A Review on Methods and Applications in Multimodal Deep Learning

SUMMAIRA JABEEN and XI LI, College of Computer Science, Zhejiang University, China
MUHAMMAD SHOIB AMIN, School of Software Engineering, East China Normal University, China
OMAR Bourahla, Songyuan Li, and ABDUL JABBAR, College of Computer Science, Zhejiang University, China

Deep Learning has implemented a wide range of applications and has become increasingly popular in recent years. The goal of multimodal deep learning (MMDL) is to create models that can process and link information using various modalities. Despite the extensive development made for unimodal learning, it still cannot cover all the aspects of human learning. Multimodal learning helps to understand and analyze better when various senses are engaged in the processing of information. This article focuses on multiple types of modalities, i.e., image, video, text, audio, body gestures, facial expressions, physiological signals, flow, RGB, pose, depth, mesh, and point cloud. Detailed analysis of the baseline approaches and an in-depth study of recent advancements during the past five years (2017 to 2021) in multimodal deep learning applications has been provided. A fine-grained taxonomy of various multimodal deep learning methods is proposed, elaborating on different applications in more depth. Last, main issues are highlighted separately for each domain, along with their possible future research directions.

CCS Concepts: • Computing methodologies → Machine learning; • Information systems → Multimedia and multimodal retrieval;

Additional Key Words and Phrases: Deep learning, multimedia, multimodal learning, datasets, neural networks, survey

ACM Reference format:
Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. 2023. A Review on Methods and Applications in Multimodal Deep Learning. ACM Trans. Multimedia Comput. Commun. Appl. 19, 2s, Article 76 (February 2023), 41 pages.
https://doi.org/10.1145/3545572

This work is supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR19F020004, National Key Research and Development Program of China under Grant 2020AAA0107400, National Natural Science Foundation of China under Grant U20A20222, National Science Foundation for Distinguished Young Scholars under Grant 6225605, Ant Group, and sponsored by CAAI-HUAWEI MindSpore Open Fund. Authors’ addresses: S. Jabeen, O. Bourahla, S. Li, and A. Jabbar, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; emails: {11821129, bourahla, leizungjyun, Jabbar}@zju.edu.cn; X. Li (corresponding author), College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; Shanghai Institute for Advanced Study, Zhejiang University Shanghai 201203 China; Zhejiang-Singapore Innovation and AI Joint Research Lab, Shanghai 201203, China and Shanghai AI Laboratory, Shanghai 201203, China; email: xilizju@zju.edu.cn; M. S. Amin, School of Software Engineering, East China Normal University, 3663 North Zhongshan Road, Shanghai, China; email: 52184501030@stc.ecnu.edu.cn. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. © 2023 Association for Computing Machinery. 1551-6857/2023/02-ART76 $15.00 https://doi.org/10.1145/3545572

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 2s, Article 76. Publication date: February 2023.
1 INTRODUCTION

Multimodal learning proposes that we are able to remember and understand more when engaging multiple senses during the learning process. MMDL technically contains different aspects and challenges such as representation, translation, alignment, fusion, co-learning when learning from two or more modalities [33, 57]. This information from multiple sources is contextually related and occasionally provides the additional necessary information to one another, revealing features that would not be viewable when working with individual modalities. MMDL models combine heterogeneous data from multiple sources, allowing for more appropriate predictions [124]. Extracting and presenting relevant information from multimodal data remains an inspirational motive for MMDL research. Merging various modalities to optimize effectiveness is still an appealing challenge. Furthermore, the accuracy and flexibility of multimodal systems are not optimum due to the insufficiency of labeled data.

The recent advances and trends of MMDL are from Audio-visual speech recognition (AVSR) [124, 161], multimedia content indexing and retrieval [8, 13, 34, 127], understanding human multimodal behaviors during social interaction, multimodal emotion recognition [16, 25, 61, 171], image and video captioning [36, 78], Visual Question-Answering (VQA) [91], multimedia retrieval [128] to health analysis [155], and so on. In this article, we analyzed the latest MMDL models to propose typical models and techniques for advancing the field forward. Various modalities, i.e., image, video, text, audio, body gestures, facial expressions, physiological signals, flow, RGB, pose, depth, mesh, and point cloud, are focused. The main goal of MMDL is to construct a model that can process information from different modalities and relate it. MMDL methods and applications are categorized into multiple groups, and the impact of feature extractor, deep learning architecture, datasets, and evaluation metrics are analyzed for each group. Moreover, key features of each model are also discussed to highlight the main contribution of the work model.

1.1 Contribution and Relevance to Other Surveys

Recently, numerous surveys have been published relating to the topic of multimodal learning. A summarized list of these review articles is presented and analyzed in Table 1. A summary of modalities used and applications discussed is shown in the table. Our contributions to this article are described in Section 1.1.1. Most of the baseline surveys only focus on models using Image, Video, Text, and Audio modalities, while our survey article is focused on some additional models and groups of applications using body gestures, facial expressions, and physiological signals. Comparative analysis on experimental results based on feature extractors, architectures, standard datasets, and evaluation metrics are also presented to elaborate the performance of various models.

1.1.1 Our Contributions. All these literature surveys provide a review on only a specific domain of multimodal learning. Some authors only discussed methods and applications of representation learning and some on fusion learning. Others discussed representation and fusion learning methods together using few modalities such as vision or text. As mentioned earlier, Mogadala et al. [100] explained different models of MMDL using image, video, and text modalities. However, we have taken one step further: Along with these modalities, this article provides a technical review of various models using audio, body gestures, facial expressions, and physiological signals modalities. Our primary focus is distinctive in that we seek to survey the literature from up-to-date deep learning concepts using more modalities. The main contributions of our article are listed below:

- We propose a novel fine-grained taxonomy of various MMDL applications, which elaborates different groups of applications in more depth.
Table 1. Analysis of Baseline Literature Surveys

| Paper                  | Year | Pub. | I | T | A | V | BG | FE | PS | DM | F | RGB | P | D | M | Applications                                                                 |
|------------------------|------|------|---|---|---|---|----|----|----|----|---|-----|---|---|---|-----------------------------------------------------------------------------|
| Zhang et al. [162]     | 2020 | IEEE | ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Image Description, Image-to-Text generation, VQA                            |
| Bisk et al. [15]       | 2020 | arXiv| ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Not Available                                                               |
| Gao et al. [41]        | 2020 | NC(MIT)| ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Caption Generation, Storytelling, QA, Dialogue, Reasoning, Referring, Expression, Entailment, Visual Generations, Navigation, Machine Translation |
| Mogadala et al. [100]  | 2019 | arXiv| ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Multimodal matching, Multimodal classification, Multimodal interaction, Multimedia content indexing, Multimedia Content retrieval. |
| Zhang et al. [163]     | 2019 | IEEE | ✓ | ✓ | ⬜ | ✓ |    |    |    |    |   |     |   |   |   | Multimedia description                                                    |
| Guo et al. [48]        | 2019 | IEEE | ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Video Classification, Event Detection, VQA, Text-to-image synthesis, Transfer Loading, Speech recognition, Media Description, Multimedia Retrieval |
| Baltrusaitis et al. [11]| 2019 | IEEE | ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Cross Media Retrieval, NLP, Video Analysis, Recommended System              |
| Li et al. [84]         | 2019 | IEEE | ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   |   | Human Activity Recognition, Medical Applications, Autonomous Systems       |
| Ramachandram et al. [118]| 2017| IEEE | ✓ | ✓ | ✓ | ✓ |    |    |    |    |   |     |   |   | ✓ | Multimodal Image Description, Multimodal Video Description, Multimodal VQA, Multimodal Speech Synthesis, Multimodal Emotion Recognition, Multimodal Event Detection, Multimodal Action Recognition, Multimodal Cross Modal Retrieval, Multimodal Natural Sound Synthesis, Multimodal Talking Face Generation |
| Ours                   | –    | –    | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |     |   |   |   | Multimodal Image Description, Multimodal Video Description, Multimodal VQA, Multimodal Speech Synthesis, Multimodal Emotion Recognition, Multimodal Event Detection, Multimodal Action Recognition, Multimodal Cross Modal Retrieval, Multimodal Natural Sound Synthesis, Multimodal Talking Face Generation |

where I = Image, T = Text, A = Audio, V = Video, BG = Body Gesture, FE = Facial Expression, PS = Physiological Signals, DM = DNA & MRI, F = Flow, P = Pose, D = Depth, M = Mesh.

- We provide an explicit overview of various models using the modalities of image, video, text, audio, body gestures, facial expressions, physiological signals, flow, RGB, pose, depth, mesh, and point cloud.
- We provide a comparative analysis of the recent baseline MMDL models with the perspective of architecture, multimedia, dataset, and evaluation metric used along with model features. Comparative analysis helps researchers to get directions in future research to improve domain performance using a specific or combination of feature extractors and architectures.
- We provide quantitative experimental results of MMDL models on different benchmark datasets.
- We highlight the issues and limitations of MMDL models and discuss their possible future research directions.

### 1.2 Structure of Survey

The rest of the article is structured as follows: Some background of MMDL is discussed in Section 2. Section 3 presents various MMDL methods and applications. MMDL datasets and evaluation metrics are discussed in Section 4. Comparative analysis on experimental results of MMDL application is discussed in Section 5. In Section 6, we provide a brief discussion and propose future research directions in this active area. Section 7 concludes the survey article. Due to page limits, details of related MMDL architecture are provided in Section 1 of supplementary materials. The structure of the article is shown in Figure 1.

### 2 BACKGROUND

A lot of ML approaches have been developed over the past decade to deal with multimodal data. Multimodal machine learning leads to a wide range of applications: from Audio-visual speech recognition, multimedia content indexing and retrieval, understanding human multimodal behaviors, emotion recognition, multimodal affect recognition, image and video captioning, VQA, multimedia retrieval, to health analysis, and so on. A list of abbreviations is presented in Appendix A.

\(^1\)Section 1 in supplementary materials contains the complete details about architectures used in MMDL methods.
Multimodal applications came into existence in the 1970s and were categorized into four eras [101]. This section discusses the core challenges of multimodal learning, and the description of four eras is explained further. In this article, we concentrate mainly on the deep learning era for multimodal applications.

### 2.1 Challenges of Multimodal Learning

The research domain of MMDL presents some specific challenges for researchers due to heterogeneous data. Learning from multiple sources allows recording correspondences among various modalities and an in-depth comprehension of natural phenomena. In multimodal learning, extracting and combining features from multiple sources contributes to making larger-scale predictions. Data from various sources are contextually linked and occasionally give supplementary information to one another, revealing patterns that would not be noticeable when functioning with individual modalities separately. Such systems combine disparate, heterogeneous data from multiple sensors, allowing for more accurate predictions. Five main technical challenges of multimodal deep learning [11] are summarized in this section, which are listed below.

- **Multimodal Representation**: The process of representing information from several media in a tensor or vector form is known as multimodal representation (MMR). As information from several media frequently contains both redundant and complementary data, the goal is to represent data in a meaningful and efficient manner. Dealing with missing data, various levels of noise, and data combinations from several media are all challenges linked with MMR. Joint and Coordinated representations are used to deal with these representation challenges.

- **Multimodal Translation**: The problem of translating or mapping information from one modality to another is addressed by multimodal translation (MMT). Consider the tasks such as producing an image from a caption or producing a caption for an image. Due to data heterogeneity, it is not possible to create one perfect description of an image. One of the most challenging aspects of multimodal translation is assessing translation quality. The results quality for speech synthesis and video or image description is highly subjective, and there is frequently no single correct translation.

- **Multimodal Alignment**: The process of developing correspondence among information from two or more media for the same event is known as multimodal alignment (MMA). A
framework must measure similarities among different media and cope with long-term dependencies to align them. Other challenges in the MMA task include the formation of better similarity metrics, the lack of annotated datasets, and various correct alignments.

- **Multimodal Fusion**: The process of combining data from two or more media to perform regression or classification is known as **multimodal fusion** (MMF). The huge volume of data and improved processing capacity enhance MMF approaches. The main challenge with MMF approaches is to deal with redundant data from this huge volume of extracted data. Another key challenge is determining how to properly examine mutually enriching and complementary information from many media. MMF systems function similarly to the brain in that they integrate numerous sources of data for meaningful interpretation and subsequent decision-making. Sentiment analysis is one of the commonly known examples of multimodal fusion; three modalities (Visual, acoustic, and language) combine to predict sentiment.

- **Multimodal Co-learning**: The process of transmitting information/knowledge among modalities is addressed by **multimodal co-learning** (MMC). Transmitting information/knowledge from a resource-rich modality is very effective for creating a model in a limited resource modality with noisy inputs, a lack of annotated data, and unreliable labels. Parallel, non-parallel, and hybrid MMC approaches are used to transfer knowledge between modalities. In parallel MMC, media shares the set of instances parallelly. Co-training and transfer learning are popular processes used for parallel co-learning approaches. Non-parallel MMC is used to share concepts or categories, while sharing media instances is not required. Conceptual grounding, **Zero-Shot Learning (ZSL)**, and transfer learning are processes used for non-parallel co-learning approaches. A shared dataset or media connects two non-parallel media in a hybrid data environment. This connection of media is used to share the representations of one media to another. The MMC is addressed by various ZSL, MMR learning, co-training, and conceptual learning approaches. MMC covers a wide range of applications, including semantic similarity estimation, AVSR, visual categorization, action identification, and so on.

### 2.2 Eras of Multimodal Learning

Prior research on multimodal learning can be categorized into four eras, listed below.

- **Behavioral Era (BE)**: The BE starts from the 1970s to the late 1980s. During this era, A Lazarus [77] proposed Multimodal behavior therapy based on seven different personality dimensions using inter-related modalities. Mulligan and Shaw [102] presents multimodal signal detection, wherein the signal detection task, the pooling of information is examined from other modalities, i.e., auditory and visual stimuli. Bahrick [10] found that infants detected bimodal temporal structure specifying object elasticity and rigidity and temporal synchrony between sights and sound of object. Hoffman-Plotkin [56] proposed a “Multimodal assessment of behavioral and cognitive deficits in abused and neglected pre-schoolers.”

- **Computational Era (CE)**: The CE starts from the late 1980s to 2000. During this era, Petajan [115] proposed “Automatic Lipreading to Enhance Speech Recognition.” Yuhas et al. [161] increase the performance of automatic speech recognition system even in noisy environments by using **neural networks (NN)**. The inspiration for these approaches was McGurk effect [96]—an interaction between vision and hearing during understanding of speech. **Hidden Markov Model (HMM)** [70] has become so popular in speech recognition because of the inherent statistical framework, the ease and accessibility of training algorithms to estimate model parameters from finite speech datasets, flexibility of the resulting recognition
system, in which the size, type, or architecture of the models can be easily changed to suit specific words, sounds.

- **Interaction Era (IE):** The IE starts from 2000 to 2010. During this era, the understanding of human multimodal behaviors is achieved during social interactions. One of the first milestones is the AMI Meeting Corpus [23] (a multimodal dataset) with over 100 hours of meeting recordings, all thoroughly annotated and transcribed. AMI seeks to create a repository of meetings and to evaluate conversations. CHIL project motive [136] uses ML methods to extract nonverbal human behaviors automatically. Under CHIL, different people-tracking systems were built using audio, video, or both modalities. **CALO Meeting Assistant (CALO-MA) [134]** real-time meeting recognition system takes the speech models. CALO-MA architecture includes real-time and offline speech transcription, action item recognition, question-answer pair identification, decision extraction, and summarization. **Social Signal Processing (SSP) [110]** aims at understanding and modelling social interactions and offering similar capabilities to computers in human-computer interaction scenarios.

- **Deep Learning Era (DLE):** The DLE starts in 2010 to date. This deep learning era is the main focus of our review article. We comprehensively discuss different groups of methods and applications proposed during this era in Section 3, and datasets and evaluation metrics in Section 2 of supplementary materials proposed during this era.

## 3 MMDL METHODS AND APPLICATIONS

Various methods and applications are designed using multimodal deep learning techniques. In this article, these methods and applications are grouped with relevance and dominance across multiple research areas. The taxonomy diagram of these applications is presented in the Figure 2.

### 3.1 Multimodal Image Description (MMID)

Image Description is mostly used to generate a textual description of visual contents provided through input image. During the deep learning era, two different fields are merged to perform
image descriptions, i.e., CV and NLP. In this process, two main kinds of modalities are used, i.e., image and text. The image description’s general structure diagram is shown in Figure 3. MMID faces multimodal representation, translation, alignment, and co-learning challenges, which are discussed in Section 2.1. MMID faces multimodal representation challenges due to representing aggregate data from disparate resources, managing missing information, and dealing with varying degrees of noise while representing many modalities. The capability to describe information in a meaningful manner is critical in multimodal situations and serves as the foundation of any model. In the MMID translation challenge, there are multiple ways to provide the correct description of an image, yet there may not be a single ideal translation or description. Evaluating better translation is a subjective task, which is evaluated by a team of human experts on certain parameters. In the MMID alignment challenge, identifying correspondences and relationships between components of occurrences from many modalities is a challenging task. It is not easy to find correspondence between description parts and image areas.

Image description frameworks are categorized into Retrieval-based, Template-based, and DL-based image descriptions. Retrieval and Template-based image descriptions are two of the earliest techniques for describing visual contents from images. In this article, DL-based image description techniques are explained in detail, which are further categorized into encoder-decoder-based, semantic concept-based, and attention-based image descriptions. Different image description approaches are analyzed comparatively according to architectures, multimedia, publication year, datasets, and evaluation metrics in Table 2.

3.1.1 Encoder-Decoder-based Image Description (EDID). EDID plays a vital role in image-captioning tasks using DL architectures. CNN architectures are used mainly as encoder parts to extract and encode data from images, and RNN architectures are used as decoder part to decode and generate captions. Wu and Hu [149] proposed a Cascade Recurrent Neural Network (CRNN) for image description. For the learning of visual language interactions, a cascade network is adopted by CRNN from the forward and backward directions. Chen et al. [27] proposed referenced-based LSTM model for image description task. In this model, training images are used as reference for proposed framework, to minimize the misrecognitions for the description task.
Jiang et al. [67] proposed a recurrent fusion network for the task of image-captioning on the basis of encoder-decoder. In this network, CNN architecture is used to extract information from input image, and RNN architecture is used to generate the description in the form of text. Guo et al. [47] proposed a multi-style image-captioning framework using CNN, GAN, LSTM, and GRU architectures. In this framework, five different captioning styles are introduced for the image: romantic, negative, positive, factual and humorous styles. He et al. [54] proposed an image caption generation framework with the guidance of Part of Speech (PoS). In the word generation part, PoS tags are fed into LSTM as guidance to generate more effective image captions. Feng et al. [39] proposed an unsupervised image-captioning framework. In this model, first attempt is made to do captioning of the image without any labeled image-sentence pairs. Ji et al. [65] introduced a globally enhanced transformation network for encoder and decoder. In this network, the encoder is used to extract global features from the image, and the global adaptive controller at the decoder side is used for controlled image description.

### 3.1.2 Semantic Concept-based Image Description (SCID)
A collection of semantic concepts extracted from the image are selectively addressed by SCID approaches. These concepts are extracted at the encoding stage along with other features of an image, then merged into hidden states of language models, and output is used to generate descriptions of images based on semantic concepts. Wang et al. [142] proposed an attribute-based image caption generation framework. Visual features are extracted by using salient semantic attributes and are passed as input to LSTM’s encoder. Zhang et al. [165] proposed a semantic guided visual attention mechanism-based image captioning model. Fully Convolutional Network (FCN) is primarily intended for semantic segmentation, especially for dense pixel-level feature extractions and semantic labeling in the form of spatial grid. Cao et al. [22] proposed a semantic-based model for image description. In this model, semantic attention-based guidance is used for LSTM architecture to produce a description of an image. Cheng et al. [30] proposed a multi-stage visual semantic attention mechanism-based image description model. In this approach, top-down and bottom-up attention modules are combined to control the visual and semantic-level information for producing fine-grained descriptions of the image.
Chen et al. [26] proposed a model to improve the accuracy of image captioning by introducing **Verb-specific Semantic Roles (VSR)**. This model targets the activity and entities roles involved in that particular action to extract and generate the most specific information from image.

### 3.1.3 Attention-based Image Description (AID)

AID plays a vital role, because it helps the image description process by focusing on distinct regions of the image according to their context. In recent years, various techniques have been proposed to better describe an image by applying an attention mechanism. Some of these attention mechanism-based image descriptions techniques are: Li et al. [81] proposed a new framework for describing images by using local and global attention mechanisms. Selective object-level features are combined with image-level features according to the context using local and global attention mechanisms. Anderson et al. [4] proposed bottom-up and top-down attention-based framework for image description to encourage deeper image understanding and reasoning. Liu et al. proposed a dual attention mechanism-based framework to describe an image for Chinese [87] and English [88] languages. The textual attention mechanism is used to improve the data credibility, and the visual attention mechanism is used to a deep understanding of image features. Wang et al. [139] proposed an E2E-DL approach for image description using a semantic attention mechanism. In this approach, features are extracted from specific image regions using an attention mechanism for producing corresponding descriptions. Wei et al. [147] proposed an image description framework by using multi-attention mechanism to extract local and non-local feature representations. Jiang et al. [66] proposed a multi-gate expansion of self-attention mechanism. In this network, attention mechanism is expanded by adding self-gated module and attention weight gate module to eliminate the irrelevant information from description.

### 3.2 Multimodal Video Description (MMVD)

Like image description, video description is used to generate a textual description of visual contents provided through input video. Here, DL approaches for the description of visual contents from videos are discussed in detail. Advancements in this field open up many opportunities in various application domains. During this process, mainly two types of modalities are used, i.e., video stream and text. MMVD faces multimodal representation, translation, alignment, fusion, and co-learning challenges due to the heterogeneous nature of data, redundancy of data in various video frames, and distinct level of noise. Videos translation with the perspective of verbs and nouns is still challenging to produce a compact and comprehensive description of visual frames. In multimodal alignment challenges, it is not easy to incorporate the alignment of verbs according to actions presented in scenes of a frame. We can enhance the quality of MMVD approaches by incorporating sound/speech features from a given visual source. We need to take care of all multimodal challenges while describing long videos with multiple sentences. The general structure diagram of the video description is shown in Figures 4(a) and (b). In this research, the video description approach is categorized based on the following architectural combinations for visual feature extraction and text generation. These approaches are comparatively analyzed according to architectures, multimedia, datasets, and evaluation metrics in Table 3.

#### 3.2.1 CNN-RNN Architectures

Most broadly used architecture combination in the domain of video description is CNN-RNN. Figure 4(a) presents the general view of the video description process by using CNN-RNN architectures, where at the visual extraction (encoder) stage, variants of CNN architectures are used, and at the sentence generation (decoder) stage, variants of RNN architectures are used. During deep learning era, several authors proposed techniques for describing videos that are based on this encoder, decoder combination. Krishna et al. [74] proposed a video description technique using action/event detection by applying a dense captioning mechanism. This is the first framework to detect and describe several events, but it did not significantly
improve video captioning. Wang et al. [138] proposed a reconstruction network for video description using an encoder-decoder reconstructor architecture, which utilizes both forward flow (from video to sentence) and backward flow (from sentence to video). Pei et al. [113] proposed an attention mechanism-based encoder-decoder framework for video description. An additional memory-based decoder is used to enhance the quality of video description. Aafaq et al. [1] proposed video captioning and capitalized on Spatio-temporal dynamics of videos to extract high-level semantics using 2D and 3D CNNs hierarchically, and GRU is used for the text generation part. Liu et al. [89] proposed SibNet, a sibling convolutional network for video description. Two architectures are used simultaneously to encode video, i.e., the content-branch to encode visual features and the semantic-branch to encode semantic features. Perez-Martin et al. [114] improve the visual captioning quality by implementing visual syntactic embedding. A PoS tagging structure is used to extract the syntactic representations from video and guide the decoder with these temporal compositions to achieve accurate descriptions.

3.2.2 RNN-RNN Architectures. During the DL era, RNN-RNN is also a popular architectural combination, because many authors contribute a lot by proposing various methods using this combination. Authors extract the visual content of the video by using RNN architectures instead of CNN. Figure 4(b) presents the general view of the video description process by using RNN-RNN architectures. Both visual extraction (encoder) and sentence generation (decoder) stage variants of RNN architectures are used. Rahman et al. [117] proposed a video-captioning framework that modifies the generated context using spatial hard pull and stacked attention mechanisms. This approach illustrates that mounting an attention layer for a multi-layer encoder will result in a more semantically correct description. Fang et al. [38] proposed a framework to generate commonsense captions of the input video. Commonsense description seeks to identify and describe the latent aspects of video such as effects, attributes, and intentions. A new dataset, Video-to-Commonsense (V2C), is also proposed for this framework. Zhang et al. [166] proposed an encoder-decoder-based framework for dense video captioning. A graph-based summarization and partition modules are used to enhance the word relationship between context and event found in video.
3.2.3 Deep Reinforcement Learning (DRL) Architectures. DRL is a learning mechanism where machines can learn intelligently from actions like human beings can learn from their experiences. In it, an agent is penalized or rewarded based on actions that bring the model closer to the target outcome. The general structure diagram of the video description DRL is shown in Figure 5(a). The main contributions of authors using DRL architectures are: Wang et al. [143] proposed a hierarchical-based reinforcement learning (HRL) model for describing a video. In this framework, a high-level module is designed to sub-goals and low-level workshop module recognizes actions to fulfill these goals. Chen et al. [29] proposed a framework based on RL for choosing informative frames from an input video. Fewer frames are required to generate video description in this approach. Li and Gong [80] proposed an E2E multitask RL framework for video description. The proposed method combines RL with attribute prediction during the training process, which results in improved video description generation. Mun et al. [103] proposed a framework where an event sequence generation network is used to monitor the series of events for generated captions from the video. Zhang et al. [164] proposed a reconstruction network for a description of visual contents, which operates on both forward flow (from video to sentence) and backward flow (from sentence to video). Xu et al. [153] proposed a polishing network that utilizes the RL technique to refine the generated captions. This framework consists of word denoising and grammar-checking networks for fine-tuning generated sentences. Wei et al. [146] proposed a framework for better exploration of RL events to generate more accurate and detailed video captions.

3.3 Multimodal Visual Question Answering (MMVQA)

VQA is an emerging technique that has piqued the interest of both the CV and NLP groups. It is a field of research about creating an AI System capable of answering natural language questions. Extracted features from input image/video and question are processed and combined to answer the question about the image, as presented in Figure 5(b). MMVQA faces multimodal representation, alignment, fusion, and co-learning challenges. In the multimodal representation challenge, capturing shared semantics from visual or text inputs is a challenging task. Implicit alignment is employed as an intermediary step to enhance the performance of the VQA task. The selection of feature extractors, such as ResNet, VGGNet, Faster RCNN, GoogleLeNet, or word embeddings, plays a vital role in extracting features from images or text. Model faces presented challenges
Co-attention links visual areas and words for sentence grounding, resulting in synchronization among distinct media for VQA tasks.

VQA is more complex as compared to other vision and language functions, such as text-to-image retrieval, video captioning, image captioning, and so on, because: (1) Questions asked in VQA are not specific or predetermined. (2) Visual information in VQA is at a high degree of dimensionality. Usually, VQA required a more thorough and detailed understanding of an image/video. (3) VQA solves several CV sub-tasks. Many authors contribute to the field of VQA by using various DL techniques. These methods are grouped and presented into three groups, i.e., multimodal joint-embedding models, multimodal attention-based models, and multimodal external knowledge-based models. Various methods of these models are comparatively analyzed in Table 4.

3.3.1 Multimodal Joint-Embedding Models (MMJEMs). MMJEMs join and learn representations of multiple modalities in a common feature space. This rationale is improved further in VQA by performing more reasoning over modalities than image/video description. Ben-Younes et al. [14] proposed a MUTAN framework for VQA. A tensor-based Tucker decomposition model is used with a low-rank matrix constraint to parameterize the bi-linear relations between visual and text interpretations. Desta et al. [35] proposed a framework that merges the visual features and language with abstract reasoning. High-level abstract facts extracted from an image optimize the reasoning process. Cadene et al. [21] proposed an E2E reasoning network for VQA. This research’s main contribution is introducing the MuRel cell, which produces an interaction between question and corresponding image regions. Patro et al. [112] proposed a joint answer and textual explanation generation model. A collaborative correlated (encoder, generator, and correlated) module is used to ensure that answer and its generated explanation are correct and coherent. Lobry et al. [90] proposed a VQA framework for remote sensing data, which can be useful for land-cover classification tasks. Fang et al. [38] proposed an open-ended VQA framework for videos using commonsense reasoning in the language part, where questions are asked about effects, intents, and attributes.

3.3.2 Multimodal Attention-based Models (MMAMs). During the encoding stage, a general encoder-decoder can feed some noisy and unnecessary information at the prediction phase.
Multimodal External Knowledge Bases Models (MMEKMs). Traditional multimodal joint embedding and attention-based models only learn from the information that is present in training sets. Existing datasets do not cover all events/activities of the real world. Therefore, MMEKMs are vital to coping with real-world scenarios. Performance of VQA task is more increasing by linking knowledge bases (KBs) databases to VQA task. Freebase [17], DBpedia [9], WordNet [98],
ConceptNet [86], and WebChild [130] are extensively used KBs. A robust VQA framework requires access to broad information content from KBs. It has been effectively integrated into the VQA task by embedding the various entities and relations.

During the DL era, various external KB methods are proposed for VQA tasks. Wang et al. [140] proposed another framework for the VQA task named Fact-based VQA (FVQA), which uses data-driven approaches and LSTM architecture to map image/question queries. FVQA framework used DBPedia, ConceptNet, and WebChild KB. Narasimhan and Schwing [104] proposed a framework for the VQA task using external knowledge resources that contain a set of facts. This framework can answer both fact-based and visual-based questions. Marino et al. [93] proposed an outside knowledge dataset for VQA, which contains more than 14,000 questions. This dataset contains several categories, such as sports, science and technology, history, and so on. This dataset requires external resources to answer, instead of only understanding the question and image features. Basu et al. [12] proposed a commonsense-based VQA framework. In this framework, the image’s visual contents are extracted and understood by the YOLO framework and represented in the answer set program. Semantic relations features and additional commonsense knowledge answer the complex questions for natural language reasoning. Yu et al. [157] proposed a framework in which visual contents of an image are extracted and processed in multiple perspectives of knowledge graph, such as semantic, visual, and factual perspectives.

### 3.4 Multimodal Speech Synthesis (MMSS)

The most important aspect of human behavior is communication (write/speak). Humans can communicate using natural language by text and speech, representing the written and vocalized form of natural language, respectively. The latest research in language and speech processing helps systems talk like a human being. Speech synthesis is the complicated process of generating natural language spoken by a machine. Natural language text modality is converted into its respective spoken waveform modality in real-time by the Text To Speech (TTS) system. Various applications are introduced in the real world using speech synthesis, such as human-computer interactive systems, screen readers, telecommunication and multimedia applications, talking toy games, and so on.

MMSS faces multimodal representation, alignment, and fusion challenges. In MMSS, the visual representation of lip movements is fused with the audio signal to anticipate uttered words. The data from multiple media may have differing noise structure and predictive strength, with perhaps incomplete information in at least one of the modalities. Alignment of phonemes according to their transcription is a technical task, mostly requiring the team of experts for mapping of phonemes according to their transcriptions. The main research objective of TTS systems these days is to produce a sound like a human’s. Therefore, various aspects are used for the evaluation of the TTS system’s quality, such as naturalness (quality with the perspective of generated speech timing structure, rendering emotions, and pronunciation), intelligibility (in a sentence, quality of each word being produced), synthetic speech preferences (choice of a listener in terms of voice and signal quality for better TTS system) and human perception factors like comprehensibility (understanding quality of received messages). Articulatory TTS, Concatenative TTS, Formant TTS, Parametric TTS, and Deep Learning TTS are the main categories of the speech synthesis process. This section discusses recent research trends and advancements of Deep Learning TTS.

#### 3.4.1 Deep Learning TTS (DLTTS)

In DLTTS frameworks, DNN architectures model the relationship between text and their acoustic realizations. The main advantage of DLTTS is the development of its extensive features without human prepossessing. Also, the naturalness and intelligibility of speech are improved using these systems. Text-to-speech synthesis process is explained
in the general structure diagram of Deep Learning Text To Speech frameworks using DNN architectures, shown in the Figure 6. Comparative analysis of these approaches is shown in Table 5.

Wang et al. [145] proposed “Tacotron,” a sequence 2 sequence TTS framework that synthesizes speech from text and audio pairs. Encoder embeds the text that extracts its sequential representations. The attention-based decoder process these representations, and after that, post-processing architecture generates the synthesized waveforms. In another research, “Deep Voice” model using DNN architecture is proposed by Arik et al. [7], which synthesizes audio from characters. This model consists of five significant blocks for the production of synthesized speech from text. The computational speed is increased compared to existing baseline models, because the model can train without human involvement. Gribiansky et al. [45] proposed a Deep Voice-2 architecture. This framework is designed to improve existing state-of-the-art methods, i.e., Tacotron and Deep Voice-1, by extending multi-speaker TTS through low-dimension trainable speaker embedding. In a third version of Deep Voice, Ping et al. [116] proposed a neural TTS system based on the fully convolutional model with an attention mechanism. This model performs parallel computations by adapting Griffin-Lim spectrogram inversion, WORLD, and WaveNet vocoder speech synthesis.

Fig. 6. General structure diagram of Deep Learning Text to Speech framework using DNN architecture.

Table 5. Comparative Analysis of Speech Synthesis Models

| Model                | Year | Architecture               | Media | Dataset                      | EM     | Model Features                                                                 |
|----------------------|------|---------------------------|-------|------------------------------|--------|--------------------------------------------------------------------------------|
| Tacotron [45]        | 2017 | GRU/RNN                   | T,A   | Internal English speeches    | MOS    |
| DeepVoice [7]        | 2017 | GRU/RNN                   | TA    | MOS                          |
| DeepVoice [145]      | 2017 | GRU/RNN                   | TA    | MOS                          |
| Parallel RNN [50]    | 2018 | CNN                       | A     | North American English, Speech |
| Parallel RNN [50]    | 2018 | CNN                       | A     | MOS                          |
| Speech22 [35]        | 2018 | CNN, RNN/STTM             | AV    | MOS                          |
| ADLV [40]            | 2018 | CNN, RNN/STTM             | TA    | MOS                          |
| DeepVoice3 [116]     | 2018 | CNN, RNN/STTM             | AV    | MOS                          |
| Parallel RNN [50]    | 2018 | CNN, RNN/STTM             | AV    | MOS                          |
| DeepVoice3 [116]     | 2018 | CNN, RNN/STTM             | AV    | MOS                          |

where DLTTS = Deep Learning Text To Speech, MMTFG = Multimodal Talking Face Generation, T = Text, A = Audio, V = Video, EM = Evaluation Metrics, Acc = Accuracy.

This framework is designed to improve existing state-of-the-art methods, i.e., Tacotron and Deep Voice-1, by extending multi-speaker TTS through low-dimension trainable speaker embedding. In a third version of Deep Voice, Ping et al. [116] proposed a neural TTS system based on the fully convolutional model with an attention mechanism. This model performs parallel computations by adapting Griffin-Lim spectrogram inversion, WORLD, and WaveNet vocoder speech synthesis.

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 2s, Article 76. Publication date: February 2023.
“Parallel WaveNet,” an advanced version of WaveNet proposed by Oord et al. [107], uses the probability density distribution method to train networks. A teacher and a student WaveNet are used parallelly in this model. Arik et al. [6] proposed a neural voice cloning system that learns human voice from fewer samples. For that purpose, two techniques, i.e., speaker adaptation and encoding, are used together. Taigman et al. [129] proposed a VoiceLoop framework for a TTS system. This model can deal with un-constrained voice samples without the need for linguistic characteristics or aligned phonemes. This framework transformed the text into speech from voices using a short-shifting memory buffer. Shen et al. [125] proposed “Tacotron2.” It is a neural TTS architecture used to synthesize speech directly from the text. A recurrent-based sequence-to-sequence feature prediction network can map characters to spectrogram and then these spectrograms are used to synthesize waveforms by using a modified version of WaveNet vocoder. Tao and Busso [131] proposed a speech recognition system using multitask learning mechanism. The proposed design takes into account the temporal dynamics across and within modalities, resulting in an enticing and feasible fusion method. Parallel Tacotron is another brilliant invention in recent times for neural TTS approach proposed by Elias et al. [37]. During inference and training processes, this approach is highly parallelizable to achieve optimum synthesis on modern hardware. One-to-many mapping nature of VAE enhances the performance of TTS and also improves its naturalness.

### 3.4.2 Multimodal Talking Face Generation (MMTFG)

MMTFG is the process of generating a series of face pictures that correlate to an audio clip. Coordinating a stationary image with audio is critical for a range of tasks in the entertainment industry, including visual dubbing, the rapid manufacturing of short video clips, and digital character animation. Controlling head positions when creating lip-synchronized clips with audio is quite challenging. Zhou et al. [168] proposed a DAVS framework for the talking face generation (TFG) task. Good-quality talking faces are generated by combining AVSR with audio-visual synchronization from audio speech or video stream as an input. In a later work, Zhou et al. [169] proposed another framework named PC-AVS; the main objective of this framework is to produce the pose-controllable stream of faces. A raw face picture is used as an identifier reference for the model to produce the desired output of talking faces. Hao et al. [51] proposed a framework for TFG that generates the visual stream of eyes controllable talking face. The framework controls the blinking action of eyes using frame replacement and eye conversion approaches.

### 3.5 Other MMDL Applications

#### 3.5.1 Multimodal Emotion Recognition (MMER)

Emotions of human is one of the approaches for expressing feelings. MMER is enormously vital for enhancing the interaction experience between humans and computers. The task of ML is to empower computers to learn and identify new inputs from training datasets, hence it can be used efficiently to detect, process, respond, understand, and recognize emotions of human through computer training. Thus, the primary purpose of affective computing is to provide machines/systems emotional intelligence ability. It has several research areas, such as learning, health, education, communication, gaming, personalized user interface, virtual reality, and information retrieval. Multimodal emotion recognition framework can be developed on an AI/ML-based prototype designed to extract and process emotion information from various modalities, such as speech, text, image, video, facial expression, body gesture, body posture, and physiological signals. MMER faces multimodal representation, alignment, fusion, and co-learning challenges. MMER is accomplished by applying an attention approach to match the transcript of audio signals and text rather than explicitly matching data. The MMC objectives were not tested in the attention-based solution, and it might be interesting to see how the framework functions in the absence of a modality. It is very challenging to design models that not only utilize complementary information but also require supplementary information. The general structure
diagram of multiplicative multimodal emotion recognition using facial, textual, and speech cues is shown in Figure 7.

During the DL era, various authors contributed to emotion recognition by using different architecture and multiple modalities; like Huang et al. [60], who proposed a fusion method for emotion recognition using two modalities, i.e., facial expression and electroencephalogram (EEG) signals. An NN classifier detects happiness, neutral, sadness, and fear states of emotions. Nguyen et al. [106] proposed another approach for emotion recognition by using audio and video streams. In this method, a combination of C3D and DBN architectures is used to model spatio-temporal information and representation of video audio streams. Nguyen et al. [105] proposed another approach for emotion recognition by using audio and video streams. In this method, a combination of C3D, MCB, and DBN architectures is used to model spatio-temporal information and representation of video audio streams. Tripathi et al. [133] proposed a multimodal emotion recognition framework using various modalities, such as text, speech, face expression, and hands movement on the IEMOCAP dataset. The fusion of modalities is performed only at the final layer to improve emotion recognition performance. Hazarika et al. [53] proposed an emotion recognition framework from video conversations. This method can generate self and interpersonal affective summaries from conversations by using contextual information extracted from videos. They also proposed another framework for detecting emotions using the attention mechanism [52]. In this framework, audio and textual modalities are used for the detection of emotions. Jaiswal et al. [64] analyze the change of emotional expressions under various stress levels of an individual. The performance of this task is affected by the degree of lexical or acoustic features. Chong et al. [31] proposed a new online-chatting system called “EmoChat” to automatically recognize the user’s emotions and attach identified emotion automatically and the sent message within a short period. During the chatting, users can know each other’s emotions in this network. Li et al. [82] proposed a multi-step deep system for reliable detection of emotions by using collected data that contains invalid data as well. Neural networks are used to filter this invalid data from videos and physiological signals by using continuity and semantic compatibility. Lai et al. [76] proposed a model for emotion recognition in interactive conversations. In this model, different RNN architectures with variable contextual window sizes differentiate various aspects of contexts in
| Year | Model | Features | Dataset | Model Features | Architecture |
|------|-------|----------|---------|----------------|--------------|
| 2020 | CNN/InceptionResnetV2 | Audio-visual | LUMED-2, DEAP | F1-Score | CNN/VGG19 |
| 2019 | CRBM, CDBN, SVM | Textual | Wikipedia, XMediaNet | NMI | CNN, 3D-CNN, RNN/GRU |
| 2019 | CNN, Mini-Xception, CNN/VGG19 | Audio-visual | IEMOCAP, AVEC | AUC | CNN, OpenSMILE, MFCC |
| 2021 | Mono2Binaural | Acoustic | FAIR-Play, YT-CLEAN, Topk | WAcc, R, PS | CNN, OpenSMILE, MFCC |
| 2021 | Mono2Binaural | Acoustic | SBU Kinect Interaction, Pose Estimation Map | WAcc, R | CNN, OpenSMILE, MFCC |
| 2019 | TextCNN | Textual | FER2013, RECOLA | F1-Score | CNN/VGGNet |
| 2019 | CNN/VGGNet | Visual | CMU-MOSI, POM | Acc | CNN/VGG19, GRU, CNN |
| 2019 | CNN/VGGNet | Visual | IEMOCAP, AVEC | Acc | CNN/VGG19, GRU, CNN |

Conversations to improve the accuracy. Huan et al. [58] proposed a model using attention mechanism. In this method, for better use of visual, textual, and audio features for video emotion detection, bidirectional GRU is cascaded with an attention mechanism. The attention function is used to work with various contextual states of multiple modalities in real time. Cimtay et al. [32] proposed a hybrid fusion method for emotion recognition using three modalities, i.e., facial expression, galvanic skin response (GSR), and EEG signals. This model can identify the actual emotion state when it is prominent or concealed because of natural deceptive face behavior.

In this section, recent experimental analysis for emotion recognition and analysis using multimodal DL are presented. Comparative analysis of these techniques is shown in Table 6. This literature review on emotion detection is based on text, audio, video, physiological signals, facial expressions and body gestures modalities. It is clearly shown by these empirical study that automatic emotion analysis is feasible and can be very beneficial in increasing the accuracy of the system response and enabling the subject’s emotional state to be anticipated more rapidly. Cognitive assessment and their physical response are also analyzed along with primary emotional states.

### 3.5.2 Multimodal Event Detection (MMED)

Due to the popularity of media sharing on the internet, users can easily share their events, activities, and ideas anytime. The aim of MMED systems is to find actions and events from multiple modalities, such as images, videos, audio, text, and so on. According to statistics, Million of tweets are posted per day, and similarly, YouTube users post more than 30,000 hours of videos per hour. Hence, in many CV applications, automatic event- and action-detection mechanisms are required from this large volume of user-generated
videos. Finding events and actions from this extensive collection of data is a complex and challenging task. It has various applications in the real world, such as disease surveillance, governance, commerce, and so on, and also helps internet users to understand and captures happenings around the world. MMED faces multimodal representation, fusion, and co-learning challenges. In alignment challenge for MMED, signals between sparse events and dense continuous signals might possibly not be aligned properly. Multimodal representation and fusion is difficult for MMED models that are designed on the basis of modalities that contain different levels and different types of noises at different points. It is challenging to deal with missing modality at test time. Like, consider a scenario that while recording a video for emotion detection, the camera stops working for a little span or it becomes darker for a while, but audio signal is clear. In this case, visual modality is missing for detection of emotion/facial expression from source video. A summary of various MMED methods is presented in Table 6.

Researchers have contributed a lot during the DL era and proposed many methods for event and action detection using multiple modalities. Like, Gao et al. [43] proposed a method for event classification via social tracking and deep learning in microblogs. Images and text media are fed to a multi-instance deep network to classify events in this framework. Huang et al. [59] proposed an unsupervised method for the detection of anomaly events from crowded scenes using DL architectures. In this model, visual, motion map, and energy features are extracted from video frames. A multimodal fusion network utilized these features for the detection of anomaly events by using the SVM model. In another research to detect salient events from videos, Koutras et al. [73] employed CNN architecture using audio and video modalities. In this model, a CNN architecture based on C3D nets is used to detect events from a visual stream, and a 2D-CNN architecture is used to detect events from the audio stream. Experimental results show the improvement of the proposed method performance over the baseline methods for salient event detection. Yang et al. [154] proposed a framework for event detection from multiple data domains, such as social and news media, to detect real-world events. A unified multi-view data representation is built for image and text modalities from social and news domains in this framework. Class-wise residual units are formulated to identify the multimedia events.

3.5.3 Miscellaneous MMDL Applications (MMAapps). During the past decades, researchers have contributed to various fields of multimodal applications using many modalities. In this section, we discuss the miscellaneous recent contributions by researchers in the field of multimodal action recognition, multimodal cross-modal retrieval, multimodal sound synthesis, and multimodal talking face generation.

Multimodal Action Recognition (MMAC) is one of the most prominent research areas in both image processing and computer vision societies. The basic objective of MMAC is to create automated solutions that can interpret and describe the actions or activities detected in the given scene. Researchers use RGB, audio, flow, depth, and pose modalities in multimodal action recognition. Kazakos et al. [72] proposed a Temporal Binding Network (TBN) for recognition of egocentric action from audio-visual frames. Incorporating audio modality enhances the accuracy of egocentric action recognition with visual modalities to identify motion and appearances from given frames. Yudistira and Kurita [160] proposed a Corrnet architecture that records multimodal spatiotemporal correlations for the action recognition task. RGB frames are passed to a spatial CNN stream, and flow field frames are passed to a temporal CNN stream; the output of these streams is used to predict the action using the Shannon fusion approach. Wu et al. [148] proposed a spatiotemporal learning framework using 3D-CNN architecture to recognize actions from videos. A two-stream 3D-CNN architecture uses supplementary information from RGB and depth videos using pose and depth modalities to enhance the action recognition performance. Garcia et al. [44]
proposed a DMCL framework for recognition of actions using RGB, depth, flow modalities. DMCL framework is designed that learns from multimodal data and strengthens the missing modalities at test time. The knowledge distillation approach is used to train weaker networks using stronger networks.

The goal of **Multimodal Cross-Modal Retrieval (MMCMR)** is to use one form of media as the inquiry to find related information from another media. The problem of **Cross-Modal Retrieval (CMR)** is determining how to quantify the content similarities across multiple media of information, which is often known as the heterogeneity gap. It is very difficult for traditional approaches to represent various representations from different modalities effectively; as a result, deep learning methods are used by researchers in the recent era for CMR methods to mitigate the traditional approaches challenges. Zhen et al. [167] proposed a Deep supervised CMR approach in which the author uses a common representation area to compare various modalities samples. This common representation eliminates cross-modality discrepancies. Jing et al. [68] proposed a framework for 3D-CMR to jointly train the representations from different media with metadata and choose the optimum features from the representations. Different loss functions, i.e., cross-modal center, cross-entropy, and mean-square-error, are used to update the framework. Wang et al. [144] proposed a DRSL hybrid framework for the CMR task by joining the relation network modules. Similarities between text and image modalities are selected by ranking relational pair-wise samples.

**Multimodal Natural Sound Synthesis (MMNSS)** is the process of electrical generation of sound in the absence of acoustic input. The machine is configured to create samples or numbers that represent the intended sound’s pressure component. Gao and Grauman [42] proposed a Mono2Binaural framework that converts the monaural signal to a binaural signal by taking visual cues. Mono2Binaural is a multimodal framework that works on audio and video modalities. Zhou et al. [170] proposed a visual-to-sound framework that produces natural sounds from visual and optical flow frames. This visual-to-sound framework is helpful for visually impaired persons. Wan et al. [137] proposed a framework that synthesizes audio signals to generate scene images using GAN architecture. The framework can produce good-quality scenes from sound with the perspective of both objective and subjective evaluations.

### 4 MMDL DATASETS AND EVALUATION METRIC

A wide variety of datasets is available for multimodal deep learning methods. These datasets are a major driving force for accelerated development in the research area of various MMDL models. In this review, the datasets used in multiple methods discussed in Section 3 of the main article are briefly described below.

#### 4.1 Datasets Used for Image Description

In this literature, various datasets are used for image description task, which are explained here-with. Flickr was first created in 2004 by Ludicorp. It is an American online community for hosting images and video services. Billions of high-resolution as well as professional images are available publicly in the online community. Different versions of Flicker datasets are available for research, i.e., Flickr-8k [55], Flickr-8kCN [83], Flicker-30k [156], and FlickrStyle10k [40]. A brief description of these datasets is given below, and their dataset splits are summarized in Table 7. **PASCAL dataset**: PASCAL Visual Object Classes (PASCAL-VOC) challenges from 2005 to 2012 provide datasets for object class recognition. Later, Rashtchian et al. [119] proposed a UIUC PASCAL sentence dataset used for image classification, object detection, and object segmentation tasks. This dataset contains 1,000 images, and five human written descriptions in English language are generated for each image.
Table 7. Dataset Splits of Flickr group

| Flickr8k | Flickr8k-CN |
|----------|-------------|
| Images   | Caption/image | Total Captions | Images   | Caption/image | Total Captions |
| Training | 6,000 | 5 | 30,000 | 6,000 | 5 | 30,000 |
| Validation | 1,000 | 5 | 5,000 | 1,000 | 5 | 5,000 |
| Test | 1,000 | 5 | 5,000 | 1,000 | 5 | 5,000 |
| Total | 8,000 | 5 | 40,000 | 8,000 | 5 | 40,000 |

Table 8. MS-COCO Dataset Statistics

| MS-COCO | MS-COCO Caption |
|----------|-----------------|
| Images | Categories | Label Instances | Captions/ Image | Images | Categories | Label Instances | Captions/ Image |
| Training | 164,000 | 91 | 2,500,000 | 5 | 166,000 | 80 | 1,500,000 | 5 |
| Validation | 82,000 | 91 | 2,500,000 | 5 | 82,000 | 80 | 1,500,000 | 5 |
| Test | 82,000 | 91 | 2,500,000 | 5 | 82,000 | 80 | 1,500,000 | 5 |
| Total | 328,000 | 91 | 2,500,000 | 5 | 338,000 | 80 | 1,500,000 | 5 |

Table 9. SentiCap Dataset Statistics

| SentiCap Dataset Splits | Number of Images | Number of Sentences | Number of Sentiments |
|-------------------------|-----------------|---------------------|----------------------|
| Positive Subset Training | 998 | 2,873 | 3 |
| Positive Subset Testing | 673 | 2,019 | 3 |
| Negative Subset Training | 997 | 2,468 | 3 |
| Negative Subset Testing | 503 | 1,509 | 3 |

Another dataset for image description task is SBU Captioned Photo, which contains one million filtered images from Flickr with associated descriptions in the form of text. Ordonez et al. [108] made the first attempt to prepare a large dataset for the image description task. The analysis shows some issues with SBU Captioned Photo dataset’s utility, like captions containing irrelevant information (name, location of objects) compared to image content or brief descriptions for each image. And Microsoft-Common Objects in Context (MS-COCO) dataset contains 328,000 images used for object recognition, captioning, and segmentation [85]. MS-COCO dataset is categorized into 91 common objects categories, 82 out of 91 categories include more than 5,000 label instances resulting in 2,500,000 label instances. This dataset contains more label instances as per category, and each image contains five descriptions. The dataset is split into a training set containing 164,000 images paired with five captions per image. And validation and test sets comprise 82,000 images, each with five captions per image. MS-COCO Captions [28] and MS-COCO QA [121] are other variants of this dataset, whose statistics are summarized in Table 8.

SentiCap [95] dataset characterizes images with emotions by producing captions automatically with negative or positive sentiments. It is a subset of the MS-COCO dataset based on COCO images labeled with three negative and three positive sentiments. The positive subset includes 998 images with 2,873 sentences for training and 673 images with 2,019 sentences for testing; similarly, the negative subset contains 997 images with 2,468 sentences for training and 503 images with 1,509 sentences for testing. Statistics and splits of the SentiCap dataset are shown in Table 9.

4.2 Datasets Used for VQA Task

In this section, datasets used for VQA task are discussed. Visual Question Answering (VQA) dataset is the most frequently used dataset for the Visual Question Answering task. There are
various variants of the VQA dataset, such as VQA v1.0 [5], VQA v2.0 [150], VQA-CP [3], VQA-X [111], Ok-VQA [93], and FVQA [140]. VQA v1.0, also named as VQA-real dataset, is split into two parts: first part contains 123,287 training-validation images and 81,434 testing images from MS-COCO [5]; second part contains 50,000 abstract scenes, which are used for high-level reasoning in VQA task. Ten answers for each question from different annotators are gathered in this version. VQA v2.0 [150] is proposed to remove the limitation of inherent bias in version 1.0 by adding another image with a different answer for the same question. Therefore, the VQA v2.0 dataset size is double as compared to version 1.0. VQA v2.0 dataset contains additional complementary images to cope with the referred above limitation and hence contains approximately 443,000 (Question, Image) pairs for training, 214,000 pairs for validation, and 453,000 pairs for testing. Statistics of VQA v1.0 and 2.0 are summarized in Table 10.

DATASET FOR QUESTION ANSWERING ON REAL-WORLD IMAGES (DAQUAR) [92] was the first significant dataset for the VQA task. DAQUAR dataset contains 6,795 QA pairs for the training set and 5,673 QA pairs for the testing set on 1,449 images from the NYU-DepthV2 dataset. Visual7W [172] is another dataset used for the VQA task that contains object localization and dense annotations in an image. Visual7W dataset consists of seven questions (what, when, where, why, who, which, and how.) The questions in Visual7W are much richer, and their answers are also longer than the VQA dataset. It consists of 327,939 QA pairs on 47,300 images from the COCO dataset. Visual Genome dataset is a huge structured knowledge presentation of visual perceptions and a complete collection of question-answers and descriptions [75]. Dense annotations of attributes, relationships, and objects within each image are collected in this dataset. This dataset contains approximately 108,000 images.

COMPOSITIONAL LANGUAGE AND ELEMENTARY VISUAL REASONING DIAGNOSTICS (CLEVR) dataset [69] is used to solve complex reasoning. It performs extensive diagnostics for better understanding of reasoning capabilities. It contains 100,000 images, out of which 70,000, 15,000, and 15,000 images are used for training, testing, and validation sets, respectively. TASK DIRECTED IMAGE UNDERSTANDING CHALLENGE (TDIUC) dataset [71] contains 167,437 Images from Visual Genome and MS-COCO datasets. TDIUC dataset contains 1,654,167 QA pairs derived from three different sources, which are organized into 12 separate categories. Statistics of other VQA datasets are summarized in Table 11.

4.3 Datasets Used for Video Description

In this section, datasets used for video description task are explained briefly. TACoS-MultiLevel dataset [123] is also annotated on TACos Corpus [120] through AMT workers. TACoS-MultiLevel dataset contains 185 long indoor videos, and approximately every video is six minutes long. These videos contain various actors, small interacting objects, and activities about cooking scenarios.
Multiple AMT workers annotate the video sequence intervals by pairing them with a short sentence. Various datasets used for video description task are comparatively analyzed in Table 12.

**Microsoft Video Description (MSVD)** dataset [24] corpus contains a different crowd-sourced description of small video clips in the form of text. MSVD dataset contains 1,970 clips from YouTube annotated in the format of sentences by AMT workers. Approximately, 10–25 seconds duration video clip is used to illustrate activity in this dataset, and it also supports multilingual descriptions in the form of English, Chinese, German, and so on. Similarly, **Montreal-Video Annotation Dataset (M-VAD)** [132] consists of 48,986 video clips from 92 different movies based on descriptive video service. M-VAD dataset contains 92 filtered and 92 unfiltered movies. On average, each video clip is spanned over 6.2 seconds, and the entire running time for the whole dataset is approximately 84.6 hours. It contains 510,933 words, 48,986 paragraphs, and 55,904 sentences. The dataset is split into 38,949 clips for training, 4,888 clips for validation, and 5,149 clips for testing.

**Microsoft Research-Video to Text (MSR-VTT)** dataset [152] consists of a broader range of open domain videos used for videos’ description task. MSR-VTT contains approximately 7,180 videos, which are sub-divided into 10,000 video clips. Twenty different categories are organized to group these clips. The dataset is split into a training, validation, and testing sets containing 6,513, 497, and 2,990 videos, respectively. Twenty captions are annotated for each video by AMT workers. **Charades** dataset [126] includes 9,848 videos of various daily life indoor activities. Different 267 AMT workers are deployed to capture these activities in the form of videos from three continents. The recordings are captured based on a given script that contains descriptions of actions and objects to be recorded. Recording of these videos is accomplished through 15 different indoor scenarios. In this dataset, 157 action classes and 46 objects are used for the description of scenes. It provides 27,847 descriptions for all captured indoor scenes. **ActivityNet Captions** dataset [74] includes 100,000 dense descriptions in natural language for approximately 20,000 videos from ActivityNet [20]. The total period for these videos is about 849 hours. In this article, ActivityNet dataset is used in video description task. **Video-to-Commonsense (V2C)** dataset [38] consists of approximately 9,000 human agent videos implementing different actions from MSR-VTT dataset. Three kinds of commonsense descriptions are used to annotate these videos by recruiting two sets of AMT workers. 6,819 videos are used for the training set, and 2,906 are used for the testing set. Approximately 121,65 commonsense captions are generated for these videos. Extension of this dataset named V2C-QA is also proposed to deal with the VQA task in the perspective of commonsense knowledge.
Table 13. Comparative Analysis of Different Emotion Detection Databases

| Dataset Name     | Language | Number of subjects | Number of samples      | Emotion states |
|------------------|----------|--------------------|------------------------|---------------|
| eNTERFACE [94]   | English  | 42 (34-male, 8-female) | 1,166 video sequences  | 6             |
| FABO [46]        | English  | 23 (11-male, 12-female) | Camera recording of 1 hour/subject | 10            |
| IEMOCAP [18]     | English  | 10 (5-male, 5-female)  | 12 hours               | 5             |
| MSP-Improve [19] | English  | 12 (6-male, 6-female)  | 18 hours               | 4             |
| MuSE [63]        | English  | 28 (19-male, 9-female) | 10 hours               | 2             |
| SEMAINE [97]     | English  | 150 (57-male, 93-female) | 959 conversations      | 5             |
| RECOLA [122]     | French   | 46 (19-male, 27-female)| 7 hours                | 5             |

4.4 Datasets Used for Speech Synthesis

Datasets used for speech synthesis task are explained briefly herewith. **North American English** dataset [145] is built using a single speaker speech database. The duration of speech data in this dataset is around 24.6 hours. Professional female speakers are used to speak this speech data. This dataset is used for the speech synthesis process. **Voice Cloning Toolkit (VCTK)** dataset [135] contains read speech data from 109 native English speakers with different accents. Speakers read approximately 400 sentences chosen from the newspaper. The cumulative length of these audios is about 44 hours, with a sample rate of 48 kHz. Standard training and test splits are not provided in this dataset. **LibriSpeech** corpus [109] contains English read speech from audiobooks that are derived through the LibriVox project. Around 2,484 male and female speakers are engaged to create this corpus. The cumulative length of these audios is approximately 1,000 hours with a sample rate of 16 kHz.

**Lj Speech** dataset [62] is a publicly available speech dataset; it contains 13,100 small audio clips. These clips are recorded by a single speaker, who read passages from seven different books. The duration of each clip is different, can vary from 1 to 10 seconds. The cumulative period of these clips is around 24 hours. **Proprietary Speech** [37] dataset is used for the speech synthesis process. The cumulative duration of speech data in this dataset is around 405 hours. It includes a total of 347,872 utterances varying in three different English accents from 45 speakers. Out of these 45 speakers, 32 speakers have a US English accent, 8 have a British English accent, and 5 have an Australian English accent.

4.5 Datasets Used for Emotion Recognition

In this section, datasets used for emotion recognition task are explained briefly. Different datasets used for emotion detection task are comparatively analyzed in Table 13, where FABO [46] consists of most number of “Emotion States” as compared to other datasets. In terms of subjects (participants) for emotion detection databases, SEMAINE [97] database consists of more number of subjects, i.e., 150 (57-Males and 93-Females), for recording of samples for emotion detection tasks.

**eNTERFACE audio-visual** database [94] contains 1,166 video sequences. Out of these sequences, 902 cover male video recordings, and 264 cover female recordings. Six different emotion categories are expressed in this database, i.e., happiness, anger, fear, disgust, surprise, and sadness. Hatice Gunes and Massimo Piccardi produced the **Bimodal Face and Body Gesture Database (FABO)** database [46] at the University of Technology, Sydney. This database creates bimodal face and body expressions for the automated study of human affective activities. Various volunteers gathered visual data in a laboratory environment by directing and requesting the participants with the perspective of desired actions/movements. Ten different emotion categories are expressed in this database, i.e., neutral, happiness, anger, fear, disgust, surprise, sadness, uncertainty, anxiety, and boredom. **Interactive Emotional Dyadic Motion Capture (IEMOCAP)** [18] is a multi-speaker and multimodal database. The cumulative duration of audio-visual data in this dataset is around 12 hours. Various videos, face motions, speech, text transcriptions are included in this bunch of data. Multiple annotators are used to annotate this dataset into categorical labels.
such as happiness, frustration, anger, sadness, neutrality, and dimensional labels such as activation, valence, and dominance.

**MSP-Improve** [19] database was compiled to record naturalistic emotions from improvised circumstances. Six sessions are recorded from 12 actors, and 652 target sentences are collected containing lexical data. The duration of each session is around three hours. Data is split into 2,785 natural interactions, 4,381 improvised turns, and 620 read sentences for 8,438 utterances in total. **Multimodal Stressed Emotion (MuSE)** [63] database recognizes the interaction between emotion and stress in naturally spoken communication. This dataset contains 55 different recordings from 28 contestants. Each contestant was recorded for stressed and not-stressed sessions. The cumulative duration of this dataset is approximately 10 hours from 2,648 utterances. **Sustained Emotionally coloured Machinehuman Interaction using Nonverbal Expression (SEMAINE)** [97] has built a substantial audio-visual database to create Sensitive Artificial Listener agents that involve a person in a prolonged emotional conversation. Five high-resolution cameras and four microphones are used synchronously for high-quality recording. One hundred fifty contestants from eight different countries record 959 conversations in total. **REmote COLlaborative and Affective (RECOLA)** database [122] consists of spontaneous and natural emotions in continuous domain. Video, audio, electrocardiogram, and electro-dermal activity modalities are used in this database. The cumulative duration of the recording is around 9.5 hours from 46 French-speaking contestants.

### 4.6 MMDL Evaluation Metrics

Evaluation metrics determine the performance of machine learning or statistical models. For any project, evaluation of algorithms and ML models is essential. There are several kinds of evaluation metrics available to validate the frameworks. To test the model, single as well as multiple evaluation metrics are used. Because the performance of architecture may vary to different evaluation metrics, it may show better results using one metric but show different results using other metrics. Therefore, evaluating the performance of a model with the perspective of many evaluation metrics gives a broader view. Overview of Various Evaluation metrics used in this article is summarized and presented with its relevant dataset and MMDL application group in Table 14.

### 5 EXPERIMENTAL RESULTS COMPARISONS ON BENCHMARK DATASETS AND EVALUATION METRICS

In this section, we provide a summary of the performance and experimental results of several models that have been reported. The following subsections show the results of all the models presented in Section 3.

#### 5.1 Multimodal Image Description Results

Multimodal image description models are categorized into three parts, i.e., encoder-decoder-based image description, semantic concept-based image description, and attention-based image description. We can discuss the performance of these models on the MS-COCO widely used dataset. Comparative analysis of the MMID model’s experimental results on benchmark evaluation metrics is shown in Table 15. Visual feature extractor plays an important role in the evaluation of the model’s performance. Like for CRNN [149] model, experimental results for BLEU, CIDEr, and METEOR metrics are improved using InceptionV3 visual extractor instead of VGG16. GET [65] model achieves the best results on the MS-COCO dataset using all the evaluation metrics as compared to all EDID models. This model gets the best experimental results on encoder-decoder-based image description models and achieves the best results on standard evaluation metrics on semantic concept-based image description and attention-based image description models.
intra-layer representations are used to merge local and global features, and the global gated adaptive controller fuses the relevant information at the decoder to enhance the model’s performance. The experimental results showed the dominance of GET as compared to other models of image captioning. Experimental results on all listed MMID models are compiled and listed separately for each standard evaluation metric, as presented in Table 8. Top row shows values (percentage) of image description models on BLEU (B_1,2,3,4) metric, and bottom row shows values (percentage) on CIDEr (C), METEOR (M), ROUGE (R), SPICE (S) metrics. GET [65] model results are better than all comparative image description models on B_1, B_4, M, R metrics except results of VSR model on C and S metrics.

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 2s, Article 76. Publication date: February 2023.
5.2 Multimodal Video Description Results

In this article, multimodal video description models are categorized into CNN-to-RNN encoder-decoder group, RNN-to-RNN encoder-decoder group, and deep reinforcement learning group of applications. Most of these models’ performances are evaluated on MSVD and MSR-VTT datasets using BLEU (B), CIDEr (C), METEOR (M), and ROUGE (R) evaluation metrics, as shown in Table 16. SemSynAN [114] a CNN-to-RNN encoder-decoder-based model, achieved the best experimental results on B, C, M, and R metrics, as compared to all other video description models. This model outperforms on both MSR-VTT and MSVD datasets. In the SymSynAN model, syntactic representations extract rich visual information from input video frames. Both 2D-CNN and 3D-CNN visual feature extractors are used to extract rich visual features from video and are globally represented. MMVD guide language model by merging syntactic, visual, and semantic representations to enhance video description accuracy at the decoder stage. Experimental results of all listed MMVD models are compiled and listed separately on common MSR-VTT and MSVD datasets, as presented in Figure 9. The top row shows values (percentage) of MMVD model’s on B, C, M, and R metrics.
on the MSR-VTT dataset, while the bottom row shows values (percentage) of MMVD model’s on B, C, M, and R metrics on MSVD dataset. SemSynAN [114] model clearly shows better results than listed comparative video description models.

5.3 Multimodal Visual Question Answering Results

VQA is a system where a machine provides the answer in the form of words or short phrases by combining CV and NLP domains. In this review article, MMVQA models are categorized into MMJEM, MMAM, and MMEKM. Table 17 shows the performance of VQA models by presenting findings of experimental results conducted on the VQA and VQA family dataset on two test splits termed “test-std and test-dev” to determine the accuracy. VQA is a pretty big dataset composed of 265,016 pictures and open-ended image-related questions. A strong grasp of language, vision, and common sense knowledge is required to answer these questions. MCAN [158] achieves the best overall accuracy on test-dev and test-std, as compared to other comparative VQA models. MCAN is a multimodal attention-based network, where the MCA layer uses image features and questions self-attention for the reasoning part. This MCA layer refines the question and image representations to improve the overall accuracy of the model.

5.4 Multimodal Speech Synthesis Results

Comparative analysis of MM speech synthesis model’s results on U.S. English speech database is presented in Table 18. For evaluation of speech synthesis, experimental results Mean Opinion Score (MOS) and Mel-Cepstral Distortion (MCD) are the most commonly used subjective and objective evaluation metrics. Subjective assessment approaches are often better suited for assessing speech synthesis models, but they demand substantial resources and confront problems in terms of findings validity, reliability, and reproduction of results. In contrast, the quantitative assessment of the Text To Speech model is performed by the objective evaluation technique. The discrepancies between produced and real speech data are commonly utilized to assess the model. These assessment measures can only represent the model’s data parsing capacity to a limited level and cannot indicate the produced speech’s quality. The MOS is the most widely employed
subjective assessment technique, which evaluates naturalness by inviting participants to rate the synthetic speech. MOS higher scores indicate better speech quality. MCD is the most commonly used objective assessment method to determine the difference between actual and reproduced samples. Tacotron2 [125] DLTTS model performs better, as compared to other speech synthesis models. Parallel Tacotron [37] is also the closest model in terms of experimental results, but its relevancy is much more than Tacotron2.

6 DISCUSSION AND FUTURE DIRECTION

Issues and limitations of various methods and applications mentioned in Section 3 are highlighted below. Possible future research directions for MMID, MMVD, MMVQA, MMSS, MMER, and MMED are discussed separately.

6.1 Multimodal Image Description

In recent years, the image description domain improves a lot, but there is still some space available for improvement in some areas. Some methods did not detect prominent attributes, objects and, to some extent, does not generate related or multiple captions. Here, some issues and their possible future directions are mentioned below to make some advancement in the image description.

- The produced captions’ accuracy depends mostly on grammatically correct captions, which depends on a powerful language generation model. Image captioning approaches strongly depend on the quality and quantity of the training data. Therefore, another key challenge in the future will be to create large-scale image description datasets with accurate and detailed text descriptions in an appropriate manner.
• Existing models demonstrate their output on datasets comprising a collection of images from the same domain. Therefore, implement and test image description models on open-source datasets will be another significant research perspective in the future.
• Too much work has been done on supervised learning approaches, and training requires a vast volume of labeled data. Consequently, reinforcement and unsupervised learning approaches have less amount of work done and have a lot of space for improvement in the future.
• More advanced attention-based image-captioning mechanism or captioning based on regions/multi-region is also a research direction in the future.

6.2 Multimodal Video Description
The performance of the video description process is enhanced with the advancement of DL techniques. Even though the present video description method’s performance is already far from the human description process, this accuracy rate is decreasing steadily, and still, there is sufficient space available for enhancement in this area. Below, some issues and their possible future directions are mentioned to make some advancement in the video description.

• It is expected in the future that humans will be capable of interacting with robots as humans can interact with each other. Video dialogue is a promising area to cope with this circumstance similarly to audio dialogue (such as Hello Google, Siri, ECHO, and Alexa). In visual dialogue, the system needs to be trained to communicate with humans or robots in a conversation/dialogue manner by ignoring conversational statements’ correctness and wrongness.
• In CV, most research has focused on descriptions of visual contents without extraction of audio/speech features from the video. Existing approaches extract features from visual frames or clips for a description of the video. Therefore, extraction of these audio features may be more beneficial for the improvement of this description process. Audio features, such as the sound of water splash, cat, dog, car, guitar, and so on, provide occasional/eventual information when no visual cue is available for their existence in the video.
• Existing approaches have performed end-to-end training, resultantly more and more data is utilized to improve the accuracy of the model. But extensive dataset still does not cover the occurrences of real-world events. To improve system performance in the future, learning from data itself is a better choice and achieves its optimum computational power.
• In the video description process, mostly humans describe the visual content based on extensive prior knowledge. They do not all the time rely solely on visual content; some additional background knowledge is applied by the domain expertise as well. Augmentation of video description approaches with some existing external knowledge would be an attractive and promising technique in this research area. This technique shows some better performance in VQA. Most likely, this would dominate some accuracy improvement in this domain as well.
• Video description can be used in combination with machine translation for auto video subtitling. This combination is very beneficial for entertainment and other areas as well. So, this combination needs to be focused on making the video subtitling process easier and cost-effective in the future.

6.3 Multimodal Visual Question Answering
In recent years, the VQA task has gained tremendous attention and accelerated the development process by the contribution of various authors in this specific domain. VQA is especially enticing, because it comprises a complete AI task that considers free-form, open-ended answers to questions by extracting features from an image/video and the question asked. The accuracy of VQA research
still goes beyond answering the visual-based questions as compared to human equivalent accuracy. Some issues and their possible future directions for the VQA task are mentioned below.

- So far, the image/video feature extraction part is almost fixed to a model; like, most of the time, ImageNet Challenge model is used. In this model, image features are extracted by the split of frames into uniform boxes, which works well for object detection, but VQA needs features by tracking all objects using semantic segmentation. Therefore, some more visual feature extraction mechanisms for VQA tasks need to be explored in the future.
- Goal-oriented datasets for VQA research need to be designed in future to support real-time applications, such as instructing users to play the game, helping visually impaired people, support for data extraction from colossal data pool and robot interaction, and so on. So far, VizWiz [50] and VQA-MED [2] are two publicly available goal-oriented datasets for VQA research. Therefore, in the future, more goal-oriented datasets need to be built for mentioned above applications.
- Most baseline VQA methods have been tested using traditional accuracy measure, which is adequate for the multiple-choice format. In the future, the assessment or evaluation method for open-ended frameworks for VQA tasks is examined to improve these models’ accuracy.

### 6.4 Multimodal Speech Synthesis

The DLTTS models use distributed representations and complete context information to substitute the clustering phase of the decision tree in HMM models. Therefore, to produce a better quality of speech synthesis process compared to the traditional method, several hidden layers are used to map context features to high-dimensional acoustic features by DLTTS methods. More hidden layers inevitably raise the number of system parameters to achieve better performance. As a result, space and time complexity for system training is also increased. Therefore, for network training, a large amount of computational resources and corpora is required. Besides these achievements, there is still room available for DLTTS models in terms of quality improvement in intelligibility and speech’s naturalness. Therefore, some issues and their feature research directions are discussed below.

- DLTTS approaches usually require a huge amount of high-quality (text, speech) pairs for training, which is time-consuming and an expensive process. Therefore, in the future, data efficiency for E2E DLTTS models training is improved by the public availability of unpaired text and speech recordings on a large scale.
- Little progress is made in front-end text analysis to extract valuable context or linguistic features to minimize the gap between the text-speech synthesis process. Therefore, it is a good direction in the future to utilize specific context or linguistic information for E2E DLTTS systems.
- Parallelization will be an essential aspect for DLTTS systems to improve system efficiency, because most DNN architecture needs many calculations. Some frameworks proposed in recent years also use parallel networks for training or inference and achieve some good results, but there is still room available to achieve optimum results in the future.
- As for DLTTS application concerns, the use of speech synthesis for other real-world applications such as voice conversion or translation, cross-lingual speech conversion, audio-video speech synthesis, and so on, are good future research directions.

### 6.5 Multimodal Emotion Recognition

Analysis of automatic emotions requires some advanced modeling and recognition techniques along with AI systems. For more advanced emotion recognition systems, future research based on
AI-based automated systems constitute more scientific progress in this area. Some future directions of listed issues of multimodal emotion recognition are mentioned below.

- Existing baseline approaches are successful, but further experience, knowledge, and tools regarding the analysis and measurement of automatic non-invasive emotions are required.
- Humans use more modalities to recognize emotions and are compatible with signal processing; machines are expected to behave similarly. But the performance of automated systems is limited with a restricted set of data. Therefore, to overcome this limitation, new multimodal recordings with a more representative collection of subjects are needed to be considered in the future.
- The preprocessing complexity of physiological signals in emotion detection is a big challenge. In physiological signals, detection of emotion states through electrocardiogram, electromyography, and skin temperature is still emerging. Hence, to determine the potency of these techniques, a detailed research can be carried out in the future.

6.6 Multimodal Event Detection

Due to the exponential increase in web data, multimodal event detection has attracted significant research attention in recent years. MMED seeks to determine a collection of real-world events in a large set of social media data. Subspace learning is an efficient approach to handle the issue of heterogeneity for the learning of features from multimodal data. However, E2E learning models are more versatile, because they are structured to obtain heterogeneous data correlations directly. MMED from social data is still challenging and needs to be improved in some open issues in the future.

- The “in the obscenity of dimensionality” issue is raised with the concatenation of various feature vectors of multiple modalities. Some considerable advancement has been made in curse of dimensionality issue, but existing techniques still require some improvement in terms of achieving better learning accuracy for social platforms. Like, GAN/RNN architecture’s extensions are valuable to improve feature learning accuracy of multimodal social data.
- Textual data also coincide with audio and video media. Therefore, MMED is improved if this textual information is considered jointly with these media.
- Current research primarily focuses on the identification of events from a single social platform. A comprehensive understanding of social data events is obtained by synthesizing the information from multiple social platforms. Hence, event detection methods can be implemented simultaneously to examine social data from multiple platforms using a transfer learning strategy.

7 Conclusion

In this survey, we discussed the recent advancements and trends in MMDL. Various DL methods are categorized into different MMDL application groups and explained thoroughly using multiple media. This article focuses on an up-to-date review of numerous applications using various modalities, such as image, video, text, audio, body gestures, facial expressions, physiological signals, flow, RGB, pose, depth, mesh, and point cloud compared to previous related surveys. A novel fine-grained taxonomy of various MMDL methods is proposed. Additionally, a brief discussion on architectures, datasets, and evaluation metrics used in these MMDL methods and applications is provided. A detailed comparative analysis is provided for each group of applications by discussing the model’s architectures, media, datasets, evaluation metrics, and features. Experimental results comparisons for MMDL applications are provided on benchmark datasets and evaluation metrics. Finally, open research problems are also mentioned separately for each group of applications, and
possible research directions for the future are listed in detail. We expect that our proposed taxonomy and research directions will promote future studies in multimodal deep learning and help in a better understanding of unresolved issues in this particular area.

APPENDIX

A LIST OF ABBREVIATIONS

List of Abbreviations

| Abbreviation | Explanation |
|---------------|-------------|
| AFDBN         | Adaptive Fractional Deep Belief Network |
| AMT           | Amazon Mechanical Turk |
| ASKMS         | Adaptive Shape Kernel Based Mean Shift |
| AUC           | Area Under Curve |
| BLSTM         | Bidirectional Long Short-Term Memory |
| BRNN          | Bidirectional Recurrent Neural Network |
| CALO-MA       | CALO Meeting Assistant |
| CHIL          | Computers in the Human Interaction Loop |
| CMN           | Compositional Modular Network |
| COH           | Color Oriented Histogram |
| CRNN          | Cascade Recurrent Neural Network |
| CTT             | Concatenative Text To Speech |
| DaQUAR  | Dataset for Question Answering on Real-world images |
| DBN           | Deep Belief Network |
| DL            | Deep Learning |
| DNN           | Deep Neural Network |
| E2E           | End-to-End |
| EER           | Equal Error Rate |
| F3RBM         | Fuzzy Removing Redundancy Restricted Boltzmann Machine |
| FBNN          | Fuzzy Restricted Boltzmann Machine |
| FVQA          | Fact-based Visual Question Answering |
| GLU           | Gated Linear Unit |
| GRU           | Gated Recurrent Unit |
| HRL           | Hierarchical Reinforcement Learning |
| KB            | Knowledge Bases |
| MAE           | Mean Absolute Error |
| MCB           | Multimodal Compact Bilinear |
| MDP           | Markov Decision Process |
| ML            | Machine Learning |
| MMIDL         | Multimodal Deep Learning |
| MMER          | Multimodal Emotion Recognition |
| MAJEM         | Multimodal Joint-Embedding Models |
| MS-COCO       | Microsoft-Common Objects in Context |
| MSR-VTT       | Microsoft Video Description |
| MuSe          | Multimodal Stressed Emotion |
| NLP           | Natural Language Processing |
| NN            | Neural Networks |
| PoS           | Part of Speech |
| PTTS          | Parametric Text to Speech |
| RB            | Rule Based System |
| RP-TF         | Random-Forest Tag-Propagation |
| RNN           | Recurrent Neural Network |
| SEMAINE       | Sustained Emotionally coloured Machinehuman Interaction using Nonverbal Expression |
| SiGRU         | Simplified GRU |
| SMT           | Statistical Machine Translation |
| SST           | Single-Stream Temporal |
| SVO           | Subject, Object, Verb |
| TTS           | Text To Speech |
| VAE           | Variational Auto-Encoders |
| VDR           | Visual Dependency Representation |

| Abbreviation | Explanation |
|---------------|-------------|
| AI            | Artificial Intelligence |
| ANN           | Artificial Neural Network |
| ATTS          | Articulatory Text To Speech |
| AVSR          | Audio-Visual Speech Recognition |
| BNDMBM        | Batch Normalized Deep Boltzmann Machine |
| C3D           | Three Dimensional Convolutional Neural Networks |
| CBDN          | Competitive Deep Belief Network |
| CLEVR         | Computational Language and Visual reasoning diagnostics |
| CNN           | Convolutional Neural Network |
| CDBN          | Contractive slab and spike Convolutional Deep Boltzmann Machine |
| DBM           | Deep Boltzmann Machine |
| DDBN          | Discriminative Deep Belief Network |
| DLTTS         | Deep Learning Text To Speech |
| DRL           | Deep Reinforcement Learning |
| EDR           | Equal Detected Rate |
| EOH           | Edge Oriented Histogram |
| FTTTS         | Formant Text To Speech |
| GRRBM         | Gaussian-Bernoulli Restricted Boltzmann Machine |
| GMM           | Gaussian Mixture Model |
| HMM           | Hidden Markov Model |
| IEMOCAP       | Interactive Emotional Dyadic Motion Capture |
| LSTMB         | Long Short-Term Memory |
| mAP           | mean Average Precision |
| MCD           | Mel Cepstral Distortion |
| MFCC          | Mel Frequency Cepstral Coefficients |
| MMAM          | Multimodal Attention-Based Models |
| MMED          | Multimodal Event Detection |
| MMKEBM        | Multimodal External Knowledge Bases Models |
| MOS           | Mean Opinion Score |
| MSDBM         | Mean supervised Deep Boltzmann Machine |
| MSVD          | Microsoft Video Description |
| M-VAD         | Montreal-Video Annotation Dataset |
| NMI           | Normalized Mutual Information |
| NBDBM         | Normalized Restricted Boltzmann Machine |
| PPLX          | Perplexity metric |
| RBBM          | Restricted Boltzmann Machine |
| RECOLA        | REmote COLaborative and Affective |
| RGF-D         | Red Green Blue-Depth |
| Robust DDBM   | Robust spike-and-slab Deep Boltzmann Machine |
| SGRU          | Stacked Gated Recurrent Unit |
| SML           | Stochastic Maximum Likelihood |
| SRE           | Semantic Relation Extractor |
| STFT          | Short-Time Fourier Transform |
| TDIUC         | Task Directed Image Understanding Challenge |
| V2C           | Video-to-Commonsense |
| VCK           | Voice Cloning Toolkit |
| VQA           | Visual Question-Answering |
REFERENCES

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12487–12496.

[2] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. CLEF (Working Notes) 2, 6 (2019).

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. 4971–4980.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6077–6086.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision. 2425–2433.

[6] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. Adv. Neural Inf. Process. Syst. 31 (2018).

[7] Sercan Ö. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. In Proceedings of the International Conference on Machine Learning. PMLR, 195–204.

[8] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. Multim. Syst. 16, 6 (2010), 345–379.

[9] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In The Semantic Web. Springer, 722–735.

[10] Lorraine E. Bahrick. 1983. Infants’ perception of substance and temporal synchrony in multimodal events. Infant Behav. Devel. 6, 4 (1983), 429–451.

[11] Tadas Baltrusaitis, Chaitanya Abuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41, 2 (2019), 423–443.

[12] Kinjal Basu, Farhad Shakerin, and Gopal Gupta. 2020. AQuA: ASP-based visual question answering. In Proceedings of the International Symposium on Practical Aspects of Declarative Languages. Springer, 57–72.

[13] Christian Beecks, Jakub Lokoř, Thomas Seidl, and Tomáš Skopal. 2011. Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval. 1–8.

[14] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker fusion for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision. 2612–2620.

[15] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. arXiv preprint arXiv:2004.10151 (2020).

[16] George Boateng. 2020. Towards real-time multimodal emotion recognition among couples. In Proceedings of the International Conference on Multimodal Interaction. 748–753.

[17] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data. 1247–1250.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42, 4 (2008), 335–359.

[19] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. IEEE Trans. Affect. Comput. 8, 1 (2016), 67–80.

[20] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 961–970.

[21] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. MUREL: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1989–1998.
A Review on Methods and Applications in Multimodal Deep Learning

[22] Pengfei Cao, Zhongyi Yang, Liang Sun, Yanchun Liang, Mary Qu Yang, and Renchu Guan. 2019. Image captioning with bidirectional semantic attention-based guiding of long short-term memory. *Neural Process. Lett.* 50, 1 (2019), 103–119.

[23] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemeot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaikos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[24] David Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 190–200.

[25] Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. 2021. HEU Emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Comput. Appl.* 33, 14 (2021), 8669–8685.

[26] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16846–16856.

[27] Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. 2017. Reference-based LSTM for image captioning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[28] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[29] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2018. Less is more: Picking informative frames for video captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 358–373.

[30] Ling Cheng, Wei Wei, Xianling Mao, Yong Liu, and Chunyan Miao. 2020. Stack-VS: Stacked visual-semantic attention for image caption generation. *IEEE Access* 8 (2020), 154953–154965.

[31] Luyao Chong, Meng Jin, and Yuan He. 2019. EmoChat: Bringing multimodal emotion detection to mobile conversation. In *Proceedings of the 5th International Conference on Big Data Computing and Communications (BIGCOM)*. IEEE, 213–221.

[32] Yucel Cimtay, Erhan Ekmeckicoglou, and Seyma Caglar-Ozhan. 2020. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access* 8 (2020), 168865–168878.

[33] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. (2020), 1441–1449 pages.

[34] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco GB DE Natale. 2017. Multimodal retrieval with diversification and relevance feedback for tourist attraction images. *ACM Trans. Multim. Comput. Commun. Appl.* 13, 4 (2017), 1–24.

[35] Mikyas T. Desta, Larry Chen, and Tomasz Kornuta. 2018. Object-based reasoning in VQA. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1814–1823.

[36] Parul Diwakar. 2021. Automatic image captioning using deep learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

[37] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J. Weiss, and Yonghui Wu. 2021. Parallel Tacotron: Non-autoregressive and controllable TTS. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5709–5713.

[38] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2Commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162* (2020).

[39] Yang Feng, Lin Ma, Wei Li, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4125–4134.

[40] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. StyleNet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[41] Jing Gao, Peng Li, Zhikui Chen, and Jianping Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computat.* 32, 5 (2020), 829–864.

[42] Luohan Gao and Kristen Grauman. 2019. 2.5D visual sound. In *Adv. Neural Inf. Process. Syst.* 32, 33 (2019), 1–14.

[43] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. 2021. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2755–2764.

[44] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-speaker neural text-to-speech. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[45] Hatice Gunes and Massimo Piccardi. 2006. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06)*. IEEE, 1148–1153.
[47] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-style image captioning with unpaired stylized text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4204–4213.

[48] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. IEEE Access 7 (2019), 63373–63394.

[49] Wenyu Guo, Ying Zhang, Jufeng Yang, and Xiaojie Yuan. 2021. Re-attention for visual question answering. IEEE Trans. Image Process. 30 (2021), 6730–6743.

[50] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VisWiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3608–3617.

[51] Jiaqi Hao, Shiguang Liu, and Qing Xu. 2021. Controlling eye blink for talking face generation via eye conversion. In Proceedings of the SIGGRAPH Asia Technical Communications Conference. 1–4.

[52] Devamanyu Hazarika, Sruthi Gorantla, Soujanya Poria, and Roger Zimmermann. 2018. Self-attentive feature-level fusion for multimodal emotion detection. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 196–201.

[53] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the EMNLP Conference on Empirical Methods in Natural Language Processing. 2594–2604.

[54] Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, and Weisheng Dong. 2019. Image caption generation with part of speech guidance. Pattern Recog. Lett. 119 (2019), 229–237.

[55] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell. Res. 47 (2013), 853–899.

[56] Debbie Hoffman-Plotkin and Craig T. Twentyman. 1984. A multimodal assessment of behavioral and cognitive deficits in abused and neglected preschoolers. Child Devel. 55, 3 (1984), 794–802.

[57] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. 2021. Video multimodal anomaly event detection. IEEE Trans. Geosci. Rem. Sens. 59, 5 (2020), 4340–4354.

[58] Ruohong Huan, Jia Shu, Sheng-Lin Bao, Rong-Hua Liang, Peng Chen, and Kai-Kai Chi. 2021. Video multimodal emotion recognition based on Bi-GRU and attention fusion. Multim. Tools. Applic. 80, 6 (2021), 8213–8240.

[59] Shaoqian Huang, Dongjun Huang, and Xinmin Zhou. 2018. Learning multimodal deep representations for crowd anomaly event detection. Math. Prob. Eng. 2018 (2018).

[60] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. 2017. Fusion of facial expressions and EEG for multimodal emotion recognition. Computat. Intell. Neurosci. 2017 (2017).

[61] Yi Huang, Xiaoshan Yang, Junyu Gao, Jiato Sang, and Changsheng Xu. 2020. Knowledge-driven egocentric multimodal activity recognition. ACM Trans. Multim. Comput. Commun. Appl. 16, 4 (2020), 1–133.

[62] Keith Ito and Linda Johnson. 2017. The LJ speech dataset. Retrieved from https://keithito.com/LJ-Speech-Dataset.

[63] Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7415–7419.

[64] Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. 2019. Controlling for confounders in multimodal emotion classification via adversarial learning. In Proceedings of the International Conference on Multimodal Interaction. 174–184.

[65] Jiayi Hao, Shiguang Liu, and Ting Zhang. 2018. Recurrent fusion network for image captioning. In Proceedings of the Conference on Computer Vision and Pattern Recognition. 499–515.

[66] Weitao Jiang, Xiying Li, Haifeng Hu, Qiuliang Hu, and Bohong Liu. 2021. Multi-gate attention network for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV). 499–515.

[67] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. 2021. Cross-modal center loss for 3D cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3142–3151.

[68] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2901–2910.

[69] Biing Hwang Juang and Laurence R. Rabiner. 1991. Hidden Markov models for speech recognition. Technometrics 33, 3 (1991), 251–272.
A Review on Methods and Applications in Multimodal Deep Learning

[91] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes.

[92] YuLong, Pengjie Tang, Hanli Wang, and Jian Yu. 2021. Improving reasoning with contrastive visual information for visual question answering.

[93] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: Visual question answering for remote sensing data.

[94] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[95] HungarianLO, Yue Han, and Marius Kiefer. 2021. A multi-task approach to text-to-image generation.

[96] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices.

[97] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. 3, 1 (2011), 5–17.
[98] George A. Miller. 1995. WordNet: A lexical database for English. Commun. ACM. 38, 11 (1995), 39–41.

[99] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In Proceedings of the AAAI Conference on Artificial Intelligence. 1359–1367.

[100] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. J. Artif. Intell. Res. 71 (2021), 1183–1317.

[101] Louis-Philippe Morency. 2020. Multimodal Machine Learning (or Deep Learning for Multimodal Systems). Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/Integrative_AI_Louis_Philippe_Morency.pdf.

[102] Robert M. Mulligan and Marilyn L. Shaw. 1980. Multimodal signal detection: Independent decisions vs. integration. Percept. Psychophys. 28, 5 (1980), 471–478.

[103] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6588–6597.

[104] Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In Proceedings of the European Conference on Computer Vision (ECCV). 451–468.

[105] Dung Nguyen, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. 2018. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. Comput. Vis. Image Underst. 174 (2018), 33–42.

[106] Dung Nguyen, Kien Nguyen, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. 2017. Deep spatio-temporal features for multimodal emotion recognition. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 1215–1223.

[107] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel WaveNet: Fast high-fidelity speech synthesis. In Proceedings of the International Conference on Machine Learning. PMLR, 3918–3926.

[108] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 24 (2011).

[109] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 5206–5210.

[110] Maja Pantic, Roderick Cowie, Francesca D’Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli. 2011. Social signal processing: The research agenda. In Visual Analysis of Humans. Springer, 511–538.

[111] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. arXiv preprint arXiv:1612.04757 (2016).

[112] Badri Patro, Shivansh Patel, and Vinay Namboodiri. 2020. Robust explanations for visual question answering. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1577–1586.

[113] Wenjie Pei, Jiuyan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-attended recurrent network for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8347–8356.

[114] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. 2021. Improving video captioning with temporal composition of a visual-syntactic embedding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 3039–3049.

[115] Eric Petajan and Hans Peter Graf. 1996. Automatic lipreading research: Historic overview and current work. In Multimedia Communications and Video Coding. Springer, 265–275.

[116] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017).

[117] Md Rahman, Thasin Abedin, Khondokar S. S. Prottoy, Ayana Moshruba, Fazlul Hasan Siddiqui, et al. 2020. Semantically sensible video captioning (SSVC). arXiv preprint arXiv:2009.07335 (2020).

[118] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. IEEE Sig. Process. Mag. 34, 6 (2017), 96–108.

[119] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon’s Mechanical Turk. In Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk. 139–147.
A Review on Methods and Applications in Multimodal Deep Learning

[120] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. Trans. Assoc. Computat. Ling. 1 (2013), 25–36.

[121] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. Adv. Neural Inf. Process. Syst. 28 (2015).

[122] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 1–8.

[123] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In Proceedings of the German Conference on Pattern Recognition. Springer, 184–195.

[124] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Attiah, Vadlamani Ravi, and Alan Peters. 2020. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl.-Based syst. 194 (2020), 105596.

[125] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerrv-Ryan, et al. 2018. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4779–4783.

[126] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Proceedings of the European Conference on Computer Vision. Springer, 510–526.

[127] Cees G. M. Snoek and Marcel Worringer. 2005. Multimodal video indexing: A review of the state-of-the-art. Multim. Tools Appl. 25, 1 (2005), 5–35.

[128] Rafael Souza, André Fernandes, Thiago S. F. X. Teixeira, George Teodoro, and Renato Ferreira. 2021. Online multimedia retrieval on CPU–GPU platforms with adaptive work partition. J. Parallel Distrib. Comput. 148 (2021), 31–45.

[129] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. 2018. VoiceLoop: Voice fitting and synthesis via a phonological loop. arXiv preprint arXiv:1707.06588 (2018).

[130] Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. WebChild: Harvesting and organizing commonsense knowledge from the web. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining. 523–532.

[131] Fei Tao and Carlos Busso. 2020. End-to-end audiovisual speech recognition system with multitask learning. IEEE Trans. Multimedia. 23 (2020), 1–11.

[132] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015).

[133] Samarth Tripathi and Homayoon Beigi. 2018. Multi-modal emotion recognition on IEMOCAP with neural networks. arXiv preprint arXiv:1804.05788 (2018).

[134] Gokhan Tur, Andreas Stolcke, Lynn Voss, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Michael Frandsen, Clint Fredericksen, Martin Graciarena, et al. 2008. The CALO meeting speech recognition and understanding system. In Proceedings of the IEEE Spoken Language Technology Workshop. IEEE, 69–72.

[135] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. SUPERSEDED-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. (2016).

[136] Alex Waibel, Hartwig Steusloff, Rainer Stiefelhagen, and Kym Watson. 2009. Computers in the human interaction loop. In Computers in the Human Interaction Loop. Springer, 3–6.

[137] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. 2019. Towards audio to scene image synthesis using generative adversarial network. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 496–500.

[138] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction network for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7622–7631.

[139] Bin Wang, Cungang Wang, Qian Zhang, Ying Su, Yang Wang, and Yanyan Xu. 2020. Cross-lingual image caption generation based on visual attention model. IEEE Access 8 (2020), 104543–104554.

[140] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. FVQA: Fact-based visual question answering. IEEE Trans. Pattern Anal. Mach. Intell. 40, 10 (2017), 2413–2427.

[141] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. 2017. The VQA-machine: Learning how to use existing vision algorithms to answer new questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1173–1182.

[142] Wei Wang, Yuxuan Ding, and Chunna Tian. 2018. A novel semantic attribute-based feature for image caption generation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3081–3085.
[143] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4213–4222.

[144] Xu Wang, Peng Hu, Liangli Zhen, and Dezhong Peng. 2021. DRSL: Deep relational similarity learning for cross-modal retrieval. Inf. Sci. 546 (2021), 298–311.

[145] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017).

[146] Ran Wei, Li Mi, Yaosi Hu, and Zhenzhong Chen. 2020. Exploiting the local temporal information for video captioning. J. Vis. Commun. Image Represent. 67 (2020), 102751.

[147] Yawei Wei, Leiquan Wang, Haiwen Shao, and Chunlei Wu. 2020. Multi-attention generative adversarial network for image captioning. Neurocomputing 387 (2020), 91–99.

[148] Ran Wei, Li Mi, Yaosi Hu, and Zhenzhong Chen. 2020. Exploiting the local temporal information for video captioning. J. Vis. Commun. Image Represent. 67 (2020), 102751.

[149] Yiwei Wei, Leiquan Wang, Haiwen Shao, and Chunlei Wu. 2020. Multi-attention generative adversarial network for image captioning. Neurocomputing 387 (2020), 91–99.

[150] Hanbo Wu, Xin Ma, and Yibin Li. 2021. Spatiotemporal multimodal learning with 3D CNNs for video action recognition. IEEE Trans. Circ. Syst. Vid. Technol. 32, 3 (2021), 1250–1261.

[151] Jie Wu and Haifeng Hu. 2017. Cascade recurrent neural network for image caption generation. Electron. Lett. 53, 25 (2017), 1642–1643.

[152] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. Comput. Vis. Image Underst. 163 (2017), 21–40.

[153] Yuling Xi, Yanning Zhang, Songtao Ding, and Shaoqian Wang. 2020. Visual question answering model based on visual relationship detection. Sig. Process.: Image Commun. 80 (2020), 115648.

[154] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5288–5296.

[155] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. 2018. High-quality image captioning with fine-grained and semantic-guided visual attention. IEEE Trans. Pattern Anal. Mach. Intell. 40, 12 (2018), 2984–2993.

[156] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Luping Zhou. 2020. Dense video captioning using graph-based sentence summarization. IEEE Trans. Multim. 23 (2020), 1799–1810.
[167] Liangli Zhen, Peng Hu, Xu Wang, and Dezong Peng. 2019. Deep supervised cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10394–10403.

[168] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence. 9299–9306.

[169] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4176–4186.

[170] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3550–3558.

[171] Junjie Zhu, Yuxuan Wei, Yifan Feng, Xibin Zhao, and Yue Gao. 2019. Physiological signals-based emotion recognition via high-order correlation learning. ACM Trans. Multim. Comput. Commun. Appl. 15, 3s (2019), 1–18.

[172] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4995–5004.

Received 8 December 2021; revised 9 April 2022; accepted 31 May 2022