Deep-Learning Methods of Cross-Modal Tasks for Conceptual Design of Product Shapes: A Review

Xingang Li
Walker Department of Mechanical Engineering
University of Texas at Austin
Austin, Texas 78712
Email: xingang.li@utexas.edu

Ye Wang
Autodesk Research
Pier 9
San Francisco, California 94611
Email: ye.wang@autodesk.com

Zhenghui Sha*
Walker Department of Mechanical Engineering
University of Texas at Austin
Austin, Texas 78712
Email: zsha@austin.utexas.edu

Abstract: Conceptual design is the foundational stage of a design process that translates ill-defined design problems into low-fidelity design concepts and prototypes through design search, creation, and integration. In this stage, product shape design is one of the most paramount aspects. When applying deep learning-based methods to product shape design, two major challenges exist: 1) design data exhibit in multiple modalities, and 2) an increasing demand for creativity. With recent advances in deep learning of cross-modal tasks (DLCMT), which can transfer one design modality to another, we see opportunities to develop artificial intelligence (AI) to assist the design of product shapes in a new paradigm. In this paper, we conduct a systematic review of the retrieval, generation, and manipulation methods for DLCMT that involve three cross-modal types: text-to-3D shape, text-to-sketch, and sketch-to-3D shape. The review identifies 50 articles from a pool of 1341 papers in the fields of computer graphics, computer vision, and engineering design. We review 1) state-of-the-art DLCMT methods that can be applied to product shape design and 2) identify the key challenges, such as lack of consideration of engineering performance in the early design phase, that need to be addressed when applying DLCMT methods. In the end, we discuss the potential solutions to these challenges and propose a list of research questions that point to future directions of data-driven conceptual design.

Keywords: Conceptual Design, Product Shape Design, Deep Learning, Multi-modality, Cross-modality.

1 Introduction

The product shape is essential in the conceptual design of engineered products because it can affect both the aesthetics and the engineering performance of a product [1]. Figure 1 shows the flow of information and the key steps for the design of product shapes at the conceptual design stage [1], where the information can be categorized into three modalities: natural language (e.g., text), sketches (e.g., 2D silhouette), and 3D shapes (e.g., meshes). We call them design modalities. Generally, customer needs and engineering requirement documents are in the form of natural languages. Design sketches and drawings are effective ways of brainstorming and expressing designers' preferences. Low-fidelity design concepts and prototypes from the conceptual design stage are often represented by 3D shapes in digital format. Design Search, Design Creation, and Design Integration are the core steps of conceptual design to gather information from existing design solutions for inspiration and to develop novel design concepts to better explore the design space [1].

Early design automation methods, such as grammar- and rule-based methods, rely primarily on human design experience and knowledge to generate design alternatives [2]. In contrast, deep learning methods can learn latent design representations from data without explicit design rules or grammars, so they have been increasingly adopted in many engineering design applications. So far, however, deep learning methods have been applied mainly in the later stages of engineering design for design automation [3]. It is challenging to apply deep learning methods to the conceptual design stage (i.e., the early design stage) for several reasons. For example, data in the conceptual design stage exhibit multiple modalities, but deep learning methods are usually applied to handle a single design modality. Moreover, in conceptual design, designers often gather a large set of information for design inspiration in different design steps, but deep learning methods tend to focus on one specific design task at a time. Fi-

*Corresponding author.

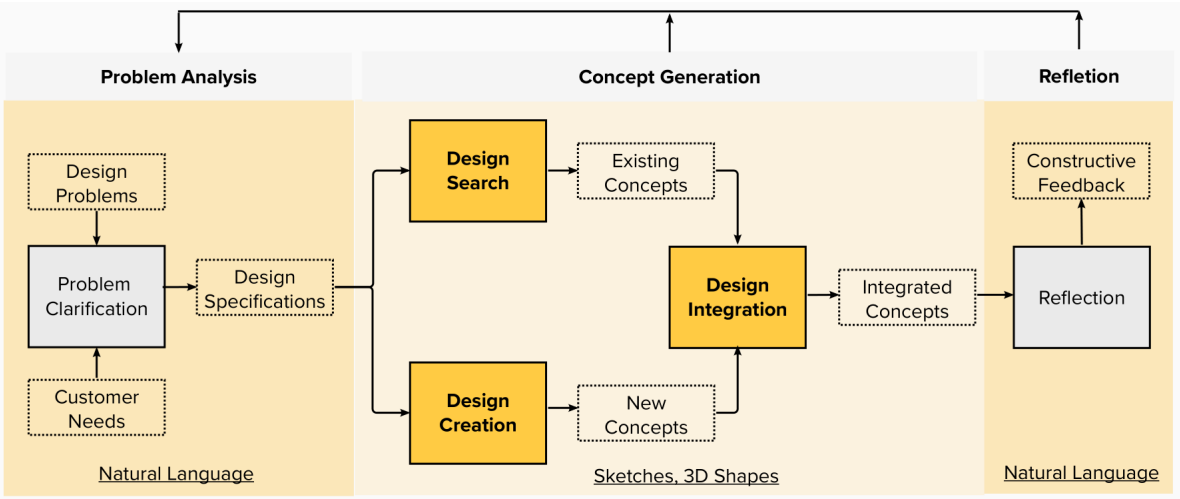

Fig. 1. Iterative conceptual design stage in the development of engineered products

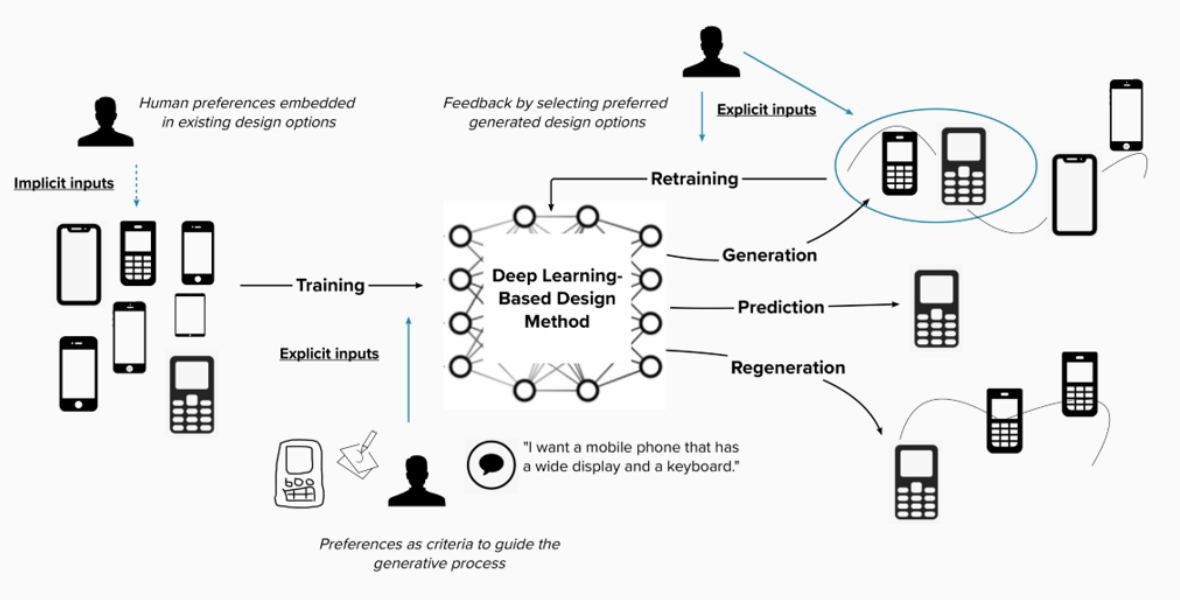

Fig. 2. Deep learning-based design process with humans in the loop

nally, human (either user or designer) input and interactions are desired in conceptual design to improve design creativity and human-centered design, but most current design methods developed using deep learning do not interact directly with human data, but only implicitly capture human preferences from training datasets, as shown in Figure 2.

With recent development in deep learning of cross-modal tasks (DLCMT) [1], we see the opportunities of applying these methods to address the aforementioned challenges, particularly in product shape design, such as car body and plane fuselage [5, 6]. DLCMT allows explicit human input in one design modality and translates it to another modality, e.g., from natural language or sketches to 3D shapes, as shown in Figure 2. In DLCMT, there are cross-modal retrieval, generation, and manipulation methods. Cross-modal retrieval methods can be used to search an existing design repository for inspiring design ideas. Cross-modal generation methods can be used to explore a design space to generate new design concepts. Lastly, cross-modal manipulation methods can further edit and manipulate existing designs to refine designs. These three categories of methods can be used in the Design Search, Design Creation, and Design Integration steps (Figure 1), respectively. In this paper, we conducted a systematic review of the state-of-the-art methods for DLCMT. Through a close examination of the existing literature, our objective is to identify the DLCMT methods and technologies that can be used to facilitate the conceptual de-

[1]DLCMT is a class of problems, aiming to translate one modality of data to another, e.g., from text to 3D shapes. To solve this problem, there is a large body of literature on cross-modal representation learning (CMRL). CMRL aims to build embeddings using information from multiple modalities (e.g., texts, audio, and images) in a common semantic space, which allows the model to compute cross-modal similarity [4]. In this paper, our review is not limited to reviewing CMRL methods but also includes other deep learning methods that can solve cross-modal problems.

sign and the challenges associated with applying them.

A total of 50 recently published journal articles and conference papers are identified and closely reviewed from the fields of computer graphics, computer vision, and engineering design. We focus on the text, sketches, and 3D shapes because they are the main design modalities in conceptual design. Specifically, we reviewed deep learning methods for three types of cross-modal tasks: text-to-sketch, text-to-3D, and sketch-to-3D. We found that most of the literature comes from computer graphics and computer vision, with few attempts at engineering design applications. This poses new challenges and opportunities for adapting the models and techniques developed to solve engineering design problems and, particularly, to bridge human input and interactions with deep learning methods in the conceptual design of engineered product shapes.

The remainder of this paper is organized as follows. Section 2 introduces background knowledge on conceptual design, design modalities, and our motivation for the review. Section 3 presents the methodology for our systematic review. We tabulate all the reviewed articles and present four statistics from the literature in Section 4. We then discuss the literature in detail and answer the research questions of the systematic review in Section 5. In the end, we propose a list of six research questions that will inform future research directions in Section 6 and conclude our work with closing remarks in Section 7.

2 Background

2.1 Conceptual Design

Conceptual design lies in the early phase of a design process in which the form and function of a product are explored [7]. In conceptual design, it is crucial to explore the design space as much as possible, and designers are demanded to generate creative designs so that the products are likely to succeed in the market [8, 9]. As shown in Figure 1, we adapt and reinterpret the five-step concept generation method in conceptual design [1]. The five steps are Problem Clarification, Design Search, Design Creation, Design Integration, and Reflection. Through these five steps, the method transfers information, such as customer needs, engineering requirements, and design ideas, to design concepts in the form of sketches and 3D shapes. The corresponding input and output of each step are represented by dotted rectangles. The process is linear in sequence from left to right, but almost always iterative. For example, feedback from Reflection could influence Problem Clarification and its subsequent steps. Each design step can also be iterative so that the design problem can be better understood, and the design space can be better explored [1].

In the conceptual design phase, the shape of a product is one of the most important considerations that are influential on the aesthetics of a product and its engineering performance [1, 10]. In this paper, we focus primarily on reviewing the DLCMT methods that can be applied for product shape design in the three concept generation steps, i.e., Design Search, Design Creation, and Design Integration, because they are the core steps for design concept exploration.

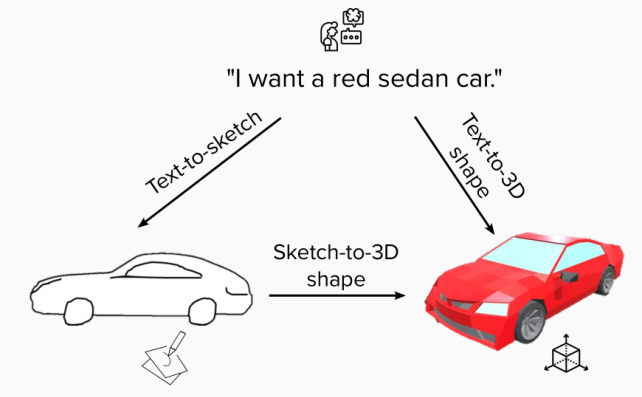

Fig. 3. Cross-modal tasks in conceptual design

2.1.1 Design Search

Design Search is the step of collecting information on existing design solutions to a design problem. In practice, several ways, such as patents, literature, and benchmarking, can be used to gather useful information [1]. By analyzing those existing products, designers can summarize their advantages and disadvantages, so that they can make necessary and customized changes to existing designs to create satisfying ones. However, the repository of existing design options could be huge, so the search process would be time-consuming and cumbersome, placing significant cognitive and physical burdens on designers. One possible solution to this problem is to use an AI-assisted search process, where designers can predefine search criteria and utilize computers to search for relevant design solutions.

2.1.2 Design Creation

Design Creation emphasizes exploring novel design concepts. Designers brainstorm ideas and explore the design space to create novel design concepts based on the knowledge of designers [1]. Design ideas are often presented as sketches and text descriptions during conceptual design [11]. Text descriptions are used to document and describe designers' ideas, while sketches can help visualize design concepts, further triggering creative design ideas [12–14]. Low-fidelity 3D models are then created for better visualization and further development. However, creating 3D models involves a lot of manual work and could be time-consuming. To facilitate the creation of novel 3D shapes, generative design methods can be used to automate the process.

2.1.3 Design Integration

The Design Integration is the step where designers aim to systematically integrate the information collected from previous steps to generate the integrated design concept(s) [1]. For product shape design, designers usually need to edit and manipulate designs collected from the Design Search and Design Creation steps. But, it can be challenging to

Table 1. The text types of natural language data used in DLCMT and the examples

Text Type	Examples
Natural language descriptions (NLD)	"It's a round glass table with two sets of wooden legs that clasp over the round glass edge".
Object names	"chairs", "cars", "planes"
Semantic keywords	"circular short", "rectangular wooden"

modify these designs computationally because their representations have certain formats (e.g., a 3D shape in voxels or point clouds or a sketch of a raster image). Some formats are not editable and must be translated into other formats, such as mesh or B-rep. Therefore, automating the modification with human inputs can significantly simplify the process.

2.2 Modalities in Conceptual Design

As shown in Figure 1, there are three main design modalities: natural language (NL), sketches, and 3D shapes, in conceptual design. In an example of car body design, as shown in Figure 3, the three modalities could be "I want a red sedan car" (NL), hand-sketching a car with desired features (sketch), and then creating a computer-aided design (CAD) model of the car (3D shape). NL allows people to convey and communicate ideas and thoughts. It is also the primary means for documentation, such as documentation of customer needs and engineering requirements. Sketches are often used to brainstorm design concepts because sketching can stimulate designers' creative imagination [12–14]. Then, a 3D shape is often built to provide better visualization and a low-fidelity prototype model for further evaluation and development of a concept.

NL data are often in the format of the text, which is usually the keyword in DLCMT methods. As shown in Table 1, there are mainly three types of text used as input in DLCMT, which are natural language descriptions (NLD), object names, and semantic keywords. 2D sketches can be represented in multiple ways, such as a pixel image [2] in static pixel space and vector image in dynamic stroke coordinate space [15, 16]. There are also generally two types of 3D sketches in the literature, and we refer to them as Type I and Type II, respectively. Type I: This kind of 3D sketch is represented in a 2D space. But compared to regular 2D sketches, they look like 3D objects. Type II: The 3D sketches that can be represented in a 3D space (either real or computational). Such a type of 3D sketch data can be captured and generated using virtual reality (VR) tools or motion sensing devices. They can also be created using 3D sketching software (e.g., SolidWorks or Autodesk). 3D shapes are typically built as B-rep models using CAD software in engineering design. However, in computer graphics and the 3D deep learning fields, 3D shapes are usually represented as meshes, point clouds, and voxel grids. Compared to CAD models, these 3D representations typically have lower fidelity with fewer geometric details and structural information because (1) coarse resolution might be used to represent the shapes due to the limitations of computational resources [17, 18], (2) certain representations are not good at representing geometric details and topological structure by nature (e.g., point clouds; see Table 2 for more information), (3) the conversion of one representation to another might lose geometric or topological information [19, 20].

2.3 Review Motivation

Our motivation for this literature review is driven by the following two major challenges posed in conceptual design. The recent advancement in DLCMT gives us opportunities to address these challenges and bring new design experiences in conceptual design.

Challenge 1: Multi-modalities. There are multiple design steps (e.g., Design Search, Design Creation, and Design Integration) in the conceptual design stage, which involve information and data with different design modalities. Designers conduct design activities with different modalities during the conceptual phase to best explore the design space and generate novel ideas [21, 22].

Deep learning methods that can be used for Design Creation have been the focus [3], but most of them are focused on handling a single design modality as pointed out by [23, 24]. Typically, these methods use unimodal data of designs either in 2D [25–27] or 3D [28–31]. In addition, there is a lack of either unimodal or cross-modal methods that are useful for Design Search and Design Integration [24].

But not until recently, we see studies in the engineering community utilizing DLCMT to assist concept creation or design evaluation [23,32,33]. DLCMT methods take into account multiple design modalities, such as texts and sketches. There are retrieval, generation, and manipulation methods for DLCMT and they can be applied to different steps in conceptual design: (1) DLCMT retrieval methods can be used for Design Search since they can search existing data and return designs that best match the query of users (e.g., returning several chairs given a query by sketch) [34]; (2) Generation methods (e.g., sketch-to-3D shape generation methods [33, 35]) can be used to automate the Design Creation process; (3) Manipulation methods can allow designers to modify the designs from another design modality. For example, using a text-to-3D manipulation method [36], designers can modify a 3D design by providing a simple text description without direct manipulation of the design, and this can significantly reduce the time for the design modification.

Challenge 2: Creativity. Design creativity is critical in conceptual design which can largely affect the success of a

[2]Images can include both sketches and natural photos. In the literature, we notice that DLCMT methods of natural photos usually use "image" while the methods of sketches use "sketch" as the keyword, respectively. Also, in engineering design, sketches are usually considered as lines and strokes. To identify DLCMT methods for engineering design, we exclude corresponding methods of "image".

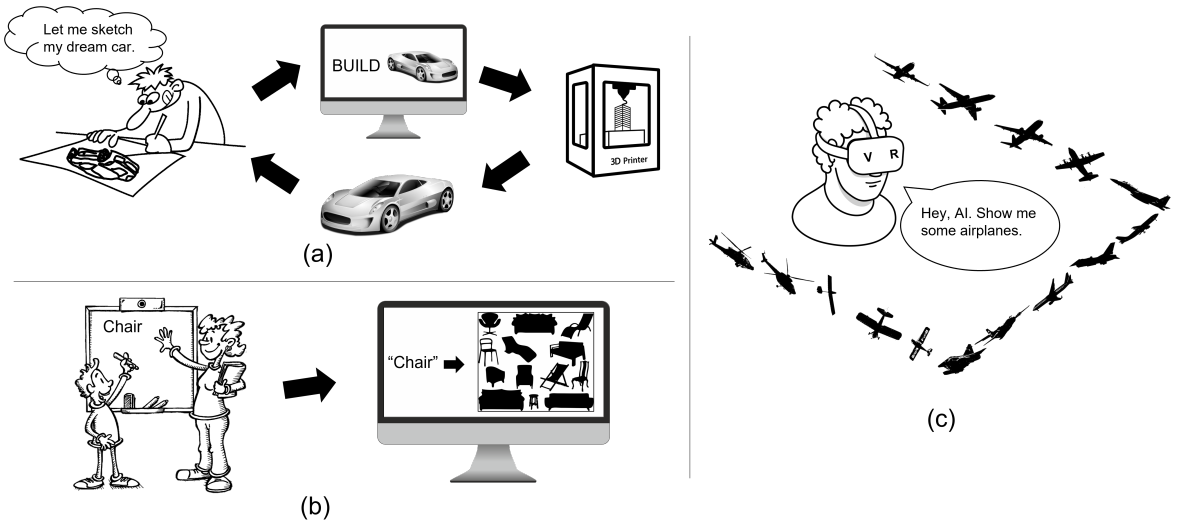

Fig. 4. Potential design applications enabled by DLCMT: (a) Democratization of product design; (b) AI-based pedagogical tools for educating and training students or novice designers; (c) Immersive design environment

product in the market. There are three main aspects (i.e., design novelty, contextual information, and human-computer interaction) that should be addressed for design creativity in the context of deep learning-based design processes.

(1) *Design novelty.* Deep learning methods (e.g., VAEs and GANs) can generate new data that are not seen in the training dataset but are still based on interpolation within the boundary of the training data. Therefore, the new designs generated from the deep learning-based design process share great similarities with the existing ones used as training data. To improve design creativity, there have been a few deep learning-based methods that focus on developing neural network architectures to generate creative designs by enabling deep-learning models' extrapolating capabilities [37,38]. These methods pose new opportunities for design because they can generate truly novel designs.
(2) *Contextual information.* On the other hand, humans have played an essential role in design creativity. However, despite advances in the development of network architecture, one observation is that human input and interaction are not much emphasized in the deep learning-based design process [3]. Burnap et al. [39] pointed out that a human's perception of the quality of the design concepts generated is often not in agreement with their numerical performance measures. The reason could be that in most deep learning-aided design processes, designers can only passively select the preferred design concepts from a set of computer-generated design options, but human designers may have contextual information [40] on a design problem which is hard to be captured by the training data.
(3) *Human-computer interaction.* As a result, there is a need to actively involve designers in a deep learning-based design process [3, 10]. Some efforts in this regard have recently been made in engineered product design. For example, the method introduced by [41] allows users to manipulate the latent space vectors learned by a GAN model to create preferred design options. Despite recent advances, we believe that design creativity can be further improved by involving humans in the design process to allow more intuitive and natural human input (e.g., text and sketch). Natural language and sketches are the most common human input in conceptual design, and DLCMT methods can intake these human inputs and transfer their modalities from one to another to promote creativity. That is manifested in the envisioned deep generative design process with humans in the loop, as shown in Figure 2. In such a process, designers can continuously supplement new design ideas during human-computer interaction to guide computers to generate creative and feasible design concepts.

In addition, there should be many design processes and applications that can be facilitated by DLCMT and we show three typical examples in Figure 4. *Design application 1:* DLCMT methods can be used to facilitate design democratization, allowing ordinary people to customize designs based on individual preferences [42]. *Design application 2:* There are also opportunities to develop AI-based pedagogical tools to teach students or train novice designers, allowing them to explore design alternatives with naive input, for example, just a simple word [43]. *Design application 3:* Immersive design uses virtual reality (VR), augmented reality (AR), and mixed reality (MR) to create a realistic digital environment in which a user is virtually immersed and can even physically interact with the digital environment [44]. The DLCMT methods can be integrated into immersive design applications to enhance the design experience in human-computer interaction.

In summary, DLCMT methods are likely to introduce new opportunities to support and enhance activities in the conceptual design stage for product shape design and be-

yond. We conduct a close examination of the existing literature aiming to identify the existing DLCMT methods and technologies that can be used for conceptual product shape design and the challenges associated with applying them. We will also discuss potential solutions to these challenges and point out future research directions.

3 Methodology

This study adopts a systematic literature review approach [45] with the procedure of formulating research questions for a review, identifying relevant studies, evaluating the quality of the studies, summarizing the studies, and interpreting the findings.

3.1 Research Questions

We are motivated to ask two research questions (RQs) according to the discussion above.

RQ 1. What DLCMT methods can be used in the following three steps of conceptual design?

(1) Design Search
(2) Design Creation
(3) Design Integration

RQ 2. What are the challenges in applying DLCMT to conceptual design and how can they be addressed?

3.2 Literature Search

3.2.1 Content Scope and Keywords

We defined the *content scope* using the following three criteria to search the literature relevant to deep learning of cross-modal tasks (DLCMT): (1) Conceptual design: Design Search, Design Creation, and Design Integration steps (highlighted in Figure 1). (2) Shape design: discrete, physical, and engineered products. (3) Design modality: text, sketch, and 3D shape.

The keywords identified and used in the literature search process are *"text-to-sketch retrieval", "text-to-sketch generation", "text-to-shape retrieval", "text-to-shape generation", "sketch-based 3D shape retrieval",* and *"sketch-based 3D shape generation"*. For "sketch-based 3D shape generation", we include the other three commonly used names: *"sketch-based 3D shape reconstruction", "sketch-based 3D shape synthesis",* and *"3D shape reconstruction from sketches"*.

The reasons for choosing these keywords come from the following aspects. (1) DLCMT between two different modalities of text, sketch, and 3D shape, should have six permutations of cross-modal. In this paper, we focus on the following three cross-modal tasks: text-to-sketch, sketch-to-3D shape, and text-to-3D shape, which are then concatenated with retrieval or generation to form the initial keywords (e.g., text-to-sketch generation). We did not include sketch-to-text, 3D shape-to-sketch, and 3D shape-to-text because sketches or 3D shapes are often the most common artifacts, and the design information flows in an order of text, sketches, and 3D shapes during the conceptual design. (2) we focus on Design Search which corresponds to retrieval methods, Design Creation which corresponds to generation methods, and Design Integration which corresponds to manipulation methods [3] . In addition, for the sketch-to-3D shape retrieval or generation methods, we made some modifications to the keywords according to the naming convention in the literature (see a comprehensive review on deep learning methods for free-hand sketch [15]). For example, we used "sketch-based 3D shape retrieval" instead of "sketch-to-3D shape retrieval" and the other three common terms introduced previously.

3.2.2 Literature Search Process

As shown in Figure 5, we finally selected 50 articles that meet our scope of review. Searches were conducted on the main databases of the literature (i.e., the *source scope*): ScienceDirect, Web of Science, Scopus, IEEExplore, ACM Digital Libraries, and Google Scholar within the time range of January 2013 to June 2022 (i.e., the *time scope*: the studies published in the past 10 years). The reason for choosing that time range is that many significant improvements in deep learning methods occurred after 2013, for example, variational autoencoders (VAEs, 2013) [46] and generative adversarial networks (GANs, 2014) [47]. Since then, they have been widely applied in various applications, including the cross-modal tasks reviewed in this paper.

The initial search yielded 1,341 seed articles, including duplicates, of which the majority (i.e., 1,304 papers) is related to two categories: sketch-based 3D shape retrieval and generation, with only 37 articles for the other four categories (i.e., text-to-sketch retrieval: 0; text-to-sketch generation: 3; text-to-3D shape retrieval: 10; and text-to-3D shape generation: 24) (see details in Table 3 in Appendix A). To make the review manageable, for the two categories of sketch-to-3D works, we decided to identify the most influential studies from those 1,304 papers using Connected Papers [4]. We found that [48] and [35] are pioneering work for deep learning-based sketch-to-3D shape retrieval and generation, respectively [24]. Therefore, they were used as the origin papers to find their most relevant work via Connected Papers (see Figure 11 in Appendix A for the two generated graphs). The search by Connected Papers identified 21 articles including [48] and [35] that meet our *content scope*.

Another finding was that the publication year of the articles in the two literature graphs turned out to be up to 2020, which could indicate that relevant articles published after

[3] We did not explicitly search for cross-modal manipulation methods because these methods cannot be found directly using specific keywords, but can be indirectly identified during the search for cross-modal retrieval and generation methods. For example, we found the work Text2Mesh [36], using the keyword "text-to-shape generation" because that keyword appears in the literature review section of the article, but the work should belong to manipulation methods after carefully reading its content. However, this might leave room for a more comprehensive review of the cross-modal manipulation methods by developing a different search strategy in the future.

[4] Access link: `https://www.connectedpapers.com/`. Connected Papers allow readers to enter an origin paper and can generate a graph of papers with the strongest connections to the origin paper by analyzing about 50,000 research papers.

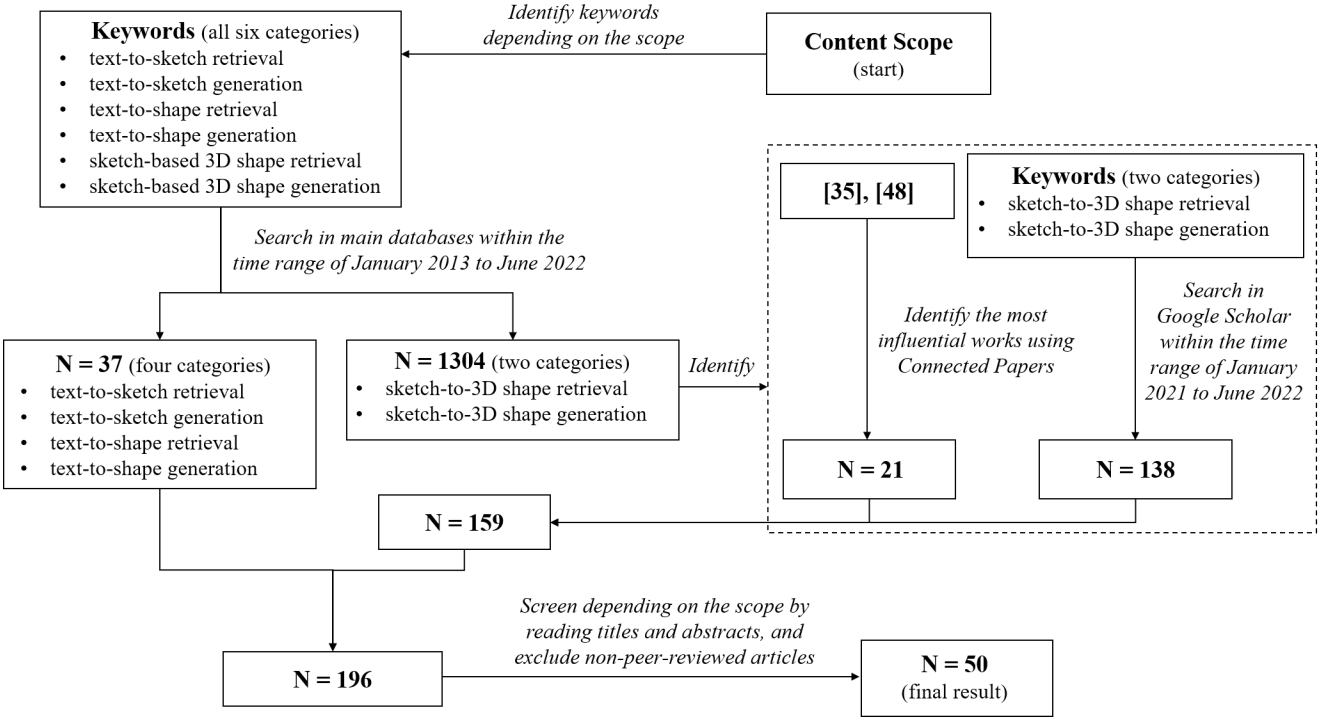

Fig. 5. Literature search process

2020 have not gained enough attention to be considered influential by Connected Papers. The finding motivated us to further find the most recent studies for these two categories, so we decided to search relevant articles within the time range from January 2021 to June 2022 in Google Scholar only, because we found that Google Scholar is more inclusive compared to other databases (i.e., the results from other databases turn out to be a subset of the results obtained from Google Scholar. See the comparison in Table 3 in Appendix A). 138 articles were found in this search process. In total, 196 papers were found to merit close examination and review.

We then reviewed the titles and abstracts of all these articles to judge their relevance to our *content scope*. We excluded 12 preprints, one Master thesis, and one Ph.D. dissertation from those 196 papers because the preprints are not peer-reviewed or officially published. Finally, 50 articles were considered the most relevant and therefore closely reviewed.

4 Summary Statistics of The Literature

We summarized all 50 articles in terms of the following variables: method type, publication year, representation of design modalities, training dataset(s), object class of the training data, generalizability, user interface, user study, and publication source in Table 4 of Appendix B which provides a complete list of these articles and the corresponding values for each of these variables. We report the statistics of four variables here, including the type of DLCMT, user interface, user study, and publication source, as an example, and introduce the others in detail in Section 5.

We did not find any work related to text-to-sketch retrieval, possibly due to the lack of interest in practical applications. We obtained 2 articles for text-to-3D shape retrieval, 6 articles for text-to-3D shape generation, 4 articles for text-to-sketch generation, 19 articles for sketch-to-3D shape retrieval, 18 articles for sketch-to-3D generation, and 5 articles for cross-modal design manipulation. Among these works, [17] can work for text-to-3D shape retrieval and generation; [50] can perform text-to-3D shape generation and manipulation; [6, 51] are shown to be capable of sketch-to-3D shape generation and manipulation.

Only 15 peer-reviewed publications are relevant to text-to-3D shape retrieval, text-to-3D shape generation, text-to-sketch generation, and cross-modal design manipulation, but

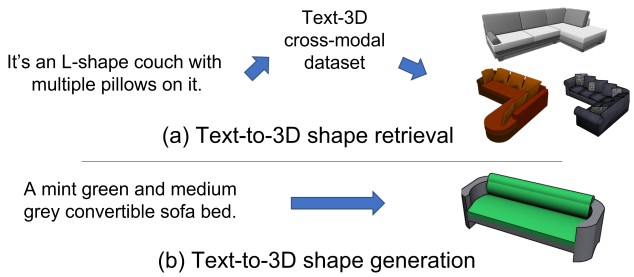

Fig. 6. Demonstration of *(a) Text-to-3D shape retrieval*: retrieving 3D shapes that best match the natural language descriptions (NLD) from a given dataset or repository; and (b) *text-to-3D shape generation*: automatically generating a 3D shape that matches the NLD. The examples of NLD and images are obtained from ShapeNet [49].

we observe a recent surging interest in these topics especially text-related ones, possibly due to advances in natural language processing (e.g., Contrastive Language-Image Pre-Training (CLIP) [52]) since our preliminary literature review [24].

There are 13 studies [6, 32, 44, 53–62] that provide user interfaces. The user interface application serves as a way to show the effectiveness of the proposed deep learning approach, which can also better facilitate human-AI interaction for creative designs. Especially, [44, 62] provide user interfaces in virtual reality (VR) and augmented reality (AR) settings, respectively, which can further improve the user experience of human-computer interaction in immersive design. Additionally, 12 studies [35, 36, 44, 53, 54, 56–61, 63] conducted user studies to further validate their methods and user applications. User studies can serve as a way to hear from human users so that researchers can improve the proposed methods from users' feedback. It can also help study human-computer interaction in a real situation.

The articles reviewed are from conference proceedings (32) and journals (18). Most DLCMT methods come from the domains of computer science and computer engineering with only two papers [32, 33] from the engineering design community.

5 Review and Discussion

In this section, we summarize our review of the papers in each of the cross-modal task categories and discuss their technical details, from which we draw insights into the challenges and opportunities of applying such methods in the engineering design field and discuss potential solutions to the challenges.

5.1 RQ 1-(1): What DLCMT Methods Can Be Used in Design Search of Conceptual Design?

5.1.1 Text-to-3D Shape Retrieval

The history of text-to-3D shape retrieval methods can be traced back to Min et al. 2004 [64], who used pure text information (query text and description associated with 3D shapes) for the 3D retrieval task, which is essentially a text-text matching.

For state-of-the-art deep learning methods as we introduce below, it is a common strategy to learn a cross-modal representation for text and 3D shapes using cross-modal representation learning techniques (see [65]) for more information. Figure 6 (a) demonstrates the process of a text-to-3D retrieval task. As a pioneering and representative work for this task, Chen et al. [17] first constructed a joint embedding of text and 3D shapes using an encoder composed of a convolution neural network (CNN) and a recurrent neural network (RNN) on text data and a 3D-CNN encoder on 3D voxel shapes. A triplet loss was applied and learning-by-association [66] was used to align the embedded representations of text and 3D shapes. They also introduced a 3D-text cross-modal dataset including two sub-datasets: 1) ShapeNet [49] (chairs and tables only) with a natural language description and 2) geometric primitives with synthetic text descriptions. However, the computational cost caused by the cubic complexity of 3D voxels limits this method to the machine learning of low-resolution voxels. Consequently, the learned joint representations will have low discriminative ability. Han et al. [67] built a $Y^2Seq2Seq$ network architecture using a Gated Recurrent Unit (GRU, one variation of RNN) to encode features of multiple-view images to represent the shape. To obtain the joint embedding of text and sketches, they trained the network using both intermodality and intramodality reconstruction losses, in addition to the triplet loss and classification loss. Therefore, the proposed network could learn more discriminative representations than [17].

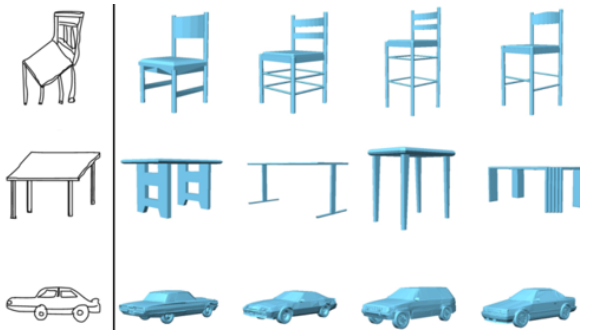

Fig. 7. *Sketch-to-3D shape retrieval* method by Wang et al. [48]. For each row, the 2D drawing is the query sketch and the 3D models are the retrieved 3D shapes from an existing dataset, Princeton Shape Benchmark (PSB) [68]. The figure is used with permission.

5.1.2 Sketch-to-3D shape Retrieval

Sketch-to-3D shape retrieval has been extensively studied using non-deep learning methods [69]. These methods usually consist of three steps: 1) automatically select multiple views from a given 3D shape in the hope that one of them is similar to the input sketch(es); 2) project the 3D shape into 2D space from the selected viewpoints; 3) match the sketch against the 2D projections based on predefined features. However, the selection of best viewpoints, as well as the design of predefined matching features, could be subjective and random, which motivates the development of deep learning-based methods that can avoid the subjective selection of views and learn features from the data of sketches and 3D shapes [48]. In light of the scope of this review, we focus on deep learning methods for sketch-to-3D shape retrieval.

Wang et al. [48] initialized the effort and proposed to learn feature representations for sketch-to-3D shape retrieval as shown in Figure 7, which avoided computing multiple views of a 3D model. They applied two Siamese CNNs [70] for views of 3D shapes and sketches, respectively, and a loss function defined on the within-domain and cross-domain similarities. To reduce the discrepancies between the sketch features and the 3D shape features, Zhu et al. [71] built a pyramid cross-domain neural network of sketches and 3D shapes. They used the network to establish a many-to-one relationship between the sketch features and a 3D shape fea-

ture. Dai et al. [72, 73] proposed a novel deep correlated holistic metric learning method with two distinct neural networks for sketch and 3D shape. Such a deep learning method mapped features from both domains into one feature space. In the construction of its loss function, both discriminative loss and correlation loss was used to increase the discrimination of features within each domain and the correlation between domains. Chen et al. [74] developed a GAN-based deep adaptation model to transform sketch features into 3D shape features, of which correlations can be enhanced by minimizing the mean discrepancy between modes. Xia et al. [75] proposed a novel semantic similarity metric learning method based on a "teacher-student" strategy by using a teacher network to guide the training of the student network. The teacher network was trained to extract the semantic features of the 3D shapes. The student network was then trained by using the pre-learned 3D shape features to learn the sketch features. Similarly, Yang et al. [76] applied a sequential learning strategy to learn 3D shape features without 2D sketches first and then used the learned features of 3D shapes to guide the learning of sketch features. During the query process, they further integrated clustering algorithms to categorize subclasses in a shape class to improve retrieval accuracy. In the methods mentioned above, deep metric learning [77] was applied to mitigate the modality discrepancy between the sketch and the 3D shape.

There are also methods that study how to represent 3D shapes more comprehensively so that 3D shapes can better correspond to sketches. Xie et al. [78] proposed a method to learn a Wasserstein barycenter of CNN features extracted from 2D projections of a 3D shape. They constructed the metric network to map sketches and the Wasserstein barycenters of 3D shapes to a common deep feature space. Then a discriminative loss was formulated to learn the deep features. The deep features learned could then be used for the sketch-to-3D shape retrieval. Chen et al. [79] proposed a novel stochastic sampling method to randomly sample rendering views of the sphere around a 3D shape and incorporated an attention network (see [80] for a comprehensive review) to exploit the importance of different views. They also developed a novel binary coding strategy to address the time-efficiency issue of sketch-to-3D shape retrieval.

Another direction to reduce the large cross-modality difference between 2D sketches and 3D shapes is to deal with noise in the sketch data. Liang et al. [81] pioneered this direction by developing a method called noise-resistant sketch feature learning with uncertainty, which achieved the new state-of-the-art for sketch-based 3D shape retrieval. Liu et al. [82] proposed a Guidance Cleaning Network to remove low-quality sketches that have much noise, which is like a data cleaning process. The authors showed superior results over state-of-the-art methods because the learning of noisy data was suppressed.

All the methods introduced above achieve state-of-the-art results on commonly used sketch-to-3D retrieval datasets, such as Princeton Shape Benchmark (PSB) [68], SHREC13 [83], and SHREC14 [84]. The multiview CNN (MVCNN) [85] has been widely used in all these methods to generate features from projection images of 3D shapes. Different from these methods aiming to retrieve objects by coarse category-level retrieval of 3D shapes given an input sketch, Qi et al. [34] introduced a novel task of fine-grained instance-level sketch-to-3D shape retrieval, with the aim of retrieving one specific 3D shape that best matches the input sketch. They created a set of paired sketch-to-3D shape data of chairs and lamps from ShapeNet [49]. Then, they built a deep joint embedding learning-based model with a novel cross-modal view attention module to learn the features of sketches and 3D shapes. As the first effort to find local image correspondences between design sketches, Navarro et al. [86] proposed a synthetic line drawing dataset rendered from 3D shapes from ShapeNet [49]. The authors obtained a learned descriptor, namely SketchZoom descriptor, for dense registration in line drawings and showed its promising application in sketch-3D shape retrieval by identifying local correspondences between sketches.

There is also interest in using CAD data in 3D shape retrieval. Qin et al. [32] developed a sketch-to-3D CAD shape retrieval approach using the variational autoencoder (VAE) and structural semantics. They created their training dataset by collecting 3D CAD models from local companies and obtained their six-view projections as sketch data. Manda et al. [87] developed a new sketch-3D CAD model dataset, CADSketchNet, from the Engineering Shape Benchmark (ESB) [88] and Mechanical Components Benchmark (MCB) [89] datasets. The authors also analyzed various deep learning-based sketch-to-3D retrieval approaches using the proposed dataset and reported the comparison results.

Efforts have also been made to bridge the semantic gap between sketches and 3D shapes to improve sketch-based 3D shape retrieval. Ye et al. [90] presented a CNN-based 3D sketch-based shape retrieval (CNN-SBR) architecture based on 3D sketch (Type II) data obtained from SketchANet [91]. Using data augmentation to prevent overfitting, they achieved a significant improvement compared to other learning-based methods. Building on previous work [90, 92], Li et al. [55] proposed a novel interactive application supported by CNN-SBR. The method used Microsoft Kinect, which can track the 3D locations of 20 joints of a human body, to track the 3D locations of a user's hand to create a 3D sketch. The proposed method was tested on a proposed dataset and achieved state-of-the-art performance in 3D sketch-based 3D shape retrieval.

The idea of utilizing a 3D sketch (Type II) as query input has been further applied to virtual reality (VR) and augmented reality (AR) settings to facilitate the immersive design. Building on the method proposed in [93], Giunchi et al. [44] designed a multimodal interface for 3D model retrieval in VR with both sketch and voice input. The authors implemented a consistent translation method between queries of 3D sketch and voice, allowing their integration during a single search session. Similarly, ShapeFindAR [62] combined both 3D sketch and textual input to enable in-situ spatial search of a 3D model repository in an AR setting. The server was built using a REST (representation state transfer) application programming interface provided by Flask, a web

framework for the Python programming language.

5.2 RQ 1-(2): What DLCMT Methods Can Be Used in Design Creation of Conceptual Design?

5.2.1 Text-to-3D Shape Generation

The task of text-to-3D shape generation is illustrated by Figure 6 (b). To accomplish this task, Jahan et al. [94] proposed a semantic label-guided shape generation approach, which can take one-hot semantic keywords as input and generate 3D voxel shapes without color and texture. The proposed method was trained using chairs, tables, and lamps obtained from the Co-Segmentation (COSEG) dataset [95] and ModelNet [96]. Based on their work on text-to-3D shape retrieval task using a joint embedding of text and 3D shape, Chen et al. [17] further combined the joint embedding model with a conditional Wasserstein GAN (WGAN) framework [97], which enables the generation of colored voxel shapes in low resolution. To improve the surface quality of the generated 3D shapes, several studies have been conducted using the proposed 3D-text cross-modal dataset by Chen et al. [17]. Li et al. [98] proposed to use class labels to guide the generation of 3D voxel shapes with the assumption that shapes with different labels (e.g., chairs and tables) have different characteristics. They added an independent classifier to the WGAN framework [97] to guide the training process. The classifier could be trained together with the generator to enable more distinctive class features in the generated 3D shapes. To further improve the quality of 3D shapes generated with color and shape, Liu et al. [50] leveraged implicit occupancy [99] as the 3D representation and proposed a word-level spatial transformer [100] to correlate shape features with semantic features of text by decoupling shape and color predictions for learning features in both texts and shapes.

The methods introduced above only support the generation of 3D shapes in individual categories (e.g., the chair category or the table category). The generalizability (the ability to generalization) of these methods remains challenging due to the unavailability and limited size of the paired data of 3D shapes and text descriptions. To improve generalizability, some researchers have tried to utilize some pre-trained models (e.g., Contrastive Language-Image Pre-Training (CLIP) [52]) and zero-shot learning techniques [101]. Sanghi et al. [43] proposed a method called CLIP-forge, which could generate 3D voxel shapes from text descriptions for ShapeNet [49] objects. It required training data (i.e., rendered images, voxel shapes, query points, and occupancy) obtained from 3D shapes without text labels. They first learned an encoding vector of a 3D geometry and then a normalizing flow model [102] of that encoding vector conditioned on a CLIP [52] feature embedding.

CLIP-Forge has good generalizability to ShapeNet [49] categories. To further improve the generalizability to classes outside common 3D shape datasets (e.g., ShapeNet [49] and ModelNet [96]), Jain et al. [103] combined Neural Radiance Field (NeRF) [104] with an image-text loss from CLIP [52] to form Dream Fields. A Dream Field is a neural 3D representation that can return a rendered 2D image given the desired viewpoint. After training, the method could generate colored 3D neural geometry from text prompts without using 3D shape data, resulting in better generalizability.

5.2.2 Text-to-Sketch Generation

Sketches can inspire design ideas [12–14], and text-to-sketch tools could help designers efficiently capture fleeting design inspirations. The generation of images from text descriptions (i.e., text-to-image synthesis/generation) has seen great progress recently [105]. Unlike text-to-image generation, text-to-sketch synthesis is more challenging and can only rely on rigid edge/stroke information without color features (i.e., pixel values) in an image [63].

Text2Sketch [106] applied a Stagewise-GAN (i.e., generative adversarial network) to encode human face attributes identified from text descriptions and transforms those attributes into sketches, which were trained on a manually annotated dataset of text-face sketches. Although the method was applied in face recognition instead of product design, it is worth being introduced here because the method is inspiring and could be applied to the design domain if a different dataset is used. Yuan et al. [63] constructed a bird sketch dataset by modifying the Caltech-UCSD Birds (CUB) dataset [107], based on which they trained a novel GAN-based model, called T2SGAN. The model featured a Conditional Layer-Instance Normalization module that could fuse the image features and sentence vectors, thus efficiently guiding the generation of sketches.

The methods mentioned above were developed for single-object sketch synthesis, and there are also methods for multi-object generation, which could be useful for generating designs part by part. An example of such methods is shown in Figure 8. Huang et al. [53] developed Sketchforme by adopting a two-step neural network: 1) a transformer-based mixture density network for the scene composer to generate high-level layouts of sketches, and 2) a sketch-RNN [16] based object sketcher to generate individual object sketches. The scene composer and the object sketcher were trained using the Visual Genome dataset [108] and the "Quick, Draw!" dataset [109], respectively. Since different datasets of text and sketches can be used, this method helped avoid the requirement for paired data of text description and sketches of an object. Based on [53], Huang et al. [54] took a further step and proposed an interactive sketch generation system called Scones. It used a Composition Proposer to propose a scene-level composition layout of objects and an

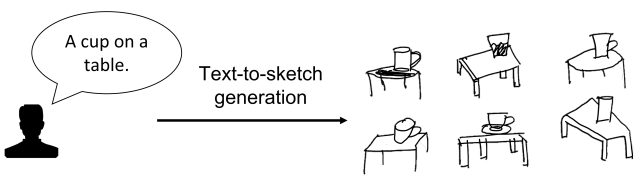

Fig. 8. Demonstration of *text-to-sketch generation*, which can generate sketches that correspond to users' natural language descriptions (NLD).

Object Generator to generate individual object sketches.

5.2.3 Sketch-to-3D Shape Generation

There are mainly two paradigms for 3D shape reconstruction from 2D sketches: the geometric-based method and the learning-based method. Sketch-based interfaces for modeling are a major branch of geometric-based methods [110] and we do not review this line of work in light of the scope of review. We also excluded some methods that apply deep learning techniques, but require predefined geometric models to guide 3D reconstruction, such as the methods presented in [58, 111]. We focus on reviewing deep learning-based methods without using predefined geometric models that require the design of rules.

Deep learning-based sketch-to-3D shape generation without any predefined geometric models was initialized by Lun et al. [35]. They proposed an encoder-multiview-decoder architecture that can extract multiview depth and normal maps from a single sketch or multiple sketches and output a 3D shape in point clouds. The resulting 3D point cloud shape can be converted to a 3D mesh shape for better visualization. 2.5D visual surface geometry (e.g., depth and normal maps) is a representation that can make a 2D image appear to have 3D qualities [112, 113]. Similarly to [35], many works use the strategy of predicting 2.5D information first to guide the generation of 3D shapes. Nozawa et al. [19] extracted depth and mask information from a single input sketch by an encoder-decoder network. Then, a lazy learning [114] method was performed to find similar samples in the dataset to synthesize a 3D shape represented by point clouds. Later, Nozawa et al. [20] extended [19] by changing the architecture with a combination of GAN and lazy learning.

To improve the surface quality of the shapes resulting from their previous work [57], Delanoy et al. [115] proposed to first predict one normal map per input 3D sketch (Type I). Then they fused all normal maps predicted from multiview sketches to the predicted 3D voxel shape to optimize the resulting surface mesh. Li et al. [56] introduced an intermediate CNN layer to model the direction of dense curvature and used an additional output confidence map along with the depth and normal maps extracted using CNNs to generate high-quality 3D mesh shapes. They also provided a user-interaction system for 3D shape design. Similar to the idea of obtaining an intermediate 2.5D representation, Yang et al. [116] proposed a skeleton-aware modeling network to generate 3D human body models using skeletons as the intermediate representation. The network can first interpret sparse joints from input sketches and then predict the Skinned Multi-Person Linear model [117] parameters based on joint-wise features. Although this work focuses on the generation of human bodies, the proposed network can inspire design researchers to consider predicting important feature points to guide the generation of 3D shapes. Li et al. [33] proposed a predictive and generative target-embedding variational autoencoder and demonstrated its effectiveness by solving a sketch-to-3D shape generation problem. The authors used a 3D extrusion shape obtained by extruding a 2D silhouette sketch as an intermediate representation, which transferred the problem to a 3D-3D prediction problem. The approach can predict a high-quality 3D mesh shape from a silhouette sketch without inner contour lines, as shown in Figure 9. In addition to the prediction function, the proposed approach can also generate numerous novel 3D mesh shapes using its generative function.

The efforts of providing an easy-to-use sketching system can be beneficial to novice users for customized design. Delanoy et al. [57] proposed an interactive sketch-to-3D generations system. They used a CNN to transform 3D sketches (Type I) to 3D voxel shapes, and another CNN as an updater to update the predicted 3D shape while users are providing more sketches. The voxel shapes can then be transferred to 3D mesh shapes. However, the output 3D shapes are low-quality due to the high memory consumption of the voxel representation. To improve the surface quality of the resulting 3D shapes, mesh and implicit field have been applied by some interaction systems. For example, Han et al. [58] proposed a novel sketching system to generate 3D mesh human faces and caricatures using a CNN-based deep regression network. The method was trained on a newly proposed dataset extended from FaceWarehouse [118]. Du et al. [59] designed a novel sketching system composed of a part generator and an automatic assembler to generate part-aware man-made objects with complex structures. They used implicit occupancy [99] as the 3D representation which can be transferred to a 3D mesh shape with detailed geometry. Similarly, Wang et al. [119] introduced a novel sketch-to-3D shape method that can segment a given sketch and build a transformation template that is then used to generate multifarious sketches. These sketches are then taken as input to an encoder-multiview-decoder network similar to [35] to generate a 3D point cloud shape. Luo et al. [60] proposed a coarse-to-fine-grained 3D mesh modeling system using 3D sketches as input for animalmorphic head design. A coarse mesh can be first generated by the input 3D sketch. Then, a novel pixel-aligned implicit learning approach is used to guide the deformation of the coarse mesh to produce a more detailed mesh. Guillard et al. [6] introduced an interactive system to reconstruct and edit 3D shapes using implicit field representation, DeepSDF [120] format, from 2D sketches using an encoder-decoder architecture, which can output mesh shapes.

The aforementioned methods are usually trained using one individual category of objects and can only deal with 3D shape generation from sketches within that specific category. To improve the generalizability of the method, Jin et al. [51] proposed a novel network consisting of a VAE (i.e., variational autoencoder) and a volumetric autoencoder to learn the joint embedding of sketches and 3D shapes using various classes of objects. The trained network has good generalizability and can be used to predict 3D voxel shapes based on 2D occluding contours. Zhang et al. [121] are the first to generate a 3D mesh shape from a single free-hand sketch. They proposed a view-aware network based on GAN to explicitly condition the process of generating 3D mesh shapes on view-

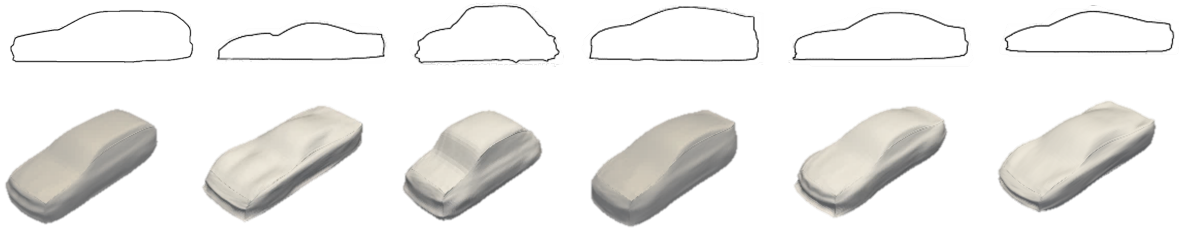

Fig. 9. *Sketch-to-3D shape generation* method by Li et al. [33]. The first row shows the input 2D silhouette sketches, and the corresponding predicted 3D mesh shapes are shown in the second row.

points. The method can improve generation quality and bring controllability to output shapes by explicitly adjusting viewpoints, which can be well generalized to out-of-distribution data.

The methods introduced above have to be trained using supervised learning, which means that the training data must be pairs of sketches and 3D shapes (i.e., labeled data). Wang et al. [122] proposed an unsupervised learning method for sketch-to-3D shape reconstruction. They embedded unpaired sketches and rendered images from 3D shapes to a common latent space by training an adaption network via autoencoder with adversarial loss. During the inference of 3D shapes from sketches, they retrieved several nearest-neighboring 3D shapes from the training dataset as prior knowledge for a 3D GAN to generate new 3D shapes that best match the input sketch. This method can only output very coarse 3D voxel shapes but provides an interesting idea based on unsupervised learning for sketch-to-3D shape generation.

In addition to the usage of popular 3D shape representations (e.g., point clouds, voxels, meshes, and implicit representation) in sketch-to-3D shape generation, new 3D representations are gaining more and more attention in this field. For example, Smirnov et al. [5, 123] proposed a novel deformable parametric template composed of Coon patches that can naturally fit into a conventional CAD modeling pipeline. The resulting 3D shapes can be easily converted to NURBS representation, allowing edits in CAD software.

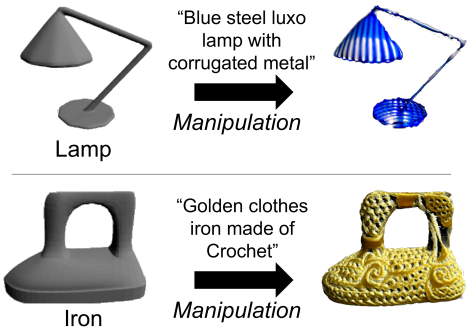

Fig. 10. *Text-to-3D shape manipulation* method, Text2Mesh by Michel et al. [36]. The method can manipulate an existing mesh shape by adding color, texture, and geometric details driven by a target natural language description. The figure is used with permission.

5.3 RQ 1-(3): What DLCMT Methods Can Be Used in Design Integration of Conceptual Design?

In this section, we introduce some works relevant to text-to-3D shape and sketch-to-3D shape integration methods. These methods allow designers to further edit and manipulate 3D designs by changing text prompts or sketches.

The sketch-to-3D shape generation method introduced by Jin et al. [51] could be further used to manipulate a given 3D voxel shape to target input sketches with the learned joint embedding space. However, it focuses on manipulating the outline of a given 3D shape. To enable manipulation of color and shape, CLIP-NeRF [61] was proposed based on CLIP [52], which has a disentangled conditional NeRF [104] architecture by introducing a shape code to deform the 3D volumetric field and an appearance code to control the colors. The method can edit a given colored 3D voxel shape to meet the target semantic description of color and shape. The text-to-3D generation method [50] can also allow intuitive manipulation of the color and shape of a generated 3D mesh shape simply by changing the input semantic keywords of color or shape.

To enable detailed edits or manipulation of geometries, in some works a differentiable renderer has been applied. Sketch2Mesh [6] introduced in Section 5.2.3 can also perform shape editing due to the integrated differentiable renderer. Using the representation power of CLIP [52], Michel et al. [36] proposed Text2Mesh (see Figure 10) to manipulate a given 3D mesh shape by predicting color and local geometric details that conform to the description of the target text.

There have been a series of DLCMT methods that can be applied to product shape design in different design steps of conceptual design. As a summary of the review, DLCMT methods indeed provide opportunities to address the two major challenges as discussed in Section 2 because they can (1) take various design modalities as input and provide methods catering to Design Search, Design Creation, and Design Integration, and (2) improve design creativity by actively involving human input [53, 54, 59, 60]. Taking advantage of these opportunities and implementing the appropriate DLCMT methods in conceptual design can therefore accelerate the search and iteration of design concepts (e.g. [17, 44, 48]) and the modification of designs (e.g., [36, 43, 51, 58]). We also observe that DLCMT methods could be particularly useful in design applications, such as design democratization, design education, and immersive design (e.g., [17, 44, 48, 62, 90]).

5.4 RQ 2: What are the Challenges in Applying DLCMT to Conceptual Design and How Can They be Addressed?

Examination of the literature has helped us identify several challenges in applying DLCMT methods to conceptual design. DLCMT has been focusing on shape synthesis, which can be applied in product shape design, as discussed above. However, Regenwetter et al. [3] state that 3D synthesis work is only tangential to engineering design because they focus more on visual appearance, rather than functional performance or manufacturability. Although we partially agree with [3] that the overlap between shape synthesis and engineering design is insignificant in light of the importance of shape design, we must admit that product shape is not the only focus in conceptual design. Other factors, such as engineering performance, system design features, and manufacturability, should also be considered and can be incorporated into the data-driven design cycle even in the early stages of the design.

In this section, we discuss in detail the challenges of applying DLCMT methods to engineering design from four aspects, including the lack of cross-modal datasets that incorporate engineering performance and manufacturability, complex systems design using DLCMT, 3D representations in DLCMT, and the generalizability of DLCMT methods.

5.4.1 The Lack of Cross-Modal Datasets that Incorporate Engineering Performance and Manufacturability

Data is the fuel for deep learning-based design methods. Data sparsity is a challenging issue for data-driven design methods, and there is generally a deficiency of big practical data [3], regardless of the data modality, to train useful and meaningful models for engineered products. Unlike the computer science community, where numerous open source unimodal or cross-modal datasets, such as [17, 49, 83, 96], are available to researchers to compare their methods with state-of-the-art methods. For example, 16 articles (e.g., [6, 17, 34, 86, 121]) use ShapeNet [49] as the training data of their methods. There is a lack of similar benchmark datasets in the engineering design field. Even if those datasets from computer science can also be beneficial to the engineering design community, they mainly focus on the shape of objects and have little emphasis on downstream engineering-related information. Using text-to-3D shape methods as an example, a user could say "I want an SUV with low fuel consumption". An SUV car shape could be easily generated, but we would not know whether the drag coefficients of the generated designs meet the requirement or not. We might ask the following question: How could a computer understand that NL description and translate it into a primitive SUV car shape taking into account the drag performance? Therefore, finding answers to this question could be an interesting research direction.

Similarly, it is also worth exploring how other downstream engineering requirements and constraints (e.g., manufacturability) can be counted when applying DLCMT to engineering design. We have not found any DLCMT methods that take into account engineering performance and manufacturability. One challenge here is the lack of such datasets. The difficulties primarily rest in the cost (either monetary or time) of running high-fidelity computational or physical experiments. Moreover, certain experimental data could be confidential for commercial or military purposes. The availability of large cross-modal datasets with engineering performance and manufacturability information could greatly ease the verification and validation of existing methods for DLCMT and promote the development of new DLCMT methods for the design of engineered products.

5.4.2 Complex Systems Design Using DLCMT

A few DLCMT studies ([53, 54, 59]) aim to generate designs part by part considering the structural relationship among components, which can be potentially applied to the design of systems. But this leaves a large space for engineering design researchers to investigate in the future. The challenges of addressing systems design using DLCMT mainly stem from the structural complexity of an engineered product, such as dependencies, constraints, and the relationship between components.

An engineered product is usually a system consisting of interconnected parts with complex dependencies. To take into account parts' dependency information, there are generally two ways to support the conceptual design of a product at the system level when applying DLCMT methods. In the first method, each component of the product is generated separately using DLCMT, and then the components are assembled either automatically using rules-based computer algorithms or manually [53, 54]. The second method is often referred to as part-aware generative design [30, 124, 125]. The objective of using DLCMT methods for part-aware design is to learn the structural relationships and dependencies between parts directly from the training data so that parts generation and assembly can be automatically completed.

Compared to the first method, the second method can save time and the cost of additional assembly steps. Those steps are often non-trivial, especially when one wants to computerize the assembly process in CAD software. In addition, part-aware generative design methods better capture the geometric details of 3D shapes [124, 125]. For example, in the transition regions between two components (e.g., the connection regions between the side rear mirrors and the car body). These geometric details may significantly influence the engineering performance (for example, aerodynamic drag) of a design.

As mentioned above, there are a few studies, i.e., text-to-sketch generation [53, 54] and sketch-to-3D [59] methods for DLCMT attempting to integrate the concept of part-aware design, but most methods treat the design object as a single monolithic part without a systems design perspective. Considering engineering applications, treating a design as a whole piece could limit the transition of the generated design shapes to later design stages, since components are usually manufactured separately. Attention has been paid to by the

engineering design community [30, 126] for part-aware design. However, how to enable part-aware design in DLCMT remains underexplored and is an important research direction.

5.4.3 3D Representations in DLCMT

Designs can be factored using different representations for storage, computation, and presentation. For example, 3D representation matters both visual quality and computational cost when implementing DLCMT, and the choice between them is often a difficult decision. Furthermore, in engineering design applications, the choice of 3D representation also influences the compatibility with downstream engineering analysis in CAD and CAE software. In what follows, we share our insight into the challenges associated with 3D representation in both aspects.

3D shapes with high visual quality and rich geometric details can help designers better understand a design concept. Voxels, point clouds, and meshes are the most commonly used representations for 3D geometry. Similar to the pixels of images, voxel grids are naturally adapted to the convolutional neural network (CNN) model, which is the major reason for its prevalence in 3D geometry learning research. The majority of the DLCMT methods (e.g., [17, 43, 57, 67, 97, 98, 122]) uses voxels for 3D shape representation. Voxel shapes are usually needed to be converted to mesh shapes for better visualization. However, the transformed mesh shapes will look coarse if the resolution of the voxel shapes is low. This could negatively influence the subjective evaluation of the shape of a design concept, and the design concept might be overlooked by designers. An intuitive way to improve the resolution of the resulting 3D voxel shapes is to use high-resolution training data, but this may not be feasible due to the limited computing resources for training the neural network. Fukamizu et al. [18] provided a two-stage strategy to synthesize high-resolution 3D voxel shapes from natural language, which could be an inspiring method for dealing with low-resolution issues. Point clouds [19, 20, 35, 127] are more efficient in representing 3D objects, but do not cover geometric details. For example, it does not encode the relationship between points and the resulting topology of an object, leading to a challenging conversion to meshes. Using meshes [56, 58, 121, 128] for 3D representation could generally alleviate the low visual quality and data storage problems, but, in the meantime, it is challenging to prepare meshes for deep learning methods due to their discrete face structures and unordered elements. Furthermore, the topology of 3D shapes cannot be easily handled using meshes. Implicit representation of 3D shapes [6, 59, 60, 120] represents the surface of a shape by a continuous volumetric field that encodes the boundary of the shape as the set at the zero level of the learned implicit 3D shape function. It can better address different topologies of 3D shapes and requires less data storage, which is a promising representation for high-resolution 3D shapes. See Table 2 for the pros and cons of applying those four representations to deep learning methods.

In addition to the above four representations, there are a few new 3D representations that are promising for handling the trade-off between the effectiveness of training neural networks and the quality of the resulting 3D shapes. Neural Radiance Field (NeRF) [61,103,104] is a method for generating novel views of scenes or objects. It can take a set of input images of an object and render the complete object by interpolating between the images. NeRF [104] is also topology-free and can be sampled at high spatial resolutions. However, 3D shapes represented by NeRF are "hidden in the black box" and we can only observe them through images rendered from different viewpoints. All the 3D representations mentioned above (i.e., voxels, point clouds, meshes, NeRF, and implicit representation) are generally not adapted to CAD software. This often brings about compatibility issues that could impede downstream editing and engineering analyses of the generated 3D shapes. To solve these problems, there are typically two ways. One way is to convert them to CAD models (e.g., converting STL/OBJ meshes to B-Rep solids). Another way is to handle the CAD shape data directly in deep learning models. Deep learning of unimodal CAD data is still an underexplored field, although some methods [129–133] and CAD datasets [134–137] have recently been introduced. DLCMT directly using CAD data [5] can be even more challenging due to the domain gap between design modalities and turns out to be a promising research direction.

Choosing the most appropriate 3D representation compatible with the adopted deep learning technique remains a challenging task. It involves considerations of data availability, data preprocessing, computational cost, visual quality of the resulting 3D shapes, data postprocessing, and the ability to adapt to later design stages.

5.4.4 Generalizability of DLCMT Methods

Finally, we noticed that efforts have been made to make the DLCMT methods more generalizable, independent of the variation between design objects (e.g., [36, 51]). There are advantages and disadvantages to generalizing the methods. On the one hand, the diversity in different methods helps address the unique nature of different design problems, so a generalized approach may not be optimal for solving a specific design problem. On the other hand, generalizability allows a method to apply to a wider range of design problems. We focus on discussing the advantages here since we observe trending efforts (e.g., [43, 103]) aiming to improve the generalizability of DLCMT methods in the review. It is challenging for deep learning methods to be generalized across multiple design problems [3]. The generalizability of a deep learning method means its ability to generalize to classes of objects beyond those used for training data. For engineering design applications, due to the sparsity of training data and the special treatment designed in the neural network architecture for a specific problem, a deep learning-based design method is difficult to generalize even in the cases where one design modality (e.g., 2D sketches or 3D shapes) is involved, let alone the generalization issues of applying DLCMT methods that involve multiple different modalities.

Table 2. Comparison of pros and cons of the three representations to deep learning methods

3D Representation	Pros	Cons
Voxels	• The data structure in fourth-order tensor makes it easy to be adapted in 3D convolution operations in deep learning methods • Can deal with 3D shapes with arbitrary topology	• Low visual quality • High computational cost because the number of the 3D representation parameters scale with the increase of spatial resolution in cubes • Cannot be directly used in engineering analyses (e.g., finite element analysis (FEA)) for performance evaluation
Point clouds	• Compatible with the output data format of common scanning software • Compact for data storage and management • Can deal with 3D shapes with arbitrary topology	• Low visual quality • No detailed geometric information about relationships between points making it hard to convert to meshes • Cannot be directly used in engineering analyses (e.g., FEA) for performance evaluation
Meshes	• High visual quality • Compact for data storage and management • Widely-accepted 3D representation in computer graphics • Compatible with downstream engineering software, such as the FEA and computational fluid dynamics (CFD) tools	• Discrete and disordered elements make it challenging to be processed by deep learning methods • Hard to deal with 3D shapes with arbitrary topology
Implicit representation	• High visual quality • Easy adaption to deep learning methods • Compact for data storage and management • Can deal with 3D shapes with arbitrary topology	• Need to use rendering techniques to extract the isosurface of the 3D shapes for visualization • Cannot be directly used in engineering analyses (e.g., FEA) for performance evaluation

Some methods [43, 103] utilize transfer learning techniques (e.g., zero-shot learning) and pre-trained models (e.g., CLIP [52]) or specially designed neural network architectures (e.g., unsupervised learning methods [119]) to improve generalizability, which could be good starting points for the engineering design community to further explore other possibilities. The challenge of generalizing the methods for DLCMT couples with other challenges and requires a community-wide effort to share datasets, create data repositories, define benchmark problems, and develop testing standards.

In summary, we have discussed the opportunities and challenges associated with applying DLCMT methods to conceptual design and proposed potential solutions to overcome the challenges with the insight gained from this literature review effort. The insights generated can potentially point to promising research directions for future studies.

6 Research Questions for Future Design Research

We notice that the opportunities and challenges identified previously are highly related to several trending topics in the engineering design community. In this section, we propose six research questions (RQs) that relate DLCMT to these trending topics: RQ (1) $\rightarrow$ design representations [138]; RQ (2) $\rightarrow$ generalizability and transferability of deep learning-based design methods [22]; RQ (3) $\rightarrow$ decision-making in AI-enabled design process [139]; RQ (4) and (5) $\rightarrow$ human-AI collaboration [23]; RQ (6) $\rightarrow$ design creativity in deep learning-based design process [37]. These RQs also point to potential research directions (see Section 7 for detail) where DLCMT can lead to. We hope these RQs can arouse a wide range of discussion and call for more efforts within the engineering design community to develop and apply DLCMT methods to address the challenges associated with conceptual design and beyond.

(1) What are the guidelines for selecting the most appropriate design representations in DLCMT?
(2) How much can the generalizability and transferability of the latent representation of multimodal data learned from DLCMT be extended across different product shape categories?
(3) Since DLCMT methods can shorten the cycle of generating designs and even connect to the downstream engineering analyses and manufacturing requirements, how could the information coming from the later design stages influence the regeneration of design concepts, and thereby a designer's decisions?
(4) DLCMT methods have the potential to facilitate the data-driven design process with humans in the loop, but how can we balance the involvement of humans and computers, and facilitate effective bidirectional human-AI communications to better stimulate designers' creativity at the human-AI interface?
(5) With the establishment of the human-AI interaction in the conceptual design based on DLCMT, what could the co-evolution between humans and AI look like?
(6) Although design creativity can be augmented by bringing humans in the loop when using DLCMT methods for product shapes generation, these methods could suffer from the limitation of data interpolation inherently rooted in data-driven design methods. Fundamental questions, such as what new mechanisms and neural network architectures can be built to enable the algorithm to extrapolate beyond the training data, thus more effectively augmenting designers' creativity, shall be further explored in the future.

7 Closing Remarks

In this paper, we conducted a systematic review of the methods for deep learning of cross-modal tasks (DLCMT), including text-to-sketch, text-to-3D shape, and sketch-to-3D shape retrieval and generation methods, for the conceptual design of product shapes. Those methods could be applied in the Design Search, Design Creation, and Design Integration steps of conceptual design. Unlike other deep learning methods applied in engineering design, DLCMT allows human input of texts and sketches, which can explicitly reflect designers' and/or users' preferences. As designers can be more actively involved in such a design process, human-computer interaction and collaboration are promoted, thereby it has a great potential to improve the conceptual design of products using a data-driven design process with humans in the loop compared to traditional design automation methods and computer-aided design methods. DLCMT could also facilitate the engineering design education and democratization of product development by allowing intuitive inputs (e.g., text descriptions and sketches), and an immersive design environment by integrating VR, AR, and MR techniques.

With the attempt to apply new 3D data representations in DLCMT and the availability of more public datasets, opportunities open up for the development of new methods for DLCMT. However, the deficiency of training datasets, trade-off in the choice of representations of 3D shapes, lack of consideration of engineering performance, manufacturability, and part-aware design, and the ability of generalization still challenge the engineering design community to apply DLCMT to engineered product design. We would like to encourage attention and efforts from the engineering design community.

There are a few limitations in the current literature review that the authors would like to acknowledge and share. First, the set of keywords used to search the literature has covered all topics in our scope of the review. However, other topics, such as shape-to-text generation (namely, shape captioning in the literature), could also be of interest to the engineering design community. Second, for the topics of sketch-to-3D shape retrieval and generation, we did not include all relevant articles, although we have covered the most influential and the most recent publications.

In the future, we will continue the review and conduct a more comprehensive analysis of the relevant works on DLCMT. Besides the review effort, we see the merit of conducting a comparative study to further understand the effects of DLCMT on the conceptual design by enabling and disabling the DLCMT-based assistance in the design process. We believe that the methods reviewed, the discussion of opportunities, challenges, potential solutions, and future research directions of applying DLCMT to conceptual product shape design can benefit the data-driven design research in the engineering design community. We hope this review effort can also facilitate the discussion and attract more attention from the engineering design community and industry stakeholders when applying DLCMT to improve the conceptual design of product shapes and beyond.

Acknowledgements

The authors gratefully acknowledge the financial support from the National Science Foundation through award 2207408.

References

[1] Ulrich, K. T., 2003. *Product design and development*. Tata McGraw-Hill Education.

[2] Chakrabarti, A., Shea, K., Stone, R., Cagan, J., Campbell, M., Hernandez, N. V., and Wood, K. L., 2011. "Computer-based design synthesis research: an overview". *Journal of Computing and Information Science in Engineering,* **11**(2).

[3] Regenwetter, L., Nobari, A. H., and Ahmed, F., 2022. "Deep generative models in engineering design: A review". *Journal of Mechanical Design,* **144**(7), p. 071704.

[4] Liu, Z., Lin, Y., and Sun, M., 2020. *Cross-Modal Representation*. Springer Singapore, Singapore, pp. 285–317.

[5] Smirnov, D., Bessmeltsev, M., and Solomon, J., 2020. "Learning manifold patch-based representations of man-made shapes". In International Conference on Learning Representations.

[6] Guillard, B., Remelli, E., Yvernay, P., and Fua, P., 2021. "Sketch2mesh: Reconstructing and editing 3d shapes from sketches". In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13023–13032.

[7] Otto, K. N., and Wood, K., 2001. *Product design: techniques in reverse engineering and new product development*. Prentice Hall, Upper Saddle River.

[8] Yang, M. C., 2009. "Observations on concept generation and sketching in engineering design". *Research in Engineering Design,* **20**(1), pp. 1–11.

[9] Hyun, K. H., and Lee, J.-H., 2018. "Balancing homogeneity and heterogeneity in design exploration by synthesizing novel design alternatives based on genetic algorithm and strategic styling decision". *Advanced Engineering Informatics,* **38**, pp. 113–128.

[10] Mountstephens, J., and Teo, J., 2020. "Progress and challenges in generative product design: A review of systems". *Computers,* **9**(4), p. 80.

[11] Ahmed, F., Ramachandran, S. K., Fuge, M. D., Hunter, S. T., and Miller, S. R., 2018. "Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel". *Journal of Mechanical Design*.

[12] Krish, S., 2011. "A practical generative design method". *Computer-Aided Design,* **43**(1), pp. 88–100.

[13] Pratt, M. J., Anderson, B. D., and Ranger, T., 2005. "Towards the standardized exchange of parameterized feature-based cad models". *Computer-Aided Design,* **37**(12), pp. 1251–1265.

[14] Menezes, A., and Lawson, B., 2006. "How designers perceive sketches". *Design studies,* **27**(5), pp. 571–585.

[15] Xu, P., Hospedales, T. M., Yin, Q., Song, Y.-Z., Xi-

ang, T., and Wang, L., 2020. "Deep learning for free-hand sketch: A survey and a toolbox". *arXiv preprint arXiv:2001.02600*.

[16] Ha, D., and Eck, D., 2018. "A neural representation of sketch drawings". In International Conference on Learning Representations.

[17] Chen, K., Choy, C. B., Savva, M., Chang, A. X., Funkhouser, T., and Savarese, S., 2018. "Text2shape: Generating shapes from natural language by learning joint embeddings". In Asian conference on computer vision, Springer, pp. 100–116.

[18] Fukamizu, K., Kondo, M., and Sakamoto, R., 2019. "Generation high resolution 3d model from natural language by generative adversarial network". *arXiv preprint arXiv:1901.07165*.

[19] Nozawa, N., Shum, H. P., Ho, E. S., and Morishima, S., 2020. "Single sketch image based 3d car shape reconstruction with deep learning and lazy learning.". In VISIGRAPP (1: GRAPP), pp. 179–190.

[20] Nozawa, N., Shum, H. P., Feng, Q., Ho, E. S., and Morishima, S., 2022. "3d car shape reconstruction from a contour sketch using gan and lazy learning". *The Visual Computer,* **38**(4), pp. 1317–1330.

[21] Wendrich, R. E., 2018. "Multiple modalities, sensoriums, experiences in blended spaces with toolness and tools for conceptual design engineering". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 51739, American Society of Mechanical Engineers, p. V01BT02A046.

[22] Song, B., Miller, S., and Ahmed, F., 2022. "Hey, ai! can you see what i see? multimodal transfer learning-based design metrics prediction for sketches with text descriptionss". In ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[23] Song, B., Zurita, N. S., Zhang, G., Stump, G., Balon, C., Miller, S., Yukish, M., Cagan, J., and McComb, C., 2020. "Toward hybrid teams: A platform to understand human-computer collaboration during the design of complex engineered systems". In Proceedings of the Design Society: DESIGN Conference, Vol. 1, Cambridge University Press, pp. 1551–1560.

[24] Li, X., Wang, Y., and Sha, Z., 2022. "Deep learning of cross-modal tasks for conceptual design of engineered products: A review". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers. *Accepted*.

[25] Chen, W., Chiu, K., and Fuge, M. D., 2020. "Airfoil design parameterization and optimization using bézier generative adversarial networks". *AIAA Journal,* **58**(11), pp. 4723–4735.

[26] Oh, S., Jung, Y., Kim, S., Lee, I., and Kang, N., 2019. "Deep generative design: Integration of topology optimization and generative models". *Journal of Mechanical Design,* **141**(11).

[27] Dering, M., Cunningham, J., Desai, R., Yukish, M. A., Simpson, T. W., and Tucker, C. S., 2018. "A physics-based virtual environment for enhancing the quality of deep generative designs". In ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[28] Shu, D., Cunningham, J., Stump, G., Miller, S. W., Yukish, M. A., Simpson, T. W., and Tucker, C. S., 2020. "3d design using generative adversarial networks and physics-based validation". *Journal of Mechanical Design,* **142**(7).

[29] Zhang, W., Yang, Z., Jiang, H., Nigam, S., Yamakawa, S., Furuhata, T., Shimada, K., and Kara, L. B., 2019. "3d shape synthesis for conceptual design and optimization using variational autoencoders". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 59186, American Society of Mechanical Engineers, p. V02AT03A017.

[30] Li, X., Xie, C., and Sha, Z., 2021. "Part-aware product design agent using deep generative network and local linear embedding". In Proceedings of the 54th Hawaii International Conference on System Sciences, p. 5250.

[31] Brock, A., Lim, T., Ritchie, J. M., and Weston, N., 2016. "Context-aware content generation for virtual environments". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 50084, American Society of Mechanical Engineers, p. V01BT02A045.

[32] Qin, F., Qiu, S., Gao, S., and Bai, J., 2022. "3d cad model retrieval based on sketch and unsupervised variational autoencoder". *Advanced Engineering Informatics,* **51**, p. 101427.

[33] Li, X., Xie, C., and Sha, Z., 2022. "A predictive and generative design approach for three-dimensional mesh shapes using target-embedding variational autoencoder". *Journal of Mechanical Design,* **144**(11), p. 114501.

[34] Qi, A., Gryaditskaya, Y., Song, J., Yang, Y., Qi, Y., Hospedales, T. M., Xiang, T., and Song, Y.-Z., 2021. "Toward fine-grained sketch-based 3d shape retrieval". *IEEE Transactions on Image Processing,* **30**, pp. 8595–8606.

[35] Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., and Wang, R., 2017. "3d shape reconstruction from sketches via multi-view convolutional networks". In 2017 International Conference on 3D Vision (3DV), IEEE, pp. 67–77.

[36] Michel, O., Bar-On, R., Liu, R., Benaim, S., and Hanocka, R., 2022. "Text2mesh: Text-driven neural stylization for meshes". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13492–13502.

[37] Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M., 2017. "Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms". *arXiv preprint arXiv:1706.07068*.

[38] Chen, W., and Ahmed, F., 2021. "Padgan: Learning to generate high-quality novel designs". *Journal of Mechanical Design,* **143**(3).

[39] Burnap, A., Liu, Y., Pan, Y., Lee, H., Gonzalez, R., and Papalambros, P. Y., 2016. "Estimating and exploring the product form design space using deep generative models". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 50107, American Society of Mechanical Engineers, p. V02AT03A013.

[40] Judd, G., and Steenkiste, P., 2003. "Providing contextual information to pervasive computing applications". In Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003)., IEEE, pp. 133–142.

[41] Valdez, S., Seepersad, C., and Kambampati, S., 2021. "A framework for interactive structural design exploration". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 85390, American Society of Mechanical Engineers, p. V03BT03A006.

[42] Starly, B., Angrish, A., and Cohen, P., 2019. "Research directions in democratizing innovation through design automation, one-click manufacturing services and intelligent machines". *arXiv preprint arXiv:1909.10476*.

[43] Sanghi, A., Chu, H., Lambourne, J. G., Wang, Y., Cheng, C.-Y., Fumero, M., and Malekshan, K. R., 2022. "Clip-forge: Towards zero-shot text-to-shape generation". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18603–18613.

[44] Giunchi, D., Sztrajman, A., James, S., and Steed, A., 2021. "Mixing modalities of 3d sketching and speech for interactive model retrieval in virtual reality". In ACM International Conference on Interactive Media Experiences, pp. 144–155.

[45] Khan, K. S., Kunz, R., Kleijnen, J., and Antes, G., 2003. "Five steps to conducting a systematic review". *Journal of the royal society of medicine,* **96**(3), pp. 118–121.

[46] Kingma, D. P., and Welling, M., 2013. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114*.

[47] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014. "Generative adversarial nets". In Advances in neural information processing systems, pp. 2672–2680.

[48] Wang, F., Kang, L., and Li, Y., 2015. "Sketch-based 3d shape retrieval using convolutional neural networks". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875–1883.

[49] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., and Su, H., 2015. "Shapenet: An information-rich 3d model repository". *arXiv preprint arXiv:1512.03012*.

[50] Liu, Z., Wang, Y., Qi, X., and Fu, C.-W., 2022. "Towards implicit text-guided 3d shape generation". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17896–17906.

[51] Jin, A., Fu, Q., and Deng, Z., 2020. "Contour-based 3d modeling through joint embedding of shapes and contours". In Symposium on Interactive 3D Graphics and Games, pp. 1–10.

[52] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. "Learning transferable visual models from natural language supervision". In International Conference on Machine Learning, PMLR, pp. 8748–8763.

[53] Huang, F., and Canny, J. F., 2019. "Sketchforme: Composing sketched scenes from text descriptions for interactive applications". In Proceedings of the 32nd annual ACM symposium on user interface software and technology, pp. 209–220.

[54] Huang, F., Schoop, E., Ha, D., and Canny, J., 2020. "Scones: towards conversational authoring of sketches". In Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 313–323.

[55] Li, B., Yuan, J., Ye, Y., Lu, Y., Zhang, C., and Tian, Q., 2021. "3d sketching for 3d object retrieval". *Multimedia Tools and Applications,* **80**(6), pp. 9569–9595.

[56] Li, C., Pan, H., Liu, Y., Tong, X., Sheffer, A., and Wang, W., 2018. "Robust flow-guided neural prediction for sketch-based freeform surface modeling". *ACM Transactions on Graphics (TOG),* **37**(6), pp. 1–12.

[57] Delanoy, J., Aubry, M., Isola, P., Efros, A. A., and Bousseau, A., 2018. "3d sketching using multi-view deep volumetric prediction". *Proceedings of the ACM on Computer Graphics and Interactive Techniques,* **1**(1), pp. 1–22.

[58] Han, X., Gao, C., and Yu, Y., 2017. "Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling". *ACM Transactions on graphics (TOG),* **36**(4), pp. 1–12.

[59] Du, D., Zhu, H., Nie, Y., Han, X., Cui, S., Yu, Y., and Liu, L., 2021. "Learning part generation and assembly for sketching man-made objects". In Computer Graphics Forum, Vol. 40, Wiley Online Library, pp. 222–233.

[60] Luo, Z., Zhou, J., Zhu, H., Du, D., Han, X., and Fu, H., 2021. "Simpmodeling: Sketching implicit field to guide mesh modeling for 3d animalmorphic head design". In The 34th Annual ACM Symposium on User Interface Software and Technology, pp. 854–863.

[61] Wang, C., Chai, M., He, M., Chen, D., and Liao, J., 2022. "Clip-nerf: Text-and-image driven manipulation of neural radiance fields". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3835–3844.

[62] Stemasov, E., Wagner, T., Gugenheimer, J., and Rukzio, E., 2022. "Shapefindar: Exploring in-situ spatial search for physical artifact retrieval using mixed reality". In CHI Conference on Human Factors in Computing Systems, pp. 1–12.

[63] Yuan, S., Dai, A., Yan, Z., Guo, Z., Liu, R., and Chen, M., 2021. "Sketchbird: Learning to generate bird sketches from text". In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2443–2452.

[64] Min, P., Kazhdan, M., and Funkhouser, T., 2004. "A comparison of text and shape matching for retrieval of online 3d models". In International Conference on Theory and Practice of Digital Libraries, Springer, pp. 209–220.

[65] Liu, Z., Lin, Y., and Sun, M., 2020. *Cross-Modal Representation*. Springer Singapore, Singapore, pp. 285–317.

[66] Haeusser, P., Mordvintsev, A., and Cremers, D., 2017. "Learning by association–a versatile semi-supervised training method for neural networks". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 89–98.

[67] Han, Z., Shang, M., Wang, X., Liu, Y.-S., and Zwicker, M., 2019. "Y2seq2seq: Cross-modal representation learning for 3d shape and text by joint reconstruction and prediction of view and word sequences". In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, pp. 126–133.

[68] Shilane, P., Min, P., Kazhdan, M., and Funkhouser, T., 2004. "The princeton shape benchmark". In Proceedings Shape Modeling Applications, 2004., IEEE, pp. 167–178.

[69] Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M. J., Johan, H., Matsuda, T., et al., 2014. "A comparison of methods for sketch-based 3d shape retrieval". *Computer Vision and Image Understanding,* **119**, pp. 57–80.

[70] Chopra, S., Hadsell, R., and LeCun, Y., 2005. "Learning a similarity metric discriminatively, with application to face verification". In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, IEEE, pp. 539–546.

[71] Zhu, F., Xie, J., and Fang, Y., 2016. "Learning cross-domain neural networks for sketch-based 3d shape retrieval". In Proceedings of the AAAI conference on artificial intelligence, Vol. 30.

[72] Dai, G., Xie, J., and Fang, Y., 2018. "Deep correlated holistic metric learning for sketch-based 3d shape retrieval". *IEEE Transactions on Image Processing,* **27**(7), pp. 3374–3386.

[73] Dai, G., Xie, J., Zhu, F., and Fang, Y., 2017. "Deep correlated metric learning for sketch-based 3d shape retrieval". In Thirty-First AAAI Conference on Artificial Intelligence.

[74] Chen, J., and Fang, Y., 2018. "Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval". In Proceedings of the European Conference on Computer Vision (ECCV), pp. 605–620.

[75] Xia, Y., Wang, S., You, L., and Zhang, J., 2021. "Semantic similarity metric learning for sketch-based 3d shape retrieval". In International Conference on Computational Science, Springer, pp. 59–69.

[76] Yang, H., Tian, Y., Yang, C., Wang, Z., Wang, L., and Li, H., 2022. "Sequential learning for sketch-based 3d model retrieval". *Multimedia Systems,* **28**(3), pp. 761–778.

[77] Kaya, M., and Bilge, H. Ş., 2019. "Deep metric learning: A survey". *Symmetry,* **11**(9), p. 1066.

[78] Xie, J., Dai, G., Zhu, F., and Fang, Y., 2017. "Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5068–5076.

[79] Chen, J., Qin, J., Liu, L., Zhu, F., Shen, F., Xie, J., and Shao, L., 2019. "Deep sketch-shape hashing with segmented 3d stochastic viewing". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 791–800.

[80] Niu, Z., Zhong, G., and Yu, H., 2021. "A review on the attention mechanism of deep learning". *Neurocomputing,* **452**, pp. 48–62.

[81] Liang, S., Dai, W., and Wei, Y., 2021. "Uncertainty learning for noise resistant sketch-based 3d shape retrieval". *IEEE Transactions on Image Processing,* **30**, pp. 8632–8643.

[82] Liu, Q., and Zhao, S., 2021. "Guidance cleaning network for sketch-based 3d shape retrieval". In Journal of Physics: Conference Series, Vol. 1961, IOP Publishing, p. 012072.

[83] Li, B., Lu, Y., Godil, A., Schreck, T., Aono, M., Johan, H., Saavedra, J. M., and Tashiro, S., 2013. *SHREC'13 track: large scale sketch-based 3D shape retrieval*.

[84] Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al., 2014. "Shrec'14 track: Extended large scale sketch-based 3d shape retrieval". In Eurographics workshop on 3D object retrieval, Vol. 2014, pp. 121–130.

[85] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E., 2015. "Multi-view convolutional neural networks for 3d shape recognition". In Proceedings of the IEEE international conference on computer vision, pp. 945–953.

[86] Navarro, P., Orlando, J. I., Delrieux, C., and Iarussi, E., 2021. "Sketchzooms: Deep multi-view descriptors for matching line drawings". In Computer Graphics Forum, Vol. 40, Wiley Online Library, pp. 410–423.

[87] Manda, B., Dhayarkar, S., Mitheran, S., Viekash, V., and Muthuganapathy, R., 2021. "'cadsketchnet'-an

annotated sketch dataset for 3d cad model retrieval with deep neural networks". *Computers & Graphics,* **99**, pp. 100–113.

[88] Jayanti, S., Kalyanaraman, Y., Iyer, N., and Ramani, K., 2006. "Developing an engineering shape benchmark for cad models". *Computer-Aided Design,* **38**(9), pp. 939–953.

[89] Kim, S., Chi, H.-g., Hu, X., Huang, Q., and Ramani, K., 2020. "A large-scale annotated mechanical components benchmark for classification and retrieval tasks with deep neural networks". In European Conference on Computer Vision, Springer, pp. 175–191.

[90] Ye, Y., Li, B., and Lu, Y., 2016. "3d sketch-based 3d model retrieval with convolutional neural network". In 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, pp. 2936–2941.

[91] Yang, Y., and Hospedales, T. M., 2015. "Deep neural networks for sketch recognition". *arXiv preprint arXiv:1501.07873,* **1**(2), p. 3.

[92] Li, B., Lu, Y., Duan, F., Dong, S., Fan, Y., Qian, L., Laga, H., Li, H., Li, Y., Lui, P., et al., 2016. "Shrec'16 track: 3d sketch-based 3d shape retrieval".

[93] Giunchi, D., James, S., and Steed, A., 2018. "3d sketching for interactive model retrieval in virtual reality". In Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, pp. 1–12.

[94] Jahan, T., Guan, Y., and van Kaick, O., 2021. "Semantics-guided latent space exploration for shape generation". In Computer Graphics Forum, Vol. 40, Wiley Online Library, pp. 115–126.

[95] Wang, Y., Asafi, S., Van Kaick, O., Zhang, H., Cohen-Or, D., and Chen, B., 2012. "Active co-analysis of a set of shapes". *ACM Transactions on Graphics (TOG),* **31**(6), pp. 1–10.

[96] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J., 2015. "3d shapenets: A deep representation for volumetric shapes". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1912–1920.

[97] Arjovsky, M., Chintala, S., and Bottou, L., 2017. "Wasserstein generative adversarial networks". In International conference on machine learning, PMLR, pp. 214–223.

[98] Li, B., Yu, Y., and Li, Y., 2020. "Lbwgan: Label based shape synthesis from text with wgans". In 2020 International Conference on Virtual Reality and Visualization (ICVRV), IEEE, pp. 47–52.

[99] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A., 2019. "Occupancy networks: Learning 3d reconstruction in function space". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4460–4470.

[100] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., 2017. "Attention is all you need". *Advances in neural information processing systems,* **30**.

[101] Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z., 2018. "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly". *IEEE transactions on pattern analysis and machine intelligence,* **41**(9), pp. 2251–2265.

[102] Dinh, L., Sohl-Dickstein, J., and Bengio, S., 2016. "Density estimation using real nvp". *arXiv preprint arXiv:1605.08803*.

[103] Jain, A., Mildenhall, B., Barron, J. T., Abbeel, P., and Poole, B., 2022. "Zero-shot text-guided object generation with dream fields". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 867–876.

[104] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R., 2020. "Nerf: Representing scenes as neural radiance fields for view synthesis". In European conference on computer vision, Springer, pp. 405–421.

[105] Frolov, S., Hinz, T., Raue, F., Hees, J., and Dengel, A., 2021. "Adversarial text-to-image synthesis: A review". *Neural Networks,* **144**, pp. 187–209.

[106] Wang, Y., Chang, L., Cheng, Y., Jin, L., Cheng, Z., Deng, X., and Duan, F., 2018. "Text2sketch: Learning face sketch from facial attribute text". In 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, pp. 669–673.

[107] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S., 2011. "The caltech-ucsd birds-200-2011 dataset".

[108] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al., 2017. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". *International journal of computer vision,* **123**(1), pp. 32–73.

[109] Jongejan, J., Rowley, H., Kawashima, T., Kim, J., and Fox-Gieg, N., 2016. "The quick, draw!-ai experiment". *Mount View, CA, accessed Feb,* **17**(2018), p. 4.

[110] Olsen, L., Samavati, F. F., Sousa, M. C., and Jorge, J. A., 2009. "Sketch-based modeling: A survey". *Computers & Graphics,* **33**(1), pp. 85–103.

[111] Nishida, G., Garcia-Dorado, I., Aliaga, D. G., Benes, B., and Bousseau, A., 2016. "Interactive sketching of urban procedural models". *ACM Transactions on Graphics (TOG),* **35**(4), pp. 1–11.

[112] He, Y., Xie, H., Zhang, C., Yang, X., and Miyata, K., 2021. "Sketch-based normal map generation with geometric sampling". In International Workshop on Advanced Imaging Technology (IWAIT) 2021, Vol. 11766, SPIE, pp. 261–266.

[113] Su, W., Du, D., Yang, X., Zhou, S., and Fu, H., 2018. "Interactive sketch-based normal map generation with deep neural networks". *Proceedings of the ACM on Computer Graphics and Interactive Techniques,* **1**(1), pp. 1–17.

[114] Aha, D. W., 2013. *Lazy learning*. Springer Science & Business Media.

[115] Delanoy, J., Coeurjolly, D., Lachaud, J.-O., and

Bousseau, A., 2019. "Combining voxel and normal predictions for multi-view 3d sketching". *Computers & Graphics,* **82**, pp. 65–72.

[116] Yang, K., Lu, J., Hu, S., and Chen, X., 2021. "Deep 3d modeling of human bodies from freehand sketching". In International Conference on Multimedia Modeling, Springer, pp. 36–48.

[117] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J., 2019. "Expressive body capture: 3d hands, face, and body from a single image". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10975–10985.

[118] Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K., 2013. "Facewarehouse: A 3d facial expression database for visual computing". *IEEE Transactions on Visualization and Computer Graphics,* **20**(3), pp. 413–425.

[119] Wang, F., Yang, Y., Zhao, B., Jiang, D., Chen, S., and Sheng, J., 2021. "Reconstructing 3d model from single-view sketch with deep neural network". *Wireless Communications and Mobile Computing,* **2021**.

[120] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S., 2019. "Deepsdf: Learning continuous signed distance functions for shape representation". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174.

[121] Zhang, S.-H., Guo, Y.-C., and Gu, Q.-W., 2021. "Sketch2model: View-aware 3d modeling from single free-hand sketches". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6012–6021.

[122] Wang, L., Qian, C., Wang, J., and Fang, Y., 2018. "Unsupervised learning of 3d model reconstruction from hand-drawn sketches". In Proceedings of the 26th ACM international conference on Multimedia, pp. 1820–1828.

[123] Smirnov, D., Bessmeltsev, M., and Solomon, J., 2019. "Deep sketch-based modeling of man-made shapes".

[124] Gao, L., Yang, J., Wu, T., Yuan, Y.-J., Fu, H., Lai, Y.-K., and Zhang, H., 2019. "Sdm-net: Deep generative network for structured deformable mesh". *ACM Transactions on Graphics (TOG),* **38**(6), pp. 1–15.

[125] Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N. J., and Guibas, L. J., 2019. "Structurenet: hierarchical graph networks for 3d shape generation". *ACM Transactions on Graphics (TOG),* **38**(6), pp. 1–19.

[126] Chen, W., and Fuge, M., 2019. "Synthesizing designs with interpart dependencies using hierarchical generative adversarial networks". *Journal of Mechanical Design,* **141**(11), p. 111403.

[127] Qi, C. R., Su, H., Mo, K., and Guibas, L. J., 2017. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 652–660.

[128] Yang, M. C., 2003. "Concept generation and sketching: Correlations with design outcome". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 37017, pp. 829–834.

[129] Wu, R., Xiao, C., and Zheng, C., 2021. "Deepcad: A deep generative network for computer-aided design models". In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6772–6782.

[130] Para, W., Bhat, S., Guerrero, P., Kelly, T., Mitra, N., Guibas, L. J., and Wonka, P., 2021. "Sketchgen: Generating constrained cad sketches". *Advances in Neural Information Processing Systems,* **34**.

[131] Ganin, Y., Bartunov, S., Li, Y., Keller, E., and Saliceti, S., 2021. "Computer-aided design as language". *Advances in Neural Information Processing Systems,* **34**.

[132] Willis, K. D., Jayaraman, P. K., Lambourne, J. G., Chu, H., and Pu, Y., 2021. "Engineering sketch generation for computer-aided design". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2105–2114.

[133] Jayaraman, P. K., Sanghi, A., Lambourne, J. G., Willis, K. D., Davies, T., Shayani, H., and Morris, N., 2021. "Uv-net: Learning from boundary representations". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11703–11712.

[134] Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., and Panozzo, D., 2019. "Abc: A big cad model dataset for geometric deep learning". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9601–9611.

[135] Seff, A., Ovadia, Y., Zhou, W., and Adams, R. P., 2020. "Sketchgraphs: A large-scale dataset for modeling relational geometry in computer-aided design". *arXiv preprint arXiv:2007.08506*.

[136] Gryaditskaya, Y., Sypesteyn, M., Hoftijzer, J. W., Pont, S. C., Durand, F., and Bousseau, A., 2019. "Opensketch: a richly-annotated dataset of product design sketches.". *ACM Trans. Graph.,* **38**(6), pp. 232–1.

[137] Regenwetter, L., Curry, B., and Ahmed, F., 2021. "Biked: A dataset and machine learning benchmarks for data-driven bicycle design". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 85383, American Society of Mechanical Engineers, p. V03AT03A019.

[138] Fuge, M. The frontiers in design representation (finder) summer school. `https://ideal.umd.edu/FinDeR/`. Accessed: 2022-10-01.

[139] Li, X., Demirel, H. O., Goldstein, M. H., and Sha, Z., 2021. "Exploring generative design thinking for engineering design and design education". In 2021 ASEE Midwest Section Conference.

[140] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.,

2014. “Microsoft coco: Common objects in context”. In European conference on computer vision, Springer, pp. 740–755.

[141] Chen, Z., and Zhang, H., 2019. “Learning implicit fields for generative shape modeling”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5939–5948.

[142] Kim, J.-H., Kitaev, N., Chen, X., Rohrbach, M., Zhang, B.-T., Tian, Y., Batra, D., and Parikh, D., 2017. “Codraw: Collaborative drawing as a testbed for grounded goal-driven communication”. *arXiv preprint arXiv:1712.05558*.

[143] Zhang, W., Wang, X., and Tang, X., 2011. “Coupled information-theoretic encoding for face photo-sketch recognition”. In CVPR 2011, IEEE, pp. 513–520.

[144] Li, J., Xu, K., Chaudhuri, S., Yumer, E., Zhang, H., and Guibas, L., 2017. “Grass: Generative recursive autoencoders for shape structures”. *ACM Transactions on Graphics (TOG),* **36**(4), pp. 1–14.

[145] Feng, Y., Zhang, Z., Zhao, X., Ji, R., and Gao, Y., 2018. “Gvcnn: Group-view convolutional neural networks for 3d shape recognition”. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 264–272.

[146] Kanezaki, A., Matsushita, Y., and Nishida, Y., 2019. “Rotationnet for joint object categorization and unsupervised pose estimation from multi-view images”. *IEEE transactions on pattern analysis and machine intelligence,* **43**(1), pp. 269–283.

[147] Shajahan, D. A., Nayel, V., and Muthuganapathy, R., 2019. “Roof classification from 3-d lidar point clouds using multiview cnn with self-attention”. *IEEE Geoscience and Remote Sensing Letters,* **17**(8), pp. 1465–1469.

[148] Qi, A., Song, Y.-Z., and Xiang, T., 2018. “Semantic embedding for sketch-based 3d shape retrieval.”. In BMVC, Vol. 3, pp. 11–12.

[149] Darom, T., and Keller, Y., 2012. “Scale-invariant features for 3-d mesh models”. *IEEE Transactions on Image Processing,* **21**(5), pp. 2758–2769.

[150] Umetani, N., 2017. “Exploring generative 3d shapes using autoencoder networks”. In *SIGGRAPH Asia 2017 technical briefs*. pp. 1–4.

[151] 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”. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 909–918.

[152] Remelli, E., Lukoianov, A., Richter, S., Guillard, B., Bagautdinov, T., Baque, P., and Fua, P., 2020. “Meshsdf: Differentiable iso-surface extraction”. *Advances in Neural Information Processing Systems,* **33**, pp. 22468–22478.

[153] Kar, A., Häne, C., and Malik, J., 2017. “Learning a multi-view stereo machine”. *Advances in neural information processing systems,* **30**.

[154] Sangkloy, P., Burnell, N., Ham, C., and Hays, J., 2016. “The sketchy database: learning to retrieve badly drawn bunnies”. *ACM Transactions on Graphics (TOG),* **35**(4), pp. 1–12.

[155] Eitz, M., Hays, J., and Alexa, M., 2012. “How do humans sketch objects?”. *ACM Transactions on graphics (TOG),* **31**(4), pp. 1–10.

[156] Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J., 2019. “Amass: Archive of motion capture as surface shapes”. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 5442–5451.

[157] Chen, X., Golovinskiy, A., and Funkhouser, T., 2009. “A benchmark for 3d mesh segmentation”. *Acm transactions on graphics (tog),* **28**(3), pp. 1–12.

[158] Park, K., Rematas, K., Farhadi, A., and Seitz, S. M., 2018. “Photoshape: Photorealistic materials for large-scale shape collections”. *arXiv preprint arXiv:1809.09761*.

[159] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V., 2017. “Carla: An open urban driving simulator”. In Conference on robot learning, PMLR, pp. 1–16.

[160] Zhou, Q., and Jacobson, A., 2016. “Thingi10k: A dataset of 10,000 3d-printing models”. *arXiv preprint arXiv:1605.04797*.

Appendix A: Details of Literature Search

As introduced in Section 3.2, Table 3 shows the number of articles found in major literature databases. In addition, we used the time range of January 2021 to June 2022 to search for the most recent studies for sketch-to-3D shape retrieval and generation, the number of which is indicated in parentheses (e.g., (35) for ShRecSk).

Figure 11 shows the articles that are most relevant to the two key articles ([35,48]) using Connected Papers (accessed in June 2022). Studies that meet the scope of our review are indicated using a quadrilateral in each sub-figure.

Table 3. Studies found in major databases using keywords of "text-to-sketch retrieval" (TSkRet), "text-to-sketch generation" (TSkG), "text-to-shape retrieval" (TShRet), "text-to-shape generation" (TShG), "sketch-based 3D shape retrieval" (SkShRet), "sketch-based 3D shape generation" (SkShG), "sketch-based 3D shape reconstruction" (SkShRec), "sketch-based 3D shape synthesis" (SkShSyn), and "3D shape reconstruction from sketches" (ShRecSk)

		Keywords (double quotation marks included)							
		TSkRet	TSkG	TShRet	TShG	SkShRet	SkShG	SkShRec	SkShSyn	ShRecSk
Database	ScienceDirect	0	0	0	0	2	0	0	0	0
	Web of Science	0	0	1	0	20	1	0	0	1
	Scopus	0	0	1	1	454	5	1	0	95
	IEEExplore	0	0	0	1	13	1	1	0	1
	ACM Digital Libraries	0	0	1	0	14	0	0	0	3
	Google Scholar	0	3	7	22	559 (96)	7 (5)	5 (2)	1 (0)	120 (35)
	Total	0	3	10	24	1062	14	7	1	220

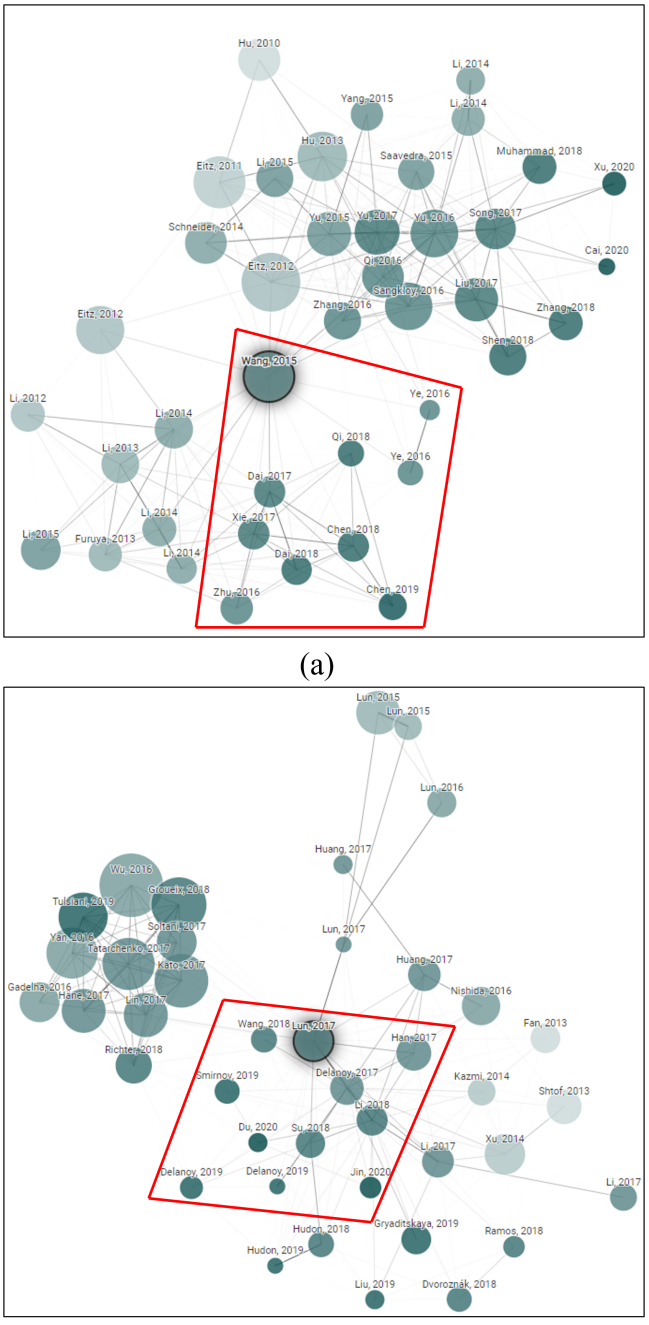

Fig. 11. (a) Studies for sketch-to-3D retrieval that are similar to [48]; (b) Studies for sketch-to-3D generation that are similar to [35]

Appendix B: Paper Summary

We summarize and tabulate all 50 articles reviewed in Table 3. There are 11 source journals and 20 conference proceedings, and their acronyms are shown below.

Nomenclature

CG	Computers & Graphics
MS	Multimedia Systems
VC	The Visual Computer
CGF	Computer Graphics Forum
AEI	Advanced Engineering Informatics
TIP	IEEE Transactions on Image Processing
MTA	Multimedia Tools and Applications
TOG	ACM Transactions on Graphics
JMD	Journal of Mechanical Design
WCMC	Wireless Communications and Mobile Computing
PACMCGIT	The Proceedings of the ACM in Computer Graphics and Interactive Techniques
MM	International Conference on Multimedia
IUI	International Conference on Intelligent User Interfaces
CHI	Conference on Human Factors in Computing Systems
I3D	Symposium on Interactive 3D Graphics and Games
MMM	International Conference on Multimedia Modeling
IMX	ACM International Conference on Interactive Media Experiences
CVPR	Computer Vision and Pattern Recognition Conference
ICCV	International Conference on Computer Vision
ECCV	European Conference on Computer Vision
ACCV	Asian Conference on Computer Vision
AAAI	Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence
ICIP	IEEE International Conference on Image Processing
UIST	Annual ACM Symposium on User Interface Software and Technology
ICCS	International Conference on Computational Science
ICPR	International Conference on Pattern Recognition
ICLR	International Conference on Learning Representations
ICVRV	International Conference on Virtual Reality and Visualization
3DIMPVT	International Conference on 3D Imaging Modeling, Processing, Visualization and Transmission
VISIGRAPP	International Joint Conference on Computer Vision Imaging and Computer Graphics Theory and Applications
ICCEIA-VR	International Conference on Computer Engineering and Innovative Application of VR

Table 4. Summary of Literature

Type of DLCMT	Reference	Year	Method	Text Type	Sketch Type	3D Representation	Dataset	Object Class	Generalizability beyond trained classes	User Interface	User Study	Publication Source
Text to 3D shape retrieval	Han et al. [67]	2019	CNN, GRU	NLD	N/A	Voxel	3D-text dataset [17]	Chairs, tables	No	No	No	Conference: AAAI
	Chen et al. [17]	2018	Text encoder (CNN, GRU), shape encoder (3DCNN)	NLD	N/A	Voxel	Propose 3D-text dataset from ShapeNet [49]	Chairs, tables, synthetic objects	No	No	No	Conference: ACCV
Text to 3D shape generation	Jain et al. [103]	2022	Network based on CLIP [52]	NLD	N/A	NeRF [104]	COCO [140]	Diverse classes	Yes	No	No	Conference: CVPR
	Sanghi et al. [43]	2022	Network based on PointNet [127], CLIP [52], OccNet [99]	Object names	N/A	Voxel	ShapeNet [49]	Diverse classes	No	No	No	Conference: CVPR
	Liu et al. [50]	2022	Shape autoencoder, word-level spatial transformer, shape generator (IMLE [141])	NLD	N/A	Implicit representation, mesh	3D-text dataset [17]	Chairs, tables	No	No	No	Conference: CVPR
	Jahan et al. [94]	2021	Shape encoder, decoder label regression network,	Semantic keywords	N/A	Implicit representation, mesh	COSEG [95], ModelNet [96]	Chairs, tables, lamps	No	No	No	Journal: CGF
	Li et al. [98]	2020	GAN based network	NLD	N/A	Voxel	3D-text dataset [17]	Chairs, tables	No	No	No	Conference: ICVRV
	Chen et al. [17]	2018	Text encoder (CNN, GRU), shape encoder (3DCNN), GAN	NLD	N/A	Voxel	Propose 3D-text dataset from ShapeNet [49]	Chairs, tables, synthetic objects	No	No	No	Conference: ACCV
Text to sketch generation	Yuan et al. [63]	2021	GAN, Bi-LSTM	NLD	Static pixel space	N/A	Propose SketchCUB from CUB [107]	Birds	No	No	Yes	Conference: CVPR
	Huang et al. [54]	2020	Composition proposer (transformer), object generator (Sketch-RNN [16])	NLD	Dynamic stroke coordinate space	N/A	CoDraw [142]	Diverse classes	Yes	Yes	Yes	Conference: IUI
	Huang et al. [53]	2019	Scene composer (transformer), object sketcher (Sketch-RNN [16])	NLD	Dynamic stroke coordinate space	N/A	Visual Genome [108], Quick, Draw! [109]	Diverse classes	Yes	Yes	Yes	Conference: UIST
	Wang et al. [106]	2018	GAN based network	NLD	Static pixel space	N/A	Propose Text2Sketch from CUFSF [143]	Human faces	No	No	No	Conference: ICIP

Type of DLCMT	Reference	Year	Method	Text Type	Sketch Type	3D Representation	Dataset	Object Class	Generalizability beyond trained classes	User Interface	User Study	Publication Source
Sketch to 3D shape retrieval	Qin et al. [32]	2022	Autoencoder (GRASS [144]), k-nearest neighbors	N/A	Static pixel space	B-Rep	Propose CAD model-sketches dataset	Diverse classes	Yes	Yes	No	Journal: AEI
	Yang et al. [76]	2022	3D model network, 2D sketch network (MVCNN [85])	N/A	Static pixel space	Mesh	SHREC13 [83], SHREC14 [84], SHREC16 [92]	Diverse classes	Yes	No	No	Journal: MS
	Qi et al. [34]	2021	Sketch encoder, shape encoder (MVCNN [85])	N/A	Static pixel space	Mesh	Propose fine-grained dataset from ShapeNet [49]	Chairs, lamps	No	No	No	Journal: TIP
	Manda et al. [87]	2021	MVCNN [85], GVCNN [145], RotationNet [146], MVCNN-SA [147]	N/A	Static pixel space	B-Rep	Propose CADSketchNet from ESB [88], MCB [89]	Diverse classes	Yes	No	No	Journal: CG
	Liang et al. [81]	2021	Sketch network, view network	N/A	Static pixel space	Mesh	SHREC13, SHREC14	Diverse classes	Yes	No	No	Journal: TIP
	Liu et al. [82]	2021	MVCNN [85], Guidance Cleaning Network	N/A	Static pixel space	Mesh	SHREC13, SHREC14	Diverse classes	Yes	No	No	Conference: ICCEIA-VR
	Xia et al. [75]	2021	Student network, teacher network (MVCNN [85])	N/A	Static pixel space	Mesh	SHREC13	Diverse classes	Yes	No	No	Conference: ICCS
	Li et al. [55]	2021	CNN based network	N/A	Type II 3D sketch	Mesh	SHREC16STB [90]	Diverse classes	Yes	Yes	No	Journal: MTA
	Navarro et al. [86]	2021	CNN based network	N/A	Static pixel space	Mesh	Propose a line drawing dataset from ShapeNet [49]	Diverse classes	Yes	No	No	Journal: CGF
	Chen et al. [79]	2019	Sketch network, segmented stochastic-viewing shape network, view attention network	N/A	Static pixel space	Mesh	SHREC13, SHREC14, PART-SHREC14 [148]	Diverse classes	Yes	No	No	Conference: CVPR
	Dai et al. [72]	2018	Source domain network, target domain network (3D-SIFT [149]	N/A	Static pixel space	Mesh	SHREC13, SHREC14, SHREC16	Diverse classes	Yes	No	No	Journal: TIP
	Chen et al. [74]	2018	MVCNN [85], GAN, metric network, cross-modality transformation network	N/A	Static pixel space	Mesh	SHREC13, SHREC14	Diverse classes	Yes	No	No	Conference: ECCV

Type of DLCMT	Reference	Year	Method	Text Type	Sketch Type	3D Representation	Dataset	Object Class	Generalizability beyond trained classes	User Interface	User Study	Publication Source
Sketch to 3D shape retrieval	Dai et al. [73]	2017	Source domain network, target domain network (3D-SIFT [149])	N/A	Static pixel space	Mesh	SHREC13 [83], SHREC14 [84]	Diverse classes	Yes	No	No	Conference: AAAI
	Xie et al. [78]	2017	CNN, metric network	N/A	Static pixel space	Mesh	SHREC13, SHREC14	Diverse classes	Yes	No	No	Conference: CVPR
	Zhu et al. [71]	2016	Cross-domain neural network, pyramid cross-domain network	N/A	Static pixel space	Mesh	SHREC14	Diverse classes	Yes	No	No	Conference: AAAI
	Ye et al. [90]	2016	CNN based network	N/A	Type II 3D sketch	Mesh	Propose SHREC16STB	Diverse classes	Yes	No	No	Conference: ICPR
	Wang et al. [48]	2015	CNN, Siamese network	N/A	Static pixel space	Mesh	PSB [68], SHREC13, SHREC14	Diverse classes	Yes	No	No	Conference: CVPR
Sketch and text to 3D shape retrieval	Stemasov et al. [62]	2022	Flask representation state transfer, HoloLens	NLD	Type II 3D sketch	Mesh, voxel	Thingiverse, MyMiniFactory	Diverse classes	Yes	Yes	No	Conference: CHI
	Giunchi et al. [44]	2021	CNN based network	NLD	Type II 3D sketch	Mesh	Propose Variational Chairs dataset from ShapeNet [49]	Chairs	No	Yes	Yes	Conference: IMX
Sketch to 3D shape generation	Li et al. [33]	2022	Target-embedding variational autoencoder	N/A	Static pixel space	Mesh	Dataset [150]	Cars, cups	No	No	No	Journal: JMD
	Nozawa et al. [20]	2022	GAN, lazy learning	N/A	Static pixel space	Point cloud, mesh	ShapeNet [49]	Cars	No	No	No	Journal: VC
	Du et al. [59]	2021	CNN, OccNet [99], 3DCNN	N/A	Static pixel space	Implicit representation, mesh	PartNet [151]	Chairs, tables, lamps	No	Yes	Yes	Journal: CGF
	Wang et al. [119]	2021	Sketch component segmentation network, transformation network, VAE	N/A	Static pixel space	Point cloud, mesh	Dataset [35]	Characters, airplanes, chairs	No	No	No	Journal: WCMC
	Guillard et al. [6]	2021	Encoder (MeshSDF [152]), decoder, differential renderer	N/A	Static pixel space	Implicit representation, mesh	ShapeNet [49]	Cars, chairs	No	Yes	No	Conference: ICCV

Type of DLCMT	Reference	Year	Method	Text Type	Sketch Type	3D Representation	Dataset	Object Class	Generalizability beyond trained classes	User Interface	User Study	Publication Source
Sketch to 3D shape generation	Zhang et al. [121]	2021	View-aware generation network (encoder, decoder), discriminator	N/A	Static pixel space	Mesh	ShapeNet-Sketch [153], Sketchy [154], TuBerlin [155]	Diverse classes	Yes	No	No	Conference: CVPR
	Yang et al. [116]	2021	CNN based network	N/A	Static pixel space	Mesh	AMASS [156]	Human bodies	No	No	No	Conference: MMM
	Luo et al. [60]	2021	Voxel-aligned implicit network, pixel-aligned implicit network	N/A	Static pixel space	Implicit representation, mesh	Propose 3DAnimalHead	Animal heads	No	Yes	Yes	Conference: UIST
	Jin et al. [51]	2020	VAE, volumetric autoencoder	N/A	Static pixel space	Voxel, mesh	PSB [68], benchmark [157]	Diverse classes	Yes	No	No	Conference: I3D
	Smirnov et al. [5]	2020	CNN based network	N/A	Static pixel space	B-Rep, mesh	ShapeNet [49]	Diverse classes	No	No	No	Conference: ICLR
	Nozawa et al. [19]	2020	Encoder-decoder, lazy learning	N/A	Static pixel space	Point cloud, mesh	ShapeNet	Cars	No	No	No	Conference: VISIGRAPP
	Smirnov et al. [123]	2019	CNN based network	N/A	Static pixel space	B-Rep, mesh	ShapeNet	Diverse classes	No	No	No	Conference: ICLR
	Delanoy et al. [115]	2019	CNN based network	N/A	Type I 3D Sketch	Voxel	COSEG [95]	Chairs, vases, synthetic shapes	No	No	No	Journal: CG
	Wang et al. [122]	2018	Autoencoder, GAN	N/A	Static pixel space	Voxel	SHREC13 [83], ShapeNet	Chairs	No	No	No	Conference: MM
	Li et al. [56]	2018	DFNet (encoder-decoder), GeomNet (encoder-decoder)	N/A	Static pixel space	Mesh	Dataset [35]	Characters	No	Yes	Yes	Journal: TOG
	Delanoy et al. [57]	2018	Singleview CNN, updater CNN	N/A	Type I 3D Sketch	Voxel	COSEG [95]	Chairs, vases, synthetic shapes	No	Yes	Yes	Journal: PACMCGIT
	Lun et al. [35]	2017	Encoder, multiview decoder	N/A	Static pixel space	Point cloud, mesh	The Models Resource, ShapeNet	Characters, airplanes, chairs	No	No	Yes	Conference: 3DIMPVT
	Han et al. [58]	2017	Deep regression network	N/A	Static pixel space	Mesh	Faceware-house [118]	Face caricatures	No	Yes	Yes	Journal: TOG

Type of DLCMT	Reference	Year	Method	Text Type	Sketch Type	3D Representation	Dataset	Object Class	Generalizability beyond trained classes	User Interface	User Study	Publication Source
Text to 3D shape manipulation	Liu et al. [50]	2022	Shape autoencoder, word-level spatial transformer, shape generator (IMLE [141])	NLD	N/A	Implicit representation, mesh	3D-text dataset [17]	Chairs, tables	No	No	No	Conference: CVPR
	Wang et al. [61]	2022	Disentangled conditional NeRF, CLIP [52], GAN	Semantic keywords, object names	N/A	NeRF [104]	Photoshapes [158], Carla [159]	Chairs, cars	No	Yes	Yes	Conference: CVPR
	Michel et al. [36]	2022	Neural style filed network, differentiable renderer, CLIP [52]	Semantic keywords, object names	N/A	Mesh	COSEG [95], Thingi10K [160], ShapeNet [49], Turbo Squid, ModelNet [96]	Diverse classes	Yes	No	Yes	Conference: CVPR
Sketch to 3D shape manipulation	Guillard et al. [6]	2021	Encoder (MeshSDF [152]), decoder, differential renderer	N/A	Static pixel space	Implicit representation, mesh	ShapeNet	Cars, chairs	No	Yes	No	Conference: ICCV
	Jin et al. [51]	2020	VAE, volumetric autoencoder	N/A	Static pixel space	Voxel, mesh	PSB [68], Benchmark [157]	Diverse classes	Yes	No	No	Conference: I3D

Figure Caption List

Table Caption List