Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis

Ankit Sisodia, Alex Burnap and Vineet Kumar*

Summer 2024

Abstract

This article develops a method to automatically discover and quantify human-interpretable visual characteristics directly from product image data. The method is generative, and can create new visual designs spanning the space of visual characteristics. It builds on disentanglement methods in deep learning using variational autoencoders, which aim to discover underlying statistically independent and interpretable visual characteristics of an object. The impossibility theorem in the deep learning literature indicates that supervision with ground truth characteristics would be required to obtain unique disentangled representations. However, these are typically unknown in real world applications, and are in fact exactly the characteristics we want to discover. Extant machine learning methods require ground truth labels for each visual characteristic, resulting in a task requiring human evaluation and judgment to both design and operationalize. In contrast, this method postulates the use of readily available product characteristics (such as brand and price) as proxy supervisory signals to enable disentanglement. This method discovers and quantifies human-interpretable and statistically independent characteristics without any specific domain knowledge on the product category. It is applied to a dataset of watches to automatically discover interpretable visual product characteristics, obtain consumer preferences over visual designs, and generate new ideal point designs targeted to specific consumer segments.

Keywords: Visual Characteristics; Generative Product Design; Disentanglement; Deep Learning

*Ankit Sisodia is an Assistant Professor of Marketing at the Daniels School Of Business at Purdue University. email: asisodia@purdue.edu. Alex Burnap is an Assistant Professor of Marketing at the Yale School of Management. email: alex.burnap@yale.edu. Vineet Kumar is an Associate Professor of Marketing at the Yale School of Management. email: vineet.kumar@yale.edu. We thank participants in seminars at Dartmouth College, Indian Institute of Management Bangalore, Indian School of Business, National University of Singapore, Nanyang Technological University, Purdue University, Santa Clara University, University of Illinois Urbana-Champaign, University of Minnesota, University of Texas Dallas, and Washington University at St. Louis. We thank Raghuram Iyengar, Oded Netzer, Artem Timoshenko, and Olivier Toubia for their comments.

INTRODUCTION

Visual product characteristics are known to be a significant driver of consumer purchase across a wide range of product categories, including automobiles, apparel, furniture, consumer technology products and even houses (Simonson and Schmitt 1997; Bloch 1995; Heitmann et al. 2020). This suggests their inclusion in quantitative marketing models for accurate forecasts of market demand, as well as segmentation and targeting for new product design. However, while demand has been traditionally modeled in marketing and economics as being based on underlying product characteristics (e.g., Lancaster (1966)), identifying and quantifying visual design characteristics remains a significant challenge. In contrast, structured product characteristics are readily characterized and quantified, e.g. in the automobile market, this may include horsepower and fuel efficiency; in housing, square footage and number of bedrooms; in apparel, size and material.[1]

We develop a method with the following aims related to visual design: a) identifying (discovering) and quantifying human-interpretable visual characteristics from product images, b) obtaining consumer preferences across a range of generated visual designs (visual conjoint), and c) generating novel "ideal point" visual designs targeted to specific consumer segments. Our method of obtaining interpretable visual characteristics could then be used in quantifying consumer preferences, demand responses, and firms' strategic choices in the visual domain. Discovery and quantification of visual characteristics is a first step in enabling these analyses. Practitioners can also use our method to generate visual designs for prototyping, visually differentiate their products from market offerings, and generate new visual designs targeted to consumer segments.

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* identify and define which visual characteristics adequately represent a product's visual form (Bloch 1995). Even after defining visual characteristics in this

[1] We use the terminology since they can be represented in structured data.

manner, 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 design in a human-interpretable manner.[2]

Generative Design for Visual Conjoint: We demonstrate how to use these quantified visual characteristics in an application of *visual conjoint analysis*. The generative aspect of our method is critically important in obtaining consumer preferences across visual characteristics, since it allows us to automatically generate images that vary the visual design separately along each of the discovered characteristics. We obtain consumer preferences over these discovered visual characteristics using a Hierarchical Bayesian (HB) model, accounting for consumer heterogeneity over observed demographic and psychographic variables. We then show how our method can be used to automatically generate novel and targeted product designs for consumer segments. Specifically, we identify two segments of consumers, and obtain segment-level "ideal points" using their estimated preferences over the disentangled 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 qualitatively show these "ideal point" visual designs are differentiated, and quantitatively show they draw choice share away from existing product offerings.

Methodological Basis: We build upon the disentanglement stream of literature in representation learning, an area of deep learning, with our primary goal of obtaining 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 [e.g. raw images] can be explained by the substantially lower dimensional and semantically meaningful [to humans] latent variables." ([] indicate our additions for clarity).

Disentanglement learning is a form of representation learning (Bengio, Courville, and Vincent 2013), and commonly builds upon variational autoencoders (VAE) (Kingma and Welling 2014).

[2]Our focus here is not on discovering *outlier* characteristics that are particularly *surprising* to humans, especially experts. Rather, it is to identify *and quantify* aspects directly from visual product images and show their use in generative design, all in an automated manner.

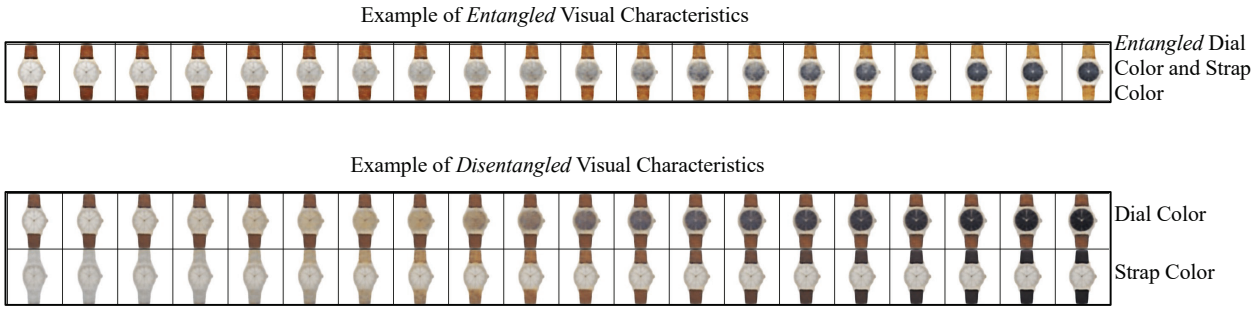
VAEs are comprised of an encoder neural net and decoder neural net, both of which are parameterized by highly nonlinear deep neural networks. The encoder neural net takes high-dimensional unstructured data (images) as input and outputs a latent low-dimensional vector of distributions (embedding of visual characteristics). The VAE uses variational inference, an approximate Bayesian approach, resulting in each of the latent (visual) characteristics represented as distributions rather than "point estimates." In contrast to (typical) deterministic neural networks such as autoencoders, this stochastic approach helps model uncertainty over visual characteristics via a smooth, continuous, compact and flexible latent embedding distribution. This modeling is important to obtain a consistent (and interpretable) representation to estimate consumer preferences, as well as smooth and controllable generation of novel visual designs through sampling of different points in the distribution. The decoder neural net takes as input the low-dimensional vector and attempts to reconstruct the original data as output. The idea underlying representation learning is that the "true" dimension of images in the data belonging to a category (e.g. a set of images of watches) is much lower than the dimensionality of the raw images.[3]

Disentanglement aims at identifying a multi-dimensional latent representation in the image data, where each dimension maps one-to-one with a human-interpretable characteristic (Bengio, Courville, and Vincent 2013; Locatello et al. 2019). With a disentangled representation, a change in one latent dimension would result in a change to only one human-interpretable visual characteristic, whereas with an entangled representation, a change in the level across one discovered latent dimension would impact *multiple* human-interpretable characteristics. Figure 1 illustrates the difference between disentangled and entangled representations.

Disentanglement learning using only images with unsupervised learning has significant limitations, due to a well-known result called the impossibility theorem (Locatello et al. 2019). Recent research recommends using supervised learning with "ground truth" visual characteristics for each

[3]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. Random images typically cannot be reduced in dimension, but images that belong to a category can typically be represented in much lower dimension. 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.

Figure 1: (Color Online) Entangled and Disentangled Visual Characteristics



Visual characteristics correspond to dimensions in latent space. Here, we see that the entangled visual characteristic changes both the dial color and strap color as its the value is changed. Disentangled characteristics corresponding to two independent characteristics for dial color and strap color, so a change in value corresponds to a change in only one visual characteristic.

data point (i.e., product image) as a supervisory signal (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 seek to learn*. Our research thus aims to extend recent machine learning developments in disentanglement methods.

Contribution: The goal of our method is to *automatically* identify and obtain a disentangled representation of *interpretable* visual characteristics in order to *generate* counterfactual visual designs targeted to consumer preferences. Our method works even in the presence of correlation between these visual characteristics in the original data. Current machine learning approaches use ground truth signals separately for each visual characteristic, which are assumed to completely and accurately capture the true underlying data generating process for images. However, the critical challenge is that ground truth is not available in typical applications, and designers expend lots of effort and resources in determining the visual characteristics for products. Our methodology aims to overcome this issue by showing that supervised disentanglement, with structured product characteristics as signals (labels), which are readily-available in typical marketing datasets, can

[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). Broadly, this manual approach requires first identifying the visual characteristics (by researcher), obtaining annotations from multiple human coders and reconciling these noisy measures to create "ground truth" labels.

both address known theoretical limitations and improve disentanglement performance to obtain human-interpretable visual characteristics. We evaluate different combinations of signals and find that using multiple signals can be beneficial for disentanglement. We also caution that the choice of supervisory signal(s) is important, with some choices leading to worse disentanglement. Finally, we also compare our method to other approaches for obtaining a low-dimensional representation in the literature, including standard and variational autoencoders in Web Appendix **??**, and find that none of the compared methods produce human-interpretable characteristics.

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 market settings. It does not require labeled data on visual characteristics, and is designed to leverage typically available structured characteristics. Second, the analyst 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 very low resolution images (like 128x128 pixels). Finally, our approach is not very computationally burdensome and can be applied in a scalable manner across different product categories.

Application and Results: We apply our proposed method on two product categories where visual design is known to be relevant. We use watches as the primary product category, and also test the method using sneakers as a second unrelated product category. The disentanglement method on the watch dataset (both images and structured product characteristics) automatically discovers and quantifies 6 *interpretable* visual characteristics of the watches. These discovered characteristics correspond to 'dial size', 'dial color', 'strap color', 'rim (bezel) color', 'dial shape', and 'knob (crown) size'.[5] We then evaluate disentanglment performance and human interpretability of the automatically discovered and quantified visual characteristics. These visual characteristics are later used for quantifying consumer preferences and generating targeted "ideal point" product designs.

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

Evaluation: We evaluate our disentanglement method relative to benchmark alternatives in 4 different ways. First, we use a metric called Unsupervised Disentanglement Ranking (UDR) from the machine learning literature (Duan et al. 2020). We compare the UDR of supervised and unsupervised disentanglement, and find that across product categories, having access to these supervisory signals based on product characteristics improves disentanglement. Second, we examine human interpretability of the discovered visual characteristics by surveying users from the US using Prolific. We generate visual designs of watches by varying one dimension of the latent representation at a time, corresponding to one visual characteristic. When respondents are asked to determine whether these *changes* are human-interpretable and what the change represents, we find that on average, 86% of respondents agree on the corresponding visual characteristic, indicating that disentanglement helps lead to human-interpretable visual characteristics. Third, we examine whether the quantified level of the visual characteristic is human-interpretable, and find that human respondents and our disentanglement algorithm agree well (85%). Fourth, we obtain consumer preferences over visual characteristics using visual conjoint analysis by separately varying each visual characteristic. We then use these estimated preferences to predict consumer choices between pairs of watch designs on a holdout sample. We find that our method's representation with only six visual characteristics obtains higher predictive accuracy than representations learned from more complex machine learning models such as pretrained deep neural nets that have been trained on millions of images. Fifth, we generate new "ideal point" product designs for two consumer segments defined using estimated preferences. We show these new products align with segment visual preferences, and steal choice share from existing products. Finally, we test the generality of the approach by using the same model architecture in a separate and completely unrelated product category of sneakers. We find that a supervised approach achieves significantly higher disentanglement performance (UDR) than the unsupervised approach. However, a different combination of supervisory signals proves to be better in the sneakers application.

LITERATURE REVIEW

Visual design is instrumental in shaping consumer preferences, perceptions of value, and experiences across a range of categories. As noted in Bloch, Brunel, and Arnold (2003), "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).

While important, it is currently challenging to characterize and study visual design from a quantitative perspective. As Orsborn, Cagan, and Boatwright (2009) summarize, "... possibly even more challenging, user feedback requires objective measurement and quantification of aesthetics and aesthetic preference." This work used 7 *researcher-defined* visual design characteristics for automobiles and then quantified these characteristics using distances between components in the automobile's physical design specifications. Likewise, Landwehr, Labroo, and Herrmann (2011) and Kang et al. (2019) both morph visual style of automobiles by identifying feature points representing key design components, while (Liu et al. 2017) also used this approach to study the impact of product appearance on demand. Recently, Dew, Ansari, and Toubia (2022) and Burnap, Hauser, and Timoshenko (2023) use generative deep models for visual morphing over visual characteristics of logos and automobiles, respectively; however, both works still required definition and quantification over interpretable visual characteristics for use by logo or automobile designers. Broadly, current approaches require human experts to both identify and quantify visual characteristics.

In conceptual contrast, there is a rich literature on methods that aim at automatic, but not interpretable, summarization of data (e.g., MDS, PCA). These methods have been extensively used in marketing (DeSarbo, Ramaswamy, and Cohen 1995). We refer readers to Web Appendix **??** for a detailed overview of connections with existing marketing methods. Our disentanglement approach aims at *both* automatic and interpretable discovery, and quantification of visual charac-

teristics. This enables their use in common marketing tasks, which in our case involves visual conjoint analysis for generating novel counterfactual "ideal point" visual designs targeted to consumer segments.

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), representation learning means "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), and even in the data-driven design of logos, exploring their influence on brand personality (Dew, Ansari, and Toubia 2022).

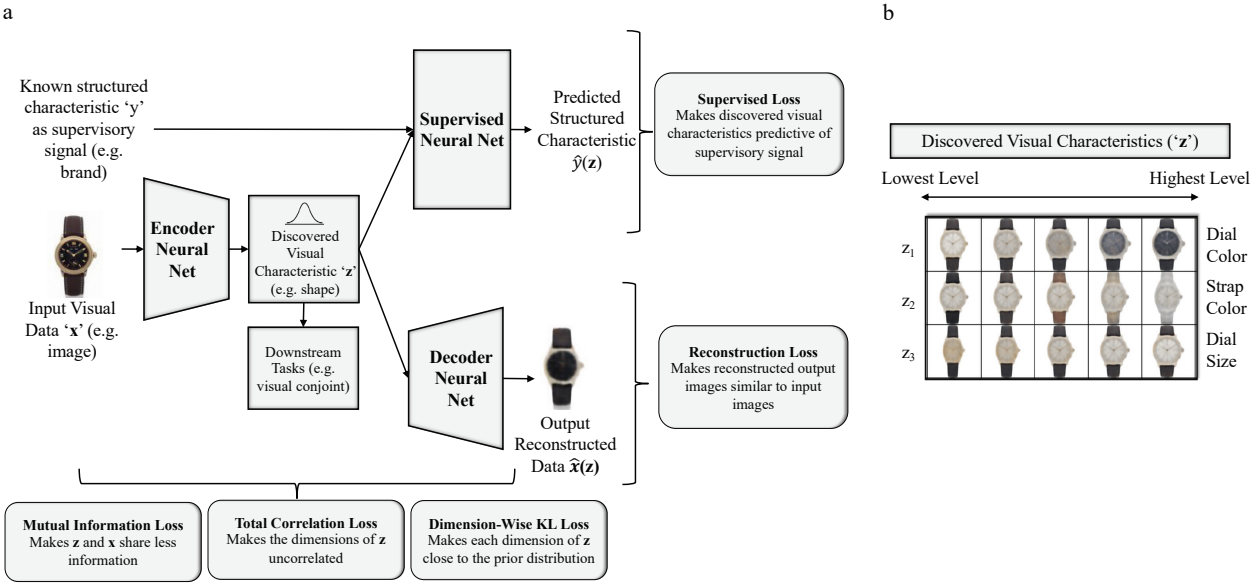
METHODOLOGY

Our proposed approach 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). An example of disentanglement with simple geometric shapes is provided in Web Appendix **??**. Disentanglement methods typically build on deep generative models such as variational autoencoders (VAE) (Kingma and Welling 2014) and generative adversarial networks (GAN) (Goodfellow et al. 2020).

The methodology developed here builds upon a VAE designed for disentanglement representation learning. Disentanglement refers to the process of decomposing complex data into independent, interpretable factors in order to better capture the true underlying relationships.[6] The

[6] Burgess et al. (2018) 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

Figure 2: (Color Online) Schematic of Proposed Disentanglement Approach



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. Note that (1) we show three discovered visual characteristics here for illustration purposes, and (2) this figure only shows disentanglement, not its later use in visual conjoint and generative visual design.

method is illustrated in the schematic depicted in Figure 2, and contains an encoder and decoder neural net. The encoder *encodes* visual data to discover a low-dimensional latent space of visual characteristics that are independent and human-interpretable. The discovered visual characteristics are then *decoded* to reconstruct visual representation of the input images using the generative model. The supervised version of the model also *predicts* a supervisory signal (e.g., brand) from the discovered visual characteristics. The model minimizes the weighted sum of 5 different type of losses — reconstruction loss, mutual 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 or a combination of product characteristics. We detail the notation used here in Table 1.

different independent latent units learn to encode different independent ground-truth generative factors of variation in the data."

Model: Supervised Variational Autoencoder with Disentanglement Losses

We first describe a variational autoencoder (VAE), the backbone model of our approach, its extension with disentanglement constraints and supervision. 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 data that can be used as supervisory signals, denoted $\mathbf{y}_i$.

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 vector
$\mathbf{Z(i)}$	Latent Space	Set of Latent Characteristics for model i
$p(\mathbf{z})$	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

The VAE uses a two-step data generating process $p(\mathbf{x}, \mathbf{z})$ (Kingma and Welling 2014). 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 $p_\theta(\mathbf{x}|\mathbf{z})$. The distribution p_θ 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. We add a supervised signal $\mathbf{y}_i$ that is predicted from the conditional distribution $p_w(\mathbf{y}|\mathbf{z})$, which is a function formed by non-linear transformation, with parameters w, of latent (visual) characteristics $\mathbf{z}$.

In practice, neural networks are estimated using optimization methods that result in point estimates of model parameters (Bengio, Courville, and Vincent 2013); in other words, they do not model uncertainty of the conditional distributions described above. Modeling the distribution of the visual characteristics $\mathbf{z}$ directly allows the characterization of distributional uncertainty over the space of possible neural networks (Blei, Kucukelbir, and McAuliffe 2017). The disentanglement approach uses the distributional aspect of modeling visual characteristics by setting distribution-level penalizations to encourage disentanglement (Kingma and Welling 2014; Chen et al. 2018). Importantly, for this paper, the modeling of distributions is critical to smooth generation of novel counterfactual images, since we are not restricted only to points that are observed in the data.

The VAE specifically builds on the variational Bayesian inference literature to incorporate neural networks within an approximate Bayesian framework (Blei, Kucukelbir, and McAuliffe 2017). In short, while the neural networks parameterizing the distributions of interest are estimated using point estimates of their parameters $(\theta, \phi, \mathbf{w})$, we learn full distributions over the 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. Given that the "true" unknown posterior $p(\mathbf{z}|\mathbf{x})$ is intractable, the variational Bayesian framework approximates this posterior to maximize a lower bound to, rather than, the likelihood of the posterior (and thus DGP) itself (Blei, Kucukelbir, and McAuliffe 2017). We adopt the conventional VAE assumption by parametrizing this approximate posterior with a multivariate Gaussian with diagonal covariance matrix 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 standard deviation 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 of the negative log-likelihood. In practice, this is specified as 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 typically in the range of hundreds of thousands to hundreds of millions depending on architectural choices (e.g., our specific 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. The reconstruction loss captures the differences between the reconstructed images generated by the decoder and the original inputs. Minimizing only this term would obtain a deep net that is able to generate images that match the input with high fidelity. The regularizer term ensures that the aggregate distribution of the latent variables does not deviate too much from the prior. This ensures that the latent space becomes structured and shares the properties of the prior distribution, such as compactness, smoothness and continuity.

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

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 $\beta > 1$. The idea is that disentangled representations are likely to be less complex and lower dimensional than entangled representations that also demonstrate statistical independence. The regularizer, which penalizes information capacity of the latent variables, therefore promotes disentanglement (Burgess et al. 2018).

Higgins et al. (2017) derive the β-VAE loss function as a constrained optimization problem.

Specifically, the goal is to minimize the reconstruction inaccuracy 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.

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

We can re-write Equation (2) as a Lagrangian under the KKT conditions (Karush 1939), where the KKT multiplier β is a regularization coefficient. This coefficient β is used as a hyperparameter to flexibly promote disentanglement, and results in the β-VAE formulation in Equation (3).

$$\min_{\theta,\phi} -\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + \beta\, KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right] \tag{3}$$

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 supervised versus unsupervised disentanglement.

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

The total loss is comprised of five terms weighted using hyperparameters, $(\alpha, \beta, \gamma, \delta)$. Adjusting these hyperparameters impacts the relative weight of each loss term and directly affects disentanglement performance. We detail the intuition for these loss terms below.[7]

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. We use L1 Loss (Absolute Error Loss) because unlike an L2 Loss (Squared Error Loss), it is more robust to outliers. Moreover, L1 loss introduces sparsity and thus, allows the model to focus on fewer important characteristics for reconstruction.

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 the 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

[7] 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 $\Omega = \{\beta\}$.

discovered latent variables $\mathbf{z}$ are independent, then the KL divergence is zero. More generally, a high penalty for the total correlation term forces the model to find statistically independent visual characteristics. A high β results in a more disentangled representation but with potentially worse reconstruction quality (and other loss terms).

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 $q(z_j)$ closer to the prior $p(z_j)$, encouraging the distribution of each latent dimension to not deviate from the prior (e.g., Gaussian) of each dimension. A high weight on this term reduces the number of discovered visual characteristics, and sets a higher bar for allowing an additional informative dimension. It ensures that each learned representations in the latent space has the desired properties of the prior distribution, such as compactness, smoothness, and continuity (Hoffman and Johnson 2016).

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 set the level of the hyperparameter δ by model selection, and note that $\delta = 0$ for the *unsupervised* disentanglement approach. Since our signals are discrete (e.g. brand), we use cross-entropy loss for the multiclass classification prediction task. Continuous signals like price are discretized using a quantile split to obtain discrete classes.

Supervised and Unsupervised Disentanglement

A key issue we examine in this research is whether structured product characteristics typically found in marketing contexts (e.g., brand, price etc.) can be used as supervisory signals to improve disentanglement, and also our ability to discover human-interpretable visual characteristics. Locatello et al. (2019), in a well-known impossibility theorem, showed that in the absence of a supervisory signal, disentangled representations are probabilistically equivalent to (an infinite set of) entangled representations. This finding implies that it is not possible to obtain a unique disentangled representation of the visual characteristics using an unsupervised approach. Locatello et al. (2020) further experimentally demonstrated that this challenge could be resolved by using

supervision with ground truth visual characteristics, in which lower supervised loss is connected to better disentanglement performance.

However, their approach of knowing ground truth for each of the visual characteristics across multiple products cannot be used for our goal of *automatic* discovery and quantification of visual characteristics. The ground truth labels corresponding to visual characteristics *are precisely what we are aiming to discover*. Moreover, we would need a researcher to apply their judgment and define visual characteristics in advance for the product category, as well as quantify each of them for the products in the dataset, implying the approach would not be automated. Our method instead posits that structured product characteristics and price might have information that correlates with visual characteristics, and using them as supervisory signals can be helpful in achieving disentanglement. Therefore, our method has a major advantage that it does not require access to ground truth characteristics.

Why might structured characteristics serve as good supervisory signals? Typical structured product characteristics commonly available in marketing data include brand, material, performance characteristics and price. Material more broadly is known to significantly affect visual appearance and consumer perceptions (Fleming 2014), e.g. being made of metal (like silver) provides a certain visual look. Similarly, brand can have a strong impact on the visual look. Consider, for instance the distinct look of a Mercedes-Benz car or a Louis Vuitton handbag. "Brand signature" is often perceptible in the visual design, especially for product categories with conspicuous consumption (Simonson and Schmitt 1997) and for luxury brands (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). Consumers can also recognize unique visual styles of brands (Ward et al. 2020). Next, consider the role of price. Many brands, especially luxury brands, maintain carefully curated pricing tiers with strong consumer associations, and in

many categories, high-priced products are viewed as having a "premium look" (Cho, Lee, and Saini 2022).

Evaluating Disentanglement Performance: To evaluate disentanglement performance, we need a metric that is applicable even when ground truth is not available, and therefore works for both supervised and unsupervised disentanglement. We evaluate disentanglement performance using a metric called Unsupervised Disentanglement Ranking (UDR) which satisfies the above requirements. UDR is a metric ranging from 0 to 1, with higher values representing more disentangled representations. UDR crucially allows for an automated way to select a model *when ground truth is not available* (Duan et al. 2020).[8]

The UDR metric is based on the assumption that representations obtained from models that are more disentangled would be more similar to each other than those from models that do not disentangle as well. This implies that given a dataset and a model, the visual characteristics learned using different random seeds (or different initial conditions) with a disentangled model should be similar, whereas every entangled representation is different in its own way and there are several ways to obtain entangled representations since the set of entangled representations is very large and potentially infinite. We note that whereas 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). If the disentanglement model is discovering the ground truth representation, then the initial parameters should not matter as much.

Defining UDR: Unsupervised Disentanglement Ranking (UDR) is defined for a pair of models i and j using Equation 6. For any pair of models i and j, UDR_{ij} is defined as a pairwise metric.

$$UDR_{ij} = \frac{1}{d_i + d_j} \left[\sum_{b \in \mathbf{Z(j)}} \frac{r_b^2}{\sum_{a \in \mathbf{Z(i)}} R(a,b)} I_{KL}(b) + \sum_{a \in \mathbf{Z(i)}} \frac{r_a^2}{\sum_{b \in \mathbf{Z(j)}} R(a,b)} I_{KL}(a) \right] \tag{6}$$

[8]Most existing metrics in the machine learning literature 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.

In the above equation, $R(a, b)$ is the correlation between the visual characteristic a that belongs to model i and the visual characteristic b that belongs to model j. We show the definition in Equation 7.

$$R(a, b) = cor(z_i(a), z_j(b)) \tag{7}$$

The term r_a is the correlation of the visual characteristic in model j that is most similar to the visual characteristic a in model i. In other words, r_a can be defined using Equation 8.

$$r_a = \max_{b \in \mathbf{Z}(j)} corR(a, b) \tag{8}$$

The right hand side of the Equation 6 has two terms inside the square bracket. The first term $\frac{r_b^2}{\sum_{a \in \mathbf{Z}(\mathrm{i})} R(a,b)}$ represents the ratio of the (squared) correlation of the visual characteristic a in model i that is most similar to visual characteristic b in model j, to the sum of the correlations across *all the visual characteristics* in model i. The squaring ensures that corner solutions or one-to-one mappings lead to higher UDR values, which is consistent with the idea of disentanglement. This term will be higher if there is a one-to-one mapping between one visual characteristic in model i and another in model j and the characteristics are statistically uncorrelated. The first term is then added across all the informative visual characteristics b of model j, which are represented by $I_{KL}(b)$ using a threshold for KL divergence between the characteristic's posterior and the prior. The second term represents the counterpart by considering one visual characteristic a that belongs to model i and then going through the corresponding process described above. Finally, we sum across all the informative visual characteristics a of model i, i.e. $I_{KL}(a)$.

We normalize this sum above by the total number of informative visual characteristics from model i and model j, denoted $(d_a + d_b)$. This is done to ensure that just having more informative characteristics does not mechanically lead to a higher UDR. Therefore, UDR_{ij} can be considered as the average correspondence in informative visual characteristics between two models i and j, and with a perfect and complete one-to-one correspondence, we will have $UDR_{ij} = 1$. We calculate the final UDR score for a particular hyperparameter configuration by averaging the UDR across all pairs of random seeds.

What Does UDR Capture? UDR captures the idea of similarity of two visual representations, which in turn are comprised of multiple visual characteristics. A pair of visual characteristics a and b from models i and j respectively, denoted $z_{i,a}$ and $z_{j,b}$ would be scored as highly similar if they axis align with each other (i.e., correlate) up to *permutation* and *sign inverse*. By permutation, we mean that the same ground truth factor c_k may be encoded by different visual characteristics within the two models $z_{i,a}$ and $z_{j,b}$ where $a \neq b$. By sign inverse, we mean that the two models may learn to encode the levels of the generative factor in the opposite order to each other, $z_{i,a} = -z_{j,b}$. Models that are identical except for sign inverse and permutation are isomorphic and equivalent from a representation learning viewpoint.

We additionally note that the UDR metric in Equation 6 is flexible enough to account for *subsetting*, i.e. non-overlapping subsets of visual characteristics that another model has learnt. While we did not observe this case in our empirical results, we found that changing the supervisory signal led to the discovery of different subsets of visual characteristics (see Web Appendix **??**). We note that differing hyperparameter settings resulted in models with different numbers of latent dimensions to be "switched off."

Operationalizing UDR: For each trained model, i.e. with $N_{seed} = 10$ random seeds, each of the representations obtained is compared pairwise with the others. Thus, we perform $\kappa = \binom{N_{seed}}{2} = 45$ pairwise comparisons with all other models trained with the same hyperparameters (β, δ), and the same vector of supervisory signals but with different seed levels. From these pairwise comparisons, we obtain UDR_{ij}, where i and j index the two models. UDR is then averaged across all combinations of i and j.

We next select informative visual characteristics and ignore uninformative visual characteristics. To implement this, we obtain 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. Rolinek, Zietlow, and Martius (2019) showed that during training, models based on VAEs enter a *polarized regime* such that

some latent variables (in our case, visual characteristics) switch off by being reduced to the prior $q_\phi(z_j) = p(z_j)$. This is due to the choice of a diagonal posterior. Typically, the dimensionality of the latent space is set higher than the expected true set of visual characteristics. This results in some of the characteristics being "switched off" or being very close to the prior distribution. These switched off characteristics are referred to as uninformative characteristics. Duan et al. (2020) showed that models with some uninformative characteristics tend to disentangle better and their unstructured (visual) characteristics are easier to semantically interpret.

Model Training, Selection, and Evaluation

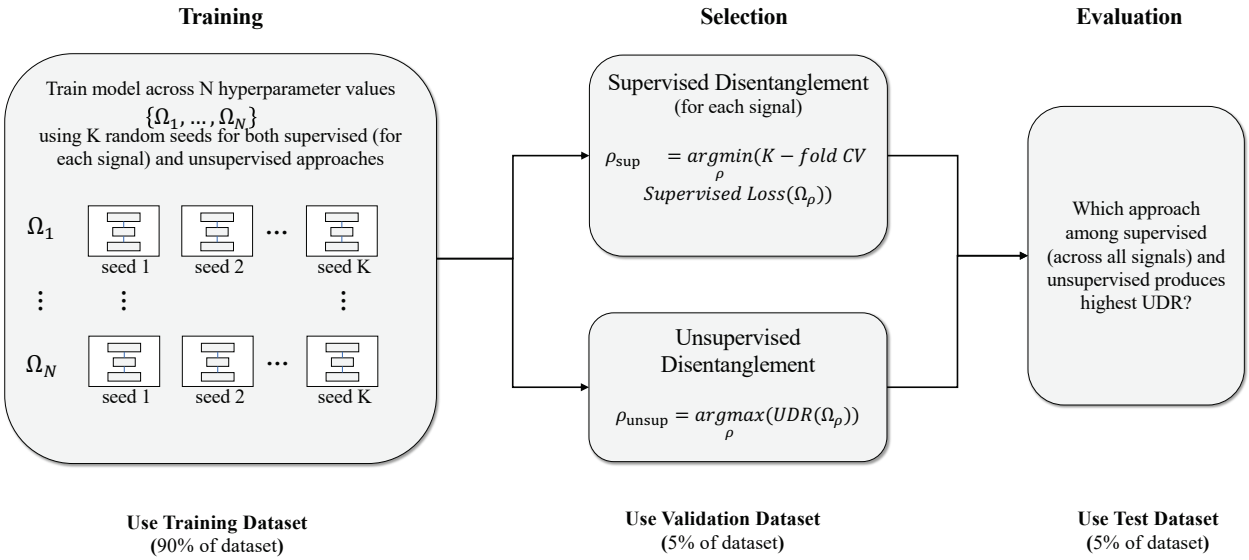
Both the supervised and unsupervised disentanglement approaches require model training (i.e., learning model parameters), model selection (i.e., choosing hyperparameters), and model evaluation (i.e., UDR disentanglement performance). However supervised and unsupervised approaches require different model training and selection steps, while the same evaluation step can we used, so we can compare them appropriately.

Model Training and Selection: We divide the dataset into (a) a training dataset for learning disentangled representations, (b) a validation dataset for model selection, and (c) a test dataset in the ratio $90 : 5 : 5$. To avoid data leakage, we ensure that each product is 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, and uses a subset of 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. Higher values of J do not result in any meaningful change in the discovered visual characteristics. We need to tune other hyperparameters including learning rate,

batch size and number of training steps or epochs.[9]

Figure 3: Model Training, Selection, & Evaluation



Notes: We train N different hyperparameter (Ω) levels for both supervised and unsupervised approaches. For supervised approaches, we choose the hyperparameter level that minimize the supervised loss $P(\hat{y}(\mathbf{z}), y)$ on the validation dataset. For the unsupervised approach, we choose the hyperparameter level that maximise the UDR. We evaluate different sets of visual characteristics learned by various approaches 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 prediction loss term).[10] In the unsupervised approach $\delta = 0$ by definition.[11] 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. For model evaluation, we compare all models using the UDR metric.

[9]The considerations for tuning hyperparameters detailed below are 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, whereas 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 $b_1 = 0.9$ and $b_2 = 0.999$; batch size as 64; number of epoch as 100.

[10]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.

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

Model Architecture: We modify the architecture used in Burgess et al. (2018) in order to use images with a resolution 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 that predicts the signals (structured product characteristics). Further details of the architecture are provided in Web Appendix **??**.

Generating New Visual Designs: We exploit the generative nature of the disentanglement learning model to controllably generate produce images. We feed the decoder of the disentanglement model a vector whose each dimension corresponds to a latent representation, $\mathbf{z}$. Recall that if the model achieves disentanglement, then $\mathbf{z}$ should be human-interpretable. More specifically, each element of the vector $\mathbf{z} = (z_1, z_2, \ldots, z_{J_{inf}})$ corresponds to a specific visual characteristic, e.g. dial color. Note that J_{inf} corresponds to the number of informative visual characteristics discovered by the model. Thus, when we choose values of the vector $\mathbf{z}$, the model is able to generate a visual design. We can thus controllably generate a wide range of visual designs corresponding to any specified vector $\mathbf{z}$. Since the decoder can take input at any point in the latent space, the model can generate novel visual designs not present in the original product image data. We show how these generated visual designs can be used for conducting visual conjoint analysis.

EMPIRICAL APPLICATION

We use our disentanglement method with an application to a dataset of watches. This dataset satisfies several desiderata detailed below. 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

Figure 4: (Color Online) Sample of Watches Auctioned at Christie's

and choice behavior (Kotler and Rath 1984). Second, we would like a market with a large number of products in order to 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

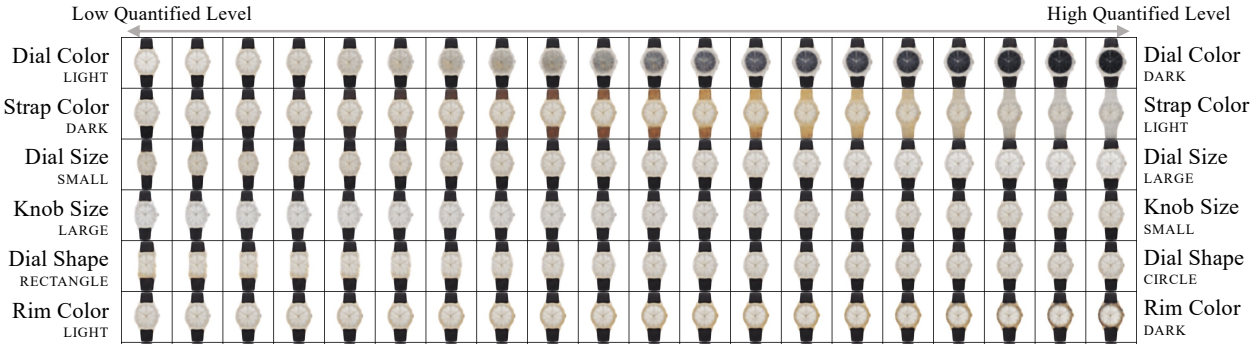
Our data includes 6,187 watches corresponding to 2,963 unique brand-models auctioned at Christie's auction house, spanning the years 2010 – 2020. The data on watches is particularly appropriate for the reasons above. For each auctioned watch in the dataset, we have its image, structured product characteristics, 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 4 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, Patek Philippe and Rolex are the four brands with the largest share of observations, while the remaining brands are coded as Others. Circa is coded as Pre-1950, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s and 2010s. Movement of a watch is classified as either mechanical, automatic or quartz. Dimensions of the watch refers to the watch diameter in case of a circular dial or the length of the longest edge in case of a rectangular dial (in millimeters). Material is coded as gold, steel, a combination of gold and steel or other materials. Summary statistics of the data are provided in Web Appendix **??**.

Results: Discovered Visual Characteristics

Figure 5 illustrates the output of the disentanglement model with supervisory signals "Brand + Circa + Movement," showing discovered visual characteristics corresponding to the signals with the highest UDR. 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 leads to no visual changes, and the posterior distribution of the discovered latent variable is almost identical to pure Gaussian noise. From ex-post human inspection (by researchers), we observe six distinct visual characteristics that are independent as well as human-interpretable. These are labeled 'dial color', 'strap color', 'dial size', 'knob (crown) size', 'dial shape', and 'rim (bezel) color'.

Figure 5: (Color Online) 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, circa and movement simultaneously.

Figure 6 shows the density plot of these discovered visual characteristics. All visual characteristics are initially modeled by a standard normal prior distribution. In the training process, each visual characteristic is encoded in the representation as a continuous distribution. If the algorithm finds a lot of variation along the visual characteristic in the image data, then we would observe the variance of that characteristic to increase. In contrast, if the algorithm finds little variation on some visual characteristic, e.g. if all watches have circular dials, the posterior distribution for this

dimension would have a low variance. It is important to note that we do not artificially constrain the scale of the visual characteristics, and allow the model to discover it from the data. Summary statistics of the visual characteristics are provided in Web Appendix **??**.

The sign (negative or positive) is arbitrary. For example, with dial size, negative might imply large dials, whereas positive might imply smaller dials. The sign might also be reversed, and both such representation would be equally valid (in fact, isomorphic). As the literature on disentanglement has pointed out, representations with different permutations (of latent dimensions) and signs are equivalent (Duan et al. 2020). Thus, the numbers corresponding to the latent dimension do represent "size" in the image in a monotonic sense. Thus, knowing the latent visual characteristic can permit the model to generate a product image with a specific size, or the inverse.

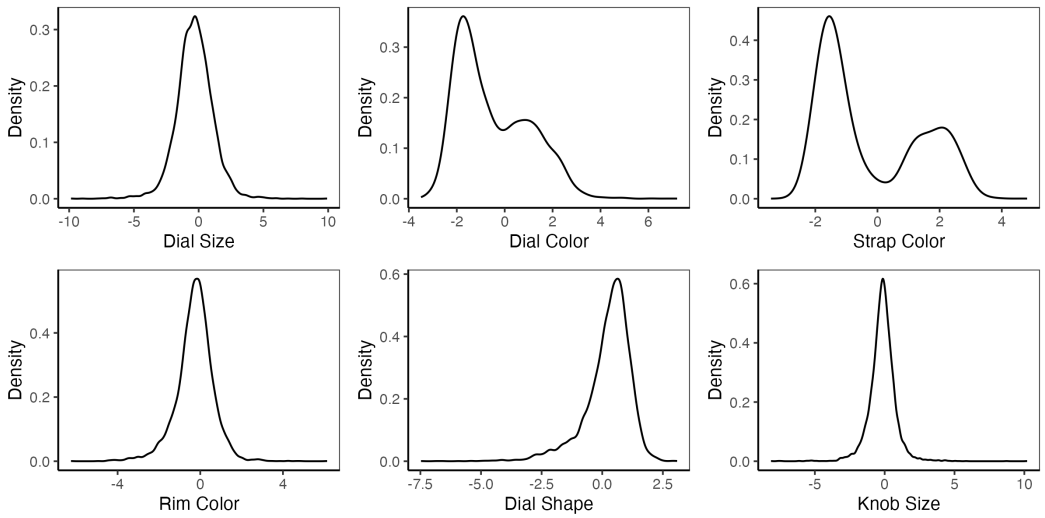
Finally, we show that the discovered visual characteristics are not highly correlated (Table 2), consistent with the goal of maintaining statistical independence across the latent dimensions. In contrast, an autoencoder is not able to find any disentangled visual characteristic and a plain-vanilla variational autoencoder finds entangled visual characteristics. Please refer to the Web Appendix **??** for these results.

Table 2: Correlations Between Visual Characteristics

	Dial Size	Dial Color	Strap Color	Rim (Bezel) Color	Dial Shape	Knob (Crown) Size
Dial Size	1.00	0.17	-0.09	0.04	0.02	0.00
Dial Color	0.17	1.00	0.06	0.01	0.03	0.01
Strap Color	-0.09	0.06	1.00	0.10	0.03	-0.06
Rim (Bezel) Color	0.04	0.01	0.10	1.00	0.07	0.04
Dial Shape	0.02	0.03	0.03	0.07	1.00	0.01
Knob (Crown) Size	0.00	0.01	-0.06	0.04	0.01	1.00

Evaluating Models with UDR: 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 comparison of different supervisory signals for disentanglement learning are detailed in the Web Appendix **??**.

Figure 6: Density of Discovered Visual characteristics (from 'Brand+Circa+Movement' Signal)



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

We find that including a combination of signals, i.e. brand, circa, and movement, was substantially better (UDR = 0.414) than the unsupervised approach (UDR = 0.131). We also note that additional supervision might not always help, because the classification problem of predicting a combination of *all* signals correctly can become more challenging. We show the discovered visual characteristics from the unsupervised approach as well as from supervised approaches corresponding to the supervisory signal combination with the lowest and the highest UDR in the Web Appendix **??**.

We find that supervision can help even in the absence of ground truth on visual characteristics by using structured product characteristics as supervisory signals. However, the specific combinations of signal(s) that would work better is likely to depend to a significant degree on the details of the empirical setting, including the product category and potentially even the resolution of the product images.

Effectiveness of Supervisory Signals in Disentanglement

We next aim to develop an understanding of why some signals might be good for supervision. Consider what is required for a signal to work well for disentanglement. Let's start with why

ground truth works, when the data images have been generated perfectly from different values of this ground truth factor. The supervised loss term in the objective uses the visual image to predict the signal, when there is only a single visual dimension that is varying in the data (image) generating process. The objective is to minimize prediction loss (e.g. MSE). If we use ground truth on a specific visual characteristic (e.g. dial color), the disentanglement algorithm is incentivized to find that visual characteristic as a discovered latent dimension, since doing so would allow it to reduce the supervised loss, all else equal.

A similar logic holds when we have a good signal that is correlated with the ground truth. The presence of the supervisory signal incentivizes the algorithm to find the specific latent dimension corresponding to the visual characteristic. Now, in conjunction with the penalization on total correlation, the algorithm is unlikely to entangle it with other factors since there is an incentive to find orthogonal (or statistically uncorrelated) latent dimensions. A higher quality (or stronger) signal would improve the incentive to find the visual characteristic that predicts the signal as a separate dimension, and also improve disentanglement by the above logic (Khemakhem et al. 2020).

With ground truth, there is a one-to-one mapping between each ground truth signal and a specific latent dimension or discovered visual characteristic. However, with imperfect signals, there is a many-to-many mapping between these signals and the true visual characteristics. Thus, while the above logic holds, there are some additional tradeoffs. For instance, if one supervisory signal impacts multiple true visual characteristics, then the algorithm would have to trade off the improvements in predictive accuracy across multiple dimensions. If there are multiple such signals that are predictive of one visual characteristic, the model would also have to weigh improvements across each of them in terms of predictive accuracy.

Broadly speaking, signals that are more strongly correlated with the visual look of the product would prove to be better signals. Signals that are more likely to strongly predict one of the visual characteristics are likely to perform well, even if they do not predict all of the visual characteristics. In contrast, signals that weakly predict multiple visual characteristics are less likely to work well.

Also, a set of signals would work better when they encode different information, i.e. each signal in the set would be strongly correlated with one separate dimension of visual characteristics, but not other dimensions. Beyond this broad logic, it is an empirical question as to which signals work better. For some product categories, brand – for instance – might work well if brand influences the look. However, in other categories, where a brand might include several different products without a common look, then brand might not be a good signal.

To evaluate whether a signal is effective, we quantify the degree to which the representation obtained separates out the visual characteristics when conditioned by different values of the signal. To implement this concept, we first select the most disentangled representation using UDR across all possible supervisory signal combinations. We then compare the distribution of these visual characteristics across different values of the signal, operationalized by the Jensen-Shannon (JS) distance. [12]

Let $\mathbf{z_{inf}}$ be the set of informative latent variables. Denote the set of values taken by a supervisory signal i as $y_i \in \mathcal{Y}_i = \{1, 2, \ldots, Y_i\}$. For example, say signal $i = 1$ is brand, and the values it can take include *Patek Philippe* and *Rolex* and *Cartier* etc. Signal $i = 2$ say is price with values *High* and *Low*. We define the *signal effectiveness* $\mathcal{S}_i$ of a supervisory signal i as:

$$\mathcal{S}_i = \frac{1}{2J_{inf}|Y_i|\left(|Y_i|-1\right)|} \sum_{k \in \mathbf{z_{inf}}} \sum_{l \in \mathcal{Y}_i} \sum_{m \in \mathcal{Y}_i : m \neq l} JS(p(z_k|y_i = l), p(z_k|y_i = m)) \tag{9}$$

The intuition is that better or *more informative signals will generate more separation* in latent visual characteristics. Consistent with this intuition, we find that Brand obtains a signal effectiveness of 0.24, whereas Price has a lower signal effectiveness of 0.13. This implies that the difference in the distribution of visual characteristics across watches corresponding to different brands is greater than the difference in the distribution of visual characteristics across low and high prices.

[12] The Jensen-Shannon distance (JS distance) is a symmetric and smoothed version of the Kullback-Leibler divergence (KL divergence). It measures the similarity between two probability distributions. Given two probability distributions P and Q, the JS distance is defined as: $\text{JS}(P, Q) = \left(\frac{1}{2}\text{KL}(P||M) + \frac{1}{2}\text{KL}(Q||M)\right)^{\frac{1}{2}}$, where $\text{KL}(X||Y)$ is the Kullback-Leibler divergence of X from Y, and $M = \frac{1}{2}(P + Q)$. Note that JS distance is always bounded between 0 and $\sqrt{\log 2}$.

Price as a Signal: Price could well be an effective signal in other empirical settings. In our case of luxury watches, the price has some unique properties that might reduce its effectiveness.[13] In other empirical settings without these specific considerations, price could well serve as one of the better signals. We show results for a separate product category, sneakers, in which price is a better signal for disentanglement in Web Appendix **??**.

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.[14] In the first survey, we evaluate the interpretability of the discovered characteristics from visual data. We show respondents an image illustrating different parts of the watch before the survey to help them understand the visual elements of the product.[15]

Next, we generated counterfactual images that vary along only one visual characteristic. For example, each watch image (see Figure 7) 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 they scan the images 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 second column of the Table 3, we report the percentage of respondents in the survey who agree with each other on which part of the watch is changing.

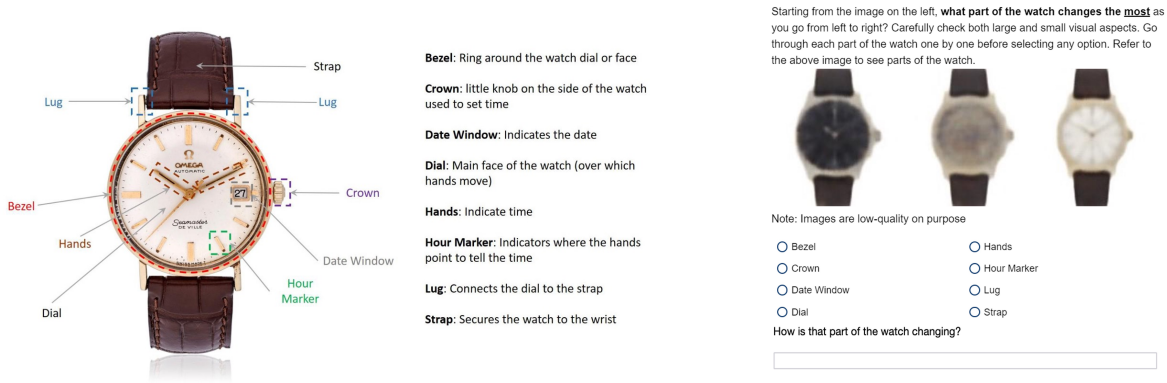
We next examine in a second survey (Figure 8) whether the quantification of the characteristics automatically determined by the method was consistent with human interpretation. We gener-

[13] First, luxury watches are expensive products, and no low cost watches were included in our dataset. Second, hammer prices are based on auction outcomes and hence can be driven by a small number of bidders. These buyers may not be price sensitive, and hence price might not be an informative signal. Third, the same model of a watch can be auctioned multiple times, leading to variation in prices within model, and therefore a noisier price signal.

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

[15] 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 7: (Color Online) Survey Question to Validate Interpretability



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 third column of Table 3.

Figure 8: (Color Online) Survey Question to Validate Quantification

Table 3: Human Interpretation of Visual Characteristics and Quantification

Visual characteristic	Interpretability Survey	Quantification Survey
Dial Size	76%	83%
Dial Color	80%	92%
Strap Color	88%	92%
Rim (Bezel) Color	79%	88%
Dial Shape	87%	68%
Knob (Crown) Size	70%	85%

In addition to comparing the supervised and unsupervised disentanglement models using UDR,

we compare the interpretability of the visual characteristics produced by them. Table **??** in the Web Appendix **??** shows that supervised disentanglement models produce more human-interpretable visual characteristics.

Robustness

We examine the robustness of the model and findings as detailed below.

Different Product Category: We evaluate the disentanglement performance of our method with an unrelated product category of sneakers. We obtain data for over 2000 sneakers from Zappos, along with the structured product characteristics of price and brand. We find that our method, without any changes in architecture, is able to disentangle three human-interpretable visual characteristics: upper color, sole color and topline shape. These results are in the Web Appendix **??**. We find that for sneakers, price serves as a relatively good supervisory signal for disentanglement.

Alternative Approach: We evaluate an alternative approach of using SHAP-learned features as an input to the disentanglement model (Lundberg and Lee 2017). The idea behind SHAP is to determine the pixels in the image that are most influential in the classification of a data point. We find that the SHAP-based approach produces fewer number of visual characteristics than our existing approach of feeding the raw image data to the disentanglement learning model. These results are in the Web Appendix **??**.

VISUAL CONJOINT ANALYSIS TO GENERATE "IDEAL POINT" PRODUCTS

We next generate new "ideal point" products that are targeted to consumer segment preferences over the 6 disentangled visual characteristics. Specifically, we developed a conjoint survey and conducted a conjoint analysis to elicit customer preferences, segmented consumers based on their preferences, and generated novel "ideal point" visual designs that maximized segment preferences over the visual characteristics. Table 4 provides a 7-step high-level overview of this application from survey design to "ideal point" generation.

Table 4: Steps in Visual Conjoint Analysis and Generative "Ideal Point" Design

Step	Description
1	Conduct a visual conjoint analysis to elicit consumer choices over 729 generated visual designs across 6 visual characteristics.
2	Estimate consumers' visual preferences using a three-tiered hierarchical Bayesian model trained on conjoint analysis data.
3	Segment consumers into two segments using the estimated consumer preference relationship between consumer covariates (demographics) and visual characteristic.
4	Define the "existing market" as the Top-10 products by utility in the overall set of 729 existing products used in the conjoint survey.
5	Define segment-level "ideal points" in visual characteristics space for the two segments. The "ideal point" for each segment is defined as the norm-scaled average preference vector of the segment.[16]
6	Generate new "ideal point" designs corresponding to above visual characteristics.
7	Evaluate model predictions of consumer preference for generated "ideal point" designs by inferring how choice shares change for each segment in the counterfactual market of "existing + ideal point" products.

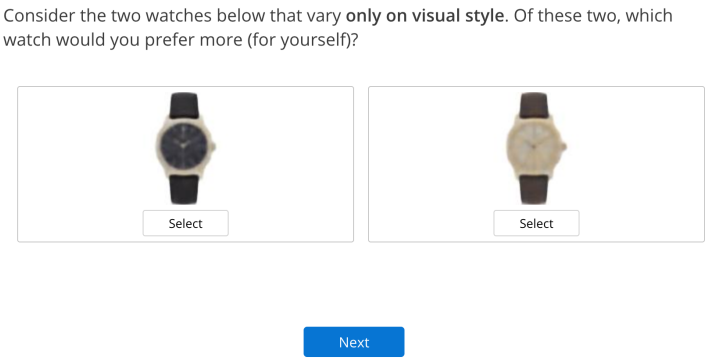
Conjoint Survey Design

We designed a choice-based conjoint (CBC) survey to elicit consumer preferences over a set of generated watches. 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 $3^6 = 729$ visual designs. We obtained CBC survey responses from 400 individuals through the Prolific platform, filtered to obtain a set of 253 respondents.[17] Each respondent evaluated 15 pairs of watches. The data includes binary responses for the 15 CBC questions, as well as respondents' covariates; namely, demographics and psychographics based on Likert responses to visual appearance.

The conjoint survey was designed with 7 survey stages. The conjoint survey stages are summarized along with their purpose in the Web Appendix **??**. Each CBC question consisted of a binary choice between two watch designs as shown in Figure 9. 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).

[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 9: (Color Online) Example choice-based conjoint (CBC) question in conjoint survey.



Conjoint Model Specification, Estimation, and Evaluation

Model Specification: We specify in Table 5 a three-level 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", "Knob (Crown) Size", "Rim (Bezel) Color"). We additionally included 6 respondent covariates denoted $\mathbf{r}$ ("Gender - Male", "Gender - Female", "Age", "Income", "Education", and "Aesthetic Importance").[18]

Table 5: Mathematical Representation of HB Conjoint Model for Visual Characteristics

Model Element	Mathematical Representation
Gaussian Hyperprior	$\mu_\Theta \sim \mathcal{N}(\mathbf{0}, \sigma^2_\Theta)$
Impact of Consumer Characteristics on Preferences	$\mathbf{\Theta} \sim \mathcal{N}(\mu_\Theta, \mathbf{\Lambda}_\Theta)$
Correlation of Preferences over Visual Characteristics	$\mathbf{\Omega}_\beta \sim \text{LKJ}(\eta)$
Preference Parameters	$\beta_i \sim \mathcal{N}(\mathbf{\Theta}^T \mathbf{r}_i, \mathbf{D}(\sigma_\beta)\mathbf{\Omega}_\beta\mathbf{D}(\sigma_\beta))$
Utility Function	$u_i^j = \beta_i^T z_j + \epsilon_{ij}$
Probability of choice j	$\psi_i(j, j') = \dfrac{\exp(u_i^j)}{\exp(u_i^j) + \exp(u_i^{j'})}$

Note that in Table 5, $\text{LKJ}(\eta)$ 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

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

parameter $\psi_i(j, j')$ 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^2_Θ, $\mathbf{\Lambda}_\Theta$, 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 between respondents' covariates and an upper-level model parameter matrix, $\mathbf{\Theta}$. 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 $\mathbf{\Theta}$ specified as a matrix of Gaussians to act as a population-level intercept term. We estimated this hierarchical model using MCMC sampling conditioned on observed consumer choices and their demographics. The HB model estimation details are in Web Appendix **??**.

Table 6: Population-averaged preference parameter β from individual-level β_i

Visual Characteristic	Preference Parameter β	Credible Interval
Dial Color	0.41	[0.23, 0.59]
Dial Shape	0.018	[-0.15, 0.18]
Strap Color	-1.7	[-1.9, -1.4]
Dial Size	0.36	[0.09, 0.6]
Knob (Crown) Size	-0.26	[-0.43, -0.095]
Rim (Bezel) Color	-0.60	[-0.77, -0.44]

Table 6 shows population-averaged estimated individual-level preference parameters β_i over the 6 visual attributes, along with credible intervals of 95% of posterior mass. Note that these are the estimated distributions of preference coefficients, not distributions of visual product characteristics. These plots were drawn by averaging individual-level respondent posteriors (averaging the posterior draws for each respondent to obtain individual-level posterior means, followed by

averaging across respondents). 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 modeling heterogeneity as well as the (implicit) assumption of the scale parameter being unity in logit estimation (Hauser, Eggers, and Selove 2019).

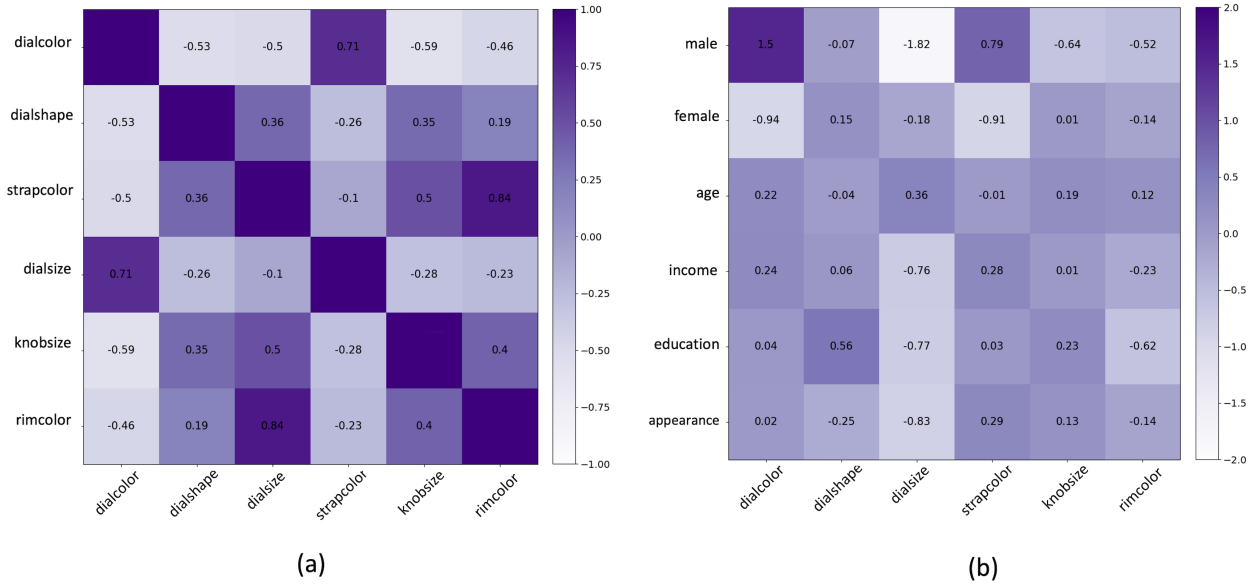
Figure 10(a) shows the correlation matrix of consumer preferences as a heatmap (i.e., normalized mean and standard deviation) over the 6 visual attributes. We find that the strongest correlation (0.70) in consumer preferences is between 'strap color' and 'rim color.' This implies that dark 'strap color' and dark 'rim color' are preferred together. The second and third strongest correlations (-0.43 and 0.41) are between 'dial shape' and 'dial color,' and 'dial size' and 'dial color,' respectively. This implies consumers prefer circular 'dial shape' with light 'dial color' and larger 'dial size' with dark 'dial color.'

We next analyze the relationship between respondents' covariates (demographics) and their preferences over visual attributes. Figure 10(b) 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 Table 5). This shows us for example how demographic variables like 'sex' correlate with visual characteristics like 'strapcolor.' START For example, respondents who indicated they are male, on average, preferred watches with a dark 'dial color', dark 'strap color,' and larger 'dial size'; older respondents prefer lighter 'strap color'; and respondents who indicated appearance is important for them (psychographic) prefer larger 'dial size.' END

Model Evaluation: We compare the predictive accuracy of our representation used along with the HB model against several benchmarks, and 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.,

Figure 10: Consumer Preferences. START (a) Correlation Matrix of Preferences across Visual Characteristic Ω_β (b) Interaction between Consumer Demographic and Visual Preferences Θ

END



DenseNets, VGG) and their performance on the *prediction accuracy metric*.[19] 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. Lastly, we benchmarked 3 nonlinear machine learning models as well as an HB model with pairwise interaction terms in an effort to assess how interactions between the visual characteristics influence consumer choice.

Table 7 reports out-of-sample hit rates. Out-of-sample splits were defined by holding out CBC conjoint tasks for each respondent (stratified splitting) as is convention in the conjoint analysis literature (Gustafsson, Herrmann, and Huber 2013) and preference learning in the machine learning literature (Fürnkranz and Hüllermeier 2010).

START We find the homogenous logit model achieves the lowest prediction accuracy, perhaps unsurprising given that out of all benchmarked models, it makes the strongest (implicit) assump-

[19]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

Model	**Out-of-Sample Hit Rate (SD)**
Disentangled Embedding + Logit Model (Homogeneity)	62.97% (2.90%)
Disentangled Embedding + Neural Net (Homogeneity)	65.81% (2.22%)
Pretrained Deep Learning Model Embedding (Observable Heterogeneity)	68.31% (1.54%)
Disentangled Embedding + Neural Net (Observable Heterogeneity)	67.52% (0.92%)
Disentangled Embedding + Random Forest (Observable Heterogeneity)	68.77% (0.90%)
Disentangled Embedding + XGBoost (Observable Heterogeneity)	69.10% (0.41%)
Disentangled Embedding + HB Model (+ Unobserved Heterogeneity)	**71.61% (1.87%)**
Disentangled Embedding + HB Model w/ Interactions (+ Unobserved Heterogeneity)	70.68% (1.35%)

Note: The homogeneity / heterogeneity we refer to in the Table above refers to *consumer-level* heterogeneity.

tions on the data and does not account for heterogeneity. The nonlinear machine learning models achieved the next highest hit rates, with random forests and XGBoost outperforming the two neural networks; namely a feedforward neural net on the visual characteristics and the ResNet50 pretrained deep learning model on the generated images. END START The HB model with a linear utility specification achieved the highest prediction accuracy, due to modeling both observed and unobserved consumer heterogeneity. END The HB MNL with interactions did not obtain a higher accuracy than the HB MNL model without the (explicit in likelihood) interaction terms. We believe this is likely due to two reasons. First, the HB model without interactions models correlations in consumer preferences across characteristics as we are estimating a full covariance matrix (i.e., not isotropic or diagonal). Second, we observed lack of convergence likely from model overparameterization. Our parametrization of the HB MNL model with interactions required us to model the (explicit) interaction parameters as being homogeneous (not conditional on covariates). Without this simplification, the number of parameters would increase substantially.[20] In short, we believe explicit modeling of interactions, specifically for this dataset, resulted in a less parsimonious model than the HB MNL model, resulting in worse out-of-sample performance. While this finding is in line with recent marketing research (Smith, Seiler, and Aggarwal 2023), this suggests more research into when conventional methods outperform machine learning methods is needed.

[20] Specifically, we would have to model the likelihood using explicit terms for the 6 main effect + 15 interaction effects, in addition to their covariance matrix, which would be of size (6 + 15) x (6 + 15).

Generating New "Ideal Point" Product Designs for Customer Segments

Developing new products and their product positioning is critical to profit-seeking firms in a competitive market (Rao et al. 2014). "Ideal points" refer to the optimal positioning of a new product in characteristic space based on preferences, often of a targeted consumer segment (DeSarbo, Ramaswamy, and Cohen 1995; Wedel and Kamakura 2000; Lee, Sudhir, and Steckel 2002). Identification of such ideal points has been extensively studied in marketing research and practice for over 50 years (Johnson 1971; Hauser and Urban 1977). The general approach involves the following steps: (a) obtain data on a consumer or segment 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 product characteristic space corresponding to the position of the maximally preferred product of the customer or segment.

We build upon this work by *generating* "ideal point" visual designs, in our case, maximally preferred watch designs for two chosen customer segments. Recent work in marketing has likewise used generative modeling to obtain preferred product designs (Dew, Ansari, and Toubia 2022; Cheng, Lee, and Tambe 2022; Burnap, Hauser, and Timoshenko 2023). The difference is that our method is based on interpretable visual characteristics *that were unknown a priori and discovered by our model*. Moreover, and critically for generative design, we can a vary any subset of them separately to create designs that span the space of visual characteristics. Interpretability is highly desirable and often required by practitioners for implementing these systems (Bloch 1995; Norman 2004).

We identify two customer segments to design "ideal point" products for from customer preferences estimated using the HB model on the conjoint survey data. Segment 1 corresponds to "affluent women" who self-reported they were female and made more than $100,000, and Segment 2 corresponds "less affluent men" who self-reported they were male and made less than $50,000. The variables and thresholds used segmentation were chosen via the HB estimated Θ matrix relating between customer covariates and visual characteristics as shown in Figure 10. We note that this segmentation thus serves as a proof-of-concept, with other segmentation approaches

possible.

Figure 11: (Color Online) Generated "Ideal Point" Watches for Two Segments

(a) Segment 1 **(b)** Segment 2

We next generate new visual designs for watches corresponding to the "ideal point" product (i.e., optimal visual characteristics) for each segments, and plot them in Figure 11. The "ideal point" refers to a point in the visual characteristic space corresponding to the maximal expected utility of consumers in a given segment, constrained to lie in a feasible portion of the visual characteristic space (DeSarbo, Ramaswamy, and Cohen 1995). "Feasible" must be defined given our (conventional) assumption of utility as an inner product between consumer preferences and (visual) product characteristics (i.e., "more is better"). We defined the "ideal point" product $\mathbf{z}_s$ for segment s as the segment's preference coefficients scaled to the average Euclidean norm ρ_c of the set of C "existing products" in the market.[21] Alternative models that instead *searches* the characteristic space via optimization methods could also be used (Michalek, Feinberg, and Papalambros 2005; Belloni et al. 2008).

The ideal point $\mathbf{z}_s$ is specified as:

$$\mathbf{z}_s = \frac{\rho_s}{\rho_c}\overline{\beta}, \quad \rho_s = \left\|\overline{\beta}\right\|_2, \quad \rho_c = \frac{1}{C}\sum_{j=1}^{C}\left\|\mathbf{z}_j\right\|_2 \tag{10}$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

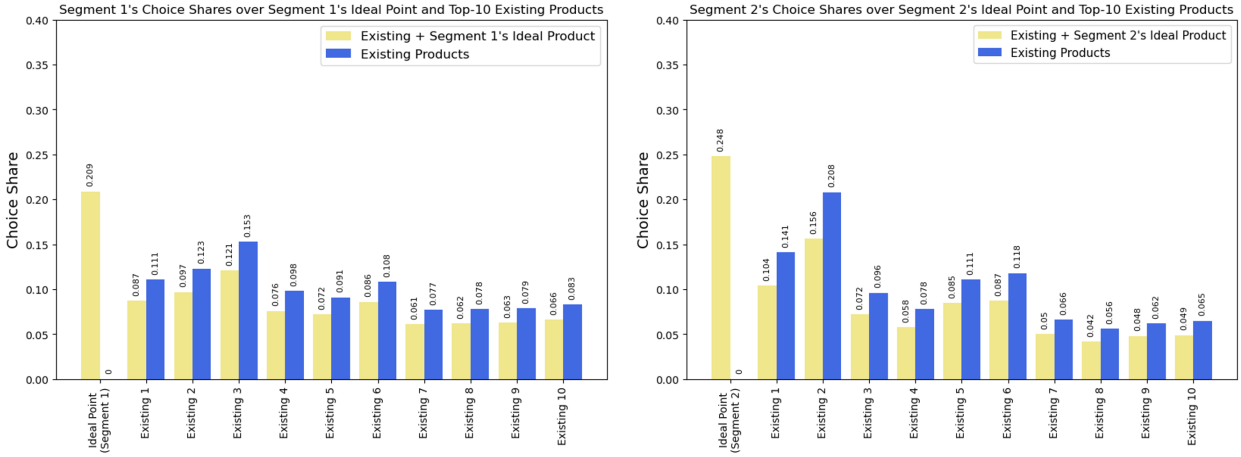
Lastly, we calculate the expected choice share of the adding the ideal point product for each segment to the market. For each segment, to simplify analysis, we assume $C = 10$ such that the

[21] This norm scaling $\frac{\rho_s}{\rho_c}$ is an assumption required for us to define an "ideal point" given that we have assumed the conventional inner-product-based utility model between consumer preference and product (visual) characteristics. Without this assumption (or a similar bound), the "ideal point" would be at infinity (Kaul and Rao 1995). Intuitively, this assumption and our definition of $\mathbf{z}_s$ is analogous to "ideal point" methods in that we are finding the location on a hypersphere in which inverse-distance from that point results in maximum utility for the segment (DeSarbo, Ramaswamy, and Cohen 1995; Hauser and Simmie 1981).

segment's consideration set consists of the Top-10 products by utility in the overall set of existing products (729 existing watches used in the conjoint survey). Since we have heterogeneity at the individual-level β_i, not every customer will have the same Top-10, so we defined the segment's Top-10 as the 10 watches that appeared most frequently when aggregated across individual customers. We note that with the given definition of ideal point, we may not always see the ideal point visual design having the highest choice share.

Figure 12 shows the change in expected choice shares for each segment's "ideal point" and the Top-10 existing products for the segment. We find 20.9% choice share for Segment 1's ideal point and 24.8% choice share for Segment 2's ideal point, signaling that the new ideal point product did indeed align with segment-level preferences. Thus, the "ideal point" generated product stole choice share from existing products for each segment. We note this analysis did not elicit (and subsequently estimate) individual consumer preferences for an "outside option," such that the analysis is limited to choice shares and not market shares.

Figure 12: Segment-Level Choice Shares With and Without Ideal Point Product



DISCUSSION AND CONCLUSION

Despite the importance of visual characteristics in marketing and business, the automatic identification and quantification of visual characteristics that represent visual design (and corresponding consumer response) has remained an open challenge. This is important as consumers have

preferences over visual design across a wide range of product characteristics (Bloch 1995). Marketing research has a long history of studying visual design, but only recently has had access to representations of visual characteristics that are realistic (e.g., images) while also being human-interpretable.

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, which are then used to generate targeted "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; our application used watch images matched to structured characteristics, but other applications may not have structured data as readily accessible. Second, although the model does not require human intervention, the data is preprocessed to ensure centering, similar size, background color, and orientation. Third, no algorithm can *guarantee* semantic interpretability of discovered visual characteristics, because that is a uniquely human ability (Locatello et al. 2019; Higgins et al. 2021). However, we validate our proposed method and find that it performs well quantitatively both with disentanglement metrics (UDR) as well as in human interpretability. Fourth, the performance of our (basic) model architecture likely varies with quality and resolution of images; richer characteristics in higher-resolution images may necessitate adjustments. Lastly, though literature heavily suggests the importance of visual stimuli in conjoint analysis (Dahan and Srinivasan 2000; Dotson et al. 2019; Sylcott, Orsborn, and Cagan 2016), future work could provide more direct comparisons between visual and traditional text descriptors.

There are several questions worthy of note for future research. First, it is important to understand the underlying reason why a particular product characteristic serves as a good super-

visory signal in any specific product category. Second, it would be useful to understand what combinations of product characteristics typically improve disentanglement the most *across* product categories, and the underlying reason. Likewise, developing neural network architectures with inductive biases for disentanglement would be valuable. Third, examining the performance of this or similar methods in other modalities like text or audio would help answer questions around practical usage for other marketing tasks. 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. Fourth, this paper raises interesting questions on when can a low-dimensional interpretable representation combined with conventional methods such as hierarchical bayesian outperform complex machine learning methods pretrained on a much broader set of image data (i.e. millions of images). Finally, 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's strategic choices involving visual design.

REFERENCES

Aaker, Jennifer L (1997), "Dimensions of Brand Personality," *Journal of Marketing Research*, 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.

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.

Belloni, Alexandre, Robert Freund, Matthew Selove, and Duncan Simester (2008), "Optimizing Product Line Designs: Efficient Methods and Comparisons," *Management Science*, 54 (9), 1544–1552.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013), "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8), 1798–1828.

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

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 (2018), "Understanding Disentangling in β-VAE," *Advances in Neural Information Processing Systems*.

Burnap, Alex, John R Hauser, and Artem Timoshenko (2023), "Product aesthetic design: A machine learning augmentation," *Marketing Science*, 42 (6), 1029–1056.

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

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

Cho, Hyun Young, Sue Hyun Lee, and Ritesh Saini (2022), "What Makes Products Look Premium? The Impact of Product Convenience on Premiumness Perception," *Psychology & Marketing*, 39 (5), 875–891.

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*.

Dahan, Ely and V Srinivasan (2000), "The Predictive Power of Internet-Based Product Concept Testing using Visual Depiction and Animation," *Journal of Product Innovation Management*, 17 (2), 99–109.

DeSarbo, Wayne S, Venkatram Ramaswamy, and Steven H Cohen (1995), "Market Segmentation with Choice-Based Conjoint Analysis," *Marketing Letters*, 6, 137–147.

Dew, Ryan, Asim Ansari, and Olivier Toubia (2022), "Letting Logos Speak: Leveraging Multiview Representation Learning for Data-Driven Branding and Logo Design," *Marketing Science*, 41 (2), 401–425.

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

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

Eastwood, Cian and Christopher KI Williams (2018), "A Framework for the Quantitative Evaluation of Disentangled Representations," *International Conference on Learning Representations*.

Fleming, Roland W (2014), "Visual Perception of Materials and Their Properties," *Vision Research*, 94, 62–75.

Fürnkranz, Johannes and Eyke Hüllermeier "Preference Learning and Ranking by Pairwise Comparison," "Preference Learning," pages 65–82, Springer (2010).

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 Patricia Simmie (1981), "Profit maximizing perceptual positions: An integrated theory for the selection of product features and price," *Management Science*, 27 (1), 33–56.

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 (2016), "Deep Residual Learning for Image Recognition," *Computer Vision and Pattern Recognition*, pages 770–778.

Heitmann, Mark, Jan R Landwehr, Thomas F Schreiner, and Harald J van Heerde (2020), "Leveraging Brand Equity for Effective Visual Product Design," *Journal of Marketing Research*, 57 (2), 257–277.

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 (2017), "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," *International Conference on Learning Representations*.

Hoffman, Matthew D and Matthew J Johnson (2016), "Elbo Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound," *Advances in Neural Information Processing Systems*.

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*.

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.

Khemakhem, Ilyes, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen (2020), "Variational Autoencoders and Nonlinear ICA: A Unifying Framework," *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217.

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

Kingma, Diederik P and Max Welling (2014), "Auto-Encoding Variational Bayes," *stat*, 1050, 1.

Kotler, Philip and G Alexander Rath (1984), "Design: A Powerful but Neglected Strategic Tool," *Journal of Business Strategy*, 5 (2), 16–21.

Kulkarni, Tejas D., William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum (2015), "Deep Convolutional Inverse Graphics Network," *Advances in Neural Information Processing Systems*.

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.

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.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020), "Visual Listening In: Extracting Brand Image Portrayed on Social Media," *Marketing Science*, 39 (4), 669–686.

Liu, Yan, Krista J Li, Haipeng Chen, and Subramanian Balachander (2017), "The Effects of Products' Aesthetic Design on Demand and Marketing-Mix Effectiveness: The Role of Segment Prototypicality and Brand Consistency," *Journal of Marketing*, 81 (1), 83–102.

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 (2015), "Deep Learning Face Attributes in the Wild," *Proceedings of International Conference on Computer Vision*.

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

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

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).

Michalek, Jeremy J, Fred M Feinberg, and Panos Y Papalambros (2005), "Linking Marketing and Engineering Product Design Decisions Via Analytical Target Cascading," *Journal of product innovation management*, 22 (1), 42–62.

Norman, Donald A (2004), *Emotional Design: Why We Love (or Hate) Everyday Things.* Civitas Books.

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," *Journal of Mechanical Design*, 131 (6), 061001.

Rao, Vithala R et al. (2014), *Applied Conjoint Analysis.* Springer.

Rolinek, Michal, Dominik Zietlow, and Georg Martius (2019), "Variational Autoencoders Pursue PCA Directions (by Accident)," *Computer Vision and Pattern Recognition*, pages 12406–12415.

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.

Simonson, Alex and Bernd H Schmitt (1997), *Marketing Aesthetics: The Strategic Management of Brands, Identity, and Image.* Simon and Schuster.

Smith, Adam N, Stephan Seiler, and Ishant Aggarwal (2023), "Optimal price targeting," *Marketing Science*, 42 (3), 476–499.

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.

Ward, Ella, Song Yang, Jenni Romaniuk, and Virginia Beal (2020), "Building a Unique Brand Identity: Measuring the Relative Ownership Potential of Brand Identity Element Types," *Journal of Brand Management*, 27, 393–407.

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.