Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053204

REVIEW

A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications

Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng, Yuming Chen and Jie Song*

Software College, Northeastern University, Shenyang, 110819, China

*Corresponding Author: Jie Song. Email: songjie@mail.neu.edu.cn

Received: 27 April 2024 Accepted: 17 June 2024 Published: 18 July 2024

ABSTRACT

Multi-modal fusion technology gradually become a fundamental task in many fields, such as autonomous driving, smart healthcare, sentiment analysis, and human-computer interaction. It is rapidly becoming the dominant research due to its powerful perception and judgment capabilities. Under complex scenes, multi-modal fusion technology utilizes the complementary characteristics of multiple data streams to fuse different data types and achieve more accurate predictions. However, achieving outstanding performance is challenging because of equipment performance limitations, missing information, and data noise. This paper comprehensively reviews existing methods based on multi-modal fusion techniques and completes a detailed and in-depth analysis. According to the data fusion stage, multi-modal fusion has four primary methods: early fusion, deep fusion, late fusion, and hybrid fusion. The paper surveys the three major multi-modal fusion technologies that can significantly enhance the effect of data fusion and further explore the applications of multi-modal fusion technology in various fields. Finally, it discusses the challenges and explores potential research opportunities. Multi-modal tasks still need intensive study because of data heterogeneity and quality. Preserving complementary information and eliminating redundant information between modalities is critical in multi-modal technology. Invalid data fusion methods may introduce extra noise and lead to worse results. This paper provides a comprehensive and detailed summary in response to these challenges.

KEYWORDS

Multi-modal fusion; representation; translation; alignment; deep learning; comparative analysis

1 Introduction

In the real world, various modal information exists from the external environment and is interrelated with each other to form a whole. Sources of multi-modal information include text, images, video, audio, sensors, and so on [1]. When a method utilizes several data types to solve the problem, it is a multi-modal method. Compared with the unimodal method, the multi-modal fusion method can effectively utilize the complementary characteristics of multiple data streams to reduce the error caused by poor data quality and the data noise between modalities, and it has rapidly become a research hotspot in many fields. The main problem in various fields is how to fuse multi-modal data more accurately and effectively and obtain better prediction accuracy. Although recent research

demonstrates the benefits of fusing multi-modal data in different applications [2], fast and effective multi-modal detection in real-world and complex environments is still challenging.

Multi-modal fusion technology has been applied in many fields, including autonomous driving, smart healthcare, sentiment analysis, data security, human-computer interaction, and other applications [3,4]. For example, automatic driving vehicles are usually equipped with a set of sensors, such as cameras and Light Detection and Ranging (LiDAR), to alleviate the perception difficulties of the automatic driving system. Automatic driving vehicles can capture scenes with overlapping perspectives to minimize visual blind spots by fusing the above multiple sensor data [5]. In smart healthcare, smart healthcare systems frequently fuse multi-modal medical signals to provide a more accurate medical diagnosis in most cases due to the complexity of diseases [6]. In the above scenarios, unimodal methods are difficult to provide precise detection. First, each modal data has its inherent shortcomings and limitations. For example, the data generated by the camera lacks depth information but has high pixels. Although the LiDAR data has the depth information, the resolution is low. It cannot identify long-distance data. Second, the unimodal task is not robust when a sensor fails or is blocked by an object, while the multi-modal task can solve the above problem. Fig. 1 shows the schematic diagram of unimodal and multi-modal architectures, respectively. Unimodal tasks only use a single data type as the model's input, while multi-modal tasks use two or more modal data as the input. The success of multi-modal fusion technology in the above fields depends on the inherent properties of different modal data and the high correlation and complementarity between them. When the unimodal information is lost, another modality can supplement the missing information to obtain a more accurate detection result.

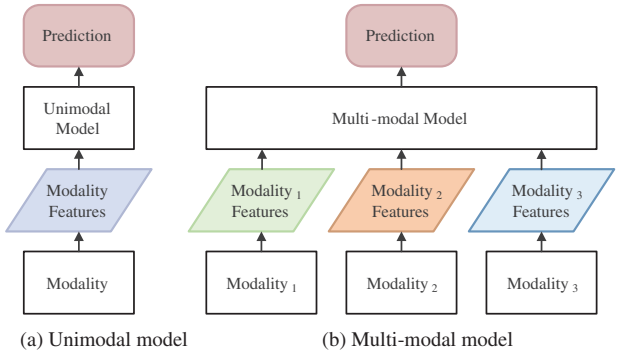


Figure 1: Model architecture

In recent years, researchers have proposed promising studies based on multi-modal fusion technology. The early application can be traced back to an audio-visual speech task proposed in 1989 [7]. When the speech signal is degraded because of data noise, this method uses a neural network to extract helpful information from the visual image and improves speech perception. Traditional multi-modal methods mainly utilize the complementarity criterion and strong correlation between modalities to maximize the consistency of different modal data. The most representative methods are the co-training, the co-regularization, and related derived methods [8]. Compared with the early traditional multi-modal methods, most current research enhances the modal fusion effect through learning-based methods. The most popular method is based on deep learning, which is used for multi-level abstract representation of data. Deep learning can accurately capture the characteristics of multi-modal data and the complementary correlation between modalities. These deep learning-based methods deal with generative and discriminant tasks by supervised and unsupervised training strategies and have made significant progress in multi-modal methods. According to different fusion

stages, the multi-modal fusion method has four main methods: early fusion, deep fusion, late fusion, and hybrid fusion. Early and deep fusion are widely accepted since they can better retain the rich information between modalities.

Although multi-modal fusion technology has achieved promising research results in many fields, there are still the following challenges: 1) Multi-modal methods can solve most unimodal problems, such as target occlusion and weather change, but it is hard to effectively utilize the information of each modality due to the heterogeneity of multi-modal data. Invalid data fusion methods may introduce extra noise and lead to worse results. 2) Due to the change in harsh environments or equipment performance limitations, the data collected by different sensors are not synchronized in a temporal or space domain. It is hard to collect data at the same time because the acquisition cycle of each sensor is independent, and sensors have different perspectives when deployed. 3) The Multi-modal model has more free weights than the unimodal model for capturing the data feature structure in the learning-based methods, which results in excessive inference time for the training model. In order to address the above problems, current research has proposed many fusion techniques, such as representation, translation, alignment, co-learning, reasoning, generation, and so on [8,9]. Representation, translation, and alignment are the three most widely used core fusion technologies for multi-modal tasks, which are detailed in this paper.

The existing survey on multi-modal fusion mainly focuses on introducing the development of frontier technologies in their respective fields [6,10]. This is because multi-modal technology relies on specific scenes and data quality. Unlike existing research, the survey describes multi-modal technology from a new perspective. It explores common issues in various fields and expands on unknown fields. The survey is more concerned with general multi-modal techniques and existing challenges. It provides a comprehensive investigation for multi-modal research and focuses on multi-modal fusion methods, techniques, and applications. Fig. 2 shows the structure of the survey paper. The survey discusses the most up-to-date multi-modal fusion techniques based on deep learning. It provides an applicability analysis of multi-modal technology in multiple application scenes by consulting plenty of literature. The survey mainly addresses three key issues: 1) What are the multi-modal fusion methods, and how are they different? 2) What are the multi-modal fusion techniques, and how do they work? 3) Which multi-modal method should be selected for the best results in different application contexts? Finally, it introduces existing challenges, future research directions, and potential solutions in multi-modal fusion technology. The main contributions of this survey are as follows:

- The survey conducts a detailed analysis of multi-modal fusion techniques and focuses on deep learning-based methods. It discusses the following four fusion stages: early fusion, deep fusion, late fusion, and hybrid fusion. It also analyzes the applicability of multi-modal fusion methods.
- The survey discusses three fusion techniques in the multi-modal field, including data representation, data mapping, and data alignment.
- The survey introduces popular multi-modal datasets in different fields and compares existing research. In addition, it discusses a series of open challenges and potential solutions in detail.

The organization of this paper is as follows: Section 2 describes the fusion methods, and Section 3 details the multi-modal fusion data. Section 4 explores multi-modal fusion techniques that can enhance the effectiveness of data fusion. Section 5 introduces relevant multi-modal applications. Finally, Section 6 discusses existing challenges and potential solutions, and Section 7 provides a conclusion.

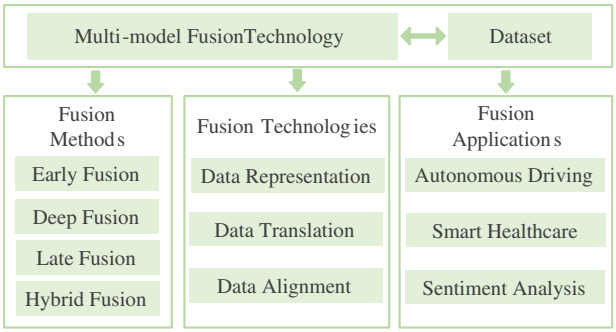


Figure 2: The structure of the survey paper

2 Fusion Methods

This section introduces the multi-modal fusion methods. The taxonomy includes early fusion, deep fusion, late fusion, and hybrid fusion, which are classified by the data fusion stage. This section analyzes the model structure and the advantages or disadvantages of each model.

Early fusion is the data level fusion and usually occurs in the input stage of each branch. The raw data is mapped to the same space through data alignment and translation technology, and then multi-modal data is fused to obtain richer and more expressive data forms. The early fusion method can quickly establish the corresponding relationship between modalities and effectively utilize the valuable information from multiple modalities, but it has more computing requirements. Fig. 3 shows the early fusion architecture, taking image and voice fusion as an example. In particular, several studies have also included deep fusion in early fusion [11]. For a network with $L+1$ layer, Eq. (1) describes the early fusion method, where M_i and M_j represent two different modalities, f_l stands for the feature mapping of a neural network at layer l, $l \in \{1, 2, \ldots, L\}$, $f_l^{M_i}$ and $f_l^{M_j}$ are the feature mapping of the two modalities M_i and M_j in the l layer of the neural network, respectively. $T_l(\cdot)$ represents the feature transformation function in the neural network layer l. Let $f_{l+1} = f_l^{M_i} \oplus f_l^{M_j}$, where $f_l^{M_i} \oplus f_l^{M_j}$ represents data fusion operations.

$$f_L = T_L\left(T_{L-1}\left(\cdots T_l\left(\cdots T_2\left(T_1\left(f_0^{M_i} \oplus f_0^{M_j}\right)\right)\right)\right)\right) \tag{1}$$

Deep fusion occurs in the feature extraction stage. It mixes the multi-modal data in the feature space to obtain the fusion features, compensates for the missing features by other modalities, and then applies fusion features to perform classification or regression tasks in the prediction stage. The fine granularity of the deep fusion method is coarser than the early fusion method [12]. Thus, the deep fusion method can reduce equipment performance requirements compared with the early fusion method. However, it has dimension explosion problems. When the feature dimension reaches a particular scale, the model's performance will decline, and information loss will increase. As shown in Fig. 4, the stage of deep fusion occurs in the backbone network. The deep fusion takes place at layer l, Eq. (2) describes the deep fusion method.

$$f_L = T_L\left(\cdots T_{l+1}\left(T_l^{M_i}\left(\cdots T_1^{M_i}\left(f_0^{M_i}\right)\right) \oplus T_l^{M_j}\left(\cdots T_1^{M_j}\left(f_0^{M_j}\right)\right)\right)\right) \tag{2}$$

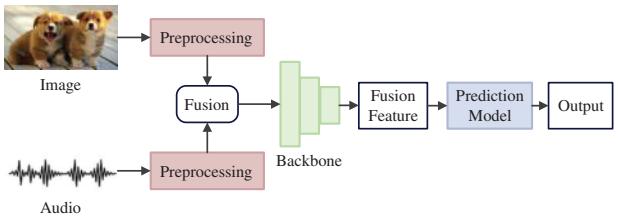


Figure 3: An example of early fusion

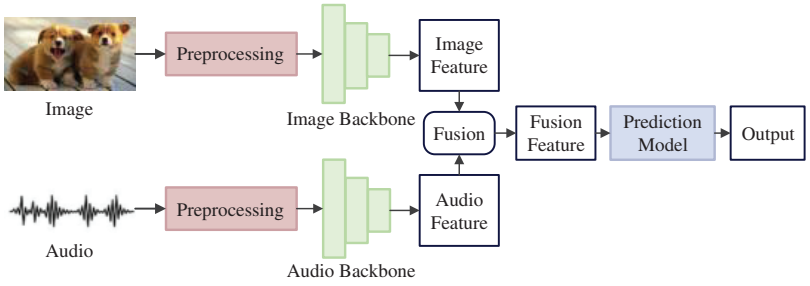


Figure 4: An example of deep fusion

Late fusion is a decision-level fusion and occurs in the prediction stage. Each modality has its separate branch for decision, and the final decision is made by fusing all the output of the decision level. The late fusion can effectively utilize the network decision information of each modality branch without considering the raw data fusion problem. However, the late fusion employs unimodal information to predict the results. Such information may be accidentally skipped or falsely detected under certain conditions due to the limitation of the data itself. In Fig. 5, the late fusion assigns a separate network branch to each modality and fuses the respective prediction results. Late fusion is an integration method that uses multi-modal information to optimize the final proposal. Eq. (3) describes the late fusion method.

$$f_L = T_L^{M_i}\left(T_{L-1}^{M_i}\left(\cdots T_1^{M_i}\left(f_0^{M_i}\right)\right)\right) \oplus T_L^{M_j}\left(T_{L-1}^{M_j}\left(\cdots T_1^{M_j}\left(f_0^{M_j}\right)\right)\right) \tag{3}$$

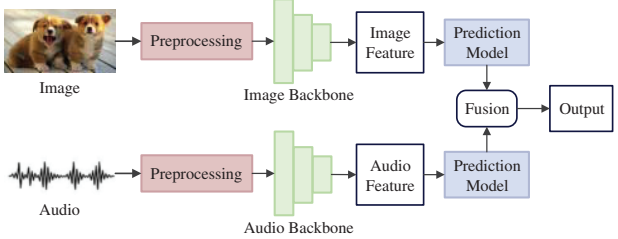


Figure 5: An example of late fusion

Hybrid fusion generally fuses the decision-level information of one branch with the data-level or feature-level information from other branches to establish a cascade relationship between multiple modalities. In exceptional cases, the simple connection of multi-modal features or the methods based on a single-level fusion cannot meet high accuracy and robustness, such as data noise and data loss. Therefore, the existing research combines the advantages of early fusion, deep fusion, and late fusion

to propose a hybrid fusion strategy [10]. Hybrid fusion combines the advantages of all three multi-modal fusion methods. It makes up for the defects of the unimodal fusion method, but this method will increase the model structure complexity and training difficulty. As shown in Fig. 6, hybrid fusion is generally dominated by at least one branch, and other modal branches provide auxiliary information to perform the final task.

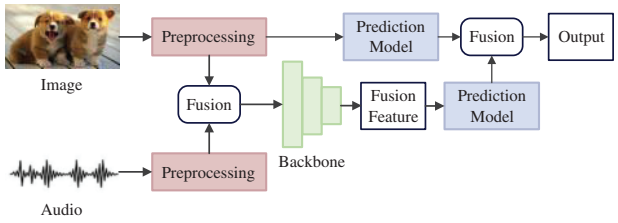


Figure 6: An example of hybrid fusion

Discussion on fusion methods. The above structure is the basic construction of multi-modal network models. In fact, most multi-modal models are more complex than these basic models for better data fusion. The most commonly used multi-modal fusion methods are Convolutional Neural Network (CNN)-based and Transformer-based methods [13]. CNN can effectively prevent over-fitting and reduce the number of parameters during image processing while preserving the original features of the image. Cai et al. [14] add geometry information to the multi-modal model, which maps multi-view image features into the BEV (bird's-eye view) space to fusion multi-view image features. Wu et al. [15] propose a virtual sparse convolution to design a fast yet effective backbone network, VirConvNet (Virtual Sparse Convolution Network). This method can discard large amounts of redundant voxels and tackle the noise problem. Although CNN has many advantages, it only focuses on local information. Thus, CNN is unable to capture long-range dependencies.

In order to address the above problem, the Transformer-based method has received widespread attention because it can utilize long-range dependencies to extract effective features. Transformers solely use the attention mechanism and dispense with recurrence and convolutions entirely. Liu et al. [16] propose a state-of-the-art transformer-based object detector, which applies knowledge distillation and exemplar replay techniques in the multi-modal model. Another multi-modal method combines the advantages of CNN and Transformer by careful design. Rong et al. [17] propose a dynamic-static feature fusion strategy containing two modules: neighborhood cross-attention and dynamic-static interaction. This strategy utilizes a dual pathway architecture to provide rich semantic information. Although this method exacerbates the complexity of multi-modal models, it significantly enhances the model accuracy.

Among the multi-modal methods, early fusion and deep fusion are used more frequently because they have the smallest granularity. The smaller the granularity, the more practical information between multi-modal data is captured. It usually brings more excellent performance. Although early fusion and deep fusion performance is high, it lacks flexibility. When using new multi-modal data to replace the input or expand the number of input channels, they can only retrain all models and require data alignment. To maximize the role of different multi-modal fusion methods, researchers must select appropriate ones according to different application scenes and existing data quality.

3 Fusion Data

This section introduces popular multi-modal datasets classified by different application fields. We have gone through 7000 papers for nearly three years and found the three most widely used fields of multi-modal technology, including autonomous driving, smart healthcare, and sentiment analysis.

Autonomous driving is one of the most widely used areas of multi-modal fusion technology. The automatic driving system uses multiple sensors (such as camera, LiDAR, and radar) to collect raw data. It uses multi-modal fusion technology to fuse the collected data and complete the perception task. In automatic driving, there are many datasets, as shown in Table 1. The most popular used datasets are KITTI[1] [18], Waymo[2] [19], and NuScenes[3] [20]. KITTI is one of the most commonly used object detection datasets in automatic driving, including 2D, 3D, and bird's eye view detection tasks. KITTI has four high-resolution cameras, Velodyne laser scanners, and the most advanced positioning system, collecting 7481 training images, 7518 test images, and corresponding point clouds. In the detection task, KITTI usually uses average precision as the evaluation index for comparison. It has three task levels: easy, mod, and hard. The public dataset of Waymo was collected by five radar sensors and five high-resolution pinhole cameras. The dataset has 798 scenarios for training, 202 scenarios for validation, and 150 scenarios for testing. Waymo has four evaluation indexes: AP/L1, APH/L1, AP/L2, and APH/L2. AP and APH represent two different detection indicators, and L1 and L2 represent objects with different detection difficulties. The NuScenes public dataset contains 1000 driving scenes, 700 for training, 150 for verification, and 150 for testing. Nuscenes must detect ten categories, including traffic cones, bicycles, pedestrians, cars, buses, etc. When calculating the AP, NuScenes uses the measurement based on the center distance instead of the traditional bounding box overlap and uses AP and TP to evaluate the detection performance.

Table 1: Popular multi-modal dataset comparison on autonomous driving

Ref./Dataset/ Year	LiDARs	Cameras	Annotated frames	3D boxes	2D boxes	Traffic scenario	Harsh environment
[18] KITTI 2012	1 Velodyne HDL-64E	2 color, 2 grayscale cameras	15 k	80 k	80 k	Highway, Urban, Suburban	–
[19] Waymo 2019	5 LiDARs	5 high-resolution pinhole cameras	230 k	12 M	9.9 M	Urban, Suburban	Night, Rain
[20] NuScenes 2019	1 Spinning 32-beams LiDRA	6 RGB cameras	40 k	1.4 M	–	Urban, Suburban	Night, Rain
[21] ApolloScape 2018	2 VUX-1HA laser scanners	2 front cameras	144 k	70 k	2.5 M	Highway, Urban, Suburban	Night
[22] A*3D 2020	1 Velodyne HDL-64E 3D-LiDAR	2 color cameras	39 k	230 k	–	Urban	Night, Rain

(Continued)

[1] www.cvlibs.net/datasets/kitti (accessed on 22/04/2024).
[2] http://www.waymo.com/open (accessed on 22/04/2024).
[3] github.com/nutonomy/nuscenes-devkit (accessed on 22/04/2024).

Table 1 (continued)

Ref./Dataset/ Year	LiDARs	Cameras	Annotated frames	3D boxes	2D boxes	Traffic scenario	Harsh environment
[23] PandaSet 2020	1 Mechanical spinning LiDAR 1 Forward-facing LiDAR	5 wide-angle cameras 1 forward-facing long-focus camera	6 k	1 M	-	Urban	Night
[24] Cirrus 2021	2 Lumiar Model H2 LiDARS	1 RGB camera	6 k	100 k	–	Urban	Night
[25] H3D 2019	1 Velodyne HDL-64E	3 color cameras	27 k	1.1 M	–	Urban	–
[26] Argoverse 2019	2 VLP-32 LiDAR	7 high-resolution ring cameras 2 front-facing stereo cameras	22 k	993 k	-	Urban	Night, Rain
[27] ONCE 2021	1 40-beam LiDAR	8 high-resolution cameras	16 k	417 k	769 k	Urban, Suburban	Night, Rain

Smart healthcare uses wearable devices, the Internet of Medical Things (IoMT), and wireless communication technology to diagnose intelligently. Due to the complexity of the disease, multi-modal medical signals are needed for diagnosis in most cases, including Electrocardiogram (ECG), Blood Pressure (BP), Arterial Blood Pressure (ABP), Electroencephalogram (EEG), Electrom-yography (EMG), Magnetic Resonance Imaging (MRI) and so on. Different medical signals have different characteristics to convey human physiology, so fusion of these medical signals can obtain better results than using a single signal. The smart healthcare field has many public datasets, as shown in Table 2. The more commonly used datasets include DEAP[4] [28], SEED[5] [29], and BRATS[6] [30]. DEAP is a multi-modal dataset for analyzing human affective states and detecting mental disease, which records the EEG and peripheral physiological signals of 32 participants by watching 40 one-minute-long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity. SEED included EEG and eye movement data of twelve participants and EEG data of three other subjects. The above datasets can be used to research patients' mental states and stress. BRATS is a large-scale brain multi-modal MR brain tumor segmentation dataset, including 8160 MRI scans of 2040 patients. Each patient contains four modal MR images with T1, T1Gd, T2, and T2-FLAIR. These images are obtained by various clinical protocols and scanners in various medical institutions and are used to develop and test the latest brain tumor segmentation algorithm.

[4] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (accessed on 22/04/2024).
[5] http://bcmi.sjtu.edu.cn/seed/index.html (accessed on 22/04/2024).
[6] http://www.braintumorsegmentation.org/ (accessed on 22/04/2024).

Table 2: Popular multi-modal dataset comparison on smart healthcare

Dataset	Year	Datatype	Scale	Size	Task
DEAP [28]	2012	Imaging	32 participants	–	• Mental state detection
SEED [29]	2015	Imaging	12 participants	–	• Mental state detection
BRATS [30]	2015	Imaging	8160 MRI	–	• Tumor segmentation
TCGA [31]	2015	Genomics, Imaging	33 distinct cancer 11,000 patients	2.5 P	• Cancer research
MedMD [32]	2023	Text, Imaging	Over 5000 distinct diseases	16 M	• Modality recognition • Disease diagnosis • Visual question answering • Report generation
CXR-Mix [33]	2023	Imaging	820,893 chest X-rays	668 k	• Disease diagnosis
PMC-VQA [34]	2023	Imaging, Text	227 k VQA pairs of 149 k images	413 k	• Visual question answering
SLAKE [35]	2021	Imaging	642 images	6 k	• Visual question answering
VinDr-Mammo [36]	2023	Imaging	5000 mammography exams	50 k	• Level assessment • Finding annotations

Sentiment analysis is a new research field that aims to enable intelligent systems to perceive, infer, and understand human sentiment. It has great commercial value, such as generating better marketing strategies [37]. Many studies in this field use physiological signals, voice, text, facial expressions, and body language better to capture human sentiment [38]. We summarized the most popular multi-modal sentiment analysis datasets in recent years, as shown in Table 3. The most popular datasets include POM[7] [39], CMU-MOSEI[8] [40], and CH-SIMS[9] [41]. POM collected 1000 film reviews from ExpoTV. Each film review is a video of the speaker evaluating a specific film and the film rating given. The rating is divided into 1 star to 5 stars, and the average length of the video is about 93 s. The dataset can be used to study the persuasion level of social networks and identify people's speech characteristics. There are 903 videos, 600 for training, 100 for verification, and 203 for testing. CMU-MOSEI is a large dataset consisting of 3228 videos from more than 1000 online YouTube speakers (57% male and 43% female). Every sentence in the dataset is marked as one of eight emotions: strong positive, positive, weakly positive, neutral, weakly negative, negative, and strong negative. The CH-SIMS dataset contains 2281 refined video clips, collecting spontaneous expressions, head poses, occlusion, and lighting from

[7] Researchers interested in the dataset can contact Sunghyun Park (park@ict.usc.edu).
[8] https://www.amir-zadeh.com/datasets (accessed on 22/04/2024).
[9] https://github.com/thuiar/MMSA (accessed on 22/04/2024).

different films, TV series, and variety shows. It has both multi-modal and independent unimodal labels. It allows researchers to study multi-modal sentiment analysis using the interaction between modalities, and it supports the use of independent unimodal labels for unimodal sentiment analysis tasks. The dataset has the following label types: positive, weakly positive, neutral, weakly negative, and negative.

Table 3: Popular multi-modal dataset comparison on sentiment analysis

Dataset	Year	Datatype	Scale	Source	Language	Topics
POM [39]	2016	Verbal, Para-verbal, Visual, audio	1000 videos	ExpoTV	English	Movie reviews
CMU-MOSEI [40]	2018	Text, Visual, Audio	3228 videos	YouTube	English	Reviews, debate, consulting
CH-SIMS [41]	2020	Text, Visual, Audio	60 videos	Movies, TV series, variety shows	Chinese	Spontaneous expressions, various head poses, occlusions, illuminations
YouTube [42]	2011	Text, Visual, Audio	47 videos	YouTube	English	Product reviews
MOSI [43]	2016	Text, Visual, Audio	93 videos	YouTube	English	Opinions, stories, reviews
MuSe-CaR [44]	2021	Text, Visual, Audio	291 videos	YouTube	English	Vehicle review
MELD [45]	2018	Text, Visual, Audio	13,000 utterances from 1433 dialogues	TV series-friends	English	Dialogues from TV series
MEMOTION 2 [46]	2022	Text, Visual	10,000 images	Reddit, Facebok	English	Politics, religion, sports
FACTIFY [47]	2022	Text, Visual	50,000 tweets	Tweeter	English	Politics, governance
WESAD [48]	2018	ECG, EDA, EMG, RESP, TEMP, ACC	15 subjects	Recorded from both a wrist- and a chest-worn device	–	Wearable stress and affect detection

4 Fusion Technologies

This section focuses on three core technologies of multi-modal fusion, which can improve the effect of data fusion, including data representation, translation, and alignment. In order to improve the multi-modal model performance, researchers need to preprocess the raw data using carefully designed methods before the data is input into the model. Reasonable use of the above three technologies can

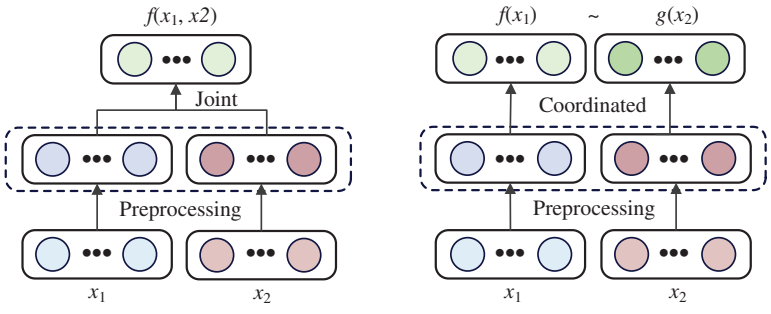
significantly improve the prediction accuracy of the multi-modal model. Finally, we review the current work in detail.

4.1 Data Representation

Cross-modal interaction and complementary information between different modalities are crucial for multi-modal tasks, but the heterogeneity of multi-modal data makes it highly challenging. Current research utilizes multi-modal representation learning to narrow the heterogeneity gap among different modalities, which plays an indispensable role in the multi-modal field. A proper method of data representation should contain essential information of data as much as possible and generate an implicit vector to represent multi-modal information. The high quality of representation learning can retain more practical information, and it helps to complete downstream tasks better. Multi-modal representation learning must consider the data noise between modalities, data loss, real-time, and efficiency. Bengio et al. [49] point out that good representation mainly has several characteristics, including data smoothing, spatiotemporal correlation, data sparsity, natural clust-ering, etc. Compared with unimodal, multi-modal representation learning faces many challenges, including data noise, data heterogeneity, data missing, information redundancy, model complexity, etc. The quality of data representation is crucial to multi-modal problems and is the basis of model training. According to the strategy of integrating different modalities, the survey divides the multi-modal representation models into two frameworks: joint representation and coordinated repress-entation. Table 4 summarizes the advantages and disadvantages of each framework, and Fig. 7 shows the structure of joint and coordinated representation.

Table 4: A summary of the advantages and disadvantages of each framework

Framework	Characteristics	Advantage	Disadvantage
Joint representation	# modalities > # representations	• Integrating information to reduce the number of separate representations	• Cannot infer individual modality
Coordinated representation	# modalities = # representations	• Maximize the cross-modal similarity or correlation • Improving multi-modal contextualization	• Restricted by the number of modalities



(a) Joint representation (b) Coordinated representation

Figure 7: Structure of joint and coordinated representations

Joint representation aims to project all unimodal representation to a shared semantic subspace and fuse multi-modal features. It mainly deals with the task of training and reasoning with multi-modal data. Common methods include addition, multiplication, and splicing [50]. Mainstream methods often utilize neural networks to construct multi-modal joint representations. The deep neural network has multi-layer properties, and each connected layer represents the data more abstractly. Thus, the last or penultimate neural layer is usually used as the data representation. In order to construct a multi-modal representation, each modality needs to start from several separate neural layers and then pass a hidden layer to project modalities in the joint space. Finally, jointing the multi-modal representation can pass multiple hidden layers or predict downstream tasks [51]. Mohammed et al. [52] propose a novel deep multi-modal multi-layer hybrid fusion network (MMHFNet). MMHFNet simultaneously contains complementary information of different modalities, vertically combines low-level features extracted from the shallow layer with high-level features extracted from the deep layer, and fully uses spatial-spectral information of different layers for multi-layer fusion. This method can adaptively generate multi-modal joint feature representation, producing better performance in accuracy and robustness.

Furthermore, joint representation tends to preserve shared semantics while ignoring the specificity information of modality. Thus, joint representation cannot guarantee complementary information and constraint relationships among different modalities at the feature fusion stage. A practical solution is adding extra regularization terms to facilitate the exchange of valuable information [53]. Regularization can discover the hidden correspondences and diversity of the multi-modal features and adjust the weights of the fusion layer dynamically. Following this strategy, promising results have emerged in many multi-modal applications, such as semantic and instance segmentation [54], gesture recognition [55], and image-text retrieval [56].

Coordinated representation learns the individual representations of each modality and coordinates modalities by defining the constraints. The popular coordinated methods are based on cross-modality similarity and cross-modality correlation [57]. Cross-modality similarity is learning a common subspace by directly measuring the distance between the vectors of different modalities. The similarity between two modalities can be calculated by the cosine similarity of two semantic vectors. It can be denoted as Eq. (4), where y_K and y_P are the semantic vectors of two modalities.

$$R(K,P) = \cos(y_K, y_P) = \frac{y_K^{\mathrm{T}} y_P}{\|y_K\| \, \|y_P\|} \tag{4}$$

Cross-modality similarity has a wide range of applications. Wu et al. [58] propose a Focal Modality-Aware Similarity-Preserving Loss method to match pedestrian images across non-overlapping camera views. It learns shared knowledge for cross-modality matching by cross-modality similarity preservation. To further extract shared knowledge, they design a modality-gated node to obtain the universal representation of both modality-specific.

Cross-modality correlation is to learn a shared subspace to maximize the correlation of different modal representation sets. Cai et al. [59] propose a novel unsupervised image fusion network (DCS-Fuse) to fuse infrared and visible images. To reduce information loss, they first learn the modal-specific features of each modality and then calculate the correlation to guide the integration of cross-modal features. Finally, this method utilizes these integrated features to reconstruct the fused image. Coordinated representation can be used in many fields, such as emotion recognition [60]. The advantage of coordinated representation is that each modality can work independently. This characteristic benefits cross-modality transfer learning and delivers information between different modalities. Finally, the comparison of multi-modal representation methods is shown in Table 5.

Table 5: The recent work on multi-modal representation

Ref.	Year	Modalities	Framework	Dataset	Task	Performance
[54]	2024	3D point cloud, Text	Joint representation	S3DIS, SUN RGB-D, and ScanNet	Semantic segmentation, instance segmentation, and object detection	mAP (S3DIS) = 62.7 mAP (SUN RGB-D) = 37 mAP (Scan Net) = 39.7
[55]	2024	Wireless signal, Visual signal		Private dataset	Human gesture recognition	Accuracy (cross-p) = 97.6 Accuracy (cross-s) = 98.1 Accuracy (cross-ps) = 97.4
[56]	2023	Image, Text		MS-COCO	Image-text retrieval problem	Accuracy (sentence) = 98.7 Accuracy (Images) = 94.2
[58]	2020	RGB images, Infrared image	Coordinated representation	SYSU-MM01	Address the RGB-IR cross-modality Re-ID problem	mAP = 44.98
[59]	2023	Infrared image, Visible image		TNO, RoadScene, and M3FD	Infrared and visible image fusion for visual object detection	Qp (TNO) = 0.94 Qp (RoadScene) = 0.93 Qp (M3FD) = 0.95
[60]	2024	EEG, Eye movement signals		SEED-CHN, SEED-GER	Emotion recognition	Accuracy (SEED-CHN) = 94.09 Accuracy (SEED-GER) = 91.62

4.2 Data Translation

Data translation is used to interconvert between modalities. It can supplement the lost information of the current modality with the information of another modality mapping and capture complementary information between different modalities. Multi-modal data translation techno-logies include neural networks, graphical models, and generative adversarial networks. The neural network is a

widely used method because of its learning ability. This method inputs the multi-modal data into the neural network and maps the different modalities to the same semantic space. The graphical model method stands for the data translation of different modalities as a graph structure. It maps the different modalities to the same semantic space by utilizing the propagation and aggregation ability of the graphical model. The generative adversarial network is a newly emerging method used for data translation and uses the confrontation between generator and discriminator to complete the translation process of multi-modal data. The generator is responsible for mapping the data of different modalities to the semantic space. The discriminator is responsible for judging whether the generated data is accurate. Finally, a better multi-modal data translation model is obtained by optimizing the confrontation process between the generator and the discriminator [61].

The modalities generation methods of data translation include a grammar-based model, encoder-decoder, and continuous generation model [9]. *Grammar-based models* rely on predefined syntax to generate specific modalities. They first detect advanced concepts from the source modalities, such as objects in the image and actions in the video. Then, these detection results are combined with the predefined grammar to generate the target modality. Kojima et al. [62] propose a system that detects human head and hand positions and combines detection results with the rule-based natural language generator to describe human behavior in video. Mitchell et al. [63] use more complex tree-based language models to generate syntax trees, which leads to more diverse descriptions. However, this method cannot capture the relationship between spatial and semantics. Thus, Elliott et al. [64] explicitly model the proximity relationship of objects to generate image descriptions. The encoder-decoder method first encodes the source modalities as a potential representation, and then the decoder uses the potential representation to generate the target modalities. The encoder-decoder model is mainly used to generate text but can also generate images and sounds. Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) usually executes the decoder and uses the encoded representation as the initial hidden state. Venugopalan et al. [65] use the pre-trained decoder LSTM for image subtitle generation to improve performance. Rohrbach et al. [66] also explore various LSTM architectures and utilize regularization techniques for video description tasks. *Continuous generation models* are used for sequence conversion. It converts from source to target modalities, generating the output online at each time step. It is mainly for text-to-speech conversion, audio-visual speech generation, and so on [67–69]. Table 6 shows the comparison of multi-modal translation methods.

Table 6: The comparison of modalities generation methods in multi-modal translation

Ref.	Year	Modalities translation	Method	Dataset	Task	Contribute
[70]	2019	Image, Text	Grammar-based	VQA, COCO	Visual question answering	• Propose a novel caption embedding module
[71]	2023	Image, Text		SyViC	Large-scale pre-trained vision & language models	• A million-scale synthetic dataset SyViC

(Continued)

Table 6 (continued)

Ref.	Year	Modalities translation	Method	Dataset	Task	Contribute
[72]	2024	Image, Text, Audio	Encoder-decoder	CMU-MOSI, CMU-MOSEI	Multi-modal sentiment analysis	• Propose a novel Disentanglement Translation Network • Propose a two-step translation method
[73]	2023	Image, Text, Audio		MOSI, MOSEI	Multi-modal sentiment analysis	• Propose a new cross-modal approach • Propose a modality reinforcement cross-attention module and noise-filtering gate module
[67]	2023	SAR and optical images	Continuous generation	SEN12MS-CR	Multi-modal remote sensing image cloud and shadow removal	• Propose a novel hierarchical spectral and structure-preserving fusion network • Propose a deep hierarchical architecture with stacked residual groups
[68]	2023	Image, Text		LRS3-T, CVSS-C	Speech-to-speech translation	• Propose the first textless audio-visual speech-to-speech translation model AV-TranSpeech • Collect a benchmark dataset LRS3-T
[69]	2024	Text, Audio		Private dataset	Speech synthesis	• Propose METTS to synthesizing bilin-gual emotional speech for each monol-ingual speaker

4.3 Data Alignment

Data alignment is used to find the corresponding relationship of elements in different modalities from the same instance. It aligns the data of different modalities in time and space and realizes the information interaction. For example, given an image and a phrase, we need to find the image region corresponding to the phrase. From a data-driven perspective, multi-modal data alignment is used to explore which elements are related, and it is essential for modeling the joint distribution across

modalities. Although multi-modal technology has many advantages, its adverse effects cannot be ignored. Fusing other modalities without constraints will bring data noise. This behavior will reduce the gain of the multi-modal model or even lower than the prediction accuracy of the unimodal model. Therefore, data alignment is one of the core issues in multi-modal research and has a wide range of applications in many fields [74].

Data alignment can be summarized as explicit and implicit alignment [75]. Explicit alignment aims to find the relationship between modalities and is mainly applied to voice-text alignment and image or video positioning. In contrast to explicit alignment, implicit alignment learns how to align the data latently during model training and is usually used as an intermediate step of another target task. It is mainly used in cross-modality retrieval, visual automatic description generation, and visual question answering [76]. Multi-modal alignment transforms the original space of modalities into a multi-modal alignment space with constraints through functional changes. Let E_s and E_t denote the sets of source entities and the corresponding target entities, where $|E_s| = |E_t|$. The multi-modal alignment method aims to find all the aligned pairs P so that the ith entity in E_s corresponds to the ith in E_t, and Eq. (5) describes P.

$$P = \{(e_1, e_2)|e_1 \equiv e_2, e_1 \in E_s, e_2 \in E_t\} \tag{5}$$

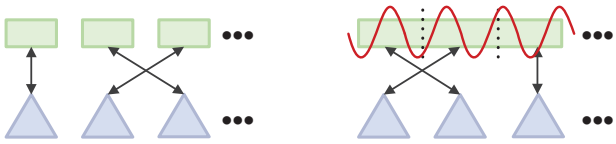
Wang et al. [77] utilize classification techniques and entity types to remove visual noises and compute a similarity matrix for alignment learning. It is denoted as Eq. (6), where $\mathrm{E}_s^{(s)}$ and $\mathrm{E}_t^{(s)}$ stand for the structural embeddings of E_s and E_t, respectively. Sim is the similarity matrix. Sim_{ij} denotes the cosine similarity between the ith entity in E_s and jth in E_t.

$$\mathrm{Sim} = \left\langle \mathrm{E}_s^{(s)}, \mathrm{E}_t^{(s)} \right\rangle \in \mathbb{R}^{|E_s| \times |E_t|} \tag{6}$$

The loss of structural modality is denoted as Eq. (7), where α and β are the temperature scales and N is the batch size.

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{\alpha} \log \left(1 + \sum_{m \neq i} e^{\alpha \mathrm{Sim}_{mi}} \right) + \frac{1}{\alpha} \log \left(1 + \sum_{n \neq i} e^{\alpha \mathrm{Sim}_{in}} \right) - \log \left(1 + \beta \mathrm{Sim}_{ii} \right) \right) \tag{7}$$

We further subdivide data alignment technology into discrete alignment and continuous alignment. The structure of discrete alignment and continuous alignment is shown in Fig. 8. Discrete alignment is mainly used to determine the association between discrete elements in different modalities. It is suitable for multi-modal tasks and can precisely segment the data into discrete elements, such as cross-modality retrieval [78]. Some recent developments have integrated it with neural networks and developed a convex relaxation method for effective learning to ease the computational difficulty [79]. The above methods are used for modal data that can be easily segmented. However, some continuous signals and spatiotemporal data are difficult or impossible to segment. For these indivisible continuous signals, the existing research proposes effective solutions based on adversarial training to solve the continuous alignment problem. For example, Liu et al. [80] present a novel Enhanced Alignment Fusion-Wasserstein Generative Adversarial Network (EAF-WGAN) for turbulent image restoration, and Munro et al. [81] design self-monitoring adversarial alignment methods for multi-modal behavior identity. Clustering methods can also group continuous data based on semantic similarity and cluster continuous original video or audio features into discrete sets [82].



(a) Discrete alignment (b) Continuous alignment

Figure 8: Structure of data alignment

Data alignment has long-term dependence and fuzzy segmentation problems. This is because of the uncertain relationship between modalities. It may be one-to-one, many-to-many, or no corresponding relationship. The data alignment method has the following challenges: 1) the length difference between different modalities. 2) Semantic differences between different modalities. In response to the above challenges, the recent work provides some practical solutions to fill in the missing information and express the subtle difference in semantic information. It can solve the problems of length mismatch and semantic difference [83]. There are also some multi-modal methods to align camera images and point clouds by using attention learning. Chen et al. [84] design a cross-attention feature alignment module to adaptively aggregate pixel-level image features of each voxel. Similarly, Li et al. [85] dynamically capture the correlation between images and LiDAR features by utilizing the cross-attention method in the fusion process. Table 7 shows the comparison of multi-modal data alignment methods.

Table 7: The recent work on multi-modal data alignment

Ref.	Year	Modalities	Network architecture	Dataset	Task	Performance
[74]	2024	Image, Text, Context captions	Transformer-based	MET-Meme, MemeCap	Multimodal emotion recognition	Accuracy (SA) = 29.17 Accuracy (SA) = 72.36 Accuracy (ID) = 44.24
[76]	2023	Image, Text, Question	AKEE	VTQA	Visual text question answering	Accuracy = 60.62
[78]	2024	Text, Molecule	Adversarial network (three fully-connected layers)	ChEBI-20	Cross-modal molecule retrieval	Hits10 = 92.1
[80]	2023	Spatiotemporal data	SNAF	Private dataset	Turbulent image restoration	SSIM = 90.47 VSI = 98.76 FSIM = 94.23

(Continued)

Table 7 (continued)

Ref.	Year	Modalities	Network architecture	Dataset	Task	Performance
[85]	2022	LiDAR point clouds, Image	LearnableAlign (three fully-connected layers)	Waymo	3D object detection	AP/L1 = 84.3

5 Fusion Applications

Multi-modal fusion technology is critical in many fields because of its rich information expression, such as autonomous driving, smart healthcare, sentiment analysis, human-computer interaction, intelligent education, etc. Fig. 9 shows the multiple fields using multi-modal techn-ology, where the number of researches is represented by bubble size. We can observe from Fig. 9 that the number of multi-modal researches has gradually increased in recent years. Then, we select the three most widely used fields to introduce the related applications of multi-modal fusion technology, including autonomous driving, smart healthcare, and sentiment analysis. First, we introduce the main tasks of the three fields respectively and then analyze the methods based on multi-modal fusion technology in each field.

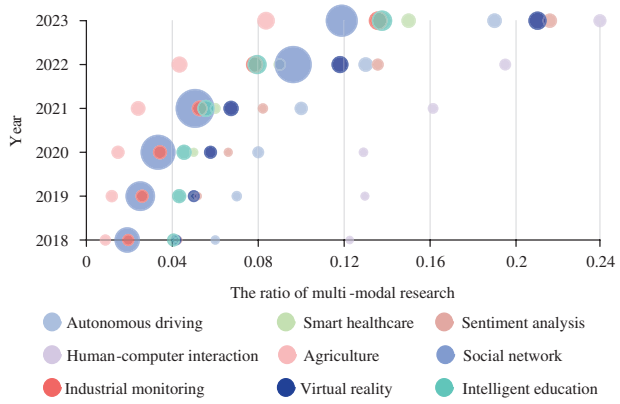


Figure 9: The summary of multi-modal applications in recent years

5.1 Autonomous Driving

Multi-modal fusion technology has made rapid progress in the perception task of autonomous driving [86]. The typical architecture for an autonomous driving system is shown in Fig. 10, consisting of three parts: Perception, Decision-making, and Performance. In order to ensure a solid and accurate perception of the environment, autonomous vehicles are usually equipped with a set of sensors (such as cameras and LiDAR). Multiple sensors are hoped to capture scenes with overlapping perspectives to minimize blind spots. Most existing methods utilize the point cloud or image data captured by LiDAR and camera to handle perception tasks, and they have achieved some results [87]. However, many studies have shown that fusing multiple data streams can obtain more significant performance advantages because of the complementary characteristics between data [88].

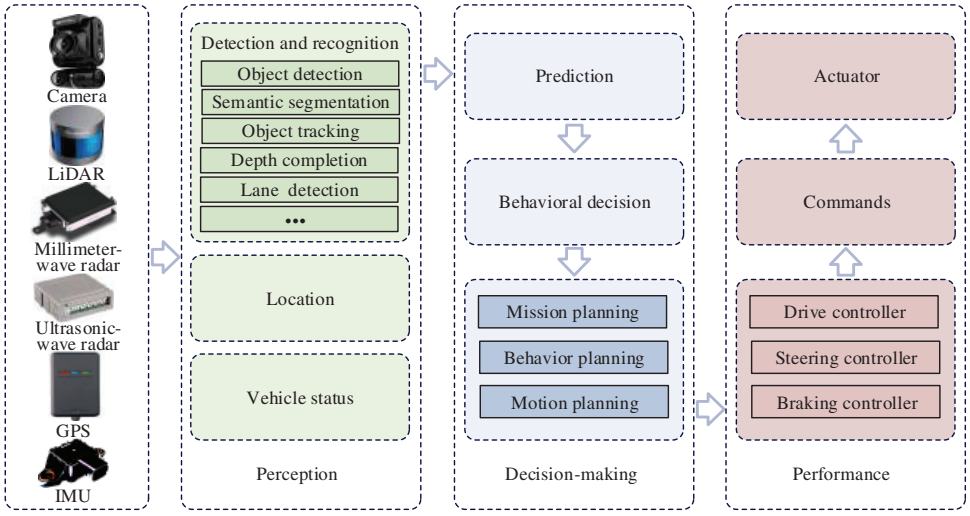


Figure 10: The typical architecture for an autonomous driving system

In autonomous driving, multi-modal fusion technology is mainly used for object detection, semantic segmentation, object tracking, and depth completion. Object detection and semantic segmentation are the most common tasks. Object detection is a traditional computer vision task that aims to locate, classify, and estimate directional boundary boxes in 3D space [89]. Identifying targets include cars, pedestrians, traffic lights, and road markings. 3D object detection is used to predict object properties, including locations, sizes, categories, etc. It can be described as Eq. (8), where $\alpha(\cdot)$ stands for a set with object states in a frame scenario, and $o_1, o_2, \ldots, o_n$ are the 3D objects. $T_{det}(\cdot)$ is the 3D detection function, and β is input data from the sensor.

$$\alpha\left(o_1, o_2, \ldots, o_n\right) = T_{det}\left(\beta\right) \tag{8}$$

According to the existing research, semantic segmentation can be divided into 2D/3D semantic and instance segmentation. 2D/3D semantic segmentation aims to predict the class label for per-pixel and per-point. Instance segmentation jointly performs semantic segmentation and object detection and expands the semantic segmentation task by distinguishing the categories of individual instances. Semantic segmentation can deal with remote sensing data of autonomous driving, such as light detection and ranging [90]. Object tracking is used to locate an object in continuous data frames [91]. Object tracking can deal with single-object and multiple-object tracking tasks and is often used in the decision-making of autonomous driving vehicles. The purpose of depth completion is to up-sample sparse irregular depth into dense regular depth. Depth completion can reduce the violent uneven distribution of scanning points in LiDAR and is helpful to downstream perception tasks.

Deep neural networks can extract appearance and geometric features from the original image. It is the most commonly used multi-modal data fusion method in autonomous driving. Vora et al. [11] fuse the semantic features of image and LiDAR points to achieve better performance in the object detection task. Meyer et al. [92] project 3D LiDAR point clouds feature into 2D images feature and utilize CNN to fuse feature representations. Although this method is efficient and cost-effective, this conversion is geometrically lossy, making it less effective for tasks that focus on scene geometry. On the contrary, Wang et al. [93] directly fuse the pseudo point cloud generated by the image branch and the original point cloud of the LiDAR branch to improve the accuracy of object detection. Although this

camera-to-LiDAR projection has better environmental perception ability and higher robustness, this conversion is semantically lossy. Thus, many methods adopt the bird's-eye view (BEV) as the unified representation of fusion [94]. BEV preserves both the geometric structure of LiDAR features and the semantic density of camera features. Unlike the above methods, Huang et al. [95] use a cascaded method to fuse features. They make good use of both original and high-level semantic information.

Most of the sensors are vulnerable to severe weather. There are several studies specifically used to combat rainstorms, fog, and other extreme weather. These methods improve the robustness of autonomous driving systems [96]. RadarNet is an early fusion method to learn the joint representation of radar and LiDAR data for 3D object detection [97]. However, radar performance is affected by adverse weather conditions, which leads to a sharp decline in prediction accuracy. Qian et al. [98] utilize complementary LiDAR to solve this problem, which is less affected by the weather. They proposed a two-stage deep fusion detector to improve the overall detection results. Many studies also propose a late fusion method, which fuses the output of LiDAR point cloud and camera image branches to make final predictions [99]. The survey summarizes some methods that ranked high in the object detection task of KITTI benchmarking in Table 8. The table shows that multi-modal research is not competitive with unimodal research in easy tasks but performs well in hard tasks. The accuracy of the multi-modal method is 79.39%, ranked first. Therefore, researchers should design appropriate multi-modal models to improve prediction accuracy in a more complex environment.

Table 8: 3D object detection on KITTI cars

Method	Year	Fusion stage	Car			GPU	Multi-modal
			Easy	Mod	Hard		
PI-RCNN [88]	2020	Early fusion	84.37	74.82	70.3	TITAN RTX	Yes
EPNet [95]	2020	Deep fusion	89.81	79.28	74.59	TITAN Xp	Yes
CLOCs [99]	2020	Late fusion	88.94	80.67	77.15	–	Yes
PFF3D [100]	2021	Early fusion	81.11	72.93	67.24	–	Yes
3D-CVF [101]	2020	Deep fusion	89.20	80.50	73.11	GTX 1080Ti	Yes
3D DualFusion [102]	2022	Deep fusion	91.01	82.40	**79.39**	–	Yes
MVX-Net (PF) [103]	2019	Early fusion	83.20	72.70	65.20	–	Yes
MMF [104]	2019	Deep fusion	86.81	76.75	68.41	–	Yes
RoIFusion [105]	2021	Deep fusion	88.09	79.36	72.51	GTX 1080Ti	Yes
GLENet-VR [106]	2022	–	**91.67**	**83.23**	78.43	NVIDIA GeForce RTX 2080Ti	No
SE-SSD [107]	2021	–	91.49	82.54	77.15	TITAN Xp	No

5.2 Smart Healthcare

A smart healthcare system involves many directions, including disease control and detection, evaluation and care, healthcare administration, patient decision-making, and medical science [108–110]. It can intelligently respond to the needs of the health environment. Some hospitals have begun to use smart beds to sense the state of patients and dynamically adjust the correct angle and posture. It can provide adequate care without nursing staff. Diseases are complex and often lead to overlapping symptoms, so multi-modal medical signal fusion plays a vital role.

Multi-modal data in the medical environment include electronic health records (EHRs), medical imaging, wearable devices, genome data, sensor data, environmental data, and behavior data. As shown in Fig. 11, medical image classification includes radiology, microscopy imaging, and visible light imaging. Multi-modal technology allows real-time measurement and analysis of multiple signals and considers different aspects of human physiology. It improves the perception experience and allows missing data to be filled out while providing accurate disease detection and prediction insights for customers and medical professionals.

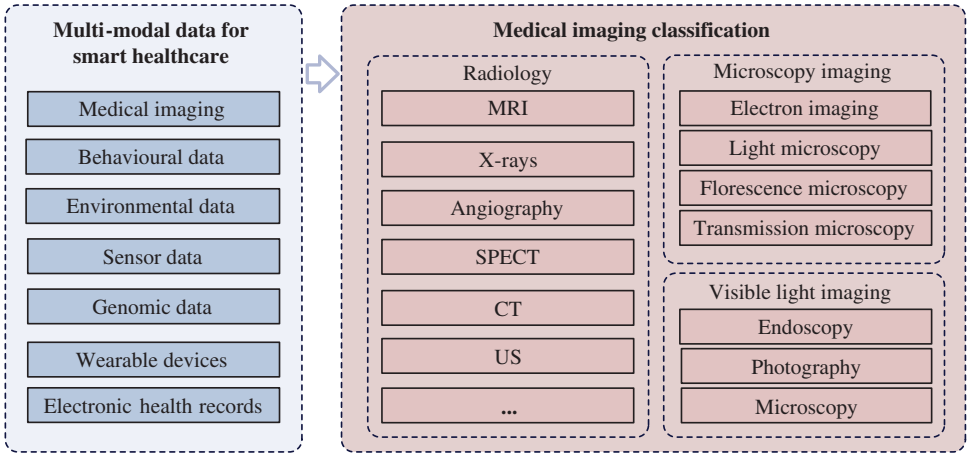


Figure 11: Multi-modal data for smart healthcare

Multi-modal technology has a wide range of applications in the medical field, including medical image recognition, health monitoring, natural language processing (NLP), and disease diagnosis and treatment. Medical image recognition has always been an essential topic in medicine, and it mainly carries out medical diagnoses using different medical images, such as nuclear magnetic resonance, X-ray, angiography, and ultrasound. Health monitoring monitors and analyzes the human body through various types of data and can analyze a more comprehensive and accurate health status. NLP mainly deals with clinical notes, reports, and records in the medical field. Medical personnel can comprehensively understand the health status of patients and enhance personalized treatment plans by incorporating NLP into the data fusion process. The multi-modal fusion method can provide more comprehensive and targeted diagnostic results in disease diagnosis and treatment.

In [111], the author proposes an advanced multi-modal medical image fusion strategy by combining non-subsampled contourlet transform (NSCT) and stationary wavelet transform (SWT) techniques, which assist radiologists in performing surgeries. However, this method will result in higher feature dimensions and make training the model more difficult. To address the above issue, Albahri et al. [112] emphasize the role of feature selection in effective decision-making and improving patient care. It can reduce dimensions and improve the accuracy of data fusion by identifying the most relevant features. Similarly, Alghwinem et al. [113] utilize feature selection techniques to extract the most relevant information from various data sources, and healthcare professionals can obtain a more comprehensive understanding of the patient's health status. This method emphasizes the interpretability of feature selection. Myronenko et al. [114] propose a 3D multi-modal brain tumor segmentation network that integrates a variational auto-encoder branch into a decoder. This method effectively solves the problem of insufficient training data. However, the training data is not always available in clinical practice. Thus, Zhou et al. [115] propose a novel brain tumor segmentation

network, which utilizes available modalities to generate 3D feature-enhanced images of missing modalities in the absence of one or more modalities. To address the ambiguity between categories in brain tumor segmentation, Liu et al. [116] introduce a context-aware network called CANet, which captures high-dimensional and discriminative features with context from convolutional space and feature interaction maps. Furthermore, they propose a context-guided attention conditional random field to fuse features selectively. Table 9 lists the multi-modal medical data fusion methods that have ranked high in the benchmark in recent years.

Table 9: The recent work on multi-modal medical data fusion

Method	Year	Fusion stage	Task	Dataset	Accuracy
MedVInT [33]	2023	Deep fusion	Medical visual question answering	PMC-VQA	42.3
SubOmiEmbed [117]	2022	Early fusion	Cancer type classification	TCGA	96.3
Open-Flamingo [118]	2022	Deep fusion	Medical visual question answering	PMC-VQA	26.4
M2I2 [119]	2022	Deep fusion	Medical visual question answering	SLAKE	81.2
BiomedGPT [120]	2023	Deep fusion	Medical visual question answering	SLAKE	86.1
BiomedCLIP [121]	2023	Deep fusion	Medical visual question answering	SLAKE	85.4

5.3 Sentiment Analysis

Sentiment analysis is an emerging field that has received significant concern, and it prefers technology applications compared with autonomous driving and smart healthcare. It is applied in many scenarios, such as public opinion analysis, psychological disease analysis, social media monitoring, etc. Sentiment analysis is divided into narrative and interactive scenarios, as shown in Fig. 12, and it aims to reveal people's views, positions, or attitudes toward a topic, person, or entity. Compared with the unimodal method, multi-modal sentiment analysis contains richer information, such as text, visual, auditory, and physiological signals. It can infer the implied sentiment polarity more accurately, like irony and exaggeration. In 2015, the multi-modal sentiment analysis survey report showed that 85% of multi-modal systems are always more accurate than unimodal systems, with an average increase of 9.83% [122]. Sentiment analysis includes sentiment polarity analysis, sentiment category analysis, and emotion degree analysis. Sentiment polarity analysis is one of the most basic tasks in multi-modal sentiment analysis.

Perez Rosas et al. [123] combine the features collected from all multi-modal data into feature vectors to generate a vector for each sentence, and they use the support vector machine (SVM) classifier to determine the sentiment classification of discourse. Like the above research, Zadeh et al. [124] propose a tensor fusion network model, which can learn the dynamic changes within and between modalities using an end-to-end method. This model uses an LSTM network with a forgetting gate to learn time-dependent language representation. The multi-modal tensors are input into the sentiment reasoning subnetwork to obtain the prediction results. Hu et al. [125] fuse multi-modal data at the

syntactic and semantic levels and introduce comparative learning better to capture the differences and consistency between sentiments and emotions.

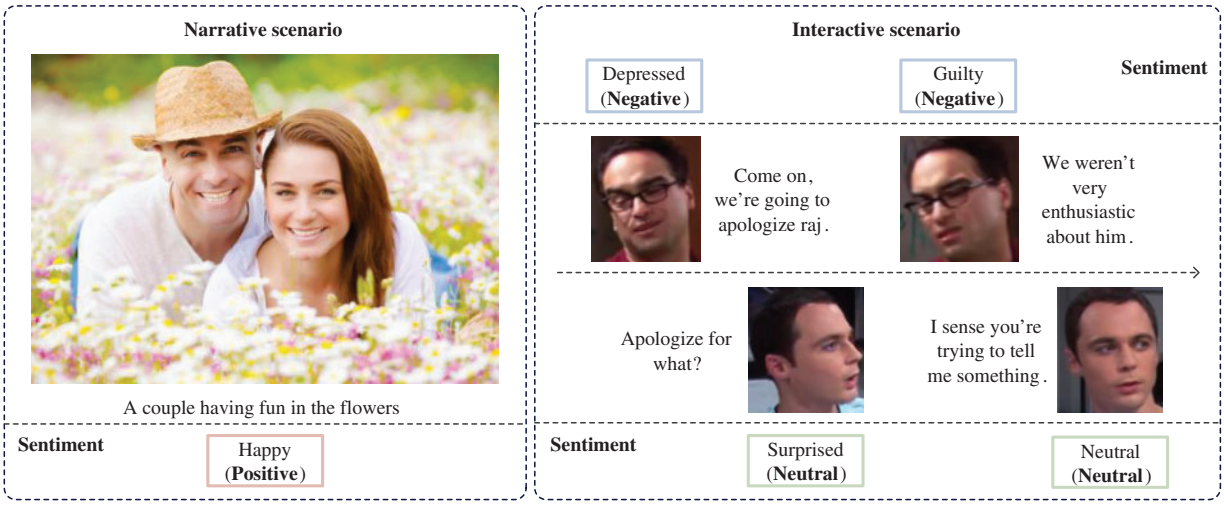


Figure 12: Narrative and interactive scenarios for multi-modal sentiment analysis

In order to reduce the computational complexity, Liu et al. [126] propose a method called Low-rank Multi-modal Fusion (LMF), which uses low-rank tensors to perform multi-modal fusion. It decomposes weights into low-rank factors and reduces the number of parameters. Han et al. [127] divide the model into fusion and Mi maximization. The MI maximization module generates MI-related losses to assist the fusion module in promoting the fusion effect and further improving the accuracy of task prediction. Yu et al. [128] propose a self-supervised Self-MM network, which divides the multi-modal sentiment analysis task into a multi-modal task and three independent unimodal subtasks. Self-MM designs a label generation module based on a self-monitoring learning strategy to obtain independent unimodal monitoring and adopt the sharing strategy to achieve a bottom representation learning network for multi-modal tasks and different unimodal tasks. Nojavanasghari et al. [129] adopt a late fusion method to train an unimodal classifier for the three modalities and then average the confidence scores of each unimodal classifier for final prediction. We compare the results of several methods in Table 10.

Table 10: The recent multi-modal fusion methods on sentiment analysis

Method	Year	Fusion stage	Dataset	Contribute	Performance
UniMSE [126]	2022	Deep fusion	MOSI	• Propose a multi-modal sentiment knowledge-sharing framework • Introduce contrastive learning between modalities and samples	Accuracy = 86.90

(Continued)

Table 10 (continued)

Method	Year	Fusion stage	Dataset	Contribute	Performance
MMIM [128]	2021	Hybrid fusion	CMU-MOSI	• Propose a hierarchical MI maximization framework for multi-modal sentiment analysis	F1 = 84.00
SPECTRA [130]	2023	Early fusion	MOSI	• Propose the first-ever speech-text dialog pre-training model for spoken dialog understanding • Design a novel temporal position prediction task to capture the speech-text alignment	Accuracy = 87.50
MMML [131]	2023	Deep fusion	CH-SIMS	• Compare different fusion methods • Examine the impact of multi-loss training within the multi-modality fusion network	F1 = 82.90
SeMUL-PCD [132]	2023	Deep fusion	CMU-MOSEI	• Propose a Multi-modal Distillation Loss calibrates the fusion network	Accuracy = 88.62
MMLatch [133]	2022	Deep fusion	CMU-MOSEI	• Propose a neural architecture for captures top-down cross-modal interactions	Accuracy = 82.40
ALMT [134]	2023	Deep fusion	CH-SIMS	• The first time explicitly tackles the adverse effects of redundant and conflicting information in auxiliary modalities	F1 = 81.57

(Continued)

Table 10 (continued)

Method	Year	Fusion stage	Dataset	Contribute	Performance
VAE-AMDT [135]	2022	Deep fusion	CMU-MOSI	• Devise a novel Adaptive Hyper-modality Learning module for representation learning • Propose a VAE-based adversarial multi-modal domain transfer for reduce the distance difference between unimodal representations	F1 = 84.20

5.4 Discussion

In addition to traditional models, multi-modal techniques also play an essential role in Large Language Models (LLMs). LLMs have larger scales, wider application scenes, stronger processing capabilities, and higher prediction accuracy than traditional models. In the past year, MultiModal Large Language Models (MM-LLMs) have undergone substantial advancements [136]. It uses cost-effective training strategies to support the inputs or outputs of multi-modal data and preserve the inherent reasoning and decision-making capabilities of LLMs. Several researchers extend LLM to images and then obtain Large Vision Language Models (LVLMs). CLIP is a vital research achievement of LVLMs [137]. This model uses contrastive learning methods to represent images and text in the same embedding space and performs excellently in multiple tasks. Ramesh et al. [138] propose the DALL-E model, which adopts the Transformer architecture and can effectively learn and express the semantic relationship between text and images. The quality and diversity of the generated images are widely recognized.

6 Open Challenges and Possible Solutions

For multi-modal problems, we should fully use the complementarity and redundancy between multiple modalities to capture valuable information. However, multi-modal tasks are still challenging because of the heterogeneity of data. In this section, we discuss the open challenges and potential solutions of multi-modal technology. We hope to provide suggestions on how to improve the performance of multi-modal tasks.

Dataset Quality. A critical bottleneck of multi-modal detection is the availability of high-quality datasets. Most existing datasets have the following problems: small scale, unbalanced categories, and marking errors. In order to increase the scale and richness of data, some datasets use partially synthesized data to build datasets [139]. However, there may be a domain gap between synthetic and real-world datasets. Although some methods exist to solve the gap between synthetic and real

data, such as generative adversarial networks [140], the related method still needs further study. It is noteworthy that the imbalance of the raw data may lead to the multi-modal model being dominated by a certain modality. Thus, we should focus on the balance of data categories in the dataset. We can also place multiple sensors in the same component to reduce parameter changes caused by turbulence and jitter.

Information Loss. It is necessary to translate the multi-modal data format in the data fusion stage because of the heterogeneity between different data types. This process will lead to the inevitable loss of information. For example, some existing methods map 3D point cloud data to 2D BEV and fuse the translated data with the existing 2D data [141]. The early fusion method can effectively use the rich information between different modalities, but it makes the feature dimension tremendously. The downstream task needs to reduce the dimension of the input data. This behavior leads to the loss of information. Therefore, designing a representation method for high-dimensional data is very important. Continuous convolution is a potential solution to extract the multi-scale convolution feature map, which reduces the loss of geometric information by capturing local information [142]. Different data fusion stages have different degrees of information loss, so we can also consider using the neural architecture search (NAS) technology to obtain the near-optimal neural structure and find the appropriate fusion stage [143].

Modality Quantity. Most existing research performs fusion operations with two modalities. This is because more modalities will lead to excessive data noise and feature dimension. However, some researches require high accuracy. For example, in autonomous driving, vehicles must accurately identify their surroundings. Therefore, how to fuse three or more modalities is worth considering. Researchers need to consider dimension explosion, data noise, and computational complexity at the same time. The following content will explain the potential solutions of data noise and computational complexity. For the dimension explosion problem, we need to find a method to reduce the number of features while avoiding too much information loss. The current research has provided many dimensionality reduction methods. The most commonly used methods are principal component analysis (PCA) and Gaussian Process Latent Variable Model (GPLVM), as well as many related variants [144]. However, they are only applicable to multi-modal tasks with several modalities. Feature filtering or clustering methods may be a potential solution. Another method is continuously testing the data of different dimensions in the calculation. We can use the results generated by the calculation to verify and adjust repeatedly until the best feature scheme is found.

Error Accumulation. The effective fusion of multi-modal data can obtain higher prediction accuracy, while the wrong fusion method will inject too much data noise. The continuous accumulation of data noise leads to the model performance degradation. Data noise is inevitable during the data fusion stage because of the heterogeneity of different modalities and information differences. We divide data noise into internal noise and external noise. Internal noise refers to the noise of data itself, also known as characteristic noise. Common noise reduction methods can be summarized as filtering-based, partial differential-based, and low-rank matrix-based methods. External noise refers to the noise generated during the multi-modal data fusion stage. At this time, researchers should consider the data quality and select effective data alignment methods and data fusion structures to reduce the data noise. Some studies use data translation technology to convert multi-modal data with each other and unify the data of different modalities [145]. We can further optimize on this basis.

Real-Time Guarantee. The multi-modal model needs to process more data than the unimodal model. Thus, the multi-modal model has more parameters and higher computational complexity. It is hard to satisfy the application scenarios with high real-time requirements. Real-time is one of the

main factors considered in multi-modal methods. Most public benchmarks have taken speed as the evaluation index [146]. We can simplify the model structure to reduce the computational com-plexity by exploring the pruning and quantization techniques of the model. In addition, some compression strategies can also effectively improve the efficiency of multi-modal model training and also be used as a potential research direction.

Dynamic Environment. In general, multi-modal data is collected dynamically so that the data distribution will change over time. When the data distribution changes, the traditional method based on deep learning can only retrain the model to adapt to the new data distribution. However, retraining the model requires a lot of computing resources and time, which is unrealistic. Therefore, we can add historical data as training samples to predict the changing trend of data in the model training stage and design a multi-modal deep learning model in the way of incremental learning to increase the accuracy of the model prediction.

Time Synchronization. Time synchronization refers to ensuring that the fused multi-modal data are aligned in time and space, and it is the most important and challenging. Currently, many algorithms are used to solve the data alignment problem [147]. Most multi-modal data only have a short time dimension, which limits the need to learn the long-term interaction model. When dealing with long-term sequences, capturing the semantic association information between modalities with the increased number of sequences is challenging. A potential solution is to use similarity measurement and knowledge graphs to reduce semantic differences and complete the data alignment tasks. On the other hand, the collected multi-modal data may generate misalignment due to the time deviation between sensors. Therefore, we can design a caching mechanism to deal with data latency.

7 Conclusion

Multi-modal fusion technology is rapidly emerging as the dominant research approach due to its superior perception and judgment abilities. In complex environments, multi-modal data fusion technology can effectively leverage its strengths to enhance accuracy and reduce ambiguities by integrating multiple sources of information. It is crucial to fuse all multi-modal information to maximize the advantages of multi-modal data fusion and push model precision to its upper bound. This paper summarizes the related work on multi-modal fusion technology in multiple fields. We focus on finding the appropriate multi-modal fusion technology to obtain better performance and provide an intuitive suggestion for researchers. We first compare the four fusion stages of multi-modal methods and introduce the three core technologies that can improve the effect of data fusion in detail. Next, we discuss the related applications of multi-modal technology by comparing the existing research methods. Finally, we analyze the existing open challenges and propose potential directions for multi-modal fusion technology.

Acknowledgement: The authors are grateful to all the editors and anonymous reviewers for their comments and suggestions and thank all the members who have contributed to this work with us.

Funding Statement: This paper is supported by the Natural Science Foundation of Liaoning Province (Grant No. 2023-MSBA-070) and the National Natural Science Foundation of China (Grant No. 62302086).

Author Contributions: Study conception and design: Tianzhe Jiao, Jie Song, Chaopeng Guo; Data collection: Tianzhe Jiao, Yuming Chen, and Xiaoyue Feng; Analysis and interpretation of methods: Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng; Draft manuscript preparation: Tianzhe Jiao, Yuming

Chen; Review and editing: Jie Song. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Dai, Z. Yan, J. Cheng, X. Duan, and G. Wang, "Analysis of multimodal data fusion from an information theory perspective," *Inf. Sci.*, vol. 623, no. 2, pp. 164–183, Apr. 2023. doi: 10.1016/j.ins.2022.12.014.

[2] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multi-modal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, pp. 2381, Feb. 2023. doi: 10.3390/s23052381.

[3] Q. Liu *et al.*, "Privacy and integrity protection for IoT multimodal data using machine learning and blockchain," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 20, no. 6, pp. 1–18, Mar. 2024. doi: 10.1145/3638769.

[4] M. Ridhun, R. S. Lewis, S. C. Misquith, S. Poojary, and K. K. Mahesh, "Multimodal human computer interaction using hand gestures and speech," in *Proc. Intell. Human Comput. Interact.*, Tashkent, Uzbekistan, Apr. 2023, pp. 63–74.

[5] D. Roy, Y. Li, T. Jian, P. Tian, K. R. Chowdhury and S. Ioannidis, "Multi-modality sensing and data fusion for multi-vehicle detection," *IEEE Trans. Multimedia*, vol. 25, pp. 2280–2295, Jan. 2023. doi: 10.1109/TMM.2022.3145663.

[6] S. U. Khan, M. A. Khan, M. Azhar, F. Khan, Y. Lee and M. Javed, "Multi-modal medical image fusion towards future research: A review," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 35, no. 8, pp. 3605–3619, Sep. 2023. doi: 10.1016/j.jksuci.2023.101733.

[7] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, Nov. 1989. doi: 10.1109/35.41402.

[8] P. P. Liang, A. Zadeh, and L. P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Comput. Surv.*, vol. 12, pp. 1–40, Apr. 2024. doi: 10.1145/3656580.

[9] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019. doi: 10.1109/TPAMI.2018.2798607.

[10] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multi-modal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, pp. 1–17, Jan. 2021. doi: 10.1016/j.imavis.202-0.104042.

[11] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 4603–4611.

[12] Y. Wang *et al.*, "Multi-modal 3D object detection in autonomous driving: A survey," *Int. J. Comput. Vis.*, vol. 131, no. 8, pp. 2122–2152, May 2023. doi: 10.1007/s11263-023-01784-z.

[13] A. Vaswani *et al.*, "Attention is all you need," in *NIPS'17: Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[14] Q. Cai, Y. Pan, T. Yao, C. W. Ngo, and T. Mei, "ObjectFusion: Multi-modal 3D object detection with object-centric fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Paris, France, Oct. 2023, pp. 18021–18030. doi: 10.1109/iccv51070.2023.01656.

[15] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 21653–21662.

[16] Y. Liu, B. Schiele, A. Vedaldi, and C. Rupprecht, "Continual detection transformer for incremental object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 23799–23808.

[17] Y. Rong, X. Wei, T. Lin, Y. Wang, and E. Kasneci, "DynStatF: An efficient feature fusion strategy for LiDAR 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 3238–3247.

[18] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision bench-mark suite," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3354–3361.

[19] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 2446–2454.

[20] H. Caesar *et al.*, "nuScenes: A multi-modal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 11621–11631.

[21] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou and R. Yang, "The apolloscape dataset for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Jul. 2019. doi: 10.1109/TPAMI.2019.2926463.

[22] Q. Pham *et al.*, "A*3D dataset: Towards autonomous driving in challenging environments," in *Proc. Int. Conf. Robotics Automat.*, Paris, France, May 2020, pp. 2267–2273. doi: 10.1109/ICRA40945.2020.9197385.

[23] P. Xiao *et al.*, "PandaSet: Advanced sensor suite dataset for autonomous driving," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Indianapolis, IN, USA, Sep. 2021, pp. 3095–3101. doi: 10.1109/ITSC48978.2021.9565009.

[24] Z. Wang *et al.*, "Cirrus: A long-range bi-pattern lidar dataset," in *Proc. Int. Conf. Robot. Automat.*, Xi'an, China, May 2021, pp. 5744–5750. doi: 10.1109/icra48506.2021.9561267.

[25] A. Patil, S. Malla, H. Gang, and Y. T. Chen, "The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes," in *Proc. Int. Conf. Robot. Automat.*, Montreal, QC, Canada, May 2019, pp. 9552–9557.

[26] M. Chang *et al.*, "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 8748–8757. doi: 10.1109/cvpr.2019.00895.

[27] J. Mao *et al.*, "One million scenes for autonomous driving: Once dataset," in *Proc. NeurIPS Datasets Benchmarks*, Dec. 2021, pp. 1–21.

[28] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012. doi: 10.1109/T-AFFC.2011.15.

[29] W. Zheng and B. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Dev.*, vol. 7, no. 3, pp. 162–175, Sep. 2015. doi: 10.1109/TAMD.2015.2431497.

[30] B. H. Menze *et al.*, "The multi-modal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015. doi: 10.1109/TMI.2014.2377694.

[31] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1, pp. 68–77, Jan. 2015. doi: 10.5114/wo.2014.47136.

[32] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," arXiv preprint arXiv:2308.02463, 2023.

[33] C. Wu, X. Zhang, Y. Wang, Y. Zhang, and W. Xie, "K-Diag: Knowledge-enhanced disease diagnosis in radiographic imaging," arXiv preprint arXiv:2302.11557, 2023.

[34] X. Zhang *et al.*, "PMC-VQA: Visual instruction tuning for medical visual question answering," arXiv preprint arXiv:2305.10415, 2023.

[35] B. Liu, L. Zhan, L. Xu, L. Ma, Y. Yang and X. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, Nice, France, Apr. 2021, pp. 1650–1654. doi: 10.1109/ISBI48211.2021.9434010.

[36] H. T. Nguyen *et al.*, "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," *Sci. Data*, vol. 10, no. 1, pp. 277, May 2023. doi: 10.1038/s41597-023-02100-7.

[37] S. Sun, C. Lou, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, Jul. 2017. doi: 10.1016/j.inffus.2016.10.004.

[38] A. Ghorbanali and M. K. Sohrabi, "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis," *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 1479–1512, Jul. 2023. doi: 10.1007/s10462-023-10555-8.

[39] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L. Morency, "Multi-modal analysis and prediction of persuasiveness in online social multimedia," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 3, pp. 1–25, Oct. 2016. doi: 10.1145/2897739.

[40] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. Morency, "Multi-modal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Assoc. Comput. Linguistics*, Melbourne, Australia, Jul. 2018, pp. 2236–2246. doi: 10.18653/v1/p18-1208.

[41] W. Yu *et al.*, "CH-SIMS: A Chinese multi-modal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. Assoc. Comput. Linguistics*, Jul. 2020, pp. 3718–3727.

[42] L. Morency, R. Mihalcea, and P. Doshi, "Towards multi-modal sentiment analysis: Harvesting opinions from the web," in *Proc. Int. Conf. Multimodal Interfaces*, Alicante, Spain, Nov. 2011, pp. 169–176. doi: 10.1145/2070481.2070509.

[43] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "MOSI: Multi-modal corpus of sentiment intensity and subjectivity analysis in online opinion videos," arXiv preprint arXiv:1606.06259, 2016.

[44] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multi-modal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1334–1350, Apr. 2021. doi: 10.1109/TAFFC.2021.3097002.

[45] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multi-modal multi-party dataset for emotion recognition in conversations," arXiv preprint arXiv:1810.02508, 2018.

[46] S. Ramamoorthy *et al.*, "Memotion 2: Dataset on sentiment and emotion analysis of memes," in *Proc. DE-FACTIFY@AAAI*, Vancouver, Canada, Feb. 2022, pp. 1–11.

[47] S. Mishra *et al.*, "FACTIFY: A multi-modal fact verification dataset," in *Proc. DE-FACTIFY@AAAI*, Vancouver, BC, Canada, Feb. 2022, pp. 1–14.

[48] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. van Laerhoven, "Introducing wesad, a multi-modal dataset for wearable stress and affect detection," in *Proc. Int. Conf. Multimodal Interact*. Boulder, CO, USA, Oct. 2018, pp. 400–408.

[49] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1798–1828, Mar. 2013. doi: 10.1109/TPAMI.2013.50.

[50] W. Hu, Y. Wang, and Y. Jia, "Multi-modal knowledge representation: A survey," in *Proc. Int. Conf. Data Sci. Cyber.*, Hefei, China, Aug. 2023, pp. 68–75. doi: 10.1109/DSC59305.2023.00020.

[51] M. Jin and J. Li, "Graph to grid: Learning deep representations for multimodal emotion recognition," in *Proc. ACM Multimedia*, New York, NY, USA, Oct. 2023, pp. 5985–5993.

[52] H. M. Mohammed, A. N. Omeroglu, and E. A. Oral, "MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection," *Expert Syst. Appl.*, vol. 223, no. 1, pp. 1–13, Aug. 2023. doi: 10.1016/j.eswa.2023.119790.

[53] X. Zheng, X. Huang, C. Ji, X. Yang, P. Sha, and L. Cheng, "Multi-modal person re-identification based on transformer relational regularization," *Inf. Fusion*, vol. 103, no. 6, pp. 1–8, Mar. 2024. doi: 10.1016/j.inffus.2023.102128.

[54] R. Huang *et al.*, "Joint representation learning for text and 3D point cloud," *Pattern Recognit.*, vol. 147, pp. 1–12, Mar. 2024. doi: 10.1016/j.patcog.2023.110086.

[55] X. Liu, S. Tang, B. Zhan, J. Wu, X. Ma, and J. Wang, "WiVi-GR: Wireless-visual joint representation-based accurate gesture recognition," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 2701–2711, Jan. 2024. doi: 10.1109/JIOT.2023.3292376.

[56] H. Sun, X. Qin, and X. Liu, "Image-text matching using multi-subspace joint representation," *Multim. Syst.*, vol. 29, no. 3, pp. 1057–1071, Jan. 2023. doi: 10.1007/s00530-022-01038-x.

[57] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, May 2019. doi: 10.1109/ACCESS.2019.2916887.

[58] A. Wu, W. S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Feb. 2020. doi: 10.1007/s11263-019-01290-1.

[59] Z. Cai, Y. Ma, J. Huang, X. Mei, and F. Fan, "Correlation-guided discriminative cross-modality features network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–18, Dec. 2023. doi: 10.1109/TIM.2024.3400307.

[60] X. Gong, Y. Dong, and T. Zhang, "CoDF-Net: Coordinated-representation decision fusion network for emotion recognition with EEG and eye movement signals," *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 4, pp. 1213–1226, Apr. 2024. doi: 10.1007/s13042-023-01964-w.

[61] Z. Fang *et al.*, "UWAT-GAN: Fundus fluorescein angiography synthesis via ultra-wide-angle transformation multi-scale GAN," in *Proc. Med. Imag. Comput. Comput. Assisted Intervention*, Vancouver, BC, Canada, Oct. 2023, pp. 745–755. doi: 10.1007/978-3-031-43990-2_70.

[62] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, Nov. 2002. doi: 10.1023/A:1020346032608.

[63] M. Mitchell *et al.,* "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Chapter Assoc. for Comput. Linguistics*, Avignon, France, Apr. 2012, pp. 747–756.

[64] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. 2013 Conf. Empirical Methods in Natural Lang. Processs*, Seattle, WA, USA, Oct. 2013, pp. 1292–1302.

[65] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney *et al.*, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North Amer. Chapt. Assoc. Comput. Linguistics: Human Lang. Technol.*, Denver, CO, USA, Jun. 2015, pp. 1494–1504. doi: 10.3115/v1/n15-1173.

[66] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *Pattern Recog.: 37th, German Conf., GCPR 2015*, Aachen, Germany, Oct., 2015, pp. 209–221. doi: 10.1007/978-3-319-24947-6_17.

[67] Y. Li, F. Wei, Y. Zhang, W. Chen, and J. Ma, "HS2P: Hierarchical spectral and structure-preserving usion network for multimodal remote sensing image cloud and shadow removal," *Inf. Fusion.*, vol. 94, pp. 215–228, Jun. 2023. doi: 10.1016/j.inffus.2023.02.002.

[68] R. Huang *et al.*, "AV-TranSpeech: Audio-visual robust speech-to-speech translation," in *Proc. Assoc. Comput. Linguistics*, Toronto, ON, Canada, Jul. 2023, pp. 8590–8604.

[69] X. Zhu *et al.*, "METTS: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1506–1518, Feb. 2024. doi: 10.1109/TASLP.2024.3363444.

[70] J. Wu, Z. Hu, and R. Mooney, "Generating question relevant captions to aid visual question answering," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 3585–3594.

[71] P. Cascante-Bonilla *et al.*, "Going beyond nouns with vision & language models using synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Paris, France, Oct. 2023, pp. 20098–20108. doi: 10.1109/iccv51070.2023.01844.

[72] Y. Zeng, W. Yan, S. Mai, and H. Hu, "Disentanglement translation network for multimodal sentiment analysis," *Inf. Fusion*, vol. 102, no. 11, pp. 1–12, Feb. 2024. doi: 10.1016/j.inffus.2023.102031.

[73] F. Wang *et al.*, "TEDT: Transformer-based encoding-decoding translation network for multimodal sentiment analysis," *Cogn. Comput.*, vol. 15, no. 1, pp. 289–303, Jan. 2023. doi: 10.1007/s12559-022-10073-9.

[74] L. Zhang *et al.*, "CAMEL: Capturing metaphorical alignment with context disentangling for multimodal emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, Mar. 2024, pp. 9341–9349. doi: 10.1609/aaai.v38i8.28787.

[75] W. Elisa *et al.*, "Multi-modal machine learning in image-based and clinical biomedicine: Survey and prospects," arXiv preprint arXiv:2311.02332, 2023.

[76] J. Yu *et al.*, "Answer-based entity extraction and alignment for visual text question answering," in *Proc. ACM Multimedia*, Ottawa, ON, Canada, Oct. 2023, pp. 9487–9491.

[77] M. Wang, Y. Shi, H. Yang, Z. Zhang, Z. Lin and Y. Zheng, "Probing the impacts of visual context in multimodal entity alignment," *Data Sci. Eng.*, vol. 8, no. 2, pp. 124–134, Apr. 2023. doi: 10.1007/s41019-023-00208-9.

[78] W. Zhao, D. Zhou, B. Cao, K. Zhang, and J. Chen, "Adversarial modality alignment network for cross-modal molecule retrieval," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 278–289, Jan. 2024. doi: 10.1109/TAI.2023.3254518.

[79] Y. Sun, L. Lei, and L. Liu, "Structural regression fusion for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, no. 1, pp. 1–18, Jul. 2023. doi: 10.1109/TGRS.2023.3335418.

[80] X. Liu *et al.*, "EAF-WGAN: Enhanced alignment fusion-wasserstein generative adversarial network for turbulent image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5605–5616, Oct. 2023. doi: 10.1109/TCSVT.2023.3262685.

[81] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proc. EEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 122–132.

[82] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Republic of Korea, Oct. 2019, pp. 7464–7473.

[83] Y. Zhou, G. Yang, Y. Zhou, D. Ding, and J. Zhao, "Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition," in *Med. Imag. Comput. Comput. Assisted Intervention–MICCAI 2023: 26th Int. Conf.*, Vancouver, BC, Canada, Oct. 8–12, 2023, pp. 704–713. doi: 10.1007/978-3-031-43990-2_66.

[84] Z. Chen *et al.*, "AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection," in *Proc. Thirty-First Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 827–833.

[85] Y. Li *et al.*, "DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 17161–17170. doi: 10.1109/CVPR52688.2022.01667.

[86] Y. Cui *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2021. doi: 10.1109/TITS.2020.3023541.

[87] Y. Li *et al.*, "Deep learning for lidar points clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2020. doi: 10.1109/TNNLS.2020.3015992.

[88] L. Xie *et al.*, "PI-RCNN: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module," in *Proc. AAAI Conf. Artif. Intell.*, Palo Alto, CA, USA, Apr. 2020, pp. 12460–12467.

[89] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 77–85. doi: 10.1109/CVPR.2017.16.

[90] X. Ma, X. Zhang, M. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, Mar. 2024. doi: 10.1109/TGRS.2024.3373033.

[91] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista and J. D. Ser, "Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows," *Inf. Fusion*, vol. 96, no. 22, pp. 281–296, Aug. 2023. doi: 10.1016/j.inffus.2023.02.005.

[92] G. P. Meyer, J. Charland, D. Hegde, A. Laddha, and C. Vallespi-Gonzalez, "Sensor fusion for joint 3D object detection and semantic segmentation," in *2019 IEEE/CVF Conf. Comput. Vision Pattern Recogn. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 1230–1237.

[93] J. Wang *et al.*, "KDA3D: Key-point densification and multi-attention guidance for 3D object detection," *Remote Sens.*, vol. 12, no. 11, pp. 1895, Jun. 2020. doi: 10.3390/rs12111895.

[94] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, and D. Tao, "SimDistill: Simulated multi-modal distillation for BEV 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, Canada, Mar. 2024, pp. 7460–7468.

[95] T. Huang, Z. Liu, X. Chen, and X. Bai, "EPNet: Enhancing point features with image semantics for 3D object detection," in *Proc. Eur. Conf. Comput. Vision (ECCV) 2020*, Glasgow, UK, Aug. 2020, pp. 35–52.

[96] M. Bijelic *et al.*, "Seeing through fog without seeing fog: Deep multi-modal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 11679–11689. doi: 10.1109/CVPR42600.2020.01170.

[97] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: Exploiting radar for robust perception of dynamic objects," in *Proc. Eur. Conf. Comput. Visson (ECCV) 2020*, Glasgow, UK, Dec. 2020, pp. 496–512.

[98] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multi-modal vehicle detection in foggy weather using complementary lidar and radar signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 444–453.

[99] S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR object candidates fusion for 3D object detection," in *Proc. 2020 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Las Vegas, NV, USA, Oct. 2020, pp. 10386–10393. 10.1109/IROS45743.2020.9341791.

[100] L. Wen and K. Jo, "Fast and accurate 3D object detection for lidar-camera-based autonomous vehicles using one shared voxel-based backbone," *IEEE Access*, vol. 9, pp. 22080–22089, Jan. 2021. doi: 10.1109/ACCESS.2021.3055491.

[101] J. H. Yoo, Y. Kim, J. S. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection," in *Proc. Comput. Vis.-ECCV 2020*, Glasgow, UK, Aug. 2020, pp. 720–736.

[102] Y. Kim, K. Park, M. Kim, D. Kum, and J. Won Choi, "3D dual-fusion: Dual-domain dual-query camera-LiDAR fusion for 3D object detection," arXiv preprint arXiv:2211.13529, 2023.

[103] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-Net: Multi-modal voxelnet for 3D object detection," in *Proc. Int. Conf. Robot. Automat*, Montreal, QC, Canada, May 2019, pp. 7276–7282. doi: 10.1109/ICRA.2019.8794195.

[104] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 7345–7353.

[105] C. Chen, L. Z. Fragonara, and A. Tsourdos, "RoIFusion: 3D object detection from lidar and vision," *IEEE Access*, vol. 9, pp. 51710–51721, Apr. 2021. doi: 10.1109/ACCESS.2021.3070379.

[106] Y. Zhang, Q. Zhang, Z. Zhu, J. Hou, and Y. Yuan, "GLENet: Boosting 3D object detectors with generative label uncertainty estimation," *Int. J. Comput. Vis.*, vol. 131, no. 12, pp. 3332–3352, Jul. 2023. doi: 10.1007/s11263-023-01869-9.

[107] W. Zheng, W. Tang, L. Jiang, and C. W. Fu, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14494–14503.

[108] X. Li *et al.*, "A muti-modal feature fusion method based on deep learning for predicting immunotherapy response," *J. Theor. Biol.*, vol. 586, no. 670, pp. 1–9, Jun. 2024. doi: 10.1016/j.jtbi.2024.111816.

[109] X. Zhang *et al.*, "Dual-view learning based on images and sequences for molecular property prediction," *IEEE J. Biomed. Health Inf.*, vol. 28, no. 3, pp. 1564–1574, Mar. 2024. doi: 10.1109/JBHI.2023.3347794.

[110] X. Yan, R. Zheng, J. Chen, and M. Li, "scNCL: Transferring labels from scRNA-seq to scATAC-seq data with neighborhood contrastive regularization," *Bioinform*, vol. 39, no. 8, pp. 1−9, Aug. 2023. doi: 10.1093/bioinformatics/btad505.

[111] S. D. Ramlal, J. Sachdeva, C. K. Ahuja, and N. Khandelwal, "An improved multi-modal medical image fusion scheme based on hybrid combination of nonsubsampled contourlet transform and stationary wavelet transform," *Int. J. Imaging Syst. Technol.*, vol. 29, no. 2, pp. 146–160, Jun. 2019. doi: 10.1002/ima.22310.

[112] A. Albahri *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, no. 10, pp. 156–191, Aug. 2023. doi: 10.1016/j.inffus.2023.03.008.

[113] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 133–152, Nov. 2023. doi: 10.1109/TAFFC.2020.3035535.

[114] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. BrainLes@Medical Imag. Comput. Comput. Assisted Intervention (MICCAI)*, Granada, Spain, Jan. 2018, pp. 311–320.

[115] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities," *Neurocomputing*, vol. 466, no. 13, pp. 102–112, Nov. 2021. doi: 10.1016/j.neucom.2021.09.032.

[116] Z. Liu *et al.*, "CANet: Context aware network for brain glioma segmentation," *IEEE Trans. Med. Imaging*, vol. 40, no. 7, pp. 1763–1777, Mar. 2021. doi: 10.1109/TMI.2021.3065918.

[117] S. Hashim, A. Muhammad, and A. Karthik Nandakumar, "SubOmiEmbed: Self-supervised representation learning of multi-omics data for cancer type classification," in *Proc. Int. Conf. Bioinf. Comput. Biol.*, Hangzhou, China, Jun. 2022, pp. 66–72. doi: 10.1109/ICBCB55259.2022.9802478.

[118] J. B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Proc. Neural Inf. Process. Syst. (NeurIPS) 2022*, New Orleans, LA, USA, Dec. 2022, pp. 1–21.

[119] P. Li, G. Liu, L. Tan, J. Liao, and S. Zhong, "Self-supervised visionlanguage pretraining for medical visual question answering," in *Proc. Int. Symp. Biomed. Imag.*, Cartagena, Colombia, Sep. 2023, pp. 1–5. doi: 10.1109/ISBI53787.2023.10230743.

[120] K. Zhang *et al.*, "BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multi-modal tasks," arXiv preprint arXiv:2305.17100, 2023.

[121] S. Zhang *et al.*, "BiomedCLIP: A multi-modal biomedical foundation model pretrained from fifteen million scientific image-text pairs," arXiv preprint arXiv:2303.00915, 2024.

[122] S. K. D'mello and J. Kory, "A review and meta-analysis of multi-modal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Feb. 2015. doi: 10.1145/2682899.

[123] V. Pérez-Rosas, R. Mihalcea, and L. Morency, "Utterance-level multi-modal sentiment analysis," in *Proc. Assoc. Comput. Linguistics*, Sofia, Bulgaria, Aug. 2013, pp. 973–982.

[124] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multi-modal sentiment analysis," arXiv preprint arXiv:1707.07250, 2017.

[125] G. Hu, T. Lin, Y. Zhao, G. Lu, Y. Wu and Y. Li, "UniMSE: Towards unified multi-modal sentiment analysis and emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 7837–7851.

[126] Z. Liu *et al.*, "Efficient low-rank multi-modal fusion with modality-specific factors," in *Proc. Assoc. Comput. Linguistics*, Melbourne, Australia, Jul. 2018, pp. 2247–2256.

[127] W. Han, H. Chen, and S. Poria, "Improving multi-modal fusion with hierarchical mutual information maximization for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, Dominican Republic, Nov. 2021, pp. 9180–9192.

[128] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multi-modal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, Virtual, May 2021, pp. 10790–10797.

[129] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, and L. Morency, "Deep multi-modal fusion for persuasiveness prediction," in *Proc. ACM Int. Conf. Multimodal Interact.*, Tokyo, Japan, Nov. 2016, pp. 284–288. doi: 10.1145/2993148.2993176.

[130] T. Yu *et al.*, "Speech-text dialog pre-training for spoken dialog understanding with explicit cross-modal alignment," in *Proc. Assoc. Comput. Linguistics*, Toronto, ON, Canada, Jul. 2023, pp. 7900–7913.
[131] Z. Wu, Z. Gong, J. Koo, and J. Hirschberg, "Multi-modality multi-loss fusion network," arXiv preprint arXiv:2308.00264, 2023.
[132] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, "Multi-label emotion analysis in conversation via multi-modal knowledge distillation," in *Proc. ACM Int. Conf. Multimedia*, Ottawa, Canada, Oct. 2023, pp. 6090–6100.
[133] G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "MMLatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 4573–4577.doi: 10.1109/ICASSP43922.2022.9746418.
[134] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu and T. Yu, "Learning language-guided adaptive hyper-modality representation for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Singapore, Dec. 2023, pp. 756–767.
[135] Y. Wang, J. Wu, K. Furumai, S. Wada, and S. Kurihara, "VAE-based adversarial multi-modal domain transfer for video-level sentiment analysis," *IEEE Access*, vol. 10, pp. 51315–51324, May 2022. doi: 10.1109/ACCESS.2022.3174215.
[136] D. Zhang *et al.*, "MM-LLMs: Recent advances in multimodal large language models," arXiv preprint arXiv:2401.13601, 2024.
[137] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8748–8763.
[138] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8821–8831.
[139] A. Kar *et al.*, "Meta-Sim: Learning to generate synthetic datasets," in *2019 IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Seoul, Republic of Korea, Feb. 2020, pp. 4550–4559.
[140] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020. doi: 10.1145/3422622.
[141] B. Yang, W. Luo, and R. Urtasun, "PIXOR: Real-time 3D object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7652–7660.
[142] S. Wang, S. Suo, W. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2589–2597.
[143] H. Tang *et al.*, "Searching efficient 3D architectures with sparse point-voxel convolution," in *Proc. Comput. Vision–Eur. Conf. Comput. Vision (ECCV) 2020*, Glasgow, UK, Nov. 2020, pp. 685–702.
[144] F. Ficuciello, P. Falco, and S. Calinon, "A brief survey on the role of dimensionality reduction in manipulation learning and control," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2608–2615, Jul. 2018. doi: 10.1109/LRA.2018.2818933.
[145] Z. Liu *et al.*, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-Eye view representation," in *Proc. Int. Conf. Robot. Automat.*, London, UK, Jul. 2023, pp. 2774–2781.
[146] T. Yin, X. Zhou, and P. Krähenbühl, "Multi-modal virtual point 3D detection," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2021, pp. 16494–16507.
[147] X. Wu *et al.*, "Sparse fuse dense: Towards high quality 3D detection with depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, New Orleans, LA, USA, Jun. 2022, pp. 5408–5417.