Multi-modal Machine Learning in Engineering Design: A Review and Future Directions

Multi-modal machine learning (MMML), which involves integrating multiple modalities of data and their corresponding processing methods, has demonstrated promising results in various practical applications, such as text-to-image translation. This review paper summarizes the recent progress and challenges in using MMML for engineering design tasks. First, we introduce the different data modalities commonly used as design representations and involved in MMML, including text, 2D pixel data (e.g., images and sketches), and 3D shape data (e.g., voxels, point clouds, and meshes). We then provide an overview of the various approaches and techniques used for representing, fusing, aligning, synthesizing, and co-learning multi-modal data as five fundamental concepts of MMML. Next, we review the state-of-the-art capabilities of MMML that potentially apply to engineering design tasks, including design knowledge retrieval, design evaluation, and design synthesis. We also highlight the potential benefits and limitations of using MMML in these contexts. Finally, we discuss the challenges and future directions in using MMML for engineering design, such as the need for large labeled multi-modal design datasets, robust and scalable algorithms, integrating domain knowledge, and handling data heterogeneity and noise. Overall, this review paper provides a comprehensive overview of the current state and prospects of MMML for engineering design applications.

1 INTRODUCTION

We perceive and interact with the world around us in multiple ways - by seeing, hearing, feeling, smelling, and so forth. Information is often delivered and communicated to the senses of the interpreter using a certain *medium* or multiple *media*, such as images and sound. A *modality* refers to the *medium* through which an object exists or is experienced. Common data modalities include [1]:

(1) Natural language (both spoken and written);
(2) Visual (images, videos, sketches, renderings, 3D geometries);
(3) Auditory (e.g., voice, sounds, and music);
(4) Haptics or Touch;
(5) Smell, taste, and self-motion;
(6) Physiological signals, such as electrocardiogram, and skin conductance;
(7) Other modalities, such as infrared images, depth images, and functional magnetic resonance images.

In machine learning (ML), a modality refers to a certain type of information or the representation format in which information is stored and input to ML models. A model is multi-modal when it works with multiple such modalities. Since we often rely on multiple media to fulfill tasks in the real world (e.g., sound and visuals to understand a video), multi-modal machine learning (MMML) is needed to develop ML models that emulate humans for these tasks.

The concept of “multi-modalities” was first explored in behavioral and sensory research by David McNeill [2]. It has since been applied to computational systems for audio-visual speech recognition and multimedia information retrieval. Before the rise of deep learning (DL), MMML relied on traditional ML models such as hidden Markov models [3], shallow artificial neural networks, dynamic Bayesian networks [4], and discriminative sequential models [5]. The representation-based DL era of MMML began around 2011, thanks to the increasing availability of multi-modal data such as VQA [6] and CLEVR [7], the increased power and affordability of graphics processing units, the ability of DL to learn high-level visual features, and the vectorization of semantic features. The strides in MMML first revitalized the area of media description like image or video captioning [8], visual question answering (VQA) [9], and cross-modal information retrieval (IR) [10]. In recent years, MMML has gained significant attention. Particularly, the advent of deep generative models (DGMs) like generative adversarial networks (GANs) [11], diffusion models (DMs) [12], and their variations [13, 14] has led to the rapid development of various cross-modal DGMs, such as DALL-E [15] and Stable Diffusion [16].

1.1 Definition of Multi-Modal Machine Learning

Most traditional ML models have been trained using a single type of input data or unimodal data, which we call unimodal machine learning in this paper. For example, a model might be trained on text data, audio data, or image data but not a combination of these types. The core of unimodal ML is to learn the latent unimodal representations, based on which we can conduct various downstream ML tasks, such as classification, regression, generation, and clustering. In comparison, MMML involves training a model on multiple types of input data or multimodal data. It can be more challenging, as it requires understanding and integrating multiple different forms of data. However, it can also be more powerful, as it can exploit the complementarity, alignment, and redundancy of multi-modal data, leading to a more complete understanding of an instance and enabling cross-modal problem solving.

In this paper, we refer to an *instance* as a data point from a multi-modal dataset that is represented by multiple modalities and a *feature* as a vectorized latent representation of input data at different learning stages that reflects the characteristics of the data in the real world. Under this setting, complementarity means the information components from different modalities complement each other for describing an instance. Alignment refers to the correspondence between the information components of an instance from different modalities. Redundancy signifies that the information components from different modalities have the same meaning, which enhances the robustness of the models learning them.

Currently, most MMML efforts focus on handling two data modes, also known as bimodal machine learning (BML). However, it is expected that MMML will evolve beyond BML to involve more data modes, potentially up to six, in the near future. Additionally, as the internet continues to connect more and more cyber-physical systems, MMML may advance to what we call "many-modal learning", where models can use all kinds of data collected by various sensors. This type of many-modal learning will present new challenges, such as the curse of dimensionality and ambiguity among the different data modes. The evolution of MMML may resonate with the multi-objective and many-objective optimization literature, where algorithms developed for multi-objective problems do not work well for many-objective problems due to the large dimensionality.

Since MMML enables ML models to "see" beyond single data modes, it enhances model performance. It has the potential to transform ML research and practical applications across a variety of domains. For example, in the healthcare industry, MMML could improve the accuracy of diagnosis and treatment recommendations by incorporating data from various sources, such as medical records, images, and patient-reported symptoms. In finance, MMML could detect fraudulent activities by combining data from transactions, account activity, and social media posts. In the retail industry, MMML can potentially personalize customer recommendations by combining data from their browsing and purchase history, as well as data from social media and other sources. The impact of MMML extends beyond specific industries. It has the potential to drive innovation and increase efficiency across a range of sectors, leading to increased productivity and competitiveness of the US economy. However, the adoption of MMML in engineering design applications has been slow. In this paper, we review existing work in MMML, discuss different methods and applications, and highlight the key challenges for its applications in engineering design.

1.2 Multi-modality in Design Representation

In recent years, design researchers have made significant progress in exploring the state-of-the-art DL models for design synthesis [17, 18, 14, 19], evaluation [20], and optimization [13]. The representation of the designs learned by the ML models affects the model effectiveness [21]. A representation framework with sufficient power to capture the nature of designs for formal and functional reasoning is necessary during the design process [22]. Distinct representation modes require different design resources, such as skill and time, and result in different levels of representation resolution and fidelity, ranging from rough sketches to realistic renderings [23, 24, 25], and detailed 3D models. The selection of design representation mode impacts human perception and creative design performance, and design fixation [24, 25, 26, 27, 28, 29].

Traditionally, hand-drawing hand-drawing sketches are commonly used to express and communicate concepts at the early design stages. This approach permits efficient exploration at different abstraction levels and preserves ambiguity in this process, which is beneficial for innovation and creativity [26, 30]. Likewise, line drawings created with assistive instruments or by computer can also serve as design representations [31]. The digital revolution and advances in computer-aided design (CAD) software provide us with more options for design representation. Two-dimensional (2D) and three-dimensional (3D) representations are typically used to generate interactive and potentially complex solutions during the detailed design stage [30]. Researchers found that the adoption of CAD in concept generation facilitates the exploration in depth, rather than in breadth [27]. Through computer graphic applications (e.g., Autodesk Maya), 3D representations can be converted to simplified or realistic renderings or view silhouettes from given perspectives to represent a design concept [32].

Physical prototypes provide another representation approach. Design details can be more clearly and accurately interpreted and understood in terms of 3D spatial relationships when represented by such prototypes [30, 33]. Clay modeling is commonly used for prototyping in automobile styling design, which not only brings a design and its spirit to life but also enables intuitive design, fine-tuning, and improvement. When physical prototypes are unavailable, designers use their photographs to communicate the corresponding designs instead [31]. In addition, textual descriptions are also used separately or along with visual representations to convey design concepts [34]. In other cases, design concepts can be represented by a set of continuously valued attributes [35], such as display length and pixel density, or functions [36]. To promote information exchange during conceptual design, researchers have also explored theoretical design representation methods, such as function model [37], functional decomposition and morphology [38], problem solution network [39], and morphological charts [39]. Recently, researchers have also explored virtual reality, augmented reality, augmented virtuality, and mixed reality to represent designs [40].

In engineering design, multiple data modes are often used together to represent a design, such as sketches with textual descriptions and prototypes with tabular, verbal, or textual descriptions. Design representation evolves across different design stages, from more abstract formats to more detailed formats. Through experimental studies, researchers have suggested that adopting different representation modes at different design stages can enable a fluent exploration process [41, 42]. Despite the wide usage of multiple modalities in design practice, most ML applications supporting design research are unimodal. MMML, being able to capture the complementarity and alignment between different modalities, has great potential for a more comprehensive AI-based comprehension of designs. This may enable more accurate design evaluation to reduce the human efforts needed for manual assessment. Furthermore, MMML allows for cross-modal syntheses, such as text-to-image and image-to-shape generation, facilitating design exploration.

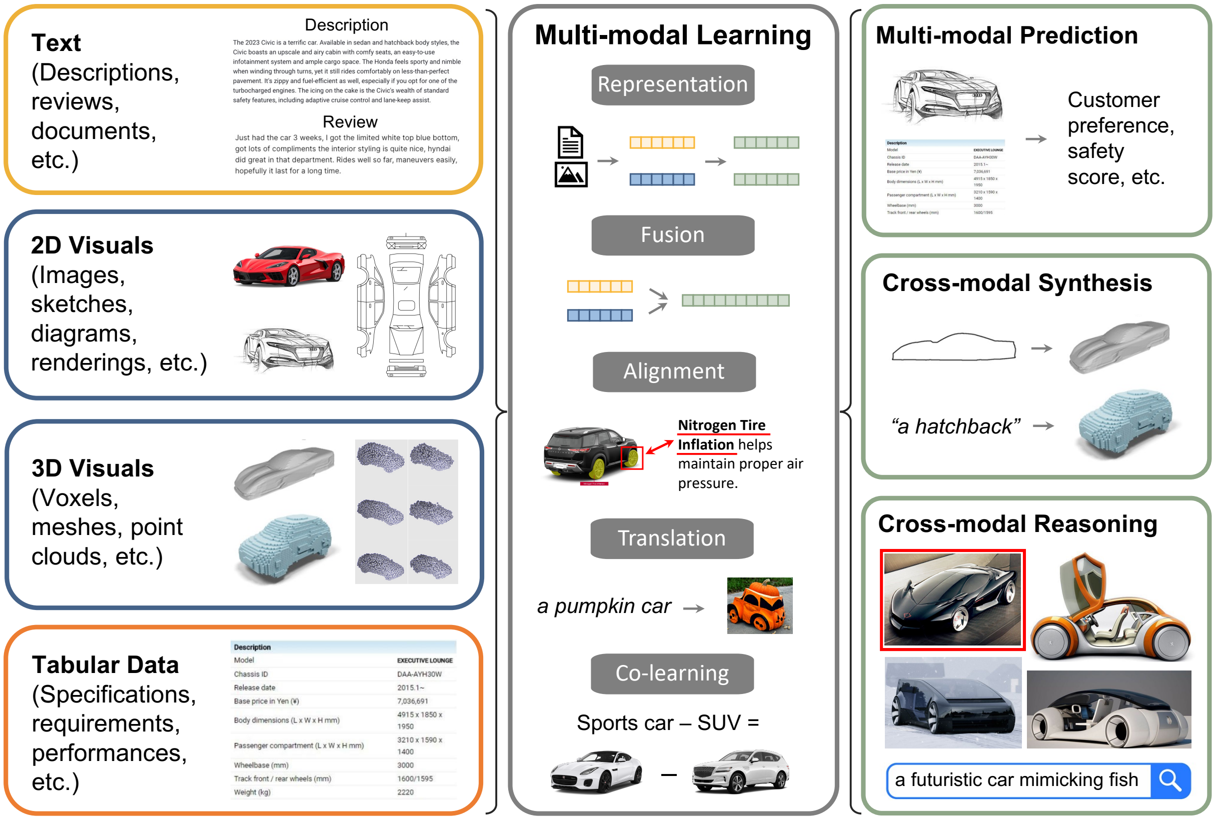
1.3 The Scope of This Paper

This paper provides a technical review of the fundamental concepts and exemplary applications of MMML. As most of the recent ML strides happen in DL, we focus our review of MMML on DL. Figure 1 illustrates the scope of this paper. Since engineering design is the target application domain, our study mainly focuses on applications involving texts, images (e.g., sketches or images), and shapes (e.g., voxels, point clouds, meshes) that have been adopted in engineering design or have the potential to be adapted to engineering design. On this basis, we also discuss the future directions and challenges of adopting MMML in engineering design. This paper differs from the prior MMML review papers from a few perspectives. Compared to the review papers published in ML [43], signal processing [44], and medical domains [45], this paper focuses more on the approaches and applications involving sketches and 3D shapes but less attention to those related to modalities less used in engineering design, such as video and audio. Moreover, the emergence of the DMs [12] and the large pre-trained multi-modal representation models (e.g., contrastive language-image pretraining (CLIP) [46]) have promoted the development of cross-modal synthesis greatly, which was not covered in the previous review papers. Compared to the review paper published in the Journal of Mechanical Design [47], this paper reviews the fundamental concepts supporting various applications, categorizes and discusses the relevant approaches, and covers a more comprehensive set of applications.

The remainder of this paper is organized as follows. Section 2 reviews the fundamental concepts of MMML and the mainstream approaches to handling them. Section 3 reviews the exemplary applications of MMML and proposes the potential for adopting or adapting them for engineering design. Section 4 discusses the challenges and opportunities that the engineering design community faces in applying MMML to engineering design. We conclude the paper by providing a summary and prospects of MMML in engineering design in Section 5.

2 FUNDAMENTAL CONCEPTS OF MULTI-MODAL MACHINE LEARNING

This section reviews the fundamental concepts of MMML, including representation, fusion, and alignment of multi-modal data, cross-modal synthesis, and co-learning. These concepts enable MMML to handle the heterogeneity and exploit the complementarity, alignment, and redundancy of multi-modal data. They are also the key ideas that differentiate MMML from unimodal ML and support various multi-modal applications. For example, multiple of these concepts may be involved in multi-modal prediction, such as representation and fu-


Text
(Descriptions, reviews, documents, etc.)
2D Visuals
(Images, sketches, diagrams, renderings, etc.)
3D Visuals
(Voxels, meshes, point clouds, etc.)
Tabular Data
(Specifications, requirements, performances, etc.)
Multi-modal Learning
Representation
Fusion
Alignment
Nitrogen Tire Inflation helps maintain proper air pressure.
Translation
a pumpkin car
Co-learning
Sports car – SUV =
Multi-modal Prediction
Customer preference, safety score, etc.
Cross-modal Synthesis
"a hatchback"
Cross-modal Reasoning
a futuristic car mimicking fish

Description

The 2023 Civic is a terrific car. Available in sedan and hatchback body styles, the Civic boasts an upscale and airy cabin with comfy seats, an easy-to-use infotainment system and ample cargo space. The Honda feels sporty and nimble when winding through turns, yet it still rides comfortably on less-than-perfect pavement. It's zippy and fuel-efficient as well, especially if you opt for one of the turbocharged engines. The icing on the cake is the Civic's wealth of standard safety features, including adaptive cruise control and lane-keep assist.

Review

Just had the car 3 weeks, I got the limited white top blue bottom, got lots of compliments the interior styling is quite nice, hyndai did great in that department. Rides well so far, maneuvers easily, hopefully it last for a long time.

Description	
Model	EXECUTIVE LOUNGE
Chassis ID	DAA-AYH30W
Release date	2015.1~
Base price in Yen (¥)	7,036,691
Body dimensions (L x W x H mm)	4915 x 1850 x 1950
Passenger compartment (L x W x H mm)	3210 x 1590 x 1400
Wheelbase (mm)	3000
Track front / rear wheels (mm)	1600/1595
Weight (kg)	2220

Figure 1. The scope of this paper is exemplified by car designs. Engineering design involves data from multiple modalities, which MMML can employ for a variety of applications

sion for capturing and integrating features from multiple data modes to exploit their complementarity, alignment for capturing the salient features, and co-learning for handling missing modalities or small datasets.

Before the review, we first describe the structures of multi-modal data briefly to help readers understand the following concepts. In MMML, multi-modal data can be parallel, non-parallel, or hybrid [43]. Parallel data comprises associated instances from multiple modalities, such as images and their captions. In contrast, non-parallel data does not require instances of multiple modalities to be associated but to share common categories, such as images from different categories and the Wikipedia pages of these categories [48]. With hybrid data, multiple modalities are not paired directly but linked indirectly through a pivot modality, i.e., each modality is partly paired with the pivot modality, such as different languages connected via English as the pivot language [49]. We begin with our discussion about multi-modal representation.

2.1 Multi-modal Representation

Information representation is the basis of any reasoning conducted by humans or computers. Information can be represented in many forms, such as physical or digital. In general, computers operate with digital representation. As for MMML, representation refers to learning vector representations of multi-modal data that can capture the complementarity, alignment, and redundancy of multiple modalities. To learn multi-modal representation, multiple unimodal representations are often first learned from the corresponding data modalities and then brought together at earlier or later learning stages.

2.1.1 *Different Representation Forms*

Multi-modal representations can be categorized into two classes: 1) joint representation and 2) coordinated representation [43]. Figure 2 exemplifies the joint representation and coordinated representation of a design instance represented by a sketch and a text description. Joint representation fuses multiple unimodal representations into a single multi-modal representation through one or multiple shared layers, enabling different modalities to complement each other (Figure 2-A). Different fusion methods will be reviewed and discussed in the following subsection. This representation is commonly used for multi-modal prediction [50, 51, 52, 53]. Coordinated representations are multiple coordinated unimodal representations learned from the associated modalities (Figure 2-B). Typically, it projects different modalities to a common subspace and maximizes the similarity [46], correlation [54, 55, 56], mutual information [57], or agreement [58] between the associated unimodal representations through the loss function. This approach facilitates capturing the correlations and the mutual information across modalities, which has been applied to cross-modal IR [10, 59, 56] and synthesis [10]. Coordinated representations can be learned at the instance level and finer levels, such as image or sentence fragments, to facilitate fine-grained reasoning [60, 61, 62, 63]. Joint representation applies to two or more modalities, while coordinated representations mostly apply between two modalities.

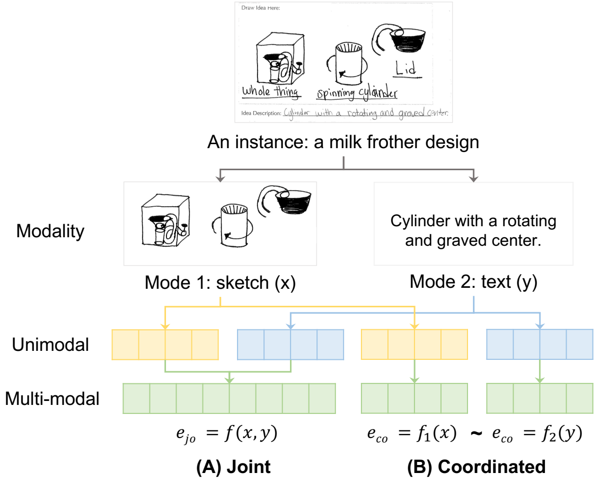

Draw Idea Here:
whole thing
spinning cylinder
Lid
Idea Description: Cylinder with a rotating and graved center.
An instance: a milk frother design
Modality
Cylinder with a rotating and graved center.
Mode 1: sketch (x)
Mode 2: text (y)
Unimodal
Multi-modal
$e_{jo} = f(x,y)$
$e_{co} = f_1(x) \sim e_{co} = f_2(y)$
(A) Joint
(B) Coordinated

Figure 2. Architectures of joint representations and coordinated representations. Joint representations are projected to the same space using all modalities as input. Coordinated representations, on the other hand, exist in their own space but are coordinated through a similarity or correlation metric.

2.1.2 Learning Process of Representations

Multi-modal representations can be learned through supervised, unsupervised, or semi-supervised learning. Since coordinated representations are learned in a task-agnostic way, different modalities can serve as the "supervision" of each other, which falls in self-supervised learning (i.e., a type of unsupervised learning) [46, 54, 55, 57, 58]. In comparison, joint representations can be obtained through either supervised or unsupervised learning. When high-quality labels are available, supervised learning captures rich intra-modal and cross-modal interactions useful for single or multiple tasks simultaneously [52, 64]. Alternatively, bimodal autoregressive transformer models can learn joint representations in a self-supervised way via pre-training tasks, such as masked content prediction [65, 66, 67, 68, 69]. Additionally, both joint representations and coordinated representations can be learned by multiple streams of DGMs (e.g., deep Boltzmann machines [50] or auto-encoders (AE) [70, 71, 56]) that are joined together or work separately in an unsupervised manner. With generative ability, this approach can fill in unexpected missing modalities [72, 50]. When labeled data is limited, but a large set of non-parallel data is available, semi-supervised models can transfer knowledge across modalities to learn better multi-modal representations [73, 74].

2.1.3 Pre-trained Bi-modal Representations from Literature

Since it is non-trivial to learn effective multi-modal representations for single tasks, researchers have released a couple of pre-trained multi-modal representations that are generalizable to various tasks. Due to the easy availability of a large amount of paired textual and visual data, most pre-trained multi-modal representations are for these two modalities. The mainstream pre-trained joint representations are learned by transformer-based autoregressive models (ARMs), such as Unicoder-AL [65], VL-BERT [66], VisualBERT [67], and B2T2 [68], LXMERT [75], ViLBERT [76], and OmniNet [77]. Benefiting from attention mechanisms, these models are good at capturing cross-modal alignment (please see the Alignment Subsection for more detail) and are commonly used for cross-modal reasoning, such as VQA [65] and image captioning [66]. In comparison, CLIP [46] is a popular coordinated representation. It is pre-trained to predict which caption goes with which image to maximize the similarities between the representations of the associated images and captions. The learned representations are transferable to various tasks and competitive with a fully supervised baseline without the need for any dataset-specific training. It has since become one of the most commonly used multi-modal representations for cross-modal synthesis tasks. Similarly, contrastive image-shape pre-training (CISP) [78] is a newly proposed coordinated representation for images and shapes. It matches images and shapes to patch embeddings using 2D and 3D convolutions and embeds the 2D and 3D patch embeddings using two transformer-based encoders, respectively. CISP has been employed for image-to-shape synthesis.

2.1.4 *Discussion on Multi-modal Representation in Engineering Design*

MMML in engineering design exploits multi-modal representations to inform the design process. While MMML can capture rich and diverse information for design tasks, it also poses unique challenges that may impact the performance and effectiveness of such models. In engineering design, design representation often needs to transform from highly abstract modes (e.g., textual descriptions), to more expressive modes (e.g., sketches), and to higher-fidelity modes (e.g., 3D models). In earlier design stages, abstract design representations are expected to avoid design fixation and help induce creativity. As the design process goes on, designers need more expressive representations to accommodate more design details reflecting the working principles or mechanisms and spatial relationships. In later design stages, high-fidelity representations are used to accurately articulate design parameters and configurations to prepare designs for downstream tasks, such as evaluation, optimization, and manufacturing. These perspectives are barely considered for MMML in other domains, which makes it more difficult to learn effective multi-modal representations for engineering design.

First, engineering design needs more effective multi-modal representation to model complex design spaces. Design representations convey structural, functional, and behavioral information about a design, making design spaces more complex than other data spaces. The information is often contained in different modalities. For example, visual representations (e.g., sketches, shapes) can demonstrate the structural and spacial relations between different components more straightforwardly, while text descriptions can describe the functions and behavior of a design more clearly. Engineering design requires MMML models to effectively integrate and analyze information from different modalities to model such design spaces and obtain a deep understanding of them.

Second, engineering design requires MMML to handle noisy design representations when learning multi-modal representations. When visualizing design concepts, especially through hand-drawing sketches, designers tend to express design concepts with their personal styles, leading to variations in design representations. Moreover, annotations are often present in design sketches and diagrams as shown in Figure 2 to clarify certain design information, which are informative but need careful processing to extract the useful information and be removed for visual feature learning. We need specific techniques to handle the stylistic differences and the annotations. Otherwise, they may disguise or mix with the conceptual differences between design instances and hinder the learning of effective multi-modal representations.

Third, engineering design demands MMML to work more effectively with small design datasets. In many cases, the available design datasets for MMML are small, which can negatively impact multi-modal representation learning. This is especially true when a design space is highly diverse and varied, as MMML models may not have sufficient examples to learn from. Missing or non-parallel data is also a common issue in engineering design. That is, certain modalities may be unavailable or difficult to collect, or different modalities of design instances are not properly synchronized or matched. Accordingly, we need proper techniques to learn effective multi-modal representations from complex, noisy, and small design datasets to promote the application of MMML to engineering design.

The learning of joint representations needs a step to fuse multiple unimodal representations. In the next section, we review the different ways of doing this, each utilizing the unimodal features to a different degree.

2.2 Multi-modal Fusion

Multi-modal fusion is the concept of joining information from two or more modalities for prediction tasks (e.g., classification or regression). Such tasks benefit from fused joint representations from three aspects: 1) the multi-modal information redundancy makes predictions more robust; 2) the multi-modal information complementarity can be captured to make predictions more accurate; 3) predictions can still be conducted when a certain modality is missing [43].

Classical methods of multi-modal fusion: Before the advent of deep neural networks (DNNs), multiple kernel learning [79] and shallow graphical models [80, 81, 82] were commonly used to fuse information from multiple modalities [83, 84]. They are more suitable when training datasets are small or model interpretability is important [43]. Since DNNs have overtaken other ML methods in many tasks in recent years, this paper focuses on multi-modal fusion in DL. Inspired by the definitions from [43, 44, 45], we categorize multi-modal fusion into three classes, namely, operation-based fusion, bilinear pooling fusion, and graph-based fusion. Figure 3 illustrates how they work.

2.2.1 *Operation-based Fusion*

The operation-based approaches integrate unimodal representations using simple operations (Figure 3-A), such as concatenation [85, 86, 87, 52], averaging [88], element-wise multiplication [89], (weighted) summation [86, 87, 90], linear combination [86], and majority voting [91]. For element-wise operations, the pre-trained unimodal representations for all modalities need to have the same dimension and be rearranged in an order suitable for such operations [92]. Operation-based fusion can be done at early or late learning stages, or in a hybrid manner. Early fusion integrates low-level features extracted from each modality, allowing for the exploitation of inter-modality correlations and interactions [92]. It is easy to implement but results in high-dimensional multimodal representations, which may cause over-fitting if insufficient training data is available. Late fusion combines high-level unimodal features or unimodal decision values [91]. It allows for more intra-modality interactions and higher flexibility of the unimodal models, enabling them to learn better unimodal representations. Late fusion can also handle unexpected missing modalities more easily than early fusion. However, it overlooks the low-level interactions between modalities. Hybrid fusion is a solution that takes advantage of both early and late fusion at the cost of more complex fusion mechanisms [86, 93]. For a specific dataset and task, the optimal fusion architecture in terms of fusion stage and operation often needs to be figured out by researchers through experiments. Researchers have proposed to exploit reinforcement learning to search the possible solution space for the optimal architecture [94] or using a surrogate model to predict promising architectures [92].

2.2.2 *Bilinear Pooling Fusion*

Bilinear pooling fusion (Figure 3-B), which is also known as tensor-based [45] or bilinear model-based fusion [95], integrates multi-modal feature vectors by calculating their outer product [95] or Kronecker product [96, 97]. This approach can capture the high-order multiplicative interactions among all modalities, leading to more expressive and predictive multi-modal representations for fine-grained recognition [95, 96]. Each feature vector is often extended with an extra value of one to preserve unimodal features [96]. However, bilinear pooling that takes vectors of n-dimensional and m-dimensional as input and outputs a joint o-dimensional feature is equal to working with a 3D tensor of $m \times n \times o$, which can be huge when the unimodal dimensions and output dimension are high. Handling such a high-dimensional tensor is complex and impractical. A few low-dimensional approximations have been proposed to optimize the trade-off between expressiveness and computation efficiency. These models either decompose the high-dimensional 3D tensor into multiple low-dimensional 2D tensors [98, 99, 100, 101, 102] or replace the high-dimensional 3D tensor with a low-dimensional 3D tensor and two 2D tensors [103, 104] or a set of low-dimensional 3D tensors [105].

2.2.3 *Graph-based Fusion*

The graph-based methods utilize the strengths of graphs in modeling relations between individual elements to fuse multi-modal features (Figure 3-C). In such graphs, each node represents an instance, and an edge indicates the relationship between a pair of instances, while the node embeddings and edge embeddings carry information from different modalities [106]. Such graphs can be learned by graph neural networks (GNNs) to update the node embeddings by passing information through the edges between the nodes. The passed information fuses both node and edge information from multiple modalities. A few studies [107, 108] used this approach to fuse image and non-image features and evidenced that it outperformed simple fusion via concatenation operation.

2.2.4 *Discussion on Multi-modal Fusion for Engineering Design*

Effective joint representations learned through multimodal fusion are the core of multi-modal prediction (e.g., classification and regression) tasks. The engineering performance evaluation of the designs represented in multiple modalities falls in this task category, which is an important aspect of engineering design. Compared to other classification or regression tasks, design evaluation relies on (1) more sophisticated structural, functional, and behavioral design features from different modalities, (2) more effective fusion to learn the interactions between the features learned from different modalities, and (3) more complex cross-modal reasoning with the multi-modal features and interactions. Effective fusion can help capture complementary information from multiple modalities, facilitating multi-modal design evaluation. Accordingly, MMML models in engineering design need to capture and fuse complex and sophisticated design features more effectively to support the corresponding reasoning for design evaluation.

Besides the approaches reviewed above, attention is also commonly used for multi-modal information fusion. Since it is a powerful mechanism to align features across

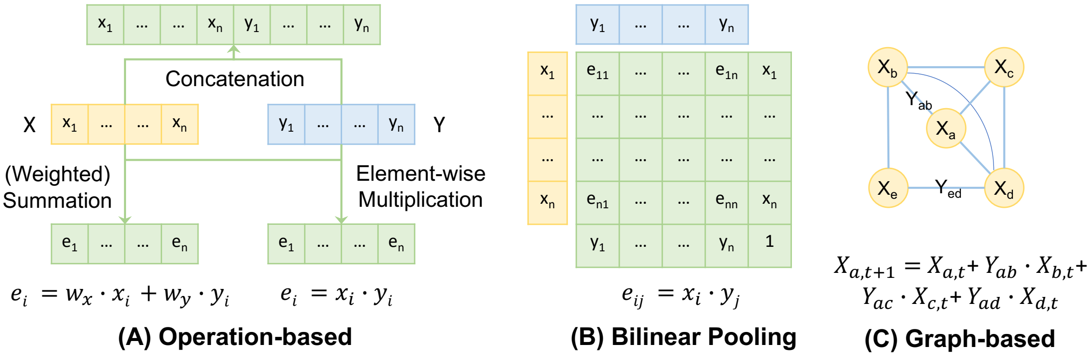

Concatenation
X
Y
(Weighted) Summation
Element-wise Multiplication
$e_i = w_x \cdot x_i + w_y \cdot y_i$
$e_i = x_i \cdot y_i$
(A) Operation-based
$e_{ij} = x_i \cdot y_j$
(B) Bilinear Pooling
$X_{a,t+1} = X_{a,t} + Y_{ab} \cdot X_{b,t} + Y_{ac} \cdot X_{c,t} + Y_{ad} \cdot X_{d,t}$
(C) Graph-based

Figure 3. Architectures of different fusion methods

modalities, it is reviewed separately in the following subsection regarding alignment.

2.3 Alignment

Alignment is defined as aligning unimodal features by finding the correlations and correspondences between elements from multiple modalities. Alignment in DL does not explicitly align features or need explicit supervision to learn the alignment [109]. In recent years, attention mechanisms have become a popular method of aligning multi-modal features [44, 45]. They can model dependencies between a query and different data elements dynamically and assign higher weights to the elements more relevant to the query [110, 111]. The query can be seen as the focus drawing attention. Multi-head attention allows us to focus on multiple data elements and preserve the important information comprehensively [110]. In MMML, cross-attention often uses queries from one modality to search for relevant features from another modality [44, 112]. Various attention mechanisms have enabled better model performance and interpretability in various tasks. We review the attention mechanisms used in MMML. Figure 4 demonstrates different attention mechanisms in aligning information of the textual descriptions and sketches.

2.3.1 Customized Attention

In DL, attention mechanisms can be customized for different data and tasks. Commonly used attention mechanisms include directional attention and symmetric attention (or co-attention), as shown in Figure 4. Directional attention uses queries from one modality to attend to another modality (Figure 4-A). In previous studies, researchers have used visual attention (i.e., using visual features to identify important semantic features) for VQA [113, 9] and image captioning [8, 114] and semantic attention (i.e., using semantic features to identify important visual features) for text-to-image synthesis [115, 116, 117]. In symmetric attention (Figure 4-B), queries from each modality are used to attend other modalities, which can highlight salient information from multiple modalities simultaneously to better reason cross-modal interactions [118, 119, 120]. Previous studies have evidenced its effectiveness in multi-modal classification [51], regression [52], IR [119, 118], and VQA [121] tasks. Additionally, multiple attention layers or stages can be stacked to capture richer alignment information and facilitate progressive reasoning. Such methods have been applied to VAQ [122, 123, 124, 8, 125, 114, 126] and text-to-image synthesis [16].

2.3.2 Multi-modal Transformer

Transformers [110] are a type of autoregressive models (ARMs) following the AE architecture built only with attention layers and feedforward layers. They were proposed for language comprehension and have become a powerful tool to learn sequential data, especially long sequences. The effectiveness of the multi-head self-attention mechanism of transformers in unimodal text learning has motivated researchers to generalize transformers to multi-modal use. Such models, such as Unicoder-AL [65], VL-BERT [66], VisualBERT [67], and B2T2 [68], take the visual tokens (e.g., regions of images) and textual tokens as input to learn the semantic, visual, and contextualized multi-modal embeddings (Figure 4-C). They are trained with different pre-trained tasks, such as masked token or image region prediction, without explicit supervision. Another strand of models extends transformers to multi-modal two-stream models, such as LXMERT [75] and ViLBERT [76]. They first derive unimodal features from unimodal tokens (e.g., semantic and visual tokens) using two streams of transformers and then

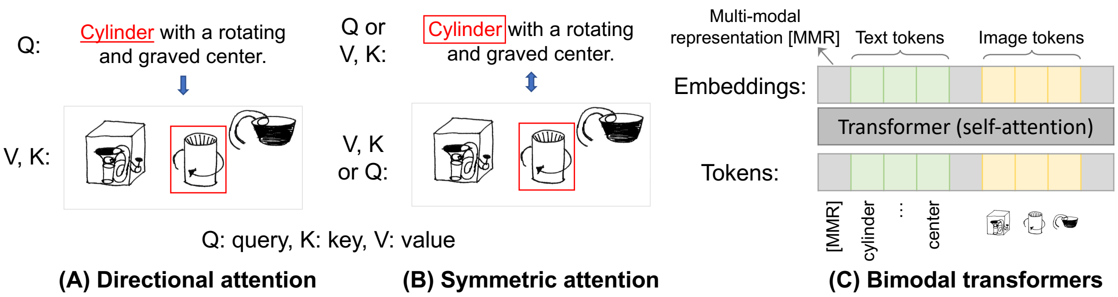

Q:
Cylinder with a rotating and graved center.
V, K:
Q or V, K:
Cylinder with a rotating and graved center.
V, K or Q:
Q: query, K: key, V: value
(A) Directional attention
(B) Symmetric attention
Multi-modal representation [MMR]
Text tokens
Image tokens
Embeddings:
Transformer (self-attention)
Tokens:
[MMR]
cylinder
...
center
(C) Bimodal transformers

Figure 4. Architectures of different alignment mechanisms

use a cross-modal attention mechanism to align the unimodal features. Data2Vec [127] takes one step further to extend bi-modal transformers to multi-modal transformers and aims at creating a general multi-modal self-supervised learning scheme. Its goal is to produce contextualized latent representations for the full input data. To accomplish this, a standard transformer with a masked view of the input in self-distillation is used.

Besides attention mechanisms, researchers have also explored other mechanisms to weigh different data elements, such as multi-modal residual networks [128], gated multi-modal units [129], and dynamic parameter layers [130]. Interested readers can refer to the specific papers for more details.

2.3.3 Discussion on Alignment in Engineering Design

The effectiveness of various attention mechanisms for capturing implicit multi-modal alignment has been evidenced in fulfilling a variety of tasks. In MMML, learning the alignment between different modalities is the basis of cross-modal synthesis. Since the current approaches can only capture implicit cross-modal alignment, there are few labeled datasets [131, 63, 74] with explicit annotated alignment to train models that can directly learn explicit cross-modal alignment from one or multiple defined perspectives. However, none of these datasets are from the engineering domain or for design applications. It is beneficial to learn explicit cross-modal alignment in a more straightforward way for applications in engineering design, since precise multi-modal design reasoning for evaluation, synthesis, and optimization relies on more sophisticated functional, structural, or behavioral alignment. Accordingly, we need well-labeled datasets to learn the finer-grained cross-modal alignment more straightforwardly to promote applications in engineering design.

Moreover, most existing models for alignment learning are constrained to aligning image and textual data, due to the rich data available from these two modalities. In engineering design, other modalities, such as sketches and shapes (e.g., voxels, meshes, and point clouds), are also commonly used to represent designs. We still lack models that can learn the alignment between these modalities and text or image representations. Additionally, while coordinated representations learn global correlation at the instance level, the attention mechanisms complement them by learning local alignment at the feature level. The effective cross-modal correlation and alignment have enabled powerful cross-modal synthesis, which is the focus of the next section.

2.4 Cross-modal Synthesis

Synthesizing information from one modality to another, known as cross-modal synthesis, is a difficult task that involves generating output in a target mode based on input in a source mode. It requires an understanding of the information in the source mode as well as the ability to produce corresponding signals or symbols in the target mode. In ML, cross-modal synthesis includes cross-modal translation and editing. The former refers to generating nonexistent samples in one modality using input from another modality, while the latter means editing existing samples according to guidance from another modality. While advances in DGMs have made this task easier, it still presents a significant challenge. In this subsection, we focus on two categories of models for cross-modal synthesis, namely, GANs and likelihood-based models. Likelihood-based models are a class of DGMs that model data distribution using likelihood functions, including multi-modal AEs, denoising diffusion models (DDMs), ARMs, and flow-based models. Figure 5 visualizes the architectures of different models using the case of synthesizing car sketches from text descriptions. We also review the metrics used to evaluate synthesis quality.

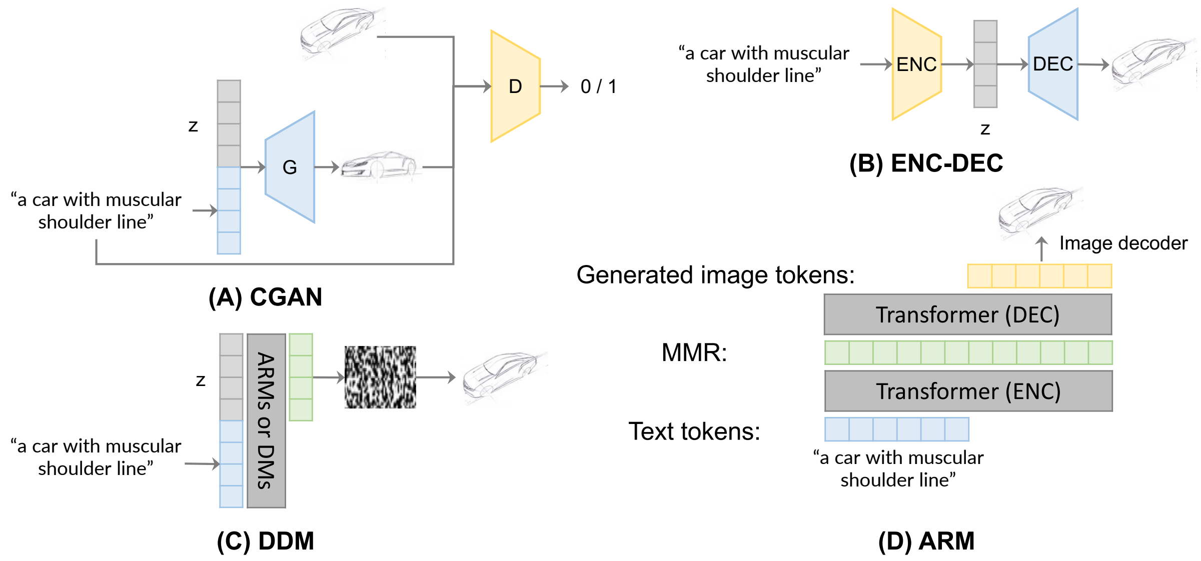


Figure 5. Architectures of different cross-modal synthesis models

2.4.1 *Conditional Generative Adversarial Networks*

Conditional generative adversarial networks (CGANs) are one of the mainstream approaches for text-to-image synthesis (Figure 5-A). A GAN model comprises a generator and a discriminator, which are alternatively trained to compete with each other. The adversarial training of the generators of GANs allows them to synthesize compelling images [11]. CGANs are GANS with both discriminators and generators conditioned on input text prompts [132, 133]. Specifically, the generator is conditioned by taking the combination of the noise sample and the text embedding as input. The discriminator views the pair of the generated image and the text input as a joint observation and judges the observation as true or fake. CGANs need to be trained on parallel data. On this basis, researchers have made efforts from different perspectives to improve image quality and the alignment between text prompts and generated images, such as using conditioning augmentation [134], stacking multiple CGANs [134, 135, 116], applying semantic attention [116] or dynamic memory [136], or adding additional loss elements or evaluation modules [135, 116, 136, 137, 138, 139, 140]. Additionally, CGANs have also been extended to take semantic concept layouts (i.e., bounding boxes with object class labels) [141, 142, 143, 144] or scene graphs [145] as input and fine-grained control to synthesize complex images with multiple objects.

2.4.2 *Multi-modal Autoencoders*

Multi-modal AEs consisting of encoder and decoder pairs are one of the most popular approaches for bidirectional cross-model synthesis (Figure 5-B). Different from unimodal AEs [146, 115, 147], the encoders of multi-modal AEs first encode instances in the source mode into latent multi-modal representations, and then the decoders generate instances in the target mode from the representations [148]. In this approach, the multi-modal representations bridge the source and target modes for text-to-image [115, 147] and text-to-geometry [146] syntheses. Multi-modal AEs can be trained on parallel data [10] or nonparallel data with pre-trained coordinated representations like CLIP [146]. Multi-modal AEs have been integrated with GANs, which can improve synthesis quality in an adversarial way [149]. As key building blocks, proper encoder and decoder models should be selected according to the source and target modes. For example, image processing mainly relies on convolutional neural networks (CNNs) [150, 151], while language processing is largely enabled by recurrent neural networks (RNNs) [152] and distributed models [153, 154]. In recent years, transformer-based models [155, 156, 157, 158] have become the default options for language processing and also been employed for image processing. shape processing is built on 3D CNNs for voxel data, GNNs for meshes, and PointNets for point clouds [159]. On the whole, the DDM, ARM, and flow-based DGMs follow the AE structure, but we review them separately as their encoder and decoder modules differ significantly.

2.4.3 Denoising Diffusion Models

DDMs are a type of latent variable DGMs that consist of a forward diffusion process and a reverse diffusion process [160] (Figure 5-C). The former corrupts training data by progressively adding noise to samples from the data distribution. The latter learns to reverse the corruption by gradually reducing noise to generate samples. Each of them can be parameterized by a set of time-dependent Gaussian transitions. Different Gaussian transition definitions result in different types of DDMs, such as denoising diffusion probabilistic models (DDPMs) [161], denoising diffusion implicit models (DDIMs) [162], and score-based models [163, 164]. During training, DDPMs and DDIMs attempt to maximize the variational lower bound of the likelihood functions representing the data distributions, while score-based models aim to minimize the losses of matching time-dependent gradients, i.e., scores. DDMs have achieved great success in conditional and unconditional image synthesis [163, 161, 165, 166] and geometry synthesis [167, 168, 169]. To improve sample efficiency of DDMs, researchers proposed to encode high-dimensional data samples into low-dimensional latent spaces and train latent diffusion models (LDMs) within the compressed latent spaces [16, 170].

For cross-modal image or shape synthesis tasks, DDMs are often conditioned through two approaches, which we term the parameter-based approach and the embedding-based approach. The parameter-based approach perturbs the means and variances of Gaussian transitions according to classifier guidance [12], similarity-based guidance [165, 166], or classifier-free guidance [171] to guide the reverse diffusion process. Classifier or similarity-based guidance requires separate classifiers or similarity evaluation modules to predict the classes or similarity scores of the samples synthesized in the current step. These modules should be noise-aware, as they also need to make predictions for intermediate noised samples. Classifier-free guidance simply takes the embedding of the guiding data as input and has been proven more effective and employed by a couple of models [165, 166, 16]. Such conditioning effect can be augmented through cross-model attention mechanisms [16]. The embedding-based approach embeds the conditioning information into the noised embeddings through separate conditioning models, such as ARMs [172] or diffusion models (DMs) [173], as depicted in Figure 5-C. This approach improves the diversity of the generated samples with minimal loss in sample realism and cross-modal correspondence.

2.4.4 Autoregressive Models

The encoders of deep generative ARMs learn distributions over sequences using the chain rule of conditional probability, whereby the decoders predict the next sequence element from the previous elements in each step. ARMs are specific to the syntheses of tokenized sequential data, such as text [74, 174, 175] and quantized images [176, 177]. Recurrent neural networks (RNNs), long short-term memory (LSTM), and transformer-based models are common ARMs for cross-modal text synthesis [178, 179]. Among them, transformer-based models have also been employed for image synthesis. The encoders and decoders of these models often work with different modalities for cross-modal synthesis. The models can also take multi-modal tokens as input to apply self-attention, enabling better cross-modal understanding and manipulation [176, 177]. However, ARMs for image synthesis generate quantized pixels one by one, impairing synthesis efficiency.

2.4.5 Other Models

Besides the models reviewed above, other models have also been explored for cross-modal synthesis. For example, flow-based models aim to explicitly learn the probability density function of real data through a sequence of invertible transformations [180]. The learned data distributions enable data generative by sampling unobserved but realistic new data points. This approach has been explored for cross-modal synthesis [181, 146]. In general, the synthesis performance of flow-based models is not on par with the other models. Additionally, implicit field models trained as classifiers have been used for shape synthesis, such as [182, 183, 184]. They take embeddings from separate shape encoders and point coordinates as input and assign a binary value to each point that indicates if this point is outside a shape or not. A shape can be inferred by sampling a set of points on its surface.

2.4.6 Model Evaluation

For cross-modal image synthesis, the quality of the generated images is often evaluated in terms of discriminability and diversity. The inception score [185] and Frechet inception distance [186] are discriminability metrics, measuring how realistic the generated images are. Previous studies have shown that they exhibit high correlations with human judgment. Multi-scale structural similarity [187] estimates the diversity of the generated images. R-precision [116] and visual-semantic similarity [137] evaluate if a generated image is semantically consistent with an input text prompt. For text-guided image editing, the generated images need to be evaluated in terms of two aspects: 1) the attribute adaption according to the text prompts and 2) the preservation of the irrelevant attributes in the original image. Cosine similarity, peak signal-to-noise ratio, and structural similarity [188] only focus on the second aspect to assess the similarity between the generated images and the text prompt. Manipulative precision is a metric considering both the above similarity and the preservation of the irrelevant attributes [189].

The synthesis quality of 3D shapes is mainly evaluated in terms of similarity. Chamfer distance [190], Earth Mover distance [190], and latent feature comparison [191] were introduced to evaluate shapes represented by point clouds. Light field descriptor [182] and Minimum Matching Distance [192] were proposed for surface-represented shapes like meshes. Intersection over union and F-score are used to evaluate the reconstruction accuracy of shapes represented by voxels [78]. Human evaluation has also been utilized to assess how realistic or consistent with the input guidance the generated samples are.

2.4.7 *Discussion*

Cross-modal synthesis is a more challenging task compared to other tasks (e.g., classification, regression). In general, CGANs have been heavily explored for such tasks and achieved higher synthesis fidelity compared to likelihood-based models. However, they need to trade off diversity for fidelity and are difficult to train. In recent years, DPMs have been exhibiting comparable or superior 2D and 3D data synthesis capabilities and attracting extensive interest [12]. In engineering design, common cross-modal synthesis models are challenged from a few other perspectives. First, the more complex and sophisticated design spaces make it more difficult to synthesize high-quality designs. Besides authenticity and conformity, synthesized designs need to be evaluated in terms of validity. The evaluation of validity involves the assessments of structural, functional, and behavioral aspects of designs, which are not considered in other domains. Accordingly, generalizable evaluation metrics and effective evaluation models are needed to make cross-modal synthesis models aware of sample "validity" in engineering design.

Second, as engineering products are designed to fulfill certain functions with given requirements and constraints, engineering performance is important to a design. However, the cross-modal synthesis models in other ML domains are not performance-aware. The performance of a design can be evaluated from a variety of aspects. For example, the evaluation of a car body design needs to consider the drag coefficient, lift coefficient, manufacturability, weight, crash safety, and so on. In traditional engineering design, these performances are often assessed through physics-based simulations or experiments. Since simulations and experiments are often time-consuming and not gradient-based, it is difficult to evaluate the generated designs in real time to inform the training of cross-modal synthesis models. Therefore, more gradient-based surrogate models for design performance evaluation should be developed and integrated into cross-modal synthesis models to support performance-aware cross-modal synthesis. Cross-modal synthesis would make more sense in engineering design when these gaps are filled.

Besides cross-modal synthesis, cross-modal correlation and alignment also allow for co-learning between modalities. We review the relevant topics in the next section.

2.5 Cross-modal Co-learning

According to the definition in [43], co-learning refers to transferring knowledge from modalities with richer resources to modalities with limited resources. It helps model data spaces of resource-limited modalities during training MMML models. In many relevant studies, only resource-limited modalities are present during test time.

2.5.1 *Cross-modal Transfer Learning*

In MMML, coordinated representations enable the transfer of data space topology (e.g., the distance between two instances) from one modality to a different modality for various tasks. Figure 6-A depicts an example of transfer the representation difference between two design instances from the text domain to the image domain to infer the unknown instance within the image domain. Such cross-modal knowledge transfer allows for zero-shot prediction in the knowledge-scarce modalities [193, 48, 72, 194, 195]. When working with hybrid multi-modal data, MMML models can use the pivot modality as the bridge for knowledge transfer between modalities without direct association, which has seen applications in machine translation [49, 196] and document transliteration [197]. When limited parallel labeled data and large amounts of non-parallel, unlabeled data are available [198], semi-supervised MMML models can be trained to transfer knowledge from label data to unlabeled data across modalities. For example, knowledge of context and feature similarities between visual objects learned from large text corpora can be transferred to facilitate image segmentation and annotation trained on limited labeled data [73]. Prior studies have also shown that MMML models trained on multi-modal data can obtain improved unimodal representations when only one modality is present during test time [178, 179].

2.5.2 Concept Grounding

Concept grounding is defined as learning semantic concepts not purely based on semantic input but also on additional modalities (e.g., vision, sound, or even olfactory perception) to mimic how we humans ground concepts through sensorimotor experience and perceptual information. In ML, a set of models project semantic and visual concepts to a common space and utilize the cross-modal association to strengthen semantic representations [199, 200]. An example of bike design is shown in Figure 6-B. With non-parallel multimodal data, the semantic and visual representations can be learned separately and then concatenated to enrich semantic representations [88, 199, 201]. For example, Vis-W2V [202] adapts the original word2vec [153] word embeddings by capturing visual notions of semantic relatedness. ViCo [203] extends GloVe [204] word embeddings to include visual co-occurrence information from Visual Genome [205].

In this paper, we extend the definition of concept grounding to include visual concepts. Visual concept grounding refers to learning visual concepts based on semantic information, besides visual information. Supervision from natural language instead of human-generated labels has been explored for visual concept grounding, which improves data efficiency in supervised learning and the generality and usability of the learned visual embeddings [206, 207, 208, 209]. In contrast, another set of models exploits weak supervision, such as Instagram hashtags or noisy labels, to learn image representations [210]. Since datasets providing weak supervision are often larger and richer in information content than the high-quality labeled datasets, representations learned by these models also exhibit high transferability. Additionally, recent studies have shown that the visual embeddings learned for image captioning can be transferred to various tasks effectively [178, 179]. It is noteworthy that grounding is beneficial only if the additional information used for concept grounding applies to the downstream tasks [179]. Otherwise, concept grounding may not result in better performance [211].

Cross-modal co-learning is also an effective tool for data augmentation. Parallel multi-modal data supports the co-training of multiple weak predictors for individual modalities. By bootstrapping each other, these co-trained predictors can discover more labeled training samples, enabling the use of inexpensive unlabeled data to augment much smaller labeled datasets [212, 213]. Such co-training processes can also identify unreliable training samples [214]. However, this may cause biases in training data and result in over-fitting [43].

2.5.3 Discussion

Cross-modal co-learning exploits complementarity, alignment, and redundancy between different data modalities. Through co-learning, knowledge residing in one modality can influence, augment, and ground representations and models of other modalities. It is also an effective technique to mitigate the challenges caused by data scarcity, which is a major issue faced by MMML in engineering design. As shown in Figure 6-A, we could have more data samples or generate data samples more easily in one modality compared to another modality. For example, it is easier to generate design concepts described by natural language, but needs more skills and time to create designs visually in general. Cross-modal transfer learning allows us to model the design space more comprehensively with the sample-rich modality (e.g., text) and generate more new design instances in the sample-scarce modality (e.g., images or sketches). The learning of effective coordinated representations is the basis of such knowledge transfer, which has been greatly empowered in recent years by the emergence of the pre-trained multimodal representations, such as CLIP. However, since such pre-trained multi-modal representations were not trained on design data particularly, the design knowledge conveyed by the learned coordinated representations is limited.

Moreover, concept grounding with multi-modal data enables us to model a design space more accurately, as exemplified in Figure 6-B. This would benefit design evaluation and better inform designers about design optimization. For example, assessing the novelty and usefulness of a group of designs relies on the accurate modeling of the associated design space. If we define a design as more novel if it is more distant from other designs, “electric bike” is the most novel design among all three within the semantics space, while “exercise bike” has the highest novelty when the semantic concepts are grounded with visual information, which is in line with human judgment.

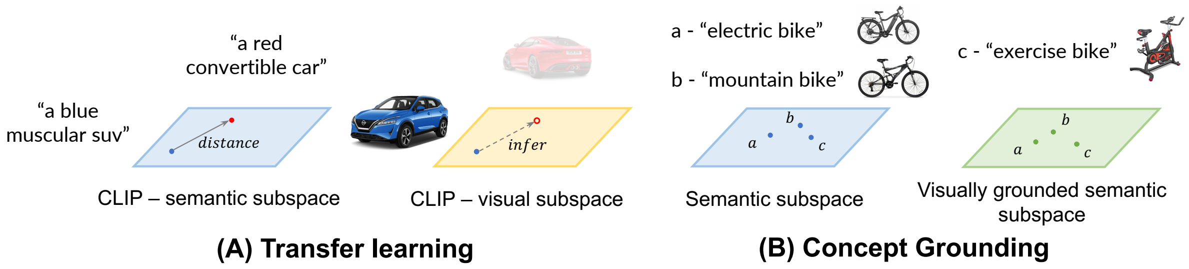


Figure 6. Examples of cross-modal co-learning. A: Two design instances exist in the semantic domain, including "a blue muscular suv" and "a red convertible car". Only the blue muscular suv exists in the visual domain. The red convertible car can be inferred by transferring the semantic distance to the visual domain. B: In the semantic domain, b - "mountain bile" is closer to c - "exercise car" than to a - "electric car". When the concepts are grounded with visual information, b is closer to a than to c, which is in line with human perception.

3 EXEMPLARY APPLICATIONS OF MULTI-MODAL MACHINE LEARNING

The fundamental concepts reviewed above are the building blocks of many multi-modal applications. In this section, we review the exemplary applications of MMML that potentially apply to engineering design. Our scope covers three major categories: 1) cross-modal synthesis, including cross-modal translation and editing, 2) multi-modal prediction, and 3) cross-modal reasoning. Since cross-modal synthesis has been reviewed as a fundamental concept in the last section, we will focus on how to apply the cross-modal synthesis models to text, image, and shape syntheses in this section. Multi-modal prediction will cover the studies that exploit MMML to improve classification and regression outcomes. Cross-modal reasoning comprises studies that rely on reasoning knowledge in one modality in response to search queries or questions present in another modality. Table 1 summarizes the fundamental concepts involved in each application category.

3.1 Cross-modal Synthesis

Cross-modal synthesis is an important part of MMML, which has attracted extensive research efforts in recent years. It is also a challenging task from three perspectives. 1) Cross-modal synthesis models need to fully understand the instance in the source mode and precisely identify its salient elements. 2) They need to produce the instance accordingly in the target mode correctly, comprehensively, and concisely. 3) The evaluation of the produced instance is difficult, as this task is open-ended. In this subsection, we review cross-modal synthesis models aiming for synthesizing text, images, and shapes.

3.1.1 Visual-to-Text Synthesis

Compared to the other modalities, visual-to-text synthesis is most explored, especially for visual captioning that generates semantic descriptions of image or video scenes. Although rule-based models were commonly used for this task before 2014 [215, 216, 217, 218, 219], we focus our review on DL-based models that comprises image encoders to capture visual features and ARM decoders to synthesize captions correspondingly. Different CNN- and transformer-based image encoders have been explored to capture visual features at the instance or finer (e.g., region, object) levels [220, 157]. RNNs and transformers are common ARM decoders. RNN-based models (e.g., LSTM) rely on memory units and recurrent connections to predict word sequences [74, 174, 175]. These models are powerful to caption simple images but lack the ability to capture the finer alignment between image regions and words in captions. In response, visual attention [8, 114] and guide vectors [19] have been applied to condition the generated texts on the visual information more tightly. Transformers have been becoming dominant since they first appeared in 2017. They rely on masked self-attention and cross-modal attention to predict caption words in a sequence and align the caption with visual features [221, 157]. Since the position encoding and attention mechanism of original transformers have been found arduous to capture the spatial relations between visual regions [222], a few variants adapt their inner architecture by adding memory modules to retain spatial relations [223, 224] or injecting spatial attention to the original attention mechanism [222]. Alternatively, another model employs dual-stream transformers to encode visual and semantic information simultaneously and a gated bilateral controller to guide the interactions between two modalities [225]. Besides ARMs, 1D CNNs have also been explored as the decoder [226, 227].

Table 1. The challenges faced by different multi-modal learning applications

	Representation	Alignment	Fusion	Synthesis	Co-learning
Cross-modal Translation	✓	✓		✓	✓
Cross-modal Editing	✓	✓	✓	✓	✓
Multi-modal Prediction	✓	✓	✓		✓
Cross-modal Reasoning	✓	✓			✓

In engineering design, sketch-, image-, or shape-to-text synthesis can be applied to automatically create natural language descriptions of designs represented visually. In recent decades, the pervasion of the internet has greatly facilitated idea sharing, saving, and discovery through websites such as Pinterest and Fusion 360 Gallery. In many cases, design ideas are shared through sketches, images, 3D models, and their renderings, with simple titles or no description at all. The absence of textual descriptions is detrimental in two aspects. First, since we mainly rely on natural language for information retrieval, the absent textual descriptions make design idea retrieval and inspiration search from such websites less guided and more time-consuming. Second, this also affects our access to multi-modal data and hinders the training of effective MMML models for applications in engineering design. The visual-to-text synthesis models reviewed above can play a role in addressing this roadblock. As of now, most attention goes to image and video captioning in other problem domains, while sketch- or shape-to-text synthesis is relatively under-explored. Since design ideas are usually created and shared using sketches and shapes, the design community should make more efforts to fill this gap.

3.1.2 *Cross-modal image Synthesis*

Text-to-2D synthesis is a challenging task that has attracted the most research interest in MMML. This task conditions image synthesis on text descriptions in different ways. In recent years, various CGANs, DDMs, and ARMs have shown great potential in synthesizing high-quality images from abstract input, such as text prompts, semantic maps, or depth maps. In this subsection, images are broadly defined to include sketches, 2D line drawings, and other 2D visuals besides normal images. Figure 7-A illustrates an example of text-to-image translation produced by Stable Diffusion [16].

CGANs witnessed the early development of text-guided image synthesis. The first few CGANs could only generate rough and low-resolution images until stacked CGANs were developed. For instance, StackGAN [134] is a two-stage CGAN. The first-stage CGAN is conditioned on input text to sketch the primitive shape and col-

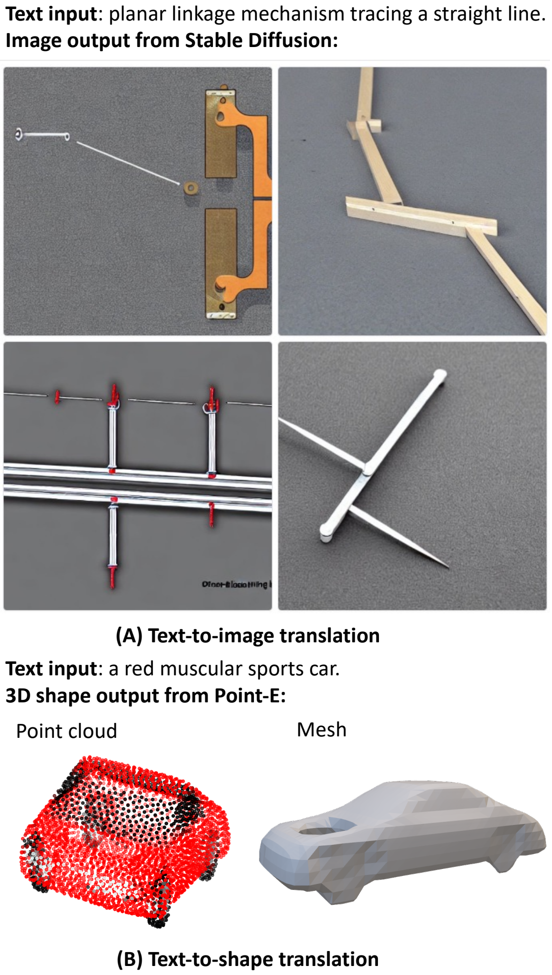

Text input: planar linkage mechanism tracing a straight line.
Image output from Stable Diffusion:
(A) Text-to-image translation
Text input: a red muscular sports car.
3D shape output from Point-E:
Point cloud
Mesh
(B) Text-to-shape translation

Figure 7. Examples of cross-modal synthesis. A: None of the planar linkage mechanisms generated by Stable Diffusion [16] meets the requirement that "tracing a straight line". B: The car generated by Point-E [172] has a hole on the hood and does not exhibit obvious "muscular sports car" features.

ors of the object, yielding low-resolution images. The second-stage CGAN is conditioned on the low-resolution image and text input to generate high-resolution photo-realistic images. It also uses a conditioning augmentation module to stabilize the training process and improve the diversity of the generated images. On this basis, DM-GAN [136] improves synthesis quality by adding a dynamic memory module to the second-stage CGAN. Alternatively, StackGAN++ [135] stacks multiple generator and discriminator branches in a tree-like structure to jointly approximate multiple distributions through different branches, further improving training stability. AttnGAN [116] further modifies StackGAN++ [135] by introducing a visual attention mechanism to correspond text tokens with image sub-regions and enforce the alignment between the text input and generated images. Other variants have also been developed to improve image resolution or alignment from different perspectives [137, 189, 228]. Different from modifying model architectures, another strand of studies improves image quality by adopting additional loss elements and evaluation modules. For example, some models use both the conditional and unconditional losses to determine whether a generated image is real and whether it matches the text description simultaneously [135, 116, 136, 137, 138]. Additionally, semantic relevance between text input and generated images [139] and the image-to-text reconstruction loss [140] have been considered in other models.

To synthesize spatially controlled or complex images, another group of CGANs is conditioned on concept layouts. Among them, the model proposed by Reed et al. [141]is able to generate images with one spatially controlled object. It first encodes the object and its spatial position into a semantic map and then employs a staked CGAN conditioned on the semantic map to create the global and local scenes in two stages. Layout2Im [142] extends this to multi-object synthesis by fusing multiple objects into a single semantic map. Other alternative models take scene graphs (i.e., graphs of text descriptors) [145] or text descriptions [144, 117] as input and convert them to semantic maps using specifically designed modules to condition the following CGANs. Instead of using semantic maps, another model uses two pathways to take the entire layout and the object class labels as input to generate the background and objects separately [143]. In general, stacked CGANs featuring two-stage processes are still limited to synthesizing low-resolution images due to computational cost. Deep fusion GAN (DF-GAN) [229] circumvents this issue using a one-stage process, which contains a text-image fusion block for deeper fusion between semantic and visual features and a target-aware discriminator. Besides CGANs, StyleGAN [230] has also been adapted for cross-modal image synthesis. StyleGAN can control the style of the synthesized images through a style-based generator that uses a matching network to learn the target style and integrate that into the synthesis process [230]. To adapt it for cross-modal image manipulation, a couple of models revise the matching network to condition image synthesis on text or image input and aim to maximize the similarity between the CLIP embeddings of the generated images and the input [231, 232, 233].

The adaption of transformer-based ARMs to the image domain is enabled by quantized image representations, which can be seen as a context-rich codebook of image constituents. They are often learned using a discrete CNN-based variational auto-encoder (VAE) consisting of a quantized encoder and a quantized decoder [147]. Vector quantized-variational auto-encoder (VQVAE) [147] and DALL-E [15] employ such ARMs for text-guided image manipulation through a two-stage training procedure. The first stage trains the discrete VAE and uses the encoder to generate quantized image representations. Then, the second stage trains an ARM decoder conditioned on text input for image generation. Vector quantized-GAN (VQGAN) [234] attaches a quantized encoder before a GAN with an ARM generator, which conditions the GAN on the encoded input to the model. A variety of variants have been developed to adapt VQVAE and VQGAN for image manipulation using textual or visual guidance, such as text prompts, category names, depth maps, semantic images, or poses [235, 176, 177, 236]. Such models combine the effectiveness of CNNs' in learning local interactions with transformers' expressiveness in handling long-range interactions and learning fine-grained cross-modal controls, facilitating the synthesis of high-resolution images.

Guided DDMs for image manipulation draw intense attention in recent years. A few emerging text-guided image synthesis models employ this approach, and many of them condition the reverse diffusion process using classifier-free guidance [165]. Guided language to image diffusion for generation and editing (GLIDE) [165] made an early attempt to compare the CLIP guidance and classifier-free guidance conditioning approaches and found that the latter achieves higher photo-realism and caption similarity. Stable Diffusion [16] employs a two-stage training process to improve the computational efficiency of DDMs. The first stage learns a 2D latent space of images for conceptual compression through a VAE, and the second stage trains a latent diffusion model (LDM) within the compressed space for semantic compression. A cross-modal attention mechanism is applied during the reverse diffusion process to improve the alignment between the generated images and the input. Imagen [237] stacks multiple DDMs to improve the quality of generated images. Besides classifier-free guidance, other conditioning approaches have also been explored. DiffusionCLIP [166] utilizes a directional CLIP loss as guidance, which allows for multi-attribute manipulation. DALL-E 2 [173] conditions the reverse diffusion process through the embedding-based approach. The authors compared using ARMs and DMs to fuse conditioning information into noised embeddings and found that the latter is more efficient and effective. It stacks two DDMs as the decoder to generate images conditioned on the noised embeddings.

Additionally, gradient optimization has also been explored for this task. For example, CLIPDraw [238] synthesizes novel drawings conforming to text input by optimizing the colors and positions of a set of Bezier curves through CLIP loss. In general, the ARMs and DDMs reviewed above can take an initial image along with conditioning information as input, making them capable of both translating instances in the source mode to images and editing images according to guidance in the source mode. The CGANs mainly aim for the translation task, while those incorporating image encoders can do image editing as well, such as ManiGAN [189], VQVAE [147], DALL-E [15], and VQGAN [234].

Cross-modal image synthesis is one of the most promising applications of MMML in engineering design. In design education, free-hand sketching or drawing design ideas is a necessity, which disqualifies many people without professional training as effective designers. However, even trained designers or engineers may have varying free-hand sketching or drawing abilities, and it is often time-consuming for them to manually create a sketch or drawing. In contrast, it is relatively easier and faster for most people to describe design ideas abstractly in natural language or using simple semantic layouts. On this basis, the models reviewed in this subsection have great potential to generate the corresponding sketches or images automatically. This application can benefit engineering design from a few perspectives. First, it can not only reduce the workload of professional designers but also make engineering design doable for non-professional designers. Since automated design idea visualization brings down the barrier to engineering design, it may invite more people to design exercises and promote large-scale design customization. Second, MMML-based design visualization may improve design creativity. Visualizing a design idea can be seen as a process of re-organizing a designer's usable knowledge, which is open-ended but limited by the person's knowledge basis. Since such cross-modal synthesis models are often trained on large multidisciplinary datasets, their "knowledge basis" could be more comprehensive compared to that of individual designers. This may enable interdisciplinary knowledge transfer, resulting in novel designs. Additionally, compared to the sketches and drawings generated by humans, images generated through MMML could be more realistic, which facilitates visual inspections and evaluations at early design stages. Since 2D visuals are often used as intermediate design representations, we need 3D design representations for down-streaming detailed design, prototyping, and manufacturing. Accordingly, we focus on cross-modal shape synthesis in the following subsection.

3.1.3 Cross-modal shape Synthesis

In engineering design, 3D representations (e.g., meshes, point clouds, voxel data) faithfully portray shapes and can accurately represent sharp, extrinsic features using a high level of detail. They are broadly adopted in the middle and late design stages to evaluate, optimize, and prototype designs in more detail and prepare a completed design for manufacturing. Due to the complexity of 3D representations, cross-modal shape synthesis is the most challenging cross-modal synthesis task as of now. Text prompts, images, 2D renderings, and depth maps are commonly used guidance for manipulating shape synthesis. In general, the DGMs for cross-modal shape synthesis share similar architectures with those for image synthesis at a high level, but work with 3D representations. Different from images that are mostly represented by structured pixels, shapes can be presented by structured voxels and unstructured meshes or point clouds. Each representation affords the adoption of different DGMs. In the following, we review the DGMs for cross-modal shape synthesis and the corresponding representations they used. Figure 7-B exemplifies the text-to-shape translation produced by Point-E [172].

Multi-modal AEs attracted early interest in image-to-shape synthesis. With the success of CNNs in synthesizing images, a group of models employs 3D CNNs to synthesize voxel-represented shapes. For example, the 3D recurrent reconstruction neural network (3D-R2N2) [239] comprises an image encoder and a 3D voxel decoder is trained with parallel data to convert one or multiple 2D views into shapes. Voxel representations are often limited to low resolutions due to high computational costs. To overcome this, Mesh R-CNN [240] integrates a mask R-CNN 2D perception module (i.e., the encoder) with a 3D CNN voxel decoder to produce coarse voxel representations, which are then converted to meshes and improved by a GCN refinement module. Multi-View Stereo (MeshMVS) [241] extends Mesh R-CNN by further conditioning the GCN refinement modules on the depth images estimated from the input images. Another strand of multi-modal AEs employs the point cloud representation. PointOutNet [242] joins an image encoder with a decoder to predict point coordinates of the output point clouds. AtlasNet [243] modifies PointOutNet [242] by imposing that the points of a point cloud are from multiple deformed parametric surfaces or a sphere for improved point cloud quality. Target-embedding VAE (TEVAE) [244] consists of a mesh VAE and a 3D extrusion (i.e., a mesh generated by extruding 2D sketches) encoder, where each mesh is represented by a feature matrix. During training, it minimizes the mesh reconstruction loss and the Euclidean loss between the embeddings from the mesh and extrusion encoders simultaneously, which allows for unconditional shape generation and sketch-to-shape translation.

CGANs and ARMs are popular for cross-modal synthesis in the 2D domain but are under-explored in the 3D domain. GANs can generate high-quality images but are limited to low resolutions for shape synthesis [245, 246]. To mitigate this issue, efforts have been made to represent shapes with 2D data. Achlioptas et al. [190] proposed to represent point coordinates of a point cloud as a matrix and train VAEs and GANs with the 2D representation. The learned latent space enables semantic editing and other manipulation. A set of work maps shapes to 2D parameter domains, then trains GANs to generate samples in the 2D domains, and finally converts them to 3D meshes [247, 248]. Rank3DGAN [249] and X-dimensional GAN (XDGAN) [250] extend this approach to conditional settings for semantic manipulation during mesh generation. ShaperCrafter [251] adapted ARMs to the 3D domain. It learns a vector-quantized grid-based implicit representation for shapes using a point VQVAE and conditions the ARM decoder on BERT embeddings of text prompts for shape synthesis and editing.

Compared to CGANs and ARMs, DDMs have been adapted to the 3D domain with greater success. Most 3D DDMs work with 3D point clouds. Luo et al. [167] proposed a DDPM working with point clouds. The model conditions the reverse diffusion process on latent representations learned via a normalizing flow module in an unsupervised way. The shape latent enables the unsupervised representation learning of different shapes. Point-Voxel Diffusion (PVD) [168] marries DDMs with point-voxel representations [170] of shapes. Besides unconditional synthesis, PVD enables conditional syntheses, such as shape completion and depth-to-shape translation. Similar to stable diffusion [16] for guided image synthesis, LION [169] employs an LDM for guided shape synthesis. It first encodes the voxel and point cloud representations of the input shape into two latent spaces using two point-voxel-CNN VAEs [170]. Then, two latent score-based DDMs are trained in the latent spaces, respectively. The latent point DDM is the main generator, which is augmented by the latent voxel representation (i.e., indicating the global shape) to boost model expressivity. When conditioned on CLIP embeddings, LION is capable of image- and text-to-3D synthesis. Point-E [172] integrates GLIDE [165] with a DDM working with point clouds to convert text prompts to 2D images and then to 3D RGD point clouds. Following the embedding-based conditioning approach, LION employs a transformer encoder to condition the noised point cloud representation on the CLIP embedding of the 2D image, enabling image- and text-to-shape synthesis. Attempts have also been made to apply DDMs to 3D voxels [78] and tetrahedral tessellation (i.e., a type of mesh) [252] representations.

Implicit field representations enable a set of cross-modal synthesis models specific to the 3D domain. An implicit field model can be trained as a classifier, such as implicit field net (IM-NET) [182] and deep signed distance function (DeepSDF) [184]. They take an embedding of a shape and point coordinates as input and assign a value to each point which indicates if this point is inside or outside the shape. Researchers have developed a set of multi-modal AEs consisting of image or text encoders and implicit shape decoders for text- or image-to-shape translation [183, 253, 254]. These models need to be trained on parallel data. Beyond that, Ibing et al. [192] proposed to train an AE with a voxel encoder and an implicit decoder to learn latent 3D space. Then, a GAN is trained in the latent space to generate new latent representations that can be decoded into shapes. The GAN can be conditioned on bounding boxes or class labels for shape manipulation. Instead of training GANs in the latent space, CLIP-Forge [146] uses a normalizing flow model to convert the latent space to a normal distribution and condition random samples from this distribution on input text prompts or images for shape manipulation. Instead of AEs, auto-decoders (ADs) have also been employed to learn latent spaces of shapes [184]. 3D-LDM [255] trains an LDM in the latent space learned by a DeepSDF-based AD to generate diverse and high-quality shapes, which can be conditioned on CLIP embeddings for shape manipulation. Similarly, multi-modal variational auto-decoders (MM-VADs) [256] train two 2D auto-decoders and a 3D implicit auto-decoder simultaneously to learn a cross-modal latent space applied to sketches, RGB views, and shapes. The cross-modal latent representation allows for manipulating shapes using sketches or RGB views.

Different from the reviewed DGMs, a few other models synthesize shapes by deforming existing shapes. For example, Pixel2Mesh [257] generates 3D meshes using a GCN to progressively deform an ellipsoid according to the perceptual features extracted from input images. Text2Mesh [258] and ClipMatrix [259] edit texture styles of input meshes according to text prompts. They emplo to predict vertex displacements and color details, which are trained to maximize the CLIP similarities between mesh 2D renderings and text prompts.

In engineering design, the generation and editing of shapes are more challenging and labor-intensive compared to that of images for both human designers and MMML models. The emergence of large pre-trained multi-modal representations (e.g, CLIP and CISP) enables the training of cross-modal shape synthesis models with non-parallel data. This mitigates the scarcity of parallel data for training such models. Overall, engineering design would benefit from cross-modal 3D synthesis from a few perspectives. First, it would accelerate design visualization and modeling in the early design stages, ranging from concept generation to system-level design, and to detail design. Effective early-stage design features the generation and examination of a rich set of design ideas, during which visualizing and concretizing design ideas are as important as conceptually conceiving them. When done manually, the creation of shapes requires professional skills and is time-consuming. High-fidelity shape synthesis guided by text prompts, images, and sketches can reduce human efforts and accelerate the design process.

Second, it would enable broader exploration in the early design stage. Traditionally, the time and efforts needed for creating 3D design representations force designers to select and focus on a limited number of design candidates quite early. When less time and effort are needed to create shapes, designers can bring a larger number of designs to later stages and closely inspect them. 3D shapes support various qualitative and quantitative design evaluations, such as visual appeal evaluation, structural inspection, and performance evaluation through finite element analysis. The detailed evaluation also opens up opportunities to further optimize existing designs or ideate new designs, further extending the exploration range. Moreover, designers can invest the spared time and effort in cognitive activities, such as analogies or inspiration search, information integration, idea conceiving, reasoning, and decision-making to plan and guide the exploration in a better way.

Third, it would boost knowledge transfer and reuse during design synthesis. Human-conducted cross-modal translation or editing is constrained by the knowledge bases and imagination of designers, which are subject to design fixation. AI-enabled design synthesis models make it possible to exploit the knowledge basis underlying large datasets. Integrating cross-modal synthesis models and pre-trained multi-modal representations makes the multi-modal knowledge buried in huge multi-modal datasets usable for cross-modal synthesis models. Compared to unimodal synthesis, such extension and fusion of the underlying knowledge basis may enable the synthesis models to generate new designs beyond the boundary of the existing solutions and transfer knowledge across disciplines or product domains. Therefore, cross-modal shape synthesis may generate more diverse and more creative designs than human designers. For example, (Refer to the figure.)

Better design concepts and final design outcomes are more likely to emerge from accelerated iterations, broader exploration, and boosted design creativity resulting from automatic cross-modal translation and editing at the early design stages. In practice, it is unusual for designers to transition a design from abstract descriptions to detailed 3D models directly. 2D sketches are often needed as an intermediate representation to support the creation of more complicated 3D models. Text-to-shape synthesis will potentially enable us to simplify and accelerate the design process by eliminating the step of sketching. However, although text-to-shape synthesis has great potential, it is still a challenging task. The shapes generated by the current models are subject to low precision and fidelity, as shown in Figure 7-B. Two reasons explain this. First, it is difficult to express 3D visual features using natural language point for point. In comparison, it is relatively easier to transfer visual features from 2D to 3D. Second, the available text-shape datasets for training text-to-shape translation models are much smaller in volume compared to paired text-2D datasets. The scarcity of training data hinders the trained models from learning a comprehensive set of common features to generate high-quality new designs. It is beneficial to follow the human design process and break this challenging task down into two steps: text-to-image synthesis and image-to-shape synthesis. As of now, the first step is better solved, as suggested by the strong capabilities of the text-to-image synthesis models reviewed above.

3.2 Multi-modal Prediction

Multi-modal prediction (e.g., classification and regression) allows us to utilize complementary information of multiple modalities to assess instances more accurately and comprehensively. For such tasks, the state-of-the-art models usually first encode information from different modalities and then fuse multiple unimodal embeddings into multi-modal representations before the prediction heads. They differ in the approaches used to encode different modalities and learning schemes.

3.2.1 Multi-modal Classification

Classification is the most common task in ML, which is also true in MMML. We have seen the applications of multi-modal classification in many domains. In HyCon [260], the authors proposed a novel method based on hybrid contrastive learning for multi-modal sentiment analysis. Specifically, the HyCon model learns both intra-modal and inter-modal interactions through both contrastive and semi-contrastive learning based on audio, visual, and text inputs. DMDE [261] is a deep multimodal design evaluation model that features a bidirectional encoder and a self-attention-based fusion model to predict overall and category-specific sentiments with inputs of design images and view hierarchy. TechDoc [106] exploits three types of information, including text, images, and document associations to predict patent classes. The authors used a pre-trained CNN and a bi-directional RNN to respectively learn image and text features and fused them through concatenation. Then, they integrated the fused features with association information through a GNN and demonstrated the multi-modal model achieved more accurate classification results than the unimodal models. Additionally, Zhou et al. proposed a novel multimodal CLIP-based method for fake news detection [262]. The model named FND-CLIP first encodes visual information with ResNet and CLIP model and textual information using BERT and a CLIP model. The multi-modal encodings are processed through weighting by cross-modal similarities and redundancy reduction. Further, their work introduces a cross-modal attention mechanism before the final classifier. Besides images, texts, videos, and audio, there are also works that use novel modalities for new tasks. For example, Deng et al. proposed a novel method called DDIMDL for drug-drug interactions based on features such as chemical substructures, targets, pathways, and enzymes [263].

3.2.2 Multi-modal Regression

Because of the complex nature of regression tasks, multi-modal information can be extremely important in increasing the accuracy and effectiveness of the models. Most works focus on proposing more effective multimodal information fusion schemes to facilitate the downstream regression heads. For example, Song et al. [20] proposed an MMML model to predict design metrics using design sketches and text descriptions. The visual and semantic embeddings are fused and aligned through a symmetric cross-modal attention mechanism. Similarly, Yuan et al. [53] employed an attention mechanism to fuse image and text descriptions to predict customer evaluation scores. Besides incorporating multi-modal data such as images and texts, other works improve the state-of-the-art methods by allowing for even more modalities. For example, Pakdamanian et al. introduced an architecture named DeepTake [264] for the novel task of driver takeover prediction. Specifically, they aimed to predict the intention, time, and quality of human driver takeover in an autonomous vehicle based on multi-modal information such as vehicle states, non-driving related tasks, biometrics, and pre-driving survey. The number of modalities present in this work is significantly higher than that in other tasks, making it both challenging and novel. The proposed method features modality-specific preprocessing and feature extraction methods. Then, the multimodal representations are fused and used by an MLP to produce the final predictions.

As reviewed and discussed in the last subsection, cross-modal synthesis has great potential for reforming our traditional design process. Different from data synthesis in the ML domain, the synthesized new designs need not only to be visually realistic but also to meet certain engineering standards, user requirements, or market preferences. Although design instances are often represented in multiple data modes, most prior surrogate models for design evaluation only utilize unimodal data, limiting evaluation comprehensiveness and accuracy. Multi-modal prediction techniques reviewed in this subsection can help overcome this issue. For example, the MMML models proposed by Song et al. [20] and Yuan et al. [53] attempted to address the challenging tasks of evaluating designs based on images and text descriptions. When compared with uni-modal schemes, both of their approaches bring a performance increase in design evaluation. Taking multi-modal design representations as input, classification models apply to categorical evaluations, such as if a generated design is valid or not, while regression models apply to the evaluation of real-valued attributes, such as drag coefficients of cars or flying ranges of drones. The input can be combinations of common design representations, such as textual description and sketch pairs for evaluating the novelty and usefulness of conceptual designs, shapes and tabular specification pairs for evaluating the performance of detailed designs, and image, tabular specification, and text description triplets for evaluating user preference and market popularity of products from e-commerce websites.

3.3 Cross-modal Knowledge Extraction

Traditional knowledge extraction mainly focuses on natural language processing to match queries with target text content. Cross-modal knowledge extraction allows for the search of relevant information across modalities from a larger range. This is done by capturing the correlations or alignment between different modalities.

3.3.1 Cross-modal Information Retrieval

Cross-modal IR aims to search instances in a target mode that are closest to queries in a source mode. According to the representations used for retrieval, we can classify cross-modal retrieval into unimodal retrieval and multi-modal retrieval. The unimodal retrieval approach only needs to learn the representations of instances in the source mode. It first retrieves instances nearest to the search query in the same mode through similarity metrics [154, 265] or k-nearest neighbor models [266, 267]. Accordingly, the counterparts of the selected instances in the target mode are returned as the retrieved candidates. This approach is simple but only applies to parallel data. Moreover, high similarities in the source model do not always lead to good cross-modal retrieval results, which can be partly overcome by carefully designing suitable metrics to rank potential candidates [154, 265]. Alternatively, the multi-modal retrieval approach projects different modalities to a common multi-modal space, within which the most relevant instances can be found based on similarities between search queries and candidates. Researchers have explored a variety of multi-modal representation spaces for this task, including manually defined intermediate space [268], coordinated semantic-visual spaces at the instance level [199, 269, 270] or the element level [61], or a joint multi-modal space through cross-modal hashing [89]. The multi-modal approach exhibits two strengths compared to the unimodal approach. First, it can learn more expressive representations that reflect both modalities, leading to better retrieval outcomes. Second, the presence of a common multi-modal representation space allows for bidirectional retrieval. In practice, more effective unimodal or multi-modal representations are required for high-quality cross-modal retrieval.

In engineering design, cross-modal IR can facilitate design knowledge gathering and design analogy and inspiration search. As mentioned before, several well-known design repositories, such as Pinterest and Fusion 360 Gallery, provide rich design precedents with very brief titles or no description at all. Since text queries are most commonly used for IR, the absence of text descriptions can be a barrier to the effective reuse of the design knowledge shared in such repositories. With the existence of annotated parallel datasets (e.g., PartNet), it is possible to train cross-model IR models, which can capture the correlations between 2D or 3D design representations and textual search queries. By transferring the learned knowledge to unimodal repositories like Pinterest and Fusion 360 Gallery, the absence of textual information can be overcome. On this basis, we can potentially achieve more efficient and accurate design knowledge retrieval.

The above discussion on each potential application also suggests that MMML may facilitate human-AI collaboration in engineering design. Design representations involve a variety of data modes with varying levels of expressiveness, abstraction, and elaborateness. For example, text descriptions and sketches as design representations are highly abstract, while shapes are more elaborate and expressive. Humans and AI are respectively more capable of working and conducting reasoning with different modalities for different tasks. With strong cognitive abilities, humans do better in guiding the human-AI hybrid exploration process and making decisions at more abstract levels, such as conceiving and describing highly abstract design ideas. This abstract information processing and reasoning is more challenging for AI to learn from data. With powerful computational power, AI is better at processing and reasoning more expressive and detailed information at more elaborated levels, such as creating or evaluating more complicated design representations (e.g., shapes). With MMML, it is possible to have humans and AI work on tasks that they are better at and apply their strengths to a better degree.

4 CHALLENGES AND OPPORTUNITIES OF MULTI-MODAL MACHINE LEARNING IN ENGINEERING DESIGN

On the one hand, MMML has the potential to revolutionize the engineering design process. By integrating multiple sources of information, including 2D visuals (e.g., images, sketches), text, 3D shapes, and others, MMML models can learn about a problem more comprehensively from multiple perspectives, leading to a deeper understanding and more informed decision-making. In engineering domains, designers often evaluate design candidates through simulations, experiments, or expert assessments, which are time-consuming, expensive, and resource-intensive in many cases. MMML will facilitate the applications of ML models to design evaluation by achieving faster, cheaper, and increasingly more accurate predictions. Moreover, MMML models can automate parts of the design generation process, which will reduce the time and effort required significantly and enable broader design exploration. Personalized and optimized designs can be created by considering individual preferences and requirements and multiple objectives and constraints, respectively. MMML is promising to enhance creativity, improve product quality, and increase efficiency, making it a valuable tool for the field.

On the other hand, MMML for engineering design faces several challenges, such as the availability of multi-modal design datasets and high-quality labels, modality-specific representation learning, scalability, algorithm compatibility, explainability, and adaptation to new modalities. Integrating different modalities of data, such as images, text, and 3D shapes, can be difficult, and the data quality of different modalities can vary, leading to biases in the model. Labeling the data for all modalities can be a time-consuming and resource-intensive task. Different modalities may require different algorithms for representation learning, making it challenging to build a unified model. MMML models can also be computationally expensive and difficult to interpret, making it challenging to understand how decisions are being made. Finally, incorporating new modalities for MMML can be challenging, as it will increase the complexity of the interactions between different modalities and require modifying the model architecture and retraining the model. Below, we discuss some of these challenges in more detail.

4.1 Large Multi-modal Design Datasets Are Needed

Unlike unimodal data, parallel multi-modal data is more difficult to collect, resulting in fewer multi-modal datasets supporting MMML. A few large parallel text-image datasets are publicly available to support semantic-visual representation learning, image captioning, and text-to-image syntheses, such as MS-COCO [271], Visual Genome [205], YFCC100M [272], JFT-300M [273], and AVA [274]. Text2Shape [275] and Text2Shape++ [251] are parallel text-shape datasets supporting semantic-geometric representation learning and text-to-shape synthesis. These two datasets only cover two object classes (tables and chairs) and only the former is publicly available so far. Many of the existing MMML models were trained on these datasets. Since the available multi-modal datasets contain little design-specific information in terms of structure, function, behavior, and performance, their applications to engineering design are limited.

The scarcity of large and high-quality multi-modal datasets is a challenge in engineering design. Most large design datasets comprise design instances in a single modality, such as sketches in Pinterest or 3D models in ShapeNet and Fusion 360 Gallery. However, MMML requires the datasets to include multiple modalities that align with each other, limiting the options for model training. DNNs with large amounts of parameters can be difficult to train without sufficient quality and quantity of training data. Transfer learning can help overcome this challenge by transferring knowledge from large external datasets [52], but these general ML datasets may not provide enough design knowledge for pre-training. To support the training of high-performing MMML models and effective knowledge transfer for complex design tasks, large multi-modal design datasets are necessary.

4.2 Design-specific Labels and Metrics Are Needed

For general ML, assigning an object class to each data sample is a commonly used approach to label datasets, supporting the training of various classification models for computer vision and language comprehension. However, engineering design involves a set of factors beyond simple classification, such as novelty, usefulness, user preference, physical constraints, and performance, to understand and assess a design solution. Training MMML models to comprehend engineering designs automatically relies on high-quality labeled data. Since simulation-, experiment-, or expert-based design labeling process is often expensive and difficult to scale, we are lacking in high-quality and large labeled design datasets. To address this challenge, the community should work together to create, organize, and maintain large design datasets with high-quality labels. This can be achieved by clarifying design requirements and goals during data curation and sharing, collecting and storing aligned design data represented in different modalities, labeling the datasets with design-related attributes (e.g., function, material, structure, weight, dimension, physics-based performances under different operating conditions, and application context), and providing pre-trained embeddings or latent representations of the data if available. These efforts will benefit the entire engineering design community greatly.

Additionally, the quality of the data samples produced by cross-modal synthesis models is often evaluated in terms of fidelity and cross-modal consistency through metrics like inception score [185], Frechet inception distance [186], Chamfer distance [190], and Earth Mover distance [190]. Since a valid design has richer meaning than fidelity and cross-modal consistency, these metrics are not informative enough for engineering design applications. The design community needs to propose generalizable design-specific evaluation metrics to assess if a synthesized design is valid in terms of function, behavior, and structure. Moreover, when more labeled design datasets are available, various surrogate models need to be developed to support various performance and constraint satisfaction evaluations and optimizations during design synthesis. This is an important step toward contextualizing MMML and general ML into engineering design.

4.3 Effective Design Representation Learning Techniques Are Needed

The most popular data modes in ML are text and image data due to their widespread availability. Accordingly, most MMML research also focuses on these two modalities. However, 2D design sketches or engineering diagrams and 3D shapes, which are commonly used design representations, are under-explored in ML. Although CNN-based image models apply to 2D design drawings, the sparsity of these sketches or diagrams hinders the performance of these models in comprehending them. Similarly, 3D CNNs and GNNs generalized to non-Euclidean spaces can learn different types of 3D data, but they are only effective in capturing local or coarse global features at low resolutions, struggling to capture fine-grained features comprehensively. Since many different types of design inspections (e.g., movement interference), performance evaluations (e.g., computational fluid dynamics and finite element analysis), and design optimizations are conducted with 3D representations, effective 3D shape learning techniques are particularly important and in demand in engineering design. This presents an opportunity for design researchers to develop specialized ML architectures for the modalities that are more common in design.

Furthermore, design representations, such as hand-drawing sketches, can be noisy and exhibit distinctive personal styles. As different designers have inconsistent abilities and preferences in expressing design ideas, similar ideas may appear in various styles and contain levels of detail. The variations in representation styles and levels of expressivity, abstraction, and elaborateness can become more substantial as the target product becomes more complex. When represented in multiple modes, some design instances may miss a certain mode as well. These challenges pose difficulties for design representation learning, as current ML models struggle to distinguish conceptual differences from representation variations, compensate for missing modes, and identify abstraction and elaborateness levels. To overcome these issues, the design community must train its own MMML models on large design datasets to make them robust to personal styles and varying abstraction and elaborateness levels for better comprehension and reasoning. Building effective human-AI hybrid teams where humans and AI can learn from and complement each other for challenging tasks could also be a future effort, as it is relatively easier for human raters to handle representation inconsistencies and missing information.

4.4 Approaches to Sophisticated Design-related Reasoning Are Needed

The state-of-the-art information reasoning in general multi-modal tasks is limited to aligning features and capturing spatial relations across modalities. This is not enough for applications in engineering design, as design reasoning often involves interrelations among different components from functional, structural, and behavioral perspectives. For example, despite their different physical structures, both rotors and fixed wings can provide lift for aircraft, while the former allows for vertical take-off and landing and the latter relies on runways to take off and land. MMML models need to capture and align these complex engineering relationships from different sources and conduct the corresponding reasoning in the design process. While various cross-modal attention mechanisms have been developed to capture these relations implicitly, the community still requires large-scale carefully annotated data and relevant MMML models that can learn such relations explicitly and conduct corresponding design reasoning.

4.5 Pre-trained and Generalizable Multi-modal Representations Are Needed

As reviewed in the last section, the emergence of pre-trained semantic-visual representation, such as CLIP, has greatly promoted the development of cross-modal synthesis models. Although such pre-trained multi-modal representations support text- or image-guided syntheses of 2D visuals or 3D shapes for general purposes, they do not provide any implications specific to engineering design. That is, these pre-trained representations struggle to capture and reflect detailed design features, requirements, and descriptions properly, which may lead to invalid outcomes when applied to design evaluation, reasoning, and synthesis. For example, in Figure 7-A, the pre-trained model fails to understand the design requirement conveyed by "planar linkage mechanism *tracing a straight line*" and none of the generated linkage mechanisms meet the requirement. To better support engineering design tasks, the design community needs to adapt the existing pre-trained models by adding more domain knowledge to them or train new representation models on large design datasets to capture more domain-specific knowledge during learning multi-modal representations.

4.6 Effective Models to Integrate Additional Modalities Are Needed

Besides the textual and visual data modes discussed above, representing, inspecting, evaluating, and optimizing engineering designs may involve additional data modes, such as olfactory, haptic, auditory, emotional, and ergonomic data. These data modes are under-explored in ML and there is a need for effective models to process such information in the design community. Bringing more data modes into MMML models for engineering design can potentially enhance the comprehensiveness and effectiveness of such models in design understanding, reasoning, evaluation, and optimization. Since most published MMML models are limited to two data modes, a future direction is to explore MMML architectures that can accommodate multiple data modes.

5 CONCLUSION

In inclusion, multimodal machine learning (MMML) has the potential to revolutionize the engineering design process by enabling better design knowledge capturing, integrating, and reasoning with multiple forms of design information. Through this review, we have reviewed the approaches to addressing the fundamental concepts of MMML, including multi-modal representation learning, information fusion, alignment, synthesis, and co-learning. On this basis, we have also discussed the current state of the art in MMML research, including its applications in cross-modal synthesis, multi-modal prediction, and cross-modal reasoning along with the potential impact on design knowledge extraction, design synthesis, evaluation, and optimization. However, MMML still faces several technical challenges, such as the scarcity of high-quality multi-modal design datasets, the difficulties in aligning complicated design features across modalities, the need for handling noisy and inconsistent design representations, and the lack of comprehensive MMML models that can incorporate various design-related data modes. In addition, there is a need for more extensive empirical case studies and evaluations to verify and demonstrate the effectiveness of MMML in real-world engineering design applications.

Despite these challenges, the future of MMML in engineering design looks promising, with the potential to significantly impact how products are designed and manufactured. Moving forward, there is a need for further research to address the remaining challenges in MMML for engineering design applications and to develop MMML models to fully realize its potential and bring about the next generation of intelligent design tools.

REFERENCES

[1] Bengio, Y., Courville, A., and Vincent, P., 2012, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **35**(8), 6, pp. 1798–1828.

[2] Petajan, E. D., 1986, "Automatic lipreading to enhance speech recognition (speech reading) — Guide books," PhD thesis, University of Illinois at Urbana-Champaign, Champaign.

[3] Huang, J., Liu, Z., Wang, Y., Chen, Y., and Wong, E., 2008, "Integration of multimodal features for video scene classification based on HMM," pp. 53–58.

[4] Singhal, A., and Brown, C. R., 1997, "Dynamic Bayes net approach to multimodal sensor fusion," *https://doi.org/10.1117/12.287628,* **3209**(22), 9, pp. 2–10.

[5] Chibelushi, C. C., Deravi, F., and Mason, J. S., 2002, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia,* **4**(1), 3, pp. 23–37.

[6] Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., and Parikh, D., 2016, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," *International Journal of Computer Vision,* **127**(4), 12, pp. 398–414.

[7] Johnson, J., Fei-Fei, L., Hariharan, B., Zitnick, C. L., Van Der Maaten, L., and Girshick, R., 2016, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017,* **2017-Janua**, 12, pp. 1988–1997.

[8] Xu, H., and Saenko, K., 2015, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **9911 LNCS**, 11, pp. 451–466.

[9] Shih, K. J., Singh, S., and Hoiem, D., 2015, "Where To Look: Focus Regions for Visual Question Answering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 11, pp. 4613–4621.

[10] Kiros, R., Salakhutdinov, R., and Zemel, R. S., 2014, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," *undefined.*

[11] Tolstikhin, I., Bousquet, O., Schölkopf, B., Thierbach, K., Bazin, P. L., de Back, W., Gavriilidis, F., Kirilina, E., Jäger, C., Morawski, M., Geyer, S., Weiskopf, N., Scherf, N., Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., Musk, E., Neuralink, Hjortsø, M. A., Wolenski, P., Ruder, S., Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., Duvenaud, D., and Doersch, C., 2014, "Generative Adversarial Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **11046 LNCS**(NeurIPS), 6, pp. 1–9.

[12] Dhariwal, P., and Nichol, A., 2021, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems,* **11**, 5, pp. 8780–8794.

[13] Nobari, A. H., Chen, W., and Ahmed, F., 2021, "Range-GAN: Range-Constrained Generative Adversarial Network for Conditioned Design Synthesis," *Proceedings of the ASME Design Engineering Technical Conference,* **3B-2021**, 3.

[14] Nobari, A. H., Chen, W., and Ahmed, F., 2021, "PcDGAN: A Continuous Conditional Diverse Generative Adversarial Network For Inverse Design," In 27th ACM SIGKDD Conference on Knowledge Discovery & Data, ACM, pp. 610–616.

[15] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I., 2021, "Zero-Shot Text-to-Image Generation,".

[16] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., 2021, "High-Resolution Image Synthesis with Latent Diffusion Models," pp. 10674–10685.

[17] Zhu, Q., Zhang, X., and Luo, J., 2023, "Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers," *Journal of Mechanical Design,* **145**(4), 4.

[18] Zhu, Q., and Luo, J., 2023, "Generative Transformers for Design Concept Generation," *Journal of Computing and Information Science in Engineering,* **23**(4), 8, pp. 1–61.

[19] Luo, J., Sarica, S., and Wood, K. L., 2021, "Guiding data-driven design ideation by knowledge distance," *Knowledge-Based Systems,* **218**, 4, p. 106873.

[20] Song, B., Associate, P., Miller, S., and Ahmed, F., 2023, "ATTENTION-ENHANCED MULTIMODAL LEARNING FOR CONCEPTUAL DESIGN EVALUATIONS," *Journal of Mechanical Design*, 1, pp. 1–38.

[21] Song, B., McComb, C., and Ahmed, F., 2022, "Assessing Machine Learnability of Image and Graph Representations for Drone Performance Prediction," *Proceedings of the Design Society,* **2**, 5, pp. 1777–1786.

[22] Gero, J. S., 1990, "Design Prototypes: A Knowledge Representation Schema for Design," *AI Magazine,* **11**(4), 12, pp. 26–26.

[23] Tseng, W. S., and Ball, L. J., 2011, "How Uncertainty Helps Sketch Interpretation in a Design Task," *Design Creativity 2010*, pp. 257–264.

[24] Häggman, A., Tsai, G., Elsen, C., Honda, T., and Yang, M. C., 2015, "Connections Between the Design Tool, Design Attributes, and User Preferences in Early Stage Design," *Journal of Mechanical Design,* **137**(7), 7.

[25] Tsai, G., and Yang, M. C., 2017, "How It Is Made Matters: Distinguishing Traits of Designs Created by Sketches, Prototypes, and CAD,".

[26] Purcell, A. T., and Gero, J. S., 1998, "Drawings and the design process: A review of protocol studies in design and other disciplines and related research in cognitive psychology," *Design Studies,* **19**(4), 10, pp. 389–430.

[27] Ullman, D. G., Wood, S., and Craig, D., 1990, "The importance of drawing in the mechanical design process," *Computers and Graphics,* **14**(2), pp. 263–274.

[28] Chang, Y. S., Chien, Y. H., Lin, H. C., Chen, M. Y., and Hsieh, H. H., 2016, "Effects of 3D CAD applications on the design creativity of students with different representational abilities," *Computers in Human Behavior,* **65**, 12, pp. 107–113.

[29] Atilola, O., Tomko, M., and Linsey, J. S., 2016, "The effects of representation on idea generation and design fixation: A study comparing sketches and function trees," *Design Studies,* **42**, 1, pp. 110–136.

[30] Hannibal, C., Brown, A., and Knight, M., 2016, "An Assessment of the Effectiveness of Sketch Representations in Early Stage Digital Design," *http://dx.doi.org/10.1260/1478077053739667,* **3**(1), 11, pp. 107–125.

[31] Atilola, O., and Linsey, J., 2015, "Representing analogies to influence fixation and creativity: A study comparing computer-aided design, photographs, and sketches," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM,* **29**(2), 4, pp. 161–171.

[32] Reid, T. N., MacDonald, E. F., and Du, P., 2013, "Impact of Product Design Representation on Cus-

tomer Judgment," *Journal of Mechanical Design,* **135**(9), 9.

[33] Yang, M. C., 2005, "A study of prototypes, design activity, and design outcome," *Design Studies,* **26**(6), 11, pp. 649–669.

[34] McKoy, F. L., Vargas-Hernández, N., Summers, J. D., and Shah, J. J., 2020, "Influence of Design Representation on Effectiveness of Idea Generation," *Proceedings of the ASME Design Engineering Technical Conference,* **4**, 11, pp. 39–48.

[35] Grace, K., Maher, M. L., Fisher, D., and Brady, K., 2014, "Data-intensive evaluation of design creativity using novelty, value, and surprise," *International Journal of Design Creativity and Innovation,* **3**(3-4), pp. 125–147.

[36] Nomaguchi, Y., Kawahara, T., Shoda, K., and Fujita, K., 2019, "Assessing Concept Novelty Potential with Lexical and Distributional Word Similarity for Innovative Design," *Proceedings of the Design Society: International Conference on Engineering Design,* **1**(1), pp. 1413–1422.

[37] Wood, K., and Otto, K., 2001, Product Design: Techniques in Reverse Engineering and New Product Development.

[38] Ulrich, K. T., and Eppinger, S. D., 2000, *Product design and development* McGraw-Hill.

[39] Fiorineschi, L., Frillici, F. S., and Rotini, F., 2018, "Issues related to missing attributes in aposteriori novelty assessments," *Proceedings of International Design Conference, DESIGN,* **3**, pp. 1067–1078.

[40] Pahl, G., Beitz, W., Feldhusen, J., and Grote, K.-H. H., 2007, *Engineering Design: A Systematic Approach* Springer London, London.

[41] Veisz, D., Namouz, E. Z., Joshi, S., and Summers, J. D., 2012, "Computer-aided design versus sketching: An exploratory case study," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM,* **26**(3), 8, pp. 317–335.

[42] BABAPOUR, M., ORNAS, V. H. A., REXFELT, O., and RAHE, U., 2014, "Media and Representations in Product Design Education," In INTERNATIONAL CONFERENCE ON ENGINEERING AND PRODUCT DESIGN EDUCATION, E. Bohemia, A. Eger, W. Eggink, A. Kovacevic, B. Parkinson, and W. Wits, eds., pp. 42–47.

[43] Baltrusaitis, T., Ahuja, C., and Morency, L. P., 2019, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **41**(2), 2, pp. 423–443.

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

[45] Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. A., and Huo, Y., 2022, "Deep Multi-modal Fusion of Image and Non-image Data in Disease Diagnosis and Prognosis: A Review,".

[46] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., 2021, "Learning Transferable Visual Models From Natural Language Supervision,".

[47] Li, X., Wang, Y., and Sha, Z., 2022, "Deep-Learning Methods of Cross-Modal Tasks for Conceptual Design of Product Shapes: A Review," *Journal of Mechanical Design*, 12, pp. 1–31.

[48] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T., 2013, "DeViSE: A Deep Visual-Semantic Embedding Model," *Advances in Neural Information Processing Systems,* **26**.

[49] Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B., 2015, "Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 10, pp. 171–178.

[50] Srivastava, N., and Salakhutdinov, R. R., 2012, "Multimodal Learning with Deep Boltzmann Machines," *Advances in Neural Information Processing Systems,* **25**.

[51] Duc Tuan, N. M., and Quang Nhat Minh, P., 2021, "Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection," *Proceedings - 2021 RIVF International Conference on Computing and Communication Technologies, RIVF 2021*, 4.

[52] 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 descriptions," In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 86267, American Society of Mechanical Engineers, p. V006T06A017.

[53] Yuan, C., Marion, T., and Moghaddam, M., 2022, "Leveraging End-User Data for Enhanced Design

Concept Evaluation: A Multimodal Deep Regression Model," *Journal of Mechanical Design,* **144**(2), 2, pp. 1–20.

[54] Andrew, G., Arora, R., Bilmes, J., and Livescu, K., 2013, Deep Canonical Correlation Analysis, 5.

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

[56] Feng, F., Wang, X., and Li, R., 2014, "Cross-modal retrieval with correspondence autoencoder," *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 11, pp. 7–16.

[57] Bachman, P., Hjelm, D., and Buchwalter, W., 2019, "Learning Representations by Maximizing Mutual Information Across Views," In NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 15535–15545.

[58] Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P., 2020, "Contrastive Learning of Medical Visual Representations from Paired Images and Text," *Proceedings of Machine Learning Research,* **182**, 10, pp. 1–24.

[59] Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L., 2013, Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 10.

[60] Karpathy, A., and Fei-Fei, L., 2014, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **39**(4), 12, pp. 664–676.

[61] Karpathy, A., Joulin, A., and Fei-Fei, L., 2014, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," *Advances in Neural Information Processing Systems,* **3**(January), 6, pp. 1889–1897.

[62] Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., and Ma, W. Y., 2019, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 6, pp. 6602–6611.

[63] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S., 2015, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," *International Journal of Computer Vision,* **123**(1), 5, pp. 74–93.

[64] Nguyen, D. K., and Okatani, T., 2018, "Multi-task Learning of Hierarchical Vision-Language Representation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 12, pp. 10484–10493.

[65] Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D., 2020, "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training," *Proceedings of the AAAI Conference on Artificial Intelligence,* **34**(07), 4, pp. 11336–11344.

[66] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J., 2019, "VL-BERT: Pre-training of Generic Visual-Linguistic Representations,".

[67] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W., 2019, "VisualBERT: A Simple and Performant Baseline for Vision and Language,".

[68] Alberti, C., Ling, J., Collins, M., and Reitter, D., 2019, "Fusion of Detected Objects in Text for Visual Question Answering," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 8, pp. 2131–2140.

[69] Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C., 2019, "VideoBERT: A Joint Model for Video and Language Representation Learning," *Proceedings of the IEEE International Conference on Computer Vision*, 4, pp. 7463–7472.

[70] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y., 2011, "Multimodal Deep Learning," In ICML 2011.

[71] Silberer, C., and Lapata, M., 2014, "Learning Grounded Meaning Representations with Autoencoders," *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference,* **1**, pp. 721–732.

[72] Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R., 2018, "Learning Factorized Multimodal Representations," *7th International Conference on Learning Representations, ICLR 2019*, 6.

[73] Socher, R., and Fei-Fei, L., 2010, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 966–973.

[74] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., and Murphy, K., 2015, "Generation and Comprehension of Unambiguous Object Descriptions," *Proceedings of the IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition, **2016-Decem**, 11, pp. 11–20.

[75] Tan, H., and Bansal, M., 2019, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 8, pp. 5100–5111.

[76] Lu, J., Batra, D., Parikh, D., and Lee, S., 2019, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," *Advances in Neural Information Processing Systems,* **32**, 8.

[77] Pramanik, S., Agrawal, P., and Hussain, A., 2019, "OmniNet: A unified architecture for multi-modal multi-task learning,".

[78] Sbrolli, C., Cudrano, P., Frosi, M., and Matteucci, M., 2022, "IC3D: Image-Conditioned 3D Diffusion for Shape Generation,".

[79] Bach, F. R., Lanckriet, G. R., and Jordan, M. I., 2004, "Multiple kernel learning, conic duality, and the SMO algorithm," *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 41–48.

[80] Lafferty, J., McCallum, A., and Pereira, F. C. N., 2001, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning,* **8**(June), pp. 282–289.

[81] Garg, A., Pavlović, V., and Rehg, J. M., 2003, "Boosted learning in dynamic bayesian networks for multimodal speaker detection," *Proceedings of the IEEE,* **91**(9), pp. 1355–1369.

[82] Ghahramani, Z., and Jordan, M. I., 1997, "Factorial Hidden Markov Models," *Machine Learning 1997 29:2,* **29**(2), pp. 245–273.

[83] Poria, S., Cambria, E., and Gelbukh, A., 2015, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis," *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 2539–2544.

[84] Gehler, P., and Nowozin, S., 2009, "On Feature Combination for Multiclass Object Classification," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 221–228.

[85] Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L. P., 2016, "Deep multimodal fusion for persuasiveness prediction," *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 10, pp. 284–288.

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

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

[88] Shutova, E., Kiela, D., and Maillard, J., 2016, "Black Holes and White Rabbits: Metaphor Identification with Visual Features," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 160–170.

[89] Cao, Y., Long, M., Wang, J., Yang, Q., and Yuy, P. S., 2016, "Deep visual-semantic hashing for cross-modal retrieval," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* **13-17-Augu**, 8, pp. 1445–1454.

[90] Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., and Bartlett, M., 2013, "Multiple kernel learning for emotion recognition in the wild," *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pp. 517–524.

[91] Morvant, E., Habrard, A., and Ayache, S., 2014, "Majority Vote of Diverse Classifiers for Late Fusion," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **8621 LNCS**, 4, pp. 153–162.

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

[93] Zhou, T., Thung, K. H., Zhu, X., and Shen, D., 2019, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Human brain mapping,* **40**(3), 2, pp. 1001–1016.

[94] Zoph, B., and Le, Q. V., 2016, "Neural Architecture Search with Reinforcement Learning," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*,

11.

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

[96] Zadeh, A., Chen, M., Cambria, E., Poria, S., and Morency, L. P., 2017, “Tensor Fusion Network for Multimodal Sentiment Analysis,” *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 7, pp. 1103–1114.

[97] Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., and Mahmood, F., 2019, “Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis,” *IEEE Transactions on Medical Imaging,* **41**(4), 12, pp. 757–770.

[98] Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., and Zhang, B.-T., 2022, Hadamard Product for Low-rank Bilinear Pooling, 7.

[99] Yu, Z., Yu, J., Fan, J., and Tao, D., 2017, “Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering,” *Proceedings of the IEEE International Conference on Computer Vision,* **2017-Octob**, 8, pp. 1839–1848.

[100] Yu, Z., Yu, J., Xiang, C., Fan, J., and Tao, D., 2017, “Beyond Bilinear: Generalized Multimodal Factorized High-order Pooling for Visual Question Answering,” *IEEE Transactions on Neural Networks and Learning Systems,* **29**(12), 8, pp. 5947–5959.

[101] Gao, Y., Beijbom, O., Zhang, N., and Darrell, T., 2015, “Compact Bilinear Pooling,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 11, pp. 317–326.

[102] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M., 2016, “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding,” *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 6, pp. 457–468.

[103] Ben-Younes, H., Cadene, R., Cord, M., and Thome, N., 2017, “MUTAN: Multimodal Tucker Fusion for Visual Question Answering,” *Proceedings of the IEEE International Conference on Computer Vision,* **2017-Octob**, 5, pp. 2631–2639.

[104] Tucker, L. R., 1966, “Some mathematical notes on three-mode factor analysis,” *Psychometrika 1966 31:3,* **31**(3), 9, pp. 279–311.

[105] Ben-Younes, H., Cadene, R., Thome, N., and Cord, M., 2019, “BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection,” *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 1, pp. 8102–8109.

[106] Jiang, S., Hu, J., Magee, C. L., and Luo, J., 2022, “Deep Learning for Technical Document Classification,” *IEEE Transactions on Engineering Management*.

[107] Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., and Rueckert, D., 2018, “Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer’s disease,” *Medical image analysis,* **48**, 8, pp. 117–130.

[108] Cao, M., Yang, M., Qin, C., Zhu, X., Chen, Y., Wang, J., and Liu, T., 2021, “Using DeepGCN to identify the autism spectrum disorder from multisite resting-state data,” *Biomedical Signal Processing and Control,* **70**, 9, p. 103015.

[109] Baltrusaitis, T., Ahuja, C., and Morency, L. P., 2017, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **41**(2), 5, pp. 423–443.

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

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

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

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

[114] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L., 2017, “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7, pp. 6077–

6086.

[115] Mansimov, E., Parisotto, E., Ba, J. L., and Salakhutdinov, R., 2015, "Generating Images from Captions with Attention," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 11.

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

[117] Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., and Gao, J., 2019, "Object-driven Text-to-Image Synthesis via Adversarial Training," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 2, pp. 12166–12174.

[118] Nam, H., Ha, J.-W., and Kim, J., 2017, "Dual Attention Networks for Multimodal Reasoning and Matching," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7, pp. 2156–2164.

[119] Lu, J., Yang, J., Batra, D., and Parikh, D., 2016, "Hierarchical Question-Image Co-Attention for Visual Question Answering," In NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 289–297.

[120] Osman, A., and Samek, W., 2018, "Dual Recurrent Attention Units for Visual Question Answering," *Computer Vision and Image Understanding,* **185**, 2, pp. 24–30.

[121] Schwartz, I., Schwing, A. G., and Hazan, T., 2017, "High-Order Attention Models for Visual Question Answering," *Advances in Neural Information Processing Systems,* **2017-Decem**, 11, pp. 3665–3675.

[122] Yang, Z., He, X., Gao, J., Deng, L., and Smola, A., 2015, "Stacked Attention Networks for Image Question Answering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 11, pp. 21–29.

[123] Fan, H., and Zhou, J., 2018, "Stacked Latent Attention for Multimodal Reasoning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 12, pp. 1072–1080.

[124] Xiong, C., Merity, S., and Socher, R., 2016, "Dynamic Memory Networks for Visual and Textual Question Answering," *33rd International Conference on Machine Learning, ICML 2016,* **5**, 3, pp. 3574–3583.

[125] Ren, S., He, K., Girshick, R., and Sun, J., 2015, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems,* **28**, 6.

[126] Lu, P., Li, H., Zhang, W., Wang, J., and Wang, X., 2017, "Co-attending Free-form Regions and Detections with Multi-modal Multiplicative Feature Embedding for Visual Question Answering," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 11, pp. 7218–7225.

[127] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M., 2022, data2vec: A general framework for self-supervised learning in speech, vision and language.

[128] Kim, J. H., Lee, S. W., Kwak, D., Heo, M. O., Kim, J., Ha, J. W., and Zhang, B. T., 2016, "Multimodal Residual Learning for Visual QA," *Advances in Neural Information Processing Systems*, 6, pp. 361–369.

[129] Arevalo, J., Solorio, T., Montes-Y-Gómez, M., and González, F. A., 2017, "Gated Multimodal Units for Information Fusion," *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2.

[130] Noh, H., Seo, P. H., and Han, B., 2015, "Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 11, pp. 30–38.

[131] Kong, C., Lin, D., Bansal, M., Urtasun, R., and Fidler, S., 2014, "What are you talking about? Text-to-image coreference," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9, pp. 3558–3565.

[132] Mirza, M., and Osindero, S., 2014, "Conditional Generative Adversarial Nets,".

[133] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H., 2016, "Generative Adversarial Text to Image Synthesis," *33rd International Conference on Machine Learning, ICML 2016,* **3**, 5, pp. 1681–1690.

[134] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D., 2016, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," pp. 5908–5916.

[135] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N., 2019, "StackGAN++: Realistic Image Synthesis with Stacked

Generative Adversarial Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **41**(08), 8, pp. 1947–1962.

[136] Zhu, M., Pan, P., Chen, W., and Yang, Y., 2019, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 4, pp. 5795–5803.

[137] Zhang, Z., Xie, Y., and Yang, L., 2018, "Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, pp. 6199–6208.

[138] Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., and Afzal, M. Z., 2017, "TAC-GAN - Text Conditioned Auxiliary Classifier Generative Adversarial Network," In Proc. CVPR.

[139] Cha, M., Gwon, Y. L., and Kung, H. T., 2018, "Adversarial Learning of Semantic Relevance in Text to Image Synthesis," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 12, pp. 3272–3279.

[140] Qiao, T., Zhang, J., Xu, D., and Tao, D., 2019, "MirrorGAN: Learning Text-to-image Generation by Redescription," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 3, pp. 1505–1514.

[141] Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., and Lee, H., 2016, "Learning What and Where to Draw," *Advances in Neural Information Processing Systems*, 10, pp. 217–225.

[142] Zhao, B., Meng, L., Yin, W., and Sigal, L., 2018, "Image Generation from Layout," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 11, pp. 8576–8585.

[143] Hinz, T., Heinrich, S., and Wermter, S., 2019, "Generating Multiple Objects at Spatially Distinct Locations," *7th International Conference on Learning Representations, ICLR 2019*, 1.

[144] Hong, S., Yang, D., Choi, J., and Lee, H., 2018, "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, pp. 7986–7994.

[145] Johnson, J., Gupta, A., and Fei-Fei, L., 2018, "Image Generation from Scene Graphs," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4, pp. 1219–1228.

[146] Sanghi, A., Chu, H., Lambourne, J. G., Wang, Y., Cheng, C.-Y., Fumero, M., and Malekshan, K. R., 2021, "CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation,".

[147] van den Oord DeepMind, A., Vinyals DeepMind, O., and Kavukcuoglu DeepMind, K., 2017, "Neural Discrete Representation Learning," *Advances in Neural Information Processing Systems,* **30**.

[148] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A., 2014, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 12.

[149] Shetty, R., Rohrbach, M., Hendricks, L. A., Fritz, M., and Schiele, B., 2017, "Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training," *Proceedings of the IEEE International Conference on Computer Vision,* **2017-Octob**, 3, pp. 4155–4164.

[150] Ajit, A., Acharya, K., and Samanta, A., 2020, "A Review of Convolutional Neural Networks," *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, 2.

[151] Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J., 2021, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, 6, pp. 1–21.

[152] Fathi, E., and Maleki Shoja, B., 2018, "Deep Neural Networks for Natural Language Processing," *Handbook of Statistics,* **38**, 1, pp. 229–316.

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

[154] Yagcioglu, S., Erdem, E., Erdem, A., and Çakici, R., 2015, "A Distributed Representation Based Query Expansion Approach for Image Captioning," *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference,* **2**, pp. 106–111.

[155] Cordonnier, J.-B., Loukas, A., and Jaggi, M., 2019, "On the Relationship between Self-Attention and Convolutional Layers,".

[156] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., 2020, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,".

[157] Wang, Y., Xu, J., and Sun, Y., 2022, "End-to-End Transformer Based Model for Image Captioning," *Proceedings of the AAAI Conference on Artificial Intelligence,* **36**(3), 3, pp. 2585–2594.

[158] Kalyan, K. S., Rajasekharan, A., and Sangeetha, S., 2021, "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing,".

[159] Qi, C. R., Su, H., Mo, K., and Guibas, L. J., 2016, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 12, pp. 77–85.

[160] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S., 2015, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," *32nd International Conference on Machine Learning, ICML 2015,* **3**, 3, pp. 2246–2255.

[161] Ho, J., Jain, A., and Abbeel, P., 2020, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems,* **2020-Decem**, 6.

[162] Song, J., Meng, C., and Ermon, S., 2020, "Denoising Diffusion Implicit Models,".

[163] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B., 2020, "Score-Based Generative Modeling through Stochastic Differential Equations,".

[164] Vahdat, A., Kreis, K., and Kautz, J., 2021, "Score-based Generative Modeling in Latent Space," *Advances in Neural Information Processing Systems,* **14**, 6, pp. 11287–11302.

[165] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M., 2021, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,".

[166] Kim, G., Kwon, T., and Ye, J. C., 2021, "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation,".

[167] Luo, S., and Hu, W., 2021, "Diffusion Probabilistic Models for 3D Point Cloud Generation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3, pp. 2836–2844.

[168] Zhou, L., Du, Y., and Wu, J., 2021, "3D Shape Generation and Completion through Point-Voxel Diffusion," *Proceedings of the IEEE International Conference on Computer Vision*, 4, pp. 5806–5815.

[169] Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., and Kreis, K., 2022, "LION: Latent Point Diffusion Models for 3D Shape Generation,".

[170] Liu, Z., Tang, H., Lin, Y., and Han, S., 2019, "Point-Voxel CNN for Efficient 3D Deep Learning," *Advances in Neural Information Processing Systems,* **32**, 7.

[171] Ho, J., and Salimans, T., 2022, "Classifier-Free Diffusion Guidance,".

[172] Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., and Chen, M., 2022, "Point-E: A System for Generating 3D Point Clouds from Complex Prompts,".

[173] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M., 2022, "Hierarchical Text-Conditional Image Generation with CLIP Latents,".

[174] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., 2014, "Show and Tell: A Neural Image Caption Generator," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **07-12-June**, 11, pp. 3156–3164.

[175] Rohrbach, A., Rohrbach, M., and Schiele, B., 2015, "The Long-Short Story of Movie Description," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **9358**, 6, pp. 209–221.

[176] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y., 2022, "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation,".

[177] Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., and Tang, J., 2021, "CogView: Mastering Text-to-Image Generation via Transformers," *Advances in Neural Information Processing Systems,* **24**, 5, pp. 19822–19835.

[178] Desai, K., and Johnson, J., 2020, "VirTex: Learning Visual Representations from Textual Annotations," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6, pp. 11157–11168.

[179] Bulent Sariyildiz, M., Perez, J., Larlus, D., Sariyildiz, M. B., Perez, J., and Larlus, D., 2020, "Learning Visual Representations with Caption Annotations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **12353 LNCS**, 8, pp. 153–170.

[180] Dinh, L., Sohl-Dickstein Google, J., Samy, B., and Google Brain, B., 2016, Density estimation using Real NVP, 7.

[181] Wei, Y., Vosselman, G., and Yang, M. Y., 2022, "Flow-based gan for 3d point cloud generation from a single image,".

[182] Chen, Z., and Zhang, H., 2018, "Learning Implicit Fields for Generative Shape Modeling," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 12, pp. 5932–5941.

[183] Liu, S., Saito, S., Chen, W., and Li, H., 2019, "Learning to Infer Implicit Surfaces without 3D Supervision," *Advances in Neural Information Processing Systems,* **32**, 11.

[184] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S., 2019, "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 1, pp. 165–174.

[185] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X., 2016, "Improved Techniques for Training GANs," *Advances in Neural Information Processing Systems*, 6, pp. 2234–2242.

[186] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., 2017, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Advances in Neural Information Processing Systems,* **2017-Decem**, 6, pp. 6627–6638.

[187] Odena, A., Olah, C., and Shlens, J., 2016, "Conditional Image Synthesis With Auxiliary Classifier GANs," *34th International Conference on Machine Learning, ICML 2017,* **6**, 10, pp. 4043–4055.

[188] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., 2004, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing,* **13**(4), 4, pp. 600–612.

[189] Li, B., Qi, X., Lukasiewicz, T., and Torr, P. H., 2019, "ManiGAN: Text-Guided Image Manipulation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 12, pp. 7877–7886.

[190] Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L., 2017, "Learning Representations and Generative Models for 3D Point Clouds," *35th International Conference on Machine Learning, ICML 2018,* **1**, 7, pp. 67–85.

[191] Shu, D., Park, S. W., and Kwon, J., 2019, "3D Point Cloud Generative Adversarial Network Based on Tree Structured Graph Convolutions," *Proceedings of the IEEE International Conference on Computer Vision,* **2019-Octob**, 5, pp. 3858–3867.

[192] Ibing, M., Lim, I., and Kobbelt, L., 2021, "3D Shape Generation with Grid-based Implicit Functions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7, pp. 13554–13563.

[193] Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C. D., and Ng, A. Y., 2013, "Zero-Shot Learning Through Cross-Modal Transfer," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1.

[194] Ba, J. L., Swersky, K., Fidler, S., and Salakhutdinov, R. "Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions,".

[195] Reed, S., Akata, Z., Lee, H., and Schiele, B., 2016, "Learning Deep Representations of Fine-grained Visual Descriptions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 5, pp. 49–58.

[196] Nakov, P., and Ng, H. T., 2009, Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages.

[197] Khapra, M. M., Kumaran, A., and Bhattacharyya, P., 2010, Everybody loves a rich cousin: An empirical study of transliteration through bridge languages.

[198] Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., and Darrell, T., 2015, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2016-Decem**, 11, pp. 1–10.

[199] Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y., 2014, "Grounded Compositional Semantics for Finding and Describing Images with Sentences," *Transactions of the Association for Computational Linguistics,* **2**, 12, pp. 207–218.

[200] Feng, Y., and Lapata, M., 2010, Visual Information in Semantic Representation.

[201] Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K., 2012, Distributional Semantics in Technicolor.

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

[203] Gupta, T., Schwing, A., and Hoiem, D., 2019, "ViCo: Word Embeddings from Visual Co-occurrences," *Proceedings of the IEEE International Conference on Computer Vision,* **2019-Octob**, 8, pp. 7424–7433.

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

[205] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L., 2016, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision,* **123**(1), 2, pp. 32–73.

[206] Mori, Y., Takahashi, H., and Oka, R., 1999, "Image-to-word transformation based on dividing,".

[207] Quattoni, A., Collins, M., and Darrell, T., 2007, "Learning visual representations using images with captions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[208] Joulin, A., van Der Maaten, L., Jabri, A., and Vasilache, N., 2015, "Learning Visual Features from Large Weakly Supervised Data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **9911 LNCS**, 11, pp. 67–84.

[209] Li, A., Jabri, A., Joulin, A., and Maaten, L. V. D., 2016, "Learning Visual N-Grams from Web Data," *Proceedings of the IEEE International Conference on Computer Vision,* **2017-Octob**, 12, pp. 4193–4202.

[210] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L., 2018, "Exploring the Limits of Weakly Supervised Pretraining," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **11206 LNCS**, 5, pp. 185–201.

[211] Kiela, D., Bulat, L., and Clark, S., 2015, "Grounding Semantics in Olfactory Perception," *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference,* **2**, pp. 231–236.

[212] Blum, A., and Mitchell, T., 1998, "Combining labeled and unlabeled data with co-training," *Proceedings of the Annual ACM Conference on Computational Learning Theory*, pp. 92–100.

[213] Levin, A., Viola, P., and Freund, Y., 2003, "Unsupervised improvement of visual detectors using co-training," *Proceedings of the IEEE International Conference on Computer Vision,* **1**, pp. 626–633.

[214] Christoudias, C. M., Urtasun, R., and Darrell, T., 2012, "Multi-View Learning in the Presence of View Disagreement,".

[215] Kojima, A., Tamura, T., and Fukunaga, K., 2002, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions," *International Journal of Computer Vision 2002 50:2,* **50**(2), pp. 171–184.

[216] Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., and Saenko, K., 2013, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2712–2719.

[217] Yang, Y., Teo, C., III, H. D., and Aloimonos, Y., 2011, Corpus-Guided Sentence Generation of Natural Images.

[218] Zitnick, C. L., and Parikh, D., 2013, "Bringing semantics into focus using visual abstraction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3009–3016.

[219] Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S. C., 2010, "I2T: Image parsing to text description," *Proceedings of the IEEE,* **98**(8), pp. 1485–1508.

[220] Girshick, R., 2015, "Fast R-CNN," In IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448.

[221] Cornia, M., Baraldi, L., and Cucchiara, R., 2022, "Explaining transformer-based image captioning models: An empirical analysis," *AI Communications,* **35**(2), 1, pp. 111–129.

[222] Herdade, S., Kappeler, A., Boakye, K., and Soares,

J., 2019, "Image Captioning: Transforming Objects into Words," *Advances in Neural Information Processing Systems,* **32**.

[223] Huang, L., Wang, W., Chen, J., and Wei, X.-Y., 2019, "Attention on Attention for Image Captioning,".

[224] He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., and Pugeault, N., 2020, "Image Captioning through Image Transformer," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **12625 LNCS**, 4, pp. 153–169.

[225] Li, G., Zhu, L., Liu, P., and Yang, Y., 2019, "Entangled transformer for image captioning," *Proceedings of the IEEE International Conference on Computer Vision,* **2019-Octob**, 10, pp. 8927–8936.

[226] Aneja, J., Deshpande, A., and Schwing, A. G., 2017, "Convolutional Image Captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 11, pp. 5561–5570.

[227] Deshpande, A., Aneja, J., Wang, L., Schwing, A. G., and Forsyth, D., 2018, "Fast, Diverse and Accurate Image Captioning Guided By Part-of-Speech," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 5, pp. 10687–10696.

[228] Li, B., Qi, X., Lukasiewicz, T., and Torr, P. H., 2019, "Controllable Text-to-Image Generation," *Advances in Neural Information Processing Systems,* **32**, 9.

[229] Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K., and Xu, C., 2020, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis,".

[230] Karras, T., Laine, S., and Aila, T., 2018, "A Style-Based Generator Architecture for Generative Adversarial Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **43**(12), 12, pp. 4217–4228.

[231] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D., 2021, "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," *Proceedings of the IEEE International Conference on Computer Vision*, 3, pp. 2065–2074.

[232] Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D., 2021, "StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators," *ACM Transactions on Graphics,* **41**(4), 8.

[233] Chefer, H., Benaim, S., Paiss, R., and Wolf, L., 2021, "Image-Based CLIP-Guided Essence Transfer,".

[234] Esser, P., Rombach, R., and Ommer, B., 2020, "Taming Transformers for High-Resolution Image Synthesis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 12, pp. 12868–12878.

[235] Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E., and Allen Hamilton, B., 2022, "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance,".

[236] Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y., 2021, "Vector-quantized Image Modeling with Improved VQGAN,".

[237] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M., 2022, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,".

[238] Frans, K., Soros, L. B., and Witkowski, O., 2021, "CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders,".

[239] Choy, C. B., Xu, D., Gwak, J. Y., Chen, K., and Savarese, S., 2016, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **9912 LNCS**, pp. 628–644.

[240] Gkioxari, G., Johnson, J., and Malik, J., 2019, "Mesh R-CNN," *Proceedings of the IEEE International Conference on Computer Vision,* **2019-Octob**, 6, pp. 9784–9794.

[241] Shrestha, R., Fan, Z., Su, Q., Dai, Z., Zhu, S., and Tan, P., 2020, "MeshMVS: Multi-View Stereo Guided Mesh Reconstruction," *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, 10, pp. 1290–1300.

[242] Fan, H., Su, H., and Guibas, L., 2016, "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017,* **2017-Janua**, 12, pp. 2463–2471.

[243] Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., and Aubry, M., 2018, "AtlasNet: A Papier-M\^ach\'e Approach to Learning 3D Surface Generation,".

[244] 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), 11.

[245] Wu, J., Zhang, C., Xue, T., Freeman, W. T., and Tenenbaum, J. B., 2016, "Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling," *Advances in Neural Information Processing Systems*, 10, pp. 82–90.

[246] Khan, S. H., Guo, Y., Hayat, M., and Barnes, N., 2019, "Unsupervised Primitive Discovery for Improved 3D Generative Modeling," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* **2019-June**, 6, pp. 9731–9740.

[247] Maron, H., Galun, M., Aigerman, N., Trope, M., Dym, N., Yumer, E., Kim, V. G., and Lipman, Y., 2017, "Convolutional neural networks on surfaces via seamless toric covers," *ACM Transactions on Graphics (TOG),* **36**(4), 7.

[248] Ben-Hamu, H., Maron, H., Kezurer, I., Avineri, G., and Lipman, Y., 2018, "Multi-chart Generative Surface Modeling," *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018*, 6.

[249] Saquil, Y., Xu, Q. C., Yang, Y. L., and Hall, P., 2020, "Rank3DGAN: Semantic mesh generation using relative attributes," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 5586–5594.

[250] Alhaija, H. A., Dirik, A., Knörig, A., Fidler, S., and Shugrina, M., 2022, "XDGAN: Multi-Modal 3D Shape Generation in 2D Space,".

[251] Fu, R., Zhan, X., Chen, Y., Ritchie, D., and Sridhar, S., 2022, "ShapeCrafter: A Recursive Text-Conditioned 3D Shape Generation Model,".

[252] Kalischek, N., Peters, T., Wegner, J. D., and Schindler, K., 2022, "Tetrahedral Diffusion Models for 3D Shape Generation,".

[253] Alwala, K. V., Gupta, A., and Tulsiani, S., 2022, "Pre-train, Self-train, Distill: A simple recipe for Supersizing 3D Reconstruction," pp. 3763–3772.

[254] Liu, Z., Dai, P., Li, R., Qi, X., and Fu, C.-W., 2022, "ISS: Image as Stepping Stone for Text-Guided 3D Shape Generation,".

[255] Nam, G., Khlifi, M., Rodriguez, A., Tono, A., Zhou, L., and Guerrero, P., 2022, "3D-LDM: Neural Implicit 3D Shape Generation with Latent Diffusion Models,".

[256] Cheng, Z., Chai, M., Ren, J., Lee, H.-Y., Olszewski, K., Huang, Z., Maji, S., and Tulyakov, S., 2022, "Cross-Modal 3D Shape Generation and Manipulation," pp. 303–321.

[257] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y. G., 2018, "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **11215 LNCS**, 4, pp. 55–71.

[258] Michel, O., Bar-On, R., Liu, R., Benaim, S., and Hanocka, R., 2021, "Text2Mesh: Text-Driven Neural Stylization for Meshes,".

[259] Jetchev, N., 2021, "ClipMatrix: Text-controlled Creation of 3D Textured Meshes,".

[260] Mai, S., Zeng, Y., Zheng, S., and Hu, H., 2022, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, pp. 1–1.

[261] Yuan, C., Marion, T., and Moghaddam, M., 2021, "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model," *Journal of Mechanical Design,* **144**(2), 09 021403.

[262] Zhou, Y., Ying, Q., Qian, Z., Li, S., and Zhang, X., 2022, Multimodal fake news detection via clip-guided learning.

[263] Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., and Liu, S., 2020, "A multimodal deep learning framework for predicting drug–drug interaction events," *Bioinformatics,* **36**(15), 05, pp. 4316–4322.

[264] Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., and Feng, L., 2021, "Deeptake: Prediction of driver takeover behavior using multimodal data," In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery.

[265] Ordonez, V., Kulkarni, G., and Berg, T. L., 2011, "Im2Text: Describing Images Using 1 Million Captioned Photographs," In NIPS 12.

[266] Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M., 2015, "Language Models for Image Captioning: The Quirks and What Works," *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference,* **2**, 5, pp. 100–105.

[267] Kwon, E., Huang, F., and Goucher-Lambert, K., 2022, "Enabling multi-modal search for inspirational design stimuli using deep learning," *AI EDAM,* **36**, 7, p. e22.

[268] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D., 2010, "Every picture tells a story: Generating sen-

tences from images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **6314 LNCS**(PART 4), pp. 15–29.

[269] Xu, R., Xiong, C., Chen, W., and Corso, J. J., 2015, "Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework," *Proceedings of the AAAI Conference on Artificial Intelligence,* **29**(1), 2, pp. 2346–2352.

[270] Hodosh, M., Young, P., and Hockenmaier, J., 2013, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *Journal of Artificial Intelligence Research,* **47**, 8, pp. 853–899.

[271] 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," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **8693 LNCS**(PART 5), 5, pp. 740–755.

[272] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J., 2015, "YFCC100M: The New Data in Multimedia Research," *Communications of the ACM,* **59**(2), 3, pp. 64–73.

[273] Sun, C., Shrivastava, A., Singh, S., and Gupta, A., 2017, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *Proceedings of the IEEE International Conference on Computer Vision,* **2017-Octob**, 7, pp. 843–852.

[274] Murray, N., Marchesotti, L., and Perronnin, F., 2012, "AVA: A large-scale database for aesthetic visual analysis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415.

[275] 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," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **11363 LNCS**, 12, pp. 100–116.