Deep Multimodal Data Fusion

FEI ZHAO, The University of Alabama at Birmingham, Birmingham, AL, USA
CHENGCUI ZHANG, The University of Alabama at Birmingham, Birmingham, AL, USA
BAOCHENG GENG, The University of Alabama at Birmingham, Birmingham, AL, USA

Multimodal Artificial Intelligence (Multimodal AI), in general, involves various types of data (e.g., images, texts, or data collected from different sensors), feature engineering (e.g., extraction, combination/fusion), and decision-making (e.g., majority vote). As architectures become more and more sophisticated, multimodal neural networks can integrate feature extraction, feature fusion, and decision-making processes into one single model. The boundaries between those processes are increasingly blurred. The conventional multimodal data fusion taxonomy (e.g., early/late fusion), based on which the fusion occurs in, is no longer suitable for the modern deep learning era. Therefore, based on the main-stream techniques used, we propose a new fine-grained taxonomy grouping the state-of-the-art (SOTA) models into five classes: Encoder-Decoder methods, Attention Mechanism methods, Graph Neural Network methods, Generative Neural Network methods, and other Constraint-based methods. Most existing surveys on multimodal data fusion are only focused on one specific task with a combination of two specific modalities. Unlike those, this survey covers a broader combination of modalities, including Vision + Language (e.g., videos, texts), Vision + Sensors (e.g., images, LiDAR), and so on, and their corresponding tasks (e.g., video captioning, object detection). Moreover, a comparison among these methods is provided, as well as challenges and future directions in this area.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Computer vision**; **Machine learning;**

Additional Key Words and Phrases: Data fusion, neural networks, multimodal deep learning

ACM Reference Format:
Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multimodal Data Fusion. *ACM Comput. Surv.* 56, 9, Article 216 (April 2024), 36 pages. https://doi.org/10.1145/3649447

1 INTRODUCTION

Data, without a doubt, is an extremely important catalyst in technological development, especially in **Artificial Intelligence** (**AI**) field. In the last 20 years, the amount of data generated in this period accounts for about 90% of all data available in the world. Moreover, the rate of data growth is still accelerating. The explosion of data provides an unprecedented chance for AI to thrive.

With the advancement of sensor technologies, not only the amount and quality of data is increased and enhanced, but the diversity of data is also skyrocketing. The data captured from different sensors provide people with distinct "views" or "perspectives" of the same objects, activities, or

Authors' addresses: F. Zhao, The University of Alabama at Birmingham, University Hall 4105, 1402 10th Ave. S., Birmingham, AL, 35294, USA; e-mail: larry5@uab.edu; C. Zhang, The University of Alabama at Birmingham, University Hall 4143, 1402 10th Ave. S., Birmingham, AL, 35294, USA; e-mail: czhang02@uab.edu; B. Geng, The University of Alabama at Birmingham, University Hall 4147, 1402 10th Ave. S., Birmingham, AL, 35294, USA; e-mail: bgeng@uab.edu.

ACM 0360-0300/2024/04-ART216
https://doi.org/10.1145/3649447

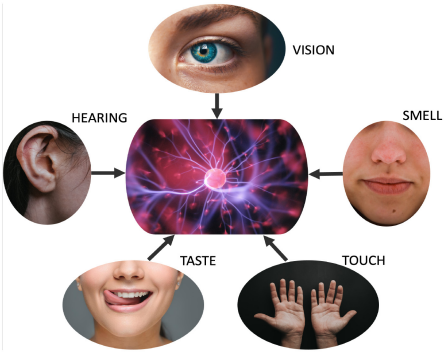


Fig. 1. The world has been projected into multiple dimensions/domains.

phenomena. In other words, people are able to observe the same objects, activities, or phenomena in different "dimensions" or "domains" by using different sensors. These new "views" help people obtain a better understanding of the world. For example, 100 years ago, in the medical field, it was extremely difficult for physicians to diagnose whether a patient has a lung tumor due to the limited way of observing organs. After the invention of the first **computerized tomography** (**CT**) scanner based on X-ray technology, the data captured from the machine provide much richer information about lungs, enabling physicians to make diagnoses based on CT images alone. With the advancement of technology, **magnetic resonance imaging** (**MRI**), a medical imaging technique that uses strong magnetic fields and radio waves, has been used to detect tumors as well. Nowadays, physicians are able to access multimodal data including CT, MRI, and blood test data, and so on. The accuracy of diagnosis based on the combination of these data is much higher, compared with that based on a single modality alone, e.g., CT, or MRI only. This is because the complementary and redundant information among CT, MRI, and blood test data can help physicians build a more comprehensive view of an observed object, activity, or phenomenon. Evolution of AI also follows a similar path. In its infancy, AI only focuses on solving problems using a single modality. Nowadays, AI tools have become increasingly capable of solving real-world problems by using multimodality.

What is multimodality? In reality, when we experience the world, we see objects, hear sounds, feel textures, smell odors, and taste flavors [11]. The world is represented by information in different mediums, e.g., vision, sounds, and textures. A visualization is shown in Figure 1. Our receptors such as eyes and ears, help us capture the information. Then, our brain will be able to fuse the information from different receptors to form a prediction or a decision. The information obtained from each source/medium can be viewed as one modality. When the number of modalities is greater than one, we call it multimodality. However, instead of using eyes and ears, machines highly depend on sensors such as RGB cameras, microphones, or other types of sensors, as shown in Figure 2. Each sensor can map the observed objects/activities into its own dimension. In other words, the observed objects/activities can be projected into the dimension of each sensor. Then, machines or robots can collect the data from each sensor and make a prediction or decision based on them. In the industry, there are numerous applications taking advantage of multimodality. For example, autonomous vehicle, which is one of the hottest topics since the 2020s, is a typical application relying on multimodality. Such a system requires multiple types of data from different sensors, e.g., LiDAR sensors, Radar sensors, cameras, and GPS. The model will fuse these data to make real-time predictions. In the medical field, more and more applications rely on the fusion of medical imaging and electronic health records to enable models to analyze imaging findings in the clinical context, e.g., CT and MRI fusion.

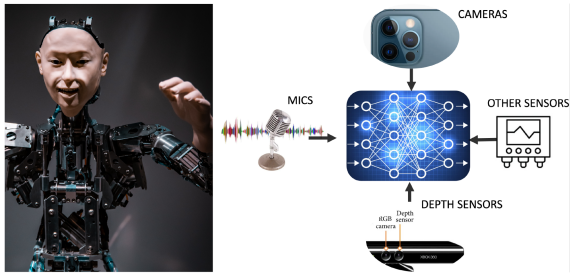


Fig. 2. The world has been projected into multiple dimensions/domains by different types of sensors.

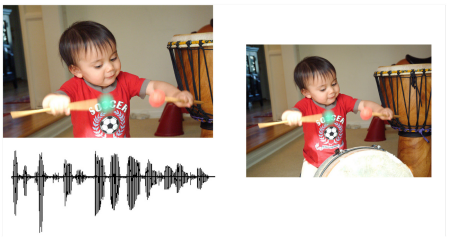


Fig. 3. The kid is striking a drum. Even if the drum is not visible, based on the vision and audio information, we can still recognize the activity correctly.

Why do we need multimodality? In general, multimodal data refer to the data collected from different sensors, e.g., CT images, MRI images, and blood test data for the cancer diagnosis, RGB data and LiDAR data for autonomous driving system, RGB data and infrared data for skeleton detection of Kinect [28]. For the same observed object or activity, the data from different modalities can have distinct expressions and perspectives. Although the characteristics of these data can be independent and distinct, they often overlap semantically. This phenomenon is called information redundancy. Furthermore, information from different modalities can be complementary. Humans can unconsciously fuse the multimodal data, obtain knowledge, and make predictions. The complementary and redundant information extracted from multimodalities can help humans form a comprehensive understanding of the world. As the example shown in Figure 3, when a kid is drumming, even if we cannot see the drum, we are still able to recognize a drum that is being struck based on the sounds. In this process, we unconsciously fuse the vision and acoustic data, and extract the complementary information of them, to make a correct prediction. If there is only one modality available, e.g., vision modality with the drum object out of sight, we can only tell that a kid is waving two sticks. With only the sound available, we would only be able to tell that a drum is being struck without knowing who is drumming. Therefore, in general, the independent interpretation based on individual modality only presents partial information of the observed activity. However, the multimodality-based interpretation can deliver the "fuller picture" of the observed activity, which can be more robust and reliable than single-modality-based models. For instance, autonomous vehicles containing multiple sensors such as RGB cameras and LiDAR sensors, need to detect objects on the road in extreme weather conditions where visibility is near zero, e.g., dense fog or heavy rain. A multimodal-based model can still detect objects while the pure-vision-based models cannot. However, it is extremely hard for machines to understand and figure out how to fuse and take advantages of the complementary nature of multimodal data to improve the prediction/classification accuracy.

How to fuse multimodal data? In the 1990s, as traditional **Machine Learning** (**ML**), a subclass of AI, flourished, ML-based models for addressing multimodal problems began to thrive. It became common for the machine to extract knowledge from multimodal data and make decisions. However, back then most of the works were focused on feature engineering, e.g., how to obtain a better representation for each modality. During that time, many modality-specific hand-crafted feature extractors were proposed, which greatly rely on prior knowledge of the specific tasks and the corresponding data. Since these feature extractors work independently, they can hardly capture the complementary and redundant nature of multimodalities. Therefore, such a feature engineering process inevitably results in a loss of information before the features are sent to the ML-based model. This leads to a negative impact on the performance of the traditional ML-based models. Although traditional ML-based models have the ability to analyze multimodal information, there is a long way to achieve the ultimate goal of AI, which is to mimic humans or even surpass human performance. Therefore, how to fuse the data in a way that can automatically learn the complementary and redundant information and minimize the manual interference remains a problem in the traditional ML field.

Deep learning is a sub-field of ML. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [88]. Its key advantage is that the hierarchical representations can be learned in an automated way, which does not require domain knowledge or human effort. For example, to the data: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, a two-layer neural network can be defined as the combination of matrices $\mathbf{W}$, and non-linear function $\sigma(\cdot)$ as shown in Equation (1). After training process, we can find $\mathbf{W}$ for which $\hat{\mathbf{y}}_\mathbf{i}$ is close to $\mathbf{y}_\mathbf{i}$ for all $i \leq N$. As the depth of the model continues to increase, so does its ability of feature representation.

$$\hat{\mathbf{y}}_\mathbf{i}(\mathbf{x}_\mathbf{i}) = \mathbf{W_2} \cdot \sigma(\mathbf{W_1}\mathbf{x}_\mathbf{i}). \tag{1}$$

Since 2010, multimodal data fusion has entered the stage of deep learning in an all-around way. Deep learning-based multimodal data fusion methods have demonstrated outstanding results in various applications. For video-audio-based multimodal data fusion, the works from [35, 37, 51, 163] address the emotion recognition problem by using deep learning techniques, including convolutional neural networks, **long short-term memory** (**LSTM**) networks, attention mechanisms, and so on. Also, for video-text multimodal data fusion, the works from [41, 56, 68, 107, 123, 124, 195] address the text-to-video retrieval task by using Transformer, BERT, attention mechanism, adversarial learning, and a combination of them. There are various other multimodal tasks, e.g., **visual question answering** (**VQA**) (text-image: [154, 220], text-video: [82, 223]), RGB-depth object segmentation [31, 39], medical data analysis [181, 185], and image captioning [216, 237]. Compared to traditional ML-based methods, **deep neural network** (**DNN**)-based methods show superior performance on representation learning and modality fusion if the amount of the training data is large enough. Furthermore, DNN is able to execute feature engineering by itself, which means a hierarchical representation can be automatically learned from data, instead of manually designing or handcrafting modality-specific features. Traditionally, the methods of multimodal data fusion are classified into four categories, based on the conventional fusion taxonomy shown in Figure 4, including early fusion, intermediate fusion, late fusion, and hybrid fusion: (1) early fusion: The raw data or pre-processed data obtained from each modality are fused before being sent to the model; (2) intermediate fusion: the features extracted from different modalities are fused together and sent to the model for decision making; (3) late fusion (also known as "decision fuse"): the individual decisions obtained from each modality are fused to form the final prediction, e.g., majority vote or weighted average, or a meta ML model on top of individual decisions. (4) hybrid fusion: a combination of early, intermediate, and late fusion. With large amounts

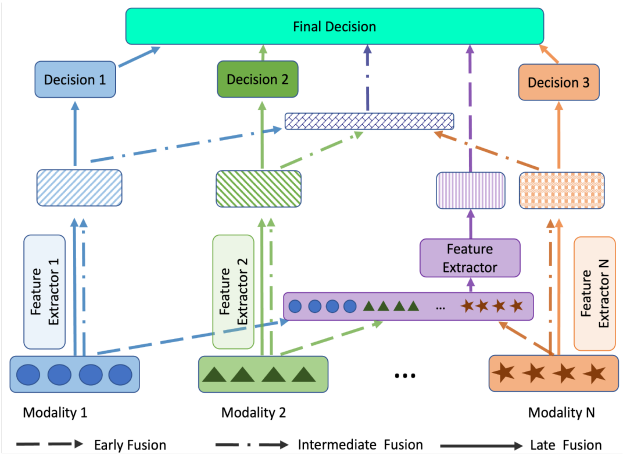


Fig. 4. The conventional taxonomy categorizes fusion methods into three classes.

of multimodal data available, the need for more advanced methods (VS handpicked ways of fusion) to fuse them has grown unprecedentedly. However, this conventional fusion taxonomy can only provide basic guidance for multimodal data fusion. In order to extract the richer representation from multimodal data, the architecture of DNN becomes more and more sophisticated, which no longer extracts features from each modality separately and independently. Instead, representation learning, modality fusing, and decision making are interlaced in most cases. Therefore, there is no need to specify exactly in which part of the network the multimodal data fusion occurs. The method of fusing multimodal data has changed from traditional explicit ways, e.g., early fusion, intermediate fusion, and late fusion, to more implicit ways. To force the DNN to learn how to extract complementary and redundant information of multimodal data, researchers have invented various constraints on DNN, including specifically designed network architectures and regularizations on loss functions, and so on. Therefore, the development of deep learning has significantly reshaped the landscape of multimodal data fusion, revealing the inadequacies of the traditional taxonomy of fusion methods. The inherent complexity of deep learning architectures often interlaces representation learning, modality fusing, and decision-making, defying the simplistic categorizations of the past. Furthermore, the shift from explicit to more implicit fusion methods, exemplified by attention mechanisms, has challenged the static nature of traditional fusion strategies. Techniques such as **graph neural networks (GNNs)** and **generative neural networks (GenNNs)** introduce novel ways of handling and fusing data that are not aligned with the early-to-late fusion framework. Additionally, the dynamic and adaptive fusion capabilities of deep models, coupled with the challenges posed by large-scale data, necessitate more sophisticated fusion methods than the conventional categories can encapsulate. Recognizing these complexities and the rapid evolution, it becomes imperative to introduce a taxonomy that delves deeper, capturing the subtleties of contemporary fusion methods.

For multimodal data fusion, there are several recent surveys available in the science community. Gao et al. [46] provide a review on multimodal neural networks and SOTA architectures. However, the review is only focused on a narrow research area: the object recognition task for RGB–depth images. Moreover, this survey is limited to the convolutional neural networks. Zhang et al. [235] present a survey on deep multimodal fusion. However, the authors categorize the models using the conventional taxonomy: early fusion, late fusion, and hybrid fusion. Furthermore, this survey is focused on the image segmentation task only. Abdu et al. [2] provide a literature review of

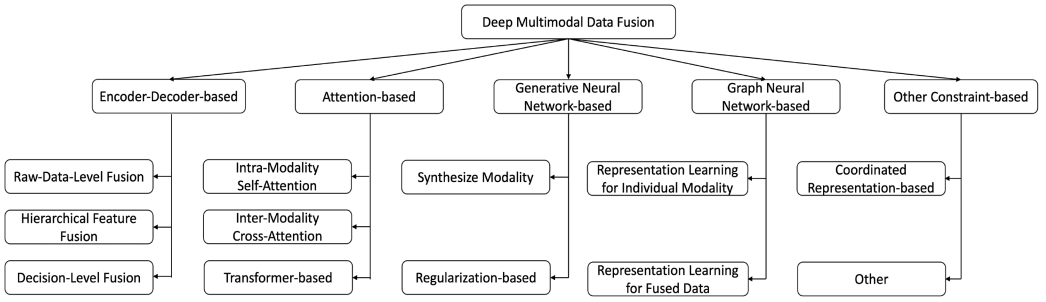


Fig. 5. The diagram of our proposed fine-grained taxonomy of deep multimodal data fusion models.

multimodal sentiment analysis using deep learning approaches. It categorizes the deep learning-based approaches into three classes: early fusion, late fusion, and temporal-based fusion. However, similar to the above surveys, this review is narrowly focused on sentiment analysis. Gao et al. [45] provide a survey on multimodal data fusion. It introduces the basic concepts of deep learning and several architectures of deep multimodal models, including stacked autoencoder-based methods, recurrent neural networks-based methods, convolutional neural network-based methods, and so on. However, it does not include the SOTA large pre-trained models and GNNs-based methods, e.g., the BERT model. Meng et al. [121] present a review of ML for data fusion. It emphasizes the traditional ML techniques instead of deep learning techniques. Also, the authors classify the methods into three different categories: signal-level fusion, feature-level fusion, and decision-level fusion. The way of categorizing the fusion methods is similar to that of the conventional taxonomy: early fusion, intermediate fusion, and late fusion, which is not new to the community. There are several other reviews [4, 128, 227] in the field of multimodality, most of which focus on a specific combination of modalities, e.g., RGB-depth images.

Therefore, in this article, we provide a comprehensive survey and categorization of deep multimodal data fusion. The contributions of this review are three-fold:

— We provide a novel fine-grained taxonomy of the deep multimodal data fusion models, diverging from existing surveys that categorize fusion methods according to conventional taxonomies such as early, intermediate, late, and hybrid fusion. In this survey, we explore the latest advances and group the SOTA fusion methods into five categories: Encoder-Decoder Methods, Attention Mechanism Methods, GNN Methods, GenNN Methods, and other Constraint-based Methods, as shown in Figure 5.
— We provide a comprehensive review of deep multimodal data fusion consisting of various modalities, including Vision+Language, Vision+Other Sensors, and so on. Compared to the existing surveys [2, 4, 45, 46, 121, 128, 227, 235, 243] that usually focus on one single task (such as multimodal object recognition) with one specific combination of two modalities (such as RGB+depth data), this survey owns a broader scope covering various modalities and their corresponding tasks, including multimodal object segmentation, multimodal sentiment analysis, VQA, and video captioning, and so on.
— We explore the new trends of deep multimodal data fusion, and compare and contrast SOTA models. Some outdated methods, such as deep belief networks, are excluded from this review. However, the large pre-trained models, which are rising stars of deep learning, are included in the review, e.g., Transformer-based pre-trained models.

The rest of this article is organized as follows. Section 2 introduces Encoder-Decoder-based fusion methods, in which the methods are grouped into three sub-classes. Section 3 presents the SOTA Attention mechanisms used in multimodal data fusion. In this section, the large pre-trained

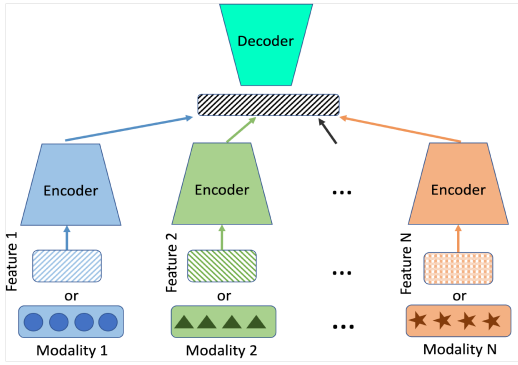


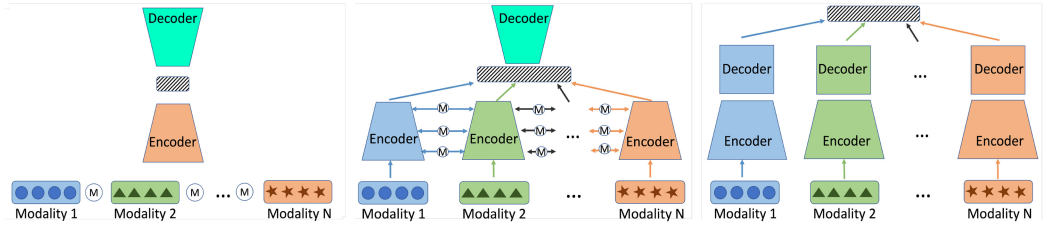
Fig. 6. The general structure of Encoder-Decoder method to fuse multimodal data. The input data of each encoder can be the raw data of each modality or the features of each modality. The encoders can be independent or share weights. The decoder can contain upsampling or downsampling operations, dependent on specific tasks.

models are introduced. In Section 4, we introduce GNN-based methods. In Section 5, we introduce GenNN-based methods, in which two main roles of GenNN-based methods in multimodal tasks are presented. Section 6 presents the other constraints adopted in SOTA deep multimodal models such as Tensor-based Fusion. In Section 7, the current notable tasks, applications, and datasets in multimodal data fusion will be introduced. Sections 8 and 9 discuss the future directions of multimodal data fusion and the conclusion of this survey.

2 ENCODER-DECODER-BASED FUSION

Encoder-Decoder architecture has been successfully adopted in single-modal tasks such as image segmentation, language translation, data reduction, and denoising. In such an architecture, the entire network can be divided into two major parts: the encoder part and the decoder part. The encoder part usually works as the high-level feature extractor, which projects the input data into a latent space with relatively lower dimensions compared to the original input data. In other words, the input data will be transformed into its latent representation by the encoder. During this process, the important semantic information of the input data will be preserved, while the noise in the input data will be removed. After the encoding process, the decoder will generate a "prediction" from the latent representation of the input data. For example, in a semantic segmentation task, the expected output of the decoder can be a semantic segmentation map with the same resolution as the input data. In a seq-2-seq language translation task, the output can be the expected sequence in the target language. In data denoising tasks, most works use a decoder to reconstruct the raw input data.

Owing to the strong representation learning ability and good flexibility of the network architecture of Encoder-Decoder models, Encoder-Decoder has been adopted in more and more deep multimodal data fusion models in recent years. Based on the differences in terms of the modalities and tasks, the architectures of multimodal data fusion models vary from each other widely. In this survey, we summarize the general idea of the Encoder-Decoder fusion methods and discard some of the task-specific fusion strategies that cannot be generalized. The general structure of the Encoder-Decoder fusion is shown in Figure 6. As we can see, the high-level features obtained from different individual modalities are projected into a latent space. Then, the task-specific decoder will generate the prediction from the learned latent representation of the input multimodal data. In real scenarios, there exists plenty of variations of this structure. We categorize them into 3 sub-classes: raw-data-level fusion, hierarchical feature fusion, and decision-level fusion.



(a) The raw-data-level fusion (b) The hierarchical feature fusion (c) The decision-level fusion

Fig. 7. Visualizations of different methods.

2.1 Raw-data-level Fusion

In this fusion, the raw data of each modality or the data obtained from the independent pre-processing of each modality will be integrated at the input level. Then, the formed input vector of the multimodalities will be sent to one encoder for extracting high-level features. The data from individual modalities are fused at a low level (e.g., the input level), and only one encoder is applied to extract the high-level features of multimodal data. For example, for the image segmentation task, Couprie et al. [27] propose the first deep learning-based multimodal fusion model. In this work, the authors fuse the multimodal data via a concatenation operation, in which the RGB image and the depth image are concatenated along the channel axis. Similarly, Liu et al. [109] concatenate RGB image and depth image together. The authors utilize depth information to assist color information in detecting salient objects with a lower computational cost compared to the double-stream network which consists of two separated sub-networks dealing with RGB data and depth data, respectively. The key advantages of this fusion are that (1) it can maximally preserve the original information of each modality, and (2) the design of a single backbone of the networks minimizes the computational cost. However, with the increasing number of modalities, the dimension of the merged input data will be extremely high. Therefore, usually, this fusion is only used for fusing the data from two modalities. It is worth mentioning that the raw-data-level combination is only suitable for homogeneous data. When it comes to heterogeneous data, e.g., text data+RGB image [207], data pre-processing is required, such as word embedding for text data. The visualization of raw-data-level fusion is shown in Figure 7(a). The Merge operation ("M") in Figure 7(a) is introduced in Figure 8, which usually involves element-wise Addition or Multiplication, Concatenation, and Cross Product.

2.2 Hierarchical Feature Fusion

Owing to the powerful ability of hierarchical representation learning of DNNs, unlike the raw-data-level fusion, many works use a well-designed architecture of networks forcing the model to fuse multimodal hierarchical features at different levels. The motivation of this fusion method is that fusing and aggregating the data from different levels of abstraction better leverages the multi-level features extracted from hierarchical deep networks, which can collectively improve the performance of the models. There are plenty of applications adopting this fusion method. For example, [69] and [171] did not simply stack depth image (or thermal image) with RGB image to form a four channels input data. The authors propose a strategy that fuses the hierarchical features of the two modalities at different levels by using element-wise summation. In this method, the RGB sub-network can be viewed as the backbone of feature extraction. The hierarchical features obtained from the other sub-network (e.g., thermal encoder and depth image encoder) will be fused with the RGB features by element-wise summation. Unlike the above works that use RGB

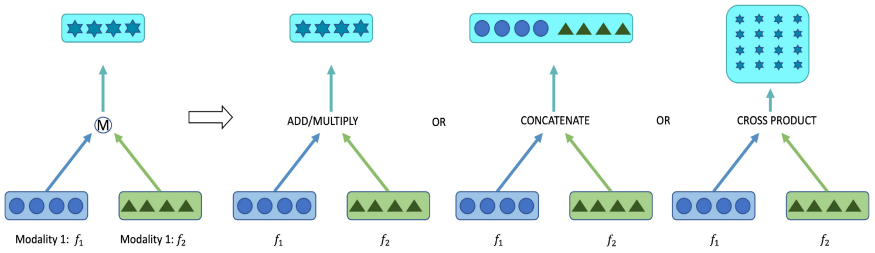


Fig. 8. The Merge operations: element-wise Addition or Multiplication, Concatenation, and Cross Product.

branch as the backbone of feature extraction, Hu et al. [64] build a new fusion sub-network as the backbone. The RGB and depth hierarchical features with the same resolution will be fused with the corresponding features of the fusion branch via an element-wise summation. In hierarchical fusion-based networks, the connection of features from different levels can help the model capture the cross-modality relationship. There are some other variations of this fusion method, e.g., in the medical field, Venugopalan et al. [179] propose a naive feature fusion network that integrates the data of three modalities by concatenating the last layers of the encoders. Hong et al. [59] propose an encoder-decoder-based fusion model for hyperspectral and LiDAR data, in which the fusion happens in the decoder of the network. In referring image segmentation task, [97] and [65] propose fusion models in which the RGB image and the referring expression are fed into convolutional neural networks or recurrent neural networks to generate their feature representations independently, and then fuse these features in the decoding stage. Similarly, in scene understanding task, [173] and [226] hierarchically fuse the features (e.g., low level, middle level, and high level features) from different modalities to improve the model performance, while [170] fuses the high level features together. The key advantages of this type of fusion include (1) the flexibility of the fusion architecture—one can decide where the fusion happens and how many hierarchical features are fused for specific tasks, (2) easy to be combined with attention mechanism—the connection between multimodality hierarchical features at the same level can be upgraded with attention mechanism, which will be introduced in Section 3. This allows researchers to leverage the relationship among different modalities to enhance the performance of the fusion model. One of the major drawbacks of this fusion method is that the individual sub-networks for different modalities demand substantial computational resources when the number of modalities is relatively high. Therefore, this fusion method is usually adopted for the fusion of two or three modalities. The visualization of hierarchical feature fusion is shown in Figure 7(b).

2.3 Decision-level Fusion

Different from the above hierarchical feature fusion strategy, which provides much flexibility on architecture design, decision-level fusion is relatively straightforward, less flexible, but easy to implement. The fusion operation in this method is fixed at the end of the decoder or classifier of individual sub-networks, which means that the cross-modal information is exchanged in the last layers or the penultimate layers of decoders. It provides limited interpretability of the multimodal interactions. For classification tasks, the final fusion can be accomplished by using a classic majority vote or the weights learned from multilayer perceptron. For regression tasks, usually a linear regressor is trained to fuse the predictions of individual modalities. For example, Zhang et al. [234] integrate the outputs of two individual decoders by concatenating them along the channel axis. Similarly, Aygün et al. [9] address the brain tumor segmentation problem by using decision-level fusion. The advantages of this method are (1) it can be used to explore the relative contribution weight of each

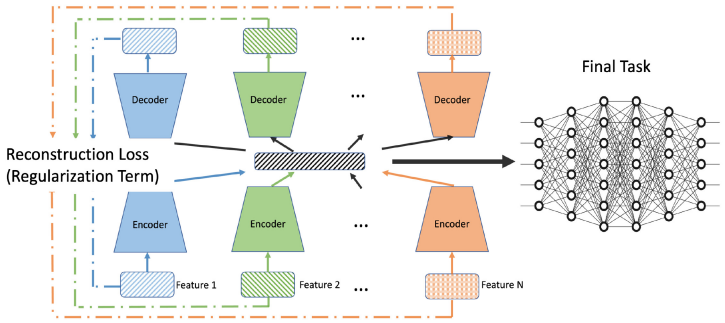


Fig. 9. The general idea of using auto-encoder method to fuse the multimodal data.

modality to generating the final decision, (2) easy to tell whether the prediction result from each modality is correct, and (3) the networks are easy to design and implement. However, the obvious drawbacks, compared to hierarchical feature fusion, are that (1) the performance of the whole networks can be limited by one modality (e.g., the branch for one modality does not work properly and generates a wrong prediction severely affecting the final prediction negatively), and (2) less flexibility in how multimodal information can be fused. The visualization of decision-level fusion is shown in Figure 7(c).

So far, we have explored three typical architectures of Encoder-Decoder fusion. In real scenarios, there can be numerous variants of them. For example, for the segmentation task, the encoder performs feature extraction with downsampling, and the decoder applies upsampling to compensate for the downsampling effect of the encoder. However, for some of the other tasks, e.g., discrimination tasks, which do not require resolution recovery, the upsampling operation will not be required in the decoders. Then, the decoder sub-networks can be replaced with generic convolutional neural networks or stacked fully connected layers to handle discrimination tasks. Another common variation is reconstruction loss-based autoencoder-decoder (AE) fusion. The general architecture is shown in Figure 9. The final loss function of AE-based fusion consists of reconstruction losses and task-specific losses. In the medical field, Wang et al. [181] propose an AE-based Drug-Target interaction prediction model. It uses two separate encoder-decoder sub-networks to extract the independent latent representations for the two input modalities. Then, the learned representations are concatenated for the downstream tasks. Similarly, Bera et al. [16] propose two AE sub-networks for representation learning. This method is straightforward. However, the reconstruction of each modality only based on independent high-level features cannot explore the cross-modality information. Also, the simple concatenation operation provides weak constraints for exploiting the cross-modal relations. Differently, Hong et al. [59] propose an AE-based segmentation model, in which the reconstruction of each modality is based on the common latent representations of both modalities. It means the model maps the input data from two different modalities into a common space to obtain a new representation of the input data, and then, reconstructs each modality from the learned representations. This design forces the model to exploit the cross-modal relationships. The two reconstruction losses work as regularization terms in the final loss function. Bendre et al. [14] propose a multimodal **variational autoencoder** (**VAE** [83]) architecture, which can learn the shared latent space of image features. The model concatenates multimodal data to form a single embedding before passing it to the VAE for learning the latent space. Khattar et al. [81] propose a VAE-based end-to-end architecture for fake news classification problem. The model concatenates the textual features and visual features together to form the embedding. Then, the autoencoder reconstructs the word embedding features and visual features.

In this section, we have reviewed the encoder-decoder architecture-based fusion models. As one of the most famous generative structures, the high flexibility of neural network architectures enables researchers to build more powerful networks to extract features and fuse the features from different modalities. However, during the fusion process, the multiplication, concatenation, cross product, or element-wise addition, is relatively naive and brute force. More and more researchers are spending efforts on how to merge features from different modalities so that the relationship among different modalities can be easily explored.

3 ATTENTION-BASED FUSION

The concept of attention mechanism is firstly introduced by [10]. Back then it was only used to improve the encoder-decoder-based neural machine translation system in the **natural language processing** (**NLP**) field. Since Vaswani et al. [177] published their groundbreaking work "Transformers" in the article "Attention Is All You Need", the attention mechanism has become one of the hottest topics in the deep learning community. Attention mechanisms enable models to assign different weights to each part of the input data so that the model can extract significant information that is more relevant and critical to the current task. It can help the model make more accurate predictions without increasing the computational cost drastically.

Lots of variations of the attention mechanism have been proposed, such as channel attention and spatial-attention in the **computer vision** (**CV**) field, e.g., [233, 239], and self-attention and multi-head attention in the NLP field, e.g., [177]. These variations enable the attention mechanism not only to work for NLP tasks but also to be used to address problems in other fields. For example, scaled dot-product attention can be used in various tasks. The input consists of queries, keys, and values. It computes the dot products of the query Q with all keys K, divides each by $\sqrt{d_k}$ (in which d_k is the dimension of keys), and applies a softmax function to obtain the weights on the values [177]. The equation is shown as Equation (2):

$$\textbf{Attention}(Q, K, V) = \textbf{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

Recently, the attention mechanism has become one of the main tools for multimodal data fusion tasks. The attention mechanism-based multimodal models can be categorized into three classes:

3.1 Intra-modality Self-attention

The general structure is shown in Figure 10(a). The motivation of this method is to force the model to exploit the intra-modality relationship. The attention operation can be dot-product-based [85], or additive gate-based [134], and so on. This means that for a given modality, the attention operation solely considers data from that specific modality. In the context of the Transformer model [177], the Key (K), Query (Q), and Value (V) tensors used in attention computation are identical and all derived from the same modality or sequence, as shown in the left part of Figure 10(c). This ensures that the attention process solely concentrates on the data for each single modality, allowing for a focused and undiluted analysis of intra-modality relationships. This method is commonly used in multimodality tasks. For example, Gao et al. [47] propose the intra-inter modality attention module-based model to address the VQA task. The authors applied the intra-modality attention mechanism to enhance the feature learning ability of the sub-network of each modality. Similarly, Malinowski et al. [115] propose a hard-attention-based multimodal fusion method, which produces a binary mask over spatial locations determining which features are passed on to downstream processing. Meanwhile, soft-attention mechanisms have gained widespread popularity in multimodal fusion tasks, primarily due to their differentiability, which stands in stark contrast to hard-attention mechanisms. For the multimodal image segmentation task, Mohla et al. [126]

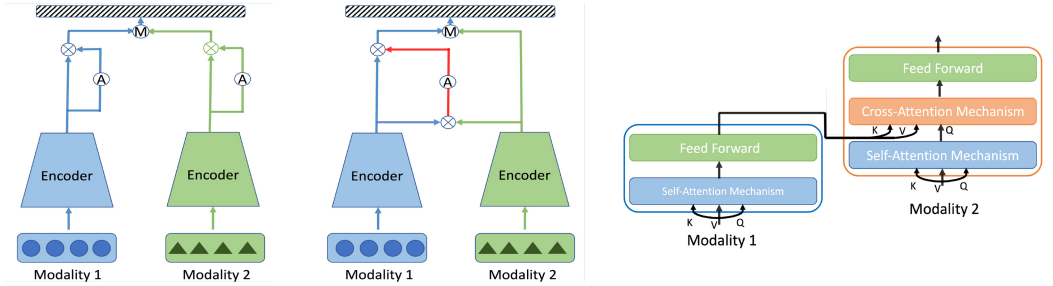


Fig. 10. Illustration of different attention mechanisms and fusion architectures. (a) shows the attention mechanism focusing on intra-modality relationship. (b) shows the attention mechanism focusing on inter-modality relationship. (c) shows Transformer-based architecture consisting of intra-modality self-attention and inter-modality cross-attention.

proficiently implemented soft-attention mechanisms, including channel- and spatial- attention mechanisms, on two specialized sub-networks: the **Hyperspectral Imagery** (**HSI**) sub-network and the LiDAR sub-network. This approach accentuates the relative importance of crucial areas within each modality, with the channel-attention mechanism in the HSI sub-network focusing on the most informative channels, and the spatial-attention mechanism in the LiDAR sub-network emphasizing the most contributive spatial regions.

The intra-modality self-attention mechanism offers several advantages, including its flexibility, ease of implementation, and relatively low computational cost, primarily because it circumvents the intricate analysis required to discern differences and exploit correlation between various modalities. However, by focusing solely on intra-modality relationships, this approach might overlook valuable complementarity among various modalities that can enhance the model performance.

3.2 Inter-modality Cross-attention

The general structure is shown in Figure 10(b). Complementary to the intra-modality self-attention, the inter-modality cross-attention mechanism focuses on exploiting the relationship among different modalities. The attention scores are computed using multimodal data. This means that each attention operation considers data from multiple modalities. In the context of the Transformer model [177], the Query (Q) tensor and the Key (K) and Value (V) tensors used in attention computation are derived from two or more different modalities or sequences, as shown in the right part of Figure 10(c). Since some modality streams can contain more information for the task at hand than the others, the obtained attention weights can be applied to the more informative modalities only. It will produce an attention-pooled feature for one modality conditioned on another modality. For example, Zhang et al. [228] apply the dot product attention mechanism to explore the inter-modality relationship between text and image features. Mohla et al. [126] propose a multimodal model based on spatial attention and channel attention. In addition to the intra-modality attention mechanism, the authors adopted an inter-modality attention mechanism to exploit the cross-modality relations between the LiDAR modality and the HSI modality. It can be viewed as LiDAR-guided HSI attention networks. Similarly, Hu et al. [65] propose a bi-directional inter-modal cross-attention module, in which the authors create a vision-guided linguistic attention module and a language-guided visual attention module to exploit the inter-modality relationship between the vision modality and the linguistic modality. Differently, in some combinations of multimodalities, it is hard to tell whether one of them is relatively more informative. Therefore, lots of works applied the attention scores on all

the modalities instead of on the most informative modality only. For example, Wu et al. [200] propose a co-attention-based multimodal fake news detection model. In the model, before each fusion operation, they enhance each modality with the other modality by using the co-attention mechanism. The stacked multiple co-attention layers force the model to fuse the multimodal features and to learn inter-dependencies among them. Similarly, Sun et al. [169] propose an inter-modality cross-attention mechanism specifically designed to learn the associations between audio and text modalities, calculating dot products of the Query and Key of audio and text in a crossed way. The cross-attention module guides one modality to attend to the other, updating the features in a manner that reflects the learned inter-modal associations, thereby enriching the understanding of the latent relationships between the modalities. Furthermore, Lu et al. [111] develop a distinctive cross-attention mechanism, integrating channel-attention and feature-intersection mechanisms. This approach facilitates dynamic information interaction among disparate modalities, allowing the model to emphasize more representative features. Later, Yoon et al. [215] proposed a sophisticated multimodal encoder that leverages cross-attention to incorporate visual and acoustic representations. This approach delves deep into understanding the latent information interwoven between the modalities, providing a holistic view and enriched representation of the multimodal data.

The inter-modality cross-attention mechanism, while powerful, presents certain challenges in its applications. The nature of inter-modality cross-attention, which seeks to manage and leverage relationships between different modalities, can inherently introduce computational and structural complexities. This becomes particularly pronounced as the number of modalities increases, demanding more computational resources and intricate management. Meanwhile, the efficacy of the inter-modality cross-attention mechanism is deeply intertwined with the quality and relevance of the modalities it deals with. Consequently, modalities of poor quality or those that are misaligned can significantly impede the optimal performance of the attention mechanism, leading to subpar results.

The intra-modality self-attention and inter-modality cross-attention mentioned above can be easily adopted in DNN with great flexibility. They can work together or be mixed with other types of multimodal fusion methods. For example, Gao et al. [47] propose an intra-inter modality attention module-based model to address the VQA task, in which the intra-attention enhances individual modality features, and the inter-attention captures interactions among various modalities. Other than that, there are many innovative works of self-attention to explore cross-modality relations and reduce computational cost. By concatenating representations of different modalities into a latent space and applying self-attention to these new representations, models can effectively exploit cross-modality relations. For instance, the attention mechanism proposed by Ye et al. [211] adaptively shifts focus to salient words in the query expression and significant portions of the input image.

However, traditional attention mechanisms, despite their efficacy in discerning local patterns and relationships, encounter limitations in recognizing long-range dependencies within data. This is attributed to the localized nature of their receptive fields, which potentially hampers the model's capacity to assimilate information from distant segments of the input. To counteract this limitation, the concept of non-local attention [186] has been introduced. This innovative approach is structured to contemplate relationships throughout the entire input space, thereby enabling models to apprehend and utilize long-range dependencies effectively. A notable implementation of non-local attention is the work by Yuan et al. [221], where non-local attention-based networks are leveraged for the fusion of homogeneous multimodal image data, such as the integration of MRI and PET or the amalgamation of infrared and visible images. Unlike localized attention mechanisms, non-local attention transcends the constraints of proximity and provides a holistic perspective of the input space, making it a valuable asset in the advancement of multimodal data fusion techniques.

Table 1. The List of the Large Pre-trained Models

Model Name and Ref.	Year	Transformer Archt.	Modality	Tasks
LXMERT [172]	2019	Multi-Transformers	Image + Text	VQA
Uniter [24]	2020	Uni-Transformer	Image + Text	VQA, VCR, ITR
UniT [62]	2021	Multi-Transformers	Image + Text	OD, VQA
ClipBERT [90]	2021	Uni-Transformer	Video + Text	VR
VLM [201]	2021	Multi-Transformers	Video + Text	VQA, VR, VC
VATT [3]	2021	Uni-Transformer	Video + Text + Audio	AC, AEC
DeCEMBERT [174]	2021	Uni-Transformer	Video + Text	VR, VC
interBERT [101]	2020	Uni-Transformer	Image + Text	ITM
Pixel-BERT [68]	2020	Uni-Transformer	Image + Text	VQA, ITR
B2T2 [5]	2019	Uni-Transformer	Image + Text	VCR
VLC-BERT [147]	2023	Uni-Transformers	Video + Text	VQA
Unicoder-VL [92]	2020	Multi-Transformers	Image + Text	AC, VC
HERO [94]	2020	Uni-Transformers	Video + Text	VQA, VR, VC
ActBERT [244]	2020	Multi-Transformers	Video + Text	VR, VX, AS
UniVL [113]	2020	Multi-Transformers	Video + Text	VR, VC, SA
ImageBERT [139]	2020	Uni-Transformer	Image + Text	ITM
VisualBERT [96]	2019	Uni-Transformer	Image + Text	VQA, VCR
ViLBERT [112]	2019	Multi-Transformers	Image + Text	VQA
VideoBERT [168]	2019	Uni-Transformer	Video + Text	AC
CBT [167]	2019	Multi-Transformers	Video + Text	AC, VC, AS
Zorro [148]	2023	Uni-Transformers	Video + Audio	AC
X-llm [22]	2023	Multi-Transformers	Image + Text	VQA,VC
MIST [43]	2023	Uni-Transformers	Video + Text	VQA
NExT-GPT [199]	2023	Uni-Transformers	Video + Text	VQA,VC, VCR

SA: sentiment analysis, ITR: image-text retrieval, VCR: visual commonsense reasoning, AS: action segmentation, VR: video retrieval, VQA: visual question answering, VC: video captioning, OD: object detection, ITM: image text matching, AC: action classification, AEC: audio event classification

3.3 Transformer-based Methods

Building on the idea of non-local attention, the Transformer architecture [177] has emerged as a groundbreaking solution. At the heart of the Transformer is the self-attention mechanism, which allows each output element to consider all input sequences (or image patch embeddings) simultaneously, effectively capturing both local and long-range dependencies without being bound by the constraints of traditional convolutional or recurrent layers. This global perspective, combined with the architecture scalability, has made Transformers particularly suited for tasks that benefit from understanding intricate relationships across data. Transformer-based large pre-trained models have been dominant on lots of multimodal data fusion tasks, such as interBERT [101] and videoBERT [168]. The state-of-the-art large pre-trained models are summarized and shown in Table 1.

The Transformer model is first introduced by [177]. It is a combination of Encoder-Decoder architecture and Attention mechanism [70], as shown in Figure 10(c). In the Encoder, there are stacked self-attention blocks, in which K, Q, V of the scaled dot product attention mechanism are from the same tensor, to explore the intra-modality relationship of the input. In the Decoder, there are stacked self-attention blocks and a cross-attention block, in which the K, Q, V are from different modalities, e.g., Q is from the second modality , while K, V are from the first modality. These self-attention and cross-attention blocks help the model capture the intra- and inter-relationship within and among multimodalies efficiently. Currently, the Transformer-based large pre-trained models can be divided into two categories: (1) uni-Transformer architecture: in this architecture, the input data from different modalities will be jointly processed by a single encoder or several stacked encoders, such as VideoBERT[168], HERO [94], NExT-GPT [199], ClipBERT [90], and

DeCEMBERT [174]; (2) multi-Transformers architecture: in this architecture, the input data from different modalities will be encoded separately by modality-specific Transformers before being jointly modelled, such as X-llm [22], UniVL [113], and ActBERT [244]. The large pre-trained models are able to learn the comprehensive representations for multimodalities and obtain competitive performance on downstream tasks. However, currently, most large pre-trained models are focused on the vision-language field. For the other types of modality, the resource of large pre-trained models is still limited. Therefore, several works build their own transformer-based model for the specific tasks that do not have any pre-trained transformer-based models available. For example, the work in [203] leverages the correlation between the MRI and the acoustical signals for vocal tract deformation task by using a cross-modal transformer-based architecture. Similarily, Hsu et al. [60] propose a multimodal transformer to capture the long-range dependency among multimodal data, e.g., text, image, numerical data, and categorical data, by the self attention mechanism.

4 GRAPH NEURAL NETWORK-BASED FUSION

So far, we have reviewed the Encoder-Decoder-based fusion and Attention-based fusion. The models of these methods have achieved great success in capturing hidden patterns from the data within Euclidean space. However, it is difficult for them to address the data that is generated from non-Euclidean domains, represented as graphs with complex relationships and interdependencies between objects [241]. Recently, there are an increasing number of applications based on GNNs addressing the multimodal problems related to graph data, e.g., VQA task [77, 95, 100, 129, 190, 214, 218], image captioning task [42, 182, 182, 208–210], cross-modal retrieval task [30, 204, 217], RGB-depth scene classification task [222], multimodal recommendation task [193, 194], neuroimaging-based disease classification task [158], 3D object localization with natural language descriptions [21], and object segmentation in the 3D scenes according to the query sentences [66]. GNN has emerged as a potent tool to process and integrate data structured as graphs, especially when modalities are inherently relational or interconnected. Among GNN, **Graph Convolutional Networks (GCNs)** stand out, leveraging convolutional layers adapted for graph data to aggregate information from neighboring nodes, thereby facilitating the fusion of spatially localized features across modalities. Another notable subtype, **Graph Attention Networks (GATs)**, introduces attention mechanisms to the graph structure. By dynamically weighing the importance of neighboring nodes, GAT offers a refined focus on relevant parts of the graph, enhancing the fusion process by capturing intricate patterns and relationships across diverse data sources.

The general strategy of applying GNN to multimodal data fusion can be categorized into two different classes:

4.1 Representation Learning for Individual Modalities

The visualization is shown in Figure 11(a). In this strategy, the GNN is used to extract the new representation from the graph data only, which means the sub-branches consisting of non-graph structured data will not use GNN for feature extraction. Then, the learned representations from different modalities will be integrated. For example, Lotfi et al. [110] propose a multimodal data fusion method for detecting rumor conversations, in which the authors model the user graph and reply tree by using GCN independently. Then, the features obtained from the user graph and the reply tree are concatenated and sent to a fully connected layer to detect the rumor conversation. Unlike the above work that fuses the GCN embeddings from different modalities by using a naive concatenation operation only, Qian et al. [142] adopt an inter-modal attention mechanism to exploit the relationships between visual features and textual features. In this work, the authors use GCN and VGG-19 [162] to obtain the textual representation and the visual representation of each post, respectively. During the feature extraction process, the authors use text-guided vision attention to

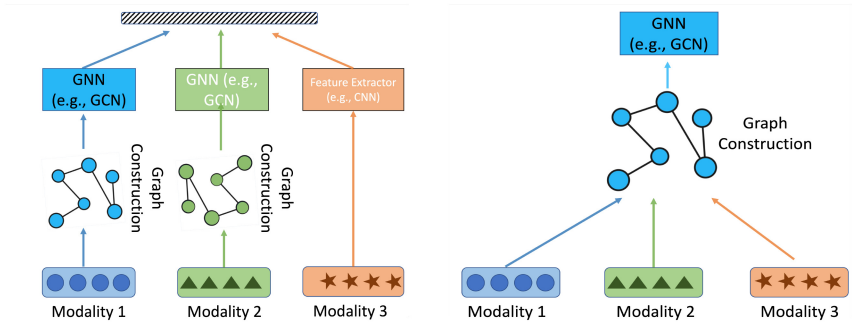


Fig. 11. Illustration of GNN-based multimodal data fusion and integration. (a) shows a general schema of GNN-based multimodal data fusion. (b) shows how the integration of multimodal data occurs in the graph construction process.

enforce an interaction between the textual feature and the visual feature. It can determine which visual feature deserves more attention. Beyond methods based on GCN, there are many other types of GNN-based methods. For instance, Yang et al. [208] introduce a GNN model utilizing multi-head attention to embed scene graphs effectively, thus to enhance the performance of image captioning tasks. This method focuses on capturing intricate relationships within scene graphs to generate more accurate and contextually relevant captions for images. Similarly, the works of Tao et al. [175] and Jia et al. [75] introduce multimodal GATs tailored for personalized recommendations. Those works conduct information propagation within individual single-modal graphs and use attention mechanisms to identify varying importance scores of different modalities to user preference. The GNN not only works well for language data and vision data but also has achieved success on biomedical data and chemical data. For instance, Wang et al. [185] propose multi-omics GCNs for a biomedical classification task, in which the model utilizes GCN to extract independent features from three modalities, i.e., mRNA expression data, DNA methylation data, and microRNA expression data, separately. Then, they create a correlation matrix based on these features to exploit the latent cross-modality correlations that help to improve the learning performance.

4.2 Representation Learning for the Fused Data

The visualization is shown in Figure 11(b). The key operation of this fusion strategy is graph construction. In general, unlike the previous strategy which can have multiple sub-networks or sub-modality branches, this strategy fuses the multimodal data in the graph construction before the representation learning process. For example, Hu et al. [61] propose a multimodal fusion model via deep graph convolution network for emotion recognition in conversation. The key contribution of this article is that it creates acoustic nodes, textual nodes, and visual nodes for utterances. Then, any two nodes in the same modality are connected in the graph. The edge between them is called the intra-modal edge. Furthermore, each node is connected with the nodes which correspond to the same utterance but from different modalities. The edges connecting them are called inter-modal edges. This operation forces the model to exploit the intra-modality and inter-modality relationships. Then, the stacked GCN is adopted to yield the high-level node representations. Finally, these new representations will be concatenated and sent through a fully connected layer to form the predictions. Similarly, Wang et al. [188] propose a knowledge-driven multimodal GCN to detect fake news. Three modalities are involved in this application: text, image, and knowledge concept. The

words in the text are taken as the graph nodes and the relationship between words is taken as edges. Unlike the work in [142] that uses VGG-19 to extract high-level visual representation and concatenate it with the textual representation, this work utilizes YOLOv3 [149] pre-trained model to detect semantic objects in images. Then, it treats the textual labels of detected objects in images as words that occur in the text content. Therefore, these textual labels of images are used in the graph construction. This operation forces the interaction between different modalities to happen. Then, GCN is applied to yield embedding vectors for these nodes based on the properties of their neighborhoods. In addition to GCN-based methods, a variety of approaches employ GAT to address diverse challenges. For instance, Jiang et al. [93] introduce a GAT-based network designed to perform cross-modal feature complementation and multimodal emotion classification tasks. The authors utilize three uni-modal encoders to encode the uni-modal features initially. Subsequently, a GAT-based cross-modal feature complementation module is deployed to gather both long-distance intra-modal contextual information and inter-modal interactive information. This method excels in preserving the consistency and diversity of multimodal features, pinpointing essential intra-modal contextual information and inter-modal interaction information, and mitigating the heterogeneity gap prevalent in multimodal data. Similarly, Ding et al. [33] propose a GAT-based fusion model. This model integrates a multimodal GAT with a temporal convolution network to discern the spatial-temporal correlations inherent in multimodal time series. The authors leverage various attention mechanisms, including self-attention and cross-attention, to explicitly model the correlations between different modalities, providing a deep understanding of the interactions and relationships within multimodal data.

The advantages of GNN-based fusion models, compared to other fusion methods, include (1) their ability to directly process graph-structured data by deep learning technologies without requiring the projection of data into Euclidean space, and (2) their intuitive way of exploiting relationships among nodes in the graph-structured data and can be extended to exploit the intra-modality and inter-modality relationships in multimodal problems. However, the drawback of GNN-based fusion models is that the graph construction process is usually highly dependent on prior knowledge of the characteristics of the specific input data and task. It is time- and space-consuming and not easy to be generalized. So far, we have reviewed the Encoder-Decoder-based fusion, Attention-based fusion, and GNN-based fusion. All of them can leverage the relationship among different modalities to improve the performance of multimodal networks. However, such fusion methods have difficulty dealing with missing data issues.

5 GENERATIVE NEURAL NETWORK-BASED FUSION

GenNN serve as a foundational pillar in the domain of deep learning, particularly for tasks centered around data generation, reconstruction, and modeling. These networks are designed to capture and replicate the underlying distributions of data, making them invaluable for a myriad of applications, from image synthesis to time-series forecasting. GenNN-based networks typically encompass a range of architectures, including but not limited to, **Generative Adversarial Networks** (**GANs**), VAEs, Flow-based and Diffusion-based models. The primary objective of GenNN is to generate data that closely mirrors real-world distributions, either by directly modeling these distributions or by learning to transform simpler distributions into more complex ones. Their versatility and capability to generate high-quality data have led to their widespread adoption in both single-modal and multimodal tasks, addressing challenges such as data imputation, augmentation, and fusion. Given their generative prowess, generative models have become instrumental in scenarios where real data is scarce, noisy, or incomplete, providing a robust mechanism to supplement and enhance existing datasets.

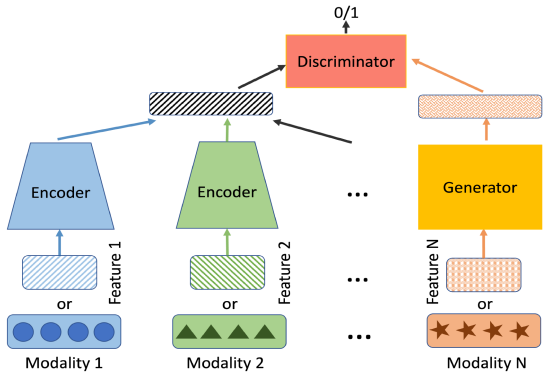


Fig. 12. The general architecture of GenNN-based methods.

For multimodal data fusion tasks, the data collection process is relatively more sophisticated, compared to single-modal tasks. Missing data is one of the common issues of multimodal datasets. Here, we define the missing data issue as the issue caused by the multimodal sample, in which at least one modality is missing. In most works related to multimodality, the researchers simply assume that the dataset is without any missing data, or apply data cleaning for removing all the samples containing missing data. The consequence is that most of those works only work well theoretically and cannot be adapted to real scenarios. To address the missing data issues of multimodal data fusion, GenNNs have emerged as powerful tools, particularly for tasks that involve data generation, reconstruction, and fusion. In the context of multimodal data fusion, GenNN-based networks offer a robust framework to address challenges such as missing data and the synthesis of new data modalities.

The intuition behind them is that the model can synthesize the missing modality based on the other modalities. The general idea is shown in Figure 12. For example, Gao et al. [48] propose a model TPA-GAN, which is used to synthesize the **positron emission tomography** (**PET**) data from MRI data. The generator of the TPA-GAN consists of the typical encoder-decoder architecture, pyramid structure, and an attention mechanism. Once the synthesized PET data is generated, a discriminator will discriminate whether the PET data is "Real" or "Fake". The authors use the TPA-GAN model to impute the missing PET data of samples. Then, they propose a PT-DCN model, an encoder-decoder-based model, to fuse the MRI and PET data and form the final classification of **Alzheimer's disease** (**AD**). In this work, the authors create two models separately: (1) TPA-GAN model to impute the missing PET images, and (2) PT-DCN model to fuse the multimodal data and complete the classification task with the imputed multimodal images. Unlike the above work imputing the missing data separately from the fusion process, Wang et al. [183] propose a generative multi-view action recognition model. Within the fusion model, two generators $G1$ and $G2$ generate representations conditionally based on the other subspace. This approach enhances the model robustness by employing adversarial training and naturally handles the incomplete view case by imputing the missing data.

Exploring along a different direction, there are a lot of multimodal applications using GenNN as a regularizer to leverage the semantic correlations among modalities. For example, Sahu and Vechtomova [153] propose a novel GAN-based multimodal data fusion model, in which the authors leverage a complementing modality to regularize the learned latent space of multimodalities. Peng and Qi [137] propose a GAN-based cross-modal representation learning model. In their method, a cross-modal GAN architecture is proposed to model the joint distribution over the data of different modalities. The authors proposed four discriminators: two intra-modality discriminators to keep

the semantic consistency within each modality, and two inter-modality discriminators to exploit the cross-modal correlation. Similar to the above works, Wang et al. [189] propose a VAE-based adversarial multimodal framework that reduces the distance difference between unimodal representations and transfers all unimodal representations to a joint embedding space, highlighting important sentimental representations over time and modality.

Furthermore, the capabilities of GenNN in synthesizing and modifying one modality based on other modalities have spurred the development of a myriad of tasks and architectures, such as Diffusion-based models, VAE-based models, and Flow-based models. These models are particularly prominent in tasks such as text-conditional image generation and image style transfer. For instance, Fan et al. [36] introduce a Flow-based model composed of a series of fully invertible basic blocks, enabling lossless image reconstruction based on both text and image inputs. This model exemplifies the innovative integration of different modalities to achieve coherent and high-fidelity reconstructions. Similarly, the works by [34] and [145] leverage discrete VAEs to compress each RGB image into an n × n grid of image tokens. These tokens, along with text tokens, are subsequently fused in downstream networks, such as Transformers, illustrating the versatility of VAE in handling and integrating diverse modalities. Following the pioneering works of Song et al. [165] and Ho et al. [58], who proposed diffusion probabilistic models to synthesize realistic images from textual prompts, a surge of Diffusion-based methods has emerged in the realm of multimodal fusion. Examples include the Diffusion-based text-to-speech model by [73] and various Diffusion-based image generation models such as DALL-E2 [144], Imagen [152], and Stable diffusion [151]. These models underscore the growing significance and adaptability of Diffusion-based methods in synthesizing and fusing multimodal information. These advancements illustrate the expanding horizons of GenNN in multimodal fusion, showcasing their potential in synthesizing diverse modalities and fostering innovations in multimodal interactions and representations.

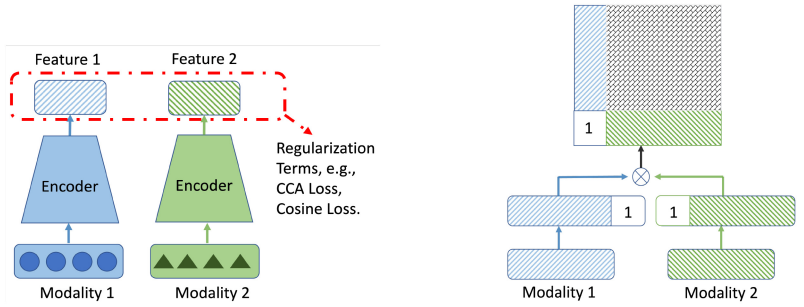
In summary, GenNN-based models can be used to address the missing data issues for multimodal tasks and can also be treated as a regularizer to leverage the semantic correlations among multimodalities. However, when it comes to exploiting the intra- and inter-modality relationships among multiple modalities to improve the performance of the models, the architecture flexibility of the GenNN-based networks is relatively low, and a lot of training skills are required. In this regard, the Attention Mechanism attracts more attention in the scientific community. Because the Attention Mechanism has a strong ability to reveal intra- and inter-relationship among different modalities, it has been widely used in multimodal data fusion.

6 OTHER CONSTRAINT-BASED METHODS

Most of the fusion strategies that we reviewed above are based on joint representations, which means that the input multimodal data will be mapped into a common latent space. The model will learn a joint representation of the input data. However, there is another line of method named coordinated representation-based framework, which learns separated but coordinated representations of each modality under certain constraints.

As Figure 13(a) shows, coordinated representation architecture handles individual modality separately, but enforces certain similarity constraints on them to bring them into a coordinated space [15]. The learned representations of each modality can be compared against each other by using **canonical-correlation analysis** (**CCA**) constraint, cosine distance constraint, L2 distance constraint, or other constraints [26, 57]. These similarity constraints will serve as a regularization term in the loss function.

Other than the regularization-based methods, Zadeh et al. [224] propose the first tensor-based fusion network. It mainly considers both the inter- and intra-modality relationships. As shown in Figure 13(b), the method expands each modality by 1 dimension, then, calculates the Cartesian



(a) Coordinated representations (b) Tensor fusion mechanism

Fig. 13. (a) shows the visualization of coordinated representation framework with certain constraints to keep the learned representations semantically consistent. (b) shows the tensor fusion mechanism focusing on inter and intra-modality relationship.

product of different modalities. Taking two modalities as an example, the authors perform the outer product (tensor product) after obtaining the features. It can be seen that after the expansion, the feature correlation between the two modalities is obtained, and the information of each specific modality is preserved. For three modalities, the method not only obtains feature correlations for two-modal and three-modal combinations, but also preserves the features of each specific modality. However, with the increase in the number of multimodalities, the feature dimension will explode quickly. Moreover, when the number of parameters is too large, it easily increases the risk of overfitting. Furthermore, there are variations of this work, e.g., [108]. In this work, the authors propose a low-rank multimodal fusion method, which performs multimodal fusion using low-rank tensors to improve efficiency.

Differently, Wang et al. [187] propose a channel-exchanging-network, which dynamically exchanges channels in different modal sub-networks. To be specific, the channel exchanging process is self-guided by individual channel importance measured by the magnitude of batch-normalization scaling factor during training. This method is parameter-free. However, this model only works on homogeneous data. The performance of heterogeneous multimodal data is limited.

One of the major drawbacks of these coordinated representation-based fusions is that they are suitable for scenarios where there are two/three input modalities. When the number of modalities is larger than three, the network architecture can be too complicated, and the performance cannot be guaranteed.

So far, we have reviewed the Encoder-Decoder-based methods, Attention mechanism-based methods, GNN-based methods, GenNN-based methods, and other Constraint-based methods. Currently, most SOTA algorithms of multimodal data fusion are designed for a combination of two specific modalities only. Therefore, we make a comparison among them according to how well they can be generalized for a higher number of modalities (the number of modalities is bigger than two.) (1) Encoder-Decoder-based methods. The architectures of networks are relatively flexible. It is easy to integrate a new modality into Encoder-Decoder-based models, e.g., adding a new sub-encoder branch for a new modality [23]. Moreover, the weights of all the sub-encoders can be shared. When the number of sub-encoders increases, the computational cost will not increase drastically [87]. (2) GenNN-based methods encompass a variety of architectures, including but not limited to, GAN, VAE, Diffusion, and Flow-based models. As the diversity and the number of input modalities increase, the architecture of GenNN-based multimodal networks tends to become increasingly intricate. This complexity arises due to the necessity to incorporate distinct

Table 2. A Comparison among the Five Categories of Fusion Methods

Methods	Generalization (Num of Modalities ≥ 3, Easy, Medium, Hard)
Encoder-Decoder-based Methods	**Easy**. Easy to add sub-branches for new modalities; encoders' weights can be shared.
GenNN-based Methods	**Hard**. Adding a new modality often requires the integration of additional generative components. Balancing and training a network with multiple such components can be increasingly intricate and computationally demanding.
Attention-based Methods	**Easy**. Easy to be adopted to explore the inter- and intra-modality relationships among multi-modalities regardless of the number of input modalities.
GNN-based Methods	**Medium**. Adding one new modality will lead to the reconstruction of the graph-structured data. The aggregation functions of networks might need to be modified as well.
Other Constraint-based Methods	**Hard**. Easy to add sub-branches for new modalities to learn separated but coordinated representations for each modality. Weights of the sub-networks cannot be shared so that the computational cost will increase drastically with the increase of sub-branches.

Table 3. A Quantitative Comparison among the Fusion Methods

Tasks	**Dataset**	**Encoder-Decoder** Metric: Min [method] – Max [method]	**Attention Mechanism** Metric: Min [method] – Max [method]	**Graph Networks** Metric: Min [method] – Max [method]	**Generative Networks** Metric: Min [method] – Max [method]
Referring Image Segmentation	ReferIt [80]	IoU: 52.81 [119]–63.63 [98]	IoU: 63.80 [211]–**72.97** [104]	IoU: 65.53 [67]	IoU: 60.31 [143]
Video Captioning	YouCookII [242]	METEOR: 19.91 [117]	METEOR: 18.23 [138]–**27.09** [155]	METEOR: 21.77 [74]–22.1 [54]	METEOR: 19.77 [225]
		BLEU-4: 11.43 [117]	BLEU-4: 9.82 [138]–**21.88** [155]	BLEU-4: 11.74 [74]–14.0 [54]	BLEU-4: 12.14 [225]
Visual Question Answering	VQA 2.0 [53]	ACC: 68.14 [125]	ACC: 67.34 [156]–**71.94** [157]	ACC: 65.89 [132]–71.29 [205]	ACC: 57.87– [25]
Object Recognition	SUN RGB-D [164]	ACC: 54.60 [240]	ACC: 60.8 [236]–61.4 [238]	ACC: 57.0 [141]–**64.4** [206]	–
		mIoU: 42.8 [240]–50.7 [64]	mIoU: 47.8 [236]–51.8 [238]	mIoU: 45.9 [141]–**51.8** [206]	–
Face Anti Proofing	CASIA-SURF [230]	ACER(%): 3.8 [232]	ACER(%): 0.74 [32]–**0.20** [180]	–	ACER(%): 2.4 [103]
Person Re-Identification	RGBD-ID [13]	mAP(%): 14.42 [212]–15.95 [196]	mAP(%): : 27.91 [76]–40.52 [140]	mAP(%): **53.02** [213]	mAP(%): 29.2 [191]–31.49 [29]
BLEU-4 [136], METEOR[12], CIDEr[178], Accuracy (ACC), Intersection over Union (IoU), Average Classification Error Rate (ACER), mean Average Precision (mAP)					

generative structures or modules for each new modality, each requiring specialized handling and processing. For instance, in GAN-based models, each additional modality typically necessitates the integration of a new generator and a new discriminator into the networks. The inherent variability in the data distribution and representation across different modalities necessitates meticulous fine-tuning and optimization to ensure coherent and meaningful fusion, making the generalization to multiple modalities non-trivial. (3) Attention mechanism-based methods. The Attention mechanism is widely used for exploring the inter- and intra-modality relationships between two modalities. When the number of input modalities increases, the attention mechanism can still work well in exploring the correlation within each modality or across modalities without increasing the computational cost drastically. (4) GNN-based methods. The graph construction process is the most time-consuming part of this method. When the number of input modalities increases, in general, a more sophisticated graph construction process is needed. However, GNN network structure for extracting new representations of nodes will not be changed in most cases. (5) Other Constraint-based methods. Most of them are based on coordinated representation learning. The key idea of them is using sub-networks to learn separated but coordinated representations for each modality under certain constraints. And the weights of each sub-branch cannot be shared with each other, i.e., when the number of modalities increases, the number of sub-branches will be increased correspondingly. The computational cost will also increase drastically. Therefore, the other Constraint-based methods are usually applied to problems containing only two modalities. The comparison is summarized in Table 2. Unlike GenNN and other Constraint-based methods which can work well only on few (e.g., 2) modalities, the Encoder-Decoder-based methods and the Attention mechanism-based methods can be easily generalized to three or more modalities. Furthermore, these two types of method can work together very well, e.g., [19] and [198].

In order to have a more granular comparison to discern the nuanced differences in their performances across diverse tasks, we have compiled Table 3, providing a detailed quantitative comparison among the fusion methods across diverse tasks, with the best performers highlighted for each

Table 4. Public Datasets for Common Multimodal Data Fusion Tasks

Tasks	Data Type	Dataset
Referring Image Segmentation	Image + Text	ReferIt [80], Google-Ref [118], UNC [219], CLEVR-Ref+ [106], VGPhraseCut [197], ScanRefer [21], ClevrTex [79]
Video Captioning/Retrieval	Video + Text	Howto100M [122], Alivol-10M [89], YouCookII [242], Charades [160], TGIF [99], MSR-VTT [202], Didemo [6]
Visual Question Answering	Vision + Text	video-text: TVQA [91], VQA 1.0 [7], VQA 2.0 [53], image-text: DAQUAR [116], COCO-QA [150],FM-IQA [44], Visual Genome [86], CLEVR [78],FVQA [184]
Object Recognition	RGB + Depth	SUN RGB-D [164], NYU Depth V1 and V2 [130, 161], RobotPKU [105], B3DO [71], BIWI [127],
Object Detection/ Semantic Segmentation	RGB+LiDAR	KITTI [49], Urban [17], KAIST [72], RobortCar [114], US3D [40] College [146], H3D [84]
Human Action Recognition	RGB + Inertial Sensors	MHAD [133], C-MHAD [192], UTD-MHAD [20], CAS-MHAD [55], NCTU-MFD [166]
Face Anti Proofing	RGB + Thermal	CASIA-SURF [230], CASIA-SURF Cefa [102] Speakingfaces [1], ThermalFace [8], TTVF [63], UL-FMTV [52],
Person Re-Identification	RGB+Thermal	RegDB [131], SYSU-MM01 [196], BIWI RGBD-ID [127], RGBD-ID [13], SUCVL RGBD-ID [120], RobotPKU [105], KinectREID [135]

task. Firstly, models utilizing a foundational encoder-decoder architecture, even in the absence of supplementary methods such as attention mechanisms, have demonstrated reasonable efficacy across a spectrum of tasks. This underscores the inherent robustness and versatility of the encoder-decoder paradigm in multimodal fusion tasks, highlighting its capability to integrate seamlessly a variety of modalities. Secondly, attention mechanisms have emerged as a predominant and versatile strategy, reflected by their widespread adoption and adaptability across a myriad of tasks. Most models leveraging attention mechanisms have manifested superior performance in various domains, underscoring the efficacy of attention mechanisms in discerning and emphasizing salient features and relationships within the data. Thirdly, GNN-based methods, while exhibiting substantial potential and outperforming attention-based models in specific tasks such as object recognition and person re-identification, are inherently constrained by the suitability of the task to graph representations. Their performance accentuates their ability to encapsulate complex inter-relational structures within the data, but their applicability is not universal and is contingent on the nature of the task. Lastly, GenNN-based methods, similar to graph networks, excel in particular contexts but are not universally applicable. Their performance is contingent on the alignment of the task with their architectural strengths, indicating a need for task-specific considerations in selecting an appropriate model.

7 APPLICATIONS AND DATASET

Currently, there are plenty of applications related to deep multimodal data fusion in the scientific community and industry. With the increase in the variety of modalities, the types of downstream tasks of multimodal data fusion are increasing. Here, we choose some of the trending applications related to multimodal data fusion and categorize them into three classes: Vision and Language, Vision and Sensors, and Others. The datasets of these tasks are summarized and shown in Table 4.

7.1 Vision and Language

In the last decade, with the great success of deep learning in non-interdisciplinary research, e.g., CV and NLP, more and more attention is geared towards addressing interdisciplinary problems by using deep learning. In the vision and language fusion, many classical tasks have been proposed: (1) VQA, which is proposed by [7]. The objective of this task is to generate a textual answer for a given image/video and its corresponding textual question. In this task, the multimodal model is supposed to be able to integrate the semantic information of the visual data and the textual data. The expected answer (ground truth) can be words, binary values, phrases, or the number of objects

in the visual data, and so on. (2) Image/video captioning. The objective of this task is to generate a textual description for a given image/video. The description is supposed to summarize the content of visual data. (3) Referring image/video segmentation. The objective of this task is to generate the segmentation of a certain entity referred to by a natural linguistic expression. (4) **Image-text matching (ITM)** or image-text retrieval. The goal of this task is to find the semantically closest item of one modality to the query in the other modality. The query can be a sample in visual modality or textual modality. (5) Reference expression comprehension. The objective of this task is to generate the bounding boxes in the image corresponding to the given text.

7.2 Vision and Sensors

In recent years, autonomous driving systems based on multimodalities have attracted attention in the industry. There are an increasing number of applications related to the fusion of RGB modality and other modalities, e.g., LiDAR, Radar, and infrared data. Here, we introduce the two most famous tasks of the fusion of vision data and other sensors: (1) Multimodal semantic segmentation task. The objective of this task is to generate the semantic segment of objects based on the multimodal input data. In this task, the combination of input data can vary, such as RGB-depth data, RGB-infrared data, RGB-LiDAR data, RGB-Radar data, RGB-depth-LiDAR data, RGB-skeleton data, and CT-MRI data. (2) Multimodal object detection task. In this task, the model is supposed to generate the bounding box of each instance in the multimodal input data. (3) Multimodal sentiment analysis. Unlike the conventional single-modal data-based sentiment analysis models, e.g., text-based models or image-based models, the model is supposed to be able to fuse the multimodal input data and predict the sentiment of the sample. (4) Multimodal person re-identification task. Compared to the traditional person re-identification task, the major difference is that the input data can be the combinations of different modalities, e.g., RGB+depth data, RGB+thermal data, and RGB+depth+thermal data. (5) multimodal human activity recognition task. This task is an integral part of a variety of applications, such as video gaming and health monitoring. The input data can be RGB+depth data, RGB+infrared data and RGB+inertial data, and so on. (6) Multimodal Face Anti-Spoofing Task. In this task, a model is designed to distinguish between genuine and spoof face presentations by leveraging multimodal data, such as RGB images, depth information, and infrared images. The fusion of different modalities enables the model to capture various facial features and anomalies that are not discernible in single-modal data, enhancing the robustness and accuracy of anti-spoofing systems. For instance, the combination of RGB and infrared data can help in detecting the presence of masks or other artifacts used for spoofing, while the integration of depth information can reveal abnormal facial structures or the absence of three-dimensional features in a spoof attempt. (7) Multimodal person re-identification task. This task involves identifying individuals across different camera views or sessions by fusing data from multiple modalities such as RGB images, depth data, and thermal imagery. Unlike traditional person re-identification tasks that rely solely on RGB data, multimodal person re-identification models integrate diverse sensory inputs to extract more comprehensive and discriminative features of individuals. For example, the fusion of RGB and thermal data can be particularly effective in environments with varying lighting conditions, allowing the model to recognize individuals based on their thermal signatures. Similarly, incorporating depth data can provide additional structural information about the individual, enhancing the model's ability to match persons in different poses or viewpoints.

7.3 Other

Other than the tasks mentioned above, there are many other miscellaneous multimodal data fusion tasks. For example, in the biomedical field, there are drug-protein/protein-protein interaction

prediction task, classification tasks for different diseases, e.g., AD and cancer subtype. In the chemistry field, there are multimodal toxicity classification, multimodal retention index prediction, and multi-sensor-based gas type classification, and so on.

8 FUTURE DIRECTIONS

Deep learning-based multimodal data fusion has seen a rapid growth in the recent decade. However, there are still several research gaps:

8.1 Missing Modality Challenges

In real scenarios, there are two sub-classes: missing modality issue and noisy modality issue. The missing modality issue is that at least one modality is absent in a multimodal sample. The noisy modality issue is that at least one modality's data is noisy or misaligned. Most SOTA methods are based on the assumption that there is no missing data issue in the dataset. Many SOTA deep data fusion models based on this assumption can only work well under ideal conditions. Currently, there are several works addressing missing modality issues, e.g., [179] proposes a GAN-based network to reconstruct the missing PET modality from available MRI data in samples; the work in [176] averages the learned latent vector of individual modality from independent encoders, and reconstructs each modality from the averaged latent vector. However, these methods naively assume that each modality is as same informative as the others and contributes to the final task equally. In the future, (1) to create adaptive fusion models, which can learn the relative importance of different modalities for the final task, can be a potential direction of improvement in such models. The adaptive fusion models should be able to automatically drop the influence of less informative modality on the task, and vice versa. (2) To develop fully distributed deep multimodal fusion. In the centralized setting, local features are transmitted or relayed to a fusion center where the final decision is made. Devices or sensors may enter or leave the network dynamically, leading to unpredictable changes in network size and topology. Sensors or devices may disappear permanently either due to damage to the nodes or drained batteries. Moreover, connectivity/communication between the devices of each modality and the fusion center is rarely perfect because of bandwidth constrains and energy limitations [18, 231]. This motivates us to consider a fully decentralized deep multimodal fusion framework without a fusion center, where all the modalities continuously update and exchange the local decisions/features with their neighbors to reach a consensus.

8.2 Lack of Data

Multimodal data fusion is an emerging research field of artificial intelligence. The publicly available multimodal datasets are still limited. As we know, the performance of deep learning-based models is typically dependent on the number of samples used during the training process. A high-quality and large-scale dataset will hugely help the model learn accurate and comprehensive representations of the observed objects or activities. Therefore, creating larger and higher-quality multimodal datasets is one of the critical tasks to move the field forward.

8.3 Lack of Large Pre-trained Models

Large pre-trained models are able to learn more comprehensive representations for multimodalities. With transfer learning, the well-trained large pre-trained models can have competitive performance on downstream tasks compared to task-specific designed models. However, currently, existing large pre-trained multimodal models only focus on the interdisciplinary field of CV and NLP. In the future, creating large pre-trained multimodal models for other interdisciplinary fields can be a potential direction in multimodal data fusion.

8.4 Interpretability of Models

Despite the incredible success of data driven approaches in different disciplines, there are some drawbacks of deep learning models that limit their applicability. For instance, it typically requires a massive training data and intensive computational resources to learn a desirable mapping, which are hardly available in communication degraded or energy constrained environments. Moreover, DNN are commonly designed to be black boxes that do not explain how to understand and characterize the prediction result and confident intervals [159]. In order to address this challenge, there are two potential ways: (1) Combine statistical signal processing framework and deep learning for improved scalibilty and interpretability. Statistical signal processing and model-based approaches employ concrete mathematical formulations that represent the underlying physics, prior information and additional domain knowledge of the problem, thus offering interpretability, flexibility, versatility, scalibility, and robustness [159, 229]. In deep multimodal fusion under various piratical constraints, the network topology, traffic, and channel conditions may change dynamically, and it is worthwhile to investigate the combination of complementary strengths of both statistical signal processing and deep learning approaches. Therefore, to combine statistical signal processing framework and deep learning for improved scalibilty and interpretability can be a promising direction of multimodal data fusion. (2) Inference and decision making with humans in the loop. In critical high-stake situations, incorporating human cognitive strengths and expertise in heterogeneous multimodality networks is imperative to improve decision quality and to enhance situational awareness. The next generation of smart manufacturing, intelligent IoT, augmented reality, and remote diagnosis systems would require the seamless integration of the human and the machine/deep learning algorithm in the same environment to understand and solve problems. For example, in natural disaster or pandemic early warning, alert and response systems where human lives and assets depend on a small detail being observed or missed, automatic ML based decision making may not be sufficient, and it is necessary to incorporate human(s) in the loop of decision making, intelligence gathering, policy making and control. Modeling and analysis of human decision making is still a challenging problem, as it needs to consider several factors including cognitive biases of humans, mechanisms to handle behavior uncertainties of humans, as well as human machine interactions, in contrast to decision making processes consisting of only machine agents [38, 50].

9 CONCLUSION

In this survey, we introduce the background and review the contemporary models of deep multimodal data fusion. We provide a novel fine-grained taxonomy which groups SOTA multimodal data fusion methods into five categories: Encoder-Decoder Methods, Attention Mechanism Methods, GNN Methods, GenNN Methods, and other Constraint-based Methods. In addition, a wide range of applications and datasets related to multimodal fusion are also included. Finally, the future research directions for deep multimodal data fusion are explored.

REFERENCES

[1] Madina Abdrakhmanova, Askat Kuzdeuov, Sheikh Jarju, Yerbolat Khassanov, Michael Lewis, and Huseyin Atakan Varol. 2021. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *Sensors* 21, 10 (2021), 3465.

[2] Sarah A. Abdu, Ahmed H. Yousef, and Ashraf Salem. 2021. Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion* 76, C (2021), 204–226.

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems* 34 (2021), 24206–24221.

[4] Furqan Alam, Rashid Mehmood, Iyad Katib, Nasser N. Albogami, and Aiiad Albeshri. 2017. Data fusion and IoT for smart ubiquitous environments: A survey. *IEEE Access* 5 (2017), 9533–9554.

[5] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2131–2140.

[6] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*. 5803–5812.

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

[8] Roshanak Ashrafi, Mona Azarbayjani, and Hamed Tabkhi. 2022. Charlotte-ThermalFace: A fully annotated thermal infrared face dataset with various environmental conditions and distances. *Infrared Physics and Technology* 124 (2022), 104209.

[9] Mehmet Aygün, Yusuf Hüseyin Şahin, and Gözde Ünal. 2018. Multi modal convolutional neural networks for brain tumor segmentation. arXiv:1809.06191. Retrieved from https://arxiv.org/abs/1809.06191

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. Retrieved from https://arxiv.org/abs/1409.0473

[11] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[12] Satanjeev Banerjee and Alon Lavie. 2005. 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*. 65–72.

[13] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. 2012. Re-identification with rgb-d sensors. In *Proceedings of the Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Part I 12*. Springer, 433–442.

[14] Nihar Bendre, Kevin Desai, and Peyman Najafirad. 2021. Generalized zero-shot learning using multimodal variational auto-encoder with semantic concepts. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1284–1288.

[15] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[16] Padmalochan Bera and Shobh. 2021. ModCGAN: A multimodal approach to detect new malware. In *Proceedings of the 2021 International Conference on Cyber Situational Awareness, Data Analytics, and Assessment (CyberSA)*. IEEE, 1–2.

[17] José-Luis Blanco-Claraco, Francisco-Angel Moreno-Duenas, and Javier González-Jiménez. 2014. The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. *The International Journal of Robotics Research* 33, 2 (2014), 207–214.

[18] Müjdat Cetin, Lei Chen, John W. Fisher, Alexander T. Ihler, Randolph L. Moses, Martin J. Wainwright, and Alan S. Willsky. 2006. Distributed fusion in sensor networks. *IEEE Signal Processing Magazine* 23, 4 (2006), 42–55.

[19] Chongqing Chen, Dezhi Han, and Jun Wang. 2020. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access* 8 (2020), 35662–35671.

[20] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 168–172.

[21] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision*. Springer, 202–221.

[22] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv:2305.04160. Retrieved from https://arxiv.org/abs/2305.04160

[23] Huai Chen, Yuxiao Qi, Yong Yin, Tengxiang Li, Xiaoqing Liu, Xiuli Li, Guanzhong Gong, and Lisheng Wang. 2020. MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma. *Neurocomputing* 394 (2020), 27–40.

[24] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 104–120.

[25] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11681–11690.

[26] Jonghyun Choi, Hyunjong Cho, Jungsuk Kwac, and Larry S. Davis. 2014. Toward sparse coding on cosine distance. In *Proceedings of the 2014 22nd International Conference on Pattern Recognition*. IEEE, 4423–4428.

[27] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. 2013. Indoor semantic segmentation using depth information. arXiv:1301.3572. Retrieved from https://arxiv.org/abs/1301.3572
[28] Leandro Cruz, Djalma Lucio, and Luiz Velho. 2012. Kinect and rgbd images: Challenges and applications. In *Proceedings of the 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*. IEEE, 36–49.
[29] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the IJCAI*. 6.
[30] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* 27, 8 (2018), 3893–3903.
[31] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. 2019. RFBNet: Deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation. arXiv:1907.00135. Retrieved from https://arxiv.org/abs/1907.00135
[32] Pengchao Deng, Chenyang Ge, Hao Wei, Yuan Sun, and Xin Qiao. 2023. Attention-aware dual-stream network for multimodal face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 18 (2023), 4258–4271.
[33] Chaoyue Ding, Shiliang Sun, and Jing Zhao. 2023. MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection. *Information Fusion* 89 (2023), 527–536.
[34] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021), 19822–19835.
[35] Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, and Wolfgang Minker. 2020. An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. arXiv:2010.03692. Retrieved from https://arxiv.org/abs/2010.03692
[36] Weichen Fan, Jinghuan Chen, Jiabin Ma, Jun Hou, and Shuai Yi. 2022. Styleflow for content-fixed image to image translation. arXiv:2207.01909. Retrieved from https://arxiv.org/abs/2207.01909
[37] Zeinab Farhoudi and Saeed Setayeshi. 2021. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. *Speech Communication* 127 (2021), 92–103.
[38] Charles Findling and Valentin Wyart. 2021. Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences* 38 (2021), 124–132.
[39] Fahimeh Fooladgar and Shohreh Kasaei. 2019. Multi-modal attention-based fusion model for semantic segmentation of RGB-depth images. arXiv:1912.11691. Retrieved from https://arxiv.org/abs/1912.11691
[40] K. Foster, G. Christie, and M. Brown. 2020. IEEE Dataport: Urban semantic 3D dataset. https://doi.org/10.21227/9frn-7208
[41] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*. Springer, 214–229.
[42] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12746–12756.
[43] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. MIST: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14773–14783.
[44] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in Neural Information Processing Systems* 28 (2015), 2296–2304.
[45] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.
[46] Mingliang Gao, Jun Jiang, Guofeng Zou, Vijay John, and Zheng Liu. 2019. RGB-D-based object recognition using multimodal convolutional neural networks: A survey. *IEEE Access* 7 (2019), 43110–43136.
[47] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6639–6648.
[48] Xingyu Gao, Feng Shi, Dinggang Shen, and Manhua Liu. 2022. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in alzheimers disease. *IEEE Journal of Biomedical and Health Informatics* 26, 1 (2022), 36–43.
[49] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.
[50] Baocheng Geng and Pramod K. Varshney. 2019. On decision making in human-machine networks. In *Proceedings of the 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 37–45.

[51] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. 2019. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 552–558.

[52] Reza Shoja Ghiass, Hakim Bendada, and Xavier Maldague. 2018. Université laval face motion and time-lapse video database (ul-fmtv). In *Proceedings of the 14th International Conference on Quantitative Infrared Thermography*.

[53] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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*. 6904–6913.

[54] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. 2023. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18941–18951.

[55] Yanan Guo, Dapeng Tao, Weifeng Liu, and Jun Cheng. 2016. Multiview cauchy estimator feature embedding for depth and inertial sensor-based human action recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 4 (2016), 617–627.

[56] Ning Han, Jingjing Chen, Hao Zhang, Huanwen Wang, and Hao Chen. 2022. Adversarial multi-grained embedding network for cross-modal text-video retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2 (2022), 1–23.

[57] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664.

[58] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[59] Danfeng Hong, Lianru Gao, Renlong Hang, Bing Zhang, and Jocelyn Chanussot. 2022. Deep encoder-decoder networks for classification of hyperspectral and LiDAR data. *IEEE Geoscience and Remote Sensing Letters Letters* 19 (2022), 1–5.

[60] Chih-Chung Hsu, Pi-Ju Tsai, Ting-Chun Yeh, and Xiu-Yu Hou. 2022. A comprehensive study of spatiotemporal feature learning for social medial popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7130–7134.

[61] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. 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)*. Online, 5666–5675.

[62] Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1439–1449.

[63] Shuowen Hu, Jonghyun Choi, Alex L. Chan, and William Robson Schwartz. 2015. Thermal-to-visible face recognition using partial least squares. *JOSA A* 32, 3 (2015), 431–442.

[64] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. 2019. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1440–1444.

[65] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. 2020. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4424–4433.

[66] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1610–1618.

[67] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. 2020. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10488–10497.

[68] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv:2004.00849. Retrieved from https://arxiv.org/abs/2004.00849

[69] Shang-Wei Hung, Shao-Yuan Lo, and Hsueh-Ming Hang. 2019. Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2374–2378.

[70] O. Iosifova, I. Iosifov, and O. Rolik. 2020. Techniques and components for natural language processing. *MoMLeT&DS* 2631, I (2020), 57–67.

[71] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. 2013. A category-level 3d object dataset: Putting the kinect to work. In *Proceedings of the Consumer Depth Cameras for Computer Vision*. Springer, 141–165.

[72] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. 2019. Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research* 38, 6 (2019), 642–657.
[73] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. 2021. Diff-tts: A denoising diffusion model for text-to-speech. arXiv:2104.01409. Retrieved from https://arxiv.org/abs/2104.01409
[74] Lei Ji, Rongcheng Tu, Kevin Lin, Lijuan Wang, and Nan Duan. 2022. Multimodal graph neural network for video procedural captioning. *Neurocomputing* 488 (2022), 88–96. DOI: https://doi.org/10.1016/j.neucom.2022.02.062
[75] Xiangen Jia, Min Jiang, Yihong Dong, Feng Zhu, Haocai Lin, Yu Xin, and Huahui Chen. 2023. Multimodal heterogeneous graph attention network. *Neural Computing and Applications* 35, 4 (2023), 3357–3372.
[76] Jianguo Jiang, Kaiyuan Jin, Meibin Qi, Qian Wang, Jingjing Wu, and Cuiqun Chen. 2020. A cross-modal multi-granularity attention network for RGB-IR person re-identification. *Neurocomputing* 406 (2020), 59–67.
[77] Lei Jiang and Zuqiang Meng. 2023. Knowledge-based visual question answering using multi-modal semantic graph. *Electronics* 12, 6 (2023), 1390.
[78] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2901–2910.
[79] Laurynas Karazija, Iro Laina, and Christian Rupprecht. 2021. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. arXiv:2111.10265. Retrieved from https://arxiv.org/abs/2111.10265
[80] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 787–798.
[81] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *Proceedings of the World Wide Web Conference*. 2915–2921.
[82] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. 2018. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 673–688.
[83] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv:1312.6114. Retrieved from https://arxiv.org/abs/1312.6114
[84] Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 1 (2021), 100001.
[85] D. N. Krishna. 2021. Using large pre-trained models with cross-modal attention for multi-modal emotion recognition. arXiv:2108.09669. Retrieved from https://arxiv.org/abs/2108.09669
[86] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 706–715.
[87] Ryohei Kuga, Asako Kanezaki, Masaki Samejima, Yusuke Sugano, and Yasuyuki Matsushita. 2017. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 403–411.
[88] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[89] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2567–2576.
[90] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. 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*. 7331–7341.
[91] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. Tvqa: Localized, compositional video question answering. arXiv:1809.01696. Retrieved from https://arxiv.org/abs/1809.01696
[92] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11336–11344.
[93] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2024. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia* 26 (2024), 77–89.
[94] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv:2005.00200. Retrieved from https://arxiv.org/abs/2005.00200
[95] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10313–10322.

[96] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv:1908.03557. Retrieved from https://arxiv.org/abs/1908.03557

[97] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5745–5753.

[98] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5745–5753. DOI: https://doi.org/10.1109/CVPR.2018.00602

[99] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4641–4650.

[100] Yaoyuan Liang, Xin Wang, Xuguang Duan, and Wenwu Zhu. 2021. Multi-modal contextual graph neural network for text visual question answering. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 3491–3498.

[101] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. Interbert: Vision-and-language interaction for multi-modal pretraining. arXiv:2003.13198. Retrieved from https://arxiv.org/abs/2003.13198

[102] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. 2021. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1179–1187.

[103] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z. Li. 2021. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2759–2772.

[104] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. 2023. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing* 32 (2023), 3054–3065.

[105] Hong Liu, Liang Hu, and Liqian Ma. 2017. Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology* 2, 1 (2017), 48–55.

[106] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4185–4194.

[107] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11915–11925.

[108] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. arXiv:1806.00064. Retrieved from https://arxiv.org/abs/1806.00064

[109] Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. 2019. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing* 363 (2019), 46–57.

[110] Serveh Lotfi, Mitra Mirzarezaee, Mehdi Hosseinzadeh, and Vahid Seydi. 2021. Detection of rumor conversations in Twitter using graph convolutional networks. *Applied Intelligence* 51, 7 (2021), 4774–4787.

[111] Houhong Lu, Yangyang Zhu, Ming Yin, Guofu Yin, and Luofeng Xie. 2022. Multimodal fusion convolutional neural network with cross-attention mechanism for internal defect detection of magnetic tile. *IEEE Access* 10 (2022), 60876–60886.

[112] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019), 13–23.

[113] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv:2002.06353. Retrieved from https://arxiv.org/abs/2002.06353

[114] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 2017. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research* 36, 1 (2017), 3–15.

[115] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. 2018. Learning visual question answering by bootstrapping hard attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–20.

[116] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems* 27 (2014), 1682–1690.

[117] Xin Man, Deqiang Ouyang, Xiangpeng Li, Jingkuan Song, and Jie Shao. 2022. Scenario-aware recurrent transformer for goal-directed video captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 4 (2022), 1–17.

[118] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11–20.

[119] Edgar Margffoy-Tuay, Juan C. Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 630–645.

[120] KAMAL UDDIN MD. 2021. Cross-modal and multi-modal person re-identification with RGB-D sensors. *Array* 12 (2021), 100089.

[121] Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. 2020. A survey on machine learning for data fusion. *Information Fusion* 57 (2020), 115–129.

[122] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

[123] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 19–27.

[124] Niluthpol C. Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2019. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval* 8, 1 (2019), 3–18.

[125] Safaa Abdullahi Moallim Mohamud, Amin Jalali, and Minho Lee. 2023. Encoder–decoder cycle for visual question answering based on perception-action cycle. *Pattern Recognition* 144 (2023), 109848.

[126] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. 2020. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 92–93.

[127] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. 2014. One-shot person re-identification with a consumer depth camera. *Person Re-Identification* (2014), 161–181.

[128] Chems Eddine Louahem M'Sabah, Ahmed Bouziane, and Youcef Ferdi. 2021. A survey on deep learning methods for cancer diagnosis using multimodal data fusion. In *Proceedings of the 2021 International Conference on e-Health and Bioengineering (EHB)*. IEEE, 1–4.

[129] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in Neural Information Processing Systems* 31 (2018), 2654–2665.

[130] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Proceedings of the ECCV*.

[131] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17, 3 (2017), 605.

[132] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. *Advances in Neural Information Processing Systems* 31 (2018), 8334–8343.

[133] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 53–60.

[134] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Ruecker. 2018. Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999. Retrieved from https://arxiv.org/abs/1804.03999

[135] Federico Pala, Riccardo Satta, Giorgio Fumera, and Fabio Roli. 2015. Multimodal person reidentification using RGB-D cameras. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 4 (2015), 788–799.

[136] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

[137] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1 (2019), 1–24.

[138] Jeripothula Prudviraj, Malipatel Indrakaran Reddy, Chalavadi Vishnu, and Chalavadi Krishna Mohan. 2022. AAP-MIT: Attentive atrous pyramid network and memory incorporated transformer for multisentence video description. *IEEE Transactions on Image Processing* 31 (2022), 5559–5569.

[139] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv:2001.07966. Retrieved from https://arxiv.org/abs/2001.07966

[140] Meibin Qi, Suzhi Wang, Guanghong Huang, Jianguo Jiang, Jingjing Wu, and Cuiqun Chen. 2021. Mask-guided dual attention-aware network for visible-infrared person re-identification. *Multimedia Tools and Applications* 80 (2021), 17645–17666.

[141] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 2017. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 5199–5208.

[142] Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3 (2021), 1–23.
[143] Shuang Qiu, Yao Zhao, Jianbo Jiao, Yunchao Wei, and Shikui Wei. 2019. Referring image segmentation by generative adversarial learning. *IEEE Transactions on Multimedia* 22, 5 (2019), 1333–1344.
[144] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
[145] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8821–8831.
[146] Milad Ramezani, Yiduo Wang, Marco Camurri, David Wisth, Matias Mattamala, and Maurice Fallon. 2020. The newer college dataset: Handheld lidar, inertial and vision with ground truth. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4353–4360.
[147] Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. VLC-BERT: Visual question answering with contextualized commonsense knowledge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1155–1165.
[148] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, and Andrew Zisserma. 2023. Zorro: The masked multimodal transformer. arXiv:2301.09595. Retrieved from https://arxiv.org/abs/2301.09595
[149] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. arXiv:1804.02767. Retrieved from https://arxiv.org/abs/1804.02767
[150] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in Neural Information Processing Systems* 28 (2015), 2953–2961.
[151] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.
[152] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.
[153] Gaurav Sahu and Olga Vechtomova. 2019. Adaptive fusion techniques for multimodal data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online, 3156–3166.
[154] Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. *Advances in Neural Information Processing Systems* 33 (2020), 3070–3081.
[155] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end Generative Pretraining for Multimodal Video Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, Los Alamitos, CA, 17938–17947.
[156] Xiang Shen, Dezhi Han, Chin-Chen Chang, and Liang Zong. 2022. Dual self-guided attention with sparse question networks for visual question answering. *IEICE TRANSACTIONS on Information and Systems* 105, 4 (2022), 785–796.
[157] Xiang Shen, Dezhi Han, Zihan Guo, Chongqing Chen, Jie Hua, and Gaofeng Luo. 2023. Local self-attention in transformer for visual question answering. *Applied Intelligence* 53, 13 (2023), 16706–16723.
[158] Gen Shi, Yifan Zhu, Wenjin Liu, and Xuesong Li. 2021. A heterogeneous graph based framework for multimodal neuroimaging fusion learning. arXiv:2110.08465. Retrieved from https://arxiv.org/abs/2110.08465
[159] Nir Shlezinger, Jay Whang, Yonina C. Eldar, and Alexandros G Dimakis. 2020. Model-based deep learning. *Proc. IEEE* 111, 5 (2023), 465–499.
[160] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. arXiv:1804.09626. Retrieved from https://arxiv.org/abs/1804.09626
[161] N. Silberman and R. Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*.
[162] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556
[163] Lovejit Singh. 2022. Deep bi-directional LSTM network with CNN features for human emotion recognition in audio-video signals. *International Journal of Swarm Intelligence* 7, 1 (2022), 110–122.
[164] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 567–576.

[165] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019), 11918–11930.
[166] Sabrina I. Soraya, Shao-Ping Chuang, Yu-Chee Tseng, Tsì-Uí Ík, and Yu-Tai Ching. 2019. A comprehensive multisensor dataset employing RGBD camera, inertial sensor and web camera. In *Proceedings of the 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 1–4.
[167] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Learning video representations using contrastive bidirectional transformer. arXiv:1906.05743. Retrieved from https://arxiv.org/abs/1906.05743
[168] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7464–7473.
[169] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross-and self-attention network for speech emotion recognition. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4275–4279.
[170] Siyang Sun, Xiong Xiong, and Yun Zheng. 2022. Two stage multi-modal modeling for video interaction analysis in deep video understanding challenge. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7040–7044.
[171] Yuxiang Sun, Weixun Zuo, and Ming Liu. 2019. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters* 4, 3 (2019), 2576–2583.
[172] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, 5100–5111.
[173] YunPeng Tan, Fangyu Liu, BoWei Li, Zheng Zhang, and Bo Zhang. 2022. An efficient multi-view multimodal data processing framework for social media popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7200–7204.
[174] Zineng Tang, Jie Lei, and Mohit Bansal. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2415–2426.
[175] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing and Management* 57, 5 (2020), 102277.
[176] Gijs van Tulder and Marleen de Bruijne. 2018. Learning cross-modality representations from multi-modal images. *IEEE Transactions on Medical Imaging* 38, 2 (2018), 638–648.
[177] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008.
[178] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.
[179] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* 11, 1 (2021), 1–13.
[180] Guoqing Wang, Chuanxin Lan, Hu Han, Shiguang Shan, and Xilin Chen. 2019. Multi-modal face presentation attack detection via spatial and channel attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
[181] Huiqing Wang, Jingjing Wang, Chunlin Dong, Yuanyuan Lian, Dan Liu, and Zhiliang Yan. 2020. A novel approach for drug-target interactions prediction based on multimodal deep autoencoder. *Frontiers in Pharmacology* 10 (2020), 1592.
[182] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition* 98 (2020), 107075.
[183] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. 2019. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6212–6221.
[184] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 10 (2017), 2413–2427.
[185] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. 2021. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications* 12, 1 (2021), 1–13.
[186] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.
[187] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems* 33 (2020), 4835–4845.

[188] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 540–547.

[189] Yanan Wang, Jianming Wu, Kazuaki Furumai, Shinya Wada, and Satoshi Kurihara. 2022. VAE-based adversarial multimodal domain transfer for video-level sentiment analysis. *IEEE Access* 10 (2022), 51315–51324.

[190] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2022. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. arXiv:2205.11501. Retrieved from https://arxiv.org/abs/2205.11501

[191] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 618–626.

[192] Haoran Wei, Pranav Chopada, and Nasser Kehtarnavaz. 2020. C-MHAD: Continuous multimodal human action dataset of simultaneous video and inertial sensing. *Sensors* 20, 10 (2020), 2905.

[193] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.

[194] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[195] Michael Wray, Hazel Doughty, and Dima Damen. 2021. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3650–3660.

[196] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 5380–5389.

[197] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10216–10225.

[198] Chunlei Wu, Yiwei Wei, Xiaoliang Chu, Sun Weichen, Fei Su, and Leiquan Wang. 2018. Hierarchical attention-based multimodal fusion for video captioning. *Neurocomputing* 315 (2018), 362–370.

[199] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519. Retrieved from https://arxiv.org/abs/2309.05519

[200] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.

[201] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. VLM: Task-agnostic video-language model pre-training for video understanding. arXiv:2105.09996. Retrieved from https://arxiv.org/abs/2105.09996

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

[203] Kele Xu, Ming Feng, and Weiquan Huang. 2022. Seeing Speech: Magnetic Resonance Imaging-Based Vocal Tract Deformation Visualization Using Cross-Modal Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6947–6949.

[204] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. 2019. Graph Convolutional Network Hashing for Cross-Modal Retrieval. In *Proceedings of the Ijcai*. 982–988.

[205] Zhaoyang Xu, Jinguang Gu, Maofu Liu, Guangyou Zhou, Haidong Fu, and Chen Qiu. 2023. A question-guided multi-hop reasoning graph network for visual question answering. *Information Processing and Management* 60, 2 (2023), 103207.

[206] Enquan Yang, Wujie Zhou, Xiaohong Qian, Jingsheng Lei, and Lu Yu. 2023. DRNet: Dual-stage refinement network with boundary inference for RGB-D semantic segmentation of indoor scenes. *Engineering Applications of Artificial Intelligence* 125 (2023), 106729.

[207] Qi Yang, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. 2022. Personality-Driven Social Multimedia Content Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7290–7299.

[208] Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Ming Yan, Fei Huang, Zhangzikang Li, and Yu Zhang. 2023. Transforming Visual Scene Graphs to Image Captions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, 12427–12440.

[209] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10685–10694.

[210] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 684–699.
[211] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10502–10511.
[212] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[213] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Part XVII 16*. Springer, 229–247.
[214] Chengxiang Yin, Kun Wu, Zhengping Che, Bo Jiang, Zhiyuan Xu, and Jian Tang. 2021. Hierarchical graph attention network for few-shot visual-semantic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2177–2186.
[215] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12226–12234.
[216] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4467–4480.
[217] Jing Yu, Yuhang Lu, Weifeng Zhang, Zengchang Qin, Yanbing Liu, and Yue Hu. 2020. Learning cross-modal correlations by exploring inter-word semantics and stacked co-attention. *Pattern Recognition Letters* 130 (2020), 189–198.
[218] Jing Yu, Chenghao Yang, Zengchang Qin, Zhuoqian Yang, Yue Hu, and Zhiguo Shi. 2019. Semantic modeling of textual relationships in cross-modal retrieval. In *Proceedings of the International Conference on Knowledge Science, Engineering and Management*. Springer, 24–32.
[219] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*. Springer, 69–85.
[220] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems* 29, 12 (2018), 5947–5959.
[221] Yu Yuan, Jiaqi Wu, Zhongliang Jing, Henry Leung, and Han Pan. 2022. Multimodal Image Fusion based on Hybrid CNN-Transformer and Non-local Cross-modal Attention. arXiv:2210.09847. Retrieved from https://arxiv.org/abs/2210.09847
[222] Yuan Yuan, Zhitong Xiong, and Qi Wang. 2019. Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9176–9184.
[223] Zhaoquan Yuan, Siyuan Sun, Lixin Duan, Changsheng Li, Xiao Wu, and Changsheng Xu. 2020. Adversarial multimodal network for movie story question answering. *IEEE Transactions on Multimedia* 23 (2020), 1744–1756.
[224] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 1103–1114.
[225] Yawen Zeng, Yiru Wang, Dongliang Liao, Gongfu Li, Jin Xu, Xiangmin Xu, Bo Liu, and Hong Man. 2024. Contrastive topic-enhanced network for video captioning. *Expert Systems with Applications* 237 (2024), 121601.
[226] Beibei Zhang, Yaqun Fang, Tongwei Ren, and Gangshan Wu. 2022. Multimodal analysis for deep video understanding with video language transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7165–7169.
[227] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion* 76 (2021), 323–336.
[228] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. 2022. Can language understand depth?. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6868–6874.
[229] Shan Zhang, Baocheng Geng, Pramod K. Varshney, and Muralidhar Rangaswamy. 2019. Fusion of deep neural networks for activity recognition: A regular vine copula based approach. In *Proceedings of the 2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 1–7.
[230] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. 2020. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2, 2 (2020), 182–193.
[231] Shan Zhang, Pranay Sharma, Baocheng Geng, and Pramod K. Varshney. 2022. Distributed estimation in large scale wireless sensor networks via a two step group-based approach. arXiv:2203.09567. Retrieved from https://arxiv.org/abs/2203.09567
[232] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z. Li. 2019. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 919–928.

[233] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 286–301.

[234] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. 2019. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. In *Proceedings of the VISIGRAPP (5: VISAPP)*. 336–343.

[235] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. 2021. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing* 105 (2021), 104042.

[236] Yang Zhang, Chenyun Xiong, Junjie Liu, Xuhui Ye, and Guodong Sun. 2023. Spatial-information guided adaptive context-aware network for efficient rgb-d semantic segmentation. *IEEE Sensors Journal* 23, 19 (2023), 23512–23521.

[237] Dexin Zhao, Zhi Chang, and Shutao Guo. 2019. A multimodal fusion approach for image captioning. *Neurocomputing* 329 (2019), 476–485.

[238] Qiankun Zhao, Yingcai Wan, Jiqian Xu, and Lijin Fang. 2023. Cross-modal attention fusion network for RGB-D semantic segmentation. *Neurocomputing* 548 (2023), 126389.

[239] Xiaolei Zhao, Jing Zhang, Jimiao Tian, Li Zhuo, and Jie Zhang. 2020. Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. *Remote Sensing* 12, 11 (2020), 1887.

[240] Zhijia Zheng, Donghan Xie, Chunlin Chen, and Zhangqing Zhu. 2020. Multi-resolution cascaded network with depth-similar residual module for real-time semantic segmentation on RGB-D images. In *Proceedings of the 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 1–6.

[241] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.

[242] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 7590–7598.

[243] Tongxue Zhou, Su Ruan, and Stéphane Canu. 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* 3 (2019), 100004.

[244] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8746–8755.

Received 6 November 2022; revised 25 October 2023; accepted 31 January 2024