# Image Visual Realism: From Human Perception to Machine Computation

Shaojing Fan, Tian-Tsong Ng, Bryan L. Koenig, Jonathan S. Herberg, Ming Jiang, *Student Member, IEEE,* Zhiqi Shen, Qi Zhao, *Member, IEEE*

**Abstract**—Visual realism is defined as the extent to which an image appears to people as a photo rather than computer generated. Assessing visual realism is important in applications like computer graphics rendering and photo retouching. However, current realism evaluation approaches use either labor-intensive human judgments or automated algorithms largely dependent on comparing renderings to reference images. We develop a reference-free computational framework for visual realism prediction to overcome these constraints. First, we construct a benchmark dataset of 2520 images with comprehensive human annotated attributes. From statistical modeling on this data, we identify image attributes most relevant for visual realism. We propose both empirically-based (guided by our statistical modeling of human data) and CNN-learned features to predict visual realism of images. Our framework has the following advantages: (1) it creates an interpretable and concise empirical model that characterizes human perception of visual realism; (2) it links computational features to latent factors of human image perception.

**Index Terms**—Visual realism, human psychophysics, statistical modeling, convolutional neural network.

## 1 Introduction

Visual realism is important for computer graphics (CG) rendering, image forensics, and photo retouching. Predicting human perception of visual realism has many applications, such as CG quality evaluation and immersion level control in virtual reality entertainment. Currently, researchers assess visual realism in two ways: automated computational prediction [1], [2], [3], and subjective human judgment [4], [5], [6], [7], [8].

For automated prediction, *reference-based image quality metrics (IQMs)* such as mean squared error (MSE) [9] and structural similarity index [10] are used to quantitatively calculate the distortions induced by global illumination and artifacts given an ideal reference image [1], [2], [3], [11]. Algorithms built on IQMs are objective and usually efficient in predicting CG quality, but they are tuned for certain types of artifacts and thus are not easily generalizable to new data. Also, in many circumstances, ideal reference images are unavailable. Furthermore, IQMs are infrequently evaluated relative to human perception [12].

For subjective human judgment, some researchers have conducted psychophysics experiments to measure how visually realistic their rendered images/scenes are judged to be as compared to the original counterparts [4], [5], [6], [7], [8]. Such evaluation is often labor-intensive as it requires sufficient numbers of both participants and stimuli to substantiate the findings. Furthermore, care is needed with regards to factors that can introduce subjective biases, such as experiment environment, image presentation, and participant characteristics [13].

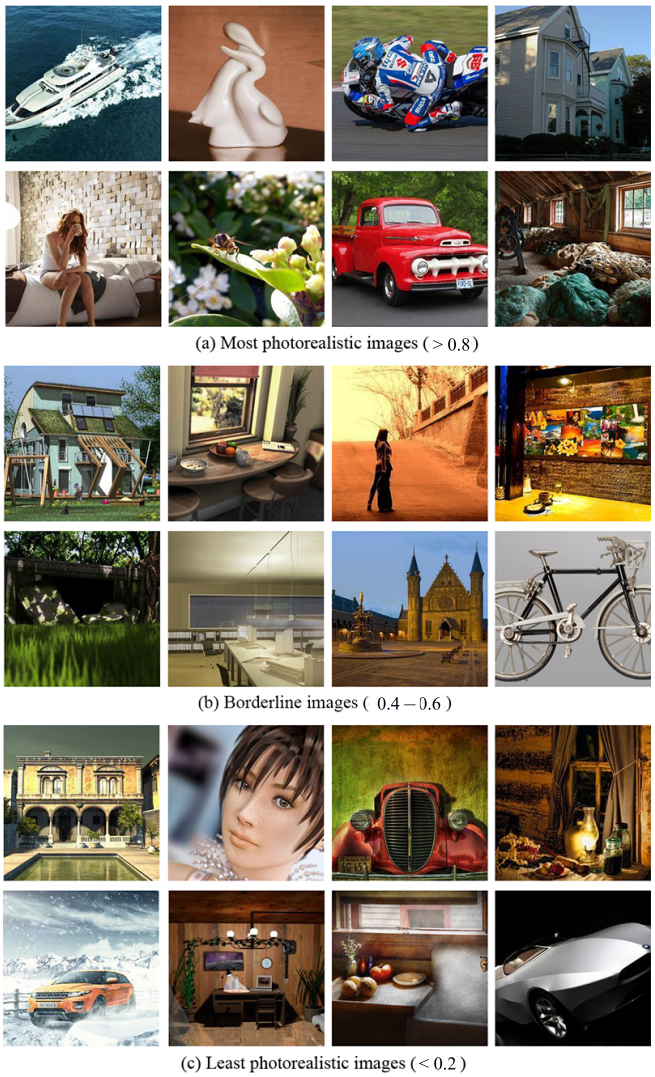
Thus, for methodologies that utilize human judgment for visual realism, the key bottleneck is labor cost, whereas automated computational prediction is typically limited in terms of its dependence on reference images. Our goal is to understand how humans perceive visual realism and to employ such understanding in computational models. To achieve this goal, we create a comprehensive dataset, *i.e.*, the Visual Realism Dataset, for which each image has an empirical realism score as well as extensive attributes labels (see Fig. 1 and Table 1). Based on the dataset, we develop computational models for realism assessment grounded by image realism ratings, rather than a model that works by detecting image artifacts. Fig. 2 illustrates the details of our framework. First, we construct our benchmark dataset with intensive human annotations. We then analyze the human data with psychometrics and signal detection theory [14]. The statistical analyses indicate five latent factors in the human data, which we label "realism", "naturalness", "attraction", "oddness" and "face". Based on the correlational structure for these factors, we compute the relation of realism perception to other visual perceptual dimensions. We develop both empirically-based (guided by our empirical modeling of human data) and unsupervisedly learned features, and then compare their performance to established state-of-the-art alternatives. The generalizability of our features and data are further evaluated using the Washington 3D Scene Dataset [8].

We summarize the main contributions of the work as follows:

1) **We establish a new benchmark dataset—the Visual Realism Dataset—for the study of visual realism.** The dataset is composed of diverse CG and photo images, and each image has both human-labeled visual realism scores and human-annotated attributes. The benchmark dataset and code are available to the public for research purposes [15]. Fig. 1 shows example images from our dataset sampled across the

- *S. Fan and Z. Shen are with the Smart Systems Institute, National University of Singapore, Singapore 119613. E-mail: {idmfs, idmshenz}@nus.edu.sg.*
- *T. Ng is with Institute for Infocomm Research, Singapore 138632. E-mail: ttng@i2r.a-star.edu.sg.*
- *B. Koenig is with the Department of Psychology, Southern Utah University, Cedar City, UT 84720, United States. E-mail: bryanleekoenig@gmail.com.*
- *J. Herberg is with the Department of Psychology, National University of Singapore, Singapore 117583, and Institute of High Performance Computing, Singapore 138632. E-mail: jonathan.herberg@gmail.com.*
- *M. Jiang and Q. Zhao are with the Department of Computer Science and Engineering, University of Minnesota, and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583. E-mail: mjiang@umn.edu, qzhao@cs.umn.edu.*
- *Corresponding author: Q. Zhao.*

*Manuscript received August 26, 2017.*



(a) Most photorealistic images ( > 0.8 )

(b) Borderline images ( 0.4 – 0.6 )

(c) Least photorealistic images ( < 0.2 )

Fig. 1. Images of different realism levels from the Visual Realism Dataset. In each row, the two images on the left are computer generated, whereas the two on the right are photos. The number in parentheses represents the realism score (the proportion of participants who rated each image as a photo rather than as CG).

range of realism scores.

2) **We develop practical, reference-free computational models** that predict visual realism. We propose empirically-based features motivated by characteristics of human perception as it relates to visual realism. We also generate CNN-learned features by fine-tuning a CNN model on the Visual Realism Dataset. our experiments demonstrate that both our empirically-based and CNN-learned features outperform other state-of-the-art methods. Experiments on another dataset demonstrate the generalizability of our empirically-based features and dataset.
3) **We introduce a strategy that identifies image attributes that relate to human perception of visual realism.** We use exploratory factor analysis followed by confirmatory factor analysis on the human annotated data, resulting in a latent structure of multi-dimensional human perception of digital images.
4) **We discover that realism perception is affected by observer characteristics.** Based on psychophysics analyses using signal detection theory [14], we find that the "expertise effect" [16], [17] extends to the realism perception of general scenes. We report that viewer gender also affects realism perception.

The current research extends our previous work [18] with fundamental improvements and new contributions. Among the four contributions summarized above, (4) is entirely new in the current work. For (1), previously only part of the benchmark dataset was annotated, but currently the entire dataset has attribute annotations. For (2), we extend the work beyond our original development of empirically-based features, and investigate new CNN-learned features with a CNN-based model. We additionally extend our experiments to a new dataset to test the generalizability of our features and data. For (3), our new strategy represents a significant improvement over our previous method. Whereas the previous method of greedy feature selection was driven by maximizing the regression objective function, we now apply exploratory and confirmatory factor analysis on the entire set of human attributes, allowing for a broader set of performance tasks, such as regression and binary classification.

The remainder of the paper is organized as follows. Sec. 2 describes related work. Sec. 3 introduces how we build our dataset and conduct psychophysics data analyses and empirical modeling. In Sec. 4 we describe our computational models and evaluate their performance on the Visual Realism Dataset and their generalizability on a new dataset, the Washington 3D Scene Dataset. In Sec. 5 we conclude by highlighting our main findings, limitations and potential future directions.

## 2 Related Works

Our interdisciplinary research employs methods and findings from computer vision, computer graphics, and psychology.

### 2.1 Predicting high-level image attributes

Computer vision researchers often link lower-level image features with higher-level attributes, such as aesthetics [19], [20], [21], [22], interestingness [23], memorability [24], visual sentiment [25], [26], and visual realism [27], [7], [18]. Recently, the resurgence of deep neural networks has substantially improved the prediction of high-level image attributes [28], [29], [30], [31], [32]. Although the resulting computer models perform considerably well in predicting such attributes, few insights are provided to explain *why* they actually work.

Computer graphics researchers have used subjective CG quality assessment since the early 1980s. One common approach has been to setup experiments in which participants judge between a real scene and its CG replica generated with various parameter settings [33], [7], [8]. Since such tests in effect strive to cause human observers to believe a CG image they see is real, they are sometimes referred to as *Visual Turing Tests*. The main shortcomings of such tests are that they are labor intensive and influenced by various factors related to participant cognitive characteristics, such as expertise and own-race sensitivity [13], [17]. Findings from our realism judgment study (Sec. 3.2.1) include insights into the impact of human bias, in particular in the form of expertise and gender effects, on realism perception.

### 2.2 Definition and preliminary considerations for image visual realism

**Visual realism defined:** The concept of *visual realism* is similar to *CG fidelity*, or *photo-realism* [4], [5], [6] in that it is defined

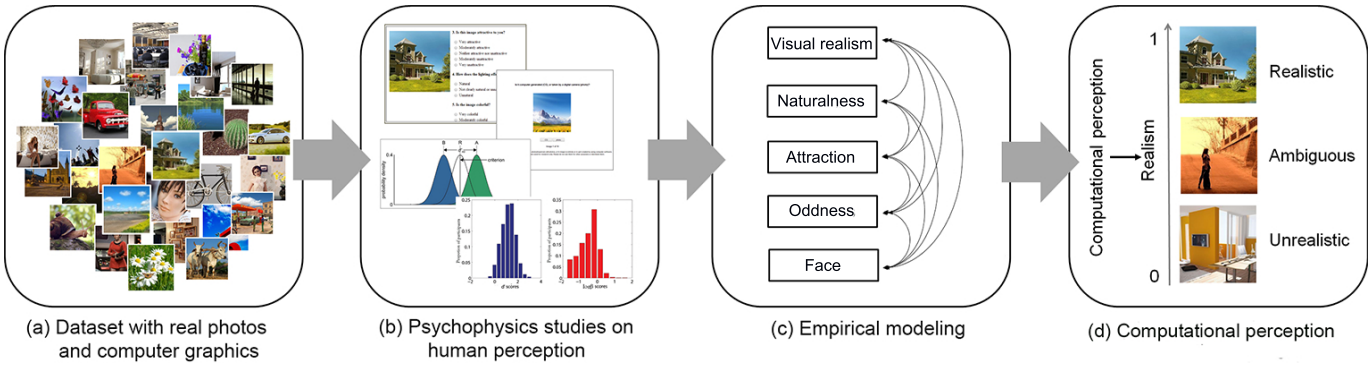



Fig. 2. An overview of our framework. First, we created a dataset of both real photos and computer graphics (a). Second, we performed psychophysics experiments and signal-detection-theory analysis to understand human perception of realism (b). We then performed factor analysis to empirically model human perception (c). Finally, we designed computer algorithms to predict visual realism (d).

as the degree to which a viewed image "produces the same visual response as the scene" (page 2, [34]). However, visual realism is more general than photo-realism in that it covers a larger span of types of images (photos, computer graphics, matte paintings) and does not rely on a reference photo/scene on which to compare renderings.

**Image attributes that influence visual realism:** Many characteristics influence CG fidelity and image realism [5]. Lighting and illumination strongly affect CG fidelity [4], [6], and they are recognized as vital element in CG rendering [35], [36]. The most important attributes for realism of composite images are illumination, color, and saturation [27], [7]. Resolution is also important for realism perception [37]. Several perceptual metrics have been proposed for evaluating photo retouching [38]. These findings on CG fidelity and realism provide foundational blocks for automated realism assessment. The current research focuses on images that are general in content (*e.g.* natural scenes, objects) and source (*i.e.* camera, physics-based rendering, and image-based rendering). Thus, our current research explores a wide range of image attributes on a large and diverse data set.

**Human characteristics that influence visual realism:** Human observers are more discerning regarding the realism of same-race faces [17]. In addition, experts (*e.g.*, CG designers), relative to laypersons, better utilize shading information than color information when judging realism of face images [17]. These findings indicate that image-observer ethnicity similarity and expertise are important factors for human face realism perception. Analogous findings have been found in the psychology of human face perception, referred to as the own-race effect [39], [40], [41], [42] and the expertise effect [43], [44], [45], [16]. The current research evaluates how much these effects extend to perception of visual realism for general images (see Sec.3.2.1).

### 2.3 Related datasets

A related dataset that we employ for model validation purposes (see Sec. 4.6) is the Washington 3D Scene Dataset, which consists of 100 photos and their image-based rendered replicas. The data consists of, for each photo-and-replica pair, average participant judgments as to which image appears more realistic [8].

In the context of visual realism research, the CG community is interested in the visual perception of computer graphics, while the computer vision community tends to examine classification accuracy for photos versus their "impostors", such as CG and composite images. Such classification is based on image characteristics related to image realism [46], [47], [27], [48]. Several computer vision benchmark datasets have been created, notably the Columbia CG and Photo Dataset [49] and the Columbia Image Splicing Dataset [50]. These datasets are useful for classification but not realism assessment. This is because they provide only image-class labels such as CG or photo, intact or spliced, and the image classes cannot be translated into visual realism. For example, a CG image may appear very realistic (see Fig. 1).

Lalonde and colleagues provide a composite image dataset annotated with human judgments of realism, which is composed only of photos (*i.e.*, no CG renderings) [27]. In contrast, we utilize a comprehensive dataset that consists of CG images developed with diverse rendering styles and of photos with various retouching levels. These images also have human-judged realism scores associated with them, allowing for quantitative realism assessment (see Sec. 3).

## 3 The Visual Realism Dataset—Construction, Psychophysics Studies and Data Analysis

We created a benchmark dataset (the "Visual Realism Dataset") for which we measured the visual realism of each image. This dataset is an important prerequisite for both empirical and computational modeling. In this section, we elaborate the methods used in selecting images, collecting human ground truth, and analyzing the data to gain deeper insights into human perception of visual realism.

### 3.1 Dataset overview

We assembled a set of 2520 diverse images. Images were selected in pairs of a CG image and a photo that depicted similar scenes. We considered matte painting images to be CG images in our database. A matte painting image is composed of a base plate, which can be a photograph or moving footage, with CG images or animations superimposed on top of it [51]. We did not include obvious CG images like cartoons. Furthermore, we excluded images with apparent artifacts. We also excluded images with unrealistic scenes, like spaceships flying in a city. All images were scaled and cropped about their centers to 256 $\times$ 256 pixels. Fig. 1 shows example images. Dataset content is illustrated in Fig. 3. Further description of image collection can be found in supplementary

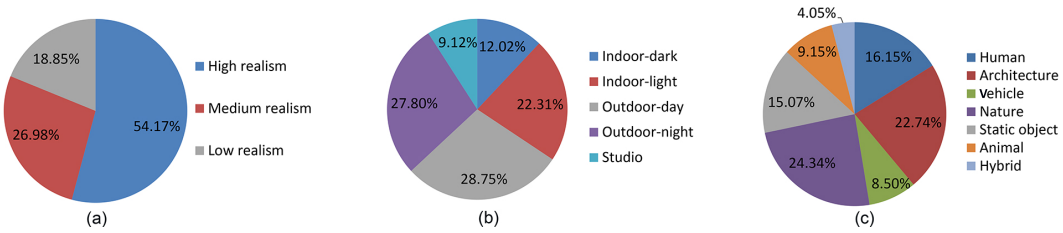



Fig. 3. An overview of statistics regarding (a) realism, (b) lighting, and (c) scene category for the Visual Realism Dataset. In (a), high realism, medium realism, and low realism indicate realism scores between the ranges of $(.67,\ 1]$, $(.33,\ .67]$, $[0,\ .33]$, respectively.

material. The dataset with all annotations can be downloaded from our project website [15].

## 3.2 Psychophysics Study I: realism judgment

The biggest difference between our benchmark dataset and previous datasets like [49] is that ours contains human perception data. The data were collected on Amazon Mechanical Turk (MTurk) as a large-scale psychophysics experiment. Psychophysics Study I is inspired by computer graphics research in which humans judged between real scenes and their CG renderings [4], [5], [6].

### 3.2.1 Experiment procedure

Participants from MTurk reported their subjective perceptions through a Visual Turing Test. They viewed a web-page which presented the title, "Real or fake? Distinguish between computer graphics and real photos". Participants viewed a sequence of images. Their task was to judge each image as "CG" or "photo" then click the corresponding button. In order to explore the impact of cognitive factors on human perception of visual realism, they were encouraged to provide some background information, including their gender and experience related to computer graphics or image processing (selecting from four options: "layperson", "graphic designer or having extensive experience on graphic design", "photographer or photographer enthusiast", or "game player").

Stringent criteria were used to ensure data validity. We required our participants to have >95% approval rate and <15% abandonment rate in MTurk system. Similar to [37], we excluded participants who responded randomly or without good-faith effort (*e.g.*, continuously if they pressed the same key on every trial). In total, the data for 21 participants were excluded, leaving the data for 1292 participants for inclusion in the analyses. On average, each image was scored by 31 participants.

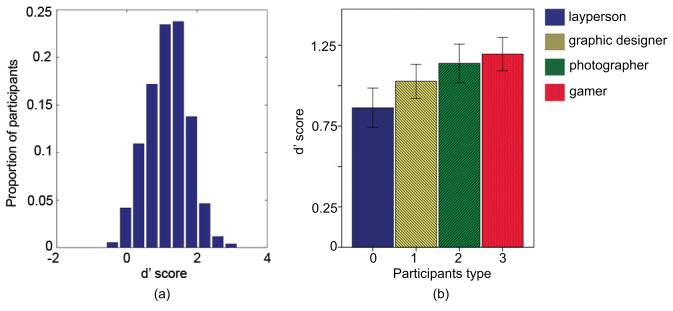



Fig. 4. (a) Distribution of human sensitivity on realism judgment ($d'$) across all participants. Higher values of $d'$ represent higher sensitivity. $d'$ values near zero indicate chance performance. (b) Participants with CG experience (gamers, photographers, graphic designers) have higher sensitivity than laypersons on realism judgment.

### 3.2.2 Data analysis

**Ground truth realism scores:** We define the *realism score* as the proportion of participants who judged the image to be a photo rather than CG. Unsurprisingly, realism scores are higher for photos ($M = .83$, $SD = .17$) than CG images ($M = .45$, $SD = .72$), $t(2518) = 42.41$, $p < .001$[1]. The average image realism score is .64. Both types of images span the the range of possible realism scores (see Fig. 1 and 3).

**Human sensitivity:** To better understand human perception of visual realism, we analyze participant performance using *signal detection theory* [14], a common analysis tool in psychophysics and biology. It offers a method of modeling the decision making process for someone who decides whether items are members of different classes. In signal detection theory, a key metric is the sensitivity index ($d'$), which indicates how much separation there is between the signal and noise distributions, relative to their variability. That is, the difference between the signal and noise distribution means is divided by the square root of their average variance. More formally, under the assumption of normality for both distributions, with the signal mean and standard deviation labeled as $\mu_S$ and $\sigma_S$, and the noise mean and standard deviation as $\mu_N$ and $\sigma_N$, $d'$ is defined as:

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}(\sigma_S^2 + \sigma_N^2)}} \tag{1}$$

In our study, we define photo as the *signal* (class member) and CG as *noise* (not class member). We compute an estimate of $d'$ for each participant from measurements of the participants' hit rate (proportion of photos correctly classified as such) and false-alarm rate (proportion of CG images incorrectly classified as photos), calculated as follows:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate) \tag{2}$$

where function $Z(p), p \in [0, 1]$, is the inverse of the cumulative distribution function for the Gaussian distribution. Thus higher values of $d'$ represent higher sensitivity. $d'$ values near zero indicate chance performance. Fig. 4 (a) shows the distribution of $d'$ across participants. The distribution indicates that participants generally have a positive $d'$ ($M = 1.20$, $SD = .63$), suggesting that they discriminate photos from CG images at a higher rate than that expected by chance. In total, 96.82% of the participants obtain above-chance performance.

**Expertise & gender effects:** Our four self-reported participant expertise categories are: laypersons (432 participants), graphic

1. The result of a $t$-test is presented as, "$t$(df) = $t$-value, $p$ = $p$-value". If a $p$ value is smaller than the conventional significance level threshold of .05, we reject the null hypothesis of no difference between the means.

TABLE 1
Image attributes (Attr), related survey items, attribute categories, and their Spearman's rank correlations ($\rho$) with ground truth image realism scores (from Psychophysics Study I). Meaningful and statistically significant correlations ($|\rho| > .15$, $p < .001$) are highlighted in bold. Numbers in parentheses are participants' mean ratings for each attribute standardized to a scale of 0 to 1.

| Attr | Survey item | Category | $\rho$ | Attr | Survey item | Category | $\rho$ |
|---|---|---|---|---|---|---|---|
| $a_1$ | Appears to be a photograph? ( .73) | Realism | **.78**$^*$ | $a_{21}$ | Clean scene and objects? ( .83) | Layout | .07 |
| $a_2$ | Familiar with the scene? ( .65) | Familiarity | **.29**$^*$ | $a_{22}$ | Makes you happy? ( .63) | Emotions | **.16**$^*$ |
| $a_3$ | Familiar with the objects? ( .77) | Familiarity | **.17**$^*$ | $a_{23}$ | Makes you sad? ( .10) | Emotions | -.04 |
| $a_4$ | Unusual or strange? ( .30) | Familiarity | **-.28**$^*$ | $a_{24}$ | Exciting? ( .60) | Emotions | -.03 |
| $a_5$ | Mysterious? ( .34) | Familiarity | **-.26**$^*$ | $a_{25}$ | Lighting effect natural? ( .75) | Illumination | **.47**$^*$ |
| $a_6$ | Objects appearance natural? ( .76) | Familiarity | **.34**$^*$ | $a_{26}$ | Shadows in the image? ( .58) | Illumination | **-.21**$^*$ |
| $a_7$ | Object combinations natural? ( .76) | Familiarity | **.22**$^*$ | $a_{27}$ | How sharp are the shadows? ( .43) | Illumination | -.05 |
| $a_8$ | Contain fine details? ( .58) | Texture | -.02 | $a_{28}$ | Contain living objects? ( .38) | Semantics | .09 |
| $a_9$ | Color appearance natural? ( .82) | Color | **.45**$^*$ | $a_{29}$ | Dynamic scene? ( .37) | Semantics | -.03 |
| $a_{10}$ | Colors go well together? ( .88) | Color | **.20**$^*$ | $a_{30}$ | Is there a storyline? ( .45) | Semantics | **-.17**$^*$ |
| $a_{11}$ | Colorful? ( .56) | Color | .11 | $a_{31}$ | Number of unique objects ( .61) | Semantics | -.05 |
| $a_{12}$ | Image quality ( .70) | Quality | .08 | $a_{32}$ | Total number of objects ( .70) | Semantics | -.07 |
| $a_{13}$ | Image sharpness ( .72) | Quality | .13 | $a_{33}$ | Number of people ( .48) | Human semantics | .01 |
| $a_{14}$ | Expert photography? ( .59) | Aesthetics | **.33**$^*$ | $a_{34}$ | Face visible? ( .21) | Human semantics | **.20**$^*$ |
| $a_{15}$ | Attractive to you? ( .71) | Aesthetics | .06 | $a_{35}$ | Is the person attractive? ( .15) | Human semantics | -.13 |
| $a_{16}$ | Close-range or distant-view? ( .64) | Layout | .01 | $a_{36}$ | Making eye contact with viewer? ( .15) | Human semantics | .14 |
| $a_{17}$ | Have objects of focus? ( .71) | Layout | .05 | $a_{37}$ | Posing for the image? ( .23) | Human semantics | -.12 |
| $a_{18}$ | Neat space? ( .71) | Layout | .12 | $a_{38}$ | Human activities ( .49) | Human semantics | .01 |
| $a_{19}$ | Empty space? ( .49) | Layout | .04 | $a_{39}$ | Human expressions ( .41) | Human semantics | .03 |
| $a_{20}$ | Perspective natural? ( .75) | Layout | **.36**$^*$ | $a_{40}$ | Expression genuine? ( .44) | Human semantics | **.35**$^*$ |

$^*$ $p < .001$ ($p$-value is corrected based on Bonferroni correction.)

designers (119 participants), photographers / photography enthusiasts (216 participants), and game players (525 participants). To equate the category sizes for comparison purposes, we randomly select 119 participants from each category (allowing for the largest common number of participants from each group). We compared $d'$ across these expertise categories by applying standard statistical techniques. We first conducted an omnibus analysis of variance (ANOVA) and then follow up with Tukey HSD post hoc tests on $d'$ to identify significant effects[52].

Participant groups differ on $d'$, $F(3, 472 = 6.61$ , $p < .001$[2]. Tukey post hoc tests indicate that gamers and photographers have significantly higher sensitivity (larger values of $d'$) than laypersons, $p$s $< .05$. Graphic designers, photographers, and gamers do not significantly differ on $d'$ (see Fig. 4 (b)).

Male participants have a higher mean $d'$ than female participants, $t(373) = 3.76$, $p < .001$, suggesting a gender effect. More men than women are gamers (367 vs. 156), so to test whether the gender effect is a byproduct of the gaming effect, we compare the performance between female game players and male game players, as well as female laypersons and male laypersons. We randomly select the same number of participants from each gender ($n$ = 156 from each for the game players, and $n$ = 205 from each for the laypersons). Male game players have a significantly higher $d'$ than female game players, $t(309) = 3.53$, $p < .001$. Male laypersons also have a higher $d'$ than female laypersons, $t(430) = 4.18$, $p < .001$. These findings suggest that the greater sensitivity on realism judgments of men compared to women is not a byproduct of the gaming effect.

These findings indicate that observer characteristics such as expertise and gender influence visual realism perception on images of general scenes. This extends previous findings of expertise effects for face image realism judgments [13], [17].

2. ANOVA results are presented as, "$F$(dfcondition, dferror) = Fvalue, $p$ = $p$ value". If a $p$ value is smaller than the conventional significance level threshold of .05, we reject the null hypothesis of no difference between the means.

### 3.3 Psychophysics Study II: attribute annotation

To identify factors related to the human perception of visual realism, we obtained participant judgments on a wide set of image attributes in Psychophysics Study II. The design builds on psychology and neuroscience research on human emotion [53], [54] and human memory [55], [56].

#### *3.3.1 Experiment procedure*

We recruited a separate group of MTurk workers to annotate the images (see Table 1; the complete questionnaire is in the supplementary material). We selected the annotation list based on previous research [4], [6], [57], [13], [24]. Each participant annotated up to 5 images. As for Study I, we excluded participants who gave a random response pattern ($n$ = 39), leaving 5762 participants for inclusion in the analyses. We also had images labeled via LabelMe [58], an online annotation tool ($a_{31-32}$ in Table 1).

#### *3.3.2 Data analysis*

We investigate the relationship between image attributes and visual realism. We use Spearman's rank-order correlations ($\rho$) and one-way ANOVAs [52] to assess such relations. The results are shown in Table. 1 and discussed below.

**Realism ratings:** Whereas in Psychophysics Study I (see Sec. 3.2.1) we have participants make a binary decision of whether each image is a photo or CG, for Psychophysics Study II we have participants rate the extent to which images appear to be a photograph versus computer generated ($a_1$, Table. 1) on a five-point scale (1 = computer generated, 5 = photograph). The ratings strongly correlate with the human realism scores from Psychophysics Study I ($\rho$ = .78, $p < .001$). Since different participants are in each study, $\rho$ = .78 demonstrates substantial consistency of human perception of visual realism over both types of measures. However, it is not a perfect positive correlation,

indicating perhaps subjectivity of realism perception or method-related differences. For the following analyses of the relationships between image attributes and realism, we use the realism score from the Psychophysics Study II, as the attributes and realism scores are from the same group of participants.

**Familiarity/Naturalness**: Various aspects of familiarity for the images are rated by our participants ($a_2$-$a_7$). We observe small correlations between the degrees of aspect familiarity and realism ($\rho$s $\geq$ .17). This is consistent with previous findings that images involving common objects correspond with memory representations, causing them to appear more realistic [27], [59].

**Color and illumination:** Color naturalness ($a_9$) and color combination ($a_{10}$) correlate moderately with realism ($\rho$ = .45, and $\rho$ = .20, respectively), which is consistent with previous findings on image composites [27], [7]. Naturalness of lighting ($a_{25}$) has a moderate correlation with realism ($\rho$ = .47), indicating illumination is important for realism. This accords with previous research [4], [6], [13].

**Attraction**: The extent to which an image appears to be the work of an expert photographer ($a_{14}$), an aesthetics attribute which can be labeled as image attraction, moderately correlate with realism ($\rho$ = .33). The correlation is negative ($\rho$ = $-$.21) for images with realism scores greater than .7. This means that highly realistic images do not appear to originate from expert photography, which is consistent with prior research on human skin rendering [60] that found maximal attractiveness and extreme realism were opposing perceptions. Despite the somewhat non-linear relationship between aesthetics and realism, subsequent analyses use linear regression for simplicity and consistency with the analyses used for the other attributes.

**Objects:** Object count and unique object count ($a_{31-32}$) are uncorrelated with realism ($|\rho|$s $\leq$ .07). We further perform one-way ANOVAs to test the effect of scene category and object type (for detailed scene categories see Fig. 3). There are effects of both scene category and object type, $Fs(12, 2507) > 4.81$ , $ps < .05$. Interestingly, face visibility and expression genuineness both positively correlate with realism ($\rho$s $\geq$ .20).

### 3.4 Empirical modeling

To investigate what major perceptual factors are related to visual realism, we conducted an exploratory factor analysis (EFA) on the variables listed in Table 1, followed by a confirmatory factor analysis (CFA) [61]. With EFA, the aim is to identify latent constructs in terms of linear combinations of the measured variables. A CFA conducted following an EFA tests the fit of the attributes identified by the EFA [61]. Attributes with poor fits, or loadings, are eliminated. The resulting latent factors contributing to visual realism, along with their correlations, are shown in Fig. 5.

We apply two common indices to measure the fit of the model to the data. The first is the Comparative Fit Index ($CFI$), which compares a chi-square for the fit of a target model to the chi-square for the fit of an independence model, *i.e.*, one in which the variables are uncorrelated. Higher $CFI$s indicate better model fit. Values that approach .90 indicate acceptable fit [61]. Another model fit metric is Root Mean Square Error of Approximation ($RMSEA$), which estimates the amount of error of approximation per model degree of freedom and takes sample size into account. Smaller $RMSEA$ values suggest better model fit. A value of .10 or less is indicative of acceptable model fit [61]. Our CFA model has acceptable fit, $CFI = .93$, $RMSEA = .095$.

As shown in Fig. 5, the statistical analyses indicate five latent factors in the human data, which we label "visual realism", "naturalness", "attraction", "oddness" and "face". Note that in Fig. 5, "visual realism" refers to a latent factor and "realism_annotation" refers to the human annotation of realism, which in turn loads on the visual realism factor.

We note that the positive correlation of the face factor with the visual realism factor is unsurprising. The Visual Realism Dataset includes only realistic face images and excludes obviously CG images like cartoons (see Sec. 3.1). Note, however, that a similar percentage of photos and CG images (7.46% and 7.86%, respectively) include visible faces, so the correlation does not reflect a confound wherein faces occur more often in photos than CG images. Indeed, another dataset with less realistic faces (*e.g.*, cartoon) might produce a negative correlation between face visibility and realism.

The correlational structure of latent factors identified four visual-perceptual dimensions that correlated with the visual realism factor. This inspired us to identify computational measures analogous to those four latent factors (*i.e.*, "naturalness", "attraction", "oddness" and "face") to use as predictors in the empirically-based modeling (see Sec. 4).

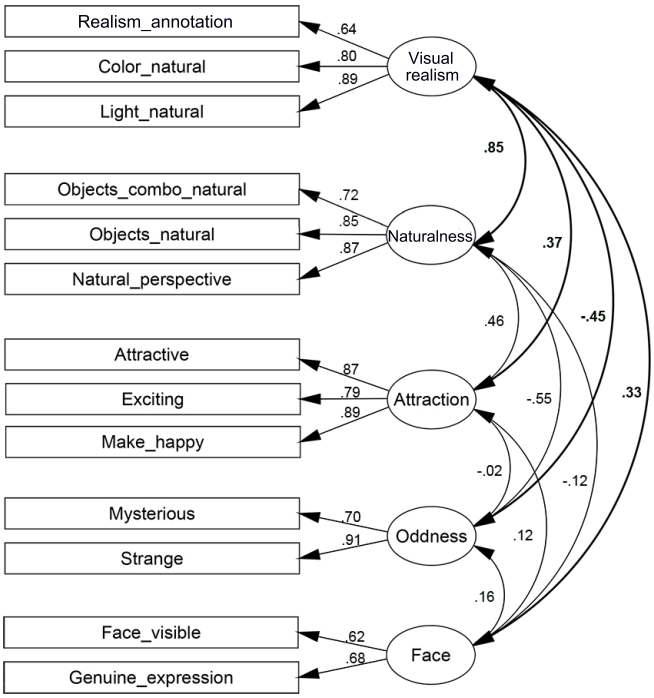



Fig. 5. Results of exploratory and confirmatory factor analysis on Visual Realism Dataset. The correlation coefficients between four latent factors ("naturalness", "attraction", "oddness" and "face") and "visual realism" are highlighted in bold.

## 4 COMPUTATIONAL MODELING

In this section, we first describe our empirically-based models: we construct empirically-based features that encode human perceptual factors relating to visual realism; we then use these features to train two types of classifiers for realism assessment: Support Vector Machine (SVM) [62] and Multi-layer Perceptron (MLP) [63], [64]. We additionally train and develop convolutional neural network

(CNN) models. We compare their performance with other state-of-the-art methods on the Visual Realism Dataset. We also test the generalizability of our features on a new dataset—Washington 3D Scene Dataset. The implementations for feature computation as well as the MLP and CNN models for both classification and regression are available on our project website [15].

### 4.1 Empirically-based model

In this subsection, we introduce our empirically-based model, which is motivated from our human studies. We first introduce the design of image features that encode human perception. We then test the features on two different types of classifiers: SVM and MLP.

#### *4.1.1 Empirically-based features*

Here we propose features that computationally capture the key factors that related to realism perception: naturalness, attraction, oddness and face. Two steps are performed: (i) exploration of various feature measurement strategies to capture each factor for visual realism, and (ii) combining these features in an optimal way. The results of these two steps are a combined computational feature set that automatically capture the perceptual factors related to visual realism of images. The details on feature computation can be found in the supplementary material.

**Naturalness:** We measure naturalness in four ways. First, we measure image similarity through content-based image retrieval. We use $10,000$ images from the SIMPLIcity dataset as a predetermined anchor database of images with common scenes and objects [65]. We then compute the image similarity by using color, illumination and texture information [66], and perform a robust content-based matching with the anchor database. The familiarity measure is denoted by the distances of the top 50 matches. Second, as suggested by [27] and [59], an image may look more realistic if its coloring coheres with memory representations. Therefore, we include color compatibility as a measure for color naturalness. We further compute color name features learned from real-world images [67] to better represent daily color compositions. Finally, we model naturalness by computing image statistics derived from the local patch ($3\times3$) structures and image power spectrum [68].

**Attraction:** Attraction is closely related to an image's aesthetics. We apply Ke's method for extracting aesthetics features, which incorporates image properties such as values in HSV space, edges distributions, blur, and contrast [19]. We also use local self-similarity geometric patterns (SSIM) to represent content symmetry, which is often utilized as a measure of aesthetics [69]. We densely sample the SSIM descriptors with a grid spacing of 4 and learned a dictionary of size 100. We use 2-level spatial pyramid pooling on the descriptors.

**Oddness:** We model oddness by applying the Local Outlier Factor (LOF) algorithm (see [70]) to global image descriptors, a method described in [23]. In anomaly detection, the LOF is an algorithm for finding anomalous data points in a feature space by measuring, for a given data point, the local deviation with respect to its neighbors. We employ a 10-distance neighborhood and three types of features for anomaly detection: the raw RGB pixel values, GIST descriptors [71], and Spatial Pyramids on SIFT histograms [72]. Intuitively, GIST summarizes the rough description of the spatial layout [71] while SIFT is a powerful local feature descriptor [72]. By applying LOF, we are able to identify images with unusual descriptor values, which usually correspond to images with unusual spatial layout or texture patterns [73], [23].

**Face:** Our empirical modeling demonstrates that the presence of human faces is an important factor for visual realism. Therefore we also include a face detector [74], [75]. We extract two types of face features: number of faces in an image, and the relative average size of the faces in the image.

#### *4.1.2 Experimental settings*

We test our empirically-based features on two different classifiers. First, we input the features above into a Support Vector Machine (SVM) [62] (referred to as "EF-SVM", EF represents Empirically-based Features). Kernel summation is used to fuse the features for a balance of performance and efficiency [76]. We use grid search to select cost, RBF kernel parameter $\gamma$, and $\epsilon$ hyperparameters.

To test if the performance is affected by the type of classifiers, we further use our features to train a Multi-layer Perceptron (MLP) (referred to as "EF-MLP"). Similar to CNN, MLP is a feedforward neural network with one or more layers between input and output layer, trained with the backpropagation learning algorithm [63], [64]. Our MLP model is developed using Keras with a Tensorflow backend. We concatenate our features to form a 15256-dimensional vector as input. The MLP structure contains two fully connected layers, which include 2048 and 256 neurons, respectively. The MLP's hidden layers are ReLU activated. The dropout rate is 0.5 and a softmax classifier is present in the output layer. Binary cross entropy is used as the loss function. The whole training includes 300 epochs with stochastic gradient descent. We use a batch size of 256 for each epoch. The learning rate is set to 0.001. A momentum of 0.9 and a weight decay of $10^{-7}$ are used. In each epoch, the network is validated against the validation set of about 500 images to monitor convergence and overfitting. We stop learning when the objective function does not improve on the validation set. We train the network in a single NVIDIA GTX Titan X GPU, and it takes approximately 30 minutes to finish the training.

In both experiments, we randomly split the images from the Visual Realism Dataset into 80% as a training set and 20% as a test set. We use a five-fold cross-validation and repeat the validation five times to determine the result.

### 4.2 CNN-based model

Recently, CNNs have been increasingly used in high-level image understanding [29], [28], [22]. Thus we also train a deep convolutional neural network (CNN) model (referred to as "VR-CNN", VR represents Visual Realism). The CNN model is trained using Keras with a Tensorflow backend [79], [80]. We initialize the training to the pretrained parameters for VGG-19 on ImageNet [81]. The parameters of the CNN are then learned end-to-end on the training images with stochastic gradient descent. Limited by the number of images in the Visual Realism Dataset, we freeze the weights of layers before the last max pooling layer (pool5). We use a batch size of 64 for each epoch. A momentum of 0.9 and a weight decay of $10^{-7}$ are used. The learning rate begins at 0.001. The whole training contains 300 epochs. In each epoch, the network is validated against the validation set of about 500 images to monitor the convergence and overfitting. We stop learning when the objective function does not improve on the validation set. We

TABLE 2
Experimental results of realism prediction (regression) and image type classification on Visual Realism Dataset. $\rho\_s$ and $A\_s$ are respectively the Spearman's rank correlation, and area under ROC curve on selected subset of uniformly-distributed scores, $\rho\_w$ and $A\_w$ are those on whole dataset. The results on the first row are from human annotations, the rest are from computational predictions. The best result from computational features on each evaluation metric is highlighted in bold.

| Category | Feature / model | Regressor / classifier | Regression | | Classification | |
|---|---|---|---|---|---|---|
| | | | $\rho\_s$ | $\rho\_w$ | $A\_s$ | $A\_w$ |
| Human | Human annotation | Human | .65[1] | .64[1] | .79[2] | .88[2] |
| EF-SVM | Empirically-based model | SVM | **.43** | **.53** | .72 | **.79** |
| EF-MLP | Empirically-based model | MLP | .38 | .47 | **.73** | .78 |
| VR-CNN | CNN-based model | Neural network | .38 | .51 | **.73** | .78 |
| Signal features | Wavelet [46] | SVM | .16 | .20 | .56 | .63 |
| | Geometry feature [48] | | .31 | .47 | .64 | .74 |
| | Camera noise [47] | | .04 | .06 | .53 | .50 |
| | Color compatibility [27] | | .20 | .23 | .57 | .61 |
| Object/scene features | SIFT [72] | SVM | .28 | .34 | .61 | .66 |
| | GIST [71] | | .16 | .23 | .58 | .61 |
| | HOG2x2 [77] | | .28 | .33 | .58 | .66 |
| | LBP [78] | | .25 | .30 | .59 | .64 |
| Our CVPR model [18] | Empirically-based model | SVM | .41 | .51 | .68 | .77 |

[1] This result is the split half consistency among participants from Psychophysics Study II.
[2] This result is from Psychophysics Study I.

train the network in a single NVIDIA GTX Titan X GPU, and it takes approximately 3 hours to finish the training. We use five-fold cross validation to get the performance on all images—for each time, we randomly selected 80% of the images from the Visual Realism Dataset as a training set, with the rest 20% as a testing set. We repeat the validation five times to determine the result.

We evaluate our empirically-based models and CNN-based model on the Visual Realism Dataset, in terms of their ability to 1) predict image realism scores, 2) accurately classify images as photo vs. CG. For 1), we use human realism scores from Psychophysics Study I as ground truth. For 2), image-type labels (*i.e.*, photo and CG) are used as ground truth.

### 4.3 Evaluation methods

To evaluate realism prediction (*i.e.*, regression) performance, we use the Spearman's rank correlation coefficient [82], [28]. To evaluate classification performance, we adopt the criterion of area under the Receiver Operating Characteristic (ROC) curve [83].

We design two additional measures to evaluate performance. Realism scores are not uniformly distributed in our dataset, perhaps in part due to judgment bias in humans (see Sec.3.2.1). To test if the non-uniformity of the realism score distribution affect prediction performance, we compose a subset of images purposefully selected so they had realism scores distributed as uniformly as possible for both photos and CG images, over the entire realism score range. We test all computer models on both the whole dataset as well as the uniform subset.

Our attribute annotation experiment indicates variability in human visual realism sensitivity (see Fig. 4). We use the non-parametric sign-test to evaluate how well our predicted results will t with human realism perception, on each separate image. In brief, the sign-test checks, for a series of comparisons, whether the number of positive versus negative values differs from chance (modeled by a binomial distribution drawn from a population with no relationship). The null hypothesis is that data in a vector $X$ come from a continuous distribution with median $m$. We used the visual realism scores provided by each of the 10 human judges from Psychophysics Study II to tune the null model. Values are shuffled 10 times, and each time we pick up one score $r$ and perform a two-sided sign test against each of the remaining 9 scores, to test if the 9 scores ($X$) follow the null distribution with median $r$. We then use the computationally predicted score $p$ (computed from annotated attributes, our models, and from other established algorithms, for comparison) for the same image, and test if $X$ follows the distribution with median $p$. Finally, we calculate the percentage of images for which the null hypothesis cannot be rejected. The higher the rate, the better the performance (*i.e.*, less it deviates from human performance).

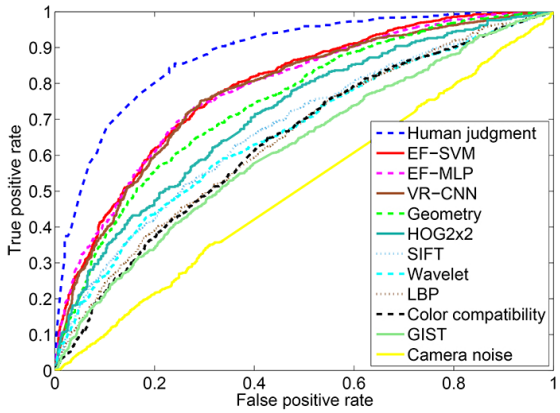



Fig. 6. ROC curve of binary image classification of the whole dataset. Our computer models (EF-SVM, EF-MLP, VR-CNN) outperform other comparing methods, yet still fall short of human performance.

### 4.4 Experimental results

The experimental results are shown in Table 2 and Fig. 6-8. We compare the prediction and classification performance with the performance based on signal processing features commonly used in CG and photo classification, which include high-order correlations of wavelet coefficients [84], physics-motivated geometry structure [85], camera noise [47], and color compatibility [27]. We further test some well known object and scene features such as SIFT

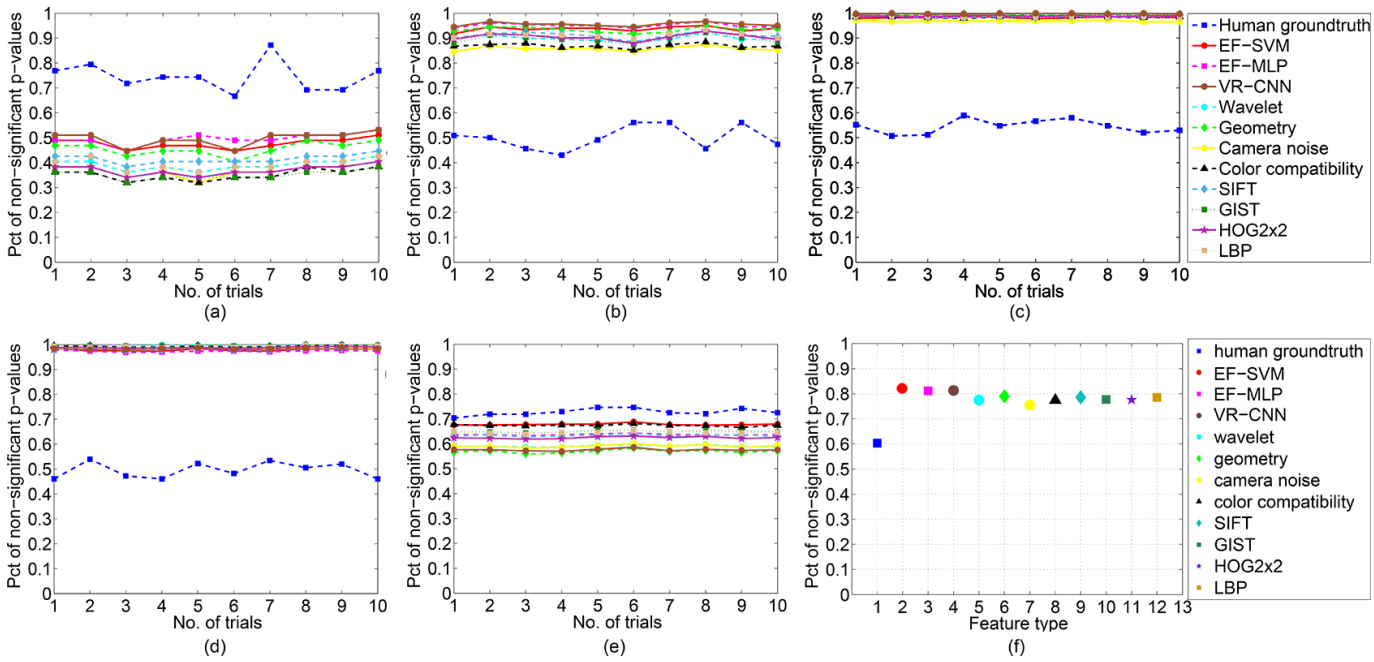



Fig. 7. Evaluation results using sign tests on the Visual Realism Dataset. The vertical axis denotes percentages of images whose regression scores were within the range of human scores, which means that these are the percentage of images for which we could not reject the null hypothesis of no difference between humans and predictor. Figures (a)-(e) show results on images with realism score between $0 \sim 0.2$, $0.2 \sim 0.4$, $0.4 \sim 0.6$, $0.6 \sim 0.8$, $0.8 \sim 1$, respectively. Figure (f) shows the results on all images.

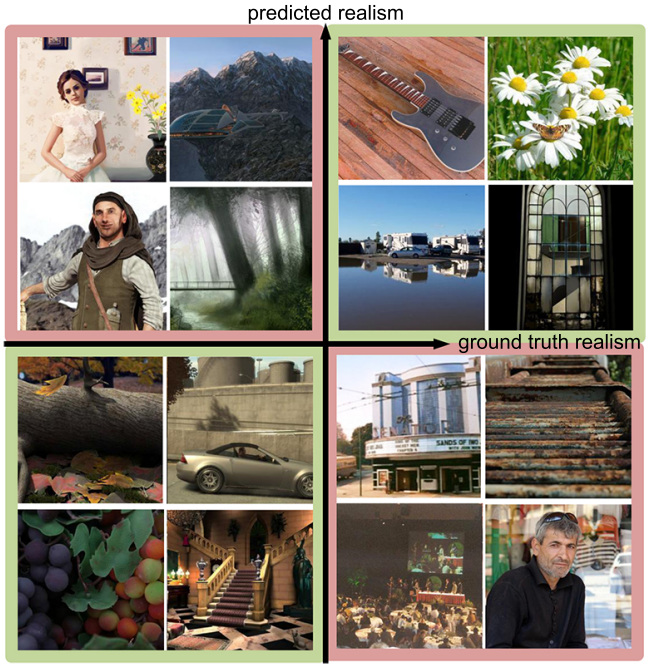



Fig. 8. Example images predicted by our EF-SVM model. The green and red background colors represent correct and false predictions respectively. Our model overpredicts realism for images with unusual scenes (*e.g.*, CG character on the upper left quadrant), whereas it under-predicts realism for images of common scenes but with unusual illumination (*e.g.*, the glass reflectance in the lower right image).

[72], GIST [71], HOG2x2 [77], and LBP [78], computed using an open-source library [86]. Finally, we also compare the regression and classification performance with our previous CVPR work [18]. The results suggest the followings.

**Comparison on overall performance:** As shown in Table 2, the proposed methods (EF-SVM, EF-MLP and VR-CNN) perform considerably better than other methods for both regression and classification. EF-SVM performs either comparable (for classification) or better (for regression) compared with other proposed methods (EF-MLP and VR-CNN), possibly due to the fact that the empirically-based features are carefully designed for encoding human perception of realism and works well for this particular task. It may also suggest that the large margin classifier SVM is more suitable here given the limited sample size. The advantage of EF-SVM is more prominent in regression than classification. This might be because for the classification task, we use image types (*i.e.*, photo and CG) as labels, which could be relatively easily distinguished using certain low-level and human-indiscernible features (many of which are provided by the comparing methods), which apparently diminishes the advantage of our empirically-based features. Whereas for the regression task, we use human perceptual realism scores as ground truth. Compared with classifying photo vs. CG, the regression task of predicting realism scores requires subtler understanding and descriptions of human perception, thus benefiting more from the designed features rather than the automatically learned ones (from VR-CNN) or those from the other comparing methods. Regarding the camera noise feature type, there may be higher sensitivity to image compression and post-processing, accounting for some of its poor classification and prediction performance. In summary, the high performance of our empirically-based features demonstrates that understanding human perception helps create better computational models.

**Comparison with our previous CVPR model:** Our EF-SVM outperforms our previous CVPR model on both regression and classification tasks. This suggests that our new set of features better represents human perception of realism. This is because the previous CVPR model is based on greedy feature, whose

feature selection was driven by maximizing the regression objective function, whereas we now apply exploratory and confirmatory factor analysis on the entire set of human attributes, allowing for a broader set of performance tasks, such as regression and binary classification.

**Whole dataset vs. subset:** The performance of most computer algorithms is generally lower on the subset. The performance drop is understandable as due to realism scores being more distributed (making the prediction task more challenging) and a smaller sample size of images. This suggests that machine learning performance might be improved by increasing the sample size.

**Performance at different realism levels:** As shown in Fig. 7, computer models are better at identifying borderline images, but when it comes to clearly real images humans do better. This may be because humans have a higher daily interaction with the real-life objects and scenes that tend to be captured by realistic looking photos, leading to a higher rate of classifying such images as real, whereas computer algorithms do not have such a bias. For images across all realism ranges, The proposed methods (EF-SVM, EF-MLP, and VR-CNN) are the best among the computational models (see Fig. 7 (f)). Their advantage is most obvious in detecting images of low realism (see Fig. 7 (a)).

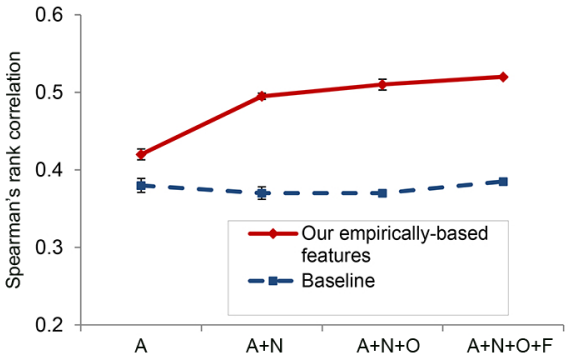



Fig. 9. Realism prediction performance ($+/-SD$) as a function of feature dimension, increased in an incremental manner for the three components: attraction (A), naturalness (N), oddness (O) and face (F). Baseline is obtained by randomly selecting dimensions from our empirically-based features, CNN-learned features, and geometry features.

### 4.5 Comparison with baseline

To check that the superior performance of our empirically-based features is not due only to a combination from diverse methods, but relies on the structure of the empirical model, we compare the performance of our combined feature set to that of a baseline containing the same number of feature dimensions for realism prediction. We construct a pool of 15,256 dimensional features. These features consist of the features in the empirically-based feature set, and the features from exhibited adequate performance as described in the previous sections: *i.e.*, the CNN feature and geometry feature. Baseline performance is obtained by randomly selecting features from the pool.

We aggregate our four components in an incremental manner, starting from the component with smallest feature dimension, until reaching the full combination (see Fig. 9). This process is repeated five times. As shown in Fig. 9, the performance of our feature set exceeds that of the baseline, and its performance improves when adding feature dimensions, whereas baseline does not. This indicates that our feature combination is meaningful.

### 4.6 Generalization to new dataset

In this subsection we extend our experiments on another dataset—Washington 3D Scene Dataset [8]. The aim is to test how well our features and dataset can be generalized for new images. This includes tests on two aspects: feature generalizability and dataset generalizability.

The Washington 3D Scene Dataset [8] consists of 100 Flickr photos—all of outdoor architectures—and their image-based rendered replicas. Each image was resized/rendered to four different resolutions such that the resultant images were 100, 200, 400, and 600 pixels in the smaller dimension (see Fig 10 (a,b,d,e)). Recall that the images in the Visual Realism Dataset are scaled and cropped about their centers to 256 $\times$ 256 pixels. Thus some images only show a partial scene (*e.g.* fourth and fifth images in the first row in Fig. 1). For consistency, we do the same post-processing on the images from Washington 3D Scene Dataset to form images with size 256 $\times$ 256 (see Fig.10 (c)). Due to the small number of images (200) in Washington 3D Scene Dataset, it is almost impossible to train a CNN without overfitting, so we only use the empirically-based features in the following experiments.

#### *4.6.1 Feature generalizability*

First, we compare the performance of our empirically-based feature set on image type classification for the Washington 3D Scene Dataset to the performance of human judges. The human judge classification performance for this dataset is reported in [8]. As shown in Fig. 11, across the four resolutions, the performance of our feature set follows the same trend as for human performance: as resolution decreases, performance decreases. While humans outperform our model on resolutions of 200 pixels or higher, our findings suggest that our empirically-based features are generalizable to the new dataset.

#### *4.6.2 Dataset generalizability*

We use the model trained with empirically-based features and the CNN-based model from the Visual Realism Dataset to predict the image types in Washington 3D Scene Dataset. We obtain a classification accuracy of 53.30% for the CNN-based model, and 54.50% for the model using empirically-based features. This indicates that training on Visual Realism Dataset alone is not sufficient for good performance on classifying images in the Washington 3D Scene Dataset. However, we show below that the classification performance can be substantially improved by adding a small number of training images from it into the training set.

We use *Sequential Minimal Optimization (SMO)* [87], an improved training algorithm for SVM that is used by the popular LIBSVM tool [62]. We iteratively add images from Washington 3D Scene Dataset to the Visual Realism Dataset to form a series of training sets. Using each training set, we then predict the remaining images from Washington 3D Scene Dataset. We also reverse the procedure to incrementally add images from the Visual Realism Dataset to Washington 3D Scene Dataset. To keep the tests comparable, we randomly select 200 images from the Visual Realism Dataset in the reversed procedure. We repeat five-fold cross-validation thirty times and compute the average.

As shown in Fig. 12, a strong boost in classification accuracy on images in the Washington 3D Scene Dataset results by adding

(a) 100px (b) 200px (c) 256px (d) 400px (e) 600px

Fig. 10. Test images at different resolutions (a,b,d,e) in the Washington 3D Scene Dataset. For each pair at a given resolution, the reference photo is on the left, and the rendered image is on the right. The 256-pixel resolution (c) was created by us for the feature generalizability experiment.

its images to the Visual Realism Dataset Base. With 100 images added (less than 4% of the whole set), the classification accuracy is improved significantly (73.63% vs. 52.70%). This suggests that our benchmark dataset provides a good base for generalizable visual realism perception. Furthermore, the performance of the Visual Realism Dataset Base increases faster than that on the Washington 3D Scene Dataset Base (refer to the steeper slope of the red line in Fig. 12). This indicates that, as a base dataset, the Visual Realism Dataset outperforms the Washington 3D Scene Dataset in the classification task. This may be due to the high diversity in the Visual Realism Dataset in terms of both image semantics and rendering type.

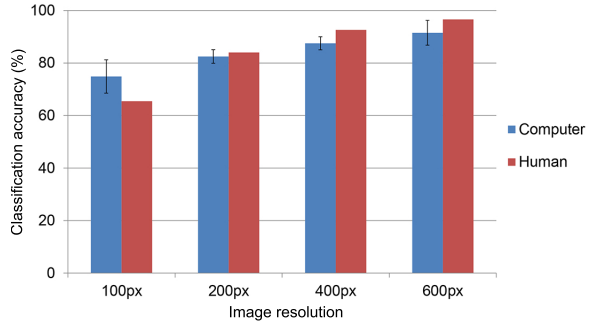



Fig. 11. Image type classification accuracy ($+/-SD$) at four resolutions. Human classification accuracy is from [8], who did not provide variability information for the human judgments.

### 4.7 Summary

Using insights from our psychophysics experiments, we develop an empirically-based feature set for realism prediction on the Visual Realism Dataset. The proposed methods (EF-SVM, EF-MLP and VR-CNN) perform considerably better than other methods on both realism prediction and image type classification, and for both parametric measures (ROC curve) and non-parametric measures (Spearman's rank correlation and sign test).

We test the performance of our empirically-based features on another independent dataset—the Washington 3D Scene Dataset. We demonstrate that our empirically-based features are generalizable to the new dataset. We further demonstrate that adding a small proportion of images from a new dataset to the Visual Realism Dataset can boost classification performance on the new dataset, suggesting the generalizability for our benchmark dataset.

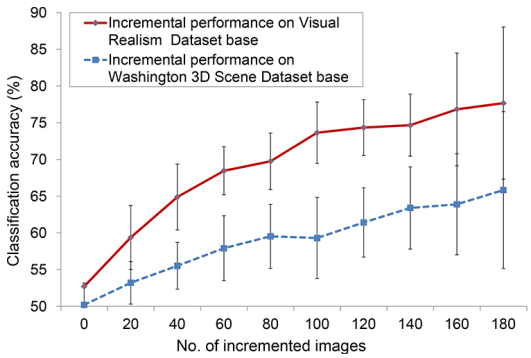



Fig. 12. Classification accuracy ($+/-SD$) with incremental training data on two datasets. X-axis stands for the number of images added to the corresponding dataset base from the other dataset. Chance performance is 50% (base line).

## 5 Conclusions and Future Work

In this paper we propose a comprehensive visual realism dataset that includes human annotated labels of extensive image attributes. We perform statistical modeling on human data to identify image attributes that are most related to visual realism. We develop both empirically-based and CNN-based models for realism prediction and image type classification. We demonstrate the generalizability of our empirically-based features and our benchmark dataset by testing them on a new dataset.

In comparison to human performance, the model trained with SVM using our empirically-based features (EF-SVM) over-predicts realism for images with specific content (*e.g.*, CG character), whereas it under-predicts realism for images of common scenes but with extreme illuminations or image quality (see Fig. 8). Future investigation of model prediction and classification performance on salient objects [88], [89] and patch distinctness [90] may generate useful insights for visual realism modeling.

The current research distinguishes itself from other investigations into visual realism modeling by its cross-disciplinary integration of methods from psychology, computer vision, and computer graphics. The analysis framework and the resulting findings not only provide unique contributions toward understanding human visual realism perception, but also has a variety of related applications, such as image forensics and immersion level control in virtual reality entertainment.

## ACKNOWLEDGMENTS

We thank Qi Shan and his colleagues from University of Washington for sharing their dataset. We thank following people for their contribution to this work in one way or another: Dr. Cheston Y.-C Tan, Mr. Karianto Leman, Mr. Zhang Fan, Dr. Chu Xinqi, Dr. Wang Hee Lin, Prof. Liu Zhen, and all the reviewers for our previous papers on this topic. This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative, and a University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).

## REFERENCES

[1] C. Wang, J. Gao, and H.-W. Shen, "Parallel multiresolution volume rendering of large data sets with error-guided load balancing," in *Proceedings of the 5th Eurographics conference on Parallel Graphics and Visualization*, pp. 23–30, Eurographics Association, 2004.

[2] B. Moon, N. Carr, and S.-E. Yoon, "Adaptive rendering based on weighted local regression," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 5, p. 170, 2014.

[3] W. Jakob, E. d'Eon, O. Jakob, and S. Marschner, "A comprehensive framework for rendering layered materials," *ACM Transactions on Graphics*, vol. 33, no. 4, 2014.

[4] G. Meyer, H. Rushmeier, M. Cohen, D. Greenberg, and K. Torrance, "An experimental evaluation of computer graphics imagery," *ACM Transactions on Graphics*, 1986.

[5] P. Rademacher, J. Lengyel, E. Cutrell, and T. Whitted, "Measuring the perception of visual realism in images," in *Rendering Techniques 2001*, pp. 235–247, Springer, 2001.

[6] A. McNamara, "Exploring perceptual equivalence between real and simulated imagery," in *ACM symposium on Applied perception in graphics and visualization*.

[7] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier, "Understanding and improving the realism of image composites," *ACM Transactions on Graphics, 2012*, vol. 31, no. 4, pp. 84:1–84:10, 2012.

[8] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. M. Seitz, "The visual turing test for scene reconstruction," in *3DTV-Conference, 2013 International Conference on*, pp. 25–32, IEEE, 2013.

[9] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 94–109, 2012.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, pp. 1398–1402, IEEE, 2003.

[11] T. S. Cho, C. L. Zitnick, N. Joshi, S. B. Kang, R. Szeliski, and W. T. Freeman, "Image restoration by matching gradient distributions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 683–694, 2012.

[12] M. Cadik, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, p. 147, 2012.

[13] S. Fan, T.-T. Ng, J. Herberg, B. Koenig, and S. Xin, "Real or fake?: Human judgments about photographs and computer-generated images of faces," in *Technical Briefs, ACM SIGGRAPH Asia*, 2012.

[14] T. Wickens, *Elementary signal detection theory*. Oxford University Press, USA, 2001.

[15] "Human perception of image visual realism." https://www.nus-sesame.top/visualrealism/, 2017.

[16] T. J. McKeeff, R. W. McGugin, F. Tong, and I. Gauthier, "Expertise increases the functional overlap between face and object perception," *Cognition*, vol. 117, no. 3, pp. 355–360, 2010.

[17] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig, "Human perception of visual realism for photo and computer-generated face images," *ACM Transactions on Applied Perception (TAP)*, vol. 11, no. 2, p. 7, 2014.

[18] S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, C. Y.-C. Tan, and R. Wang, "An automated estimator of image visual realism based on human cognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 4201–4208, IEEE, 2014.

[19] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, vol. 1, pp. 419–426, IEEE, 2006.

[20] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *ICIP*, pp. 105–108, IEEE, 2008.

[21] T. O. Aydin, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 21, no. 1, pp. 31–42, 2015.

[22] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," *arXiv preprint arXiv:1704.00248*, 2017.

[23] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, "The interestingness of images," in *ICCV*, pp. 1633–1640, IEEE, 2013.

[24] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1469–1482, 2014.

[25] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM Multimedia*, pp. 83–92, ACM, 2010.

[26] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM Multimedia*, pp. 223–232, 2013.

[27] J. Lalonde and A. Efros, "Using color compatibility for assessing image realism," in *ICCV*, 2007.

[28] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *ICCV*, 2015.

[29] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Learning a discriminative model for the perception of realism in composite images," *arXiv preprint arXiv:1510.00477*, 2015.

[30] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

[31] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, 2017.

[32] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[33] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala, "Visual equivalence: towards a new standard for image fidelity," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 76, 2007.

[34] J. A. Ferwerda, "Three varieties of realism in computer graphics," in *Proceedings SPIE Human Vision and Electronic Imaging*, vol. 3, pp. 290–297, 2003.

[35] P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *ACM SIGGRAPH 2008 classes*, p. 32, ACM, 2008.

[36] C.-H. Hung, T.-P. Wu, Y. Matsushita, L. Xu, J. Jia, and C.-K. Tang, "Photometric stereo in the wild," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 302–309, IEEE, 2015.

[37] H. Farid and M. J. Bravo, "Perceptual discrimination of computer generated and photographic faces," *Digital Investigation*, vol. 8, no. 3, pp. 226–235, 2012.

[38] E. Kee and H. Farid, "A perceptual metric for photo retouching," *proceedings of the national academy of sciences*, vol. 108, no. 50, pp. 19907–19912, 2011.

[39] C. A. Meissner and J. C. Brigham, "Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review.," *Psychology, Public Policy, and Law*, vol. 7, no. 1, p. 3, 2001.

[40] C. Michel, B. Rossion, J. Han, C.-S. Chung, and R. Caldara, "Holistic processing is finely tuned for faces of one's own race," *Psychological Science*, vol. 17, no. 7, pp. 608–615, 2006.

[41] M. J. Bernstein, S. G. Young, and K. Hugenberg, "The cross-category effect mere social categorization is sufficient to elicit an own-group bias in face recognition," *Psychological Science*, vol. 18, no. 8, pp. 706–712, 2007.

[42] V. Natu, D. Raboy, and A. J. O'Toole, "Neural correlates of own-and other-race face perception: Spatial and temporal response differences," *NeuroImage*, vol. 54, no. 3, pp. 2547–2555, 2011.

[43] I. Gauthier, M. Tarr, *et al.*, "Becoming a "greeble" expert: exploring mechanisms for face recognition," *Vision research*, vol. 37, no. 12, pp. 1673–1682, 1997.

[44] B. Rossion, I. Gauthier, V. Goffaux, M. Tarr, and M. Crommelinck, "Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses," *Psychological Science*, vol. 13, no. 3, pp. 250–257, 2002.

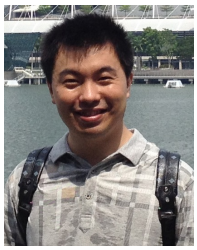
[45] C. M. Bukach, I. Gauthier, and M. J. Tarr, "Beyond faces and modularity: The power of an expertise framework," *Trends in cognitive sciences*, vol. 10, no. 4, pp. 159–166, 2006.
[46] S. Lyu and H. Farid, "How realistic is photorealistic?," *IEEE Transactions on Signal Processing*, 2005.
[47] E. Dirik, S. Bayram, H. Sencar, and N. Memon, "New features to identify computer generated images," in *ICIP, 2007*, vol. 4, pp. IV–433, IEEE, 2007.
[48] T.-T. Ng and S.-F. Chang, "Discrimination of computer synthesized or recaptured images from real images," in *Digital Image Forensics*, 2013.
[49] T.-T. Ng, S.-F. Chang, J. Hsu, and M. Pepeljugoski, "Columbia photographic images and photorealistic computer graphics dataset," *Columbia University, Advent Technical Report*, pp. 205–2004, 2005.
[50] T.-T. Ng, S.-F. Chang, and Q. Sun, "A data set of authentic and spliced image blocks," *Columbia University, ADVENT Technical Report*, pp. 203–2004, 2004.
[51] C. Barron, "Matte painting in the digital age," in *ACM SIGGRAPH 98 Conference abstracts and applications*, p. 318, ACM, 1998.
[52] R. Bailey, *Design of comparative experiments*, vol. 25. Cambridge University Press, 2008.
[53] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[54] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, no. 4, pp. 626–630, 2005.
[55] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, 2015.
[56] W. A. Bainbridge, D. D. Dilks, and A. Oliva, "Memorability: A stimulus-driven perceptual neural signature distinctive from memory," *NeuroImage*, vol. 149, pp. 141–152, 2017.
[57] P. J. Lang and M. M. Bradley, "The international affective picture system (iaps) in the study of emotion and attention," *Handbook of emotion elicitation and assessment*, vol. 29, 2007.
[58] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
[59] S. Y. Choi, M. Luo, M. Pointer, and P. Rhodes, "Investigation of large display color image appearance-lll: Modeling image naturalness," *JIST*, vol. 53, no. 3, pp. 31104–1, 2009.
[60] F. Giard and M. J. Guitton, "Beauty or realism: The dimensions of skin from cognitive sciences to computer graphics," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1748–1752, 2010.
[61] R. Kline, *Principles and Practice of Structural Equation Modeling*. Guilford Press, 2011.
[62] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011.
[63] H. Bourlard and C. J. Wellekens, "Links between markov models and multilayer perceptrons," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167–1178, 1990.
[64] S. Haykin and N. Network, "Neural networks: A comprehensive foundation," *Neural Networks*, vol. 2, no. 2004, p. 41, 2004.
[65] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *TPAMI*, 2001.
[66] Kirk, "Content based image retrieval." https://github.com/kirk86/ImageRetrieval, 2013.
[67] J. Van De Weijer, C. Schmid, and J. Verbeek, "Learning color names from real-world images," in *CVPR, 2007*, pp. 1–8, IEEE, 2007.
[68] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.
[69] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR, 2007*.
[70] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, pp. 93–104, ACM, 2000.
[71] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, 2001.
[72] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR, 2006*.
[73] T. Hu and S. Y. Sung, "Detecting pattern-based outliers," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3059–3068, 2003.
[74] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pp. 214–219, IEEE, 1998.
[75] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–I, IEEE, 2001.
[76] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.
[77] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*.
[78] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, 2002.
[79] F. Chollet *et al.*, "Keras," 2015.
[80] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[82] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Wiley StatsRef: Statistics Reference Online*, 2006.
[83] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[84] H. Farid and S. Lyu, "Higher-order wavelet statistics and their application to digital forensics," in *Computer Vision and Pattern Recognition Workshop, 2003.*, vol. 8, pp. 94–94, IEEE, 2003.
[85] T.-T. Ng, S.-F. Chang, J. Hsu, L. Xie, and M.-P. Tsui, "Physics-motivated features for distinguishing photographic images and computer graphics," in *ACM Multimedia*, 2005.
[86] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *NIPS*, 2012.
[87] J. Platt *et al.*, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.
[88] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014.
[89] K. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *TPAMI*.
[90] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1139–1146, 2013.

**Shaojing Fan** is a postdoctoral research fellow at the Sensor-enhanced Social Media (SeSaMe) Centre in the Smart Systems Institute, National University of Singapore (NUS). Prior to joining NUS, she was a senior research engineer at Institute for Infocomm Research, part of Singapore's Agency for Science, Technology, and Research. She received the B.E. and M.E. degree in Communication and Information System from South China University of Technology. She finished her D.Phil. at Institute for Infocomm Research, Singapore, and Ningbo University, China. Her main research interests includes cognitive vision, computer vision, and experimental psychology.

**Tian-Tsong Ng** is a research scientist at the Institute for Infocomm Research, part of Singapore's Agency for Science, Technology, and Research. He received his M.Phil. in Signal Processing from Cambridge University in 2001 and his Ph.D. in Electrical Engineering from Columbia University in 2007. His research interest lies in the application of advanced signal processing methods to discover the structure of image formation which helps solving problems in computer vision, graphics and image forensics.

**Bryan L. Koenig** received a B.A., with a major in Psychology and a minor in Latin, from St. John's University in 1998. He received an M.A. in General/Experimental Psychology from the College of William and Mary in 2005. In 2009 he earned a PhD in Social Psychology with a minor in Experimental Statistics from New Mexico State University. For three years he then worked as a research scientist at the Institute of High Performance Computing, part of Singapore's Agency for Science, Technology, and Research. Until recently, he was an adjunct instructor at Washington University in St. Louis. He is now an assistant professor in the Department of Psychology at Southern Utah University. He does research on social perception, emotions, morality, and evolutionary psychology.

**Jonathan S. Herberg** has worked over five years as a research scientist (cognitive psychologist) at the Institute of High Performance Computing, part of Singapore's Agency for Science, Technology, and Research. He obtained his Ph.D. in Psychology from Peabody's Cognition and Cognitive Neuroscience Program at Vanderbilt University. His research includes experimental and educational psychology, collaborative learning, human-computer interaction, psychometrics and predictive analytics.



**Ming Jiang** is a postdoctoral associate at the Department of Computer Science and Engineering, University of Minnesota. He is interested in computer vision, cognitive vision, psychophysics and computational neuroscience. His studies focus on computational models of visual attention. He obtained his Ph.D. degree in electrical and computer engineering from National University of Singapore, and his M.E. and B.E. degrees in computer science from Zhejiang University. He is a member of the IEEE.

**Zhiqi Shen** is a research engineer at the Sensor-enhanced Social Media (SeSaMe) Centre in the Smart Systems Institute, National University of Singapore. He was an intern student at Institute for Infocomm Research, part of Singapore's Agency for Science, Technology, and Research from 2014 to 2015. He received his B.E degree in Network Engineering from Ningbo University of Technology in 2015. His research interest lies in deep learning for computer vision and pattern recognition.

**Qi Zhao** is an assistant professor in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. Her main research interests include computer vision, machine learning, cognitive neuroscience, and mental disorders. She received her Ph.D. in computer engineering from the University of California, Santa Cruz in 2009. She was a postdoctoral researcher in the Computation & Neural Systems, and Division of Biology at the California Institute of Technology from 2009 to 2011. Prior to joining the University of Minnesota, Qi was an assistant professor in the Department of Electrical and Computer Engineering and the Department of Ophthalmology at the National University of Singapore. She has published more than 40 journal and conference papers in top computer vision, machine learning, and cognitive neuroscience venues, and edited a book with Springer, titled Computational and Cognitive Neuroscience of Vision, that provides a systematic and comprehensive overview of vision from various perspectives, ranging from neuroscience to cognition, and from computational principles to engineering developments. She is a member of the IEEE.