What Characterizes Personalities of Graphic Designs?

NANXUAN ZHAO, City University of Hong Kong
YING CAO, City University of Hong Kong
RYNSON W.H. LAU, City University of Hong Kong

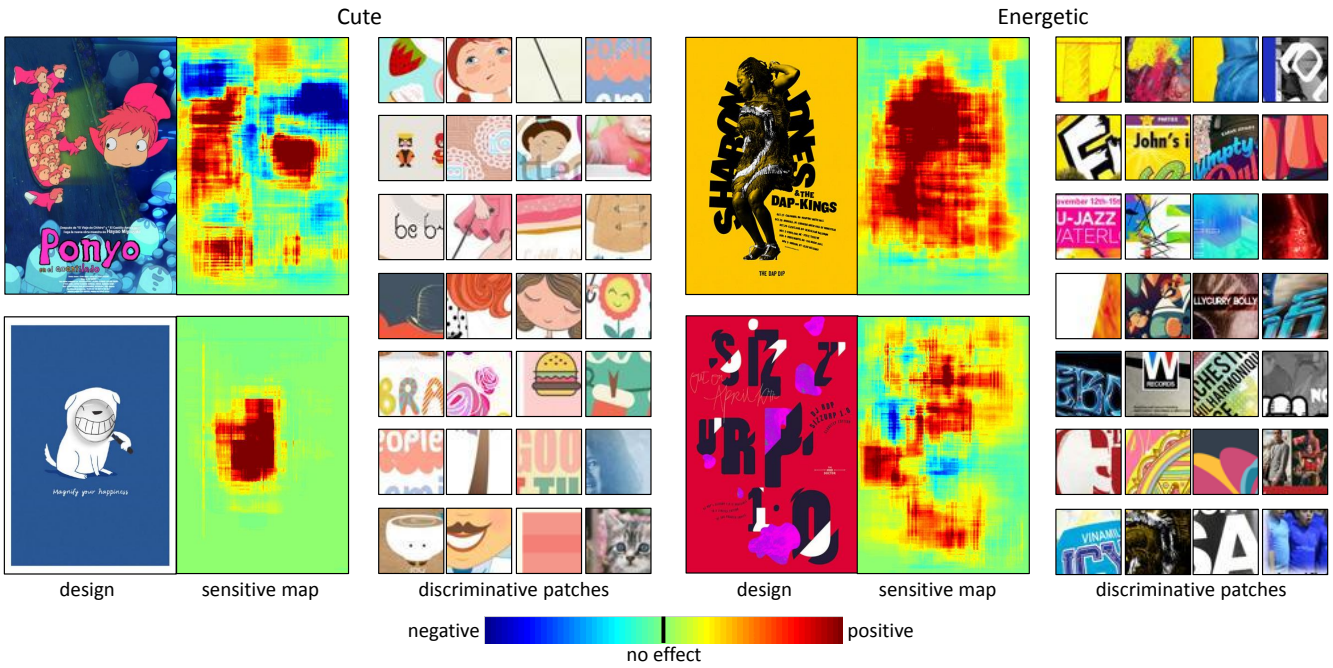


Fig. 1. Our model can automatically discover what characterizes the personality of a graphic design. For each of the two personality labels ("Cute" and "Energetic") shown above, we show two design examples together with their corresponding personality-based sensitive maps. This sensitive map measures the contribution of each local region of a design example to a particular personality, with red indicating positive effect while blue indicating negative effect. For each personality, we also show a set of discriminative patches that contribute most to the personality. These patches are mined from a collection of designs that are ranked top by our model based on the personality. ©Hayao Miyazaki, I Love Doodle, Scott Williams and Andrea Dell'Anna.

Graphic designers often manipulate the overall look and feel of their designs to convey certain personalities (e.g., cute, mysterious and romantic) to impress potential audiences and achieve business goals. However, understanding the factors that determine the personality of a design is challenging, as a graphic design is often a result of thousands of decisions on numerous factors, such as font, color, image, and layout. In this paper, we aim to answer the question of what characterizes the personality of a graphic design. To this end, we propose a deep learning framework for exploring the effects of various design factors on the perceived personalities of graphic designs. Our framework learns a convolutional neural network (called *personality scoring network*) to estimate the personality scores of graphic designs by ranking the crawled web data. Our personality scoring network automatically learns a visual representation that captures the semantics necessary to predict graphic design personality. With our personality scoring network, we systematically and quantitatively investigate how various design factors (e.g., color, font, and layout) affect design personality across different scales (from pixels, regions to elements). We also demonstrate a number of practical application scenarios of our network, including element-level design suggestion and example-based personality transfer.

CCS Concepts: • **Computing methodologies** → **Perception**; Neural networks;

Additional Key Words and Phrases: Graphic design, Personality, Deep learning

ACM Reference Format:
Nanxuan Zhao, Ying Cao, and Rynson W.H. Lau. 2018. What Characterizes Personalities of Graphic Designs?. *ACM Trans. Graph.* 37, 4, Article 116 (August 2018), 15 pages. https://doi.org/10.1145/3197517.3201355

Ying Cao is the corresponding author. This work was led by Rynson Lau.
Authors' addresses: Nanxuan Zhao, Department of Computer Science, City University of Hong Kong, nanxuzhao2-c@my.cityu.edu.hk; Ying Cao, Department of Computer Science, City University of Hong Kong, caoying59@gmail.com; Rynson W.H. Lau, Department of Computer Science, City University of Hong Kong, rynson.lau@cityu.edu.hk.

0730-0301/2018/8-ART116 $15.00
https://doi.org/10.1145/3197517.3201355

1 INTRODUCTION

Graphic design, as a visual communication medium, is ubiquitous in our daily life, including newspapers, magazines, packaging, posters and websites. It is especially created to convey certain ideas and messages to its audiences via a combination of images, symbols, and text. To achieve this goal, designers often deliberately adjust the overall look and feel of their designs to give people an instant impression that may last for a long time, even before they start reading the contents [Nauert 2011; Phillips and Chaparro 2009; Reinecke et al. 2013]. This look and feel of a design is known as the design's *personality*, which signatures the design and should match with the design objectives [Gross 2015; Jensen 2013]. A design's personality can often be described using a set of adjectives. For example, a fitness website or poster should look powerful, fresh and well-organized. A business card should look elegant, formal and respectful. A design that can reflect a proper personality has the ability to attract the right clients and repel the wrong ones, while building up a distinctive identity and setting it apart from competitors [Gross 2015; Walter 2012]. However, understanding what makes a design embrace a certain personality is challenging, as a design is a result of thousands of decisions on numerous factors, including color, font, image, and layout. Given such a large design decision space, how to select these factors such that they will interact with each other to manifest a certain personality is a non-trivial task even for experienced designers.

In this paper, we want to answer a question: what makes a graphic design possess a certain personality? Such a study may provide insights into the best practices for expressing personalities in a real-world graphic design process and also inform computational methods for facilitating graphic design. Our work focuses on one common yet important type of graphic design, posters, as they exhibit a diverse range of personalities in order to immediately catch viewers' attention.

To this end, we propose a convolutional neural network (CNN), which, given a graphic design and a personality label, predicts a score that indicates the degree of the personality possessed by the design. We refer to our CNN as a *personality scoring network*. Instead of using hand-crafted features with limited representation power, our network can automatically learn the most relevant features for predicting human perception of design personality. However, the personality score of a design is unknown a priori, and is very subjective and subtle for human to annotate reliably. Thus, we propose to supervise our network in a ranking formulation to learn by comparing pairs of graphic designs in terms of personality. This allows us to train our model for predicting personality scores without the need for score-level supervision. To train our network, we leverage a large number of graphic designs with personality tags from the web, with little human intervention. This makes it possible to scale our method to massive amounts of data.

With our personality scoring network, we have conducted several qualitative and quantitative experiments, in an attempt to glean insights into what makes graphic designs embrace certain personalities. We first construct a personality-based sensitive map to visually reveal which part of a design that contributes most/least to its personality. We then study how design personality correlates with some key design factors including color, font and space, which have a great impact on human perception of a design [Cousins 2015; Kliever 2015]. Furthermore, we show that our personality scoring network and our learned high-level discriminative features can benefit two personality-aware graphic design tasks: 1) Element-level design suggestion: given a target element on a design, our method will suggest the properties of the element in order to enhance the personality of the design. 2) Example-based personality transfer: our method automatically modifies a source design to match it with a reference design in terms of personality. In summary, our major contributions are:

- We propose a deep ranking framework to learn a model for estimating personality scores of graphic designs (Section 4). Our framework is trained from web data with minimal human supervision.
- We perform quantitative and qualitative analyses using our learned model, which offer a systematic understanding of what makes a graphic design convey a certain personality (Section 6).
- We present two novel and practical personality-based design applications enabled by our model (Section 7).

2 RELATED WORK

To our knowledge, we are the first to seek for understanding of what characterizes the personality of a graphic design. Our work bears some high-level similarity to a recent work [Doersch et al. 2015], which mines the discriminative image patches that characterize a city from a set of geo-tagged web images. While we also aim to find discriminative design ingredients that are important for conveying a personality from web data with weak supervision, unlike their patch-level analysis, our network allows us to analyze across different scales from pixels, patches to elements. In addition, instead of using hand-crafted features (e.g., HOG and color as in their work), we take advantage of the CNN to automatically learn the most relevant features to our task. In the remainder of this section, we review some previous works that are relevant to ours.

2.1 Semantic Attributes

Personality can be regarded as a specific class of semantic attributes, which is a set of words to describe visual or functional properties of objects. In this regard, our work can be related to emerging research efforts on using semantic attributes as high-level, linguistic descriptions to guide image searching [Parikh and Grauman 2011] and editing [Laffont et al. 2014], font selection [O'Donovan et al. 2014b], 3D shape editing [Yumer et al. 2015], material appearance editing [Serrano et al. 2016], and 3D avatar generation [Streuber et al. 2016]. To model the attributes, these methods proposed to learn a function for mapping an object to a continuous score indicating the strength of an attribute, from crowdsourced data. Our focus is on modeling personalities of graphic designs, which has not been investigated before. In addition, to learn our personality scoring function, we propose to take advantage of web data harvested from an image search engine to avoid expensive crowdsourcing. Karayev et al. [2013] constructed a large-scale dataset of images with style annotations for visual style recognition. In contrast to this work

that predicts the presence of a style attribute on an image, we aim to regress a continuous strength of an attribute, and demonstrate the advantage of the regression formulation over naive classification in our experiments.

2.2 Graphic Design

Various graphic design factors, color [Jahanian et al. 2017; Lin et al. 2013; O'Donovan et al. 2011], font [O'Donovan et al. 2014b], illustrations [Garces et al. 2014, 2017], and layout [Cao et al. 2012, 2014; O'Donovan et al. 2014a] have been studied individually. To study graphic design as a whole, Ritchie et al. [2011] designed a style-based exploration tool for webpages using low-level features (e.g., mean color, number of words on a page). Pang et al. [2016] utilized a set of features to characterize the temporal behaviors of user attention on webpages. Chaudhuri et al. [2013] proposed a part-based assembly approach for 3D shape creation based on semantic attributes, and extended it to webpages based on both global and local features. Saleh et al. [2015] learned a style similarity metric for searching infographics based on the features extracted from bitmap input. All of the above works rely on hand-crafted features, which have limited representation power and are specific to their own problems. In contrast, we learn powerful and generic design features that characterize the personalities of graphic designs automatically with a CNN.

A few recent works on graphic designs have applied CNNs to learn features for different tasks. Bylinskii et al. [2017] developed fully convolutional networks to predict the importance maps of both graphic designs and data visualizations. Redi et al. [2017] designed a deep classifier to predict the aesthetic level of non-photographic images. We leverage the CNN to model perceptual personality on graphic designs and systematically investigate its relation with various design factors, which has not been studied by the previous works.

2.3 Deep Ranking Network

Deep ranking networks have been proposed to learn from ranked labels with deep neural networks. They have been applied to both 2D images [Chang et al. 2016; Gygli et al. 2016; Wang et al. 2014] and 3D shapes [Lau et al. 2016] with great success. A loss function is often formulated on triplets [Wang et al. 2014; Zhao et al. 2015] or pairs [Gong et al. 2013; Gygli et al. 2016]. In the triplet approach, given a triplet (query, positive and negative), a loss function is designed such that the positive will be closer to the query in the learned embedding space than the negative. This approach has been applied to learn the fine-grained image similarity [Wang et al. 2014] and hash function for image retrieval [Zhao et al. 2015]. The loss function of the pairwise approach is defined over pairs of examples, and requires one example to be ranked higher than the other. For example, Gygli et al. [2016] learned a network to produce a ranked list of video segments according to their suitability as animated GIFs by requiring the GIF segments to score higher than the non-GIF segments. Our work builds upon this kind of learning frameworks, using ranked graphic design pairs obtained from the web, but integrates a semantic embedding network into the deep ranking network to learn the semantics-aware features.

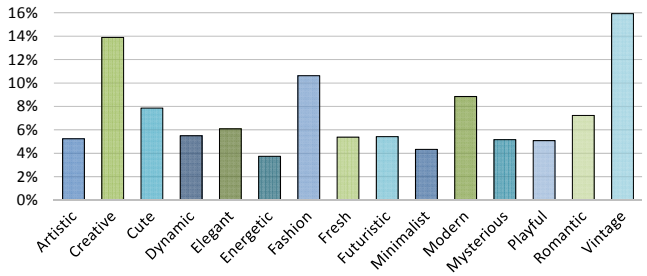


Fig. 2. Distribution of different personality labels in our poster dataset.

3 DATA COLLECTION

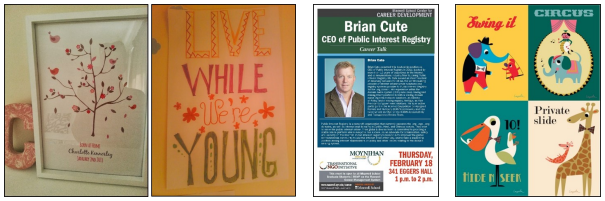
To train our model, we construct a dataset of posters with personality labels, by querying the Google image search engine with a list of personality labels as keywords. In this section, we first describe how we select the list of personality labels. We then discuss how we collect and post-process our dataset.

3.1 Personality Selection

There is a long list of personality labels that can be used to describe designs. However, not all of them are commonly used and easily perceptible by people. To select the labels for our work, we have first collected an initial list of 40 personality labels that are frequently used in several design books and a popular design blog [Gross 2015]. We ensure that each of these initial labels has at least appeared in two different sources. After manually clustering the labels with similar meaning (e.g., creative and unconventional), we have further reduced the list by keeping only the labels that have higher interest values in Google Trends (since 2004) and more relevant design results returned from Google image search. We have finally ended up with 15 personality labels, which are listed in Figure 2.

3.2 Poster Dataset

To build our poster dataset, we collect poster images with metadata (e.g., title, tags and descriptions) by using each of the 15 personality labels (e.g., "cute") and "poster" as the keyword to query the Google image search engine. We only keep the images in portrait layout, which is most commonly used in poster design. We then filter our initial poster dataset by removing: (1) duplicate images; (2) low-resolution images (less than 200 in width or 300 in height); (3) images without any tags that tell their personalities (to avoid selecting irrelevant posters, even though this would over-filter out some relevant posters). In addition, we only keep at most 5 visually similar posters from the same source to increase the diversity of our dataset. We regard two poster images as similar if the cosine distance between their 4096-dimensional features is smaller than 0.006. The features are taken from the penultimate layer of a 16-layer VGG [Simonyan and Zisserman 2014] pre-trained on the ImageNet [Deng et al. 2009]. All images are resized to 224×224 before feeding to the VGG. Finally, we manually discard images if: (1) they are photographs or drawings; (2) their personality labels appear as nouns on them, e.g., a person's name; (3) they contain



(a) (b) (c)

Fig. 3. Examples excluded from our poster dataset: (a) A photograph or drawing. (b) The personality label ("Cute") appears as a person's name. (c) More than one design in a single image ©Ingela Arrhenuis.

multiple designs on a single image. Figure 3 shows some examples that are excluded from our dataset. In total, we have collected 5,075 posters in our dataset.

3.3 Personality Label Assignment

For each poster in our dataset, we assign one or several labels to it from the set of 15 personality labels. To do this, we rely on the metadata associated with the posters, which are usually carefully chosen by the designers to describe the properties of the posters. A personality label is assigned to a poster if it appears in the metadata of the poster. A single poster can have more than one label, as it may possess different personalities. It is worth noting that all the labels are assigned automatically from the metadata, without any human intervention. Figure 2 shows the distribution of personality labels in our poster dataset.

4 SEMANTICS-AWARE DEEP RANKING NETWORK

4.1 Problem Formulation

Our goal is to learn a personality scoring function that, given a graphic design and a personality label, outputs a continuous score indicating the degree of the specified personality that the design possesses. With the binary personality labels in hand, a straightforward solution to this problem is to build a classifier to distinguish between different personalities for a given design, and use the class probabilities as the personality scores [Izadinia et al. 2015]. Unfortunately, since personalities are rather subjective, there are no clear boundary between different personalities (i.e., personality labels can be ambiguous), causing the classification formulation to be inadequate for our problem. As there are various degrees of a personality, we address our problem in a regression formulation.

To handle the problem that ground-truth personality scores are not available for training, we adopt a ranking formulation to learn our personality scoring function, by comparing pairs of posters according to their personality labels. Our hypothesis behind this design is that the comparison of design pairs in terms of their labels should be a more reliable supervision signal than using the label of a single design alone. We justify the advantage of our ranking framework over vanilla classification in Section 5.

Our ranking framework assumes that, given a personality label l, designs D^+ containing the label should be ranked higher than designs D^- without the label. Formally, given a pair of designs and a personality label (d^+, d^-, l), where $d^+ \in D^+$ and $d^- \in D^-$, we would like to learn a personality scoring function p that maps an input design d to its personality score $p_l(d)$, such that:

$$p_l(d^+) > p_l(d^-). \tag{1}$$

For this purpose, we build a semantics-aware deep ranking network, in which the personality scoring network is used to model p and trained on the web data collected in Section 3.

4.2 Network Architecture

Figure 4(a) illustrates the architecture of our semantics-aware deep ranking network. It consists of two personality scoring networks sharing the same weights. Each personality scoring network (Figure 4(b)) takes a personality label l and a single design d as input, and outputs the personality score for d. The personality label is encoded using a 1-of-K representation and fed into a semantic embedding network to obtain a semantic embedding vector $\mathcal{S}$, while design d is fed into a design feature network to extract a design feature vector $\mathcal{F}$. Finally, the design feature vector and the semantic embedding vector are concatenated and sent through a scoring network to predict a personality score p for design d.

4.2.1 Semantic Embedding Network. To enable end-to-end training for various personalities, we explicitly integrate a semantic embedding network into our deep ranking network to specify which label is currently used to rank input designs. Given a personality label l, we convert it into a one-hot vector, and then send it to a small subnetwork with 2 hidden Fully-Connected (FC) layers with REctified Linear Unit (ReLU) as activation function. Each FC layer has 64 units. Finally, we obtain a 64-dimensional semantic embedding vector $\mathcal{S}$.

4.2.2 Design Feature Network. This network is designed to extract features from the input design. Instead of using a pre-trained model such as VGG [Simonyan and Zisserman 2014], we design our own model since graphic designs have different visual characteristics from natural images. The basic block is a 3×3 convolutional layer with a ReLU activation function. For the convolutional layers, zero padding and 1×1 stride are used to keep the output size the same as the input. A 4×4 max pooling layer with a stride of 4×4 is added after each convolutional layer to reduce the spatial resolution of the feature maps. Finally, a FC layer is added at the end of the network to aggregate local features extracted by the convolutional layers into a global design feature vector $\mathcal{F}$. We set the the dimension of the design feature vector to 256, to balance the significance of semantic and design vectors. A dropout layer is used to prevent overfitting throughout the architecture.

4.2.3 Semantic Scoring Network. The design feature vector $\mathcal{F}$ and semantic embedding vector $\mathcal{S}$ are concatenated and fed into a 3-layer Multi-Layer Perceptron (MLP) network to predict the personality score. The numbers of units for the first two hidden layers are 256 and 128, respectively. ReLU and dropout are used in both layers. The final output layer has one unit for personality score.

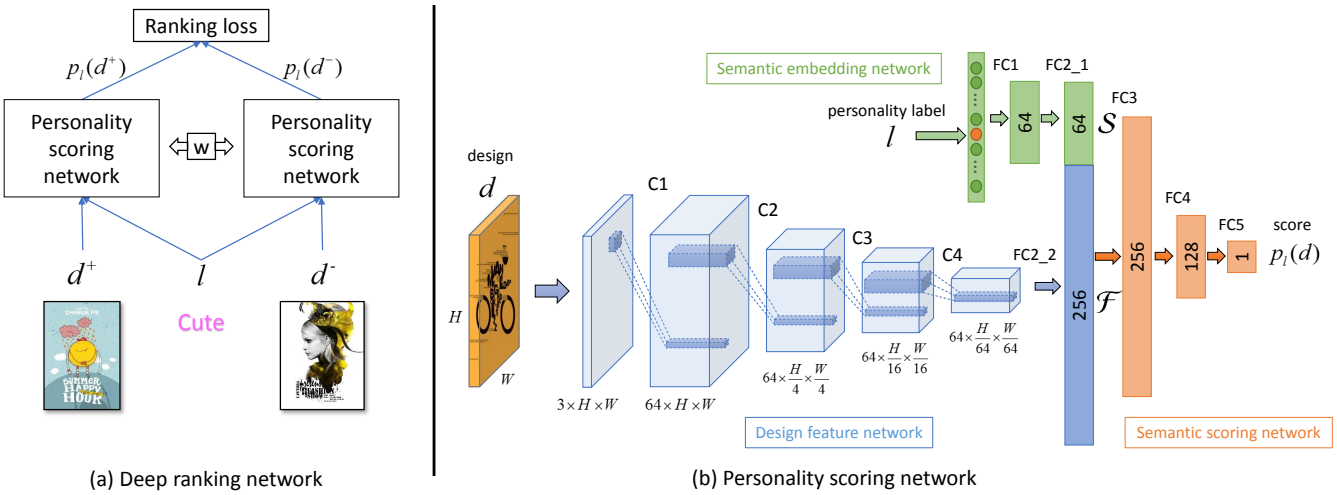


Fig. 4. The architecture of our semantics-aware deep ranking network (a). It comprises two personality scoring networks with shared weights. For each personality scoring network (b), given a design, a design feature vector $\mathcal{F}$ is first extracted using a design feature network. The personality label is encoded by a semantic embedding network as a semantic embedding vector $\mathcal{S}$. The design feature vector and semantic embedding vector are then concatenated and fed into a scoring network to output a personality score. (C - convolution layer and FC - fully connected layer) ©Charlie P's and DEVIANT ART user giraff-e.

4.3 Loss Function

We define the following hinge loss for a pair of designs (d^+, d^-) and a personality label l:

$$H_l(d^+, d^-) = max(0, m - p_l(d^+) + p_l(d^-)), \tag{2}$$

where m is a margin hyperparameter. This hinge loss imposes a ranking constraint to enforce that d^+ (with l) scores higher than d^- (without l) by a margin m.

Our final loss function over the training dataset $\mathcal{D}$ is:

$$\mathcal{L}(l, \mathcal{D}, \mathcal{W}) = \sum_{l\in L}\sum_{d^+\in D^+}\sum_{d^-\in D^-} H_l(d^+, d^-) + \lambda \|\mathcal{W}\|_2^2, \tag{3}$$

where $\|\mathcal{W}\|_2^2$ is the l_2 norm regularizer to prevent overfitting. λ is a regularization parameter. $\mathcal{W}$ are the parameters of the personality scoring function p. Eq. 3 computes the total loss over $\mathcal{D}$.

4.4 Training

To train our network, we scale all input posters to a resolution of 300×200. The final network has a total of 429,825 parameters, and is trained end-to-end using standard backpropagation. We use the ADADELTA optimizer [Zeiler 2012] for optimization, which sets the learning rate adaptively. We use a dropout rate of 0.5 for all the dropout layers in the design feature network and the semantic scoring network. The weight decay λ is set to 0.005. We randomly split the poster dataset with a train-to-validation ratio of 9:1. For each personality label l, we obtain the training design pairs by randomly sampling 170 positive designs (with l) and 3,400 negative designs (without l) from the training set. We then combine them exhaustively to balance the positive-negative pairs. In total, we obtain 8.67M triplets (positive design, negative design, label) for training. The validation triplets are generated in the same way with 20 positive designs and 400 negative designs. We train the network using a batch size of 32 for 300,000 iterations, which takes about 124 hours on a PC with an i7 3GHz CPU, 24GB RAM and a Titan X GPU.

5 MODEL EVALUATION

To evaluate the effectiveness of our semantics-aware deep ranking network, we compare it with several baselines.

5.1 Evaluation Dataset

To evaluate our model, we have collected a new set of pairwise ranking data from 615 posters to form an evaluation dataset. To obtain more reliable ground truth rankings, we use crowdsourcing similar to previous works [Garces et al. 2014; O'Donovan et al. 2014b]. For our evaluation dataset, there are about 189,000 comparisons for each personality, in which many of them are non-discriminative. Since having a large percentage of non-discriminative queries in a questionnaire may have a negative effect on the quality of the results [Lun et al. 2015], we first asked 3 professional designers to give a score from 1 (strongly disagree) to 5 (strongly agree) on each of the 615 posters to indicate if they agree that the poster has each of the 15 personalities. When generating pairs for comparison, those pairs with a larger score difference would have a higher chance to be selected. Finally, we selected 2,000 pairs for each personality, resulting in a total of 30,000 comparisons.

We asked the workers in Amazon Mechanical Turk (AMT) to rank each pair of posters given a personality, by answering the questions like "which poster is more cute?" in two-alternative forced choice (2AFC) manner (i.e., we only allow them to choose one or the other, in order to better measure small differences [O'Donovan et al. 2014b]). Each HIT consists of 40 different comparisons for a single personality. We further duplicated five randomly chosen questions by swapping the presentation order of posters in each pair to reject unreliable workers. Workers need to answer four of them consistently for us to accept their data. Our study involved a total

of 1,281 workers. (56.4% were female. 11.5% claimed to be design experts.) 18.8% of HITs were rejected. We show the screenshot of the AMT study in S1.1 of the supplemental.

For evaluation, we discard the rankings with high disagreement (less than 70% agreement). For each remaining ranking, we take the majority of votes as the ground truth. A ranking prediction is regarded as correct if the rank obtained by the two predicted scores agrees with the ground truth.

5.2 Baselines

We compare our deep ranking method (**Ours**) to an existing learn-to-rank method (**RankSVM**) trained on various features, and classification-based approaches.

5.2.1 RankSVM. It is commonly used to learn relative attributes [Parikh and Grauman 2011]. Since RankSVM needs to be trained separately to compute the weight vectors for each personality label, we train a total of 15 RankSVM models for the 15 personalities, using different features, including hand-crafted and deep features. For the hand-crafted features, following the work on measuring style similarity between infographics [Saleh et al. 2015], we use color and HOG. For deep features, we use the deep features from VGG and our model.

- **Color and luminance histogram (*Color*).** We compute color and luminance histogram features, with 10 bins for each of RGB color and luminance channels.
- **HOG.** We extract HOG features with a cell size of 16 [Dalal and Triggs 2005], and use PCA to reduce the dimensionality of HOG to 230.
- **VGG features (*VGG*).** We extract 4096-dimensional feature from the penultimate layer of a 16-layer VGG pre-trained on the ImageNet [Deng et al. 2009]. All posters are resized to 224×224 before feeding to VGG.
- **Our deep feature ($\mathcal{F}$).** We use the output of the last hidden layer of our design feature network as deep features.

5.2.2 Binary Classification (Multiple). For each personality label, we train a vanilla binary classification network separately to predict the presence of this label. The architecture is almost the same as our personality scoring network except that the semantic embedding network is removed and a sigmoid activation function is used in the last layer. We use cross-entropy loss to train each network. To balance positive and negative posters, we sample the same number of positive and negative posters in each mini-batch. For fair comparison, all other settings are kept the same as our deep ranking network. During test time, the output probabilities are used as the personality scores. In other words, a higher probability means a higher score.

5.2.3 Binary Classification (Single). We train an end-to-end binary classification network, which takes as input a poster and a personality label and predicts if the poster possesses the personality. The architecture is same as our personality scoring network, except that the last layer is replaced by a sigmoid layer. The cross-entropy loss is used during training. In each mini batch, we sample the same number of positive and negative posters as in Binary classification (Multiple) above.

5.2.4 Multi-label Classification. We train a network to predict the presence of all personalities given a poster. We use the design feature network, followed by the semantic scoring network in the multi-label classification architecture. The sigmoid activation function is used in the last layer with 15 units, each of which predicts the presence of a personality. We use binary cross-entropy as the loss function. As the numbers of posters with different personalities are quite different, we set the weights to be inversely proportional to their label counts in the loss function.

5.3 Results

Table 1 shows the accuracy of different methods on our evaluation dataset. Our deep ranking method outperforms all other methods by a large margin, particularly for the personalities such as "Artistic", "Minimalist" and "Romantic". It is interesting to note that RankSVM performs better using our design feature vector than both the hand-crafted features and pre-trained VGG features in most cases. This implies that our model can learn a high-quality discriminative design representation. The discrimination power of our design features also helps RankSVM converge faster (~200x faster than using ***Color***).

5.3.1 The Role of the Semantic Embedding Network. Multi-label classification shows a relatively high performance on some personalities such as "Cute" and "Playful", but the average performance is low. Instead, with the semantic embedding network, both binary classification and our model can learn better discriminative, semantics-aware feature representation for different personalities, resulting in higher average performances. Binary classification (Single) has slightly worse performance than Binary classification (Multiple). However, training multiple classifiers separately would complicates the modeling process and increase the number of hyper-parameters to tune, in comparison to a single, unified model enabled by the semantic embedding network. These confirm the benefits of explicitly modeling semantic information using a network in our problem.

5.3.2 The Role of the Ranking Formulation. The major difference between our model and binary classification is the choice of the loss formulation. Our model outperforms binary classification, which justifies the advantage of our ranking formulation in the problem of presonality score prediction.

5.3.3 Human Performance. We report the human accuracy (***Human***), as an upper bound performance, on our evaluation dataset (agreement > 70%) in Table 1. For a given pair, a human choice is considered as correct if it agrees with the ground-truth (i.e., majority). The accuracy can also be regarded as human consistency. An average consistency of 94.36% on the evaluation dataset (90.67% over all the crowdsourced rankings) indicates that humans are fairly consistent in evaluating the personalities of graphic designs. For the personalities with higher human consistency, the compared designs often differ greatly in terms of certain features, resulting in a higher model prediction accuracy. However, some of the personalities, such as "Creative", "Elegant", "Modern", and "Playful", show lower human consistencies, resulting in a lower model prediction accuracy. One possible reason is that, compared with others, these personalities are more subjective and can convey different meanings under different context. For example, "Modern" can refer to an 80-year-old

Table 1. Ranking prediction accuracy (percentage of correctly predicted rankings) of different methods on different personality labels. The highest accuracy of each row is highlighted in bold and underlined.

Personality	Accuracy (%)								
	RankSVM (*Color*)	RankSVM (*Color+HOG*)	RankSVM (*VGG*)	RankSVM (Our $\mathcal{F}$)	Multi-label Classif.	Binary Classif. (Single)	Binary Classif. (Multiple)	Ours	Human
Artistic	55.82	56.86	57.89	75.73	53.07	63.51	49.50	**76.71**	93.46
Creative	59.69	56.52	62.05	71.18	51.93	50.19	51.95	**72.55**	92.08
Cute	81.59	66.34	**85.89**	77.02	84.37	83.33	83.65	79.14	94.08
Dynamic	53.41	59.91	67.09	76.34	50.24	75.00	77.00	**79.81**	92.91
Elegant	60.01	54.85	58.35	65.88	59.95	60.97	53.25	**67.54**	90.33
Energetic	50.05	58.94	73.49	76.67	48.71	77.16	79.3	**81.47**	95.43
Fashion	66.81	60.11	69.47	79.72	58.89	73.68	67.95	**86.26**	95.19
Fresh	85.63	70.15	75.86	77.92	71.05	80.19	**90.20**	84.36	96.47
Futuristic	78.85	69.05	80.02	84.76	74.21	73.79	83.65	**86.20**	96.18
Minimalist	59.11	79.87	93.68	89.88	50.53	76.98	85.65	**96.05**	96.56
Modern	58.63	61.20	62.33	63.38	56.31	64.04	58.20	**66.03**	91.28
Mysterious	80.26	75.33	78.80	75.16	78.09	77.28	82.10	**84.06**	94.33
Playful	70.11	66.71	63.90	69.75	**71.49**	67.66	64.90	69.04	93.25
Romantic	65.40	73.28	92.63	88.43	51.16	62.37	78.40	**92.47**	98.66
Vintage	71.65	77.11	83.84	87.60	44.46	72.92	84.70	**92.32**	95.16
Average	66.47	65.75	73.69	77.30	60.30	70.60	72.69	**80.93**	94.36

innovation that is still in use today or a cutting-edge contemporary design. How to interpret the meaning of these personalities depends upon personal knowledge and experiences, which could vary across different participants.

5.3.4 Visualization of the Rankings. Figure 5 shows top 3 and bottom 3 designs ranked using the personality scores predicted by our model on the evaluation dataset. In each case, our model gives higher ranks to designs with higher degrees on a specific personality. Here, we only show results for four personality labels. Refer to S1 of the supplemental for results for other personality labels.

5.3.5 Additional Results. To evaluate the effectiveness of our learned deep features, we visualize them using t-SNE [Maaten and Hinton 2008] and perform personality-based design retrieval. To demonstrate the generality of our framework, we have also trained our deep ranking network on another popular type of graphic design, webpages. Refer to S1 of the supplemental for more details.

6 MODEL-BASED ANALYSIS

To understand what visual features of graphic designs help distinguish different personalities, we perform qualitative and quantitative analyses with the learned personality scoring network across different levels, from pixel, region to element. For pixel- and region-level analyses, we build a personality-based sensitive map, revealing the locations that are important for conveying a given design personality. For element-level analysis, we conduct several experiments to find out how the perceived design personality is correlated with key design factors including color, font, and space.

6.1 Sensitive Map

We build a personality-aware sensitive map based on the method in [Zeiler and Fergus 2014], which was proposed to visualize how different locations of an image affect its classification. We aim at visualizing how different locations of a design affect its personality. We first slide an occluding window (48×48 in our setting) across a design. At each location, we replace all pixels within the window with the mean color of the design, and send the resulting design to our personality scoring network to generate a new personality score. We then subtract the new score from the score of the original design as the pixel value of the window center in the sensitive map. In this way, we obtain a pixel-wise sensitive map that shows the areas of a graphic design that contribute most (positive values) or least (negative values) to a specific personality. The sensitive map also allows us to mine discriminative patches from a collection of designs for a particular personality. In particular, for a given personality, we first rank designs in our dataset based on their personality scores in descending order. For each of the top-ranked designs, we then take the image patch whose absence causes the greatest drop in personality score, to form a set of discriminative patches.

Figure 6 shows some results. For each personality, the left column shows two posters with their sensitive maps, while the right column shows the discriminative patches mined from our dataset. These results exhibit some interesting patterns. For example, "Artistic" posters tend to have more decorative patterns, and "Romantic" posters make extensive use of intimate interaction between men and women. However, the discriminative patches extracted for "Creative" do not show obvious patterns. This is because a lot of "Creative" posters rely on clever composition of common objects to convey the sense of creativity. For example, the lower poster under "Creative" in Figure 6 conveys creativity by placing a blade under the rock.

Fig. 5. Comparison of the top ranked and bottom ranked designs predicted by our model for different personalities. For each personality, the top ranked designs are outlined in green boxes, while the bottom ranked designs are outlined in orange boxes. ©Ryan Swanson, Daniel Danger, 20th Century Fox, Artist Posters Collection at the Library of Congress, Walt Disney Animation Studios, Balinisteanu Iulian, 2007 Chris Diston, gggrafik design, RVCA, Warner Bros. Pictures, Dawn Hudson, Paramount Pictures, Albert Exergian, Krzysztof Iwański, Brunner, Eiichiro Oda, DEVIANT ART user PheoniX-VII and cstm.

We have compared our sensitive map with the importance map for measuring the importance of elements in a design [Bylinskii et al. 2017]. The importance map is computed using a fully convolutional network trained on human-labeled data. Figure 7 shows that not all important areas in a poster contribute to a target personality. For example, the poster for "Minimalist" focuses more on negative spaces. Any fancy elements may decrease the feeling of "Minimalist" in a design. For "Romantic", the interaction between lovers is a key element, rather than the text that the importance map detects as salient. Our sensitive map can intuitively tell which part of a design is important or detrimental to conveying a particular personality. As such, it can be used by graphic designers to iteratively improve their designs, as well as serve as a building block for some personality-aware graphic design systems, such as element-level design suggestion introduced in Section 7.1. The sensitive maps and discriminative patches of other personalities can be found in S2 of the supplemental.

6.2 Design Factors on Personality

A proper setting of design factors (e.g., color and font) is essential to a design, as they can draw user attention, set the mood, affect the personality, and even influence user decisions. Color, font and space are among the most important design factors commonly used by designers to manipulate the personality and mood of a design [Cousins 2015; Kliever 2015]. Thus, we study how these three factors affect design personality. Since the focus of our paper is on graphic design, we mainly study some design-specific factors. There are still many other factors, such as image contents, that may have huge effects on the personality of a design, which is an interesting topic for future work.

6.2.1 Color. We have shown in Section 5 that color plays an indispensable role in personality prediction. To find out how the change of color affects the personality score, we compute the Spearman rank correlation (ρ) between the mean of each HSV color channel and the personality score on our training set. Results show that Mean Hue and Saturation have a weak correlation with personality in general. While Mean Hue has the highest absolute correlation score on "Mysterious" ($\rho = 0.1861$), Mean Saturation has the highest absolute correlation score on "Fashion" ($\rho = -0.2566$). In contrast, Mean Value shows the stronger correlation with personality. Figure 8 shows the color distribution of designs under three personalities with strong correlation. Linear least-square fitting lines are added to highlight the correlation trends. To find out the preferred color for a personality, we quantize each channel in the HSV color space into 20 bins and run the linear RankSVM on them individually. The weights are displayed as a color stripe on top of each diagram, aligned with the values on the x-axis. For a value range, more yellow indicates a more positive effect, while more red indicates a more negative effect. For example, the stripe on Mean Hue of "Fresh" goes from orange to yellow then red, while the actual Mean Hue transitions from red to green to blue. This suggests that green can help a design gain a higher score for "Fresh" than red and blue.

If we take a closer look at Figure 8, we can see that warm colors (e.g., red to yellow to green) and pink-magenta play an important role in a design to exhibit "Cute". For "Fresh", as expected, orange and green dominate in the designs. Besides, the increasing trend of the least-square fitting lines for Mean Value in "Cute" and "Fresh" indicates that they prefer bright colors. This agrees with our common sense that a design will appear to be more vivid and vibrant with brighter colors. "Futuristic" shows a strong preference on darker and colder colors (e.g., blue and purple). This is because the future is often mysterious. This observation is consistent with a previous study [Shedroff and Noessel 2012] on color of future screen used in science fiction for decades. They conjecture that blue is chosen because of its rareness in nature. The correlation plots of other personalities can be found in S3.1 of the supplemental.

6.2.2 Font. The choice of font plays an important role in setting the mood and increasing visual interest of a graphic design. To study the type of fonts that a particular design personality prefers, we use 200 fonts, $\mathbf{F} = \{f_i\}_{i=1}^{200}$ [O'Donovan et al. 2014b], and select 15 posters, $\mathbf{D} = \{D_j\}_{j=1}^{15}$, from our dataset. Each of the selected posters has a few text elements. None of them have complex visual contents, so as to emphasize on the font. For each design D_j, we

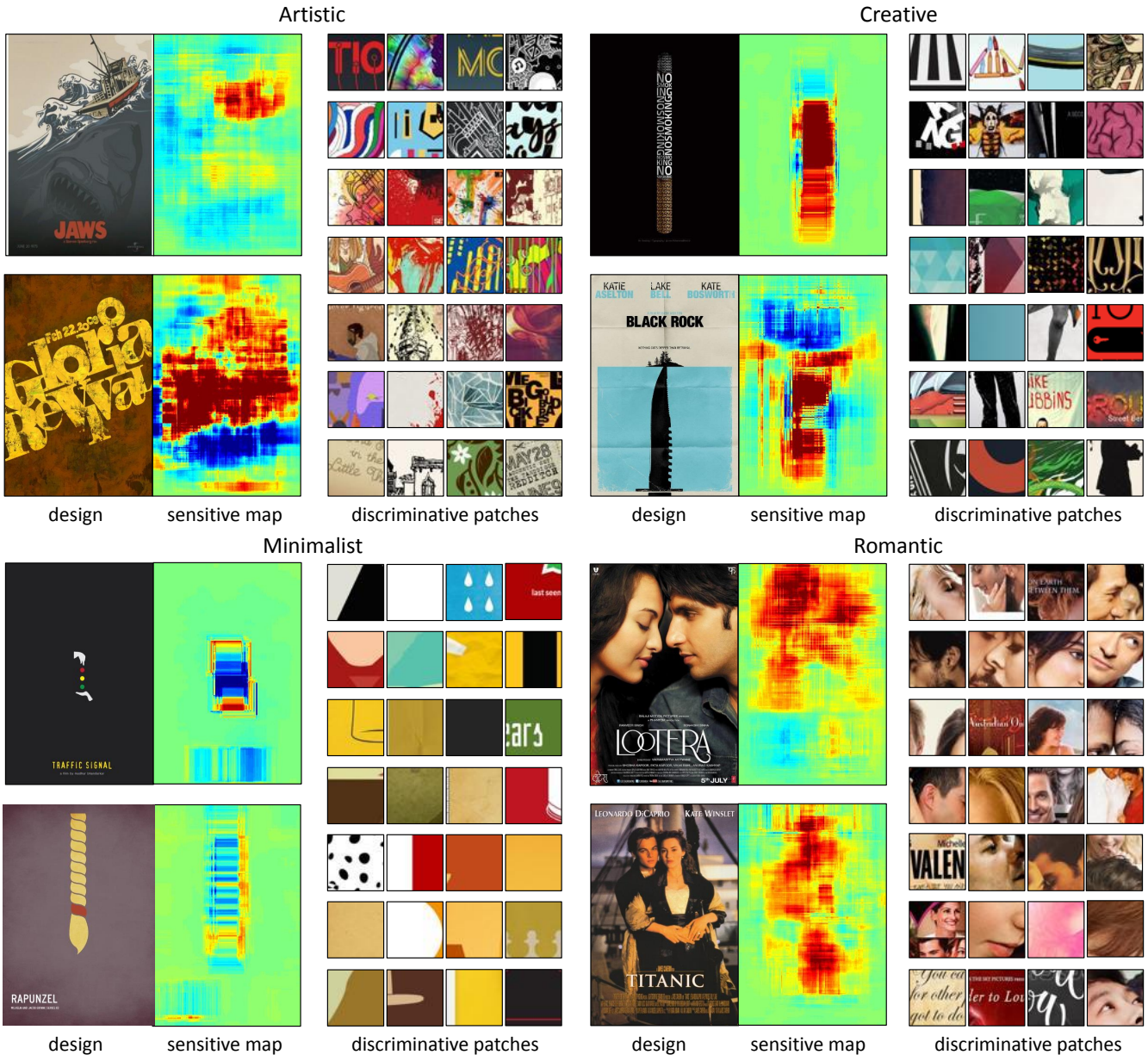


Fig. 6. Sensitive maps and discriminative patches for various personalities. For each personality, input designs and their sensitive maps are shown on the left. A sensitive map is used to show the locations on a design with positive (red) or negative (blue) impacts to a given personality. The discriminative patches are shown on the right, which are extracted from our poster dataset to characterize a given personality. ©Universal Pictures, Dennis Cho, LD Entertainment, Christian Jackson, Percept Picture Company, Paramount Pictures and Balaji Motion Pictures.

replace the original font of its largest text element with each of the fonts in $\mathbf{F}$, resulting in 200 new designs, $\mathbf{D}_j^{\mathbf{F}} = \{D_j^{f_i}\}_{f_i \in \mathbf{F}}$. Given a personality, we apply our model to predict the personality scores for $\{\mathbf{D}_j^{\mathbf{F}}, j = 1, \ldots, 15\}$, which are then normalized using the minimum and maximum values from our dataset. We next compute an average score for each font from the average personality score of the 15 new designs (for the 15 personality labels) that use the font. The fonts with higher average scores are considered as more important to the personality, and vice versa. Figure 9 shows the top 5 and bottom 5 fonts for three representative personalities. We can see that for each personality, the most and least important fonts exhibit distinctive visual properties. In particular, for "Elegant", the most important

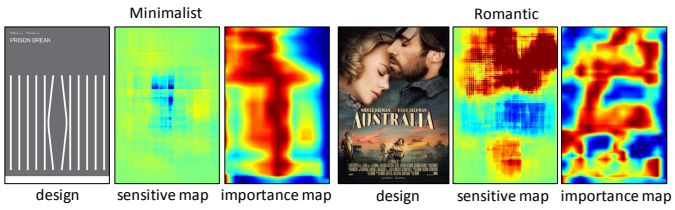


Fig. 7. Comparison of our sensitive map and the importance map [Bylinskii et al. 2017]. For each design, our sensitive maps can properly find the areas that contribute positively/negatively to a given personality. ©Albert Exergian and 20th Century Fox.

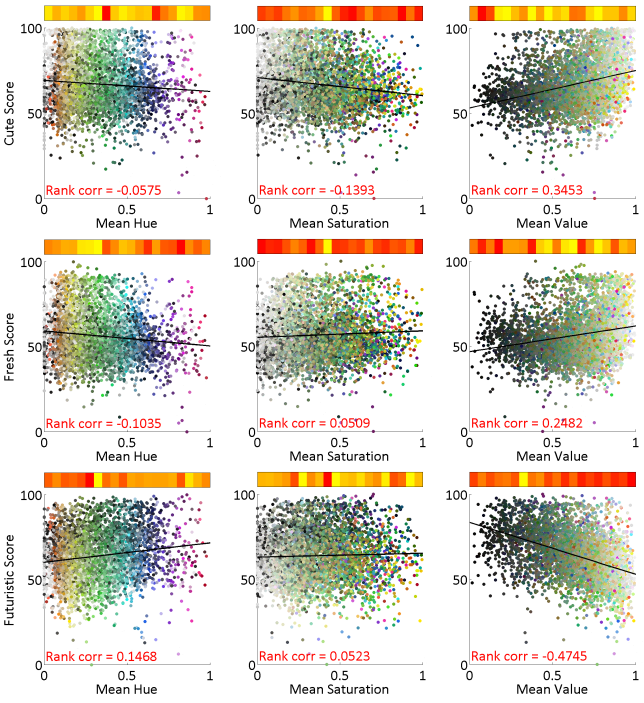


Fig. 8. Correlation between HSV colors and three personalities: "Cute", "Fresh" and "Futuristic". In each diagram, a dot corresponds to one design represented by its mean color. A linear least-square fitting line (in black) is also shown, along with the Spearman rank correlation at the bottom-left corner. The weights of the linear RankSVM model are shown as a stripe above the diagram, where a block being more yellow (or red) indicates that the corresponding value range on the x-axis has a more positive (or negative) effect to a specific personality.

fonts are more cursive, while the least important fonts are wider. "Fresh" prefers serif and thin fonts over bold fonts. For "Romantic", the top 5 fonts are more cursive and wider than the bottom ones.

We look further into how the font properties may affect the perceived personality of a design. To this end, we consider 7 concrete font properties used in [O'Donovan et al. 2014b], including "angular", "cursive", "italic", "serif", "sharp", "thin", and "wide". Given a font property and a personality, we compute a rank correlation for each design D_i between the font property scores of $\mathbf{F}$ by [O'Donovan et al. 2014b] and the personality scores of $\mathbf{D}_j^{\mathbf{F}}$. Figure 10 shows the average rank correlation coefficients[1] across different designs on three font properties. We find that the weight-related font properties, e.g., "wide" and "thin", have a higher influence upon design personality, while "cursive" has relatively less effect. More specifically, "Cute" and "Vintage" designs show a stronger preference on "wide" fonts, whereas "Elegant" and "Minimalist" designs tend to use "thin" fonts. Refer to S3.2 of the supplemental for the results of other font properties.

Fig. 9. Comparison of the most important and least important fonts for different personalities. For each personality, the most important fonts are shown at the top row, while the least important fonts are at the bottom row.

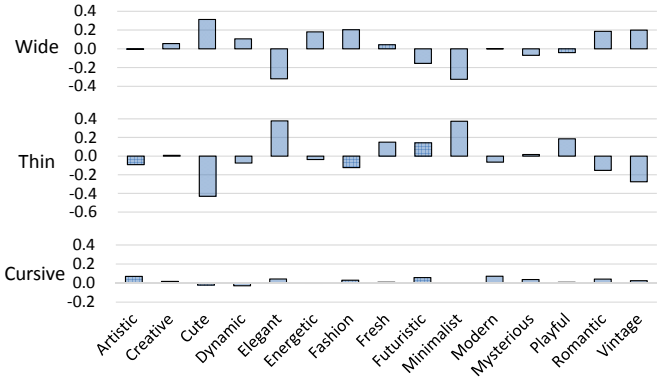


Fig. 10. Correlation between font properties and graphic design personalities. For each font property ("Wide", "Thin" or "Cursive"), the bars denote the average rank correlation coefficients between the property and individual personalities.

6.2.3 Space (or Negative Space). Negative space is an essential element in a design to give readers a visual break [Cousins 2015; Kliever 2015]. We would like to explore the relationship between space and personality. Here, we examine two important factors: (1) percentage of negative space, and (2) symmetry of negative space. To estimate the amount of negative space more accurately, we manually labeled 231 designs selected randomly from our dataset, and draw a mask to indicate the negative space of each design. We then compute the percentage of negative space and symmetry of negative space for each design. To measure the symmetry of negative space, we calculate the percentage of negative space (in pixels) with

[1]To aggregate rank correlation coefficients, we first convert rank correlation coefficients to Fisher's z coefficients, which are then averaged and converted back to rank correlation coefficients [Silver and Dunlap 1987].

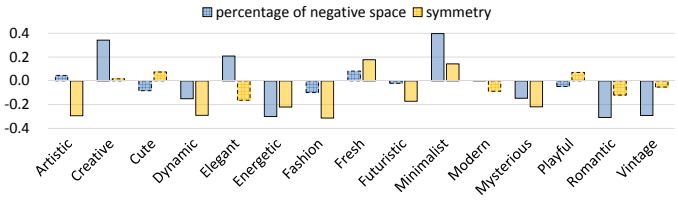


Fig. 11. Spearman rank correlation of personality with percentage of negative space (blue bar) and symmetry of negative space (yellow bar). The bars with dotted outline are not statistically significant ($p > 0.05$).

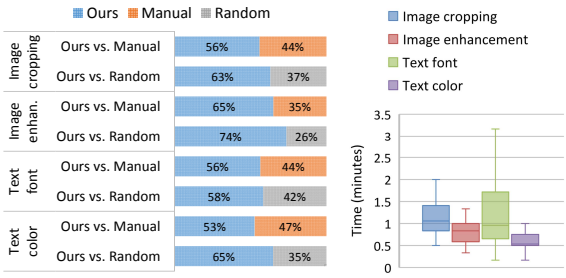


Fig. 12. Results of the user study on element-level design suggestion. Left: For each task, we compare our method (Ours) against manual method (Manual) and random method (Random) using 2AFC pairwise comparisons, and show the percentage of preferred votes by participants. Right: Average time that participants spent on each task.

symmetrical counterparts about the x-axis, y-axis, or center of the design. Figure 11 shows the results. We can see that the percentage of negative space can affect the personality of a design. For example, a design tends to be more "Minimalist" or "Creative" as the amount of negative space increases, while it is more "Energetic" as the amount of negative space decreases. The symmetry of negative space has negative effect on most personalities. For example, a design tends to be more "Dynamic" as the negative space becomes less symmetrical. This finding agrees with the design principles used by artists [Bradley 2010].

7 APPLICATIONS

Given our network for predicting design personality scores and the high-level features learned by our network, we have explored two novel applications: element-level design suggestion and example-based personality transfer.

7.1 Element-level Design Suggestion

The personality of a design is usually achieved by exploring complex property space of basic elements (e.g., images and texts) on it. It would therefore be very useful to provide designers with guidance on how to modify the properties of design elements in order to enhance a specific design personality. To this end, we present an interactive suggestion approach to allow designers to set design element properties to improve a specified personality of design.

7.1.1 Algorithm. Given an input design, the user first selects a personality and a target design element. Our method will then suggest a list of property values for the target element, ordered by their likelihood of enhancing the selected personality. Currently, our method supports four types of basic yet important properties for images/texts: (1) *image cropping*: crops a target image to fit a user-specified cropping region; (2) *image enhancement*: enhances a target image by changing its sharpness, brightness, contrast, and saturation; (3) *text font*: selects a font for some target texts; and (4) *text color*: selects a color for some target texts.

Our method works in a brute-force manner by listing possible property values and ranking the designs with the modified values according to the personality scores predicted by our model. The scores for each personality are normalized using the minimum and maximum values from our dataset. To make our method tractable, for font selection, we consider 300 fonts randomly selected from Google Fonts. For other properties with a large search space, we discretize the search space into a set of candidate values and adopt an iterative sampling approach for computational efficiency. Refer to S4.1 of the supplemental for details.

For the resulting designs to remain visually pleasing and functionally valid, we impose some constraints on text properties. Specifically, we exclude the colors that will make text look similar to its surrounding colors, which will certainly impair its visibility. We also adjust the size of the text to ensure that it fits within the design border during font selection.

7.1.2 Evaluation. We present several user studies to evaluate the quality of the results generated by our method.

Comparison to baselines. We first compare our results (**Ours**) with those generated by novices manually (**Manual**) and selected randomly (**Random**). For each of our 15 personalities, we create 6 design cases, resulting in a total of 90 design cases. Each design case is assigned with one of the four tasks: text font selection, text color selection, image cropping and image enhancement. For each design case, given a personality, a design element and one of the four tasks, we generate 3 suggestions by each of the three methods: (1) Ours: we use top 3 designs suggested by our method. (2) Manual: we recruited 18 graduate students with no prior training on graphic design. PowerPoint was chosen as the editing tool, as they were all familiar with it. Each participant was required to complete 15 design cases, one for each of the 15 personalities. Each design case was performed by 3 different participants. (3) Random: we randomly sample 3 suggestions from the whole search space uniformly.

To evaluate the designs from the three methods, we asked AMT workers to compare the results of Ours against those of Manual and Random through pairwise comparisons in a 2AFC manner. For each design case, we create 9 pairs of results by Ours and Manual and 9 pairs by Ours and Random, resulting in a total of 1,620 unique comparisons. Each comparison was evaluated by 6 different workers. Refer to S4.2 of the supplemental for more details.

Figure 12 summarizes the results. We can see that our results are significantly better than those of the other two methods for image cropping, image enhancement and font selection tasks ($p < 0.05$, chi-squared test). For text color selection, while our method has higher preference than the manual method, the preference is only marginally significant ($p = 0.078$, chi-squared test). This may be

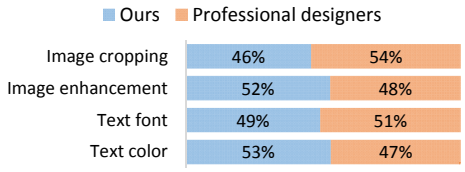


Fig. 13. Results of pairwise comparison of designs by ours and professional designers. For each task, we show the percentage of preferred votes by participants for each method. Overall, our results are comparable to those of the professional designers (all preferences are not statistically significant, according to a chi-squared test with $p > 0.05$).

because selecting a color for some text is a relatively easy task for human, compared with the other three tasks. We show some results by our and manual methods for the four tasks in Figures 14 and 15. Refer to S4.3 of the supplemental for more results.

We also report the time taken by the participants (Manual) on different tasks in Figure 12 (right). The image cropping and text font selection tasks often require a subtle trial-and-error process, and took a much longer time for human to complete (1 minute on average). The other two tasks took more than 0.5 minute. In contrast, our method completes each of the tasks using only one or a few seconds, depending on the complexity of the search space, e.g., image resolution for image cropping. The results above show that our method can help select proper element–wise properties to better convey design personality in significantly less time.

Comparison to professional designers. We also would like to know how our results compare with those by professional designers. We therefore conduct another experiment on the same four tasks as before, with 36 design cases. For text font and color selection, we collected 18 designs with explicit personality tags from two graphic design websites (Canva and Freepik), and separated the design elements into different editable layers. For each of these designs, we then manually selected one of the major text blocks with only one color but without any special effects (e.g., shadow) as the target text element. Note that for the collected designs, the font and color of the target text elements are known, which are regarded as being created by professional designers. For image cropping and enhancement, since the original images used to create the collected designs are typically not available, we manually created another 18 different design cases. For these 18 design cases, we recruited 3 professional graphic designers. Each of them was asked to edit 6 design cases. They were allowed to use any editing tools, e.g., Adobe Photoshop. All 15 personalities were involved in these 36 design cases. The experiment was performed in the same way as in the previous study (i.e., Ours vs. Manual), except that the candidate font set is extended to the all Google Fonts to allow our model to have more diverse choices.

For evaluation, we run a perceptual study on AMT, where 50 participants were asked to perform pairwise comparisons of the results by our method and professional designers. Each participant evaluated all the design cases, plus 5 duplicated cases for consistency check. As shown in Figure 13, our method is on par with the designers on all the tasks. Qualitative comparisons can be found in S4.3 of the supplemental.

7.2 Example-based Personality Transfer

Designing with examples is a common way used by most people, especially novices, during a design process, by referencing some design examples on the use of design elements, setting of their properties, and laying out the elements. However, transferring the personality across designs is non-trivial, as it is often difficult to find examples of a specific personality. As the design feature representation learned by our network is personality-aware (see Section 5), the distance between two designs in the feature space reflects their personality similarity. Hence, we explore the use of our design feature representation for personality transfer.

7.2.1 Algorithm. Given a reference design and a source design, we aim to transfer the personality of the reference design to the source design, by adjusting the properties (i.e., element size / position, text font, image sharpness / brightness / contrast / saturation) in the source design. In addition, we also allow users to specify some constraints (e.g., changeable elements, relative importance of elements and if overlapping between elements is allowed), and our output design will conform with these constraints. We formulate this adjustment as a constrained optimization problem. Formally, let $\mathcal{E}$ and $\mathcal{X}_\theta$ be the reference and source designs, respectively. We denote the θ as a vector encoding property configurations of all elements of the source design. For the user-specified constraints, we express them as equality/inequality relations that are abstracted as hard constraints: $\mathcal{H}(\theta) = 0$. Our personality transfer is achieved by solving the following optimization problem:

$$\arg\min_{\theta} \alpha_1 E_{trans}(\mathcal{E}, \mathcal{X}_\theta) + \alpha_2 E_{prior}(\theta) \quad \text{s.t. } \mathcal{H}(\theta) = 0, \tag{4}$$

where $E_{prior}(\theta)$ is a prior term to encourage the source design to conform with user-specified design guidelines. (For simplicity, we only use layout balance here.) $E_{trans}(\mathcal{E}, \mathcal{X}_\theta)$ is a transfer term to force the source and reference designs to be close in the design feature space $\mathcal{F}$:

$$E_{trans}(\mathcal{E}, \mathcal{X}_\theta) = d(\mathcal{F}_\mathcal{E}, \mathcal{F}_{\mathcal{X}_\theta}), \tag{5}$$

where d is a cosine distance function. Since the objective function is highly non-linear and multi-modal, we optimize it using the Metropolis-Hastings algorithm of the Markov Chain Monte Carlo (MCMC) method [Hastings 1970; Metropolis et al. 1953] similar to [Merrell et al. 2011] to efficiently explore the solution space. Refer to S5 of the supplemental for the algorithmic details.

7.2.2 Evaluation. Figure 16 shows two example results. For the top example, we can see that the background image and text elements are adjusted to produce a more "Vintage" design. For the bottom example, the rearrangement of the elements results in a more "Dynamic" design. More results can be found in S5.4 of the supplemental.

We evaluate the effectiveness of our method in a user study, where we compare our results against the source designs and the results by novices manually. We created 15 design cases for our study. For novice results, we recruited 15 participants with little experience

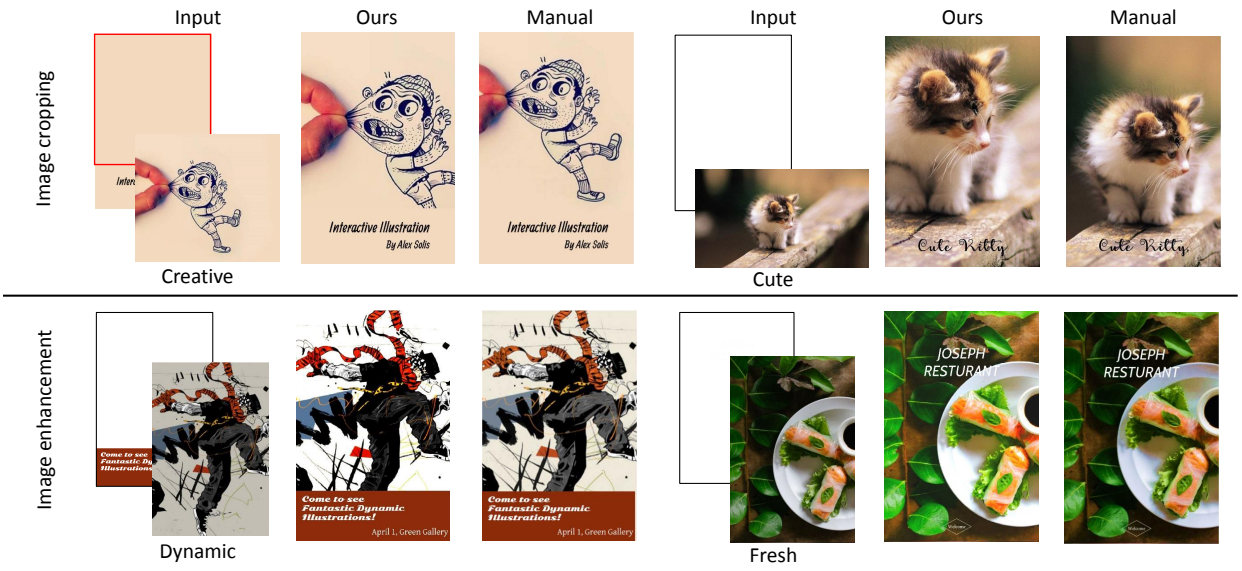


Fig. 14. Element-level design suggestion results on image cropping and image enhancement. For each design case, we show the input design, image and personality on the left. Top results by our method (Ours) and results by novices (Manual) are shown on the right.

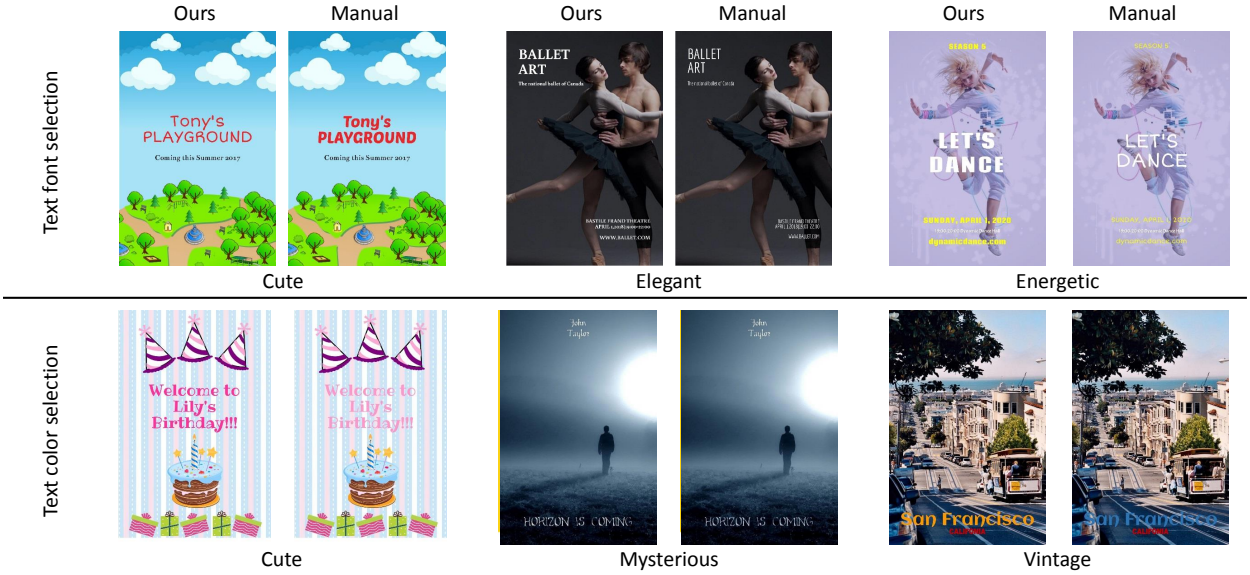


Fig. 15. Element-level design suggestion results on text font and color selection. For each design case, we show the top results by our method (Ours) and results by the manual method (Manual).

on graphic design from a local university. For each design case, given a reference design and a source design, the participants were instructed to edit the source by changing its design elements, in order to match its personality with that of the reference. For a fair comparison, the participants were only allowed to change the element properties that our optimizer operates on. Each design case was done by 5 different participants.

We use AMT to evaluate the results. For each design case, we display a reference design, along with three candidate designs: a source design (Original), the results by our method (Ours) and novices (Manual). AMT workers are asked to select which of the three candidates is more similar to the reference in terms of a given personality and which one is more visually pleasing. There were a total of 75 comparisons, each of which was evaluated by 30 different workers. The results are shown in Figure 17. Our results are significantly better ($p < 0.05$, chi-squared test) than those by the other two, on both personality similarity and visual aesthetics. This demonstrates that

Fig. 16. Example-based personality transfer. Given a reference design and a source design, our method can generate a plausible new design that matches the personality of the reference design. ©Vintagetravel and Estuary Clinic.

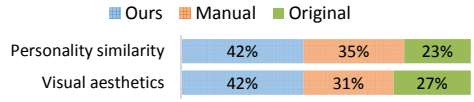


Fig. 17. Results of the user study on example-based personality transfer. We show the percentage of preference votes for the source designs (Original), results by our method (Ours) and novices manually (Manual), in terms of personality similarity and visual aesthetics. Ours is preferred over Original and Manual significantly in both aspects ($p < 0.05$, chi-squared test).

our method can effectively transfer personality from one design to another, while improving the visual quality of the designs.

8 CONCLUSION

In this paper, we have proposed a semantics-aware deep ranking framework to investigate the personalities of graphic designs. With our framework. we learn a model to predict design personality scores and generic high-level design representation. Our framework can be learned from web search results with minimal human supervision. We have shown that our learned model allows us to perform comprehensive analysis at both region-level and element-level, to understand what contribute to the perceived personalities of graphic designs. In addition, we have also shown that our learned model and representation can enable two novel personality-based graphic design applications. In particular, our element-level design suggestion allows users to quickly select desirable properties for design elements to better convey a personality. Our example-based personality transfer can automatically modify a design to match the personality of another design. To encourage future works, we release our dataset and code at our website[2].

Limitations and future works. First, in this work, we only model the personality of a single type of graphic design. It would be interesting to model different types of graphic designs jointly, in order to study whether different types of graphic designs (e.g., advertisement, magazine and webpages) share some common features to convey their personalities. To this end, our model can be trained on and applied to a dataset with mixed types of graphic designs to discover the most commonly-used features across different types of graphic designs for expressing a particular personality. Second, our design element suggestion application only considers a limited set of basic operations on design elements, such as changing font types and cropping an image using a rectangle region. In practice, to enhance the personality of a graphic design, designers often make use of other sophisticated operations, such as applying special effects to font (e.g., glowing) and cropping an image with irregular shapes (e.g., ellipse). We plan to incorporate such advanced operations into our method in the future. Third, our proposed model is not limited to graphic design personality. We believe that it has the potential to be used to discover other intrinsic features that characterize human perception on graphic designs, such as "What makes a product design look expensive?" and "What makes a graphic design memorable?".

ACKNOWLEDGMENTS

We thank the anonymous reviewers for the insightful and constructive comments, and NVIDIA for generous donation of a Titan X Pascal GPU card for our experiments. This work is in part supported by two SRG grants from City University of Hong Kong (Ref. 7004676 and 7004889).

REFERENCES

Steven Bradley. 2010. How To Use Space In Design. http://vanseodesign.com/web-design/design-space/. (2010).
Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann. 2017. Learning Visual Importance for Graphic Designs and Data Visualizations. In *ACM UIST*.
Ying Cao, Antoni Chan, and Rynson Lau. 2012. Automatic stylistic manga layout. *ACM TOG* 31, 6 (2012), 141.
Ying Cao, Rynson Lau, and Antoni Chan. 2014. Look Over Here: Attention-Directing Composition of Manga Elements. *ACM TOG* 33, 4 (2014).
Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. 2016. Automatic Triage for a Photo Series. *ACM TOG* 35, 4, Article 148 (2016). https://doi.org/10.1145/2897824.2925908
Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: content creation with semantic attributes. In *ACM UIST*. 193–202.
Carrie Cousins. 2015. How Color, Type and Space Can Impact Mood. https://designshack.net/articles/graphics/how-color-type-and-space-can-impact-mood/. (2015).
Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*. 886–893.
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L.i Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE CVPR*.
Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2015. What makes Paris look like Paris? *ACM TOG* 58, 12 (2015).
Elena Garces, Aseem Agarwala, Diego Gutierrez, and Aaron Hertzmann. 2014. A similarity measure for illustration style. *ACM TOG* 33, 4 (2014).
Elena Garces, Aseem Agarwala, Aaron Hertzmann, and Diego Gutierrez. 2017. Style-based exploration of illustration datasets. *Multimedia Tools and Applications* 76, 11 (2017), 13067–13086.

[2]http://nxzhao.com/projects/design_personality/

Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. 2013. Deep convolutional ranking for multilabel image annotation. *arXiv:1312.4894* (2013).
Rebecca Gross. 2015. What It Means to Design With Personality: 25 Awesome Case Studies. https://designschool.canva.com/blog/graphic-art/. (2015).
Michael Gygli, Yale Song, and Liangliang Cao. 2016. Video2gif: Automatic generation of animated gifs from video. In *Proc. IEEE CVPR*. 1001–1009.
Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.
Hamid Izadinia, Bryan Russell, Ali Farhadi, Matthew Hoffman, and Aaron Hertzmann. 2015. Deep classifiers from image tags in the wild. In *Proc. ACM MM Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. 13–18.
Ali Jahanian, Shaiyan Keshvari, SVN Vishwanathan, and Jan Allebach. 2017. Colors–Messengers of Concepts: Visual Design Mining for Learning Color Semantics. *ACM TOCHI* 24, 1 (2017), 2.
Kara Jensen. 2013. What is the "Look and Feel" of a Website? And Why It's Important. https://www.bopdesign.com/bop-blog/2013/11/what-is-the-look-and-feel-of-a-website-and-why-its-important/. (2013).
Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2013. Recognizing image style. *arXiv:1311.3715* (2013).
Janie Kliever. 2015. Designing for Engagement: How Color, Type and Space Can Impact The Mood Of Your Design. https://designschool.canva.com/blog/design-for-engagement/. (2015).
PierreYves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. 2014. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM TOG* 33, 4 (2014).
Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. 2016. Tactile Mesh Saliency. *ACM TOG* 35, 4 (2016).
Sharon Lin, Daniel Ritchie, Matthew Fisher, and Pat Hanrahan. 2013. Probabilistic color-by-numbers: Suggesting pattern colorizations using factor graphs. *ACM TOG* 32, 4 (2013).
Zhaoliang Lun, Evangelos Kalogerakis, and Alla Sheffer. 2015. Elements of Style: Learning Perceptual Shape Style Similarity. *ACM TOG* 34, 4 (2015).
Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.
Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. 2011. Interactive furniture layout using interior design guidelines. *ACM TOG* 30, 4 (2011), 87.
Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21, 6 (1953), 1087–1092.
Rick Nauert. 2011. Why First Impressions Are Difficult to Change: Study. http://www.livescience.com/10429-impressions-difficult-change-study.html. (2011).
Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2011. Color compatibility from large datasets. *ACM TOG* 30, 4 (2011).
Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014a. Learning layouts for single-page graphic designs. *IEEE TVCG* 20, 8 (2014), 1200–1213.
Peter O'Donovan, Jānis Lībeks, Aseem Agarwala, and Aaron Hertzmann. 2014b. Exploratory font selection using crowdsourced attributes. *ACM TOG* 33, 4 (2014).
Xufang. Pang, Ying Cao, Rynson Lau, and Antoni Chan. 2016. Directing user attention via visual flow on web designs. *ACM TOG* 35, 6 (2016).
Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Proc. IEEE ICCV*. 503–510.
Christine Phillips and B Chaparro. 2009. Visual appeal vs. usability: which one influences user perceptions of a website more. *Usability News* (2009), 1–9.
Miriam Redi, Frank Liu, and Neil O'Hare. 2017. Bridging the Aesthetic Gap: The Wild Beauty of Web Imagery. In *ACM ICMR*. 242–250.
Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Gajos. 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *ACM SIGCHI*. 2049–2058.
Daniel Ritchie, Ankita Kejriwal, and Scott Klemmer. 2011. d. tour: Style-based exploration of design example galleries. In *ACM UIST*. 165–174.
Babak Saleh, Mira Dontcheva, Aaron Hertzmann, and Zhicheng Liu. 2015. Learning style similarity for searching infographics. In *Proc. GI*. 59–64.
Ana Serrano, Diego Gutierrez, Karol Myszkowski, Hans-Peter Seidel, and Belen Masia. 2016. An intuitive control space for material appearance. *ACM TOG* 35, 6 (2016), 186.
Nathan Shedroff and Christopher Noessel. 2012. *Make it so: interaction design lessons from science fiction.* Rosenfeld Media.
Clayton Silver and William Dunlap. 1987. Averaging correlation coefficients: should Fisher's z transformation be used? *Journal of Applied Psychology* 72, 1 (1987), 146.
Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).
Stephan Streuber, M Quiros-Ramirez, Matthew Hill, Carina Hahn, Silvia Zuffi, Alice O'Toole, and Michael Black. 2016. Body talk: Crowdshaping realistic 3D avatars with words. *ACM TOG* 35, 4 (2016).
Aarron Walter. 2012. Redesigning With Personality. https://www.smashingmagazine.com/2012/03/redesigning-with-personality/. (2012).
Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proc. IEEE CVPR*. 1386–1393.
Mehmet Yumer, Siddhartha Chaudhuri, Jessica Hodgins, and LeventBurak Kara. 2015. Semantic shape editing using deformation handles. *ACM TOG* 34, 4 (2015).
Matthew Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv:1212.5701* (2012).
Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proc. IEEE ECCV*. 818–833.
Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *Proc. IEEE CVPR*. 1556–1564.