Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis (Web Appendix)

Ankit Sisodia, Alex Burnap and Vineet Kumar*

2024

Table of Contents

*Ankit Sisodia is an Assistant Professor of Marketing at the Daniels School Of Business at Purdue University. email: asisodia@purdue.edu. Alex Burnap is an Assistant Professor of Marketing at the Yale School of Management. email: alex.burnap@yale.edu. Vineet Kumar is an Associate Professor of Marketing at the Yale School of Management. email: vineet.kumar@yale.edu.

A Discovery of Visual Characteristics across Models

We compare the visual characteristics discovered by our disentanglement approach with benchmark models like Autoencoders, Variational Autoencoders and Unsupervised Disentanglement.

Comparison among Disentanglement Models: Figure A.1 show the discovered visual characteristics learned by the supervised approaches corresponding to three cases: supervised with high UDR, supervised with low UDR and the unsupervised approach. We compare the human interpretability of the visual characteristics obtained from the supervised disentanglement approach with the ones obtained from the unsupervised approach using consumer surveys. In these surveys, we ask consumers whether they are able to interpret the discovered visual characteristics. From Table A.1, we can see that on average consumers are better able to interpret the visual characteristics from the supervised approach as compared with the unsupervised approach. Thus, the results of these survey validate that supervision helps us obtain more disentangled visual characteristics in addition to just using the UDR metric.[1]

Table A.1: Human Interpretation of Visual Characteristics

Visual Characteristic	Mean [95% CI]		% Improvement
	Supervised	Unsupervised	
Dial Color	0.80 [0.70, 0.89]	0.81 [0.72, 0.90]	–[†]
Dial Size	0.76 [0.66, 0.86]	0.78 [0.69, 0.88]	–[†]
Strap Color	0.88 [0.80, 0.96]	0.90 [0.83, 0.97]	–[†]
Rim Color	0.79 [0.69, 0.88]	0.42 [0.30, 0.54]	88.1%
Dial Shape	0.87 [0.79, 0.95]	0.49 [0.37, 0.61]	90.5%
Knob Size	0.70 [0.59, 0.80]	0.56 [0.44, 0.68]	25.0%
Across All 6 Char	0.80 [0.76, 0.84]	0.67 [0.62, 0.71]	21.2%

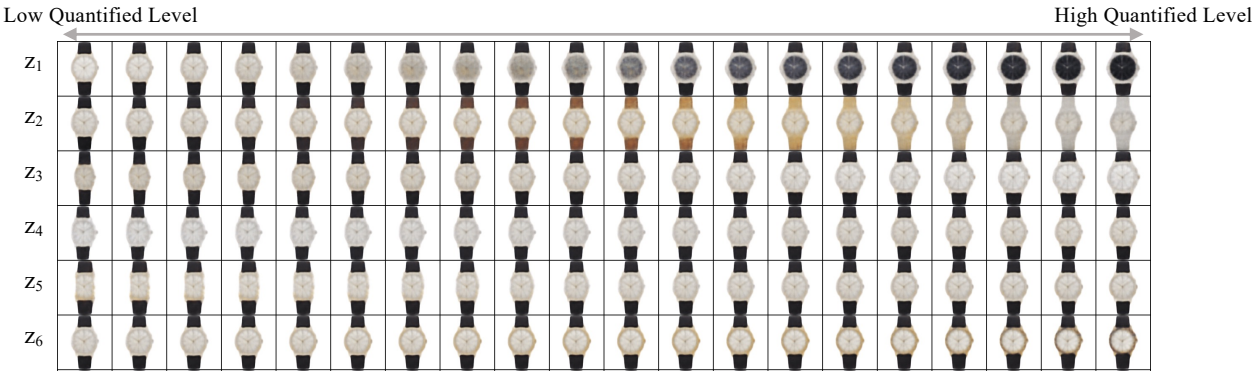
[†] The mean interpretability of the visual characteristic in the supervised approach overlaps with the 95% CI of the interpretability of the unsupervised.

We note that although we find supervision helps disentanglement, and that supervision is required for overcoming the "impossibility theorem" discussed in Locatello et al. (2019), unsupervised disentanglement has a known ability to discover some visual characteristics. This observation helped spur the drive towards more control over the VAE objective by decomposing it into

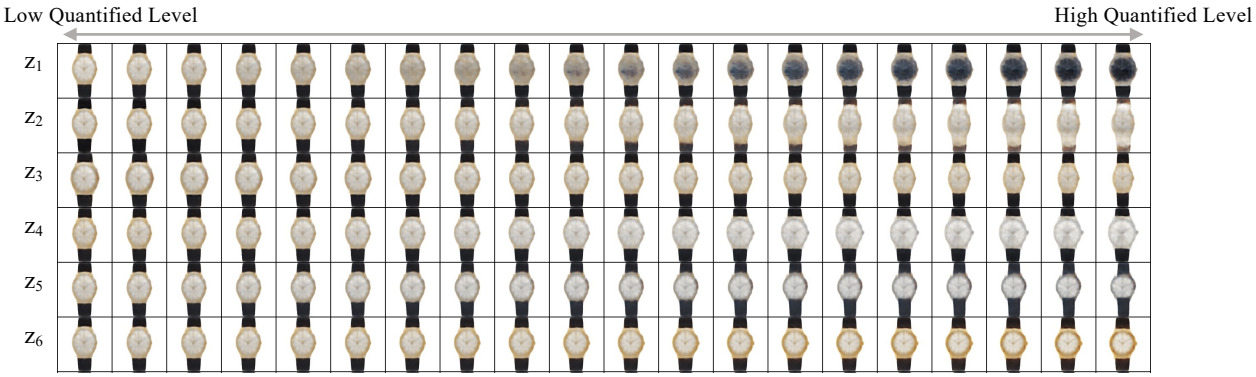
[1] To assess the variability and reliability of our sample estimates, we employed a bootstrap resampling method.

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

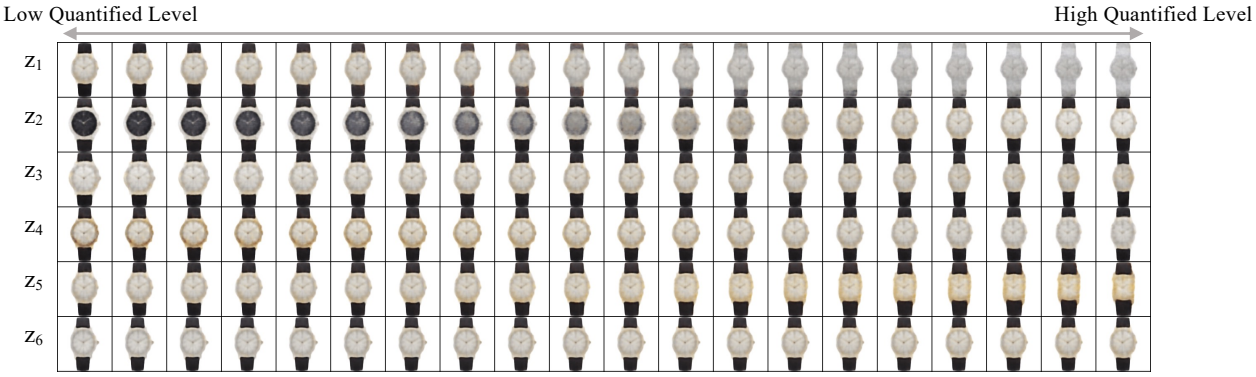
(a) High UDR: 'Brand', 'Circa' & 'Movement' Supervisory Signal



(b) Low UDR: 'Circa' Supervisory Signal



(c) Unsupervised Approach



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

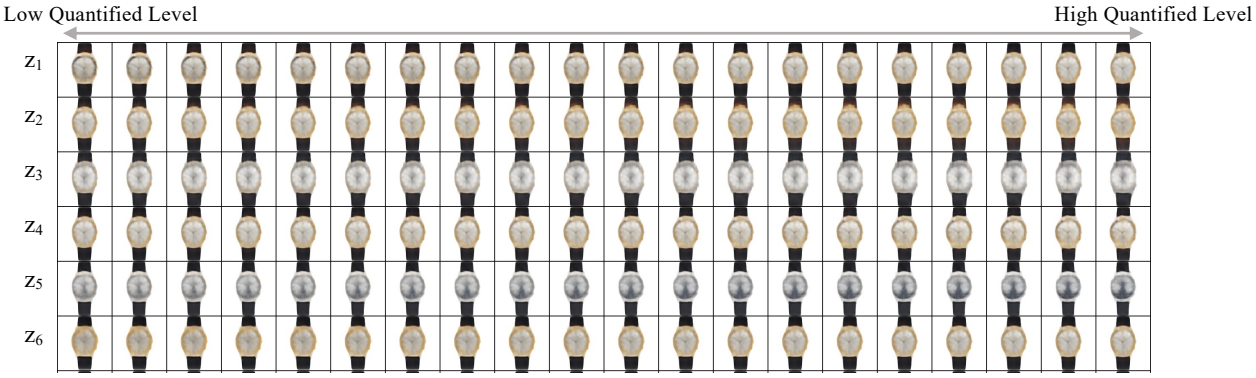
terms that explicitly control disentanglement (Chen et al. 2018; Hoffman and Johnson 2016). In short, several factors may be at play, including (1) the common prior assumption of isotropic Gaussian has no off-diagonal covariance terms, promoting uncorrelatedness of the embedding; and (2) VAEs pursue PCA direction (locally) (Rolinek, Zietlow, and Martius 2019). Further intuition for why unsupervised disentanglement can work in practice at all is well-discussed in Mathieu et al. (2019).

Comparison with Benchmark Models: We obtain the visual characteristics discovered by an an autoencoder (AE) and a variational autoencoder (VAE) to serve as reference to the disentanglement model. Figure A.2 gives the output of discovered visual characteristics from an autoencoder and a variational autoencoder. We show the top six visual characteristics based on the KL divergence value of the difference between the posterior and the Gaussian prior.

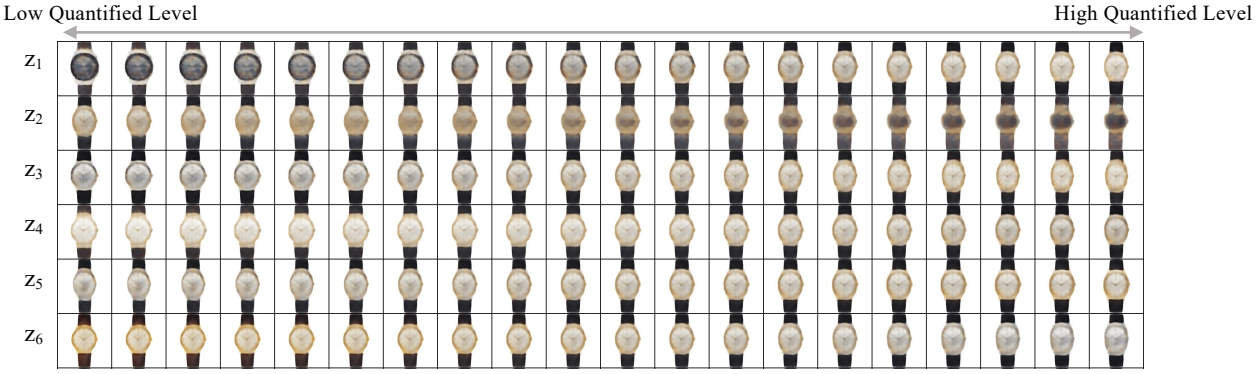
We cannot interpret any of the visual characteristics discovered by the AE. Note that these characteristics are not uninformative because their KL divergence is not close to 0. We find that the VAE leads to entanglement. By entangled, we mean that when any one entangled characteristic is changed while others are fixed, the watch image changes in more than one interpretable visual characteristic. This is unlike a disentangled model in which there is a one-to-one mapping between visual characteristics and latent factors of variation.

Figure A.2: Discovered Visual characteristics from Different Methods

(a) Autoencoder



(b) Variational Autoencoder



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

B Connections with Existing Marketing Methods

We include a high-level comparison of the methods in Table B.1.

Table B.1: Comparison of Methods

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

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

An autoencoder (AE) (Baldi and Hornik 1989; Rumelhart, Hinton, and Williams 1986) is a nonlinear method that focuses on reconstructing the original high-dimensional data

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

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

Comparison of Generative Methods: The two broad classes of generative models are based on variational autoencoders (VAEs) (Kingma and Welling 2014) and generative adversarial networks (GAN) [2] (Goodfellow et al. 2020). Most state-of-the-art disentangled *representation learning* methods are based on VAEs. VAEs are comprised of two models – the encoder neural net and the decoder neural net. The encoder neural net compresses high-dimensional input data to a lower-dimensional latent vector (latent characteristics), followed by inputting the latent vector to the decoder neural net

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

which outputs a reconstruction of the original input data. VAEs balance having both a low reconstruction error between the input and output data (e.g., images, text), as well as a KL-divergence of the latent space distribution (latent characteristics) from a researcher-defined prior distribution (e.g., Gaussian). The KL-divergence term acts as a regularizer on the latent space, such that it has desired structure (smoothness, compactness). VAEs are parameterized in both the encoder neural net and decoder neural net using neural networks whose parameters are learned jointly.

Table B.2: Comparison between VAE and GAN based methods

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

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

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[3], InfoGAN explicitly constrains a small subset of the 'noise' variables to have high mutual information with generated data. Several VAE-based methods have proven to be superior (Kim and Mnih 2018; Chen et al. 2018) than InfoGAN. Recent methods based on StyleGAN (Karras, Laine, and Aila 2019) such as Info-StyleGAN (Nie et al. 2020) are able to perform disentanglement at a much higher resolution (1024×1024) unlike the VAE-based methods. However, unlike InfoGAN, Info-StyleGAN suffers from the need for human labels or pretrained models, which can be expensive to obtain (Voynov and Babenko 2020).

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

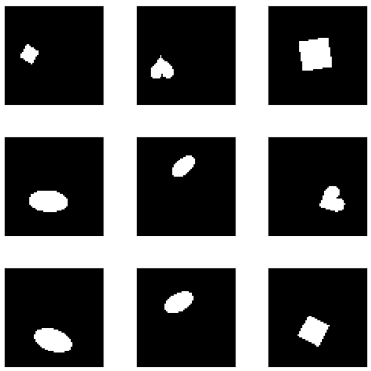
[3]Moreover, GANs tend to suffer from training instability. Common failure modes are vanishing gradients, mode collapse, and failure to converge.

because of its suitability to our research question.

C Disentanglement with a Simple Geometric Shape

Consider the dataset of 2D objects dSprites (Higgins et al. 2017). Each image in this data (see Figure C.1) shows an object of a specific shape, size and color at a specific location in the image. Across images, we can see different possible combinations of these visual characteristics. The objective of disentanglement is to separate out these independent factors of variation to obtain object shape, position, size, and color as the 4 latent dimensions discovered by the disentanglement model. The advantage of disentanglement is that, even when the dimensionality of the latent space is increased to a large number, it will only discover these true factors of variation (shape, size, color and position).

Figure C.1: Sample of dSprites Images



D Disentanglement in a Different Product Category – Sneakers

Our data includes sneakers sold at a fashion e-commerce firm. For each sneaker in the dataset, we have its image, brand, and price. Figure D.1 shows a sample of sneaker images in our dataset. We obtained the dataset of sneakers sold on a fashion e-commerce firm in March 2023. These shoes were classified as sneakers by the retailer. Overall, our dataset includes 2,227 unique sneaker models with an average of 2.5 images per sneaker model. The size of the overall dataset includes 5,575 images. We only included the side view of sneakers in order to focus on the variation in the shape of the sneakers. Finally, we specifically used grayscale images because each sneaker model with the same shape comes in multiple colors. We preprocessed each image to have the size of 128x128 dimensions to keep the images consistent with the watch category. A total of 247 unique brands are present in the data. Skechers, Vans, New Balance, adidas and ASICS are the five brands with the largest share of observations. Table D.1 provides summary statistics of the sneakers.

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

Figure D.1: Sample of Sneakers

From Table D.2, we show that price serves as the most effective supervisory signal for learning human-interpretable visual characteristics for sneakers. To understand why price is the most effective supervisory signal, we calculate the Signal Effectiveness score that relies on the intuition that better or more informative signals will generate more separation in latent visual characteristics. Consistent with this intuition, in sneakers, the Signal Effectiveness Score for brands is 0.26

Table D.1: Summary Statistics of Structured characteristics of Sneakers

Statistic	Mean	SD	Min	Max
Brand (Skechers)	0.09	0.29	0	1
Brand (Vans)	0.08	0.28	0	1
Brand (New Balance)	0.07	0.26	0	1
Brand (adidas)	0.06	0.24	0	1
Brand (ASICS)	0.05	0.22	0	1
...				
Brand (Others)	0.14	0.34	0	1
Price (in $s)	112.30	46.45	30.00	650.00

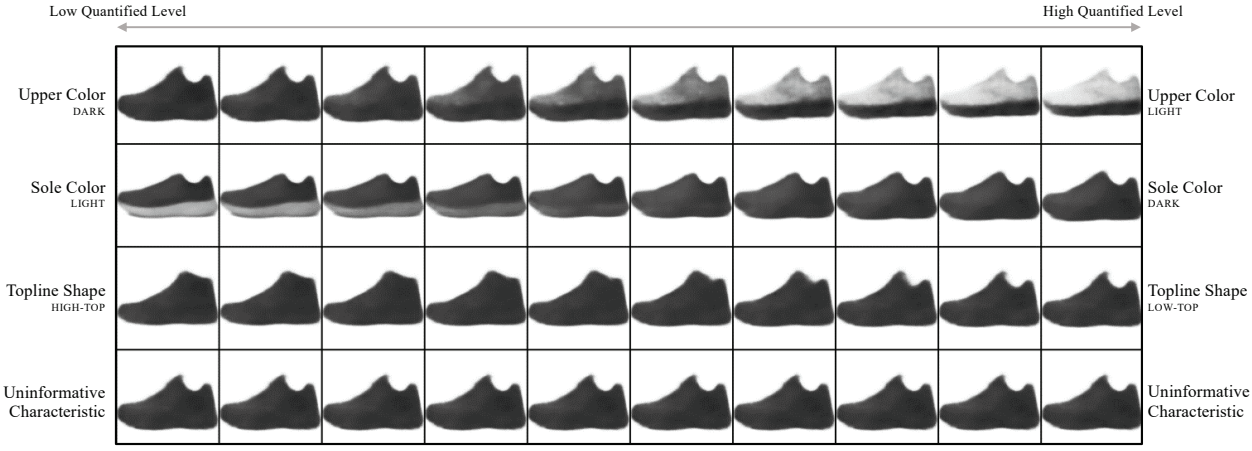
compared with 0.32 for discrete prices.

Table D.2: Comparison of Different Supervisory Approaches

Number of Signals	Supervisory Signals	UDR
0	Unsupervised	0.126
1	Brand	0.093
1	Price (5 Discrete Classes)	0.267

Figure D.2 gives an output of discovered visual characteristics corresponding to the supervisory signals 'price'. In each row of the figure, we show how the sneaker image changes based on changes in levels of one visual characteristic, while keeping all the other characteristics fixed. We only show three visual characteristics as rest of the characteristics are found to be uninformative i.e. the KL divergence of the posterior was not much different from the Gaussian prior. Traversing along an uninformative characteristic leads to no visual change, and we show one uninformative characteristic for reference.

Figure D.2: Discovered Visual Characteristics of Sneakers

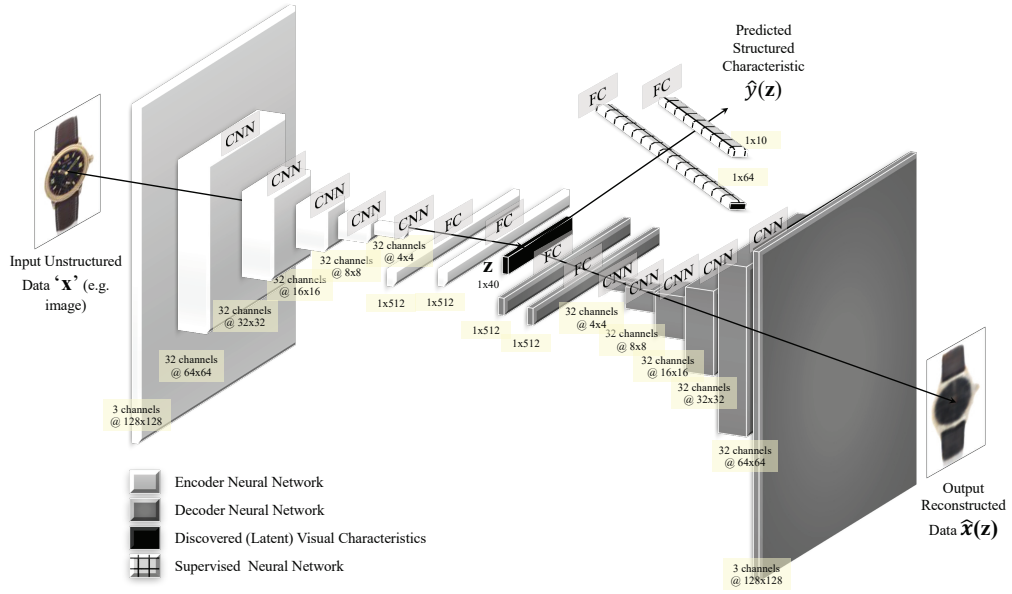


Notes: Latent traversals along a *focal sneaker* used to visualise the semantic meaning encoded by single visual characteristic learnt by a trained model. In each row, the quantitative level of a single characteristic is varied keeping the other characteristics fixed. The resulting reconstruction is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

E Model Architecture

The model architecture is detailed in Figure E.1. 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 $b_1 = 0.9$ and $b_2 = 0.999$. We set batch size equal to 64. We train the model for 100 epochs. Portions of our codebase were built on elements sourced from the disentangling-vae open source project (Dubois et al. 2019).

Figure E.1: Model Architecture



F Summary Statistics of Structured Characteristics of Auctioned Watches

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

Table F.1: Summary Statistics of Structured Characteristics of Auctioned Watches

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

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

G Summary Statistics of Visual Characteristics of Auctioned Watches

Table G.1 details the summary statistics of the visual characteristic levels learned.

Table G.1: Summary Statistics of Discovered Visual Characteristics

Visual characteristic	Mean	SD	Min	Max
Dial Size	−0.32	1.42	−9.86	9.92
Dial Color	−0.50	1.52	−3.49	7.20
Strap Color	−0.24	1.67	−3.43	4.82
Rim (Bezel) Color	−0.26	0.90	−6.26	6.14
Dial Shape	0.24	0.95	−7.48	3.09
Knob (Crown) Size	−0.17	0.95	−8.14	10.20

H Watches: UDR and Hyperparameters for Different Supervisory Signals

Table H.1 lists the UDR corresponding to each combination of supervisory signals. Table H.2 lists the hyperparameters obtained for each combination of supervisory signals.

Table H.1: Comparison of Different Supervisory Approaches (at Optimal Hyperparameter Weights for Each Signal)

Number of Signals	Supervisory Signals	UDR
0	Unsupervised	0.131
1	Brand	0.316
1	Circa	0.111
1	Material	0.130
1	Movement	0.122
1	Price (2 Discrete Classes)	0.122
2	Brand & Circa	0.382
2	Brand & Material	0.349
2	Brand & Movement	0.123
2	Brand & Price	0.120
2	Circa & Material	0.209
2	Circa & Movement	0.338
2	Circa & Price	0.103
2	Material & Movement	0.119
2	Material & Price	0.108
2	Movement & Price	0.140
3	Brand, Circa & Material	0.260
3	**Brand, Circa & Movement**	**0.414**
3	Brand, Circa & Price	0.342
3	Brand, Material & Movement	0.206
3	Brand, Material & Price	0.299
3	Brand, Movement & Price	0.273
3	Circa, Material & Movement	0.364
3	Circa, Material & Price	0.230
3	Circa, Movement & Price	0.224
3	Material, Movement & Price	0.080
4	Brand, Circa, Material & Movement	0.242
4	Brand, Circa, Material & Price	0.293
4	Brand, Circa, Movement & Price	0.279
4	Brand, Material, Movement & Price	0.262
4	Circa, Material, Movement & Price	0.322
5	Brand, Circa, Material, Movement & Price	0.321

Table H.2: Optimal Hyperparameters Obtained by Model Selection Criteria

Approach	Signal	# Signals	β	δ
Unsupervised	–	0	18	0
Supervised	Brand	1	18	50
Supervised	Circa	1	4	35
Supervised	Material	1	6	25
Supervised	Movement	1	4	20
Supervised	Price	1	1	16
Supervised	Price (with 2 classes)	1	22	45
Supervised	Brand and Circa	2	48	5
Supervised	Brand and Material	2	50	1
Supervised	Brand and Movement	2	6	20
Supervised	Brand and Price	2	6	25
Supervised	Circa and Material	2	36	1
Supervised	Circa and Movement	2	50	5
Supervised	Circa and Price	2	4	18
Supervised	Material and Movement	2	6	10
Supervised	Material and Price	2	4	20
Supervised	Movement and Price	2	12	20
Supervised	Brand, Circa and Material	3	48	1
Supervised	Brand, Circa and Movement	3	50	1
Supervised	Brand, Circa and Price	3	50	1
Supervised	Brand, Material and Movement	3	40	1
Supervised	Brand, Material and Price	3	50	1
Supervised	Brand, Movement and Price	3	48	1
Supervised	Circa, Material and Movement	3	48	1
Supervised	Circa, Material and Price	3	46	1
Supervised	Circa, Movement and Price	3	42	1
Supervised	Material, Movement and Price	3	1	1
Supervised	Brand, Circa, Material and Movement	4	44	1
Supervised	Brand, Circa, Material and Price	4	50	1
Supervised	Brand, Circa, Movement and Price	4	50	1
Supervised	Brand, Material, Movement and Price	4	44	1
Supervised	Circa, Material, Movement and Price	4	44	1
Supervised	Brand, Circa, Material, Movement and Price	5	44	1

I Using Shapley Values (SHAP) for Disentanglement

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

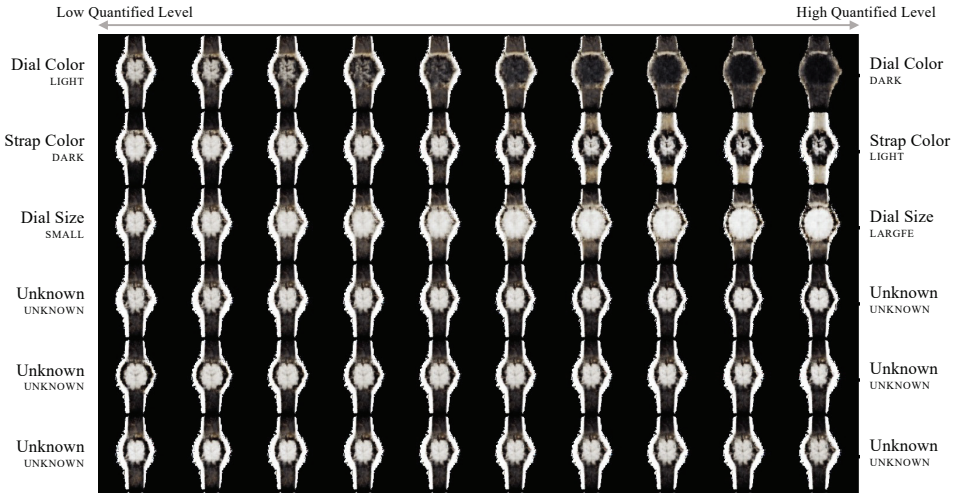
In this approach, we first train a deep learning model to predict the supervisory signal (e,g. brand) from images. Next, we calculate SHAP values to identify which features of the deep learning model drive the model's results (Lundberg and Lee 2017). The SHapley Additive exPlanations (SHAP) technique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

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

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

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

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



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

J Conjoint Analysis: Survey Design and Model Estimation

Conjoint Survey Stages The conjoint survey stages are summarized along with their purpose in Table J.1.

Table J.1: Conjoint Survey Design Elements

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

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

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

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

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

Estimation of HB Conjoint Analysis Model We estimated posterior distributions of HB model parameters $\{\{\beta_i\}_{i=1}^N, \Theta, \mu_\Theta, \mathbf{\Lambda}_\beta\}$ with Markov chain Monte Carlo (MCMC) sampling using the No-U-Turn (NUTS) sampler (Hoffman, Gelman et al. 2014). Sampling consisted of 1 chain,[4] each with 2,000 draws of which 2,000 were used for sampler tuning. Convergence of MCMC chains was determined via acceptance criteria of the sampler and its targets (65%), and chain divergences from trace plots (less than 5% draws diverging). Hyperparameter values for prior distributions were determined from overlap of prior draws with posterior draws, and by using both in-sample and out-of-sample hit rates.

[4] Note that we obtained better prediction accuracy numbers with more chains and across parallel GPUs, but report those with a single deterministic chain for replicability.

Our codebase was written in Python using the PyMC library (Patil, Huard, and Fonnesbeck 2010) which leverages the Jax, NumPyro, and Aesara graph compilation libraries to achieve graphical processing unit (GPU) acceleration. Estimation using 1 RTX8000 takes around 15 minutes, with estimation using multiple GPUs approximately dividing the computational time by the number of GPUs, but we note this is heavily dependent on GPU system configuration with CUDA/OpenCL kernel libraries.

References

Baldi, Pierre and Kurt Hornik (1989), "Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima," *Neural Networks*, 2 (1), 53–58.

Bengio, Yoshua, Aaron C Courville, and Pascal Vincent (2012), "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives," *CoRR, abs/1206.5538*, 1 (2665), 2012.

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

Chen, Xi, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel (2016), "Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, pages 2180–2188.

Dubois, Yann, Alexandros Kastanos, Dave Lines, and Bart Melman "Disentangling VAE," `http://github.com/YannDubs/disentangling-vae/` (2019).

Gabbay, Aviv, Niv Cohen, and Yedid Hoshen (2021), "An Image is Worth More than a Thousand Words: Towards Disentanglement in the Wild," *Advances in Neural Information Processing Systems*, 34.

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

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

Hoffman, Matthew D, Andrew Gelman et al. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.," *Journal of Machine Learning Research*, 15 (1), 1593–1623.

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

Kappe, Eelco and Stefan Stremersch (2016), "Drug Detailing and Doctors' Prescription Decisions: The Role of Information Content in the Face of Competitive Entry," *Marketing Science*, 35 (6), 915–933.

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

Karras, Tero, Samuli Laine, and Timo Aila (2019), "A Style-Based Generator Architecture for Generative Adversarial Networks," *Computer Vision and Pattern Recognition*, pages 4401–4410.

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

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

Lee, Thomas Y and Eric T Bradlow (2011), "Automated Marketing Research using Online Customer Reviews," *Journal of Marketing Research*, 48 (5), 881–894.

Lee, Wonkwang, Donggyun Kim, Seunghoon Hong, and Honglak Lee (2020), "High-Fidelity Synthesis with Disentangled Representation," *European Conference on Computer Vision*, pages 157–174.

Linting, Mariëlle, Jacqueline J Meulman, Patrick JF Groenen, and Anita J van der Koojj (2007), "Nonlinear Principal Components Analysis: Introduction and Application.," *Psychological Methods*, 12 (3), 336.

Liu, Xiao, Param Vir Singh, and Kannan Srinivasan (2016), "A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing," *Marketing Science*, 35 (3), 363–388.

Locatello, Francesco, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard

Schölkopf, and Olivier Frederic Bachem (2019), "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," *International Conference on Machine Learning*, pages 4114–4124.

Locatello, Francesco, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen (2020), "Weakly-Supervised Disentanglement Without Compromises," *International Conference on Machine Learning*, pages 6348–6359.

Lundberg, Scott M and Su-In Lee (2017), "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, 30.

Mathieu, Emile, Tom Rainforth, Nana Siddharth, and Yee Whye Teh "Disentangling Disentanglement in Variational Autoencoders," "International conference on machine learning," pages 4402–4412, PMLR (2019).

Mika, Sebastian, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch (1998), "Kernel PCA and De-noising in Feature Spaces," *Advances in Neural Information Processing Systems*, 11.

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

Patil, Anand, David Huard, and Christopher J Fonnesbeck (2010), "PyMC: Bayesian Stochastic Modelling in Python," *Journal of Statistical Software*, 35 (4), 1.

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

Roweis, Sam and Zoubin Ghahramani (1999), "A Unifying Review of Linear Gaussian Models," *Neural Computation*, 11 (2), 305–345 00715.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986), "Learning Representations by Back-Propagating Errors," *Nature*, 323 (6088), 533–536.

Shapley, Lloyd S (1997), "A Value for n-Person Games," *Classics in Game Theory*, 69.

Voynov, Andrey and Artem Babenko (2020), "Unsupervised Discovery of Interpretable Directions in the GAN Latent Space," *International Conference on Machine Learning*, pages 9786–9796.

Williams, Christopher (2000), "On a Connection Between Kernel PCA and Metric Multidimensional Scaling," *Advances in Neural Information Processing Systems*, 13.