Market Structure Mapping with Disentangled Visual Characteristics

(Authors' names blinded for peer review)

Demand models typically use structured data for estimating the value of product characteristics. However, for several product categories such as automobiles, consumers emphasize that visual characteristics of the product are significant demand drivers. Since visual characteristics are typically in high-dimensional unstructured data (e.g., product images), this poses a challenge to incorporate them in demand models. We introduce a method that enables estimation of demand using visual characteristics, by building on the BLP demand model with recent advances in disentangled representation learning. Our method also overcomes the challenge of not having supervised signals, which are required for good disentanglement, by using the demand model as supervisory signal. We discover independent and human interpretable visual characteristics directly from product image data, while simultaneously estimating equilibrium demand in a competitive automobile market in the UK. We conduct a counterfactual analysis using a recent dramatic change in the visual design language of BMW cars, and show our predicted results align with actual changes in BMW market share. To our best knowledge, this work is the first to link visual product characteristics with demand–in other words, to quantify the economic value of design.

Key words: visual analytics, deep learning, demand models

1. Question

Research Questions

1. (Methodological) Does market outcome data help in disentanglement, in addition to structured product characteristics?
2. (Substantive) - Within a brand and product category (segment in case of a car), are products close in the structured space also close in the visual space?
3. (Substantive) - Across firms, is the competition in the visual space a strategic substitute or complement to the competition in the structured space?
4. (Substantive) - Does differentiation across product categories (or segments in car market) increase when visual information is included?
5. (Additional) Across firms, is the competition in the visual space a strategic substitute or complement to the competition in the structured space?
6. (Additional) How much can I change my design without losing brand recognition? (Design Reach)

Approaches

1. Across all make-models, calculate the average pairwise distance to other models in structured as well as visual space. Then calculate rank correlation (spearman) and non-rank corrleation (pearson) within the same segment as well as across all segments.
2. Within a segment and across all segments, calculate how much area is covered by each make as a percentage of the overall area in the structured space as well as visual space.
3. Within a particular segment in a particular year, find the centroid for each make in the structured as well visual space. Then calculate the mean distance to all other centroids in that segment in both spaces. Find the closest make to each make in a particular segment and overall.

Validation

1. Netzer's Text Data / Car Switching Data / Charlie Murry's 2nd Choice Data
2. Interpretability & Quantification - similar to watches for characteristic discovery and quantification
3. Quantification using only real image for distance validation

Notes

1. We will not talk about the generative aspect.
2. Convert all color images to grayscale in the paper.

Immediate Work

1. Full dump of all the existing raw images in the dataset.
2. CSV with latent visual characteristics of data
3. Full dump of all the existing reconstructions in the dataset

Loss Function

1. Multi-Task Learning (Confirm how to combine multiple losses. For example, two different cross entropy losses or a combination of cross entropy and mean squared error loss)

Signals

1. Unsupervised
2. Market Outcome (F.E. from BLP)
3. Price
4. Choose one of MPG / HPWT / Space
5. Combination of Market Outcome + Price + (one of MPG / HPWT / Space)

Exterior look/design is the top reason shoppers avoid a particular vehicle (30%), followed by cost (17%).

–JD Power Avoider Study 2015

2. Introduction

Visual characteristics comprising product form are often a primary factor in a product's market success (Jindal et al. 2016, Veryzer Jr 1993). Visual characteristics are designed by firms for everything from communicating intended product differentiation and segmentation (Bloch 1995, Homburg et al. 2015), signaling brand equity (Aaker 1997), and of course, to making appealing products that consumers choose (Creusen and Schoormans 2005, Norman 2013). Therefore, it is crucial for firms to comprehend not only the relative positioning of their product in the easily measurable performance characteristic space but also in the challenging-to-quantify visual characteristic space. Furthermore, it is important for firms to understand the reasons for the proximity or disparity of two products within this visual characteristic space.

In order to achieve this, it necessitates the quantification of visual design characteristics. This task is inherently challenging, considering that even low-fidelity product form representations, such as wireframes or silhouettes, might require hundreds of highly interdependent variables such as Bezier curve control points. This challenge scales to the millions of variables when using product images.[1] This implies the requirement of not just identifying these visual characteristics, but also the need to quantify them accurately.

While it is feasible for researchers to manually define the visual characteristics of interest, such an approach would necessitate the exercise of researcher judgment across each category, introducing potential biases. To circumvent this, we propose a general methodology applicable across any product category that obviates the need for researcher intervention.

Our objective is ... identify human interpretable visual characteristics and then create a market structure map that uses both visual and structured

We show that a market structure map based on either of these characteristics is incomplete picture Our method is able to uncover insights into why certain products x y z

We compare the insights obtained from a market structuer map created using visual alone, with performance alone and using both.

We describe the method for finding visual characteristics ¡ state of research in deep learning disentanglement ¿ ¡ methodology ¿ ¡ results ¿

[1] For example, even a 1000 pixel × 1000 pixel black and image is 1 million variables, each having a value of 0 or 1.

3. Literature Review

Our work is situated in multiple streams of literature across machine learning and marketing. First is the stream on the use of unstructured data, specifically images, within marketing. Typically, such work has relied on obtaining information from images using deep learning based methods. The second stream is based on advances in machine learning for unstructured data in "disentangled" representation learning. Finally, the application in the paper is related to the stream on market structure mapping. We discuss each of these in detail below.

3.1. Empirical Models with Unstructured Data in Marketing

Unstructured data (text, images etc.) are rich in content and very high dimensional [2], which makes it challenging to tractably incorporate in marketing models. One common approach in marketing literature is for the researcher to define a specific set of attributes of interest. Typically in this approach, after defining a set of attributes, researchers use deep learning based methods to quantify those attributes. For example, in a study on AirBnB images, Zhang et al. (2022) finds photographic attributes of images such as color, composition and aspect ratio and combines them with photo quality which is quantified using a deep learning method. Similarly, Zhang and Luo (2022) finds photographic attributes of images such as color, composition and figure-ground relationship of images uploaded by customers to Yelp to study their effect on restaurant exit, and Malik et al. (2019) specifies interpretable properties of profile pictures from an online professional social network such as photographic quality (e.g., blur, exposure etc.) and facial characteristics (e.g., beard, lip makeup etc.) to study the beauty premium in career progression.

Overall, in this approach, identifying the selected characteristics in advance relies on domain knowledge and researcher judgment. This process creates a challenge if researchers select an unimportant visual characteristics or more importantly omit an important one.[3]

A second approach is to use classical statistical methods like principal component analysis (PCA) or autoencoders (Bengio et al. 2013), which can reduce the dimensionality automatically. However the disadvantage is that the characteristics obtained are generally

[2] For instance, images are high-dimensional data since even a modest-sized image of 1,000×1,000 pixels exists in a 1,000,000-dimensional space.

[3] For example, a researcher studying the effect of hair color on a worker's chance of getting promoted might conclude that hair color has explanatory power. However, hair color is correlated with race and the explanation behind the worker's promotion might possibly be due to race.

not interpretable. Also, multiple true characteristics might be collapsed into one dimension, for example in PCA, into a single eigenvector. Other empirical work using image data in marketing includes the study of how a brand's visual identity can be inferred using online images (Liu et al. 2020), the use of images to augment how designers develop products (Burnap et al. 2022, Dew et al. 2022), and to predict return rates using product images (Dzyabura et al. 2019). In a study conducted by **?**, photographs of the frontal facades of various car models were captured and subsequently processed through morphing software to generate morphs for each model. This was accomplished by identifying feature points that denoted the salient elements of each design. The same methodology was employed by Liu et al. (2017), who investigated the influence of product aesthetics on consumer demand. It is imperative to note that these methodologies predominantly rely on human experts for the identification and quantification of the visual attributes in question.

In stark contrast, our proposed methodology leverages disentanglement learning to autonomously extract and quantify the visual characteristics, thereby circumventing the necessity for manual intervention. Our work is distinct with our use of human interpretabe visual to create market structure maps.

3.2. Disentangled Representation Learning

Representation learning is a machine learning sub-field that theorizes that high-dimensional data is generated from low-dimensional factors. According to Bengio et al. (2013), the goal is to learn data representations that simplify the extraction of useful information for building predictive models. This paper focuses on a branch of representation learning called disentangled representation learning, which aims to isolate meaningful factors of variation in data Bengio et al. (2013). Take the dSprites dataset (Higgins et al., 2017) as an example, which contains 2D images of objects with different shapes, sizes, colors, and positions. Disentangled representation learning seeks to separate these factors, identifying shape, size, color, and position as four latent dimensions. Notably, disentanglement identifies only the real factors of variation, regardless of the dimensionality of the latent space. We use this method to automatically learn (discovery) and quantify human-interpretable visual characteristics of products without human labeling or intervention.

One key challenge of any disentanglement method is that, with purely unsupervised methods, there is no theoretical guarantee for learning unique disentangled representations Locatello et al. (2019). In other words, we need some form of relevant supervision to identify independent and semantically interpretable visual characteristics. To address this challenge, 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. However, since our aim to identify the visual characteristics automatically without specifying them in advance, we can not use this approach. Instead, in this paper, we address this theoretical challenge by obtaining alternative supervisory signals derived from a classic model of market equilibrium (Berry et al. 1995). The model of market equilibrium provides us with product fixed effects that can be used as supervisory signals. Implicitly, we assume that product fixed effects capture consumer preferences on unobservable product characteristics including visual characteristics. We compare this alternative source for supervisory signals with signals obtained from structured product characteristics to study whether they alone or together allow us to overcome the well-known challenge in the deep learning literature.

3.3. Market Structure Mapping

Market structure mapping is one of the primary and commonly used methods in competitive marketing strategy (Rao et al. 1986). Firms use it to understand the relative positions of their products with respect to their rivals in order to inform brand positioning; new product development; and product, advertising, and pricing strategies (Urban et al. 1984, DeSarbo et al. 1993, Bergen and Peteraf 2002, Lattin et al. 2003, DeSarbo et al. 2006). Market structure methods in marketing have used a variety of data sources, such as panel-level scanner data (Erdem 1996), consumer search data (Ringel and Skiera 2016, Kim et al. 2011), online product reviews (Lee and Bradlow 2011, Tirunillai and Tellis 2014), social media engagement data (Liu et al. 2020, Yang et al. 2022), co-occurrence of products in shopping baskets (Gabel et al. 2019) as well as co-occurrence of products in product reviews (Netzer et al. 2012).

The input to commonly used methods to map the market structure is a similarity matrix between all pairs of products. Existing work on market structure map has ignored visual appearance of products. Although it can be overcome using the survey approach by asking consumers to provide visual similarity scores between different pairs of products, it would still be very inefficient and not automatic because we would need to cover all n^2 pairs and would also suffer from a bias from human raters if they focus on different aspects of design. Moreover, the absence of underlying visual characteristics in this approach would not inform managers on the reason why two products are located close together or further apart on the market structure map. In this work, we discover interpretable visual product characteristics from product images to inform the market structure mapping process. We compare the market structure maps based on structured as well as visual space and identify products that are less differentiated in the structured space but more differentiated in the visual space and vice versa.

4. Methodology

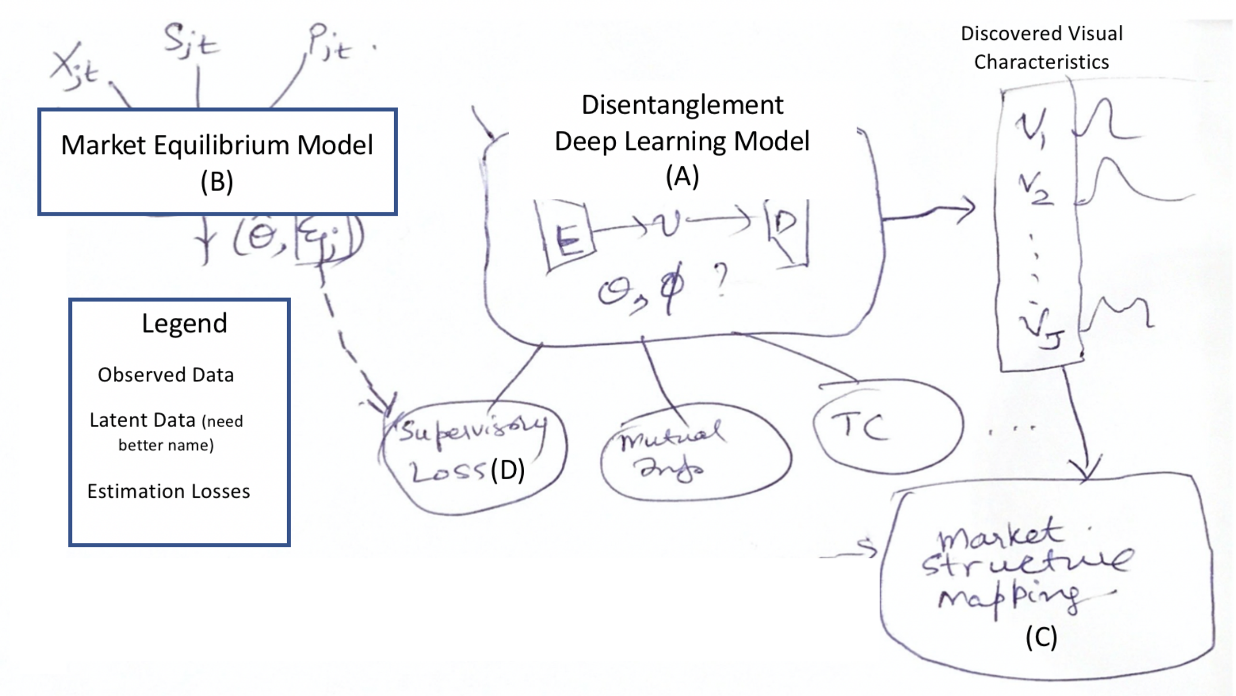
Ale / Vineet to look at methodology

1. Motivation for supervision (impossibility theorem)
2. Where does supervisory signal come from? Sources of supervisory signals. CS lit (Ground truth) but we don't know what we are looking for. CS lit also says that noisy signals that are correlated with ground truth on visual char can also work.
3. Previous paper suggested str pdt char
4. Augment that with consumer choices (use a lot of information that marketers don't observe but consumers do). Part of that includes unobs pdt quality, reliability, visual appearance, brand etc. Idea (if consumers are making choices based on visual pdt char, then consumer choices are correlated with visual char).
5. But consumer choices are also dependent on other pdt char + price. So how do we get a cleaner or less noisy signal that is not just consumer choices. BLP allows to separate out what is observable to the researcher and what is unobservable to the researcher. BLP gives us product-market fixed effect for every year but we include the fixed effect at a make-model or a product level.
6. Make-Model Fixed Effect v Market Share (0.63 v 0.61): This doesn't need to be in the section but backs up our hypothesis that xi should be better than share.
7.

Source will have a branch : BLP + Xs

Our method for discovering and mapping the visual product characteristics has three components as illustrated in Figure TODO. The first component (A) is a disentanglement based approach to obtain visual product characteristics, which promotes independence and human interpretability. Disentanglement methods require the ground truth of the visual characteristics as supervisory signals, which is what we aim to discover and hence they are

Figure 1 **Overview of Methodology**


X_{jt}
S_{jt}
P_{jt}
Market Equilibrium Model (B)
$f(\theta, \xi_j)$
Legend
Observed Data
Latent Data (need better name)
Estimation Losses
Disentanglement Deep Learning Model (A)
E → v → D
θ, ϕ ?
Discovered Visual Characteristics
V_1
V_2
V_J
Supervisory Loss (D)
Mutual Info
TC
. . .
Market Structure Mapping (C)

not available. This requires us to obtain supervisory signals from an alternative source. The second component (B) provides supervisory signals from a model of market equilibrium, which uses traditional structured data, as well as from structured product characteristics directly. Finally, the third component (C) uses both the structured characteristics and the discovered visual characteristics to create competitive market structure maps.

4.1. Disentanglement Model

Our approach to extract human interpretable quantified visual characteristics is based on disentangled representation learning. These methods discover low dimensional representations, in our case visual attributes, that are statistically independent and more likely to be human interpretable (Bengio et al. 2013). The majority of techniques for disentanglement are based on deep generative models, including variational autoencoders (VAE) (Kingma and Welling 2014) and generative adversarial networks (GAN) (Goodfellow et al. 2020). We build a disentanglement model based on a VAE, whose table of notation is laid out in Table 1.

Ankit: This section needs to be re-written. Cite other paper and include the new points. Write a shorter revision.

Table 1 Table of Notation in Disentanglement Model

Symbol	Category	Meaning
$\mathbf{m}$	Input Data	Product image
$\mathbf{y}$	Input Data	Supervisory signal(s)
$\widehat{\mathbf{m}}$	Output Data	Reconstructed image
$\widehat{\mathbf{y}}$	Output Data	Predicted Supervisory Signal(s)
$\mathbf{v}$	Latent Space	Visual characteristic vector
$\mathbf{v}_{\text{inf}}$	Subset of Latent Space	Informative visual characteristic
$\rho(\mathbf{v})$	Model	Prior distribution
$\rho_\theta(\mathbf{m}\vert\mathbf{v})$	Decoder Neural Net	Conditional Probability of Generating Image Data given Latent Space
$q_\phi(\mathbf{v}\vert\mathbf{m})$	Encoder Neural Net	Conditional Probability of Latent Space given Image Data
$\rho_\psi(\mathbf{y}\vert\mathbf{v})$	Supervisory Neural Net	Conditional Probability of Supervisory Signal given Latent Space
θ	Weights of Neural Net	Decoder's parameters
ϕ	Weights of Neural Net	Encoder's parameters
ψ	Weights of Neural Net	Supervisory Net's parameters
$\mathbf{E}_{q_\phi(\mathbf{v}\vert\mathbf{m})}[\log \rho_\theta(\mathbf{m}\vert\mathbf{v})]$	Loss Function	Reconstruction Loss
$I_q(\mathbf{v},\mathbf{m})$	Loss Function	Mutual Information Loss
$KL\left[q(\mathbf{v})\vert\vert\prod_{j=1}^{J} q(v_j)\right]$	Loss Function	Total Correlation Loss
$\sum_{j=1}^{J} KL\left[q(v_j)\vert\vert p(v_j)\right]$	Loss Function	Dimension KL Divergence Loss
$P(\hat{y}(\mathbf{v}),y)$	Loss Function	Supervised Loss
$\mathcal{L}(\theta,\phi,\psi;\mathbf{m},\mathbf{v},\mathbf{y})$	Loss Function	Total Loss
J	Hyperparameter	Dimensionality of latent space
λ_1	Hyperparameter	Weight on Total Correlation Loss
λ_2	Hyperparameter	Weight on Supervised Loss

We have a dataset $\mathbf{m}$ of images. We assume that they are generated from a distribution parameterized by visual characteristics $\mathbf{v}$. The encoder takes images $\mathbf{m}$ as input and reduces their dimensionality to obtain the discovered visual characteristics $\mathbf{v}$. The decoder takes $\mathbf{v}$ as input and outputs a reconstruction $\widehat{\mathbf{m}}$ of the original image $\mathbf{m}$. Both the encoder and decoder are deep neural networks parameterized by corresponding model parameters ϕ and θ. The generative model is a combination of the prior $\rho(\mathbf{v})$ set to an isotropic unit Gaussian $\mathcal{N}(0,1)$ and a decoder neural net $\rho_\theta(\mathbf{m}|\mathbf{v})$. The true posterior is intractable as in variational Bayesian inference (Blei et al. 2017) and so it is approximated as $\log q_\phi(\mathbf{v}|\mathbf{m}) = \log \mathcal{N}(\mathbf{v};\boldsymbol{\mu_d},\boldsymbol{\sigma_d}^2\mathbf{I})$ where $\boldsymbol{\mu_d}$ and $\boldsymbol{\sigma_d}$ are the mean and the s.d. of the approximate posterior. The loss for the original VAE is written in Equation (1). We refer readers to Kingma and Welling (2014) for its detailed derivation.

$$\underbrace{L(\theta,\phi;\mathbf{m},\mathbf{v})}_{\text{VAE Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{v}|\mathbf{m})}\left[\log \rho_\theta(\mathbf{m}|\mathbf{v})\right]}_{\text{Reconstruction Loss}} + \underbrace{KL\left[q_\phi(\mathbf{v}|\mathbf{m})||\rho(\mathbf{v})\right]}_{\text{Regularizer Term}} \tag{1}$$

In Equation (2), we decompose the regularizer term in Equation (1) into three terms (Chen et al. 2018, Hoffman and Johnson 2016, Kim and Mnih 2018). We follow the $\beta-$TCVAE method (Chen et al. 2018) by imposing a heavier penalty on the total correlation loss term. We provide an intuition behind each of the loss terms in Table 2. In addition to the losses detailed in Equation (2), Table 2 includes a supervised loss.

$$\begin{aligned}\underbrace{L(\theta,\phi;\mathbf{m},\mathbf{v})}_{\text{Open Loop Loss}} &= \underbrace{\mathbf{E}_{q_\phi(\mathbf{v}|\mathbf{m})}\left[\log \rho_\theta(\mathbf{m}|\mathbf{v})\right]}_{\substack{\text{Reconstruction}\\ \text{Loss}}} + \underbrace{I_q(\mathbf{v},\mathbf{m})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} \\ &+ \lambda_1 \underbrace{KL\left[q(\mathbf{v})||\prod_{j=1}^{J} q(v_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}} + \underbrace{\sum_{j=1}^{J} KL\left[q(v_j)||\rho(v_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}}\end{aligned} \tag{2}$$

Fixed effects from BLP; anything that is not captured in structured / what do consumers care about? prices, str pdt char, visual design, unobs quality like reliability, brand (xijt captures last 3 in fixed effects)

Our methodological contribution is the use of supervised learning via a model of market equilibrium to help overcome a key challenge in disentangled representation learning. Specifically, Locatello et al. (2019) proved there are no theoretical guarantees to learning disentangled representations without supervision to align disentangled representations with their ground truth. It implies that disentangled and entangled representations are equivalent in the absence of a supervisory signal. As a result, the deep learning literature has focused attention on improving disentanglement methods using benchmark datasets with known ground truth labels corresponding to each visual characteristics Locatello et al. (2020). However, *these ground truth labels are precisely the visual characteristics we aim to discover.*

We overcome this challenge by leveraging combination of a model of market equilibrium (Berry et al. 1995) as well as structured product characteristics to provide supervisory signals. In particular, first we describe a demand system comprising a competitive market of firms and heterogeneous consumers that allows us to estimate fixed effects corresponding to each product. We assume that the product fixed effects captures consumer preferences over disentangled human interpretable visual characteristics and thus allows to serve as

Table 2 Loss terms for proposed method

Loss Term	**Notation**	**Intuition**
Reconstruction Loss	$\mathbf{E}_{q_\phi(\mathbf{v}\vert\mathbf{m})}\left[\log \rho_\theta(\mathbf{m}\vert\mathbf{v})\right]$	Penalizing this term encourages the input data $\mathbf{m}$ to be as similar to the reconstructed output $\hat{\mathbf{m}}(\mathbf{v})$ as possible. This means that we want the discovered visual characteristics to have the necessary information so that the reconstructed output is as close as possible to the input image.
Mutual Information Loss	$I_q(\mathbf{v},\mathbf{m})$	Penalizing this term means encouraging the visual characteristics $\mathbf{v}$ store as little information about the product image $\mathbf{m}$ as possible from an information-theoretic point of view (Achille and Soatto 2018). Although it seems counter intuitive that penalizing this term or discouraging the visual characteristics to share as little mutual information with the product image would encourage disentangled representation, but since it allows the visual characteristics to not store any nuisance information, it promotes visual characteristics to discard nuisance information.
Total Correlation Loss	$KL\left[q(\mathbf{v})\vert\vert\prod_{j=1}^{J} q(v_j)\right]$	Penalizing this term encourages the discovered visual characteristics $\mathbf{v}$ to be statistically independent (Watanabe 1960). A λ_11 penalty on this term means that we are promoting the KL divergence to be as close to zero as possible. When the KL divergence is zero, then the discovered visual characteristics are statistically independent. By imposing a heavier penalty on this term, we seek to promote statistically independent discovered visual characteristics that also aid in achieving high reconstruction accuracy instead of only finding any set of visual characteristics that aid in achieving high reconstruction accuracy.
Dimension-Wise KL Loss	$\sum_{j=1}^{J} KL\left[q(v_j)\vert\vert\rho(v_j)\right]$	Penalizing this term encourages the distribution of each visual characteristic of every datum to be close the prior distribution. The prior is typically assumed to be Gaussian. This term promotes a continuous latent space, which allows generation from a smooth and compact region of latent space.
Supervised Loss	$MSE(\widehat{\mathbf{y}(\mathbf{v})},\mathbf{y})$	Penalizing this term encourages the discovered visual characteristics $\mathbf{v}$ to obtain high accuracy in predicting $\mathbf{y}$. When the signal is discrete, we use cross-entropy loss for the multiclass classification prediction task, and for a continuous signal, we use mean squared loss for the regression prediction task.

supervisory signals for the disentanglement-based deep learning method. [4]. We use the product fixed effects in combination with structured product characteristics to obtain a vector of supervisory signals. Next, we train a disentanglement-based machine learning model to find visual characteristics from images of products that can predict the vector of supervisory signals, i.e. fixed effects obtained from the demand model and structured product characteristics. Our approach is motivated by the result from Locatello et al. (2020) that even weak supervision with noisy metrics of ground truth is able to achieve good disentanglement.

[4] Product fixed effects would also capture consumer preferences over other product unobservable characteristics such as quality etc.

Why Product Fixed Effects? We posit that consumers have preferences over visual product characteristics, which in turn impacts their choices, which finally affects market outcomes. Thus, metrics based on market outcomes are likely to be correlated with visual product characteristics, implying that they can serve as relevant supervisory signals to help discover and obtain disentangled visual characteristics. However, since there might be other factors that impact market outcomes as well, we use a model of market equilibrium to isolate the unobserved product characteristic, which includes the impact of all unstructured product characteristics (among other factors).[5] Thus, rather than just using the market shares corresponding to each product as the supervisory signal, we use the estimated product-level effects as the signal.

While it may seem we have several possible visual characteristics to discover but only one signal for each product, the machine learning literature has show that even weak supervision with a signal that has some correlation to ground truth will help obtain a disentangled representation in practice. This underlying motivation leads us to include the estimated product-level effects obtained from the market equilibrium model, since the effect is likely *weakly correlated* with each of the visual characteristics. Other aspects (e.g. unobserved quality) may also affect this variable.[6]

Why Structured Product Characteristics? Consider why specific structured product characteristics might work to supervise visual characteristics. Typical structured characteristics commonly available in marketing data include brand, material, performance characteristics and price. First, consider a characteristic like material, e.g. silver that provides a certain visual look to a product. Material more broadly is known to significantly affect visual appearance and consumer perceptions (Fleming 2014). Second, a product characteristic like brand is likely to strongly impact visual look of a product. Consider, for instance the distinct look of a Mercedes-Benz car or a Louis Vuitton handbag. The signature of the brand design is often visibly present and apparent from the product's appearance to consumers, especially for product categories with visible consumption (Simonson and Schmitt 1997, Liu et al. 2020, Ferraro et al. 2013) or luxury brands (Megehee and Spake 2012, Lee et al. 2018). Further, existing marketing research has shown that brands have different

[5] It is also possible to use only the demand model to obtain these product-level fixed effects.

[6] However, this approach might be problematic if we want to discover visual characteristics where consumers have no preferences at all over such characteristics, since the signal would be uncorrelated with visual characteristics in that case.

personalities (Aaker 1997) that can be expressed through their product-related characteristics, product category associations, brand name, symbol or logo, advertising style, price, distribution channel and user imagery (Batra et al. 1993, Liu et al. 2020). Third, consider the role of price, which is strictly speaking not a product characteristic, since it can be set by the retailer. However, many brands, especially luxury brands, maintain carefully curated pricing tiers with strong consumer associations. For example, even when BMW cars share mechanical components, the visual appeal of the product line at different price points is quite distinct, with the 3 Series and the 5 Series looking distinctly different than the 4 Series and the 7 Series.[7] In other cases, a structured characteristic like engine size might impact the size of the front grille for engineering reasons. Overall, for a wide set of product characteristics, we can see the mechanism by which product characteristics from structured data or price might be correlated with and hence predictive of visual characteristics. Thus, we posit that these could serve as useful supervisory signals for disentangling visual characteristics.

Of course, in practice, which of the characteristics turn out to be useful supervisory signals may be highly dependent on the empirical setting at hand. We test whether the product fixed effects alone or structured product characteristics alone or a combination of both of them serve as superior supervisory signals.

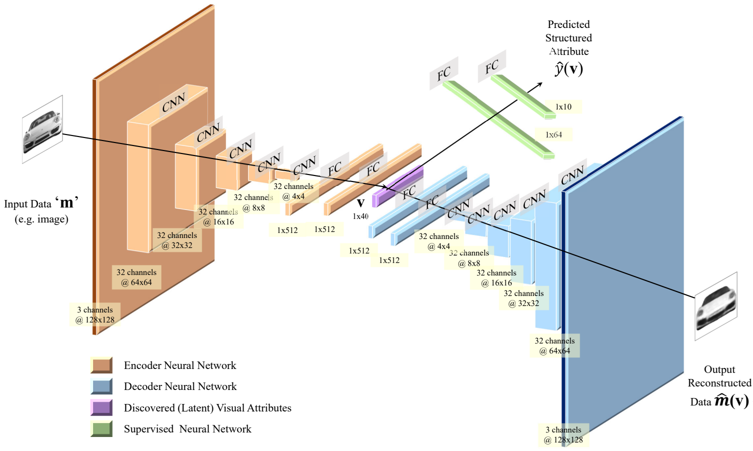
The modified loss equation is specified in Equation (3). The supervisory signal vector $\mathbf{y}$ is predicted from the conditional distribution $\rho_\psi(\mathbf{y}|\mathbf{v})$ where the supervised neural net is parameterized by ψ.

$$\underbrace{L(\theta,\phi;\mathbf{m},\mathbf{v})}_{\text{Disentanglement Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{v}|\mathbf{m})}[\log p_\theta(\mathbf{m}|\mathbf{v})]}_{\substack{\text{Reconstruction}\\ \text{Loss}}} + \underbrace{I_q(\mathbf{v},\mathbf{m})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} + \lambda_1 \underbrace{KL\left[q(\mathbf{v})||\prod_{j=1}^{J} q(v_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}}$$

[7] We quote BMW's Design Chief Domagoj Dukec (https://www.motor1.com/news/581345/bmw-design-boss-some-cars-polarizing/).

> Two-thirds of customers want *an elegant and harmonious aesthetic*, which is why volume sellers like the 3 Series and 5 Series play it safe in terms of design. However, for the remaining 33 percent of customers who want to stand out from the crowd, cars like the 4 Series and the new 7 Series cater to those who *really want to polarize.* These customers *want a more irrational car, and they're willing to pay more for that emotional expression, and to really make a statement.*

Figure 2 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 200 epochs.

$$+ \underbrace{\sum_{j=1}^{J} KL\left[q(v_j)||p(v_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \lambda_2 \underbrace{P(\widehat{\mathbf{y}(\mathbf{v})}, \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}} \tag{3}$$

4.1.1. Neural Net Architecture Figure 2 shows the detailed neural net architecture. Our architecture is a modified version of the one used in Burgess et al. (2017). We modify the architecture to use images of 128×128 pixels as well as to incorporate a demand model. We use Convolutional Neural Net (CNNs) to construct the encoder neural net because we are working with images. We stack a sequence of CNN layers in the encoder neural net so that we learn high-level concepts for images. We then use 2 fully-connected (FC) layers to first flatten the output of the sequence of CNN layers and then reduce the number of dimensions in order to learn a maximum of J visual characteristics. The decoder neural net is simply the transpose of the encoder neural net, and is designed to reconstruct the image from the 20-dimensional latent visual characteristics. Finally, we feed the discovered visual characteristics to predict the vector of supervisory signals that serve as labels.

Why can't λ_1 and λ_2 be estimated? Any ML literature with 3 cites. Why is there a separation between paramters and hyperparams? What is the underlying logic?

4.1.2. Hyperparameter Selection We make modeling choices in the form of several hyperparameters, which impact the estimation process but are not parameters estimated with the model (e.g., number of training epochs). First, we have a hyperparameter λ_1, which is the weight on the total correlation loss within the disentanglement loss (Chen et al. 2018). Second we have a hyperparameter λ_2, which represents the weight of the supervised loss when incorporated into the loss of the disentanglement model. For instance, a higher value of λ_2 will weigh or prioritize that visual characteristics are able to predict the vector of supervisory signal(s) more relative to other loss terms like mutual information or reconstruction loss, which could reduce the quality of disentanglement. On the other hand, a zero value on λ_2 implies that we are not addressing the impossibility theorem and thus have no theoretical guarantees to find disentangled visual characteristics. Following Locatello et al. (2020), we select the hyperparameter λ_1 and λ_2 corresponding to the lowest 10-fold cross-validated supervised loss for each vector of supervisory signal.

We use Unsupervised Disentanglement Ranking (UDR), a metric proposed by Duan et al. (2020), to compare the quality of disentangled representations produced by different vectors of supervisory signals. UDR is an automated method that 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). The UDR metric measures the robustness of disentangled representations to variance at different starting points. It relies on the assumption that for a particular dataset, a disentangling VAE will converge on the same disentangled representation (up to *permutation*, *sign inverse*, and *subsetting*). We detail the steps involved in selecting hyperparameters as well as the intuition behind the hyperparameters in Table 3.[8]We select the vector of supervisory signal corresponding to the highest UDR. We describe the steps to calculate UDR below in Table 4.

4.2. Demand Model

We use a well-known model of oligopoly market equilibrium to obtain the product fixed effect corresponding to each product, which is then used as a supervisory signal. We detail

[8]

1. Permutation: The same ground truth factor may be encoded by a model with two different seed values at a different index position.
2. Sign inverse: A model with two different seed values may learn to encode the values of the generative factor in the opposite order to each other.
3. Subsetting: A model from one seed value may learn a subset of the factors that the a model with a different seed value has learned. This is because different seed values may encourage a different number of latents to be switched off in the two models.

Table 3 Hyperparameter Selection

Step	Description
1	Set all the hyperparameters except λ_1 and λ_2
2	Set[‡1] batch size, number of visual characteristics, learning rate, and the number of epochs
3	Obtain disentangled representations for every combination of $(\lambda_1, \lambda_2) \in \Lambda_1 \times \Lambda_2$ and λ_2 for each vector of supervisory signal
4	Select λ_1 and λ_2 corresponding to lowest supervised loss for each vector of supervisory signal
5	Calculate Unsupervised Disentanglement Ranking (UDR)[‡2] Duan et al. (2020) and select supervisory signal with the highest UDR
6	Obtain the learned visual characteristics $\mathbf{v}$

[‡1] Intuition for Hyperparameters:

1. Batch Size: On the one hand, if a very low value for batch size is used, then the model takes longer to converge. On the other hand, if a very high value for batch size is used, then the model loses its generalizability beyond the training set.
2. Number of Visual Characteristics: On the one hand, if a very low number of visual characteristics is specified, then the model would force multiple factors of variation to be coded into a single visual characteristic. On the other hand, if a very high number of visual characteristics is specified, then the model would encourage a single factor of variation would be split into multiple visual characteristics.
3. Learning Rate: On the one hand, if a very low learning rate is used, then the model can get stuck on a local minima. On the other hand, if a very high learning rate is used, then the model may overshoot the minima.
4. Number of Epochs: On the one hand, training for a very low number of epochs may lead the model not to converge. On the other hand, training for a very high number of epochs may lead to overfitting the training data.

[‡2] See Table 4 for details related to UDR

the BLP demand model (Berry et al. 1995) below with the specification laid out in Berry et al. (1999). We detail the table of notation for the demand model in Table 5.

Make it in to 2 columns - maybe demand and supply

Consumers: In each market $t = 1, \ldots, T$, there are J_t differentiated goods and I_t consumers. For each market, we observe average quantities, prices and product characteristics for all J_t products. The indirect utility of consumer i from purchasing product j in market t is a function of observed product characteristics $\mathbf{x_{jt}}$, unobserved product-market characteristics ξ_{jt}, price p_{jt}, consumer characteristics $\nu_{\mathbf{it}}$. y_{it} is the income of the consumer i in market t, ν_{it}^k represents consumers i's taste for characteristic k in market t, and finally, ϵ_{ijt}

Table 4 UDR Algorithm

Step	Description
1	For each trained model $\tau(\lambda, \mathbf{y})$, perform $\kappa = 10$ pairwise comparisons
2	Pairwise comparisons: models trained with the same λ and $\mathbf{y}$ but with different seed values
2	Calculate the $UDR_{\tau_{s_1}\tau_{s_2}}$, where τ_{s_1} and τ_{s_2} index the model τ learned with two different seed values
3	Calculate $UDR_{\tau_{s_1}\tau_{s_2}}$ score as similarity matrix $R_{\tau_{s_1}\tau_{s_2}}$ where each entry is the Spearman correlation between the responses of individual latent units of the two models.
4	Calculate absolute value of the similarity matrix as $\lvert R_{\tau_{s_1}\tau_{s_2}} \rvert$
5	Compute the score $UDR_{\tau_{s_1}\tau_{s_2}}$ for each pair of models[‡]
6	Compute the final score UDR_{τ} for model τ by taking the median of $UDR_{\tau_{s_1}\tau_{s_2}}$

[‡] $UDR_{\tau_{s_1}\tau_{s_2}} = \frac{1}{v_{\inf_a} + v_{\inf_b}} \left[\Sigma_b \frac{r_a^2 I_{KL}(b)}{\Sigma_a R(a,b)} + \Sigma_a \frac{r_b^2 I_{KL}(a)}{\Sigma_b R(a,b)} \right]$, where a and b index the latent units of models τ_{seed_1} and τ_{seed_2}, respectively, $r_a = max_a R(a,b)$ and $r_b = max_b R(a,b)$. I_{KL} indicates an *informative* visual characteristics within a model and $v_{\inf}$ is the number of such characteristics: $v_{\inf_a} = \Sigma_a I_{KL}(a)$ and $v_{\inf_b} = \Sigma_b I_{KL}(b)$

Table 5 Table of Notation in Demand Model

Symbol	**Meaning**
j	Products
t	Markets
i	Consumers
f	Firms
T	Number of Markets
J_t	Number of products in market t
I_t	Number of consumers in market t
F_t	Number of firms in market t
ζ	Model Parameters
ζ_1	Linear demand-side parameters
ζ_2	Non-linear common parameters
ζ_3	Linear supply-side parameters
p_{jt}	Price
c_{jt}	Marginal Cost
x_{jt}	Observed product characteristic
U_{ujt}	Indirect utility
δ_{jt}	Mean utility
μ_{ijt}	Heterogeneous utility
ϵ_{ijt}	Idiosyncratic taste shock
d_{ijt}	Choice indicator
s_{ijt}	Choice probability
s_{jt}	Market share
ξ_{jt}	Demand-side structural error
$\tilde{\xi}_j$	Product-level fixed effects
ω_{jt}	Supply-side structural error
Z^D	Demand Instruments
Z^S	Supply Instruments
W	Weighting matrix
g	Sample Moments

denotes a mean-zero idiosyncratic taste shock. The indirect utility is specified in Equation (4).

$$U_{ijt} = \mathbf{x_{jt}}\overline{\beta} - \alpha p_{jt}/y_{it} + \xi_{jt} + \sum_k (\sigma_\beta^k x_{jt}^k \nu_{it}^k) + \epsilon_{ijt} \tag{4}$$

Price p_{jt} is typically endogeneous, and based on the unobserved product-market characteristics ξ_{jt}, and hence correlated with it. The unobserved product-market characteristics ξ_{jt} can reflect hard to quantify aspects of the product such as quality or style. The unobserved product characteristics can be decomposed into product fixed effect $\tilde{\xi}_j$ and rest of the unobserved product-market characteristics $\Delta\xi_{jt}$. This decomposition is written in Equation (5).

Bring more attention to the equation.

$$\xi_{jt} = \Delta\xi_{jt} + \tilde{\xi}_j \tag{5}$$

This decomposition is important since the product fixed effects $\tilde{\xi}_j$ is used as the supervisory signal for the disentanglement learning model. If the visual data at the model level is time varying at a lower frequency (e.g. every 5 years) than the sales data (every year), then it would be possible to model separate fixed effects at this lower frequency. We do not use such data in our estimation, hence model just product-level fixed effects. Each consumer i in market t has unit demand, and chooses from the set $J_t = \{0, 1, \ldots, J_t\}$, including the outside good denoted by $j = 0$, which represents no purchase and is given by $U_{i0t} = \epsilon_{i0t}$. Consumers select the alternative (including outside good) with the highest utility:

$$d_{ijt} = \begin{cases} 1 \text{ if } U_{ijt} > U_{ilt} \text{ for all } l \neq j \\ 0 \qquad\quad \text{otherwise} \end{cases} \tag{6}$$

Note that as in BLP, we can decompose the indirect utility in Equation (4) into a mean utility, $\delta_{\mathbf{jt}}$, and a deviation from that mean, μ_{ijt}.

$$\begin{aligned} \delta_{jt}(\mathbf{x_{jt}}, p_{jt}, \Delta\xi_{jt}, \tilde{\xi}_j; \zeta_{\mathbf{1}}) &= \mathbf{x_{jt}}\overline{\beta} + \Delta\xi_{jt} + \tilde{\xi}_j \\ \mu_{ijt}(\mathbf{x_{jt}}, p_{jt}, \nu_{\mathbf{ijt}}, y_i; \zeta_{\mathbf{2}}) &= -\alpha p_{jt}/y_{it} + \sum_k (\sigma_\beta^k x_{jt}^k \nu_{it}^k) + \epsilon_{ijt} \end{aligned} \tag{7}$$

The parameter vector is denoted $\zeta = (\zeta_{\mathbf{1}}, \zeta_{\mathbf{2}})$. The vector $\zeta_{\mathbf{1}}$ contain the linear parameters or the mean preference on $\mathbf{x_{jt}}$, i.e. $\overline{\beta}$. These preferences are common across all consumers.

The vector $\boldsymbol{\zeta_2}$ contain the nonlinear parameters or the standard deviation from mean preference i.e. σ_β as well as the term on the price α. These nonlinear parameters introduce heterogeneity in preferences over structured product characteristics.

Using the standard assumption that ϵ_{ijt} are i.i.d. with the Type I extreme value distribution, the probability s_{ijt} that consumer i chooses product j in market t and aggregate product market shares are given by equation (8) below.

$$s_{ijt} = \frac{\exp(\delta_{jt} + \mu_{ijt})}{\Sigma_{l\in J_t}\exp(\delta_{lt} + \mu_{ilt})} \qquad \text{and} \qquad s_{jt} = \int \frac{\exp(\delta_{jt} + \mu_{ijt})}{\Sigma_{l\in J_t}\exp(\delta_{lt} + \mu_{ilt})} dFi \tag{8}$$

Firms: We assume that automobile firms, indexed by f and part of a set F_t, play a static, full information, simultaneous move pricing game each period. Firms choose the price levels of all their models (products) with the objective of maximizing overall profit. We specify a constant marginal cost c_{jt} for a product j in market t. The pricing first order condition for vehicle j is given by Equation (9).

$$s_{jt} + \Sigma_{j\in J_t}(p_{jt} - c_{jt})\frac{\partial s_{jt}}{\partial p_{jt}} = 0 \tag{9}$$

We parameterize the marginal costs as written below in Equation (10).

$$c_{jt} = \mathbf{x_{jt}}\boldsymbol{\gamma_1} + \mathbf{w_{jt}}\boldsymbol{\gamma_2} + \omega_{jt} \tag{10}$$

where $\mathbf{x_{jt}}$ are product characteristics, $\mathbf{w_{jt}}$ are observable cost-shifters and ω_{jt} are unobserved cost-shifters. We can estimate the marginal costs for each product when we solve the supply model jointly with the demand model.

Instruments: In this demand model, we assume that a consumer's utility depends up on the observed product characteristics as well as unobserved (to the researcher) product characteristics. Firms observe these unobserved product characteristics and set then set prices, which implies that price is endogenous and necessitates the use of instruments.

We use BLP instruments in our analysis, other possibilities are detailed in **??**.

$$Z_{BLP} = \{1, x_{jt}, w_{jt}, \Sigma_{j\in J_t\ \{j\}}1, \Sigma_{j\notin J_t}1, \Sigma_{j\in J_t\ \{j\}}x_{jt}, \Sigma_{j\notin J_t}x_{jt}\} \tag{11}$$

With the addition of demand instruments Z^D_{jt}, we construct demand-side moment conditions of the form $E[\tilde{\xi}_{jt}Z^D_{jt}] = 0$. Similarly, we also construct supply-side moment conditions of the form $E[\omega_{jt}Z^S_{jt}] = 0$ using supply instruments Z^S_{jt}.

GMM Estimation: We construct a GMM estimator using both supply-side and demand-side moment conditions.

$$g(\theta) = \begin{bmatrix} \frac{1}{N}\Sigma_{jt} E[\xi_{jt} Z^D_{jt}] \\ \frac{1}{N}\Sigma_{jt} E[\omega_{jt} Z^S_{jt}] \end{bmatrix} \tag{12}$$

We construct a nonlinear GMM estimator for ζ with some weighting matrix W in Equation (13). We solve this problem twice. First, we obtain a consistent estimate of W and then an efficient GMM estimator.

$$\widehat{\theta} = \underset{\theta}{min} \;\; g(\zeta)' W g(\zeta) \tag{13}$$

5. Empirical Setting

add some text

5.1. Data

We compiled a data set covering 2008 through 2017 consisting of automobile characteristics, market shares and their images from the United Kingdom (UK). We obtain information on sales (in 1000's) and images of the automobiles from DVM-CAR (Huang et al. 2021). Market research studies have shown that up to 70% of consumers identify and judge automobiles by the appearance of headlights and grille located on the face of the automobile.[9] So we only select the images of the front face of the automobiles and ignore other views. Since our sales data comes at the make-model level, we choose the average of product characteristic across trims. We use the make-model fixed effect learned from the BLP demand model in addition to the structured product characteristics to construct supervisory signals for the disentanglement model.

We collected manufacturer suggested retail prices (MSRP), and characteristics of all automobiles sold in the UK from 2008-2017 from Parker's. We have product characteristics for weight, horsepower, length, width, and miles per gallon. The price variable is the list price (in £1000's) for the entry-level trim. Prices in all years are deflated to 2015 UK using the consumer price index. We supplemented the Parker's information with additional information, including vehicle country of production and company ownership information. We also supplemented additional information from the Office of National Statics, UK. We gathered the price of ultra low sulphur petrol per gallon and ultra low sulphur diesel per

[9] URL: https://www.wsj.com/articles/SB114195150869994250

gallon as well as the number of households in the UK. Similar to Berry et al. (1995), we calculated miles per UK pound (MP£) as miles per gallon divided by the price per gallon. We measure the market size as the number of households in the UK. We use 'HP/Weight', 'MP£', and 'Space' to construct BLP instruments.

In Table 6, we display summary statistics for the products at the make-model-year level. There are 2439 observations in our sample and a total of 379 distinct models. The variables include quantity (in units of 1000), price (in £000 units), the ratio of horsepower to weight (in HP per 10 lbs.), the number of ten mile increments one could drive for one £ of gasoline (MP£), tens of miles per gallon (MPG), and size (measured as length times width). We provide sales-weighted means for each variable. We see that automobiles have improved in terms of both power and fuel efficiency over these ten years.

Table 6 Descriptive Statistics of Structured Data

Market	No. of Observations	Quantity	Price	HP/Wt	Space	MPG	MP£
2008	233	6.158	21.398	0.416	1.245	4.578	0.760
2009	247	6.159	21.089	0.411	1.229	4.838	0.905
2010	243	6.655	21.584	0.414	1.247	5.022	0.837
2011	231	6.876	21.784	0.421	1.261	5.183	0.782
2012	244	7.028	21.533	0.422	1.264	5.422	0.825
2013	241	8.075	21.351	0.423	1.268	5.573	0.878
2014	251	8.608	21.697	0.431	1.277	5.702	0.962
2015	251	9.148	22.754	0.443	1.290	5.787	1.126
2016	253	9.225	24.067	0.457	1.305	5.692	1.149
2017	245	8.687	24.834	0.465	1.318	5.502	1.053
All	2439	7.685	22.352	0.433	1.274	5.391	0.948

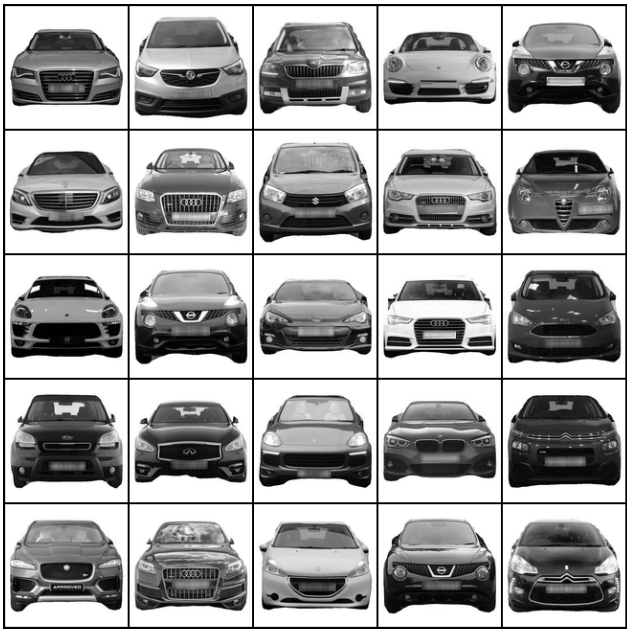
In Figure 3, we display images of 25 automobiles present in our dataset. Note that, we converted color images of size 128 × 128 to grayscale for our study (sales are also not available separately by color). Moreover, our goal is to extract visual characteristics that are related to the shape of the automobile and not related to the color. For each image, we have its associated make, model, year, structured product characteristics and price.

6. Results

6.1. Discovered Visual Characteristics

We learn the visual characteristics of each make-model sold in the UK between 2008 and 2017 using disentanglement representation learning. We compare the unsupervised approach to learn visual characteristics with supervised approaches. In the supervised approach, we train the learned visual characteristics to predict the supervisory signal associated with each make-model. We use the following supervisory signals:

Figure 3 Sample of Automobile Images



1. ’Price’ of the make-model
2. Combination of structured product characteristics of the make-model (‘HP/Weight’, ‘MPG’, ‘Space’)
3. Make-model fixed effects learned from the demand model[10]

For model training, we initially set model hyperparameters including the number of epochs=200, batch size=64, number of latent space dimensions=20, learning rate=0.0005, and the threshold value on KL loss term=0.10. We sweep over a grid of values to select the hyperparameters corresponding to the weight on the total correlation term λ_1, and the weight on the demand loss term λ_2 . Specifically, we vary λ_1 as $[1,5,10,20,30,40,50]$ and vary λ_2 as $[0,1,5,10,20,30,40,50]$. We follow the approach specified in JMR Paper to select the hyperparameters. For supervised approach, we select the hyperparameters λ_1 and λ_2 that lead to the lowest supervised loss on the validation dataset. Next, we use the UDR metric described in the methodology section to compare the supervised approaches with the unsupervised approach. Note that higher values of the UDR metric correspond to better disentanglement and discovery of independent visual characteristics. From Table 7, we find that the visual characteristics learned from supervising on ‘Make-Model Fixed Effects’ obtained from the demand model achieve the best disentanglement in terms of UDR.

[10] We provide the estimates of the demand model in Appendix ???.

Table 7 Comparison of Different Supervisory Approaches

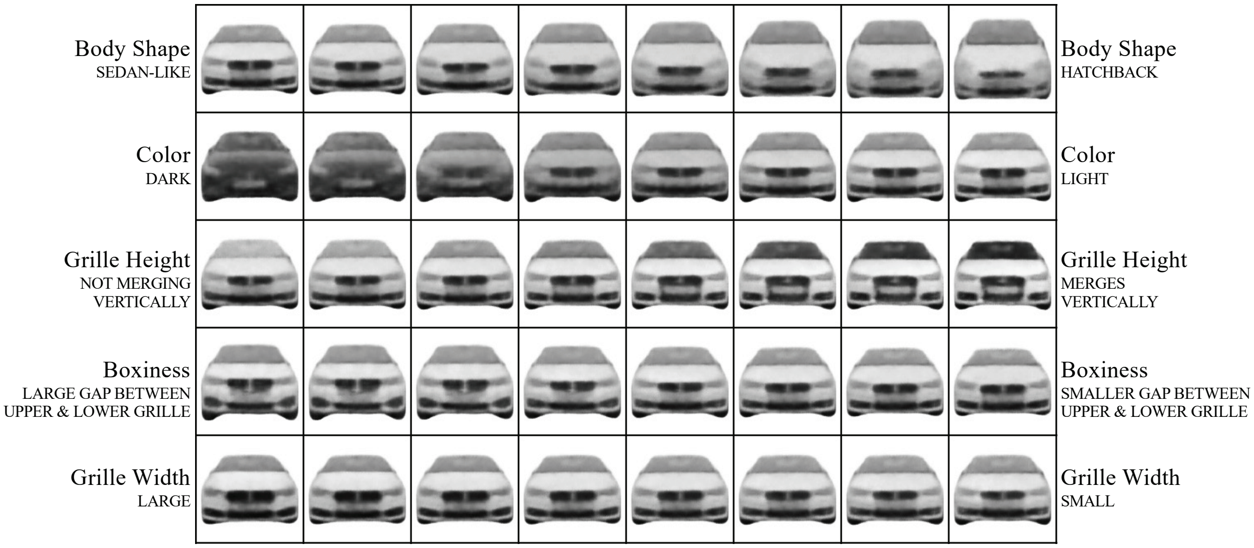
Number of Signals	Supervisory Signals	λ_1	λ_2	UDR
1	Make-Model FE	50	40	0.642
0	Unsupervised β-TCVAE	50	0	0.608
3	HP/Weight, MPG, Space	50	20	0.614
1	Price	50	50	0.587
1	Unsupervised VAE	1	0	0.071
1	Unsupervised AE	0	0	0.073

We show the discovered visual characteristics in Figure 4 corresponding to the model with $\lambda_1 = 50$, $\lambda_2 = 40$ and 'Make-Model Fixed Effects' supervisory signal. Each row in the image corresponds to a visual characteristics. In each row, we change the value of one visual characteristic while fixing the value of all the other characteristics. Note that since we use a generative deep learning based method, we are able to change the underlying learnt visual characteristics and generate counterfactual images. The ability to generate counterfactual images allows us to interpret each visual characteristics as it is able to isolate the effect of change only in one visual characteristic while keeping the other characteristics fixed. We find five visual characteristics of a automobile's front view to be informative. Rest of the visual characteristics were uninformative i.e. changing the visual characteristic produces no change in the image. We interpret these visual characteristics as written below.

1. Body Shape: Automobiles scoring low on this characteristic are sedan-like and those scoring high on this characteristic are hatchback-like.
2. Color: Automobiles scoring low on this characteristic are darker and vice-versa.
3. Grille Height: As the score of this visual characteristic increases, the top and bottom part of the grille begins to merge.
4. Boxiness: Automobiles scoring low on this characteristic have high degree of boxiness and vice-versa. define boxiness precisely here
5. Grille Width: Automobiles scoring low on this characteristic have a wider grille and vice-versa.

In Table 8, we show the correlation matrix between all the structured product characteristics and visual product characteristics. First, we find that visual product characteristics are uncorrelated with each other except a weak correlation between the boxiness and body shape. Second, we find that structured product characteristics are correlated with each other. Finally, we find that structured product characteristics and visual product characteristics are weakly correlated with each other.

Figure 4 Discovered Visual Characteristics



Left to Right: Vary one visual characteristic, keeping all others fixed

Table 8 Correlation Matrix

	Structured Characteristics				Visual Characteristics			
	Price	MPG	HP/Weight	Space	Boxiness	Body Shape	Grille Height	Grille Width
Price	1.00							
MPG	-0.60	1.00						
HP/Weight	0.74	-0.48	1.00					
Space	0.67	-0.47	0.36	1.00				
Boxiness	0.05	0.04	0.27	0.08	1.00			
Body Shape	-0.51	0.26	-0.53	-0.36	-0.07	1.00		
Grille Height	0.14	0.05	0.17	0.06	0.02	-0.08	1.00	
Grille Width	-0.11	0.09	-0.08	-0.17	-0.06	-0.07	0.00	1.00

6.2. Visual Market Structure Map

A market structure map is a graphical representation of the positioning of products that exist within a given market. It is a strategic tool often used by businesses to better understand the competitive landscape and to help inform marketing, product development, and overall business strategy. By mapping the market, companies can identify gaps in the market, potential opportunities for new products, and the relative positioning of their competitors.

We use the multidimensional scaling to map the market structure. Each product in this approach assumes that a product is a vector of characteristics in a high-dimensional space. It uses a 'dissimilarity matrix' – a collection of pairwise 'distances' or dissimilarities between the products — as an input. MDS transforms these pairwise dissimilarities into a lower-dimensional (often two-dimensional for ease of visualization) space, aiming to preserve the relative distances as much as possible. It means that different products appear farther apart on the map, while similar products appear close together. The goal of this approach is to reduce the complexity of the data while preserving as much of the original information about product similarities (or dissimilarities) as possible. The details of the MDS algorithm are in Appendix ???.

In MDS, the dimensions that result from the analysis do not have a predefined meaning like the original variables used in the data. Interpreting the dimensions in MDS can be somewhat subjective. We can use the coordinates of the items in the MDS space as the dependent variables, and the known characteristics (e.g., price and brand) as the independent variables. It can help quantify how well each characteristic explains the variation along each MDS dimension.

We use Multidimensional Scaling (MDS) to draw market structure maps based on only structured product characteristics, only visual characteristics and using a combination of structured as well as visual product characteristics. We use 'Price', 'HP/Weight', 'MPG' and 'Space' as structured product characteristics; 'Body Shape', 'Boxiness', 'Grille Height', and 'Grille Width' as visual product characteristics. We normalize each characteristic of a make-model by subtracting the mean value of the characteristic across make-models and then dividing it by the standard deviation of the characteristic across make-models before we implement MDS. Note that an alternative method to create a market structure to locate products in the visual characteristic space is to conduct a consumer survey about perceptions of visual similarity. The survey provides a dissimilarity matrix that can be used to construct market structure maps. The main disadvantage of the alternative approach is it is not possible to understand why the products are visually similar. This is because it uses the pairwise distance as a metric with no knowledge of the underlying human interpretable visual characteristics. Our method allows us to understand why two products are visually similar by looking at their underlying human interpretable visual characteristics. Moreover, the survey based approach is also not automatic because it requires a survey to cover all n^2 pairs. Further, the survey based approach suffers from a bias from human raters because they may focus on different aspects of design.

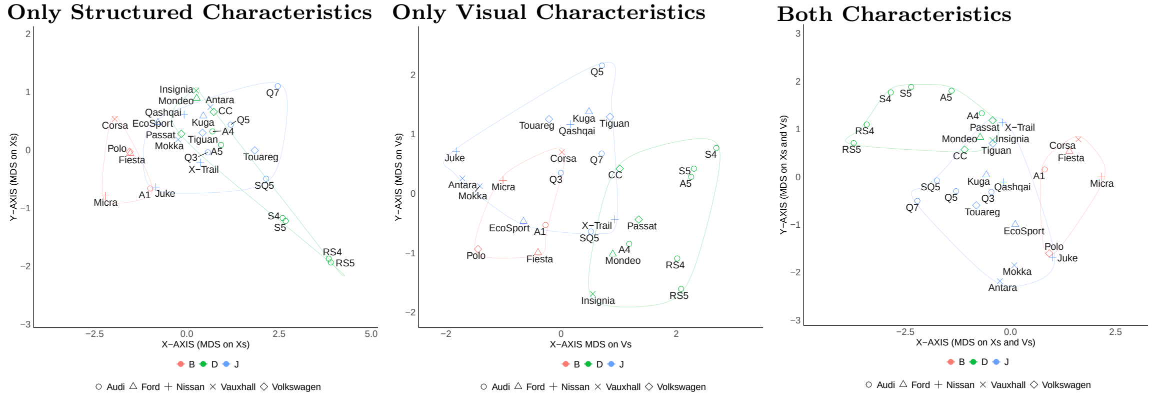
Figure 5 shows the market structure of automobiles belonging to the '2013' market. We select B, D and J segment because they have high market share (Segment A (Minicars): 9%, Segment B (Subcompact): 24%, Segment C (Compact): 23%, Segment D (Mid-Size):

9%, Segment E (Mid-Size Luxury): 3%, Segment J (SUV): 27%, Segment M (MPV): 4%) and high variation in characteristics across segments. We show models belonging to the top 5 makes by market share for B, D and J segment in the '2013' market in the maps. Table 9 shows the structured as well as visual product characteristics for the models belonging to Segment B, D and J of the top 5 makes.

Table 9 Product Characteristics

	Structured Characteristics				Visual Characteristics			
Make Model	**Price**	**MPG**	**HPWT**	**Space**	**Boxiness**	**Body Shape**	**Grille Height**	**Grille Width**
Segment B (Subcompact)								
Audi A1	19.33	5.98	0.49	1.17	-0.21	0.21	-0.16	-0.66
Ford Fiesta	16.18	6.37	0.41	1.22	0.95	1.47	-0.14	-0.34
Nissan Micra	12.73	5.72	0.39	0.98	-0.61	1.78	-0.21	0.12
Vauxhall Corsa	14.22	5.82	0.31	1.20	-0.53	1.27	1.31	0.54
Volkswagen Polo	15.29	5.53	0.37	1.17	0.17	-0.22	-3.67	-0.78
Segment D (Mid-Size)								
Audi A4	31.47	5.27	0.51	1.49	1.38	0.06	1.52	-0.20
Audi A5	34.44	5.10	0.54	1.47	0.86	-2.21	1.53	0.42
Audi RS4	56.57	2.60	1.12	1.49	1.78	-1.60	1.40	-0.45
Audi RS5	64.84	2.60	1.11	1.46	1.59	-2.56	1.03	-1.21
Audi S4	41.11	3.60	0.86	1.49	1.11	-3.11	0.99	0.99
Audi S5	44.84	3.48	0.85	1.46	1.19	-1.46	1.99	0.84
Ford Mondeo	23.90	5.34	0.44	1.56	1.06	-0.01	1.23	-0.55
Vauxhall Insignia	25.51	5.53	0.43	1.57	1.79	0.87	0.93	-0.73
Volkswagen CC	28.44	5.05	0.48	1.56	0.75	-1.56	-0.71	0.69
Volkswagen Passat	24.69	5.10	0.42	1.35	0.74	-0.75	1.53	-0.23
Segment J (SUV)								
Audi Q2	27.79	5.58	0.47	1.31	-2.21	-2.39	0.92	0.64
Audi Q3	31.88	5.16	0.47	1.37	-3.00	-2.52	1.85	-0.66
Audi Q5	40.71	4.78	0.61	1.55	-2.44	-1.83	1.33	0.36
Audi Q7	59.43	3.97	0.56	1.73	-2.65	-1.09	0.76	-0.33
Audi SQ5	49.44	3.40	0.85	1.55	-0.57	-0.75	1.19	-1.17
Audi SQ7	77.99	3.80	0.83	1.74	0.01	-0.33	0.30	-0.79
Ford EcoSport	17.08	5.32	0.39	1.28	-0.65	-0.17	-1.02	-0.86
Ford Edge	36.96	4.72	0.44	1.63	-1.34	-0.56	-0.90	-0.10
Ford Kuga	29.29	4.33	0.42	1.47	-1.41	0.13	-1.36	0.56
Nissan Juke	20.41	5.08	0.49	1.27	-1.81	-0.48	-2.40	0.54
Nissan Qashqai	24.59	5.77	0.39	1.41	-0.94	-0.67	-1.29	0.06
Nissan X-Trail	28.19	5.06	0.43	1.32	1.14	0.56	-0.82	-0.32
Vauxhall Mokka	23.94	5.15	0.45	1.35	-2.65	0.04	-0.64	-1.35
Volkswagen Tiguan	30.42	4.39	0.45	1.42	-0.48	-1.41	0.24	0.52
Volkswagen Touareg	44.87	4.20	0.50	1.64	-0.24	-1.95	-1.24	0.42

Figure 5 (Color Online) Segment B, D & J: Market Structure Map



Only Structured Characteristics

Only Visual Characteristics

Both Characteristics

6.2.1. Insights

Does visual and structured characteristics provide two independent dimensions of variation or does one restrict the other because they are strongly correlated? We operationalize this by calculating the average pairwise distance of each make-model to rival make-models in a particular segment in the structured space as well as the visual space. Next, we calculate the correlation between the average pairwise distances in the structured space and the visual space. Table 10 reports these correlations for the UK automobile market in 2013. A low correlation indicates similarity in structured space does not relate well to the visual similarity. Across all segments, we find that when products are close in structured space, visual space allows the competition to be relaxed.

Table 10 Correlation Between Distances in Structured Space & Distances in Visual Space

Segment	Correlation
A (Minicars)	-0.078
B (Subcompact)	-0.080
C (Compact)	0.145
D (Mid-size)	0.074
E (Mid-size Luxury)	0.154
J (SUV)	0.102
M (MPV)	0.031

Does differentiation across segments increase when visual information is included? From the market structure map using only structured product characteristics, we note that Segment B is clearly separated from both Segment D and J. However, Segment D and J overlap. From the market structure map using only visual product characteristics, we note that Segment B and J overlap. However, Segment D is separated from Segment B and J. Interestingly, when we account for both structured and visual product characteristics, all the

3 Segments separate out much more. This means that if one considers only type of characteristic, then the market appears more competitive. However, if one includes both the characteristics, then the market is less competitive. In other words, differentiation across product categories increases when visual information is included.

We verify this empirically by calculating the ratio of non-overlapping area to the total area bounded by the vertices in each of the three market structure maps. We find that in a map using only structured characteristics, 15.1% of the area is overlapping. Interestingly, when we create a market structure map using both structured and visual characteristics, then only 1.5% of the area is overlapping. Note that 39.4% of the area is overlapping in the market structure map created using only visual characteristics.[11]

Is the competition in the visual space a strategic substitute or complement to the competition in the structured space? We study this by finding the centroid for each make within a particular segment in a particular year. We refer to them as own-make centroid. Next, we calculate the distance to all other rival-make centroids in that segment. Finally, we find the closest make for each make in a particular segment. We do this for both structured space as well as visual space. Table 11 shows the closest makes in the structured space as well as visual space for the top 10 selling makes in the J segment in the '2013' market.

Table 11 Closest Within-Segment Rivals in Structured Space & Visual Space

Make	Quantity Sold	Closest Rival in Structured Space	Closest Rival in Visual Space
Segment B (Subcompact)			
Ford	113390	Volkswagen	Peugeot
Vauxhall	76413	Dacia	Fiat
Volkswagen	39453	Ford	Mitsubishi
Peugeot	37896	Volkswagen	Ford
MINI	31062	Suzuki	DS
Segment D (Mid-Size)			
BMW	39612	Audi	Volkswagen
Mercedes-Benz	31068	Infiniti	BMW
Audi	29955	BMW	Lexus
Vauxhall	24382	Ford	Suzuki
Volkswagen	17964	SKODA	BMW
Segment J (SUV)			
Nissan	82431	Kia	Jeep
Kia	22480	Nissan	Hyundai
Audi	21888	BMW	BMW
Vauxhall	20461	Chevrolet	Ssangyong
BMW	19336	Audi	Audi

[11] Please note that the area calculations are an approximation of those in Figure 5. We construct a convex hull around the coordinates and then calculate the area

Within Make-Segment Product Differentiation: Role of Visual Space We operationalize this by comparing the area share of each make in a particular segment in the market structure maps created using only structured space with the ones using only visual space. We consider the top 4 most selling models of a make in a segment. We show the area share of makes belonging to the Segment J in the '2013' market. On the one hand, we can see that Audi has models close together in the visual space but far apart in the structured space. On the other hand, we can see that Nissan has models close together in the structured space but far apart in the visual space.

Table 12 Area Share of a Make in Structured Space & Visual Space

Make	Quantity Sold	Models	Area Share (Structured)	Area Share (Visual)
Segment J (SUV)				
Audi	21888	4	17.90%	9.43%
BMW	19336	4	6.35%	5.51%
Jeep	1842	4	9.38%	21.92%
Kia	22480	3	3.87%	4.84%
Mitsubishi	5375	3	0.84%	1.23%
Nissan	82431	4	7.41%	17.33%

7. Discussion & Conclusion

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.

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.

Bergen M, Peteraf MA (2002) Competitor identification and competitor analysis: a broad-based managerial approach. *Managerial and decision economics* 23(4-5):157–169.

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

Berry S, Levinsohn J, Pakes A (1999) Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review* 89(3):400–430.

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

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

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 (2022) Product aesthetic design: A machine learning augmentation. *Marketing Science* .

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

Creusen ME, Schoormans JP (2005) The different roles of product appearance in consumer choice. *Journal of product innovation management* 22(1):63–81.

DeSarbo WS, Grewal R, Wind J (2006) Who competes with whom? a demand-based perspective for identifying and representing asymmetric competition. *Strategic Management Journal* 27(2):101–129.

DeSarbo WS, Manrai AK, Manrai LA (1993) Non-spatial tree models for the assessment of competitive market structure: an integrated review of the marketing and psychometric literature. *Handbooks in operations research and management science* 5:193–257.

Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science* 41(2):401–425.

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.*

Erdem T (1996) A dynamic analysis of market structure based on panel data. *Marketing science* 15(4):359–378.

Ferraro R, Kirmani A, Matherly T (2013) Look at me! look at me! conspicuous brand usage, self-brand connection, and dilution. *Journal of Marketing Research* 50(4):477–488.

Fleming RW (2014) Visual perception of materials and their properties. *Vision research* 94:62–75.

Gabel S, Guhl D, Klapper D (2019) P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research* 56(4):557–580.

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.

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.*

Homburg C, Schwemmle M, Kuehnl C (2015) New product design: Concept, measurement, and consequences. *Journal of marketing* 79(3):41–56.

Huang J, Chen B, Luo L, Yue S, Ounis I (2021) Dvm-car: A large-scale automotive dataset for visual marketing research and applications. *arXiv preprint arXiv:2109.00881* .

Jindal RP, Sarangee KR, Echambadi R, Lee S (2016) Designed to succeed: Dimensions of product design and their impact on market share. *Journal of Marketing* 80(4):72–89.

Kim H, Mnih A (2018) Disentangling by factorising. *ICML*, 2649–2658.

Kim JB, Albuquerque P, Bronnenberg BJ (2011) Mapping online consumer search. *Journal of Marketing research* 48(1):13–27.

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

Lattin JM, Carroll JD, Green PE (2003) *Analyzing multivariate data*, volume 1 (Thomson Brooks/Cole Pacific Grove, CA).

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

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5):881–894.

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

Liu Y, Li KJ, Chen H, Balachander S (2017) The effects of products' aesthetic design on demand and marketing-mix effectiveness: The role of segment prototypicality and brand consistency. *Journal of Marketing* 81(1):83–102.

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

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

Megehee CM, Spake DF (2012) Consumer enactments of archetypes using luxury brands. *Journal of business research* 65(10):1434–1442.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3):521–543.

Norman D (2013) *The design of everyday things: Revised and expanded edition* (Basic books).

Rao VR, Sabavala DJ, et al. (1986) Measurement and use of market response functions for allocating marketing resources. *(No Title)* .

Ringel DM, Skiera B (2016) Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science* 35(3):511–534.

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

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of marketing research* 51(4):463–479.

Urban GL, Johnson PL, Hauser JR (1984) Testing competitive market structures. *Marketing Science* 3(2):83–112.

Veryzer Jr RW (1993) Aesthetic response and the influence of design principles on product preferences. *Advances in Consumer research* 20(1).

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

Yang Y, Zhang K, Kannan P (2022) Identifying market structure: A deep network representation learning of social engagement. *Journal of Marketing* 86(4):37–56.

Zhang M, Luo L (2022) Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Management Science* .

Zhang S, Lee D, Singh PV, Srinivasan K (2022) What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science* .

Electronic Companion Supplement

Appendix A: Demand Model Estimate

In this section, we estimate a demand model to understand the consumer preferences on structured product characteristics as well as obtain the fixed effects for each make-model. The make-model fixed effects capture the time-invariant consumer preferences on each product's unobservable characteristics such as quality and visual style. To make our model simple, we only introduce heterogeneity on two terms: price and constant. We cluster the standard errors at the make-model level. From Table EC.1, we can see that all the estimates are precise by looking at the standard errors of our estimates. Our estimates are in line with economic intuition. We find that on average consumers prefer cars with more power, higher fuel efficiency as well as larger space. We also find that consumers with higher income have lower price sensitivity.

Table EC.1 Parameter Estimates of Model of Market Equilibrium

	Variable	Parameter Estimate	Standard Errors
Means ($\overline{\beta}'s$)			
	HP/Weight	6.00	(1.30)
	MP£	2.50	(0.26)
	Space	3.00	(0.82)
Standard Deviation ($\sigma'_{\beta}s$)			
	Constant	4.60	(0.94)
Term on Price			
	(-p/y)	-17.00	(2.20)
Supply-Side Terms			
	Constant	3.60	(0.08)
	ln(HP/Weight)	0.74	(0.06)
	ln(MPG)	-0.24	(0.07)
	ln(Space)	1.60	(0.08)
	Trend	0.005	(0.003)

Should we also write other BLP specifications, get different estimates for make-model fixed effects and then do everything again

Appendix B: Other Posterior Traversals

Posterior Traversals for AE/VAE/Unsupervised

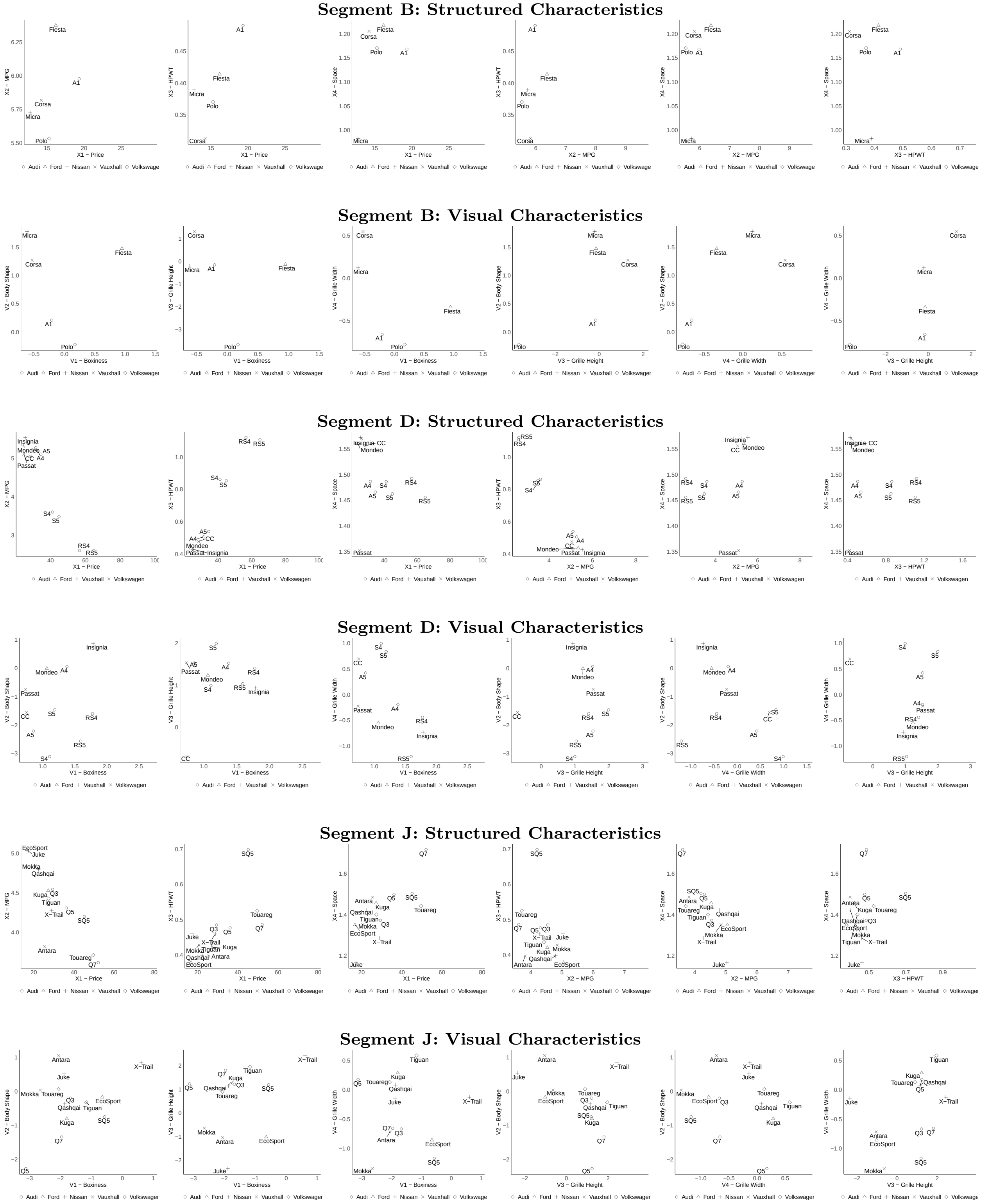
Appendix C: Market Structure Maps with Other Segments

Placeholder for putting market structure maps for other segments

Appendix D: Maps Two Char By a Time

Finally, we aim to understand why make-models are located close together or further apart in market structure maps. The use of interpretable product characteristics (both structured and visual) allows us to do so. A survey based approach to understand visual similarity between a pair of make-models would not allow us to understand the reason for why make-modes are located close together or further apart in market structure maps. Figure EC.1 plots 6 characteristic by characteristic for both structured as well as visual product characteristics for Segment B, D and J.

Figure EC.1 Segment-Wise Characteristic Map



Segment B: Structured Characteristics

Segment B: Visual Characteristics

Segment D: Structured Characteristics

Segment D: Visual Characteristics

Segment J: Structured Characteristics

Segment J: Visual Characteristics