Design and Evaluation of Product Aesthetics:
A Human-Machine Hybrid Approach

by

Alex Burnap

John R. Hauser

and

Artem Timoshenko

July 2019

Alex Burnap is a Postdoctoral Fellow in Marketing, MIT Management School, Massachusetts Institute of Technology, E62-366, 77 Massachusetts Avenue, Cambridge, MA 02139, (405) 880-3660, aburnap@mit.edu.

John R. Hauser is the Kirin Professor of Marketing, MIT Management School, Massachusetts Institute of Technology, E62-538, 77 Massachusetts Avenue, Cambridge, MA 02139, (617) 253-2929, hauser@mit.edu.

Artem Timoshenko is an Assistant Professor of Marketing at Kellogg School of Management, Northwestern University, 2211 Campus Drive, Suite 5391, Evanston, IL 60208, (617) 803-5630, artem.timoshenko@kellogg.northwestern.edu.

Design and Evaluation of Product Aesthetics:
A Human-Machine Hybrid Approach

Abstract

Aesthetics are critically important to market acceptance in many product categories. In the automotive industry in particular, an improved aesthetic design can boost sales by 30% or more. Firms invest heavily in designing and testing new product aesthetics. A single automotive "theme clinic" costs between $100,000 and $1,000,000, and hundreds are conducted annually. We use machine learning to augment human judgment when designing and testing new product aesthetics. The model combines a probabilistic variational autoencoder (VAE) and adversarial components from generative adversarial networks (GAN), along with modeling assumptions that address managerial requirements for firm adoption. We train our model with data from an automotive partner—7,000 images evaluated by targeted consumers and 180,000 high-quality unrated images. Our model predicts well the appeal of new aesthetic designs—38% improvement relative to a baseline and substantial improvement over both conventional machine learning models and pretrained deep learning models. New automotive designs are generated in a controllable manner for the design team to consider, which we also empirically verify are appealing to consumers. These results, combining human and machine inputs for practical managerial usage, suggest that machine learning offers significant opportunity to augment aesthetic design.

Keywords: *Aesthetics, Generative Adversarial Networks, Generating New Products*, *Machine Learning, Prelaunch Forecasting, Product Development*, *Variational Autoencoders.*

1. Introduction

Consumers consistently rank aesthetics among the three most important attributes in product choice (Bloch 1995; Creusen and Schoormans 2005). For example, the visual design of the original iPod was judged to be a critical factor in its market acceptance (Reppel, Szmigin, and Gruber 2006). In categories such as home appliances, aesthetics help firms establish product differentiation beyond functional attributes (Bloch 1995; Crilly, Moultrie, and Clarkson 2004; Person et al. 2007); for instance, the Dyson DC01 used transparent design to communicate its complexity to consumers, helping it become the best-selling vacuum in the U.K. (Noble and Kumar 2010). Firms use aesthetics to strategically position and enhance brand recognition (Aaker and Keller 1990; Karjalainen and Snelders 2010; Keller 2003); trade dress violations (non-functional attributes that signal brand identity) are hotly contested in Lanham Act (§43A) litigation. Aesthetics pervade marketing—visually-appealing products and packaging drive consumers to choose one product over another, especially at the point of purchase in crowded brick-and-mortar stores, supermarkets, and online retailers (Clement 2007; Orth and Malkewitz 2008).

Developing product aesthetics can require substantial investment, yet returns on investment are found across markets—a study of 93 firms across nine product categories found that firms that heavily invested into aesthetic design had 32% higher earnings than industry averages (Hertenstein, Platt, and Veryzer 2005). Marketing and product managers routinely manage the aesthetic design of products, services, and digital marketplaces. In this paper, we propose a methodology to improve the process of aesthetic product design and testing. While the basic concepts of the proposed methodology are applicable across markets, we demonstrate the approach in the automotive industry.

In the automotive industry, product aesthetics can explain up to 60% of uncertainty in purchase decision for certain segments (Kreuzbauer and Malter 2005). Automotive design significantly affects its market performance (Cho, Hasija, and Sosa 2015; Jindal et al. 2016; Rubera 2015), in large part by influencing consumer consideration (Liu et al. 2017; Palazzolo and Feinberg 2015). For example, the redesign of the Buick Enclave in 2008 commanded a 30% increase in MSRP over the Buick Rendezvous it replaced, and the redesign of the 2005 Volkswagen Beetle resulted in a 54% market share gain in just a single year (Kreuzbauer and Malter 2005; Blonigen, Knittel, and Soderbery 2013). On the other hand, the visual appeal of the 2001 Pontiac Aztec was cited as a reason for its lack of market success (Vlasic 2011). Not surprisingly, automotive firms invest heavily in design and redesign—$1.25 billion on average with up to $5.7 billion invested in critical designs (Blonigen, Knittel, and Soderbery 2013; Pauwels et al. 2004; Rubera 2015).

1

Traditionally, human judgment enters the aesthetic product design pipeline in at least two ways. While there are established aesthetic heuristics and cognitive design principles (Coates 2003; Crilly, Moultrie, and Clarkson 2004; Norman 2004), aesthetic design is often generated and screened by design teams who have an "eye" for visual design. Design teams are powerful within organizations; their aesthetic judgments are hard to overrule (Vlasic 2011).

Human judgment also affects aesthetics through consumer evaluations. Firms often ask consumers to evaluate alternative designs in laboratory test markets, A/B tests, or "theme clinics." In a typical automotive theme clinic, a few hundred targeted consumers are recruited and brought to a central location to evaluate aesthetic designs. Consumers view the aesthetic designs and rate them on established benchmarks such as semantic scales for sporty, appealing, innovative, and luxurious (Coates 2003; Manoogian II 2013). Theme clinics are costly. Automotive firms typically invest between $100,000 and $1,000,000 per theme clinic for a single new vehicle design. With multiple aesthetic designs per vehicle and over a hundred vehicles in its worldwide product line, General Motors alone spends tens of millions of dollars on theme clinics. With the additional hidden costs incurred when designers routinely and manually screen hundreds of aesthetic designs to narrow down alternatives for a theme clinic, costs can exceed $100 million for a single manufacturer.

In this paper, we augment human judgment with machine learning tools to address both aspects of aesthetic design: (1) the generation and (2) the testing of new aesthetic designs. For testing, we predict how consumers would rate product aesthetics directly from visual images. Specifically, we use an *encoding model* to represent visual designs (images) using a smaller number of 500 "features," and train a *predictive model* that predicts aesthetic ratings based on those features. We use the predictive model to screen newly proposed aesthetic designs so that only the highest-potential designs are tested in theme clinics (or their equivalent for non-automotive applications).

For generation, we create new designs to inspire the design team. We develop a *generative model* that creates new product designs that are predicted to be aesthetically appealing while satisfying product attributes desired by the designers (e.g., "Cadillac-like"). This gives the design team a controllable tool which they may use to morph through the design space. While some images are nonsensical, other generated images inspire creative aesthetic design by bending product forms.

Our focus is *augmentation* of human expertise and creativity, rather than its *automation*. Our focus addresses several key managerial issues related to integrating machine intelligence into existing human workflows. This combination of machine learning and human judgment (machine-human hybrid) leads to better designs and a reduction in the costs of design generation and testing.
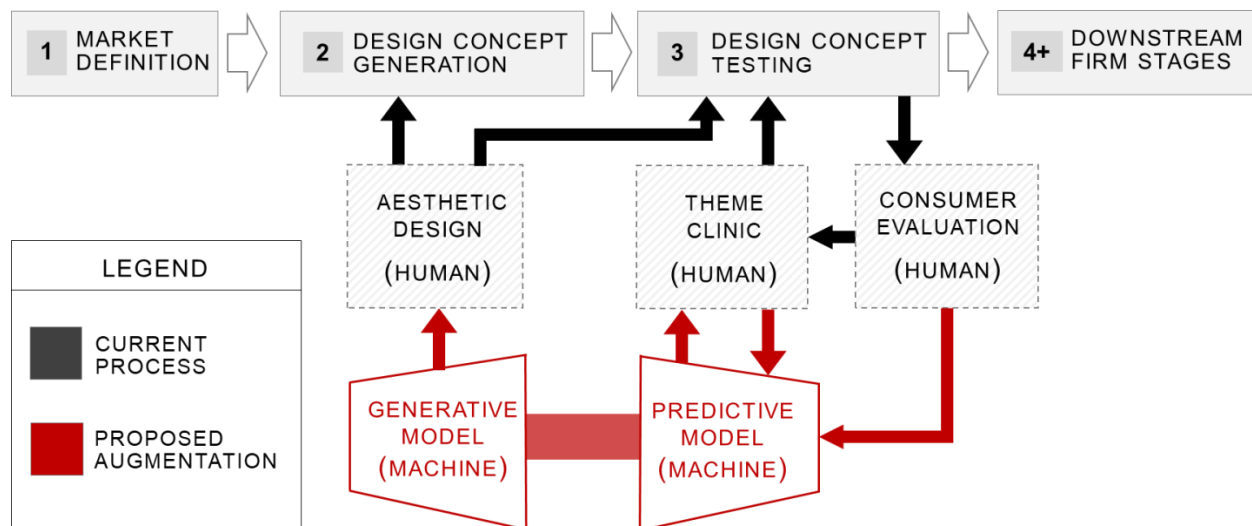
2. Conceptual Model of Aesthetic Product Design

2.1. Augmenting the Design Process

Figure 1 summarizes the current design process and the proposed role of machine learning augmentation. Consider first the current process as shown in the first two rows (coded as black). The process begins with a market definition that is external to the aesthetic design efforts. For example, Apple targeted smartphones with a touchscreen; Zenni Optical aimed to develop prescription sports glasses; IKEA targeted affordable yet aesthetically pleasing furniture. In automotive markets, firms target particular segments such as luxury compact utility vehicles (currently the Cadillac XT5, Buick Envision, Volvo XC60, BMW X3, and others). Market definition provides soft constraints on aesthetics based on the targeted consumers and the firm's capabilities (Box 1).

Aesthetic designers then create hundreds to thousands of freehand sketches that are converted to 2D images (Box 2). For example, Dyson and General Motors may generate several hundred sketches per new product model, while IKEA may generate fewer given its product line variety and turnover (Bouchard, Aoussat, and Duchamp 2006; Toffoletto 2013). The human design team next screens potential designs to a smaller set of testable design concepts in a process known as "downselecting" (Box 3). Consumers evaluate the remaining designs in theme clinics resulting in more screening. Successful designs are advanced downstream for further development, including engineering, manufacturing, and marketing communications (advertising, social media, websites, etc.). The process is highly iterative and asynchronous across both generation and testing; our augmentation applies to all iterations.

Figure 1. Augmenting Aesthetic Design with Machine Learning

3

The red trapezoids and arrows highlight the machine learning augmentation. After the machine learning models are trained, we use them to augment both design concept testing and generation. In testing, we evaluate design concept images using the *predictive model* to help eliminate designs likely to score low in theme clinics. This enables faster iteration as the design team focuses on the remaining high-potential images.

By focusing on high-potential images, theme clinics are more efficient and effective. The resulting aesthetic designs are likely more profitable when introduced to the market. The firm benefits from shorter product development times by more rapidly sending final designs to downstream firm stages. Cost reductions result from the corresponding lower product "drop rate" (i.e., the share of design concepts later terminated in downstream firm stages) (Cooper 1990; Danneels and Kleinschmidtb 2001). As a final added benefit, rigorous quantitative evaluation of aesthetics helps "shield" designs from downstream changes driven by engineering, manufacturing, or accounting (Hartley 1996; Manoogian II 2013; Vlasic 2011).

The *generative model* creates many designs that are realistic and aesthetically appealing, and can stretch designs in ways that spark creativity among designers (Coates 2003; Martindale 1990). The designers can control the generative model through specifying designer attributes. Example attributes are 'red,' 'Cadillac-like,' 'Sport Utility Vehicle (SUV),' '2015 vintage,' or 'viewed from the side.' (The vintage variable is important when training the model. It captures the changing nature of fashion in the automotive industry (Hekkert, Snelders, and Wieringen 2003; Martindale 1990).

This intended use is motivated by our experience with real design teams—human creativity is augmentable but currently nowhere near automatable. Designers must innovate product designs while anticipating changes in consumer demand; a challenging task given the often significant time it takes to develop a new product, e.g., Adidas apparel 12-18 months, GM vehicles 36+ months (*Adidas AG 2017 Annual Report* 2017; Manoogian II 2013). Our focus on augmentation guides the modeling decisions in our approach; the final aesthetic decisions remain with the design team.

2.2. Technical Challenges in Augmenting Aesthetic Design

Both the predictive model and the generative model require that we address critical challenges in working with product aesthetics images and with the traditional aesthetic design pipeline. First, images are inherently high dimensional. Even modest quality images are 100 x 100 pixels for each of red, green, and blue colors together comprising 30,000 variables—far too many to be input to conventional choice models. Previous work has instead primarily represented aesthetics in choice models using characteristic lines (Chan, Mihm, and Sosa 2018; Ranscombe et al. 2012), landmark points (Landwehr,

4

Labroo, and Herrmann 2011), silhouettes (Orsborn, Cagan, and Boatwright 2009; Reid, Gonzalez, and Papalambros 2010), and Bezier curves (Kang et al. 2016); or aggregated numbers such as J.D. Power APEAL or online reviews (Cho, Hasija, and Sosa 2015; Homburg, Schwemmle, and Kuehnl 2015; Jindal et al. 2016; Pauwels et al. 2004). We focus on working with images, the industry standard, due to their realism in representing the design. Design concept realism greatly affects consumer evaluation and enhances accuracy and precision in attribute-based product evaluation and forecasting (Dotson et al. 2016; Reid, MacDonald, and Du 2013; Hauser, Eggers, and Selove 2019). Choice modeling with images has only recently been enabled by deep learning, including relevant work in aesthetic prediction (Pan et al. 2017; Vasileva et al. 2018) and aesthetic generation (Kang et al. 2017; Sbai et al. 2019). Our work fundamentally differs from these works as we tackle key managerial issues required for adoption within companies; specifically, limited data sizes, and modeling decisions that augment rather than automate existing human workflows within firms.

Second, gathering consumer evaluations is costly and this results in limited training data. In our application, we are fortunate to have 7,000 aesthetic ratings by consumers for vehicles, but even that would be insufficient to estimate a predictive model with 30,000 variables. To address limitations on the size of the training sample, we need to either (a) reduce the number of explanatory variables using dimensionality reduction or (b) increase the amount of training data to make the predictive model feasible. We do both by training an *encoding model* to embed images in a low-dimensional space. Embeddings reduce the dimensionality, and allow us to combine the relatively small (and expensive for the firm) sample of labelled training data (images with aesthetic ratings) with a much larger sample of unlabeled training data (images without ratings). Success depends upon whether the dimensionality reduction preserves the important information from the full images while still allowing us to predict human aesthetic judgment and generate perceptually realistic new designs.

Embeddings using a neural network have seen recent adoption in marketing science; for example, Timoshenko and Hauser (2019) embed textual data to identify consumer needs; Sun, Ghose, and Liu (2018) embed consumer purchasing behavior; Liu, Lee, and Srinivasan (2018) embed product reviews to predict sales conversion; Zhang et al. (2017) embed Airbnb photos to predict demand; Liu, Dzyabura, and Mizik (2017) embed social media images to predict identity; Chakaborthy, Kim and Sudhir (2019) embed reviews to identify sentiment and missing evaluations. In our case, good embeddings are those which reduce dimensionality to low-dimensional "features" while still allowing us to rate product aesthetics and generate realistic new designs. (A "feature" in our context is an abstract representation inherent to embeddings. It is <u>not</u> an attribute such as "round headlight" or "an 8° slope.")

5

Third, aesthetic evaluations are holistic (Berlyne 1971; Bloch 1995; Crilly, Moultrie, and Clarkson 2004; Martindale 1990). Aesthetic images cannot be decomposed as is done in marketing-science methods such as conjoint analysis (Orme and Chrzan 2017). Design aspects are interdependent; we cannot expect consumers to evaluate design aspects separately. For example, even for pre-defined aesthetic aspects, a consumer cannot evaluate the aesthetics of a new BMW X3 design as an additive sum of the shape and position of headlights, the slope of the hood, and the height of the beltline. Rather, the Gestalt interplay of all design elements, including subtle elements such as the "Hoffmeister kink," drive consumer evaluations of qualitative attributes such as appealing, sporty, aggressive, luxurious, or modern (Coates 2003). The same interdependence is true for embedded features, thus we use deep neural networks for the encoder, predictive, and generative models to capture complex non-linear interactions among the reduced embedded "features."

Fourth, the design process is highly iterative and asynchronous. Multiple design teams (and sub-teams) in both design concept testing and generation must be able to use the machine learning augmentation for their corresponding roles. For example, the design team may split into sub-teams to work on several promising design concepts in parallel, while the current theme clinics may be testing entirely different design concepts. To enhance parallel development, we require an ability to separate the predictive and generative models into standalone components. At the same time, we require the ability to combine these models for usage and training.

3. Overview of a Machine Learning Approach to Augment Aesthetic Product Design

Let X_i be the (height x width x color) matrix of pixels of image i. The proposed approach requires two inputs: product images labeled with aesthetic ratings evaluated by consumers, $\{(X_i, y_i) | i = 1..N\}$, and unlabeled product images without ratings but with product attributes, $\{(X_j, \vec{a}_j) | j = 1..M\}$. For example, y_i might be the average consumer evaluation of the aesthetic appeal of product design i, and $\vec{a}_j$ may be attributes (if known) such as color and brand. In a typical application, the firm would only have consumer evaluations for a small fraction of images, $N \ll M$. Our two high-level goals are to test new product aesthetics by predicting consumer ratings, $\hat{y}_{new}$, for new product images created by the design team, X_{new}, as well as generate new product designs, $\hat{X}_{gen}$, according to attributes desired by the design team, $\vec{a}_{gen}$, and with predicted aesthetic ratings, $\hat{y}_{gen}$, that indicate consumer acceptance.

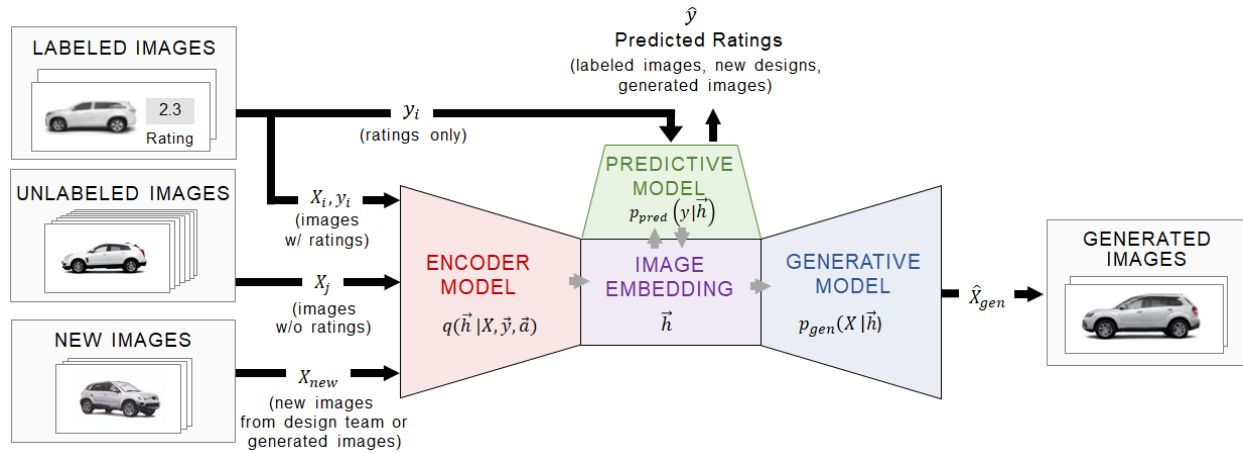**Figure 2.** Proposed Machine Learning Augmentation Approach

Figure 2 summarizes the general input and output flow of the proposed machine learning augmentation for aesthetic design. For every product design i, the encoding model in Figure 2 (red trapezoid on the left) uses a function, $q\left(\vec{h}_i \mid X_i, y_i, \vec{a}_i, \vec{\beta}\right)$ to embed the 196,608-dimensional image, X_i, its aesthetic rating (if known), y_i, and its attributes (if known), $\vec{a}_i$, to a 512-dimensional embedding, $\vec{h}_i$. Note the 196,608-dimensional images are due to being 256 x 256 x 3 (height x width x color), while the 512-dimensional embedding size is chosen by us. The embedding (purple square in the middle of Figure 2) makes it tractable to use the predictive model, $p_{pred}(\hat{y}_i \mid \vec{h}_i, \vec{\beta})$, (green trapezoid near the top of Figure 2) to predict ratings for new images and for generated images. The embedding also makes it tractable to use the generative model, $p_{gen}(X_i \mid \vec{h}_i, \vec{\beta})$, (blue trapezoid on the right of Figure 2) to create new high-potential images, X_{gen}. When generating images, we sample $\vec{h}$ such that $p_{pred}(\hat{y} \mid \vec{h}, \vec{\beta})$ is large, and produce either a set of images or a video displaying images as we change $\vec{h}$ smoothly.

The three models in the proposed approach—the generative model, encoding model, and predictive model—are formulated as probability distributions which are connected by the embedding. We learn an embedding *distribution* for each product design rather than a point estimate. Each product design i has its own embedding distribution $\vec{h}_i \sim q\left(\vec{h}_i \mid X_i, y_i, \vec{a}_i, \vec{\beta}\right)$. Learning the parameters of the distribution leverages the variational inference literature (Blei, Kucukelbir, and McAuliffe 2017; Jordan et al. 1999), thereby enabling Bayesian parameter estimation at data sizes otherwise intractable for MCMC sampling. Variational approaches have seen recent adoption in marketing; for example, Braun and McAuliffe (2010) derive variational bounds for the mixed multinomial logit model; Ishigaki, Terui, Sato, and Allenby (2018) estimate a latent topic choice model for sparse transaction data; Ansari, Li, and Zhang (2018) estimate a topic model for a recommender system. In our case, a variational Bayesian

7

approach improves predictive performance and allows better interpolation by the generative model.

We use deep neural networks (augmented through adversarial training) to parametrize all three models. The loss functions for the deep neural networks are based on maximizing the data log likelihood—minimizing loss is the same as choosing parameters that lead to distributions that have high data likelihood. We develop the deep neural network model by first using conditional probability to separate the log-likelihood function into components that are based on prediction, generation, and encoding (§4). We then approximate the true log-likelihood with a lower bound based on the observed data (evidence lower bound). Finally, we assume priors from a parametric family of probability distributions and choose parameters to maximize the approximate log-likelihood based on the observed images and ratings. The overall combined model is maximized jointly and is based on those images that have ratings and those that do not have ratings. For the remainder of paper, when it is clear in context, we refer to the predictive model, generative model, and encoding model as the predictor, generator, and encoder, respectively.

4. Proposed Approach: Variational Autoencoder and Generative Adversarial Network

4.1. Embeddings Are Chosen to Approximate the Joint Distribution of Images and Ratings

Our goals are to augment design concept testing by predicting aesthetic ratings, $\hat{y}_{new}$, for new design images, X_{new}, as well as augment design concept generation by creating new designs, X_{gen}, given attributes, $\vec{a}_{gen}$, desired by the design team.

Let $\vec{\beta}$ be a vector of model parameters. To ease notation, we temporarily write all parameters and likelihoods for a single datum i. We seek a joint distribution, $p(y_i, X_i | \vec{a}_i, \vec{\beta})$, for the ratings and images conditioned on the design attributes and the parameters. The joint distribution can be decomposed into a predictive model and a generative model by the laws of conditional probability: $p(y_i, X_i | \vec{a}_i, \vec{\beta}) = p_{pred}(y_i | X_i, \vec{a}_i, \vec{\beta}) p_{gen}(X_i | \vec{a}_i, \vec{\beta})$. Unfortunately, representing and estimating the two conditional distributions is not feasible when product images are high dimensional.

To address high-dimensionality, we approximate the true joint distribution using *embeddings*. In our case, the embedding compresses information from high-dimensional images, aesthetic ratings, and product attributes to enable tractable predictive and generative models. We choose embeddings such that the embeddings-based likelihood closely approximates the true but intractable likelihood, $p(y_i, X_i | \vec{a}_i, \vec{\beta})$.

We estimate an embedding posterior distribution, $q_{enc}(\vec{h}_i | X_i, y_i, \vec{a}_i, \vec{\beta})$, for each product design i such that $\vec{h}_i$ has substantially fewer dimensions (e.g., 512) than X_i (e.g., 196,608), yet retains most of

8

the information contained in the images, ratings, and product attributes. To infer embeddings, we use variational Bayes methods to approximate the true joint log-likelihood, $\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$, with an approximate log-likelihood, $\ell^i_{approx}(\vec{\beta})$, dependent on the embeddings, $\vec{h}_i$ (Blei, Kucukelbir, and McAuliffe 2017; Jordan et al. 1999). Note that while this technique could introduce new data-dependent parameters, $\vec{\beta}_i$, we implicitly include them in the set of "global" parameters, $\vec{\beta}$, as they will later be efficiently learned with a deep learning model (amortized inference, Hoffman et al. 2012).

To obtain this approximation, we first condition the true log likelihood, $\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$, on the embeddings via marginalization, which leads to the logarithm of an expectation. We then approximate the logarithm of the expectation by the expectation of the logarithm. By Jensen's Inequality, the approximation is a lower bound to the true log likelihood. Rearranging terms we arrive at the approximate likelihood in Equation 1. Given an image, X_i, its rating, y_i, and its attributes, $\vec{a}_i$, we seek to maximize $\ell^i_{approx}(\vec{\beta})$ and thus approximately maximize $\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$. Appendix A2 provides a more-detailed derivation that combines standard variational Bayes theory with predictive models. If $D_{KL}(\cdot || \cdot)$ signifies the Kullback-Leibler (KL) divergence, then:

(1)
$$
\begin{aligned}
\ell^i_{approx}(\vec{\beta}) &= \mathrm{E}_{\vec{h}_i}\big[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) + \log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) - \log q_{enc}(\vec{h}_i | X_i, y_i, \vec{a}_i, \vec{\beta}) \\
&\quad + \log p_{prior}(\vec{h}_i | \vec{a}_i)\big] \\
&= E_{\vec{h}_i}\big[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) + \log p_{gen}(X_i | \vec{h}_i, \vec{\beta})\big] \\
&\quad - D_{KL}\left(q_{enc}(\vec{h}_i | X_i, y_i, \vec{a}_i, \vec{\beta}) || p_{prior}(\vec{h}_i | \vec{a}_i)\right)
\end{aligned}
$$

Equation 1 is intuitive. The first term seeks to maximize our ability to predict aesthetic ratings based on the embeddings. The second term seeks to reproduce images based on the embeddings. The last term is the negative of KL divergence from the embedding (distribution) to the prior on the embeddings. The KL term acts to regularize the embedding such that the distributions of encoded images do not stray too far from the prior. The KL divergence prevents the encoder from "cheating" by giving each image a unique sub-region of the embedding space, effectively memorizing training data and hurting generalization performance. Note that for this derivation we have assumed conditional independence between images, ratings, and attributes given the embedding. This allows our model to use these data separately, as guided by the challenges discussed in §2.2 with regards to labelled data and the multiple stakeholders involved in current design workflows.

Because the approximate log-likelihood, $\ell^i_{approx}(\vec{\beta})$, is a lower bound to the true log likelihood,

9

$\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$, the approximation is commonly called the "evidence lower bound" (ELBO; Jordan et al. 1999). We derived Equation 1 for every datum i, which leads to an overall full data approximate log-likelihood, $\mathcal{L}(\vec{\beta})$, which we maximize over parameters, $\vec{\beta}$, using the observed images and ratings:

$$(2) \qquad \mathcal{L}(\vec{\beta}) = \sum_i \ell^i_{approx}(\vec{\beta})$$

Equation 2 assumes global parameters for (1) predicting ratings based on embeddings and (2) generating images based on embeddings. If we did not make the assumption of global parameters, then the model would not be useful for either prediction or generation. We also specify a common prior describing the desired probabilistic structure of embedding (distributions). We make Equations 1 & 2 "variational" by assuming that each datum i has its own embedding distribution defined by distributional model parameters conditioned on the images and attributes. Finally, we note that the embedding portions of Equations 1 & 2 are equivalent to those derived from minimizing the KL divergence from (1) a true (unknown) distribution of the embeddings conditioned on the images and attributes to (2) the best variational distribution of the embeddings conditioned on the images and attributes (Jordan et al. 1999).

4.2 Decomposition Defines the Predictive, Generative, and Encoding Models

Because of the non-linear relationship between images, ratings, and embeddings we use neural networks for the predictive, generative, and embedding models. To train these neural networks feasibly, we decompose the global parameter space ($\vec{\beta}$) into parameters for the predictive model ($\vec{\beta}_P$), the generative model ($\vec{\beta}_G$), and encoding model ($\vec{\beta}_E$). We do so by using conditional probability to separate Equation 2 into its component models. Decomposition enables us to train and use each neural network separately and iteratively, which helps to overcome a key managerial issue, namely, that the model must be usable in an asynchronous and iterative environment by teams from industrial design, marketing, brand strategy, and so on. Specifically:

$$\mathcal{L}(\vec{\beta}) = \mathcal{L}_{pred}(\vec{\beta}_P) + \mathcal{L}_{gen}(\vec{\beta}_G) + \mathcal{L}_{enc}(\vec{\beta}_E)$$

$$(3) \qquad \mathcal{L}_{pred}(\vec{\beta}_P) = \sum_{i \in rated} E_{\vec{h}_i}[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}_P)]$$

$$\mathcal{L}_{gen}(\vec{\beta}_G) = \sum_{i \in rated,unrated} E_{\vec{h}_i}[\log p_{gen}(X_i | \vec{h}_i, \vec{\beta}_G)]$$

$$\mathcal{L}_{enc}\big(\vec{\beta}_E\big) = - \sum_{i \,\in\, rated,unrated} \Big\{ D_{KL}\Big(q_{enc}(\vec{h}_i|X_i, y_i, \vec{a}_i, \vec{\beta}_E)||p_{prior}(\vec{h}_i|\vec{a}_i)\Big)$$

$$+ \; D_{KL}\big(q_{attr}(\vec{\pi}_i|X_i, \vec{\beta}_E)|| \, p_{attr}(\vec{\pi}_i|\vec{a}_i)\big)\Big\}$$

This decomposition is possible since the embedding "bottlenecks" information contained in images, ratings, and attributes that is necessary to both predict ratings and generate new designs (Shu, Bui, and Ghavamzadeh 2016). We further promote separation in practice depending on the data regime, specifically, the presence of rating and attribute information. With unlabeled (i.e., unrated) data, we focus on learning images and attributes without rating dependency. Because the generative likelihood and the encoder likelihood do not depend upon the ratings or the parameters of the predictive model, we use all labeled <u>and</u> unlabeled images in the neural network for the embeddings.

For (relatively) less expensive unlabeled data, we have access to product attributes (e.g., color, brand). Accordingly, the encoder log-likelihood, $\mathcal{L}_{enc}\big(\vec{\beta}_E\big)$, differs from that in Equation 1 because we also learn the relationship between images and their attributes. Thus, in addition to the embeddings, $\vec{h}_i$, we use variational inference to encode attribute information with $\vec{\pi}_i$. Following the same reasoning used to derive Equation 1, we obtain the last KL divergence term in the encoder log-likelihood. This term acts as a classifier for attributes, $\vec{a}_i$, when the attributes are known. We use the classifier to predict attributes when they are unknown or ambiguous (e.g., when generating new designs). See also Appendix A3 and Keng (2017).

4.3. Deep Neural Networks to Parametrize Variational Probability Models

We choose probability distributions for the predictive, generative, and encoding models in Equation 3 using the framework of variational autoencoders (VAE; Kingma and Welling 2013). We later augment the VAE with concepts from adversarial learning as used in Generative Adversarial Networks (GANs, Goodfellow et al. 2014).

Predictive Model. We use a deep neural network, $f_P\big(\vec{h}_i, \vec{\beta}_P\big)$, that maps the embeddings to a prediction on the rating of interest, say a 1-to-5 rating on "appeal." Information about the full images and attributes is summarized in the embeddings. For the predictive model, we choose Laplace distributions with means y_i and unit diversity, $p_{pred}(y_i|\vec{h}, \vec{\beta}_P) = \frac{1}{2}e^{-\left|y_i - f_P(\vec{h}_i, \vec{\beta}_P)\right|}$. Define $\hat{y} = f_P\big(\vec{h}_i, \vec{\beta}_P\big)$ as the predicted rating from the neural network, then the predictive distribution implies that we minimize the mean absolute error of predicted versus true ratings.

$$(4) \quad \mathcal{L}_{pred}(\vec{\beta}_P) = \sum_{i \in rated} E_{\vec{h}_i}[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}_P)] = -\sum_{i \in rated} |y_i - \hat{y}_i|$$

Generative Model. We use a deep neural network, $f_G(\vec{h}_i ; \vec{\beta}_G)$ that generates candidate images from an embedding. We train the model on observed images, but use it for both observed and new images. We choose the generative model to be a high-dimensional Laplace distribution with means, X_i, and unit diversity, $p_{gen}(X_i | \vec{h}_i, \vec{\beta}_G) \propto e^{-|X_i - f_G(\vec{h}_i, \vec{\beta}_G)|}$. Define $\hat{X}_i$ as the predicted image from the neural network with elements $\hat{x}_{id}$ for each of the $3D$ pixels in that image (3D because there are three colors).

For many products, and vehicles in particular, we help the generative model by using "masks." A mask defines the general nature of a product, say "SUV-like" and is represented as a matrix with the same height and width as the product image. The mask's D pixels are binary values describing the presence of the product of a general shape. Masks focus generative models on product designs rather than other potentially unrelated information in product images. In the generator, masks are analogous to a 4th color; they are predicted by the same deep neural network using the (now-augmented) parameters, $\vec{\beta}_G$. Let M be the mask with elements m_d and let $\widehat{M}_i$ be the predicted mask with elements $\hat{m}_{id}$. Then, following the same arguments used for pixels, the log-likelihood function for the generative model is:

$$\mathcal{L}_{gen}(\vec{\beta}_G) = \sum_{i \in rated,unrated} E_{\vec{h}_i}[\log p_{gen}(X_i, M_i | \vec{h}_i, \vec{\beta}_G)]$$

$$(5)$$

$$= -\sum_{i \in rated,unrated} \left\{ \frac{1}{3D} \sum_d |x_{id} - \hat{x}_{id}| + \frac{1}{D} \sum_d |m_d - \hat{m}_{id}| \right\}$$

Encoding Model. We use a deep neural network, $f_E(X_i, \vec{a}_i, \vec{\beta}_E)$, to map images, X_i, and product attributes, $\vec{a}_i$, to an embedding, $\vec{h}_i$, with distributional parameters, $\vec{\mu}_i(X_i, \vec{a}_i)$ and $\vec{\sigma}_i(X_i, \vec{a}_i)$. Note that we omit the (labeled) rating dependency of the encoder as discussed in §4.2, as well as implicitly include per-datum variational distributional parameters in $\vec{\beta}_E$. The estimation is "amortized" into a single neural network (Shu et al. 2018).

For the embedding variational family, we choose multivariate Gaussian mixture distributions with mixture components depending on product attributes, $\vec{a}_i$, (e.g., 'SUV'). In other words, an embedding, $\vec{h}_i$, has a Gaussian mixture marginal distribution, but $\vec{h}_i | \vec{a}_i$ has a single Gaussian conditional distribution given attributes (Dilokthanakul et al. 2016). This expands a representation capacity of our

12

model (Ranganath, Tran, and Blei 2015), without resorting to more complicated autoregressive and flow-based methods (Chen et al. 2016).

We further assume each K-dimensional Gaussian has diagonal covariance, thereby factorizing into K conditionally independent Gaussians in which K is the dimensionality of the embeddings. If k indexes the elements of the embedding, then the variational assumption implies that $q_{enc}(\vec{h}_i \mid X_i, \vec{a}_i, \vec{\beta}_E) \propto \prod_{k=1}^{K} \sigma_{ik}^{-1} e^{-\frac{1}{2}(h_{ik}-\mu_{ik})^2 \sigma_{ik}^{-2}}$ where the $\vec{\mu}_i$ and $\vec{\sigma}_i$ are functions of X_i, $\vec{a}_i$, and $\vec{\beta}_E$. Using known formulae for $D_{KL}(\mathcal{N}(\mu_{ik}, \sigma_{ik}) \mid\mid \mathcal{N}(0,1))$, we obtain a simpler representation for the first divergence term in the encoder log-likelihood (Kingma and Welling 2013).

For the second divergence term in the encoder log-likelihood, we show in Appendix A3 that we may approximate $D_{KL}\big(q_{attr}(\vec{\pi}_i|X_i, \vec{\beta}_E)\mid\mid p_{attr}(\vec{\pi}_i|\vec{a}_i)\big) \cong constant - \log q_{attr}(\vec{a}_i|X_i, \vec{\beta}_E)$. This results in the encoder acting as a multinomial classifier, $q_{attr}(\vec{a}_i|X_i, \vec{\beta}_E)$, to predict attributes from product images. Specifically, we have C multinomial distributions, where C is the number of attributes (e.g., brand, bodytype) and ℓ_c is the number of levels of attribute c (e.g., Cadillac).

The encoder neural net for $\vec{\mu}_i$ and $\vec{\sigma}_i$ thus also produces $\vec{\pi}_i$, from which we draw Dirichlet probabilities, $\hat{a}_i = q_{enc}(\vec{\pi}_i|X_i, \vec{\beta}_E)$, using a soft-max function. We recognize $E_{\vec{\pi}_i}\big[\log q_{enc}(\vec{\pi}_i|X_i, \vec{\beta}_E)\big]$ as the cross-entropy for a draw of the attributes, $\vec{a}_i$, from the multinomial probabilities, $\hat{a}_i$. This provides the second term in Equation 6. This term rewards the encoder for its ability to predict the attributes based on the images. During training, this term encourages the encoder to learn attributes, while during prediction (when we do not know attributes) this term allows us to estimate unknown product attributes, $\hat{a}_i$, by sampling from the multinomial distribution indexed by $\vec{\pi}_i$. Putting both terms together we obtain:

$$(6) \qquad \mathcal{L}_{enc}(\vec{\beta}_E) = \sum_{i \in rated, unrated} \left\{ \sum_{k=1}^{K} \left[\frac{1}{2}(\mu_{ki}^2 + \sigma_{ki}^2) - \log \sigma_{ki} - \frac{1}{2} \right] + \sum_{c=1}^{C} \sum_{\ell=1}^{\ell_c} a_{ic\ell} \log \hat{a}_{ic\ell} \right\}$$

4.4. Adversarial Formulation, Generative Adversarial Network (GAN) Modification

Equations 3 through 6 specify a set of log-likelihood functions for the neural networks that together compose the VAE in Figure 2. The log-likelihoods balance the ability to predict ratings, generate new images, and represent images by embeddings. But to be effective for machine learning augmentation, we require that the embeddings encode and generate images well. For example, if we are generating luxury crossover utility vehicles (luxury CUVs), the images should look like well-designed luxury CUVs. After extensive experimentation and tuning, we found it necessary to augment the VAE

using the concept of adversarial training found in generative adversarial networks (GANs). We note our extension is similar to concurrent work by Heljakka, Solin, and Kannala (2019).

The basic idea is that we reward the generative model for generating images that are different than the prior, but we reward the encoder if it can distinguish a generated image from an observed image because it is too far from the prior. The GAN is implemented with competing objectives—a term in the generative model is the negative of a term in the encoder. Because the generator and encoder are trained iteratively, the generator and encoder reach a min-max solution to a two-player game. That is, we converge to a fixed point where generated images and actual images are both encoded to the same embedding space (Goodfellow et al. 2014).

We augment the variational autoencoder with the adversarial training concept from GANs, by adding the following term to the likelihood of the encoder to reward the encoder for distinguishing generated images from observed images. We subtract the same term from the log-likelihood of the generator to reward the generator to produce images that are hard to distinguish from observed images. These terms do no simply cancel out as we iteratively train the encoder and predictor while fixing the generator, and vice versa. We use the methods of §4.4 to simplify this expression with normally distributed prior and posterior distributions to obtain an analog to Equation 6.

$$(7) \qquad \sum_{i \,\in generated\ images} D_{KL}\left(q_{enc}(\vec{h_i}|X_i, \vec{a_i}, \vec{\beta_E})||p_{prior}(\vec{h_i}|\vec{a_i})\right)$$

This approach to adversarial training differs from conventional GANs in that we are not learning an implicit generative model by rewarding a "discriminator" (i.e., the analogue of the encoder in our work) to classify images as real or generated. We instead perform adversarial training in the embedding space much like feature matching and perceptual similarity approaches (Larsen, Sønderby, and Winther 2015). We maintain the probabilistic interpretation using the Kullback-Leibler divergence term rather than previously-used feature matching on point estimates conceptually similar to Huang et al. 2018).

4.5. Synthesizing the VAE, Adversarial (GAN-like), and Neural Network Loss Function Perspective

§§4.1-4.3 used the perspective of VAEs and probability models to derive an approximate separable likelihood function, the components of which are based on deep neural networks. These VAE components may be viewed as combining semi-supervised VAEs (Kingma et al. 2014) with conditional VAEs (Sohn, Lee, and Yan 2015). In §4.4 we augmented the VAE perspective with adversarial methods (GANs) to encourage the generator to generate recognizable new images while encouraging the encoder to recognize generated images as generated rather than observed. In contrast to typical VAE-GAN

approaches (Larsen, Sønderby, and Winther 2015; Zhao et al. 2016; Berthelot et al. 2017), adversarial autoencoders (Makhzani et al. 2015) and adversarial generative encoders (Heljakka, Solin, and Kannala 2018; Ulyanov, Vedaldi, and Lempitsky 2018), we retain the probabilistic interpretation of the VAE viewpoint as our adversarial losses are motivated by divergences of the embedding distribution. The combined approach improves embeddings and minimizes "posterior collapse" in VAEs; our model converges and the embeddings differ from the priors.

Table 1 summarizes the log-likelihood functions as loss functions comparable to more typical neural network frameworks. We do so recognizing that $loss(\vec{\beta}_P, \vec{\beta}_G, \vec{\beta}_E) = -\mathcal{L}(\vec{\beta}_P, \vec{\beta}_G, \vec{\beta}_E)$. Table 1 further indicates which images are included in summations.

5. Moving from Theory to Practical Implementation

The basic structure of our machine learning augmentation is derived by using deep neural networks to approximate separable log-likelihood functions. From this perspective, our augmentation is a VAE modified with an adversarial (GAN) component. We differ from most VAEs and GANs because we include as well information from product ratings, product attributes, and masks, all three of which are necessary to enhance predictive ability and generate realistic new aesthetic designs. But given the challenges discussed in §2, we must further customize and "tune" the model.

5.1. Custom Deep Learning Architecture

Many deep learning models are trained for one application (e.g., object detection) and modified for another application (e.g., aesthetic rating prediction), often referred to as "transfer learning" (Evgeniou and Pontil 2004). Such customized models have proven effective on image data for identifying product returns and brand identity (Dzyabura et al. 2019; Liu, Dzyabura, and Mizik 2019), and on textual data for identifying customer needs and predicting sales conversion (Liu, Lee, and Srinivasan 2019; Timoshenko and Hauser 2019). Unfortunately, we were unable to identify any pre-trained models that *generate* new highly-rated automotive images and which would be controllable in a manner that may be used by designers (e.g., morph Brand A sedan with Brand B sport-utility vehicle). The demands of interconnecting attributes, masks, ratings, high-resolution images, and adversarial training for generation together require a custom deep learning architecture trained on unique data. Because the custom deep learning model required substantial "tuning," we were careful to hold out data for model evaluation. We randomly split data into training, validation, and testing sets using a seeded random number generator for reproducibility. Validation data were used to set model hyperparameters (e.g., learning rates) and monitor training progress. Testing data were used only for model evaluation.
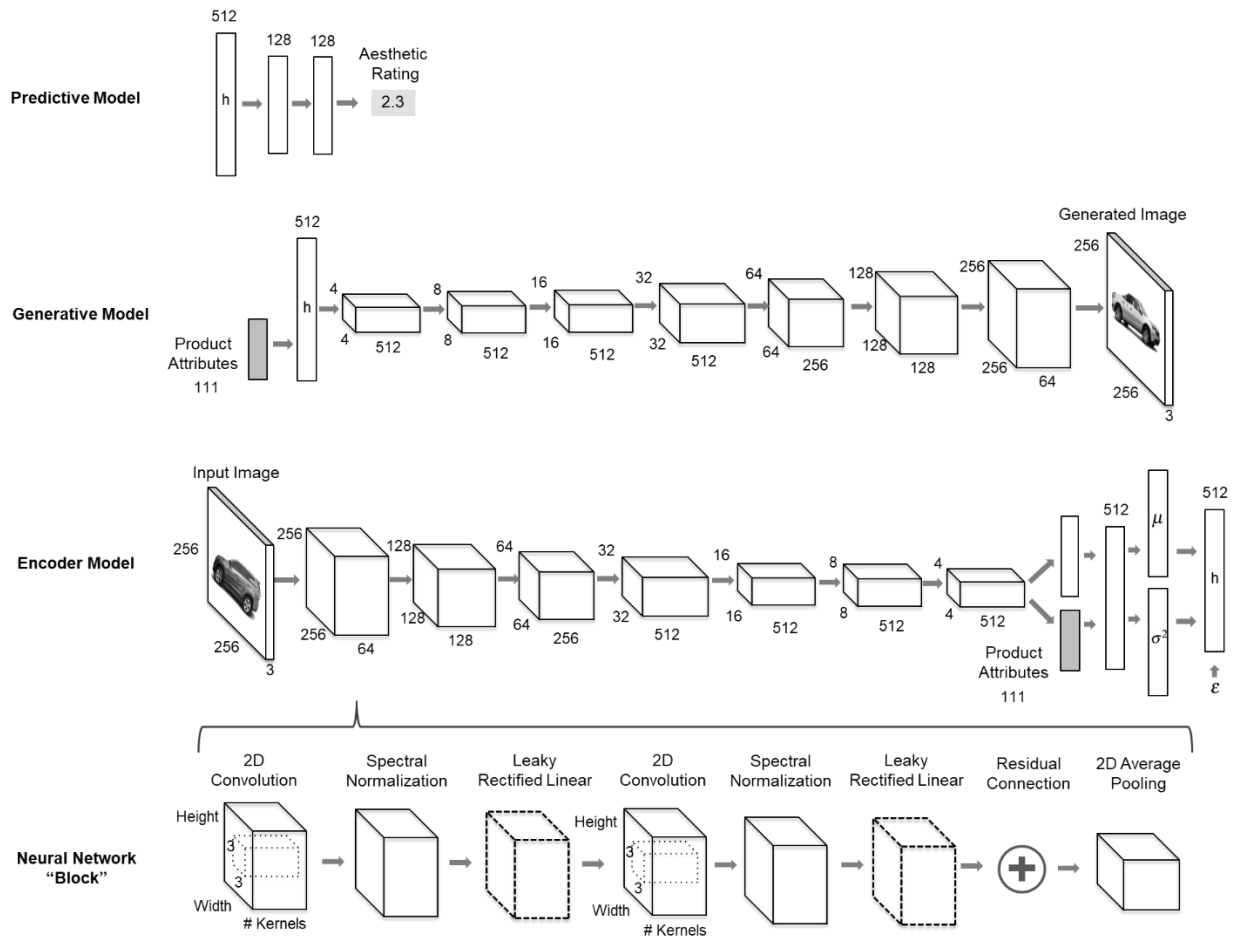
15

Table 1. Predictive, Generative, and Encoding Loss Functions (where loss = − log-likelihood)

Loss Function	Intuition
Predictive Model (summed over rated images)	
$|y_i - \hat{y}_i|$	Laplace term rewards the predictor for predicting ratings.
Generative Model (summed over rated, unrated images, or, when indicated, generated images)	
$\frac{1}{3D}\sum_d |x_{id} - \hat{x}_{id}|$	Laplace term rewards generator for predicting images that are close to true images.
$\frac{1}{D}\sum_d |m_{id} - \hat{m}_{id}|$	Laplace term rewards generator for predicting masks that correctly segment the design within the image.
$\sum_{k=1}^{K}\left[\frac{1}{2}(\mu_{kg}^2 + \sigma_{kg}^2) - \log\sigma_{kg} - \frac{1}{2}\right]$	Adversarial term rewards the generator for images that are drawn from the same distribution as observed images. Summed over <u>generated</u> images only (g).
Encoding Model (summed over rated, unrated images, or, when indicated, generated images)	
$\sum_{k=1}^{K}\left[\frac{1}{2}(\mu_{ki}^2 + \sigma_{ki}^2) - \log\sigma_{ki} - \frac{1}{2}\right]$, $\vec{\mu}_i, \vec{\sigma}_i = f_E(X_i, \vec{a}_i, \vec{\beta}_E)$	KL divergence rewards the encoder for embeddings close to the prior.
$-\sum_{c=1}^{C}\sum_{\ell=1}^{\ell_c} a_{ic\ell} \log \hat{a}_{ic\ell'}$, $\hat{a}_i = f_E(X_i, \vec{\beta}_E)$	Cross entropy rewards encoder for predicting attributes from images.
$-\sum_{k=1}^{K}\left[\frac{1}{2}(\mu_{kg}^2 + \sigma_{kg}^2) - \log\sigma_{kg} - \frac{1}{2}\right]$	Adversarial term rewards encoder for distinguishing generated images from observed images. Summed over <u>generated images</u> only (g).

Figure 3 summarizes the deep neural network architectures for the predictive, generative, and encoding models. Each architecture is based on many modeling decisions (and tuning). To simplify the description of the architectures, Figure 3 uses "blocks" of neural network "layers." The layers in each block are given at the top of Figure 3. Each of these layers (e.g., 2D convolution) performs the indicated operations on "features" from the previous layers. For example, 2D convolution applies two-dimensional filters to the input in an analogy to neuronal receptive fields (Olshausen and Field 1996). We use spectral normalization as a regularization technique to control the magnitude of gradients

16

during model training (see §5.3). A leaky rectified linear layer acts as a nonlinear activation function to transform values from the previous layer, thereby enabling the neural network to learn complex feature interactions. A residual connection simply adds the original input from the first layer of the block to the now transformed features at the end of the block. In doing so, the intermediary layers learn the residual error from the previous block. 2D average pooling reduces the dimensionality of the previous layer by down-sampling patches of 2D features to a single value.

Figure 3. Deep Neural Network Architectures for Predictive, Generative, and Encoding Models

5.2. Progressive Training

The rows in Figure 3 corresponding to the generative and encoding models enhance generation and encoding with a "progressive structure" as detailed in (Karras et al. 2017). In a progressive structure, we start at low resolutions, then increase resolution in stages and retrain the blocks until we reach the target image resolution of 256 x 256 pixels (three colors plus the mask). We label the resolution stages by height and width, say 4 x 4 pixels or 32 x 32 pixels. When the model is trained, we obtain lower

17

resolution images by down-sampling from the full image. The input and output image resolutions are given by the corresponding block. For example, we obtain 32 x 32 images as the 4th block in the encoder model. During generation, at each stage in the progressive structure, the immediately lower resolution image is blended with the stage's resolution according to an annealing schedule, in which the trained image progressively includes less of the low resolution image. Both the encoder and generator sample for multiple iterations before moving to next resolution—the number of samples is a hyper-parameter set using the validation data. The advantage of a progressive structure is that, as the model learns to encode or generate, it maintains knowledge from previous images effectively. Rather than starting model training from a completely blank slate at full resolution, the model has already learned information about the product design at lower resolutions.

5.3. Stabilization and Tuning of Model Training

Deep learning models do not always have well-behaved parameter estimation procedures. This is especially true for the models in Figure 3, which feature complex deep learning models that balance multiple competing and/or adversarial loss function terms. We therefore seek to avoid unbounded "back-propagated" gradients that lead to catastrophic failures in model training (e.g., outputting images of only black pixels). We stabilize training with the spectral normalization layer in each block. This promotes Lipschitz continuity to regularize the model by dividing the raw output weights of each layer by the largest singular value of the matrix of weights. See Miyato et al. 2018. We also manually enforce both soft and hard constraints on portions of the model architecture. For example, we bound the variance of the Gaussian random variables in the KL-divergence terms and use floating point safeguards to ameliorate the potential numerical instability introduced in §4 by the logarithms and various L_p-norms.

We tune model training by scaling the contribution of each loss term in Table 1 to its corresponding encoder, generator, or predictor loss with user-defined multiplicative scaling terms. This creates a balancing act between the seven loss terms given their endogenous interdependency during training. Further tuning is often domain specific (e.g., the scaled importance of the vehicle body style versus relative to its orientation). Empirically we found that the KL-divergence loss terms were best scaled to 1/10 or less relative to the loss terms for reconstructing images reconstruction loss, as well as annealed from zero at start of training to a maximum value after model convergence.

We stabilize model training by tuning the rate at which the competing models are updated. Empirically, we found that the generative model must be updated substantially more often than the encoding model; from 10 times as much to as little as two. We found that this ratio depends empirically

18

on the stage of model training. This may be unique to our application which combines likelihood-based VAE with adversarial training. Typically, generative GAN models are updated less often than encoding models (Arjovsky, Chintala, and Bottou 2017) and benefit from unequal learning rates (Heusel et al. 2017). By contrast our training ratio favors minimizing $D_{KL}(p|q)$ rather than $D_{KL}(q|p)$.

5.4. Gradient Backpropagation Using Local Reparameterization

We train the model by minimizing the loss functions in Table 1 with first-order stochastic gradient methods using mini-batches of training data. Specifically, we use the Adam stochastic gradient optimizer (Kingma and Ba 2015). Stochastic gradient methods are justified given their empirical performance and scalability via the backpropagation algorithm. Backpropagation simplifies an otherwise large calculation of a multi-parameter gradient to an equivalent series of smaller iterative gradient calculations. Gradients for a given loss in Table 1 propagate from the layer calculating the loss backwards to "earlier" layers, thereby taking advantage of the compositional structure of network layers and the chain rule of differentiation (Lagrange 1797).

To use gradient methods, we employ the "reparameterization trick" used in Kingma and Welling (2013). We rewrite the otherwise intractable gradient of an expectation over the embedding distribution (Equation 3 and Equation 7) to an equivalent tractable formulation by splitting the stochastic Gaussian embedding distribution into a deterministic neural net and an independent additive stochastic term. With this simplification we more easily compute an unbiased estimate of the gradient using Monte Carlo samples of the independent additive term. We similarly use this reparametrization trick when we do not have access to product attributes during training and inference. In this case we use a relaxation of the otherwise non-differentiable categorical attribute variables called the Gumbel-Softmax (Jang, Gu, and Poole 2016).

6. Case Study: Aesthetic Design of U.S. Automotive Market SUV/CUVs

Our machine learning augmentation has two goals (1) predict consumer evaluations of aesthetic designs (ratings) for a sample of proposed designs created and tested by an automotive partner and (2) generate new aesthetic designs that are likely to be rated highly. As an initial test of the proposed model (Figure 2) we evaluate (1) whether the trained VAE/GAN model can predict ratings of SUVs and CUVs that were evaluated in our partner's theme clinics and (2) whether high-predicted-appeal designs generated by the VAE/GAN model are rated highly and low-predicted-appeal designs are rated poorly.

6.1. Data: Images Rated in Theme Clinics

We obtained aesthetic ratings for images of the 203 unique SUVs/CUVs from model years 2010-

2014 (MY2010-14) tested in our partner's theme clinics. Our partner collected ratings from 178 targeted consumers in one of their theme clinics. Following established procedures for their theme clinics, our partner targeted consumers who were rigorously screened and selected to be interested in purchasing in the target category. Respondents were incentive aligned using our partners' standard methods, both fiscally and with knowledge that their input would guide future aesthetic design.

Ratings were obtained from a web-based survey co-located with the theme clinic. Warm-up questions motivated respondents (credibly) that their ratings would affect aesthetic design and introduced our partner's standard pairwise-semantic-differential rating-scale (most unappealing to most appealing). We anchored respondents' ratings by asking each respondent to rate five pre-chosen pairs of images—prechosen in pretests to be most divisive on the pairwise ratings scale. The five image pairs were the same for all respondents, but randomly counterbalanced between respondents.

Respondents each rated ten sequential pages of SUVs/CUVs in which each page contained five images. To test respondent consistency, the 2nd and 8th page and the 3rd and 9th page contained the same images randomly ordered. After extensive pretesting, the survey was implemented by our automotive partner. To maintain consistency among images and mitigate image-color biases, all images were reduced to greyscale and shown with a side viewpoint. (Unrated and generated images are three-color.) Appendix A4 provides an example rating page.

To maintain data quality prior to any further analysis, we eliminated respondents who were judged to be inconsistent based on Krippendorf's α where $\alpha = 1 -$ (observed disagreement among like images)/(expected disagreement due to chance). Krippendorf's α is a generalization of other interrater reliability measures such as Fleiss' κ and Cohen's κ (Krippendorff 2011). The cutoff was $\alpha = 0.75$, which eliminated 38 respondents (21%).

Ratings from consistent users were aggregated to a mean value for each of the 203 unique SUVs in MY2010-14, which were assigned to corresponding vehicle viewpoints to result in a full data of 7,308 rated images by giving the same mean rating to all viewpoints. We randomly split the full data into training, validation, and testing data at a ratio of 50%:25%:25%. (Same-model SUVs and CUVs remained together to ensure that validation or testing information did not leak into the training data.)

6.2. Data: Unrated Full-Color Images

Although high-quality full-color images are readily available on the web, almost all of these images are copyrighted and could not be used without permission to train our model. Fortunately, high-quality images are available from aggregators and are used often by automotive firms in their marketing communications. The typical cost is about $50,000 per month. We obtained 180,000 unrated SUV/CUV

20

images from our partner firm and a proprietary (unnamed) aggregator. All images were rescaled to 256 x 256 pixel images and included several attributes. We used conventional computer vision tools to obtain masks and attributes that were not coded by the firm, such as color and viewpoint.

7. Evaluation of the Machine Learning Augmentation

7.1. Predictive Ability

We evaluate the model on the 1,836 rated images in the held-out testing set. Table 2 reports the mean absolute error (MAE) for predicted-versus-actual ratings on the held-out data. Our model predicts well with a MAE of 0.372 of a scale point. Figure 4 provides example predictions for eight SUVs/CUVs. To put the MAE of the predicted model in perspective, we compare its predictive ability to a series of baseline models that vary from naïve to sophisticated, we used the same training and validation data to develop the benchmarks and spend considerable effort optimizing the sophisticated benchmarks so that the benchmarks provide a meaningful comparison.

Naïve baselines. The most naïve baseline is that respondents select the scale midpoint. This baseline represents zero information. A less naïve baseline uses global information from the training respondents by assuming that all variation about the median (average) rating is noise.

Sophisticated benchmark 1: Random Forest and Computer Vision Features. Computer vision and machine learning have a long history of processing high-dimensional image and video data for object detection and image segmentation. Conventional methods reduce high-dimensional visual data to a small set of "hand engineered" features chosen based on experience of successful prediction. These features are used in conventional machine learning methods such as support vector machines.

Our benchmark uses three types of hand-engineered features from computer vision: (1) Histogram of oriented gradients (HOG) features encode edge and shape information. HOG features divide the image into a grid of image patches, calculate the gradients of each patch, and bin these gradients into a histogram. Edge orientation and shape intensity are contained in the gradients' direction and magnitude values. (2) A downscaled version of the image itself (e.g., 256 x 256 to 32 x 32). And (3) histograms of color values for each red, green, and blue image channel. These features are used in a random forest with 100 trees. We chose a random forest because it performed best when tested against other common machine learning approaches: support vector machines, Gaussian process regression, and L1/L2-regularized linear regression.

Sophisticated benchmark 2: Pre-trained Deep Learning Model. Many researchers in marketing (and machine learning) use "pre-trained" open-source neural networks trained for one prediction task and repurposed for another prediction task. As a sophisticated benchmark, we used the pre-trained

21

VGG16 deep learning model trained on the ImageNet database. This benchmark outperformed other common pre-trained models (e.g., ResNet50, InceptionV3) for our prediction task. The VGG16 model is a pyramid of sixteen stacked layers (13 convolutional and 3 fully connected) that sequentially reduce images in size until they are classified in the last layer.
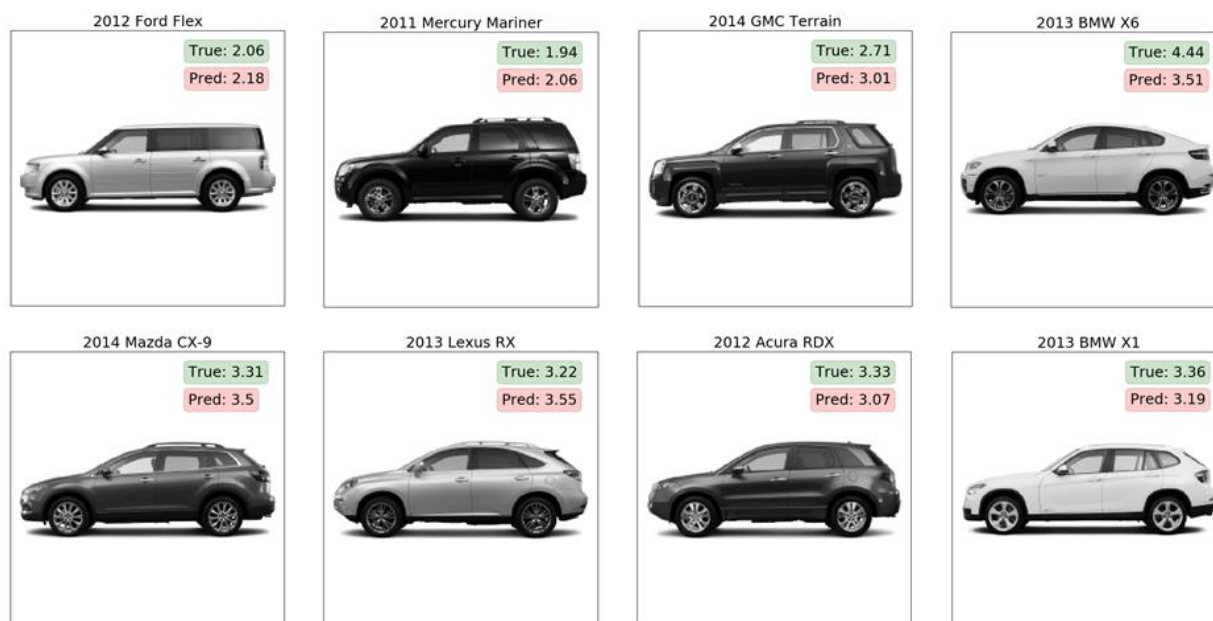
The initial layers of VGG16 transform pixels to edges and lines found in visual images. For our benchmark, we maintain the initial layers and replace the last classification layer with two batch-normalized fully-connected rectified-linear layers followed by a regression layer. (This architecture was chosen using validation data.) We train the model in two steps. We first freeze the pre-trained layers and use our data to train the new layers. We then fine-tune the entire neural network with scaled-down back-propagated gradients. The two-step procedure improves the benchmark's prediction.

Results. The proposed machine learning augmentation outperforms both the naïve baselines and the sophisticated benchmarks. The improvements are substantial. The sophisticated benchmarks do well relative to the naïve benchmarks and the proposed VAE/GAN deep-network does well relative to the sophisticated benchmarks.

Table 2. Predictive Test of Machine Learning Augmentation versus Baselines and Benchmarks

Prediction Model	Mean Absolute Error (std. dev.)	Improvement
Baseline: Scale Midpoint (Zero Information)	0.601 (n/a)	0.0%
Baseline: Median Rating in Training Images (Best Uniform)	0.603 (n/a)	~0.0 %
Benchmark: Computer Vision Features and Random Forest (Conventional Machine Learning)	0.450 (0.003)	25.3 %
Benchmark: VGG16 with Fine-Tuned Final Layers (Pre-trained Deep Learning)	0.434 (0.008)	28.0%
Proposed Machine Learning Augmentation (Custom Deep Learning)	0.372 (0.004)	38.3 %

22

Figure 4. Examples of Predictive Accuracy for Machine Learning Augmentation

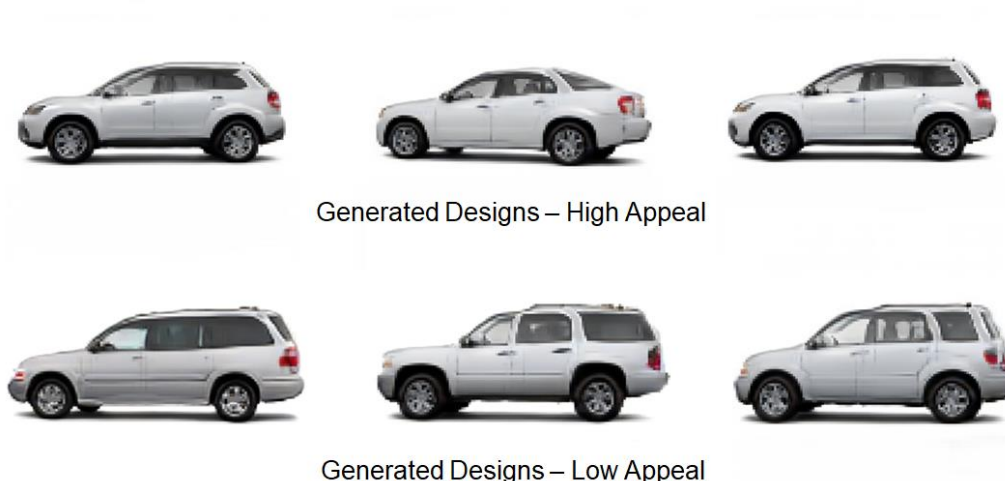

7.2. Generative Capability

From embeddings to images. Our first test is whether or not the embeddings (and attributes) can create credible product aesthetic images. We begin by sampling points in the embedding space, conditioned on desired attributes, then moving smoothly around that space. Each sampled point is on a hypersphere, such that interpolation is smooth in spherical coordinates. For each point in the embedding space, we generate a high-dimensional image. We judge that the images are realistic and are valuable inputs to the design team given their ability to be controllably morphed. We provide a demonstration video at https://vimeo.com/334094197.

Consumer evaluations. To test consumer reaction, we generated 50 targeted images: 25 were predicted by the VAE/GAN to be rated highly and 25 were predicted to be rated poorly. Figure 5 provides examples of generated images of each type. For consistency, we generated each image to be a light gray SUV/CUV from the side view. We added diversity with masks from other body styles (e.g., sedans, hatchbacks). The generated designs were morphed slightly to ensure plausibility by respondents and to mitigate the effect of biases (Lopez, Miller, and Tucker 2019). Using the methods of §6.1, we used 181 respondents from a professional Internet panel (ProdegeMR, at $4 per respondent) to rate these images. Using industry standards, we screened respondents to be SUV/CUV "intenders." To maintain quality, we pretested the survey carefully. Prior to any data analysis, following best market-research practices, we used "trap" questions to eliminate professional respondents, consistency checks to eliminate inattentive respondents, and response timings (and pattern checking) to eliminate

23

respondents who were just "clicking through." In total, we eliminated 53% of the initial 385 respondents who began the survey.

Respondents generally preferred images that were predicted to be rated high over images that were predicted to be rated low. In our data, respondents confirmed our predictions 74.0% of the time, well above random. We are encouraged by this initial test.

Figure 5. Example Generated Images (See also https://vimeo.com/334094197)

Generated Designs – High Appeal

Generated Designs – Low Appeal

8. Conclusion

8.1. Discussion and Summary

Deep learning methods are beginning to affect all aspects of marketing science, sometimes with methods customized to the challenge, sometimes by using tuned pre-trained models. Many of these methods rely on hard-to-quantify unstructured data such as natural language or images. We focus on consumers' aesthetic judgments by using state-of-the-art machine learning to augment human judgment. The augmentation comes in two forms. First, we predict consumer evaluations of new potential aesthetic designs from images. Second, we controllably generate new SUV/CUV images to enhance creative design.

These goals are challenging. We developed a machine learning augmentation that combines many concepts. We derive a version of the semi-supervised VAE from a probabilistic perspective to maximize the likelihood that an image is encoded well and rated as predicted, while using a low-dimensional embedding to "bottleneck" information between a predictive and generative model. Within the VAE framework, we include attributes to carry information about images and masks to constrain target images to be realistic. We add adversarial training concepts from the GAN literature to improve

24

training of realistic generated images while retaining VAE probabilistic interpretation so that the output does not collapse to known images. Finally, we use a variety of engineering ideas (e.g., spectral normalization, progressive training, residual connections, adaptively balanced training losses) to tune the deep learning model so that it is able to properly converge during training. The end result is a model that predicts image ratings substantially better than other machine learning benchmarks and controllably generates images that are perceived as valuable boosts to creativity.

An important contribution is our modeling decisions that address key managerial issues required for using machine intelligence within existing design processes. These contributions are a key difference of our work when contrasted with recent deep learning approaches; e.g., (Kang et al. 2017; Pan et al. 2017; Vasileva et al. 2018, Sbai et al. 2019) and the review by (Deng, Loy, and Tang 2017). We address the delicate integration between machine intelligence and existing human workflows, a point stressed heavily in our working interactions with marketing and design professionals. We focus on *augmenting* rather *automating* human expertise and human creativity, and ensuring all models are meaningfully controllable by the respective teams within the firm. For example, conditional independence assumptions enable full model decomposition into the predictive and generative models, allowing iterative and asynchronous training and usage by teams. Another example includes deriving our work from the VAE probabilistic framework rather than GANs, trading off generative realism for generative diversity for creative design.

The second managerial issue we address is the relatively limited amount of rated data on product images and aesthetic ratings. This was due to the expense required to collect rated data as well as show many unique designs exist; our rated data on new-SUV/CUV design images were only 203 unique SUV/CUVs over five years of theme clinics. If we were to train the deep learning model on these images alone, the embeddings and the generative capability would be weak if not impossible. We overcome this issue using semi-supervised learning to combine the expensive rated "small data," with the inexpensive and significantly larger "big data" of unrated images. Combining small data and big data made it feasible to train a deep learning model that does well on the theme-clinic-based data found in firms.

8.2. Limitations and Further Research

The SUV/CUV application is a proof-of-concept developed to augment a real automotive design team. New vehicles take many years and $1-5 billion in investment (Blonigen, Knittel, and Soderbery 2013). Over time we will learn whether machine learning augmentation has documented monetary benefits. Directly assessing the financial value of aesthetics is challenging given its interrelatedness with

25

confounding factors such as functional attributes, brand identity, and aesthetic trends (Person, Snelders, and Schoormans 2016). Recent promising work into disentangling these factors includes those that explicitly control of covariation in functional and form attributes (Kang et al. 2016; Zhang et al. 2019), as well as those that temporally model aesthetic trends (Yoganarasimhan 2017). For now, our work relies on predictive statistics and human judgment for validation. We hope that other researchers will test machine learning augmentation in other industries, perhaps industries in which the new product development cycle is shorter and less expensive.

Our model was engineered to be effective for vehicles after a heavy degree of engineering and model tuning, with much of the engineering challenge coming from the combination of VAE and GAN concepts. Many recent deep learning advances have focused on design generation using GANs, and, as a result, lack the latent space structure needed for the predictive model. When generation alone is the primary goal, GANs tend to outperform other models including VAEs and flow-based approaches; we direct readers to recent GAN applications for aesthetics (Kang et al. 2017; Pan et al. 2017; Sbai et al. 2019).

A natural next step is to assess the degree to which the proposed approach augments designer creativity. While perhaps less straightforward to measure, similar questions have seen recent marketing interest in applications such as idea generation (Toubia and Netzer 2017). Assuming our efforts continue to be guided by real design needs and managerial problems, this trend of various research streams being integrated bodes well for a future of hybrid human and machine intelligence.

26

Appendices

A1. Summary of Notation

i, j	indexes product images
d	indexes dimension of high-dimensional image space (i.e., a pixel)
k	indexes dimension of low-dimensional embedding
c, l	indexes c^{th} product attribute (e.g., brand) with l attribute levels (e.g., Cadillac)
X_i	product image i
x_{id}	pixel d of product image i
$\widehat{X}_i$	product image (generated)
$\hat{x}_{id}$	pixel d of product image i (generated)
y_i	aesthetic rating (known)
$\hat{y}_i$	aesthetic rating (predicted)
$\vec{h}_i$	product embedding vector
$\vec{a}_i$	product attributes vector (if known)
$\hat{a}_i$	product attributes vector (predicted)
M_i	product mask (if known)
m_{id}	pixel d of mask for product i
$\widehat{M}_i$	product mask (generated)
$\hat{m}_{id}$	pixel d of mask for product i (generated)

A2. Derivation of Approximate Log-Likelihood in Equations 1

§4.1 decomposed $p(y_i, X_i | \vec{a}_i, \vec{\beta})$, the joint likelihood of a product image, X_i, and its aesthetic rating, y_i, conditioned on product attributes, $\vec{a}_i$, into the predictive model and generative model. We seek to maximize the log-likelihood, $\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$, for every datum i in our training data with respect to the model parameters $\vec{\beta}$. However, parameter estimation and inference of this high-dimensional likelihood is an intractable problem. Instead we approximate the likelihood using variational inference. We first assume the existence of low-dimensional latent embeddings, $\vec{h}_i$. We introduce $\vec{h}_i$ via marginalization of $\vec{h}_i$ over the joint density in the second line of (A.1). We expand this density to the predictive model and generative model as well as a prior over the product embedding.

(A.1)
$$\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$$
$$= \log \int p(y_i, X_i, \vec{h}_i | \vec{a}_i, \vec{\beta}) \, d\vec{h}$$

27

$$= \log \int p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta}) \, d\vec{h}$$

We seek to learn an embedding *distribution* rather than just a point estimate of $\vec{h}_i$. We do not explicitly assume this form of the new joint density with the introduced product embedding, $\vec{h}_i$, and instead introduce a tractable distribution which we will use to approximate it, $q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})$, resulting in the "encoder model."

(A.2)
$$\log \int p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta}) \, d\vec{h}$$

$$= \log \int \frac{p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta}) d\vec{h}$$

$$= \log \mathrm{E}_{\vec{h}_i \sim q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \left[\frac{p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \right]$$

Our new goal is to find the best encoder model, $q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})$, for each datum i from an assumed *family* of tractable densities. In other words, we aim to estimate hyperparameters of the latent product embedding, $\vec{h}_i$, which index a unique element within an assumed variational distribution family. This results in unique variational parameters for each datum i which we implicitly include in the global set of model parameters, $\vec{\beta}$.

Estimating these parameters using sampling techniques (e.g., MCMC) is intractable, hence we cast sampling as an optimization problem using a lower bound of the expectation in Equation (A.2) via Jensen's inequality. This approximation is known as the "evidence lower bound," which is less than or equal to the intractable high-dimensional joint density, $\log p(y_i, X_i|\vec{a}_i, \vec{\beta})$.

(A.3)
$$\log \mathrm{E}_{\vec{h}_i \sim q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \left[\frac{p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \right]$$

$$\geq \mathrm{E}_{\vec{h}_i \sim q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \left[\log \frac{p_{pred}(y_i|\vec{h}_i, \vec{\beta}) p_{gen}(X_i|\vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i|\vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i|X_i, \vec{a}_i, \vec{\beta})} \right]$$

With the logarithm moved inside the expectation, we decompose the joint density into three separate terms: the predictive model, the generative models, and the ratio of the encoder and prior model. Under the expectation of the encoder model, this last term ends up being Kullback-Leibler

divergence between the encoder and the prior over the embedding, $D_{KL}(\log q_{enc} || p_{prior})$.

$$
\begin{aligned}
& E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log \frac{p_{pred}(y_i | \vec{h}_i, \vec{\beta}) p_{gen}(X_i | \vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \right] \\
& = E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) \right] + E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) \right] \\
& \qquad - E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log \frac{q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})}{p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})} \right] \\
& = E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) \right] + E_{\vec{h}_i \sim q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \left[\log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) \right] \\
& \qquad - D_{KL}\left(\log q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta}) || p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) \right) \\
& = \ell_{approx}^i(\vec{\beta})
\end{aligned}
$$

(A.4)

These three terms comprise the approximation, $\ell_{approx}^i(\vec{\beta})$, that we maximize. Since the Kullback-Leibler divergence term is negative, maximizing the overall approximation includes minimizing distributional dissimilarity between the posterior of the embedding given by the encoder model, $q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})$, and the distributional prior that we choose, $p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})$. If we minimize this divergence to zero, the approximate likelihood is equal to the true likelihood, i.e., $\log p(y_i, X_i | \vec{a}_i, \vec{\beta}) = \ell_{approx}^i(\vec{\beta})$. Thus, maximizing $\ell_{approx}^i(\vec{\beta})$ lower bounds the previously intractable likelihood maximization of $\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$.

A3. Derivation of Approximate Log-Likelihood from Equations 3

We derive Equation 3 using the same reasoning as in Appendix A2. The only difference between Equation 1 and Equation 3 is the additional term, $D_{KL}(q_{attr}(\vec{\pi}_i | X_i, \vec{\beta}_E) | p_{attr}(\vec{\pi}_i | \vec{a}_i))$. This term is the product attribute classifier with associated cross entropy loss term given in Table 1. We need this term to train the encoder model to be able to predict product attributes when they are not in the data, in particular, during prediction of new designs from generative model.

Adding the latent variables for parameters of the multinomial attribute classifier, $\vec{\pi}_i$, results in a double integral and a corresponding expectation over the joint density of both $\vec{h}_i$ and $\vec{\pi}_i$. Our assumptions on factorization of the latent terms splits into two KL-divergence terms in the last line of the derivation. See Keng (2017) for additional discussion on the relation between KL-divergence and the cross entropy loss term.

$$\log p(y_i, X_i | \vec{a}_i, \vec{\beta})$$

$$= \log \iint p(y_i, X_i, \vec{h}_i, \vec{\pi}_i | \vec{a}_i, \vec{\beta})\, \mathrm{d}\vec{h}d\vec{\pi}$$

$$= \log \iint p_{pred}(y_i | \vec{h}_i, \vec{\beta}) p_{gen}(X_i | \vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{h}_i, \vec{\beta})\, \mathrm{d}\vec{h}d\vec{\pi}$$

$$= \log \iint \frac{p_{pred}(y_i | \vec{h}_i, \vec{\beta}) p_{gen}(X_i | \vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{h}_i, \vec{\beta})}{q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta}) \mathrm{d}\vec{h}d\vec{\pi}$$

$$= \log \mathrm{E}_{\vec{h}_i, \vec{\pi}_i \sim q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\frac{p_{pred}(y_i | \vec{h}_i, \vec{\beta}) p_{gen}(X_i | \vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{h}_i, \vec{\beta})}{q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} \right]$$

$$\geq \mathrm{E}_{\vec{h}_i, \vec{\pi}_i \sim q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log \frac{p_{pred}(y_i | \vec{h}_i, \vec{\beta}) p_{gen}(X_i | \vec{h}_i, \vec{\beta}) p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{h}_i, \vec{\beta})}{q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} \right]$$

$$= \mathrm{E}_{\vec{h}_i, \vec{\pi}_i \sim q(\vec{h}_i, \vec{\pi}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) + \log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) + \log \frac{p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})}{q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})} \right.$$

$$\text{(A.4)} \qquad \left. + \log \frac{p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{\beta})}{q_{attr}(\vec{\pi}_i | X_i, \vec{\beta})} \right]$$

$$= \mathrm{E}_{\vec{\pi}_i \sim q(\vec{\pi}_i | X_i, \vec{\beta})} \left[\mathrm{E}_{\vec{h}_i \sim q(\vec{h}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) + \log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) - \log \frac{q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})}{p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})} \right. \right.$$

$$\left. \left. - \log \frac{q_{attr}(\vec{\pi}_i | X_i, \vec{\beta})}{p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{\beta})} \right] \right]$$

$$= \mathrm{E}_{\vec{h}_i \sim q(\vec{h}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) + \log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) - \log \frac{q_{enc}(\vec{h}_i | X_i, \vec{a}_i, \vec{\beta})}{p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta})} \right]$$

$$- \mathrm{E}_{\vec{\pi}_i \sim q(\vec{\pi}_i | X_i, \vec{\beta})} \left[\log \frac{q_{attr}(\vec{\pi}_i | X_i, \vec{\beta})}{p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{\beta})} \right]$$

$$= \mathrm{E}_{\vec{h}_i \sim q(\vec{h}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log p_{pred}(y_i | \vec{h}_i, \vec{\beta}) \right] + \mathrm{E}_{\vec{h}_i \sim q(\vec{h}_i | \vec{a}_i, X_i, \vec{\beta})} \left[\log p_{gen}(X_i | \vec{h}_i, \vec{\beta}) \right]$$

$$- \mathrm{D}_{KL} \left[q(\vec{h}_i | \vec{a}_i, X_i, \vec{\beta}) || p_{prior}(\vec{h}_i | \vec{a}_i, \vec{\beta}) \right] - \mathrm{D}_{KL} \left(q_{attr}(\vec{\pi}_i | X_i, \vec{\beta}) | p_{attr}(\vec{\pi}_i | \vec{a}_i, \vec{\beta}) \right)$$

30

A4. Example Rating Page from Aesthetic Rating Survey used in Theme Clinic

References

Aaker DA, Keller KL (1990) Consumer evaluations of brand extensions. *Journal of Marketing* 54(1):27–41.

Adidas AG (2017) *Adidas AG 2017 Annual Report*.

Ansari A, Li Y, Zhang JZ (2018) Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach. *Marketing Science* 37(6):987–1008.

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

Berlyne DE (1971) *Aesthetics and Psychobiology* (Appleton-Century-Crofts, East Norwalk, CT, US).

Berthelot D, Schumm T, Metz L (2017) BEGAN: Boundary Equilibrium Generative Adversarial Networks. arXiv:1703.10717 [cs, stat].

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.

Blonigen BA, Knittel CR, Soderbery A (2013) *Keeping it Fresh: Strategic Product Redesigns and Welfare* (National Bureau of Economic Research).

Bouchard C, Aoussat A, Duchamp R (2006) Role of sketching in conceptual design of car styling. *Journal of Design Research* 5(1):116.

Braun M, McAuliffe J (2010) Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association* 105(489):324–335.

Chakraborty I, Kim M, Sudhir K (2019) Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Attribute Self-Selection. *Cowles Foundation Discussion Paper No. 2176*.

Chan TH, Mihm J, Sosa ME (2018) On styles in product design: An analysis of U.S. design patents. *Management Science* 64(3):1230–1249.

Chen X, Kingma DP, Salimans T, Duan Y, Dhariwal P, Schulman J, Sutskever I, Abbeel P (2016) Variational lossy autoencoder. *arXiv:1611.02731 [cs, stat]*.

Cho H, Hasija S, Sosa M (2015) *How Important is Design for the Automobile Value Chain?* (Social Science Research Network, Rochester, NY).

Clement J (2007) Visual influence on in-store buying decisions: An eye-track experiment on the visual influence of packaging design. *Journal of Marketing Management* 23(9–10):917–928.

Coates D (2003) *Watches Tell More Than Time: Product Design, Information, and the Quest for Elegance* (McGraw-Hill London).

Cooper RG (1990) Stage-gate systems: A new tool for managing new products. *Business horizons* 33(3):44–54.

Creusen MEH, Schoormans JPL (2005) The different roles of product appearance in consumer choice. *Journal of Product Innovation Management* 22(1):63–81.

Crilly N, Moultrie J, Clarkson PJ (2004) Seeing things: Consumer response to the visual domain in product design. *Design Studies* 25(6):547–577.

Danneels E, Kleinschmidtb EJ (2001) Product innovativeness from the firm's perspective: Its dimensions and their relation with project selection and performance. *Journal of Product Innovation Management* 18(6):357–373.

Deng Y, Loy CC, Tang X (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, *34*(4), 80-106.

Dilokthanakul N, Mediano PAM, Garnelo M, Lee MCH, Salimbeni H, Arulkumaran K, Shanahan M (2016) Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv:1611.02648 [cs, stat]*.

Dotson J, Beltramo M, Feit E, Smith R (2016). Modeling the Effect of Images on Product Choices. *SSRN Electronic Journal*.

Dzyabura D, Hauser JR, El Kihal S, Ibragimov M (2018) Leveraging the power of images in predicting product return rates. *SSRN Electronic Journal*.

Evgeniou T, Pontil M (2004) Regularized multi–task learning. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (ACM), 109–117.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*. 2672-2680.

Hartley J (1996) Brands Through the Lens of Style. *Quest and Associates.*

Hauser J, Eggers F, Selove M (2018). The Strategic Implications of Scale in Choice-Based Conjoint Analysis.

Hekkert P, Snelders D, Wieringen PC (2003) 'Most advanced, yet acceptable': Typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology* 94(1):111–124.

Heljakka A, Solin A, Kannala J (2018) Pioneer networks: Progressively growing generative autoencoder. *arXiv:1807.03026 [cs, stat]*.

Heljakka A, Solin A, Kannala J (2019) Towards photographic image manipulation with balanced growing of generative autoencoders. *arXiv:1904.06145 [cs, stat]*.

Hertenstein JH, Platt MB, Veryzer RW (2005) The impact of industrial design effectiveness on corporate financial performance*. *Journal of Product Innovation Management* 22(1):3–21.

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*. 6626–6637.

Hoffman M, Blei D, Wang C, Paisley J (2013). Stochastic variational inference. *Journal of Machine Learning Research* 14(1):1303-1347.

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

Huang H, Li Z, He R, Sun Z, Tan T (2018) IntroVAE: Introspective variational autoencoders for photographic image synthesis. *Advances in Neural Information Processing Systems*. 31(1):52–63.

Ishigaki T, Terui N, Sato T, Allenby G (2018). Personalized market response analysis for a wide variety of products from sparse transaction data. *International Journal of Data Science and Analytics* 5(4):233-248.

Jang E, Gu S, Poole B (2016) Categorical reparameterization with gumbel-softmax. *arXiv:1611.01144 [cs, stat]*.

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.

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. *Machine learning* 37(2):183–233.

Kang N, Ren Y, Feinberg FM, Papalambros PY (2016) Form + function: Optimizing aesthetic product design via adaptive, geometrized preference elicitation. Working Paper - University of Michigan.

Kang WC, Fang C, Wang Z, McAuley J (2017) Visually-aware fashion recommendation and design with generative image models. *2017 IEEE International Conference on Data Mining (ICDM)*. 207–216.

Keng (2017) Semi-supervised Learning with Variational Autoencoders. Available: https://bit.ly/2O9RvF8.

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

Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Keller KL (2003) Brand Synthesis: The multidimensionality of brand knowledge. *Journal of Consumer Research* 29(4):595–600.

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. *International Conference on Learning Representation*. (San Diego, California).

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

Kingma DP, Welling M (2013) Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kreuzbauer R, Malter AJ (2005) Embodied cognition and new product design: Changing product form to influence brand categorization. *Journal of Product Innovation Management* 22(2):165–176.

Krippendorff K (2011) Computing Krippendorff's alpha-reliability.

Landwehr JR, Labroo AA, Herrmann A (2011) Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts. *Marketing Science* 30(3):416–429.

Larsen ABL, Sønderby SK, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

Liu L, Dzyabura D, Mizik N (2017) Visual listening in: Extracting brand image portrayed on social media.

Liu X, Lee D, Srinivasan K (2018) Large scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

Lopez C, Miller S, Tucker C (2019). Exploring biases between human and machine generated designs. *Journal of Mechanical Design*, *141*(2), 021104.

Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Manoogian J II (2013) Vehicle design process used at general motors.

Martindale C (1990) *The Clockwork Muse: The Predictability of Artistic Change* (Basic Books, New York, NY).

Miyato T, Kataoka T, Koyama M, Yoshida Y (2018) Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Noble CH, Kumar M (2010) Exploring the appeal of product design: A grounded, value-based model of key design elements and relationships. *Journal of Product Innovation Management* 27(5):640–657.

Norman DA (2004) *Emotional Design: Why We Love (or Hate) Everyday Things* (Basic books, New York, NY).

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607.

Orsborn S, Cagan J, Boatwright P (2009) Quantifying aesthetic form preference in a utility function. *Journal of Mechanical Design* 131(6):061001.

Orth UR, Malkewitz K (2008) Holistic package design and consumer brand impressions. *Journal of marketing*, *72*(3):64-81.

Orme B, Chrzan K (2017). Becoming an Expert in Conjoint Analysis: Choice Modelling for Pros. Sawtooth Software.

Palazzolo M, Feinberg F (2015) Modeling consideration set substitution. Working Paper - University of Michigan.

Pan Y, Burnap A, Hartley J, Gonzalez R, Papalambros PY (2017) Deep design: Product aesthetics for heterogeneous markets. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. (ACM, New York, NY, USA), 1961–1970.

Pauwels K, Silva-Risso J, Srinivasan S, Hanssens DM (2004) New products, sales promotions, and firm value: The case of the automobile industry. *Journal of Marketing* 68(4):142–156.

Person O, Snelders D, Karjalainen TM, Schoormans J (2007) Complementing intuition: insights on styling as a strategic tool. *Journal of Marketing Management* 23(9–10):901–916.

Person O, Snelders D, Schoormans J (2016) Assessing the performance of styling activities: An interview study with industry professionals in style-sensitive companies. *Design Studies* 42:33–55.

Ranganath R, Tran D, Blei DM (2015) Hierarchical variational models. *arXiv:1511.02386 [cs, stat]*.

Ranscombe C, Hicks B, Mullineux G, Singh B (2012) Visually decomposing vehicle images: Exploring the influence of different aesthetic features on consumer perception of brand. *Design Studies* 33(4):319–341.

Reid T, Gonzalez R, Papalambros PY (2010) Quantification of perceived environmental friendliness for vehicle silhouette design. *Journal of Mechanical Design* 132(10):101010.

Reid T, MacDonald E, Du P (2013) Impact of product design representation on customer judgment. *Journal of Mechanical Design* 135(9):091008.

Reppel AE, Szmigin I, Gruber T (2006) The iPod phenomenon: identifying a market leader's secrets through qualitative marketing research. *The Journal of Product and Brand Management; Santa Barbara* 15(4):239–249.

Rubera G (2015) Design innovativeness and product sales' evolution. *Marketing Science* 34(1):98–115.

Sbai O, Elhoseiny M, Bordes A, LeCun Y, Couprie C (2019) DesIGN: Design inspiration from generative networks. Leal-Taixé L, Roth S, eds. *Computer Vision – ECCV 2018 Workshops*. (Springer International Publishing, Cham), 37–44.

Shu R, Bui HH, Zhao S, Kochenderfer MJ, Ermon S (2018) Amortized inference regularization. *arXiv:1805.08913 [cs, stat]*.

Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*. 3483–3491.

Sun C, Ghose A, Liu X (2018) Automating Online-Offline Data Multi-View Merger for Integrated Marketing. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3265400.

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Science* 38(1):1-20.

Toffoletto G (2013) The strategic value of design. A model derived from the existing literature and six case studies of design driven organizations. Politecnico di Milano.

Toubia O, Netzer O (2017) Idea generation, creativity, and prototypicality. *Marketing Science* 36(1):1–20.

Ulyanov D, Vedaldi A, Lempitsky V (2018) It takes (only) two: Adversarial generator-encoder networks. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vasileva, M, Plummer B, Dusad K, Rajpal S, Kumar R, Forsyth D (2018). Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 390-405.

Vlasic B (2011) *Once upon a car: The fall and resurrection of America's big three auto makers–GM, Ford, and Chrysler* (William Morrow).

Yoganarasimhan H (2017) Identifying the presence and cause of fashion cycles in data. *Journal of Marketing Research* 54(1):5–26.

Zhang W, Yang Z, Jiang H, Nigam S, Yamakawa S, Furuhata T, Shimada K, Kara LB (2019) 3D shape synthesis for conceptual design and optimization using variational autoencoders. *arXiv:1904.07964 [cs, stat]*.

Zhao J, Mathieu M, LeCun Y (2016) Energy-based Generative Adversarial Network. arXiv:1609.03126.