Automatically Discovering Unknown Product Attributes Impacting Consumer Preferences

Ankit Sisodia, Alex Burnap, Vineet Kumar

Yale School of Management, ankit.sisodia@yale.edu, alex.burnap@yale.edu, vineet.kumar@yale.edu

January 2022

Marketing models typically focus on understanding how structured product attributes impact consumer preferences. However, obtaining attributes present in unstructured data (e.g. text or images), although important, rely on human (expert) judgment. Our research building on recent advances in disentangled representations (with statistical independence and semantic meaning) in deep learning aims to discover such attributes from unstructured data *automatically*, without human intervention. The recent deep learning literature has emphasized supervision of the discovered attributes on ground truth, as unsupervised approaches are not theoretically guaranteed to discover unique disentangled representations. Our approach does not require ground truth on the visual attributes, which is assumed in most of the literature. We instead use readily available structured product attributes as supervisory signals, and identify which signals work best. Our approach is practically useful since we do not need to specify the number of attributes or their meaning, yet we discover semantically interpretable and statistically independent attributes. We apply this approach to automatically discover visual product attributes of high-end watches auctioned at Christie's, and discover 6 semantically interpretable visual attributes providing a disentangled representation. We find that supervisory signals such as 'brand' promote disentanglement relative to the unsupervised approach, but surprisingly 'price' does not.

Key words: Attribute Discovery, Deep Learning, Disentanglement

1. Introduction

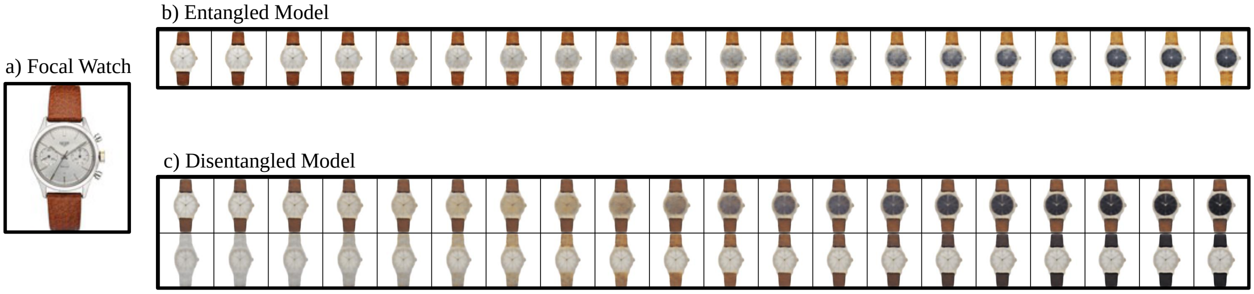
Product attributes form the basis of consumer choice and willingness to pay (WTP) for almost all products and services. At the market level, hedonic demand theory posits that market demand for a product is the aggregation of demand over its underlying product attributes (Lancaster 1966). Unsurprisingly, attributes form the foundation for quantitative models used in marketing and economics, for tasks ranging from quantifying consumer preferences (Guadagni and Little 1983), pricing products and services (Mahajan et al. 1982) as well as modeling competitive markets (Berry et al. 1995). For these tasks, the relevant set of attributes for the model must be defined in advance; for example, for a car this may include horsepower, fuel efficiency, and towing capacity.

However, attributes are typically *manually* defined using human judgment. The (human) modeler defines attributes using a combination of intuition and expertise, pre-defined attributes governed by availability of structured data, and qualitative methods based on consumer input like Voice of the Customer (Griffin and Hauser 1993). In some categories, the set of relevant attributes may be obvious and previously determined; however, in other categories, attributes potentially unknown to the researcher may strongly impact consumer valuation, and even the number of such unknown attributes may not be apparent. For example, attributes capturing a product's visual aesthetics are critical drivers of purchase for categories like fashion goods or automobiles, yet enunciating *why* a product looks appealing is challenging for researchers and consumers alike (Berlyne 1973).

Research Goal: Our research aims to *automatically* discover multiple (visual) attributes directly from unstructured *big data* (e.g., images), in which the attributes are semantically interpretable (i.e. make sense to humans). Our goal is complementary to widely-adopted yet *manual* approaches for attribute discovery, which either involve researcher judgment or consumer input through surveys and focus groups. It may be helpful to understand why it is desirable to have an automated method for attribute discovery. First, existing methods based on human judgment are inherently not scalable to marketing channels flush with unstructured data (e.g., a brand's social media, a product's textual reviews), potentially missing information required for defining attributes. Second, prior research has found that even experienced users or researchers may not be able to define such attributes (Berlyne 1973). Third, when the underlying attributes changes (e.g. a brand undergoes a design change or new products enter the market), it can be quite costly to repeat the process to obtain the new attributes (Karjalainen and Snelders 2010).

We automatically discover attributes using a methodology built on *disentangled* representation learning, an emerging area of deep learning that aims at identifying independent yet semantically-meaningful factors of variation within data (Bengio et al. 2013). By entanglement, we mean that a change in value across one discovered attribute impacts multiple semantically-interpretable

Figure 1 Example of Attribute Entanglement and Disentanglement



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

attributes, whereas a disentangled representation would result in a change to only one semantically-interpretable attribute, i.e. a one-to-one mapping. In our case, we disentangle (or separate out) attributes that are present but hidden and unknown in unstructured big data commonly found in marketing (e.g., images, text, videos), with the hope that we discover attributes beyond those typically found in structured data, e.g., brand. Figure 1 illustrates the difference between disentangled and entangled representations.

Advantages of Our Approach: Our approach provides practical advantages to both academics and practitioners. First, our disentanglement-based approach is designed to work with unstructured *big data* that would be practically obtainable in real managerial contexts (e.g., product images, textual reviews). Thus, marketing data like sales or consumer choices (typical dependent variables Y) and structured product characteristics (typically explanatory variables X) can now be augmented with *discovered* attributes hidden in textual reviews or product images (unstructured data). Second, the researcher does not need to define these (unknown) attributes in advance, and does not even need to specify the number of such attributes that must be discovered. Third, our approach also automatically determines the quantitative levels associated with each of the discovered unstructured (visual) attributes. Fourth, the output of the machine learning part of our approach can be embedded in a downstream task, e.g. a demand estimation model, to obtain a quantitative evaluation of how much each discovered unstructured (visual) attribute impacts an economic primitive (e.g., willingness to pay). Thus, our work complements recent work in marketing that has used deep learning methods to extract researcher-defined attributes from unstructured data such as images (Zhang et al. 2021b,a, Zhang and Luo 2018, Troncoso and Luo 2020) or directly used the unstructured data to find an outcome of interest such as return rate (Dzyabura et al. 2019) or brand attributes (Liu et al. 2020).

Methodology: Our work is connected with and situated in the context of multiple streams of deep learning literature. We provide an overview here and refer readers to Appendix A of the

Supplement for a comprehensive literature review. Recent advances in disentanglement within deep learning have primarily been developed using either generative adversarial networks (GANs) (Goodfellow et al. 2020) or variational autoencoders (Kingma and Welling 2014). InfoGAN (Chen et al. 2016) was perhaps the first GAN-based disentanglement method, and took an information-theoretic based approach that has been employed in more recent models such as Info-StyleGAN (Nie et al. 2020) built using StyleGAN (Karras et al. 2019). Generally, while GAN-based methods have much higher generative quality, they tend to learn comparably more entangled representations than VAE-based methods. A detailed comparison of the characteristics of both GANs and VAEs is provided in Appendix A.3 of the Supplement.

Since our primary goal is to infer disentangled representations from unstructured data rather than generating realistic unstructured data output, we base our disentanglement approach upon a VAE, which includes an encoder neural net and decoder neural net, both of which are parametrized by highly nonlinear deep neural networks. The encoder neural net takes high-dimensional unstructured data (e.g., images) as input and outputs a latent low-dimensional vector (i.e., discovered attributes), whereas the decoder neural net takes as input the low-dimensional vector and attempts to reconstruct the original data as output.[1] Deep autoencoders have found recent application in business and marketing; for example, Dew et al. (2021) used VAEs to study logo design and Malik et al. (2019) used conditional adversarial autoencoders to study the impact of beauty premium in human faces on career outcomes. However, these studies did not aim to obtain disentangled representations. To promote disentanglement, we instead penalize the *total correlation* of the discovered attributes such that they are statistically independent (Chen et al. 2018, Hoffman and Johnson 2016, Kim and Mnih 2018). Ideally, changes along any one of the discovered attributes leads the reconstructed image to change visually only along that one attribute.

A key challenge of *any* disentanglement approach is that there is no theoretical guarantee to discover a set of disentangled unstructured (visual) attributes even if independence is enforced, unless the disentanglement algorithm is supervised using ground truth labels of the unstructured (visual) attributes (Locatello et al. 2019). In other words, the labels (supervisory signals) we would need are the same attributes we are trying to discover in the first place.[2] Existing methods in deep learning have shown that learned attributes can be disentangled if unstructured data is labeled even if the labeling is partial, incomplete, or imprecise (Locatello et al. 2020). However, such labeling requires humans, contrasting with our goal of automatically disentangling attributes from data

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

[2] A bijective function of the attributes would work as well (Khemakhem et al. 2020).

typically found in marketing applications. Further, our work is motivated by marketing applications which typically do not have even partial ground truth of unstructured product attributes. Our method instead leverages supervision on readily available, complete, and precise structured data that are collected in marketing data sets. Specifically, the encoder neural net is additionally connected to a supervised neural net, thereby connecting the discovered visual attributes to structured attributes (e.g., brand). This enables us to study the issue of supervised vs unsupervised disentanglement; in particular, how typical structured attributes found in marketing may be used to improve disentanglement, and thus, automated attribute discovery.

Application and Results: We apply the proposed approach in the visual domain, with the goal of automatically discovering visual attributes of luxury watches. Our empirical context uses unique data from auctions of high-end luxury watches by Christie's spanning a 10-year period (2010 — 2020). We chose this application for a few underlying reasons. First, watches represent a product category where visual and design aspects captured in the images are likely to play an important role in consumer valuation and choice behavior (Kotler and Rath 1984). Second, as typical with marketing data, we have structured data appropriately matched up with the images. Third, the auction mechanism provides prices that represent the true valuation (WTP) of the highest bidder, which is an economic primitive of interest (Milgrom and Weber 1982).

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

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

Our results are in part unexpected. In our comparison of supervisory signals for disentanglement, we find that 'brand' helps but 'price' hurts disentanglement performance relative to an unsupervised approach. This is surprising as 'price' is one of, if not the most significant economic primitives affecting product design. In other words, since 'price' is often assumed as function of product attributes (and consumer preferences over those attributes), we expected using it as a supervisory signal would improve disentanglement—instead, it resulted in worse disentanglement than no supervision at all. Lastly, we provide an illustrative application in Appendix E of the Supplement of how one might use discovered visual attributes in a typical marketing application. In the application, we estimate the impact of visual attributes on consumer willingness-to-pay.

Contributions: We provide an automated approach that does not rely on human judgment to discover product attributes from unstructured data typically found in marketing. Such attributes could be used in competitive analysis, product positioning and pricing decisions by firms. In addition to the direct value of such discovery, and understanding how they impact consumers, they can have an indirect impact on structured attributes. When inferring economic valuation of structured attributes, we could have omitted variables (from unstructured data), leading to biased inference due to unobserved correlation between unstructured (visual) and structured attributes.

From a methodological perspective, our paper makes contributions on the issue of supervised versus unsupervised disentanglement in representation learning. First, we show that attributes can be discovered without access to ground truth from visual or unstructured data. We show how structured data typically available in marketing applications may be used as supervisory signals for obtaining better disentanglement. This aspect is useful as the machine learning literature often assumes the presence of ground truth supervisory signals, which are seldom available in real managerial applications. Second, and equally important, we demonstrate that just the idea of using any supervised signal might not work and may indeed backfire. The machine learning literature has focused on using a supervised approach due to known theoretical challenges with recovering a unique latent representation via unsupervised disentanglement. However, our research points out that, in practice, supervised learning may not be a panacea and that the choice of supervisory signal is critically important. In fact, many supervisory signals actually lead to *worse* disentanglement than using no supervision at all (i.e., unsupervised disentanglement).

Limitations: Our approach has several limitations worth noting and addressing in future research. First, it requires structured data to be matched to corresponding unstructured data. In our application the watch images are matched to corresponding structured characteristics and auction price, but other applications may not have such structured data. Second, although the algorithm does not require human intervention, the data is preprocessed once to ensure centering, similar size, background color, and orientation. Third, no algorithm can *guarantee* semantic interpretability for newly discovered features, because that is a uniquely human ability (Locatello et al. 2019, Higgins et al. 2021). However, in practice we observe that it performs well in a realistic and practical setting. It would be useful to test the semantic interpretability of the different models considered in our paper with humans to more thoroughly quantify semantic interpretability.

2. Methodology

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

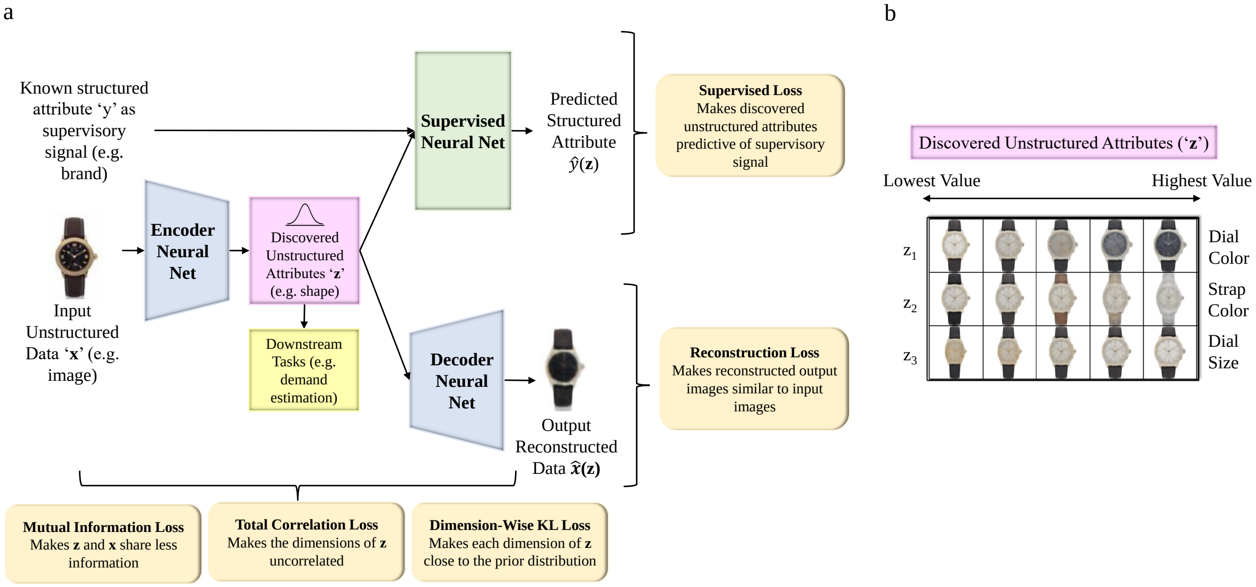
Our method is illustrated in Figure 2. We *encode* unstructured data (e.g. text or images) to discover unstructured (visual) attributes that are independent, low-dimensional and semantically interpretable (e.g., shape) and then *decode* the discovered unstructured (visual) attributes to reconstructed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

2.1. Model: Supervised Variational Autoencoder with Disentanglement Losses

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

We refer to $p_\theta(\mathbf{x}|\mathbf{z})$ as the decoder neural net, $q_\phi(\mathbf{z}|\mathbf{x})$ as the encoder neural net, and $p_{\mathbf{w}}(y|\mathbf{z})$ as the supervised neural net. As in variational Bayesian inference (Blei et al. 2017) the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is intractable, so we follow the original VAE assumption that the true posterior can

Figure 2 Schematic Illustration of Proposed Approach


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

Notes: **a**: The encoder neural net maps an input image into low-dimensional unstructured (visual) data attributes, 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 values of discovered attributes to visualise the semantic meaning encoded by single disentangled visual attribute of a trained model. In each row the value of a single visual attribute is varied while the other attributes are fixed. The resulting effect on the reconstruction is visualised. We show three discovered visual attributes here for illustration purposes.

be approximated using a variational family of Gaussians with diagonal covariance $\log q_\phi(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and the s.d. of the approximate posterior (Kingma and Welling 2014). We simultaneously train the encoder neural net, the decoder neural net, and the supervised neural net, by minimizing a variational bound to the negative log-likelihood. In practice, this results in a loss minimization problem to find point estimates of the neural network parameters, $(\theta, \phi, \mathbf{w})$, while inferring a full distribution over the discovered attributes, $\mathbf{z}_i \in \mathbb{R}^J$. The parameter space or number of weights of the deep neural networks in intended applications are often in the range of hundreds of thousands to hundreds of millions depending on architectural decisions (e.g., our chosen architecture in Appendix B of the Supplement 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 by starting with the loss for the original VAE in Equation (1), and refer readers to Kingma and Welling (2014) for its detailed derivation.

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

To learn disentangled representations, a recent model denoted β-VAE (Higgins et al. 2017) extends Equation 1 by imposing a heavier penalty on the regularizer term using an adjustable hyperparameer $\beta > 1$.[3] Intuitively, β-VAE uses the hyperparameter β to sacrifice reconstruction accuracy in order to learn more disentangled representations. We adopt this decomposition and further extend it by decomposing the regularizer term in Equation (1) into three terms (Chen et al. 2018, Hoffman and Johnson 2016, Kim and Mnih 2018). These three terms enable us to directly and separately control disentanglement constraints of the model as follows in Equation (4).

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

We finally add a supervised loss term to enforce the discovered attributes to help predict the supervised signal y in Equation (5). This enables us to study whether using typical structured data (e.g., ‘brand’) in a supervised approach helps improve disentanglement, and moreover, compare supervised disentanglement versus unsupervised disentanglement.

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

Our model's total loss is comprised of five loss terms weighted using the scaling hyperparameters, $(\alpha,\beta,\gamma,\delta)$. Adjusting these hyperparameters critically affects disentanglement performance

[3] Higgins et al. (2017) derive the β-VAE loss function as a constrained optimization problem. Specifically, the goal is to maximize the reconstruction accuracy subject to the inferred visual attributes being matched to a prior isotropic unit Gaussian distribution. This can be seen in Equation 2 where ϵ specifies the strength of the applied constraint.

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

We can re-write Equation 2 as a Lagrangian under the KKT conditions (Kuhn and Tucker 2014, Karush 1939), where the KKT multiplier β is a regularization coefficient. This explicit coefficient β is used as a hyperparameter (set by the researcher) to promote disentanglement, and results in the β-VAE formulation in Equation 3.

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

by adjusting the relative weight of each of the five loss terms, for which we detail the intuition below:[4]

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

Mutual Information Loss: $I_q(\mathbf{z},\mathbf{x})$ is the mutual information between the discovered unstructured (visual) attribute $\mathbf{z}$ and the product image $\mathbf{x}$. From an information-theoretic perspective (Achille and Soatto 2018), penalizing this term reduces the minimum amount of information about $\mathbf{x}$ stored in $\mathbf{z}$ that is sufficient to reconstruct the data by ensuring $\mathbf{z}$ does not store nuisance information. A low α would result in $\mathbf{z}$ storing nuisance information, whereas a high α results in loss of sufficient information needed for reconstruction. We follow the β-TCVAE approach (Chen et al. 2018) and impose $\alpha = 1$ and not more on this term to encourage the visual attributes to store the minimum amount of information about the raw data that is sufficient to reconstruct the raw data while not compromising on the reconstruction accuracy.

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

Dimension-Wise KL Loss: The dimension-wise KL loss term, $\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]$, penalizes the objective to push $q(z_j)$ to the prior $p(z_j)$ encouraging the probabilistic structure imposed by the parametric assumptions of the prior (e.g., Gaussian). This term promotes continuity in the latent space, which allows generation from a smooth and compact region of latent space. We follow the β-TCVAE approach (Chen et al. 2018) and impose $\gamma = 1$ to encourage the individual visual attributes to not deviate too much from the prior distribution while not compromising on the reconstruction accuracy.

Supervised Loss: Penalizing the supervised loss $P(\hat{y}(\mathbf{z}), y)$, where $\hat{y}(\mathbf{z}) \sim p_{\mathbf{w}}(y|\mathbf{z})$ prioritizes the discovered visual attributes $\mathbf{z}$ to obtain high accuracy in predicting y. We find the value of the

[4] 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 values 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 values of the hyperparameter set $\Omega = \{\beta\}$.

hyperparameter δ for the supervised disentanglement approach by model selection and set $\delta = 0$ for the unsupervised disentanglement approach. When the signal is discrete (e.g. brand), we use cross-entropy loss for the multiclass classification prediction task, and for a continuous signal (e.g. price), we use mean squared loss for the regression prediction task.

2.2. Supervised Disentanglement vs Unsupervised Disentanglement

A key issue we examine in this work is whether structured data variables typically found in marketing contexts (e.g., brand) can be used as supervisory signals to improve disentanglement, and thus our ability to discover unstructured (visual) attributes. (Locatello et al. 2019) showed that this is challenging in theory, as there is no guarantee for finding a unique disentangled representation using an unsupervised approach.[5] Locatello et al. (2020) further showed that this challenge could be resolved by using *supervision* with ground truth attributes, in which lower supervised loss is correlated with a high score on disentanglement performance metrics. However, their approach (and several related in machine learning) are not aligned with our goal of attribute discovery for several reasons. First, needing ground truth labels of the attributes conflicts with our goals as these labels are that of the attributes we are trying to discover in the first place. Second, if the approach requires humans to (partially) label attributes, then the approach is not fully automated. Third, disentanglement performance metrics used in machine learning are generally only usable for synthetic datasets with access to ground truth attributes (Higgins et al. 2021).

Our work instead takes an empirical viewpoint that is practical to marketing contexts. We do not assume access to ground truth attributes, and consequently, we measure disentanglement performance using two evaluation methods: (1) supervised classification accuracy on held-out data, and (2) Unsupervised Disentanglement Ranking (UDR). UDR is a metric proposed by Duan et al. (2020) to work in real-world data where ground truth is not available, and is based on heuristics of good disentanglement rather than theoretical guarantees. This metric posits that for a particular dataset and a particular VAE-based disentangled representation learning model, the unstructured (visual) attributes learned using different seed values should be similar, whereas every entangled representation is different in its own way. Specifically, UDR expects two disentangled representations learned from the same model on the same dataset with two different seeds to be similar up to permutation and sign inverse. Appendix C.2 of the Supplement has details on how UDR is operationalized.

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

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

We investigate how well each of the (six) supervisory signals lead to better disentanglement than the one learned by the purely unsupervised approach. We select informative visual attributes and ignore uninformative visual attributes by calculating the $KL\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]$ for each visual attribute and then select attributes with KL divergence above a threshold (Duan et al. 2020). Variation across an uninformative attribute would produce little to zero visual change in the image.[6] Both the supervised and unsupervised disentanglement approaches require model training (i.e., how model parameters are estimated), model selection (i.e., how model hyperparameters are chosen), and model evaluation (i.e., how good is a selected model's resulting disentanglement). However supervised and unsupervised approaches require different model training and selection steps, while having the same evaluation step (so the comparison between unsupervised and supervised approaches is apples-to-apples). We therefore describe these steps separately for the supervised and unsupervised approaches in Appendix C of the Supplement.

3. Empirical Application

We consider an application of our proposed approach in the visual domain. Examples of visually important aspects of products that impact consumer demand include a product's design, packaging and even promotion materials. Understanding their impact on consumer demand is of considerable interest (Kang et al. 2019, Burnap et al. 2019). Existing methods either ignore visual attributes completely, or collapse all visual (and other) unobservable attributes to form an unobserved product characteristic (Cho et al. 2015).

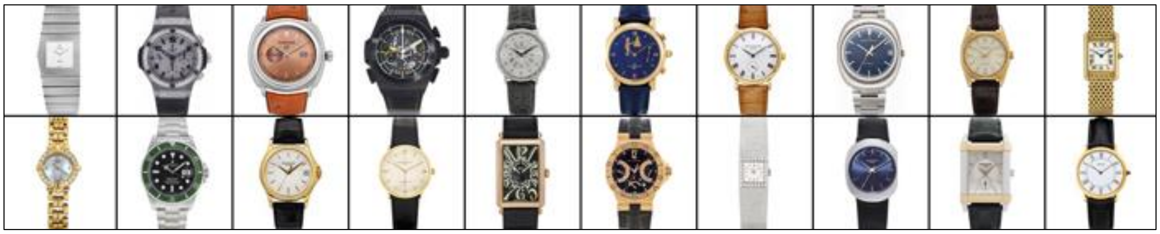
3.1. Data

Our data includes watches auctioned at Christie's auction house, spanning the years 2010 to 2020. We choose this data for two main reasons. First, visual attributes of watches are important considerations for consumers in this market. Second, the auction mechanism leads to a truthful revelation of the buyer's willingness to pay (WTP) for the watches.

For each auctioned watch in the dataset, we have its image, auction attributes, structured product attributes, and the hammer price paid at the auction (i.e., the willingness to pay). Structured attributes 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. Auction attributes are year of the auction and location of the auction. Figure 3 shows a sample of watch images in our dataset. From visual inspection, one can observe visual difference amongst

[6] Rolinek et al. (2019) showed that during training, models based on VAEs enter a *polarised regime* where many unstructured (visual) attributes are switched off by being reduced to the prior $q_\phi(z_j) = p(z_j)$. Entering this polarized regime allows the models to disentangle. These switched off attributes are referred to as uninformative attributes. Duan et al. (2020) showed that models with fewer uninformative attributes do not disentangle well and their unstructured (visual) attributes are hard to semantically interpret.

Figure 3 **Sample of Watches Auctioned at Christie's**



the watches in the size of the dial, shape of the dial, color of the dial, and color of the strap. The hammer price corresponding to a consumer's willingness to pay (in $1000s) are in constant 2000 dollars, adjusted for inflation using the Consumer Price Index.

The data includes both online and offline auctions. The offline locations are Dubai, Hong Kong and New York City. A total of 199 unique brands are present in the data. Audemar's Piguet, Cartier, Patel Philippe and Rolex are the four brands with the largest share of observations, while the remaining brands are coded as Others. Circa is coded as Pre-1950, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s and 2010s. Movement of a watch is classified as either mechanical, automatic or quartz. Dimensions of the watch refers to the watch diameter in case of a circular dial or the length of the longest edge in case of a rectangular dial (in millimeters). Material is coded as gold, steel, a combination of gold and steel or other materials. We create a time trend variable using the auction dates. Table 1 provides summary statistics of the auctioned watches.

3.2. Model: Training, Selection and Evaluation

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

3.3. Results: Supervised vs Unsupervised Disentanglement

We show an example output in Figure 4 of the discovered visual attributes corresponding to supervisory signals 'brand' and 'price' as well as the 'unsupervised approach'. In each row of the

Table 1 Summary Statistics of Structured Attributes of Auctioned Watches

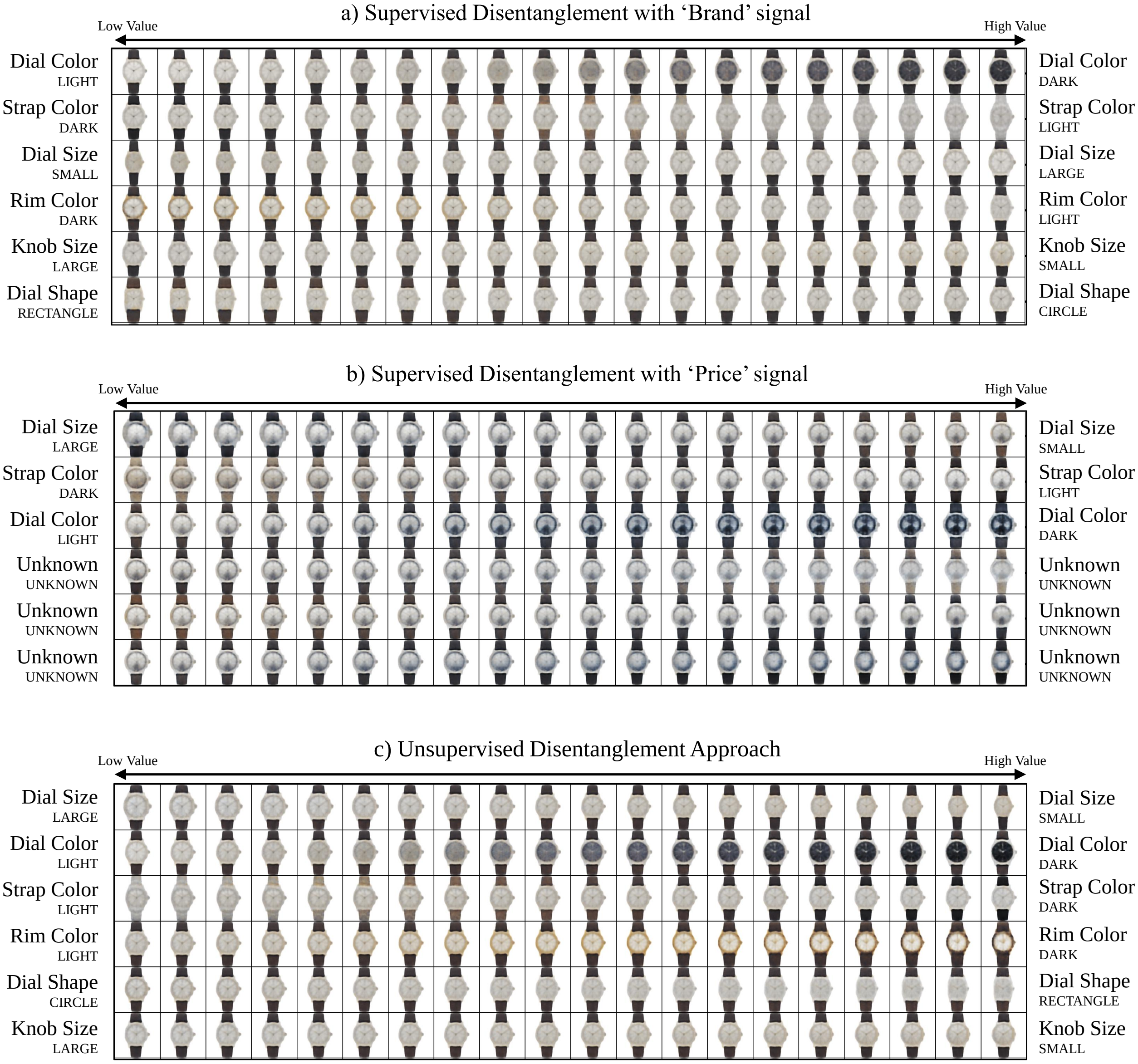
Statistic	Mean	SD	Min	Max
Auction Year	2015	2.99	2010	2020
Auction Location (Dubai)	0.13	0.34	0	1
Auction Location (Hong Kong)	0.37	0.48	0	1
Auction Location (New York City)	0.20	0.40	0	1
Auction Location (Online)	0.30	0.46	0	1
Brand (Audemar's Piguet)	0.06	0.24	0	1
Brand (Cartier)	0.07	0.25	0	1
Brand (Patek Philippe)	0.20	0.40	0	1
Brand (Rolex)	0.18	0.38	0	1
Brand (Others)	0.49	0.50	0	1
Circa (Pre-1950s)	0.05	0.21	0	1
Circa (1950s)	0.05	0.22	0	1
Circa (1960s)	0.07	0.26	0	1
Circa (1970s)	0.10	0.30	0	1
Circa (1980s)	0.08	0.26	0	1
Circa (1990s)	0.19	0.39	0	1
Circa (2000s)	0.33	0.47	0	1
Circa (2010s)	0.14	0.35	0	1
Movement (Automatic)	0.54	0.50	0	1
Movement (Mechanical)	0.36	0.48	0	1
Movement (Quartz)	0.11	0.31	0	1
Watch Dimensions (in mm)	36.21	6.83	9	62
Material (Gold)	0.60	0.49	0	1
Material (Gold and Steel)	0.05	0.22	0	1
Material (Steel)	0.28	0.45	0	1
Material (Others)	0.07	0.25	0	1
Timetrend (in days)	2,095	1,066	0	3,830
Hammer Price (in $000s)	23.25	55.18	1.00	950.20

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

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

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

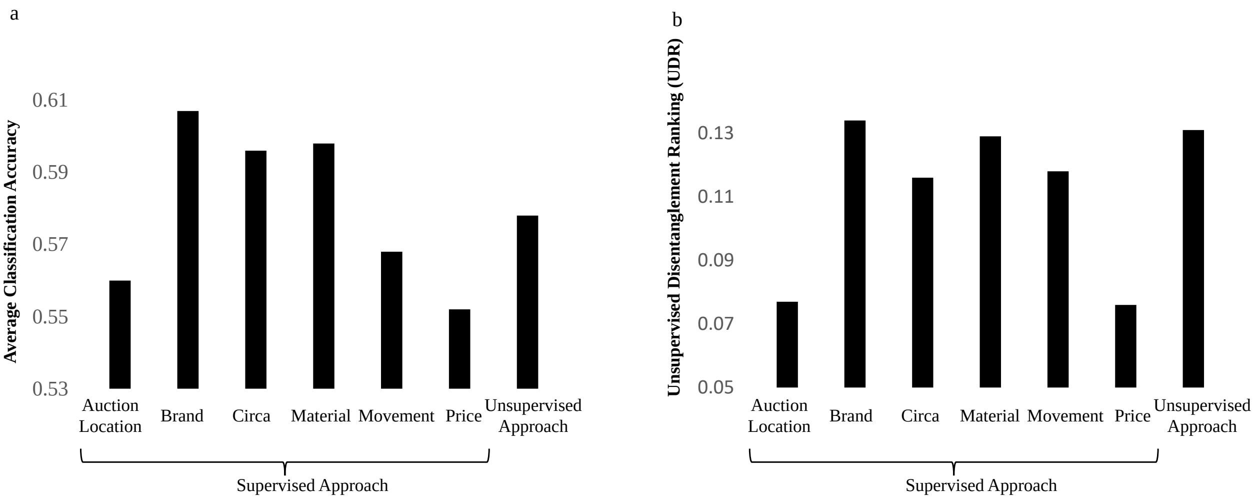
Figure 4 **Discovered Visual Attributes from Supervised and Unsupervised Approaches**



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

discovered visual attributes from each of the seven disentanglement models (six supervised disentanglement models and single unsupervised disentanglement model) are able to predict structured product attributes - namely brand, year of manufacture or *circa*, type of movement associated with the watch and the materials used in the watch. Using the trained models, we classify the watches in the test set. We then calculate the average accuracy for each supervisory signal. Second, we eval-

Figure 5 Disentanglement Performance of Supervised & Unsupervised Approaches



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

uate the models based on the UDR metric, where a higher UDR is preferable since it corresponds to a more disentangled representation.

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

From both these evaluation approaches, we find that supervision with a typical dependent variable in marketing such as 'price' does not necessarily lead to to the best disentanglement, whereas 'brand' serves as the best signal. Perhaps even more surprising, unsupervised disentanglement *in practice* leads to better disentanglement than price and other supervisory signals. In other words, while supervised signals are necessary for guaranteed disentanglement *in theory*, in practice we find that several supervisory signals lead to worse performance than unsupervised disentanglement. Further, conditional on using supervised methods, deep learning literature (Locatello et al. 2020) has assumed ground truth on the visual attributes as the supervisory signals. We use other structured attributes as the supervisory signal because obtaining ground truth on real-world datasets is not feasible. We show that supervising on structured attributes helps in discovering disentangled visual attributes. Thus, supervision can help even in the absence of ground truth on visual attributes.

It is useful to understand why unsupervised approaches to disentangle work in practice even though they are proven to have no theoretical guarantees. Rolinek et al. (2019) had shown that the loss function used in unsupervised approaches does not in itself encourage disentanglement. Indeed Locatello et al. (2019) showed that any rotationally invariant prior makes disentangled representations learnt in an unsupervised setting unidentifiable when optimizing the loss function for unsupervised approaches. However, Rolinek et al. (2019) showed that the interactions between the reconstruction objective and the enhanced pressure to match a diagonal prior created by the modified objectives of the disentangling VAEs force the decoder neural net to pursue orthogonal representations. During training, models based on VAEs enter a polarised regime where many unstructured (visual) attributes are switched off by being reduced to the prior $q_\phi(z_j) = p(z_j)$. Entering this polarized regime and ensuring sufficient dimensionality of the latent space are critical in allowing the models to disentangle.

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

4. Discussion and Conclusion

Marketers have typically defined attributes based on structured characteristics of products, e.g. brand, size and material. However, product attributes (especially visual ones) hidden in unstructured data play an important but underexamined role in marketing. We have proposed a methodology to automatically identify visual attributes from unstructured image data, which builds upon the disentanglement literature in machine learning. Our approach leverages typically available marketing variables like brand to act as a supervisory signal to obtain better disentanglement performance, in terms of both semantic interpretability and accuracy. We have demonstrated the potential value of our disentanglement deep learning method in an application examining how such unstructured (visual) attributes impact consumer preferences and market price obtained in an auction setting.

The theoretical result of Locatello et al. (2019) showed that multiple attribute representations are probable in the absence of a supervisory signal, leading to uncertainty about the true representation and resulted in the machine learning literature focusing mostly on supervised approaches for disentanglement. Our research examines this assertion empirically and finds that in practice, for automatic attribute discovery (in the absence of human intervention), supervised disentanglement can lead to worse results than unsupervised disentanglement. Rather, our results point to the idea that the choice of supervisory signal can be equally important. This finding was unexpected, particularly for the supervisory signal price (consumer WTP), as marketers often use WTP as the primary economic primitive in models in marketing and economics. Thus, even when the supervisory signal likely contains ground truth information about the visual attributes (i.e., if visual attributes impact WTP), if the supervision is weak, it may actually bias the process of recovery of unknown attributes and lead to worse disentanglement.

Our research has several limitations, some of which could serve as useful directions for future work. First, it would be helpful to understand the semantic interpretability more thoroughly by surveying consumers about what discovered attributes they viewed as more or less interpretable, and also which ones they had expected versus which ones were novel. It would also be useful to take this model to applications involving settings with firm choices, e.g. price setting or promotion. Second, our method can readily be extended to other modalities of data beyond images, e.g. text and voice or music data. Third, 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 better another, e.g. the presence of text might help disentangle images better. Finally, it would also be helpful to obtain more insight into the specific conditions under which certain signals perform better than unsupervised while others make disentanglement worse.

References

Aaker JL (1997) Dimensions of brand personality. *Journal of Marketing Rresearch* 34(3):347–356.

Achille A, Soatto S (2018) Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research* 19(1):1947–1980.

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. *International Conference on Machine Learning*, 214–223.

Batra R, Lehmann D, Singh D (1993) The brand personality component of brand goodwill: Some antecedents and consequences. *Brand Equity & Advertising: Advertising's Role in Building Strong Brands* 83–96.

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

Berlyne DE (1973) Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism* 31(4).

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518):859–877.

Burgess C, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, Lerchner A (2017) Understanding disentangling in β-vae. *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems.*

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

Chen RTQ, Li X, Grosse RB, Duvenaud DK (2018) Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 2615–2625.

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 2180–2188.

Cheung B, Livezey JA, Bansal AK, Olshausen BA (2015) Discovering hidden factors of variation in deep networks. *Workshop at International Conference on Learning Representations.*

Cho H, Hasija S, Sosa M (2015) How Important is Design for the Automobile Value Chain? *Available at SSRN 2683913* .

Dew R, Ansari A, Toubia O (2021) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science* .

Duan S, Matthey L, Saraiva A, Watters N, Burgess C, Lerchner A, Higgins I (2020) Unsupervised model selection for variational disentangled representation learning. *International Conference on Learning Representations.*

Dzyabura D, El Kihal S, Hauser JR, Ibragimov M (2019) Leveraging the power of images in managing product return rates. *Available at SSRN 3209307* .

Eastwood C, Williams CK (2018) A framework for the quantitative evaluation of disentangled representations. *International Conference on Learning Representations*.

Gabbay A, Cohen N, Hoshen Y (2021) An image is worth more than a thousand words: Towards disentanglement in the wild. *Advances in Neural Information Processing Systems* 34.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. *Communications of the ACM* 63(11):139–144.

Green PE, Krieger AM, Wind Y (2001) Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* 31(3_supplement):S56–S73.

Griffin A, Hauser JR (1993) The voice of the customer. *Marketing Science* 12(1):1–27.

Guadagni PM, Little JD (1983) A logit model of brand choice calibrated on scanner data. *Marketing Science* 2(3):203–238.

Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, Tsao D, Botvinick M (2021) Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications* 12(1):1–14.

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*.

Hoffman MD, Johnson MJ (2016) Elbo surgery: yet another way to carve up the variational evidence lower bound. *Workshop in Advances in Approximate Bayesian Inference, Neural Information Processing Systems*.

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

Kang N, Ren Y, Feinberg F, Papalambros P (2019) Form + function: Optimizing aesthetic product design via adaptive, geometrized preference elicitation. *arXiv preprint arXiv:1912.05047* .

Karjalainen TM, Snelders D (2010) Designing visual recognition for the brand. *Journal of Product Innovation Management* 27(1):6–22.

Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T (2020) Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, volume 33, 12104–12114.

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

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

Khemakhem I, Kingma D, Monti R, Hyvarinen A (2020) Variational autoencoders and nonlinear ica: A unifying framework. *International Conference on Artificial Intelligence and Statistics*, 2207–2217 (PMLR).

Kim H, Mnih A (2018) Disentangling by factorising. *International Conference on Machine Learning*, 2649–2658.

Kingma DP, Mohamed S, Rezende DJ, Welling M (2014) Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 3581–3589.

Kingma DP, Welling M (2014) Auto-encoding variational bayes. *International Conference on Learning Representations*.

Klys J, Snell J, Zemel RS (2018) Learning latent subspaces in variational autoencoders. *Advances in Neural Information Processing Systems*, 6444–6454.

Kotler P, Rath GA (1984) Design: A powerful but neglected strategic tool. *Journal of Business Strategy* 5(2):16–21.

Kuhn HW, Tucker AW (2014) Nonlinear programming. *Traces and emergence of nonlinear programming*, 247–258.

Kulkarni TD, Whitney WF, Kohli P, Tenenbaum JB (2015) Deep convolutional inverse graphics network. *Advances in Neural Information Processing Systems*, 2539–2547.

Kumar A, Sattigeri P, Balakrishnan A (2017) Variational inference of disentangled latent concepts from unlabeled observations. *International Conference on Learning Representations*.

Lancaster KJ (1966) A new approach to consumer theory. *Journal of Political Economy* 74(2):132–157.

Lee W, Kim D, Hong S, Lee H (2020) High-fidelity synthesis with disentangled representation. *European Conference on Computer Vision*, 157–174 (Springer).

Liu L, Dzyabura D, Mizik N (2020) Visual listening in: Extracting brand image portrayed on social media. *Marketing Science* 39(4):669–686.

Locatello F, Bauer S, Lučić M, Rätsch G, Gelly S, Schölkopf B, Bachem OF (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, 4114–4124.

Locatello F, Poole B, Rätsch G, Schölkopf B, Bachem O, Tschannen M (2020) Weakly-supervised disentanglement without compromises. *International Conference on Machine Learning*, 6348–6359 (PMLR).

Locatello F, Tschannen M, Bauer S, Rätsch G, Schölkopf B, Bachem O (2020) Disentangling factors of variations using few labels. *International Conference on Learning Representations*.

Mahajan V, Green PE, Goldberg SM (1982) A conjoint model for measuring self-and cross-price/demand relationships. *Journal of Marketing Research* 19(3):334–342.

Malik N, Singh PV, Srinivasan K (2019) A dynamic analysis of beauty premium. *Available at SSRN 3208162* .

Mathieu M, Zhao J, Sprechmann P, Ramesh A, LeCun Y (2016) Disentangling factors of variation in deep representations using adversarial training. *Advances in Neural Information Processing Systems*, 5040–5048.

Milgrom PR, Weber RJ (1982) A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society* 1089–1122.

Nie W, Karras T, Garg A, Debnath S, Patney A, Patel A, Anandkumar A (2020) Semi-supervised stylegan for disentanglement learning. *International Conference on Machine Learning*, 7360–7369 (PMLR).

Reed S, Sohn K, Zhang Y, Lee H (2014) Learning to disentangle factors of variation with manifold interaction. *International Conference on Machine Learning*, 1431–1439.

Ridgeway K, Mozer MC (2018) Learning deep disentangled embeddings with the f-statistic loss. *Advances in Neural Information Processing Systems*, 185–194.

Rolinek M, Zietlow D, Martius G (2019) Variational autoencoders pursue pca directions (by accident). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12406–12415.

Roweis S, Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Computation* 11(2):305–345, 00715.

Siddharth N, Paige B, van de Meent JW, Desmaison A, Goodman ND, Kohli P, Wood F, Torr PH (2017) Learning disentangled representations with semi-supervised deep generative models. *Advances in Neural Information Processing Systems*, 5925–5935.

Troncoso I, Luo L (2020) Look the part? the role of profile pictures in online labor markets. *Available at SSRN 3709554* .

Voynov A, Babenko A (2020) Unsupervised discovery of interpretable directions in the gan latent space. *International Conference on Machine Learning*, 9786–9796 (PMLR).

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

Zhang M, Luo L (2018) Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Available at SSRN 3108288* .

Zhang S, Lee D, Singh PV, Srinivasan K (2021a) What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science* .

Zhang S, Mehta N, Singh PV, Srinivasan K (2021b) Frontiers: Can an artificial intelligence algorithm mitigate racial economic inequality? an analysis in the context of airbnb. *Marketing Science* 40(5):813–820.

Electronic Companion Supplement

Appendix A: Literature

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

A.1. Attribute Discovery

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

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

A.2. Disentangled Representation Learning

Our work builds on a stream of literature in machine learning known as *disentangled* representation learning, which aims to separate distinct informative factors of variation in the data (Bengio et al. 2013). For example, a model to extract disentangled representations trained on a dataset of 3D objects might learn independent factors of variation corresponding to object identity, position, scale, lighting and color.

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

A.3. Generative Modeling using GANs and VAEs

The two broad classes of generative models are based on variational autoencoders (VAEs) (Kingma and Welling 2014) and generative adversarial networks (GAN) [7] (Goodfellow et al. 2020). Most state-of-the-art disentangled *representation learning* methods are based on VAEs. VAEs are comprised of two models – the encoder neural net and the decoder neural net. The encoder neural net compresses high-dimensional input data to a lower-dimensional latent vector (latent attributes), followed by inputting the latent vector to the decoder neural net which outputs a reconstruction of the original input data. VAEs balance having both a low reconstruction error between the input and output data (e.g., images, text), as well as a KL-divergence of the latent space distribution (latent attributes) from a researcher-defined prior distribution (e.g., Gaussian). The KL-divergence term acts as a regularizer on the latent space, such that it has desired structure (smoothness, compactness). VAEs are parametrized in both the encoder neural net and decoder neural net using neural networks whose parameters are learned jointly.

Several methods based on GANs have also been used for disentanglement. InfoGAN was one of the first scalable unsupervised methods for learning disentangled representations (Chen et al. 2016). While GANs are typically less suited relative to VAEs for representation learning, as GANs traditionally do not infer a representation[8], InfoGAN explicitly constrains a small subset of the 'noise' variables to have high mutual information with generated data. Several VAE-based methods have proven to be superior (Kim and Mnih 2018, Chen et al. 2018) than InfoGAN. Recent methods based on StyleGAN (Karras et al. 2019) such as Info-StyleGAN (Nie et al. 2020) are able to perform disentanglement at a much higher resolution (1024×1024) unlike the VAE-based methods. However, unlike InfoGAN, Info-StyleGAN suffers from the need for human labels or pretrained models, which can be expensive to obtain (Voynov and Babenko 2020).

We choose a VAE-based approach over a GAN-based approach for several reasons. First, our goal is to propose an easy-to-train method that can be used by researchers as well as practitioners (Lee et al. 2020). Second, our goal of discovering unique (visual) attributes that are semantically meaningful and independent of each other requires high disentanglement performance, but reconstruction accuracy is not our primary goal (Lee et al. 2020). GANs suffer from lower disentanglement performance because they focus on localized concepts but not global concepts of the image (Gabbay et al. 2021). On the other hand, discovered attributes from VAEs are much more globally distributed as compared with GANs. This allows the VAE-based methods to discover few important and semantically interpretable unstructured (visual) attributes that can represent the input raw data. Third, one of the benefits of our approach is that we are able to not just discover disentangle attributes, but infer the values of these attributes for all dataums in the data. This enables use in downstream marketing tasks that require attribute values, for example, econometric modeling to estimate the impact of discovered attributes on WTP (see Appendix E). GANs do not conventionally infer a representation of the data, and hence do not have this benefit. Finally, VAEs often require less data to train in comparison with GANs (Karras et al. 2019). Thus, even though GANs can provide much better reconstruction and work

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

[8] Moreover, GANs tend to suffer from training instability. Some of the common failure modes are vanishing gradients, mode collapse and failure to converge.

better for small and detailed objects (Locatello et al. 2020), we choose a VAE-based approach because of its suitability to our research question. Table EC.2 summarizes the recent disentanglement methods and Table EC.3 summarizes metrics to measure disentanglement.

Table EC.1 Comparison between VAE and GAN based methods

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

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

Table EC.2 Recent Disentanglement Literature

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

Table EC.3 Various Disentanglement Metrics

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

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

A.4. Disentanglement with and without Supervision

A key challenge in *any* disentangled representation learning approach is whether a unique set of discoverable attributes exists and whether it can be learned. This is especially relevant to real-world unstructured data that exists in the field of marketing where ground truth cannot be known. Contextualizing results from independent component analysis (Hyvärinen and Pajunen 1999) to disentanglement learning, Locatello et al.

(2019) showed that there is no theoretical guarantee for learning independent attributes using an unsupervised disentanglement approach. To address this concern, Locatello et al. (2020) showed that a small number of labelled examples with even potentially imprecise and incomplete labels is sufficient to perform model selection to learn disentangled representations.

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

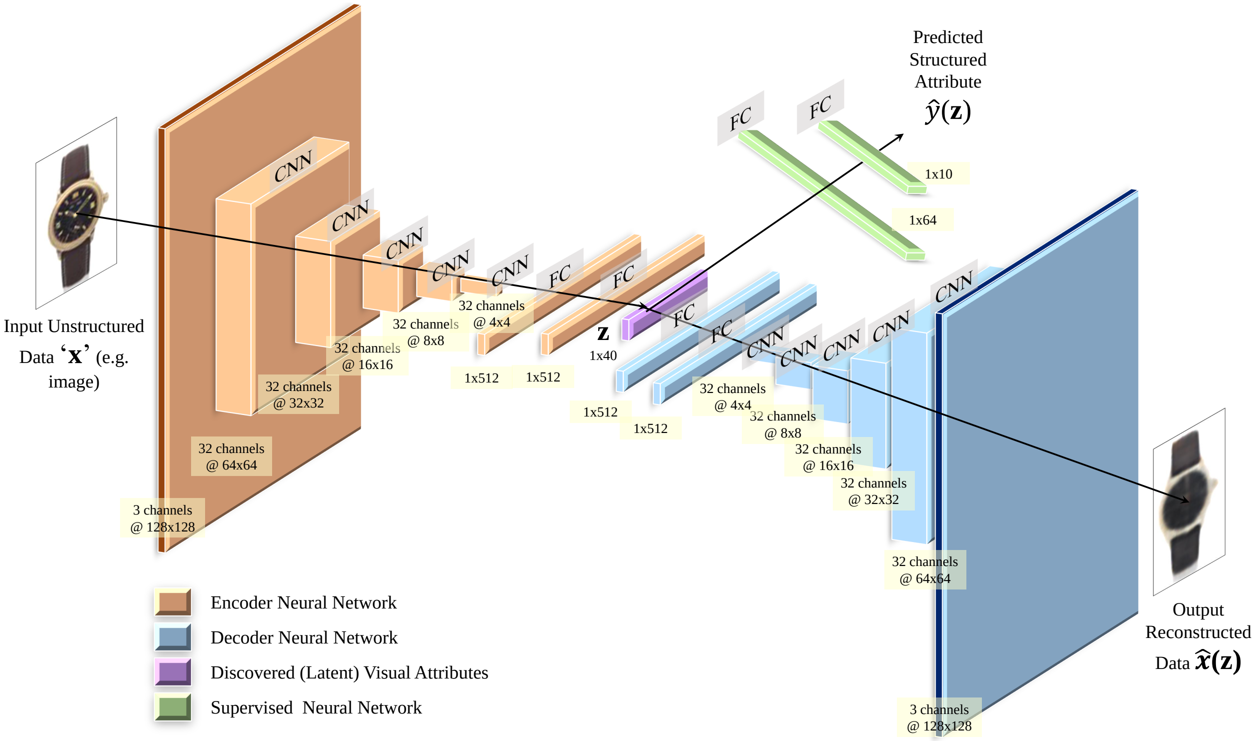
Appendix B: Model Architecture

Before we build the model architecture, we decide the number of latent codes J or the maximum number of attributes that our model aims to find. On the one hand, a small J might combine multiple visual attributes into one which results in entanglement. On the other hand, when J is large, the model discovers redundant or irrelevant attributes or it might even break up a true attribute across multiple dimensions. We choose $J = 20$ to balance these considerations, based on our empirical setting. Next, we describe the model architecture.

Figure EC.1 shows the detailed model architecture. We modify the architecture used in Burgess et al. (2017) in order to use images of 128×128 pixels as well as to incorporate a supervised neural net. Since we provide an application of our proposed method in the visual domain, we use Convolutional Neural Net (CNNs) to construct the encoder neural net. In the encoder neural net, we stack a sequence of CNN layers in order to learn high-level concepts for images. Finally, we introduce 2 fully-connected (FC) layers to first flatten the output of the sequence of CNN layers and then reduce the number of dimensions in order to learn a maximum of J visual attributes. 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 attributes. Finally, we connect fully connected layers to the discovered visual attributes to create the supervised neural net in order to predict the structured attribute.

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

Figure EC.1 Model Architecture



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

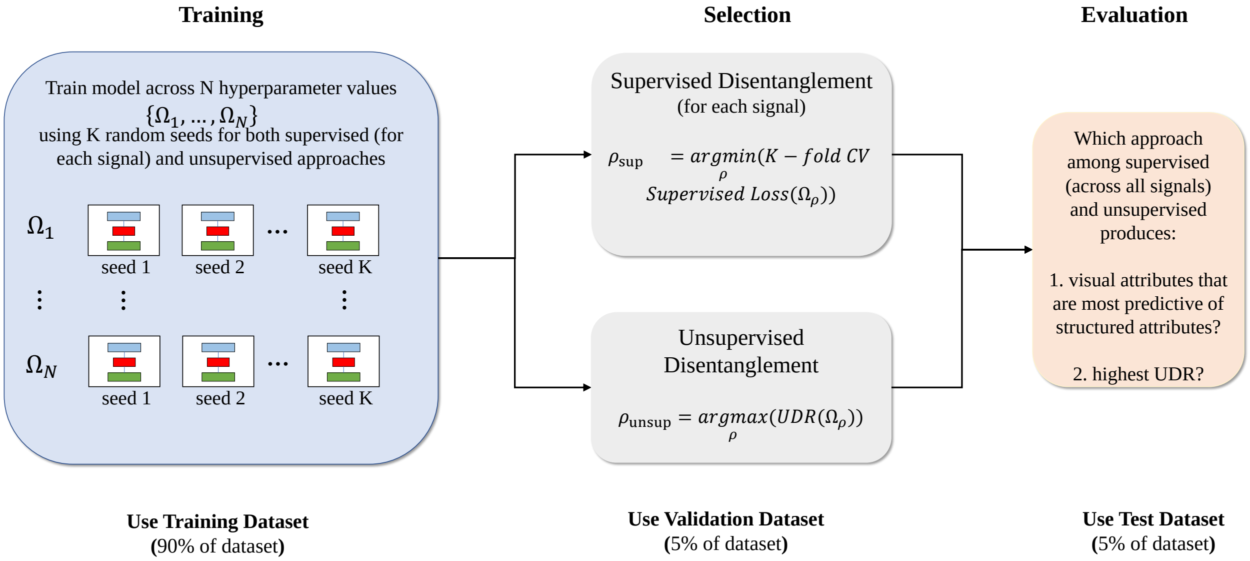
Appendix C: Modeling Details: Training, Selection, and Evaluation

We detail below the process for training the models given the initial weights and hyperparameters, followed by model selection using cross-validation or UDR metrics, and finally model evaluation.

C.1. Model Training and Model Selection

We divide the dataset into a training dataset for training (or learning) disentangled representations, a validation dataset for model selection using the Unsupervised Disentanglement Ranking (UDR) and a test dataset to compare various supervised approaches and the unsupervised approach in the ratio 90:5:5. Figure EC.2 provides a schematic diagram for the model training and selection for the supervised and the unsupervised approaches. The training process takes in the unstructured data (watch images) as input, and chooses a structured watch attribute (e.g., brand) as the supervisory signal to the model (only for the supervised approach). We fix the hyperparameters corresponding to α, γ, batch size, latent space dimensions J, type of optimizer, optimizer parameters such as learning rate, type of decoder neural net and number of training steps (Locatello et al. 2020).

Figure EC.2 Model Training, Selection & Evaluation


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

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

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

Unsupervised Approach: We sweep over hyperparameters corresponding to β (weight on the total correlation loss term). In the unsupervised approach $\delta = 0$ by definition. We use Unsupervised Disentanglement Ranking (UDR), a metric proposed by Duan et al. (2020), for model selection purposes. We choose this metric because it allows for an automated way to select a model and does not require access to the ground truth data generative process unlike other metrics such as β-VAE metric (Higgins et al. 2017), the FactorVAE metric (Kim and Mnih 2018), Mutual Information Gap (MIG) (Chen et al. 2018) and DCI Disentanglement scores (Eastwood and Williams 2018). We select the hyperparameter setting corresponding to the highest UDR. Appendix C.2 has details on how UDR is calculated. Table EC.4 lists the hyperparameters obtained for the unsupervised approach. Similar to the supervised approach, we use the chosen trained model to extract the discovered unstructured (visual) attributes on the test dataset as well.

C.2. Model Evaluation

One of the contributions of our paper is identify the class of marketing signals which help in disentangling factors of variation. We also compare the use of such supervisory signals with an unsupervised approach.

Table EC.4 Hyperparameters Obtained by Model Selection Criteria

Disentanglement Approach	Signal	β	δ
Supervised	Auction Location	1	60
Supervised	Brand	18	50
Supervised	Circa	4	35
Supervised	Material	6	25
Supervised	Movement	4	20
Supervised	Price	1	16
Unsupervised	–	18	0

We evaluate the model along two dimensions: (a) performance in predicting the set of structured attributes, and (b) unsupervised disentanglement ranking (UDR).

Performance on Predicting Structured Attributes: A good disentangled representation should be developed to help in a variety of downstream tasks (Bengio et al. 2013). Based on this logic, we compare various supervisory signals by their ability to discover attributes that can be used to classify the watches according to different structured attributes. In the downstream classification models, we train the discovered disentangled attributes to predict a particular structured attribute on the training dataset. Next, we predict the structured product attribute from the test set using the trained classification model. For each set of discovered disentangled attributes corresponding to a particular supervisory signal, we calculate the average accuracy across different classification tasks. Finally, we select the supervisory signal that provides the highest average accuracy.

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

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

For each trained model, we perform $\kappa = 45$ pairwise comparisons with all other models trained with the same β value but with different seed values and calculated the UDR_{ij}, where i and j index the two models. Each UDR_{ij} score is calculated by computing the similarity matrix R_{ij}, where each entry is the Spearman correlation between the responses of individual latent units of the two models. The absolute value of the similarity matrix is then taken $|R_{ij}|$ and the final score UDR_{ij} for each pair of models is calculated according to the Equation (EC.1). However, since this approach is an unsupervised method, it does not have theoretical guarantees to disentangle as shown by Locatello et al. (2019).

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

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

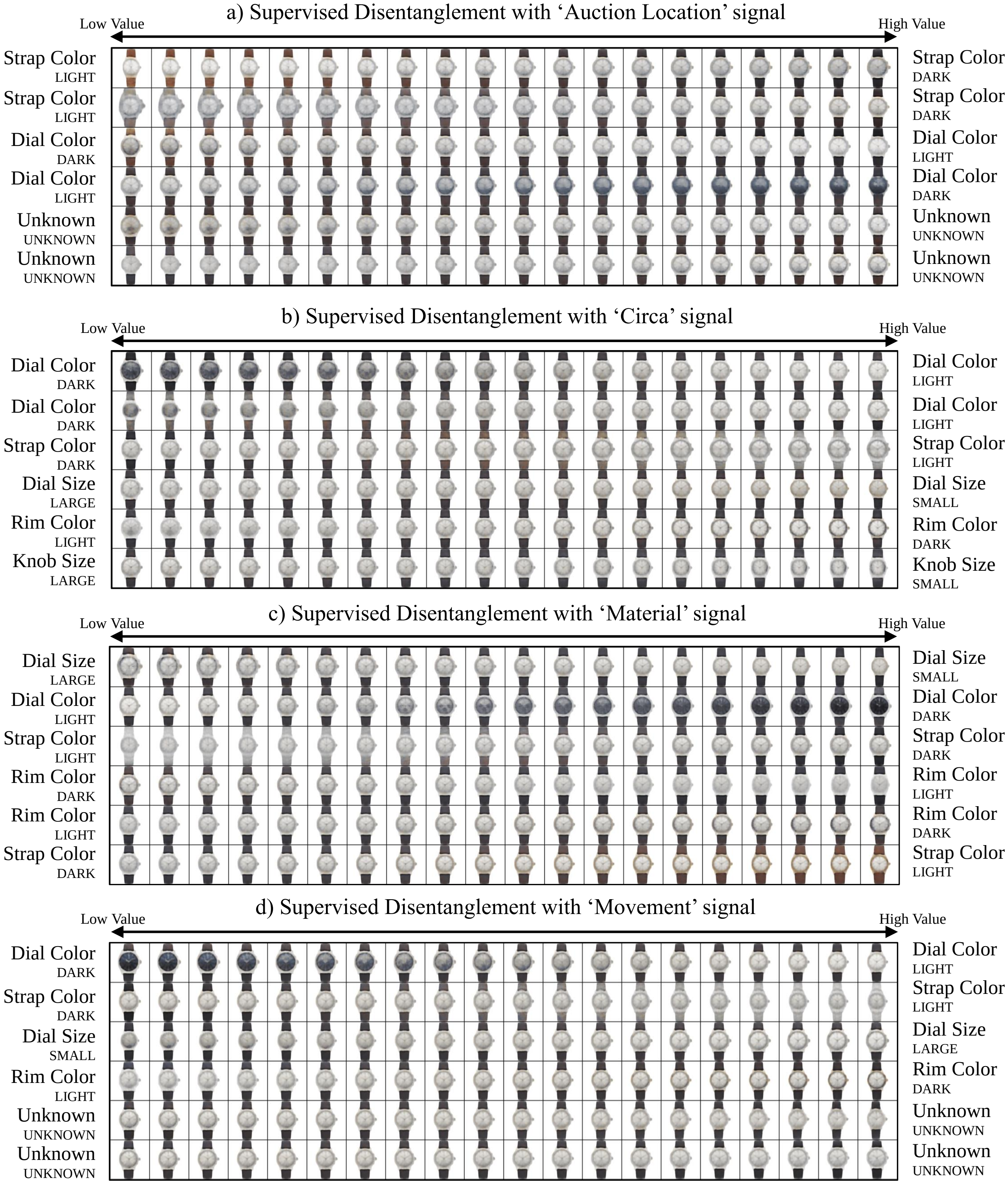
Appendix D: Results: Discovered Visual Attributes

Figure EC.3 shows the discovered visual attributes. It contains the attributes learnt by the below supervisory signals.

1. Location of the Auction
2. Circa or Decade of Manufacture
3. Material used in the watch
4. Type of Movement

Visual attributes learnt by the unsupervised approach as well as supervising on brand and price are in Figure 4.

Figure EC.3 Discovered Visual Attributes from other Supervised Approaches



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

Appendix E: Example "Downstream Task": Impact of Visual Attributes on WTP

We provide an example of using disentanglement in a marketing-related downstream task. In particular, we show how discovered disentangled visual attributes can be applied in standard marketing model of consumer willingness-to-pay (WTP). In our setting, watches are auctioned according to the English auction mechanism in which bidders iteratively submit successively higher bids and the final bidder wins the item in return for a payment equaling her final bid. It is equivalent to a second-price sealed-bid auction in which the winner will be the bidder who values the object most highly, and the price paid will be the value of the object to the bidder who values it second-most highly. The dominant strategy of each bidder is to bid their true value of the watch (Milgrom and Weber 1982). As a consequence, the data allows us to study how the discovered visual attributes affect consumers' WTP. We use the specifications described in Equation (EC.2) with parameter vector $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4, \mathbf{w}_5)$. We deliberately use a simple linear specification to predict the hammer price wtp_i of the watch for each auction i for the sake of interpretability of the results.

$$\log(wtp_i) = \mathbf{w}_1^T \mathbf{s}_i + \mathbf{w}_2^T \mathbf{z}_i + \mathbf{w}_3^T \mathbf{a}_i + \mathbf{w}_4^T \tau_i + \mathbf{w}_5^T \mathbf{z}_i * \tau_i + \epsilon_i \tag{EC.2}$$

where wtp_i is the hammer price of the auctioned watch expressed in constant 2000 dollars; $\mathbf{s}_i$ is the vector of structured product attributes that includes brand of the watch, year of manufacture or *circa* of the watch, type of movement associated with the watch, dimensions of the watch and the materials used in the watch; $\mathbf{z}_i$ is the vector of informative visual attributes that includes size of the dial, dial color, strap color, dial shape, size of the knob and rim color; $\mathbf{a}_i$ is the vector of auction attributes that includes auction location; and τ_i is the timetrend; and ϵ_i is the error term. Table 1 and Table EC.6 has the summary statistics for the structured attributes and visual attributes respectively.

We specify a series of nested linear models based on the linear specification. The first two linear specifications do not include discovered visual attributes. They serve as the benchmark models. In the first model (1), we only include structured product attributes. In the second model (2), we include structured product attributes, auction location fixed effects and timetrend. In the third model (3), we only include visual attributes. In the fourth model (4), we include structured product attributes and visual attributes. In the fifth model (5), we include structured product attributes, visual attributes, auction location fixed effects and timetrend. Finally, in the sixth model (6), we include structured product attributes, visual attributes, auction location fixed effects, timetrend as well as an interaction between visual attributes and timetrend.

Table EC.5 has the results from the models specified by Equation EC.2. From Model (1), we can see that structured product attributes are only able to explain 32% of the variation in the hammer prices (as characterized by R^2). From Model (2), we see that the structured product attributes, auction location fixed effects and timetrend explain $R^2 = 41\%$ of the variation in the hammer prices. From the results of model (3), we see that visual attributes are able to explain $R^2 = 4\%$ of the variation in the hammer prices. This is noteworthy because in most marketing models, researchers use only structured attributes and in models in marketing and economics (Guadagni and Little 1983, Berry et al. 1995), researchers aggregate across all the visual (and other) attributes to form one attribute that is the unobserved product quality. This leads to two issues. First is the obvious problem of aggregation, which implies the researcher cannot semantically interpret or identify the source of this term. Second, since such an attribute is treated as a structural error, we cannot model consumer heterogeneity across the attribute. In contrast, in our model these unobserved visual attributes are recovered as part of our algorithm, and have semantic interpretation.

Next, we include both structured product attributes and discovered visual attributes in model (4). We observe that structured product attributes and visual attributes together are able to explain $R^2 = 32\%$ of the variation in hammer prices. In model (5), we include auction location fixed effects and timetrend as well. Together, they explain $R^2 = 42\%$ of the variation in the hammer prices. In model (6), we also include the interaction between timetrend and visual attributes and together they explain $R^2 = 43\%$ of the variation in the hammer prices.

Our results show that watches auctioned at Hong Kong fetch the highest hammer price followed by New York City, Dubai, and then those auctioned online. Watches with the brand Patek Philippe have the highest hammer prices, followed by Rolex, Audemar's Piguet, and then Cartier compared with the baseline of other brands. We find that watches manufactured in 2010s have the highest hammer prices, followed by 2000s, 1990s and then others compared with the baseline of pre-1950s. Mechanical watches fetch higher hammer prices followed by quartz watches and then finally automatic watches. Watches with Steel have the least hammer prices, followed by a combination of Gold and Steel, and then Gold compared with the baseline of other materials. Much of these results are consistent with prior knowledge about the market for high-end watches. We find that a larger watch dimension leads to a higher hammer price. We also find that auctioned watches have an increasing valuation over time.

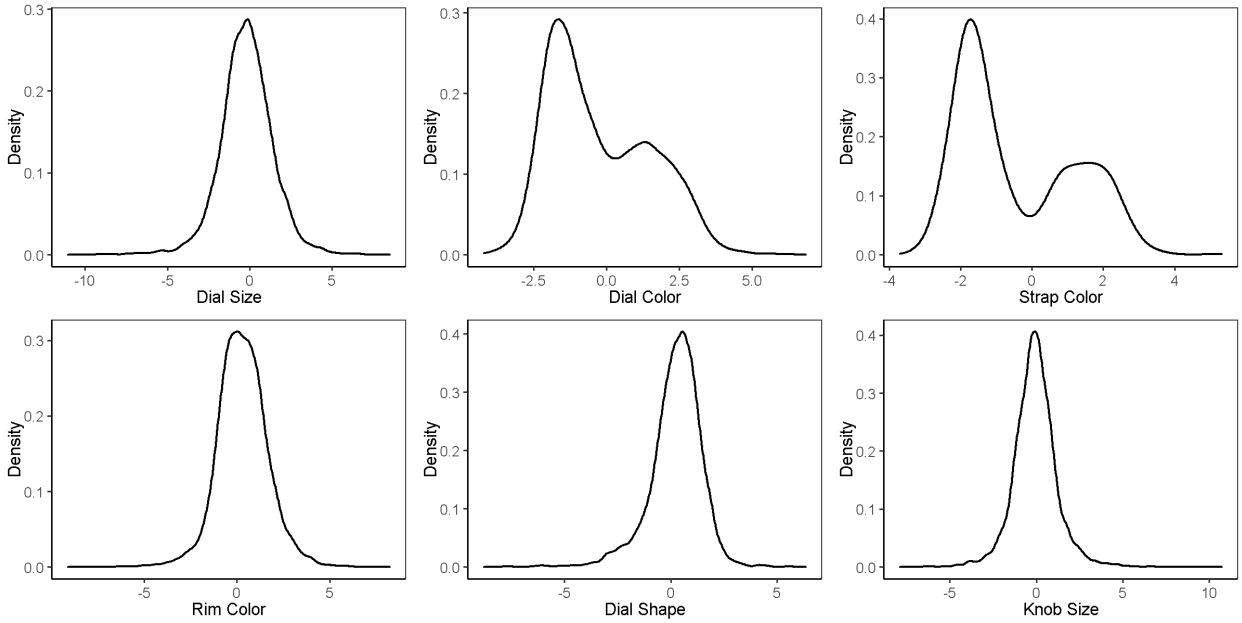
We find that 'knob size' and 'rim color' are not statistically significant for any of the six models. Other visual attributes are statistically significant for at least one model specification. We use the model specification (5) to interpret the results for visual attributes. We find that:

- watches with larger dial size fetch higher price in comparison with smaller sized watches.
- watches with a darker dial color fetch higher price as opposed to watches with lighter dial color.
- watches with a more circular dial have a higher hammer price as compared to a rectangular dial.

Table EC.5: Results

	Dependent variable:					
	log10(Willingness to Pay (Highest Bid in Auction))					
	Without Visual Attributes			With Visual Attributes		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.88*** (0.05)	2.37*** (0.05)	3.94*** (0.01)	2.95*** (0.06)	2.57*** (0.06)	2.59*** (0.06)
Auction Location						
Dubai		0.39*** (0.02)			0.40*** (0.02)	0.40*** (0.02)
Hong Kong		0.55*** (0.02)			0.56*** (0.02)	0.55*** (0.02)
New York City		0.50*** (0.02)			0.51*** (0.02)	0.51*** (0.02)
Brand						
Audemar's Piguet	0.26*** (0.02)	0.23*** (0.02)		0.27*** (0.02)	0.23*** (0.02)	0.23*** (0.02)
Cartier	0.08*** (0.02)	0.08*** (0.02)		0.11*** (0.02)	0.12*** (0.02)	0.12*** (0.02)
Patek Philippe	0.48*** (0.02)	0.42*** (0.01)		0.49*** (0.02)	0.44*** (0.01)	0.44*** (0.01)
Rolex	0.43*** (0.02)	0.38*** (0.02)		0.41*** (0.02)	0.36*** (0.02)	0.35*** (0.02)
Circa						
1950s	0.11*** (0.04)	0.10*** (0.03)		0.10*** (0.04)	0.08** (0.03)	0.08** (0.03)
1960s	0.11*** (0.03)	0.08*** (0.03)		0.08** (0.03)	0.06* (0.03)	0.06* (0.03)
1970s	0.05 (0.03)	0.03 (0.03)		0.03 (0.03)	0.003 (0.03)	0.004 (0.03)
1980s	0.08** (0.03)	0.05 (0.03)		0.06* (0.03)	0.03 (0.03)	0.03 (0.03)
1990s	0.16*** (0.03)	0.12*** (0.03)		0.14*** (0.03)	0.10*** (0.03)	0.10*** (0.03)
2000s	0.22*** (0.03)	0.16*** (0.03)		0.20*** (0.03)	0.14*** (0.03)	0.14*** (0.03)
2010s	0.35*** (0.03)	0.26*** (0.03)		0.33*** (0.03)	0.23*** (0.03)	0.24*** (0.03)
Movement						
Automatic	−0.07*** (0.02)	−0.03 (0.02)		−0.07*** (0.02)	−0.04** (0.02)	−0.04** (0.02)
Mechanical	0.18*** (0.02)	0.17*** (0.02)		0.18*** (0.02)	0.17*** (0.02)	0.17*** (0.02)
Material						
Gold	−0.16*** (0.02)	−0.13*** (0.02)		−0.15*** (0.02)	−0.12*** (0.02)	−0.12*** (0.02)
Gold and Steel	−0.30*** (0.03)	−0.25*** (0.03)		−0.31*** (0.03)	−0.25*** (0.03)	−0.26*** (0.03)
Steel	−0.40*** (0.02)	−0.34*** (0.02)		−0.42*** (0.02)	−0.35*** (0.02)	−0.35*** (0.02)
Watch Dimensions (in mm)	0.02*** (0.001)	0.02*** (0.001)		0.02*** (0.001)	0.02*** (0.001)	0.02*** (0.001)
Timetrend (in days)		0.0001*** (0.0000)			0.0001*** (0.0000)	0.0001*** (0.0000)
Visual Attributes						
Dial Size			0.04*** (0.004)	0.01** (0.004)	0.03*** (0.004)	0.05*** (0.001)
Dial Color			−0.01 (0.004)	0.005 (0.004)	0.01*** (0.004)	−0.01*** (0.01)
Strap Color			−0.03*** (0.004)	0.01*** (0.004)	0.01* (0.004)	−0.03*** (0.01)
Rim Color			0.002 (0.005)	−0.01 (0.004)	−0.01* (0.004)	0.001 (0.01)
Dial Shape			0.04*** (0.01)	0.04*** (0.005)	0.04*** (0.005)	0.03 (0.01)
Knob Size			0.004 (0.01)	−0.01 (0.005)	−0.005 (0.004)	0.01* (0.01)
TimeTrend * Visual Attributes						
TimeTrend * Dial Size						−0.0000*** (0.0000)
TimeTrend * Dial Color						0.0000*** (0.0000)
TimeTrend * Strap Color						0.0000*** (0.0000)
TimeTrend * Rim Color						−0.0000 (0.0000)
TimeTrend * Dial Shape						0.0000 (0.0000)
TimeTrend * Knob Size						−0.0000* (0.0000)
Observations	6,187	6,187	6,187	6,187	6,187	6,187
R^2	0.32	0.41	0.04	0.32	0.42	0.43
Adjusted R^2	0.31	0.41	0.03	0.32	0.42	0.43
Residual Std. Error	0.43	0.40	0.51	0.43	0.39	0.39
F Statistic	168.13***	206.79***	38.34***	128.91***	167.92***	140.72***
Note:						*p<0.1; **p<0.05; ***p<0.01

Figure EC.4 Histogram of Discovered Visual Attributes (from 'Brand' Signal)



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

Appendix F: Quantitative Analysis of Individual-Level Discovered Attributes

Table EC.6 has the summary statistics of the visual attribute levels learned by using the supervisory signal 'brand'. Figure EC.4 shows the histogram of these discovered visual attributes. We see that the distribution of 'dial color' and 'strap color' do not seem to follow a standard normal distribution. This is because the method does not enforce any conventional parameterization on the distribution of the visual attributes of our data. The histogram also shows that the algorithm is able to find continuous as well as discrete visual attributes. While 'dial size', 'dial color', 'rim color' and 'dial shape' can be interpreted as continuous visual attributes with a distribution close to gaussian, 'dial color' and 'strap color' seem to be discrete visual attributes. A watch's 'dial color' or 'strap color' could come from one of two gaussian distributions.

Table EC.6 Summary Statistics of Discovered Visual Attributes (from 'Brand' Signal)

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