A Comparative Review on Multi-modal Sensors Fusion based on Deep Learning

Article *in* Signal Processing · July 2023

DOI: 10.1016/j.sigpro.2023.109165

3 authors, including:

Jing Liang
University of Electronic Science and Technology of China
94 PUBLICATIONS **882** CITATIONS
SEE PROFILE

Fangqi Zhu
The DEI Group
19 PUBLICATIONS **270** CITATIONS
SEE PROFILE

Signal Processing 213 (2023) 109165

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

A comparative review on multi-modal sensors fusion based on deep learning

Qin Tang [a], Jing Liang [*,a], Fangqi Zhu [b]

[a] The School of Information and Communication Engineering, University of Electronic Science and Technology of China, Xiyuan Ave, Chengdu, Sichuan 611731, China
[b] Seagate Technology, Longmont, CO, 80501, United States

ARTICLE INFO

Keywords:
Multi-model data fusion
Deep learning
Inference mechanisms

ABSTRACT

The wide deployment of multi-modal sensors in various areas generates vast amounts of data with characteristics of high volume, wide variety, and high integrity. However, traditional data fusion methods face immense challenges when dealing with multi-modal data containing abundant intermodality and cross-modality information. Deep learning has the ability to automatically extract and understand the potential association of multi-modal information. Despite this, there is a lack of a comprehensive review of the inherent inference mechanisms of deep learning for multi-modal sensor fusion. This work investigates up-to-date developments in multi-modal sensor fusion via deep learning to provide a broad picture of data fusion needs and technologies. It compares the characteristics of multi-modal data for various sensors, summarizes background concepts about data fusion and deep learning, and carefully reviews a large number of investigations in four inference mechanisms: adaptive learning, deep generative, deep discriminative, and algorithms unrolling. The pros and cons of the above methodologies are presented, and several popular application domains are discussed, including medical imaging, autonomous driving, remote sensing, and robotics. A large collection of multi-modal datasets published in recent years is presented, and several tables that quantitatively compare and summarize the performance of fusion algorithms are provided. Finally, by acknowledging the limitations of current research, we establish potential open challenges and future directions as guidance for deep learning-based multi-sensor fusion.

1. Introduction

The era of data explosion necessitates more efficient collection, transformation, processing, storage, and communication of high volumes of heterogeneous data. These data provide us with a comprehensive and three-dimensional understanding of objects due to their multi-source, multi-modal, and multi-domain characteristics. Instead of relying on experience or intuition, we now have greater confidence in explaining phenomena, drawing conclusions, and making decisions based on various real-world data. Consequently, it is crucial to deal with the massive, complementary, and redundant data from multiple sensors by applying unified rules to obtain a harmonious cognition of the objects of interest.

Enormous quantities of data are gathered from diverse sensors capturing different modalities, including video, radio detection and ranging (radar), infrared, light detection and ranging (Lidar), inertial measurement units (IMUs), and others. One of the key challenges is to construct an efficient data representation and structure that can incorporate information from these disparate sources. However, the raw data obtained from these sensors is heterogeneous, complex, imperfect, and voluminous. Multi-modal sensor fusion technologies can address these challenges by extracting and combining multi-modal information to obtain more reliable, informative, and precise data [1].

Multi-modal sensor fusion involves a complex and comprehensive process of handling multi-source data. Classic signal processing and estimation techniques such as the Kalman filter, particle filter, and expectation maximization have provided an essential theoretical foundation for data fusion technology. Subsequently, other data fusion approaches have emerged, including statistical inference (e.g., Bayesian reasoning, maximum likelihood estimation), Dempster-Shafer theory, random finite set-based methods, Markov random field, information-theoretic fusion (e.g., minimum description length and entropy methods), and fuzzy logic. These techniques have contributed significantly to the development of information fusion technology [2–4].

The rapid advancement in processing hardware, such as computing resources and platforms like graphic processing units (GPUs), tensor

* Corresponding author.
E-mail addresses: tangqin@std.uestc.edu.cn (Q. Tang), liangjing@uestc.edu.cn (J. Liang).

https://doi.org/10.1016/j.sigpro.2023.109165
Received 10 April 2023; Received in revised form 4 June 2023; Accepted 19 June 2023
Available online 3 July 2023

processing units (TPUs), cloud-based processing platforms, and related big data processing techniques, has created opportunities to handle large datasets for multi-sensor data fusion. Recently, deep learning (DL) has emerged as a prominent method for processing big datasets, exhibiting the following three characteristics. Firstly, DL has excellent nonlinear mapping capabilities, enabling it to automatically build complex models that are difficult to hand-design. Secondly, DL-based models have deep layers, allowing them to generate high-dimensional representations, which are more comprehensive than previous plain feature extraction methods. Finally, the flexibility of DL models enables their application in multiple domains, such as computer vision, autonomous driving, robotics, medical diagnosis, and industrial manufacturing. As a result, DL is highly anticipated to enhance the overall performance of multi-sensor data fusion algorithms [5–8].

Several surveys on multi-modal sensor fusion have been published in recent years. For instance, LikLau et al. [9] proposed a multi-perspective classification of data fusion to evaluate smart city applications and applied the proposed classification to selected applications such as monitoring, control, resource management, and anomaly detection, among others, in each smart city domain. Deng et al. [10] conducted a review of coverage optimization problems from the perspective of sensor data fusion. They classified the current coverage optimization problems and summarized existing coverage models based on data fusion. Ghamisi et al. [11] provided a comprehensive review dedicated to image-level fusion, including point cloud, hyperspectral, Lidar, and multi-temporal data. Qiu et al. [12] attempted a comprehensive survey of multi-sensor applications, including the latest unsupervised learning and transfer learning, but only for use in human activity recognition applications. Ahmad et al. [13] surveyed and systematically analyzed the evolution of traditional algorithms to DL-based methods for the field of hyperspectral image classification, providing guidance and guidelines for future prospects of the community. Ramachandram et al. [14] summarized advances in deep multi-modal learning, including regularization strategies and optimization methods for multi-modal fusion structures.

However, most of these surveys have focused on specific issues, and there is a lack of comprehensive reviews on the latest techniques and applications. In the past few years, DL models such as deep generation models, transformer, and algorithm unrolling-based methods have advanced rapidly, significantly enhancing multi-modal sensor fusion technologies [15–19]. Therefore, there is a need for more comprehensive reviews that keep up with the latest advancements in the field and their applications. This survey investigates the state-of-the-art multi--modal sensors fusion based on DL, as shown in Fig. 1. The main contributions of this paper are:

- We conduct an extensive review of DL-based multi-sensor data fusion algorithms and techniques. The DL-based inference mechanism for multi-modal sensor fusion is categorized into five groups: adaptive learning, deep generative, deep discriminative, algorithm unrolling, and transformer models. We analyze these recent multi-modal fusion methodologies in detail, comparing their differences, advantages, and limitations.
- We discuss existing DL-based multi-modal fusion algorithms from the perspective of various applications, including medical diagnosis, autonomous driving, remote sensing, and intelligent robotics. We also evaluate their performance in public datasets using popular metrics, analyzing the advantages and limitations of different models in qualitative and quantitative experiments. Likewise, we present 24 multi-modal datasets for diverse applications.
- We identify the existing research challenges in DL-based inference mechanisms for sensor fusion and propose potential future directions, providing forward-looking guidance for researchers in this field.

The remainder of the paper is organized as follows. Section II outlines the background and concepts for multi-sources data and DL multi-sensor fusion. Section III reviews the current literature on data fusion via DL. All the literature is reviewed regarding application scenes, structured and semi-structured data from multiple sources, and technical

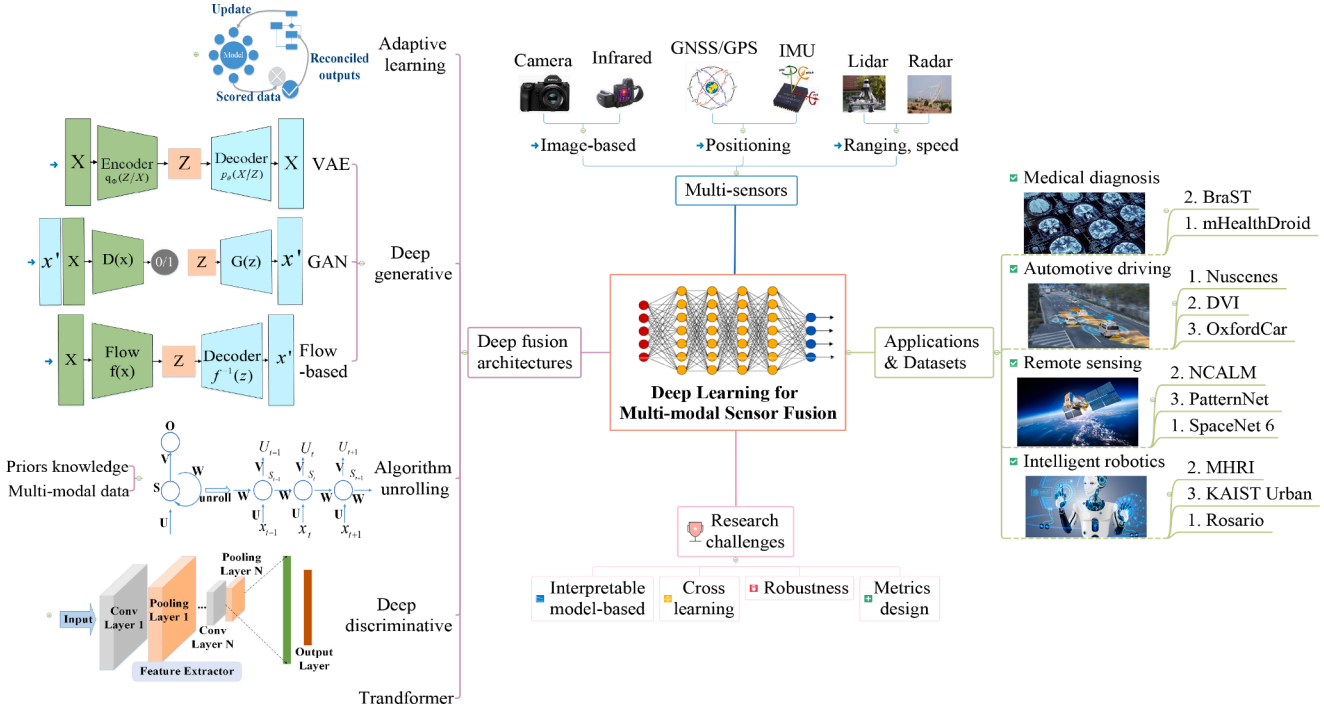


Fig. 1. The overall structure of this paper.

Table 1
A comparison of heterogeneous sensors.

Performance index \ Sensor	Lidar	Radar	Camera	Infrared	IMU	GPS	SAR	
Accuracy	Range	-250m-400m	<1000m	<100m	<100m	300	global	100km-500km
	Angle	360°	±60°	45°	45°	360°	360°	30°-60°
	Time	fast (10ms)	fast (1ms)	fast	fast	slow	slow	slow
	Resolution	>1mm	determined by the signal bandwidth	related to pixel size	0.5-1.5 mRad	1°/h	1∼3m	0.3∼30m
Adaptability	Lighting	excellent	excellent	poor	excellent	excellent	excellent	excellent
	Weather	average	average	average	good	good	average	excellent
	Temperature, humidity, dirt	good	average	average	good	good	average	good
Advantage	•high precision •superior 3D modelling capability	•effective in all weather and light conditions •low compute resource needed •distance and speed measured	•wide field of view •long-range •low cost	•wide measurement range •short response time	•positioning range is the whole scene	•high precision •the error does not diverge over time	•all-day, all-weather •long distance	
Limitation	• very expensive • large form factor • less effective in certain harsh weather conditions	• expensive (long range) • cannot detect traffic signs • low resolution	• ineffective in harsh weather • ineffective in low light • less effective in low contrast	• requires full visibility • short distance spread	• positioning accuracy is average • the divergence of the error over time	• requires full visibility • the positioning range cannot cover indoors	• cannot distinguish details • poor networking capability	
Application	• autonomous driving • robotics • medical monitoring • atmospheric monitoring • smart city	• autonomous driving • military • medical monitoring • atmospheric monitoring • smart city	• autonomous driving • computer vision • security monitoring • robotics • internet of things	• security monitoring • military • medical equipment • car night vision • consumer electronics	• vehicle navigation • unmanned aerial vehicle (UAV) • Robots	• positioning • speed measurement • Timing	• environment monitoring • internet of things • remote sensing	

advantages. Section IV presents applications and public datasets. Section V proposes some principal challenges and future directions. Section VI summarises this work.

2. Background and problem formulation for multi-sensors fusion

This section offers an overview of the sources of multi-modal data, background concepts, and classification relevant to multi-sensor fusion. Additionally, it provides a brief introduction to DL for data fusion and its popular models.

2.1. Multi-modal data from multi-sensors

Each sensor type, or "modality," has its inherent strengths and weaknesses. Different modalities provide critical information with corresponding weights in the fusion process [5,20–23]. With this in mind, we compare the performance indicators, perception parameters, pros, and cons of various sensors in tasks, as depicted in Table 1.

Lidar is an active sensor that emits its own illumination source to measure the distance to objects. The reflected energy is detected and measured to calculate the distance traveled using the speed of light [24]. Lidar uses different types of scattering, such as Rayleigh scattering, Mie scattering, Raman scattering, and fluorescence, for various applications, including high-resolution mapping, robotics, control, and navigation [25]. Lidar's advantage is obtaining point cloud output that reflects the spatial structure, which can be fused with other sensors to provide a better picture of the objects and environment. However, Lidar has difficulty reconstructing point cloud data in poor weather conditions, such as heavy rain.

Radar is a detection sensor that uses electromagnetic waves in the radio or microwave domain to determine the distance, angle, or velocity of objects. Radar is used in various domains where positioning is crucial due to its ability to detect objects at long ranges [26–28]. Unlike other electromagnetic wavelengths, such as visible light, radar can penetrate through weather phenomena like fog, clouds, rain, snow, and sleet, which block visible light. Modern radar systems use digital signal processing and machine learning technologies to extract useful information from high levels of noise.

Likewise, various sensors can be used for proximate and distant sensing, including cameras, Global Positioning System (GPS), IMU, and odometers [29,30]. However, weather conditions such as fog, sun, rain, snow, or darkness can reduce visibility and affect the quality of the images and videos. As a result, alternative measuring sensors are being considered, including airborne and spaceborne geophysical measurements, such as gravity recovery and climate experiment satellite missions and airborne electromagnetic surveys. Point cloud data representing elevation can be obtained using airborne Lidar and terrestrial laser scanning (TLS) [23]. The integration of temporal information with spatial and spectral/backscattering information of remotely sensed data is also possible. The quantity of synthetic aperture radar (SAR) sensors and satellite/airborne-based hyperspectral sensors has significantly increased, extending optical-sensing capabilities [31,32]. Furthermore, the number of stationary/moving platforms for sensors-equipped has also increased, including terrestrial-based, space-based, and moving vehicle-based platforms.

As there is a growing demand for various tasks, it should be guaranteed that the combination of the multi-modal sensors chosen can overcome the limitations of individual sensors. To achieve successful multimodal data fusion, several key properties must be taken into consideration: 1) Consistency: the different modalities of data need to be consistent and coherent to ensure that the fused results are meaningful and accurate; 2) Complementarity: multi-source data should provide information that is relevant across modalities, and the fusion process can help fill data gaps, reduce uncertainty, and improve overall performance; 3) Compatibility: algorithmic models, data formats, and tasks should be considered for compatibility when fusing data from different modalities. By considering these characteristics, multi-modal sensor

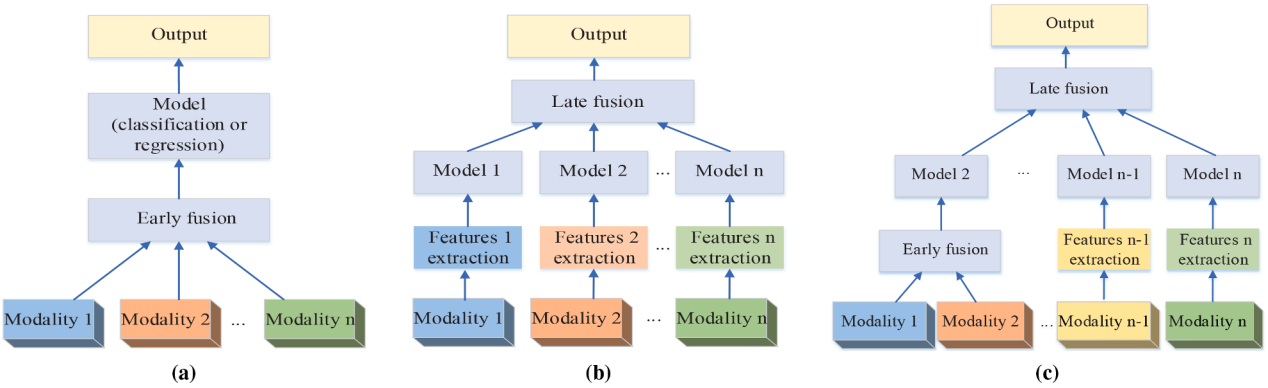


Fig. 2. The three types of multi-sensor data fusion frameworks. (a) Early fusion (data or feature-level); (b) Late fusion (decision-level); (c) Hybrid fusion.

fusion can improve the accuracy, robustness, and reliability of the system.

2.2. Deep learning for multi-sensors fusion

The concept of multi-sensor data fusion was first defined by the Joint Directors of Laboratories (JDL) [33] as a multilevel, multifaceted process that involves the automatic detection, association, correlation, estimation, and combination of data and information from several sources. The primary motivation behind data fusion from multiple sensors is to leverage the heterogeneous characteristics of different sources and formulate a unified, coherent, and reasonable representation that can facilitate better decision-making.

Traditionally, multi-sensor fusion has been performed by independently modeling each observation using a recursive filter and integrating the resulting information into a common state estimation [34–37]. These methods suggest that the optimal combination of information from different modalities can improve the overall accuracy of estimation. Earlier formulations include a Bayesian filter and multi-modal measurement models, which quantify the probability that the current measurement value of each modality is in a given predicted state [38].

The abundance of data from different sensors, both in terms of quantity and modal, presents both challenges and opportunities. However, advances in computing power and DL technology have made efforts to make data more available and complimentary. The classification of data fusion can be divided into early fusion (data or feature-level), late fusion (decision-level), and hybrid fusion. We list them below, and their specific structures are shown in Fig. 2.

- Early fusion involves the integration of feature vectors obtained from multiple sensor sources, connecting the representation of each modality before input to DL models. However, it is limited in its ability to directly handle heterogeneous sensor data at the data-level fusion stage. In contrast, feature-level fusion involves the extraction of feature vectors from the captured data, which are then immediately integrated. This approach allows for more flexible handling of heterogeneous sensor data.
- Late fusion, also known as decision-level fusion, involves training different modalities separately and then integrating the output results of multiple models. Alternatively, specific operations such as classification, regression, and logical decision-making may be utilized based on the acquired data features, and higher-level decisions are made based on the requirements of mission-critical applications. Unlike early and feature-level fusion methods, the fusion process in late fusion has nothing to do with features, and errors from multiple models are usually independent of each other.
- Hybrid fusion combines the three aforementioned fusion mechanisms to provide increased diversity and flexibility. It leverages the advantages of early and late fusion methods while also increasing the structural complexity and training difficulty of the model. The effectiveness of the hybrid fusion approach depends on the rationality of the combined strategy, which is a fundamental factor in improving performance.

In summary, each of the three fusion methods has its own advantages and limitations. Early fusion can better capture the relationship between features, while late fusion can better address overfitting problems but does not allow for training all data simultaneously. Early fusion performs better than late fusion when the correlation between modalities is relatively high. If the modalities are largely uncorrelated, such as when the dimension and sampling rate are not correlated, the late fusion method is more suitable. Although the hybrid fusion approach is flexible, researchers must choose a suitable fusion rule based on specific application problems and research content.

Table 2
A comparison between DL and conventional approaches for multi-modal sensor fusion.

Properties	Deep Multimodal Learning	Conventional Multimodal Learning
Data preprocessing	End-to-end training requires almost no preprocessing of input data.	Some early fusion technologies require data preprocessing.
Feature representation	Both modality (features) representation and shared (fusion) representations are learned from data.	Features are manually modeled based on prior knowledge of the problem and data.
Dimensionality reduction	The model implicitly reduces dimensionality through hidden layers.	Features selection and dimensionality reduction are usually performed explicitly and iteratively.
Data and modality expansion	All fusion models can be easily expanded in terms of data size and the number of modalities.	Early fusion (data-level fusion) may be challenging and not scalable; later fusion rules may need to be redefined.
Fusion stage	Support early, late or intermediate fusion.	Usually, early or late fusion is performed.
Fusion architecture design	The fusion architecture can be self-learned during training.	The fusion architecture is usually artificially designed.
The amount of data required	Deeper and more complex model networks usually require large amounts of training data.	It may not require that much training data.
Computational cost	A powerful graphics processing unit (such as GPU) is required to obtain a reasonable training time.	The GPU may provide acceleration, but it is not important.
Hyperparameter	Many hyperparameter adjustments are essential for the most advanced performance.	It usually does not have that many hyperparameters.

A comparison of multi-modal sensor fusion based on DL approaches and traditional methods is presented in Table 2. Deep multi-modal learning approaches offer several advantages over conventional methods, often providing improved performance for multi-modal data problems.

3. The deep learning based inference mechanism for multi-sensors fusion

Among the different stages of the fusion, the inherent inference mechanism demonstrates how the different sources of the data are organized, reformulated, and merged to represent the data flow to reveal the ground truth. This section mainly discusses the following four inference mechanisms: adaptive learning, deep generative, deep discriminative, algorithm unrolling, and transformer models. To conclude, Table 3 compares most of the algorithms in this section about their input modalities, fusion types, application scenarios, employed models, DL methods, primary challenges to overcome, and performance based on unified metrics.

3.1. Adaptive learning models

In early research, adaptive learning was mainly achieved through adaptive filtering, which was used to handle input data with unknown statistical characteristics. Adaptive filtering typically consisted of a "training process" and a "tracking process." The training process tuned parameters through online optimization, while the tracking process changed states and related parameters with the statistical variations of the input [61–63]. Vercauteren et al. [64] proposed a decentralized multi-sensor fusion algorithm using the Bayesian filtering framework for online inference in sensor networks. The authors introduced a sigma-point information filter by combining the sigma-point filter methodology and the information filter framework. Their method provides accurate and robust collaborative inference under Gaussian approximation. Maggio et al. [65] proposed a tracking algorithm based on an adaptive multi-feature statistical target model. The algorithm fused extracted features from the source data of the target into a single particle filter, and weighed their contributions using reliability metrics from the distribution of particles in each state space. This approach improved the flexibility of the representation by exploiting the complementarities of the failure modes of multiple descriptors.

Table 3
Summary and comparison of DL methods for multi-modal data fusion.

References	Applications	Input modalities	Fusion models	Fusion levels	Challenges to overcome	DL methods	Performance			
							Q	*E*	*R*	*T*
[39]	WSNs, target tracking	Mobile WSNs	Radio-fingerprints, accelerations	Kernel-based + KF	Early-level	⋄	Y	P	N	N
[40]	Vehicle location	Data registration	Images, velocities	R-CNN + KF	Early-level	⋄	H	C	N	H
[41]	Autonomous robots	Data registration	Appearance, depth, motion	CNN + experts	Feature-level	⋄	H	P	C	H
[42]	Object tracking	Multi-feature images	CNN + CF	Feature-level	Back-propagating gradients	⋄	H	Y	C	P
[43]	Visual tracking	Multi-features images	VGG + CF	Feature-level	Dynamic fusion	⋄	H	Y	P	P
[44]	Rotating machinery	Multi-features accelerometers	SAE + DBN	Feature-level	Supervise missing data	□	H	Y	P	P
[45]	Mage reconstructions	Edge detection, colorization, segmentation	MVAE	Feature-level	Dynamic fusion	□	H	Y	P	P
[46]	Structured output prediction	Multi-features images	VAE-based	Feature-level	Dynamic fusion	□	H	C	Y	P
[47]	Trajectory prediction	Multi-agent behavior data	CVAE	Feature-level	Architectural considerations	□	Y	Y	N	H
[48]	Visual saliency prediction	Saliency and image	SalGAN	Feature-level	Mode-collapse problem	□	P	Y	N	H
[49]	Sports analytics	Depth, IMU data	cGAN	Feature-level	Outdoor ground truth	□	Y	H	N	H
[50]	Vision computing	Multi-source images	Flow-based	Feature-level	Dynamic fusion	□	Y	H	N	N
[51]	Gesture recognition	Skeletal dynamics, depth, RGB images	DBN + 3DCNN + HMM	Feature-level	Multi-modal time series data	△	Y	H	N	P
[52]	Image recoloring	Multi-features natural images	CNN-based	Feature-level	Network architecture	△	Y	C	C	P
[53]	Indoor scene recognition	Depth, RGB images	CNN-based	Feature-level	Network robustness	△	Y	H	N	P
[54]	Disease diagnosis	MRI, PET data	SIFT + CCA	Hybrid-level	Network robustness	△	Y	P	H	P
[55]	Situational understanding	Video and audio data	CNN-based	Decision-level	Explanation quality	△	Y	Y	C	P
[56]	Object tracking	Visible and infrared images	DCF-based	Decision-level	Data registration	△	Y	Y	C	C
[16]	Image enhancement	Visible and infrared images	CNN-based	Feature-level	Weights and speed	★	Y	H	C	P
[57]	Compressive spectral imaging	HS, MS images	LADMM-Net	Feature-level	Learnable transforms	★	H	H	N	P
[58]	Uncertainty quantification	RGB images and Lidar data	Transformer	Feature-level	Learnable transforms	★	H	H	Y	N
[59]	Uncertainty quantification	Dense scene flow data	PINN	Hybrid-level	Learnable transforms	★	H	H	N	P
[60]	Graph signal denoising	Multiple noisy graph signals	GNN + trend filter	Signal-level	Network robustness	★	H	Y	Y	N

⋄: Adaptive learning-based;□: Deep generative-based; △: Deep discriminative-based; ★: Algorithm unrolling-based; *Q*: Quality; *E*: Efficiency; *R*: Robustness; *T*: Tested with real-world dataset. Y: Well-considered in theory; H: Well-analysis; P: Rich experiment; C: Concluded but is not adequate; N: Not mentioned.

The integration of DL with adaptive filtering has received considerable attention in recent years. Mahfouz et al. [39] proposed a novel DL-based method that combines a Kalman filter with accelerations of a moving target and radio fingerprints of received signal strength indicators (RSSIs) from a wireless sensor network (WSN) to estimate the real-time position of the target. The Kalman filter is updated using a kernel-based ridge regression-based iterative algorithm to predict the target's position, achieving high accuracy and robustness. Similarly, Gao et al. [66] proposed a DL-based approach that employs an adaptive Kalman filter (AKF) embedded with an Elman neural network to detect the position of interest. The Sage-Husa AKF was utilized to estimate the statistical characteristics of the noise and adapt the filter parameter, effectively overcoming noise variations [67]. Moreover, the Elman neural network-based method can improve the accuracy of the Kalman filter and compensate for the AKF estimated errors, resulting in more precise and reliable position estimation.

Adaptive DL can also be achieved by combining convolutional neural networks (CNNs) with filtering [68]. Zhang et al. [40] proposed a tracking framework that fuses two modalities, images and instantaneous velocities of vehicles. This method integrates Kalman filter for velocity estimation while fusing the inputs of both texture and color of the vehicles. Another example is the fusion scheme based on a mixture of deep CNN and expert system for object detection introduced by Mees et al. [41]. Their method trains CNN and expert networks without prior information to adaptively fuse high-level vision features and the weight of expert classifier outputs. These approaches demonstrate the potential of adaptive DL in enhancing the performance of data fusion and analysis.

Collaborative filters (CFs) are capable of capturing object or signal interactions and associations. Combining CFs with deep learning (DL) frameworks has been the focus of many recent studies [41–43,69]. Valmadre et al. [42] proposed the CFNet, which rewrites the CF into a differentiable neural network layer and integrates it with the feature extraction network for end-to-end optimization. The VOT2017 competition champion algorithm uses fine-tuning of the network model to adaptively learn depth features suitable for CFs, which are then introduced into the C-COT tracking framework [70].

In addition, CFs can serve as the counterparts of convolution filters in DL to encode the holistic representation of objects. Zheng et al. [43] proposed a robust visual tracking method via multi-task deep dual correlation filters. This model jointly trains the CF and the network parameters, and alternately optimizes to obtain multi-level features. It explores the interdependence between multi-level features and captures object appearance changes. These studies demonstrate the potential of combining CFs with DL frameworks to enhance the performance of object tracking and recognition.

3.2. Generative models

To achieve efficient multi-modal sensor fusion, one of the major challenges is learning representations and formulating inference spaces while considering the complexity and reducing the redundancy of heterogeneous data. Generative modeling using neural networks originated in the 1980s with the purpose of learning about data without supervision. However, the essential label information is available through inferencing about implicit relationships, making it clear that generative models potentially provide advantages in multi-sensor fusion. The common idea of generative modeling stems through training a generative model whose samples $\tilde{\boldsymbol{x}} \sim p_\theta(\tilde{\boldsymbol{x}})$ come from the same distribution as the training data distribution, $\boldsymbol{x} \sim p_d(\boldsymbol{x})$.

Research in this field has been divided into several interconnected approaches: variational autoencoders (VAEs) [71], generative adversarial networks (GANs) [72], and flow-based models (autoregressive models and normalizing flows) [73]. The popular approaches and their corresponding characteristics in multi-modal sensor fusion are described below.

3.2.1. Variational auto-encoders

Inferring the hidden variables is usually only compliant for simple linear models. For nonlinear models, one has to take to approximate Bayesian methods. The VAEs is combining neural networks and variational inference to latent-variable models, as shown in Fig. 3. VAEs is expected to describe the distribution $\tilde{p}(x)$ of a batch of samples $x = \{x_1, \ldots, x_n\}$, by decomposing hidden variables z, as follows:

$$p\left(x\right) = \int p\left(x\middle|z\right)p\left(z\right)dz, \quad p\left(x,z\right) = p\left(x\middle|z\right)p\left(z\right) \tag{1}$$

Then, fitting $p(x,z)$ with $p(x|z)$, and $p(z)$ with model $q(z)$ fitting, $q(z)$ is usually defined as the standard normal distribution. Theoretically, the purpose of learning is to maximize the marginal likelihood of the information:

$$\begin{aligned} q(x|z) &= \underset{q(x|z)}{\operatorname{argmax}} \int \tilde{p}(x)\ln\left(\int q\left(x\middle|z\right)q\left(z\right)dz\right)dx \\ &= \underset{q(x|z)}{\operatorname{argmax}} \mathbb{E}_{x\sim\tilde{p}(x)}\left[\ln\left(\int q\left(x\middle|z\right)q\left(z\right)dz\right)\right]. \end{aligned} \tag{2}$$

The kullback–Leibler (KL) divergence is introduced to replace the intractable term in the integral:

$$\begin{aligned} \mathscr{L} &= \mathbb{E}_{x\sim\tilde{p}(x)}\left[-\int p(z|x)\ln q(x|z)dz + \int p(z|x)\ln\frac{p(z|x)}{q(z)}dz\right] \\ &= \mathbb{E}_{x\sim\tilde{p}(x)}\left[\mathbb{E}_{z\sim p(z|x)}[-\ln q(x|z)] + \mathbb{E}_{z\sim p(z|x)}\left[\ln\frac{p(z|x)}{q(z)}\right]\right] \\ &= \mathbb{E}_{x\sim\tilde{p}(x)}\left[\mathbb{E}_{z\sim p(z|x)}[-\ln q(x|z)] + KL(p(z|x)\parallel q(z))\right]. \end{aligned} \tag{3}$$

The above formula (3) is the evidence lower bound (ELBO) [74], which is defined via an inference network $q_\phi(z|x)$ for serving as a tractable importance distribution. $\mathscr{L}$ has a lower bound $-\mathbb{E}_{x\sim\tilde{p}(x)}[\ln\tilde{p}(x)]$, so the relative quality of the generative model can be compared by comparing

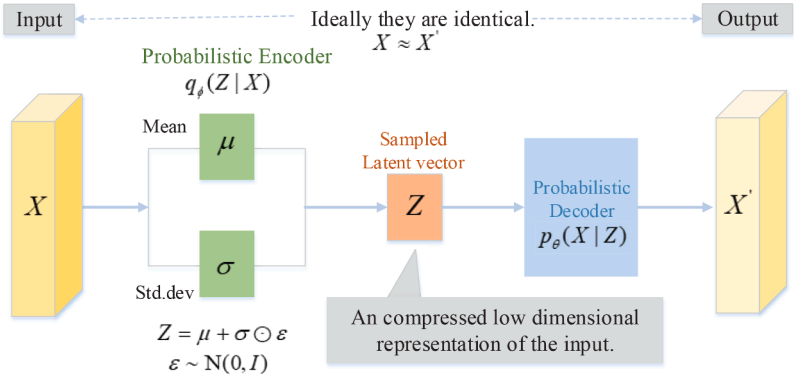


Fig. 3. The general representation of VAE.

the closeness between $\mathscr{L}$ and $-\mathbb{E}_{x\sim\tilde{p}(x)}[\ln\tilde{p}(x)]$. The smaller the overall $\mathscr{L}$, the closer the generative distribution is to the real distribution.

VAE-based approaches that learn joint representations of multi-modal data can produce deeper and more valuable representations [75]. Chen et al. [44] proposed a method for fusing data from multiple accelerometers to improve the fault diagnosis reliability of rotating machinery. Features in time-domain and frequency-domain are extracted from different sensors and fused into two-layer sparse auto-encoders (SAEs) unsupervised neural networks. Early generative approaches for multi-modal input either do not learn a joint distribution or require additional calculations to supervise missing data. Wu and Goodman [45] proposed a multi-modal VAE (MVAE) method for weakly-supervised learning from multi-modal data. The authors assume a generative model of the from $p_\theta(x_1, x_2, \ldots, x_N, z) = p(z)p_\theta(x_1|z)p_\theta(x_2|z)\ldots p_\theta(x_N|z)$, $x_1, \ldots, x_N$ and z denote the N modalities and the common latent variable respectively. Their method is a hybrid of a product-of-experts inference system and a sub-sampled training model for solving the problem of multi-modal information inference. In particular, the MVAE model shares parameters to efficiently learn under any combination of missing modalities.

Similarly, Kurle et al. [46] transformed the raw dataset into tuples $\{\mathbf{x}^{(n)} = (\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_M^{(n)})\}_{n=1}^N$, where M denotes the index of the source, and formulated a multi-source learning framework based on a variational autoencoder. Each encoder in the proposed framework is marked as a different information source $\mathbf{x}_m^{(n)} \in \mathbf{R}^{D_m}$. The multi-sources are fused via shared latent variables by calculating divergence measures in individual sources' posterior approximations. It is crucial to optimize the importance sampling based on ELBO to prevent the generative models from generating averaged conditional samples. In the field of robots, Ivanovic et al. [47] proposed a conditional VAE (CVAE) approach for human behavior prediction, which can produce a multi-modal probability distribution over future human trajectories conditioned on past interactions and candidate robot future actions.

3.2.2. Generative adversarial networks

Another main type of deep generative model that leverages latent information for sensor fusion is the GAN [72,76]. Unlike VAEs, GANs consist of two independent networks: a generator network $G: \mathbb{R}^m \to \mathbb{R}^n$ that receives a latent variable $z \sim p_z(z)$, and a discriminator network $D: \mathbb{R}^n \to [0,1]$ that estimates the probability that a sample comes from the data distribution $x \sim p_d(x)$ in order to produce realistic outputs, as shown in Fig. 4. The generator outputs samples that are then evaluated by the discriminator. This process is formulated as the following min-max optimization problem:

$$\min_G \max_D V(D,G),$$
$$V(D,G) = \mathbb{E}_{x\sim p_d(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \tag{4}$$

The first term of $V(D,G)$ is the entropy of the data from the real distribution $p_d(x)$ passing through the discriminator, representing the best-case scenario. The discriminator attempts to maximize this term to 1. The second term is the entropy of the data from a random input $p(z)$ passing through the generator G, which produces a fake sample that is then evaluated by the discriminator D to determine its authenticity. Here, the discriminator D aims to minimize it to 0.

Mini batch stochastic gradient descent training of generative adversarial nets [72]. Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]. \tag{5}$$

Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right). \tag{6}$$

GANs minimizes Jensen-Shannon (JS) divergence between the $p_\theta(\tilde{x})$ given by the model and the $p_d(x)$ shown by the data in terms of training programs.

GANs have the capability to generate high-quality images while capturing the semantic attributes of the training images. To address the mode-collapse problem, which limits the variety of modes that a generator network can learn to produce, researchers have proposed various methods to improve GAN performance. For instance, Radford et al. [77] introduced deep convolutional generative adversarial networks (DCGANs), which are a strong candidate for unsupervised learning due to their architectural constraints. Pan et al. [48] proposed SalGAN, a data-driven metric-based saliency prediction method trained with an adversarial loss function. SalGAN consists of two networks that aim to generate saliency maps that resemble the ground truth. Despite these advances, mode-collapse remains an issue for GANs [78].

To overcome this problem, Chen et al. [79] proposed an InfoGAN architecture that incorporates a regularization parameter to maximize the mutual information between a small subset of input latent variables. Hong et al. [49] developed a tailored conditional GAN framework to track a player's swing in 3D space using inexpensive tools such as depth sensors and IMUs. The regression and subject classification networks in this framework train in parallel to learn a relationship between the fusion of depth and IMU sensor data and the ground truth, while also learning the ability to discriminate between swings belonging to different subjects. These advances in GAN technology have the potential to enhance unsupervised learning and generate more diverse and

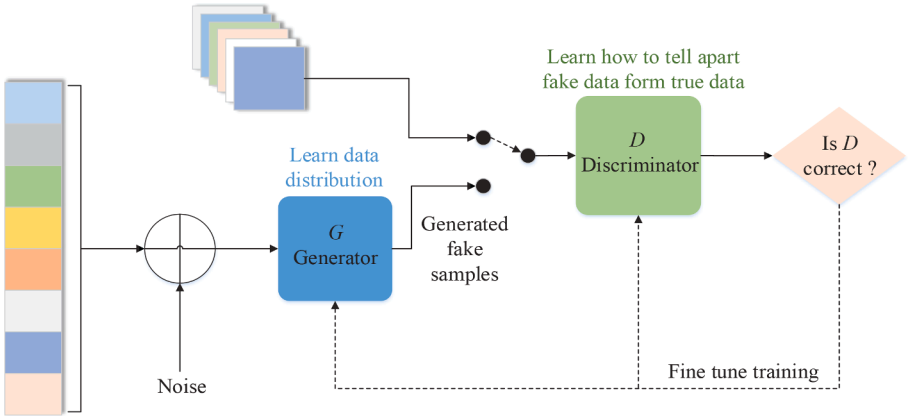


Fig. 4. The general representation of GANs.

realistic images.

3.2.3. Flow-based generative models

Flow-based models explicitly learn the probability density function of input data through the use of "distribution flows." These flows are powerful statistical tools for density estimation [73,80,81]. Accurate estimation of $p_\theta(x)$ enables the completion of various downstream tasks, including the efficient generation of novel yet realistic data points, prediction of the rarity of future events, inference of latent variables, and the completion of partial data samples.

The flow-based generative model depicted in Fig. 5 shares a common objective with GANs. It is well-established that the generator is typically a neural network with a substantial number of parameters, and obtaining a specific expression for P_G is notably challenging. In contrast, flow-based generative models directly address the aforementioned formula (4), representing the most significant distinction between these models and GANs. Given a chain of probability density functions, we can determine the relationship between each consecutive pair of variables. By expanding the equation of the output variable $\mathbf{x}$ step by step, we can trace back to the initial distribution z_0.

$$\mathbf{x} = \mathbf{z}_K = f_K \circ f_{K-1} \circ \ldots \circ f_1(\mathbf{z}_0). \tag{7}$$

The path taken by the random variables $\mathbf{z}_i = f_i(\mathbf{z}_{i-1})$ is known as the flow, and the successive distributions π_i form a normalizing flow. To perform the computation in the Eq. (7), the transformation function $f_i, i = 1, \ldots, K$ must satisfy two conditions [80]:

Condition 1: "It is easily invertible."

$$\begin{cases} \mathbf{y}_{1:d} & = \mathbf{x}_{1:d}, \\ \mathbf{y}_{d+1:D} & = \mathbf{x}_{d+1:D} \odot \exp(s(\mathbf{x}_{1:d})) + t(\mathbf{x}_{1:d}). \end{cases} \tag{8}$$

$$\Leftrightarrow$$

$$\begin{cases} \mathbf{x}_{1:d} & = \mathbf{y}_{1:d}, \\ \mathbf{x}_{d+1:D} & = (\mathbf{y}_{d+1:D} - t(\mathbf{y}_{1:d})) \odot \exp(-s(\mathbf{y}_{1:d})). \end{cases} \tag{9}$$

Condition 2: "Its Jacobian determinant is easy to compute." The Jacobian is a lower triangular matrix.

$$\det(\mathbf{J}) = \prod_{j=1}^{D-d} \exp(s(\mathbf{x}_{1:d}))_j = \exp\left(\sum_{j=1}^{D-d} s(\mathbf{x}_{1:d})_j\right). \tag{10}$$

The exact log-likelihood of input data $\log p(\mathbf{x})$ becomes tractable. As a result, the training criterion of flow-based generative model is simply the negative log-likelihood (NLL) over the training dataset D:

$$\mathscr{L}(\mathscr{D}) = -\frac{1}{|\mathscr{D}|}\sum_{\mathbf{x}\in\mathscr{D}} \log p(\mathbf{x}). \tag{11}$$

Flow-based inference processes have the potential to enable various tasks such as data generation, data forecasting, and latent variables inference. Wang et al. [50] proposed a deep image fusion network (DFN) that uses a dual-deep CNN-based fusion framework to guide the output generations directly by integrating layered feature maps. The network utilizes a weighted gradient flow-based fusion strategy to drive network optimization and formulate the fused outputs. This approach explicitly fuses salient representations from different sources, which distinguishes it from other implicitly generative models.

Simultaneous localization and mapping (SLAM) has been a widely studied problem in the robotics and computer vision communities for years. Good motion prediction and feature correspondence are vital for the SLAM problem, especially when visual information is ambiguous. Yan et al. [82] discussed how state-of-the-art SLAM algorithms benefit from flow-based methods and summarized recently emerging techniques, including learning techniques, sensor fusion, and continuous-time trajectory modeling. Multi-sensor fusion is a major advantage of SLAM, as it allows for incorporating other sensors such as inertial and event-based cameras to compensate for the shortcomings of visual observations. Moreover, the high frame rate is well-suited for flow-based SLAM problems, as it ensures a good linear characteristic.

The similarities of these generative models are that they adopt random noise (usually Gaussian distribution) in the data generated modeling and measure the difference of the distribution between noise and training data when modeling the distribution. Each method needs to make trade-offs on balance, including run time, heterogeneity, and architectural restrictions.

3.3. Discriminative models

Discriminative models, also referred to as conditional models, directly map inputs to outputs. As mentioned in the previous subsection, generative models focus on explaining how data is generated using Bayesian theorem. In contrast, discriminative models aim to identify the decision boundary by modeling the conditional probability to distinguish one class from another in a dataset. There are two primary types of discriminative models:

(1) Directly model the mapping from the input space X to the output space Y, that is, learning function h,

$$h : X \to Y, \quad s.t.\ y = h(x) \tag{12}$$

(2) Modeling the conditional probability $P(y|x)$, and then classify it according to the Bayesian risk minimization criterion,

$$y = \arg\max_{y\in\{-1,1\}} P(y|x). \tag{13}$$

This type of discriminative model focuses on predicting data labels without making any assumptions about the data points. They cannot learn from unlabelled data nor generate new data

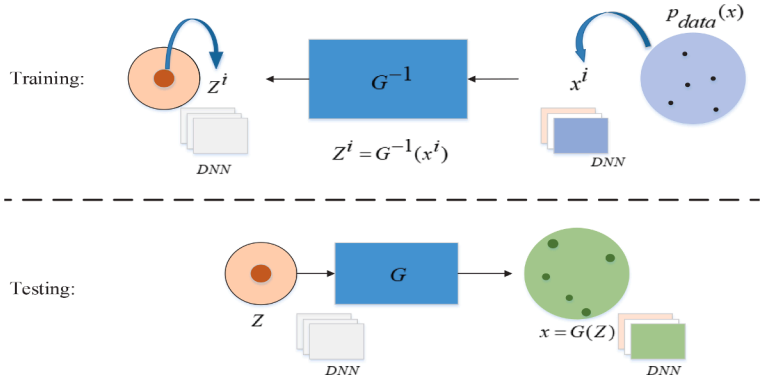


Fig. 5. The general representation of flow.

instances, but they are generally more robust to outliers in supervised tasks. Deep discriminative models have been proposed for many multi-modal data fusion applications, including temporal data, video, image, text, and feature dependencies that require effective learning.

An example of the application of deep discriminative models is in human activity recognition. Wu et al. [51] proposed a novel method called deep dynamic neural network (DDNN) for multi-modal human gesture recognition. This semi-supervised hierarchical dynamic framework is based on hidden Markov model (HMM) for simultaneous gesture segmentation and recognition, using bone joint information, depth, and RGB images as multimodal inputs. Unlike traditional methods that rely on complex manual feature construction, their approach uses Gauss-Bernoulli deep belief network (DBN) to deal with skeleton dynamics, and a 3D convolutional neural network to manage and fuse batches of depth and RGB images. This purely data-driven approach demonstrated the potential for deep learning techniques to explore multi-modal data further.

Deep discriminative models have also been applied to image feature fusion for various tasks. For instance, Yan et al. [52] proposed a trainable end-to-end deep discriminative model for distinguishing natural images from recolored images by utilizing a CNN-based deep structure that includes three feature extraction modules and a feature fusion block. They also used inter-channel images, illumination maps, and input images as inputs to improve the effectiveness of forgery detection in their proposed system.

The problem of RGB-D scene recognition has become more prominent with the widespread development and application of depth sensors [83, 84]. Zhu et al. [53] proposed a discriminative multi-modality framework for RGB-D scene recognition fusion that considers both inter-modal and intra-modal correlations of all samples and regularizes learned features to make them discriminative and compact. The proposed model uses CNN as the basic layer and the results of the multi-modal layer can be back-propagated to the lower CNN layer, updating the parameters in each network layer iteratively until convergence.

Discriminative deep learning models for multi-modal data fusion have been proposed for various applications, such as target tracking [56, 85], sparse representation [86], situational understanding [55], and disease diagnosis [54].

3.4. Algorithm unrolling models

Integrating domain knowledge into deep learning-based inference structures can be advantageous for addressing specific problems or datasets, such as structural data, data on specific manifolds, and complex multi-modal data with high-order interconnections. This approach provides an alternative perspective to tackle data fusion problems.

In the 2010s, Gregor et al. [87] proposed learning fast approximations of sparse coding, which extended the computational graph of the iterative algorithm used to solve the traditional sparse coding problem into a DNN, and provided a proximal unrolling framework. The unrolling method treats the iterative optimization of a given continuous model as a dynamic system. To implement the unrolling method, a learnable unrolling unit must first be designed from domain knowledge and embedded into the iteration loop.

Figure 6 provides an illustration of the algorithm unrolling process. Typically, algorithm unrolling involves repetitive analytic operations, represented abstractly as an iterative function h (left). By mapping each iteration $h^l, l = 0, \ldots, L-1$ into a single network-layer $z^l, l = 0, \ldots, L$, and stacking a finite number of layers together, an unrolling deep network can be generated (right). The parameter vector $\theta^l, l = 0, \ldots, L-1$ (which includes model parameters and regularization coefficients) is updated during each iteration h. These parameters can be determined using real datasets by training the network end-to-end instead of cross-validation or analytical derivation.

In the field of multi-modal sensor fusion, Zhao et al. [16] proposed a method for fusing infrared and visible images, called "Algorithm Unrolling Image Fusion (AUIF)," which combines the prior information of traditional optimization models and the strong feature extraction capability of DL. The AUIF model begins with the iterative formulas of two traditional optimization models and is designed to accomplish two-scale decomposition, separating the low-frequency background information and the high-frequency texture and gradient information from the source image. The number of parameters required for the AUIF model is lower than that required for conventional ones.

Juan et al. [57] proposed a DL architecture called LADMM-Net based on the algorithm unrolling method for various spectral image fusion tasks. This architecture overcomes the high running time of hyper-spectral (HS) and multi-spectral (MS) compression measurements in traditional methods and addresses the shortcomings of choosing representation transformation. The proposed architecture casts each iteration of a linearized version of the multiplier alternating direction algorithm into a CNN-based structure to form a deep network. This method also estimates the high-frequency image components contained in the auxiliary variables and Lagrangian multipliers, and is evaluated on the Harvard and CAVE datasets. Similar optimization ideas crop up by being embedded in RNN [88,89].

To summarize, algorithm unrolling involves three key steps: designing an iterative format, embedding learnable modules, and theoretical analysis. Current research has primarily focused on the first two steps. However, the use of learnable modules to perform expansion and calculation of the iterative format can result in inaccurate results and compromise the theoretical properties of the original numerical algorithm. Therefore, it is essential to conduct further theoretical analysis of the unrolling process to ensure its reliability and effectiveness.

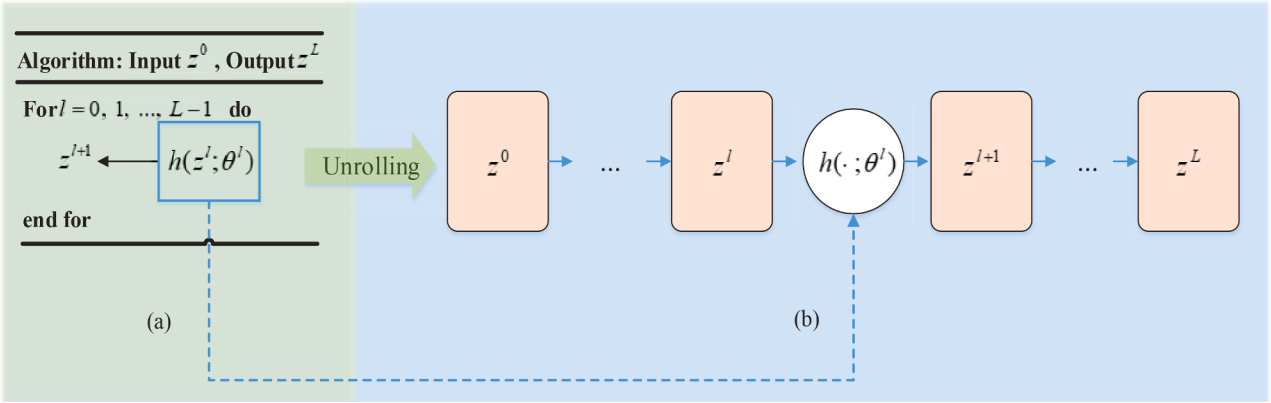


Fig. 6. The schematic diagram of algorithm unrolling.

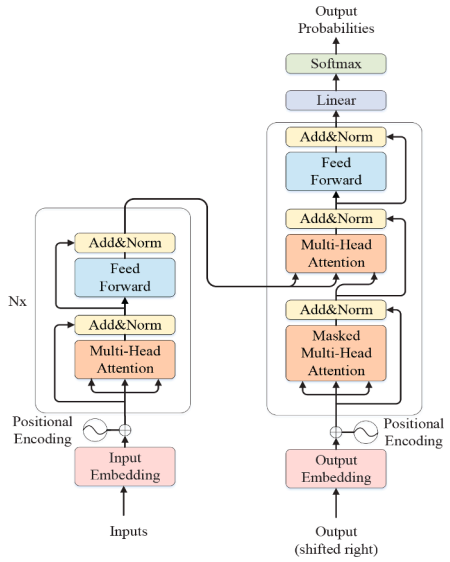


Fig. 7. The architecture of Transformer model [90].

3.5. Transformer models

The Transformer model was originally proposed by Vaswani et al. [90] to enhance the efficiency of machine translation through its utilization of the self-attention mechanism and position encoding. The architecture of the Transformer is shown in Fig. 7.

Each Transformer block can be expressed as:

$$\begin{aligned} X^{'} &= \text{LayerNorm}(\text{MultiheadAttention}(X, X)) + X, \\ X_B &= \text{LayerNorm}(\text{PositionFFN}(X^{'})) + X^{'}. \end{aligned} \tag{14}$$

where X is the input of the Transformer block and X_B is the output of the Transformer block. Note that the MultiheadAttention() function accepts two argument tensors, one for query and the other for key-values. If the first argument and second argument are the same input tensor, this is the MultiheadSelfAttention mechanism.

Recent advancements in multi-modal derivatives of large language models have highlighted the potential of Transformers in multi-modal fusion models [91–93]. Prakash et al. [58] developed TransFuser, a multi-modal fusion Transformer that utilizes multiple self-attention modules and processes single-view RGB images and Lidar data to generate a concise representation of the 3D environment. Similarly, Yasuda et al. [94] proposed MultiTrans, a Transformer-based multi-sensor fusion system that can handle multiple sensors and extract a joint representation by utilizing self-attention to identify useful features across different modalities.

Moreover, the Transformer's long-term dependency modeling ability has been extended in the field of rotating machinery fault diagnosis by Weng et al. [95] developed a multisensor fusion Transformer (MsFT) that utilizes a multisensor embedding generator to adaptively fuse multiple signals and generate learnable embeddings with positional information. MsFT also employs an improved Transformer with a local learning unit that addresses the limitations of traditional Transformers in local feature extraction.

The issue of transferability poses a significant challenge for multi-modal learning using Transformers, as it involves the task of transferring models across various datasets and applications. Multi-modal Transformers face two key efficiency problems: Firstly, due to their large parameter capacity, they require extensive training data, making them data-hungry. Secondly, their time and memory complexity increases quadratically with the input sequence length, due to self-attention, which limits their efficiency.

4. Applications and public dataset for multi-sensors fusion

This section discusses several popular applications based on deep multi-sensor fusion learning, including medical diagnosis, automotive driving, remote sensing, and intelligent robotics. We also present popular datasets of the above applications. Moreover, we list numerous public datasets commonly used for multi-sensor deep fusion technology, as shown in Table 4.

4.1. Medical diagnosis

Continuous updates to medical imaging equipment enrich image modalities and significantly enhance the accuracy of disease diagnosis in clinical settings. Medical imaging modalities such as ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and digital subtraction provide intuitive image information for specific parts of the human body. Wells et al. [113] from Harvard Medical School had highlighted the use of DL for medical image analysis tasks as a development trend in this field.

However, due to differences in methods and devices, the features generated from these images are heterogeneous. For example, CT is effective in displaying skeletal information but may not present structural details such as soft tissue clearly. MR images are suitable for soft tissue details, but bone detection results have outstanding defects. PET images provide a wealth of metabolic information with low resolution.

Table 4
Public multi-modal dataset for DL data fusion.

Dataset Name	Modalities of Data	Application	Reference	Year
KITTI	Video, IMU, 3D-Lidar, GPS data	Autonomous driving	Geiger et al. [96]	2013
ChaLearn	RGB-D, audio, skeletal pose	Human activity recognition	Escalera et al. [97]	2014
mHealthDroid	Electrocardiogram (eCG), magnetometer and gyroscopes data	Health monitoring	Banos et al. [98]	2015
Pinterest Multi	Images and text	Multi-modal word embeddings	Mao et al. [99]	2016
RobotCar	Cameras, Lidar, GPS and IMU data	Autonomous driving	Maddern et al. [100]	2016
MHRI	Multi RGB-D, video and audio	Human-robot interaction	Pablo et al. [101]	2017
FCVID	Video and audio	Action recognition	Jiang et al. [102]	2017
KAIST Urban	Lidar, fiber optic gyro (FOG) and IMU data	Robotics, SLAM	Jeong et al.[103]	2018
NCALM	MS-Lidar and HS data	Land use and cover classification	Le Saux et al.[104]	2018
ApolloScape	Camera images and point clouds	Autonomous vehicles navigation	Zhou et al. [105]	2019
Rosario	Wheel odometry, IMU, stereo camera and GPS-RTK data	Localization and mapping in agricultural environments	Taihú et al. [106]	2019
Nuscenes	Camera, radar and Lidar data	Autonomous driving	Caesar et al. [21]	2020
Muse	Lidar, camera, WiFi, RGB-D, UWB-Range and IMU data	Robot, SLAM	Lavish et al. [107]	2020
SpaceNet 6	HS-, MS-, panchromatic, infrared and SAR images	Object detection and segmentation in remote sensing	Shermeyer et al. [108]	2020
Fusion-DHL	WiFi, IMU, and floor plan data	Indoor location	Herath et al. [109]	2021
Stcrowd	Lidar point clouds and images	Pedestrian detection, tracking	Cong et al. [110]	2022
ActionSense	RGB cameras, depth camera, and microphones	Scene reasoning, action planning	DelPreto et al. [111]	2022
WEAR	IMU data and video	Human activity recognition	Bock et al. [112]	2023

Multi-image fusion technologies based on DL are effectively used to reconstruct three-dimensional images using two-dimensional slices to create intuitive and precise effects. Current medical image analysis via DL focuses on multi-modal medical image fusion methods for classification and recognition, localization and detection, tissue organs, and lesion segmentation.

Various approaches have been developed for the inference mechanism in medical image analysis [127–130], including pixel-level image fusion [23], convolutional sparse representation (CSR) [126,129,131], stacked auto-encoder (SAE) [132,133], CNN, and deep Boltzmann machine (DBM) [134,135]. One of the earliest applications of DL in medical image analysis is image screening, which involves image-level classification that takes one or more inspection images as input and predicts the symptom type or diagnosis variables for various diseases or their severity. DL models initially focused on SAE, DBM networks, and unsupervised pre-training methods to analyze neuroimaging, such as Alzheimer's disease (AD) or mild cognitive impairment (MCI) [136].

Suk et al. [137,138] used DBM and SAE, respectively, to extract potential hierarchical features from 3D neuroimaging image blocks and construct an AD/MCI diagnosis model. Both methods were verified on the ADNI dataset, and the results showed that the classification performance of the model using SAE is better than that of DBM. CNN has become a standard technology in image inspection classification. Gao et al. [139] fused two 2D CNNs to extract temporal and spatial information features of echo cardiograms and classify the echo cardiogram viewpoint to assist in diagnosing heart disease. Some works combine CNN with RNN, such as Gao et al., who use CNN to extract low-level local feature information in slit-lamp images and combine RNN to extract high-level features for classifying nuclear cataracts.

Some scholars combine CNN with other DL models for diagnosis classification. For example, Kallenberg et al. [140] combined the characteristics of CNN and SAE and use an unsupervised pre-training convolutional sparsely auto-encoder (CSAE) to achieve breast density image segmentation and breast risk assessment. Van et al. [141] used convolutional restricted Boltzmann machines (CRBM) by combining the distinguishability of CNN and the generative characteristics of RBM to analyze lung CT.

Classifying each pixel is key to detecting an image of interest or lesion. Multi-processing stream CNN can integrate different perspective profile information or multi-modal image data for this purpose [147]. Albarqouni et al. [148] used a multi-scale CNN scheme to detect mitosis in breast cancer pathological images. Chen et al. [149] used two 2D deep features to approximate the features of 3D medical images and combined them with a supported vector machine (SVM) classifier to achieve the automatic detection of cerebral microbleeds (CMBs) with sensitivity-weighted imaging (SWI). Unlike the traditional CNN method that takes the original image as input, Li et al. [150] used Sobel edge contour features and Gabor texture features as input data. They then used CNN for feature fusion and depth feature extraction, which improved the accuracy of C-arm X-ray image automatic detection of the lumbar spine.

In addition, objective metrics for image fusion are divided into subjective and objective evaluation methods. Subjective evaluation relies on personal judgment based on human eyes, but emotional factors can affect the evaluation result. Objective evaluation quantitatively simulates the human visual system (HVS) for image quality perception by measuring relevant indicators. The existing objective indicators are categorized into four categories: information theory-based, image feature-based, image structure similarity-based, and human perception-based. Various indicators are elaborated, as shown in Table 5.

To compare and evaluate the performance of existing methods, Singh et al. [126] conducted numerous experiments on 14 pairs of CT-MR and 29 pairs of MR-SPECT multi-modal images taken from the Harvard Whole Brain Atlas [118]. The experimental results are presented in tables for quantitative comparison. Table 6 provides a comprehensive quantitative comparison of the complete CT-MR dataset concerning the average and standard deviation of all performance metrics. Table 7 presents an assessment of all fusion methods by computing the quantitative metrics of 29 pairs of MR-SPECT images.

4.2. Automative driving

Automated driving cars rely on environmental sensors such as radar, camera, ultrasonic, and Lidar to perceive their surroundings for safety functions. However, each sensor has limitations and cannot provide complete information about the vehicle's surroundings. Additionally, the different ways of recording information among sensors often result in complementarities [151,152]. Sensor fusion technologies act as a "brain" in autonomous driving systems that try to prevent blind spots from emerging when perceiving the vehicle's surroundings.

The multi-modal fusion framework, as shown in Fig. 8, involves feature extraction and multi-channel feature fusion in the multi-modal fusion module, taking into account the inconsistency of the feature space of the multi-sensor data. The fusion process combines prior knowledge from consecutive frames to enhance the overall system's ability to recognize the environment. Hence, the accuracy and robustness of tasks such as target detection, target tracking, and SLAM are improved in abnormal exposure, rain, and fog scenes.

Target detection methods using multi-modal data fusion have been continuously proposed. Cho et al. [171] proposed an expert fusion method of radar point cloud, Lidar point cloud, and camera image for vehicle detection and tracking. Schlosser et al. [172] suggested a fusion method based on HHA (horizontal disparity, height above ground, and angle) sampling using Lidar point cloud and camera image to form a

Table 5
Metrics for performance evaluation in image fusion.

Category	Evaluation parameters	Expressions	↑/↓
Information theory-based [114]	Entropy (EN)	$EN = -\sum_{i=0}^{L-1} p(i)\log_2 p(i)$ Entropy values refer to the amount of information present in the fused image.	↑
	Mutual information (MI)	$MI = I_{FR}(f;r) + I_{FS}(f;s)$ where $I_{FX}(f;x) = \sum_{f,x} p_{FX}(f,x)\log_2 \frac{p_{FX}(f,a)}{p_X(x)p_F(f)}$ MI values mean the amount of feature information in the fused image transferred from the source image.	↑
Structural similarity-based[115]	Root mean squared error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathrm{Ir}(i,j) - If(i,j))^2}$ The root mean square error of the fused image and the original image	↓
Image feature-based [116]	Edge intensity (EI)	$EI(F) = \frac{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} s_x(i,j)^2 + s_y(i,j)^2}}{M * N}$ where s_x, s_y are the results of *Sobal* operator convolution.	↑
	Standard deviation (SD)	$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ where μ is average value $\bar{x}$.	↑
Human perception inspired [117]	Chen-Blum metric (Q_{CB})	$Q_C(x,y) = \lambda_T(x,y)Q_{TF}(x,y) + \lambda_V(x,y)Q_{VF}(x,y)$ Two reference images of the original image are required.	↑

↑: is used to indicate that higher values correspond to better performance. ↓: is used to indicate that lower values correspond to better performance.

Table 6
Averaged evaluation metric values for 14 pairs of CT-MR neurological images (EN, MI, etc. are introduced in Table 5).

Fusion methods	Reference	Performance measures			
		EN	MI	SD	EI
Source MR	Harvard Whole Brain Atlas [118]	3.8803 ± 0.3366	–	59.3509 ± 4.9843	–
Source CT	Harvard Whole Brain Atlas [118]	3.0647 ± 0.3035	–	83.9079 ± 5.4953	–
NSCT-based	Das et al. [119]	4.8081 ± 0.3713	3.0086 ± 0.1102	86.9047 ± 4.4341	0.4785 ± 0.0490
GFF	Li et al. [120]	4.6501 ± 0.2949	1.3093 ± 0.2232	64.0659 ± 3.7542	0.4458 ± 0.0530
MST-SR	Liu et al. [121]	5.1369 ± 0.2105	2.8894 ± 0.1578	82.8205 ± 3.9480	0.4824 ± 0.0416
Adaptive-PCNN	Ganasala et al. [122]	4.9135 ± 0.2426	2.8656 ± 0.1751	84.0917 ± 4.3443	0.5199 ± 0.0513
Fuzzy-PCNN	Yang et al. [123]	5.0273 ± 0.2126	2.9026 ± 0.2145	84.9058 ± 4.3424	0.5477 ± 0.0436
CT-MR fusion	Singh et al. [124]	5.1529 ± 0.2378	3.1594 ± 0.1637	88.5082 ± 4.4684	0.5578 ± 0.0465
NSCT-Type2 Fuzzy	Yang et al. [125]	4.8686 ± 0.4170	2.9393 ± 0.1469	85.8544 ± 4.2374	0.5659 ± 0.1328
MMISF	Singh et al. [126]	5.5652 ± 0.2011	3.4027 ± 0.2128	89.4922 ± 3.8922	0.5881 ± 0.0438

Table 7
Averaged evaluation metric values for 22 pairs of MR-SPECT neurological images (EN, STD, etc. are introduced in Table 5).

Fusion methods	Reference	Performance measures				
		EN	STD	MI	SSIM	EI
Region-SPV	Sudheer and Bindu [142]	4.1225 ± 0.6795	68.212 ± 8.8127	0.4114 ± 0.0222	0.3152 ± 0.0635	0.6669 ± 0.1316
Directive-NSCT	Bhatnagar et al. [143]	4.3566 ± 0.6876	65.146 ± 7.5963	0.1948 ± 0.0135	0.2455 ± 0.0495	0.1394 ± 0.0253
GFF	Li et al. [120]	4.3049 ± 0.6636	57.343 ± 9.1360	0.3855 ± 0.0515	0.3777 ± 0.0607	0.6648 ± 0.0621
OMP	Yang and Li [144]	4.4827 ± 0.7329	65.131 ± 8.3147	0.4130 ± 0.0472	0.3852 ± 0.0632	0.6708 ± 0.0619
MST-SR	Liu et al. [121]	4.6985 ± 0.6626	65.822 ± 7.4958	0.4440 ± 0.0302	0.3973 ± 0.0646	0.6773 ± 0.0650
Adaptive-PCNN	Ganasalaet al. [122]	4.7641 ± 0.6204	68.739 ± 8.2240	0.4031 ± 0.0516	0.3783 ± 0.0406	0.6753 ± 0.0673
Fuzzy-PCNN	Yang et al. [123]	4.5966 ± 0.6623	67.262 ± 8.5285	0.4447 ± 0.0349	0.4020 ± 0.0404	0.6842 ± 0.0645
NSCT-based	Yang et al. [145]	4.3564 ± 0.6686	68.009 ± 9.0919	0.4493 ± 0.0348	0.3376 ± 0.0598	0.6879 ± 0.0621
HMSD-GDGF	Zhu et al. [146]	4.8645 ± 0.6922	67.468 ± 7.3426	0.2132 ± 0.0211	0.3988 ± 0.1305	0.2362 ± 0.0366
NSCT-Type2 Fuzzy	Yang et al. [125]	4.8688 ± 0.6712	69.1135 ± 7.2812	0.2593 ± 0.0379	0.4165 ± 0.1227	0.6169 ± 0.1221
MMISF	Singh et al. [126]	5.1978 ± 0.5431	69.511 ± 8.6855	0.4928 ± 0.0150	0.4748 ± 0.0458	0.7258 ± 0.0606

6-channel extended image for pedestrian detection. DOU et al. [173] proposed a fusion method of forwarding projection for the depth map and camera image to assist 3D detection on the point cloud through images. Other methods perform object detection after fusing point cloud with the image. Liang et al. [174] designed the dense fusion module to fuse data from four branches and then combined it with the output features of the point cloud and image feature extraction network. Liu et al. [175] fused low-level and high-level feature maps to enhance the positioning information of high-level features for semantic segmentation.

To overcome the challenge of effectively aligning transformed features from multiple modalities, Li et al. [176] proposed a method called InverseAug, that inverses geometric-related augmentations, such as rotation, to achieve accurate geometric alignment between Lidar points and image pixels. They also proposed LearnableAlign, which leverages cross-attention to dynamically capture the correlations between image and Lidar features during fusion.

Additionally, to handle inferior image conditions, Bai et al. [177] proposed TransFusion, a robust solution for Lidar-camera fusion with a soft-association mechanism. This method consists of convolutional backbones and a detection head based on a transformer decoder. The decoder adaptively fuses object queries with useful image features, leveraging both spatial and contextual relationships.

Although fully autonomous driving has not yet been realized, assisted driving vehicles are already equipped with various safety functions to protect drivers and passengers. These functions include front collision warning (FCW), emergency automatic braking (AEB), Lane change collision warning (LCW), lane departure warning (LDW), and lane hold assist (LKA), among others. Pedestrian and vehicle detection technologies are also rapidly evolving. Table 8 illustrates the types of sensors

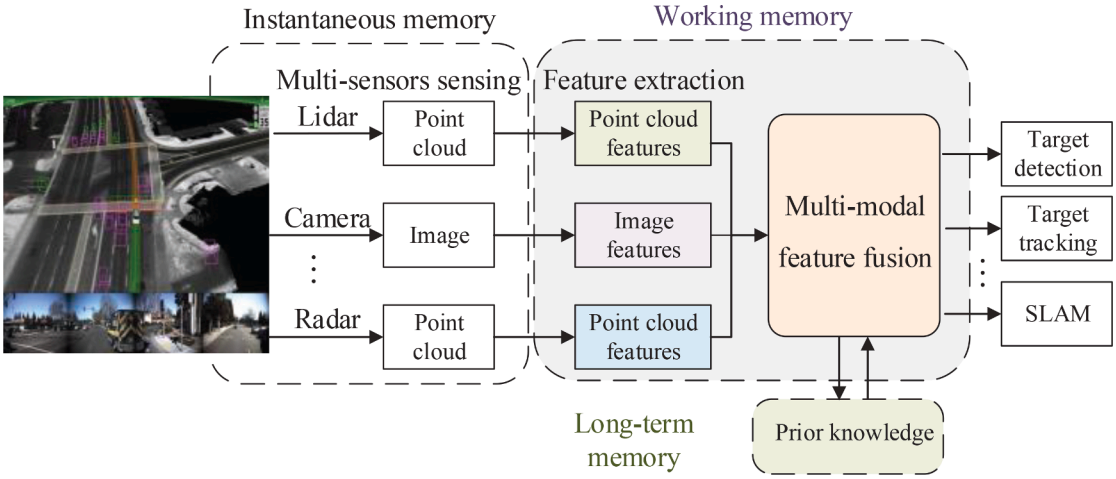


Fig. 8. The multi-modal fusion framework in autonomous driving.

Table 8
Sensors required to achieve different automatic driving functions.

Function	Ranging		Ranging accuracy (m)	Viewing angle (∼°)	Angular resolution (∼°)	Update frequency	Sensor data (Hz)
	Tracking (m)	Classification (m)					
FCW, AEB	50 ∼ 80	25 ∼ 50	0.5	30	⩽0.25	25	Lidar, radar
Pedestrian protection	40	30	0.1	60	0.25	12.5	Camera, Lidar
LCW, LCA	30	15	0.2	50	0.1	12.5	Camera
LDW, LKA	100	–	0.2	50	0.1	25	Camera
Start inhibit	5	5	0.2	⩽180	2	< 12.5	Camera
No signal light port assist	⩽190	⩽50	0.3	250	⩽0.25	12.5	Camera, Lidar, radar
With signal light port assist	⩽80	⩽20	0.3	250	⩽0.25	12.5	Camera, Lidar, radar
Collision protection	20	–	0.1	60	1	> 25	Radar
ACC	⩽200	⩽20	0.3	10 ∼ 20	⩽0.3	< 12.5	Radar
S&G	50	⩽20	0.1	⩽180	1	< 12.5	Camera, Lidar
Auxiliary system	5	–	0.1	⩽180	2	< 12.5	Ultrasound, camera

required to achieve various automatic driving functions. Assisted driving vehicles rely on the fusion of information from heterogeneous sensors to understand their environment and plan their behavior.

The CLEAR MOT metrics [178] are widely used to evaluate the accuracy of detection and tracking systems in automated driving. These metrics include multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), mostly tracked (MT), mostly lost (ML), identity switches (IDS), and fragmentation (FRAG), which reflect the tracking orientation characteristics of the system. Higher values of MT and lower values of ML, IDS, and FRAG indicate improved continuous tracking and reduced trajectory error. Table 9 provides a comparison of publicly available methods in the KITTI tracking benchmark.

4.3. Remote sensing

This subsection categorizes existing fusion methods in remote sensing into two groups: heterogeneous image fusion and heterogeneous sensor fusion. The former includes spatial-spectral fusion (such as pan-sharpening, HS, and MS fusion) and spatiotemporal fusion. The latter includes fusion methods such as Lidar-optical, SAR-optical, and RS-GBD, which integrate data from different imaging mechanisms. Feature-level and decision-level fusion are commonly used in heterogeneous fusion due to the differences in the imaging mechanisms. We provide examples of various fusion sub-fields and summarize classic literature in each direction to assist readers. Table 10 presents the performance of current algorithms for remote sensing image classification in the MUUFL dataset [179].

Optical imaging is used to capture panchromatic, MS, and HS images. In the fusion process, MS images can supply spectral information to panchromatic images or spatial information to HS images. SAR echoes record the backscattered energy of ground objects, offering insights into surface characteristics and dielectric properties. Thermal infrared remote sensing images can reveal the temperature distribution of ground objects but tend to have lower spatial resolution. The integration of multi-modal remote sensing images has facilitated the advancement of DL algorithms in remote sensing, particularly in tasks such as land cover and use [190], object detection [191,192], soil moisture estimation [193], and scene classification [194,195]. Currently, most DL-based multi-modal fusion techniques in remote sensing are focused on these applications.

MS, HS, and Lidar images are typically fused through feature-level and decision-level techniques [196–200]. Chen et al. [196] utilized CNN to extract detailed features from HS images and Lidar data, such as time, spectrum, angle, and terrain features, followed by a classifier to

Table 9
Comparison of publicly available methods in the KITTI tracking benchmark.

Method	Input modality	Metrics					
		MOTA ↑	MOTP ↑	MT ↑	ML ↓	IDS ↓	FRAG ↓
CEM [153]	2D images + target dynamics	51.94%	77.11%	20.00%	31.54%	125	396
RMOT [154]	2D camera images + motion context	52.42%	75.18%	21.69%	31.85%	50	376
TBD [155]	Vehicle tracklets + semantic scene labels + scene flow, etc.	55.07%	78.35%	20.46%	32.62%	31	529
mbodSSP [156]	Video sequence	56.03%	77.52%	23.23%	27.23%	**0**	699
SCEA [157]	Video sequence + structural motion	57.03%	78.84%	26.92%	26.62%	17	461
ODAMOT [158]	Video images	59.23%	75.45%	27.08%	15.54%	389	1274
NOMT-HM [159]	Video sequence + target dynamics	61.17%	78.65%	33.85%	28.00%	28	241
LP-SSVM [160]	Video sequence	61.77%	76.93%	35.54%	21.69%	16	422
RMOTT [154]	2D camera images + motion context	65.83%	75.42%	40.15%	9.69%	209	727
NOMT [159]	Video sequence + point trajectories + target dynamics	66.60%	78.17%	41.08%	25.23%	13	**150**
DCO-X * [161]	Video sequence + point trajectories	68.11%	78.85%	37.54%	14.15%	318	959
NOMT-HM * [159]	Video sequence + target dynamics	75.20%	80.02%	50.00%	13.54%	105	351
SCEA* [162]	Video sequence	75.58%	79.39%	53.08%	11.54%	104	448
MDP [163]	Multi-video frame	76.59%	82.10%	52.15%	13.38%	130	387
MCMOT-CPD [164]	Video sequence	78.90%	82.13%	52.31%	11.69%	228	536
DSM [165]	RGB images + Lidar point clouds	76.15%	83.42%	60.00%	8.31%	296	868
Autotrack [166]	RGB images + Lidar point clouds + GPS/IMU data	**82.25%**	80.52%	**72.62%**	**3.54%**	1025	1402
Complexer-YOLO [167]	Visual point-wise + semantic point clouds	75.70%	78.46%	58.00%	5.08%	1186	2092
PointTrackNet [168]	Lidar point clouds	68.23%	76.57%	60.62%	12.31%	111	725
2D-3DModel [169]	Image + world-space information	75.39%	79.25%	49.85%	10.31%	165	660
Cross-MMT [170]	2D RGB + 3D point clouds	79.93%	**84.77%**	66.00%	10.00%	278	716

↑: is used to indicate that higher values correspond to better performance. ↓: is used to indicate that lower values correspond to better performance.

Table 10
OA, AA and κ values on the MUUFL dataset (in %) by fusing HSI and Lidar data.

Class	RF[180]		SVM [181]		RNN[182]		2-CNN[183]		FusAtNet[184]		CoupledNet[185]		HWRN[186]		HybridSN[187]		EndNet[188]		MFT[18]		MorphCNN[189]	
	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L	H	H+L
1	98.08	98.09	98.08	98.35	96.07	96.43	97.33	97.58	98.74	98.15	97.68	97.88	98.33	98.16	97.12	98.12	89.03	91.82	97.61	97.90	98.37	97.46
2	76.50	77.19	57.32	55.52	80.69	79.29	82.27	82.29	79.78	80.32	76.15	83.50	79.51	74.26	86.16	82.27	77.11	84.61	92.51	92.11	90.08	90.38
3	85.89	85.71	73.87	72.89	85.25	86.25	86.12	91.20	86.67	89.08	86.12	86.92	88.22	88.36	85.28	89.79	78.66	74.13	92.12	91.80	90.74	89.72
4	83.97	83.91	48.53	39.53	87.55	88.87	90.83	92.10	85.93	88.24	91.93	92.10	94.00	93.31	87.14	93.37	77.80	86.39	92.83	91.59	93.83	96.13
5	92.82	93.23	81.45	80.12	88.00	90.10	92.72	95.57	93.90	94.23	94.25	92.47	95.68	94.27	92.08	92.79	79.03	82.15	94.31	95.60	94.49	95.94
6	89.16	88.71	28.44	30.92	78.78	82.84	81.71	91.19	75.62	89.84	65.46	95.26	74.04	88.48	75.62	94.35	90.14	96.61	88.56	88.19	97.06	97.97
7	82.60	82.69	58.65	54.40	85.05	85.66	87.88	90.66	85.80	87.88	87.74	85.57	89.67	90.19	90.19	91.65	77.03	82.39	92.68	90.27	86.23	92.92
8	90.46	90.38	85.96	86.37	87.45	87.11	96.28	96.13	93.25	95.85	96.76	97.23	97.42	96.96	92.51	92.45	93.32	94.97	97.08	97.26	96.92	96.23
9	44.83	44.83	25.45	18.61	59.49	61.62	65.27	56.00	55.92	58.20	57.14	57.52	60.79	63.98	60.56	59.27	30.64	50.78	59.80	61.35	30.09	39.74
10	13.21	16.66	7.47	07.47	28.73	42.52	45.40	00.00	09.77	27.58	18.96	17.81	39.65	37.35	29.88	25.28	27.20	31.22	12.45	17.43	00.00	00.00
11	55.29	55.29	00.00	00.00	79.60	55.29	65.49	68.62	80.78	85.49	75.68	74.11	91.76	89.41	62.35	65.88	76.82	86.79	71.09	72.79	65.88	82.35
OA	90.18	90.27	81.65	80.74	89.30	89.78	92.09	93.05	91.77	92.64	91.64	92.42	93.18	92.66	91.41	92.59	83.36	86.31	94.18	94.34	93.16	93.45
AA	73.89	74.25	51.38	49.47	77.88	77.82	81.03	78.30	76.92	81.35	77.08	80.03	82.64	83.16	78.08	80.47	72.43	78.35	81.00	81.48	76.70	79.89
$\kappa(\times 100)$	86.88	87.01	74.33	72.89	85.86	86.47	89.53	90.79	88.97	90.18	88.89	89.93	90.94	90.25	88.62	90.19	78.36	82.17	92.30	92.51	90.90	91.33

Overall accuracy (OA): represents the proportion of correctly classified test samples versus all test samples. Average accuracy (AA): represents the average of class-wise accuracy. κ: reflects the degree of agreement between the generated classification maps of the considered model and the provided ground truth.

obtain a fusion classification result. Several studies have explored the fusion of HS and Lidar data for complex environmental monitoring tasks, such as estimating above-ground wetland vegetation biomass using MS, HS, and Lidar data, estimating shallow water depth and turbidity through analysis of the voxelized water depth FWL method [201], and classifying plant function types and bare soil cover by combining UAV multi-spectrum and structural motion photogrammetry [197]. These studies provide valuable insights into the effective integration of MS, HS, and Lidar data for complex environmental monitoring tasks.

In recent years, a variety of technologies have emerged for fusing remote sensing data. Bhagat et al. [202] proposed a space-constrained adversarial method that extracts deep features from the spatial characteristics of visual images and spectral aspects of infrared images to obtain a more detailed and comprehensive remote sensing feature representation. Cui et al. [203] proposed a cross-modal image-matching network called CMM-Net, which considers specific and shared information to obtain an invariant modality feature representation. The consistent modality feature representation of thermal infrared and visible light images is learned, and high-level semantic information between different modalities is extracted. These studies provided valuable insights into fusing remote sensing data and improving image quality and feature representation for target detection and ground object classification.

Wang et al. [204] were the first to utilize Transformers for semantic segmentation in high-resolution remote sensing imagery. Meanwhile, Xu et al. [205] found that the vision Transformer (ViT) models produced satisfactory results in HS image classification tasks. Roy et al. [18] proposed a method for using feature embeddings from additional multimodal data sources in HSI classification. By tokenizing the data into CLS and HSI patch tokens and introducing a novel attention mechanism, their approach improved information exchange between HSI tokens and the CLS token, highlighting the potential benefits of integrating other sources of multimodal data into transformer-based models. To address the diversity of objects and the cross-modal gap between different images, He et al. [206] proposed a Transformer-Induced Hierarchical Graph Network (GraFNet), which exploits the structural information of land cover categories to learn joint representations and introduces an attentive heterogeneous information aggregation mechanism to capture modality-specific object-object interaction patterns in a topology-aware environment. Additionally, GraFNet employs modality hierarchical dependency modeling to improve cross-modal compatibility.

Furthermore, the IEEE Geoscience and Remote Sensing Society (IEEE GRSS) has been organizing an annual multi-source remote sensing data fusion competition since 2006 to encourage and promote research in this field.

4.4. Intelligent robots

Understanding human behavior is crucial for effective human-computer interaction and for developing intelligent robots that can navigate and interact with the world around them [207–211]. These robots are designed to recognize various characteristics of their environment, such as size, color, distance, material, shape, direction, and sound, using sensors such as vision sensors, distance measuring sensors, auditory sensors, and tactile sensors. In recent years, multi-modal sensor fusion algorithms have emerged as a promising area of robotics, with significant advancements being made in this field. These algorithms have numerous applications, including interactive robots [212], medical robots [213], industrial robots [214], and human-robot co-driving [215].

Intelligent interactive robots are increasingly deployed in various domains, such as education, entertainment, public services, and smart homes. To enhance their functionality and capabilities, researchers have proposed various techniques and methods. One such approach is the

deep supervised learning method proposed by Cuayahuitl et al. [210], which involves learning perception and interaction, as well as automatic simulation, by inputting hundreds of sample images, multi-modal dialogues, and physical demonstrations of robot operation. Another recent development in this field is the context-aware companion chat robot technology proposed by Kuo et al. [216]. This approach uses RNNs with gated recurrent units (GRU) instead of the traditional long and short-term memory (LSTM) structure. The VGG16 model is selected as the image information feature extractor, and multi-sensor image information is fused with the sound signal to provide an appropriate response to the user.

In the field of smart medical services, various approaches had been proposed to enhance the performance of medical human-computer interaction scenarios. Lin et al. [217] designed a multi-sensor fusion approach for body sensor networks (BSNs) using interpretable neural networks. The aim is to improve the performance of fusion decision-making in medical scenarios. Meanwhile, Qi et al. [218] proposed a multi-layer RNN composed of an LSTM module and a dropout layer (LSTM-RNN) for remote operation of surgical robots. Moreover, the field of medical rehabilitation has also seen various research, including rehabilitation robots [219] and intelligent prostheses [220].

Industrial intelligent robots rely on a fusion of video images, sounds, electromagnetics, and other sensory inputs to reason and perform tasks such as materials handling, parts manufacturing, inspection, and assembly. Wei et al. [221] proposed a method to reduce false alarm detection from Lidar to enhance industrial automation environments and avoid robot collision avoidance. The method involves projecting the beacon into the optical image. The resulting projection is then fused with the Lidar feature space to verify detection. This approach demonstrates how sensing data fusion can be leveraged to improve the performance and safety of industrial intelligent robots.

The SLAM technology is particularly useful in real-time navigation for robots, allowing them to overcome localization and mapping problems and move effortlessly. The field of multi-sensor fusion has seen significant research progress in this area, with Lidar, camera, and IMUs being used as auxiliary sensors for SLAM. For example, Brossard et al. [222] have developed a network that uses deep learning and variational inference to correct the system's dynamic propagation and observation model, thereby addressing the drift problem of fast mobile robots. Such approaches demonstrate the potential of multi-sensor fusion in enabling accurate and efficient real-time navigation for robots.

Sim-to-real (Sim2Real) refers to the process of transferring learned motion strategies or control policies from a simulation environment to the real world. Typically, motor skills or control strategies are acquired through training and learning in a simulated environment. Then, multi-modal data captured by various sensors is input to reinforce skill replay or strategy control, allowing the robot to make optimal autonomous decisions in complex and unstructured environments. Table 11 presents the characteristics of five commonly used simulation platforms for robot motion control, including Gazebo [223], PyBullet [224], MuJoCo [225], V-REP [226], and Webots [227]. These platforms provide a means to simulate robot behavior and test control strategies before deploying them in the real world.

5. Open challenges and future prospects

Despite the extensive efforts and progress made in the field of multi-modal sensor fusion using DL techniques, a deeper exploration of several open challenges and potential research directions is necessary to achieve further advancements.

5.1. Open challenges

- Multi-modal fusion via DL promotes flexible expression and demonstrates how multi-source data flows across deep layers. However, these methods are typically chosen based on specific problems that have their own complexity and real-time requirements. For instance, deep generative methods enable learning of representations and formulation of fusion spaces at the middle hidden depths[44,49,50], while deep discriminative models directly map multi-inputs to single-fusing-outputs for classification tasks [55,56]. Both adaptive learning and algorithm unrolling have the advantage of injecting interpretable domain knowledge into DL-based iterative inference [40,57]. Choosing an appropriate multi-modal cognitive fusion framework adaptively is a challenging task.
- Multi-modal sensor fusion methods at the feature-level are overwhelmingly represented and account for more than 50% of all reviewed articles [42–50]. In contrast, the effectiveness of other types of fusion has not been fully explored. It is undeniable that deep learning has, in most cases, reduced the need for manual feature design. However, some works extract information and fuse it at the decision-level by independently modeling each modality, which can eliminate bias from heterogeneous features that may be learned [55, 56]. Hybrid-level fusion has been used in image and video analysis

Table 11
Comparison of commonly used robot simulation platforms.

Simulation platform	Language	OS	Physics engine	3D rendering engine	ROS compatibility	Scenarios
Gazebo	C+/Python	Linux/Mac OSX	ODE/Bullet/Simbody/ DART	OGRE	★★★★★	Autonomous navigation, multi-agent interactive control, etc.
PyBullet	Python	Linux/Mac OSX/ Windows	Bullet	TinyRend erer	★★★	Algorithm testing and validation in the robot continuous control tasks.
MuJoCo	C/C+/Python	Linux/Mac OSX/ Windows	MuJoCo	OpenGL	★★	Robot attitude control and dynamic mechanical analysis in a structured environment.
V-REP	Matlab/C/C+/Python/ Java	Linux/Mac OSX/ Windows	ODE/Bullet/Vortex/ Newton	Internal/External	★★★★	Robot motion planning in industrial scenarios, such as object grabbing, etc.
Webots	Matlab/C/C+/Python/ Java	Linux/Mac OSX/ Windows	ODE	OGRE	★★★	Bionic foot robot motion simulation, multi-robot collaborative control.

[54]. Overall, each method is only applied in a few scenarios, which significantly complicates generalization.

- While most multi-modal sensor fusion models consider fusion quality improvement to be a crucial element, there is often inadequate discussion on computational efficiency, as shown in Table 3. With some works not even evaluating this vital property due to the computational complexity of DL-based methods [40,46,49]. It is essential to discuss data quality and fusion quality systematically, and the corresponding metrics and criteria are also worth investigating.
- Throughout the survey, we have noticed that the concern for robustness in fusion methods has been under-considered. Based on Table 3, only a few pieces of research considered "Robustness," while some mentioned it but did not provide adequate information. They did not prove with experimental results whether the models are stable in an unstable environment [16,53,59]. Simply improving fusion accuracy and quality while ignoring robustness may result in a defective model.

5.2. *Future prospects*

Given the above open challenges, we propose some potential future prospects for readers below.

- DL models are highly effective at modeling complex, nonlinear, and large-scale data that can be challenging to express directly with traditional model-based methods. However, inference mechanisms are often more interpretable that rely on domain knowledge. Therefore, it is worthwhile to further explore the use of structural embedding to integrate data with domain information.
- DL-based multi-modal data fusion has numerous applications, each with specific requirements tailored to the task. While multi-modal data with diverse features can enhance our understanding of the world, it is common for some modal data to be missing in practical scenarios. Cross-modal learning offers a promising solution for transferring knowledge learned from multimodal data regions to scenarios where modal data is partially missing.
- In situations where partial modalities or partial abnormal burst inputs are missing, the fusion program may continue to operate instead of ceasing, which can negatively impact the model's performance. It is critical to ensure that the fusion model has reliable robustness to handle abnormal data. Partial modalities and adversarial examples can be used as input into the model to test its robustness since they introduce subtle interference to verify the model's resilience. Further research on adversarial examples is beneficial for advancing DL.
- Ground truth is often difficult to obtain in various areas such as object tracking, image fusion, etc., making it extremely challenging to evaluate the quality of the fused result. Therefore, the image fusion field needs to design no-reference metrics with greater characterization ability. These proposed metrics can be used on one hand to construct the loss function to guide higher-quality fusion. On the other hand, the newly designed metrics can also be used to fairly evaluate the fused results and encourage further research in the field of fusion.

6. Conclusion

In this work, numerous opportunities and challenges are thrown down to the area of data fusion. It comprehensively investigates the state-of-the-art advances in multi-modal data fusion based on prevalent DL, focusing mainly on model inference, technical differences, performance, limitations, and applications of existing works. Although the domain of multi-modal sensor fusion via DL is developing, there is still much potential for further research from the perspectives of both theories and applications. We hope that this review will provide readers with new prospects to explore suitable fusion techniques for their applications.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jing Liang reports financial support was provided by Institute of Scientific and Technical Information of Sichuan Province. Jing Liang reports financial support was provided by The 111 Project.

Data availability

No data was used for the research described in the article.

Acknowledgment

This work was supported by Sichuan Science and Technology Program under Grant 2023NSFSC0450, and the 111 Project under Grant B17008.

References

[1] D.L. Hall, J. Llinas, An introduction to multisensor data fusion, Proc. IEEE 85 (1) (1997) 6–23.
[2] Y. Bar-Shalom, P.K. Willett, X. Tian, Tracking and Data Fusion vol. 11, YBS Publishing Storrs, CT, USA, 2011.
[3] F. Castanedo, A Review of Data Fusion Techniques, Sci. World J. (2013) 1–19, https://doi.org/10.1155/2013/704504.
[4] Q. Tang, J. Liang, Maneuvering multitargets tracking system using surveillance multisensors, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–12.
[5] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129.
[6] I.M. Pires, N.M. Garcia, N. Pombo, F. Flórez-Revuelta, From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices, Sensors 16 (2) (2016) 184.
[7] F. Zhu, Q. Liang, Ocrnn: an orthogonal constrained recurrent neural network for sleep analysis based on eeg data, Ad Hoc Netw. 104 (2020) 102178.
[8] F. Zhu, Q. Liang, Rethink of orthographic constraints on RNN and its application in acoustic sensor data modeling, IEEE Internet Things J. 9 (3) (2021) 1962–1975.
[9] B.P.L. Lau, S.H. Marakkalage, Y. Zhou, N.U. Hassan, C. Yuen, M. Zhang, U.-X. Tan, A survey of data fusion in smart city applications, Inf. Fusion 52 (2019) 357–374.
[10] X. Deng, Y. Jiang, L.T. Yang, M. Lin, L. Yi, M. Wang, Data fusion based coverage optimization in heterogeneous sensor networks: a survey, Inf. Fusion 52 (2019) 90–105.
[11] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P.M. Atkinson, J.A. Benediktsson, Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art, IEEE Geosci. Remote Sens. Mag. 7 (1) (2019) 6–39.
[12] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, G. Fortino, Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges, Inf. Fusion 80 (2022) 241–265.
[13] M. Ahmad, S. Shabbir, S.K. Roy, D. Hong, X. Wu, J. Yao, A.M. Khan, M. Mazzara, S. Distefano, J. Chanussot, Hyperspectral image classification-traditional to deep models: a survey for future prospects, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2022) 968–999.
[14] D. Ramachandram, G.W. Taylor, Deep multimodal learning: a survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108.
[15] E. de Bézenac, S.S. Rangapuram, K. Benidis, M. Bohlke-Schneider, R. Kurle, L. Stella, H. Hasson, P. Gallinari, T. Januschowski, Normalizing Kalman filters for multivariate time series analysis. NeurIPS, 2020.
[16] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, J. Liu, Efficient and model-based infrared and visible image fusion via algorithm unrolling, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1186–1196.
[17] G. Revach, N. Shlezinger, X. Ni, A.L. Escoriza, R.J. van Sloun, Y.C. Eldar, KalmanNet: neural network aided Kalman filtering for partially known dynamics, arXiv preprint arXiv:2107.10043(2021).
[18] S.K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, J. Chanussot, Multimodal fusion transformer for remote sensing image classification, arXiv preprint arXiv:2203.16952(2022).
[19] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, S. Yang, Transformer based conditional GAN for multimodal image fusion, IEEE Trans. Multimed. (2023) 1–14, https://doi.org/10.1109/TMM.2023.3243659.
[20] M. Appel, F. Lahn, W. Buytaert, E. Pebesma, Open and scalable analytics of large earth observation datasets: from scenes to multidimensional arrays using SciDB and GDAL, ISPRS J. Photogramm. Remote Sens. 138 (2018) 47–56.
[21] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, Nuscenes: a multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11621–11631.

[22] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, A. Wallace, Radiate: a radar dataset for automotive perception in bad weather. 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 1–7.

[23] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: recent advances and future prospects, Inf. Fusion 42 (2018) 158–173.

[24] J.P. Dakin, R. Brown, Handbook of Optoelectronics: Concepts, Devices, and Techniques (Volume One), CRC Press, 2017.

[25] C. Badue, R. Guidolini, R.V. Carneiro, P. Azevedo, V.B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T.M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, A.F. De Souza, Self-driving cars: a survey, Expert Syst. Appl. 165 (2021) 113816.

[26] P. Bahl, V.N. Padmanabhan, Radar: an in-building RF-based user location and tracking system. Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064) vol. 2, IEEE, 2000, pp. 775–784.

[27] P.A. Torrione, C.S. Throckmorton, L.M. Collins, Performance of an adaptive feature-based processor for a wideband ground penetrating radar system, IEEE Trans. Aerosp. Electron. Syst. 42 (2) (2006) 644–658.

[28] K.L. Bell, J.T. Johnson, G.E. Smith, C.J. Baker, M. Rangaswamy, Cognitive radar for target tracking using a software defined radar system. 2015 IEEE Radar Conference (RadarCon), IEEE, 2015, pp. 1394–1399.

[29] M.M. Atia, A.R. Hilal, C. Stellings, E. Hartwell, J. Toonstra, W.B. Miners, O. A. Basir, A low-cost lane-determination system using GNSS/IMU fusion and HMM-based multistage map matching, IEEE Trans. Intell. Transp. Syst. 18 (11) (2017) 3027–3037.

[30] S. Hazra, A. Santra, Robust gesture recognition using millimetric-wave radar system, IEEE Sens. Lett. 2 (4) (2018) 1–4.

[31] M. Eslami, A. Mohammadzadeh, Developing a spectral-based strategy for urban object detection from airborne hyperspectral TIR and visible data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (2015) 1–9.

[32] J. Zhu, J. Hu, S. Jia, X. Jia, Q. Li, Multiple 3-D feature fusion framework for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. PP (2018) 1–14.

[33] F.E. White, Data Fusion Lexicon. Technical Report, Joint Directors of Labs Washington DC, 1991.

[34] F. Caron, E. Duflos, D. Pomorski, P. Vanheeghe, GPS/IMU data fusion using multisensor Kalman filtering: introduction of contextual aspects, Inf. Fusion 7 (2) (2006) 221–230.

[35] P. Hebert, N. Hudson, J. Ma, T. Howard, T. Fuchs, M. Bajracharya, J. Burdick, Combined shape, appearance and silhouette for simultaneous manipulator and object tracking. 2012 IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 2405–2412.

[36] Q. Tang, F. Zhu, J. Liang, Interactive multi-model tracking of a highly maneuvering target using mspdaf with least squares virtual fusion. 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2019, pp. 1–5.

[37] A. Basit, M. Tufail, M. Rehan, An adaptive gain based approach for event-triggered state estimation with unknown parameters and sensor nonlinearities over wireless sensor networks, ISA Trans. 129 (2022) 41–54.

[38] J. Ilonen, J. Bohg, V. Kyrki, Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing, Int. J. Robot. Res. 33 (2) (2014) 321–341.

[39] S. Mahfouz, F. Mourad-Chehade, P. Honeine, J. Farah, H. Snoussi, Target tracking using machine learning and Kalman filter in wireless sensor networks, IEEE Sens. J. 14 (10) (2014) 3715–3725.

[40] Y. Zhang, B. Song, X. Du, M. Guizani, Vehicle tracking using surveillance with multimodal data fusion, IEEE Trans. Intell. Transp. Syst. 19 (7) (2018) 2353–2361.

[41] O. Mees, A. Eitel, W. Burgard, Choosing smartly: adaptive multimodal fusion for object detection in changing environments. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 151–156.

[42] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H. Torr, End-to-end representation learning for correlation filter based tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2805–2813.

[43] Y. Zheng, X. Liu, X. Cheng, K. Zhang, Y. Wu, S. Chen, Multi-task deep dual correlation filters for visual tracking, IEEE Trans. Image Process. 29 (2020) 9614–9626.

[44] Z. Chen, W. Li, Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network, IEEE Trans. Instrum. Meas. 66 (7) (2017) 1693–1702.

[45] M. Wu, N. Goodman, Multimodal generative models for scalable weakly-supervised learning, Adv. Neural Inf. Process. Syst. 31 (2018) 5575– 5585.

[46] R. Kurle, S. Günnemann, P. van der Smagt, Multi-source neural variational inference, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 4114–4121.

[47] B. Ivanovic, K. Leung, E. Schmerling, M. Pavone, Multimodal deep generative models for trajectory prediction: a conditional variational autoencoder approach, IEEE Robot. Autom. Lett. 6 (2) (2021) 295–302.

[48] J. Pan, C.C. Ferrer, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: visual saliency prediction with generative adversarial networks, arXiv preprint arXiv:1701.01081(2017).

[49] H. Jia, J. Liu, Y. Wu, T. Bednarz, L. Yao, W. Hu, Condor: mobile golf swing tracking via sensor fusion using conditional generative adversarial networks. EWSN, 2021, pp. 31–42.

[50] M. Wang, X. Liu, H. Jin, A generative image fusion approach based on supervised deep convolution network driven by weighted gradient flow, Image Vis. Comput. 86 (2019) 1–16.

[51] D. Wu, L. Pigou, P.-J. Kindermans, N.D.-H. Le, L. Shao, J. Dambre, J.-M. Odobez, Deep dynamic neural networks for multimodal gesture segmentation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1583–1597.

[52] Y. Yan, W. Ren, X. Cao, Recolored image detection via a deep discriminative model, IEEE Trans. Inf. Forensics Secur. 14 (1) (2019) 5–17.

[53] H. Zhu, J.-B. Weibel, S. Lu, Discriminative multi-modal feature fusion for RGBD indoor scene recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2969–2976.

[54] B. Lei, S. Chen, D. Ni, T. Wang, Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion, Front. Aging Neurosci. 8 (2016) 77.

[55] H. Taylor, L. Hiley, J. Furby, A. Preece, D. Braines, VADR: discriminative multimodal explanations for situational understanding. 2020 IEEE 23rd International Conference on Information Fusion (FUSION), IEEE, 2020, pp. 1–8.

[56] X. Yun, Y. Sun, X. Yang, N. Lu, Discriminative fusion correlation learning for visible and infrared tracking, Math. Probl. Eng. 2019 (2019) 1–11.

[57] J.M. Ramirez, J.I. Martínez-Torre, H. Arguello, LADMM-net: an unrolled deep network for spectral image fusion from compressive data, Signal Process. 189 (2021) 108239.

[58] A. Prakash, K. Chitta, A. Geiger, Multi-modal fusion transformer for end-to-end autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7077–7087.

[59] Y. Kittenplon, Y.C. Eldar, D. Raviv, Flowstep3D: model unrolling for self-supervised scene flow estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4114–4123.

[60] S. Chen, Y.C. Eldar, L. Zhao, Graph unrolling networks: interpretable neural networks for graph signal denoising, IEEE Trans. Signal Process. 69 (2021) 3699–3713.

[61] S. Särkkä, A. Vehtari, J. Lampinen, Rao-blackwellized particle filter for multiple target tracking, Inf. Fusion 8 (1) (2007) 2–15.

[62] S.S. Saab, Z.S. Nakad, A standalone RFID indoor positioning system using passive tags, IEEE Trans. Ind. Electron. 58 (5) (2011) 1961–1970.

[63] M. Zorzi, Robust Kalman filtering under model perturbations, IEEE Trans. Autom. Control 62 (6) (2017) 2902–2907.

[64] T. Vercauteren, X. Wang, Decentralized sigma-point information filters for target tracking in collaborative sensor networks, IEEE Trans. Signal Process. 53 (8) (2005) 2997–3009.

[65] E. Maggio, F. Smerladi, A. Cavallaro, Adaptive multifeature tracking in a particle filtering framework, IEEE Trans. Circuits Syst. Video Technol. 17 (10) (2007) 1348–1359.

[66] X. Gao, D. You, S. Katayama, Seam tracking monitoring based on adaptive Kalman filter embedded ELMAN neural network during high-power fiber laser welding, IEEE Trans. Ind. Electron. 59 (11) (2012) 4315–4325.

[67] K. Szabat, T. Orlowska-Kowalska, Performance improvement of industrial drives with mechanical elasticity using nonlinear adaptive Kalman filter, IEEE Trans. Ind. Electron. 55 (3) (2008) 1075–1084.

[68] J. Wang, J. Li, Y. Shi, J. Lai, X. Tan, Am^{3}net: adaptive mutual-learning-based multimodal data fusion network, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2022) 5411–5426.

[69] M. Fu, H. Qu, Z. Yi, L. Lu, Y. Liu, A novel deep learning-based collaborative filtering model for recommendation system, IEEE Trans. Cybern. 49 (3) (2019) 1084–1096.

[70] E. Gundogdu, A.A. Alatan, Good features to correlate for visual tracking, IEEE Trans. Image Process. 27 (5) (2018) 2526–2540.

[71] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114(2014).

[72] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784(2014).

[73] L. Dinh, D. Krueger, Y. Bengio, Nice: non-linear independent components estimation, arXiv preprint arXiv:1410.8516(2015).

[74] D.P. Kingma, M. Welling, An introduction to variational autoencoders, arXiv preprint arXiv:1906.02691(2019).

[75] R.J. Piechocki, X. Wang, M.J. Bocus, Multimodal sensor fusion in the latent representation space, Sci. Rep. 13 (1) (2023) 2005.

[76] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[77] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).

[78] I. Goodfellow, Nips 2016 tutorial: generative adversarial networks, arXiv preprint arXiv:1701.00160(2017).

[79] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: interpretable representation learning by information maximizing generative adversarial nets, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems vol. 29, Curran Associates, Inc., 2016.

[80] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using real NVP, arXiv preprint arXiv:1605.08803(2017).

[81] D.P. Kingma, P. Dhariwal, Glow: generative flow with invertible 1×1 convolutions, Adv. Neural Inf. Process. Syst. 31 (2018).

[82] Z. Yan, H. Zha, Flow-based slam: from geometry computation to learning, Virtual Real. Intell. Hardw. 1 (5) (2019) 435–460.

[83] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: a RGB-D scene understanding benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.

[84] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, Y. Liu, Understand scene categories by objects: a semantic regularized scene classifier using convolutional neural networks. 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 2318–2325.

[85] Z. Fu, F. Angelini, S.M. Naqvi, J.A. Chambers, GM-PHD filter based online multiple human tracking using deep discriminative correlation matching. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4299–4303.

[86] Q. Zhu, X. Xu, N. Yuan, Z. Zhang, D. Guan, S.-J. Huang, D. Zhang, Latent correlation embedded discriminative multi-modal data fusion, Signal Process. 171 (2020) 107466.

[87] K. Gregor, Y. LeCun, Learning fast approximations of sparse coding. Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010, pp. 399–406.

[88] S.A.H. Hosseini, B. Yaman, S. Moeller, M. Hong, M. Akçakaya, Dense recurrent neural networks for accelerated MRI: history-cognizant unrolling of optimization algorithms, IEEE J. Sel. Top. Signal Process. 14 (6) (2020) 1280–1291.

[89] A. Mehranian, A.J. Reader, Model-based deep learning pet image reconstruction using forward–backward splitting expectation–maximization, IEEE Trans. Radiat. Plasma Med. Sci. 5 (1) (2020) 54–64.

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998– 6008.

[91] M. Tsimpoukelli, J.L. Menick, S. Cabi, S. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, Adv. Neural Inf. Process. Syst. 34 (2021) 200–212.

[92] Y.-L. Sung, J. Cho, M. Bansal, Vl-adapter: parameter-efficient transfer learning for vision-and-language tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5227–5237.

[93] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Adv. Neural Inf. Process. Syst. 35 (2022) 23716–23736.

[94] M. Yasuda, Y. Ohishi, S. Saito, N. Harado, Multi-view and multi-modal event detection utilizing transformer-based multi-sensor fusion. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 4638–4642.

[95] C. Weng, B. Lu, Q. Gu, X. Zhao, A novel multisensor fusion transformer and its application into rotating machinery fault diagnosis, IEEE Trans. Instrum. Meas. 72 (2023) 1–12.

[96] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the KITTI dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.

[97] S. Escalera, X. Baró, J. Gonzalez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H.J. Escalante, J. Shotton, I. Guyon, Chalearn looking at people challenge 2014: dataset and results. European Conference on Computer Vision, Springer, 2014, pp. 459–473.

[98] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J.A. Holgado-Terriza, S. Lee, H. Pomares, I. Rojas, Design, implementation and validation of a novel open framework for agile development of mobile health applications, Biomed. Eng. Online 14 (2) (2015) 1–20.

[99] J. Mao, J. Xu, Y. Jing, A. Yuille, Training and evaluating multimodal word embeddings with large-scale web annotated images, arXiv preprint arXiv:1611.08321(2016).

[100] W. Maddern, G. Pascoe, C. Linegar, P. Newman, 1 year, 1000 km: the oxford robotcar dataset, Int. J. Robot. Res. 36 (1) (2017) 3–15.

[101] P. Azagra, F. Golemo, Y. Mollard, M. Lopes, J. Civera, A.C. Murillo, A multimodal dataset for object model learning from natural human-robot interaction. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 6134–6141.

[102] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, S.-F. Chang, Exploiting feature and class relationships in video categorization with regularized deep neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2017) 352–364.

[103] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, A. Kim, Complex urban Lidar data set. 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 6344–6351.

[104] B. Le Saux, N. Yokoya, R. Hänsch, S. Prasad, 2018 IEEE GRSS data fusion contest: multimodal land use classification [technical committees], IEEE Geosci. Remote Sens. Mag. 6 (1) (2018) 52–54.

[105] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, D. Manocha, Trafficpredict: trajectory prediction for heterogeneous traffic-agents. Proceedings of the AAAI Conference on Artificial Intelligence vol. 33, 2019, pp. 6120–6127.

[106] T. Pire, M. Mujica, J. Civera, E. Kofman, The Rosario dataset: multisensor data for localization and mapping in agricultural environments, 2019. arXiv:1809.06413.

[107] R. Hanten, C. Schulz, A. Zwiener, A. Zell, MuSe: Multi-Sensor Integration Strategies Applied to Sequential Monte Carlo Methods. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019, pp. 7798–7804, https://doi.org/10.1109/IROS40897.2019.8967893.

[108] J. Shermeyer, D. Hogan, J. Brown, A. Van Etten, N. Weir, F. Pacifici, R. Hansch, A. Bastidas, S. Soenen, T. Bacastow, et al., Spacenet 6: multi-sensor all weather mapping dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 196–197.

[109] S. Herath, S. Irandoust, B. Chen, Y. Qian, P. Kim, Y. Furukawa, Fusion-DHL: WIFI, IMU, and floorplan fusion for dense history of locations in indoor environments, arXiv preprint arXiv:2105.08837(2021).

[110] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha, Y. Ma, Stcrowd: a multimodal dataset for pedestrian perception in crowded scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19608–19617.

[111] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, D. Rus, Actionsense: a multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment, Adv. Neural Inf. Process. Syst. 35 (2022) 13800–13813.

[112] M. Bock, M. Moeller, K. Van Laerhoven, H. Kuehne, Wear: a multimodal dataset for wearable and egocentric video activity recognition, arXiv preprint arXiv:2304.05088(2023).

[113] W.M. Wells III, Medical image analysis–past, present, and future, 2016.

[114] M. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (2011) 744–756.

[115] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, Image Process., IEEE Trans. 13 (2004) 600–612.

[116] B. Rajalingam, R. Priya, R. Bhavani, Hybrid multimodal medical image fusion using combination of transform techniques for disease analysis, Procedia Comput. Sci. 152 (2019) 150–157.

[117] Y. Chen, R. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (2009) 1421–1432.

[118] D. Summers, Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html, J. Neurol., Neurosurg. Psychiatry 74 (3) (2003) 288.

[119] S. Das, M.K. Kundu, NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency, Med. Biol. Eng. Comput. 50 (10) (2012) 1105–1114.

[120] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.

[121] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.

[122] P. Ganasala, V. Kumar, Feature-motivated simplified adaptive PCNN-based medical image fusion algorithm in NSST domain, J. Digit. Imaging 29 (1) (2016) 73–85.

[123] Y. Yang, Y. Que, S.-Y. Huang, P. Lin, Technique for multi-focus image fusion based on fuzzy-adaptive pulse-coupled neural network, Signal Image Video Process. 11 (3) (2017) 439–446.

[124] S. Singh, R.S. Anand, D. Gupta, CT and MR image information fusion scheme using a cascaded framework in ripplet and NSST domain, IET Image Proc. 12 (5) (2018) 696–707.

[125] Y. Yang, Y. Que, S. Huang, P. Lin, Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain, IEEE Sens. J. 16 (10) (2016) 3735–3745.

[126] S. Singh, R.S. Anand, Multimodal medical image sensor fusion model using sparse K-SVD dictionary learning in nonsubsampled shearlet domain, IEEE Trans. Instrum. Meas. 69 (2) (2020) 593–607.

[127] U. Asif, M. Bennamoun, F.A. Sohel, A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling, IEEE Trans. Pattern Anal. Mach. Intell. 40 (9) (2018) 2051–2065.

[128] X. Chen, E. Konukoglu, Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders, 2018. arXiv:1806.04972.

[129] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang, Medical image fusion via convolutional sparsity based morphological component analysis, IEEE Signal Process Lett 26 (3) (2019) 485–489.

[130] H. Huang, D. Zheng, H. Chen, Y. Wang, C. Chen, L. Xu, G. Li, Y. Wang, X. He, W. Li, Fusion of CT images and clinical variables based on deep learning for predicting invasiveness risk of stage I lung adenocarcinoma, Med. Phys. 49 (10) (2022) 6384–6394.

[131] Y. Liu, X. Chen, R.K. Ward, Z. Jane Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.

[132] M. Ahmad, L. Tawalbeh, N. Sayyah, L. Shaikha, R. Safadi, R. Sahyoun, The jordanians' perception of the association between foods and other risk factors with cancer, Int. J. Cancer Res. Prev. 10 (2–3) (2017) 243–252.

[133] R. Thirukovalluru, S. Dixit, R.K. Sevakula, N.K. Verma, A. Salour, Generating feature sets for fault diagnosis using denoising stacked auto-encoder. 2016 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2016, pp. 1–7.

[134] S. Saadat, M. Pickering, D. Perriman, J. Scarvell, P. Smith, Fast and robust multi-modal image registration for 3D knee kinematics, vol. 2017-December, 2017, pp. 1–5.

[135] D. Ye, J.Y. Hsi Fuh, Y. Zhang, G.S. Hong, K. Zhu, In situ monitoring of selective laser melting using plume and spatter signatures by deep belief networks, ISA Trans. 81 (2018) 96–104.

[136] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[137] H.-I. Suk, S.-W. Lee, D. Shen, A.D.N. Initiative, et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage 101 (2014) 569–582.

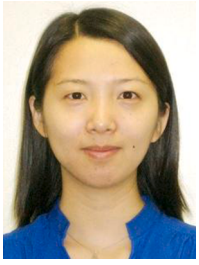
[138] H.-I. Suk, S.-W. Lee, D. Shen, Latent feature representation with stacked auto-encoder for AD/MCI diagnosis, Brain Struct. Funct. 220 (2) (2015) 841–859.

[139] X. Gao, W. Li, M. Loomes, L. Wang, A fused deep learning architecture for viewpoint classification of echocardiography, Inf. Fusion 36 (2017) 103–113.

[140] M. Kallenberg, K. Petersen, M. Nielsen, A.Y. Ng, P. Diao, C. Igel, C.M. Vachon, K. Holland, R.R. Winkel, N. Karssemeijer, et al., Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring, IEEE Trans. Med. Imaging 35 (5) (2016) 1322–1331.
[141] G. van Tulder, M. de Bruijne, Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted Boltzmann machines, IEEE Trans. Med. Imaging 35 (5) (2016) 1262–1272.
[142] A. Sudheer, C.H. Bindu, Region based multi-focus image fusion using the spectral parameter variance. 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), IEEE, 2016, pp. 1306–1310.
[143] G. Bhatnagar, Q.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimed. 15 (5) (2013) 1014–1024.
[144] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (1) (2012) 10–19.
[145] Y. Yang, S. Tong, S. Huang, P. Lin, Multifocus image fusion based on NSCT and focused area detection, IEEE Sens. J. 15 (5) (2014) 2824–2838.
[146] J. Zhu, W. Jin, L. Li, Z. Han, X. Wang, Multiscale infrared and visible image fusion using gradient domain guided image filtering, Infrared Phys. Technol. 89 (2018) 8–19.
[147] A. Teramoto, H. Fujita, O. Yamamuro, T. Tamaki, Automated detection of pulmonary nodules in pet/ct images: ensemble false-positive reduction using a convolutional neural network technique, Med. Phys. 43 (6Part1) (2016) 2821–2827.
[148] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images, IEEE Trans. Med. Imaging 35 (5) (2016) 1313–1321.
[149] H. Chen, L. Yu, Q. Dou, L. Shi, V.C. Mok, P.A. Heng, Automatic detection of cerebral microbleeds via deep learning based 3D feature representation. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 764–767.
[150] Y. Li, W. Liang, Y. Zhang, H. An, J. Tan, Automatic lumbar vertebrae detection based on feature fusion deep learning for partial occluded c-arm x-ray images. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 647–650.
[151] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1341–1360.
[152] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, K. Chowdhury, Deep learning on multimodal sensor data at the wireless edge for vehicular network, IEEE Trans. Veh. Technol. 71 (7) (2022) 7639–7655.
[153] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (01) (2014) 58–72.
[154] J.H. Yoon, M.-H. Yang, J. Lim, K.-J. Yoon, Bayesian multi-object tracking using motion context from multiple objects. 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015, pp. 33–40.
[155] A. Geiger, M. Lauer, C. Wojek, C. Stiller, R. Urtasun, 3D traffic scene understanding from movable platforms, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2013) 1012–1025.
[156] P. Lenz, A. Geiger, R. Urtasun, Followme: efficient online min-cost flow tracking with bounded memory and computation. Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4364–4372.
[157] J.H. Yoon, C.-R. Lee, M.-H. Yang, K.-J. Yoon, Online multi-object tracking via structural constraint event aggregation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1392–1400.
[158] A. Gaidon, E. Vig, Online domain adaptation for multi-object tracking, 2018, US Patent 9,984,315.
[159] W. Choi, Near-online multi-target tracking with aggregated local flow descriptor. Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3029–3037.
[160] S. Wang, C.C. Fowlkes, Learning optimal parameters for multi-target tracking with contextual interactions, Int. J. Comput. Vis. 122 (3) (2017) 484–501.
[161] A. Milan, K. Schindler, S. Roth, Detection- and trajectory-level exclusion in multiple object tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[162] J.H. Yoon, C.-R. Lee, M.-H. Yang, K.-J. Yoon, Online multi-object tracking via structural constraint event aggregation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[163] Y. Xiang, A. Alahi, S. Savarese, Learning to track: Online multi-object tracking by decision making. Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4705–4713.
[164] B. Lee, E. Erdenee, S. Jin, M.Y. Nam, Y.G. Jung, P.K. Rhee, Multi-class multi-object tracking using changing point detection. European Conference on Computer Vision, Springer, 2016, pp. 68–83.
[165] D. Frossard, R. Urtasun, End-to-end learning of multi-sensor 3D tracking by detection. 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 635–642.
[166] K. Burnett, S. Samavi, S. Waslander, T. Barfoot, A. Schoellig, Autotrack: a lightweight object detection and tracking system for the SAE autodrive challenge. 2019 16th Conference on Computer and Robot Vision (CRV), IEEE, 2019, pp. 209–216.
[167] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, H. Michael Gross, Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
[168] S. Wang, Y. Sun, C. Liu, M. Liu, Pointtracknet: an end-to-end network for 3-d object detection and tracking from point clouds, IEEE Robot. Autom. Lett. 5 (2) (2020) 3206–3212.
[169] A. Osep, W. Mehner, M. Mathias, B. Leibe, Combined image-and world-space tracking in traffic scenes. 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 1988–1995.
[170] Y. Zhong, S. You, U. Neumann, Modeling cross-modal interaction in a multi-detector, multi-modal tracking framework. Proceedings of the Asian Conference on Computer Vision (ACCV), 2020.
[171] H. Cho, Y.-W. Seo, B.V. Kumar, R.R. Rajkumar, A multi-sensor fusion system for moving object detection and tracking in urban driving environments. 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 1836–1843.
[172] J. Schlosser, C.K. Chow, Z. Kira, Fusing Lidar and images for pedestrian detection using convolutional neural networks. 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 2198–2205.
[173] J. Dou, J. Xue, J. Fang, SEG-voxelnet for 3D vehicle detection from RGB and Lidar data. 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 4362–4368.
[174] M. Liang, B. Yang, Y. Chen, R. Hu, R. Urtasun, Multi-task multi-sensor fusion for 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7345–7353.
[175] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[176] Y. Li, A.W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, et al., Deepfusion: Lidar-camera deep fusion for multi-modal 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17182–17191.
[177] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, Transfusion: robust Lidar-camera fusion for 3D object detection with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1090–1099.
[178] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear MOT metrics, EURASIP J. Image Video Process. 2008 (2008) 1–10.
[179] P. Gader, A. Zare, R. Close, J. Aitken, G. Tuell, Muufl Gulfport Hyperspectral and Lidar Airborne Data Set. Tech. Rep, Univ. Florida, Gainesville, FL, USA, 2013.
[180] M. Ahmad, S. Shabbir, S.K. Roy, D. Hong, X. Wu, J. Yao, A.M. Khan, M. Mazzara, S. Distefano, J. Chanussot, Hyperspectral image classification-traditional to deep models: a survey for future prospects, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2021) 968–999.
[181] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, IEEE Trans. Geosci. Remote Sens. 42 (8) (2004) 1778–1790.
[182] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder-decoder approaches, arXiv preprint arXiv:1409.1259(2014).
[183] X. Wu, D. Hong, J. Chanussot, Convolutional neural networks for multimodal remote sensing data classification, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–10.
[184] S. Mohla, S. Pande, B. Banerjee, S. Chaudhuri, Fusatnet: dual attention based spectrospatial multimodal fusion network for hyperspectral and Lidar classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 92–93.
[185] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, Q. Liu, Classification of hyperspectral and Lidar data using coupled CNNs, IEEE Trans. Geosci. Remote Sens. 58 (7) (2020) 4939–4950.
[186] X. Zhao, R. Tao, W. Li, Multisource remote sensing data classification using deep hierarchical random walk networks. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2187–2191.
[187] S.K. Roy, G. Krishna, S.R. Dubey, B.B. Chaudhuri, Hybridsn: exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification, IEEE Geosci. Remote Sens. Lett. 17 (2) (2019) 277–281.
[188] D. Hong, L. Gao, R. Hang, B. Zhang, J. Chanussot, Deep encoder–decoder networks for classification of hyperspectral and Lidar data, IEEE Geosci. Remote Sens. Lett. 19 (2020) 1–5.
[189] S.K. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, J. Chanussot, Hyperspectral and Lidar data classification using joint CNNs and morphological feature learning, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16.
[190] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, B. Le Saux, Advanced multi-sensor optical remote sensing for urban land use and land cover classification: outcome of the 2018 IEEE GRSS data fusion contest, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (6) (2019) 1709–1724.
[191] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 54 (12) (2016) 7405–7415.
[192] G. Cheng, Y. Si, H. Hong, X. Yao, L. Guo, Cross-scale feature fusion for object detection in optical remote sensing images, IEEE Geosci. Remote Sens. Lett. 18 (3) (2021) 431–435.
[193] A. Singh, K. Gaurav, Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images, Sci. Rep. 13 (1) (2023) 2251.
[194] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, More diverse means better: multimodal deep learning meets remote-sensing imagery classification, IEEE Trans. Geosci. Remote Sens. 59 (5) (2020) 4340–4354.

[195] B. Yuan, L. Han, X. Gu, H. Yan, Multi-deep features fusion for high-resolution remote sensing image scene classification, Neural Comput. Appl. 33 (6) (2021) 2047–2063.
[196] B. Chen, B. Huang, B. Xu, Multi-source remotely sensed data fusion for improving land cover classification, ISPRS J. Photogramm. Remote Sens. 124 (2017) 27–39.
[197] J.B. Sankey, T.T. Sankey, J. Li, S. Ravi, G. Wang, J. Caster, A. Kasprak, Quantifying plant-soil-nutrient dynamics in rangelands: fusion of UAV hyperspectral-Lidar, UAV multispectral-photogrammetry, and ground-based Lidar-digital photography in a shrub-encroached desert grassland, Remote Sens. Environ. 253 (2021) 112223.
[198] F. Rodríguez-Puerta, R. Alonso Ponce, F. Pérez-Rodríguez, B. Águeda, S. Martín-García, R. Martínez-Rodrigo, I. Lizarralde, Comparison of machine learning algorithms for wildland-urban interface fuelbreak planning integrating ALS and UAV-borne Lidar data and multispectral images, Drones 4 (2) (2020) 21.
[199] R. Hänsch, O. Hellwich, Fusion of multispectral Lidar, hyperspectral, and RGB data for urban land cover classification, IEEE Geosci. Remote Sens. Lett. 18 (2) (2020) 366–370.
[200] Z. Xiang, L. Xiao, J. Yang, W. Liao, W. Philips, Detail-injection-model-inspired deep fusion network for pansharpening, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[201] Y. Du, J. Wang, Z. Liu, H. Yu, Z. Li, H. Cheng, Evaluation on spaceborne multispectral images, airborne hyperspectral, and Lidar data for extracting spatial distribution and estimating aboveground biomass of wetland vegetation *Suaeda salsa*, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (1) (2018) 200–209.
[202] S. Bhagat, S.D. Joshi, B. Lall, S. Gupta, Multimodal sensor fusion using symmetric skip autoencoder via an adversarial regulariser, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 1146–1157.
[203] S. Cui, A. Ma, Y. Wan, Y. Zhong, B. Luo, M. Xu, Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets, IEEE Trans. Geosci. Remote Sens. (2021) 1–14, https://doi.org/10.1109/TGRS.2021.3099506.
[204] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, S. Fang, A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[205] K. Xu, P. Deng, H. Huang, Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[206] Q. He, X. Sun, W. Diao, Z. Yan, D. Yin, K. Fu, Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing, ISPRS J. Photogramm. Remote Sens. 193 (2022) 90–103.
[207] B. Yao, Z. Zhou, L. Wang, W. Xu, Q. Liu, A. Liu, Sensorless and adaptive admittance control of industrial robot in physical human- robot interaction, Robot. Comput. Integr. Manuf. 51 (2018) 158–168.
[208] R. Huang, H. Cheng, J. Qiu, J. Zhang, Learning physical human–robot interaction with coupled cooperative primitives for a lower exoskeleton, IEEE Trans. Autom. Sci. Eng. 16 (4) (2019) 1566–1574.
[209] H. Wang, S. Li, L. Song, L. Cui, P. Wang, An enhanced intelligent diagnosis method based on multi-sensor image fusion via improved deep learning network, IEEE Trans. Instrum. Meas. 69 (6) (2020) 2648–2657.
[210] H. Cuayáhuitl, A data-efficient deep learning approach for deployable multimodal social robots, Neurocomputing 396 (2020) 587–598.
[211] N. Saito, T. Ogata, S. Funabashi, H. Mori, S. Sugano, How to select and use tools?: active perception of target objects using multimodal deep learning, IEEE Robot. Autom. Lett. 6 (2) (2021) 2517–2524.
[212] M. Thosar, S. Zug, A.M. Skaria, A. Jain, A review of knowledge bases for service robots in household environments. AIC, 2018, pp. 98–110.
[213] Y. Zhang, M. Lu, A review of recent advancements in soft and flexible robots for medical applications, Int. J. Med. Robot. Comput. Assist. Surg. 16 (3) (2020) e2096.
[214] P. Li, X. Liu, Common sensors in industrial robots: a review. Journal of Physics: Conference Series vol. 1267, IOP Publishing, 2019, p. 012036.
[215] L. Mora, X. Wu, A. Panori, Mind the gap: developments in autonomous driving research and the sustainability challenge, J. Clean. Prod. 275 (2020) 124087.
[216] P.-H. Kuo, S.-T. Lin, J. Hu, C.-J. Huang, Multi-sensor context-aware based chatbot model: an application of humanoid companion robot, Sensors 21 (15) (2021) 5132.
[217] K. Lin, Y. Li, J. Sun, D. Zhou, Q. Zhang, Multi-sensor fusion for body sensor network in medical human–robot interaction scenario, Inf. Fusion 57 (2020) 15–26.
[218] W. Qi, S.E. Ovur, Z. Li, A. Marzullo, R. Song, Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network, IEEE Robot. Autom. Lett. 6 (3) (2021) 6039–6045.
[219] P. Zhang, J. Zhang, Deep learning analysis based on multi-sensor fusion data for hemiplegia rehabilitation training system for stoke patients, Robotica (2021) 1–18.
[220] G. Li, S. Liu, L. Wang, R. Zhu, Skin-inspired quadruple tactile sensors integrated on a robot hand enable object recognition, Sci. Robot. 5 (49) (2020).
[221] P. Wei, L. Cagle, T. Reza, J. Ball, J. Gafford, Lidar and camera detection fusion in a real-time industrial multi-sensor collision avoidance system, Electronics 7 (6) (2018) 84.
[222] M. Brossard, S. Bonnabel, Learning wheel odometry and IMU errors for localization. 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 291–297.
[223] N. Koenig, A. Howard, Design and use paradigms for gazebo, an open-source multi-robot simulator. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566) vol. 3, IEEE, 2004, pp. 2149–2154.
[224] E. Coumans, Y. Bai, Pybullet, a python module for physics simulation for games, robotics and machine learning (2016).
[225] E. Todorov, T. Erez, Y. Tassa, Mujoco: a physics engine for model-based control. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2012, pp. 5026–5033.
[226] E. Rohmer, S.P. Singh, M. Freese, V-rep: a versatile and scalable robot simulation framework. 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2013, pp. 1321–1326.
[227] O. Michel, Cyberbotics Ltd. webots: professional mobile robot simulation, Int. J. Adv. Robot. Syst. 1 (1) (2004) 5.

Qin Tang has been working toward the successive post-graduate and doctoral programs in the School of Information and Communication Engineering at the University of Electronic Science and Technology of China, since September 2017. She won the Best Paper Award from the 7th International Conference on Communications, Signal Processing, and Systems, in July 2018. Her research interests include multi-modal data fusion, target cognition, and machine learning.



Jing Liang (SM'16) received the B.S. and M.Sc degrees from Beijing University of Posts and Telecommunications, China, in 2003 and 2006, respectively and She got the Ph.D. degree from the University of Texas at Arlington in August 2009, all in electrical engineering. Currently, she is a professor in the School of Information and Communication Engineering at the University of Electronic Science and Technology of China. Her current research interests include radar sensor networks, collaborative and distributed signal processing, wireless communication and networks, and machine learning. She serves as the Guest Editor of IEEE Internet of Things Journal, Pattern Recognition (Elsevier), EURASIP Journal of Wireless Communications and Networking. She is the TPC co-chair of several international conferences. Dr. Liang is the recipient of Hundred Talent Plan, Sichuan, China, and the key project leader of the National Natural Science Foundation of China.

Fangqi Zhu (M'21) received the B.Sc and M.Sc degrees in Electrical Engineering from the University of Electronic Science and Technology of China, Chengdu, in 2013 and 2016, respectively. He got his Ph.D. in Electrical Engineering from the University of Texas at Arlington in 2020. His current research interests include structured DL, statistical inference and optimization, signal processing and anomaly detection in manufacturing. He is the reviewer of several conferences and journals.