computers

MDPI

Review

A Short Survey on Deep Learning for Multimodal Integration: Applications, Future Perspectives and Challenges

Giovanna Maria Dimitri

Dipartimento di Ingegneria Dell'Informazione e Scienze Matematiche (DIISM), Università degli Studi di Siena, 53100 Siena, Italy; giovanna.dimitri@unisi.it

Abstract: Deep learning has achieved state-of-the-art performances in several research applications nowadays: from computer vision to bioinformatics, from object detection to image generation. In the context of such newly developed deep-learning approaches, we can define the concept of multi-modality. The objective of this research field is to implement methodologies which can use several modalities as input features to perform predictions. In this, there is a strong analogy with respect to what happens with human cognition, since we rely on several different senses to make decisions. In this article, we present a short survey on multimodal integration using deep-learning methods. In a first instance, we comprehensively review the concept of multimodality, describing it from a two-dimensional perspective. First, we provide, in fact, a taxonomical description of the multimodality concept. Secondly, we define the second multimodality dimension as the one describing the fusion approaches in multimodal deep learning. Eventually, we describe four applications of multimodal deep learning to the following fields of research: speech recognition, sentiment analysis, forensic applications and image processing.

Keywords: deep learning; multi-modal; integration; fusion

Citation: Dimitri, G.M. A Short Survey on Deep Learning for Multimodal Integration: Applications, Future Perspectives and Challenges. *Computers* **2022**, *11*, 163. https://doi.org/10.3390/computers11110163

Academic Editor: Robertas Damaševičius

Received: 25 October 2022
Accepted: 15 November 2022
Published: 18 November 2022

1. Introduction

In recent years, deep-learning (DL) approaches have opened up the path to addressing tasks which have been considered, so far, almost impossible to tackle [1]. Research challenges which previously, in fact, had to rely on feature-extraction steps, were then efficiently and successfully addressed through the use of non-features-engineering-based DL methods, constituting a significant revolution in the field of artificial intelligence (AI).

To give some examples, successful applications of DL can be found in several different fields of applications, such as, for instance, computer vision [2–4], bioinformatics [5] and natural language processing [6,7].

In all of these cases of applications, the strength of deep neural networks (DNN) relies on the possibility of processing directly to structured data, such as, for instance, images or graphs, without the need of a pre-processing features-extraction step [8].

In this context, the field of multimodal DL has recently gained more attention. Several DL fields of research were moved, in fact, towards trying to build systems that could mimic as much as possible the learning behaviour of a human being. Trying to answer such a research question, multimodality becomes one of the main research directions to be pursued. Our world is, indeed, inherently multimodal [9]. We, in fact, constantly rely on several multiple senses (e.g., audio, visual, touch) and combine their knowledge to make our decisions. Each modality [10] we come into contact with, brings, in fact, knowledge to us [11]. If, therefore, we want to have DL systems more similar to humans [12], we need to move towards the multimodality framework [13]. To offer a simple example, we could think of a basic concept, such as the concept related to the animal dog. When, as children, we built our prototypical concept concerning the animal dog, we certainly relied on all of our sense [14]: vision (looking at dog images as well as the real animal and in this

way learning how the animal should look), information about the sound produced by the animal (for instance, when the dog was barking), as well as other details such as the way in which the animal interacts with the surrounding world. A graphical example of such a multimodal cognitive process is represented in Figure 1.

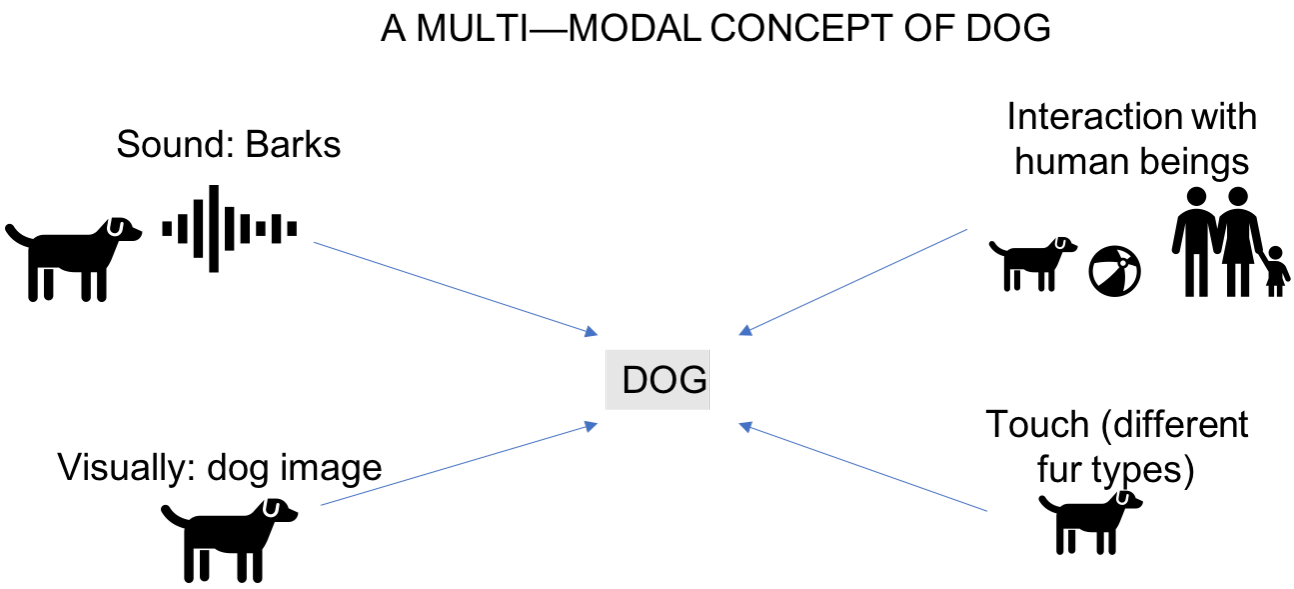

Figure 1. Conceptual representation of the animal dog. As humans, we rely on several of our senses to learn such a concept. Fusing together all of these pieces of information, we construct a concept and we then become able to recognize the animal.

The goal of this paper is to provide an overview and survey of DL-based approaches for multi-modal integration. The design challenges of the present study were manifold. A unique consensus for the definition of multimodality is not present in the literature. Moreover, applications are widespread between different research fields. Therefore, we decided to focus on four different fields of applications, in such a way that the examples provided could be explanatory and useful for introducing the importance of multimodal. Our survey offers a novel perspective on multimodality definition. For the first time in the literature, to the best of our knowledge, we defined multimodality as a two-dimensional concept: the first dimension concerning what we defined as a multimodal taxonomy and the second dimension focusing on the fusion steps in multimodal architecture.

The paper is structured as follows. In Section 2, we introduced the general principle of multimodality. In Section 3, DL applications are presented, in particular, four applications related to the research fields of: speech recognition, sentiment analysis, deep-fake recognition and computer vision. Moreover, in Section 4, we summarize and discuss future perspectives on the field.

2. Multimodality: A Definition

The world surrounding us can be defined as inherently multimodal. In our day-to-day learning experiences, we collect information using all of our five senses to be able to gather data to later use in our personal cognition experience. Similarly, a multimodal dataset, in machine learning, can be defined as a dataset in which multiple data types, related to the same concept, have been collected. The word multimodality, in fact, indicates the presence of several different types of data to be used as input in the implemented DL model [15]. If a uni-modal approach can, in fact, provide only a partial view of the research problem at hand, using multiple data sources (i.e., a multimodal approach) could actually significantly improve the learning process [16].

More formally, in this review, we introduce the definition of multimodality as a two-dimensional concept. In particular, the two dimensions can be defined as follows:

1 ***Dimension 1 Taxonomy****: concerns a taxonomical definition of the concept of multimodality*
2 ***Dimension 2 Fusion****: instead, focuses on the types of fusion to be used for the multimodal features*

To define the taxonomy of multimodality, i.e., dimension 1, we used the formalism of the seminal paper [15]. In particular, the taxonomical description of multimodality is based on the definition of five key concepts, which we will later describe in Section 2.1.

In contrast, the second dimension focuses on when and how the multimodal features-merging process takes place. If this is performed in a pre-processing step, we call it *early fusion*. If, instead, this process is left to a later step, taking place after the classification or prediction performed by the DL architecture, it is called *late fusion*. A halfway merging solution could happen, in which features are fused during the model deployment, or half at the beginning and half at the end of the learning process. This latter is named *hybrid fusion*. We will describe this dimension in depth in Section 2.2.

The relationship between the two multimodality dimensions is graphically represented in Figure 2.

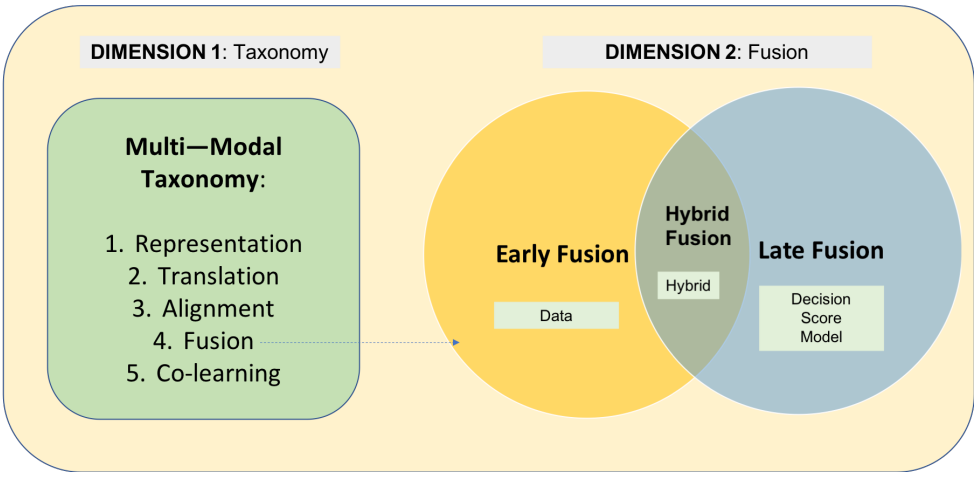

Figure 2. Figure showing the relationship exiting between the two defined multimodality dimensions, taxonomy and fusion, described in Section 2.

2.1. Dimension 1: Taxonomy

The taxonomy description of multimodality we report here was originally proposed in [15]. However, to the best of our knowledge, this is the first time in the literature that multimodality taxonomy is defined as one of the two multimodality dimensions. The seminal paper [15] defines five fundamental research challenges in multimodality.

In particular [15], the authors state that with a taxonomical definition of multimodality, we can provide a new vision of it, going beyond the more usual description of fusion.

The five taxonomical multimodal research challenges are [15]:

1 **Representation**: this concept addresses the need of representing and summarizing in a proper way the input features, irrespective of the differences existing between the modalities. To give an example, language is often represented as symbolic, while visual modalities are often represented as signals.
2 **Translation**: this point concerns the capability of translating one modality to another. This task is particularly crucial, as it determines that the relationship between modalities might exist. If this is not case, then this should be carefully taken into account in the multimodality-DL implementation.
3 **Alignment**: in this case, this concept refers to the need to identify the direct relationships between one or more sub-element parts of the same input data. In [15], the authors assess the need of measuring the similarity between the modalities used in the experimental setting.

4 **Fusion**: this research challenge tackles, instead, the need of joining information in two or more modalities, to perform a certain prediction task. We will expand on this in the definition of Section 2.2.
5 **Co-learning**: this research challenge entails the need of transferring knowledge between the modalities used in the experimental settings. This is said to be particularly relevant [15] for several types of DL algorithms such as, for example, zero-shot learning, which is a new DL approach in which models are tested on examples which were never seen during training. Moreover, this research aspect is said to be particularly relevant when one of the modalities is characterized by a limited amount of resources (for example, a limited amount of labelled data). Several deep-learning extensions have been proposed for the co-learning research field. A relevant overview concerning this concept is reported in [17].

2.2. Dimension 2: Fusion

The second multimodality dimension we defined concentrates on the fusion process. More specifically, dimension 2 focuses on the study of *when* and *if* the various input modalities will be merged during the learning process. In particular, we could more formally define the following fusion paradigms [16,18]:

1 **Early fusion**: such fusion happens when the multimodal input data is fused prior to the application of the AI model. More specifically, the fusion process takes places as at the initial step, before the dataset is used as input to the DL algorithm. More specifically, we could say that the fusion process happens directly on the raw data. If, instead of the raw data, a pre-processing feature-extraction step is performed, then we say that the merging step is performed at a feature level.
2 **Late fusion**: in this case, the fusion step takes place after the application of the AI algorithm. In this case, data are processed in a separate multimodal way.
More specifically, such an approach considers the different modalities as single streams. The drawback is that the possible conditional relationship existing among the different modalities are not considered during the learning process. Merging strategies could be of different types:
3 **Hybrid fusion**: this fusion takes place when the multimodal input data is fused both before and after the application of the relevant AI algorithm. In particular, this approach stands in the middle between late and early fusion, taking place halfway in the DNN model considered. This approach could be, for example, particularly suitable when there are modalities with consistent dimensions which need to be merged, as well as when the modalities are of a very different nature, and, therefore, require a first pre-processing and a later merging procedure during the training process.

We could further use the distinction presented in [19] to define the following five types of fusion methods.

These five types can be considered as sub-types of the three aforementioned fusion modalities:

1 **Data-level fusion**: this is defined as the case in which the modalities are fused before the learning process. In other words, features are computed independently for each single modality, and the fusion step takes place when they are already in the feature-extracted form [20,21]. This is part of the early fusion modality approach.
2 **Decision-level fusion**: in this case, the decision scores are firstly computed separately for each AI model applied. Subsequently, the individual decision scores are fused into one decision, using, for example, an ensemble of methods, or majority voting approaches [22,23]. This is part of the late fusion-modality approach.
3 **Score-level fusion**: in this case, the fusion mechanism happens at the level of the probabilistic output scores of neural networks. It is different from decision-level fusion, as, in this case, a proper fusion of scores is performed, while in the previous case, majority voting and ensembling approaches were applied [24]. A single decision is

then taken only after the fused scores have been produced. This is part of the late fusion-modality approach.

4 **Hybrid-level fusion**: in this case, the characteristics of the features- and decision-level fusion are merged. In particular, a comparison takes place between the decisions taken by single classifiers and the decisions taken by classifiers when modalities are fused at a feature level. Such approaches, therefore, rely on the idea of improving the performances of single classifiers with the use of features-future engineering [25,26]. This is part of the hybrid-modality approach.

5 **Model-level fusion**: in this case, the fusion takes place during the training and learning procedure of the AI model implemented [27,28]. This is part of the hybrid-modality approach.

3. Multimodality Deep-Learning Applications

Several multimodal DL architectures have been proposed [9]. In this section, we will consider four fields of the application of multimodality to DL. In particular, we will present examples for the following research fields: speech recognition, sentiment analysis, DeepFake forensic detection and image reconstruction or segmentation.

In Figure 3, we present a graphical summary abstract of the four areas of application presented in our paper. The following subsections are organized as follows. In Section 3.1, we will describe applications of multimodal approaches for speech recognition. In Section 3.2, we will describe application concerning the detection of sentiments in audio–visual–text datasets. Later, in Section 3.3, applications to the forensic field will be described. Eventually, in Section 3.4, we will describe computer-vision applications with particular reference to image reconstruction and segmentation. An overall summary of the applications and papers reviewed in this Section is presented in Table 1. In the table, we report the model used, together with the features and the relevant paper references. In the discussion of the applications and of the models, we will mainly describe the performances obtained and the input multimodal features used.

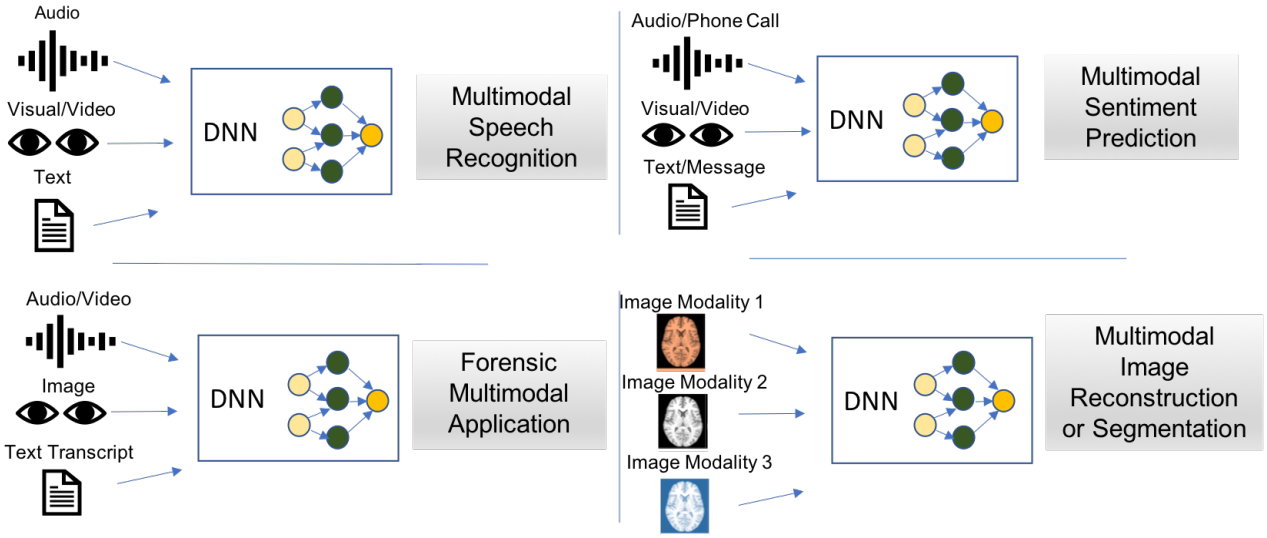

Figure 3. Example of workflows for multimodal speech recognition, sentiment analysis, multi-modal image reconstruction and forensic multimodal application.

Table 1. Table summarizing the various applications listed in Section 3. The first column reports the application type and second column the DL model used. The third column includes a brief description of the input multimodal features and the last column a reference to the relevant paper.

Application	Model	Multi-Modal Features	Reference
Speech Recognition	HMM	visual + audio	McGurk et al. [29]
	Restricted Boltzamn machines	audio + video	Ngiam et al. [30]
	Deep Boltzman Machines	image + text	Srivastava et al. [31]
	Multimodal Compact Bilinear pooling (MCB)	visual + text	Fukui et al. [32]
	RNN	audio + visual	Makino et al. [33]
	Attention Based DL	audio + visual	Petridis et al. [34]
	Seq2Seq	audio + visual	Zhou et al. [35]
	DL with gating layer	audio + visual	Tao et al. [36]
	Bidirectional Gated Recurrent Units	audio + visual	Petridis et al. [37]
Sentiment Analysis	Convolutional Deep Belief Network	face + body gesture + voice + physiological signals	Ranganathan et al. [38]
	cLSTM-MMA	visual + textual	Pan et al. [39]
	transformers	audio + visual + text	Khare et al. [40]
	LSTM	lyric + audio	Liu et al. [41]
	CNN	visual + text + audio	Cambria et al. [42]
	CNN	image + text	Lee et al. [43]
	FC	audio + video	Ortega et al. [44]
	LSTM and FC	physiological	Dhaouadi et al. [45]
	FC	physiological	Bizzego et al. [46]
	FC	text + audio + video	Ray et al. [47]
Forensic Applications	MLP, SincNet, Xecption	audio + video	Lomnitz et al. [48]
	LSTM + MLP	multimodal spectral features	Lewis et al. [49]
	Siamese Networks	audio + video	Mittal et al. [50]
	Xception, VGG, SincNet	video + audio	Khalid et al. [51]
	CNN based	video + audio	Cai et al. [52]
Computer Vision	multiple DNN	brain MRI scans	Zhang et al. [53]
	visible,X-ray,thermal,infrared radiation	Mask R-CNN	Pemasiri et al. [54]
	multiple DL	several sensors	Feng et al. [55]
	GAN	multispectral	Hong et al. [56]
	CNN separable convolutions	brain MRI T1 + MRI T2 + CT	Dimitri et al. [57]
	Multi-Modal U-Net	T2W and FLAIR brain scans	Falvo et al. [58]

3.1. Audio-Visual Multimodal Applications for Speech Recognition

Speech recognition is a research field in which multimodality can play a fundamental role. Indeed, one of the earliest examples of multimodal DL applications concerns the use of audio-visual features for speech recognition. Such systems are defined as AVSR, i.e., audio-visual speech recognition systems.

The underlying motivation [15] behind the use of multimodality in AVSR was, in fact, theorised early by the so called McGurk effect described in [29], in which the interaction between hearing and vision during speech was analysed. Among the various findings, the authors in the paper reported [15,29] experimental results concerning the importance of using both audio and visual features to perform speech recognition. More specifically, during the experiments performed, the fact was noticed that when a human being was hearing the words */ba-ba/* at the same time as watching the person's lips pronouncing */ga-/ga/*, the final perception was the sound */da-da/*.

Among the oldest speech-language recognition works, we can find architectures based on hidden Markov models (HMM) approaches, particularly popular at the time of the seminal paper concerning the McGurk effect [29]. However, nowadays, state-of-the-art performances, in multimodal speech-language recognition, are obtained using DL approaches [30].

For instance, in [30], which is one of the first examples of applications of DL for speech recognition, the authors combined features from video and audio, showing how the use of several different modalities could consistently improve the learning process (for instance, in the paper, audio and video). In particular, the work used Restricted Boltzmann machines to extract features for audio-visual classification. In more detail, a multimodal architecture was trained over the concatenated audio and video extracted features. The authors obtained state-of-the-art performances using benchmark datasets, and showed the importance of a multimodal approach through ablation studies.

Similarly, in [31], the authors proposed a Deep Boltzmann Machine approach with the aim of building a generative model for multimodal data. The model proposed was based on a bi-modal dataset made of images and text or, alternatively, audio and video. The experiments achieved good classification performances compared to the state of the art models such as support vector machines (SVMs).

Several other papers can be included in this line of research. For example, in [32], the authors addressed the issue of visual question answering. In the paper, the authors stressed the need to have a multi-modal input feature. In particular, such features could come from visual or textual representations, and could be successfully merged through the concatenation of visual and textual information. The method proposed is named Multimodal Compact Bilinear pooling (MCB). In the paper, the authors present a comprehensive assessment of multimodal performances, using a benchmark dataset for visual answering: in particular, the Visual7W dataset and the VQA challenge) [32]. The authors showed that the model proposed could significantly improve the state-of-the-art performances on the Visual7W dataset and on the Visual Question Answering challenge [32].

Another interesting application of recurrent neural networks (RNNs) for speech recognition is reported in [33]. In the work, an audio-visual speech recognition system is presented. For this purpose, a large dataset was collected and released, with more than 31k hours of audio-visual training material. Moreover, in [34], the authors provided a novel hybrid-attention-based DL model, in which a connectionist temporal classification approach was combined with a sequence-to-sequence model for character-level recognition.

Similarly, in [37], the authors present an end-to-end model which is based upon residual networks and bidirectional gated recurrent units. This is one of the first works in which image pixel information is fused with audio waveforms to perform context-word recognition. An attention model was also used in [35], where the authors provided a Seq2Seq DL attention-based model and used it for speech-recognition purposes, reaching state-of-the-art performances. Moreover, in [36], the authors proposed a DL approach with a gating layer, to avoid having the drop in performance that happens when visual features are merged with clean speech features.

3.2. Multimodal Applications for Sentiment Analysis

Multimodal sentiment analysis concerns the automatic analysis of different types of data (for example, visual or text) to predict emotions and feelings. Such a task is inherently a particularly challenging one, and the use of several sources and modalities could, in principle, play a fundamental role.

Even for a human being, recognizing an emotion only from a text message could be a challenging task. However, if the text message is accompanied by a vocal communication, the delivering of the message sentiment intent could become clearer. Even better is if there is the addition of some visual information, such as, for example, a video or a live interaction, between the two participants to the conversation. In other words, in this way, the message is conveyed through multiple modalities, and a certain clearer message-sentiment comprehension is delivered to the listener.

Several works present applications of DL for emotion and sentiment analysis, using a multimodal approach. For instance, in [38], the authors present a multimodal database emoF-BVP where the face, body gestures, voice and physiological signals of actors playing

different 23 different emotions were recorded. The multimodal database is enriched by facial and posture tracking, and a set of experiments was performed on the proposed dataset, showing promising improvements with respect to the state of the art. In particular, in this work [38], the authors proposed a model named convolutional deep belief network (CDBN), which is able to efficiently learn salient multimodal features to express emotions. The model is able to return consistent predictions when low-intensity or subtle expressions are detected, improving significantly the state-of-the-art performances.

Several other examples exist in the literature in which multimodal emotion-recognition systems made use of DL techniques [59].

In most of the recent works for automatic sentiment analysis, the features taken into consideration were mostly audio and text or image and text.

For example, in [39], the authors proposed a hybrid fusion process, based a on multimodal emotion assessment, using visual and text cues. The architecture proposed made use of a late-fusion approach. Moreover, in their work [39], the authors proposed a novel multimodal-attention-based mechanism, which they named cLST-MMA. The attention mechanism implemented could help in fusing relevant information from the different modalities, performing comprehensive experiments and reaching state-of-the-art performances in the benchmark IEMOCAP dataset for emotion recognition.

Analogously, several examples in the literature present the use of audio and text [40–42,59,60], and promising performances have been reported for these experiments.

However, in the field of multimodal applications for sentiment analysis detection, another common combination of features is represented by the use of visual features (i.e., videos) and text, for example, in [43] (images and text) or [44] (video and text).

Another interesting possible application of multimodality to the field of sentiment analysis is through the use of physiological monitored parameters in humans. For instance, in [45], the authors used physiological signals monitored on gamers to predict sentiments and stress levels. In particular, the monitored signals were electrocardiography (ECG), electrodermal activity (EDA), and electromyography (EMG), which were measured using non-invasive wearable sensors. In this way, real-time gamers' emotions and stress levels could be successfully retrieved. In particular, the authors showed the used of long short-term memory models together with other relevant neural networks models provide insights into the evaluation of both approaches to predict the real-time stress of gamers'.

Similarly, in [46], the authors collected a dataset composed of physiological parameters, in particular, ECG, electrodermal activity data and other types of physiological parameters that can be monitored through wearable sensors. Such signals were collected from 232 subjects using four different acquisition devices. The authors, moreover, applied a DNN-based approach to predict emotions from the collected physiological signals.

It is worth noticing, in this context, that much attention has been devoted to the collection of corpora and datasets, which can be used for the purpose of multimodal emotion recognition and classification. Such datasets are crucial for DL algorithms training, and, therefore, their collection is extremely important. An example of this is [47], where a multimodal, novel, publicly available dataset for sarcasm detection was released. Another interesting example is [61], where a multimodal-sensor-based dataset was collected and released.

3.3. Forensic Applications: Multimodal Deepfake Detection

In this subsection, we will consider a particularly recent and important application of computer engineering forensic, i.e., the detection and analysis of a deepfake.

Deepfake has become a wide spread phenomenon in recent years. It is possible to define a deepfake as content generated by AI, which seems authentic if evaluated through the eyes of a human being [62].

The word deepfake comes from the combination of the words *deep learning* and *fake*. If the deepfake is created with malicious intent, this can result in [62] a video of a person in

which their face or body has been digitally altered so that they appear to be someone else and could, for instance, be used to spread false information [63]. One important challenge becomes, therefore, the automatic detection of deepfakes. This task is usually performed through the use of DNN classifiers, capable of classifying correctly deepfakes vs pristine images [64]. The use of multimodal data in this context, for instance audio, video and images, becomes, therefore, crucial for the success of this detection task.

For example, in [48], a multimodal deepfake-detection solution was proposed in which single images together with temporal images were included in a single framework, obtaining very good performances.

Similarly, in [49], the same Facebook DeepFake challenge was used in the experimental framework implemented. In particular, the authors proposed a hybrid DL approach, based on spatial, spectral and temporal input features, with the aim of predicting real and fake videos, in a multimodal fashion. The experiments showed performances reaching 61.95% accuracy when tested on the Facebook Deepfake Detection Challenge (DFDC) dataset.

Moreover, in [50], the authors exploited audio and video features for deepfake detection. The method is based on the use of a Siamese network, and a triplet loss. To extract information for learning, in fact, the similarity between two videos was analysed, obtaining a classifier which could make the decision if a video was fake or not. The authors obtained an AUC of 84.4% on the DeepFake-TIMIT Dataset and an AUC of 96.6% on the DF-TIMIT dataset, two common benchmarks used for deepfake detection. In their work, the authors exploited, for the first time in the literature, the use of audio and video modalities, and used them to perceive and detect the presence of a deepfake.

Similarly, in [51], the authors used a benchmark called Audio-Video Multimodal Deepfake Detection Dataset (FakeAVCeleb) containing videos as well as synthesised fake audios to be used for deepfake classification. In the paper, the authors extensively compared one modality to multimodality, highlighting the superiority of a multimodal approach for performing deepfake detection.

Several other notable examples of deepfake detection can be found in the literature, such as, for instance, [52]. It has to be highlighted the importance, also in this case, of the construction of datasets and benchmarks. An example in this context is [65]. In this case, the authors collected not only deepfake videos, but also lip-synced fake audios.

3.4. Computer Vision and Multimodality: Image Segmentation and Reconstruction

Computer vision is a research area in which DL has played a fundamental role. In the last few years, a plethora of DL applications to computer visions have been proposed. This has led to surprising results. For instance, we could mention one of the latest research areas which has developed concerns images or video generation from text prompts [66,67]. In this section, we will present the role that multimodality can play when considering the two computer-vision tasks of image segmentation and reconstruction.

In the computer-vision area, however, one important remark should be made.

If, until now, the term modality has referred to the cases in which different types of input were used (e.g., audio + text, audio + video), in the case of images, the term multimodal could also refer to different modalities of images. For example, in biomedical imaging, brain scans could be of the modality of MRI or CT scans.

Concerning image segmentation, it consists in the task of labelling each pixel of an image with the relative object to which it belongs to [2]. In this context, DL has played a fundamental role, finding applications in several different research areas [68]. The development of well-known DL architectures, such as SegNet [69], DeepLab [70] or Xception [71] nets, have dramatically improved the state-of-the-art performances achieved by CNN-based architectures for segmentation tasks. In this context, multimodal approaches can play a fundamental role [68].

For instance, in [55], the authors report experiments performed for autonomous driving purposes. In particular, to achieve robust performances in scene understanding for

autonomous driving, a multimodal, in particular, a multi-sensor approach, was used to obtain the relevant scene segmentation map. Particular attention in the study was devoted to the fusion step [55].

In [54], instead, we can find a relevant example in which multimodal image segmentation of biomedical images is performed. In particular, the authors proposed a so-called modality invariant method, in order to obtain human body-part segmentation. In this case, the modality is intended to be semantically referred to different types of scans: visible images, X-ray images, thermal images (also named heatmaps) and infrared-radiation (IR) images. In the experiments, the authors first considered pairs of modalities, and later developed an architecture devoted to the understanding of a further relationship between multiple modalities. Subsequently, segmentation was performed using a CNN-based architecture.

Another notable example of multimodal biomedical image segmentation (in which multimodal is intended to mean different types of images) concerns the application of DL architectures for brain-tumour segmentation. Several notable examples have been proposed, using the publicly released and available dataset Brats [53]. Many challenges have, in fact, been proposed during the years, with the addition of different image modalities, which have brought improvements in the segmentation performances obtained.

A comprehensive table reporting performances and comparisons between Brats challenge participants is reported in [53]. Furthermore, another important and interesting field of application of image segmentation is the one referring to multispectral and satellite images [56,72,73]. In addition, in this case, a multimodal approach can be applied. For instance, in [56], the authors address the problem of semantic segmentation in large-scale urban satellite images with a scarce amount of cross-modality data. The experiments of the novel GAN-based architectures were evaluated on two multimodal (hyperspectral and multispectral) image datasets, achieving state-of-the-art performances.

We will now continue with the exploration of some case studies used in the field of image reconstruction.

Notable examples of such applications can be found in the biomedical-imaging field. For example, in [57], the authors proposed a multimodal brain MRI reconstruction architecture based on separable convolutional blocks. Several different brain-scan types were used as input. The quality of the reconstructed brain scans was assessed through image-quality indicators such as SNR (signal-to-noise ratio). The fusion was performed concatenating the latent embedding dimension space. Several other examples could be listed in this line of research, such as, for instance, [58,74,75].

4. Discussion and Conclusions

Deep-learning approaches have recently shown their capabilities in outperforming standard machine-learning methods in several tasks. The field of multimodal deep learning is a newly developing field. In the context of multimodal architecture, multiple data modalities are fused and used in the learning process, mimicking human behaviour during the cognitive learning process. The importance of multimodality has proved to be manifold and in several different tasks. In our brief overview, we present applications in computer-vision, forensic, natural-language processing tasks and sentiment analysis. We profoundly analysed four fields in which multimodal DL has been successfully applied. The four fields should give a broad overview of the numerous research applications that multimodality can have. We also reported a summary and comparison of the DL methodologies used. Evaluation on the performances were reported on the various methodologies. We considered a plethora of different methods, in order to provide the reader with a short applicative overview of the experimental fields analysed.

Future challenges are manifold. One of the most important matters constitutes the requirements of a novel dataset, i.e., benchmarks needed for training DL architectures.

We believe that the field is in continuous expansion and the promising results will allow the development of novel techniques and DL models in the next few years.

Funding: This research received no external funding.

Data Availability Statement: The data has been present in main text.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural networks
MLP	Multi-layer perceptrons
DL	Deep learning
ML	Machine learning
AI	Artificial intelligence
CNN	Convolutional neural networks
RNN	Recurrent neural network

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
2. Cheng, H.D.; Jiang, X.H.; Sun, Y.; Wang, J. Color image segmentation: Advances and prospects. *Pattern Recognit.* **2001**, *34*, 2259–2281. [CrossRef]
3. Dimitri, G.M.; Spasov, S.; Duggento, A.; Passamonti, L.; Toschi, N. Unsupervised stratification in neuroimaging through deep latent embeddings. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 1568–1571.
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
5. Cicaloni, V.; Spiga, O.; Dimitri, G.M.; Maiocchi, R.; Millucci, L.; Giustarini, D.; Bernardini, G.; Bernini, A.; Marzocchi, B.; Braconi, D.; et al. Interactive alkaptonuria database: Investigating clinical data to improve patient care in a rare disease. *FASEB J.* **2019**, *33*, 12696–12703. [CrossRef]
6. Iqbal, T.; Qureshi, S. The survey: Text generation models in deep learning. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *34*, 2515–2528. [CrossRef]
7. He, X.; Deng, L. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Process. Mag.* **2017**, *34*, 109–116. [CrossRef]
8. Bianchini, M.; Dimitri, G.M.; Maggini, M.; Scarselli, F. Deep neural networks for structured data. In *Computational Intelligence for Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 29–51.
9. Summaira, J.; Li, X.; Shoib, A.M.; Li, S.; Abdul, J. Recent Advances and Trends in Multimodal Deep Learning: A Review. *arXiv* **2021**, arXiv:2105.11087.
10. Van Leeuwen, T. Multimodality. In *The Routledge Handbook of Applied Linguistics*; Routledge: London, UK, 2011; pp. 668–682.
11. Jewitt, C.; Bezemer, J.; O'Halloran, K. *Introducing Multimodality*; Routledge: London, UK, 2016.
12. Bateman, J.; Wildfeuer, J.; Hiippala, T. *Multimodality: Foundations, Research and Analysis–A Problem-Oriented Introduction*; Walter de Gruyter GmbH & Co KG: Berlin, Germany, 2017.
13. Bernsen, N.O. Multimodality theory. In *Multimodal User Interfaces*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 5–29.
14. Bertelson, P.; De Gelder, B. The psychology of multimodal perception. *Crossmodal Space Crossmodal Attention, Spence*; Spence, C., Driver, J., Eds.; Oxford University Press: Oxford, UK, 2004; pp. 141–177.
15. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
16. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6.
17. Rahate, A.; Walambe, R.; Ramanna, S.; Kotecha, K. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* **2022**, *81*, 203–239. [CrossRef]
18. Snoek, C.G.; Worring, M.; Smeulders, A.W. Early versus late fusion in semantic video analysis. In Proceedings of the 13th annual ACM international conference on Multimedia, Singapore, 6–11 November 2005; pp. 399–402.
19. D'mello, S.K.; Kory, J. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.* **2015**, *47*, 1–36. [CrossRef]

20. Castellano, G.; Kessous, L.; Caridakis, G. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 92–103.
21. D'mello, S.K.; Graesser, A. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Model. User Adapt. Interact.* **2010**, *20*, 147–187. [CrossRef]
22. Kanluan, I.; Grimm, M.; Kroschel, K. Audio-visual emotion recognition using an emotion space concept. In Proceedings of the 2008 16th European Signal Processing Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 1–5.
23. Salur, M.U.; Aydın, İ. A soft voting ensemble learning-based approach for multimodal sentiment analysis. *Neural Comput. Appl.* **2022**, *34*, 18391–18406. [CrossRef]
24. Aizi, K.; Ouslim, M. Score level fusion in multi-biometric identification based on zones of interest. *J. King Saud-Univ.-Comput. Inf. Sci.* **2022**, *34*, 1498–1509. [CrossRef]
25. Mansoorizadeh, M.; Moghaddam Charkari, N. Multimodal information fusion application to human emotion recognition from face and speech. *Multimed. Tools Appl.* **2010**, *49*, 277–297. [CrossRef]
26. Chetty, G.; Wagner, M.; Goecke, R. A multilevel fusion approach for audiovisual emotion recognition. *Emot. Recognit. Pattern Anal. Approach* **2015**, *2015*, 437–460.
27. Metallinou, A.; Wollmer, M.; Katsamanis, A.; Eyben, F.; Schuller, B.; Narayanan, S. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affect. Comput.* **2012**, *3*, 184–198. [CrossRef]
28. Giacobe, N.A. Application of the JDL data fusion process model for cyber security. In Proceedings of the Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2010, SPIE, Orlando, FL, USA, 7–8 April 2010; Volume 7710, pp. 209–218.
29. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef]
30. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the ICML, Bellevue, WA, USA, 28 June–2 July 2011.
31. Srivastava, N.; Salakhutdinov, R.R. Multimodal learning with deep boltzmann machines. *Adv. Neural Inf. Process. Syst.* **2012**, *2012*, 25.
32. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv* **2016**, arXiv:1606.01847.
33. Makino, T.; Liao, H.; Assael, Y.; Shillingford, B.; Garcia, B.; Braga, O.; Siohan, O. Recurrent neural network transducer for audio-visual speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 905–912.
34. Petridis, S.; Stafylakis, T.; Ma, P.; Tzimiropoulos, G.; Pantic, M. Audio-visual speech recognition with a hybrid ctc/attention architecture. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), 2018; pp. 513–520.
35. Zhou, P.; Yang, W.; Chen, W.; Wang, Y.; Jia, J. Modality attention for end-to-end audio-visual speech recognition. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6565–6569.
36. Tao, F.; Busso, C. Gating neural network for large vocabulary audiovisual speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1290–1302. [CrossRef]
37. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. In Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552.
38. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
39. Pan, Z.; Luo, Z.; Yang, J.; Li, H. Multi-modal attention for speech emotion recognition. *arXiv* **2020**, arXiv:2009.04107.
40. Khare, A.; Parthasarathy, S.; Sundaram, S. Self-supervised learning with cross-modal transformers for emotion recognition. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Virtual, 19–22 January 2021; pp. 381–388.
41. Liu, G.; Tan, Z. Research on multi-modal music emotion classification based on audio and lyric. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; Volume 1, pp. 2331–2335.
42. Cambria, E.; Hazarika, D.; Poria, S.; Hussain, A.; Subramanyam, R. Benchmarking multimodal sentiment analysis. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing, Budapest, Hungary, 17–23 April 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 166–179.
43. Lee, J.H.; Kim, H.J.; Cheong, Y.G. A multi-modal approach for emotion recognition of tv drama characters using image and text. In Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Korea, 19–22 February 2020; pp. 420–424.
44. Ortega, J.D.; Senoussaoui, M.; Granger, E.; Pedersoli, M.; Cardinal, P.; Koerich, A.L. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv* **2019**, arXiv:1907.03196.
45. Dhaouadi, S.; Khelifa, M.M.B. A multimodal physiological-based stress recognition: Deep Learning models' evaluation in gamers' monitoring application. In Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2–5 September 2020; pp. 1–6.

46. Bizzego, A.; Gabrieli, G.; Esposito, G. Deep neural networks and transfer learning on a multivariate physiological signal Dataset. *Bioengineering* **2021**, *8*, 35. [CrossRef]
47. Ray, A.; Mishra, S.; Nunna, A.; Bhattacharyya, P. A Multimodal Corpus for Emotion Recognition in Sarcasm. *arXiv* **2022**, arXiv:2206.02119.
48. Lomnitz, M.; Hampel-Arias, Z.; Sandesara, V.; Hu, S. Multimodal Approach for DeepFake Detection. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020.
49. Lewis, J.K.; Toubal, I.E.; Chen, H.; Sandesera, V.; Lomnitz, M.; Hampel-Arias, Z.; Prasad, C.; Palaniappan, K. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020.
50. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM international Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2823–2832.
51. Khalid, H.; Kim, M.; Tariq, S.; Woo, S.S. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection, Virtual, 24 October 2021; pp. 7–15.
52. Cai, Z.; Stefanov, K.; Dhall, A.; Hayat, M. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. *arXiv* **2022**, arXiv:2204.06228.
53. Zhang, W.; Wu, Y.; Yang, B.; Hu, S.; Wu, L.; Dhelim, S. Overview of multi-modal brain tumor mr image segmentation. *Healthcare* **2021**, *9*, 1051. [CrossRef]
54. Pemasiri, A.; Nguyen, K.; Sridharan, S.; Fookes, C. Multi-modal semantic image segmentation. *Comput. Vis. Image Underst.* **2021**, *202*, 103085. [CrossRef]
55. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360. [CrossRef]
56. Hong, D.; Yao, J.; Meng, D.; Xu, Z.; Chanussot, J. Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5103–5113. [CrossRef]
57. Dimitri, G.M.; Spasov, S.; Duggento, A.; Passamonti, L.; Lió, P.; Toschi, N. Multimodal and multicontrast image fusion via deep generative models. *Inf. Fusion* **2022**, *88*, 146–160. [CrossRef]
58. Falvo, A.; Comminiello, D.; Scardapane, S.; Scarpiniti, M.; Uncini, A. A multimodal deep network for the reconstruction of T2W MR images. In *Progresses in Artificial Intelligence and Neural Systems*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 423–431.
59. Abdullah, S.M.S.A.; Ameen, S.Y.A.; Sadeeq, M.A.; Zeebaree, S. Multimodal emotion recognition using deep learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 52–58. [CrossRef]
60. Parry, J.; Palaz, D.; Clarke, G.; Lecomte, P.; Mead, R.; Berger, M.; Hofer, G. Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1656–1660.
61. Park, C.Y.; Cha, N.; Kang, S.; Kim, A.; Khandoker, A.H.; Hadjileontiadis, L.; Oh, A.; Jeong, Y.; Lee, U. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Sci. Data* **2020**, *7*, 1–16. [CrossRef]
62. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *Acm Comput. Surv.* **2021**, *54*, 1–41. [CrossRef]
63. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 1, p. 38.
64. Amerini, I.; Galteri, L.; Caldelli, R.; Del Bimbo, A. Deepfake video detection through optical flow based cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Virtual, 11–17 October 2019.
65. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**, arXiv:2108.05080.
66. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning. PMLR, Baltimore, MA, USA, 17–23 July 2021; pp. 8821–8831.
67. Qiao, T.; Zhang, J.; Xu, D.; Tao, D. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1505–1514.
68. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
69. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
70. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
71. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
72. Abady, L.; Dimitri, G.; Barni, M. Detection and Localization of GAN Manipulated Multi-spectral Satellite Images. In Proceedings of the ESANN 2022 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium and Online Event, 5–7 October 2022.

73. Yuan, K.; Zhuang, X.; Schaefer, G.; Feng, J.; Guan, L.; Fang, H. Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7422–7434. [CrossRef]
74. Abbessi, R.; Verrier, N.; Taddese, A.M.; Laroche, S.; Debailleul, M.; Lo, M.; Courbot, J.B.; Haeberlé, O. Multimodal image reconstruction from tomographic diffraction microscopy data. *J. Microsc.* 2022, *online ahead of print*. [CrossRef]
75. Filipović, M.; Barat, E.; Dautremer, T.; Comtat, C.; Stute, S. PET reconstruction of the posterior image probability, including multimodal images. *IEEE Trans. Med. Imaging* **2018**, *38*, 1643–1654. [CrossRef]