The Visual Computer (2022) 38:2939–2970
https://doi.org/10.1007/s00371-021-02166-7

SURVEY

A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets

Khaled Bayoudh[1] · Raja Knani[2] · Fayçal Hamdaoui[3] · Abdellatif Mtibaa[1]

Accepted: 15 May 2021 / Published online: 10 June 2021

Abstract

The research progress in multimodal learning has grown rapidly over the last decade in several areas, especially in computer vision. The growing potential of multimodal data streams and deep learning algorithms has contributed to the increasing universality of deep multimodal learning. This involves the development of models capable of processing and analyzing the multimodal information uniformly. Unstructured real-world data can inherently take many forms, also known as modalities, often including visual and textual content. Extracting relevant patterns from this kind of data is still a motivating goal for researchers in deep learning. In this paper, we seek to improve the understanding of key concepts and algorithms of deep multimodal learning for the computer vision community by exploring how to generate deep models that consider the integration and combination of heterogeneous visual cues across sensory modalities. In particular, we summarize six perspectives from the current literature on deep multimodal learning, namely: multimodal data representation, multimodal fusion (i.e., both traditional and deep learning-based schemes), multitask learning, multimodal alignment, multimodal transfer learning, and zero-shot learning. We also survey current multimodal applications and present a collection of benchmark datasets for solving problems in various vision domains. Finally, we highlight the limitations and challenges of deep multimodal learning and provide insights and directions for future research.

Keywords Applications · Computer vision · Datasets · Deep learning · Sensory modalities · Multimodal learning

1 Introduction

In recent years, much progress has been made in the field of artificial intelligence thanks to the implementation of machine learning methods. In general, these methods involve a variety of intelligent algorithms for pattern recognition and data processing. Usually, several sensors with specific characteristics are employed to obtain and analyze global and local patterns in a uniform way. These sensors are generally very versatile in terms of coverage, size, manufacturing cost, and accuracy. Besides, the availability of vast amounts of data (big data), coupled with significant technological advances and substantial improvements in hardware implementation techniques, has led the machine learning community to turn to deep learning to find sustainable solutions to a given problem. Deep learning, also known as representation-based learning [2], is a particular approach to machine learning that is gaining popularity due to its predictive power and portability. The work presented in [3] showed a technical transition from machine learning to deep learning by systematically highlighting the main concepts, algorithms, and

✉ Khaled Bayoudh
khaled.isimm@gmail.com

Raja Knani
knani.raja@gmail.com

Fayçal Hamdaoui
faycel_hamdaoui@yahoo.fr

Abdellatif Mtibaa
abdellatif.mtibaa@enim.rnu.tn

[1] Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia

[2] Physics Department, Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia

[3] Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Control, Electrical Systems and Environment (LASEE), National Engineering School of Monastir, University of Monastir, Monastir, Tunisia

trends in deep learning. In practice, the extraction and synthesis of rich information from a multidimensional data space require the use of an intermediate mechanism to facilitate decision making in intelligent systems. Deep learning has been used in many practices, and it has been shown that its performance can be greatly improved in several disciplines, including computer vision. This line of research is part of the rich field of deep learning, which typically deals with visual information of different types and scales to perform complex tasks. Currently, the deep learning algorithms have demonstrated their potential and applicability in other active areas such as natural language processing, machine translation, and speech recognition, performing comparably or even better than humans.

A large number of computer vision researchers focus each year on developing vision systems that enable machines to mimic human behavior. For example, some intelligent machines can use computer vision technology to simultaneously map their behavior, detect potential obstacles, and track their location. By applying computer vision to multimodal applications, complex operational processes can be automated and made more efficient. Here, the key challenge is to extract visual attributes from one or more data streams (also called modalities) with different shapes and dimensions by learning how to fuse the extracted heterogeneous features and project them into a common representation space, which is referred to as deep multimodal learning in this work.

In many cases, a set of heterogeneous cues from multiple modalities and sensors can provide additional knowledge that reflects the contextual nature of a given task. In the arena of multimodality, a given modality depends on how specific media and related features are structured within a conceptual architecture. Such modalities may include textual, visual, and auditory modalities, involving specific ways or mechanisms to encode heterogeneous information harmoniously.

In this study, we mainly focused on visual modalities, such as images as a set of discrete signals from a variety of image sensors. The environment in which we live generally includes many modalities in which we can see objects, hear tones, feel textures, smell aromas, and so on. For example, the audiovisual modalities are complementary to each other, where the acoustic and visual attributes come from two different physical entities. However, combining different modalities or data sources to improve performance is still often an attractive task from one standpoint, but in practice, it makes little sense to distinguish between noise, concepts, and conflicts between data sources. Moreover, the lack of labeled multimodal data in the current literature can lead to reduced flexibility and accuracy, often requiring cooperation between different modalities. In this paper, we reviewed recent deep multimodal learning techniques to put forward typical frameworks and models to advance the field. These networks show the utility of learning hierarchical representations directly from raw data to achieve maximum performance on many heterogeneous datasets. Thus, it will be possible to design intelligent systems that can quickly answer questions, reason, and discuss what is seen in different views in different scenarios. Classically, there are three general approaches to multimodal data fusion: early fusion, late fusion, and hybrid fusion.

In addition to surveys of recent advances in deep multimodal learning itself, we also discussed the main methods of multimodal fusion and reviewed the latest advanced applications and multimodal datasets popular in the computer vision community.

The remainder of this paper is organized as follows. In Sect. 2, we discuss the differences between similar previous studies and our work. Section 3 reviews recent advances in deep multimodal algorithms, the motivation behind them, and commonly used fusion techniques, with a focus on deep learning-based algorithms. In Sects. 4 and 5, we present more advanced multimodal applications and benchmark datasets that are very popular in the computer vision community. In Sect. 6, we discuss the limitations and challenges of vision-based deep multimodal learning. The final section then summarizes the whole paper and points out a roadmap for future research.

2 Comparison with previous surveys

In recent years, the computer vision community has paid more attention to deep learning algorithms due to their exceptional capabilities compared to traditional handcrafted methods. A considerable amount of work has been conducted under the general topic of deep learning in a variety of application domains. In particular, these include several excellent surveys of global deep learning models, techniques, trends, and applications [4,180,182], a survey of deep learning algorithms in the computer vision community [179], a survey that focuses directly on the problem of deep object detection and its recent advances [181], and a survey of deep learning models including the generative adversarial network and its related challenges and applications [19]. Nonetheless, the applications discussed in these surveys include only a single modality as a data source for data-driven learning. However, most modern machine learning applications involve more than one modality (e.g., visual and textual modalities), such as embodied question answering, vision-and-language navigation, etc. Therefore, it is of vital importance to learn more complex and cross-modal information from different sources, types, and data distributions. This is where deep multimodal learning comes into play.

From the early works of speech recognition to recent advances in language- and vision-based tasks, deep multimodal learning technologies have demonstrated significant progress in improving cognitive performance and interoperability of prediction models in a variety of ways. To date, deep multimodal learning has been the most important evolution in the field of multimodal machine learning using deep learning paradigm and multimodal big data computing environments. In recent years, many pieces of research based on multimodal machine learning have been proposed [37], but to the best of our knowledge, there is no recent work that directly addresses the latest advances in deep multimodal learning particularly for the computer vision community. A thorough review and synthesis of existing work in this domain, especially for researchers pursuing this topic, is essential for further progress in the field of deep learning. However, there is still relatively little recent work directly addressing this research area [32–37]. Since multimodal learning is not a new topic, there is considerable overlap between this work and the surveys of [32–37], which needs to be highlighted and discussed.

Recently, the valuable works of [32,33] considered several multimodal practices that apply only to specific multimodal use cases and applications, such as emotion recognition [32], human activity and context recognition [33]. More specifically, they highlighted the impact of multimodal feature representation and multilevel fusion on system performance and the state-of-the-art in each of these application areas.

Furthermore, some cutting-edge works [34,36] have been proposed in recent years that address the mechanism of integrating and fusing multimodal representations inside deep learning architectures by showing the reader the possibilities this opens up for the artificial intelligence community. Likewise, Guo et al. [35] provided a comprehensive overview of deep multimodal learning frameworks and models, focusing on one of the main challenges of multimodal learning, namely multimodal representation. They summarized the main issues, advantages, and disadvantages for each framework and typical model. Another excellent survey paper was recently published by Baltrušaitis et al. [37], which reviews recent developments in multimodal machine learning and expresses them in a general taxonomic way. Here, the authors identified five levels of multimodal data combination: representation, translation, alignment, fusion, and co-learning. It is important to note here that, unlike our survey, which focused primarily on computer vision tasks, the study published by Baltrušaitis et al. [37] was aimed mainly at both the natural language processing and computer vision communities. In this article, we reviewed recent advances in deep multimodal learning and organized them into six topics: multimodal data representation, multimodal fusion (i.e., both traditional and deep learning-based schemes), multitask learning, multimodal alignment, multimodal transfer learning, and zero-shot learning. Beyond the above work, we focused primarily on cutting-edge applications of deep multimodal learning in the field of computer vision and related popular datasets. Moreover, most of the papers we reviewed are recent and have been published in high-quality conferences and journals such as the visual computer, ICCV, and CVPR. A comprehensive overview of multimodal technologies—their limitations, perspectives, trends, and challenges—is also provided in this article to deepen and improve the understanding of the main directions for future progress in the field. In summary, our survey is similar to the closest works [35,37], which discuss recent advances in deep multimodal learning with a special focus on computer vision applications. The surveys we discussed are summarized in Table 1.

3 Deep multimodal learning architectures

In this section, we discuss deep multimodal learning and its main algorithms. To do so, we first briefly review the history of deep learning and then focus on the main motivations behind this research to answer the question of how to reduce heterogeneity biases across different modalities. We then outline the perspective of multimodal representation and what distinguishes it from the unimodal space. We next introduce recent approaches for combining modalities. Next, we highlight the difference between multimodal learning and multitask learning. Finally, we discuss multimodal alignment, multimodal transfer learning, and zero-shot learning in detail in Sects. 3.6, 3.7, and 3.8, respectively.

3.1 Brief history of deep learning

Historically, artificial neural networks date back to the 1950s and the efforts of psychologists to gain a better understanding of how the human brain works, including the work of F. Rosenblat [8]. In 1960, F. Rosenblat [8] proposed a perceptron as part of supervised learning algorithms that is used to compute a set of activations, meaning that for a given neuron and input vector, it performs the sum weighted by a set of weights, adds a bias, and applies an activation function. An activation function (e.g., sigmoid, tanH, etc.), also called nonlinearity, uses the derived patterns to perform its nonlinear transformation. As a deep variant of the perceptron, a multilayer perceptron, originally designed by [9] in 1986, is a special class of feed-forward neural networks. Structurally, it is a stack of single-layer perceptrons. In other words, this structure gives the meaning of "deep" that a network can be defined by its depth (i.e., the number of hidden layers). Typically, a multilayer perceptron with one or two hidden layers does not require much data to learn informative features due to the reduced number of parameters to be trained. A multilayer perceptron can be considered as a deep neural network if the number of hidden layers is greater than one, as confirmed by [10,11]. In this regard, many more advances in the

Table 1 Summary of reviewed deep multimodal learning surveys

Refs.	Year	Publication	Scope	Multimodality?
[4]	2015	Nature	A comprehensive overview of deep learning and related applications	✗
[19]	2018	IEEE Signal Processing Magazine	An overview of generative adversarial networks and related challenges in their theory and application	✗
[179]	2016	Neurocomputing	A review of deep learning algorithms in computer vision for image classification, object detection, image retrieval, semantic segmentation and human pose estimation	✗
[180]	2018	IEEE Access	A survey of deep learning: platforms, applications and trends	✗
[181]	2019	arXiv	A survey of deep learning and its recent advances for object detection	✗
[182]	2018	ACM Comput. Surv.	A survey of deep learning: algorithms, techniques, and applications	✗
[32]	2019	Book	A survey on multimodal emotion detection and recognition	✓
[33]	2018	Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies	A survey on multimodal deep learning for activity and context detection	✓
[34]	2017	IEEE Signal Processing Magazine	A survey of recent progress and trends in deep multimodal learning	✓
[35]	2019	IEEE Access	A comprehensive survey of deep multimodal learning and its frameworks	✓
[36]	2015	Proceedings of the IEEE	A comprehensive survey of methods, challenges, and prospects for multimodal data fusion	✓
[37]	2017	IEEE Transactions on Pattern Analysis and Machine Intelligence	A survey and taxonomy on multimodal machine learning algorithms	✓

field are likely to follow, such as the convolutional neural networks of LeCun et al. [21] in 1998 and the spectacular deep network results of Krizhevsky et al. [7] in 2012, opening the door to many real-world domains including computer vision.

3.2 Motivation

Recently, the amount of visual data has exploded due to the widespread use of available low-cost sensors, leading to superior performance in many computer vision tasks (see Fig. 1). Such visual data can include still images, video sequences, etc., which can be used as the basis for constructing multimodal models. Unlike the static image, the video stream provides a large amount of meaningful information that takes into account the spatiotemporal appearance of successive frames, so it can be easily used and analyzed for various real-world use cases, such as video synthesis and description [68], and facial expression recognition [123]. The spatiotemporal concept refers to the temporal and spatial processing of a series of video sequences with variable duration. In multimodal learning analytics, the audio-visual-textual features are extracted from a video sequence to learn joint features covering the three modalities. Efficient learning of large datasets at multiple levels of representation leads to faster content analysis and recognition of the millions of videos produced daily. The main reason for using multimodal data sources is that it is possible to extract complementary and richer information coming from multiple sensors, which can provide much more optimistic results than a single input. Some monomodal learning systems have significantly increased their robustness and accuracy, but in many use cases, there are shortcomings in terms of the universality of different feature levels and inaccuracies due to noise and missing concepts. The success of deep multimodal learning techniques has been driven by many factors that have led many researchers to adopt these methods to improve model performance. These factors include large volumes of widely usable multimodal datasets, more powerful computers with fast GPUs, and high-quality feature representation at multiple scales. Here, a practical challenge for the deep learning community is to strengthen correlation and redundancy between modalities through typical models and powerful mechanisms.

3.3 Multimodal representation

Multi-sensory perception primarily encompasses a wide range of interacting modalities, including audio and video.

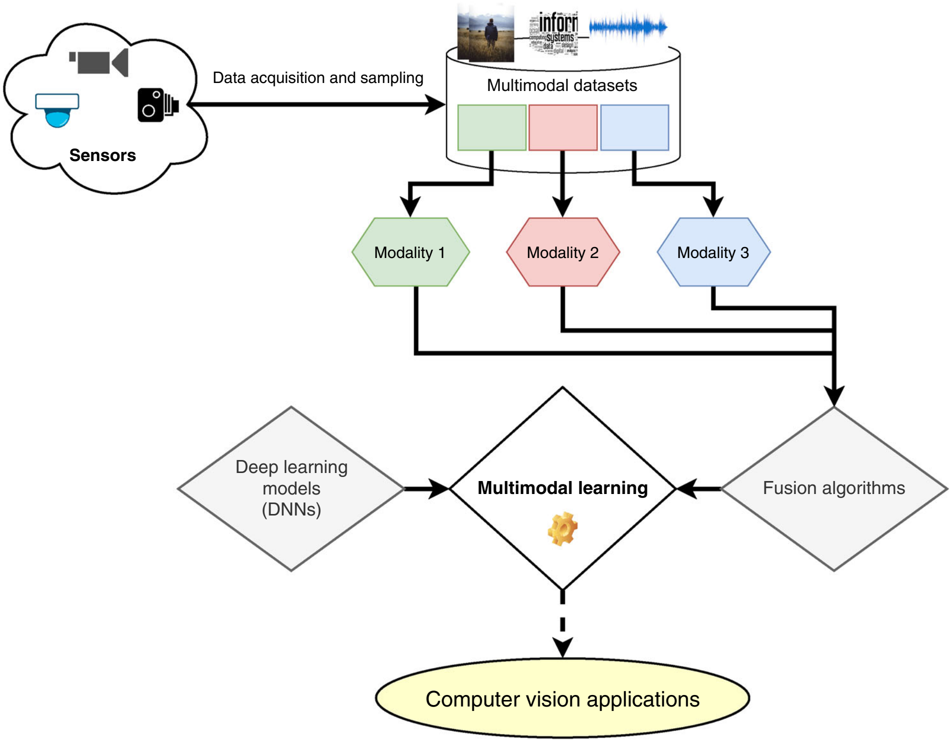

Fig. 1 An example of a multimodal pipeline that includes three different modalities

For simplicity, we consider the following temporal multimodal problem, where both audio and video modalities are exploited in a video recognition task (emotion recognition). First, let us consider two input streams of different modalities: $X_a = \left\{\chi_1^n, \ldots, \chi_T^n\right\}$ and $X_v = \left\{\chi_1^m, \ldots, \chi_T^m\right\}$, where χ_t^n and χ_t^m refer to the n- and m-dimensional feature vectors of the X_a and X_v modalities occurring at time t, respectively. Next, we combine the two modalities at time t and consider the two unimodal output distributions at different levels of representations. Given ground truth labels $Z = \left\{Z^1, \ldots, Z^T\right\}$, we aim here to train a multimodal learning model M that maps both X_a and X_v into the same categorical set of Z. Each parameter of the input audio stream χ_a^T and video stream χ_v^T is synchronized differently in time and space, where $\chi_a^T \in \mathbb{R}^i$ and $\chi_v^T \in \mathbb{R}^j$, respectively. Here, we can construct two separate unimodal networks from X_a and X_v, denoted, respectively, by N_a and N_v, where $N_a : X_a \rightarrow Y$, $N_v : X_v \rightarrow Y$, and $M = N_a \bigoplus N_v$. Y denotes the predicted class label of the training samples generated by the output of the constructed networks and $\bigoplus$ indicates the fusion operation. The generated multimodal network M can then recognize the most discriminating patterns in the streaming data by learning a common representation that integrates relevant concepts from both modalities. Figure 2 shows a schematic diagram of the application of the described multimodal problem to the video emotion recognition task.

Therefore, it is necessary to consider the extent to which any such dynamic entity will be able to take advantage of this type of information from several redundant sources. Learning multimodal representation from heterogeneous signals poses a real challenge for the deep learning community. Typically, inter- and intra-modal learning involves the ability to represent an object of interest from different perspectives, in a complementary and semantic context where multimodal information is fed into the network. Another crucial advantage of inter- and intra-modal interaction is the discriminating power of the perceptual model for multisensory stimuli by exploiting the potential synergies between modalities and their intrinsic representations [112]. Furthermore, multimodal learning involves a significant improvement in perceptual cognition, as many of our senses are involved in the process of treatment information from several modalities. Nevertheless, it is essential to learn how to interpret the input signals and summarize their multimodal nature to construct aggregate feature maps across multiple dimensions. In the multimodality theory, obtaining contextual representation from more than one modality has become a vital challenge, which has been termed in this study as the multimodal representation.

Typically, monomodal representation involves a linear or nonlinear mapping of an individual input stream (e.g., image, video, or sound, etc.) into a high-level semantic representation. The multimodal representation leverages the correlation power of each monomodal sensation by aggregating their spatial outputs. Thus, the deep learning model must be adapted to accurately represent the structure and representation space of the source and target modality. For example,

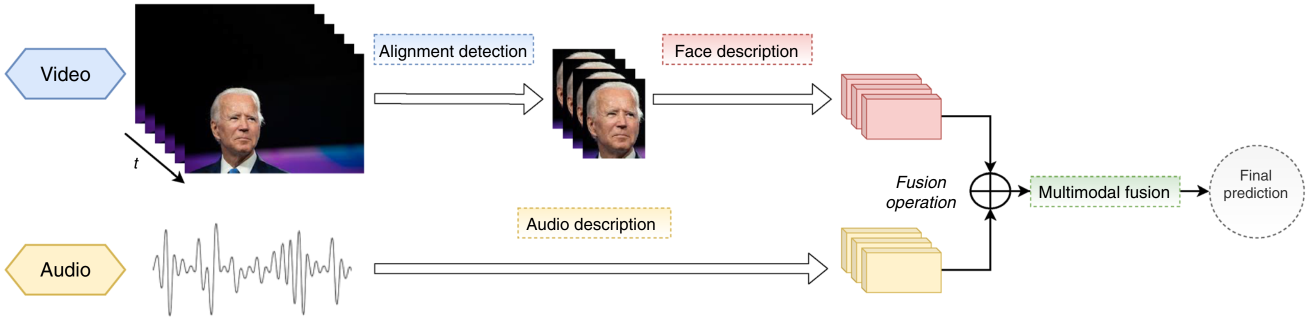

Fig. 2 A schematic illustration of the method used: The visual modality (video) involves the extraction of facial regions of interest followed by a visual mapping representation scheme. The obtained representations are then temporally fused into a common space. Additionally, the audio descriptions are also generated. The two modalities are then combined using a multimodal fusion operation to predict the target class label (emotion) of the test sample

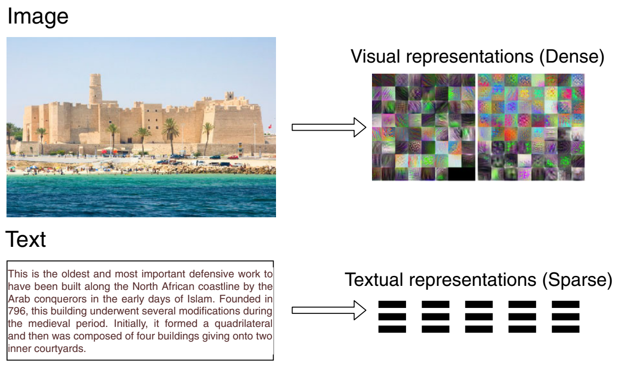
Image
Visual representations (Dense)
Text
Textual representations (Sparse)

This is the oldest and most important defensive work to have been built along the North African coastline by the Arab conquerors in the early days of Islam. Founded in 796, this building underwent several modifications during the medieval period. Initially, it formed a quadrilateral and then was composed of four buildings giving onto two inner courtyards.

Fig. 3 Difference between visual and textual representation

a 2D image may be represented by its visual patterns, making it difficult to characterize this data structure using natural modality or other non-visual concepts. As shown in Fig. 3, the textual representation (i.e., a word embedding) is very sparse when compared to the image one, which makes it very challenging to combine these two different representations into a unified model. As another example, when the driver of a car is driving autonomously, he probably has a LiDAR camera and other embedded sensors (e.g., depth sensors, etc) [81] to perceive his surroundings. Here, poor weather conditions can affect the visual perception of the environment. Moreover, the high dimensionality of the state space poses a major challenge, since the vehicle can mobilize in both structured and unstructured locations. However, an RGB image is encoded as a discrete space in the form of grid pixels, making it difficult to combine visual and non-visual cues. Therefore, learning a joint embedding is crucial for exploiting the synergies of multimodal data to construct shared representation spaces. This implies the emphasis on multimodal fusion approaches, which will be discussed in the next subsection.

3.4 Fusion algorithms

The most critical aspect of the combinatorial approach is the flexibility to represent data at different levels of abstraction. By using an intermediate formalism, the learned information can be combined into two or more modalities for a particular hypothesis. In this subsection, we describe common methods for combining multiple modalities, ranging from the conventional to the modern methods.

3.4.1 Conventional methods

3.4.1.1 Typical techniques based

To improve the generalization performance of complex cognitive systems, it is necessary to capture and fuse an appropriate set of informative features from multiple modalities using typical techniques. Traditionally, they range from early to hybrid fusion schemes (see Fig. 4):

– *Early fusion:* low-level features that are directly extracted from each modality will be fused before being classified.
– *Late fusion:* also called "decision fusion", which consists of classifying features extracted from separate modalities before fusing them.
– *Hybrid fusion:* also known as "intermediate fusion", which consists of combining multimodal features of early and late fusion before making a decision.

Feature-level fusion (i.e., early fusion) provides a richness of information from heterogeneous data. The extracted features often lack homogeneity due to the diversity of modalities and disparities in their appearance. Also, this fusion process can generate a single large representation that can lead to prediction errors. In the case of a late fusion, such techniques as majority vote [38] and low-rank multimodal fusion [39] may be used to aggregate the final prediction scores of several classifiers. Thus, each modality independently takes the

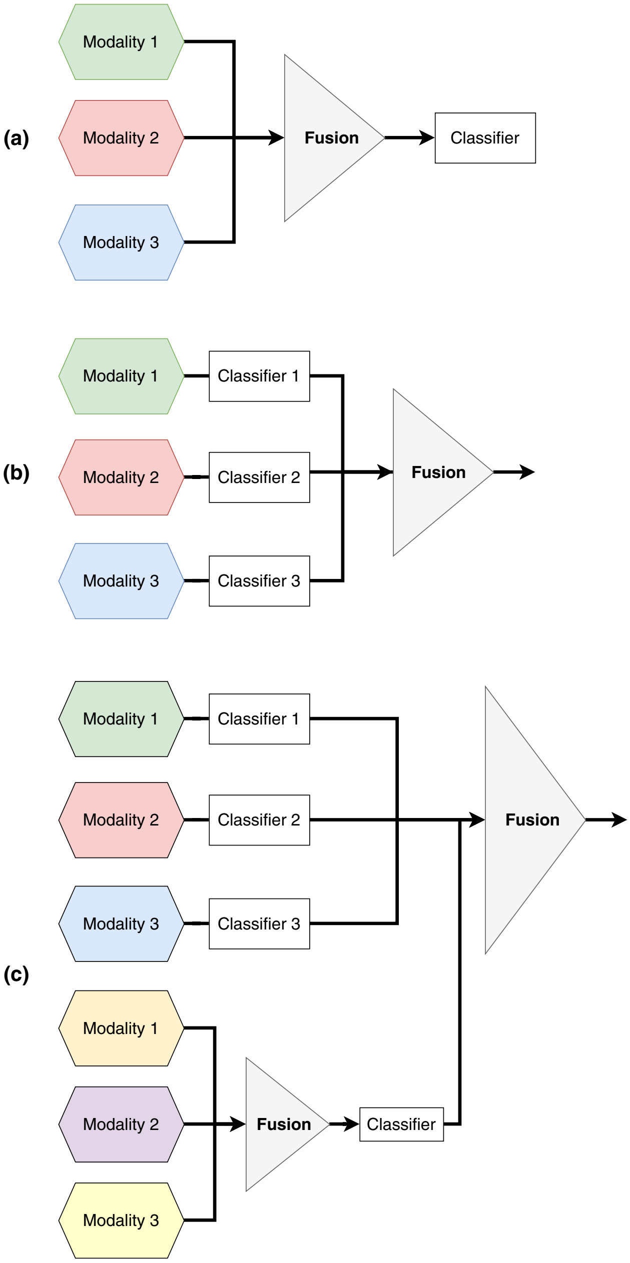

Fig. 4 Conventional methods for multimodal data fusion: **a** Early fusion, **b** Late fusion, **c** Hybrid fusion

decision, which can reduce the overall performance of the integration process. In the case of intermediate fusion, the spatial combination of intermediate representations of the different data streams usually produced with varying scales and dimensions, making them more challenging to merge. To overcome this challenge, the authors of [124] designed a simple fusion scheme, called multimodal transfer module (MMTM), to transfer and hierarchically aggregate shared knowledge from multiple modalities in CNN networks.

3.4.1.2 Kernel based

Since a long time ago, the support vector machine [40] classifier has been introduced as a learning algorithm for a wide range of classification tasks. Indeed, SVM is one of the most popular linear classifiers that are based on learning a single kernel function through the handling of linear tasks, such as discrimination and regression problems. The main idea of an SVM is to separate the feature space into two classes of data with a hard margin. Kernel-based methods are among the most commonly used techniques for performing fusion due to their proven robustness and reliability. For more details, we invite the reader to consult the work of Gönen et al. [41] that focused on the taxonomy of multi-kernel learning algorithms. These kernels are intended to make use of the similarities and discrepancies across training samples as well as a wide variety of data sources. In other words, these modular learning methods are used for multimodal data analysis. Recently, a growing number of studies have focused, in particular, on the potential of these kernels for multi-source-based learning for improving performance. In this sense, a wide range of kernel-based methods have been proposed to summarize information from multiple sources using a variety of input data. In this regard, Gönen et al. [41] pioneered multiple kernel learning (MKL) algorithms that seek to combine multimodal data that have distinct representations of similarity. MKL is the process of learning a classifier through multiple kernels and data sources. Also, it aims to extract the joint correlation of several kernels in a linear or nonlinear manner. Similarly, Aiolli et al. [42] proposed the MKL-based algorithm, called EasyMKL, which combines a series of kernels to maximize the segregation of representations and extract the strong correlation between feature spaces to improve the performance of the classification task. An alternative model, called convolutional recurrent multiple kernel learning (CRMKL), based on the MKL framework for emotion recognition and sentiment analysis is reported by Wen et al. [43]. In [43], the MKL algorithm is used to combine multiple features that are extracted from deep networks.

3.4.1.3 Graphical models based

One of the most common probabilistic graphical models (PGMs) includes the hidden Markov model (HMM) [44]. It is an unsupervised and generative model. It has a series of potential states and transition probabilities. In the Markov chain, the transition from one state to another leads to the generation of observed sequences in which the observations are part of a state set. A transition formalizes how it is possible to move from one state to another and for each one there is a probability distribution of being borrowed. The states are hidden, but the first state generates a visible state from a given one. The main property of Markov chains is that the probabilities depend only on the previous state of the model.

In HMM, a kind of generalization of mixing densities defined by each state is involved, as confirmed by Ghahramani et al. [45]. Specifically, Ghahramani et al. [45] introduced the factorial HMM (FHMM) which consists of combining the state transition matrix of HMMs with the distributed representations of vector quantizer (VQ) [46]. According to [46], VQ is a conventional technique for quantifying and generalizing dynamic mixing models. FHMM addresses the limited representational power of the latent variables of HMM by presenting the hidden state under a certain weighted appearance. Likewise, Gael et al. [47] proposed the non-parametric FHMM, called iFHMM, by introducing a new stochastic process for latent feature representation of time series.

In summary, the PGM model can be considered a robust tool for generating missing channels by learning the most representative inter-modal features in an unsupervised manner. One of the drawbacks of the graphical model is the high cost of the training and inference process.

3.4.1.4 Canonical correlation analysis based

In general, a fusion scheme can construct a single multimodal feature representation for each processing stage. However, it is also straightforward to place constraints on the extracted unimodal features [37]. Canonical correlation analysis (CCA) [201] is a very popular statistical method that attempts to maximize the semantic relationship between two unimodal representations so that complex nonlinear transformations of the two data perspectives can be effectively learned. Formally, it can be formulated as follows:

$$\left(v1^{*}, v2^{*}\right) = \underset{v1,v2}{\operatorname{argmax}}\; corr(v1^{T}X1, v2^{T}X2), \tag{1}$$

where $X1$ and $X2$ stand for unimodal representations, $v1$ and $v2$ for two vectors of a given length, and *corr* for the correlation function. A deep variant of CCA can also be used to maximize the correlation between unimodal representations, as suggested by the authors of [202]. Similarly, Chandar et al. [203] proposed a correlation neural network, called CorrNet, which is based on a constrained encoder/decoder structure to maximize the correlation of internal representations when projected onto a common subspace. Engilberge et al. [204] introduced a weaker constraint on the joint embedding space using a cosine similarity measure. Besides, Shahroudy et al. [205] constructed a unimodal representation using a hierarchical factorization scheme that is limited to representing redundant feature parts and other completely orthogonal parts.

3.4.2 Deep learning methods

3.4.2.1 Deep belief networks based

Deep belief network (DBN) is part of the graphical generative deep model [15]. They form a deeper variant of the restricted Boltzmann machine (RBM) by combining it together. In other words, a DBN consists of stacking a series of RBM where the hidden layer of the first RBM is the visible layer of the higher hierarchies. Structurally, a DBN model has a dense structure similar to that of a shallow multilayer perceptron. The first RBM is designed to systematically reconstruct its input signal in which its hidden layer will be handled as the visible layer for the second one. However, all hidden representations are learned globally at each level of DBN. Note that DBN is one of the strongest alternatives to overcome the vanishing gradient problem through a stack of RBM units. Like a single RBM, DBN involves discovering latent features in the raw data. It can be further trained in a supervised fashion to perform the classification of the detected hidden representations.

Compared to other supervised deep models, DBN requires only a very small set of labeled data to perform weight training, which leads to a high level of usefulness in many multimodal tasks. For instance, Srivastava et al. [206] proposed a multimodal generative model based on the concept of deep Boltzmann machine (DBM) which learns a set of multimodal features by filling in the conditional distribution of data on a space of multimodal inputs such as image, text, and audio. Specifically, the purpose of training a multimodal DBN model is to improve the prediction accuracy of both unimodal and multimodal systems by generating a set of multimodal features that are semantically similar to the original input data so that they can be easily derived even if some modalities are missing. Figure 5 illustrates a multimodal DBN architecture that takes as input two different modalities (image and text) with different statistical distributions to map the original data from a high-dimensional space to a high-level abstract representation space. After extracting the high-level representation from each modality, an RBM network is then used to learn the joint distribution. The image and text modalities are modeled using two DBMs, each consisting of two hidden layers. Formally, the joint representation can be expressed as follows:

$$P(v_i|\theta) = \sum_{h^1,h^2} P(v_i, h^1, h^2|\theta), \tag{2}$$

where v_i refers to the input visual and textual modalities, θ to the network parameters, and h to the hidden layer of each modality.

In a multimodal context, the advantage of using multimodal DBN models lies in their sensitivity and stability in both supervised, semi-supervised and unsupervised learning protocols. These models allow for better modeling of very complex and discriminating patterns from multiple input modalities. Despite these advantages, these models have a few limitations. For instance, they largely ignore the spatiotemporal cues of multimodal data streams, making the inference process computationally intensive.

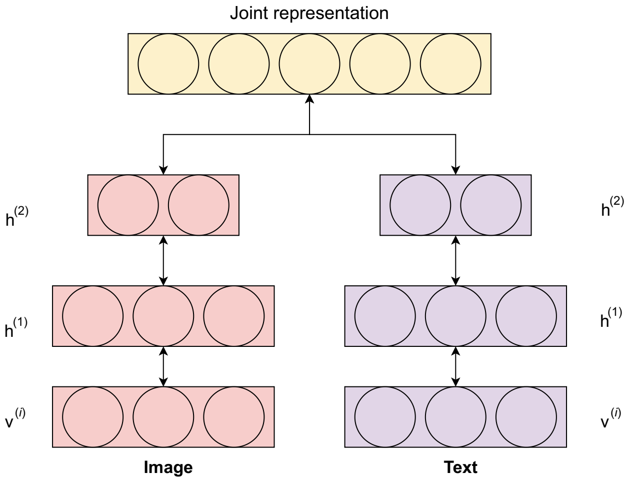

Fig. 5 Structure of a bimodal DBN

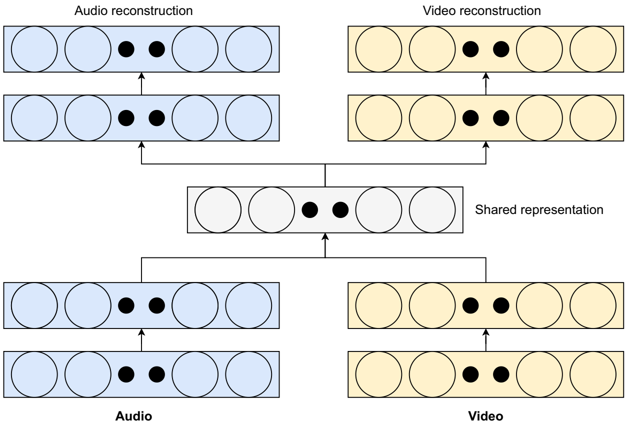

Fig. 6 Structure of a bimodal AE

3.4.2.2 Deep autoencoders based

Deep autoencoders (DAEs) [207] are a class of unsupervised neural networks that are designed to learn a compressed representation of input signals. Conceptually, they consist of two coupled modules: the encoding module (encoder) and the decoding module (decoder). On the one hand, the encoding module consists of several processing layers to map high-dimensional input data into a low-dimensional space (i.e., latent space vectors). On the other hand, the decoding module takes these latent representations as input and decodes them in order to reconstruct the input data. These models have recently drawn attention from the multimodal learning community due to their great potential for reducing data dimensionality and, thus, increasing the performance of training algorithms. For instance, Bhatt et al. [208] proposed a DAE-based multimodal data reconstruction scheme that uses knowledge from different modalities to obtain robust unimodal representations and projects them onto a common subspace. Similar to the work of Bhatt, Liu et al. [209] proposed the integration of multimodal stacked contractive AEs (SCAEs) to learn cross-modality features across multiple modalities even when one of them is missing, intending to minimize the reconstruction loss function and avoid the overfitting problem. The loss function can be formulated as follows:

$$\text{Loss}_{\text{reconst}} = \sum_{i=1}^{M} (\|x_i - \hat{x}_i\|_2^2) + \|y_i - \hat{y}_i\|_2^2). \tag{3}$$

Here, (x_i, y_i) denotes a pair of two inputs, and $(\hat{x}_i, \hat{y}_i)$ represent their reconstructed outputs.

Several other typical models based on stacked AEs (SAEs) have been proposed to learn coherent joint representations across modalities. For example, the authors of [210–212] designed multimodal systems based on SAEs, where the encoder side of the architecture represents and compresses each unimodal feature separately, and the decoder side constructs the latent (shared) representation of the inputs in a unsupervised manner. Figure 6 shows the coupling mechanism of two separate AEs (bimodal AE) for both modalities (audio and video) into a jointly shared representation hierarchy where the encoder and decoder components are independent of each other. As a powerful tool for feature extraction and dimensionality reduction, the DAE aims to learn how to efficiently represent manifolds where the training data is unbalanced or lacking. One of the main drawbacks of DAEs is that many hidden parameters have to be trained, and the inference process is time-consuming. Moreover, they also miss some spatiotemporal details in multimodal data.

3.4.2.3 Convolutional neural networks based

Convolutional neural networks (CNNs or ConvNets) are a class of deep feed-forward neural networks whose main purpose is to extract spatial patterns from visual input signals [20,22]. More specifically, such models tend to model a series of nonlinear transformations by generating very abstract and informative features from highly complex datasets. The main properties that distinguish CNNs from other models include their ability to capture local connectivity between units, to share weights across layers, and to block a sequence of hidden layers [4]. The architecture is based on hierarchical filtering operations, i.e., using convolution layers followed by activation functions, etc. Once the convolution layers are linearly stacked, the growth of the receptive field size (i.e., kernel size) of the neural layers can be simulated by a max-pooling operation, which implies a reduction in the spatial size of the feature map. After applying a series of convolution and pooling operations, the hidden representation learned from the model must be predicted. For this purpose, at least one fully connected layer (also called dense layer) is used that concatenates all previous activation maps.

Since its introduction by Krizhevsky et al. [7] in 2012, the CNN model has been successfully applied to a wide range of multimodal applications, such as image dehazing [239,240] and human activity recognition [241]. An adaptive multimodal mapping between two visual modalities (e.g., images and sentences) typically requires strong representations of the individual modalities [213]. In particular, CNNs have demonstrated powerful generalization capabilities to learn how to represent visual appearance features from static data. Recently, with the advent of robust and low-cost RGB-D sensors such as the Kinect, the computer vision community has turned its attention to integrating RGB images and corresponding depth maps (2.5D) into multimodal architectures as shown in Fig. 7. For instance, Couprie et al. [214] proposed a bimodal CNN architecture for multiscale feature extraction from RGB-D datasets, which are taken as four-channel frames (blue, green, red, and depth). Similarly, Madhuranga et al. [215] used CNN models for video recognition purposes by extracting silhouettes from depth sequences and then fusing the depth information with audio descriptions for activity of daily living (ADL) recognition. Zhang et al. [217] proposed to use multicolumn CNNs to extract visual features from the face and eye images for the gaze point estimation problem. Here, the regression depth of the facial landmarks is estimated from the facial images and the relative depth of facial keypoints is predicted by global optimization. To perform image classification directly, the authors of [217,218] suggested the possibility of using multi-stream CNNs (i.e., two or more stream CNNs) to extract robust features from a final hidden layer and then project them onto a common representation space. However, the most commonly adopted approaches involve concatenating a set of pre-trained features derived from the huge ImageNet dataset to generate a multimodal representation [216].

Formally, let f_i^j be the feature map of j modalities and i be the current spatial location, where $j = \{1, 2, \ldots, N\}$. As shown in Fig. 7, in our case $N = 2$, since the feature maps $FC2$ (RGB) and $FC2$ (D) were taken separately from the RGB and depth paths. The fused feature map F_i^{fusion}, which is a weighted sum of the unimodal representations, can be calculated as follows:

$$F_i^{\text{fusion}} = \sum_{j=1}^{N} w_i^j f_i^j. \tag{4}$$

Here, w_i^j denotes the weight vectors that can be computed as follows:

$$w_i^j = \frac{\exp(f_i^j)}{\sum_{k=1}^{N} \exp(f_i^k)}. \tag{5}$$

In summary, a multimodal CNN serves as a powerful feature extractor that learns local cross-modal features from visual modalities. It is also capable of modeling spatial cues from multimodal data streams with an increased number of parameters. However, it requires a large-scale multimodal dataset to converge optimally during training, and the inference process is time-consuming.

3.4.2.4 Recurrent neural networks based

Recurrent neural networks (RNNs) [12] are a popular type of deep neural network architectures for processing sequential data of varying lengths. They learn to map input activations to the next hierarchy level and then transfer hidden states to the outputs using the recurrent feedback, which gives them the capacity to learn useful features from the previous states, unlike other deep feedforward networks such as CNNs, DBNs, etc. It also can handle time series and dynamic media such as text and video sequences. By using the backpropagation algorithm, the RNN function takes an input vector and a previous hidden state as input to capture the temporal dependence between objects. After training, the RNN function is fixed at a certain level of stability and can then be used over time.

However, the vanilla RNN model is typically incapable of capturing long-term dependencies in sequential data since they have no internal memory. To this end, several popular variants have been developed to efficiently handle this constraint and the gradient vanishing problem with impressive results, including long short-term memory (LSTM) [13] and gated recurrent linear units (GRU) [14]. In terms of computational efficiency, GRU is a lightweight variant of LSTM since it can modulate the information flow without using its internal memory units.

In addition to their use for unimodal tasks, RNNs have proved useful in many multimodal problems that require modeling long- and short-range dependencies across the input sequence, such as semantic segmentation [219] and image captioning [220]. For instance, Abdulnabi et al. [219] proposed a multimodal RNN architecture designed for semantic scene segmentation using RGB and depth channels. They integrated two parallel RNNs to efficiently extract robust cross-modal features from each modality. Zhao et al. [220] proposed an RNN-based multimodal fusion scheme to generate captions by analyzing distributional correlations between images and sentences. Recently, several new multimodal approaches based on RNN variants have been proposed and have achieved outstanding results in many vision applications. For example, Li et al. [221] designed a GRU-based embedding framework to describe the content of an image. They used GRU to generate a description of variable length from a given image. Similarly, Sano et al. [222] proposed a multimodal BiLSTM for ambulatory sleep detec-

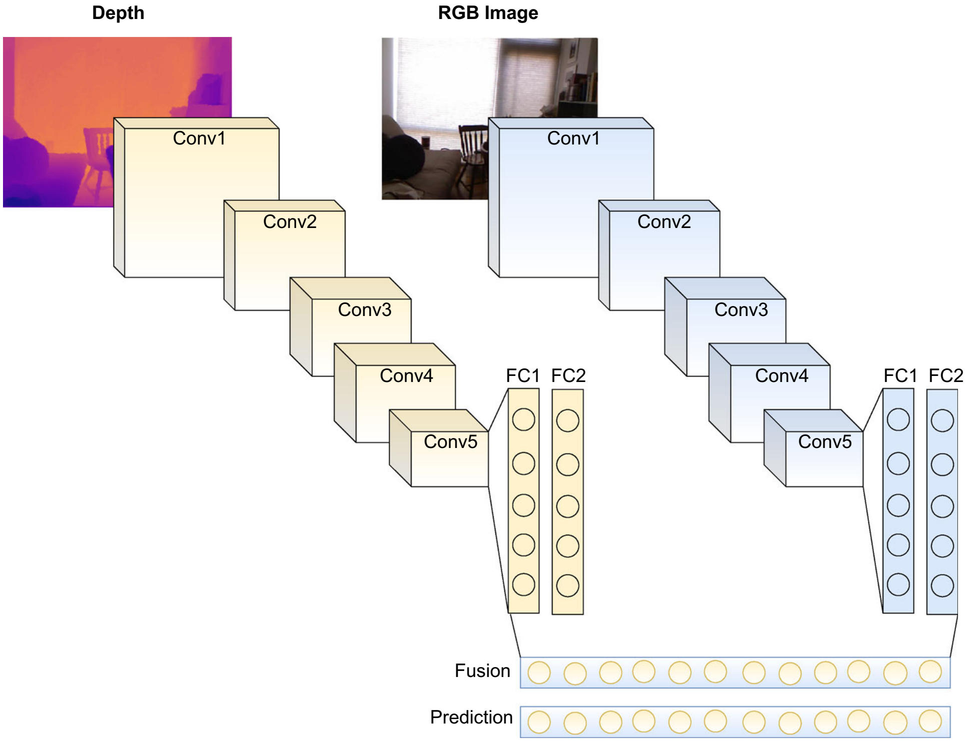

Fig. 7 Structure of a bimodal CNN

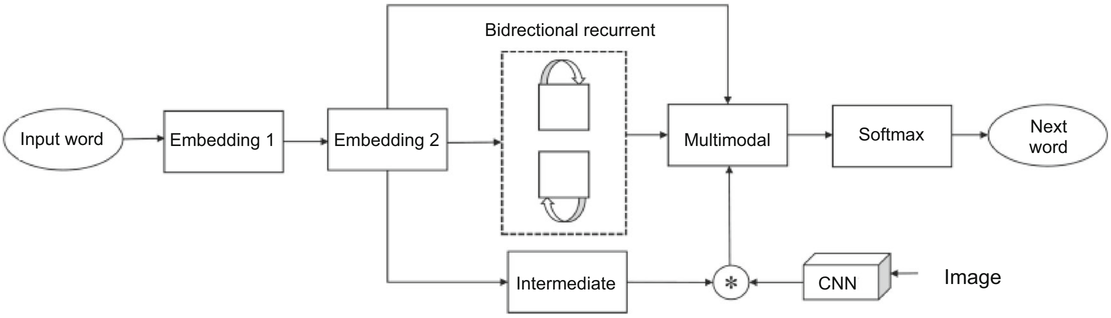

Fig. 8 A schematic illustration of bidirectional multimodal RNN (m-RNN) [223]

tion. In this case, BiLSTM was used to extract features from the wearable device and synthesize temporal information.

Figure 8 illustrates a multimodal m-RNN architecture that incorporates both word embeddings and visual features using a bidirectional recurrent mechanism and a pre-trained CNN. As can be seen, m-RNN consists of three components: a language network component, a vision network component, and a multimodal layer component. The multimodal layer here maps semantic information across sub-networks by temporally learning word embeddings and visual features. Formally, it can be expressed as follows:

$$m(t) = f(v_w.w(t),\ v_r.r(t),\ v_I.I), \tag{6}$$

where $f(.)$ denotes the activation function, w and r consist of the word embedding feature and the hidden states in both directions of the recurrent layer and I represent the visual features.

In summary, the multimodal RNN model is a robust tool for analyzing both short- and long-term dependencies of multimodal data sequences using the backpropagation algorithm. However, the model has a slow convergence rate due to the high computational cost in the hidden state transfer function.

3.4.2.5 Generative adversarial networks based

Generative adversarial networks (GANs) are part of deep generative architectures, designed to learn the data distribution through the adversarial learning. Historically, they were first developed by Goodfellow et al. [16], which demonstrated the ability to generate realistic and reasonably impractical representations from noisy data domains. Structurally, GAN is a unified network consisting of two sub-networks, a generator network (G) and a discriminator network (D), which interact continuously during the learning process. The principle of its operation is as follows: The generator network takes as input the latent distribution space (i.e., a random noise (z)) and generates an artificial sample. The discriminator takes the true sample and those generated by the generator and tries to predict whether the input sample is true (false) or not. Hence, it is a binary classification problem, where the output must be between 0 (generated) and 1 (true). In other words, the generator's main task is to generate a realistic image, while the discriminator's task is to determine whether the generated image is true or false. Subsequently, they should use an objective function to represent the distance between the distribution of generated samples (p_z) and the distribution of real ones (p_{data}). The adversarial training strategy consists of using a minimax objective function $V(G, D)$, which can be expressed as follows:

$$\min_G \max_D \; V(G,D) = \mathbb{E}_{x \sim p_{\text{data}}} |\log(D(x))| + \mathbb{E}_{z \sim p_z} |\log(1 - D(x))| \tag{7}$$

Since their development in 2014, generative adversarial training algorithms have been widely used in various unimodal applications such as scene generation [17], image-to-image translation [18], and image super-resolution [224, 225]. To obtain the latest advances in super-resolution algorithms for a variety of remote sensing applications, we invite the reader to refer to the excellent survey article by Rohith et al. [226].

In addition to its use in unimodal applications, the generative adversarial learning paradigm has recently been widely adopted in multimodal arenas, where two or more modalities are involved, such as image captioning [227] and image retrieval [228]. In recent years, GAN-based schemes have been receiving a lot of attention and interest in the field of multimodal vision. For example, Xu et al. [229] proposed a fine-grained text-image generation framework using an attentional GAN model to create high-quality images from text. Similarly, Huang et al. [230] proposed an unsupervised image-to-image translation architecture that is based on the idea that the image style of one domain can be mapped into the styles of many domains. In [231], Toriya et al. addressed the task of image alignment between a pair of multimodal images by mapping the appearance features of the first modality to the other using GAN models. Here, GANs were used as a means to apply keypoint-mapping techniques to multimodal images. Figure 9 shows a simplified diagram of a multimodal GAN.

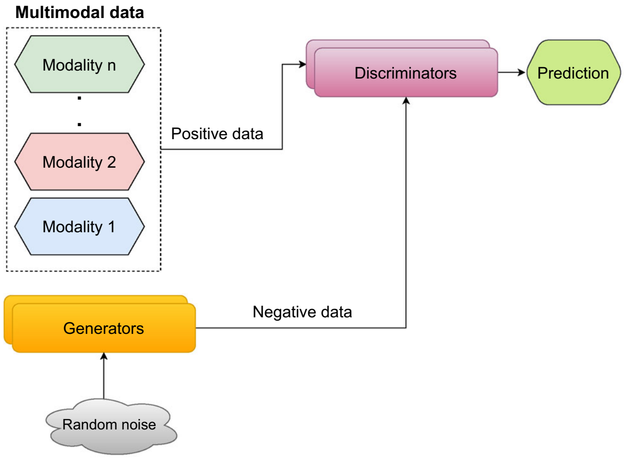

Fig. 9 A schematic illustration of multimodal GAN

In summary, unsupervised GAN is one of the most powerful generative models that can address scenarios where training data is lacking or some hidden concepts are missing. However, it is extremely tricky to train the network when generating discrete distributions, and the process itself is unstable compared to other generative networks. Moreover, the function that this network seeks to optimize is an adversarial loss function without any normalization.

3.4.2.6 Attention mechanism based

In recent years, the attention mechanism (AM) has become one of the most challenging tasks in computer vision and machine translation [232]. The idea of AM is to focus on a particular position in the input image by computing the weighted sum of feature vectors and mapping them into a final contextual representation. In other words, it learns how to reduce some irrelevant attributes from a set of feature vectors. In the multimodal analysis, an attentional model can be designed to combine multiple modalities, each with its internal representation (e.g., spatial features, motion features, etc.). That is, when a set of features is derived from spatiotemporal cues, these variable-length vectors are semantically combined into a single fixed-length vector. Furthermore, an AM can be integrated into RNN models to improve the generalization capability of the former by capturing the most representative and discriminating patterns from heterogeneous datasets. A formalism for integrating an AM into the basic RNN model was developed by Bahdanau et al. [1]. Since the encoding side of an RNN generates a fixed-length feature vector from its input sequence, this can lead to very tedious and time-consuming parameter tuning. Therefore, the AM acts as a contextual bridge between the encoding and decoding sides of an RNN to pay attention only to a particular position in the input representation.

Consider as an example of neural machine translation [1] (see Fig. 10), where an encoder is trained to map a sequence of input vectors $x = (x_1, \ldots, x_{T_x})$ into a fixed-length vector c and a decoder to predict the next word $(y_{t'})$ from previous predicted ones $\{y_1, \ldots, y_{t'-1}\}$. Here, c refers to an encoded vector produced by a sequence of hidden states that can be expressed as follows:

$$c = q(h_1, \ldots, h_{T_x}), \tag{8}$$

where q denotes some activation functions. The hidden state h_t $(h_t \in \mathbb{R}^n)$ at time step t can be formulated as:

$$h_t = f(x_t, \ldots, h_{t-1}). \tag{9}$$

The context vector c_i can then be computed as a weighted sum of a sequence of annotations $\{h_1, \ldots, h_{T_x}\}$ as follows:

$$c_i = \sum_{j=1}^{T_x} \sigma_{ij} h_j, \tag{10}$$

where the alignment weight σ_{ij} of each annotation h_j can be calculated as:

$$\sigma_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \tag{11}$$

and $e_{ij} = a(s_{i-1}, h_j)$. s_{i-1} is the hidden state at the $(i-1)$-th position of the input sequence.

Since its introduction, the AM has gained wide adoption in the computer vision community due to its spectral capabilities for many multimodal applications such as video description [233,234], salient object detection [235], etc. For example, Hori et al. [233] proposed a multimodal attention framework for video captioning and sentence generation based on the encoder–decoder structure using RNNs. In particular, the multimodal attention model was used as a way to integrate audio, image, and motion features by selecting the most relevant context vector from each modality. In [236], Yang et al. suggested the use of stacked attention networks to search for image regions that correlate with a query answer and identify representative features of a given question more precisely. More recently, Guo et al. [237] introduced a normalized variant of the self-attention mechanism, called normalised self-attention (NSA), which aims to encode and decode the image and caption features and normalize the distribution of internal activations during training.

In summary, the multimodal AM provides a robust solution for cross-modal data fusion by selecting the local fine-grained salient features in a multidimensional space and filtering out any hidden noise. However, the only weakness of AM is that the training algorithm is unstable, which may affect the predictive power of the decision-making system.

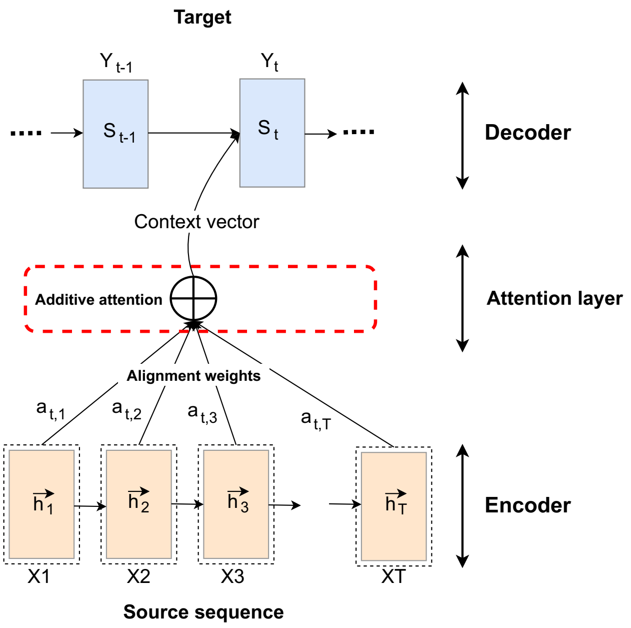

Fig. 10 A schematic illustration of the attention-based machine translation model

Furthermore, the number of parameters to be trained is huge compared to other deep networks such as RNNs, CNNs, etc.

3.5 Multitask learning

More recently, multitask learning (MTL) [108,109] has become an increasingly popular topic in the deep learning community. Specifically, the MTL paradigm frequently arises in a context close to multimodal concepts. In contrast to single-task learning, the idea behind this paradigm is to learn a shared representation that can be used to respond to several tasks in order to ensure better generalizability. Although, there are some similarities between the fusion methods discussed in Sect. 3.4 and the methods used to perform multi-tasks simultaneously. What they have in common is that the sharing of the structure between all tasks can be learned jointly to improve performance. The conventional typology of the MTL approach consists of two subtasks:

- Hard parameter sharing [110]: It consists of extracting a generic representation for different tasks using the same parameters. It is usually applied to avoid overfitting problems.
- Soft parameter sharing [111]: It consists of extracting a set of feature vectors and simultaneously drawing similarity relationships between them.

Figure 11 shows a meta-architecture for the two-task case. As can be seen, there are six intermediate layers in total, one shared input layer (*bottom*), two task-specific output layers

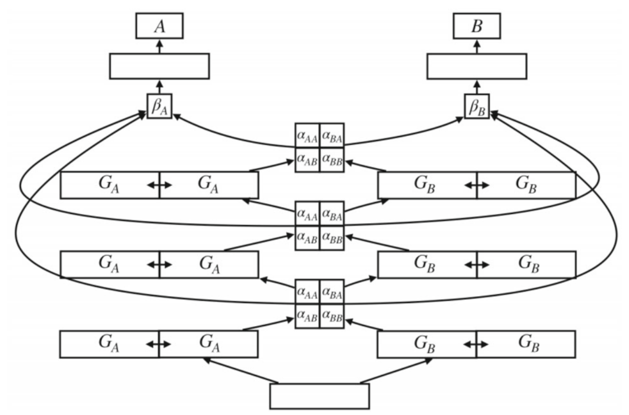

Fig. 11 A meta-architecture in the case of two tasks A and B [109]

(top), and three hidden layers per task, each divided into two G-subspaces. Typically, MTL contributes to the performance of the target task based on knowledge gained from auxiliary tasks.

3.6 Multimodal alignment

Multimodal alignment consists of linearly linking the features of two or more different modalities. Its areas of application include medical image registration [169], machine translation [1], etc. Specifically, multimodal image alignment provides a spatial mapping capability between images taken by sensors of different modalities, which may be categorized into feature-based [167,168] and patch-based [165,166] methods. Feature-based methods detect and extract a set of matching features that should be structurally consistent to describe their spatial patterns. Patch-based methods first split each image into local patches and then consider the similarity between them by computing their cross-correlation and combination. Generally, the alignment task can be divided into two subtasks: the attentional alignment task [170,171] and the semantic alignment task [172,173]. The attentional alignment task is based on the attentional mapping between the features of the input modality and the target one, while the semantic alignment task takes the form of an alignment method that directly provides alignment capabilities to a predictive model. The most popular use of semantic alignment is to create a dataset with associated labels and then generate a semantically aligned dataset. Both of these tasks have proven effective in multimodal alignment, where attentional alignment features are better able to take into account the long-term dependencies between different concepts.

3.7 Multimodal transfer learning

Typically, training a deep model from scratch requires a large amount of labeled data to achieve an acceptable level of performance. A more common solution is to find an efficient method that transfers knowledge already derived from another trained model onto a huge dataset (e.g., 1000k-ImageNet) [198]. Transfer learning (TL) [70] is one of the model regularization techniques that have proven their effectiveness for training deep models with a limited amount of available data and avoiding overfitting problems. Transferring knowledge from a pre-trained model associated with a sensory modality to a new task or similar domain facilitates the learning and fine-tuning of a target model using a target dataset.

The technique can accelerate the entire learning process by reducing inference time and computational complexity. Moreover, the learning process can learn the data distribution in a non-parallel manner and ensure its synchronization over time. It can also learn rich and informative representations by using cooperative interactions among modalities. Moreover, it can improve the quality of the information transferred by eliminating any latent noise and conflict [113,115–117]. For example, Palaskar et al. [113] proposed a multimodal integration pipeline that loads the parameters of a pre-trained model on the source dataset (transcript and video) to initialize the training of the target dataset (summary, question, and video). They used hierarchical attention [114] as a merging mechanism that can be used to generate a synthesis vector from multimodal sources. An example of a multimodal transfer learning pipeline based on the fine-tuning mechanism is shown in Fig. 12. It can be seen that a deep model is first pre-trained on a source domain, the learned parameters are then shifted to different modalities (i.e., fine-tuned models) and finally blended into the target domain using fusion techniques.

3.8 Zero-shot learning

In practice, the amount of labeled data samples for effective model training is often insufficient to recognize all possible object categories in an image (i.e., seen and unseen classes). This is why zero-shot learning [130] takes place. This supervised learning approach opens up many valuable applications such as object detection [131], object classification and retrieval of videos [141], especially when appropriate datasets are missing. In other words, it addresses multi-class learning problems when some classes do not have sufficient training data. However, during the learning process, additional visual and semantic features such as word embeddings [132], visual attributes [133], or descriptions [134] can be assigned to both seen and unseen classes. In the context of multimodality, a multimodal mapping scheme typically combines visual and semantic attributes using only data related to the seen classes. The objective is to project a set of synthesized features in order to make the model more generalizable toward the recognition of the unseen class in test samples [135]. Such methods tend to use GAN models to

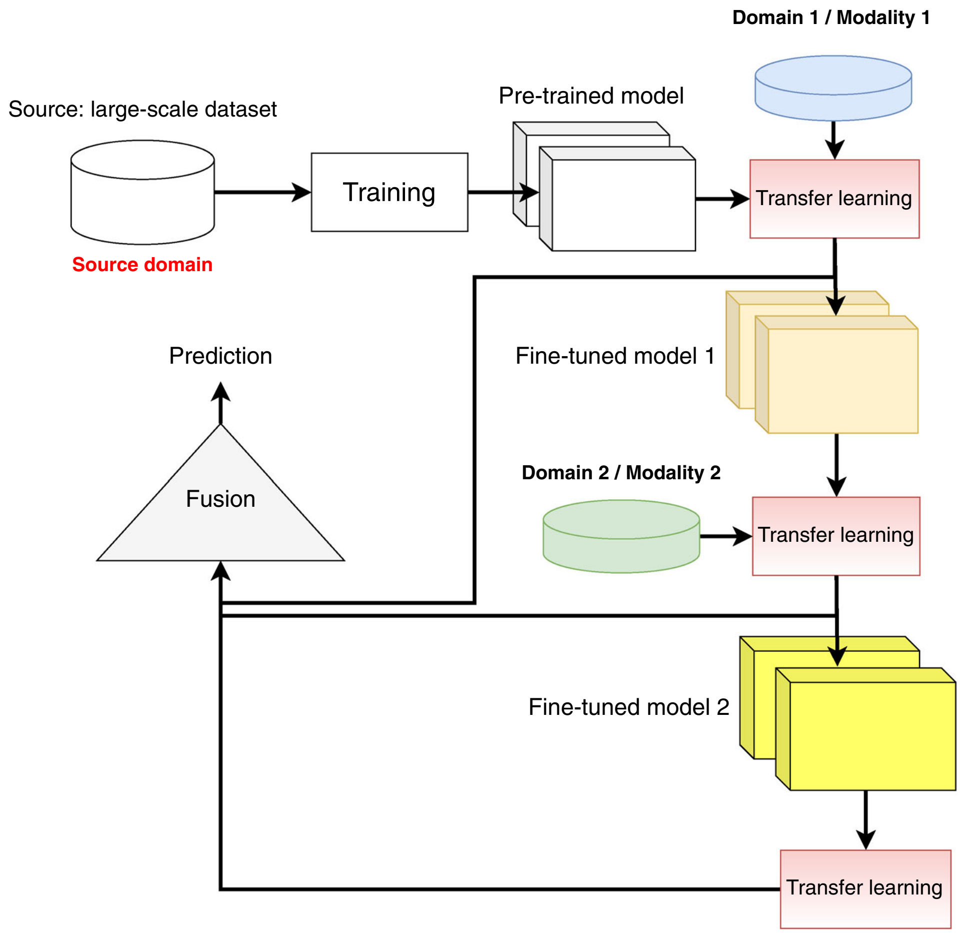

Fig. 12 An illustration of an example of a multimodal transfer learning process

synthesize and reconstruct the visual features of the unseen classes, resulting in high accuracy classification and ensuring a balance between seen and unseen class labels [136,137].

4 Tasks and applications

When modeling multimodal data, several compromises have to be made between system performance, computational burden, and processing speed. Also, many other factors must be regarded when selecting a deep model due to its sensitivity and complexity. In general, multimodality has been employed in many vision tasks and applications, such as face recognition and image retrieval. Table 2 summarizes the reviewed multimodal applications, their technical details, and the best results obtained according to evaluation metrics such as accuracy (ACC) and precision (PREC). In the following, we first describe the core tasks of computer vision, followed by a comprehensive discussion of each application and its intent.

4.1 Generic computer vision tasks

4.1.1 Object detection

Object detection tasks generally consist of identifying rectangular windows (i.e., bounding boxes) in the image (i.e., object localization) and assigning class labels to them (i.e., object classification), through a process of patch extraction and representation (i.e., region of interest (RoI)). The localization process aims at defining the coordinates and position of the patch. In order to classify each object instance, a patch proposal strategy may be applied before the final prediction step. In practice, there are several possible detection methods. The most typical of these is to apply the classifier to an arbitrary region of the image or to a range of different shapes and scales. In the case of detecting patches, the same techniques used in traditional computer vision, such as the sliding window (SW) fashion, can be easily applied when patches are generated in SW mode, neural networks can be used to predict the target information. However, due to their complexity, this type of solution is not cost-effective, both in terms of training duration and memory consumption. In order to significantly reduce this complexity, the deep learning community has pioneered a new generation of CNN-based frameworks. Recent literature has focused on this challenging task: In [67], Jiao et al. studied a variety of deep object detectors, ranging from one-stage detectors to two-stage detectors.

4.1.1.1 One-stage detectors

Monomodal based The overfeat architecture [24] consists of several processing steps, each of which is dedicated to the extraction of multi-scale feature maps by applying the dense SW method to efficiently perform the object detection task. To significantly increase the processing speed of object detection pipelines, Redmon et al. [25] implemented a one-stage lightweight detection strategy called YOLO (You Only Look Once). This approach treats the object detection task as a regression problem, analyzing the entire input image and simultaneously predicting the bounding box coordinates and associated class labels. However, in some vision applications, such as autonomous driving, security, video surveillance, etc., real-time conditions become necessary. In this respect, two-stage detectors are generally slow in terms of real-time processing. In contrast, SSD (single-shot multibox detector) [78] has reduced the needs of the patches' proposal network and, thus, accelerated the object detection pipeline. It can learn multi-scale feature representation from multi-resolution images. Its capability to detect objects at different scales enables it to enhance the robustness of the entire chain. Like most object detectors, the SSD detector consists of two processing stages: extracting the feature map through the VGG16 model and detecting the object by applying a convolutional filter through the $Conv4-3$ layer. As similar to the principle of YOLO and SSD detectors, RetinaNet [79] takes only one stage to detect dense objects by producing multi-scale semantic feature maps using a feature pyramid network (FPN) backbone and the ResNet model. To deal with the class imbalance in the training phase, a novel loss function called "focal loss" is considered by [79]. This function allows training a one-stage detector with high accuracy by reducing the level of artifacts.

Multimodal based High-precision object recognition systems with multiple sensors are aware of external noise and environmental sensitivity (e.g., lighting variations, occlusion, etc.). More recently, the availability of low-cost and robust sensors (e.g., RGB-D sensors, stereo, etc.) has encouraged the computer vision community to focus on combining the RGB modality with other sensing modalities. According to experimental results, it has been shown that the use of depth information [183,184], optical flow information [185], and LiDAR point clouds [186] in addition to conventional RGB data can improve the performance of one-stage based detection systems.

4.1.1.2 Two-stage detectors

Monomodal based The R-CNN detector [74] employs the patch proposal procedure using the selective search [80] strategy and applies the SVM classifier to classify any potential proposals. Fast R-CNN was introduced in [75] to improve the detection efficiency of R-CNN. The principle of Fast R-CNN is as follows: it first feeds the input image into the CNN network, extracts a set of feature vectors, applies a patch proposal mechanism, generates potential candidate regions using the RoI pooling layer, reshapes them to a fixed size, and then performs the final object detection prediction. As an efficient extension of fast R-CNN, Faster R-CNN [76] serves to use a deep CNN as a proposal generator. It has an internal strategy for proposing patches called region proposal network (RPN). Simultaneously, RPN carries out classification and localization regression to generate a set of RoIs. The primary objective is to improve the localization task and the overall performance of the decision system. In other words, the first network uses prior information about being an object, and the second one (at the end of the classifier) that deals with this information for each class. The feature pyramid network (FPN) detector [77] consists of a pyramidal structure that allows the learning of hierarchical feature maps extracted at each level of representation. According to [77], learning multi-scale representations is very slow and requires a lot of memory. However, FPN can generate pyramidal representations with a higher semantic resolution than traditional pyramidal designs.

Multimodal based As mentioned before, two-stage detectors are generally based on a combination of a CNN model to perform classification and a patch proposal module to generate candidate regions like RPNs. These techniques have proven effective for the accurate detection of multiple objects under normal and extreme environmental conditions. However, multi-object detection in both indoor and outdoor environments under varying environmental and lighting conditions remains one of the major challenges facing the computer vision community. Furthermore, a better trade-off between accuracy and computational efficiency in two-stage object detection remains an open question [84]. The question may be addressed more effectively by combining two or more sensory modalities simultaneously. However, the most common approach is to concatenate heterogeneous features from different modalities to generate an artificial multimodal representation. The recent literature has shown that it is attractive to learn shared representations from the complementarity and synergies between several modalities for increasing the discriminatory power of models [190]. Such modalities may include visual RGB-D [187], audio-visual data [188], visible and thermal data [189], etc.

4.1.1.3 Multi-stage detectors

Monomodal based Cascade R-CNN [26] is one of the most effective multi-stage detectors that have proven their robustness over one and two-stage methods. It is a cascaded version of R-CNN aimed at achieving a better compromise between object localization and classification. This framework has proven its capability in overcoming some of the main challenges of object detection, including overtraining problems [5,6] and false alarm distribution caused by the patches' proposal stage. In other words, the trained model may be over-specialized on the training data and can no longer generalize on the test data. The problem can be solved by stopping the learning process before reaching a poor convergence rate, increasing the data distribution in various ways, etc.
Multimodal based More recently, only a few multimodal-based multi-stage detection frameworks [191–193] have been developed and have achieved outstanding detection performance on benchmark datasets.

4.1.2 Visual tracking

For decades, visual tracking has been one of the major challenges for the computer vision community. The objective is to observe the motion of a given object in real time. A tracker can predict the trajectory of a given rigid object from a chronologically ordered sequence of frames. The task has attracted a lot of interest because of its enormous relevance in many real-world applications, including video surveillance [82], autonomous driving [83], etc. Over the last few decades, most deep learning-based object tracking systems have been based on CNN architectures [84,139]. For example, in 1995, Nowlan et al. [85] implemented the first tracking system that tracks hand gestures in a sequence of frames using a CNN model. Multi-object tracking (MOT) has been extensively explored in recent literature for a wide range of applications [86,138]. Indeed, MOT (tracking-by-detection) is another aspect of the generic object tracking task. However, MOT methods are mainly designed to optimize the dynamic matching of the objects of interest detected in each frame. To date, the majority of the existing tracking algorithms have yet to be adapted to various factors, such as illumination and scale variation, occlusions, etc [178]. Multimodal MOT is a universal aspect of MOT aimed at ensuring the accuracy of autonomous systems by mapping the motion sequence of dynamic objects [194]. To date, several multimodal variants of MOT have been proposed to improve the speed and accuracy of visual tracking by using multiple data sources, e.g., thermal, shortwave infrared, and hyperspectral data [195], RGB and thermal data [196], RGB and infrared data [197], etc.

4.1.3 Semantic segmentation

In image processing, image segmentation is a process of grouping pixels of the image together according to particular criteria. Semantic segmentation consists of assigning a class label to each pixel of a segmented region. Several studies have provided an overview of the different techniques used for semantic segmentation of visual data, including the works of [27,28]. Scene segmentation is a subtask of semantic segmentation that enables intelligent systems to perceive and interact in their surrounding environment [27,66]. The image can be split into non-overlapping regions according to particular criteria, such as pixel and edge detection and points of interest. Some algorithms are then used to define inter-class correlations for these regions.

Monomodal based Over the last few years, the fully convolutional network (FCN) [29] has become one of the robust models for a wide range of image types (multimedia, aerial, medical, etc.). The network consists of replacing the final dense layers with convolution layers, hence the reason for its name "FCN". However, the convolutional side (i.e., the feature extraction side) of the FCN generates low-resolution representations which lead to fairly fuzzy object boundaries and noisy segmentations. Consequently, this requires the use of a posteriori regularizations to smooth out the segmentation results, such as conditional random field (CRF) networks [69]. As a light variant of semantic segmentation, instance segmentation yields a semantic mask for each object instance in the image. For this purpose, some methods have been developed, including Mask-RCNN [30], Hybrid Task Cascade (HTC) [31], etc. For instance, the Mask R-CNN model offers the possibility of locating instances of objects with class labels and segmenting them with semantic masks. Scene parsing is a visual recognition process that is based on semantic segmentation and deep architectures. A scene can be parsed into a series of regions labeled for each pixel that is mapped to semantic classes. The task is highly useful in several real-time applications, such as self-driving cars, traffic scene analysis, etc. However, fine-grained visual labeling and multi-scale feature distortions pose the main challenges in scene parsing.

Multimodal based More recently, it has been shown in the literature that the accuracy of scene parsing can be improved by combining several detection modalities instead of a single one [91]. Many different methods are available, such as soft correspondences [94], 3D scene analysis from RGB-D data [95], to ensure dense and accurate scene parsing of indoor and outdoor environments.

4.2 Multimodal applications

4.2.1 Human recognition

In recent years, a wide range of deep learning techniques has been developed that focus on human recognition in videos. Human recognition seeks to identify the same target at different points in space-time derived from complex scenes. Some studies have attempted to enhance the quality of person recognition from two data sources (audio-visual data) using DBN and DBM [72] models, which have allowed several types of representation to be combined and coordinated. Some of these works include [48], [73]. According to Salakhutdinov et al. [72], a DBM is a generative model that includes several layers of hidden variables. In [48], the structure of deep multimodal Boltzmann machines (DMBM) [71] is similar to that of DBM, but it can admit more than one modality. Therefore, each modality will be covered individually using adaptive approaches. After joining the multi-domain features, the high-level classification will be performed by an individual classifier. In [73], Koo et al. developed a multimodal human recognition framework based on face and body information extracted from deep CNNs. They employed the late fusion policy to merge the high-level features across the different modalities.

4.2.2 Face recognition

Face recognition has long been extremely important, ranging from conventional approaches that involve the extraction and selection of handcrafted features, such as Viola and Jones detectors [49] to the automatic extraction and training of end-to-end hierarchical features from raw data. This process has been widely used in biometric systems for control and monitoring purposes. The most biometric systems rely on three modes of operation: enrolment, authentication (verification), and identification [92]. However, most facial recognition systems, including biometric systems, suffer from a restriction in terms of universality and variations in the appearance of visual patterns. End-to-end training of multimodal facial representations can effectively help to overcome this limitation. Multimodal facial recognition systems can integrate complex representations derived from multiple modalities at different scales and levels (e.g., feature level, decision level, score level, rank level, etc.). Note that face detection, face identification, and face reconstruction are subtasks of face recognition [50]. Numerous works in the literature have demonstrated the benefits of multimodal recognition systems. In [51], Ding et al. proposed a new late fusion policy using CNNs for multimodal facial feature extraction and SAEs for dimensional reduction. The authors of [93] introduced a biometric system that combines biometric traits from different modalities (face and iris) to establish an individual's identity.

4.2.3 Image retrieval

Content-based image research (CBIR), commonly known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) [54], is the process of recovering visual content (e.g., colors, edges, textures, etc) stored in datasets by learning their visual representations. The retrieval procedure leads to the generation of metadata (i.e., keywords, tags, labels, and so on). The CBIR mechanism can be simulated in two fundamental phases: the offline database indexing phase and the online retrieval step. During the indexing stage, image signatures will be generated and stored in a database. In the retrieval phase, the image to be recovered will be treated as a query and the matching process will reconcile this image signature with that stored in the database. Over the last few years, several cross-modal image retrieval tasks, e.g., text-to-image retrieval [100], sketch-to-image retrieval [101], cross-view image retrieval [102], composing text and image-to-image [103], etc. have been covered in the literature.

4.2.4 Gesture recognition

Gesture recognition is one of the most sophisticated tasks of computer vision. The task has already gained the attention of the deep learning community for many reasons. In particular, its potential is to facilitate human–computer interaction and detect motion in real time. As gestures become more diversified and enriched, our instinctive intelligence will recognize basic actions and associate them with generic behaviors. The challenge of action recognition is mainly related to the difficulty of extracting body silhouettes from foreground rigid objects to focus on their emotions [96]. Occlusions that occur between different object parts can lead to a significant decrease in performance. However, various factors, such as variations in speed, scale, noise, and object position, can significantly affect the recognition process. Some real-world applications of gesture recognition include driver assistance, smart surveillance, human–machine interaction, etc. Regarding the multimodal dimensions of gesture recognition, the authors of [97] proposed a multi-stream architecture based on the RNN (LSTM) model to capture spatial-temporal features from gesture data. In [98], the authors developed a multimodal gesture recognition system using the 3D Residual CNN (ResC3D) model [99] trained on an RGB-D dataset. The features extracted by the ResC3D model are then combined with a canonical correlation scheme to ensure consistency in the fusion process. Likewise, Abavisani et al. [200] developed a fusion approach to derive knowledge from multiple modalities in individual unimodal 3D CNN networks.

4.2.5 Image captioning

Recently, image captioning has become an active research topic in the field of multimodal vision, i.e., the automatic generation of text captions to describe the content of images. In a supervised learning way, training of model parameters is provided by a set of labeled learning examples in the form of an image and its related captions. The task has also been demonstrated its ability for application in a variety of real-world systems, including social media recommendation, image indexing, image annotation, etc. Most recently, Biten et al. [52] combined both visual and textual data to generate captions across two stages: template caption generation stage and entity insertion stage. Similarly, Peri et al. [53] proposed a multimodal framework that encodes both images and captions using CNN and RNN as an intermediate level representation and then decodes these multimodal representations into a new caption that is similar to the input. The authors of [128] presented an unsupervised image captioning framework based on a new alignment method that allows the simultaneous integration of visual and textual streams through semantic learning of multimodal embeddings of the language and vision domains. Moreover, a multimodal model can also aggregate motion information [174], acoustic information [175], temporal information [176], etc. from successive frames to assign a caption for each one. We invite the reader to read the survey of Liu et al. [177] to learn more about the methods, techniques, and challenges of image captioning.

4.2.6 Vision-and-language navigation

Visual-and-language navigation (VLN) [87,88,118–121] is a multimodal task that has become increasingly popular in recent years. The idea behind VLN is to combine several active domains (i.e., natural language, vision, and action) to enable robots (intelligent agents) to navigate easily in unstructured environments. A key innovation in this area is the synthesis of heterogeneous data into multiple modalities using natural language commands to navigate through crowded locations and visual cues to perceive the surroundings. It seeks to establish an interaction between visual patterns and natural language concepts by merging these modalities into a single representation.

4.2.7 Embodied question answering

Embodied question answering (EQA) [89,90,122] is an emerging multimodal task in which an intelligent agent acts intelligently in a three-dimensional environment in order to respond to a given question. To this end, the agent must first explore its environment, capture visual information, and then answer the question posed. In [90], the authors proposed the multi-target embodied question answering (MT-EQA) task as a generalization of EQA. In contrast to EQA, MT-EQA considered some questions related to multiple targets, where an agent has to navigate toward various locations to answer a question asked (Fig. 13a).

4.2.8 Video question answering

Currently, video question answering (VQA) [125–127,129, 143] is one of the promising lines of research for reasoning the correct answer to a particular question, based on the spatiotemporal visual content of video sequences. To answer that question, we need to consider the correlation between features in the spatial and temporal dimensions (Fig. 13b). The VQA task can be conceptually divided into three subtasks. The first task is to identify the endpoints of the problem in the natural domain, while the second task is to capture the correlation of the problem in the spatial domain. The third task consists of reasoning about how this correlation varies in space over time. Typically, video sequences contain audio-visual information of substantially different structures and visual appearance, which requires reasoning schemes that take into account the spatiotemporal nature of the data. To this end, increased attention has been paid to these challenges by developing a wide range of spatiotemporal reasoning mechanisms. Currently, the most common existing methods use attention [125,127,129] and memory [126] mechanisms to efficiently learn visual artifacts and the semantic correlations that allow questions to be answered accurately. These techniques are more effective for spatial-temporal video representation and reasoning as they increase the memorization and discrimination capacity of models.

4.2.9 Style transfer

Neural style transfer (NST), also known as style transfer, has recently gained momentum following the publication of the works of Gatys et al. [156]. Gatys et al. [156] demonstrated that visual features of models could be combined to represent image styles. It arises in a context of strong growth in DNNs for several applications, including art and painting [157,158]. For example, Lian et al. [157] proposed a style transfer-based method that takes any natural portrait of a human and transforms it into Picasso's cubism style. Informally, style transfer is an optimization-based technique that renders the content of an existing image (content image) in the style of another image (style image). Figure 14 depicts an example of transferring the style of a specific painting to a scene image using the DeepArts tool [162]. In practice, style transfer involves applying a particular artistic style to a content image. For

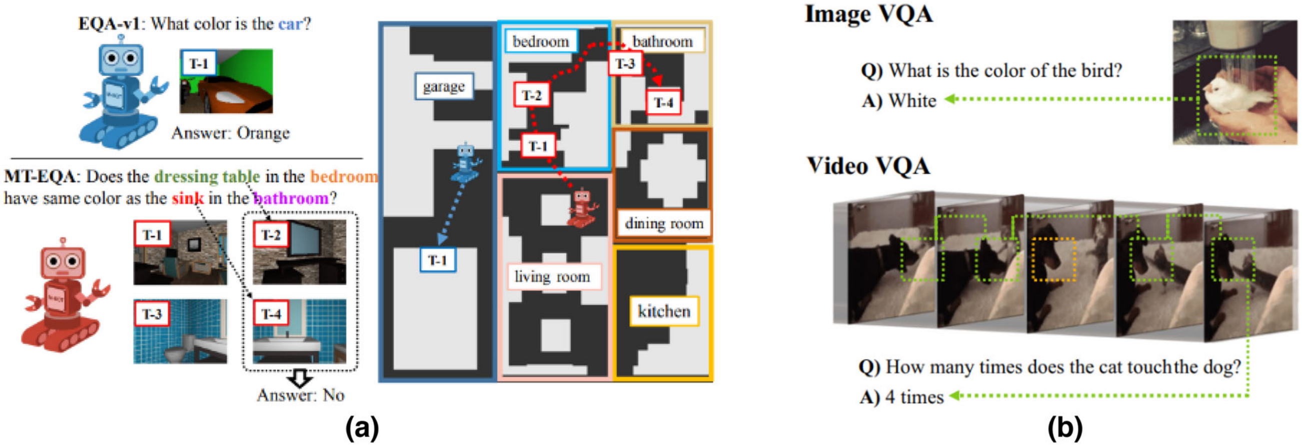
EQA-v1: What color is the car?
T-1
Answer: Orange
MT-EQA: Does the dressing table in the bedroom have same color as the sink in the bathroom?
T-1
T-2
T-3
T-4
Answer: No
garage
bedroom
bathroom
dining room
living room
kitchen
(a)
Image VQA
Q) What is the color of the bird?
A) White
Video VQA
Q) How many times does the cat touchthe dog?
A) 4 times
(b)

Fig. 13 Difference in results between EQA and VQA tasks: **a** EQA [90], **b** VQA [129]

this, a loss function must be specified and minimized. It is essentially a weighted sum of the error (loss) between the input content and the output image and the loss between the original style and the applied style [161]. Over the last few years, some research has been published to improve this tool by considering the mapping of semantic patterns in content and style images, from which the multimodal style transfer (MST) emerges [159,160]. The authors of [159] proposed a universal graph-based style transfer to transform multimodal features by matching style patterns and semantic content and appearance in a way that avoids the lack of flexibility in real-world scenarios. The use of the graph cut technique allows a better matching between content features and style clusters, which was formulated as an energy minimization issue. In [160], Wang et al. introduced a residual CNN architecture and loss network to transfer the artistic style of the input picture across multiple scales and dimensions. Specifically, the residual network receives an image as input and learns to produce multi-scale representations as output. These representations are then separately considered as inputs to the loss network so that a stylization loss can be computed for each one.

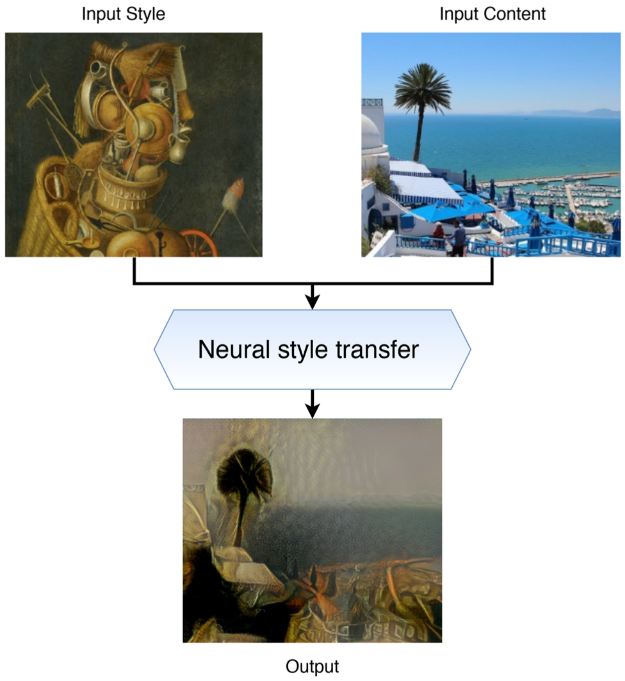

Fig. 14 Example of NST algorithm output to transform the style of a painting to a given image

4.2.10 Medical data analysis

In recent years, deep learning algorithms have been developed to save time and dependability during patient care by improving clinical accuracy and detecting abnormalities in medical images [104,140]. As one application, retinal image registration [142] is an increasingly challenging task of medical image analysis that is receiving more and more attention from the computer vision and healthcare communities. In [142], Lee et al. proposed a new CNN-based retinal image registration method to learn multimodal features simultaneously from several imaging modalities. This method consists of combining CNN features and small patches taken from multiple imaging modalities (e.g., FA, OCT fundus, etc.) and then implementing learning and optimization processes to achieve greater registration accuracy. In unsupervised mode, it is possible to encode complex visual patterns over two input imaging modalities (3D MR and TRUS) without the need for explicit labels [144]. In clinical practice, the different imaging modalities (e.g., computed tomography (CT), chest X-rays, etc.) provide rich and informative features that allow for a more accurate diagnosis in the early stages of the disease [238]. More recently, the scientific community has already taken an active interest in this topic to fight against the emerging Coronavirus, known as COVID-19 (SARS-CoV-2)

[105]. To date, the COVID-19 pandemic has spread rapidly in most countries of the world, endangering people's lives. Deep learning techniques and the availability of medical data contributed considerably to tackling the pandemic. The latest literature indicates that the combination of multimodal data can predict and screen for this virus more accurately [106,145]. However, many studies still need to be undertaken in the future.

4.2.11 Autonomous systems

Up to now, deep learning has proven to be a powerful tool for generating multimodal data suitable for robotics and autonomous systems [146]. These systems involve, for example, the interaction of sophisticated perception/vision and haptic sensors (e.g., monocular cameras, stereo cameras, and so on) [147], the merging of depth and color information from RGB-D cameras [148], and so on. Figure 15 shows an autonomous vehicle with several on-board sensors, including a camera and several radars and LiDARs. Most existing approaches combine RGB data with infrared images or 3D LiDAR points [164] to improve the sensitivity of perception systems, which can be suitable to all conditions and scenarios. For instance, RGB-D cameras (e.g., Microsoft Kinect, Asus Xtion, and so on) can provide color and pixel-wise depth information, characterizing the distance of visual objects in a complex scene [199]. Among the advantages of these types of sensors are their low computational cost, their long-range, their ability to have an internal mechanism to limit the impact of bad weather, etc. [149]. More recently, some automated systems, such as mobile robots, have been used in manufacturing environments. However, in a manufacturing context, these systems are usually already routinely programmed with repetitive actions that lack the capacity for autonomy. They also depend on an unstructured environment for autonomous decision making (e.g., navigation, localization, and environment mapping (SLAM)).

For decades, visual SLAM (simultaneous localization and mapping) has been an active area of research in the robotics and computer vision communities [148,150]. The challenge lies both in locating a robot and mapping its surrounding environment. Several methods have been reported to improve the mapping accuracy of real-time scenarios in unstructured and large-scale environments. Some of these methods include descriptor-based monocular cameras with ORB-SLAM [151], stereovision with ORB-SLAM2 [152], and photometric error-based methods such as LSD-SLAM [153] or DSO [154]. However, there are still many challenges facing these data-driven automated systems, particularly for intelligent perception and mapping. Some of these challenges are reflected in the fact that large amounts of data are required to train models. Therefore, large-scale datasets are required to ensure that systems produce the desired outcomes. As a result, more powerful feature extractors will require more parameters and, therefore, more learning data. For instance, Caesar et al. [155] demonstrated how generalization performance could be greatly improved when developing a multimodal dataset, called nuScenes, which is acquired by a wide range of remote sensors, including six cameras, five radars, and one LiDAR. The dataset consists of 1000 scenes in total, each about 20 s long and fully labeled with 3D bounding boxes that cover 23 classes and eight attributes.

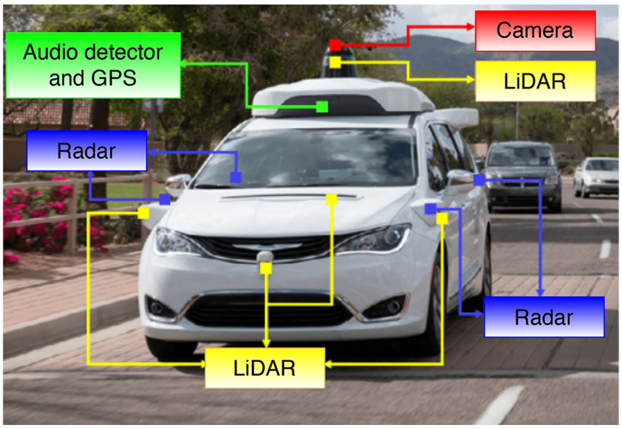

Fig. 15 Waymo self-driving car equipped with several on-board sensors [163]

5 Popular visual multimodal datasets

A growing trend towards deep multimodal learning has been fuelled by the availability of high-dimensional multisource datasets obtained from various sensors, including RGB-D cameras (depth sensors). Multimodal data acquisition is increasingly used in many research disciplines. The deep multimodal analysis relies on a large amount of heterogeneous sensor data to achieve high performance and avoid overfitting problems. Until now, a series of benchmark datasets have been developed for the training and validation of deep multimodal learning algorithms. This opens up the question of which ones should be chosen and how they can be used for benchmarking purposes with state-of-the-art methods. To answer this question, in this section we present a selection of multimodal datasets commonly used in vision applications, including RGB-D and RGB flow datasets. Typically, optical flow information is used to capture the motion of moving objects in a video sequence. It was originally developed by Horn et al. [65], formulated as a two-dimensional vector flow that captures spatio-temporal motion variations in images under fairly controlled conditions in both indoor and outdoor environments. The emphasis on these modalities (RGB, depth, and flow data) is based on the fact that for many vision-based multimodal problems, it has been shown that the fusion of optical flow and depth information with RGB yields the best performance [242,243]. A selection of RGB-D and RGB flow datasets and their detailed information is given in Table 3, so that researchers can easily choose the right dataset for their needs. Table 3 shows the typical computer vision tasks, such as object recognition and semantic segmentation, along with their respective benchmark datasets.

All datasets listed in Table 3 will be detailed in the following paragraphs:

- **RGB-D Object:** According to the original paper [55], the larger-scale RGB-D object dataset consists of RGB videos and depth sequences of 300 object instances in 51 categories from multiple view angles for a total of 250,000 images.
- **BigBIRD:** The dataset was originally introduced by [56]. It contains 125 objects, 600 RGB-D point clouds, and 600 12 megapixel images taken by two sensors: Kinect and DSLR cameras.
- **A large dataset of object scans:** It includes more than 10,000 scanned and reconstructed objects in nine categories acquired by PrimeSense Carmine cameras.
- **RGB-D Semantic Segmentation:** The dataset has originally been proposed in [58], it was acquired by the Kinect RGB-D sensor. It contains six categories such as juice bottles, coffee cans and boxes of salt, etc. On the one hand, the training set contains three 3D models for each category. On the other hand, the testing set includes 16 objects scenes.
- **RGB-D Scenes v.1:** The dataset contains eight scenes in which each scene corresponds to a single video sequence of several RGB-D images.
- **RGBD Scenes v.2:** The dataset contains 14 scenes of video sequences including furniture that have been acquired by the Kinect device.
- **NYU:** There are two versions of the dataset (NYU-v1 and NYU-v2) that were recorded by the Kinect sensor. On the one hand, NYU-v1 dataset contains 64 different indoor scenes and 108617 unlabelled images. On the other hand, NYU-v2 Dataset includes 464 different indoor scenes and 407024 unlabeled images.
- **RGB-D People:** This dataset was initially introduced by [60], it consists of more than 3000 RGB-D images captured from Kinect sensors.
- **SceneNet RGB-D:** This dataset contains 5M RGB-D images extracted from a total of 16895 configurations.
- **Kinetics-400:** It consists of a massive dataset of YouTube video URLs that includes a diverse set of human actions. The dataset includes more than 300,000 video sequences across 400 classes of human action.
- **Scene Flow:** The dataset includes over 39,000 high-resolution frames from synthetic video sequences. It combines a wide range of data types such as RGB stereo rendering, optical flow maps, and so on.
- **MPI-Sintel:** The dataset consists of 1040 annotated optical flow and corresponding RGB images from very long sequences.

6 Discussion, limitations, and challenges

Over the last few decades, the deep learning paradigm has proven its ability to outperform human expertise in many practices. Deep learning algorithms involve a sequence of multiple layers of nonlinear processing units that are used to extract and transform feature vectors coming from raw data. Up to now, the deep learning community is still seeking a better trade-off between complex model structuring, computational power requirements, and real-time processing capability. Among its assets, computer vision seeks to give machines the visual capabilities of human beings thanks to deep learning algorithms that are fed with information from a wide range of sensors. In recent years, the trend toward its use in a fairly wide range of applications has become increasingly evident. Therefore, it is necessary to develop applications that can automatically predict the target information. However, most current scene-content analysis methods are still limited in their ability to deal with information that is not usable in real-life contexts. But this field is very interesting for the scientific and industrial communities. This aspect of uncertainty underlines the need to propose innovative and practical methods under very similar conditions to those used in practice. In general, capturing multimodal data streams under different acquisition conditions and increasing the data volume makes it easier to recognize visual content. Deep learning models are often robust strategies for dealing with the linear and nonlinear combination of multimodal data. Despite the impressive results of deep multimodal learning, no absolute conclusions can be drawn in this regard. Considering this exponential growth, the main challenges of multimodal learning methods are the following:

- **Dimensionality and data conflict:** Confusion between various data sources is a challenge for future analysis. The multimodal data is usually available in various formats. This variation makes it difficult to extract valuable information from the data. However, multimodal information generally has a large dimension. In other words, acquiring and processing a large amount of multimodal data is costly in terms of computation complexity and memory consumption. Moreover, the synchronization of temporal data allows maximizing the correlation between the features of several levels of representation. However, feature-level fusion is more flexible than decision-level

Table 2 Summary of the multimodal applications reviewed, their related technical details, and best results achieved

References	Year	Application	Sensing modality/data sources	Fusion scheme	Dataset/best results
[73]	2018	Person recognition	Face and body information	Late fusion (Score-level fusion)	DFB-DB1 (EER = 1.52%)
					ChokePoint (EER = 0.58%)
[51]	2015	Face recognition	Holistic face + Rendered frontal pose data	Late fusion	LFW (ACC = 98.43%)
					CASIA-WebFace (ACC = 99.02%)
[93]	2020	Face recognition	Biometric traits (face and iris)	Feature concatenation	CASIA-ORL (ACC = 99.16%)
					CASIA-FERET (ACC = 99.33%)
[100]	2016	Image retrieval	Visual + Textual	Joint embeddings	Flickr30K (mAP = 47.72%; R@10 = 79.9%)
					MSCOCO (R@10 = 86.9%)
[101]	2016	Image retrieval	Photos + Sketches	Joint embeddings	Fine-grained SBIR Database (R@5 = 19.8%)
[102]	2015	Image retrieval	Cross-view image pairs	Alignment	A dataset of 78k pairs of Google street-view images (AP = 41.9%)
[103]	2019	Image retrieval	Visual + Textual	Feature concatenation	Fashion-200k (R@50 = 63.8%)
					MIT-State (R@10 = 43.1%)
					CS (R@1 = 73.7%)
[97]	2015	Gesture recognition	RGB + D	Recurrent fusion, Late fusion, and Early fusion	SKIG (ACC = 97.8%)
[98]	2017	Gesture recognition	RGB + D	A canonical correlation scheme	Chalearn LAP IsoGD (ACC = 67.71%)
[200]	2019	Gesture recognition	RGB + D + Opt. flow	A spatio-temporal semantic alignment loss (SSA)	VIVA hand gestures (ACC = 86.08%)
					EgoGesture (ACC = 93.87%)
					NVGestures (ACC = 86.93%)
[52]	2019	Image captioning	Visual + Textual	RNN + Attention mechanism	GoodNews (Bleu-1 = 8.92%)
[53]	2019	Image captioning	Visual + Textual + Acoustic	Alignment	MSCOCO (R@10 = 91.6%)
					Flickr30K (R@10 = 79.0%)
[128]	2019	Image captioning	Visual + Textual	Alignment	MSCOCO (BLUE-1 = 61.7%)
[174]	2019	Image captioning	Visual + Textual	Gated fusion network	MSR-VTT (BLUE-1 = 81.2%)
					MSVD (BLUE-4 = 53.9%)
[175]	2019	Image captioning	Visual + Acoustic	GRU Encoder-Decoder	Proposed dataset (BLUE-1 = 36.9%)
[176]	2020	Image captioning	Visual + Textual (Spatio-temporal data)	Object-aware knowledge distillation mechanism	MSR-VTT (BLUE-4 = 40.5%)
					MSVD (BLUE-4 = 52.2%)

Table 2 continued

References	Year	Application	Sensing modality/data sources	Fusion scheme	Dataset/best results
[87]	2018	Vision-and-language navigation	Visual + Textual (instructions)	Attention mechanism + LSTM	R2R (SPL = 18%)
[88]	2019	Vision-and-language navigation	Visual + Textual	Attention mechanism + Language Encoder	R2R (SPL = 38%)
[118]	2020	Vision-and-language navigation	Visual + Textual (instructions)	Domain adaptation	R2R (Performance gap = 8.6)
					R4R (Performance gap = 23.9)
					CVDN (Performance gap = 3.55)
[119]	2020	Vision-and-language navigation	Visual + Textual (instructions)	Early fusion + Late fusion	R2R (SPL = 59%)
[120]	2020	Vision-and-language navigation	Visual + Textual (instructions)	Attention mechanism + Feature concatenation	VLN-CE (SPL = 35%)
[121]	2019	Vision-and-language navigation	Visual + Textual (instructions)	Encoder-decoder + Multiplicative attention mechanism	ASKNAV (Success rate = 52.26%)
[89]	2018	Embodied question answering	Visual + Textual (questions)	Attention mechanism + Alignment	EQA-v1 (MR = 3.22)
[90]	2019	Embodied question answering	Visual + Textual (questions)	Feature concatenation	EQA-v1 (ACC = 61.45%)
[122]	2019	Embodied question answering	Visual + Textual (questions)	Alignment	VideoNavQA (ACC = 64.08%)
[125]	2019	Video question answering	Visual + Textual (questions)	Bilinear fusion	TDIUC (ACC = 88.20%)
					VQA-CP (ACC = 39.54%)
					VQA-v2 (ACC = 65.14%)
[126]	2019	Video question answering	Visual + Textual (questions)	Alignment	TGIF-QA (ACC = 53.8%)
					MSVD-QA (ACC = 33.7%)
					MSRVTT-QA (ACC = 33.00%)
					Youtube2Text-QA (ACC = 82.5%)
[127]	2020	Video question answering	Visual + Textual (questions)	Hierarchical Conditional Relation Networks (HCRN)	MSRVTT-QA (ACC = 35.6%)
					MSVD-QA (ACC = 36.1%)
[129]	2019	Video question answering	Visual + Textual (questions)	Dual-LSTM + Spatial and temporal attention	TGIF-QA (*l*2 distance = 4.22)
[159]	2019	Style transfer	Content + Style	Graph based matching	A dataset of images from MSCOCO and WikiArt (PV = 33.45%)
[160]	2017	Style transfer	Content + Style	Hierarchical feature concatenation	A dataset of images from MSCOCO (PS = 0.54s)

Table 2 continued

References	Year	Application	Sensing modality/data sources	Fusion scheme	Dataset/best results
[142]	2019	Medical data analysis	RGB + FA + OCT	Alignment	Color fundus-FA (ACC = 90.10%) Color fundus-OCT (ACC = 84.59%)
[144]	2018	Medical data analysis	MR + TRUS	Alignment	A total of 763 sets of data from the National Institutes of Health and Mount Sinai Hospital (TRE = 3.48mm)
[145]	2020	Medical data analysis	X-Ray + Ultrasound + CT	Data fusion	A dataset of X-Ray, CT and Ultrasound images (PREC = 100%)
[155]	2019	Autonomous systems	RGB + LiDAR + Radar	Feature concatenation	nuScenes (mAP = 28.9%; NDS = 44.9%)
[199]	2019	Autonomous systems	RGB + D + Inertial measurements	Filter-based approaches or nonlinear optimization approaches	KITTI Odometry (t_{rel} =1.78%; r_{rel} = 0.95%) AirSim (t_{rel} = 4.53%; r_{rel} =8.75%)

ACC accuracy, *MR* mean rank, *SPL* success weighted by path length, *mAP* mean average precision, *AP* average precision, R@i recall for setting i, *PREC* precision, *PV* percentage of the votes, *PS* processing speed, *TRE* target registration error, *NDS* nuScenes detection score, *ATE* absolute trajectory error, *Trel* average translational error percentage, *Rrel* rotational error

fusion due to the homogeneity of data samples. As mentioned before, some dimensionality reduction algorithms (e.g., k-NN, PCA, etc) and models already exist that compress (encode) input signal or extract a reduced set of low dimensional patterns to facilitate their analysis and further processing.

- **Data availability:** One of the most significant challenges of deep multimodal learning is the large amount of data required to learn discriminative feature maps. The amount of multimodal data significantly affects the overall performance of the vision system. In some cases, the number of training samples for a given dataset may not be sufficient to effectively train a deeper or wider network. However, networks trained with a limited number of examples can no longer generalize well to a new dataset. As mentioned earlier, several methods have been used to increase the size of the dataset by generating additional learning samples. One of the most common techniques includes data augmentation, which is a transformation process that is applied to the input data to increase the size of the data to make it more invariant. Also, AE can address missing patterns by generating intermediate shared representations from the input data and showing intra- and inter-pattern correlations.
- **Real-time processing and scalability:** Multimodal real-time data processing should be considered to improve the performance of deep learning architectures. Current trends focus on proposing complex architectures to build new real-time processing systems with a better trade-off between accuracy and efficiency. However, the need to reduce computing capacity remains the main challenge, which can lead to a deterioration in the overall accuracy of training algorithms. Vision-based multimodal algorithms constantly require new technological developments from year to year (e.g., cloud computing technologies, local GPU devices, etc.) to enable the growing scalability needed to handle the next generation of multimodal applications. For instance, the edge/cloud computing solution for mulitmodal analysis provides an effortless way to create and handle multimodal datasets for training and deploying models [107]. In practice, autonomous vehicles, healthcare robots, and other real-time embedded systems consume more hardware resources, storage, and battery than other emerging technologies, resulting in a lack of adaptation to future needs.

7 Conclusion

This study provided a comprehensive overview of recent multimodal deep learning in the computer vision community. The focus of this survey is on the analogy between inter- and intra-modal learning when dealing with heteroge-

Table 3 A selection of the frequently used multimodal datasets in the literature

Reference	Year	Dataset	Modality	Main tasks	Size
[55]	2011	RGB-D Object	RGB + D	Object recognition	Contains 300 object instances under 51 categories from different angles for a total of 250,000 RGB-D images
[56]	2014	BigBIRD	RGB + D	Object recognition	Contains 125 objects, 600 RGB-D point clouds, and 600 12 megapixel images
[57]	2016	A large dataset of object scans	RGB + D	Object recognition	Contains more than 10,000 scanned and reconstructed objects in 9 categories
[58]	2011	RGB-D Semantic Segmentation	RGB + D	Semantic segmentation	Contains 3 3D models for 6 categories and 16 test object scenes
[55]	2011	RGB-D Scenes v.1	RGB + D	Object recognition Semantic segmentation	Contains 8 video scenes from several RGB-D images
[55]	2014	RGB-D Scenes v.2	RGB + D	Object recognition Semantic segmentation	Contains 14 scenes of video sequences
[59]	2011	NYU v1-v2	RGB + D	Semantic segmentation	NYU-v1 contains 64 different indoor scenes and 108617 unlabelled images. NYU-v2 contains 464 different indoor scenes and 407024 unlabeled images
[60]	2011	RGB-D People	RGB + D	Object recognition	Contains more than 3000 RGB-D images
[61]	2016	SceneNet RGB-D	RGB + D	Semantic segmentation Instance segmentation Object detection	Contains 5M RGB-D images
[62]	2017	Kinetics-400	RGB + Opt. flow	Motion recognition	Contains more than 300,000 video sequences in 400 classes
[63]	2016	Scene Flow	RGB + Opt. flow	Object segmentation	Contains over 39,000 high resolution images
[64]	2012	MPI-Sintel	RGB + Opt. flow	Semantic segmentation Object recognition	Contains 1040 annotated optical flow and matching RGB images

neous data. In this context, we provide a brief history of deep learning, summarize typical deep learning concepts and algorithms that have evolved from shallow networks to deeper networks such as RNN, DBN, and DAE, and show their role in multimodal fusion. We also provide an overview of multimodal datasets commonly used in the literature (RGB-D and RGB-Opt. flow) and report a methodological analysis of computer vision problems and multimodal applications. Vision-based implementation strategies are also discussed in detail to improve comprehension of the multimodal algorithm's ability to make fast and efficient decisions. The survey also presented state-of-the-art and widely used methods for producing uniform and multimodal distributions across different modalities. Furthermore, it is important to note that each multimodal problem requires a specific fusion strategy, ranging from traditional methods to deep learning techniques. Nevertheless, choosing the right fusion of different schemes remains a vital challenge for the computer vision community in terms of accuracy and efficiency.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 (2016)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798–1828 (2013)
3. Bayoudh, K.: From machine learning to deep learning, (1st ed.), Ebook, ISBN: 9781387465606 (2017)
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
5. Lawrence, S., Giles, C.L.: Overfitting and neural networks: conjugate gradient and backpropagation. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, pp. 114–119 (2000)

6. Bilbao, I., Bilbao, J.: Overfitting problem and the over-training in the era of data: particularly for artificial neural networks. In: 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 173–177 (2017)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2017)
8. Rosenblatt, F.: Perceptron simulation experiments. Proc. IRE **48**, 301–309 (1960)
9. Van Der Malsburg, C.: Frank Rosenblatt: principles of neurodynamics–perceptrons and the theory of brain mechanisms. Brain Theory, 245–248 (1986)
10. Huang, Y, Sun, S, Duan, X, Chen, Z.: A study on deep neural networks framework. In: IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 1519–1522 (2016)
11. Sheela, K.G. Deepa, S.N.: Review on methods to fix number of hidden neurons in neural networks. Math. Problems. Eng. 2013(25740) (2013)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
14. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (IndRNN): building a longer and deeper RNN. arXiv:1803.04831 (2018)
15. Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**, 1527–1554 (2006)
16. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv:1406.2661 (2014)
17. Turkoglu, M.O., Thong, W., Spreeuwers, L., Kicanaoglu, B.: A layer-based sequential framework for scene generation with GANs. arXiv:1902.00671 (2019)
18. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv:1611.07004 (2018)
19. Creswell, A., et al.: Generative adversarial networks: an overview. IEEE Signal Process. Mag. **35**, 53–65 (2018)
20. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev (2020)
21. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2015)
23. Stone, J.V.: Principal component analysis and factor analysis. In: Independent Component Analysis: A Tutorial Introduction, MITP, pp. 129–135 (2004)
24. Sermanet, P. et al.: OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229 (2014)
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. arXiv:1506.02640 (2016)
26. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. arXiv:1906.09756 (2019)
27. Thoma, M.: A survey of semantic segmentation. arXiv:1602.06541 (2016)
28. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. Int. J. Multimed. Infom. Retr. **7**, 87–93 (2018)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. arXiv:1411.4038 (2015)
30. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv:1703.06870 (2018)
31. Chen, K. et al.: Hybrid task cascade for instance segmentation. arXiv:1901.07518 (2019)
32. Marechal, C. et al.: Survey on AI-based multimodal methods for emotion detection. In: High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet, pp. 307–324 (2019)
33. Radu, V., et al.: Multimodal deep learning for activity and context recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**, 157:1–157:27 (2018)
34. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. **34**, 96–108 (2017)
35. Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: a survey. IEEE Access **7**, 63373–63394 (2019)
36. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE **103**(9), 1449–1477 (2015)
37. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 (2017)
38. Morvant, E., Habrard, A., Ayache, S.: Majority vote of diverse classifiers for late fusion. In: Structural, Syntactic, and Statistical Pattern Recognition, pp. 153–162 (2014)
39. Liu, Z. et al.: Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. arXiv:1806.00064 (2018)
40. Zhang, D., Zhai, X.: SVM-based spectrum sensing in cognitive radio. In: 7th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2011)
41. Gönen, M., Alpaydın, E.: Multiple Kernel learning algorithms. J. Mach. Learn. Res. **12**, 2211–2268 (2011)
42. Aiolli, F., Donini, M.: EasyMKL: a scalable multiple kernel learning algorithm. Neurocomputing **169**, 215–224 (2015)
43. Wen, H., et al.: Multi-modal multiple kernel learning for accurate identification of Tourette syndrome children. Pattern Recognit. **63**, 601–611 (2017)
44. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
45. Ghahramani, Z., Jordan, M.I.: Factorial hidden Markov models. Mach. Learn. **29**, 245–273 (1997)
46. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Trans. Commun. **28**, 84–95 (1980)
47. Gael, J.V., Teh, Y.W., Ghahramani, Z.: The infinite factorial hidden Markov model. In: Proceedings of the 21st International Conference on Neural Information Processing Systems, pp. 1697–1704 (2008)
48. Alam, M. R., Bennamoun, M., Togneri, R., Sohel, F.: A deep neural network for audio-visual person recognition. In: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–6 (2015)
49. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**, 137–154 (2004)
50. Wang, M., Deng, W.: Deep Face Recognition: A Survey. arXiv:1804.06655 (2019)
51. Ding, C., Tao, D.: Robust face recognition via multimodal deep face representation. IEEE Trans. Multimed. **17**, 2049–2058 (2015)
52. Biten, A.F., Gomez, L., Rusiñol, M., Karatzas, D.: Good News, Everyone! Context driven entity-aware captioning for news images. arXiv:1904.01475 (2019)
53. Peri, D., Sah, S., Ptucha, R.: Show, Translate and Tell. arXiv:1903.06275 (2019)
54. Duan, G., Yang, J., Yang, Y.: Content-based image retrieval research. Phys. Proc. **22**, 471–477 (2011)
55. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: IEEE International Conference on Robotics and Automation, pp. 1817–1824 (2011)

56. Singh, A., Sha, J., Narayan, K.S., Achim, T., Abbeel, P.: BigBIRD: A large-scale 3D database of object instances. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 509–516 (2014)
57. Choi, S., Zhou, Q.-Y., Miller, S., Koltun, V.: A Large Dataset of Object Scans. arXiv:1602.02481 (2016)
58. Tombari, F., Di Stefano, L., Giardino, S.: Online learning for automatic segmentation of 3D data. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4857–4864 (2011)
59. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: International Conference on Computer Vision Workshops (2011)
60. Spinello, L., Arras, K.O.: People detection in RGB-D data. In: Intelligent and Robotic Systems (2011)
61. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R.: SceneNet: Understanding Real World Indoor Scenes With Synthetic Data. arXiv:1511.07041 (2015)
62. Kay, W. et al.: The Kinetics Human Action Video Dataset. arXiv:1705.06950 (2017)
63. Mayer, N. et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4040–4048 (2016)
64. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. Comput. Vis. ECCV **2012**, 611–625 (2012)
65. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
66. Wang, W., Fu, Y., Pan, Z., Li, X., Zhuang, Y.: Real-time driving scene semantic segmentation. IEEE Access **8**, 36776–36788 (2020)
67. Jiao, L., et al.: A survey of deep learning-based object detection. IEEE Access **7**, 128837–128868 (2019)
68. Dilawari, A., Khan, M.U.G.: ASoVS: abstractive summarization of video sequences. IEEE Access **7**, 29253–29263 (2019)
69. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
70. Shao, L., Zhu, F., Li, X.: Transfer learning for visual categorization: a survey. IEEE Trans. Neural Netw. Learn. Syst. **26**, 1019–1034 (2015)
71. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep Boltzmann machines. J. Mach. Learn. Res. **15**(1), 2949–2980 (2014)
72. Salakhutdinov, R., Hinton, G.: Deep Boltzmann machines. In: Artificial Intelligence and Statistics, pp. 448–455 (2009)
73. Koo, J.H., Cho, S.W., Baek, N.R., Kim, M.C., Park, K.R.: CNN-based multimodal human recognition in surveillance environments. Sensors **18**, 3040 (2018)
74. Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524 (2014)
75. Girshick, R.: Fast R-CNN. arXiv:1504.08083 (2015)
76. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv:1506.01497 (2016)
77. Lin, T.-Y. et al.: Feature pyramid networks for object detection. arXiv:1612.03144 (2017)
78. Liu, W. et al.: SSD: single shot multibox detector, pp. 21–37. arXiv:1512.02325 (2016)
79. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv:1708.02002 (2018)
80. Uijlings, J.R., Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**, 154–171 (2013)
81. Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., López, A.M.: Multimodal end-to-end autonomous driving. arXiv:1906.03199 (2019)
82. 1.Mohanapriya, D., Mahesh, K.: Chapter 5—an efficient framework for object tracking in video surveillance. In: The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, pp. 65–74 (2020)
83. Rangesh, A., Trivedi, M.M.: No blind spots: full-surround multi-object tracking for autonomous vehicles using cameras and LiDARs. IEEE Trans. Intelli. Veh. **4**, 588–599 (2019)
84. Liu, L., et al.: Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**, 261–318 (2020)
85. Nowlan, S., Platt, J.: A convolutional neural network hand tracker. In: Advances in Neural Information Processing Systems, pp. 901–908 (1995)
86. Ciaparrone, G., et al.: Deep learning in video multi-object tracking: a survey. Neurocomputing **381**, 61–88 (2020)
87. Anderson, P. et al.: Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3674–3683 (2018)
88. Wang, X. et al.: Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. arXiv:1811.10092 (2019)
89. Das, A. et al.: Embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–10 (2018)
90. Yu, L. et al.: Multi-target embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6309–6318 (2019)
91. Wang, A., Lu, J., Wang, G., Cai, J., Cham, T.-J.: Multi-modal unsupervised feature learning for RGB-D scene labeling. In: Computer Vision—ECCV, pp. 453–467 (2014)
92. Dargan, S., Kumar, M.: A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Syst. Appl. **143**, 113114 (2020)
93. Ammour, B., Boubchir, L., Bouden, T., Ramdani, M.: Face-Iris multimodal biometric identification system. Electronics **9**, 85 (2020)
94. Namin, S.T., Najafi, M., Salzmann, M., Petersson, L.: Cutting edge: soft correspondences in multimodal scene parsing. In: IEEE International Conference on Computer Vision (ICCV), pp. 1188–1196 (2015)
95. Zou, C., Guo, R., Li, Z., Hoiem, D.: Complete 3D scene parsing from an RGBD image. Int. J. Comput. Vis. **127**, 143–162 (2019)
96. Escalera, S., Athitsos, V., Guyon, I.: Challenges in multimodal gesture recognition. J. Mach. Learn. Res. **17**, 1–54 (2016)
97. Nishida, N., Nakayama, H.: Multimodal gesture recognition using multi-stream recurrent neural network. In: Revised Selected Papers of the 7th Pacific-Rim Symposium on Image and Video Technology, pp. 682–694 (2015)
98. Miao, Q. et al.: Multimodal gesture recognition based on the ResC3D network. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 3047–3055 (2017)
99. Tran, D., Ray, J., Shou, Z., Chang, S.-F., Paluri, M.: ConvNet architecture search for spatiotemporal feature learning. arXiv:1708.05038 (2017)
100. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5005–5013 (2016)
101. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans. Graph. **35**, 119:1–119:12 (2016)

102. Lin, T.-Y., Yin Cui, Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5007–5015 (2015)
103. Vo, N. et al.: Composing text and image for image retrieval—an empirical odyssey. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6439–6448 (2019)
104. Xu, Y.: Deep learning in multimodal medical image analysis. In: Health Information Science, pp. 193–200 (2019)
105. Shi, F., et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev. Biomed. Eng. **1**, 2020 (2020)
106. Santosh, K.C.: AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. J. Med. Syst. **44**, 93 (2020)
107. Wang, X., et al.: Convergence of edge computing and deep learning: a comprehensive survey. IEEE Commun. Surv. Tutorials **1**, 2020 (2020)
108. Ruder, S.: An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098 (2017)
109. Ruder, S., Bingel, J., Augenstein, I., Søgaard, A.: Latent Multi-task Architecture Learning. arXiv:1705.08142 (2018)
110. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
111. Duong, L., Cohn, T., Bird, S., Cook, P. low resource dependency parsing: cross-lingual parameter sharing in a neural network parser. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 845–850 (2015)
112. Peng, Y., et al.: CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network. IEEE Trans. Multimed. **20**(2), 405–420 (2017)
113. Palaskar, S., Sanabria, R., Metze, F.: Transfer learning for multi-modal dialog. Comput. Speech Lang. **64**, 101093 (2020)
114. Libovický, J., Helcl, J.: Attention strategies for multi-source sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers), pp. 196–202 (2017)
115. He, G., et al.: Classification-aware semi-supervised domain adaptation. In: CVPR, pp. 964–965 (2020)
116. Rao, R., et al.: Quality and relevance metrics for selection of multimodal pretraining data. In: CVPR, pp. 956–957 (2020)
117. Bucci, S., Loghmani, M.R., Caputo, B.: Multimodal Deep Domain Adaptation. arXiv:1807.11697 (2018)
118. Zhang, Y., Tan, H., Bansal, M.: Diagnosing the Environment Bias in Vision-and-Language Navigation. arXiv:2005.03086 (2020)
119. Landi, F., et al.: Perceive, Transform, and Act: Multi-Modal Attention Networks for Vision-and-Language Navigation. arXiv:1911.12377 (2020)
120. Krantz, et al.: Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. arXiv:2004.02857 (2020)
121. Nguyen, K., et al.: Vision-based Navigation with Language-based Assistance via Imitation Learning with Indirect Intervention. arXiv:1812.04155 (2019)
122. Cangea, et al.: VideoNavQA: Bridging the Gap between Visual and Embodied Question Answering. arXiv:1908.04950 (2019)
123. Zarbakhsh, P., Demirel, H.: 4D facial expression recognition using multimodal time series analysis of geometric landmark-based deformations. Vis. Comput. **36**, 951–965 (2020)
124. Joze, H.R.V., et al.: MMTM: multimodal transfer module for CNN fusion. In: CVPR, pp. 13289–13299 (2020)
125. Cadene, et al.: MUREL: multimodal relational reasoning for visual question answering. In: CVPR, pp. 1989–1998 (2019)
126. Fan, C. et al.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR, pp. 1999–2007 (2019)
127. Le, et al.: Hierarchical Conditional Relation Networks for Video Question Answering. arXiv:2002.10698 (2020)
128. Laina, I., et al.: Towards unsupervised image captioning with shared multimodal embeddings. In: ICCV, pp. 7414–7424 (2019)
129. Jang, Y., et al.: Video question answering with spatio-temporal reasoning. Int. J. Comput. Vis. **127**, 1385–1412 (2019)
130. Wang, W., et al.: A survey of zero-shot learning: settings, methods, and applications. ACM Trans. Intell. Syst. Technol. **10**, 13:1–13:37 (2019)
131. Wei, L., et al.: A single-shot multi-level feature reused neural network for object detection. Vis. Comput. (2020). https://doi.org/10.1007/s00371-019-01787-3
132. Hascoet, T., et al.: Semantic embeddings of generic objects for zero-shot learning. J. Image Video Proc. **2019**, 13 (2019)
133. Liu, Y., et al.: Attribute attention for semantic disambiguation in zero-shot learning. In: ICCV, pp. 6697–6706 (2019)
134. Li, K., et al.: Rethinking zero-shot learning: a conditional visual classification perspective. In: ICCV, pp. 3582–3591 (2019)
135. Liu, Y., Tuytelaars, T.: A: deep multi-modal explanation model for zero-shot learning. IEEE Trans. Image Process. **29**, 4788–4803 (2020)
136. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR, pp. 5542–5551 (2018)
137. Kumar, Y. et al.: Harnessing GANs for Zero-shot Learning of New Classes in Visual Speech Recognition. arXiv:1901.10139 (2020)
138. Zhang, X., et al.: Online multi-object tracking with pedestrian re-identification and occlusion processing. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01854-0
139. Abbass, M.Y., et al.: Efficient object tracking using hierarchical convolutional features model and correlation filters. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01833-5
140. Xi, P.: An integrated approach for medical abnormality detection using deep patch convolutional neural networks. Vis. Comput. **36**, 1869–1882 (2020)
141. Parida, K., et al.: Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In: CVPR, pp. 3251–3260 (2020)
142. Lee, J. A., et al.: Deep step pattern representation for multimodal retinal image registration. In: CVPR, pp. 5077–5086 (2019)
143. Hashemi Hosseinabad, S., Safayani, M., Mirzaei, A.: Multiple answers to a question: a new approach for visual question answering. Vis. Comput. (2020). https://doi.org/10.1007/s00371-019-01786-4
144. Yan, P., et al.: Adversarial image registration with application for mr and trus image fusion. arXiv:1804.11024 (2018)
145. Horry, Michael. J. et al.: COVID-19 Detection through Transfer Learning using Multimodal Imaging Data. IEEE Access 1 (2020) https://doi.org/10.1109/ACCESS.2020.3016780
146. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. J. Field Robot. **37**, 362–386 (2020)
147. Metzger, A., Drewing, K.: Memory influences haptic perception of softness. Sci. Rep. **9**, 14383 (2019)
148. Guclu, O., Can, A.B.: Integrating global and local image features for enhanced loop closure detection in RGB-D SLAM systems. Vis. Comput. **36**, 1271–1290 (2020)
149. Van Brummelen, J., et al.: Autonomous vehicle perception: the technology of today and tomorrow. Transp. Res. C Emerg. Technol. **89**, 384–406 (2018)
150. He, M., et al.: A review of monocular visual odometry. Vis. Comput. **36**, 1053–1065 (2020)
151. Liu, S., et al.: Accurate and robust monocular SLAM with omni-directional cameras. Sensors **19**, 4494 (2019)

152. Mur-Artal, R., Tardos, J.D.: ORB-SLAM2: an open-source SLAM system for monocular. Stereo RGB-D Cameras (2016). https://doi.org/10.1109/TRO.2017.2705103
153. Engel, J., et al.: LSD-SLAM: large-scale direct monocular SLAM. In: Computer Vision—ECCV, pp. 834–849 (2014)
154. Engel, J., et al.: Direct Sparse Odometry. arXiv:1607.02565 (2016)
155. Caesar, H., et al.: nuScenes: a multimodal dataset for autonomous driving. In: CVPR, pp. 11621–11631 (2020)
156. Gatys, L., et al.: A Neural Algorithm of Artistic Style. arXiv:1508.06576 (2015)
157. Lian, G., Zhang, K.: Transformation of portraits to Picasso's cubism style. Vis. Comput. **36**, 799–807 (2020)
158. Wang, L., et al.: Photographic style transfer. Vis. Comput. **36**, 317–331 (2020)
159. Zhang, Y. et al.: Multimodal style transfer via graph cuts. In: ICCV, pp. 5943–5951 (2019)
160. Wang, X., et al.: Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer. arXiv:1612.01895 (2017)
161. Jing, Y., et al.: Neural Style Transfer: A Review. arXiv:1705.04058 (2018)
162. DeepArts: turn your photos into art. https://deepart.io (2020). Accessed 18 Aug 2020
163. Waymo: Waymo safety report: On the road to fully self-driving. https://waymo.com/safety (2020). Accessed 18 Aug 2020
164. Wang, Z., Wu, Y., Niu, Q.: Multi-sensor fusion in automated driving: a survey. IEEE Access **8**, 2847–2868 (2020)
165. Ščupáková, K., et al.: A patch-based super resolution algorithm for improving image resolution in clinical mass spectrometry. Sci. Rep. **9**, 2915 (2019)
166. Bashiri, F.S., et al.: Multi-modal medical image registration with full or partial data: a manifold learning approach. J. Imag. **5**, 5 (2019)
167. Chen, C., et al. Progressive Feature Alignment for Unsupervised Domain Adaptation. arXiv:1811.08585 (2019)
168. Jin, X., et al.: Feature Alignment and Restoration for Domain Generalization and Adaptation. arXiv:2006.12009 (2020)
169. Guan, S.-Y., et al.: A review of point feature based medical image registration. Chin. J. Mech. Eng. **31**, 76 (2018)
170. Dapogny, A., et al.: Deep Entwined Learning Head Pose and Face Alignment Inside an Attentional Cascade with Doubly-Conditional fusion. arXiv:2004.06558 (2020)
171. Yue, L., et al.: Attentional alignment network. In: BMVC (2018)
172. Liu, Z., et al.: Semantic Alignment: Finding Semantically Consistent Ground-truth for Facial Landmark Detection. arXiv:1903.10661 (2019)
173. Hao, F., et al.: Collect and select: semantic alignment metric learning for few-shot learning. In: CVPR, pp. 8460–8469 (2019)
174. Wang, B., et al.: Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network. arXiv:1908.10072 (2019)
175. Wu, M., et al.: Audio caption: listen and tell. In: ICASSP, pp. 830–834 (2019) https://doi.org/10.1109/ICASSP.2019.8682377
176. Pan, B., et al. Spatio-temporal graph for video captioning with knowledge distillation. In: CVPR, pp. 10870–10879 (2020)
177. Liu, X., Xu, Q., Wang, N.: A survey on deep neural network-based image captioning. Vis. Comput. **35**, 445–470 (2019)
178. Abbass, M.Y., et al.: A survey on online learning for visual tracking. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01848-y
179. Guo, Y., et al.: Deep learning for visual understanding: a review. Neurocomputing **187**, 27–48 (2016)
180. Hatcher, W.G., Yu, W.: A survey of deep learning: platforms, applications and emerging research trends. IEEE Access **6**, 24411–24432 (2018)
181. Wu, X., Sahoo, D. Hoi, S.C.H.: Recent Advances in Deep Learning for Object Detection. arXiv:1908.03673 (2019)
182. Pouyanfar, S., et al.: A survey on deep learning: algorithms, techniques, and applications. ACM Comput. Surv. **51**, 92:1–92:36 (2018)
183. Ophoff, T., et al.: Exploring RGB+depth fusion for real-time object detection. Sensors **19**, 866 (2019)
184. Luo, Q., et al.: 3D-SSD: learning hierarchical features from RGB-D images for amodal 3D object detection. Neurocomputing **378**, 364–374 (2020)
185. Zhang, S., et al.: Video object detection base on rgb and optical flow analysis. In: CCHI, pp. 280–284 (2019). https://doi.org/10.1109/CCHI.2019.8901921
186. Simon, M., et al.: Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds. In: CVPRW, pp. 1190–1199 (2019). https://doi.org/10.1109/CVPRW.2019.00158
187. Tu, S., et al.: Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images. Precision Agric. **21**, 1072–1091 (2020)
188. Li, J., et al.: Facial expression recognition with faster R-CNN. Proc. Comput. Sci. **107**, 135–140 (2017)
189. Liu, S.: Enhanced situation awareness through CNN-based deep multimodal image fusion. OE **59**, 053103 (2020)
190. Michael, Y.B., Rosenhahn, V.M.: Multimodal Scene Understanding, 1st edn. Academic Press, London (2019)
191. Djuric, N., et al.: MultiXNet: Multiclass Multistage Multimodal Motion Prediction. arXiv:2006.02000 (2020)
192. Asvadi, A., et al.: Multimodal vehicle detection: fusing 3D-LIDAR and color camera data. Pattern Recogn. Lett. **115**, 20–29 (2018)
193. Mahmud, T., et al.: A novel multi-stage training approach for human activity recognition from multimodal wearable sensor data using deep neural network. IEEE Sens. J. (2020). https://doi.org/10.1109/JSEN.2020.3015781
194. Zhang, W., et al.: Robust Multi-Modality Multi-Object Tracking. arXiv:1909.03850 (2019)
195. Kandylakis, Z., et al.: Fusing multimodal video data for detecting moving objects/targets in challenging indoor and outdoor scenes. Remote Sens. **11**, 446 (2019)
196. Yang, R., et al.: Learning target-oriented dual attention for robust RGB-T tracking. In: ICIP, pp. 3975–3979 (2019). https://doi.org/10.1109/ICIP.2019.8803528
197. Lan, X., et al.: Modality-correlation-aware sparse representation for RGB-infrared object tracking. Pattern Recogn. Lett. **130**, 12–20 (2020)
198. Bayoudh, K., et al.: Transfer learning based hybrid 2D–3D CNN for traffic sign recognition and semantic road detection applied in advanced driver assistance systems. Appl. Intell. (2020). https://doi.org/10.1007/s10489-020-01801-5
199. Shamwell, E.J., et al.: Unsupervised deep visual-inertial odometry with online error correction for RGB-D imagery. IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/TPAMI.2019.2909895
200. Abavisani, M., et al.: Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training. arXiv:1812.06145 (2019)
201. Yang, X., et al.: A survey on canonical correlation analysis. IEEE Trans. Knowl. Data Eng. **1**, 2019 (2019)
202. Hardoon, D.R., et al.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**, 2639–2664 (2004)
203. Chandar, S., et al.: Correlational neural networks. Neural Comput. **28**, 257–285 (2016)
204. Engilberge, M., et al.: Finding beans in burgers: deep semantic-visual embedding with localization. In: CVPR, pp. 3984–3993 (2018)

205. Shahroudy, A., et al.: Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos. arXiv:1603.07120 (2016)
206. Srivastava, N., et al.: Multimodal learning with deep Boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2014)
207. Bank, D., et al.: Autoencoders. arXiv:2003.05991 (2020)
208. Bhatt, G., Jha, P., Raman, B.: Representation learning using step-based deep multi-modal autoencoders. Pattern Recogn. **95**, 12–23 (2019)
209. Liu, Y., Feng, X., Zhou, Z.: Multimodal video classification with stacked contractive autoencoders. Signal Process. **120**, 761–766 (2016)
210. Kim, J., Chung, K.: Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. IEEE Access **8**, 104933–104943 (2020)
211. Singh, V., et al.: Feature learning using stacked autoencoder for shared and multimodal fusion of medical images. In: Computational Intelligence: Theories, Applications and Future Directions, pp. 53–66 (2019)
212. Said, A. B., et al.: Multimodal deep learning approach for joint EEG-EMG data compression and classification. In: IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6 (2017)
213. Ma, L., et al.: Multimodal convolutional neural networks for matching image and sentence. In: IEEE International Conference on Computer Vision (ICCV), pp. 2623–2631 (2015)
214. Couprie, C., et al.: Toward real-time indoor semantic segmentation using depth information. J. Mach. Learn. Res. (2014)
215. Madhuranga, D., et al.: Real-time multimodal ADL recognition using convolution neural networks. Vis. Comput. (2020)
216. Gao, M., et al.: RGB-D-based object recognition using multimodal convolutional neural networks: a survey. IEEE Access **7**, 43110–43136 (2019)
217. Zhang, Z., et al.: RGB-D-based gaze point estimation via multi-column CNNs and facial landmarks global optimization. Vis. Comput. (2020)
218. Singh, R., et al.: Combining CNN streams of dynamic image and depth data for action recognition. Multimed. Syst. **26**, 313–322 (2020)
219. Abdulnabi, A.H., et al.: Multimodal recurrent neural networks with information transfer layers for indoor scene labeling. IEEE Trans. Multimed. **20**, 1656–1671 (2018)
220. Zhao, D., et al.: A multimodal fusion approach for image captioning. Neurocomputing **329**, 476–485 (2019)
221. Li, X., et al.: Multi-modal gated recurrent units for image description. Multimed. Tools Appl. **77**, 29847–29869 (2018)
222. Sano, A., et al.: Multimodal ambulatory sleep detection using lstm recurrent neural networks. IEEE J. Biomed. Health Inform. **23**, 1607–1617 (2019)
223. Shu, Y., et al.: Bidirectional multimodal recurrent neural networks with refined visual features for image captioning. In: Internet Multimedia Computing and Service, pp. 75–84 (2018)
224. Song, H., et al.: S^{2}RGANS: sonar-image super-resolution based on generative adversarial network. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01986-3
225. Ma, T., Tian, W.: Back-projection-based progressive growing generative adversarial network for single image super-resolution. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01843-3
226. Rohith, G., Kumar, L.S.: Paradigm shifts in super-resolution techniques for remote sensing applications. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01957-8
227. Jia, X., et al.: TICS: text-image-based semantic CAPTCHA synthesis via multi-condition adversarial learning. Vis. Comput. (2021). https://doi.org/10.1007/s00371-021-02061-1
228. Fan, X., et al.: Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal Person re-identification. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-02015-z
229. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)
230. Huang, X., et al.: Multimodal unsupervised image-to-image translation. In: CVPR, pp. 172–189 (2018)
231. Toriya, H., et al.: SAR2OPT: image alignment between multimodal images using generative adversarial networks. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 923–926 (2019)
232. Chaudhari, S., et al.: An Attentive Survey of Attention Models. arXiv:1904.02874 (2020)
233. Hori, C., et al.: Attention-based multimodal fusion for video description. In: IEEE International Conference on Computer Vision (ICCV), pp. 4203–4212 (2017)
234. Huang, X., Wang, M., Gong, M.: Fine-grained talking face generation with video reinterpretation. Vis. Comput. **37**, 95–105 (2021)
235. Liu, Z., et al.: Multi-level progressive parallel attention guided salient object detection for RGB-D images. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01821-9
236. Yang, Z., et al.: Stacked attention networks for image question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–29 (2016)
237. Guo, L., et al.: Normalized and geometry-aware self-attention network for image captioning. In: CVPR, pp. 10327–10336 (2020)
238. Bayoudh, K., et al.: Hybrid-COVID: a novel hybrid 2D/3D CNN based on cross-domain adaptation approach for COVID-19 screening from chest X-ray images. Phys. Eng. Sci. Med. **43**, 1415–1431 (2020)
239. Zhang, S., et al.: Joint learning of image detail and transmission map for single image dehazing. Vis. Comput. **36**, 305–316 (2020)
240. Zhang, S., He, F.: DRCDN: learning deep residual convolutional dehazing networks. Vis. Comput. **36**, 1797–1808 (2020)
241. Basly, H., et al.: DTR-HAR: deep temporal residual representation for human activity recognition. Vis. Comput. (2021). https://doi.org/10.1007/s00371-021-02064-y
242. Zhou, T., et al.: RGB-D salient object detection: a survey. Comp. Vis. Med. (2021). https://doi.org/10.1007/s41095-020-0199-z
243. Savian, S., et al.: Optical flow estimation with deep learning, a survey on recent advances. In: Deep Biometrics, pp. 257–287 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Khaled Bayoudh received a Bachelor's degree in Computer Science from the Higher Institute of Computer Science and Mathematics of Monastir (ISIMM), University of Monastir, Monastir, Tunisia, in 2014. Then, he graduated with a Master's degree in Highway and Traffic Engineering: Curricular Reform for Mediterranean Area (HiT4Med) from the National Engineering School of Sousse (ENISo), University of Sousse, Sousse, Tunisia, in 2017. In 2018, he received the M1 Master's degree in Software Engineering from ISIMM. He is currently a PhD student at the National School of Engineering of Monastir (ENIM), and a researcher in the Electronics and Micro-electronics Laboratory (EμE) at the Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia. His research focuses on Artificial Intelligence, Machine Learning, Deep Learning, Multimodal and Hybrid Learning, Intelligent Systems, and so on.

Raja Knani obtained a Master's degree in Micro and Nanoelectronics from the FSM, University of Monastir, Monastir, Tunisia, in 2014. She is currently a PhD student, and a researcher in the Electronics and Microelectronics Laboratory (EμE) at the FSM, University of Monastir, Monastir, Tunisia. She is interested particularly in Artificial Intelligence, Human-computer interaction, Gesture recognition and tracking, and so on.

Fayçal Hamdaoui received the Electrical Engineering degree from the National School of Engineering of Monastir (ENIM), University of Monastir, Tunisia, in 2010. In July 2011, he graduated with a Master diploma and in 2015 with a PhD Degree, both in the Electrical Engineering and both from ENIM. He is currently an Associate Professor at ENIM and a researcher in the Laboratory of Control, Electrical Systems and Environment (LASEE) at ENIM. His research interests are use of Artificial Intelligence (Deep Learning and Machine Learning), soft computing on image and video processing, embedded systems, SoC and SoPC programming.

Abdellatif Mtibaa is currently full Professor in Micro-Electronics, Hardware Design and Embedded System with Electrical Department at the National School of Engineering of Monastir and Head of Circuits Systems Reconfigurable-ENIM-Group at Electronic and microelectronic Laboratory. He holds a Diploma in Electrical Engineering in 1985 and received his PhD degree in Electrical Engineering in 2000. His current research interests include System on Programmable Chip, high level synthesis, rapid prototyping and reconfigurable architecture for real-time multimedia applications. Dr. Abdellatif MTIBAA has authored/co-authored over 200 papers in international journals and conferences. He served on the technical program committees for several international conferences. He also served as a co-organizer of several international conferences.