Proceedings of the ASME 2022 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2022
August 14-17, 2022, St. Louis, Missouri, USA

DRAFT DETC-91269

HEY, AI! CAN YOU SEE WHAT I SEE? MULTIMODAL TRANSFER LEARNING-BASED DESIGN METRICS PREDICTION FOR SKETCHES WITH TEXT DESCRIPTIONS

Binyang Song*
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: binyangs@mit.edu

Scarlett Miller
School of Engineering Design and Innovation
The Pennsylvania State University
State College, Pennsylvania, 16802
Email: shm13@psu.edu

Faez Ahmed
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: faez@mit.edu

ABSTRACT

Measuring design creativity is an indispensable component of innovation in engineering design. Properly assessing the creativity of a design requires a rigorous evaluation of the outputs. Traditional methods to evaluate designs are slow, expensive, and difficult to scale because they rely on human expert input. An alternative approach is to use computational methods to evaluate designs. However, most existing methods have limited utility because they are constrained to unimodal design representations (e.g., texts or sketches) and small datasets. To overcome these limitations, we propose a multimodal transfer learning-based machine learning model to predict five design metrics: drawing quality, uniqueness, elegance, usefulness, and creativity. The proposed model utilizes knowledge from large external datasets through transfer learning and simultaneously processes text and sketch data from early-phase concepts through multimodal learning. Through six unimodal models using only texts or sketches, we show that transfer learning improves the predictive validity of text learning and sketch learning by 2%-18% and 9%-24%, respectively, for design metric evaluation. By comparing our multimodal model with the best unimodal models, we demonstrate that joining unimodal text and sketch learning models further increases the predictive validity of the approach by 4%-10%. The proposed models are generalizable to many application contexts beyond design concepts. Our findings highlight the importance of analyzing designs from multiple perspectives for design assessment. Finally, we discuss the challenges and opportunities in developing AI models for design metric evaluation.

1. INTRODUCTION

Design creativity evaluation is a core component of the innovation process in engineering design [1, 2]. Numerous design ideas are generated at early design stages, which creates the need for an effective and efficient creativity evaluation to facilitate informed decision making and boost designers' creative and innovative behaviors [3, 4]. However, design creativity evaluation is challenging for two primary reasons. First, design ideas are often complex, involving heterogeneous representations, such as sketches, texts, and 3-dimensional (3D) models. Second, compared to more straightforward tasks (e.g., classification and object recognition), it requires a profound and comprehensive understanding of object functions, behaviors, structures, and aesthetics based on the information scattered in multiple representations. Accordingly, many of the existing methods rely on intensive human inputs [5–7], which makes the methods inherently subjective, resource-demanding, unscalable, and subject to human fatigue [8]. With the advances in artificial intelligence (AI), prior work has shown promise for leveraging AI to evaluate simple unimodal design ideas, such as sketches [9–11] and

*Address all correspondence to this author.

texts [9, 12]. In contrast, the use of AI for evaluating multimodal design ideas remains largely unexplored. Moreover, training artificial neural networks (ANNs) for complex tasks (e.g., design metric evaluation) requires a large volume of labeled data. As mentioned above, manual evaluation for labeling design ideas is expensive and time-consuming. Therefore, the available labeled datasets are often small in volume, which poses an additional challenge to train ANNs for this purpose.

To address the identified challenges, we propose a multimodal transfer learning model for predicting design metrics in this paper. Multimodal learning enables us to learn across multiple modalities to solve more complex tasks that combine more than one data mode [13]. Transfer learning improves the performance of models trained on small datasets by transferring knowledge learned from other related domains where data is readily available [14]. We develop and validate the proposed models using a set of milk frother design ideas represented by sketches and text descriptions along with their design ratings provided by experts. The contributions of this paper are:

1. We compare three text embedding models to predict five design metrics (i.e., creativity, uniqueness, drawing quality, elegance, and usefulness) by capturing semantic features from design descriptions and show that the transformer-based bidirectional encoder representations from transformers (BERT) [15] model performs the best;
2. We compare three convolutional neural network-based sketch embedding models for design metric prediction, which capture visual features from design sketches, and show that the pre-trained Inception [16] model performs the best;
3. We show that transfer learning from large external datasets is effective in improving design metric prediction performance for small datasets by using an Inception model pre-trained on 50 million sketches from the QuickDraw dataset;
4. We develop a multimodal learning model that embeds sketches and texts simultaneously for design metric prediction and show that it performs significantly better than all text- or sketch-based machine learning models;
5. We compare the predictability of these design metrics and show that sketches are more informative for predicting drawing quality, elegance, and usefulness, while design descriptions are more informative for predicting uniqueness.

The remainder of this paper is organized as follows. Section 2 provides a detailed review of the relevant building blocks of the proposed model. The labeled milk frother design ideas for training the model, the associated data pre-processing modules, and the key components of the multimodal transfer learning model are introduced in Section 3. Section 4 reports and discusses the multimodal transfer learning model performance and summarizes the challenges and opportunities in AI-based design metric evaluation. Section 5 concludes this paper by highlighting the findings and contributions of this paper.

2. BACKGROUND

Design ideas are most commonly available as free-hand sketches and their text descriptions. Accordingly, we review the relevant approaches of embedding sketches and texts in this section. We also discuss transfer learning for tackling the common issue of small datasets in design and survey the prior work on machine learning approaches for multimodal data.

2.1 Sketch and Text Embeddings

In this study, the sketches are drawn on paper solution sheets, which can be stored and processed as static pixel-based spaces. We focus on convolutional neural networks (CNNs) [17]-based models for encoding sketches due to their effectiveness in embedding pixel data (e.g., images and sketches). While these models are mostly used for images, our focus is on free-hand sketches, which are fundamentally different from realistic photo images. Free-hand sketches have both unique challenges (e.g., highly sparse, abstract, and designer-dependent) and advantages (e.g., lack of background and use of iconic representation) [18]. Previous studies have employed both customized CNNs [19] and standardized CNNs (e.g., ResNet [20], VGG [20], and Inception [21]) for sketch classification and similarity search. Researchers also studied the differences in hyperparameters between CNNs for encoding images and sketches, respectively [22]. Optimized CNN models have also been explored by incorporating an additional channel for learning shape [23] or contour [24] information of sketches to improve model performance. Additionally, when created using touchscreen devices, sketches can also be rendered as dynamic stroke coordinate spaces or graph spaces. This affords the use of different ANNs, such as recurrent neural networks (RNNs) [25] and graph neural networks (GNNs) [26]. While most prior work on sketches focuses on classifying sketches into categories, we study a more difficult learning task: predicting design ratings, which can be viewed as a regression problem.

To analyze text data, modern machine learning primarily focuses on natural language processing (NLP) methods that encode text data as continuous vectors. Among such methods, we concentrate the review on transformer-based language embedding models [27]. These models have proven exceedingly effective in many benchmark NLP applications(e.g., translation and search), compared to prior text embedding models such as word2vec [28], Global vector for word representation [29], and bidirectional long short-term memory [30]. The transformer-based models hold an edge from two perspectives: (1) the transformer encoder learns input sequences of words bidirectionally by reading entire sequences at once instead of unidirectionally,

which enables it to understand the contexts of single words more comprehensively; and (2) the transformer utilizes a self-attention mechanism to strengthen the learning of contextual relations between words [27]. Trained on multiple large text databases for multiple tasks, universal sentence encoder (USE) [3] and BERT [15] are the most popular transformer-based text embedding models. Compared to USE using unidirectional transformer decoders, BERT adopts bidirectional transformer decoders and a deeper network architecture [15], leading to better embeddings in general. As most other models were designed for generating word-level embeddings, the transformer-based models also surpass them in terms of generating sentence-level embeddings [31]. In our work, we use both USE and BERT models to encode design descriptions and predict the overall ratings.

2.2 Transfer Learning

Labeled design datasets are often small in size because it is expensive to evaluate designs through expert-based approaches or simulation. Transfer learning provides a promising approach to tackle this issue. Transfer learning is a machine learning technique that aims at improving the performance of target models within target domains by transferring the knowledge learned from different but related source domains [14]. With knowledge transferred, the learning performance can be improved while less labeled data is required in a target domain. Knowledge transfer from one domain to another relies on the similarities and relevance between the domains [32]. However, knowledge transfer is not always beneficial (e.g., the pronunciation transfer from Spanish to French is misleading). Negative transfer occurs when transfer learning has negative impacts [33]. Positive transfer learning is built on capturing the transferable and beneficial knowledge elements across domains [33]. In deep learning, the most common approach of transferring knowledge is to share or transfer the parameters of the models trained on one or multiple source domains to the target domain. The bridge for knowledge transfer is the same pivot features shared by the source data and the target data [14].

Transfer learning has been proven effective in many machine learning tasks such as image translation. Researchers in the engineering design domain have explored this topic as well. For example, Whalen and Mueller [34] utilized knowledge transfer between different topologies of truss designs in constructing a surrogate model for design evaluation. Cheng et al. [35] leveraged complementary knowledge transferred among multiple optimization tasks for coevolutionary multitasking to realize concurrent global optimization. Raina et al. [36] proposed a probabilistic mechanism to transfer design strategies from humans to computers and across design problems. Pandita et al. [37] transferred knowledge from legacy data to estimate process parameters for new materials and machines in additive manufacturing (AM). Huang et al. [38] conducted multi-fidelity surrogate modeling of melt pool in AM by transferring knowledge from high-fidelity simulation data to a model trained on low-fidelity data. This paper demonstrates the positive knowledge transfer from large generic text and sketch datasets to our target milk frother design dataset.

2.3 Multimodal Learning

Because a single design can be represented in multiple modes, deep learning models that can capture design features simultaneously from multimodal representations are required. To leverage the complementarity of multimodal data, unimodal models are first constructed to encode different design representations separately. Then, they are fused to generate joint embeddings through shared layers at earlier or later training stages. Researchers classify different fusion mechanisms employed in previous studies into three categories [13]. The first uses simple operations, such as concatenation [39, 40] and weighted sum [40, 41]. For weighted sum, the pre-trained embeddings for all modalities need to have the same dimension and be rearranged in an order suitable for element-wise addition [42]. The second category employs attention mechanisms to generate joint embeddings, which can dynamically learn the alignment between multimodal features by attending the features from one modality using the features from another modality [43]. Multiple attention heads can be applied to preserve more comprehensive information [44]. The third category learns joint embeddings through a parallel reconstruction process, which trains multiple unimodal reconstruction models, such as auto-encoders (AEs), simultaneously with shared layers [45,46]. By minimizing the reconstruction loss, the correlation and mutual information between the unimodal models increase [47, 48]. Additionally, bilinear pooling-based approaches have also been adopted to fuse embeddings from multiple modalities, for which interested readers can refer to [49].

As an emerging topic, researchers have explored multimodal learning for a variety of tasks, such as text-to-image or image-to-text generation [50], text or image classification [51]. However, its application in engineering design is quite limited. Recently, Yuan, Mation, and Moghaddam [52] reported an attention-based multimodel model learning from images and texts for design concept evaluation. Their model analyzed detailed orthographic product images and textual product descriptions to predict product user ratings. Unlike the design metrics evaluated in this paper, user ratings reflect the sentimental feedback towards a design rather than the technical assessment, such as creativity and usefulness. Additionally, this study focuses on design idea evaluation at an earlier stage when little design detail is available.

Based on the state-of-the-art machine learning techniques reviewed above, we propose a pioneer multimodal transfer learning model for design metric prediction using texts and sketches.

3. DATA AND METHOD

This section introduces the data used and the multimodal transfer learning model proposed for design metric prediction in this study. We pose the design metric prediction problem as a regression task. The multimodal transfer learning model can be broken into four modules, as depicted in Fig. 1. First, the raw design data is pre-processed and converted into sketches and texts processable for computers. The second and third modules entail two unimodal learning models that learn features from the sketches and texts, respectively. Both are pre-trained to transfer knowledge from a respective large dataset. The two unimodal transfer learning models are joined together in the last module to construct the multimodal transfer learning model. Each module is introduced separately.

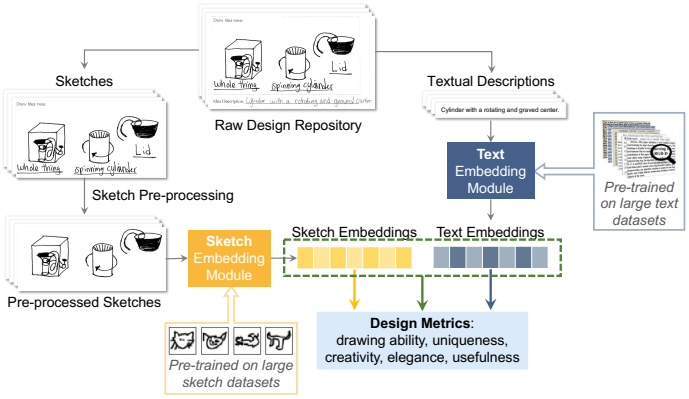

FIGURE 1. THE OUTLINE OF THE PROPOSED MULTIMODAL LEARNING MODEL FOR DESIGN CREATIVITY EVALUATION.

3.1 Data

The multimodal transfer learning model is developed based on a repository of 1,112 milk frother design ideas[1] from prior work [53]. These design ideas were generated in response to a design challenge that asked participants[2] to design an innovative device that froths milk in a short amount of time. Each design idea was created and recorded freehand on a solution sheet, as seen on the left of Fig. 2. The solution sheet consists of a drawing area denoted by "Draw Idea Here" for sketching the idea and a description area starting with "Idea Description" for adding a text description. The collected ideas were then scanned into electronic files. Following the consensual assessment technique (CAT) [5], two design experts were asked to evaluate the design ideas in terms of five design metrics: (i) drawing quality, (ii) uniqueness, (iii) usefulness, (iv) creativity, and (v) elegance. Particularly, drawing quality reflects clean lines, accurate proportions, appropriate shading, etc. of a sketch. Uniqueness refers to how original or surprising the idea is. Usefulness refers to how logical, practical, valuable, and understandable the idea is. Creativity is defined by ideas that are both of high quality and novelty. Elegance indicates the simplicity, clear insight, and concise presentation of an idea [54]. Among them, creativity, uniqueness, and usefulness are commonly used CAT creativity metrics, while drawing quality and elegance are also commonly used for evaluating design sketches. The inter-rater reliability test shows that the ratings from the two experts achieve a median Spearman correlation of 0.76 for the five metrics, with the highest value being 0.88 and the lowest value being 0.44. Twenty-six out of the 1,112 labeled milk frother design ideas are excluded from this study due to low data quality after scanning or missing any of the five design metrics. As a result, 1,086 design ideas constitute the dataset for training and validating the proposed model.

3.2 Data Pre-Processing

The raw data involves two representation modes: handwritten text descriptions and freehand sketches. The text descriptions are first extracted from the solution sheets manually and recorded as strings readable by computers. The text descriptions refer to the texts written in the description area, excluding the text annotations within the drawing area. The text descriptions' maximal, minimal, and average word lengths are 54, 1, and 9, respectively. The main tasks of data pre-processing reside in sketch pre-processing, aiming to remove irrelevant information and noise from the drawing area. The whole process comprises five steps using the OpenCV Python package. The flowchart on the right of Fig. 2 summarizes the process and shows the output from each step.

Crop Sketches Since the sketches (original size = $1,650 \times 1,275$) were scanned from the solution sheets, the positions of the drawing area vary across the sketch files. To locate the drawing area in each file, we first locate the position of the printed letter "D" from the string "Draw Idea Here". Then, the drawing area is defined as a fixed region surrounding the letter "D". Accordingly, we crop the original sketches to a smaller size ($525 \times 1,030$) to only retain the drawing area. After this, the printed text "Draw Idea Here" is removed from each sketch as irrelevant information by filling the corresponding area with white.

Convert to Grayscale In our dataset, most sketches were drawn with black pens, while a small portion of the sketches were drawn using color pens, often one color for the entire sketch. Due to the variability of colors in scanning the sketches, we convert all sketches into grayscale.

Remove Hand-Written Texts The presence of text annotations in the drawing area can distract the sketch learning model from learning pivot features from the sketches. Accordingly, we

[1]Set 1: https://sites.psu.edu/creativitymetrics/2018/07/18/milkfrother/. Set 2: https://sites.psu.edu/creativitymetrics/2018/08/23/milk-frother-industry/

[2]The participants are first-year engineering design students.

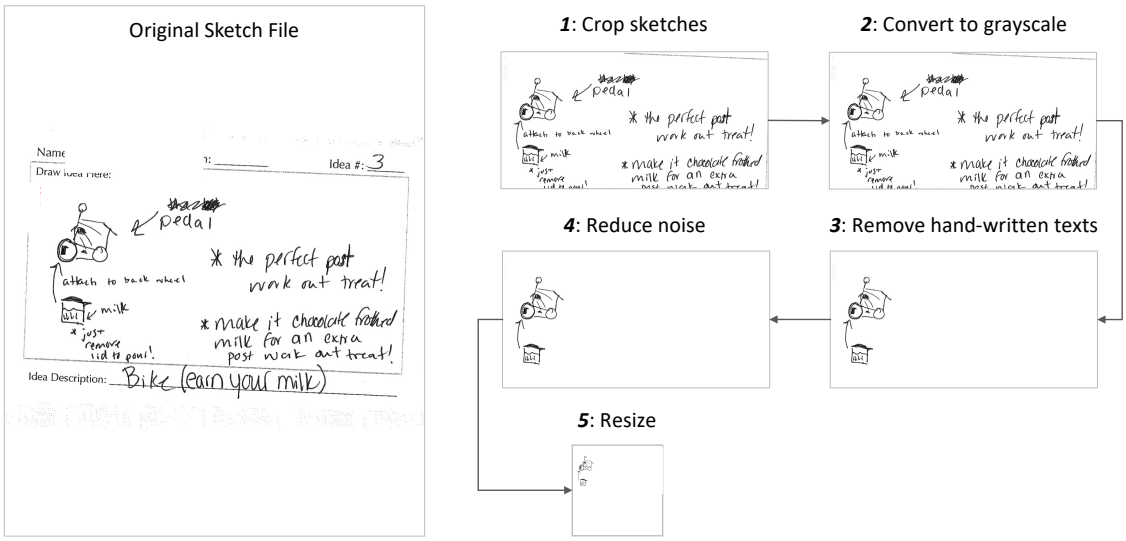

FIGURE 2. THE FLOWCHART OF SKETCH PRE-PROCESSING FOR AN EXAMPLE SKETCH FROM OUR DATASET.

develop a semi-automatic model to remove hand-written texts from the drawing area. First, contours within the drawing area are detected based on pixel gradients. Since the contours containing sketches and the contours containing texts are likely to have different features (e.g., shapes and average pixel values), the model automatically detects these features and removes the contours deemed as texts. Then, we manually remove the text annotations that the model can not capture.

Reduce Noise Since the sketches were drawn freehand and scanned to electronic files, there is noise in the drawing area. We use a denoising model to remove Gaussian noise from the sketches [55].

Resize Sketches For training CNNs, a large sketch size requires a high computational cost. In contrast, a small sketch size can cause significant information loss during resizing the original sketches, impairing the performance of the prediction model. Considering this trade-off, we resize the sketches ($525 \times 1,030$) to 158×309, i.e., 30 percent of the original size[3]. Furthermore, to enable knowledge transfer from the source domain to the target domain, the data sample from both domains should have the same shape. In this study, the source domain is defined by the images from the ImageNet dataset and the sketches from the QuickDraw dataset, which are square. Accordingly, we modify the sketches in our dataset to a square shape (309×309) by filling the bottom area (151×309) with white background.

[3]For reference, ImageNet and QuickDraw are commonly used datasets for training CNNs. The sizes of the images or sketches from them are 256×256 (or 224×224) and 28×28, respectively.

3.3 Models

In this subsection, we first introduce six unimodal learning models (three for text and three for sketches) for embedding the texts and sketches, respectively. We then explain how they are joined to construct the multimodal transfer learning model.

Text Learning Models We experiment with three different models for text embedding, including a simple frequency-based model and two transformer-based models, as depicted in Fig. 3. The first uses term frequency-inverse document frequency (TF-IDF) [56] to encode the text descriptions without transfer learning, named as *Text-TFIDF*. Its output embeddings are vectors with a dimension equal to the number of the unique terms in the vocabulary of the entire dataset, indicating the importance of each term to each text description. The last two adopt the transformer-based USE [3] and BERT [15] text embedding modules and are called *Text-USE* and *Text-BERT*, respectively. They both embed the text descriptions and output embeddings with a fixed dimension of 512 according to the design of transformer encoders [27]. The USE and BERT modules are pre-trained on multiple large text datasets, such as Wikipedia, for multiple tasks. This enables knowledge transfer from the large datasets to our target dataset. In each model, a dense layer, a dropout layer, and a dense output layer are sequentially connected to the text embedding module. The output layer employs Rectified Linear Unit (ReLU) activation function for the regression task in this study. The final outputs are the predicted values of the design metrics. In the training process of Text-USE and Text-BERT, we freeze the pre-trained USE and BERT modules, and only the weights of the last few layers are trainable. This enables the model to learn effectively from the small dataset while retaining knowledge from prior training.

CNN-based Sketch Embedding Models Three CNN-based

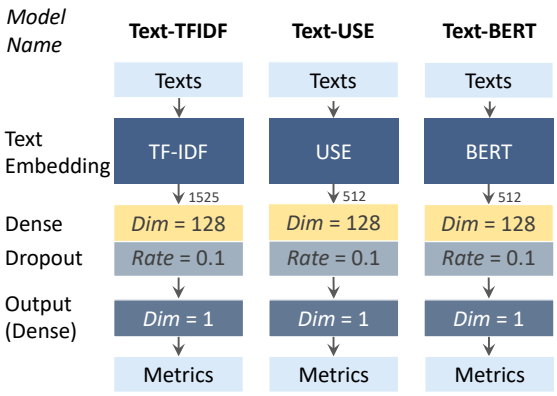

FIGURE 3. THE ARCHITECTURE OF THREE TEXT LEARNING MODELS WE USE IN THIS WORK.

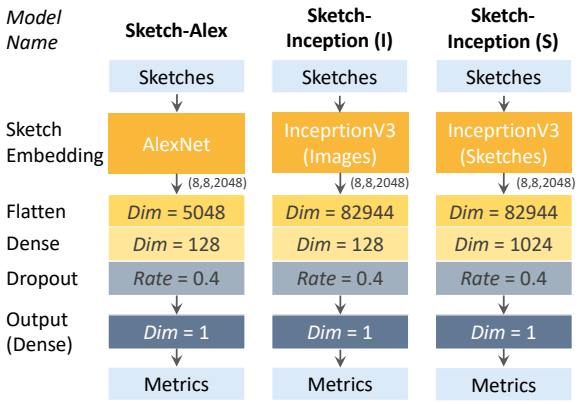

FIGURE 4. THE ARCHITECTURE OF THE CNN MODELS FOR SKETCH LEARNING.

architectures are explored for sketch embedding, as visualized in Fig. 4. The differences across the models manifest in three aspects: (1) The first model adopts the AlexNet CNN architecture [57] (called *Sketch-Alex*). The second and third models use the InceptionV3 CNN architecture [16] and are named *Sketch-Inception (I)* and *Sketch-Inception (S)*. (2) Sketch-Alex is trained on the target data directly without transfer learning, whereas Sketch-Inception (I) and Sketch-Inception (S) are first pre-trained on two large datasets and transfer the learned knowledge to the target data. (3) The InceptionV3 module in Sketch-Inception (I) is pre-trained on 14 million images from the ImageNet dataset, while the InceptionV3 module in Sketch-Inception (S) is pre-trained on 50 million sketches from the QuickDraw dataset. We use the former pre-trained model from the Python Keras API, while we locally train the latter. During fine-tuning, we further adapt the weights of the pre-trained modules to our target task. The knowledge learned from the large external datasets is transferred to our target dataset through such a process. Similarly, a flatten layer, a dense layer, a dropout layer, and a dense output layer using the ReLU activation function follow each CNN embedding module sequentially. We determine the hyperparameters of these layers through experiments.

Multimodal Transfer Learning Model We then integrate the best text and sketch learning models among all alternative models to construct the multimodal learning model. In this study, Text-BERT and Sketch-Inception (S)) perform best for text and sketch learning, respectively. We integrate the two models by concatenating their outputs and adding a dense layer with the ReLU activation function to produce the final output. The architectures of the multimodal learning models are shown in Fig. 5. Likewise, the multimodal learning model is initialized with the pre-trained weights from the unimodal models to transfer the learned knowledge from the unimodal data to the multimodal data. During training, the trainable weights of both unimodal models are fine-tuned jointly to leverage the complementarity between different representations. The joint model takes advantage of both transfer learning and multimodal learning, which is termed a multimodal transfer learning model in this paper.

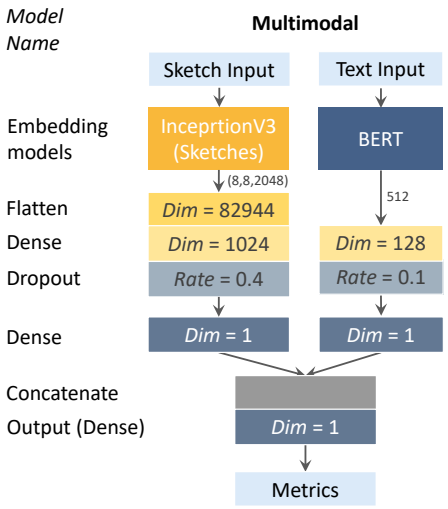

FIGURE 5. THE ARCHITECTURE OF THE MULTIMODAL TRANSFER LEARNING MODEL

4. RESULTS AND DISCUSSION

In this section, the performances of the unimodal text and sketch learning models and the multimodal learning model are reported and discussed to validate the efficacy of the proposed model. Specifically, we evaluate the performance of each model in terms of its explanatory power for the variabilities of the five design metrics, i.e., the coefficient of determination (R^2 value)

in statistics. The statistical significance of the differences in explanatory power between the models is also assessed [58]. To train and test the models, we split the 1,086 milk frother design ideas with the expert-assessed design metrics into training, validation, and test sets following the ratio of 0.8:0.1:0.1. The distribution stratification of the design metrics is maintained during the data split. Since each model is trained for predicting each design metric separately, the design idea split is generated for each design metric uniquely according to its specific stratification. All models for predicting the same metric are trained and tested on the same split. Through a series of pilot experiments, we choose a batch size of 24 and a learning rate starting from 5×10^{-5} with a decaying rate of 0.9 per 4 epochs to train all models. The maximal number of training epochs is set to 300, while the training process can be ended beforehand if the validation loss does not decrease for 50 consecutive epochs. We train each model 15 times for the statistical significance test.

4.1 Effect of Transfer Learning

In handling the small milk frother design dataset, transfer learning improves model performance prominently for both text and sketch learning. We first compare the performance of the text learning models. The maximal and average explanatory power, i.e., R^2 values, achieved by the models are plotted in Fig. 6. The R^2 values indicate that both transformer-based transfer learning models (Text-USE and Text-BERT) outperform the Text-TFIDF model that is directly trained without transfer learning for all metrics. Text-TFIDF hardly captures informative features from the text data, while Text-USE and Text-BERT using pre-trained text embeddings are more effective in capturing the semantic features for design metric prediction. Since the BERT text embedding module utilizes bidirectional transformers and a deeper architecture [15], it can capture more delicate features for predicting most of the design metrics, especially uniqueness. Comparing Text-TFIDF and Text-BERT, we note that transfer learning increases the R^2 values by 0.02-0.18 for text learning.

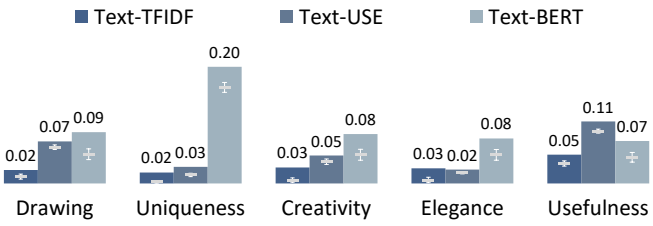

FIGURE 6. THE PERFORMANCES OF THE TEXT LEARNING MODELS. THE COLUMNS SHOW THE MAXIMAL R^2 VALUES, WHILE THE MARKERS ON THE COLUMNS INDICATE THE MEAN R^2 VALUES WITH 1 STANDARD ERROR BAR.

Next, we discuss the results for models trained on sketches. Fig. 7 shows the maximal and average R^2 values of the sketch learning models — Sketch-Alex, Sketch-Inception (I), and Sketch-Inception (S). The relatively low R^2 values of Sketch-Alex imply that the AlexNet without transfer learning is not effective in capturing meaningful features for design metric prediction. In comparison, InceptionV3-based models with transfer learning can capture more features informative for the given task. Moreover, Sketch-Inception (S) achieves higher R^2 values than Sketch-Inception (I) does for all design metrics, indicating more effective knowledge transfer by the former than by the latter. This is because that Sketch-Inception (S) is pre-trained on sketches from the QuickDraw dataset, which are more similar to the milk frother sketches than the images from the ImageNet dataset, on which Sketch-Inception (I) is pre-trained. According to the average R^2 values shown in Fig. 7, changing the source domain of transfer learning from images to sketches (Inception (S) - Inception (I)) improves the model performances more prominently than introducing knowledge transfer from images to our target sketches (Inception (I) - Alex). The differences between Sketch-Inception (S) and Sketch-Alex indicate that transfer learning increases the R^2 values by 0.09-0.24 for sketch learning.

Additionally, we note that the BERT-based text model is better at predicting uniqueness (the highest R^2 value across all metrics in Fig. 6) than all metrics. This finding may indicate that written descriptions, often short, may contain semantic features that describe what makes a design unique. Sketch-based models (especially, Sketch-Inception (S)) are better at predicting drawing quality (the highest R^2 values across all metrics in Fig. 7) than all other metrics. This observation makes sense, as one expects that a sketch itself will contain more information about the drawing quality of the sketch. Unlike drawing quality, the other metrics may have more dependence on the entire design concept underlying the multimodal representations. Future work will analyze the expert evaluation process to better understand what type of information experts use from each representation mode to evaluate different metrics. We will also validate these observations through computational models.

The results indicate the positive transfer of knowledge from the external datasets to our dataset for both text and sketch learning. While training the models, we observe three strengths of transfer learning. First, transfer learning enables the applications of more complex models to small datasets. In general, complex models can capture more delicate features but rely on larger datasets to train, which constrains their applications to small datasets. For example, it is challenging to train a generalizable InceptionV3 model with 23,885,392 trainable parameters using our dataset containing 1,086 sketches from scratch. Initialized with weights pre-trained on large datasets, the complex models can start from good conditions and are more likely to learn meaningful features from small datasets. Second, transfer

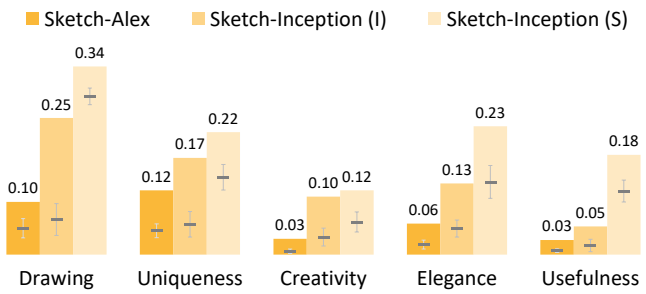

FIGURE 7. THE PERFORMANCES OF THE SKETCH LEARNING MODELS. THE COLUMNS SHOW THE MAXIMAL R^2 VALUES, WHILE THE MARKERS ON THE COLUMNS INDICATE THE MEAN R^2 VALUES WITH 1 STANDARD ERROR BAR.

learning improves model performance in handling small datasets. Pre-trained on large source datasets, the models can effectively capture generalizable pivot features, which are difficult to learn from small target datasets. The knowledge transfer from the source datasets to the target dataset improves model performance. Third, transfer learning saves computational costs. Although it often takes a long time to pre-train a complex model on a large dataset, the pre-trained models can be transferred to different relevant datasets and for multiple tasks. Compared to training from scratch, the fine-tuning process is much faster. Particularly, this study provides a large pre-trained InceptionV3 model for sketch learning tasks to the design and machine learning communities. The model is accessible from GitHub[4].

Moreover, the results also inform the application of transfer learning. The selections of the model architecture (USE versus BERT) and the source dataset (ImageNet versus QuickDraw) influence the effectiveness of knowledge transfer. Besides knowledge transfer across domains, we also explored two other types of knowledge transfer in this study. This first transfer is from unimodal data to multimodal data, which is beneficial in this study. The second is knowledge transfer across tasks, such as predicting different design metrics. We learn from our pilot experiments that the multi-task model that predicts all design metrics simultaneously is inferior to the models that predict each design metric separately, indicating a negative transfer between the tasks. A possible reason is that the evaluation of different metrics relies on different pivot features, as indicated by the low average correlation coefficient (0.092) between the metrics.

4.2 Effect of Multimodal Learning

The joint learning of the multimodal representations enables the model to capture more informative features and perform better for design metric prediction, compared to unimodal learning from only sketches or texts. The R^2 values of the multimodal transfer learning model are prominently higher than that of the best unimodal text and sketch learning models, as shown in Fig. 8. The results indicate that the sketches and texts complement each other when they are learned jointly, enabling the multimodal learning model to capture more informative features for design metric prediction. Moreover, the comparison between the best text and sketch models suggests that the sketch representation is more informative than the text representation for predicting most design metrics, especially drawing quality, elegance, and usefulness. Two intuitive reasons explain this. First, the evaluation of drawing quality and elegance naturally relies more on visual features than semantic features. Second, sketches are highly illustrative and can easily transcend language barriers to communicate complex information, such as interactions between design elements and interactions with the environment [18], which benefits design metric evaluation. It is worth noting that the informativeness of sketches and text descriptions may vary across different design domains or settings. Although sketches are more expressive than texts in this study, they only embed visual features. The semantic features embedded in texts can complement the visual features. By joining them together, multimodal learning can capture the interactions between the two types of features, which leads to an increase in R^2 values of 0.04-0.10.

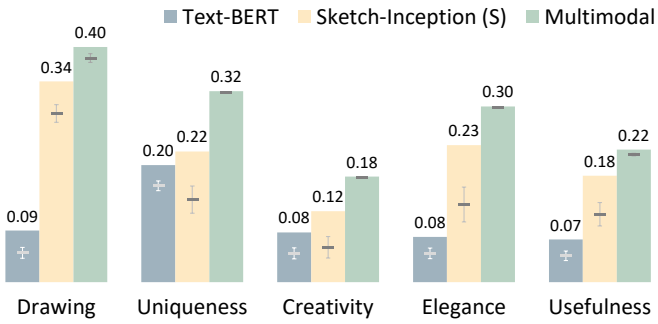

FIGURE 8. THE PERFORMANCE COMPARISON BETWEEN THE UNIMODAL AND MULTIMODAL LEARNING MODELS. THE COLUMNS SHOW THE MAXIMAL R^2 VALUES, WHILE THE MARKERS ON THE COLUMNS INDICATE THE MEAN R^2 VALUES WITH 1 STANDARD ERROR BAR.

Furthermore, the R^2 values shown in Fig. 8 also inform the predictability of the design metrics using design descriptions and sketches. Specifically, the drawing quality, uniqueness, and elegance present higher predictability (R^2 values $\geq$ 0.3) than the creativity and usefulness. According to the definitions of these metrics introduced in Section 3, the assessment of the high-predictability metrics is more straightforward, while assessing creativity and usefulness requires a deeper understanding

[4]https://github.com/likeshine/sketch-text-multimodal-transfer-learning

and more abstract logical inference. These results are consistent across different representation methods. These findings indicate that ANNs may be less effective in handling abstract tasks that need a broader understanding of the context, such as predicting creativity. In addition to design metric prediction, the proposed multimodal transfer learning model can be generalized to design evaluation in broader contexts, such as assignment and exam assessment for design courses.

Compared to the results reported in a prior study [59] aiming at the same task, the proposed multimodal transferring learning model exhibits significantly higher explanatory power for the variabilities of the design metrics. The prior study [59] aimed at utilizing the less resource-demanding Shah, Vargas-Hernandez, and Smith (SVS) [6] features to predict the more resource-demanding CAT metrics. It required significant human inputs to review the design ideas for labeling the SVS features. The authors represented each design idea using a vector that is the one-hot encoding of the corresponding SVS features and employed three regression models to predict the CAT metrics. We compare the R^2 values achieved by our multimodal transfer learning model and the highest R^2 values achieved by the three regression models that use SVS features, as Shown in Table 1. The results suggest that our model outperforms the SVS-feature regression models prominently and requires much less human effort.

TABLE 1. THE COMPARISON IN R^2 VALUES BETWEEN THE MULTIMODEL TRANSFER LEARNING MODEL AND THE BEST REGRESSION MODELS USING SVS FEATURES

	Multimodal	**Regression Models**
Drawing	0.40	0.12
Uniqueness	0.32	0.07
Creativity	0.18	0
Elegance	0.30	0.16
Usefulness	0.22	0.17

The approach of applying transfer learning and multimodal learning to design ideas represented in multiple modes can be generalized to fulfil other design tasks. Further studies in other contexts are needed to test the generalizability of the trained models weights from the milk frother domain to other design domains for design metric prediction. However, the model architectures developed in this study can be adapted to various tasks.

4.3 Challenges in Design Creativity Evaluation through Sketches and Texts

The magnitude of the R^2 values achieved by the proposed model indicates that there is still plenty of room for improvement. This subsection summarizes the challenges and potential opportunities in design metric prediction using AI.

Most Design Data Is Noisy and May Lack Information First, for training deep learning models, sketches and text descriptions for representing design ideas are noisy. Different designers have inconsistent abilities and preferences in expressing design ideas using sketches and text descriptions. Similar design ideas can be communicated in distinct styles and with distinct levels of information details. Moreover, designers often represent designs in multiple modes, but a certain mode may be missing from some designs. Comparatively, human raters can handle these representation inconsistencies and missing information relatively easily; however, they are subject to fatigue. Meanwhile, it is still challenging for the current AI to distinguish conceptual differences from the undesired representation inconsistencies and make up for the missing information. In the future, efforts should be made to build effective human-AI hybrid teams where humans and AI can learn from and augment each other for this challenging task.

Most Design Datasets Are Small in Size The second challenge is the lack of large and high-quality labeled datasets to train high-performing AI. As neural networks go deep with huge amounts of trainable weights, it is almost impossible to obtain high-performing AI models without large, closely related training datasets. This study has demonstrated the efficacy of transfer learning in handling small datasets and the superiority of QuickDraw as the source domain. However, the QuickDraw dataset, comprising of simple human-made doodles, is far from an ideal source domain for transfer learning in this task. Sketches in QuickDraw are simple and labeled according to their natural categories. They cannot effectively support model training to capture complex features (e.g., object functions, behaviors, structures, and aesthetics), which design metric evaluation and various other design tasks rely on. Since the complex features of engineering designs have not been studied on a large scale through machine learning approaches, little effort has been made to construct a large labeled dataset for training high-performing AI. Therefore, a large labeled design idea dataset will greatly benefit the development of AI models to support engineering design. Through transfer learning, the constructed dataset can be leveraged for a variety of design tasks.

Design Evaluations May Consider Multiple Modes Third, it is difficult to encode complex design ideas represented in multiple modes comprehensively into one expressive embedding. This study makes an attempt to embed design ideas using multimodal models learning from texts and sketches. Further challenges reside in: (1) including the textual annotations within the drawing area into the multimodal embedding, (2) adding repre-

sentations in new modalities, such as 3D models, into the joint embeddings, and (3) developing advanced information fusion approaches to align and relate features learned from multiple modalities. Future studies can focus on the state-of-the-art methods for image segmentation and annotation, attention-based information fusion, and multimodal learning for embedding complex design ideas more effectively.

5. CONCLUSION

In this study, we develop and validate a multimodal transfer learning model for predicting design creativity metrics. This model enables simultaneous learning from design representations in multiple modes and knowledge transfer from external datasets to our target dataset. As the building blocks of the multimodal model, we construct and compare three text learning models and three sketch learning models, respectively. By leveraging the power of AI, this model sheds light on efficient and scalable design evaluation. We demonstrate the model using a set of milk frother design ideas represented by text descriptions and sketches to predict five design metrics: (i) drawing quality, (ii) uniqueness, (iii) usefulness, (iv) creativity, and (v) elegance.

The results of this study lead to five key findings: (1) The BERT model outperforms the other models in capturing semantic features from design descriptions for design metric prediction. (2) The Inception model pre-trained on sketches outperforms the other models in capturing visual features from design sketches for design metric prediction. (3) Transferring knowledge from large external datasets to our dataset benefits design metric prediction. (4) The multimodal model that simultaneously learns from sketches and texts surpasses the unimodal models that learn from only sketches or texts for design metric prediction. (5) Drawing quality, uniqueness, and elegance present higher predictability than creativity and usefulness based on the model learning from design descriptions and sketches. The proposed multimodal transfer learning model architecture can be generalized to broader application contexts. We also contribute a pre-trained Inception model that applies to various sketch-based tasks to the design and machine learning societies. Additionally, the challenges and opportunities in developing AI models for design evaluation are discussed to inform future research.

REFERENCES

[1] Hammedi, W., Van Riel, A. C., and Sasovova, Z., 2011. "Antecedents and Consequences of Reflexivity in New Product Idea Screening*". *Journal of Product Innovation Management,* ***28***(5), 9, pp. 662–679.

[2] Miller, S. R., Hunter, S. T., Starkey, E., Ramachandran, S., Ahmed, F., and Fuge, M., 2021. "How Should We Measure Creativity in Engineering Design? A Comparison Between Social Science and Engineering Approaches". *Journal of Mechanical Design,* ***143***(3), 3.

[3] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil Google Research Mountain View, R., 2018. "Universal Sentence Encoder". *AAAI*, 3, pp. 16026–16028.

[4] Sarkar, P., and Chakrabarti, A., 2014. "Ideas generated in conceptual design and their effects on creativity". *Research in Engineering Design,* ***25***(3), pp. 185–201.

[5] Amabile, T. M., 1996. *Creativity in Context : Update to The Social Psychology of Creativity*. Routledge, New York, 1.

[6] Shah, J. J., Vargas-hernandez, N., and Smith, S. M., 2003. "Metrics for measuring ideation effectiveness". *Design Studies,* ***24***(2), 3, pp. 111–134.

[7] Oman, S. K., Tumer, I. Y., Wood, K., and Seepersad, C., 2013. "A comparison of creativity and innovation metrics and sample validation through in-class design projects". *Research in Engineering Design,* ***24***(1), 9, pp. 65–92.

[8] Ling, G., Mollaun, P., and Xi, X., 2014. "A study on the impact of fatigue on human raters when scoring speaking responses". *Language Testing,* ***31***(4), 5, pp. 479–499.

[9] Chaudhuri, N. B., Dhar, D., and Yammiyavar, P. G., 2020. "A computational model for subjective evaluation of novelty in descriptive aptitude". *International Journal of Technology and Design Education 2020*, 11, pp. 1–38.

[10] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019. "Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel". *Journal of Mechanical Design,* ***141***(2), 2.

[11] Ahmed, F., Fuge, M., Hunter, S., and Miller, S., 2018. "Unpacking Subjective Creativity Ratings: Using Embeddings to Explain and Measure Idea Novelty". *Proceedings of the ASME Design Engineering Technical Conference,* ***7***, 11.

[12] Ahmed, F., and Fuge, M., 2017. "Capturing winning ideas in online design communities". In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, Association for Computing Machinery, p. 1675–1687.

[13] Zhang, C., Yang, Z., He, X., and Deng, L., 2019. "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications". *IEEE Journal on Selected Topics in Signal Processing,* ***14***(3), 11, pp. 478–493.

[14] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q., 2021. "A Comprehensive Survey on Transfer Learning". *Proceedings of the IEEE,* ***109***(1), 1, pp. 43–76.

[15] Devlin, J., Chang, M.-W. W., Lee, K., and Toutanova, K., 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding". In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies - Proceedings of the Conference, Vol. 1, Association for Computational Linguistics, pp. 4171–4186.

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016. "Rethinking the Inception Architecture for Computer Vision". In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-Decem, pp. 2818–2826.

[17] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE,* ***86***(11), pp. 2278–2323.

[18] Xu, P., Hospedales, T. M., Yin, Q., Song, Y.-Z., Xiang, T., and Wang, L., 2022. "Deep Learning for Free-Hand Sketch: A Survey". *IEEE transactions on pattern analysis and machine intelligence*, 1.

[19] Seddati, O., Dupont, S., and Mahmoudi, S., 2015. "DeepSketch: Deep convolutional neural networks for sketch recognition and similarity search". *Proceedings - International Workshop on Content-Based Multimedia Indexing,* ***2015-July***, 7.

[20] Lu, W., and Report, E. T., 2017. Free-hand Sketch Recognition Classification. Tech. rep., Stanford University.

[21] Jahan, N., Nesa, A., and Layek, M. A., 2021. "Parkinson's Disease Detection Using CNN Architectures withTransfer Learning". In 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Institute of Electrical and Electronics Engineers (IEEE), pp. 1–5.

[22] Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., and Hospedales, T. M., 2016. "Sketch-a-Net: A Deep Neural Network that Beats Humans". *International Journal of Computer Vision 2016 122:3,* ***122***(3), 7, pp. 411–425.

[23] Zhang, X., Huang, Y., Zou, Q., Pei, Y., Zhang, R., and Wang, S., 2020. "A Hybrid convolutional neural network for sketch recognition". *Pattern Recognition Letters,* ***130***, 2, pp. 73–82.

[24] Zhang, L., 2021. "Hand-drawn sketch recognition with a double-channel convolutional neural network". *EURASIP Journal on Advances in Signal Processing 2021 2021:1,* ***2021***(1), 8, pp. 1–12.

[25] Ha, D., and Eck, D., 2017. "A Neural Representation of Sketch Drawings". *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 4.

[26] Yang, L., Wei, X., Tong, S. J., Zhou, K., Zheng, Y., Zhuang, J., and Fu, H., 2021. "SketchGNN: Semantic Sketch Segmentation with Graph Neural Networks". *ACM Trans. Graph,* ***37***(111), p. 2021.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017. "Attention is all you need". In Advances in Neural Information Processing Systems, Vol. 2017-Decem, Neural information processing systems foundation, pp. 5999–6009.

[28] Mikolov, T., Chen, K., Corrado, G. S., Dean, J., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., 2013. "Distributed Representations of Words and Phrases and their Compositionality". *Advances in Neural Information Processing Systems*, 10, pp. 1–9.

[29] Pennington, J., Socher, R., and Manning, C. D., 2014. "GloVe: Global Vectors for Word Representation". *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543.

[30] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., 2018. "Deep Contextualized Word Representations". *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference,* ***1***, pp. 2227–2237.

[31] Zan, Z., Li, L., Liu, J., and Zhou, D., 2020. "Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images". *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval,* ***20***, 6, pp. 117–125.

[32] Pan, S. J., and Yang, Q., 2010. "A survey on transfer learning". *IEEE Transactions on Knowledge and Data Engineering,* ***22***(10), pp. 1345–1359.

[33] Wang, Z., Dai, Z., Poczos, B., and Carbonell, J., 2019. Characterizing and Avoiding Negative Transfer.

[34] Whalen, E., and Mueller, C., 2022. "Toward Reusable Surrogate Models: Graph-Based Transfer Learning on Trusses". *Journal of Mechanical Design,* ***144***(2), 2.

[35] Cheng, M. Y., Gupta, A., Ong, Y. S., and Ni, Z. W., 2017. "Coevolutionary multitasking for concurrent global optimization: With case studies in complex engineering design". *Engineering Applications of Artificial Intelligence,* ***64***, 9, pp. 13–24.

[36] Raina, A., Cagan, J., and McComb, C., 2019. "Transferring design strategies from human to computer and across design problems". *Journal of Mechanical Design, Transactions of the ASME,* ***141***(11), 11.

[37] Pandita, P., Ghosh, S., Gupta, V. K., Meshkov, A., and Wang, L., 2022. "Application of Deep Transfer Learning and Uncertainty Quantification for Process Identification in Powder Bed Fusion". *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg,* ***8***(1), 3.

[38] Huang, X., Hu, Z., Xie, T., Wang, Z., Chen, L., and Zhou, Q., 2021. "Point-Cloud Neural Network Using Transfer Learning-Based Multi-Fidelity Method for Thermal Field Prediction in Additive Manufacturing". *Proceedings of the ASME Design Engineering Technical Conference,* ***3A-2021***, 11.

[39] Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L. P., 2016. "Deep multimodal fusion

for persuasiveness prediction". *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 10, pp. 284–288.

[40] Anastasopoulos, A., Kumar, S., and Liao, H., 2019. "Neural Language Modeling with Visual Features". *undefined*, 3.

[41] Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F., 2019. "CentralNet: A multilayer approach for multimodal fusion". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* ***11134 LNCS***, pp. 575–589.

[42] Perez-Rua, J. M., Vielzeuf, V., Pateux, S., Baccouche, M., and Jurie, F., 2019. "MFAS: Multimodal fusion architecture search". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* ***2019-June***, 6, pp. 6959–6968.

[43] Bahdanau, D., Cho, K., and Bengio, Y., 2014. "Neural Machine Translation by Jointly Learning to Align and Translate". *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9.

[44] Graves, A., Wayne, G., and Danihelka, I., 2014. "Neural Turing Machines". *arXiv preprint arXiv:1410.5401.*, 10.

[45] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., Salakhutdinov, R., Zemel, R., and Bengio, Y., 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

[46] Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L., 2016. "Visual7W: Grounded question answering in images". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* ***2016-Decem***, 12, pp. 4995–5004.

[47] Kottur, S., Vedantam, R., Moura, J. M. F., and Parikh, D., 2016. "VisualWord2Vec (Vis-W2V): Learning Visually Grounded Word Embeddings Using Abstract Scenes". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6, pp. 4985–4994.

[48] Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., and Luo, J., 2017. "Deep Multimodal Representation Learning from Temporal Data". In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5447–5455.

[49] Tenenbaum, J. B., and Freeman, W. T., 2000. "Separating style and content with bilinear models". *Neural Computation,* ***12***(6), pp. 1247–1283.

[50] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X., 2018. "AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1316–1324.

[51] Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., and Afzal, M. Z., 2017. "TAC-GAN - Text Conditioned Auxiliary Classifier Generative Adversarial Network". *arXiv preprint arXiv:1703.06412*, 3.

[52] Yuan, C., Marion, T., and Moghaddam, M., 2022. "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model". *Journal of Mechanical Design,* ***144***(2), 2, pp. 1–20.

[53] Toh, C. A., and Miller, S. R., 2016. "Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection". *Research in Engineering Design,* ***27***(1), pp. 73–89.

[54] Zheng, X., and Miller, S. R., 2019. "Is Ownership Bias Bad? The Influence of Idea Goodness and Creativity on Design Professionals Concept Selection Practices". *Journal of Mechanical Design, Transactions of the ASME,* ***141***(2), 2.

[55] Buades, A., Coll, B., and Morel, J.-M., 2011. "Non-Local Means Denoising". *Image Processing On Line,* ***1***, 9, pp. 208–212.

[56] Liu, C. Z., Sheng, Y. X., Wei, Z. Q., and Yang, Y. Q., 2018. "Research of Text Classification Based on Improved TF-IDF Algorithm". *2018 IEEE International Conference of Intelligent Robotic and Control Engineering, IRCE 2018*, 10, pp. 69–73.

[57] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. "ImageNet Classification with Deep Convolutional Neural Networks". In Advances in Neural Information Processing Systems 25 (NIPS 2012), Vol. 25, pp. 1–9.

[58] Dror, R., Shlomov, S., and Reichart, R., 2020. "Deep dominance - How to properly compare deep neural models". In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Association for Computational Linguistics, pp. 2773–2785.

[59] Edwards, K., Miller, S. R., and Ahmed, F., 2021. "If A Picture Is Worth 1000 Words , Is A Word Worth 1000 Features For". In Proceedings of the ASME Design Engineering Technical Conference, Virtual Conference.

Appendix A: Statistical Significance of Performance Differences

The statistical significance of the differences in explanatory power between a pair of models (e.g., model A vs. model B) is assessed using a metric named Almost Stochastic Order (ASO) [58], which was proposed specifically for comparing ANNs and reported superior to common significance metrics, such as p-value. The ASO values fall in the range [0, 1], and a value lower than 0.5 means that the former model (i.e., model A) is stochastically dominant over the latter model (i.e., model B). The lower the value is, the more confident we are that the former is better than the latter. In contrast, a value higher than 0.5 means the latter model outperforms the former one. Following the convention of p-value, we deem an ASO value lower than 0.05 or higher than 0.95 indicates a statistical significant

difference. The ASO values lower than 0.01 or higher than 0.99 are highlighted by **, while the other values lower than 0.05 or higher than 0.95 are highlighted using *. The pairwise performance comparisons between the text models for predicting each design metric are listed in Table 2. In most cases, the ASO values are lower than 0.5 and close or equal to 0, indicating that the former sketch models performed better. Text-USE performs better in predicting drawing quality and usefulness than Text-BERT.

TABLE 2. STATISTICAL SIGNIFICANCE OF THE DIFFERENCES IN EXPLANATORY POWER BETWEEN THE TEXT LEARNING MODELS: D-DRAWING, UN-UNIQUENESS, C-CREATIVITY, E-ELEGANCE, US-USEFULNESS

	D	UN	C	E	US
BERT vs. USE	1**	0**	0.17	0**	1**
BERT vs. TFIDF	0**	0**	0**	0**	0.37
USE vs. TFIDF	0**	0**	0**	0.01**	0**

Table 3 summarizes the pairwise performance comparisons between the sketch models in terms of each design metric. In all cases, we note that the values are lower than 0.5 and close or equal to 0, indicating that the former sketch models performed significantly better in most cases.

TABLE 3. THE STATISTICAL SIGNIFICANCE OF THE DIFFERENCES IN EXPLANATORY POWER BETWEEN THE SKETCH LEARNING MODELS: D-DRAWING, UN-UNIQUENESS, C-CREATIVITY, E-ELEGANCE, US-USEFULNESS

	D	UN	C	E	US
Inception (S) vs. Inception (I)	0**	0**	0**	0**	0**
Inception (S) vs. Alex	0**	0**	0**	0**	0**
Inception (I) vs. Alex	0.06	0.15	0**	0**	0.01*

Table 4 shows the comparisons between the unimodal and multimodal learning models for predicting each design metric. In all cases, except two, the ASO values are equal to 0, indicating that the former models performed significantly better. For predicting uniqueness and creativity, we find no statistical difference between the sketch-based Inception (S) model and the text-based BERT model.

TABLE 4. THE STATISTICAL SIGNIFICANCE OF THE DIFFERENCES IN EXPLANATORY POWER BETWEEN THE UNIMODAL AND MULTIMODAL LEARNING MODELS: D-DRAWING, UN-UNIQUENESS, C-CREATIVITY, E-ELEGANCE, US-USEFULNESS

	D	UN	C	E	US
Multimodal vs. Inception (S)	0**	0**	0**	0**	0**
Multimodal vs. BERT	0**	0**	0**	0**	0**
Inception (S) vs. BERT	0**	0.67	0.41	0**	0**