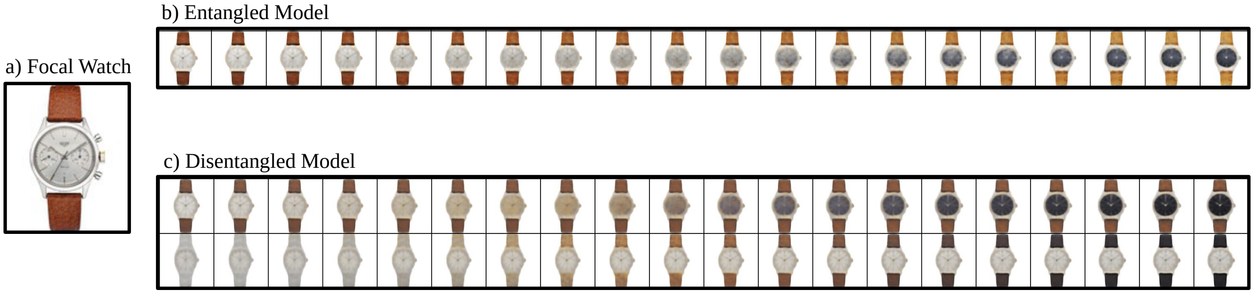
Advantages: Our approach has a number of practical advantages. First, the method is designed to work with unstructured *image data* that would be practically obtainable in real managerial settings. It does not require labeled data on visual characteristics, and is designed to leverage typically available structured characteristics. Second, the researcher does not define the (unknown) visual characteristics in advance, and does not even need to specify the number of such characteristics that must be discovered. Third, our method is also flexible with regard to image quality, and works with low resolution images (like 128x128 pixels). Finally, our approach can be applied in a scalable manner across product categories using the same architecture with minimal hyperparameter tuning.

Application and Results: We apply our proposed method on watches as the primary product category, and also test the method using sneakers. Recall that our goal is to automatically obtain interpretable visual characteristics. Our disentanglement method automatically discovers and quantifies 6 *interpretable* visual characteristics of the watches. These discovered characteristics correspond to 'dial size', 'dial color', 'strap color', 'dial shape', 'knob (crown) size', and 'rim (bezel) color'.[5]

Disentanglement aims at identifying multi-dimensional latent representation in the image data, where each dimension maps one-to-one with a human interpretable characteristics (Bengio, Courville, and Vincent 2013). With a disentangled representation, a change in one latent dimension would result in a change to only one human interpretable characteristic. Entanglement (in contrast) implies that a change in the level across one discovered latent dimension impacts *multiple* human interpretable characteristics. Figure 1 illustrates the difference between disentangled and entangled representations.

[5] The visual depiction and description of the parts of a watch are available at the website: https://bespokeunit.com/watches/watch-parts-guide/

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.

Evaluation: We evaluate our disentanglement method relative to benchmark alternatives in a number of ways. First, we measure *disentanglement performance* relative to unsupervised disentanglement. We find that across product categories (watches and sneakers), having access to supervisory signals based on product characteristics improves disentanglement. To demonstrate this, we use a metric called Unsupervised Disentanglement Ranking (UDR) from the machine learning literature (Duan et al. 2020).

Next, to assess human *interpretability*, we conduct a survey with 99 individuals from the US using Prolific. We generate watch visual designs by varying one dimension of the latent representation at a time. We asked respondents to determine if changes along that specific dimension generates watches that vary along a specific human interpretable visual characteristic. We find that on average, 86% of respondents agree on the corresponding visual characteristic of the product that is changing. The high level of agreement is consistent with disentanglement leading to human interpretable visual characteristics.

To validate the *quantification of characteristics*, we also examine whether the quantified level of the characteristic is human interpretable. We test this aspect by showing respondents 2 pairs of watches (randomly drawn from the disentangled representations), where each pair varies on the same visual characteristic (e.g. dial color) but in different degrees. We then ask 300 respondents to evaluate which pair is visually "more similar." We measure the divergence between the human responses and the algorithm's quantification of the characteristics, to evaluate whether humans and algorithms view the *semantic meaning of the quantification* of characteristics similarly. We find that human respondents and the algorithm match 85% of the time along this similarity metric, reflecting that the algorithm's quantification is human interpretable.

Finally, we obtain a higher predictive accuracy for consumer choices over generated visual designs as detailed in the visual conjoint application below.

Visual Conjoint Application: We use the obtained visual characteristics in conjoint analysis. Consumers are presented with multiple alternatives that span the visual design space. A crucial benefit of our method is the ability to generate different designs by varying one or more visual characteristics at a time. Existing approaches to create visual designs are costly to implement limiting the number of alternatives that can be used in any visual conjoint analysis study (Sylcott, Orsborn, and Cagan 2016). In contrast, we can controllably *generate* a large dataset of counterfactual design at scale, in order to gain a deeper understanding of which designs are preferred by consumers and attribute their preference to each of the visual characteristics.

We estimate individual-level preferences using a Hierarchical Bayesian (HB) model that accounts for consumer heterogeneity. We evaluate how well the proposed approach predicts consumers' choice preferences against a benchmark pretrained deep learning model based on ResNet50 fine-tuned on the watch image data.

Our results show that consumers have preferences over human interpretable visual characteristics discovered by our method, and that these characteristics can be used to quantify and predict consumer choice. Specifically, we find the HB model with visual characteristics discovered by our method achieves a higher predictive accuracy (72.33%) than the benchmark uninterpretable pretrained deep learning model (68.31%).

We next show how our method can be used to automatically generate novel and targeted product designs for consumer segments. Specifically, we aim to identify two segments of consumers. We obtain segment-level "ideal points" over the 6 discovered visual characteristics. We then use the generative capability of the method to generate novel designs corresponding to each segment's most preferred watch design.

We test the generality of the approach by using the same model architecture (with a couple of hyperparameter updates) in a separate and unrelated product category of sneakers. We find again that a supervisory signal (price) achieves significantly higher disentanglement performance (UDR) than the unsupervised approach.

Contribution: Our paper contributes on the issue of using supervision for disentanglement by using structured product characteristics. Our approach is quite different from the ML approach of using ground truth. Clearly, ground truth signals capture exactly the true underlying data generating process for the product images separately for each visual characteristic and each product image. Thus, adding ground truth as a supervisory signal would always enhance disentanglement. However, the critical challenge is that ground truth is not available in typical business applications. We evaluate different combinations of signals and find that using multiple signals can be beneficial for disentanglement. We also caution that supervised learning may not be a panacea and that the choice of supervisory signal(s) is important, with some choices leading to worse disentanglement. A key aspect of our methodology is the ability to generate novel designs based on the interpretable visual characteristics discovered, which provides the foundation to conducting visual conjoint analysis for products based on these characteristics.

LITERATURE REVIEW

Visual design is instrumental in shaping consumer preferences, perceptions of value, and experiences across a range of categories. As Bloch, Brunel, and Arnold (2003) says, “Vegetable peelers, wireless phones, car-washing buckets, and lawn tractors are all being designed with attention to the aesthetic value of their appearance.” Brands follow a process of incorporating visual design including identifying and selecting visual elements and implementing them to impact consumer experiences (Simonson and Schmitt 1997). Other research has found a positive relationship between aesthetic appeal and usability (Tractinsky, Katz, and Ikar 2000). It is therefore important to understand the value of visual design.

However, it is quite challenging to characterize and therefore study visual design from a quantitative perspective. As Orsborn, Cagan, and Boatwright (2009) say, "... possibly even more challenging, user feedback requires objective measurement and quantification of aesthetics and aesthetic preference." Here, the authors identify and choose 7 specific visual design characteristics for automobiles (specifically SUVs) and then quantify these characteristics by using the distance between various physical components present in their car design specifications. Another study by Landwehr, Labroo, and Herrmann (2011) photographed the frontal designs of car models and then used morphing software to create morphs for each car model by identifying feature points that represent the key components of each design. Liu et al. (2017) also used this approach to study the impact of product appearance on demand. Broadly, these approaches require human experts to both identify and quantify the visual characteristics. In contrast, our approach obtains and quantifies the visual characteristics automatically using disentanglement learning.

Representation Learning and Disentanglement Representation learning is a sub-field of machine learning that posits that the data generating process for real-world high-dimensional data arises from low-dimensional factors. According to Bengio, Courville, and Vincent (2013), "learning representations of the data that make it easier to extract useful information when building classifiers or other predictors." The literature has focused on the properties and the value of different representations for different feature extraction and prediction applications. Representation learning has found success in a wide variety of applications such as natural language processing (Liu, Lin, and Sun 2020), speech recognition (Conneau et al. 2020), causal learning (Schölkopf et al. 2021) etc.

Our work builds on a stream of literature in representation learning known as *disentangled* representation learning, which aims to separate distinct informative factors of variation in the data (Bengio, Courville, and Vincent 2013). Consider the dataset of 2D objects dSprites (Higgins et al. 2017). Each image in this data shows an object of a specific shape, size and color at a specific location in the image. Across images, we can see different possible combinations of these visual characteristics. The objective of disentanglement is to separate out these independent factors of variation to obtain object shape, position, size, and color as the 4 latent dimensions discovered by the disentanglement model. The advantage of disentanglement is that, even when the dimensionality of the latent space is increased to a large number, it will only discover these true factors of variation (shape, size, color and position).

METHODOLOGY

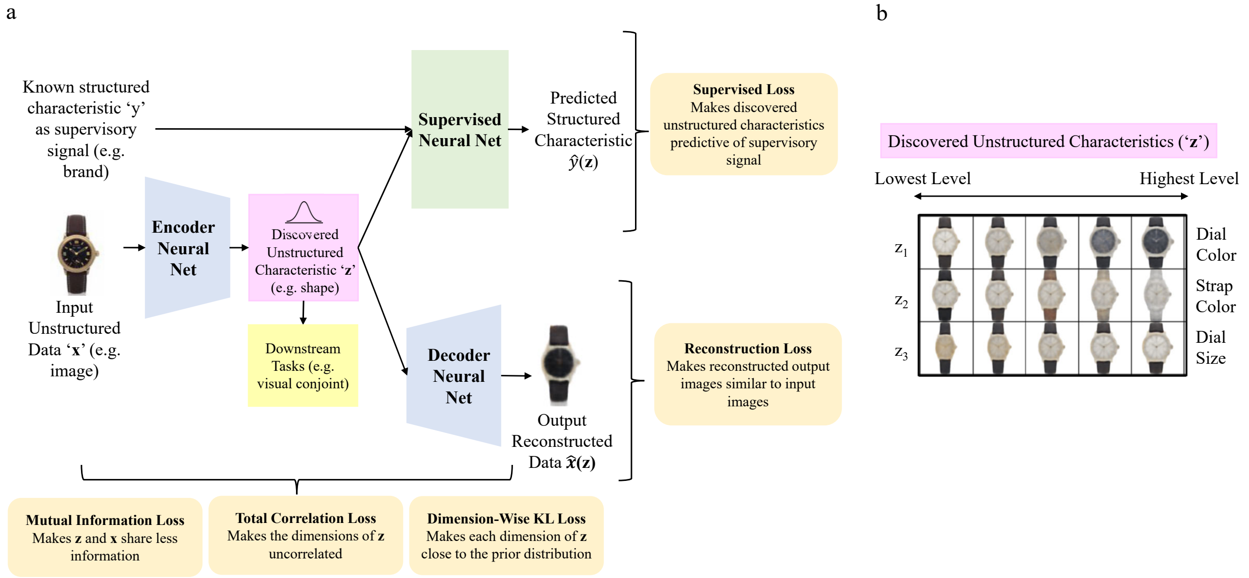
Our proposed approach builds on recent advances in disentangled representation learning, a stream of machine learning focused on learning lower-dimensional representations of high-dimensional data. Most disentanglement methods are built on deep generative models such as variational autoencoders (VAE) (Kingma and Welling 2014) and generative adversarial networks (GAN) (Goodfellow et al. 2020).

Our model builds upon a VAE designed for disentanglement, which belongs to representation learning methods. Representation learning transforms the data into a more informative potentially lower-dimensional format to extract meaningful features from raw data for use in predictive modeling or other tasks (Bengio, Courville, and Vincent 2013). Disentanglement refers to the process of decomposing complex data into independent, interpretable factors in order to better capture the true underlying relationships.[6]

Our method is illustrated in the schematic depicted in Figure 2. The model *encodes* visual data to discover a low-dimensional latent space of visual characteristics that are independent and human interpretable. The model then *decodes* the discovered visual characteristics to reconstruct visual representation of the input images. The model also *predicts* a supervisory signal (e.g., typical marketing structured data such as brand) from the discovered visual characteristics. The model minimizes the weighted sum of 5 different type of losses — reconstruction loss, mutual

[6] Burgess et al. (2017) describes this in more detail: “A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors (Bengio, Courville, and Vincent 2013). For example, a model trained on a dataset of 3D objects might learn independent latent units sensitive to single independent data generative factors, such as object identity, position, scale, lighting or colour, similar to an inverse graphics model (Kulkarni et al. 2015). A disentangled representation is therefore factorised and often interpretable, whereby different independent latent units learn to encode different independent ground-truth generative factors of variation in the data.”

Figure 2: Schematic of Proposed Approach


a
Known structured characteristic 'y' as supervisory signal (e.g. brand)
Supervised Neural Net
Predicted Structured Characteristic $\hat{y}(\mathbf{z})$
Supervised Loss
Makes discovered unstructured characteristics predictive of supervisory signal
Input Unstructured Data 'x' (e.g. image)
Encoder Neural Net
Discovered Unstructured Characteristic 'z' (e.g. shape)
Downstream Tasks (e.g. visual conjoint)
Decoder Neural Net
Output Reconstructed Data $\hat{\mathbf{x}}(\mathbf{z})$
Reconstruction Loss
Makes reconstructed output images similar to input images
Mutual Information Loss
Makes z and x share less information
Total Correlation Loss
Makes the dimensions of z uncorrelated
Dimension-Wise KL Loss
Makes each dimension of z close to the prior distribution
b
Discovered Unstructured Characteristics ('z')
Lowest Level
Highest Level
z_1
z_2
z_3
Dial Color
Strap Color
Dial Size

Notes: **a**: The encoder neural net maps an input image into low-dimensional visual 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.

information loss, total correlation loss, dimension-wise Kullbeck-Leibler (KL) loss and supervised loss. Note that the supervisory signal can be just one product characteristic from structured data or a combination of product characteristics. We detail the notation used here in Table 1.

Table 1: Table of Notation for Disentanglement Model

Symbol	**Category**	**Meaning**
$\mathbf{x}$	Input Data	Product image
$\mathbf{y}$	Input Data	Supervisory signal(s)
$\widehat{\mathbf{x}}$	Output Data	Reconstructed image
$\widehat{\mathbf{y}}$	Output Data	Predicted Supervisory Signal(s)
$\mathbf{z}$	Latent Space	Visual characteristic vector
$\mathbf{z}_{\text{inf}}$	Subset of Latent Space	Informative visual characteristic
$p(\mathbf{z})$	Model	Prior distribution
$p_\theta(\mathbf{x}\|\mathbf{z})$	Decoder Neural Net	Conditional Probability of Generating Image Data given Latent Space
$q_\phi(\mathbf{z}\|\mathbf{x})$	Encoder Neural Net	Conditional Probability of Latent Space given Image Data
$p_w(\mathbf{y}\|\mathbf{z})$	Supervisory Neural Net	Conditional Probability of Supervisory Signal given Latent Space
θ	Weights of Neural Net	Decoder's parameters
ϕ	Weights of Neural Net	Encoder's parameters
w	Weights of Neural Net	Supervisory Net's parameters
$\mathbf{E}_{q_\phi(\mathbf{z}\|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\|\mathbf{z})\right]$	Loss Function	Reconstruction Loss
$I_q(\mathbf{z},\mathbf{x})$	Loss Function	Mutual Information Loss
$KL\left[q(\mathbf{z})\|\|\prod_{j=1}^{J} q(z_j)\right]$	Loss Function	Total Correlation Loss
$\sum_{j=1}^{J} KL\left[q(z_j)\|\|p(z_j)\right]$	Loss Function	Dimension KL Divergence Loss
$P(\hat{y}(\mathbf{z}), y)$	Loss Function	Supervised Loss
$\mathcal{L}(\theta,\phi,\beta;\mathbf{x},\mathbf{z})$	Loss Function	Total Loss
J	Hyperparameter	Dimensionality of latent space
α	Hyperparameter	Weight on Mutual Information Loss
β	Hyperparameter	Weight on Total Correlation Loss
γ	Hyperparameter	Weight on Dimension KL Divergence Loss
δ	Hyperparameter	Weight on Supervised Loss

Model: Supervised Variational Autoencoder with Disentanglement Losses

We first describe a variational autoencoder (VAE) and subsequently describe how it is extended with disentanglement constraints and supervision using structured data. We denote the observed dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ where the i-th observation is a high-dimensional product image $\mathbf{x}_i$ and its corresponding vector of supervised signals $\mathbf{y}_i$. The VAE uses 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 number of characteristics to be discovered (or the size of the latent space). In the second step, the original product image $\mathbf{x}_i$ is reconstructed as $\widehat{\mathbf{x}}_i$ using the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z}) = f_1(\mathbf{x}; \mathbf{z}, \theta)$. The distribution $f_1(\mathbf{x}; \mathbf{z}, \theta)$ is specified as 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 $\mathbf{y}_i$ is predicted from the conditional distribution $p_w(\mathbf{y}|\mathbf{z}) = f_2(\mathbf{y}; \mathbf{z}, \mathbf{w})$, where $f_2(\mathbf{y}; \mathbf{z}, \mathbf{w})$ is a function formed by non-linear transformation, with parameters $\mathbf{w}$, of latent (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 $p_\mathbf{w}(\mathbf{y}|\mathbf{z})$ 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 specified as $\log q_\phi(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$, 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 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.

$$\underbrace{L(\theta,\phi,\mathbf{w};\mathbf{x},\mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction Loss}} + \underbrace{KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\text{Regularizer Term}} \tag{1}$$

To learn disentangled representations, the β-VAE model (Higgins et al. 2017) extends Equation (1) by imposing a heavier penalty on the regularizer term using an adjustable hyperparameer $\beta > 1$.[7] Intuitively, β-VAE uses the hyperparameter β to sacrifice reconstruction accuracy in order to learn more disentangled representations. This framework is adapted and further extended 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 (4).

$$\underbrace{KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\substack{\text{Regularizer Term}\\ \text{of Total Loss}}} = \underbrace{I_q(\mathbf{z},\mathbf{x})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} + \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}} + \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}} \tag{4}$$

Finally, we add a supervised loss term to enforce the discovered characteristics to help predict the supervisory signal(s) $\mathbf{y}$ in Equation (5). This enables us to study whether using typical structured data (e.g., 'brand') with a supervised model helps improve disentanglement, and to compare

[7] 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 (2) where ϵ specifies the strength of the applied constraint.

$$\max_{\theta,\phi} \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \text{ subject to } KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right] < \epsilon \tag{2}$$

We can re-write Equation (2) 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 (3).

$$\mathcal{L}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) \geq \mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) = \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta(KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]) \tag{3}$$

supervised versus unsupervised disentanglement.

$$\underbrace{L(\theta,\phi,\mathbf{w});\mathbf{x},\mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log p_{\theta}(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction}\\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z},\mathbf{x})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}} + \gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}),\mathbf{y})}_{\substack{\text{Supervised}\\ \text{Loss}}} \tag{5}$$

Our model's total loss is comprised of five loss terms weighted using hyperparameters, $(\alpha,\beta,\gamma,\delta)$. Adjusting these hyperparameters critically affects disentanglement performance by adjusting the relative weight of each of the five loss terms, and we detail the intuition for these loss terms below:[8]

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

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 visual characteristic $\mathbf{z}$ and the product image $\mathbf{x}$. From an information-theoretic perspective (Achille and Soatto 2018), penalizing this term reduces the amount of information about $\mathbf{x}$ stored in $\mathbf{z}$. The information needs to be sufficient to reconstruct the data while avoiding storing nuisance information, minimizing copying of the input data. A low α would result in $\mathbf{z}$ storing nuisance information, whereas a high α could result in loss of sufficient information needed for reconstruction.

Total Correlation Loss: The total correlation loss, $KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]$, 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

[8] 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 (4). 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 $\Omega = \{\beta\}$.

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.

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 latent dimension to not deviate from the prior (e.g., Gaussian). A high weight on this term reduces the number of discovered visual characteristics. This term also promotes continuity in the latent space, which allows generation from a smooth and compact region of latent space.

Supervised Loss: Penalizing the supervised loss $P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})$, where $\hat{\mathbf{y}}(\mathbf{z}) \sim p_{\mathbf{w}}(\mathbf{y}|\mathbf{z})$ prioritizes the discovered visual characteristics $\mathbf{z}$ to obtain high accuracy in predicting $\mathbf{y}$. 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. When two or more signals are combined, we discretize the continuous signals and combine them with other discrete signals (if any) and use cross-entropy loss.

Supervised Disentanglement vs Unsupervised Disentanglement

A key issue we examine in this work is whether structured product characteristics typically found in marketing contexts (e.g., brand) can be used as supervisory signals to improve disentanglement, and thus our ability to discover human interpretable visual characteristics. Locatello et al. (2019) showed that in the absence of a supervisory signal, disentangled representations are probabilistically equivalent to entangled representations. This finding implies that it is not possible to obtain a unique disentangled representation using an unsupervised approach. 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 is not suitable for our goal of visual characteristic discovery for several reasons. First, needing ground truth labels of the characteristics conflicts with our goals as these labels are what we are trying to discover in the first place. Second, if the approach requires humans to (even partially) label characteristics, then the approach is not fully automated.

Our research here instead posits that structured product characteristics and price might have information that correlates with visual characteristics, and we therefore use them as supervisory signals. Therefore, our method does not require access to ground truth characteristics.

Why might structured characteristics serve as good supervisory signals? Consider why specific structured product characteristics might work to supervise visual characteristics. Typical structured characteristics commonly available in marketing data include brand, material, performance characteristics and price. First, consider a characteristic like material, e.g. silver that provides a certain visual look to a product. Material more broadly is known to significantly affect visual appearance and consumer perceptions (Fleming 2014). Second, a product characteristic like brand is likely to strongly impact visual look of a product. Consider, for instance the distinct look of a Mercedes-Benz car or a Louis Vuitton handbag. The signature of the brand design is often visibly present and apparent from the product's appearance to consumers, especially for product categories with visible consumption (Simonson and Schmitt 1997; Liu, Dzyabura, and Mizik 2020; Ferraro, Kirmani, and Matherly 2013) or luxury brands (Megehee and Spake 2012; Lee, Hur, and Watkins 2018). 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). Third, consider the role of price, which is strictly speaking not a product characteristic, since it can be set by the retailer. However, many brands, especially luxury brands, maintain carefully curated pricing tiers with strong consumer associations.

Evaluating Disentanglement Performance: We investigate the disentanglement performance measured by Unsupervised Disentanglement Ranking (UDR) across all available supervisory signals and the unsupervised approach.

UDR is a metric based on heuristics of good disentanglement, which allows for an automated way to select a model when ground truth is not available (Duan et al. 2020). 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) require access to the ground truth data generating process and are therefore not suitable for our empirical setting.[9]

The UDR metric posits that for a particular dataset and a particular VAE-based disentangled representation learning model, the 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 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. We compare the UDR for each of the set of supervisory signals with the unsupervised approach, and select the combination of supervisory signals that obtains the highest UDR.

The key idea behind UDR 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*.[10] By

[9] Estermann, Marks, and Yanik (2020) details the value of UDR, which we quote below: “There are no labels available for many real-life applications and for some data, generative factors of interest are hard or impossible for humans to annotate. Recently, Duan et al. [8] defined a new, unsupervised heuristic for evaluating the disentanglement performance of models, based on the assumption that models that disentangle well are more likely to be similar to each other than the ones that do not disentangle [16, 17, 18, 19]. They demonstrate that this Unsupervised Disentanglement Ranking (UDR) correlates well with metrics that rely on previously annotated labels across various models and datasets [8].”

[10] Ridgeway and Mozer (2018) and Eastwood and Williams (2018) introduced three properties that should be present in a disentangled representation. Duan et al. (2020) showed that the UDR metric has a high correlation with the metrics that measure these properties on datasets when the ground truth was available.

1 Modularity: One factor of the latent representation is only influenced by a change in one generative factor.

2 Compactness: One generative factor is only sensitive to a change in one latent code dimension

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 latent dimensions to be switched off in the two models.[11]

Model Training, Selection, and Evaluation

Both the supervised and unsupervised disentanglement approaches require model training (i.e., how model parameters are learned), model selection (i.e., how model hyperparameters are chosen), and model evaluation (i.e., the selected model's disentanglement performance). However supervised and unsupervised approaches require different model training and selection steps, while having the same evaluation step so we can compare them appropriately.

Model Training and Selection: We divide the dataset into a training dataset for learning disentangled representations, a validation dataset for model selection and a test dataset in the ratio 90:5:5. To avoid data leakage, each product was present only in one of the above subsets. Figure 3 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,

3 Explicitness: The amount of information captured by the latent code representation about the factors of variation.

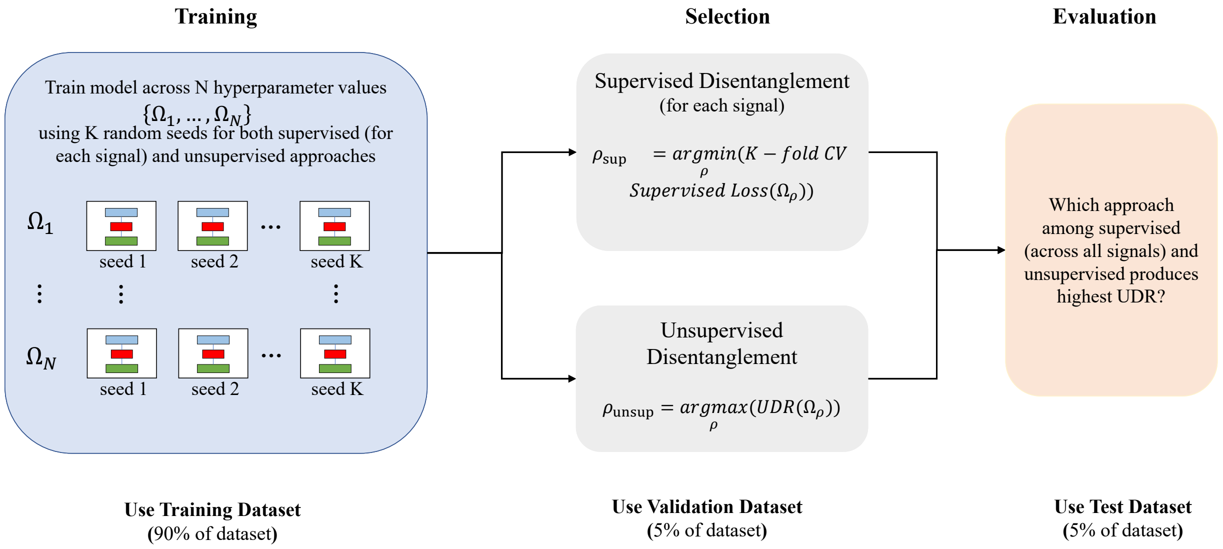
[11] For each trained model, we perform $\kappa = 45$ pairwise comparisons with all other models trained with the same β level and δ level but with different seed levels and calculated the UDR_{ij}, where i and j index the two models. Each UDR_{ij} score is calculated by computing the similarity matrix 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 (6).

$$UDR_{ij} = \frac{1}{d_a + d_b}\left[\Sigma_b \frac{r_a^2 I_{KL}(b)}{\Sigma_a R(a,b)} + \Sigma_a \frac{r_b^2 I_{KL}(a)}{\Sigma_b R(a,b)}\right] \tag{6}$$

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. 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. Variation across an uninformative characteristic would produce little to zero visual change in the image.

and uses a structured watch characteristics (e.g., brand) as the supervisory signal to the model. We fix the hyperparameters based on suggestions in the literature (Locatello et al. 2020; Chen et al. 2018). The number of latent codes J represents the number of characteristics that our model aims to find. A very low J might miss important characteristics, whereas a high value of J might lead to more uninformative characteristics. We choose $J = 20$ to balance these considerations, based on our empirical setting. We need to tune other hyperparameters including learning rate, batch size and number of training steps or epochs.[12]

Figure 3: Model Training, Selection, & Evaluation


Training
Selection
Evaluation
Train model across N hyperparameter values
$\{\Omega_1, \ldots, \Omega_N\}$
using K random seeds for both supervised (for each signal) and unsupervised approaches
Ω_1
seed 1
seed 2
seed K
Ω_N
seed 1
seed 2
seed K
Supervised Disentanglement
(for each signal)
$\rho_{\text{sup}} = \underset{\rho}{argmin}(K - fold\ CV\ Supervised\ Loss(\Omega_\rho))$
Unsupervised Disentanglement
$\rho_{\text{unsup}} = \underset{\rho}{argmax}(UDR(\Omega_\rho))$
Which approach among supervised (across all signals) and unsupervised produces highest UDR?
Use Training Dataset
(90% of dataset)
Use Validation Dataset
(5% of dataset)
Use Test Dataset
(5% of dataset)

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 using the UDR metric.

In order to select the model with appropriate hyperparameters, we sweep over levels of hyperparameters corresponding to β (weight on the total correlation loss term) and δ (weight on the

[12] Considerations for tuning hyperparameters detailed next is common to all deep learning models. 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. Specifically, we choose the number of random seeds used as 1 to 10; Adam optimizer with learning rate 5e-4 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$; batch size as 64; number of epoch as 100.

prediction loss term).[13] In the unsupervised approach $\delta = 0$ by definition.[14] Finally, we retrain the model on the entire training dataset with the selected hyperparameters, and then use the trained model to extract discovered visual characteristics on the test dataset.

Model Evaluation: We compare all supervisory signals along with the unsupervised approach using the UDR metric.

Model Architecture

The model architecture is detailed Figure 4. 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. We use Convolutional Neural Net (CNNs) to construct the encoder neural net, where we stack a sequence of CNN layers 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 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 include fully connected layers to the discovered visual characteristics to create the supervised neural net in order to predict the structured characteristics that serve as labels.

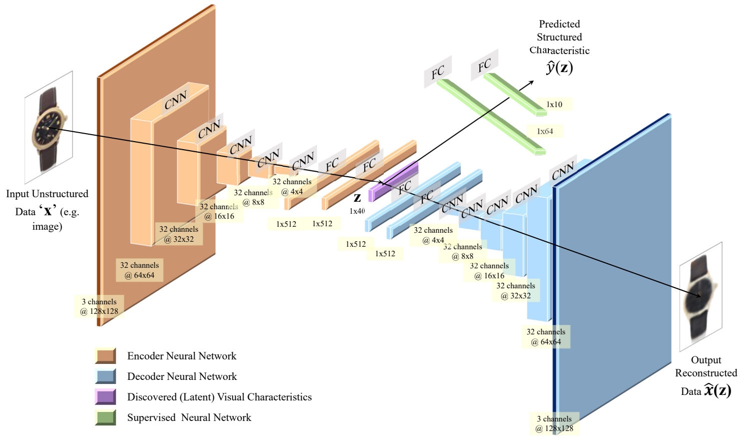
EMPIRICAL APPLICATION

We consider an application of our proposed approach using watches, which have several useful features. Typically, categories with the following characteristics would be appropriate to use with our method. First, we would like 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, we would like a market with a large number of products in order to

[13] For each β and δ level, following Locatello et al. (2020), we select the hyperparameter setting corresponding to the lowest 10-fold cross-validated supervised loss for supervised model selection.

[14] We use Unsupervised Disentanglement Ranking (UDR) for unsupervised model selection.

Figure 4: 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.

train the deep learning algorithm. Third, as with typical marketing data, we need to have a set of structured characteristics appropriately matched up with the images. Finally, for our validation exercise, human respondents need to be familiar with the product category in order to evaluate the interpretability of the discovered visual characteristics.

Data

Our data includes 6,187 watches auctioned at Christie's auction house, spanning the years 2010 – 2020. The data on watches is particularly appropriate for the reasons above. For each auctioned watch in the dataset, we have its image, structured product characteristics, and the hammer price paid at the auction. 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 5 shows a sample of watch images in our dataset. The hammer price (in $1000s) are in inflation-adjusted year 2000 dollars.

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

Figure 5: Sample of Watches Auctioned at Christie's

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.

The model evaluation step compares the set of supervised models and the unsupervised model to evaluate the model with the best disentanglement, or the highest UDR metric. The results of the quantitative evaluations of disentanglement performance are detailed in Table 2.

Table 2: Comparison of Different Supervisory Approaches

Number of Signals	Supervisory Signals	UDR
2	Brand & Material	0.363
2	Circa & Movement	0.357
2	Brand & Circa	0.309
3	Brand, Material & Movement	0.242
2	Circa & Material	0.184
1	Brand	0.135
0	Unsupervised	0.131
1	Material	0.128
2	Material & Movement	0.122
2	Brand & Movement	0.121
1	Movement	0.116
1	Circa	0.112
1	Price	0.076

In this particular dataset of watches, including a combination of signals, i.e. brand and material, was significantly better (UDR = 0.363) than the unsupervised approach (UDR = 0.131). We also note that the unsupervised approach can be better than some supervisory approaches. It is important to understand why adding more variables as supervisory signals might not always benefit disentanglement goals, and might even result in lower disentanglement (or more entanglement).

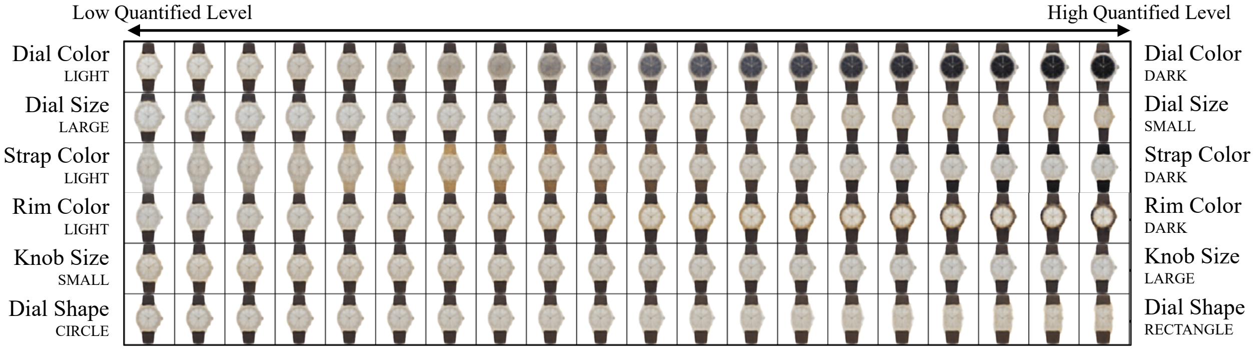
Recall that disentanglement is measured using UDR. This goes somewhat contrary to the intuition that adding more data generally helps prediction, and deserves explanation. While the previous statement is true when adding explanatory variables, and we get potential better prediction, here the situation is different. We are adding an additional dependent variable, so the model weight parameters and even possibly the hyperparameters could be quite different. If the signal is very weakly correlated with the visual characteristics, then we might find that the disentanglement deep net tries to train its weights in order to obtain features that try to predict the chosen supervisory signals. Thus, it is likely to distort from the ground truth in cases when the supervisory signal has very little information about the ground truth. Thus, in theory, a disentanglement model with a supervisory signal might be worse on disentanglement than an unsupervised version, and similarly, adding more signals could well hurt disentanglement. We also observe this in practice with the watches dataset, where some signals (like price) are possibly not as correlated with visual characteristics, and result in lower disentanglement performance than the unsupervised case. Also, using Brand+Material is better than a combination of 3 signals.

The disentanglement literature has assumed ground truth on the visual characteristics as the supervisory signals (Locatello et al. 2020). We use structured characteristics as supervisory signals since 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. However, the specific combinations of signal(s) that would work better is likely to depend on the empirical setting.

Results: Discovered Visual Characteristics

Figure 6 illustrates the output of the disentanglement model with supervisory signals Brand+Material, showing discovered visual characteristics. Each row of the figure demonstrates how the watch design changes based on changes in levels of *one specific* discovered visual characteristic, while keeping all the other characteristics fixed. We only show 6 visual characteristics as the others were found to be uninformative. By uninformative, we mean that traversing along those dimensions

Figure 6: Discovered Visual characteristics



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. Discovered visual characteristics are learned by supervising the characteristics to predict both the brand and material simultaneously.

leads to no visual changes, and the distribution of the latent variable is almost identical to pure Gaussian noise. From ex-post human inspection (by researchers), we observe that we are able to obtain six distinct visual characteristics that are independent as well as human interpretable. These are ‘dial color’, ‘dial size’, ‘strap color’, ‘rim color’, ‘knob size’ and ‘dial shape’.

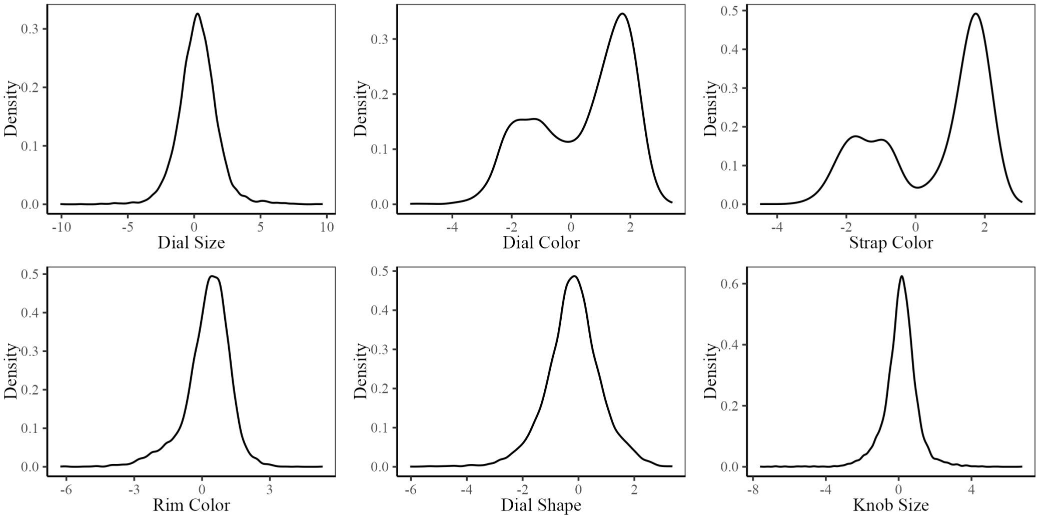
Table 3 details the summary statistics of the visual characteristic levels learned by using the supervisory signal ‘brand’. Figure 7 shows the density plot of these discovered visual characteristics. The method does not enforce any restrictions on the distribution of the visual characteristics of our data, and we observe some deviate significantly from the normal prior (e.g. ‘dial color’). A watch’s ‘dial color’ or ‘strap color’ could come from any one of the mixtures of Gaussian distributions. The density plot shows that the method is able to find a variety of distributions of visual characteristics. Finally, we show that the discovered visual characteristics are highly uncorrelated in Table 4, consistent with the method’s inclusion of a greater weight on the total correlation loss.

We next provide evidence on why Brand+Material served as a good signal. We motivated the use of brand as a supervisory signal with the observation that watches from different brands are likely be visually different. A brand uses visual aesthetics to differentiate itself. Similarly, material is known to significantly affect visual appearance and its perception. These reasons could explain why the visual characteristics would have a different distribution for each brand and material.

Table 3: Summary Statistics of Discovered Visual characteristics (from 'Brand+Material' Signal)

Visual characteristic	Mean	SD	Min	Max
Dial Size	0.28	1.49	-11.08	9.68
Dial Color	0.38	1.59	-5.42	3.42
Strap Color	0.50	1.58	-4.50	3.08
Rim Color	0.25	1.01	-6.27	5.33
Dial Shape	-0.19	0.99	-6.03	3.36
Knob Size	0.11	0.93	-7.61	6.79

Figure 7: Density of Discovered Visual characteristics (from 'Brand+Material' 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.

Table 4: Correlations Between Visual Characteristics

	Dial Size	Dial Color	Strap Color	Rim Color	Dial Shape	Knob Size
Dial Size	1.00	0.17	-0.08	-0.03	-0.02	0.00
Dial Color	0.17	1.00	0.03	-0.00	0.09	-0.02
Strap Color	-0.08	0.03	1.00	-0.11	-0.03	-0.04
Rim Color	-0.03	-0.00	-0.11	1.00	0.09	-0.01
Dial Shape	-0.02	0.09	-0.03	0.09	1.00	0.05
Knob Size	0.00	-0.02	-0.04	-0.01	0.05	1.00

Validation of Discovered Visual Characteristics

We would like to evaluate whether the visual characteristics discovered by the disentanglement model are human interpretable, both qualitatively and quantitatively. We conducted two surveys to validate that humans (a) identify the distinct characteristics and (b) are consistent with our model in their quantitative evaluation.[15] In the first survey, we evaluate the interpretability of the discovered characteristics from visual data. We present respondents with a image showing the different parts of the watch before conducting the survey to help them understand the visually distinct elements of the product.[16]

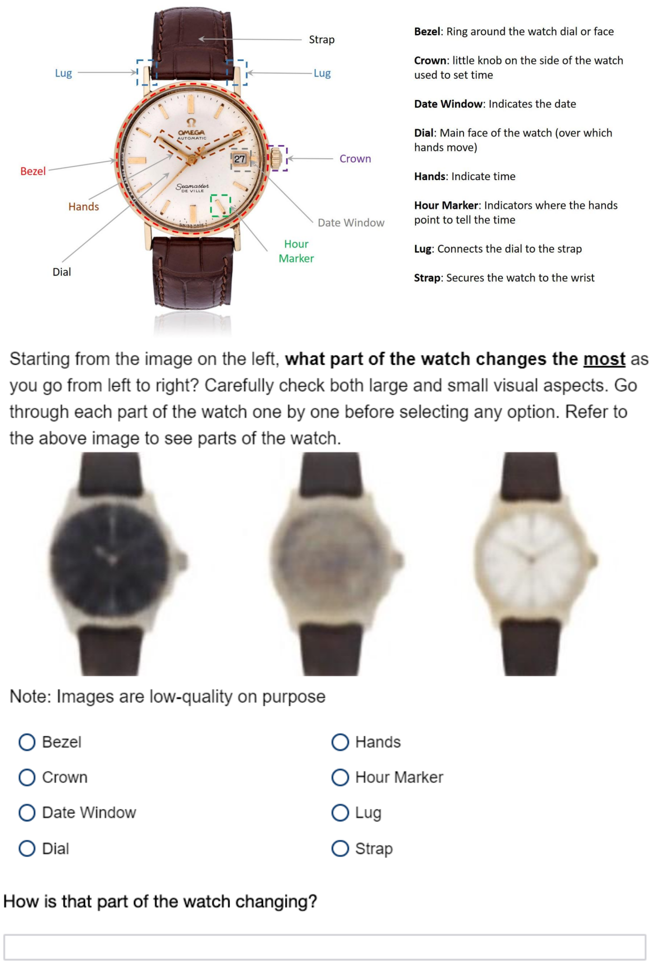
Next, we generated counterfactual images that vary along only one visual characteristic. For example, each watch image (see Figure 8) is generated by fixing all except one focal visual characteristic, and *only changing the level of the focal visual characteristic*. We ask 99 respondents to identify *which part* of the watch is changing as move from left to right, and *how* that part was changing. We find that the average agreement among respondents was 86%, with a range from 73%–96%, despite the low image resolution. In the first column of the Table 5, we report the percentage of respondents of the survey who agree with each other on which part of the watch is changing.

We next examine in a second survey (Figure 9) whether the quantification of the characteristics automatically determined by the method was consistent with human interpretation. We gener-

[15] We choose respondents based in the US who are fluent in English. For both surveys, we employ an attention check.

[16] We obtained the parts of the watch from the URL: `https://bespokeunit.com/watches/watch-parts-guide/`. This was shown in all survey screens.

Figure 8: Survey Question to Validate Interpretability



Bezel: Ring around the watch dial or face

Crown: little knob on the side of the watch used to set time

Date Window: Indicates the date

Dial: Main face of the watch (over which hands move)

Hands: Indicate time

Hour Marker: Indicators where the hands point to tell the time

Lug: Connects the dial to the strap

Strap: Secures the watch to the wrist

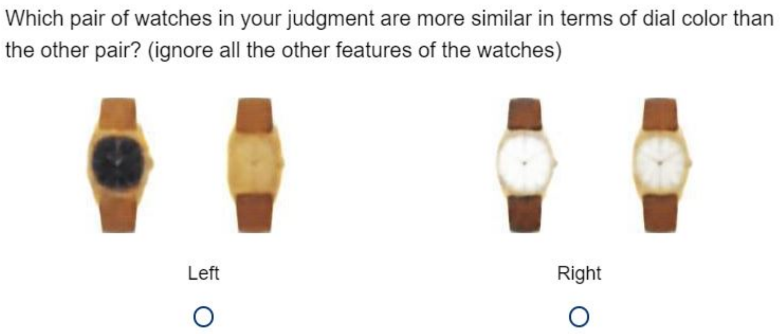
Starting from the image on the left, **what part of the watch changes the <u>most</u>** as you go from left to right? Carefully check both large and small visual aspects. Go through each part of the watch one by one before selecting any option. Refer to the above image to see parts of the watch.

Note: Images are low-quality on purpose

○ Bezel	○ Hands
○ Crown	○ Hour Marker
○ Date Window	○ Lug
○ Dial	○ Strap

How is that part of the watch changing?

Figure 9: Survey Question to Validate Quantification



Which pair of watches in your judgment are more similar in terms of dial color than the other pair? (ignore all the other features of the watches)

Left ○

Right ○

ated several pairs of watch images that differed only along one visual characteristic. We ask 300 respondents to select the pair of watches that are more similar, which represents an ordinal evaluation. We evaluate whether the responses matched with our algorithm's quantification. We find that a strong majority (average of 85%) agree with the algorithm's quantification scale for the visual characteristics, as detailed in the second column of Table 5.

Table 5: Human Interpretation of Visual Characteristics and Quantification

Visual characteristic	Interpretability Survey	Quantification Survey
Dial Size	81%	83%
Dial Color	84%	92%
Strap Color	96%	92%
Rim Color	90%	88%
Dial Shape	91%	68%
Knob Size	73%	85%

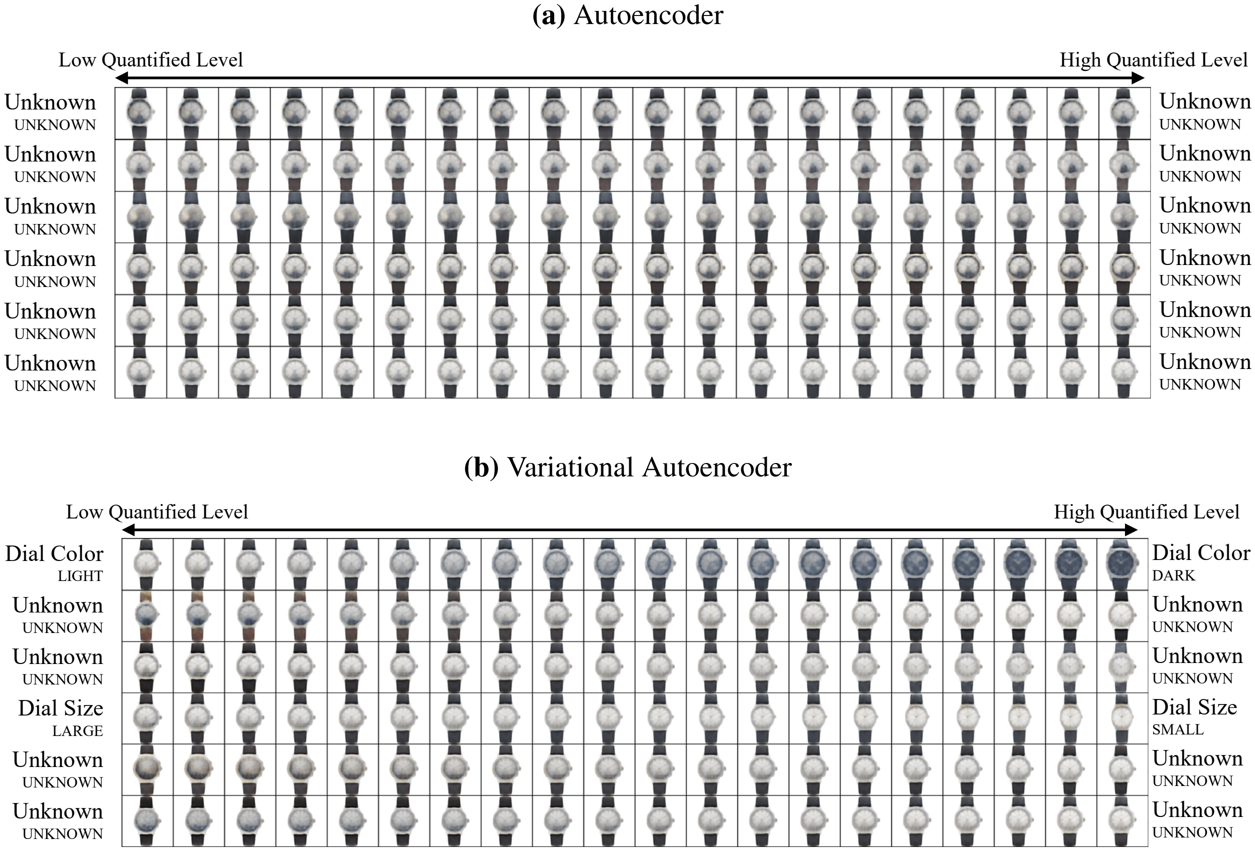
Discovery with Autoencoders and Variational Autoencoders

We obtain the visual characteristics discovered by an an autoencoder (AE) and a variational autoencoder (VAE) to serve as reference to the disentanglement model. We observe that an AE does not recover any useful visual characteristics, the vAE only recovers 2, dial color and dial size.

Figure 10 gives output of discovered visual characteristics from an autoencoder and a variational autoencoder. We show the top six visual characteristics based on the KL divergence value of the difference between the posterior and the Gaussian prior.

We cannot interpret any of the visual characteristics discovered by the AE. Note that these characteristics are not uninformative because their KL divergence is not close to 0. We are able to interpret two characteristics discovered by a VAE: dial color and dial size. However, all the other visual characteristics appear to be entangled. By entangled, we mean that when any one entangled characteristic is kept fixed and other characteristics are changed, the watch image changes in more than one interpretable way. Moreover, we can see that the information about the changing dial color is contained in multiple visual characteristics. This is unlike a disentangled model in which each visual characteristic captures a unique factor of variation.

Figure 10: Discovered Visual characteristics from Different Methods


(a) Autoencoder

(b) Variational Autoencoder

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 Autoencoder. **b**: Discovered visual characteristics learned by Variational Autoencoder.

We applied our method to obtain visual characteristics for the sneakers category using the same model architecture. These results are in the Appendix. We also evaluate that an alternative approach of using SHAP-learned features as an input to the disentanglement model produces fewer number of visual characteristics. These results are also in the Appendix.

MANAGERIAL APPLICATION: VISUAL CONJOINT ANALYSIS

We designed a choice-based conjoint (CBC) analysis survey for eliciting customer preferences for novel visual designs for watches, generated by the disentanglement model. Generated watch designs were created by sampling 3 levels – low, medium, and high – of the posterior distributions of the 6 discovered visual characteristics, resulting in $3^6 = 729$ generated watch designs.

We obtained CBC survey responses from 400 individuals through the Prolific platform. [17] The filtering of respondents resulted in a final set of 253 respondents. The data collected consisted of each filtered respondent's binary choices for the 15 CBC questions, as well as their covariates; namely, demographics and Likert responses to visual appearance.

Conjoint Survey Design

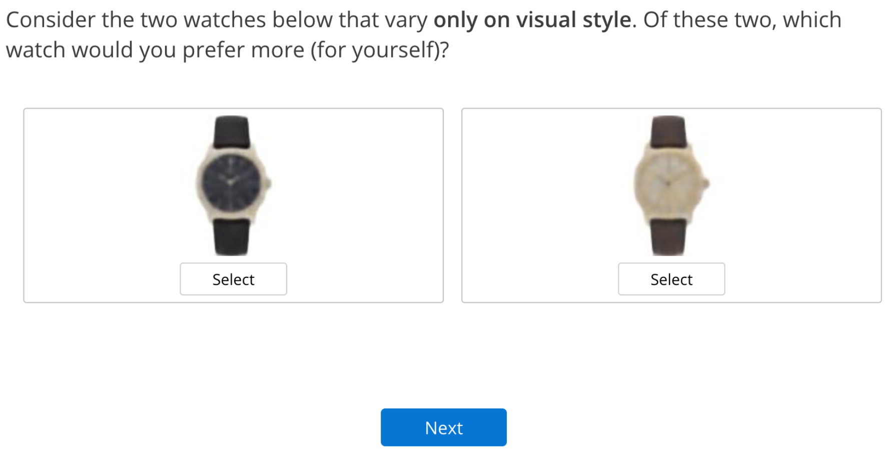
The conjoint survey was designed with 7 survey stages summarized along with their purpose in Table 6. Each CBC question consisted of a binary choice between two watch designs as shown in Figure 11. The CBC design ensured all unique product designs were enumerated while also sampling pairs of product images that spanned the visual attribute space for statistical efficiency (i.e., D-optimality) (McCullough 2002).

Conjoint Model Specification, Estimation, and Evaluation

Model Specification: We specify a Hierarchical Bayesian (HB) model (Lenk et al. 1996) to estimate and infer individual-level preferences elicited from the conjoint survey over the 6 discovered visual characteristics denoted $\mathbf{z}$ ("Dial Color", "Dial Shape", "Strap Color", "Dial Size",

[17] Respondents were filtered post-hoc for a number of reasons: (a) they did not pass the Instructional Manipulation Check (IMC) attention check (Oppenheimer, Meyvis, and Davidenko 2009), (b) they gave inconsistent responses to repeated questions, (c) they did not wear a watch, or (d) they answered "Prefer not to say" for any of the demographic questions.

Figure 11: Example choice-based conjoint (CBC) question in conjoint survey.



Consider the two watches below that vary **only on visual style**. Of these two, which watch would you prefer more (for yourself)?

Table 6: Conjoint Survey Design Elements

Stage	Name	Purpose
1	Introduction	Explain purpose of study and obtain consent.[1]
2	Category Identification	Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to.[2]
3	Instructional Manipulation Check (IMC)	Attention check "trap question" for post-hoc respondent filtering.
4	Choice-Based Conjoint (CBC) Instructions	Explain upcoming conjoint choice question tasks with instructions to choose based only on visual style.[4]
5	"Warm Up" CBC Practice	Help respondents understand the range of watch designs before making real choices.
6	15 CBC questions	Elicit respondent choice of preferred watch design
7	Respondent Information	Obtain demographic and psychographic variables[7]

[1] Respondents were also instructed to be as "consistent" in their choices as possible, with a monetary incentive of $2 for consistency (in addition to $3 for completion).

[2] Respondents saw a set of 4 blurry images for each of the 3 product categories (automobiles, shoes, and watches) similar to the generated watch designs from the disentanglement model. They were then asked for a one word description of the images. We find that greater than 99% of respondents identify the product category depicted in the images. We also used generated watch designs and find that 97% of respondents identify the product category as watches.

[4] Respondents were instructed to choose between two possible watch designs based only on visual style. No other information such as price or other product characteristics were provided.

[7] Respondents demographic variables (e.g., age, gender, income, education) as well as Likert and psychographic questions about how important visual appearance was to the respondent were obtained.

"Knob Size", "Rim Color"). We additionally included 6 respondent covariates denoted $\mathbf{r}$ ("Gender - Male", "Gender - Female", "Age", "Income", "Education", and "Aesthetic Importance").[18]

$$
\begin{aligned}
\mu_\Theta &\sim \mathcal{N}(\mathbf{0}, \sigma_\Theta^2) \\
\mathbf{\Theta} &\sim \mathcal{N}(\mu_\Theta, \mathbf{\Lambda}_\Theta) \\
\mathbf{\Omega}_\beta &\sim \mathrm{LKJ}(\eta) \\
\mathbf{\Lambda}_\beta &= \mathbf{D}(\sigma_\beta)\mathbf{\Omega}_\beta\mathbf{D}(\sigma_\beta) \qquad (7) \\
\beta_i &\sim \mathcal{N}(\mathbf{\Theta}^T\mathbf{r}_i, \mathbf{\Lambda}_\beta) \\
u_i^j &= z_j\beta_i + \epsilon_{ij} \\
y_i^{j,j'} &\sim \mathrm{Bernoulli}(\omega_i(j,j')) \\
\text{where} \quad \omega_i(j,j') &= \frac{\exp(u_i^j)}{\exp(u_i^j) + \exp(u_i^{j'})}
\end{aligned}
$$

where $\mathrm{LKJ}(\eta)$ is a Cholesky factorization of the correlation matrix $\mathbf{\Omega}_\beta$ of the individual "part-worth" preference vector over visual characteristics (Lewandowski, Kurowicka, and Joe 2009). $\mathbf{D}(\cdot)$ denotes a diagonal matrix, $\mathbf{r}_i$ are consumer covariates, u_i^j is the utility customer i gets from watch design j, and ϵ_{ij} is a Gumbel random variable. The Bernoulli probability parameter $\omega_i(j,j')$ is specified by the logit function, and $\{j,j'\}_i$ denotes the set of all pairwise choice comparisons for watches $j,j' \in J$ that customer i chose over in the conjoint survey. Note that σ_Θ^2, $\mathbf{\Lambda}_\Theta$, η are researcher-defined hyperparameters chosen via model selection using prediction accuracy on the validation data split as the evaluation metric.

We tested a variety of parametric HB model specifications including Gaussian mixture priors before settling on a variant of the conventional HB model specification, namely, a unimodal population-level prior, β, over individual-level "part-worth" coefficient vectors, β_i. The mean of the consumer preference "part-worth" vector was accordingly modeled as the inner product be-

[18] These 6 covariates were selected from the full set of respondent covariates for model parsimony via initial correlation analysis and pretesting. Gender covariates were one-hot encoded, while the remaining four covariates were re-coded from the conjoint survey as real values normalized in the range [-1, 1].

tween respondents' covariates and an upper-level model parameter matrix, $\boldsymbol{\Theta}$. We specified the full covariance matrix over the visual attributes, with the prior drawn from a Cholesky factorization of the covariance matrix for numerical stability, and imposed positive semi-definiteness during sampling (Lewandowski, Kurowicka, and Joe 2009). Lastly, we included a third-level prior over $\boldsymbol{\Theta}$ specified as a matrix of Normals to act as a population-level intercept term. The full model specification is given in Equation (7).

Model Estimation and Parameter Posteriors: We estimated posterior distributions of HB model parameters $\{\{\beta_i\}_{i=1}^{N}, \Theta, \mu_\Theta, \boldsymbol{\Lambda}_\beta\}$ with Markov chain Monte Carlo (MCMC) sampling using the No-U-Turn (NUTS) sampler (Hoffman, Gelman et al. 2014).[19] Hyperparameter values for prior distributions were determined by comparing overlap of prior draws with posterior draws, and by using both in-sample and out-of-sample hit rates.

Figure 12: Posterior Distributions of Population-Level Preference Coefficients β

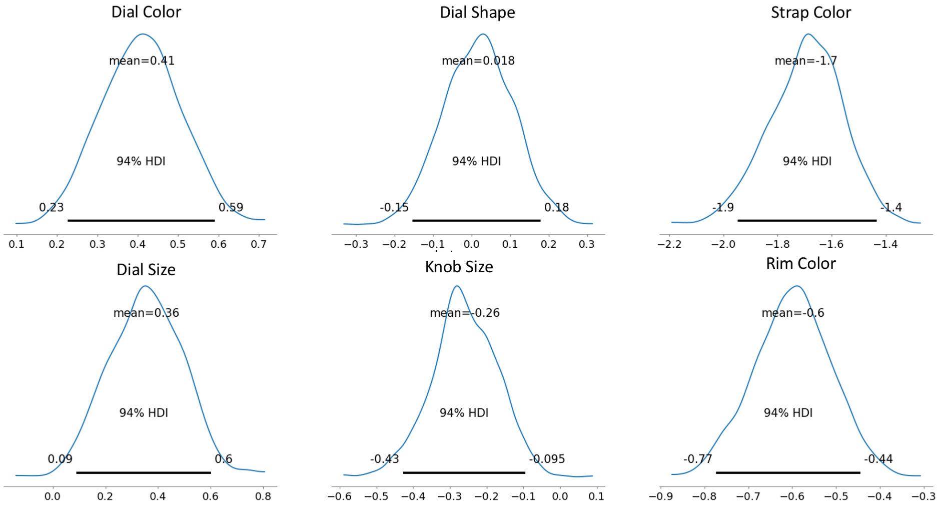
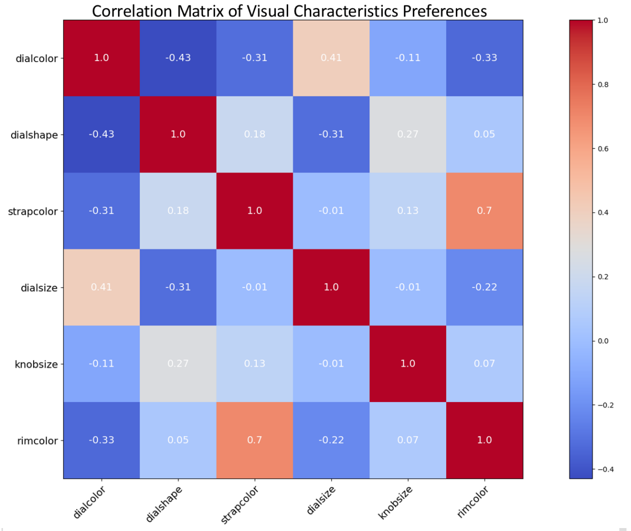


Figure 12 shows posterior plots of population-level preference parameters over the 6 visual attributes. These plots are drawn by averaging individual-level respondent posteriors. For robustness, we compared the mean of these posteriors to a homogeneous logit model and found qualitatively similar results (same effect signs), noting that the magnitudes are different due to

[19] Sampling consisted of 8 parallel chains, each with 10,000 draws of which 5,000 were used for sampler tuning. Convergence of MCMC chains was determined via acceptance criteria of the sampler and its targets (80%), and chain divergences from trace plots (less than 5% draws diverging).

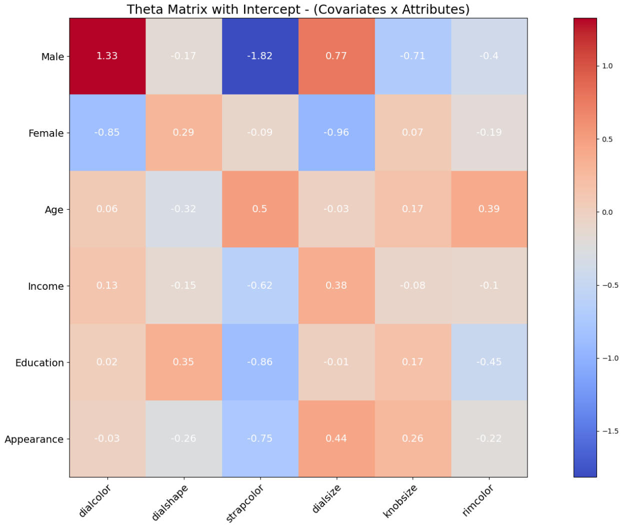
Figure 13: Correlation of population-level preference parameters β



modeling heterogeneity as well as the (implicit) assumption of the scale parameter being unity in logit estimation (Hauser, Eggers, and Selove 2019). Figure 13 shows the corresponding correlation matrix as a heatmap (i.e., normalized mean and standard deviation) over the 6 visual attributes. We find that the strongest correlation (0.70) in consumer preferences is between Strap Color and Rim Color. This implies that, for example, black strap and gold rim color are preferred together as part of a contrasting visual design for the watch. The second and third strongest correlations (-0.43 and 0.41) are those between Dial Shape and Dial Color, and Dial Size and Dial Color, respectively. This implies that consumers prefer, for instance, circular watches with white dial color and larger dial size with black dial color.

We next analyze the relationship between respondents' covariates (demographics) and their preferences over visual attributes. Figure 14 shows a heatmap of the expectation of $\mathbf{\Theta}+\mu_{\Theta}$, namely, the matrix $\mathbf{\Theta}$ plus an intercept term μ_{Θ} from the 3rd-level Gaussian hyperprior (see Equation (7)). Intuitively, this point estimate shows how heterogeneous preferences for watch images correlate with respondent covariates. We note that on average, women are more likely to prefer watches with white dial color and brown strap color, while men are more likely to prefer black dial color and black strap color. We also note that older respondents, particularly those above 55 years of age, were more likely to prefer gold dial color with gold strap color. Lastly, we note that this preference for gold watches was reversed when respondents were in higher income brackets.

Figure 14: Heatmap of expectation of theta matrix relating consumer covariates with preferences over visual characteristics.



Model Evaluation: We compare the predictive accuracy of our representation used along with the HB model against benchmarks, including a baseline logit model and a pretrained deep learning model. We evaluated the models on hit rates for respondents' binary choices among watch visual designs. The first benchmark was a homogenous logit model without respondent covariate variables. The second benchmark was a pretrained deep learning model that included covariate variables to model respondent heterogeneity. We chose the ResNet50 architecture (He et al. 2016) after pre-testing a variety of pretrained network architectures (e.g., DenseNets, VGG) and their performance on the *prediction accuracy metric*.[20] Transfer learning to our conjoint choice task was achieved by "freezing" parameters in the "bottom" layers of neural network, removing the "top" classification layer, and adding new layers on top to train for conjoint choice prediction. These new layers consisted of two nonlinear layers of size 64 before input into a final logit layer for classification.

Table 7 reports these hit rates on a single training and testing split of the data, as defined by

[20] ResNet50 consists of 50 layers consisting of 48 convolutional layers, each with batch normalization, rectified linear, and residual connection between layers. We used pretrained parameters originally estimated on the ImageNet benchmark dataset.

Table 7: Conjoint Model Accuracy (Generated Watches)

Model	**Out-of-Sample Hit Rate (Std. Dev.)**
Logit Model (Homogeneous)	63.16% (2.34%)
Pretrained Deep Learning Model (Heterogeneous)	68.31% (1.54%)
HB Model (Heterogeneous)	72.33% (0.85%)

holding out CBC conjoint tasks for each respondent (stratified splitting) as is convention in the conjoint analysis literature (Gustafsson, Herrmann, and Huber 2013). We find that the pretrained deep learning benchmark performs better than the logit model, as expected especially since it includes additional information on consumer covariates. However, surprisingly, the HB model with a linear utility specification achieves a higher predictive accuracy than the pretrained deep learning model, despite using only 6 visual characteristics to represent the complete visual design of watches.

Generating New "Ideal Point" Watches for Customer Segments

As Orsborn, Cagan, and Boatwright (2009) says, "Even if researchers choose the correct semantics to test, and even if respondents accurately record their responses on these semantic scales, the results on the semantic scales must be translated back into a product shape, where the designer must take the consumers' numerical scores for a set of semantics and translate that into a form which consumers will find desirable."

"Ideal points" refer to the optimal positioning of a product in characteristic space based on preferences of a selected consumer segment. Identification of such ideal points has extensively studied in marketing research and practice (Johnson 1971; Hauser and Urban 1977; DeSarbo, Ramaswamy, and Cohen 1995; Wedel and Kamakura 2000; Lee, Sudhir, and Steckel 2002). The general approach involves the following steps: (a) obtain data on a consumer segment's stated or revealed preferences over a set of existing products that are represented by product characteristics, (b) estimate a predictive model of preferences over these characteristics, and (c) identify new points in characteristic space corresponding to the position of the maximally preferred product of the segment.

We build upon this work by *generating* “ideal point” visual designs, in our case, maximally preferred watch designs for two customer segments. These two segments were identified using the HB model estimated on the conjoint survey data. Specifically, customer segments were defined by singular value decomposition of the expectation of the Θ matrix, denoted $\bar{\mathbf{\Theta}}$ and shown in Figure 14, which relates preferences over visual characteristics with respondent covariates.

$$\bar{\mathbf{\Theta}} = \mathbf{U}_\Theta \Sigma_\Theta \mathbf{V}_\Theta \tag{8}$$

$$\bar{\mathbf{\Theta}}_s = \mathbf{U}_\Theta \mathbf{D}(\sigma_s) \mathbf{V}_\Theta \tag{9}$$

where $\mathbf{U}_\Theta$ is a $J \times J$ unitary matrix corresponding to the J discovered visual characteristics, Σ_Θ is a $J \times R$ diagonal matrix of singular values σ_s, $\mathbf{V}_{\mathbf{\Theta}}$ is a $R \times R$ unitary matrix corresponding to the R consumer covariates, $\bar{\mathbf{\Theta}}_s$ is the segment-level matrix relating segment preferences over visual characteristics and covariates, and $\mathbf{D}$ denotes a diagonal matrix of the same order as Σ_Θ.

Intuitively, the segmentation definition in Equation (9) identifies segments that are most prominent in the population. Larger segments are those defined by eigenvectors with larger singular values than smaller segments with smaller singular values. Similar intuition may be gained by recognizing that the conventional HB specification of a unimodal Gaussian population preference parametrizes individual-level heterogeneity as deviances from the population, see e.g., Evgeniou, Pontil, and Toubia (2007). Eigenvectors of the row-space thus correspond to “principal” deviances of individual-level heterogeneity. Aggregating these deviances across individuals obtains a representation of principal “tastes” differences occurring in the population.

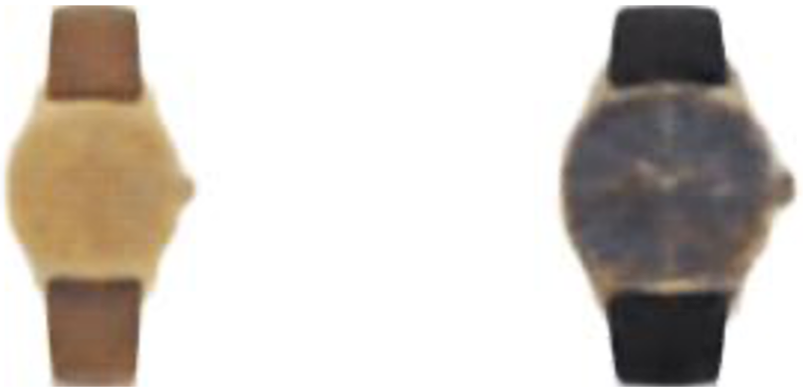
We choose to analyze two customer segments by choosing the two eigenvectors with the largest singular values; in other words, the two most prominent segments in the population. The corresponding segment-level preference “part-worths” of segment s are a random vector denoted β_s. We define the “ideal point” of segment s as the expectation of β_s over its posterior as well as all consumers i,

$$\begin{aligned} E\left[\beta_s\right] &= \sum_i \mathbf{U}_\Theta \mathbf{D}(\sigma_s) \mathbf{V}_\Theta \qquad (10) \\ &= \sum_i \sigma_s \mathbf{u}_s \mathbf{v}_s^T \mathbf{r}_i \end{aligned}$$

where $\mathbf{u}_s$ and $\mathbf{v}_s$ are the s-th row and column vectors of $\mathbf{U}_\Theta$ and $\mathbf{V}_\Theta$, respectively.

We find that Segment 1 corresponds to consumers that are more likely to be female, younger, moderately affluent, less educated, and attach an above average importance to visual appearance relative to the population. On average, this segment prefers watches that have white dial color, are smaller and more rectangular, have brown strap color, and a rim color matching the dial color. Segment 2 corresponds to consumer that were more likely to be male, older, more educated, and attach slightly above average importance to visual appearance relative to the population. On average, this segment prefers watches that have black dial color, are larger and more circular, and with a rim color that contrasts with the dial color.

Figure 15: Generated "Ideal Point" Watches for Two Segments



Segment 1:
"Ideal Point" Watch Design

Segment 2:
"Ideal Point" Watch Design

We next generated new watches corresponding to the "ideal point" (i.e., optimal visual characteristics) for the two segments as obtained in Equation (10). Figure 15 shows the "ideal point" watches of the two segments, as defined by the mean of the segment-level preference vector posterior, β_s.[21]

DISCUSSION AND CONCLUSION

Despite the importance of visual characteristics in marketing and business, to date, there has been no comprehensive approach to automatically identify the characteristics that contribute to the visual design. This is an important issue because consumers are known to have preferences over visual design across a wide range of product characteristics. So, understanding the impact of visual design on consumer demand is of considerable interest (Kang et al. 2019; Burnap, Hauser, and Timoshenko 2019; Liu et al. 2017).

Our research develops a methodology to automatically discover and quantify visual design characteristics using a combination of unstructured product image data, in conjunction with structured product characteristics and price. In contrast to ML methods which require ground truth, we use structured characteristics to supervise the disentanglement model to enhance its performance. The discovered characteristics are disentangled, and interpretable by humans. Moreover, we can generate novel counterfactual designs by varying the levels of the discovered characteristics one at a time. We use this flexibility to conduct visual conjoint design and obtain consumer preferences over visual characteristics. These are then used to obtain distinct "ideal point" visual designs.

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

[21] Note that this implicitly assumes the "ideal point" product has the same vector norm magnitude as the preference parameters. In other words, the segment's "ideal point" is equal to the segment's preference vector β_s. See (Kaul and Rao 1995; DeSarbo, Ramaswamy, and Cohen 1995) for more details.

et al. 2021). However, we validate that in practice we observe that our proposed method performs well in a realistic and practical setting.

There are several questions worthy of examination in future research. First, it would be useful to understand what combinations of product characteristics typically improve the disentanglement the most across product categories, and the underlying reason. More insight into the specific conditions under which certain combinations of signals might produce better disentanglement would be valuable. Second, examining the performance of a similar method in other modalities like text or audio would also be helpful. Since consumer decision making is likely to depend on multiple sources of information and persuasion, it would be interesting to examine whether having one modality helps to improve the impact of another, e.g. the presence of text might help disentangle images better. Third, it would be interesting to examine how visual characteristics may be incorporated into models of demand and supply, so that we can understand both consumer preferences and firm strategic choices involving visual design.

Overall, we expect these developments in this area of disentanglement to enable many different research questions regarding visual design and consumer perceptions and preference to be explored.

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.

Aubry, Mathieu, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic "Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models," "CVPR," (2014).

Baldi, Pierre and Kurt Hornik (1989), "Neural networks and principal component analysis: Learning from examples without local minima," *Neural networks*, 2 (1), 53–58.

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.

Bengio, Yoshua, Aaron C Courville, and Pascal Vincent (2012), "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR, abs/1206.5538*, 1 (2665), 2012.

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

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.

Bloch, Peter H (1995), "Seeking the ideal form: Product design and consumer response," *Journal of marketing*, 59 (3), 16–29.

Bloch, Peter H, Frederic F Brunel, and Todd J Arnold (2003), "Individual differences in the centrality of visual product aesthetics: Concept and measurement," *Journal of consumer research*, 29 (4), 551–565.

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).

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2020), "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*.

DeSarbo, Wayne S, Venkatram Ramaswamy, and Steven H Cohen (1995), "Market segmentation with choice-based conjoint analysis," *Marketing Letters*, 6, 137–147.

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).

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

Estermann, Benjamin, Markus Marks, and Mehmet Fatih Yanik (2020), "Robust Disentanglement of a Few Factors at a Time using rPU-VAE," *Advances in Neural Information Processing Systems*, 33, 13387–13398.

Evgeniou, Theodoros, Massimiliano Pontil, and Olivier Toubia (2007), “A convex optimization approach to modeling consumer heterogeneity in conjoint estimation,” *Marketing Science*, 26 (6), 805–818.

Ferraro, Rosellina, Amna Kirmani, and Ted Matherly (2013), “Look at me! Look at me! Conspicuous brand usage, self-brand connection, and dilution,” *Journal of Marketing Research*, 50 (4), 477–488.

Fleming, Roland W (2014), “Visual perception of materials and their properties,” *Vision research*, 94, 62–75.

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.

Gustafsson, Anders, Andreas Herrmann, and Frank Huber (2013), *Conjoint measurement: Methods and applications* Springer Science & Business Media.

Hauser, John R, Felix Eggers, and Matthew Selove (2019), “The strategic implications of scale in choice-based conjoint analysis,” *Marketing Science*, 38 (6), 1059–1081.

Hauser, John R and Glen L Urban (1977), “A normative methodology for modeling consumer response to innovation,” *Operations Research*, 25 (4), 579–619.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun “Deep residual learning for image recognition,” “Proceedings of the IEEE conference on computer vision and pattern recognition,” pages 770–778 (2016).

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, Andrew Gelman et al. (2014), "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.," *J. Mach. Learn. Res.*, 15 (1), 1593–1623.

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).

Johnson, Richard M (1971), "Market segmentation: A strategic management tool," *Journal of Marketing Research*, 8 (1), 13–18.

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*.

Kappe, Eelco and Stefan Stremersch (2016), "Drug detailing and doctors' prescription decisions: the role of information content in the face of competitive entry," *Marketing Science*, 35 (6), 915–933.

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*.

Kaul, Anil and Vithala R Rao (1995), "Research for product positioning and design decisions: An integrative review," *International Journal of research in Marketing*, 12 (4), 293–320.

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

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

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).

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

Landwehr, Jan R, Aparna A Labroo, and Andreas Herrmann (2011), "Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts," *Marketing Science*, 30 (3), 416–429.

Lee, Jack KH, Karunakaran Sudhir, and Joel H Steckel (2002), "A multiple ideal point model: Capturing multiple preference effects from within an ideal point framework," *Journal of Marketing Research*, 39 (1), 73–86.

Lee, Jung Eun, Songyee Hur, and Brandi Watkins (2018), "Visual communication of luxury fashion brands on social media: effects of visual complexity and brand familiarity," *Journal of Brand Management*, 25, 449–462.

Lee, Thomas Y and Eric T Bradlow (2011), "Automated marketing research using online customer reviews," *Journal of Marketing Research*, 48 (5), 881–894.

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.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009), "Generating random correlation matrices based on vines and extended onion method," *Journal of multivariate analysis*, 100 (9), 1989–2001.

Linting, Mariëlle, Jacqueline J Meulman, Patrick JF Groenen, and Anita J van der Koojj (2007), "Nonlinear principal components analysis: introduction and application.," *Psychological methods*, 12 (3), 336.

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, Xiao, Param Vir Singh, and Kannan Srinivasan (2016), "A structured analysis of unstructured big data by leveraging cloud computing," *Marketing Science*, 35 (3), 363–388.

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.

Liu, Zhiyuan, Yankai Lin, and Maosong Sun (2020), *Representation learning for natural language processing* Springer Nature.

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

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).

Lundberg, Scott M and Su-In Lee (2017), “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, 30.

McCullough, Dick (2002), “A user’s guide to conjoint analysis.,” *Marketing Research*, 14 (2).

Megehee, Carol M and Deborah F Spake (2012), “Consumer enactments of archetypes using luxury brands,” *Journal of business research*, 65 (10), 1434–1442.

Mika, Sebastian, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch (1998), “Kernel PCA and de-noising in feature spaces,” *Advances in neural information processing systems*, 11.

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).

Oppenheimer, Daniel M, Tom Meyvis, and Nicolas Davidenko (2009), “Instructional manipulation checks: Detecting satisficing to increase statistical power,” *Journal of experimental social psychology*, 45 (4), 867–872.

Orsborn, Seth, Jonathan Cagan, and Peter Boatwright (2009), “Quantifying aesthetic form preference in a utility function,”.

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).

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

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986), “Learning representations by back-propagating errors,” *nature*, 323 (6088), 533–536.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio (2021), “Toward causal representation learning,” *Proceedings of the IEEE*, 109 (5), 612–634.

Shapley, Lloyd S (1997), “A value for n-person games,” *Classics in game theory*, 69.

Simonson, Alex and Bernd H Schmitt (1997), *Marketing aesthetics: The strategic management of brands, identity, and image* Simon and Schuster.

Sylcott, Brian, Seth Orsborn, and Jonathan Cagan (2016), “The effect of product representation in visual conjoint analysis,” *Journal of Mechanical Design*, 138 (10), 101104.

Tractinsky, Noam, Adi S Katz, and Dror Ikar (2000), “What is beautiful is usable,” *Interacting with computers*, 13 (2), 127–145.

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.

Wedel, Michel and Wagner A Kamakura (2000), *Market segmentation: Conceptual and methodological foundations* Springer Science & Business Media.

Williams, Christopher (2000), “On a connection between kernel PCA and metric multidimensional scaling,” *Advances in neural information processing systems*, 13.

APPENDIX – TABLE OF CONTENTS

CONNECTIONS WITH EXISTING MARKETING METHODS

We also include both a high-level comparison of the methods in Table A.1.

Table A.1: Comparison of Methods

Method	PCA	MDS	AE	VAE	Disentanglement
Dimensionality Reduction	Yes	Yes	Yes	Yes	Yes
Reconstruction of Existing Examples	Yes	Yes	Yes	Yes	Yes
Generation of New Examples	No	No	No	Yes	Yes
Use with Unstructured Data	Yes	Yes	Yes	Yes	Yes
Interpretability using Unstructured Data	No	No	No	No	Yes
Stochastic (S) or Deterministic (D)	D	D	D	S	S
Non-Linear Transformations	No	No	Yes	Yes	Yes

Several methods used in marketing can be used to compress high-dimensional data into a lower-dimensional representation as shown in Table A.1. The simplest and perhaps most well-known is principle component analysis (PCA). PCA assumes that the data lie on a linear subspace and captures the global linear structure in the data. PCA has been used in marketing for dimensionality reduction (Liu, Singh, and Srinivasan 2016; Kappe and Stremersch 2016) in order to make solving the models tractable. Multi-dimensional scaling (MDS) is a method that aims to minimize dissimilarity between distances in the high-dimensional data and distances in the lower-dimensional representation. MDS is a general method as "distance" can be nonlinear and even nonmetric; however, conventionally researchers assume Euclidean distances which makes it equivalent to PCA (Williams 2000). While PCA and MDS have been widely-used in marketing to reduce data dimensionality for managerial interpretation (see, e.g., (Lee and Bradlow 2011)), these methods are not well suited to capturing complex nonlinear relationships in unstructured data (Linting et al. 2007). Consequently, they are likewise not well suited for our goal of discovering interpretable visual characteristics directly from unstructured image data.

An autoencoder (AE) (Baldi and Hornik 1989; Rumelhart, Hinton, and Williams 1986) is a nonlinear method that focuses on reconstructing the original high-dimensional data (typically unstructured data such as images), while compressing the original data into a lower-dimensional representation. Autoencoders can capture complex nonlinear relationships, especially those prevalent in visual data, and thus typically outperform linear methods like PCA in terms of reconstruction accuracy (Mika et al. 1998). An AE is equivalent to PCA if it is restricted to only linear transformations (Roweis and Ghahramani 1999; Bengio, Courville, and Vincent 2012). While the AE can reconstruct the original data with medium-to-high fidelity, it cannot generate new out-of-sample data that it has never seen. Thus, similar to the case of PCA and MDS, we cannot term it as a generative model.

In contrast, a variational autoencoder (VAE) is a probabilistic generative model that similarly represents high-dimensional data using lower-dimensional latent variables (Kingma and Welling 2014). The VAE takes a Bayesian approach by learning the latent variable distributions using variational inference. While architecturally similar to the (non-generative) AE, the VAE is able to *generate new data that are similar to the input data* by sampling from its probabilistic generative model by conditioning on the latent variables. Lastly, β-TCVAE (Chen et al. 2018) builds upon VAE by: (a) promoting statistical independence in the latent space; (b) discourages data copying by minimizing mutual information between the input data and the latent space; (c) minimizes the number of truly informative dimensions. The above objectives are often conflicting, and the model uses hyperparameters that decide the weights associated with these terms.

Comparison of Generative Methods: The two broad classes of generative models are based on variational autoencoders (VAEs) (Kingma and Welling 2014) and generative adversarial networks (GAN) [22] (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, as GANs traditionally do not infer a representation[23], 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.

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

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

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 human interpretable 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 human 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: 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."

SUMMARY STATISTICS OF STRUCTURED CHARACTERISTICS OF AUCTIONED WATCHES

Table B.1 provides summary statistics of the auctioned watches.

Table B.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.

HYPERPARAMETERS TUNING – WATCHES

We list the hyperparameters obtained for both the supervised approaches and the unsupervised approach for finding visual characteristics of watches in Table C.1.

Table C.1: Hyperparameters Obtained by Model Selection Criteria

Disentanglement Approach	Signal	Number of Signals	β	δ
Supervised	Brand	1	18	50
Supervised	Circa	1	4	35
Supervised	Material	1	6	25
Supervised	Movement	1	4	20
Supervised	Price	1	1	16
Supervised	Brand and Circa	2	48	5
Supervised	Circa and Material	2	36	1
Supervised	Brand and Material	2	50	1
Supervised	Circa and Movement	2	50	5
Supervised	Brand and Movement	2	6	20
Supervised	Material and Movement	2	6	10
Supervised	Brand, Material and Movement	3	40	1
Unsupervised	–	0	18	0

USING SHAPLEY VALUES (SHAP) FOR DISENTANGLEMENT

In this section, we use an alternative approach to discover visual characteristics. The idea behind this approach is to identify select elements (pixels) of each input image that are predictive of a supervisory signal, and then use those elements as an input to the disentanglement model.

In this approach, we first train a deep learning model to predict the supervisory signal (e,g. brand) from images. Next, we calculate SHAP values to identify which features of the deep learning model drive the model's results (Lundberg and Lee 2017). The SHapley Additive exPlanations (SHAP) technique utilizes game theory to interpret the results of machine learning models. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (Shapley 1997). SHAP values of each feature captures the contribution of each feature to overall model predictions. It is calculated by estimating differences between models with subsets of the feature space and then averaging across samples.

We calculate SHAP values to rank the features based on their contribution to the model's output. The higher the SHAP value for a feature, the more significant its contribution. We then sort the SHAP values in descending order to select the pixels corresponding to the top features using the SHAP values as a mask. These image subsamples are used as an input to the disentanglement-based VAE model. Figure D.1 shows a sample of images fed to the disentanglement-based VAE model using this approach.

Figure D.2 gives example output of discovered visual characteristics from this approach. In each row of the figure, we show how the watch image changes based on changes in levels of one selected visual characteristic, while keeping all the other characteristics fixed. We show the top six visual characteristics based on the KL divergence value of the difference between the posterior and the Gaussian prior. We can only interpret the first three visual characteristics. The next three visual characteristics appear to be entangled. By entangled, we mean that when any one entangled characteristic is kept fixed and other characteristics are changed, the watch image changes in more than one interpretable way. Note that these characteristics are not uninformative because their KL divergence is not close to 0.

Figure D.1: Sample of images from SHAP-based approach

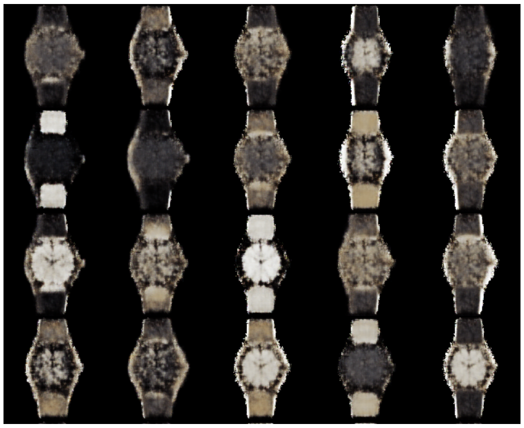
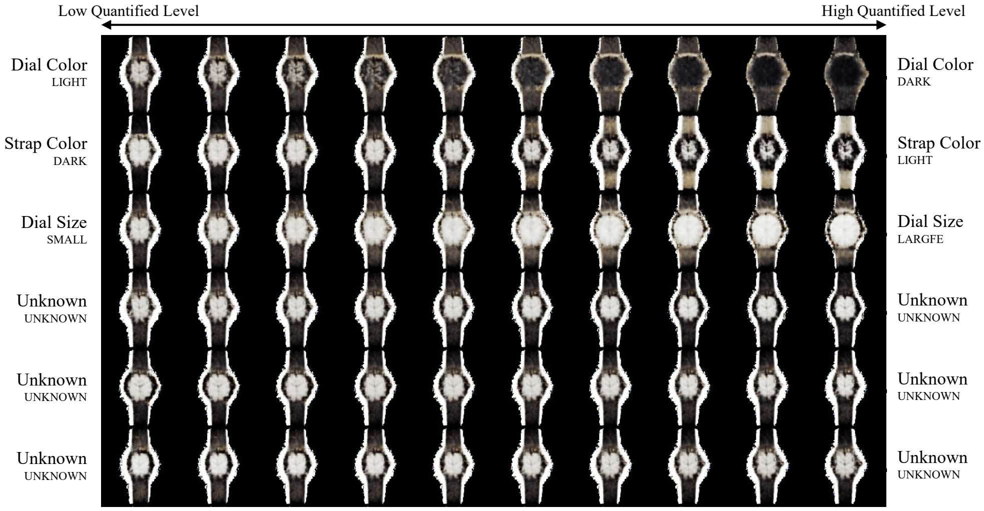


Figure D.2: Discovered Visual Characteristics using SHAP-based approach



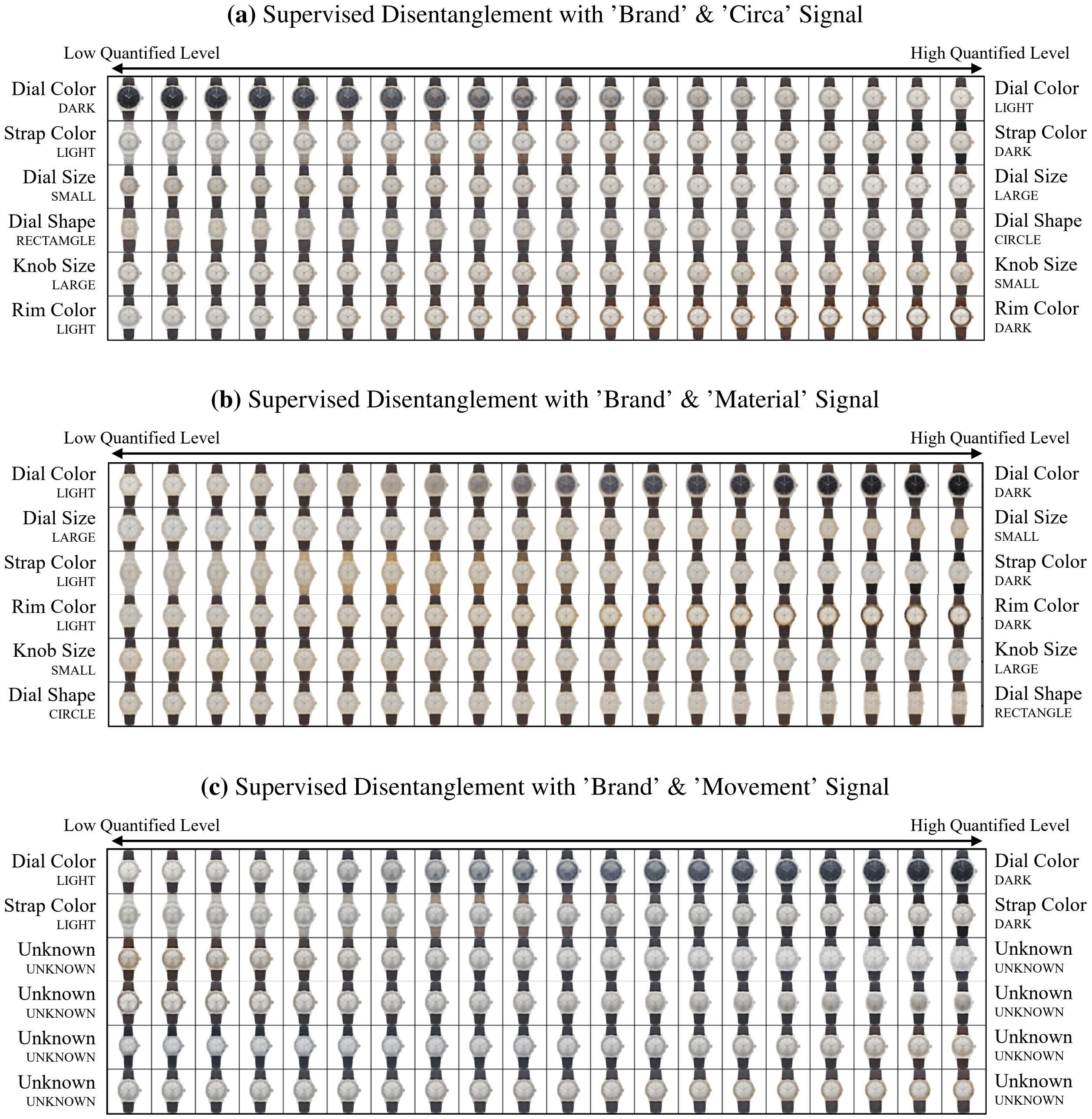
Notes: Latent traversals along a *focal watch* used to visualise the semantic meaning encoded by single visual characteristic learnt by a trained model. In each row, the quantitative level of a single characteristic is varied keeping the other characteristics fixed. The resulting reconstruction is visualized.

DISCOVERED VISUAL CHARACTERISTICS WITH DIFFERENT SUPERVISORY SIGNALS

Figures E.1 to E.3 show the discovered visual characteristics learned by using various combinations of the supervisory signals (Brand, Circa, Material and Movement).

Overall, we find that combining two structured product characteristics as supervisory signals generally achieves a higher UDR than a single product characteristic as a signal. We also find that the combination of Brand+Material achieves the best disentanglement. This combination performs even better than including 3 product characteristics as supervisory signals.

Figure E.1: Discovered Visual characteristics from Multiple Supervisory Signals



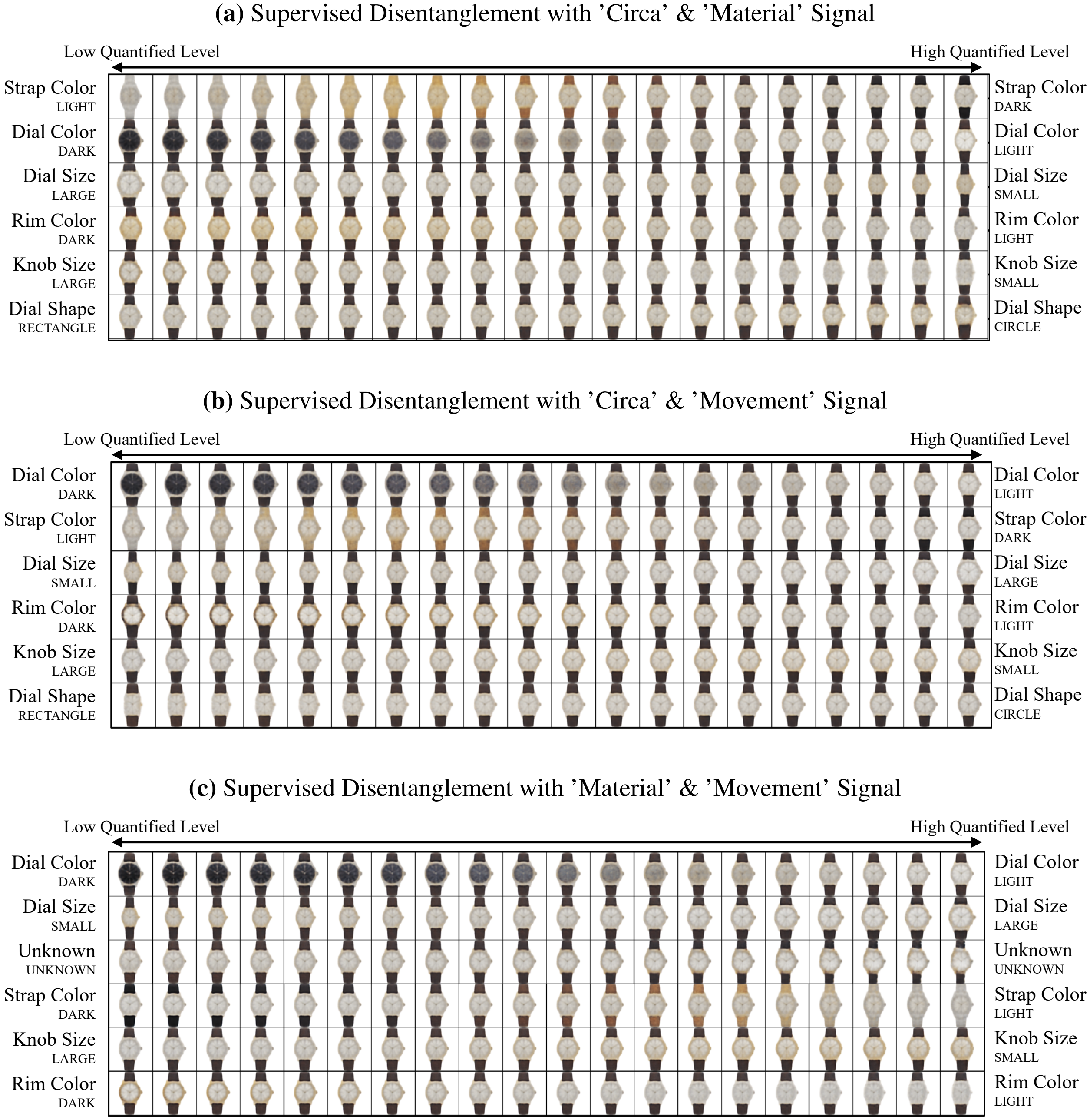
(a) Supervised Disentanglement with ’Brand’ & ’Circa’ Signal

(b) Supervised Disentanglement with ’Brand’ & ’Material’ Signal

(c) Supervised Disentanglement with ’Brand’ & ’Movement’ Signal

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 and circa simultaneously. **b**: Discovered visual characteristics learned by supervising the characteristics to predict the brand and material simultaneously. **c**: Discovered visual characteristics learned by supervising the characteristics to predict the brand and movement simultaneously.

Figure E.2: Discovered Visual characteristics from Multiple Supervisory Signals



(a) Supervised Disentanglement with ’Circa’ & ’Material’ Signal

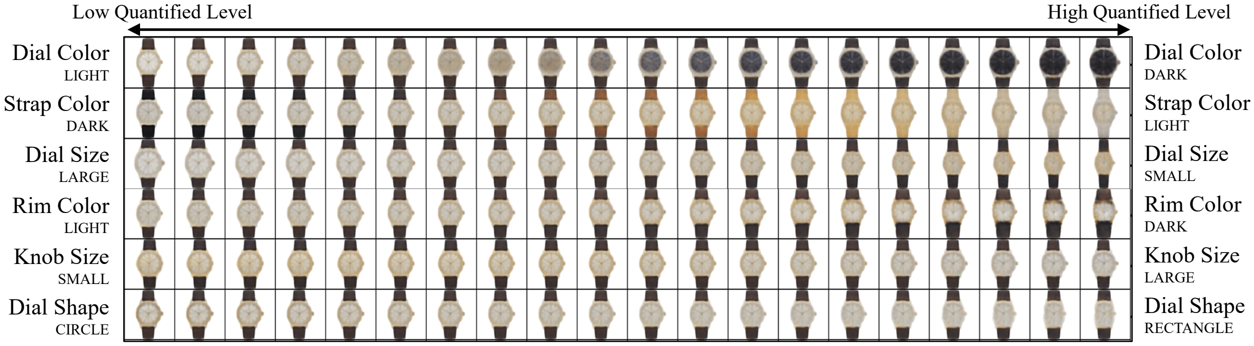
(b) Supervised Disentanglement with ’Circa’ & ’Movement’ Signal

(c) Supervised Disentanglement with ’Material’ & ’Movement’ Signal

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 circa and material simultaneously. **b**: Discovered visual characteristics learned by supervising the characteristics to predict the circa and movement simultaneously. **c**: Discovered visual characteristics learned by supervising the characteristics to predict the material and movement simultaneously.

Figure E.3: Discovered Visual characteristics from Multiple Supervisory Signals

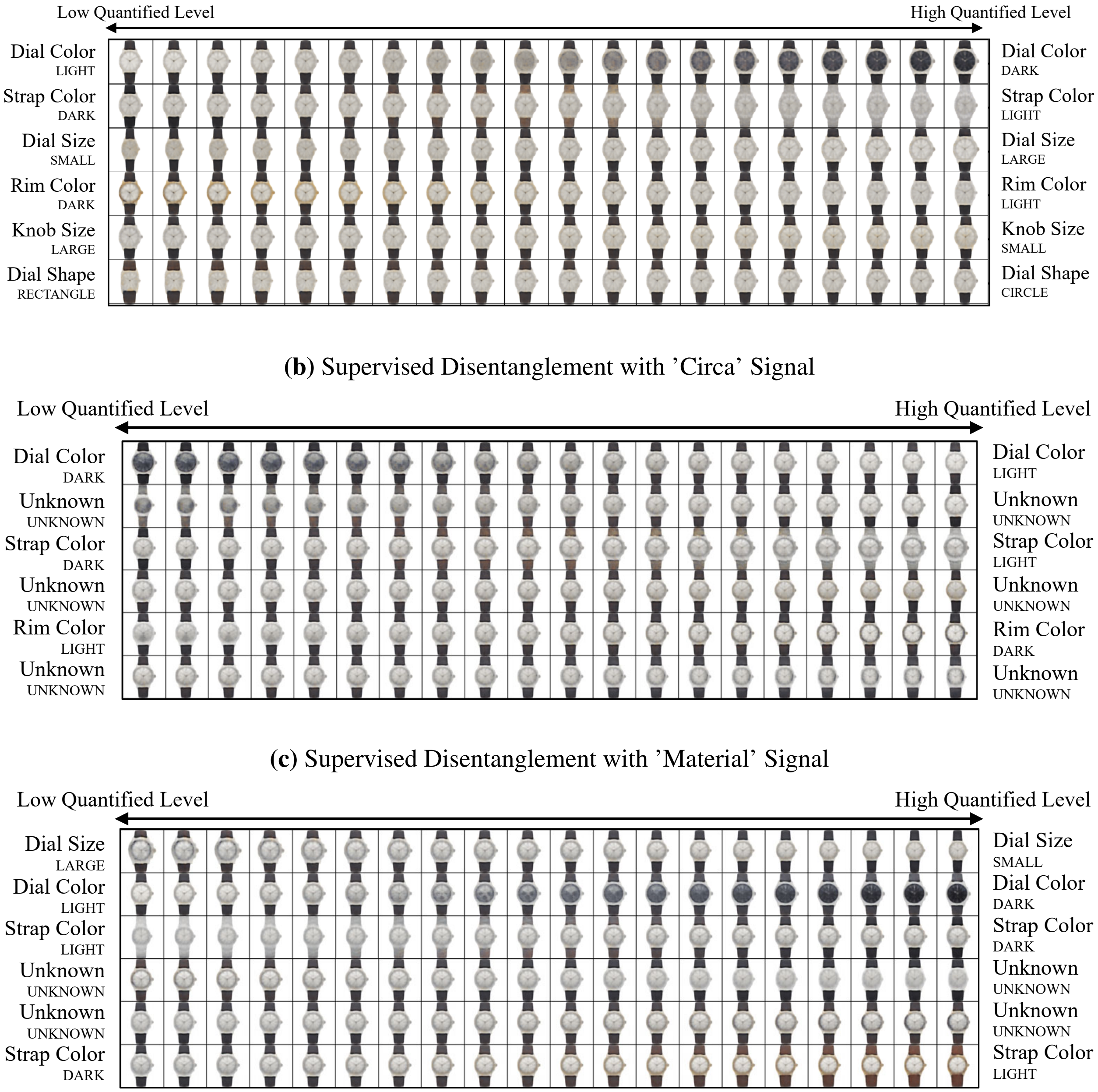
(a) Supervised Disentanglement with 'Brand', 'Material' & 'Movement' Signal



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, material and movement simultaneously.

Figure Figures E.4 to E.6 show the discovered visual characteristics learned by using ‘Brand’, ‘Circa’, ‘Material’, ‘Movement’, and ‘Price’ as supervisory signals as well as an unsupervised approach.

Figure E.4: Discovered Visual characteristics from Single Supervisory Signals



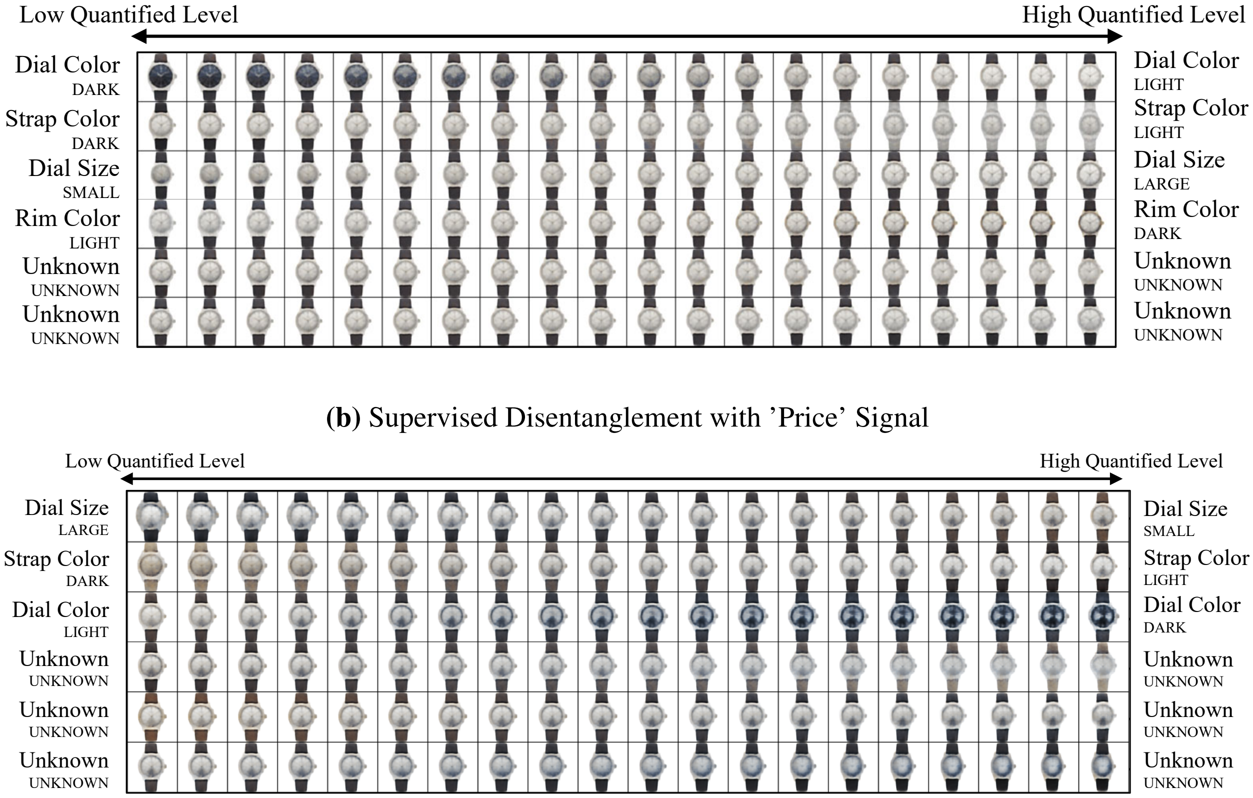
(a) Supervised Disentanglement with 'Brand' Signal

(b) Supervised Disentanglement with 'Circa' Signal

(c) Supervised Disentanglement with 'Material' Signal

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 circa simultaneously. **c**: Discovered visual characteristics learned by supervising the characteristics to predict the material simultaneously.

Figure E.5: Discovered Visual characteristics from Single Supervisory Signals



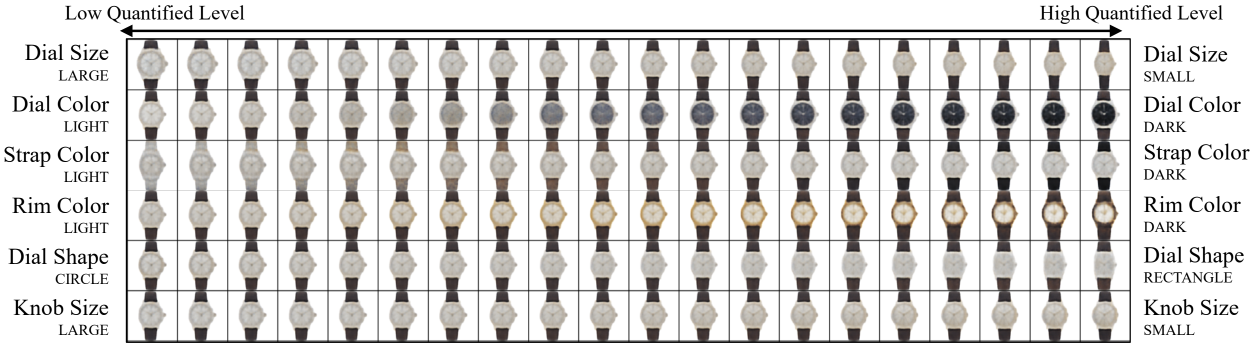
(a) Supervised Disentanglement with 'Movement' Signal

(b) Supervised Disentanglement with 'Price' Signal

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 movement simultaneously. **b**: Discovered visual characteristics learned by supervising the characteristics to predict the circa and movement simultaneously. **c**: Discovered visual characteristics learned by supervising the characteristics to predict the price simultaneously.

Figure E.6: Discovered Visual characteristics from Unsupervised Approach

(a) Unsupervised Disentanglement



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 the unsupervised approach.

DISENTANGLEMENT IN A DIFFERENT PRODUCT CATEGORY – SNEAKERS

Our data includes sneakers sold at Zappos. For each sneaker in the dataset, we have its image, brand, and price. Figure F.1 shows a sample of sneaker images in our dataset. We obtained the dataset of sneakers sold on Zappos in March 2023. These shoes were classified as sneakers by the retailer. Overall, our dataset includes 2,227 unique sneaker models with an average of 2.5 images per sneaker model. The size of the overall dataset includes 5575 images. We only included the side view of sneakers in order to focus on the variation in the shape of the sneakers. Finally, we specifically used grayscale images because each sneaker model with the same shape comes in multiple colors. We preprocessed each image to have the size of 128x128 dimensions to keep the images consistent with the watch category. A total of 247 unique brands are present in the data. Skechers, Vans, New Balance, adidas and ASICS are the five brands with the largest share of observations. Table F.1 provides summary statistics of the sneakers.

We use the same deep learning model architecture as well as the same hyperparameters (except the disentanglement hyperparameters β and δ) as the one used for learning visual characteristics of watches. We follow the same method for training the model, selecting the hyperparameters β and δ and then evaluating different supervisory signals for the sneakers category using Unsupervised Disentanglement Ranking (UDR).

Figure F.1: Sample of Sneakers Sold at Zappos

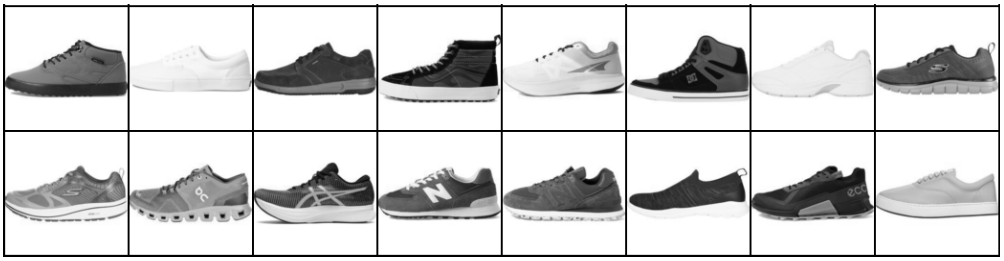


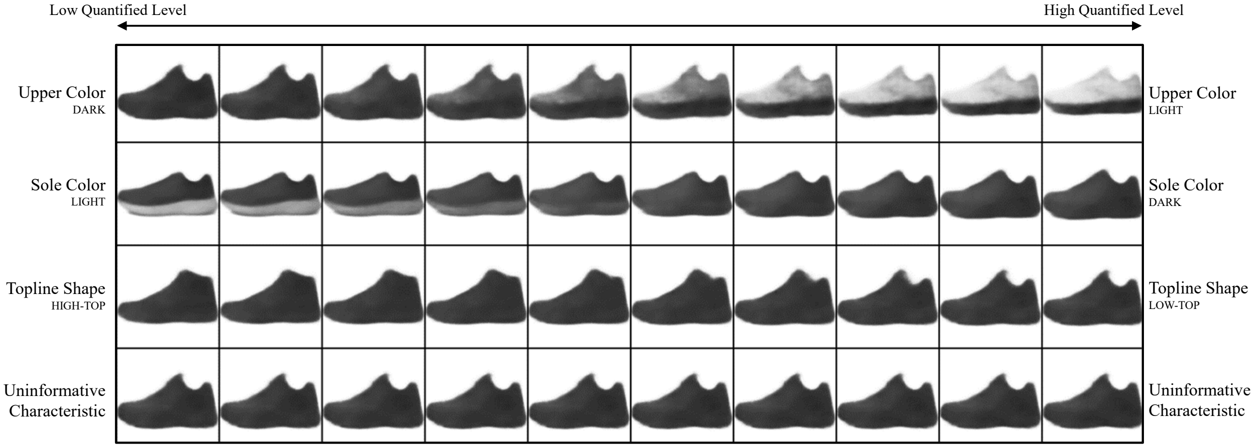
Figure F.2 gives example output of discovered visual characteristics corresponding to the supervisory signals ‘price’. In each row of the figure, we show how the sneaker image changes based on changes in levels of one visual characteristic, while keeping all the other characteristics fixed. We

Table F.1: Summary Statistics of Structured characteristics of Sneakers Sold at Zappos

Statistic	Mean	SD	Min	Max
Brand (Skechers)	0.09	0.29	0	1
Brand (Vans)	0.08	0.28	0	1
Brand (New Balance)	0.07	0.26	0	1
Brand (adidas)	0.06	0.24	0	1
Brand (ASICS)	0.05	0.22	0	1
...				
Brand (Others)	0.14	0.34	0	1
Price (in $s)	112.30	46.45	30.00	650.20

only show three visual characteristics as rest of the characteristics are found to be uninformative i.e. the KL divergence of the posterior was not much different from the Gaussian prior. Traversing along an uninformative characteristic leads to no visual change, and we show one uninformative characteristic for reference.

Figure F.2: Discovered Visual Characteristics of Sneakers



Notes: Latent traversals along a *focal sneaker* 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. Discovered visual characteristics learned by supervising the characteristics to predict the price simultaneously.

Table F.2: Comparison of Different Supervisory Approaches

Number of Signals	Supervisory Signals	UDR
1	Price[1]	0.286
0	Unsupervised	0.126
2	Brand & Price[1]	0.094
1	Brand	0.093

Price when used separately is a continuous variable. However, when used in conjunction with Brand, we discretize Price into 5 levels.