ASME Journal of Mechanical Design
Online journal at:
https://asmedigitalcollection.asme.org/mechanicaldesign

Comparing and Evaluating Human and Computationally Derived Representations of Non-Semantic Design Information

Elisa Kwon
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
e-mail: elisa.kwon@berkeley.edu

Kosa Goucher-Lambert[1]
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
e-mail: kosa@berkeley.edu

Design artifacts provide a mechanism for illustrating design information and concepts, but their effectiveness relies on alignment across design agents in what these artifacts represent. This work investigates the agreement between multi-modal representations of design artifacts by humans and artificial intelligence (AI). Design artifacts are considered to constitute stimuli designers interact with to become inspired (i.e., inspirational stimuli), for which retrieval often relies on computational methods using AI. To facilitate this process for multi-modal stimuli, a better understanding of human perspectives of non-semantic representations of design information, e.g., by form or function-based features, is motivated. This work compares and evaluates human and AI-based representations of 3D-model parts by visual and functional features. Humans and AI were found to share consistent representations of visual and functional similarities, which aligned well with coarse, but not more granular, levels of similarity. Human–AI alignment was higher for identifying low compared to high similarity parts, suggesting mutual representation of features underlying more obvious than nuanced differences. Human evaluation of part relationships in terms of belonging to the same or different categories revealed that human and AI-derived relationships similarly reflect concepts of "near" and "far." However, levels of similarity corresponding to "near" and "far" differed depending on the criteria evaluated, where "far" was associated with nearer visually than functionally related stimuli. These findings contribute to a fundamental understanding of human evaluation of information conveyed by AI-represented design artifacts needed for successful human–AI collaboration in design.
[DOI: 10.1115/1.4063567]

Keywords: data-driven design, design representation, design theory and methodology

1 Introduction

It is critical to consider the representation of design artifacts to enable effective communication and provision of design information between design agents. Not only is it necessary to align human perspectives of design artifacts, but as computational data-driven methods become increasingly utilized in design, human and computational representations also need to agree to support seamless co-design processes between humans and artificial intelligence (AI). One process in which the use of data-driven methods is especially prevalent is in the retrieval of external sources of inspiration in early-stage design [1]. In this work, we explore the representation of design artifacts by considering inspirational stimuli for design. Representation by humans and AI of visual and functional attributes of inspirational stimuli are specifically investigated.

The focus in this work on non-semantic form and function-based features of design stimuli is first motivated by a recent review by Jiang et al. on data-driven design-by-analogy (DbA). In their review, the use of modalities beyond textual data such as visual information (2D-image or 3D-model datasets) to support visual or multi-modal DbA is proposed [1]. Currently, there is limited knowledge regarding how relationships defined computationally in terms of non-textual properties of inspirational stimuli align with human representations and evaluations. One implementation of data-driven methods for the retrieval of multi-modal inspirational stimuli was explored in prior work by our team [2]. In this work, deep-neural networks modeling visual and functional relationships between 3D-model parts were used in a multi-modal search platform for inspiration discovery. In two subsequent user studies, designers using this system searched for stimuli in terms of appearance and function-based similarities to a specified input and were frequently returned results they did not expect [3].

These findings additionally motivate the aim of the present study to investigate the representation and evaluation of non-text-based measures of similarity, which have not been widely studied in

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received May 8, 2023; final manuscript received September 19, 2023; published online November 7, 2023. Assoc. Editor: Katherine Fu.

interactive settings. Increased availability, interest, and use of 2D-image and 3D-model datasets encourage the development of tools enabling the discovery of design stimuli related to an input by non-text-based features rather than semantic distances. Our approach is to compare and evaluate human and AI-based representations of non-text-based definitions of similarity to increase understanding of these less explored measures of similarity. Our aim through this work is to promote the effective transfer of information and communication between human and AI design agents when engaging with multi-modally represented design artifacts. This central research aim is explored across two research questions:

(RQ1) How *consistent* and *aligned* are human and AI-based representations of non-text-based similarities of inspirational stimuli?

(RQ2) How do *evaluations* of human and AI-based representations of non-text-based similarities of inspirational stimuli compare?

These research questions are studied considering 3D-model parts as a source of design-relevant inspiration, for which non-text-based visual and functional similarities are defined. A human subjects study was conducted ($n = 36$) consisting of a triplet rating task and a categorization task. Initial findings from this study were previously reported by Kwon and Goucher–Lambert in Refs. [4,5]. First, addressing RQ1, human assessments of the similarity of 3D-model parts by visual and functional features were collected in a triplet rating task. The alignment and consistency across human and computational representations of visual and functional similarities are evaluated. Additional insight into a comparison of these representations is revealed through qualitative analysis. To address RQ2, findings from a categorization task are analyzed in which stimuli were organized by participants based on visual or functional similarity. Categorization of stimuli is used to evaluate similarities computed in terms of human versus AI-based representations for each similarity type explored. Low levels of similarity are expected to align with different-group categorization and vice versa. Leveraging the notion of "near" and "far," typically attributed to conceptual distances (e.g., [6]), "near" distances are associated with same-group categorization and "far" with different-group categorization. Comparing how humans and AI represent and evaluate non-text-based relationships between design artifacts can support their improved transfer of information and the effective retrieval of relevant sources of inspiration for humans by AI.

2 Related Works

Motivating the study of multi-modal representations of similarity in this paper, prior work on multi-modal inspirational design stimuli and methods and tools enabling their retrieval is reviewed.

2.1 Impact of Multi-Modal Inspirational Stimuli. The impact of inspirational stimuli, most notably analogies, on design processes has been well studied due to their potential to retrieve relevant concepts from long-term memory and aid conceptual design [7]. By studying these processes, features of inspirational stimuli that can lead to beneficial outcomes such as increased novelty, feasibility, or innovativeness of ideas (e.g., [6,8,9]) can be determined. This work specifically focuses on the multi-modal representations of stimuli, which, when presented to designers, can differently influence design outcomes. Several examples of past work have investigated designers' interactions with multi-modal stimuli. Borgianni et al. studied the impact of stimulus form on idea generation by presenting textual, pictorial, or combined stimuli to designers [10,11]. Findings encourage the presentation of multiple forms of stimuli to designers due to the diversity and limited overlap of ideas generated by participants exposed to different forms of stimuli. Designers tend to prefer visual information [12,13], which Linsey et al. found can lead to increased idea novelty [14]. Han et al. suggest that images combined with unrelated semantic elements can promote creative idea generation [15], while using pictorial stimuli was found by Malaga et al. to outperform the use of words alone for enhanced creativity of ideas [16]. In general, interacting with visual stimuli can importantly trigger the formation of new mental images, which can support the generation of new design ideas [17]. These studies demonstrate the value of providing multi-modal, e.g., pictorial, stimuli to designers. In the current study, 3D-model parts are proposed as another form of inspirational stimuli containing both visual and functional attributes. Human and computational representations of relationships between these parts are investigated to support designers' interactions with these stimuli. In the next subsection, methods enabling their representation and retrieval are explored.

2.2 Enabling Retrieval of Inspirational Stimuli. To provide designers with relevant sources of inspiration, similarity relationships between designer inputs and potential stimuli need to be defined. Defining similarities to support data-driven DbA has been most widely studied in the context of deriving analogical distances between source and target domains [1]. Computational methods are often used to retrieve design stimuli with varying analogical distances to a given design problem or designer-specified input. Similarity relationships specifically relying on textual information can be derived. For instance, text-based processing has been used to define function-based similarity between design problems and solutions from patents [18], to define contextual similarity between patents [6,19] or to assign function-based topics to patents based on different semantic themes [20,21]. Semantic networks used during engineering design activities can facilitate exploration and retrieval of analogies consisting of common words, such as in WordNet or ConceptNet [22], or technology-based knowledge from patent texts in the Technology Semantic Network (TechNet) [23].

However, beyond processing textual information, there is increasing interest in using AI to represent and retrieve stimuli from 2D-image and 3D-model datasets [1]. These stimuli can support multi-modal analogy for design inspiration. Sketch-based retrieval of visually similar examples can importantly support visual analogy [24,25]. Zhang and Jin used an unsupervised deep-learning model to construct a latent space for a dataset of sketches [25]. Image-based search using visual similarity can also extract relevant examples from sources such as patent documents [26,27]. Jiang et al. constructed a convolutional neural network-based model to derive a vector space where feature vectors embed visual and technology-related information from patent images [27]. Other sketch-based user interfaces include DreamSketch, which provides designers with 3D-modeled design solutions based on early-stage 2D-sketch-based designs [28], or SketchSoup, which inputs rough sketches and generates new sets of sketches to inspire further concept generation [29]. Design ideas represented in 3D can be recognized by tools such as the InspireMe interface, which provides suggestions for new components to add to a designer's initial 3D model [30]. Kim et al. developed a co-creative sketching AI partner that provides inspirational sketches related by visual and conceptual similarity to designer-drawn sketches [31]. The effects of providing sketches with varying levels of visual and conceptual similarity to the designer's sketch were investigated [32]. In our prior work, deep learning was applied to develop deep-neural networks modeling visual and functional relationships between 3D-model parts in a large dataset [2]. These neural networks were used to construct a multi-modal search platform, through which designers' search for inspiration was examined.

Using AI to represent multi-modal stimuli in terms of non-text-based features can increase their utilization as sources of design inspiration. Also, designers' interactions with multi-modal inputs (e.g., sketch or 3D-model based) can be better enabled. Ensuring that computational methods used to define non-text-based relationships appropriately represent how humans perceive these similarities is the primary aim motivating this work. This aim is achieved by conducting a human subjects study, as described in the following section, to model human representations of visual

and functional similarities between inspirational stimuli and by comparing and evaluating human and AI-based representations.

3 Methods

To compare and evaluate human and AI-based representations of inspirational stimuli, visual, and functional similarities between 3D-model parts are explored in this work. Similarity is described by distances between stimuli in embedding spaces derived using two approaches. The first approach uses deep learning to construct neural networks modeling these relationships, resulting in *computational embedding spaces* for a large dataset of 3D-model parts (developed in prior work [2]). Presented in the current work, the second approach uses human-evaluated similarities of a selection of 3D-model parts to build *psychological embedding spaces* of parts. Psychological embedding spaces rely on visual and functional similarity assessments collected in a triplet rating task (Sec. 3.1.3). In a second task (Sec. 3.1.4), these parts were then categorized based on visual and functional similarity. Categorization outcomes are used to quantify how computed human and AI-based similarity measures are evaluated in terms of higher-level human assessments. Methods used to conduct the study, define and evaluate non-text-based similarities, and analyze post-task qualitative data are described in this section.

3.1 Experimental Design. This study consisted of two main tasks: a triplet rating task and a categorization task, each completed twice (once for each similarity type explored). For one similarity type, participants completed 25 trials of the triplet rating task followed by the categorization task. The same two tasks were then repeated for the other stimulus set. The order of similarity type (visual or functional) presented was counterbalanced across participants. After completing each set of 25 triplet ratings, participants were additionally asked to provide open-ended responses describing the specific criteria used to assess visual or functional similarity. Experimental details of the triplet rating and categorization tasks are fully described in Secs. 3.1.3 and 3.1.4. To determine which stimuli to present in these tasks, two distinct sets of 16 3D-model parts were selected from the computational embedding spaces with varying pairwise distances in either visual or functional similarity.

3.1.1 Participants. For this study, 36 participants (13 female, 22 male, and 1 non-binary) were recruited including 14 graduate students, 16 undergraduate students, and six industry professionals (with <1–9 years of experience). In prior work from the authors, any impact of expertise when engaging with inspirational stimuli was in their utilization in a structured design task (not relevant to the current study) [33]. For the tasks completed, no particular level of engineering design knowledge or experience was required, and no analysis of differences in expertise was conducted. Participants were recruited via email from among current students in Mechanical Engineering as well as participants who previously completed research studies related to engineering design. Participants were compensated with $10 for their completion of the 30-minute study. This human subjects research study was approved by the Institutional Review Board at the University of California, Berkeley.

3.1.2 Selection of Task Stimuli. The stimulus sets provided to participants in the study (see Fig. 1) were selected by considering distances between 3D-model parts in deep-learning-based computational embedding spaces. These neural networks were trained on 573,585 part instances belonging to 26,671 3D-model object assemblies across 24 object categories. To encode the visual similarity of 3D-model parts, the 128-dimensional deep-learning model used 2D snapshots from various angles of each part to understand its geometric and physical form. The 64-dimensional functional network was developed by considering neighboring parts within a part's respective object assembly such that two parts are similar if they share similar neighbors (e.g., a chair leg and back are functionally similar because a chair seat is a common neighbor). The development of these neural networks is fully described in our past work [2].

Given the size and diversity of the full dataset, candidate stimuli were restricted to "chair" and "table" object categories, specifically considering chair seats, chair backs, and tabletops (as labeled within the PartNet dataset [34]), resulting in 2043 possible parts. This was done to reduce the potential difficulty of rating similarity between and categorizing very diverse objects (e.g., bottles and tables). Although task complexity was reduced as a result, ultimately, the aim of this selection of stimuli was to encourage the assessment of similarity in terms of visual and functional features only. This aim could be better achieved without the influence of semantic information, including product category. Potential limitations of the present findings related to stimuli selection are discussed in Sec. 5.4.

The full 16-part stimulus sets selected to present in the triplet rating and categorization tasks are shown in Fig. 1, chosen based on visual similarity in Fig. 1(*a*) and functional similarity in Fig. 1(*b*). Euclidean distances between parts in the computational embedding spaces are used to represent how similar (low distance) or dissimilar (high-distance) parts are. While distances between neighbors are not constant, neighboring parts (e.g., 1 and 2) are

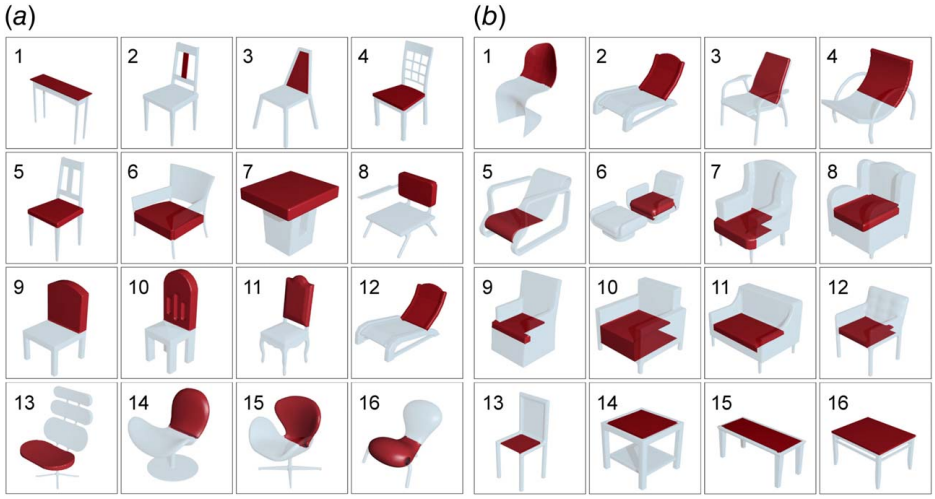

Fig. 1 Stimuli presented during study to assess (*a*) visual and (*b*) functional similarity

always nearer in terms of pairwise distance than non-neighboring parts (e.g., 1 and 5). By maintaining consistency in pairwise distances, we ensure that the stimulus sets used contain a diversity of distances where all parts belong to both low- and high-distance pairs.

3.1.3 Triplet Rating Task. Developing a psychological embedding space that models human representations of a given stimulus set requires the collection of many trials of human judgments. A common task used to elicit these judgments is a triplet rating task where one of two options is selected as being more similar to a given reference. Prior work by Nandy and Goucher–Lambert and Ahmed et al. have also used triplet similarity ratings to generate embedding spaces for human representations of design stimuli [35,36]. Preceding each triplet rating task of 25 trials, participants were told that "In the [first/second] section of this study, you will consider the [function/appearance] of parts when assessing similarity." At the beginning of each trial, participants were asked to "Select the option with the most similar part in [function/appearance] to the reference part" with two options presented, such as in the example shown in Fig. 2. Participants were instructed to make this selection based on the red-highlighted 3D-model part in the object assembly. When considering functional similarity, participants were told to consider the object the red part belongs to, other neighboring parts in the object, and that parts with high functional similarity may be used in the same object and/or neighbor similar parts. For visual similarity trials, no further detail was provided.

For the number of parts in each stimulus set (16), a total of 1680 unique triplet trials are possible. Ahmed et al. recommend that a minimum of 30% of the full stimulus set is needed to construct a robust embedding space of human representations [36]. In our study, 36 participants completed 25 triplet ratings for each stimulus set. Due to data collection errors and the exclusion of data from one participant who failed the attention check for the visual similarity triplet rating task, a total of 801 trials for visual similarity and 826 trials for functional similarity were included, constituting 48% and 49% of all potential trials.

3.1.4 Categorization of Stimuli. Following 25 trials of the triplet rating task for one similarity type, a categorization task was then conducted for the same stimulus set (presented unordered). Rather than allow participants to freely group parts, two different criteria were specified to consider for each similarity type. Participants were instructed to examine the (1) shape (e.g., geometry) and (2) size (e.g., thickness) of a part when categorizing parts by visual similarity. The (1) object the reference part belongs to and (2) neighboring parts to the reference part were criteria specified when categorizing parts by functional similarity. These criteria were selected based on knowledge of the part features learned by the computationally derived neural networks as well as the similarity evaluation criteria participants provided in pilot testing. For each stimulus set, participants constructed two sets of three or four categories for the specified criteria. By associating computed similarities derived from the previously specified methods and criteria with categorization outcomes, insight into how similarities are evaluated can be gained.

Fig. 2 Example triplet of 3D-model parts shown to participants during triplet rating task

3.2 Definition and Evaluation of Similarities Between Stimuli. Using the similarity assessments obtained in the triplet rating tasks, a psychological embedding space was constructed for each stimulus set to model human representations of parts. Similarity between parts is represented by their embedding space distances. The relationship between similarity and categorization is then used to gain insight into how computed similarities between stimuli were evaluated.

3.2.1 Construction of Psychological Embedding Spaces. Using outcomes from the triplet rating tasks, psychological embedding spaces were constructed. The PYTHON library PsiZ was used to generate these models, which specifically handles behavioral data such as triplet ratings to infer psychological embeddings.[2] The embedding techniques and development of the software package used to obtain the psychological embeddings from triplet ratings are fully described by Roads and Mozer [37]. These models include two layers: an embedding layer representing multidimensional features and a similarity kernel. The similarity kernel consists of a distance function (weighted Minkowski distance) and a similarity function (exponential decay in similarity with increased distance). The use of this two-component kernel is motivated by psychological theory and has been used to successfully represent psychological embeddings [37]. The number of dimensions for each model was determined by training models with dimensions varying from two to ten. The highest value at which validation set (10% of trials) losses stopped improving for increasing values of dimensionality was selected. The final psychological embedding spaces for both visual and functional similarity are two-dimensional with training/validation/test set losses of 0.45/0.51/0.45 and 0.39/0.43/0.49, respectively. Constructing these embedding spaces importantly enables the measurement of distances between stimuli in terms of human representations for comparison against computational representations of visual and functional attributes.

3.2.2 Definition of Similarity Between Parts. Two definitions of similarity between stimuli are considered in this work. The first definition involves directly computing Euclidean distances between all 120 pairs of parts in both the psychological and computationally derived embedding spaces. These distances are derived from the full 64 and 128-dimensional computational embeddings and two-dimensional psychological embeddings. Measuring similarity in terms of Euclidean distances implies symmetry in pairwise relationships that may not always be appropriate to maintain, as explored by Chaudhari et al. [38], following observations of asymmetric similarity in Tversky's featural theory of similarity [39]. However, in the context of 3D-shape retrieval for the properties of similarity explored, symmetric distance-based measures are considered suitable [40]. We compute Euclidean distance to represent part similarity in order to align with our prior work where it was used to retrieve nearest neighbors to users' input queries in a multi-modal search platform [2]. Prior work has shown Euclidean distance to be suitable as a simple metric to query for closest matching images or models to sketch-based queries in other implementations of retrieval tasks using deep-learning models [41]. Pairwise distances are compared when assessing consistency between human and AI-based representations of visual and functional similarity in Sec. 4.2.1.

A second definition, similarity levels, is defined to represent similarity between pairs at a higher level than pairwise distances through the process conceptualized in the example in Fig. 3. Euclidean embedding space distances are ordered by decreasing distance where lower distance between parts represents higher similarity, and vice versa. According to pairwise embedding space distances,

[2]https://github.com/psizorg/psiz

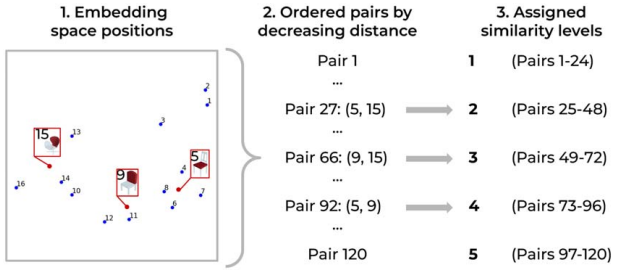

Fig. 3 Conceptual overview of process used to assign similarity levels to pairs of parts by ordered pairwise distances

pairs are assigned a similarity level between 1 and 5, where each level contains 24 pairs. Lower levels are assigned to high distance, and thus low similarity, pairs (such as 5 and 15 in the example) and higher levels to low distance, high similarity pairs (such as 5 and 9 in the example). Assignment of pairs to similarity levels (i.e., 1–5) may provide more generalizable insights beyond specific distances computed in this work. In studies investigating stimuli used for design-by-analogy, for example, retrieval criteria for "near" or "far" analogies are often expressed in terms of percentiles of similarity [8,42]. Insights pertaining to high and low levels of similarity, rather than specific distances, may be relevant toward identifying "near" and "far" stimuli in terms of non-text-based relationships, different from more commonly explored text-based analogies. The alignment of pairs assigned to similarity levels based on psychological and computational embedding space distances is investigated in Sec. 4.2.2. Similarity levels are used in the evaluation of each representational space, as described in the following subsection.

3.2.3 Evaluation of Similarity Representations Through Categorization of Parts. In the categorization task detailed in Sec. 3.1.4, each participant created three to four categories to organize stimuli in terms of given criteria. Based on these categories, every unique pair of parts (120 in total) was associated with an outcome of either being grouped together or separately. Each pair of parts was also assigned to a similarity level (1–5) consisting of 24 pairs in total. For each participant, the proportion of pairs in each similarity level categorized into different groups was computed. The mean proportion across participants was then found for each similarity level, using bootstrapping to compute 95% confidence intervals due to the small sample size. Significantly above-chance group means (where the lower limit of the 95% confidence interval is greater than 50%) indicate different-group categorization of pairs of parts with the specified similarity level. Using the notion of "near" and "far" prevalent in DbA (e.g., studies by Chan et al. [8,43]), above-chance different-group categorization is used to distinguish the boundary between similarity levels at which stimuli may be evaluated as "too far." Analogously, stimuli with a similarity level associated with below-chance different-group categorization (i.e., same-group categorization) may be "too near" to be relevant. We propose that in between these similarity levels lies the "sweet spot" referred to by Fu et al. in inspirational stimuli avoiding these extremes [6]. A similar approach was used by Cooke et al. to model the relationship between similarity and categorization of 3D objects [44]. The boundary separating parts that are "too far" are expected to be observed at low similarity levels, corresponding to parts separated by greater distances within the embedding spaces, while stimuli perceived as "too near" are expected to be related by high levels of similarity. These boundaries are used to *evaluate* each embedding space by identifying the levels of similarity at which stimuli may be "near" or "far," according to human perspectives of distance.

3.3 Analysis of Qualitative Data. Following the completion of each set of 25 triplet ratings for stimuli based on visual and functional similarity, participants provided written open-ended responses to describe the criteria they used to assess similarity. While evaluation criteria used when employing deep learning can be speculated, exact definitions for each dimension of these models are unknown. However, deeper insight can be gained regarding how humans represent visual and functional information through the qualitative post-task data obtained. This analysis may help to inform future deployment of computational methods to represent relationships based on multi-modal information by understanding the features of inspirational stimuli emphasized using each method. Criteria used across participants to evaluate visual and functional similarity were coded from the open-ended responses provided by following an inductive category formation approach [45]. Using this method of qualitative content analysis, criteria were defined to code responses where new criteria were formed if responses could not be subsumed under previously defined criteria. Multiple criteria from a participant's response could be assigned to the same codes. This process continued for all responses collected from all 36 participants and repeated for both similarity types evaluated.

Two coders, each with at least one publication in an engineering design journal, coded the full dataset. Coder 1 manually coded all responses using the described inductive category formation process. Coder 2 then validated the assignment of responses to the criteria identified by Coder 1 independently. Across both coding processes, 0.98 Cohen's Kappa inter-rater reliability was achieved for visual similarity and 0.82 for functional similarity, suggesting high consistency between coders [46]. Differences in assigned codes were discussed and resolved between coders. Details of both coding outcomes are described in Sec. 4.4 to identify human evaluation criteria used for visual and functional similarity.

4 Results

The main research aims of this work are to *compare* and *evaluate* the representation of stimuli within embedding spaces constructed using human and AI-based assessments of visual and functional similarity. First, the newly constructed psychological embedding spaces modeling human similarity assessments are presented in Sec. 4.1. Two sets of comparisons are then conducted between psychological and computational embeddings of stimuli. First considered is the consistency in defining *pairwise distances* from embedding spaces of both models (Sec. 4.2.1). In the second comparison, stimulus pairs are ordered in terms of pairwise distances and assigned to *levels of similarity* (1–5). Alignment between methods in assigning pairs to the same levels is then investigated (Sec. 4.2.2). In Sec. 4.3, to evaluate these psychological and computational similarity representations, the relationship between similarity and categorization of parts is explored. Supporting these findings, qualitative findings are presented to uncover features of non-text-based similarities that may be specific to human or computational representations (Sec. 4.4). The analyses presented in this section are based on earlier versions of these results in Refs. [4,5]. Insights from these findings can support improved agreement across human and AI design agents during engagement with multimodally represented design artifacts.

4.1 Examining Psychological Embedding Spaces. In order to compare human and AI-based representations of inspirational stimuli by non-text-based relationships, psychological embedding spaces were constructed, as described in Sec. 3.2.1. These are visualized in Fig. 4 where plotted points are numbered corresponding to parts in stimulus sets in Fig. 1. Numbering of parts reveals how relationships between parts are represented in computational spaces (detailed in Sec. 3.1.2) such that part 1 is closer in distance to 2 than 5.

Inspecting the psychological embedding spaces, several visually related stimuli separated by low Euclidean distances in the computational embedding space are more distant in Fig. 4(*a*). Part 12, for

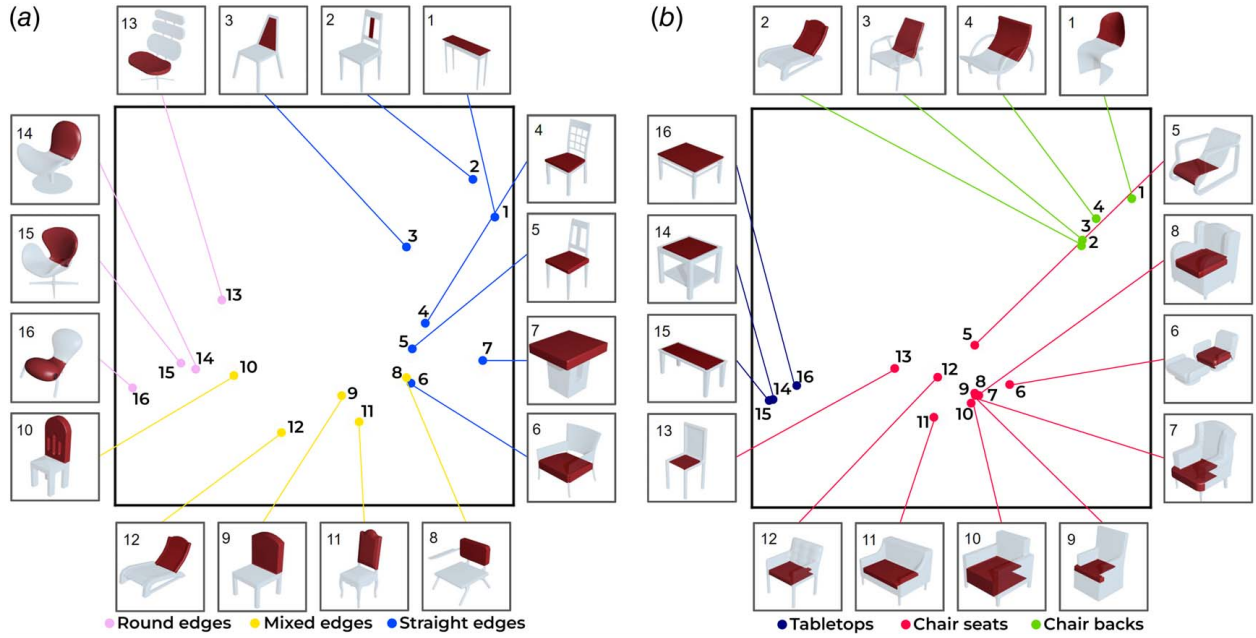

Fig. 4 Visualizations of psychological embedding spaces. (*a*) Visually similar stimuli in Fig. 1(*a*) grouped by edges. (*b*) Functionally similar stimuli in Fig. 1(*b*) grouped by object part.

instance, is closer to parts 9 and 10 in the psychological embedding space than to 13, which is a nearest neighbor to 12 in the computational embedding space. There may be low agreement between visual similarity relationships represented by both models. Parts are colored in terms of edge curvature, which is one potential criterion used to evaluate visual similarity. Criteria used by human participants to make both visual and functional similarity assessments are presented in Sec. 4.4. Computationally derived relationships appear more preserved in terms of functional similarity, as demonstrated by clusters of closely numbered parts in Fig. 4(*b*). It is evident from the separation of tabletops, chair backs, and chair seats in Fig. 4(*b*) that humans relied on the object part when making functional similarity judgments. Overall, by representing these distances using human evaluations of similarity obtained experimentally, the alignment with deep-learning methods can be determined. In the following section, further examination of the agreement between these representations is conducted.

4.2 Agreement Between Human and Computational Representations of Similarity. In response to the first research question posed in this work, one focus of this study is to determine the agreement between human and AI-based representations of visual and functional similarity. Two sets of stimuli consisting of 16 3D-model parts are considered, as shown in Fig. 1. Across these stimuli, there are 120 unique pairs in each stimulus set, where visual attributes are evaluated for one stimulus set (Fig. 1(*a*)) and functional attributes for the other (Fig. 1(*b*)). In total, four models are investigated: one computational and one psychological embedding space for each of the visual and functional

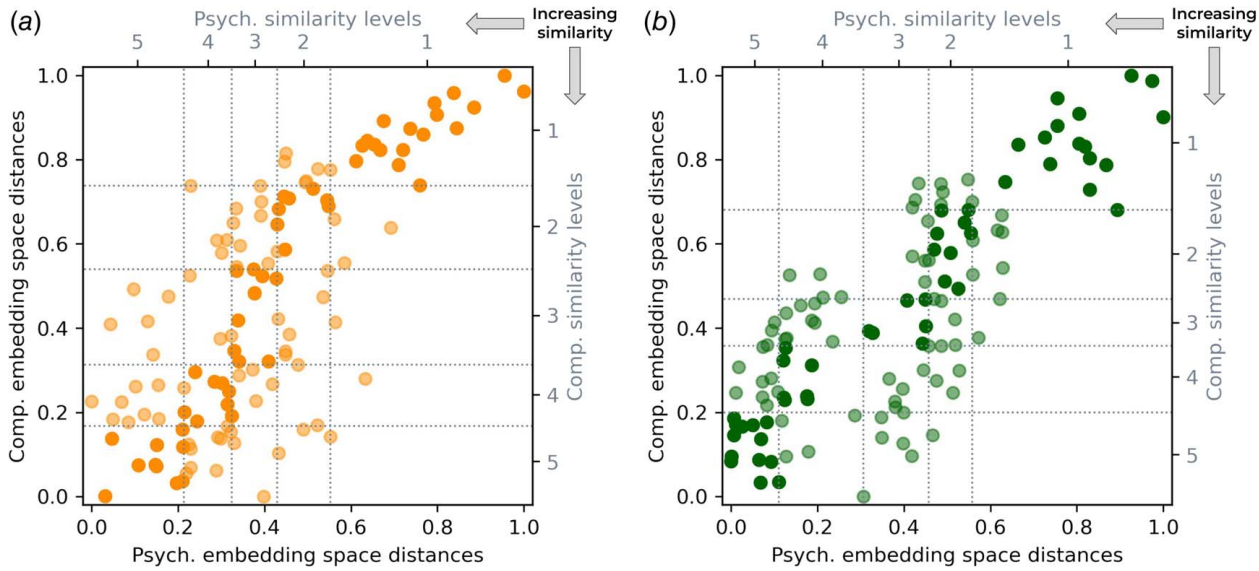

Fig. 5 Range-normalized pairwise psychological and computational embedding space distances with associated similarity levels shown for (*a*) visual and (*b*) functional similarity. Darker points indicate overlap of pairs assigned to the same similarity levels.

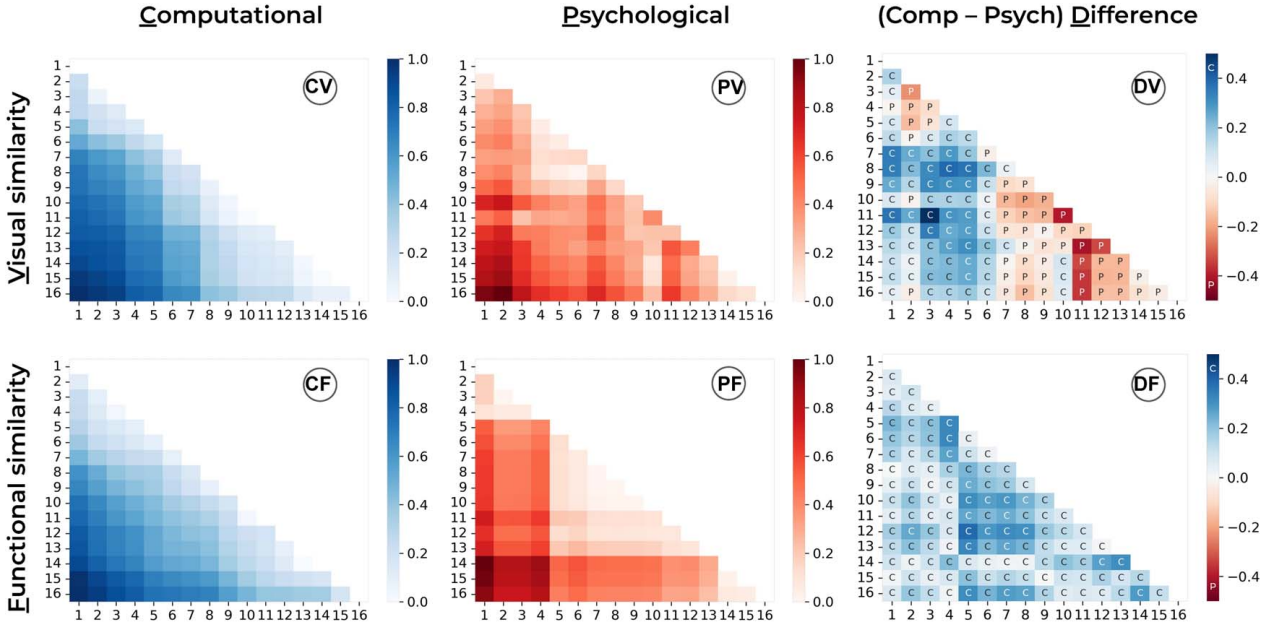

Fig. 6 Range-normalized pairwise distances between stimuli 1–16 in computational (C) and psychological (P) embedding spaces representing visual (V) and functional similarity (F). Differences in computational and psychological distances shown in (D).

similarity stimulus sets. The agreement between these representations of similarity is evaluated for each similarity type using measures of consistency and alignment. *Consistency* is assessed in terms of computing pairwise distances and *alignment* in assigning pairs to low- to high-ranging levels of similarity.

4.2.1 Consistency Between Pairwise Similarities. The first method used to describe the agreement between human and computational representations of visual and functional similarities is by comparing pairwise distances. Two metrics are used to compare range-normalized pairwise distances derived from psychological and computational embedding spaces: Pearson correlation, r, and Cronbach's alpha coefficient, α. Both metrics, in the context of inter-rater reliability, measure the consistency between raters in measuring a common dimension [46]. To compare embedding space distances, these metrics reveal whether the two models are consistent in what is being assessed, but not necessarily that the computed distances exactly agree. The relationship between range-normalized pairwise distances in psychological and computational embedding spaces is visualized in Fig. 5, for visual similarity (Fig. 5(*a*)) and functional similarity (Fig. 5(*b*)). There are significant positive correlations between pairwise distances modeling both visual similarity, $r(118) = 0.74$, $p < 0.001$, and functional similarity, $r(118) = 0.79$, $p < 0.001$. These relationships are confirmed visually by the positive linear correlations of pairwise embedding space distances. High Cronbach's alpha values are also observed for distances representing visual similarity, $\alpha = 0.82$, and functional similarity, $\alpha = 0.85$. In general, these results demonstrate that, considering all pairs, the embedding spaces are consistent in their representations of visual and functional similarity.

Range-normalized pairwise distances between stimuli labeled 1–16 are differently visualized in heatmaps in Fig. 6. Larger distances between stimuli are darker and represent lower similarity between parts. In the first column, distances derived from computational models are represented (labeled CV and CF) and from psychological models in the second column (PV and PF). Heatmaps in the third column (DV and DF) represent differences (computational—psychological) between these distances to directly compare which pairs of stimuli are represented by a larger distance in one embedding space than the other. The first row of heatmaps represents distances in terms of visual features (CV, PV, and DV) while the second row represents functional features (CF, PF, and DF). As noted in Sec. 3.1.2, stimuli were selected based on computational embedding space distances, which explains the visual consistency in heatmaps CV and CF showing increasing distances between farther separated pairs (e.g., pairs 5 and 15, compared to pairs 5 and 6).

To further investigate where there is more and less agreement between models, the third column of heatmaps (DV and DF) in Fig. 6 is examined. Squares labelled with "C" indicate pairs separated by a higher distance in the computational than psychological embedding spaces, and vice versa for squares labelled with "P". As an example, the dark "C" square in heatmap-DV shows that parts 3 (triangular chair back) and 11 (irregularly curved chair back) in the visual similarity stimulus set (Fig. 1(*a*)) are more distant in the computational embedding space. By contrast, the dark "P" squares at the intersections of parts 10 (round-edged chair back) and 11 or parts 11 and 13 (rounded chair seat) are evaluated as more distant, and less similar, by humans than by the computational model. Deeper insight into how humans made visual similarity judgments is explored in Sec. 4.4.1.

Interestingly, comparing heatmap DF to DV, no pairs in terms of function are considered more distant in the psychological than computational embedding spaces. This suggests that, across all pairs, more pairs were considered similar by humans. Parts 5 through 13 appear to be closely related by humans in functional similarity, as reflected by the presence of many lightly colored, high similarity, pairs in heatmap-PF. Inspecting Fig. 1(*b*), these parts correspond to chair seats, regardless of the type of chair the seats belong to (e.g., 1-, 2-, and 4-legged). In heatmap-CF, a range of distances is observed between parts 5–13, which is a consequence of how the stimuli were selected. The human criteria used to assess the functional similarity between these parts may be less nuanced and consider less information available in the shown stimuli. Evaluation criteria used by humans to make functional similarity judgments are detailed in Sec. 4.4.2.

4.2.2 Alignment Across Similarity Levels. In addition to the analysis of pairwise similarities, agreement between human and computational evaluations of similarities is also examined at a higher level. Rather than compare all pairwise similarities, the agreement of pairs assigned to a range of levels of similarity is

considered by using measures of percent agreement. The process of defining similarity levels is conceptualized in the example in Fig. 3, as described in Sec. 3.2.2. In Fig. 5, the relationship between pairwise distances and assigned similarity levels is shown, where overlapping pairs assigned to the same psychological and computational similarity levels are plotted darker. Percent agreement is used to measure this overlap in the number of pairs assigned, based on both embedding space distances, to the same similarity levels.

Relatively low values for percent agreement of 44% (53/120 pairs) and 43% (51/120 pairs) are observed for similarity levels assigned in terms of visual and functional similarity, respectively. This definition can be broadened to also include adjacent levels such that a pair assigned to level 2 based on distance in one embedding space is considered to agree when assigned to level 1 or 3 based on distance in the other. Accounting for adjacent levels is a popular modification to percent agreement, e.g., for ratings assigned on a 1–7 scale [46]. Adjusting for this modification, there is an 85% (102/120) overlap of pairs assigned to the same visual similarity levels, and an 82% (98/120 pairs) overlap in functional similarity levels. Alignment of pair assignment to similarity levels improves considerably when adjacent levels are included, indicating that the low percent agreement is not due to large discrepancies between embedding space distances. Computationally derived measures may therefore sufficiently represent human-evaluated similarities at a coarser view, i.e., when identifying near versus far or high versus low similarity between stimuli. However, misalignment is apparent at a more granular level, such as across five levels of similarity, which can be impactful if retrieval of stimuli at varying distances from an input is desired.

In Fig. 7, the percent agreement of pairs assigned to similarity levels in terms of psychological or computational embedding space distances is shown across similarity levels and by criteria represented (visual or functional). Adding insight to the low percent agreement observed, there appears to be variation across similarity levels. As indicated using Chi-square tests, the difference in percent agreement across similarity levels is observed to be statistically significant for both visual similarity ($\chi^2(4,\ N=120)=12.03,\ p=0.017$) and functional similarity ($\chi^2(4,\ N=120)=12.07,\ p=0.017$). The largest contribution to this difference appears to be due to the high overlap of pairs identified as sharing low similarity by both human and computational representations. These results demonstrate improved alignment for pairs sharing low similarity, suggesting that higher similarity between pairs may be driven by different factors considered by humans and AI.

Overall, high agreement between human and AI-based representations of visual and functional similarity was found, but not across all analyses. Specifically, our findings demonstrate high consistency in defining pairwise embedding space distances and high alignment in assigning pairs to broadly defined levels of low to high similarity. These results support the notion that existing AI-based models (e.g., those used in this work) represent human perspectives of visual and functional similarities effectively overall. However, successful retrieval of inspirational stimuli may rely on the alignment of similarities across more granularly defined levels not currently achieved. Observed differences and areas of misalignment therefore encourage further examination of stimulus features that may not currently be considered by computational methods. In the following subsection, we explore human evaluation of these representations through higher-level assessments.

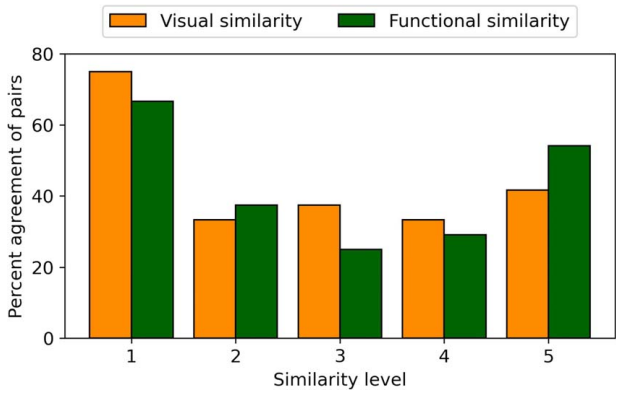

Fig. 7 Percent agreement of pairs assigned to the same levels by human and AI-based visual and functional similarities

4.3 Evaluation of Human and AI-Based Representations of Similarity via Categorization. To further assess the extent to which human and computational representations of similarity agree, human categorization of stimuli is used. Addressing the second research question posed, categories reveal how humans *evaluate* similarities between stimuli more holistically, compared to pairwise decisions. The relationship between category formation and extracted embedding space distances between stimuli reveals how distances, measuring human versus computational representations, relate to higher-level evaluations. The approach outlined in Sec. 3.2.3 is followed in this analysis. Represented in Fig. 8, mean proportions of pairs categorized in different groups are shown for each level of visual and functional similarity between parts (1 = low, 5 = high). Pairs assigned to each similarity level based on computational and psychological embedding space distances are plotted separately.

Mean proportion values and associated bootstrapped 95% confidence intervals are also detailed in Table 1. Confidence interval limits that do not cross the 50% threshold are bolded in Table 1 to indicate similarity levels associated with significant proportions of same or different-group categorization, also marked visually by boundaries in Fig. 8.

4.3.1 Categorization of Visually Similar Inspirational Stimuli. The categorization of stimuli related by visual similarity is first examined. When evaluating based on shape (Fig. 8(*a*)), parts with similarity levels up to 4 in terms of both psychological and computational embedding space distances are categorized into different groups above chance. Boundaries shown in Fig. 8(*a*) show humans and AI may consider pairs with similarity levels of 1–4 to be "too far." For parts categorized based on size (Fig. 8(*b*)), pairs with a similarity level up to 4 in the psychological embedding space and up to level 3 in the computational embedding space may be "too far." The "farther" boundary in size versus shape suggests that size is a less discriminating factor when evaluating visual similarity, according to computational embedding space distances. Combining both categorization criteria, stimuli sharing similarity up to level 4 are not grouped together and are thus associated with being "too far." These stimuli represent up to 60-80% of all pairs that may be discounted as being too visually dissimilar to be relevant, suggesting that only the most similar parts in both representational spaces may be similar enough in shape and size to group together.

4.3.2 Categorization of Functionally Similar Inspirational Stimuli. The relationship between categorization and functional similarity is also determined. Across both criteria used to categorize functional similarity, the most similar pairs in the psychological embedding space (level 5) are placed in different categories below chance (i.e., in the same categories above chance). High similarity pairs in this context may be "too near" or too obviously related to be relevant. This finding, not observed with high levels of computationally determined similarity, may reflect that these criteria align with how participants made pairwise similarity judgments.

For categories made based on the object a part belongs to (Fig. 8(*c*)), pairs with similarity levels up to 3 are considered "too far" based on both computational and psychological embedding space distances. A steep drop-off is observable between the mean proportion of pairs that are grouped in different categories with

Table 1 Proportions of pairs categorized in different groups across similarity levels

		Categorization criteria by visual similarity				Categorization criteria by functional similarity			
		Shape		Size		Object		Neighbors	
Embedding space	Similarity level	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Computational	1	0.95	(0.**89**, 1.1)	0.88	(0.**78**, 1.1)	0.96	(0.**91**, 1.1)	0.92	(0.**84**, 1.1)
	2	0.83	(0.**71**, 1.0)	0.79	(0.**67**, 0.98)	0.87	(0.**80**, 0.97)	0.88	(0.**80**, 0.99)
	3	0.76	(0.**65**, 0.93)	0.71	(0.**59**, 0.92)	0.70	(0.**59**, 0.85)	0.71	(0.**62**, 0.85)
	4	0.62	(0.**51**, 0.77)	0.58	(0.48, 0.72)	0.48	(0.38, 0.61)	0.53	(0.43, 0.67)
	5	0.39	(0.30, 0.51)	0.45	(0.35, 0.60)	0.39	(0.31, 0.50)	0.36	(0.28, 0.48)
Psychological	1	0.94	(0.**88**, 1.1)	0.83	(0.**75**, 1.0)	0.97	(0.**92**, 1.1)	0.92	(0.**84**, 1.1)
	2	0.89	(0.**81**, 1.0)	0.71	(0.**60**, 0.88)	0.94	(0.**87**, 1.1)	0.90	(0.**82**, 1.0)
	3	0.72	(0.**60**, 0.91)	0.76	(0.**67**, .90)	0.93	(0.**85**, 1.1)	0.87	(0.**78**, 0.99)
	4	0.60	(0.**52**, 0.72)	0.61	(0.**52**, 0.77)	0.40	(0.16, 0.79)	0.51	(0.32, 0.77)
	5	0.41	(0.31, 0.55)	0.50	(0.38, 0.71)	0.15	(0.02, 0.**34**)	0.19	(0.06, 0.**37**)

psychological similarity levels of 3 (0.93) compared to 4 (0.40). Boundaries separating parts that are "too far" to categorize together by neighboring parts (Fig. 8(*d*)) are also between similarity levels 3–4 in both computational and psychological embedding spaces. Across both categorization criteria, stimuli that are "too far" to be grouped together constitute 60% of pairs. These boundaries are "farther" than observed when categorizing by visual similarity, implying that parts do not need to be as "far" in visual as in functional similarity to be divided into separate groups. In terms of functional similarity, human and AI evaluations are found to be aligned. Qualitative post-task responses are analyzed in the following subsection to identify features of stimuli underlying human similarity assessments of non-text-based information.

4.4 Exploring Human Criteria for Similarity Assessments. Supporting the main findings of this work measuring the agreement between and evaluation of human and computational

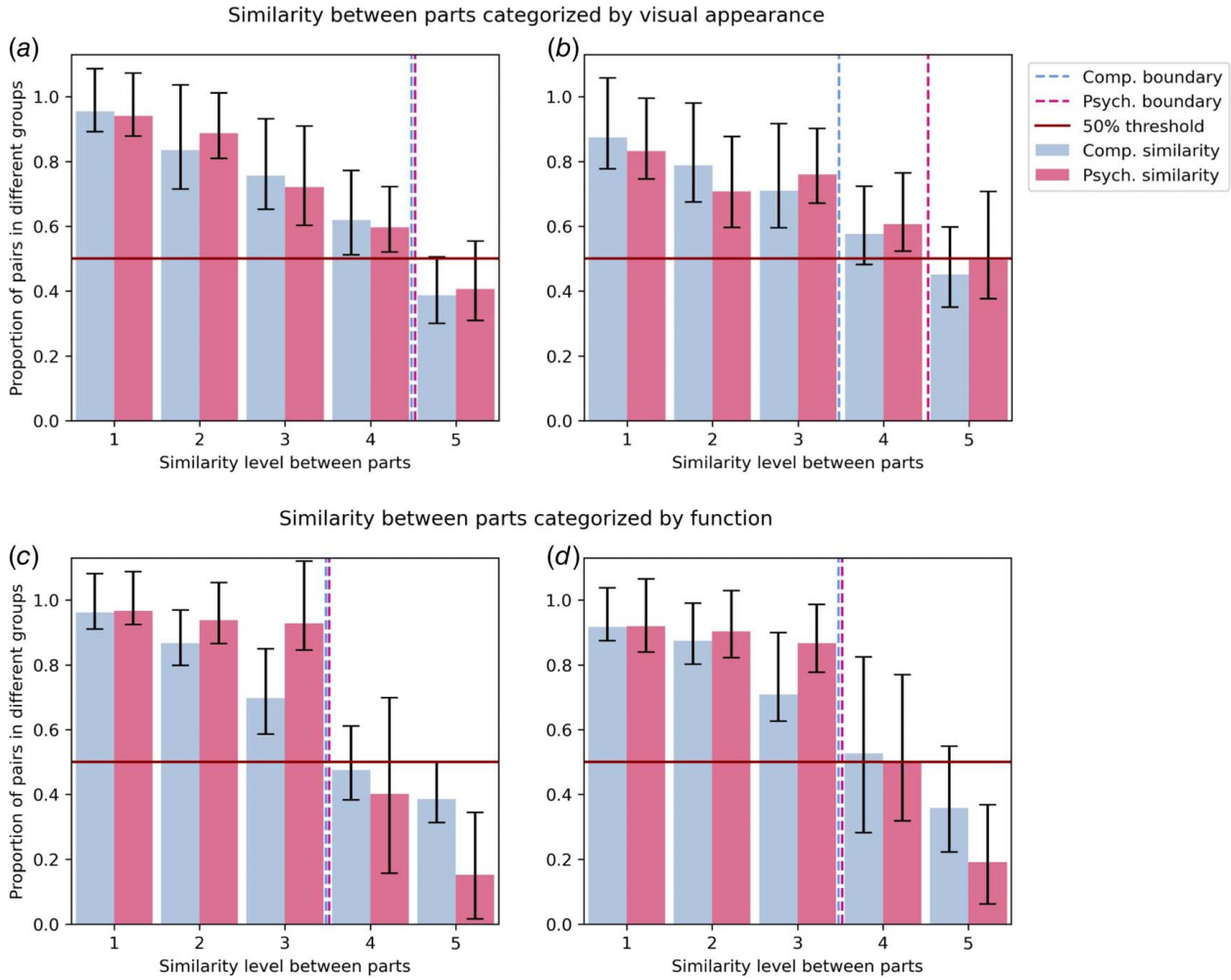

Fig. 8 Mean proportions (with bootstrapped 95% confidence intervals) of pairs in different groups when categorizing by visual similarity based on (*a*) shape and (*b*) size and by functional similarity based on (c) object belonging and (*d*) neighboring parts

representations of non-text-based characteristics of stimuli, an examination of human evaluation criteria is additionally conducted through qualitative analysis. This analysis can reveal areas where computational methods may improve to further align with human evaluation criteria, specifically when considering visual and functional similarity. The evaluation criteria discussed are obtained through the process described in Sec. 3.3.

4.4.1 Evaluation Criteria for Visual Similarity. Following a qualitative inductive category formation procedure, eight criteria for evaluating visual similarity were identified from 35 collected responses (one response was missing in data collection). Participants' responses were assigned to an average of two codes each. These criteria are listed in Table 2, where counts refer to the number of participants whose responses were coded into the relevant criteria. Multiple components of a response coded into the same criteria were counted only once.

Most frequently referenced (by 34 participants) was the shape of the part, which included references to specific shape geometries (e.g., rectangle, triangle, and circle), curvature or straightness, angularity (sharpness or roundedness), etc. Size was also highly referenced, by 24 participants, which mostly considered thickness (or flatness), dimensions (length, width, height, volume), and proportions between dimensions. More unique responses made reference to non-visual features, including the part's function, "how it interacts with the body" (coded as "user interaction"), or "what part of the chair it was on" (coded as "object", referring to the part's placement within the whole object). These criteria were provided by both participants who completed either the visual or functional similarity trials first.

As referenced in Sec. 4.2.1, parts 10 and 11 and parts 11 and 13 in the visual similarity stimulus set (Fig. 1(*a*)) are farther in distances according to human than AI-based representations. Differences in curvature of edges, angularity of corners, and continuity of surface area in these examples may contribute to greater perceived dissimilarities by humans. The boundary associated with different-group categorization of visual stimuli by size was also "nearer" in the psychological than computational embedding space. When modeling visual features using deep-neural networks, since multiple random perspectives of parts are considered, similarities e.g., edge thickness, may be differently emphasized by AI than by humans. Though the same criteria may be used (e.g., curvature or thickness), when applied to multiple perspectives compared to one isometric view, differences may be observed. The role of these criteria and information seen and emphasized by humans versus AI in contributing to their representations of visual features of stimuli is further discussed in Sec. 5.1.

4.4.2 Evaluation Criteria for Functional Similarity. To assess functional similarity between parts, responses provided by participants were coded into nine different criteria, shown in Table 3. Compared to criteria used to evaluate visual similarity, responses were more variable across 36 participants. The most frequently appearing criteria, referenced by 18 participants, was coded broadly as "interaction" and included both what might interact with the part and how. For example, body parts or objects that might be supported by the part were considered as well as the type of support provided (e.g., for vertical or horizontal loads). Interestingly, some responses were explicitly human centered when assessing part function in terms of interactions, including "I imagined how I would most often interact with the part" or "I categorized based on how a person would use it." Other criteria were more objective regarding the identity of the part (e.g., chair seat), the whole object (e.g., chair), or the primary use and function of the part (e.g., "for human seating"). These criteria align with the clustering of chair seats, chair backs, and tabletops in Fig. 1(*b*) and of pairwise distances in heatmap-PF in Fig. 6.

Visual attributes were referenced including size and shape and were acknowledged by some participants as useful if others were exhausted, with one participant stating that if other criteria did not decide the selection "the choice was mostly arbitrary and based on shape matching." More participants who completed functional similarity trials first referred to criteria based on visual features, suggesting that the ordering of tasks may have had the opposite effect than expected. Participants completing visual similarity trials first may have known not to rely on these features when assessing functional similarity. Participants' reference to visual features to form functional similarity assessments may support why a notably higher proportion of pairs with similarity level 3 in the psychological (0.87–0.93) than computational embedding space (0.70–0.71) are categorized separately (Fig. 8, Table 1). Parts may share high functional similarity in the psychological embedding space due to non-function-based features referenced in the triplet rating task, but then be categorized separately.

Physical attributes were also considered and classified under "material", including properties such as stiffness, stress fields, and softness or hardness. One participant noted that physical qualities "could influence how the user would feel using the object." A related sentiment was expressed by several responses categorized broadly as "type" in Table 3 to correspond to criteria based on whether the part appeared comfortable, provided cushioning, or in one example, "was more of a lounge type form fitting surface or if it was more of an upright type sitting on surface." These types of surfaces contrasted with those that appeared rigid and were more structural. Abstract criteria such as perceived comfort are impactful in the evaluation of the overall function served by the object part, but may be difficult to capture using AI, since corresponding visual attributes may not be obvious. Further considerations of representing abstract features of inspirational stimuli are discussed in Sec. 5.2.

Table 2 Human evaluation criteria of visual similarity

Criteria: description	Count
Shape: geometry, curvature, edges, angles	34
Size: thickness, dimensions, proportions, volume	24
Style: distinctive features, aesthetics	3
Surface area: presence of gaps, holes	2
Orientation: plane of part (vertical/horizontal)	2
Function: function of part	2
User interaction: how it interacts with body	1
Object: placement of part within object	1

Table 3 Human evaluation criteria of functional similarity

Criteria: description	Count
Interaction: how and what objects/body parts interact with part under use	18
Function: main use/purpose of part	12
Position: location of part within object	11
Shape: geometry or curvature	8
Size: thickness or flatness	8
Type: comfortable/lounge or structural/rigid	6
Object: identity of whole object	5
Material: stiffness, softness, stress fields	5
Neighboring parts: adjacent parts in object	3

5 Discussion

This work explores the representation and evaluation of non-semantic attributes of design artifacts, where 3D-model parts as inspirational stimuli are investigated through two research questions. Related to RQ1, the *consistency* and *alignment* of human and AI-based representations of visual and functional similarities were first compared. Computed pairwise similarities between stimuli were found to be consistent across both representations.

Low overall alignment of embedding space distances assigned to the same levels of similarity was found, but improved when a coarser measure of comparison was introduced. Addressing RQ2, the *evaluation* of human and AI-based representations of visual and functional similarities was examined by considering the categorization of parts with varying levels of similarity. Overall, increasing similarity between pairs was associated with decreasing proportions of pairs categorized in different groups. The levels of similarity at which stimuli were evaluated as "near" and "far" were found to mostly align for human and AI-based representations but differed for the evaluation criteria considered. Together, these findings indicate that, while consistent in defining similarity between parts, human and AI-based representations of visual and functional attributes of stimuli may differently reflect near and farness as perceived by humans. Implications for representing inspirational stimuli in terms of non-semantic information are discussed, as well as limitations of the present study and future directions.

5.1 Framing of Similarity Assessments. The first implication of representing non-text-based information of inspirational stimuli is the framing used when making similarity assessments. Humans assessed visual features of stimuli by interacting with 2D images of 3D-model parts taken from one isometric view. Instead, neural networks were trained on multiple images taken at random angles of each 3D-model part. Therefore, neural networks may equally represent the similarity between geometries and shapes from less obvious perspectives (e.g., the side edge of a chair seat) to the most common or meaningful views, from the human perspective. As noted in Table 2, the plane or orientation of the part in the shown image influenced participants' perceptions of visual similarity. Instead, the equal weighting of all perspectives in the neural networks may explain differences in pairwise computational and psychological embedding space distances between stimuli. For example, parts 10 and 11 in Fig. 1(*a*) are considered more similar by AI than parts 10 and 14, but the opposite relationships are true based on human representations (as shown in heatmap-PV in Fig. 6). While humans may have emphasized the rounded top edge in parts 10 and 14 as the most influential criteria determining their similarity, the AI-determined relationships also consider the straight edges seen from side views of each part. This retrieval of stimuli based on less obvious features of parts can lead to the discovery of seemingly distant inspiration, which may be helpful to designers but may also be distracting if too unexpected.

The issue of framing is also present in the representation of functional relationships. As presented in Sec. 4.4.2, multiple perspectives may be relevant to consider such as the interaction and relationship of the part with other parts, objects, or humans. Notably absent from the AI-based representation of functional similarity is the human-centered framing and identification of intended and afforded interactions with parts referenced by participants in this study. Instead, functional relationships are derived based on relationships to other parts within whole object assemblies. It is therefore suggested that data-driven methods should account for the framing of representation of stimuli that is most impactful or appropriate for the type of similarity modeled.

5.2 Information Captured at Varying Levels of Abstraction. A second implication of representing non-semantic attributes of inspirational stimuli is to capture information at varying levels of abstraction. In the example of 3D-model parts, several more concrete features of stimuli were represented by both humans and AI. These features included the identity of the object the part belonged to and neighboring parts within the same object assembly. While the neural networks used did not explicitly input semantic labels of parts or objects, these relationships were inferred through hierarchical information. When representing function, this concrete information regarding part and object identity was meaningful across both human and computational representations.

As revealed through qualitative insights in Sec. 4.4.2, more abstract information was also relevant. For example, participants referenced a product's style or its type in terms of level of comfort or use for lounging. This criterion incidentally aligned with computational embedding space distances since comfortable chairs (e.g., parts 6–8 in Fig. 1(*b*)) share visual attributes, which the function-based neural networks also incorporate. Prior work has relatedly employed visual information through shape grammars and 3D geometries of products to assess overall similarities in product style [47,48]. For humans, the visual style was found to be associated with a more conceptual meaning (i.e., the appearance of cushioned chairs with comfort), influencing the representation of functional relationships. Insights from this study encourage further understanding of the relationship between visual attributes and function and the use of AI to computationally define abstract, conceptual features of stimuli toward improved alignment with human representations.

5.3 Consideration of Higher-Level Evaluation of Representations. The above-discussed criteria influencing human versus AI representations of form and function-based features may impact how humans perceive distances between stimuli. We assessed these evaluations by determining the levels of similarity at which stimuli were considered too dissimilar to be categorized together. Findings from this analysis indicate the importance of considering the perception of same and different group belonging when representing design stimuli by non-semantic features since differences were observed depending on the criteria evaluated. The levels of similarity between parts associated significantly with different-group categorization were found to mostly align across human and AI-based definitions but were "nearer" when evaluating visual compared to functional attributes. As suggested in Sec. 5.1, factors impacting differences in observed categorization outcomes can be attributable to, e.g., visual features considered by AI that may be less obvious to humans or functional information known to humans, but unavailable when training AI.

It is also notable that up to 80% of stimulus pairs were found to be "too far." Definitions of "far" applied in prior work to retrieve conceptually related analogies tend to be more extreme, constituting crowd-sourced ideas occurring once or examples in the 10th percentile of text-based similarities to a design challenge [42,43]. When selecting "near" and "far" inspirational stimuli across various computed distances from a relevant source (e.g., designer's current idea or design prompt), the designer's perception of distance is proposed as an important factor to investigate. We suggest that it is important for computational methods to develop relationships between inspirational stimuli that agree with human representations across multiple measures of comparison. While human and AI-based representations were consistent across pairwise distances, misalignment was observed in assignment to five similarity levels, which may reflect more holistically how similarity relationships are understood.

5.4 Limitations and Future Work. This work presents a comparison and evaluation of human and AI-based representations of visual and functional similarity between 3D-model parts. We acknowledge the potential limitation of the present findings to the specific stimulus sets and similarity types assessed by participants. In this work, a limited set of stimuli was utilized to reduce the significant complexity of this study and to make the task of assessing the similarity of non-textual information tractable for humans. Future work might explore the generalizability of these findings to additional examples and contexts. Several features of stimuli influencing similarity assessments, e.g., the number of different objects presented or perception of comfort, may be specific to the chosen stimulus sets. As well, although the tasks conducted were not explicitly design-relevant, the design experience of participants may impact their judgments of relationships between the shown stimuli. The number of participants involved in this study may

impact the results presented in two main ways. The first potential impact is on the constructed psychological embedding spaces. Although the collection of data for 30% of all unique triplet ratings is recommended according to prior work [36], the models are trained without covering all possibilities, limiting the representativeness of the embeddings. Categorization outcomes are presented as an average of proportions out of 24 pairs across 36 participants. Bias introduced by individual differences when forming categories can impact the present results, since participants' categories are not identical. The impact of misalignment across participants in the categorization of pairs in each similarity level is observable in the wide 95% confidence intervals around mean proportions observed. Continued work utilizing these methods is suggested to involve a larger sample size and an examination of the specific stimuli belonging to formed groups.

Furthering this study, future work is encouraged to investigate additional stimuli containing multi-modal information from which to extract and define non-text-based similarities and study in a design context. By gaining more knowledge regarding how these similarities are perceived and evaluated, new sources of information and inspiration can be more effectively engaged with and utilized by designers. Toward this aim, efforts to define a more holistic definition of similarity are encouraged, which appropriately account for varied features of stimuli (semantic and non-semantic) and components of similarity. Integrating human perspectives into design support tools through the development of personalized AI models is also recommended to improve human–AI collaboration. The retrieval of and interaction with design artifacts across various forms can be enabled by computational platforms developed by the broader design research community.

6 Conclusions

Design tools increasingly enable collaboration, communication, and information transfer between humans and AI when engaging with complex design artifacts such as 3D-CAD models. The growing interest in representing design artifacts such as these across multiple modalities, in contrast to text-based labels or descriptions only, also motivated the present work. The context of use specifically considered was the retrieval of inspirational stimuli in the form of 3D-model parts. Successful human–AI collaboration in search and retrieval tasks requires that the relationship between an input specified by a human agent and the output selected by an AI agent map onto shared representations. For instance, when utilizing computational support to search for distant stimuli to inspire a design idea, for seamless integration of AI into this process, human and AI perspectives of distance should agree. By examining similarities and differences across human and AI-based representations of multi-modal design information, factors promoting and inhibiting this collaboration can be identified.

The aim of this work was to compare and evaluate representations of non-text-based features of 3D-model parts derived from human evaluations and deep-learning approaches. Using measures of consistency and alignment, high agreement between humans and AI was found for representing both visual and functional similarities, at a coarse level of analysis. In both cases, this agreement was highest for identifying pairs of stimuli sharing low similarity, suggesting that humans and AI agree on identifying obvious differences, but less on features driving increased similarity between pairs. The framing of how humans and AI assess features of parts as well as the representation of abstract information are proposed as factors that may need further consideration when modeling visual and functional similarities using computational methods. Additionally, the levels of similarity between parts associated with categorization of parts into separate groups mostly aligned when defined by human and AI-based representations and thus similarly reflected the near and farness between stimuli perceived by humans. However, stimuli categorized separately were "nearer" in visual than functional similarity, suggesting that, for different criteria, the perception of the same level of similarity can vary. Overall, this work presents an exploration of less explored, increasingly relevant, definitions of similarity of inspirational stimuli based on form and function-based attributes. Findings from this study encourage further research on representing multi-modal information to better understand and support effective representation of design artifacts across relevant design agents.

Acknowledgment

The authors thank the participants who completed the study and Dr. Forrest Huang for developing the computational models used in this work. This work was supported by the National Science Foundation under grant 2145432—CAREER. This paper is based on preliminary work published in the proceedings of the 2023 International Design Engineering and Technical Conferences [4] and the 24th International Conference on Engineering Design [5].

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Jiang, S., Hu, J., Wood, K. L., and Luo, J., 2022, "Data-Driven Design-by-Analogy: State-of-the-Art and Future Directions," ASME J. Mech. Des., **144**(2), p. 020801.

[2] Kwon, E., Huang, F., and Goucher-Lambert, K., 2022, "Enabling Multi-Modal Search for Inspirational Design Stimuli Using Deep Learning," Artif. Intell. Eng. Des. Anal. Manuf., **36**(1), p. e22.

[3] Kwon, E., Rao, V., and Goucher-Lambert, K., 2023, "Understanding Inspiration: Insights Into How Designers Discover Inspirational Stimuli Using an AI-Enabled Platform," Des. Stud., **88**(1), p. 101202.

[4] Kwon, E., and Goucher-Lambert, K., 2023, "Similarities and Differences in Human vs. Computational Representations of Non-semantic Inspirational Design Stimuli," Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Boston, MA, Aug. 20–23, PAper No. IDETC2023–115108.

[5] Kwon, E., and Goucher-Lambert, K., 2023, "Examining the Boundary Between Near and Far Design Stimuli," Proc. Des. Soc., **3**(1), pp. 1725–1734.

[6] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K., 2013, "The Meaning of "Near" and "Far": The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output," ASME J. Mech. Des., **135**(2), p. 021007.

[7] Sio, U. N., Kotovsky, K., and Cagan, J., 2015, "Fixation or Inspiration? A Meta-Analytic Review of the Role of Examples on Design Processes," Des. Stud., **39**(1), pp. 70–99.

[8] Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., and Kotovsky, K., 2011, "On the Benefits and Pitfalls of Analogies for Innovative Design: Ideation Performance Based on Analogical Distance, Commonness, and Modality of Examples," ASME J. Mech. Des., **133**(8), p. 081004.

[9] Goucher-Lambert, K., Gyory, J. T., Kotovsky, K., and Cagan, J., 2020, "Adaptive Inspirational Design Stimuli: Using Design Output to Computationally Search for Stimuli That Impact Concept Generation," ASME J. Mech. Des., **142**(9), p. 091401.

[10] Borgianni, Y., Rotini, F., and Tomassini, M., 2017, "Fostering Ideation in the Very Early Design Phases: How Textual, Pictorial and Combined Stimuli Affect Creativity," Proceedings of of the 21st International Conference on Engineering Design, Vancouver, BC, Canada, Aug. 21–25, The Design Society, pp. 139–148.

[11] Borgianni, Y., Maccioni, L., Fiorineschi, L., and Rotini, F., 2020, "Forms of Stimuli and Their Effects on Idea Generation in Terms of Creativity Metrics and Non-obviousness," Int. J. Des. Creativity Innov., **8**(3), pp. 147–164.

[12] Linsey, J. S., Clauss, E. F., Kurtoglu, T., Murphy, J. T., Wood, K. L., and Markman, A. B., 2011, "An Experimental Study of Group Idea Generation Techniques: Understanding the Roles of Idea Representation and Viewing Methods," ASME J. Mech. Des., **133**(3), p. 031008.

[13] Gonçalves, M., Cardoso, C., and Badke-Schaub, P., 2014, "What Inspires Designers? Preferences on Inspirational Approaches During Idea Generation," Des. Stud., **35**(1), pp. 29–53.

[14] Linsey, J., Wood, K., and Markman, A., 2008, "Modality and Representation in Analogy," Artif. Intell. Eng. Des. Anal. Manuf., **22**(2), pp. 85–100.
[15] Han, J., Shi, F., Chen, L., and Childs, P. R. N, 2018, "The Combinator: A Computer-Based Tool for Creative Idea Generation Based on a Simulation Approach," Des. Sci., **4**(1), p. e11.
[16] Malaga, R. A., 2000, "The Effect of Stimulus Modes and Associative Distance in Individual Creativity Support Systems," Decis. Support Syst., **29**(2), pp. 125–141.
[17] Menezes, A., and Lawson, B. R., 2006, "How Designers Perceive Sketches," Des. Stud., **27**(5), pp. 571–585.
[18] Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., and Wood, K., 2014, "Function Based Design-by-Analogy: A Functional Vector Approach to Analogical Search," ASME J. Mech. Des., **136**(10), p. 101102.
[19] Fu, K., Cagan, J., Kotovsky, K., and Wood, K., 2013, "Discovering Structure in Design Databases Through Functional and Surface Based Mapping," ASME J. Mech. Des., **135**(3), p. 031006.
[20] Song, H., Evans, J., and Fu, K., 2020, "An Exploration-Based Approach to Computationally Supported Design-by-Analogy Using D3," Artif. Intell. Eng. Des. Anal. Manuf., **34**(4), pp. 444–457.
[21] Song, H., and Fu, K., 2022, "Design-by-Analogy: Effects of Exploration-Based Approach on Analogical Retrievals and Design Outcomes," ASME J. Mech. Des., **144**(6), p. 061401.
[22] Han, J., Sarica, S., Shi, F. and Luo, J., 2022, "Semantic Networks for Engineering Design: State of the Art and Future Directions," ASME J. Mech. Des., **144**(2), p. 020802.
[23] Sarica, S., Song, B., Luo, J., and Wood, K., 2021, "Idea Generation With Technology Semantic Network," Artif. Intell. Eng. Des. Anal. Manuf., **35**(3), pp. 265–283.
[24] Zhang, Z., and Jin, Y., 2020, "An Unsupervised Deep Learning Model to Discover Visual Similarity Between Sketches for Visual Analogy Support," Proceedings of the ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Virtual, Online, Aug. 17–19, p. V008T08A003.
[25] Zhang, Z., and Jin, Y., 2021, "Toward Computer Aided Visual Analogy Support (CAVAS): Augment Designers Through Deep Learning," Proceedings of the ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Virtual, Online, Aug. 17–19, p. V006T06A057.
[26] Jiang, S., Luo, J., Ruiz-Pava, G., Hu, J., and Magee, C. L., 2020, "A Convolutional Neural Network-Based Patent Image Retrieval Method for Design Ideation," Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Virtual, Online, Aug. 17–19, p. V009T09A039.
[27] Jiang, S., Luo, J., Ruiz-Pava, G., Hu, J., and Magee, C. L., 2021, "Deriving Design Feature Vectors for Patent Images Using Convolutional Neural Networks," J. Mech. Des., **143**(6), p. 061405.
[28] Kazi, R. H., Grossman, T., Cheong, H., Hashemi, A., and Fitzmaurice, G., 2017, "DreamSketch: Early Stage 3D Design Explorations With Sketching and Generative Design," Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST'17, Quebec City, QC, Canada, Oct. 22–25, pp. 401–414.
[29] Arora, R., Darolia, I., Namboodiri, V. P., Singh, K., and Bousseau, A., 2017, "SketchSoup: Exploratory Ideation Using Design Sketches," Comput. Graph. Forum, **36**(8), pp. 302–312.
[30] Chaudhuri, S., and Koltun, V., 2010, "Data-Driven Suggestions for Creativity Support in 3D Modeling," CM Trans. Graph., **29**(6), pp. 1–10.
[31] Kim, J., Maher, M. L., and Siddiqui, S., 2021, "Collaborative Ideation Partner: Design Ideation in Human-AI Co-Creativity," Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications (CHIRA), Virtual, Online, Oct. 28–29, pp. 123–130.
[32] Kim, J., and Maher, M. L., 2023, "The Effect of AI-Based Inspiration on Human Design Ideation," Int. J. Des. Creativity Innov., **11**(2), pp. 81–98.
[33] Kwon, E., Rao, V., and Goucher-Lambert, K., 2022, "Investigating the Roles of Expertise and Modality in Designers' Search for Inspirational Stimuli," Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, St. Louis, MO, Aug. 14–17, p. V006T06A015.
[34] Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., and Su, H., 2019, "Partnet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, June 16–20.
[35] Nandy, A., and Goucher-Lambert, K., 2022, "Do Human and Computational Evaluations of Similarity Align? An Empirical Study of Product Function," ASME J. Mech. Des., **144**(10), p. 041404.
[36] Ahmed, F., Ramachandran, S., Fuge, M., Hunter, S., and Miller, S., 2019, "Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel," ASME J. Mech. Des., **141**(2), p. 021102.
[37] Roads, B., and Mozer, M., 2019, "Obtaining Psychological Embeddings Through Joint Kernel and Metric Learning," Behav. Res., **51**(5), pp. 2180–2193.
[38] Chaudhari, A. M., Bilionis, I., and Panchal, J. H., 2019, "Similarity in Engineering Design: A Knowledge-Based Approach," Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, Aug. 18–21, p. V007T06A045.
[39] Tversky, A., 1977, "Features of Similarity," Psychol. Rev., **84**(4), pp. 327–352.
[40] Tangelder, J.W.H., and Veltkamp, R.C., 2004, "A Survey of Content Based 3D Shape Retrieval Methods," Proceedings Shape Modeling Applications, Genova, Italy, June 7–9, pp. 145–156.
[41] Huang, F., Schoop, E., Ha, D., Nichols, J., and Canny, J., 2021, "Sketch-Based Creativity Support Tools Using Deep Learning," *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, Y Li, and O Hilliges, eds., Springer International Publishing, Cham, pp. 379–415.
[42] Goucher-Lambert, K., and Cagan, J., 2019, "Crowdsourcing Inspiration: Using Crowd Generated Inspirational Stimuli to Support Designer Ideation," Des. Stud., **61**(1), pp. 1–29.
[43] Chan, J., Dow, S. P., and Schunn, C., 2015, "Do the Best Design Ideas (Really) Come From Conceptually Distant Sources of Inspiration?" Des. Stud., **36**(1), pp. 31–58.
[44] Cooke, T., Jakel, F., Wallraven, C., and Bülthoff, H. H., 2007, "Multimodal Similarity and Categorization of Novel, Three-Dimensional Objects," Neuropsychologia, **45**(3), pp. 484–495.
[45] Mayring, P., 2004, *Qualitative Content Analysis*, Sage, Great Britain, 266–269.
[46] Stemler, S. E., 2004, "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability," Pract. Assess. Res. Eval., **9**(1), p. 4.
[47] Culbertson, T. D., and Simpson, T. W., 2014, "Using Shape Grammars to Identify Salient Features in Support of Product Family Design," Proceedings of the ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Buffalo, NY, Aug. 17–20, p. V007T07A036.
[48] Ranscombe, C., Kinsella, P., and Blijlevens, J., 2017, "Data-Driven Styling: Augmenting Intuition in the Product Design Process Using Holistic Styling Analysis," ASME J. Mech. Des., **139**(11), p. 111417.