applied sciences

MDPI

Review

Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend

Wenhao Chai and Gaoang Wang *

Zhejiang University-University of Illinois at Urbana-Champaign Institute, Zhejiang University, Haining 314400, China; wenhaochai.19@intl.zju.edu.cn

* Correspondence: gaoangwang@intl.zju.edu.cn

Abstract: Deep vision multimodal learning aims at combining deep visual representation learning with other modalities, such as text, sound, and data collected from other sensors. With the fast development of deep learning, vision multimodal learning has gained much interest from the community. This paper reviews the types of architectures used in multimodal learning, including feature extraction, modality aggregation, and multimodal loss functions. Then, we discuss several learning paradigms such as supervised, semi-supervised, self-supervised, and transfer learning. We also introduce several practical challenges such as missing modalities and noisy modalities. Several applications and benchmarks on vision tasks are listed to help researchers gain a deeper understanding of progress in the field. Finally, we indicate that pretraining paradigm, unified multitask framework, missing and noisy modality, and multimodal task diversity could be the future trends and challenges in the deep vision multimodal learning field. Compared with existing surveys, this paper focuses on the most recent works and provides a thorough discussion of methodology, benchmarks, and future trends.

Keywords: multimodal learning; computer vision; deep learning; introductory; survey

Citation: Chai, W.; Wang, G. Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend. *Appl. Sci.* **2022**, *12*, 6588. https://doi.org/10.3390/app12136588

Academic Editor: Rubén Usamentiaga

Received: 5 June 2022
Accepted: 28 June 2022
Published: 29 June 2022

1. Introduction

Humans perceive objects in various ways, such as sight, hearing, and touch. "Modality" is defined as a signal that an object is represented in a certain way. Common unique modality information includes text, image, video, and sound. Modality information from multiple channels can be more widely collected and recorded in modern networked society. For example, recording a short video with subtitles and posting it conveys information that includes multiple modalities such as video, text, and sound. "Multimodality" refers to modal information for the same object. As humans can use multimodal information to help make judgments, machines can also perform representation learning and its downstream tasks through input with multimodal information.

Recent work has shown that deep learning is widely used in the field of computer vision, with great success in, for example, image classification, semantic segmentation, object detection, and other vision tasks. Meanwhile, large-scale pretrained models based on deep learning are well developed for natural language processing tasks. Some researchers have found that in some real-world tasks, better performance is achieved by using information from both visual and language modalities. Deep learning models that use multimodal information tend to be better than those that use a unimodality. Huang et al. [1] has demonstrated that multimodal learning yields more accurate estimates of latent space representations.

Several surveys have been published on the topic of multimodal learning [2]. The first to thoroughly propose this concept was a widely accepted taxonomy for multimodal machine learning by [3] as representation, translation, alignment, fusion, and co-learning. This taxonomy allows researchers to better grasp the state and future trends in multimodal

learning. Among them, multimodal representation learning is the focus of this paper. Representation learning for unimodality has been a long-standing topic in deep learning. Essentially, it is the process of converting input signals into features usable by the model for machine learning [4]. Zhang et al. [5] provide a comprehensive analysis of works on deep multimodal learning from three perspectives: learning multimodal representations, fusing multimodal signals at various levels, and multimodal applications. Guo et al. [6] highlight the critical issues of newly developed technologies, such as the encoder–decoder model, generative adversarial networks, and attention mechanism in a multimodal representation learning perspective. Mogadala et al. [7] discuss multimodal problem formulation, methods, existing datasets, and metrics, including the results obtained with corresponding state-of-the-art methods.

However, the existing review articles focus on the analysis and classification of abstract theories and are not comprehensive enough to discuss the applications and existing methods, especially the new implementations in recent years. Unlike previous reviews, this paper focuses on the visual–language multimodal learning techniques that have become popular in recent years, presenting recent progress in three main parts of learning architectures, learning paradigms, and multimodal data analysis. Meanwhile, we collect larger scale multimodal tasks and benchmarks, as well as state-of-the-art (SOTA) models and metrics corresponding to each dataset. To reflect currency, innovation, and diversity, the vast majority of methods mentioned in this article are from articles included in major deep learning conferences after 2019 and have a large impact. This review focuses on specific technologies rather than comprehensive taxonomy. We hope this article will be helpful to both novice and veteran researchers in the multimodal field. Generally, the major contributions of our work can be summarized as follows:

- In terms of **methodology**, we propose a rational classification, including architectures, paradigms, and issues. We pay particular attention to recent works in this field.
- In terms of **benchmarks**, we collect larger scale multimodal tasks and benchmarks, as well as state-of-the-art models and metrics corresponding to each dataset.
- In terms of **trends**, we outline the trends according to the works in recent years and discuss the challenges that deserve attention in the future.

The rest of the paper is organized as follows shown in Figure 1. We first review different architectures used in multimodal learning in Section 2. The learning paradigms are described in Section 3. In Section 4, multimodal data analysis is provided and discussed. In Section 5, we introduce several vision-related applications in the multimodal learning field, followed by the multimodal benchmark in Section 6. In Section 7, we point out challenges and future trends in multimodal learning. Finally, we conclude our work in Section 8.

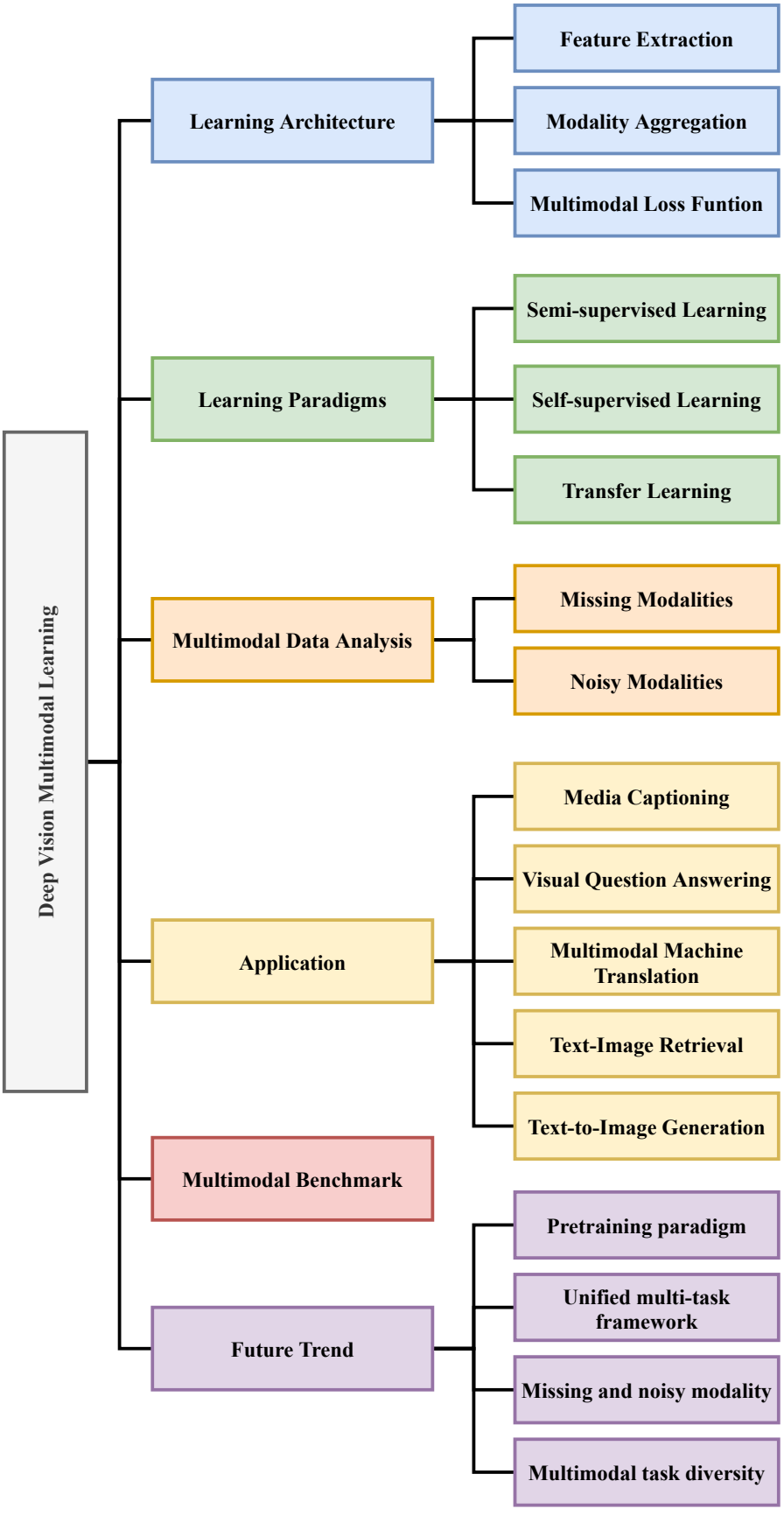

Figure 1. Outline of this survey.

2. Learning Architectures

The construction of a learning architecture and framework is the core technology of deep multimodal learning. This section discusses the design of feature extraction, modality aggregation, and multimodal loss function.

2.1. Feature Extraction

Feature extraction is the first step after the multimodal input signal enters the network. Essentially, the input signal is mapped to the corresponding feature space to form a feature vector. A well-performing representation tends to perform better on subsequent downstream subtasks. Zhang et al. [8] show that even only improving visual features significantly improves the performance across all vision–language tasks. For feature extraction, this paper mainly collects and discusses the transform-based network architecture.

Transformers [9] are a family of sequence-to-sequence deep neural networks. While they were originally designed for natural language processing (NLP), they have recently been widely used on modalities such as images [10], video [11], and audio [12]. Transformers use the self-attention mechanism to embed the input signal, then map it into the feature space for representation. An attention mechanism can be described as mapping a query and a set of key–value pairs to an output, where the query, key, value, and output are all in the form of vectors. The output is computed as a weighted sum of values, each weighted by the query's similarity function with the corresponding key. This approach [9] effectively handles the task of feature extraction from long sequences. As it has a larger receptive field, it can better capture the auto-correlation feature information in the input sequence. Transformer-based models have the following two good properties: unity and translation ability. Unity refers to the architecture in which feature extraction for different modalities can all use the transformer-based architecture of the network. Translation ability means that the feature hierarchy between different modalities can be easily aligned or transferred due to the same feature extractor architecture. These two properties are discussed in detail in the remainder of this section.

For multimodal feature extraction, transformers have demonstrated the ability of unity. In a typical multimodal learning network, text features are extracted using BERT [13], and image features are extracted using ResNet [14]. This approach results in the granularity on both sides not being aligned, as the text side is the token word and the image side is the global feature. A better modeling method should also convert the local features of the image into "visual words" so that the text words can be aligned with the "visual words". After the transformer has achieved state-of-the-art achievements on each unimodality task, the task of using the transformer for multimodal learning is taken for granted. Furthermore, because the individual feature extraction of each modality can be realized with it, a unified framework is formed. Therefore, the interaction and fusion of various modality information also become more accessible. Singh et al. [15] use a single holistic universal model for all language and visual alignment tasks. They use an image encoder transformer adopted from ViT [10] and a text encoder transformer to process unimodal information and a multimodal encoder transformer that takes as input the encoded unimodal image and text and integrates their representations for multimodal reasoning. The decoder is applied to the output of the appropriate encoder for different downstream tasks to obtain the result.

Transformers also have demonstrated the ability of translation ability. Likhosherstov et al. [16] propose co-training transformers. By training multiple tasks on a single modality and co-training on multimodalities, the model's performance is greatly improved. Akbari et al. [17] present a framework for learning multimodal representations from unlabeled data using convolution-free transformer architectures. These works show the important role of transformers in multimodal feature extraction and they will certainly become a major trend in the future.

However, due to the excessive memory and a large number of parameter requirements from transformers, existing work typically fixes or fine-tunes the language model and trains only the vision module, limiting its ability to learn cross-modal information in an

end-to-end performance. Lee et al. [18] decompose the transformers into modality-specific and modality-shared parts so that the model learns the dynamics of each modality both individually and together, which dramatically reduces the number of parameters in the model, making it easier to train and more flexible.

Another mechanism that has been used for multimodal feature extraction is called a memory network [19,20]. It is often used in conjunction with transformers, which are models that focus on different aspects. Most neural network models cannot read and write long-term memory parts and cannot be tightly coupled with inference. The method based on a memory network stores the extracted features in external memory and designs a pairing and reading algorithm, which has a good effect on improving the modal information inference of sequences and is effective on video captioning [21], vision-and-language navigation [22,23], and visual-and-textual question answering [24] tasks.

In conclusion, recent works dig out the potential value of the transformer-based network as the multimodal feature extractor. The transformer-based network could be applied to vision, language, and audio modalities. Thus, the multimodal feature extractor can be composed of a unified architectural form, which is suitable for aligning the features and transferring the knowledge between modalities.

2.2. Modality Aggregation

After extracting multimodal features, it is important to aggregate them together. The more typical fusion methods [25,26] are divided into early fusion and late fusion. Early fusion refers to directly splicing the multimodal input signals and then sending them into a unified network structure for training. Late fusion [27,28] refers to the fusion of the features of each modal signal after feature extraction. Finally, the network structure at the end is designed according to the different downstream tasks. As early and late fusion inhibit intra- and intermodal interactions, current research focuses on intermediate fusion methods, which allow these fusion operations to be placed in multiple layers of deep learning models. Different modalities can be fused simultaneously into a shared presentation layer or executed incrementally using one or more modalities at a time. Depending on the specific task and network architecture, the multidodal fusion method can be very flexible. Karpathy et al. [29] uses a "slow-fusion" network where extracted video stream features are gradually fused across multiple multimodal fusion layers, which performs better in a large-scale video stream classification task. Neverova et al. [30] consider the difference in information level between modalities and fuse the information of different modalities step by step, which are visual input modalities, then motion input modals, then audio input modalities. Ma et al. [31] propose a method to automatically search for an optimal fusion strategy regarding input data. In terms of specific fusion architecture, existing fusion methods can be divided into matrix- or MLP-based, attention-based, and other specific methods.

Matrix-based fusion. Zedeh et al. [32] use "tensor fusion" to obtain richer feature fusion methods with three types of tensors: unimodal, bimodal, and trimodal. Liu et al. propose the low-rank multimodal fusion method, which improves efficiency based on "tensor fusion". Hou et al. [33] integrate multimodal features by considering higher-order matrices.

MLP-based fusion. Xu et al. [34] propose a framework consisting of three parts: a compositional semantics language model, a deep video model, and a joint embedding model. In the joint embedding model, they minimize the distance of the outputs of the deep video model and compositional language model in the joint space and update these two models jointly. Sahu et al. [35] first encode all modalities, then use decoding to restore features, and finally calculate the loss between features.

Attention-based fusion [36]. Zedeh et al.use "delta-memory attention" and "multiview gated memory" to simultaneously capture temporal and intermodal interactions for better multiview fusion. The purpose of using the memory network is to save the multimodal interaction information at the last moment. In order to capture the interactive information between multimodalities and unimodality, the multi-interactive attention

mechanism [37] is further used; that is, textual and visual will fuse information through attention when they cross the multiple layers. Nagran et al. [38] use a shared token between two transformers so that this token becomes the communication bottleneck of different modalities to save the cost of computational attention.

In addition to these specific fusion methods, there is some auxiliary work. Perez et al. [39] help find the best fusion architecture for the target task and dataset. However, these methods often suffer from the exponential increase in dimensions and the number of modalities. The low-rank multimodal fusion method [40] performs multimodal fusion using low-rank tensors to improve efficiency. Gat et al. [41] propose a novel regularization term based on the functional entropy. Intuitively, this term encourages balancing each modality's contribution to the classification result.

In conclusion, direct concating of modal inputs is abandoned, and more work is considered to fuse them gradually while extracting features or after that. The fusion methods should be chosen specifically due to the properties of modalities and tasks.

2.3. Multimodal Loss Function

A reasonable loss function design can effectively help the network to train, but there is currently no comprehensive summary of some special loss functions used in multimodal training. This section lists some loss functions specifically designed for multimodal learning, shown in Figure 2.

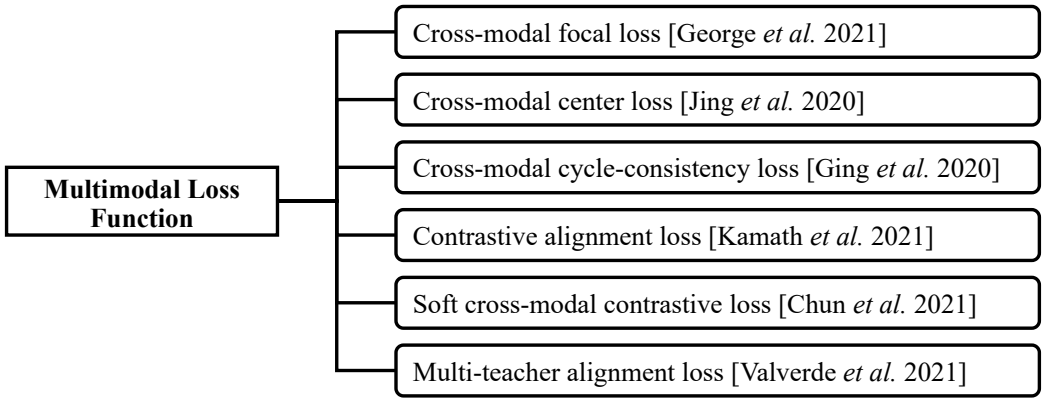

Figure 2. Multimodal loss function summary.

Cross-modal focal loss [42]. This is an application of focal loss [43] in the multimodal field. Its core idea is that when one of the channels can correctly classify a sample with high confidence, the loss contribution of the sample to the other branch can be reduced; if one channel can completely correctly classify a sample, then the other branch can no longer penalize the model. It shows that another modality mainly controls the loss value of the current modality.

Cross-modal center loss [44]. This is an application of center loss [45] in the multimodal field, which minimizes the distances of features from objects belonging to the same class across all modalities.

Cross-modal cycle-consistency loss [46]. This is used to enforce the semantic alignment between clips and sentences. It replaces the cross-modal attention units used in [47,48]. Its basic idea is to consider mapping the original modality to the target modality during cross-modal matching and find the matching target modality information with the highest similarity for specific original modality information. The matched target modal information is inversely mapped back to the original modal. Finally, the distance between the mapped value and the original value is calculated.

Contrastive alignment loss [49]. This loss function adopts InfoNCE [50] with reference to contrastive learning. It enhances the alignment of visual and textual embedded feature representations, ensuring that aligned visual feature representations and linguistic

feature representations are relatively close in feature space. This loss function does not act on the position but directly on the feature level to improve the similarity between the corresponding samples.

Soft cross-modal contrastive loss [51]. To capture pairwise similarities, HIB formulates a soft version of the contrastive loss [52] widely used for training deep metric embeddings. Soft cross-modal contrastive loss adopts the probabilistic embedding loss in previous contrastive loss, where the match probabilities are now based on the cross-modal pairs.

Multiteacher alignment loss [53]. This is a loss function used in multimodal knowledge distillation [54] that facilitates the distillation of information from multimodal teachers in a self-supervised manner. It measures the distance of the feature distribution of each modality after the feature extraction stage.

In conclusion, most of the loss functions for multimodal learning are extended by some previous unimodality loss functions, which are complemented according to the cyclic consistency or alignment between modalities.

3. Learning Paradigms

If learning architectures are likened to concrete equations and functions, learning paradigms are more like methodologies that guide problem solving. In addition to the most widely used supervised learning paradigm, other learning paradigms are also employed in the multimodal field, such as semi-supervised learning, self-supervised learning, and transfer learning. This section will discuss and summarize these three parts shown in Figure 3.

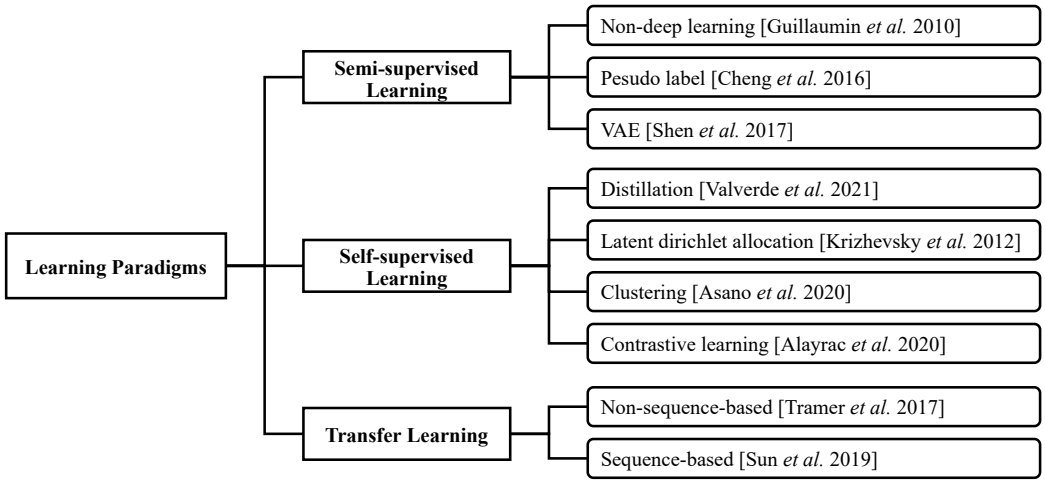

Figure 3. Multimodal learning paradigms.

3.1. Semi-Supervised Learning

Semi-supervised learning is a learning paradigm concerned with the study of how computers and natural systems such as humans learn in the presence of both labeled and unlabeled data [55]. It can use readily available unlabeled data to improve supervised learning tasks when the labeled data are scarce or expensive. The semi-supervised learning paradigm is important in multimodal learning because aligned and structured multimodal datasets are often expensive and difficult to obtain.

Guillaumin et al. [56] presents an early example of successful multimodal image classification using non-deep learning methods. Text assistance is used to assist unlabeled image classification. This suggests the potential for complementarity between modalities, which will be discussed later in this article. Cheng et al. [57,58] apply this learning paradigm to the RGB-D Object Recognition task. Their idea is to train an RGB-based and depth-based classifier separately on the labeled dataset and design a fusion module to obtain the result. For the unlabeled dataset, they first obtain the two prediction results of the RGB and depth streams, respectively, and exchange them as pseudo-labels of the other stream for training,

to achieve the purpose of semi-supervision. This method naively considers the possibility of cross-validation between modalities, but it does not necessarily work well for other multimodal forms such as text and image multimodality. In [59,60], some methods applied to vision–language mapping with the variational auto-encoding Bayes framework are extended to a semi-supervised model for an image–sentence mapping task.

3.2. Self-Supervised Learning

The self-supervised paradigm [61] can be viewed as a special form of unsupervised learning method with a supervised form, where supervision is induced by self-supervised tasks rather than preset prior knowledge. In contrast to a completely unsupervised setting, self-supervised learning uses information from the dataset itself to construct pseudo-labels. In terms of representation learning, self-supervised learning has great potential to replace fully supervised learning. For self-supervised signal representation within unimodality, Taleb et al. [62] cut an image into patches of uniform size and shuffle the order and trained the network to stitch the shuffled patches into the original image, which is similar to solving a puzzle problem. Training the network to solve the jigsaw puzzle allows the network to learn the deep features of the image in a self-supervised manner, thereby improving the performance of the network in downstream tasks such as segmentation and classification.

This learning paradigm is especially suitable for multimodal domains. This is because, in multimodal learning, not only a single modality itself will generate self-supervised signals, but also the alignment and constraints between modalities are also important sources of self-supervised signals. The rich self-supervised signals enable multimodal self-supervised learning. Tamkin et al. [63] introduce a domain-agnostic benchmark for self-supervised (DABS) multimodal learning on seven diverse domains: realistic images, multichannel sensor data, English text, speech recordings, multilingual text, chest X-rays, and images with text descriptions. It is an attempt to create the latest benchmark in the field. Valverde et al. [53] present a novel self-supervised framework consisting of multiple teachers that have diverse leverage modalities, including RGB, depth, and thermal images, to simultaneously exploit complementary cues and distill knowledge into a single audio student network. This work also proves that the single modality is a sufficiently robust one on some multimodal tasks with the training assistance of other modalities. Coen et al. [64] also train with signals transported across modalities. Gomez et al. [65] use textual information to train a CNN [66] to extract unlabeled image features. The motivation is that textual descriptions and annotations are easier to obtain than images. The first step is to learn image topics through latent Dirichlet allocation [67], and then perform parameter training of the image feature extraction network based on these topics.

In the video field, some work has also been developed in recent years. Afouras et al. [68] demonstrate that object embedding obtained from a self-supervised network facilitates a number of downstream audio-visual tasks that have previously required hand-engineered supervised pipelines. Asano et al. [69] propose a novel clustering method that allows pseudo-labeling of a video dataset without any human annotations by leveraging the natural correspondence between the audio and visual modalities. Specifically, it is to learn a clustering labeling function without access to any ground-truth label annotations. They think that each modality is equally informative in the algorithm to learn a more robust model. Alayrac et al. [70] use video and audio signals to extract features and design a contrastive loss, and then fuse the video and audio features with text features for a contrastive loss. The advantage of this method is that the parts with the same semantic level can be aligned when comparing modalities because the semantics of text are often more advanced in video and audio. Cheng et al. [71] separated the audio and video in video and determined whether they are from the same video to turn self-supervised learning into a binary classification problem. Alwassel et al. [72] conduct a comprehensive study of self-supervised clustering methods for video and audio modalities. They proposed four approaches, namely single-modality deep clustering (SDC), multihead deep clustering (MDC), concatenation deep clustering (CDC), and cross-modal deep clustering (XDC).

These approaches differentiate how intramodal and intermodal supervisory signals are utilized when the clustering algorithm iterates.

3.3. Transfer Learning

Transfer learning [73] is an indispensable part of today's deep learning field. The essence of transfer learning is to adjust the model parameters that have been trained on the source domain to the target domain. Since the dataset of downstream tasks is often relatively small in practical applications, training directly on it will lead to overfitting or difficulty in training. Taking natural language processing as an example, the way that has been developing this year is to train on large-scale datasets and then use pretrained models to transfer learning to downstream tasks. Such pretraining models often have a large amount of parameters, such as BERT [13], GPT [74], GPT-2 [75], and GPT-3 [76]. After the success of transfer learning in the field of natural language processing, various pretraining models have sprung up in the unimodality situation, such as ViT [10] in the field of computer vision and Wave2Vec [77] in the field of speech. There has been extensive work showing that they benefit downstream unimodal tasks in performance and efficiency.

This need is even more important in the multimodal field since multimodal aligned data are rare and expensive. A large number of the downstream tasks in the multimodal field rely on transfer learning, e.g., [49,53,78,79]. Hu et al. [80] share the same model parameters across all tasks instead of separately fine-tuning task-specific models and handle a much higher variety of tasks across different domains. This section describes the different methods of multimodal transfer learning. An important type of transfer learning is to unify vision and language features into a shared hidden space to generate a common representation for the source domain, then adjust the common representation to the target domain [47,81–83]. It can be subdivided into non-sequence-based, such as image–text, and sequence-based, such as video–text and video–audio.

Non-sequence-based. Rahman et al. [84] believe that non-text modalities (vision and audio) will affect the meaning of words and then affect the position of feature vectors in semantic space [85], so non-text and text jointly determine the new position of feature vectors in semantic space. It is a method of assisting transfer learning of text with information from other modalities. Gan et al. [86] propose a method to enhance the generalization ability of models using large-scale adversarial training [87], which consists of two steps of pretraining and transfer learning. It is a general framework that can be applied to any multimodal pretrained model to improve the model's generalization ability.

Sequence-based. Different from non-sequential tasks, sequential tasks represented by videos have more difficulties in transfer learning. Consecutive clips usually contain similar semantics from consecutive scenes, which means sparsely sampled clips already contain critical visual and semantic information in the video. Therefore, a small number of clips are sufficient to replace the entire video for training. Based on this, a large part of the work [88,89] is to take clips from the video for training randomly. Many approaches extract features from text input and clip input from sampled video separately and then aggregate them before the prediction layer. Lei et al. [90] propose to constrain each frame of video information with textual information such as "early fusion" and finally summarize the resulting predictions for each frame. Sun et al. [91] propose to convert video frames into discrete token sequences by applying hierarchical vector quantized features to generate a sequence of "visual words" that are aligned with the text. Furthermore, it is self-supervised by a masked language model method similar to BERT. This method of converting visual information into "visual words" is also reflected in [92], which is a good solution for aligning different modal representations in transfer learning.

In conclusion, multimodal learning allows the use of rich learning paradigms. In the absence of supervised signals, complementary and alignment information between modalities can be an alternative to self-supervision and semi-supervision. Multimodal transfer learning is also more diverse and generalizable.

4. Multimodal Data Analysis

Huang et al. [1] theoretically demonstrate that under specific effective learning methods, multimodal learning can perform better than unimodality. However, in practical situations, there are often many problems in acquiring multimodal information. In some scenarios, some modalities are missing, discontinuous, or unstructured. In the other cases, some modalities have low coincidence or contain much noise. These situations have significant adverse interference effects on the performance of deep data-based learning and are also the biggest problems that need to be overcome in the application of multimodal representation learning. This section lists two main challenges, namely missing and noisy modality and potential solutions.

4.1. Missing Modalities

The missing modality problem is defined as a situation where the data contain missing modalities during training and inference of multimodal learning. Approaches to solving this problem are mainly carried out from generative models and transfer learning. In addition, Ma et al. [93] use Bayesian meta-learning for the first time in the case of severely missing modality (about 90%).

The idea of the generative model is to use the existing modalities to generate the missing modalities for the completion of regular training. Huang et al. [94] propose that for incomplete multimodal data (such as images and text labels), during training, the features of the images are first extracted, and then they are used to generate the missing text label features. Training is carried out by using the generated text pseudo-labels as supervision. Given only an image, the model can generate the corresponding missing label features during the inference process and then obtain the full multimodal representation to perform some downstream tasks such as classification and retrieval.

Ding et al. [95,96] use the idea of transfer learning to solve the problem of missing modality. They first borrow an auxiliary database with complete modalities and then simultaneously consider knowledge transfer across databases and modalities within a unified framework. Ma et al. [31] show that the multimodal model can perform better than the unimodal in terms of any missing ratio by an optimal fusion strategy.

4.2. Noisy Modalities

There is much noise in modal information in the real world, and the information between different modalities is often not equal. For example, on multimodal sentiment analysis tasks [97], text annotations often have a high weight in the final judgment, but visual and auditory signals have low confidence and are full of noise. The mainstream method is to integrate and denoise information through multimodal co-learning, which aids the modeling of a resource-poor modality by exploiting knowledge from another resource-rich modality.

Pham et al. [98] iteratively translate and compute the loss between the two modalities while training the model, in the expectation of obtaining aligned modal representations. Moon et al. [99] collect noisy data from social media, which are composed of short captions with accompanying images. On this basis, they first extract the clear text part from the noisy caption modality as the information input of the third modality and then use the attention mechanism to fuse the information of all modalities. It is also a good idea to obtain the correspondence between modalities after denoising by clustering [69,100]. Lee et al. [101] explore the task of denoising noisy audio with the aid of a video signal. Their approach is to learn the affinity between two modalities and treat the unmatched parts as noise.

In conclusion, missing modalities and noise modalities are inevitable in real-world applications. Multimodal models have been shown to be able to outperform unimodal models in all cases theoretically and experimentally, while how to fully utilize multimodal information is also a potential topic.

5. Application

With the widespread dissemination and collection of multimodal data based on multimedia platforms such as TikTok and YouTube, multimodal representation learning also has more excellent application value in society. This section introduces the common vision-related applications of multimodal representation learning, including media captioning, visual question answering, multimodal machine translation, text–image retrieval, and text-to-image generation. They represent several major categories: vision to vision, text to text, test to vision, text–vision retrieval, and text to text under vision assistance.

5.1. Media Captioning

So-called media captioning is used to generate a corresponding caption or description for an image [102] or a video [103]. COCO [104] is one of the most comprehensive early datasets and is still widely used today. The Charades [105] dataset is collected from real and diverse examples of daily dynamic scenes. Charades-Ego [106] is a large-scale dataset of paired third- and first-person videos. Before the advent of transformers, such tasks were usually carried out by methods [107–109] based on the family of recurrent neural networks (RNNs) [110], long short-term memory (LSTM) [111], and convolutional neural networks (CNNs) [112,113]. In recent years, many methods have been proposed based on the attention mechanism [114]. There are also some works exploring the possibility of this task in weakly supervised or even unsupervised directions [115,116]. Wang et al. [117] propose a novel hierarchical reinforcement learning framework for video captioning, where a high-level manager module learns to design subgoals and a low-level worker module recognizes the primitive actions to fulfill the subgoal. Lu et al. [118] first generate the caption template and then fill it in according to the image content.

5.2. Visual Question Answering

This application was first proposed comprehensively in 2015. Given an image and a natural language question about the image, the task is to provide an accurate natural language answer [119–121]. It combines the two tasks of text-based Q&A [122,123] and describing visual content [124], and especially emphasizes the importance of vision in this task [125]. There are many datasets proposed for this task [119,126–133]. In 2016, Wu et al. [134] grouped the approaches for this application into four broad categories: joint embedding approaches [135–138], attention mechanisms [139–144], compositional models [24,145–148], and models using an external knowledge base [149,150]. A number of approaches have emerged in recent years, but most fall into these four categories.

5.3. Multimodal Machine Translation

Compared with machine translation, multimodal machine translation accepts information from multiple modalities to assist in completing translation tasks. A classic example is, given a sentence described in English, and a picture associated with it, expecting an output of a sentence in German [151,152]. There are many datasets proposed for this task [152–154]. The most likely application scenario for this task is caption translation in videos or images [155]. In recent years, many methods have been proposed based on the attention mechanism [156–160]. There are also some methods based on multiagent communication [161,162]. Considering that images and texts do not have good consistency in some scenarios, Elliott et al. [163] propose adversarial evaluation as a necessary metric. However, visual information may only be needed in particular cases. In [164–168], some methods are committed to solving this problem. Huang et al. [169] try to explore the possibility of unsupervised learning with shared visual features in different languages.

5.4. Text–Image Retrieval

Text–image retrieval is a task to explore the relationship between text and vision, mainly divided into two subtasks, text-based image retrieval and image-based text retrieval. This task was proposed and developed before deep learning was widely used [170]. Match-

ing between images and sentences is the key to text–image cross-modal retrieval. Some works jointly represent the information of vision and text modalities in a feature space and then calculate the similarity between them [107,171–174].

5.5. Text-to-Image Generation

Unlike generating text from images, there is no specific template for the task of generating images from text. Therefore, one of the most commonly used methods is generative adversarial networks (GANs) [175], and there are already some surveys [176–178] on this. Ref. [179] first introduced a GAN for this task. The following references list some variants of GANs, including CookGAN [180], ControlGAN [181], SD-GAN [182], DM-GAN [183], MirrorGAN [184], AttnGAN [185], StackGAN [186], and StackGAN++ [187].

In conclusion, the combination of modalities leads to diverse downstream tasks. However, multimodal tasks should be carefully evaluated to ensure that they have practical applications.

6. Multimodal Benchmark

This section lists widely used multimodal datasets, corresponding tasks, and state-of-the-art models, shown in Table 1. The state-of-the-art models use large-scale pretraining and a unified multitask framework to approach outstanding performance in many multimodal tasks. For generative tasks, we detail the metrics used for evaluation.

6.1. Image Captioning

The early image captioning tasks mainly used the Flickr30K [188] and Flickr8K datasets, and the images of these datasets came from the Flickr website. The most commonly used dataset is COCO Captions [189], which contains images of complex scenes between people, animals, and common daily objects. The annotations for image descriptions in the COCO Captions dataset are based on the entire image. Flickr30K Entities mark the nouns mentioned in the caption in Flickr30K, and mark the corresponding bbox. Localized Narratives [190] provides each word with a specific region in the image represented by its tracking segment, including nouns, verbs, adjectives, and prepositions. TextCaps [191] requires models to read and reason about text in images to generate descriptions about them. The images in VizWiz [192] are taken by visually impaired people using mobile phones, are of low quality and involve a wide variety of everyday activities, most of which require some text reading. The dataset aims to make more people aware of the needs of the blind, and to develop assistive technologies to solve the visual challenges in their daily life and to solve the vision problems of the blind. Nocaps [193] aims to evaluate whether a model can accurately describe objects of emerging classes in test images without corresponding training data.

As for the evaluation metrics of image captioning, BLEU [194], METEOR [195], CIDEr [196], and SPICE [197] are some of the most used.

BLEU is Bilingual Evaluation Understudy, which is calculated by modified n-gram precision. It is a precision-based metric but not recall, while short sentences always obtain a higher score. BLEU is calculated by the formula:

$$BLEU = BP \cdot \exp \sum_{n=1}^{N} w_n \log p_n \quad (1)$$

where n denotes n-gram, w_n denotes the weight and is usually set to $\frac{1}{n}$, BP denotes the brevity penalty, and p_n denotes the n-gram-level precision.

METHOR is Metric for Evaluation of Translation with Explicit ORdering. It claims to have a better correlation with human judgment when considering the order and matching of words. METHOD is a recall-based metric calculated by the formula:

$$METHOR = (1 - pen) \times F_{means} \quad (2)$$

where F_{means} denotes the weighted F-score, *pen* denotes the penalty factor to penalize the word order in the candidate translation that differs from the word order in the reference translation.

CIDEr is Consensus-based Image Description Evaluation. It calculates the cosine angle of its TF-IDF vector (note that each dimension of the vector represents an n-gram and not necessarily a word) to obtain the similarity between the candidate sentence and the reference sentence by the formula:

$$CIDEr_n(c, S) = \frac{1}{M}\sum_{i=1}^{M}\frac{g^n(c)\cdot g^n(S_i)}{||g^n(c)|| \times ||g^n(S_i)||} \tag{3}$$

where c denotes the candidate caption, S denotes the set of reference captions, n denotes the n-gram, M denotes the number of reference captions, and $g^n(\cdot)$ denotes the TF-IDF vector based on the n-gram.

SPICE is Semantic Propositional Image Caption Evaluation. It uses a graph-based semantic representation to encode objects, attributes, and relationships in a caption. It first parses the caption to be evaluated and the reference captions into syntactic dependency trees using the Probabilistic Context-Free Grammar (PCFG) dependency parser and then maps the dependency trees into scene graphs using a rule-based approach. Finally, the F-score values of objects, attributes, and relationships in the caption to be evaluated are calculated. SPICE is calculated by a series of formulas:

$$\begin{cases} SPICE(c, S) = \dfrac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)} \\ P(c, S) = \dfrac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \\ R(c, S) = \dfrac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \end{cases} \tag{4}$$

where c denotes a candidate caption, S denotes a set of reference captions, $G(\cdot)$ denotes the conversion of a text into a scene graph using some method, $T(\cdot)$ denotes the mapping of a scene graph into a set of tuples, $\otimes$ operates like an intersection, except that it is not a strict match, but is similar to METEOR in matching.

On this task, a multitask unified framework based on the transformer architecture [198,199] is a common feature of state-of-the-art models. These methods concatenate image and text features as input to the model to be pretrained and use self-attention to learn image–text semantic alignments. Li et al. [200] use the "object tags" to align the cross-domain semantics, which are the labels of objects in the image. Wang et al. [201] performs end-to-end pretraining using a unified target on a large number of weakly aligned image–text pairs, making it easy to apply transfer learning to downstream tasks.

6.2. Visual Question Answering

The most commonly used datases are VQA, VQA v2 [119,125], and its sub datasets, which are derived from realistic images. Artificially generated datasets such as CLEVR [127] have a limited variety of objects, very well-defined patterns of problems, and clean backgrounds. The VQA dataset contains open-ended questions about images. These questions require an understanding of vision, language, and commonsense knowledge. It states that a given answer is correct if it matches the more frequent answer or if it at least matches one of the possible ground-truth answers.

On this task, a multitask unified framework based on the transformer architecture is the trend of state-of-the-art models. Nam et al. [202] use the attention mechanism to model text and images separately, and use triplet loss to measure the similarity between text and images. Kazemi et al. [203] simply use a combination of multilayer CNNs and LSTMs to achieve outstanding performance as well.

6.3. Multimodal Machine Translation

The most common dataset used in multimodal machine translation tasks is Multi30k [152], which is a dataset to stimulate multilingual multimodal learning for English–German. The Multi30K dataset extends the Flickr30K dataset with translated or independent German sentences. Each image is paired with several English and German descriptions. Most current models resort to global context modeling, attention mechanism, or multimodal joint representation learning to utilize visual features. The global context modeling method uses an encoder to extract visual features as the extra input of the translation model [156]. The attention mechanism uses an attention-based weighted sum of the visual information and the source sentence embedding separately, and a gate matrix to fuse the information from these two modalities [158]. The multimodal joint representation learning method fuses the two pieces of modal information by decomposing the multimodal translation problem into two subtasks: learning translation and generating visual feature representations [204].

As the state-of-the-art model, Lin et al. [205] combine local and global visual information to learn multimodal contextual features.

6.4. Text-Image Retrieval

The most popular datasets for retrieval tasks are also COCO and Flickr30k. Both are hand-crafted datasets, with five short, descriptive, and conceptual captions for each image. For text–image retrieval, the most commonly used evaluation metric is R@K, which is the abbreviation for recall at K and is defined as the proportion of correct matchings in top-k retrieved results.

VisualSparta [206] is the first transformer-based text-to-image retrieval model that enables real-time search of very large datasets with significantly improved accuracy compared to previous state-of-the-art methods. By efficiently implementing invert indexing, VisualSparta achieves a speed advantage that is even greater on large-scale datasets.

6.5. Text-to-Image Generation

COCO is also widely used in text-to-image generation tasks. There are also several datasets containing only one broad category. Caltech-UCSD Birds (CUB) [207] has 11,788 images of 200 different categories of birds. Oxford-102 Flower [208] is a 102-category dataset consisting of 102 flower categories. The flowers are those commonly occurring in the United Kingdom. The images have large scale, pose and light variations. CelebAText-HQ [209] is a large-scale face image dataset with facial attributes, designed for text-to-face generation. In state-of-the-art models that generate images from text, large-scale pretraining is a key factor. GANs have shown extraordinary potential in text generation tasks.

As for the evaluation metrics of text-to-image generation, IS [210] and FID [211] are two of the most used.

IS is inception score. It considers both the quality and diversity of the generated images. To be specific, for a single image, the category probability distribution should be the focus, but for a set of images, the category probability distribution should be diverse. IS is calculated by the formula:

$$IS = \exp E_{x \sim p_G} KL(p(y|x)||p(y)). \tag{5}$$

where x denotes the generated image, y denotes the probability of different categories, G denotes the image generator, and KL denotes the KL-divergence.

FID is Fréchet inception distance score. Unlike IS, FID compares the generated image with the real image by computing a kind of similarity. FID is calculated by the formula:

$$FID(r, G) = ||\mu_r - \mu_g|| + T_r(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \tag{6}$$

where μ and Σ denote means and covariances, r and g denote the real images and generated images, Tr is the trace of matrix.

As the state-of-the-art model, Zhou et al. [212] propose the first work to leverage the well-aligned multimodal semantic space of the powerful pretrained CLIP model without any text data. Zhang et al. [187] consider this task as a two-stage problem including low-resolution image generation and image super-resolution. They argue that the model distribution generated from low-resolution images has a better probability of overlapping with the natural image distribution, but at the same time the need for high-resolution images is necessary.

In conclusion, existing multimodal datasets are still relatively limited due to the fact that labeled and aligned multimodal data are more difficult to obtain. We hope that more real-world-based multimodal tasks, evaluation metrics, and datasets are released.

Table 1. Multimodal benchmark.

Task and Dataset	**Model**	**Year**	**Metric**			
Image Captioning			**BLEU-4 [194]**	**METEOR [195]**	**CIDEr [196]**	**SPICE [197]**
COCO Captions [189]	OFA [199]	2022	43.5	31.9	149.6	26.1
	SimVL [201]	2021	40.6	33.4	143.3	25.4
	Oscar [200]	2020	41.7	30.6	140.0	24.5
COCO [104]	M2 Transformer [213]	2019	39.1	29.2	131.2	22.6
Flickr30k [188]	Unified VLP [198]	2019	30.1	23.0	67.4	17.0
npcaps [193]	LEMON-large [214]	2021	34.7	31.3	114.3	14.9
	SimVLM-huge [201]	2021	32.2	30.6	110.3	14.5
TextCaps [191]	LSTM-R [215]	2012	22.9	21.3	100.8	13.8
VizWiz [192]	CASIA-IVA	2020	28.3	22.1	79.1	17.9
Local Narratives [190]	LoopCAG [216]	2021	27.0	26.0	114.0	-
SciCap [217]	CNN+LSTM [217]	2021	21.9	-	-	-
Visual Question Answering			Overall	Y/N	Number	Other
VQA v1 [119]	SAAA [203]	2017	64.5	82.2	39.1	55.2
	DAN [202]	2016	64.3	83.0	39.1	53.9
VQA v2 [125]	VLMo [218]	2021	81.3	94.7	67.3	72.9
	OFA	2022	80.5	92.9	67.0	72.7
VQA-CP [140]	CSS [219]	2020	58.9	84.4	49.4	48.2
VQA-CE [220]	RandImg [220]	2021	63.3	-	-	-
COCO	MCB 7 att. [135]	2016	66.5	-	-	-
VCR [221]	VL-BERT-Large [79]	2019	75.5	-	-	-
GQA [131]	NSM [222]	2019	63.2	78.9	-	-
CLEVR [127]	NS-VQA [136]	2018	99.8	-	-	-
IconQA [223]	Patch-TRM [223]	2021	82.7	-	-	-
MSRVTT-QA [224]	Just Ask [225]	2020	41.5	-	-	-
Multimodal Machine Translation			BLEU	METEOR		
Multi30k [152]	DCCN [205]	2020	39.7	56.8		
	Caglayan [226]	2019	39.4	58.7		
	MM Transformer [160]	2020	38.7	55.7		
Text-Image Retrieval			Recall@1	Recall@5	Recall@10	
COCO Captions	VisualSparta [206]	2020	68.2	91.8	96.3	
COCO	Oscar	2020	78.2	95.8	98.3	
Flickr30k	VisualSparta	2020	57.4	82.0	88.1	
FooDI-ML [227]	ADAPT-12T [227]	2021	19.0	30.0	45.0	
WIT [228]	WIT-ALL [228]	2021	34.6	64.2	-	
Fashion IQ [229]	RTIC-GCN [230]	2021	-	-	40.6	
Text-to-Image Generation			FID [211]	IS [210]		
COCO	Lafite [212]	2021	8.1	32.3		
CUB [207]	Lafite	2021	10.5	6.0		
CelebA-HQ [209]	Lafite	2021	12.5	2.9		
Oxford Flower [208]	StackGAN++ [187]	2018	48.7	3.3		
Text Generation			BLEU-2	BLEU-3	BLEU-4	BLEU-5
COCO Captions	LeakGAN [231]	2017	0.95	0.88	0.78	0.69
	RankGAN [232]	2017	0.85	0.67	0.56	0.54

7. Future Trends

Long before deep learning, multimodal learning had already been developed to a certain extent. With the help of deep learning represented by neural networks, the future development prospects of deep multimodal learning will be broader.

7.1. Pretraining Paradigm

Most of the current deep learning methods rely on a large amount of labeled data. In order to obtain better performance, it is necessary to have more labeled data, which has become a significant bottleneck. In fact, by traditional methods, annotating large amounts of structured data and getting the training to converge to an optimal position is no less challenging than crafting a good embedding space. Collecting large amounts of aligned and labeled multimodal data has always been a significant challenge in solving multimodal learning problems such as fusion, translation, and co-learning. Unsupervised pretraining methods such as BERT [13] address this challenge well and can significantly improve overall system performance in some tasks such as multimodal representation [79], vision–text co-training [82], and cross-modal translation [233]. A multimodal model pretrained on a large-scale dataset can be easily transferred to a specific task, and such a process is one of the common means in the field of deep learning.

7.2. Unified Multitask Framework

In recent years, there has been a growing body of work on building a unified end-to-end framework for multimodal learning based on the idea of multitask learning. A typical example is [199], which accepts multimodal inputs such as images, videos, and text, and has several different modal task outputs (e.g., image generation, visual grounding, image captioning, image classification, text generation, etc.). Different combinations between modalities can produce multiple forms of tasks, and the endowed alignment links between modalities can in turn facilitate better representations. A unified multimodal framework requires strong arithmetic support but is bound to become one of the mainstream directions for multimodal learning.

7.3. Missing and Noisy Modality

As mentioned in the Multimodal Data Analysis section, there are often many data-based issues in a real-world application, which can be divided into two categories: missing and noisy modality. Due to the characteristics of multimodal learning itself, there is mutual constraint and supervision information between modalities. The multimodal method can be devised to perform well in some cases, even if some modalities are randomly missing or when there are some very noisy modalities. There is also a class of cases where the problem of unbalanced weights between modalities occurs. Peng et al. [234] point out that during multimodal training, one dominant modality can inhibit the training of another modality, resulting in the failure to realize the full potential of multimodality, which is also a matter of concern.

7.4. Multimodal Task Diversity

The mutual arrangement and combination of modalities can define many tasks in multimodal learning. Therefore, some recent works try to expand outward and are no longer satisfied with only scoring on the conventional caption tasks but instead explore new tasks with specific practical value. These new tasks focus on exploring deep multimodal learning for practical applications. The commonsense captioning task [235] aims to generate captions and perform commonsense reasoning at the same time given an input video. The multiperspective captioning task [236] considers text and image inputs from different viewpoints. The distinctive image captioning task [237] is defined to describe the unique object or context of an image to distinguish it from other semantically similar images. The diverse caption and rich image generation task [238] proposes a bidirectional image and text generation task to align rich images and their corresponding multiple different titles,

aiming to achieve multiple sentences from one image uniformly and multiple sentences to generate more suitable images.

The above four aspects of pretraining strategy, unified framework, missing and noisy modality, and multimodal task diversity are considered by us as feasible frontier directions for multimodal learning. We hope that researchers will find inspiration from them and make new contributions to the field of multimodal learning.

8. Conclusions

This paper discusses the methodology, benchmark, and trend of deep vision multimodal learning, which is supported by references to some of the more influential papers of recent years.

The methodology is divided into learning architecture, learning paradigms, and multimodal data analysis. In terms of learning architectures, we looked at different steps of a typical deep learning process: feature extraction, modality aggregation, and multimodal loss function. We are surprised that the transformer-based models have played a significant role in multimodal learning in recent years. There is endless potential to exploit the unity and flexibility of the attention mechanism across multiple modalities. As for loss function, most of the loss functions for multimodal learning are extended by some previous unimodality loss functions, which are complemented according to the cyclic consistency or alignment between modalities.

We considered some common learning paradigms in deep multimodal learning, including semi-supervised learning, self-supervised learning, and transfer learning. We find that the mutual complementation and alignment of modalities can serve as good alternatives when fully supervised signals are not available. Multimodal transfer learning is more robust than single modality. We also found that knowledge transfer between modalities can be achieved utilizing knowledge distillation, etc., which is an important research direction.

In terms of multimodal data analysis, we focus on two kinds of data issues that multimodal learning may have in a practical situation: missing modalities and noisy modalities. These two cases reflect the imbalance between the modes in different aspects. Not much work has been done in this direction, and it is a development focus on how to utilize strong and weak modalities in multimodal learning reasonably.

For multimodal learning applications and benchmarks, we collect mainstream tasks in the vision–language multimodal learning field, including media captioning, visual question answers, multimodal machine translation, text–image retrieval, and text-to-image generation. These tasks are made up of combinations of modalities and are of practical value. For each task, we list the widely used datasets and some state-of-the-art models on them and analyze the reasons for the success of these models. We hope that more meaningful tasks and datasets will be published.

After our thorough research, we have proposed some possible future research directions in the deep vision multimodal learning field: pretraining strategy, unified framework, missing and noisy modality, and multimodal task diversity. We believe that a unified large-scale multitask pretraining framework will become mainstream, as already demonstrated in natural language processing. Additionally, as we mentioned before, the task of making fuller use of unbalanced modal information and more diversity should be given full attention.

In conclusion, this review examines some of the influential work in deep vision multimodal learning in recent years and explores future trends. We hope this review will be helpful for beginners or researchers entering the field. Multimodal learning will be a significant trend of deep learning in the future, and more interested people are needed to get involved.

Author Contributions: Conceptualization, W.C. and G.W.; Writing—original draft, W.C.; Writing—review & editing, W.C.; Supervision, G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; Huang, L. What Makes Multi-modal Learning Better than Single (Provably). *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10944–10956.
2. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, Washington, DC, USA, 28 June 28–2 July 2011.
3. 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] [PubMed]
4. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
5. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [CrossRef]
6. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
7. Mogadala, A.; Kalimuthu, M.; Klakow, D. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Intell. Res.* **2021**, *71*, 1183–1317. [CrossRef]
8. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5579–5588.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
11. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
12. Gong, Y.; Chung, Y.A.; Glass, J. Ast: Audio spectrogram transformer. *arXiv* **2021**, arXiv:2104.01778.
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
15. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. FLAVA: A Foundational Language And Vision Alignment Model. *arXiv* **2021**, arXiv:2112.04482.
16. Likhosherstov, V.; Arnab, A.; Choromanski, K.; Lucic, M.; Tay, Y.; Weller, A.; Dehghani, M. PolyViT: Co-training Vision Transformers on Images, Videos and Audio. *arXiv* **2021**, arXiv:2111.12993.
17. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1–16.
18. Lee, S.; Yu, Y.; Kim, G.; Breuel, T.; Kautz, J.; Song, Y. Parameter efficient multimodal transformers for video representation learning. *arXiv* **2020**, arXiv:2012.04124.
19. Jason, W.; Sumit, C.; Antoine, B. Memory Networks. *arXiv* **2014**, arXiv:1410.3916.
20. ukhbaatar, S.; Szlam, A.; Weston, J.; Fergus, R. End-to-end memory networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–26.
21. Wang, J.; Wang, W.; Huang, Y.; Wang, L.; Tan, T. M3: Multimodal memory modelling for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7512–7520.
22. Lin, C.; Jiang, Y.; Cai, J.; Qu, L.; Haffari, G.; Yuan, Z. Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation. *arXiv* **2021**, arXiv:2111.05759.
23. Chen, S.; Guhur, P.L.; Schmid, C.; Laptev, I. History aware multimodal transformer for vision-and-language navigation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1–14.
24. Xiong, C.; Merity, S.; Socher, R. Dynamic memory networks for visual and textual question answering. In Proceedings of the International Conference on Machine Learning, New York, NY USA, 19–24 June 2016; pp. 2397–2406.
25. Boulahia, S.Y.; Amamra, A.; Madi, M.R.; Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vis. Appl.* **2021**, *32*, 1–18.
26. Khaleghi, B.; Khamis, A.; Karray, F.O.; Razavi, S.N. Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **2013**, *14*, 28–44. [CrossRef]
27. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1583–1597. [CrossRef] [PubMed]

28. Kahou, S.E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R.C.; et al. Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, Australia, 9–13 December 2013; pp. 543–550.
29. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
30. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1692–1706. [CrossRef]
31. Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; Peng, X. Are Multimodal Transformers Robust to Missing Modality? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Waikoloa, HI, USA, 3–8 January 2022; pp. 18177–18186.
32. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
33. Hou, M.; Tang, J.; Zhang, J.; Kong, W.; Zhao, Q. Deep multimodal multilinear fusion with high-order polynomial pooling. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–10.
34. Xu, R.; Xiong, C.; Chen, W.; Corso, J. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 19–25 January 2015; Volume 29.
35. Sahu, G.; Vechtomova, O. Dynamic fusion for multimodal data. *arXiv* **2019**, arXiv:1911.03821.
36. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018; Volume 32.
37. Xu, N.; Mao, W.; Chen, G. Multi-interactive memory network for aspect based multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019; Volume 33, pp. 371–378.
38. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1–14.
39. Pérez-Rúa, J.M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; Jurie, F. Mfas: Multimodal fusion architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6966–6975.
40. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
41. Gat, I.; Schwartz, I.; Schwing, A.; Hazan, T. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3197–3208.
42. George, A.; Marcel, S. Cross modal focal loss for rgbd face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7882–7891.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. Jing, L.; Vahdani, E.; Tan, J.; Tian, Y. Cross-modal center loss. *arXiv* **2020**, arXiv:2008.03561.
45. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
46. Ging, S.; Zolfaghari, M.; Pirsiavash, H.; Brox, T. Coot: Cooperative hierarchical transformer for video-text representation learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22605–22618.
47. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
48. Zhu, L.; Yang, Y. Actbert: Learning global-local video-text representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8746–8755.
49. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. MDETR-modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1780–1790.
50. Van den Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
51. Chun, S.; Oh, S.J.; De Rezende, R.S.; Kalantidis, Y.; Larlus, D. Probabilistic embeddings for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8415–8424.
52. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
53. Valverde, F.R.; Hurtado, J.V.; Valada, A. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11612–11621.
54. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
55. Zhu, X.; Goldberg, A.B. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2009**, *3*, 1–130.
56. Guillaumin, M.; Verbeek, J.; Schmid, C. Multimodal semi-supervised learning for image classification. In Proceedings of the 2010 IEEE Computer society conference on computer vision and pattern recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 902–909.

57. Cheng, Y.; Zhao, X.; Cai, R.; Li, Z.; Huang, K.; Rui, Y. Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 3345–3351.
58. Cheng, Y.; Zhao, X.; Huang, K.; Tan, T. Semi-supervised learning for rgb-d object recognition. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 2377–2382.
59. Tian, D.; Gong, M.; Zhou, D.; Shi, J.; Lei, Y. Semi-supervised multimodal hashing. *arXiv* **2017**, arXiv:1712.03404.
60. Shen, Y.; Zhang, L.; Shao, L. Semi-supervised vision-language mapping via variational learning. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1349–1354.
61. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* 2021, *Early Access*. [CrossRef]
62. Taleb, A.; Lippert, C.; Klein, T.; Nabi, M. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 661–673.
63. Tamkin, A.; Liu, V.; Lu, R.; Fein, D.; Schultz, C.; Goodman, N. DABS: A Domain-Agnostic Benchmark for Self-Supervised Learning. *arXiv* **2021**, arXiv:2111.12062.
64. Coen, M.H. Multimodal Dynamics: Self-Supervised Learning in Perceptual and Motor Systems. Ph.D. Thesis, Massachusetts Institute of Technology, Boston, MA, USA, 2006.
65. Gomez, L.; Patel, Y.; Rusinol, M.; Karatzas, D.; Jawahar, C. Self-supervised learning of visual features through embedding images into text topic spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4230–4239.
66. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
67. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
68. Afouras, T.; Owens, A.; Chung, J.S.; Zisserman, A. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 208–224.
69. Asano, Y.; Patrick, M.; Rupprecht, C.; Vedaldi, A. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4660–4671.
70. Alayrac, J.B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; Zisserman, A. Self-supervised multimodal versatile networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 25–37.
71. Cheng, Y.; Wang, R.; Pan, Z.; Feng, R.; Zhang, Y. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3884–3892.
72. Alwassel, H.; Mahajan, D.; Korbar, B.; Torresani, L.; Ghanem, B.; Tran, D. Self-supervised learning by cross-modal audio-video clustering. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9758–9770.
73. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
74. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://openai.com/blog/language-unsupervised/ (accessed on 1 June 2022).
75. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
76. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
77. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
78. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv* **2021**, arXiv:2102.04830.
79. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
80. Hu, R.; Singh, A. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1439–1449.
81. Chen, F.; Zhang, D.; Han, M.; Chen, X.; Shi, J.; Xu, S.; Xu, B. VLP: A Survey on Vision-Language Pre-training. *arXiv* **2022**, arXiv:2202.09061.
82. Li, G.; Duan, N.; Fang, Y.; Gong, M.; Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 11336–11344.
83. Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; Liu, J. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4155–4165.
84. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. *Integr. Multimodal Inf. Large Pretrained Transform.* **2020**, *2020*, 2359.

85. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019; Volume 33, pp. 7216–7223.
86. Gan, Z.; Chen, Y.C.; Li, L.; Zhu, C.; Cheng, Y.; Liu, J. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6616–6628.
87. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv* **2017**, arXiv:1705.07204.
88. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV), München, Germany, 8–14 September 2018; pp. 305–321.
89. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6202–6211.
90. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7331–7341.
91. Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; Schmid, C. Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7464–7473.
92. Tan, H.; Bansal, M. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv* **2020**, arXiv:2010.06775.
93. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. Smil: Multimodal learning with severely missing modality. *arXiv* **2021**, arXiv:2103.05677.
94. Huang, Y.; Wang, W.; Wang, L. Unconstrained multimodal multi-label learning. *IEEE Trans. Multimed.* **2015**, *17*, 1923–1935. [CrossRef]
95. Ding, Z.; Ming, S.; Fu, Y. Latent low-rank transfer subspace learning for missing modality recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27 July 2014; Volume 28.
96. Ding, Z.; Shao, M.; Fu, Y. Missing modality transfer learning via latent low-rank constraint. *IEEE Trans. Image Process.* **2015**, *24*, 4322–4334. [CrossRef]
97. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
98. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019; Volume 33, pp. 6892–6899.
99. Moon, S.; Neves, L.; Carvalho, V. Multimodal named entity disambiguation for noisy social media posts. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2000–2008.
100. Gupta, T.; Schwing, A.; Hoiem, D. Vico: Word embeddings from visual co-occurrences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7425–7434.
101. Lee, J.; Chung, S.W.; Kim, S.; Kang, H.G.; Sohn, K. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 1336–1345.
102. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
103. Rohrbach, A.; Rohrbach, M.; Tandon, N.; Schiele, B. A dataset for movie description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3202–3212.
104. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
105. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 510–526.
106. Sigurdsson, G.A.; Gupta, A.; Schmid, C.; Farhadi, A.; Alahari, K. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv* **2018**, arXiv:1804.09626.
107. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
108. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef]
109. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.W.; Salakhutdinov, R.R. Review networks for caption generation. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2369–2377.
110. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
111. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

112. Johnson, J.; Karpathy, A.; Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4565–4574.
113. Xu, H.; Li, B.; Ramanishka, V.; Sigal, L.; Saenko, K. Joint event detection and description in continuous video streams. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 396–405.
114. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
115. Laina, I.; Rupprecht, C.; Navab, N. Towards unsupervised image captioning with shared multimodal embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7414–7424.
116. Rohrbach, A.; Rohrbach, M.; Tang, S.; Joon Oh, S.; Schiele, B. Generating descriptions with grounded and co-referenced people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4979–4989.
117. Wang, X.; Chen, W.; Wu, J.; Wang, Y.F.; Wang, W.Y. Video captioning via hierarchical reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4213–4222.
118. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural baby talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7219–7228.
119. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
120. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5014–5022.
121. Yuan, X.; Côté, M.A.; Fu, J.; Lin, Z.; Pal, C.; Bengio, Y.; Trischler, A. Interactive language learning by question answering. *arXiv* **2019**, arXiv:1908.10909.
122. Fader, A.; Zettlemoyer, L.; Etzioni, O. Paraphrase-driven learning for open question answering. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 1608–1618.
123. Weston, J.; Bordes, A.; Chopra, S.; Rush, A.M.; Van Merriënboer, B.; Joulin, A.; Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv* **2015**, arXiv:1502.05698.
124. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef]
125. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
126. Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; Fidler, S. Movieqa: Understanding stories in movies through question-answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4631–4640.
127. Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2901–2910.
128. Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4999–5007.
129. Yagcioglu, S.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv* **2018**, arXiv:1809.00812.
130. Zadeh, A.; Chan, M.; Liang, P.P.; Tong, E.; Morency, L.P. Social-iq: A question answering benchmark for artificial social intelligence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8807–8817.
131. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6700–6709.
132. Talmor, A.; Yoran, O.; Catav, A.; Lahav, D.; Wang, Y.; Asai, A.; Ilharco, G.; Hajishirzi, H.; Berant, J. Multimodalqa: Complex question answering over text, tables and images. *arXiv* **2021**, arXiv:2104.06039.
133. Xu, L.; Huang, H.; Liu, J. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9878–9888.
134. Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; van den Hengel, A. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.* **2017**, *163*, 21–40. [CrossRef]

135. 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.
136. Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; Tenenbaum, J. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–12.
137. Vedantam, R.; Desai, K.; Lee, S.; Rohrbach, M.; Batra, D.; Parikh, D. Probabilistic neural symbolic models for interpretable visual question answering. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6428–6437.
138. Cadene, R.; Dancette, C.; Cord, M.; Parikh, D. Rubi: Reducing unimodal biases for visual question answering. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
139. Fan, H.; Zhou, J. Stacked latent attention for multimodal reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1072–1080.
140. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4971–4980.
141. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
142. Zhang, Y.; Hare, J.; Prügel-Bennett, A. Learning to count objects in natural images for visual question answering. *arXiv* **2018**, arXiv:1802.05766.
143. Alberti, C.; Ling, J.; Collins, M.; Reitter, D. Fusion of detected objects in text for visual question answering. *arXiv* **2019**, arXiv:1908.05054.
144. Hu, R.; Singh, A.; Darrell, T.; Rohrbach, M. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9992–10002.
145. Andreas, J.; Rohrbach, M.; Darrell, T.; Klein, D. Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 39–48.
146. Hu, R.; Andreas, J.; Rohrbach, M.; Darrell, T.; Saenko, K. Learning to reason: End-to-end module networks for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 804–813.
147. Lei, J.; Yu, L.; Bansal, M.; Berg, T.L. Tvqa: Localized, compositional video question answering. *arXiv* **2018**, arXiv:1809.01696.
148. Cadene, R.; Ben-Younes, H.; Cord, M.; Thome, N. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1989–1998.
149. Wu, Q.; Wang, P.; Shen, C.; Dick, A.; Van Den Hengel, A. Ask me anything: Free-form visual question answering based on knowledge from external sources. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4622–4630.
150. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
151. Caglayan, O.; Aransa, W.; Wang, Y.; Masana, M.; García-Martínez, M.; Bougares, F.; Barrault, L.; Van de Weijer, J. Does multimodality help human and machine for translation and image captioning? *arXiv* **2016**, arXiv:1605.09186.
152. Elliott, D.; Frank, S.; Sima'an, K.; Specia, L. Multi30k: Multilingual english-german image descriptions. *arXiv* **2016**, arXiv:1605.00459.
153. Hewitt, J.; Ippolito, D.; Callahan, B.; Kriz, R.; Wijaya, D.T.; Callison-Burch, C. Learning translations via images with a massively multilingual image dataset. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2566–2576.
154. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4581–4591.
155. Hitschler, J.; Schamoni, S.; Riezler, S. Multimodal pivots for image caption translation. *arXiv* **2016**, arXiv:1601.03916.
156. Calixto, I.; Liu, Q.; Campbell, N. Incorporating global visual features into attention-based neural machine translation. *arXiv* **2017**, arXiv:1701.06521.
157. Delbrouck, J.B.; Dupont, S. An empirical study on the effectiveness of images in multimodal neural machine translation. *arXiv* **2017**, arXiv:1707.00995.
158. Calixto, I.; Liu, Q.; Campbell, N. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv* **2017**, arXiv:1702.01287.
159. Zhou, M.; Cheng, R.; Lee, Y.J.; Yu, Z. A visual attention grounding neural model for multimodal machine translation. *arXiv* **2018**, arXiv:1808.08266.

160. Yao, S.; Wan, X. Multimodal transformer for multimodal machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, DC, USA, 5–10 July 2020; pp. 4346–4350.
161. Lee, J.; Cho, K.; Weston, J.; Kiela, D. Emergent translation in multi-agent communication. *arXiv* **2017**, arXiv:1710.06922.
162. Chen, Y.; Liu, Y.; Li, V. Zero-resource neural machine translation with multi-agent communication game. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018; Volume 32.
163. Elliott, D. Adversarial evaluation of multimodal machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2018; pp. 2974–2978.
164. Caglayan, O.; Madhyastha, P.; Specia, L.; Barrault, L. Probing the need for visual context in multimodal machine translation. *arXiv* **2019**, arXiv:1903.08678.
165. Ive, J.; Madhyastha, P.; Specia, L. Distilling translations with visual awareness. *arXiv* **2019**, arXiv:1906.07701.
166. Yang, P.; Chen, B.; Zhang, P.; Sun, X. Visual agreement regularized training for multi-modal machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 9418–9425.
167. Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; Zhao, H. Neural machine translation with universal visual representation. In Proceedings of the International Conference on Learning Representations, Formerly Addis Ababa, Ethiopia, Virtual, 6–9 May 2019.
168. Calixto, I.; Rios, M.; Aziz, W. Latent variable model for multi-modal translation. *arXiv* **2018**, arXiv:1811.00357.
169. Huang, P.Y.; Hu, J.; Chang, X.; Hauptmann, A. Unsupervised multimodal neural machine translation with pseudo visual pivoting. *arXiv* **2020**, arXiv:2005.03119.
170. Rui, Y.; Huang, T.S.; Ortega, M.; Mehrotra, S. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **1998**, *8*, 644–655.
171. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
172. Vendrov, I.; Kiros, R.; Fidler, S.; Urtasun, R. Order-embeddings of images and language. *arXiv* **2015**, arXiv:1511.06361.
173. Wang, L.; Li, Y.; Lazebnik, S. Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5005–5013.
174. Klein, B.; Lev, G.; Sadeh, G.; Wolf, L. Associating neural word embeddings with deep image representations using fisher vectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4437–4446.
175. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 11.
176. Huang, H.; Yu, P.S.; Wang, C. An introduction to image synthesis with generative adversarial nets. *arXiv* **2018**, arXiv:1803.04469.
177. Agnese, J.; Herrera, J.; Tao, H.; Zhu, X. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1345. [CrossRef]
178. Frolov, S.; Hinz, T.; Raue, F.; Hees, J.; Dengel, A. Adversarial text-to-image synthesis: A review. *Neural Netw.* **2021**, *144*, 187–209. [CrossRef]
179. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1060–1069.
180. Zhu, B.; Ngo, C.W. CookGAN: Causality based text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5519–5527.
181. Li, B.; Qi, X.; Lukasiewicz, T.; Torr, P. Controllable text-to-image generation. *arXiv* **2019**, arXiv:1909.07083.
182. Yin, G.; Liu, B.; Sheng, L.; Yu, N.; Wang, X.; Shao, J. Semantics disentangling for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2327–2336.
183. Zhu, M.; Pan, P.; Chen, W.; Yang, Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5802–5810.
184. 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.
185. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1316–1324.
186. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
187. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [CrossRef]
188. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]

189. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
190. Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; Ferrari, V. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 647–664.
191. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. Textcaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 742–758.
192. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3608–3617.
193. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019–2 November 2019; pp. 8948–8957.
194. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
195. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
196. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
197. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.
198. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 13041–13049.
199. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv* **2022**, arXiv:2202.03052.
200. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137.
201. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv* **2021**, arXiv:2108.10904.
202. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 299–307.
203. Kazemi, V.; Elqursh, A. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv* **2017**, arXiv:1704.03162.
204. Elliott, D.; Kádár, A. Imagination improves multimodal translation. *arXiv* **2017**, arXiv:1705.04350.
205. Lin, H.; Meng, F.; Su, J.; Yin, Y.; Yang, Z.; Ge, Y.; Zhou, J.; Luo, J. Dynamic context-guided capsule network for multimodal machine translation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1320–1329.
206. Lu, X.; Zhao, T.; Lee, K. VisualSparta: An Embarrassingly Simple Approach to Large-scale Text-to-Image Search with Weighted Bag-of-words. *arXiv* **2021**, arXiv:2101.00265.
207. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-Ucsd Birds-200-2011 Dataset. 2011. Available online: https://authors.library.caltech.edu/27452/ (accessed on 1 June 2022).
208. Nilsback, M.E.; Zisserman, A. Automated flower classification over a large number of classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
209. Xia, W.; Yang, Y.; Xue, J.H.; Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2256–2265.
210. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–9.
211. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–12.
212. Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; Sun, T. LAFITE: Towards Language-Free Training for Text-to-Image Generation. *arXiv* **2021**, arXiv:2111.13792.
213. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
214. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up vision-language pre-training for image captioning. *arXiv* **2021**, arXiv:2111.12233.
215. Zhu, Q.; Gao, C.; Wang, P.; Wu, Q. Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv* **2020**, arXiv:2012.05153.

216. Yan, K.; Ji, L.; Luo, H.; Zhou, M.; Duan, N.; Ma, S. Control Image Captioning Spatially and Temporally. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 1–6 August 2021; pp. 2014–2025.
217. Hsu, T.Y.; Giles, C.L.; Huang, T.H. SciCap: Generating Captions for Scientific Figures. *arXiv* **2021**, arXiv:2110.11624.
218. Wang, W.; Bao, H.; Dong, L.; Wei, F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *arXiv* **2021**, arXiv:2111.02358.
219. Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; Zhuang, Y. Counterfactual samples synthesizing for robust visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10800–10809.
220. Dancette, C.; Cadene, R.; Teney, D.; Cord, M. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1574–1583.
221. Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6720–6731.
222. Hudson, D.; Manning, C.D. Learning by abstraction: The neural state machine. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–19.
223. Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; Zhu, S.C. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. *arXiv* **2021**, arXiv:2110.13214.
224. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5288–5296.
225. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1686–1697.
226. Sulubacak, U.; Caglayan, O.; Grönroos, S.A.; Rouhe, A.; Elliott, D.; Specia, L.; Tiedemann, J. Multimodal machine translation through visuals and speech. *Mach. Transl.* **2020**, *34*, 97–147. [CrossRef]
227. Olóndriz, D.A.; Puigdevall, P.P.; Palau, A.S. FooDI-ML: A large multi-language dataset of food, drinks and groceries images and descriptions. *arXiv* **2021**, arXiv:2110.02035.
228. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 2443–2449.
229. Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; Feris, R. Fashion iq: A new dataset towards retrieving images by natural language feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11307–11317.
230. Shin, M.; Cho, Y.; Ko, B.; Gu, G. RTIC: Residual Learning for Text and Image Composition using Graph Convolutional Network. *arXiv* **2021**, arXiv:2104.03015.
231. Guo, J.; Lu, S.; Cai, H.; Zhang, W.; Yu, Y.; Wang, J. Long text generation via adversarial training with leaked information. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018; Volume 32.
232. Lin, K.; Li, D.; He, X.; Zhang, Z.; Sun, M.T. Adversarial ranking for language generation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
233. Shi, B.; Ji, L.; Liang, Y.; Duan, N.; Chen, P.; Niu, Z.; Zhou, M. Dense procedure captioning in narrated instructional videos. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6382–6391.
234. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. Balanced Multimodal Learning via On-the-fly Gradient Modulation. *arXiv* **2022**, arXiv:2203.15332.
235. Yu, W.; Liang, J.; Ji, L.; Li, L.; Fang, Y.; Xiao, N.; Duan, N. Hybrid reasoning network for video-based commonsense captioning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 5213–5221.
236. Bin, Y.; Shang, X.; Peng, B.; Ding, Y.; Chua, T.S. Multi-Perspective Video Captioning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 5110–5118.
237. Wang, J.; Xu, W.; Wang, Q.; Chan, A.B. Group-based distinctive image captioning with memory attention. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 5020–5028.
238. Huang, Y.; Liu, B.; Fu, J.; Lu, Y. A Picture is Worth a Thousand Words: A Unified System for Diverse Captions and Rich Images Generation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 2792–2794.