IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 May 2024, accepted 11 June 2024, date of publication 18 June 2024, date of current version 2 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3416370

SURVEY

Situation Awareness in AI-Based Technologies and Multimodal Systems: Architectures, Challenges and Applications

JIELI CHEN[1], KAH PHOOI SENG[1,2], (Senior Member, IEEE), JEREMY SMITH[3], (Member, IEEE), AND LI-MINN ANG[2], (Senior Member, IEEE)

[1]XJTLU Entrepreneur College (Taicang), Xian Jiaotong-Liverpool University, Taicang 215400, China
[2]School of Engineering and Science, University of Sunshine Coast, Petrie, QLD 4502, Australia
[3]Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX Liverpool, U.K.

Corresponding author: Kah Phooi Seng (Jasmine.Seng@xjtlu.edu.cn)

ABSTRACT Situation Awareness (SA) is a process of sensing, understanding and predicting the environment and is an important component in complex systems. The reception of information from the environment tends to be continuous and of a multimodal nature. AI technologies provide a more efficient and robust support by subdividing the different stages of SA objectives into tasks such as data fusion, representation, classification, and prediction. This paper provides an overview of AI and multimodal methods used to build, enhance and evaluate SA in a variety of environments and applications. Emphasis is placed on enhancing perceptual integrity and persistence. Research indicates that the integration of artificial intelligence and multimodal approaches has significantly enhanced perception and comprehension in complex systems. However, there remains a research gap in projecting future situations and effectively fusing multimodal information. This paper summarizes some of the use cases and lessons learned where AI and multimodal techniques have been used to deliver SA. Future perspectives and challenges are proposed, including more comprehensive predictions, greater interpretability, and more advanced visual information.

INDEX TERMS Artificial intelligence, situation awareness, deep learning, machine learning, reinforcement learning, multimodal fusion.

I. INTRODUCTION

Situation Awareness (SA) is a capability to perceive and understand critical factors in the environment, and further towards a set of projections of what will happen with the system in the near future. According to a widely adopted three-level SA model introduced by Endsley [1], it comprises of three stages, i.e., perception, comprehension and projection. Its applications in numerous fields, such as aviation, military, cyber security, power systems, etc., have proved its importance and rationality in risk assessment and decision making in complex systems.

SA technologies are usually orientated in one of two ways; one is to measure the operator's level of SA whereas the other is designed to provide or enhance SA. Previous SA measurement techniques, such as NASA TLX [2], SAGAT [3], SART [4], CDM [5], were often derived from aviation and were questionnaire orientated. With the introduction of sensors for collecting data (e.g., eye-tracking, physiological signals [6]), the emergence of real-time SA assessments has amplified the demand for increased capacity and faster data processing. Munir et al. [7] indicated that challenges in measuring SA are subjectivity, physiology, surveying, limitedness and coverage, respectively. Addressing these challenges invariably points to a wider range of more rational data collection. Similarly, environmentally oriented data acquisition and processing are crucial in order to provide and enhance SA. According to the three-level model of SA, a comprehensive perception of the environment determines higher dimensional understanding and more accurate predictions. The authors in [8] indicate that multimodal systems have a significant improvement in SA ability. This can also be

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Kashif Bashir.

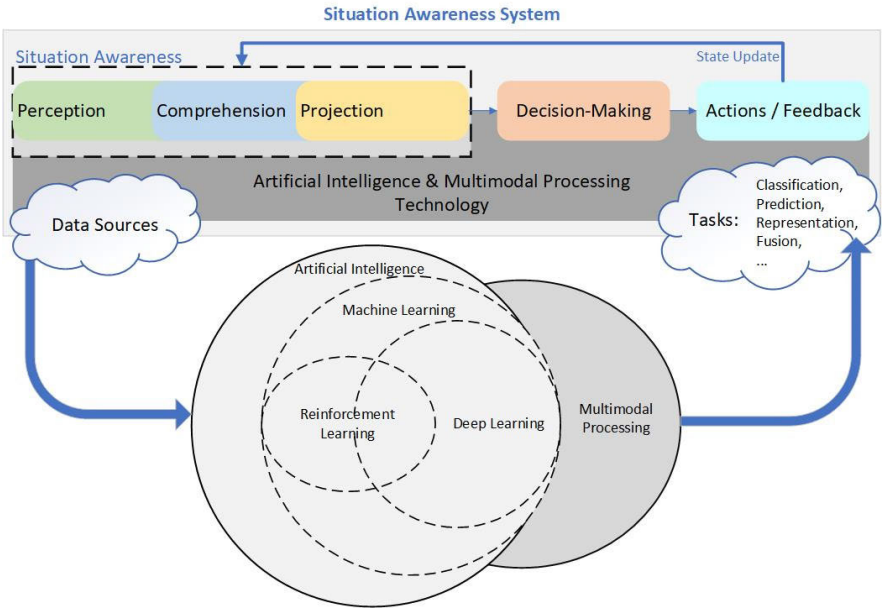

FIGURE 1. Overview of situation awareness system employing ai and multimodal technology.

verified in the way we experience our surroundings: different senses (e.g., visual, auditory, tactile, etc.) provide our brain with constant and multimodal information. In [9], the authors point out that the advantages of a multimodal approach have been demonstrated in studies both measuring and maintaining SA in automated vehicle drivers. Therefore, SA systems are designed to manage larger data volumes, wider detection ranges, longer time spans, and diverse data types. The challenge here can be the implementation of multimodal representation and fusion tailored for SA purposes.

The introduction of Artificial Intelligence is significant for the development of SA. It can improve an operator's ability to perceive an environment or an individual, while significantly reducing the workload of humans. ''Learning'' is a concept that appears frequently in AI techniques. Simply put, it is the process of recording instances and fitting a function that maps inputs to outputs. Several challenges in substituting ''learning'' functions into SA systems include (1) selecting sensible input data and transforming it into a form that can be easily processed, (2) learning effective features from the data, and (3) utilizing the learned features to accomplish perception, understanding, and prediction of the environment. These challenges are the same ones that AI technology is currently facing. In [10], the authors mention that deep learning methods face overfitting, lack of interpretability, and high demands on data quality and quantity when representing data. Fig. 1 shows an overview of the relationship between SA systems, AI and multimodal processing techniques based on the three-level model of SA. AI technologies can be subdivided into many tasks applied with SA systems, depending on the SA level and objectives. The perceptions in the SA system are the basis for everything. The authors in [11] state that based on the capabilities of the perception layer, it can be classified as low-level perception and high-level perception. Low-level perception is just responsible for data collection, presentation and understanding of low-level contextual features of classes such as time, temperature, location, etc. Correspondingly, high-level perception requires further translation of these understood features. For instance, acceleration data obtained from gyroscope measurements may determine that the user is in a 'running' state. Feature extraction, data representation and some simple classification and detection tasks are important contributions of AI technology at the perceptual level. The commonly used methods here are (1) Logistic Regression (LR) [12], Bayesian learning [13], [14], [15], [16], [17], K-Nearest Neighbors (KNN) [18], [19], Decision Trees (DT) [20], [21], Random Forest (RF), Support Vector Machine (SVM) [22] and Artificial Neural Networks based on Supervised Learning, and (2) Unsupervised learning based Principal Component Analysis (PCA), Independent Component Analysis (ICA), Kernel Density Estimation (KDE), Kullback-Leibler divergence (KLD). It is worth mentioning that the contribution of Deep Learning in data representation tasks is huge. Based on neural network architecture, multilayer perceptron (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN) can learn features in a variety of complex modalities (e.g., time-series data, images, text, audio) and train them for specific tasks. For instance, YOLO detector for target detection [23]. The comprehension stage is the focus for exemplifying the competence of an SA system, where the various detections and classifications achieved in the perception stage may be integrated to achieve a holistic perception of the environment by the entity. Multimodal techniques, such as multimodal representation and multimodal fusion, contribute to this phase. At the projection level, the contribution of AI methods is

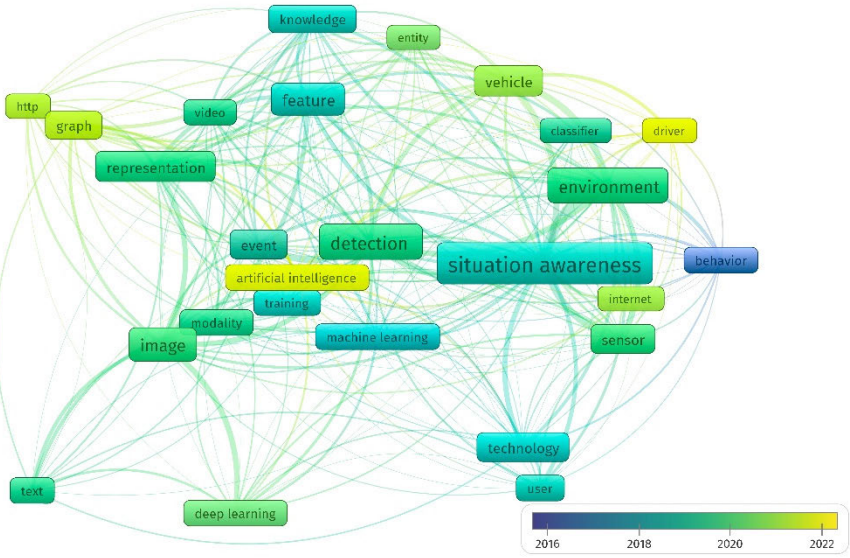

FIGURE 2. Keyword co-occurrence map of the collected works.

similar to them at the perception level. In particular, algorithms based on 'learning' may make predictions about future states through regression, classification, etc.

In order to explore the current trends and research challenges of AI and multimodal techniques applied to SA, this survey proposes the following survey objectives and discusses them in the form of a review of related work.

1) Overview of AI technologies for SA: Explore different AI methodologies, such as machine learning, deep learning, reinforcement learning, and their usage in providing and enhancing SA.
2) Multimodal Data Integration and Fusion: Investigate commonly used uni- and multi-modal data representation methods and how artificial intelligence techniques can be utilized to integrate and fuse information from multiple modalities (e.g., visual, auditory, textual) in order to improve the environmental comprehension of SA systems.
3) Multimodal AI in applications and use cases on SA: Examine specific applications and case studies where the combination of AI and multimodal technologies have been applied to improve SA in diverse domains, such as healthcare, risk management and automated systems.
4) Investigation on SA-related dataset: Generalization of datasets that can be employed to provide SA in the current state of lack of SA-targeted datasets. Extract and analyze the content, size, diversity, limitations and biases of the dataset.
5) Future directions for AI-based multimodal SA system: In-depth discussion on potential research directions, technical difficulties of SA systems enabled by AI and multimodal technologies to improve the ability of complex systems in sensing, understanding and predicting the environment.

To the best of our knowledge, there are numerous surveys related to SA. In [7], the authors review how to measure and quantify SA and describe a large number of application areas for SA, mentioning that AI technology contributes significantly to the 'prediction' phase of SA. In [24], SA between individuals, teams and systems is investigated. The authors argue that the perception of dynamic relationships between individuals and teams is also particularly important in the context of SA systems effectively perceiving the environment, which requires more developed sensor systems for improved information perception, more robust AI models for improved perception levels and decision-making capabilities. A similar point is pointed in [25], where the authors analyze the human-AI interaction in SA and argue that AI technologies need to help provide SA to complex systems in a more transparent as well as explanatory way, whereas the authors in [26] argue that this can resolve the tension between mass perception and AI systems.

In some specific areas, the concept of SA is starting to be emphasized by researchers. For example, [27] reviewed SA methods evaluated in aviation environments and suggested that SA systems should combine subjective (e.g., SAGAT) and objective (e.g., sensor signals) measurements in order to provide comprehensive perceptions, but the review made little mention of the contribution of AI technologies. For the maritime environment, it is mentioned in both [28] and [29] that AI technology-enabled SA systems can well improve the safety and efficiency of autopiloted ships, e.g., computer vision technology helps in detection and evasion as well as navigation; multimodal processing technology helps in multi-sensor fusion. In addition, SA systems in smart grid [30], [31], [32] environments provide more accurate and timely fault monitoring and emergency response with the support of AI; Internet of Things [11], [33], [34], autonomous vehicles [9], and social media [35], [36] environments are mostly

TABLE 1. Summary of studies for AI-empowered situation awareness.

Architecture	Year	Studies	Methods	Contributions	Best Outcome
Machine Learning	2009	[22]	HMM, SVM	Human activity detection	Acc. 66.51%-88.79%
	2016	[13]	KLD, Bayesian Inference	Airport taxiway obstruction detection	Higher Recall and F1-score compared to baseline
	2017	[14]	Naïve Bayes	Route recommendations	5.6%-9.8% outperformed on Acc. over baseline
	2019	[20]	RF	SA prediction on automated vehicle communication environment	Acc. 84.16%
	2020	[15]	DBN, GNG, MJPF	Abnormality detection	Acc. over 97%
	2020	[37]	BoW, Naïve Bayes	Driving style detection	6%-15% outperformed on F1 score over baseline
	2022	[21]	DT/LightGBM	SA prediction in autonomous driving	RMSE 0.121, MAE 0.096, Corr. 0.719
	2022	[17]	HDBN, GNG, MJPF	Radio environment prediction	Outperformed baseline on AUC, Acc. and RMSE.
	2022	[38]	DT/TGNA, MAB	Grid environment prediction	3.15%-6.76% outperformed on regret measure over baselines
	2023	[18]	RF, KNN, ANN	Assessment, prediction and intervention of autonomous car driver	Statistical analysis presented
	2023	[12]	SVM, LR	SA perception on pilot workload	Acc. 75%-82%
	2023	[19]	KDE, PCA	Islanding detection on microgrids	Outperformed over baseline
Deep Learning	2019	[39]	CNN, RNN\BiLSTM	Text classification for domestic violence	Acc. 89.12%-91.78%
	2019	[40]	ANN, Petri Nets	Time perception in airports, flow perception in websites	7.62% improvement over baseline on Acc.
	2020	[41]	LSTM-CNN	Internet memes classification	Class. Acc. 96.1%
	2021	[42]	CNN	Text-based crime classification	7%-8% outperformed on Acc. over baseline
	2022	[23]	CNN/YOLOX-s	Ship detection and localization	mAP 84.88%-89.42% with AR fusion
	2022	[43]	RNN/BiLSTM	Pipeline leak detection and localization	Class. Acc. 90.19%-98.13%, MAE 0.72
	2022	[44]	Auto-Encoder	Power synchronization control stability detection	Acc. 99.53%
	2022	[45]	TCN, Transformer	Network security situation prediction	MAE 0.044-0.052, RMSE 0.061-0.071
	2022	[46]	ANN	SA classification in air combat	Acc. 92.4%-93%
	2022	[47]	GCN, TCN	Vessel trajectory prediction	Outperformed baselines on displacement errors
	2022	[48]	BNN	Adverse weather detection and pilot workload perception	Acc. 66.5%
	2023	[49]	ANN	Traffic Violation Detection	Broader range and Acc. of detection than baseline
	2023	[50]	GCN	Fault detection and stability prediction in the GDT of IoE	Acc. over 90%
	2023	[51]	CNN/RetinaNet	Marine SA based on ship detection and classification	Acc. 60%-80%
	2023	[52]	LSTM	Trajectory projection for autonomous vehicle	MAE 32.99% reduction from baseline
	2023	[53]	CNN	Visual defogging to enhance SA in traffic	Outperformed over baseline
	2023	[54]	CNN	Visual support for ship sailing SA	Avg. Precision 60.4%
Reinforcement Learning	2018	[55]	ANN, TD-Learning	SA prediction on network environment	Outperformed over baseline
	2020	[56]	Q-Learning	SA detection on malicious vehicles	Outperformed over baseline
	2020	[57]	DRL	Autonomous vehicles decision-making	Outperformed over baseline
	2021	[58]	Q-Learning	Network attack detection and prediction	Outperformed over baseline
	2022	[59]	DRL, Auto-Encoder	Aircraft 3D detection tracking, air combat manoeuvre decision	Tracking success rate 89.2%, decided manoeuvre 99.1% better than baseline
	2022	[60]	TD-Learning	SA on network environment and decision making	Outperformed over baseline
	2022	[61]	DRL	Autonomous vehicles decision-making	Outperformed over baseline
	2022	[62]	DRL, CNN	Autonomous vehicles decision-making and motion-controlling	Outperformed over baseline
	2023	[63]	MAB	Insect species identification, concept drift detection	Acc. 69.6%
	2023	[64]	UCB	Optimizing live transcoding tasks using edge computing	Performance better than baseline while reducing 92.3% operating time
	2023	[65]	DRL	Air reconnaissance and trajectory decision	Outperformed over baseline
	2024	[66]	Q-Learning	Safety policy optimization for automated driving	AVR 0.018, TV 86

based on computer vision and natural language processing technologies, and there is a huge demand for innovations in AI and multimodal technologies. However, these works provide a comprehensive review of SA systems in their specific domains, few of them reveal the impact of the development of AI and multimodal technologies on SA systems and future research directions.

Compared to other SA-related reviews and earlier works, this review provide the following contributions:

- Focusing on recently published works, the paper covers more than 100 references works from various databases (e.g., IEEE Xplore, Scopus) and is complemented by summary tables in different sections. Corresponding tables are summarized for the different sections with the aim to help researchers better understand the current state-of-the-art and trends.
- Encompassing a broad spectrum of AI-enabling technologies for SA, this paper synthesizes existing literature. It thoroughly analyzes and discusses architectures, models, and principles across three main categories of methods: machine learning, deep learning, and reinforcement learning.
- Addressing the impact of data analysis on SA systems in diverse environments, this paper highlights varying data sources and processing methods. It engages in a comparative discussion of the effects of single-modal and multi-modal processing within SA systems, along with an in-depth analysis of typical modal fusion methods in multimodal situations.

The remainder of this article is organized as follows. Section II makes a comprehensive analysis of AI algorithms and architectures used for SA systems, discussing the tasks they address in SA systems in terms of three broad categories: machine learning, deep learning and reinforcement learning. Section III discusses the impact on SA systems in terms of information sources, information types, multimodal collaboration and fusion. Section IV discusses the use of 3D technologies for spatial situation awareness. The following Section V gives categorized illustrations of some use cases and applications for AI and multimodal enabled SA systems. SA-related datasets for several fields are discussed in Section VI. Finally, challenges to be addressed, future trends and a summary of the article will be discussed in Sections VII and VIII.

II. AI METHODS AND ARCHITECTURES IN SITUATION AWARENESS

Over the past decade, developments in AI have taken SA systems to a new stage. The AI methods and architectures used are diverse. However, there are numerous ways to classify AI methods, such as supervised, unsupervised and semi-supervised learning based on whether human supervision is required [10]; online and batch learning based on whether dynamic incremental learning is possible [67]; as well as classification algorithms, regression algorithms, clustering algorithms, anomaly detection algorithms and migration learning etc., depending on the task being solved [68]. In reviewing the relevant literature, this paper divides AI methods applied to SA into three broad categories: machine learning, deep learning and reinforcement learning. Against the background of gradually increasing data volume and data variety, and growing practical needs, such a categorization can broadly reflect the general trend in the development of AI algorithms and SA systems. This section is therefore divided into three subsections to discusses the important elements for AI methods and architectures in SA, with Sections II-A to II-C discussing and summarizing the literature based on each of these three categories of approaches. Table 1 illustrates a summary of AI methods and architectures in SA discussed in this section.

Given the diverse focus, task variations, dataset disparities, and inconsistent utilization of evaluation metrics in the literature on situational awareness, it becomes arduous to directly compare data outcomes across different research studies. In an effort to try to avoid potential limitations and biases in the data and results, the following metrics were extracted and compared in this paper when comparing various research results. For classification tasks, accuracy, precision, recall, and F-measure, etc. are a few commonly used metrics:

$$Class.Metrics: \begin{cases} Accuracy = \dfrac{TP+TN}{TP+TN+FP+FN} \\ Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \\ F1_{score} = 2 \cdot \dfrac{Precision \cdot Recall}{Precision + Recall} \\ AP = \displaystyle\int P(R)\, dR \\ mAP = \dfrac{\sum AP_{classes}}{num_{classes}} \\ AUC = \displaystyle\int ROC\,(FPR)\, d\,(TPR) \end{cases} \tag{1}$$

The larger the value of the above metrics, the better the performance of the system. TP, TN, FP, FN stands for true positive, true negative, false positive and false negative, respectively. 'Positive' and 'negative' represent the result of prediction, whereas 'true' and 'false' represent whether the prediction is correct or not. $P(R)$ is a curve with precision on the vertical axis and recall on the horizontal axis, and ROC denotes the receiver operating characteristic curve with True Positive Rate (TPR) on the vertical axis and False Positive Rate (FPR) on the horizontal axis. In classification tasks, accuracy represents the ratio of the number of correct predictions to the total number of samples, which can reflect the overall performance of the model in the dataset, but in the case of sample imbalance, the model may 'cheat' (e.g., completely ignoring the categories with small sample sizes), resulting in accuracy not being a suitable metric for evaluation. Precision and recall, on the other hand, are used to evaluate the model against the

predicted results and the original sample, respectively. The F1 score provides a balance between precision and recall, with relative accuracy being more suited to imbalanced datasets.

For the regression task, assuming that set $X = \{x_1, x_2, \ldots, x_N\}$ and group $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$ are the actual and predicted values, respectively, the commonly used metrics are defined below:

$$Reg.Metrics: \begin{cases} MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \hat{x}_i\right| \\ MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2 \\ RMSE = \sqrt{MSE} \\ PCC = \frac{\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)\left(y_i - \hat{y}_i\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2 \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)}} \end{cases} \tag{2}$$

where MAE, MSE, RMSE and PCC are short for Mean Absolute Error, Mean Square Error, Root Mean Square Error and Pearson Correlation Coefficient, respectively. y_i and $\hat{y}_i$ represents another set of data for calculating PCC with set X. The smaller the values of MAE, MSE and RMSE or the larger the absolute value of PCC, the stronger the regression performance of the system. They both measure the predictive power of the model by calculating the difference between the predicted and true values. MAE treats all errors equally, which makes it unable to reflect the distribution of prediction errors. On the contrary, MSE and RMSE are more sensitive to predicted values with larger errors.

It is worth mentioning that some specific metrics are used to evaluate reinforcement learning models. Cumulative rewards are the total rewards earned by the model over a period of time, which is used as a basic metric to evaluate the performance of RL models; average rewards refer to the average of the rewards earned by the model over a period of time, which can be supplemented with cumulative rewards metrics in order to understand the stability of the model. The metrics described above may not cover all cases, and specific metrics will be described subsequently for individual studies.

A. MACHINE LEARNING-BASED APPROACHES

The performance of machine learning technology on classification and decision-making tasks has been revolutionary for SA tasks.

1) MONITERING-BASED PERCEPTION

It is common practice to use sensors to monitor environmental conditions and convert them into digital signals and use machine learning methods to process. The authors in [22] investigated the use of various detectors in smart homes to determine human behaviors. Among them is the use of SVM for posture detection of targets in video tracking systems. In this learning there are three pose classifications, 'standing', 'lying down' and 'sitting'. Using

$$K\left(x_i, x_j\right) = \phi\left(x_i\right)^T \phi\left(x_j\right) \tag{3}$$

as a kernel function, where ϕ is the function mapping the training vectors to a higher dimension. After training, a 'one-against-one' classification with a voting strategy would compare the incoming targets and determine which of the three postures it belongs to. For the processing of audio modality, the authors use a neural network-based detector to determine whether there is speech activity in each frame from parameters such as the energy and frequency of the audio. For fusing and utilizing the features from these modalities, a left-right Hidden Markov Model (HMM) was used at a later stage to learn eight states in the environment, including 'individual work', 'siesta', etc. The left-right HMM employs a fully connected structure, allowing the states of each node to progress either 'left' or 'right' or remain unchanged. In this model, we denote the HMM with a set of hidden states S and observable states O:

$$S = \{s_1, s_2, \ldots, s_N\} \tag{4}$$

$$O = \{o_1, o_2, \ldots, o_M\} \tag{5}$$

The transition probability matrix A consists of transition probabilities A_{ij}, denoting the probability of transition from hidden state S_i to S_j. Additionally, the emission probability matrix B comprises probabilities B_{jk}, representing the probability of emitting observation O_k given the hidden state S_j.

$$A_{ij} = P\left(q_{t+1} = s_j \,|\, q_t = s_i\right) \tag{6}$$

$$B_{jk} = P(o_k = v_k \,|\, q_t = s_j) \tag{7}$$

where q_t and o_k denotes the sequences of states and observations.

For individual situations, and 'presentation', 'speech', etc. for multi-person situations, respectively. When this work is applied to real-time SA, overall recognition rates of close to 90% can be achieved. It is worth noting that recognition in multi-person situations is often confused, which the authors attribute to multiple situations sharing the same features.

The authors in [12] investigated the use of eye-tracking signals and electroencephalogram (EEG) for SA perception related on air traffic control officers workload. The authors recorded EEG and eye-tracking data from participating air traffic control officers and introduced the SAGAT [3] freeze probe technique and NASA-TLX assessment scores to produce training data. For feature extraction of EEG and ET data, independent component analysis (ICA), fast Fourier transform (FFT), power spectral density (PSD), and Hilbert transform, etc. were used. The authors define a two-level SA classification (Fig. 1), which first determines whether it is a low SA, and later determines whether the low SA is associated with a high workload. Seven classification algorithms are evaluated here including logistic regression (LR), Radial, polynomial and linear basis function (i.e., SVM-R, SVM-P and SVM-L), Random Forest (RF) and Artificial Neural Network (ANN). Greater than 75% accuracy was obtained using

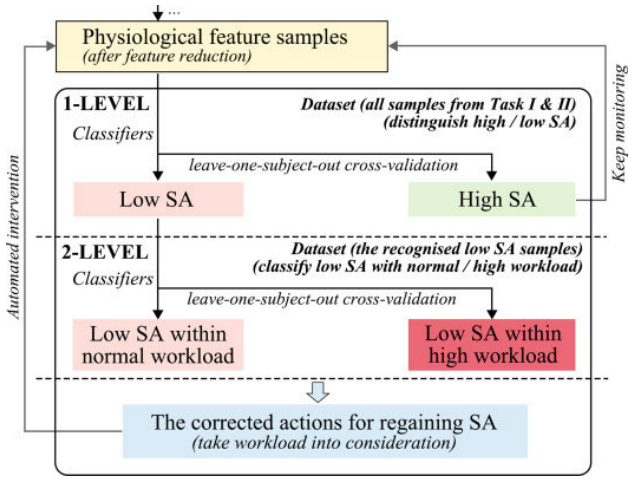

FIGURE 3. Classification of SA classes using biological data [12].

SVM-R in the first level of SA classification, whereas over 82% accuracy was achieved using LR in the second quarter of SA classification. These simulations suggest that integrating EEG and eye-tracking signals with machine learning classification algorithms can significantly improve SA perception in high-demand environments like air traffic control.

The authors in [13] explored leveraging visual inputs augmented by GPS data and high precision maps to enhance obstacle perception for unmanned aerial vehicles (UAVs) during taxi operations. The proposed method is multimodal, i.e., the camera image, the airport map and the GPS measurements. The camera images are inverse perspective mapped (IPM) to match the airport map, and then the Kullback-Leibler divergence (KLD) between the distribution of the navigation map and the camera images is minimized to obtain a calibrated navigation map and GPS measurements. The authors integrated probabilistic representations from each information source, combining Bayesian inference and the aforementioned data processing into a self-learning framework. The demonstrated experiments revealed that the self-learning method exhibited significant improvement in detecting smaller objects compared to the non-self-learning method. The authors point out that this self-learning process occurs throughout the taxing process of the UAV, so that the features of the image are constantly being learned and the weight of the navigation map resulting from the skidding process is constantly being increased as it is learned. Therefore, this Bayesian learning based approach has a high degree of interpretability.

The benefits of machine learning go beyond classification tasks. In [19], Tajdinian et al. proposed unsupervised anomaly detection algorithm based on kernel density estimation (KED) for power grid SA. KDE is used to estimate the distribution of the dataset and PCA is then employed to downscale the obtained probability density function for feature extraction. Technically, KDE is a probability-based, non-parameter technique for estimating the probability density function of a random variable. Assuming that there is a set of independent random variables:

$$X = \{x_1, x_2, \ldots, x_N\} \tag{8}$$

where X is drawn from an unknown probability distribution $PDF\ f\ (x)$, the KDE for x can be expressed as follows:

$$\hat{f}\ (x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{9}$$

where $\hat{f}\ (x)$ is the estimated PDF of the point x, K denotes the kernel function, n and h denotes the number of data points and the bandwidth parameter, respectively.

Similarly, Carlos et al. [37] not only employed machine learning method for classification task, but also innovatively applied the bag of words (BoW) model for data representation when they investigated how to use mobile phone sensors to analyze vehicle driving styles. This study only deals with unimodality, i.e., the acceleration sensor on the phone. As the acceleration is vector data and its orientation to the vehicle may be unknown, the authors used Principal Component Analysis (PCA) to downscale and manually calibrate the X and Y axes respectively during data pre-processing. It is interesting to note that acceleration data is usually a time series, whereas the data representation method BoW used by the authors is a typical natural language processing method which counts the number of times the subsequence of interest occurs in the entire sequence. When applied to acceleration data, the BoW model learns various characteristics of signal fluctuations, enabling recording, clustering, and encoding. This allows subsequent classification tasks to acquire more advanced features of the data that are not intuitive to the human observer. The classification task being an important aspect of SA of this study, the authors implemented two classifiers: (1) a binary classifier to distinguish aggressive from safe driving, and (2) a multi-class classifier to distinguish aggressive driving maneuvers. The evaluation involved four classifiers: Multilayer Perceptron Neural Network (MLP), Random Forest (RF), Gaussian Naïve Bayes Classifier (GNB), and KNN. The experiments showed that MLP performed best on the first classifier, which did not misclassify a single driving event on the authors' dataset, while GNB was slightly better than MLP on the multi-classification task, with an accuracy rate higher than 96%. It is evident that the matching of the computation method of feature vectors to the algorithm of the classifier may be important.

The machine learning models used in this stage of SA systems are more inclined to provide preliminary interpretations of the raw data, such as ICA and SVM used in [12], BoW and GNB used in [37]. They give statistically significant interpretations of the detected data in the environment, which allows some simple classification tasks to be realized and provide SA for the system. These methods are relatively easy to implement, but the limitation is a shallow understanding of the environment.

2) FROM MONITORING TO UNDERSTANDING

Using the data obtained from monitoring to hypothesize or fit the environment to a particular distribution enhances the system's understanding of the environment. Understanding is usually based on probabilities, as in the case of Bayesian networks, where various events in the environment can be modeled as a node in a directed graph, with edges between nodes representing probabilistic dependencies between events. This directed graph can be represented as $G = (N, E)$, where N and E are the set of all nodes and all edges, respectively. Let $X = x_n\,(n \in N)$ be the random variable represented by a node n in the graph, then we have:

$$P(X) = \prod_{n \in N} P\left(x_n \mid x_{pa(n)}\right) \tag{10}$$

where $P(X)$ denotes the joint probability of all random variables, whereas $pa(n)$ denotes the conditional probability of the parent of node n. Since real SA application scenarios are usually full of variations, the changing state over time may lead to the bias of traditional static Bayesian networks.

Dynamic Bayesian Networks (DBNs) introduce a temporal variable by learning the temporal dynamics of the state of each object in the training data in order to predict the object state when analyzing time series data. Mathematically, it can be expressed as:

$$P\left(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1\right) = P\left(x_t \mid x_{pa(t)}\right) \tag{11}$$

The transition between time steps is using a first-order Markov assumption, meaning that the state at time t depends only on the state at $t-1$, represented as $P\left(x_t \mid x_{t-1}\right)$.

Thekke Kanapram et al. In [15], the use of DBN as a data-driven model for anomaly detection, state prediction, and collective SA in connected and self-driving cars is investigated. Each self-driving car will train a DBN model using its own collected sensor data. In order to implement DBN state changes over time, the authors introduced the growing neural gas (GNG) clustering algorithm to learn the transition and conditional probabilities of each node in the DBN model. Each DBN network is then equipped with a Markov jump particle filter (MJPF) [16] for independently state estimation and anomaly detection. Technically, a state transition model is represented as:

$$x_t = f_{s_t}\left(x_{t-1}, v_t\right) + \epsilon_t \tag{12}$$

where f_{s_t} represents the state transition function associated with the mode s_t, u_t is the control input, and ϵ_t is the process noise.

Additionally, the observations y_t for detection at time t are related to the state x_t through a measurement function h_{s_t}:

$$y_t = h_{st}\left(x_t\right) + \delta_t \tag{13}$$

The experiment was set up with two self-driving smart vehicles equipped with LIDAR and 3D cameras. The trained vehicles were able to travel according to the learnt routes and make emergency braking decisions when detecting pedestrians crossing the road, demonstrating the performance of anomaly detection. In simulation tests, the anomaly detection of the proposed DBN model without wireless transmission loss (using the IEEE 802.11p standard) achieves an accuracy of 98.26%, and the accuracy is higher than 97% even when the transmission loss is increased, which justified the IoT enablement. Additionally, the authors in [17] proposed the use of hierarchical dynamic Bayesian networks (HDBN) to achieve SA of UAVs with respect to the radio environment, updating the transition probabilities using the GNG algorithm and applying MJPF for model inference for multi-level anomaly detection. Differently, the authors apply KLD abnormality (KLDA) and continuous level abnormality (CLA) to the discrete and continuous layers of the DBN network, respectively, in order to extend the MJPF for more accurate anomaly detection. And the model is further updated by incremental learning.

The authors in [14] propose an algorithm called FAVourite rOUte Recommendation (FAVOUR). This Bayesian learning based algorithm provides SA for multimodal route selection. To enable route selection, the input to the method introduces user preferences in addition to the necessary maps. These preferences are collected by means of a question-answer process and stored in the form of a binary comparison for easy subsequent training. The FAVOUR algorithm uses an incremental learning strategy, updating the approximation $\tilde{w}$ of user preferences once for each additional user using the equation $\tilde{w} = argmax_w p\left(w \mid T\right)$. Where $p(w)$ is the initial belief without any user preference, $p\left(T \mid w\right)$ and $p\left(w \mid T\right)$ are prior and posterior belief, respectively. The $p\left(w \mid T^i\right)$ is updated using Bayes' rule:

$$p\left(w \mid T^i\right) = p\left(T^i \mid w\right) \times \frac{p\left(w \mid T^i - 1\right)}{p\left(T^i\right)} \tag{14}$$

Given the constraints on repeatedly querying new users, the authors use a mass preference prior (MPP) approach to extract features from existing user data for the task of transfer learning. The MPP iteration incorporates the Kullback-Leibler divergence (KLD), halting when the difference between successive KLD values falls below a specified threshold. In the experiments, multiple modes of transport and route characteristics including distance, time and cost were assumed. The experiments show that the introduced MPP-based transfer learning improves accuracy by 4.3% to 10.6%. In contrast, the FAVOUR algorithm, based on Bayesian learning, achieves a maximum accuracy improvement of 23.6% compared to the traditional algorithm. It is worth noting that the introduction of transfer learning improves the model more significantly with less training data, whereas the benefits of incremental learning increase with the amount of data.

Decision trees (DT) are another effective classification and regression algorithm. In contrast to Bayesian networks, which use graph structures for modeling, the decision tree approach is based on a tree structure that divides various attributes of things and points to various possible outcomes. With supervised learning at its core, different branches are constructed artificially on sample features to form a tree

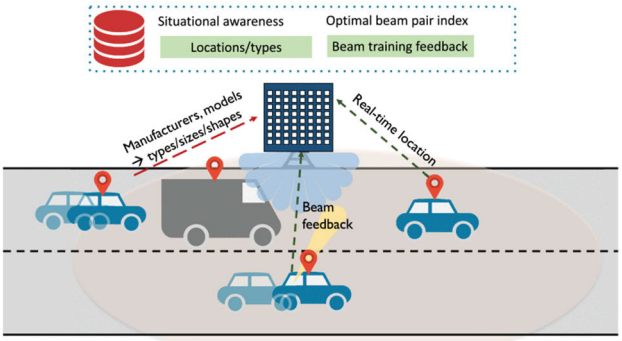

FIGURE 4. Empowering vehicle SA systems using machine learning [20].

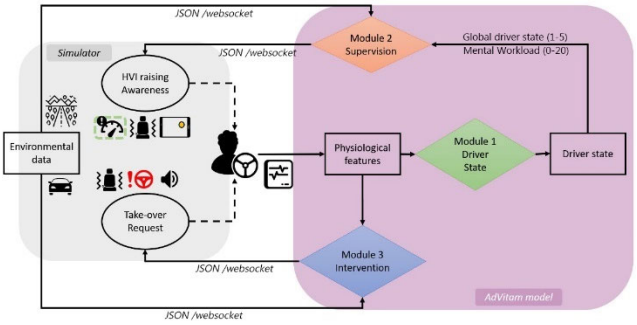

FIGURE 5. SA and corresponding actions for automated drivers using multimodal data [18].

structure. Frequently used decision tree-based algorithms are ID3 [69], C4.5 [70], CART [71], Random Forest (RF) [72] etc. Theoretically, let D be the data collected from the environment and $P(k)$ be the proportion of the k^{th} class of samples in the set D. The information entropy and conditional entropy of the data are (2) and (3), respectively.

$$H(D) = -\sum_{k=1}^{K} P(k) \log_2 P(k) \tag{15}$$

$$H(D \mid X) = \sum_{x \in X} P(x) H(D \mid X = x) \tag{16}$$

Measure the uncertainty of information D according to (2) and (3), i.e., the likelihood that an event will have different outcomes in the environment. To further determine the uncertainty of an event after obtaining a certain condition, the information gain G is obtained from (2) and (3) to be used as a selection criterion for the attribute.

$$G(D, X) = H(D) - H(D \mid A) \tag{17}$$

In SA, classifiers are commonly employed to identify events or states. Authors in [20] explore utilizing a multi-classification machine learning approach to enhance situational awareness for vehicle communication in millimeter wave environments. To ensure that the right beam is selected for communication while the vehicle is moving, the authors use the RF. The algorithm consists of an aggregation of multiple classification trees, can consider a wide range of non-linear and complex features. Predictions from individual classification trees are aggregated to determine the final prediction class. The input to this process is the location information of the vehicle and the output is a matching beam pair proposal. The SA function is implemented on the base station that communicates with the vehicle. The base station uses the vehicle information transmitted periodically by the connected vehicle with the location information to construct a location relationship map and to reason about suitable beam pair recommendations, which are then transmitted to the corresponding vehicle. The authors also evaluated RBF-SVM, gradient boosting, and found that the random forest algorithm outperformed the other algorithms by at least 14% in terms of beam alignment probability. However, the beam pairs derived using different classifiers did not differ significantly in the subsequent data transmission sessions. This underscores the significance of understanding the properties of the source modality and their impact on the SA system.

Decision trees scale well, which incorporates the concept of gradient to make the loss function fall quickly, such as Gradient Boosting Decision Tree (GBDT) [73], and its variants such as XGBoost [74] and LightGBM [75]. Zhou et al., in [21] utilized LightGBM to learn eye-tracking data in order to dynamically assess the vehicle driver's SA level and improve the takeover performance when switching between autonomous and manual driving. The authors normalized the degree of SA to a continuous variable in 0 and 1. SA prediction was modelled as a regression problem using a loss function with a combination of MSE and MAE. To construct the decision tree structure, the authors summarized 28 variables related to eye-tracking data for predicting SA. When LightGBM uses the full set of predictor variables to predict SA, it has achieved RMSE and MAE below 0.11 and below 0.087, respectively. On this basis, the authors introduced shapley additive explanations (SHAP), which uses shapley values to indicate the effect of predictor variables on SA. Utilizing only the top fourteen most influential predictor variables identified by SHAP further enhanced the performance of LightGBM.

In [18], the authors investigate the use of machine learning models for state perception, supervision, and intervention with human drivers in autonomous vehicles. The perception module used RF, ANN, and K-nearest Neighbor algorithm (KNN) to analyze the driver's physiological signal (ECG, EDA, RESP), to make predictions for two states of fatigue, one mental workload, three emotional states, and an overall SA with good performance. After obtaining the results of the several classifiers mentioned above, the fusion of modalities is then performed by thresholding, voting, and logical rules in order to reach a reconstruction of the driver's state. Therefore, a driver's state can be characterized on a scale of one to five from poor to good. The authors have designed rules in the supervision and intervention modules, respectively, which use the results obtained in the perception module to perform the appropriate actions on the driver.

Continuously learning new data is another viable approach in the face of SA's changing environment. As mentioned earlier [14] and [15] use the idea of incremental learning. The authors in [38] have investigated the use of online learning methods to achieve energy SA when IoT electric vehicles are charging and discharging to smart buildings. The method is based on Contextual Multi-Armed Bandit (CMAB) [76]. Confirmation of the building's grid operation modes, i.e., normal and abnormal, guided by sensors and state variables in smart buildings. The power distribution of the smart building will be uploaded to the cloud as a context, denoted as $e(t_n)$. The CMAB algorithm deployed in the cloud will identify $e(t_n)$ and search for similar situations in the historical context E_n as a reference. the CMAB algorithm will compute the overall probability distribution and try to make the choice with the highest 'reward'. Specifically, the reward expression $Q(t_n) \in [0, 1]$ is represented as:

$$Q(t_n) = Q^{safe} \cdot \left(d_n \left(1 - \exp\left(-Q^{ev}\right)\right)\right) \tag{18}$$

where Q^{ev} and Q^{safe} are the reward and penalty mechanisms for EV charging behavior and driver satisfaction. d_n is another penalty factor that denotes the completion rate of the energy requested from vehicles.

The authors have also defined two types of attackers i.e., internal and external attackers. They will cause privacy leakage potential to the system in the cloud and externally respectively. CMAB will take this into consideration while making a choice, i.e., it introduces a Tree based Gaussian Noise Aggregation (TGNA) algorithm while calculating the probability distribution, which randomizes the 'reward' mechanism of the CMAB algorithm. Different from the metrices shown on (1) and (2), average regret (AR) and cumulative regret (CR) is used as an important metric to evaluate the model. Based on the regret mechanism, regret increases when the system does not select the best vehicle and incurs a loss. A low value of AR and CR indicates a better scheduling strategy.

$$CR = \sum_{t=1}^{T} R(a_t) \tag{19}$$

$$AR = \frac{1}{T} \cdot CR \tag{20}$$

where a_t is one step of decision in a set $A = \{a_1, a_2, \ldots, a_T\}$, and $R(\cdot)$ denotes the quantified regret value, i.e., the difference between the current reward value and the maximum reward value, as:

$$R(a_t) = \max_{a \in A} Q(a) - Q(a_t) \tag{21}$$

After experimental comparisons, the CMAB-based model stabilized at 11% in AR, which is an improvement of 8.3% to 76.5% compared to the baseline model, while the CR decreased from 27.5% to 81.8% compared to the baseline.

To understand complex environments, more sensors are introduced [15]. The increase in the number of model parameters as well as the computational complexity allows SA systems to understand the changes in the environment over time, such as the HDBN used in [17] and CMAB used in [38]. Overall, machine learning models in this phase provide a deeper understanding of the semantic information in the environment, especially for temporal signals.

B. DEEP LEARNING-BASED APPROACHES

With the accumulation of data volume, complexity of algorithms and development of computing performance, machine learning introduced the concept of neural networks. While traditional machine learning algorithms may rely more on handcrafting in this aspect of feature engineering, neural network mechanisms allow deep learning algorithms to mimic the human brain for 'learning' capabilities. Deep learning algorithms that utilize deeper networks to mine data for information tend to outperform traditional machine learning algorithms.

1) AWARENESS BASED ON SERIALIZED DATA

The introduction of neural networks means a wider range of data to process, faster processing, and more robust feature extraction. A neural network usually consists of an input layer Y_{in}, hidden layers Y_{hid}^n and an output layer Y_{out}, defining the number of nodes in each layer to be $m_0, m_1, \ldots, m_{k-1}, m_k$, then the output vector of each layer is as follows:

$$\begin{aligned}
Y_{in} &= \left[Y_{in}^1, Y_{in}^2, \ldots, Y_{in}^{m_0}\right]^T \\
Y_{hid_1} &= \left[Y_{hid_1}^1, Y_{hid_1}^2, \ldots, Y_{hid_1}^{m_1}\right]^T \\
&\vdots \\
Y_{hid_n} &= \left[Y_{hid_n}^1, Y_{hid_n}^2, \ldots, Y_{hid_n}^{m_{k-1}}\right]^T \\
Y_{out} &= \left[Y_{out}^1, Y_{out}^2, \ldots, Y_{out}^{m_k}\right]^T
\end{aligned} \tag{22}$$

Classically, the forward propagation of layers other than the input layer can be represented as follows:

$$net_i = W_i Y_{i-1} + b_i \tag{23}$$

$$\begin{aligned}
Y_i &= f_i(net_i) \\
&= f_i\left(\left[net_i^1, net_i^2, \ldots, net_i^{m_i}\right]^T\right) \\
&= \left[Y_i^1, Y_i^2, \ldots, Y_i^{m_i}\right]^T
\end{aligned} \tag{24}$$

where W_i and b_i denotes the weight matrix and bias vector for each layer, respectively, whereas f_i is the loss function.

Benefiting from the automation of feature engineering in neural networks compared to machine learning, the process of applying neural networks to SA can be simplified as:

- Modeling and building network structures according to business scenarios;
- Training model by feeding labeled data into the neural network;
- The neural network automatically adjusts the hidden and output layer neuron weights and biases by the difference between the predicted and actual outputs;

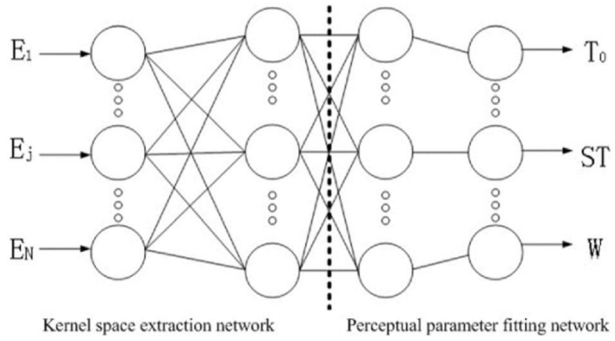

FIGURE 6. Using ANN to obtain perceptual parameters [40].

- Data prediction, feeding new data into the neural network and obtaining a predicted output.

The authors in [40] proposed an algorithm based on Artificial Neural Networks (ANN) and Petri Nets (PN) for SA of airport operations named Perceptual Petri Nets (PPN). Extending ANN using the parallel, concurrent, and asynchronous features of PN to make it suitable for scenario of airports services. The authors used various parameters from the system and various external conditions as labels and inputs, respectively, for training the ANN to obtain the desired perceptual parameters. The structure of ANN is illustrated in Fig. 4, where E_n represent values of external conditions, T_0, ST and W are negative exponential distribution, normal distribution and weight of the network, respectively. After using three influential external conditions, time, weather, and aircraft type, as input parameters for model training, the ANN achieved an accuracy rate higher than 81% on several perceptual parameters. The comparison to the baseline method (A-CDM) improved 7.62% in terms of average accuracy. In addition, PNN performs equally well on web services with high concurrency characteristics, proving its versatility.

Dantas et al. [46] suggest the use of a pre-trained ANN network to provide recommendations in their study on how to improve the SA of air combat pilots. The authors simulated 10,000 air combat scenarios and used information such as altitude, radar warning, aircraft position and type as variables for training a four-layer ANN based on expert advice. To prevent overfitting, 20% Dropout was set at each layer. This supervised learning method was tested with an accuracy over 92%, which demonstrates that ANNs as a classifier is effective in making recommendations for air combat decisions.

Zhong et al. [49] investigated the use of ANN with fuzz testing mechanism for SA of traffic violations of self-driving cars in simulated situations and proposed the AutoFuzz algorithm. The principle behind fuzz testing is to identify conditions that trigger errors in the testing system by continuously inputting random variables. Training data is gathered from various vehicle sensors (such as cameras and radar) within the simulated environment, resulting in a multimodal dataset. AutoFuzz trains an ANN classifier in each test for predicting the confidence that a randomly generated test case causes a system error (i.e., a simulated traffic accident occurs), and implements the one with the highest confidence. This structure allows for the implementation of incremental learning, where knowledge gained from previous training is used to filter out a portion of the test cases in advance to improve efficiency. Compared to a decision tree-based baseline method that also has incremental learning capabilities, AutoFuzz consistently detects 10-39% more traffic violations in a variety of simulated environments.

Incorporating neural network mechanisms into existing machine learning algorithms is one way to implement deep learning. Yiu et al. in [48] investigated the use of Bayesian neural networks (BNN) to identify potentially subjective hazardous environments from EEG data. Similarly to [12], the methodology was contextualized around the work of Air Traffic Control Officers and applied the NASA-TLX index to assess the mental state of participants for the dataset. The EEG data was similarly preprocessed using the ICA algorithm. As mentioned earlier, Bayesian networks are a process of calculating a posterior distribution based on a prior distribution using Bayesian formulas. In BNN, the estimated posterior distribution is then obtained by training the model weights of the neural network. The authors implemented a binary classification network using BNN, i.e., the outputs are only 'good visibility' and 'low visibility'. Mathematically, the collected data D can be represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\} \tag{25}$$

where x_i and y_i represents the input data and corresponding labels, respectively. The initial and prior distribution $P(\theta)$ of the BNN is set before observing any data, where θ denotes the initial weights and biases of the network. Since then, using the Bayes' theorem, the posterior distribution can be calculated by:

$$P(\theta D) = \frac{P(D\theta) P(0)}{P(D)} \tag{26}$$

where $P(\theta D)$ denotes the updated distribution of parameters after observing data.

Compared to traditional machine learning methods (DT, RF, SVM and LR), the BNN prevailed with accuracy of 66.5% and F1 score of 61.4%. Facing the problem of interpretability of deep learning, the authors used the SHAP value to imply the top ten most important features of neural networks heavy. The three highest of these features correspond to the temporal, frontal, and parietal cortices of the human brain that process auditory, visual, and integrative sensory information, respectively, reflecting the rationality of the methodology.

To better process sequential information, adding a 'recurrent' mechanism to the ANN can increase the correlation between the front and back inputs, i.e., recurrent neural network (RNN) [77]. Each time a neuron processes an output it will be multiplied by a weight and fed back into the input, in such a way that allows the network to have some 'memory' and to 'learn' over time and even 'forget knowledge.

The characteristics of RNN structures make them widely used in natural language processing tasks. The authors in [39] investigated the use of deep learning algorithms to detect domestic violence related posts and comments in social media. For feature extraction, the authors have taken two pre-trained embeddings, Word2Vec and GloVe, for comparison; the classification models involved in the comparison include deep learning algorithms such as RNN, LSTM, GRU, BiLSTM, and CNN, as well as traditional machine learning algorithms such as SVM, LR, DT, and RF. The experimental results show that GRU and BiLSTM with GloVe embedding achieved optimal indices with 91.78% and 91.29% accuracy, respectively. The experimental results show that GloVe has a higher performance on this work, the reason may be that the essence of GloVe is to construct heterogeneous co-occurrence matrices containing all the words that have appeared, which remembers the representations of these words in the context very well as compared to the predictive type of the word2vec approach. In addition, the experimental results clearly demonstrate the strong performance of classifiers using deep learning algorithms over machine learning.

The authors in [42] investigated the use of deep learning algorithms for the prediction and perception of crime types and crime risk levels for text-based criminal case summaries. In order to explain the crime risk level more intuitively, the 21 crime types involved in the classification were artificially weighted, and the crime risk rating was described as

$$\begin{aligned} CRS = {} & WC \cdot (WG + WA) \\ & + WP \cdot (WG + WA) \\ & + WM \cdot 10 \end{aligned} \tag{27}$$

where WC, WG, WA, WP, and WM are the weights of the crime type, gender, age, physical injury, and material injury, respectively. In order to classify texts and predict CRS, 568 keywords were defined and assigned to 21 crime types. After that, a four-layer DNN structure consisting of fully connected layers and a three-layer CNN structure were constructed for predicting the CRS. The experiments show that the CNN structure comprehensively outperforms the Bi-DNN in terms of accuracy, precision, recall, and F1 score, reaching 91%, 92%, 82%, and 84%, respectively. This situation also occurs when comparing SVM and Naïve Bayes.

Overall, when SA systems process time-series signals, the use of basic deep learning models, such as ANN, BNN, RNN, etc., can improve the ability to fit the environment compared to traditional machine learning models, as demonstrated in work such as [42] and [48].

2) AWARENESS BASED ON VISUAL INFORMATION

Vision is widely regarded as one of the primary senses for humans, playing a crucial role in understanding complex systems. Beskow et al. in [41] investigated the use of deep learning networks to enable the perception and classification of Internet culture (i.e., memes) for social media tweets. The authors analyzed the problem from a modal perspective: tweets contain memes information that may be embedded in text and images, including text in images. An optical character recognition (OCR) technique based on Google Tesseract is applied to extract text from images. After extracting the word embeddings of the text using GloVe, LSTM is used as a text classifier to transform the obtained word embeddings and hidden vectors into new hidden vectors. For image classification, the convolutional neural network (CNN), which is the most popular in the computer vision direction, is used. To achieve multimodal fusion, the authors modified the fully connected layer at the tail of the CNN network to ensure that its output is a vector rather than a classification result. Interestingly, the authors consider face information to be equally important. A face encoder that also outputs a vector is added to the method. Ultimately, the vectors output by the LSTM, CNN, and face encoder will be used as input to a fully connected layer, and the classification results will be output from a subsequent SoftMax layer. In their experiments, the authors on the text classifier side used LR, SVM, and Naïve Bayes to compare with LSTM, and the result is that LSTM not only has superior accuracy and F1 score, but also has a 12% to 33% improvement in recall. VGG18, ResNet18 and Inception-v3 are compared as CNN-based visual classifiers. The difference between them was not significant, with accuracy and recall close to 95%. In the multimodal case with simultaneous consideration of visual, textual and facial information has the best performance, with both accuracies, F1 score and recall higher than 96%. The authors assert that this method's performance is at least 8 times superior to that of traditional template-based classifiers, further underscoring the advantages of multimodal approaches.

It is important to mention the contribution of CNNs on object detection, an important way for providing visual perception. The latest techniques introduce concepts such as attention mechanisms [78], high-resolution representation [79], hierarchical perception [80], inter-target relations [81], etc., to improve performance. These models can be used as a backbone to provide visual perception for situational awareness systems with different requirements. In investigating how to enhance the SA of autonomous surface vehicles, Liu et al. [23] introduced the CNN-based YOLOX network as a ship detector. Among the four versions of YOLOX, the authors considered the hardware limitations, so they chose YOLOX-s with the lowest computational cost to accommodate edge devices in the maritime IoT. Four Res-Blocks (ResNet) form the backbone of this network, flanked by a Feature Pyramid Network (FPN) for fusing information from different receptive fields. The authors employed a transfer learning approach to train the YOLOX-s network, i.e., the use of the pre-trained model obtained by training on the COCO dataset. 2268 images were used as a fine-tuning training set to update the parameters while the backbone of the network was partially frozen. Thanks to the numerous sensors (e.g., LiDAR, radar, cameras, satellite navigation systems, etc.) in the maritime IoT, an augmented reality (AR)

navigation system based on multimodal fusion is proposed. The detected vessels are transformed according to the relationship between the proposed pixel coordinate system and the world coordinate system, which allows to present information such as latitude, longitude, speed and heading in the system. Similarly, the authors in [51] introduced a framework for maritime surveillance that uses a deep learning approach to implement video content analysis of videos captured by surveillance cameras on board ships. RetinaNet, YOLOv3 and YOLOv3-tiny are compared as ship detectors in the experiment. YOLOv3-tiny with fewer parameters suffers a slight loss in accuracy. Worth mentioning that when using videos as the training set, the authors specifically added videos containing occluded vessels to increase the performance of the detector.

Visual signals may be affected by the environment and degrade the imaging quality of image sensors. How to solve the problems of occlusion and blurring of visual signals is a challenge for vision-based SA. In [53], the authors propose DADFNet to overcome the effect of possible haze on visibility in images. This network is based on an encoder-decoder structure to reconstruct clear images by improving a typical U-Net. The encoder extracts high-channel semantic features from the blurred image using multiple convolutional units and maximum pooling operation, and then the image is reconstructed by the convolutional units and bilinear interpolation operator in the decoder. The authors add an adversarial loss to the traditional L1 loss to constrain the image reconstructed by the network. On the other hand, the authors in [54] introduced point cloud data detected by LiDAR and AIS data containing ship position information to assist in 2D target recognition of ships using YOLOv6. Projecting the point cloud data onto the 2D image improves the localization of target monitoring while the AIS data is enhanced in visualization. This provides enhanced support for ship navigation.

Providing visual perception for SA systems is the focus of deep learning methods to differentiate them from other methods. CNN-based methods provide semantic-level perception and understanding for SA systems.

3) DEEP LEARNING EMPOWERED PROJECTION

Prediction tasks essentially involve analyzing existing data to predict trends or characteristics of future events or unknown data.

Transformer, based on encoder-decoder structure and self-attention mechanism, is a very popular approach in recent years. Multi-head attention allows attention to be computed for each position in the input sequence, which allows the algorithm to obtain global contextual information, and is well suited for processing sequential data. It is able to identify more important and stable inputs and is superior to RNN at the feature processing level. Parallel computation is applicable during training and hence also outperforms RNN in terms of training efficiency. Mathematically, start from self-attention mechanism, query Q, key K and value V are three vectors that wanted from the input embeddings

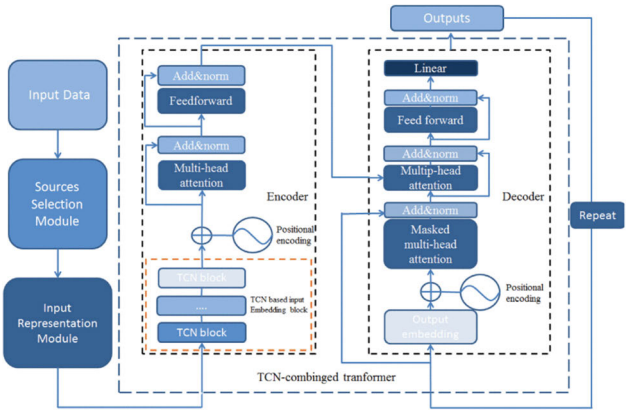

FIGURE 7. 7 SA model with temporal concept constructed using TCN as an embedding method for Transformer [45].

$X = \{x_1, x_2, \ldots, x_N\}$, and they can be represented as:

$$Query : Q = XW_Q \tag{28}$$

$$Key : K = XW_K \tag{29}$$

$$Value : V = XW_V \tag{30}$$

where W_Q, W_K, W_V are learnable weight matrices. Therefore, the attention could be calculated as:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{31}$$

In multi-head attention, 'heads' comes from the repetition of the (31), let h represents the number of heads.

$$MultiHead\,(Q, K, V) = Concat\,(H_1, H_2, \ldots, H_h) \cdot W_O \tag{32}$$

where H_h denotes each head and W_O is a learnable weight matrix for output transformation [82].

In [45], Yin et al. combined the features of Transformer and CNN for investigating the long-term SA of network conditions. Given that the embedding method in the traditional Transformer is not able to extract features directly on the time series, the authors have used Temporal Convolutional Network (TCN) instead of it to form a SA model called TCN-combined Transformer. TCN allows time series to be input directly without further coding, and subsequently Transformer calculates the correlation and periodicity between network traffic sequences to predict future sequences. Technically, the multi-head attention is edited as:

$$MH = ConvSA\,(X) \tag{33}$$

where $ConvSA()$ denotes the process of combing the TCN and original multi-head attention similar to (31).

Experiments show that the TCN-Transformer network outperforms both networks individually. The performance of RNN-based GRU and LSTM deteriorates drastically when the sequence length is too long, whereas the proposed network based on Transformer does not suffer from this problem.

In complex systems, relationships between objects often surpass the capabilities of conventional data structures like

arrays and trees. Graph structures effectively depict interrelationships and influences among objects. Introducing the idea of graphs into deep learning methods is a novel direction in recent years. The authors in [47] investigated the use of graph-based neural network algorithms for trajectory perception and prediction of ships at sea. The authors proposed the spatio-temporal multigraph convolutional network (STMGCN), comprising two key layers: the STMGC layer, based on graph convolutional networks (GCN), and the self-attentive temporal convolutional layer (SATCL), leveraging a self-attention mechanism. STMGCL consists of three STGCNs that perform spatio-temporal convolution operations on graphs composed of AIS data. The reason for using three networks in parallel is to generate embeddings from the three graphs 'social forces', 'time to nearest point of approach' and 'size of surrounding ships' respectively. The original 2-D graph convolutional operation can be defined as:

$$f_{out}(X) = \sigma\left(\sum_{h=1}^{K}\sum_{w=1}^{K} f_{in}(\psi(x,h,w)) \cdot W(h,w)\right) \quad (34)$$

where, σ denotes the activation function, σ represents the sampling function which indicates the neighbor nodes of the x position and W is the learnable weight matrix. The authors improved the operation for three graphs:

$$f_{out}\left(v_m^\theta\right) = \sigma\left(\sum_{v_n^\theta \in \mathrm{B}\left(v_m^\theta\right)} \frac{1}{Z_m^\theta\left(v_n^\theta\right)} f_{in}\left(\psi\left(v_m^\theta, v_n^\theta\right)\right) \cdot W\left(v_m^\theta, v_n^\theta\right)\right) \quad (35)$$

where v_m^θ and v_n^θ are two vertices at the time θ representing the distance relationship.

Besides, SATCL contains a TCN, a self-attention module and a fully connected network. Its role is to compute the correlation of the three graphs to enable prediction of ship trajectories. Comparing the deep learning methods such as LSTM, and GRU, the STMGCN comprehensively outperforms in average displacement error (ADE), final displacement error (FDE), and maximum displacement error (MDE). They are represented as:

$$ADE = \frac{\sum_{m\in M}\sum_{n\in N_m}\left(\varphi_m^n - \hat{\varphi}_m^n\right)^2 + \left(\lambda_m^n - \hat{\lambda}_m^n\right)^2}{\sum_{m\in M}\hat{N}_m} \quad (36)$$

$$FDE = \frac{1}{M}\sum_{m\in M}\sqrt{\left(\varphi_m^n - \hat{\varphi}_m^n\right)^2 + \left(\lambda_m^n - \hat{\lambda}_m^n\right)^2} \quad (37)$$

$$MDE = \frac{1}{M}\sum_{m\in M} max\sqrt{\left(\varphi_m^n - \hat{\varphi}_m^n\right)^2 + \left(\lambda_m^n - \hat{\lambda}_m^n\right)^2} \quad (38)$$

where $[\varphi_m^n, \lambda_m^n]$ and $[\hat{\varphi}_m^n, \hat{\lambda}_m^n]$ are respectively the Cartesian coordinates of the m^{th} vessel's ground truth trajectory at time n.

In [50], the authors innovatively combined digital twin technology with graph theory to propose graph digital twin (GDT) for SA of the energy Internet. The authors abstract multiple Phasor Measurement Units (PMUs) in the Energy Internet into graphical structures by treating the PMU as the subject of the digital twin. After GCN processing of the graph, there are two designed classifiers: (1) a binary classifier for predicting stability for the whole graph, and (2) a multi-classifier for detecting fault types for individual nodes. Multiple experiments have shown the method to be approximately 99% accurate in stability prediction and at least 94% accurate in fault type detection.

The authors in [52] proposed SA-LSTM based on LSTM to solve the task of trajectory prediction of self-driving cars under off-road road conditions. Compared to [47], this work provides a more comprehensive trajectory prediction including short, medium & long term. Among them, residual LSTM and autoregressive LSTM are used for short-term and medium-term prediction, respectively. The authors point out that spatial inference is more important in long-term prediction, so a CNN-based situational awareness extraction module is activated before the autoregressive LSTM aiming at sensing the level of risk within the environment. The authors collected over 11,000 frames of data for training and validation by building a game platform. The experimental results show that compared to vanilla LSTM, SA-LSTM obtains higher accuracy in short-medium and long-term prediction, 0.0153, 0.0260 and 0.0394, respectively. This metric improves the performance from 16.73% to 32.99% compared to baselines such as GNN and MLP.

4) MULTI-SENSOR ENHANCING AWARENESS

Multi-sensors are an effective way to enhance the sensing range. The authors in [43] investigated fluid pipeline related SA. The data source for the approach is a network of multiple wireless pressure sensors in the pipeline, and these sensing always produce time series data. In the part of detecting leaks in pipelines, the authors used sliding windows to compute three features based on similarity, namely 'leak', 'pump' and 'valve'. Bidirectional Long Short-Term Memory (BiLSTM) based on RNN structure was used as a classifier. The accuracy of the method was experimentally measured to be at least 90%, with less than 5% false discovery rate. Compared to other ML methods, the problem of false alarms and leakage is well solved. For the leak localization function, the authors calculated the time difference between anomaly-induced pressure fluctuations reaching different sensors in the wireless sensor network. Specific localization is obtained using a set propagation speed and pipe length.

Based on an encoder-decoder architecture, the authors in [44] investigated stability detection and SA of power grids using deep learning algorithms and proposed the encoder stacked classifier which is based on auto-encoders. Firstly, denoising autoencoder is used during data preprocessing in order to reduce the noise in the raw data. Multiple encoders are used to learn the distribution of the contaminated data and thus learn how to characterize the data. The same number of decoders are then used to reconstruct the feature vectors into

signal data for validation. The backpropagation algorithm will be applied to the encoder and decoder training. The trained autoencoder network will discard the decoder part as there is no longer a need to reconvert the learnt features into signal data. A fully connected layer and a Softmax classifier will replace it and perform a binary classification task. A dataset consisting of 42944 grid voltage related time series is used for training and validation. The proposed algorithm achieves 99.53% in accuracy, better than 98% of the traditional multilayer perceptron (MLP i.e., ANN). In addition, since the encoder learns how to repair the corrupted data, it has half the false positive rate of the MLP, which reflects the robustness of the algorithm.

C. REINFORCEMENT LEARNING-BASED APPROACHES

1) CONTINUOUS PROJECTION FOR SA SYSTEMS

Realistic environments are ever-changing and hard to predict, whereas the core of SA is the perception, comprehension and projection of the environment. On this basis, it is possible for the SA system to provide feedback on the response to the environment. Traditional supervised or unsupervised learning always tries to extract information in a fixed dataset. This may be incomplete for SA. Reinforcement learning (RL) provides a great way to think about problem solving. It learns optimal solutions by trial and error as the agent interacts with the environment. It is continuous and ongoing [67]. This may enhance SA from static and external to dynamic and real-time.

Temporal Difference (TD) learning is a typical RL algorithm. Similar to Monte Carlo methods, it does not need to re-establish a complete knowledge of the environment but learns the value function $V(s)$ directly from experience using TD errors. Similar to dynamic programming methods, it allows to boost on the estimation results in real time without waiting for the whole event to finish. Its updates obey the rules:

$$V(s) \leftarrow V(s) + \alpha \times TD_{error} \tag{39}$$

where α is the learning rate and

$$TD_{error} = r + \gamma V\left(s'\right) - V(s) \tag{40}$$

The authors in [55] introduced reinforcement learning to accomplish security SA for smart grids. The authors conceptualize all users within the grid, including both legitimate and malicious entities, as players (agents) within a network framework grounded in game theory. Therefore, the strategies in the game are the behaviors of the users in the network. A SA system based on TD learning and neural networks is proposed (Fig.6). The input layer of the system takes the strategies (behaviors) of all the players (users) at a single time as a single input, learns these strategies through the hidden layer of the neural network based on TD learning, and outputs a vector *Situation* $(t+1)$ that contains information such as node addresses, attack states, and attack events. Weights are updated following the second input, enabling players to learn

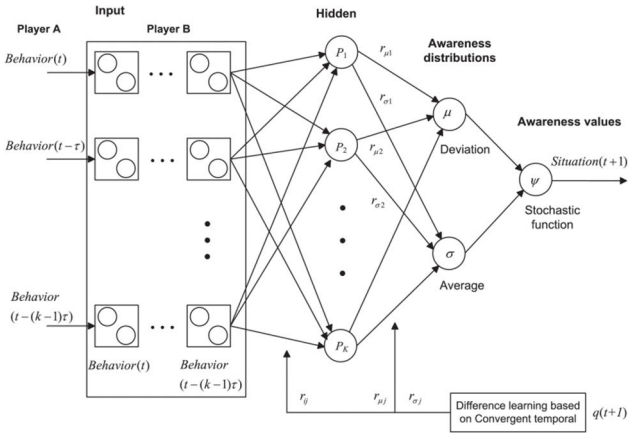

FIGURE 8. Integrating TD learning, neural networks, and game theory to SA [55].

from strategies employed by other players in the previous phase and adjust their behaviors accordingly to pursue what they perceive as the optimal strategy. The authors conducted experiments in a simulated environment. The perception rate of the proposed method is consistently higher than that of the baseline method in a setup of nine attacks.

The authors in [60] investigated the use of reinforcement learning algorithms for SA of communication environments in high-speed railway scenarios. In order to adapt to the ever-changing environment during high-speed mobility, the authors used TD learning to simulate the parameter switching mechanism during communication. Based on the communication environment, the authors define a tuple containing the state (i.e., information in the environment), action (i.e., switching parameters), reward (i.e., behavioral reasonableness) and policy (i.e., TD value). Trained according to the updated rules described in the previous section, the agent will always find the most suitable switching parameter later in the iteration.

2) FROM PROJECTION TO DECISION

Decision making is a series of interventions by an intelligent system after it has an expectation of a future state. An important contribution of Q-Learning is the introduction of the concepts of 'state' as well as 'state update'. It is an extension of TD Learning, which is the off-policy approach, i.e., it does not find the optimal policy by learning the policy directly. Its updating rule [56] is defined as:

$$Q_{new}(s,a) = (1-\alpha)\cdot Q(s,a) + \alpha\cdot\left(r + \gamma\cdot \max a'\cdot Q\left(s',a'\right)\right) \tag{41}$$

where:

- s and s' represents the current and next state.
- a and a' represents the action and possible action in state s and s', respectively.
- r denotes the reward received after taking action α.
- $\alpha(0 < \alpha < 1)$ is the learning rate of the method.
- γ $(0 < \gamma < 1)$ is the discount factor that affects future rewards.

The authors investigate the use of reinforcement learning to allow vehicles to maintain SA and trust assessment of their environment. Nodes within the vehicular network are treated as agents with varying levels of participation. In this approach, Q-learning is utilized to dynamically adjust the trust computation strategy for individual nodes, enabling adaptation to the evolving environment while ensuring consistent and reliable trust assessments. The authors define state as the agent's judgement of internal and external information, where internal information is the a priori knowledge that the agent has built up over its history, and external information is the information that the agent receives in real time. When receiving internal and external information with different views, it may affect the choice of strategy. The Q-table $Q(s, a)$ is updated according to the update rule. The authors conducted a simulation experiment involving over a thousand intersections and thousands of roads, distinguishing between two types of vehicles: normal and malicious. The smaller the share of malicious vehicles, the more accurate the system is in perceiving the environment with events (close to 100%). And the system still performs reasonably well with half of the malicious vehicles. Similarly, The authors in [58]{Citation} proposed a Q-learning based context-aware routing mutation algorithm (CQ-RM) in their research on SA for network security using reinforcement learning methods.

Deep Reinforcement Learning (DRL) were created when combining Deep Learning Networks and Reinforcement Learning methods. For example, Q-learning is usually model-free, whereas a deep learning approach is introduced, it is deep Q-learning (DQN). The authors in [59] used DQN in their study of spatial SA and autonomous maneuver decision-making in air combat. 3D perception in space relies on a YOLOv6-based 3D target detector. After commutation, the detected 3D target is given a six-degree-of-freedom attitude estimate, which is fed into the DQN as a state in Q-learning. A three-layer fully connected network forms the DQN, which is used to obtain the optimal Q-value, i.e., the policy, by means of gradient descent. Base on the Q-function represents by (41), DQN has neural network weights θ given the state s and action a. Using the Mean Squared Error (MSE) as the loss function, the target Q-Value is computed as:

$$Target = r + \gamma \cdot \max a' \cdot Q\left(s', a'; \theta^{-}\right) \tag{42}$$

where θ^{-} denotes the weights of a separate target network which are periodically updated from the main Q-Learning network.

Aircraft deployed with the algorithm achieved between 84.6% and 99.1% chance of winning a simulated air battle, demonstrating the performance of the DQN. Similarly, the authors in [65] combined Soft Actor-Critic network with ANN to empower UAVs to sense radar signals as well as make maneuver decisions.

In [57], the authors construct a DRL-based decision-making strategy for overtaking self-driving cars. Using state variables (vehicle speed, relative distance, etc.) and control actions (acceleration, lane change, etc.) as inputs, they construct a neural network to approximate and update the Q-table in the decision problem. Its network can be represented as:

$$Q(s, a, \theta) = V(s; \theta_1) + A(s, a; \theta_2) \tag{43}$$

where V and A denotes function for states and actions, respectively. θ is the learnable parameters of function V and A.

Similarly, in [61], the authors introduce transfer learning based on the DRL algorithm, i.e., the pre-trained neural network in another task (source task) is fine-tuned to the new target task, which also thought to introduce the a priori knowledge from the source task and transform it into the knowledge needed for the target task. To differentiate from the pre-trained network, the authors explored more possible decisions by adding randomized actions of the agent. In order to obtain a better state representation of the sensory data, the authors in [62] added a convolutional neural network and an attention layer before the DRL model learning parameters in order to extract global and important local information from the sensory data, respectively. Specifically, decision-making and action control belong to discrete and continuous action spaces, respectively, and their raw data are passed through the CNN layer to obtain new state representations, while the attention layer assigns different levels of attention to different vehicles at the same moment, which is a more intuitive and effective approach than the traditional RL algorithms that implement the two action spaces, decision making and vehicle control, together.

It is worth mentioning that the multi-armed bandit (MAB) mentioned in Section II-A is also a typical reinforcement learning method. In [63], Costa et al. proposed Bayes-Adaptive Contextual Multi-Armed Bandit (BA-C-MAB) algorithm in order to achieve species perception of insect-related signals. Similarly, Liu et al. [64] used a UCB-based reinforcement learning approach to risk-aware the performance of edge devices using contextual information of the devices.

In a recent study, Zhang et al. [66] discussed the use of reinforcement learning approaches to approximate the optimal bounds of safety policies in autonomous driving to address the problem of overly conservative safety policies, proposed Safe Reinforcement Learning with Dead Ends Avoidance and Recovery (DEARRL). The proposed method distinguishes the degree of danger of the state to constrain and correct the security policy. The method was evaluated using the number of constraint violations (TV) and the average constraint violation rate (AVR). Lower values of TV and AVR represent a lower number and likelihood of the method being violated in the task, i.e., higher security. Compared with the baseline method, the proposed method obtains the minimum TV (86 times) in several environments while the baseline method obtains 3330 times. In addition, the AVR value is 0.018, which is lower than the 0.519 of the baseline method, implying that its constraints on the safety states achieve effective performance.

D. PHASE SUMMARY OF AI ARCHITECTURES AND METHODS FOR SITUATION AWARENESS

The survey work in this section reveals the following key points and insights for AI architectures and methods in SA:

- Approaches based on AI technology mainly provide or enhance situation awareness for complex systems through tasks such as data representation, feature extraction, classification, and decision making, which basically covers the three-layer model of SA.
- With the shift from machine learning to deep learning, multiple works have experimentally demonstrated an increase in accuracy, e.g., [39], [42], [48].
- Reinforcement learning achieves a higher level of situational awareness for more informed decision making in complex dynamic situations through adaptive decision-making strategies, provides a reliable decision-making layer beyond the three-layer SA model to enable the shift from sensing the environment to influencing it.
- Although AI-based algorithms perform well on prediction tasks, most of the work in which AI techniques have been used does not include prediction of SA as a primary task, but rather focuses on achieving advanced perception and further understanding.
- The volume of data and the number of modalities have a decisive influence on the capability of SA. Compared to unimodal systems, situation awareness using multimodal systems usually has a stronger sensing capability, and more work has been done to achieve prediction of situations, e.g., [18], [21], [40], [55].

III. MULTIMODAL SYSTEMS IN SITUATION AWARENESS

Modalities are carriers of information, and data from multiple sources recorded simultaneously in an environment may present semantic correlations or complementarities. To achieve a more comprehensive perception in SA, simultaneously sampling the environment using different modalities enables the system to perceive the environment from various perspectives. It is possible to access patterns that are not captured by a single modality when multimodalities are considered together. For instance, a single ECG data may be able to determine whether the participant is exercising, whereas combined with accelerometer data may be able to suggest a specific type of exercise (e.g., running, cycling, etc.). This requires that the system uses a suitable multimodal fusion method. On the other hand, when designing an SA system. The effectiveness of data representation in capturing modalities within the environment depends on the careful selection of data and modalities aligned with SA goals, along with the utilization of their respective representations. Despite the many benefits of multimodal systems, multimodal approaches have not dominated all recent SA-related work, and unimodal approaches may be successful in achieving their goals in some single environments. Therefore, in Section IV-A, approaches to modal representation will be discussed from the perspectives of unimodal systems,

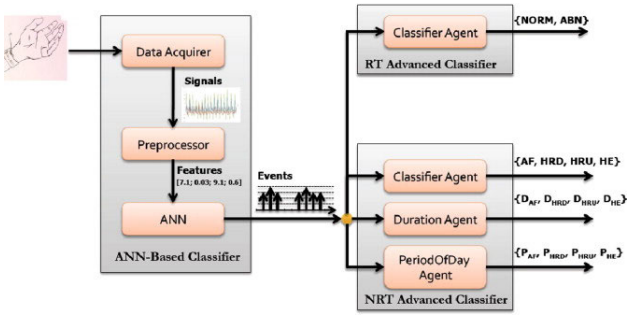

FIGURE 9. Motion data recording using wearable sensors and representation using statistical features [83].

whereas Section IV-B will delve into multimodal representation and fusion approaches, drawing from a diverse array of literature, to provide the reader with more comprehensive information through comparison. The discussion in this section is also based on the description of related work and differs from Section II in that it focuses more on data acquisition, modality selection, corresponding representations and multimodal fusion methods for implementing SA systems that cope with different environments. Table 2 illustrates a summary of different types of data and their corresponding representation methods used in SA-related works. Table 3 indicates the typical multimodal fusion methods.

A. REPRESENTION METHODS FOR UNIMODAL SA SYSTEMS

Sensor deployment is essential for sampling the environment. Accordingly, different sensors have different characteristics. The authors' work in [83] focuses on the detection of motion disorders in individuals with autism. A single accelerometer deployed on the participant's wrist was used as the primary sensor for action perception. The data stream it produces is three-dimensional, capturing acceleration changes in the x, y, and z axes. To extract features from this time-series data, the authors computed various metrics for the signals in each axis, including mean absolute value, root mean square, variance, standard deviation, waveform length, and over-zero. These features will represent the signal in terms of hard-to-observe dimensions. Movements such as arm rotation, swinging, and lifting were perceived from these features, as well as repetitive patterns of movement that are characteristic of patients with mobility disorders. The authors used an ANN-based classification network to classify the extracted features with an accuracy of over 99%. This result justifies the data acquisition and feature extraction part.

In the study of vehicle motion perception and anomaly detection, the authors used the spatial coordinates and motion trajectories of vehicles as data sources in [84]. Within a probabilistic framework, a DBN is built for modelling and representing the position information. The DBN architecture can reasonably observe the changes in position coordinates and motion trajectories produced by a moving vehicle. As time progresses, the DBN may learn correlations between certain random variables in the data to better represent the

data. The authors used Gaussian Process (GP) regression to correlate spatial coordinates and displaced movements to obtain the most probable movement patterns. Mathematically, the utilize of GP is to find a function $f(\cdot)$ that relates input X and output Y, such that:

$$Y_A = \hat{f}_A(X_A) + \eta_A \tag{44}$$

where, $\eta_A \sim \mathcal{N}(0, \sigma_A^2)$ denotes a Gaussian zero-mean white noise process and represents the estimation error. $\hat{f}_A(\cdot)$ is a function that fits the relationship between X and Y. Particularly takes agent's locations as input to estimates the expected motivations of vehicles for an activity A. It can be regarded as a probability distribution from a GP process:

$$f(X) \sim GP\left(\mu(X), \sum(X)\right) \tag{45}$$

In the same experimental setup as in [17], the agent using this method successfully determined anomalies during multiple automatic avoidance maneuvers and emergency stops by self-driving vehicles.

The authors in [85] investigate the SA of the grid. PMU, a specialized monitoring sensor for the grid, records phase information within the grid. This data is analyzed at various locations in the grid, leveraging precise time information provided by GPS. The authors concluded that there is a ubiquitous uncertainty in the system and therefore trained a Random Matrix Model (RMM) based on random matrix theory using real-time collected data as used for anomaly detection in the grid. An N-dimensional random data lifting $X_{N\times T}$ with T observations is defined. $n \in N$ represents the metrics or devices in the grid system, whereas $t \in T$ represents the timestamp, i.e., the data in the last column is the closest to real-time. $X_{N\times T}$ can be normalized as:

$$\hat{x}_i = \frac{1}{\sigma_i(x_i)}(x_i - \mu_i(x_i)) \tag{46}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{iT})$.The σ_i and μ_i denotes the standard deviation and mean calculated from every column of the random matrix $X_{N\times T}$.

Based on the data analysis using the RMM, the authors employed linear eigenvalue statistics (LES) as a method of data representation to provide an understanding of the complex system through probabilistic estimation of the data rather than exact measurements. The LES τ of the matrix $X_{N\times T}$ can be defined as:

$$\tau(\varphi, X) = \sum_{i=1}^{N} \varphi(\lambda_i) \tag{47}$$

where $\lambda_i = eigenvaluesofX$. The $\varphi(\cdot)$ is a selectable test function, such as Chebyshev Polynomials:

$$T_2 : \varphi(\lambda) = 2x^2 - 1 \tag{48}$$

$$T_3 : \varphi(\lambda) = 4x^3 - 3x \tag{49}$$

This approach uses a data-driven procedure for SA and changes in real time, so it has good flexibility and robustness.

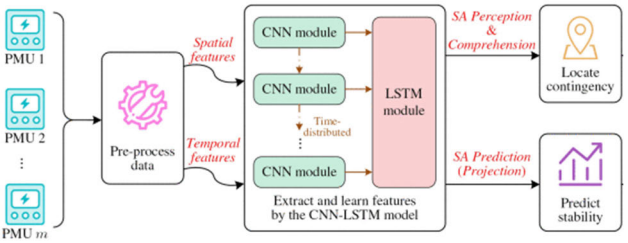

FIGURE 10. Multi-dimensional feature extraction of power grid data using AI-based methods [86].

Besides, it is model-free, which makes it low computational and has good generality.

Similarly, in recent power system related SA studies, Wang et al. [86] have used PMUs as a window for recording and analyzing grid data. In contrast, the former used LES to represent only spatial features and ignored temporal features. The authors used an AI-based approach to extract and learn temporal and spatial features from PMU data, having arrived at a more comprehensive perception, understanding and prediction. An approach combining CNN and LSTM is proposed to extract spatially relevant features and temporally relevant features in PMU data, respectively. Interestingly, multiple LSTM units are used as the tail of the network to process the continuous output of the CNN. This approach efficiently leverages the correlation of each time slice, benefiting from the processing capabilities of the LSTM units. The simultaneously obtained temporal and spatial correlation features allow the SA system to spatially determine whether an anomaly is occurring at each location where PMUs are deployed, while temporally projecting the stability of each location.

Images and videos are very common modalities in daily life, which often contain more information compared to one-dimensional sensor data and may be more intuitively understandable to humans. This makes images and videos a very typical modality for enabling SA. However, the difficulty in using images and videos as input to the system is how to use algorithms to give computers perception and understanding similar to that of humans. The authors in [87] used Reversible Jump Markov Chain Monte Carlo (RJMCMC) and Histogram of Oriented Gradients (HoG) methods to represent every single frame of low- and high-resolution videos, respectively, to detect pedestrians appearing in the video. Let G_x and G_y be the gradients in the x and y directions, respectively. The magnitude M and orientation θ of the gradient are calculated as follows:

$$M = \sqrt{G_x^2 + G_y^2} \tag{50}$$

$$\theta = \arctan\left(\frac{G_y}{G_x}\right) \tag{51}$$

The histogram within cells can be partitioned into multiple blocks by the gradient direction:

$$H_i = \sum_{pixels} M_{\theta in\ i^{th}\ block} \tag{52}$$

To achieve pedestrian tracking and crowd detection, the authors propose a hierarchical clustering method. The method employs a bottom-up strategy, i.e., each detected individual is treated as a separate cluster, which is then aggregated into more groups based on the relative distance between each cluster or group (composed of multiple clusters). For example, let k be the size of a group. G_k and e_k are the vertices of the connectivity graph and the total number of its edges. According to the rules of clustering, when a person is associated with more than half of the members of a group will be able to this person to join the group. Then:

$$\hat{e}_{k+1} = e_k + \frac{k}{2} \tag{53}$$

$$\hat{e}_{\mathrm{k}} = \begin{cases} \left(\frac{k}{2}\right)^2 & \text{if } k \text{ is even} \\ \frac{k-1}{2}\left(1 + \frac{k-1}{2}\right) & \text{if } k \text{ is odd} \end{cases} \tag{54}$$

where, $\hat{e}_{k+1}$ is the minimal number of edges in the connectivity graph G_{k+1} after a person p_i joined. Therefore, once there are two groups of people, the relationship between connectivity graphs G_p and G_q:

$$e_{p+q} \geq \hat{e}_{p+q} + \left(e_p + e_q\right) - \left(\hat{e}_p + \hat{e}_q\right) \tag{55}$$

The clustering achieves a similar subjective perception of people's concept of crowd, and experiments show that its accuracy approaches that of a human observer, which enhances the SA system's understanding of crowd videos. In order to address the problem that computer vision's scene description in a single viewpoint is of limited help to the system's understanding of the environment, the authors introduced different kinds of visual sensors deployed in different locations in [88]. Multiple UAVs and cameras deployed at the roadside are used to observe and monitor the traffic scene from multiple perspectives. The authors used a fuzzy ontology-based aggregation approach for further high-level representation of events detected by the visual sensors. Each of the sensors in the system is equipped with such data representation capabilities, but their different viewpoints may lead to differences in the obtained representations, which provide the system with a more comprehensive perception and understanding of the environment.

In recent work, neural network based modal representations have been widely used in image processing. In their study, referenced as [89], the authors utilized images from social media to compile a large dataset specifically tailored for landslide detection. They trained a CNN-based ResNet to construct an end-to-end SA system capable of real-time landslide confidence calculations on mountain-related images sourced from social networks. The trained network can perform landslide confidence calculations in real time on mountain related images obtained from social networks and represent the possible locations in the images. In addition, ViT based on Transformer architecture is also a popular image representation method recently. The authors in [90] constructed a ViT-based image representation. The authors used ResNet as image embedding. In order to cope with the complex and changing environmental disturbances and limited samples, a proposed enhanced self-attention module is used to strengthen the contour features of the target region and also reduces the collective amount of data required to train the model, enabling few-shot learning. This configuration improves the robustness of the SA system in complex environments and reduces the data acquisition requirements.

B. MULTIMODAL REPRESENTION AND FUSION FOR COMPREHENSION IN SA SYSTEM

1) REPRESENTATION IN MULTIMODAL SYSTEMS

Multimodal design brings richer information to the system, but also implies more complex structures and calculations. Earlier studies did not bring neural network-based approaches to multimodal systems. Byun et al. considered multimodal systems when building SA for indoor environments in [91]. To allow the environment to guide the lighting control, motion sensors and luminance sensors were deployed in the room for real-time monitoring. The authors designed a series of logical rules to determine whether lighting is required in the current monitoring situation. The sensor data does not need to undergo additional representation or interpretation, and the fusion of multi-sensor signals is accomplished by simple nesting of if statements. However, such a multimodal system reaches only low-level perception at the SA level.

Achieving more advanced sensing and understanding often necessitates deploying additional sensors alongside sophisticated representation and fusion methods. In [92], cameras and pressure sensors are deployed simultaneously in the room to sense and monitor the indoor environment and the patient's condition, respectively. Multiple pressure sensors are arrayed on the patient bed and each deployed pressure sensor has independent coordinates (i.e., rows and columns). The authors used the center of gravity (CoG) method for representing the pressure sensor data, and subsequently the patient's sitting posture as well as position in the bed can be determined based on the extracted features. Cameras that are deployed in high places can observe the bed as well as the neighboring area. The authors used Mixture of Gaussians (MoG) for modelling the video of invariant viewpoints, which is used to segment out the background and objects. The multimodal fusion in this system is also rule-based, but further representation of the sensor data and video allows the system to gain a more advanced perceptual capability. Based on the decision tree approach, the authors in [93] perform feature extraction and fusion of environment-related processors (e.g., smoke, brightness, temperature, etc.), vital sign sensors (e.g., blood pressure, heartbeat, body temperature, etc.) and behavioral sensors (e.g., acceleration, position, etc.) deployed indoors. Entropy calculation is introduced to evaluate the importance of each attribute and the attribute that gives the maximum gain is selected as the root of the decision tree to construct a better decision tree. To deepen the system's understanding of the data, the authors introduced the Frequent

Pattern Growth algorithm (FP-Growth) to learn the correlations in the data. Certain scenes and time periods may involve only a few sensors with the most pertinent information; for example, the presence sensor in the bathroom during morning hours holds significant perceptual value.

The authors in [94] examined the use of wearable systems to provide awareness of participants' workloads. Participants' Electroencephalogram (EEG), respiration (RSP), electrocardiogram (ECG), pulse wave through photoplethysmography (PPG), electrodermal activity (EDA) and skin temperature (SKT) were recorded by multiple sensors and made into a dataset. The signals are simultaneously recorded and filtered by a bandpass FIR filter, which defined as:

$$y[n] = \sum_{k=0}^{N-1} h[k] \cdot x[n-k] \tag{56}$$

where $x[n]$ and $y[n]$ are respectively the input and output signal at time n. $h[n]$ is the impulse response coefficients of the filter which is designed to only pass specific range of frequencies for different signals. For instance, the authors set $[0.03 - 0.5]\,Hz$ for the RSP signals and $[0.1 - 5]\,Hz$ for the PPG signals.

For data representation, the authors developed different features based on the different characteristics of each physiological signal, e.g., using the interval between two peaks in the ECG signal as a representation. Multiple SVMs make up the authors' proposed system, which, after training with the dataset, provides classification and confidence estimates for real-time input data, and also represents the perceptual awareness of the system. For three levels of SA, the represented score for perception ($score_{L1}$), comprehension ($score_{L2}$) and projection ($score_{L3}$) calculated from the cascaded SVMs are represented as:

$$score_{Ln}(k),\ n = [1,2,3] = \begin{cases} SVM_{L1}(X_{RSP}[k]) \\ SVM_{L2}(X_{ECG\&SKT}[k]) + score_{L1} \\ SVM_{L3}(X_{PPG}[k]) + score_{L1} + score_{L2} \end{cases} \tag{57}$$

where $X_{(\cdot)}[k]$ denotes the filtered and delineated signals with workload situation labels.

Different physiological data possess different importance in each perceptual class. The authors concluded that the RSP signal was the most important at the lower levels of perception, whereas the ECG, SKT and PPG features were then engaged as the participant's workload increased and the system's perceptual capabilities improved.

Additionally, eye movement data were added to physiological data in the study in [95] to enable assessment of SA levels in participants, which used the SAGAT as a criterion. The authors used metrics such as variance and root-mean-square as representations of physiological and eye movement data and used the K-NN algorithm to replace or interpolate possible outliers and missing values in the time series. The multi-sensor data representations were integrated by an ANN-based fuzzy cognitive map (FCM) to understand their concepts and the relationships between them. The classification results of the FCM then represent the perceptual capabilities of the system.

Neural network-based architectures simplify feature extraction and fusion process of multimodal data by the system to a great extent. The end-to-end characteristic makes deep learning networks easier to deploy. The application of models is usually modular. For instance, the SA system proposed in [96] utilizes sequential deep learning networks to compute the confidence level of vehicle multi-sensor signals. The author in [97] used HMM and ANN as classification models for audio and physiological signals, respectively. Additionally, In [98], the authors utilize CNN-based ResNet and Deep Spiking Neural Networks (DSNN) as feature learning methods using the built-in sound sensor, position sensor, accelerometer, and gyroscope of the cell phone as channels for indoor SA. The authors designed three ResNet networks for sound events, trained for emergency monitoring, emergency classification and normal event classification. The integrated acceleration and gyroscope signals were used as activity recognition. Two DSNN networks were used for abnormal activity detection and normal activity classification, respectively. One more DSNN network is used for pair prediction of location information. The information from multiple sensors is fused at the outputs of the three modules to provide the system with indoor situation awareness. Besides, Wang et al. [99] skillfully exploited the sensitivity of the CNN structure to spatial information by mapping the relative angular shift (RAS) and rate of change of frequency (ROCOF), two temporal signals measured from the power grid, into a single-channel grayscale image and using separately designed CNNs for feature extraction and anomaly detection. Technically, the gray image, which provide the spatial information, is generated as:

$$GD_{t(k)} = c_R \cdot R_{t(k)} + c_G \cdot G_{t(k)} + c_B \cdot B_{t(k)} \tag{58}$$

where $GD_{t(k)}$ is the gray degree at time $t \in T$. The two modalities ROCOF and RAS will be fused in this grayscale image. They are denoted as respectively:

$$ROCOF_{t(k)} = \frac{F_{t(k-\tau+1)} - F_{t(k)}}{t(k-\tau+1) - t(k)} \tag{59}$$

$$RAS^{i}_{t(k)} = VA^{i}_{t(k)} - VA^{global}_{t(k)} \tag{60}$$

where $F_{t(k)}$ is the measured frequency of the power system at time k and τ is the sampling interval. The $VA^{i}_{t(k)}$ and $VA^{global}_{t(k)}$ are respectively the voltage angle shift and global voltage angle of the i^{th} device at time k.

Social media, as a source of information besides sensors, can provide huge amounts of multimodal information and is a common scenario for multimodal SA systems involving AI-based approaches. Images and text are the most common modals in social networks. In [100], the authors designed two CNNs for feature extraction of images and text respectively.

The extracted features will represent images and text from the same tweet and be classified through the aggregation layer and the softmax layer. Similarly, the work in [101] adds an element of tweet API to the text and images. After classification and fusing the images and text using CNNs, the perceptual information is fused with a geo map using location information extracted from the API to provide a more intuitive SA using visualization. Word embedding, e.g., Word2Vec and GloVe, is a useful approach when it is necessary to get a good semantic representation for each lexical learning instead of classifying a whole passage of text. This is evidenced by the work in [39] and [41].

As datasets get larger and model parameters increase, machine learning methods excel in accuracy and can solve many problems, even rivaling humans, but their interpretability decreases. Employing rule-based algorithms offers a potential solution to mitigate the interpretability challenges encountered with machine learning methods. In [102], the authors provide an ontology-based contextual model for homecare-oriented SA systems. The authors construct separate ontologies for the patient, the environment, warnings, and social relationships, and diverse physiological sensors, environmental sensors, and textual information are represented and fused within or across ontologies. The work in [103] similarly utilizes an ontology-based model: multiple UAVs equipped with cameras deployed with tracking algorithms act as agents for sensing the environment, and the data they generate is represented in coded form and integrated into ontological statements. The authors introduce semantic Web technologies to provide sensible relevance suggestions while tagging and tracking the data. Besides, Belief-Desire-Intention (BDI) reasoning is also one of the ways to improve interpretability. The rules formulated based on the BDI mechanism use the dynamic information received by the agent as a way to update beliefs or learn prior knowledge. The agent executes predefined actions from the intention set upon receiving cases that align with those contained in the belief set. The authors in [104] have used location sensors, physiological sensors, etc. as the SA perception layer, the rules in the belief part of the BDI as a holistic understanding for a single agent, and the intent part as the decision and action layer to provide feedback and update to the environment, enabling monitoring and alerting of early symptoms of Covid-19. For instance, if an agent A_i detected from the sources that fever symptoms are happening, the belief mechanism B_i of BDI would determine whether it is in the knowledge base KB_i, the detection method can be represented as:

$$Awareness = \begin{cases} True, & B_i \sim \phi \in KB_i \\ False, & B_i \sim \phi \notin KB_i \\ True, & B_j \sim \phi \in KB_i \end{cases} \tag{61}$$

where B_j is the belief operation from another agent A_j, ϕ is the fused representation of multimodal signals.

Additionally, The Bayesian-based approach always describes the relationship between variables in terms of conditional probabilities, which may confer it better interpretability compared to neural network architectures. Rehder et al. [105] designed a Bayesian network for recognizing drivers' intention to change lanes. Sensors such as cameras and Lidar on the vehicle are integrated and the data is represented as a relationship between the vehicle and the lane, e.g., speed, orientation, distance to the vehicle in front and behind, and distance to other lanes. This information plays a role in the individual nodes in the Bayesian network, providing probability-based references. In comparison with a deep learning network with 20 hidden layers, the Bayesian network-based approach has comparable performance and better interpretability.

2) FUSION IN MULTIMODAL SYSTEMS

Multimodal fusion is necessary when the system is designed to have inputs in multiple modalities. Effective multimodal fusion enables SA systems to obtain more accurate and comprehensive information in the environment, whereas scattered aggregation may reduce the performance of detection and prediction. There are generally two types of methods for multimodal fusing, model-agnostic approaches and model-based approaches, respectively. The following discussion describes several works that have used typical multimodal fusion methods.

For model-agnostic approaches, data from multiple modalities can be spliced and fused into a single feature vector in the early stages of the process and then input to the model for subsequent tasks. This early fusion generates feature vectors that always contain redundant information, so they are usually combined with feature extraction methods to streamline the information. Hegde et al. [107] proposed that the multimodal information received by each agent in a multimodal SA system will be filtered and fused into a contextual profile in advance. multiple profiles from different agents will be subsequently integrated and exchanged information by the server. Assuming there are N devices given by the set D, M kinds of modalities, including GPS, environment audio, user speech, handwriting text and typing text, given by the set C. The contextual profile P is concatenated by D and C.

$$D = \{D_1, D_2, \ldots, D_N\}$$
$$C = \{C_1, C_2, \ldots, C_M\}$$
$$P_i = \left\{P_{i,j} : i \in N, j \in M\right\} \tag{62}$$
$$P_i = \left\{P_{i,GPS}, P_{i,EA}, P_{i,US}, P_{i,HT}, P_{i,TXT}\right\} \tag{63}$$

In [106], the authors superimposed the raw features of the biometrics from the participants' faces and ears on top of each other, referred to as stochastic biometric fusion. Further features were subsequently extracted from the fused vectors by Fisher's linear discriminant analysis and discriminability was increased using linear discriminant analysis (LDA). Finally, the inputs are fed into a K-NN based classifier to provide risk perception for the biometric based SA system. Similar to [99], the authors in [108] fuses the information from multiple modalities into a 2D signal in advance and

TABLE 2. Typical multimodal fusion methods.

Fusion	Year	study	Methods	Contributions
Early	2009	[106]	LDA, K-NN	Randomized multimodal data folding and distance-based linear discrimination
	2013	[107]	Contextual profiles	Multimodal contextual information exchange between mobile devices in a SA system
	2022	[108]	Linear projection	Projection of multiple measurements into the same two-dimensional space
	2023	[109]	Decision tree	Evaluate the importance of each modal feature in differentiating SA levels to eliminate the negative impact of redundant or irrelevant features on model performance
Late	2017	[110]	Ontological model	SA Reasoning at the semantic level by building ontological models
	2020	[111]	Graph-based clustering	Combining location and textual information to construct semantic graphs to learn location-event correlations.
	2020	[103]	Ontological model	SA Reasoning via Ontological Design Combining Location Information and Target Recognition
	2020	[112]	Attention mechanism	Selection of valid modal representations based on confidence and masking of potentially misleading representations.
	2021	[113]	Rule-based algorithms	Computing and sensing events based on multi-sensor signals and using rules to fuse and make decisions
	2022	[114]	Visualization	User-oriented interaction system for multi-sensor data and SA event visualization
	2022	[115]	Ensemble learning	Stack the probabilities of model outputs for multiple modalities and weight the outputs for different modalities.
	2023	[116]	Attention mechanism	Decision-level fusion of long-sequence multimodal information
Model-based	2016	[117]	AE	Learning intermodal correlations and the same representation of multimodalities through autoencoder (AE) structures.
	2020	[118]	VAE	Learning and aligning the representations of modalities generated by separate autoencoders at the distributional and semantic levels to obtain stronger discriminability.
	2021	[119]	ViLBERT	Learning intermodal correlations through hard parameter sharing
	2022	[120]	VAE	Learning normal and abnormal patterns in multivariate time series based on autoencoder.
	2022	[121]	RNN	Fusion of emotional intensity and temporal dimensions for long-series multimodal data
	2023	[122]	Attention mechanism	Psychologically based multilevel intra-modal self-attention and inter-modal cross-attention for understanding irony
	2023	[123], [124]	CNN	Multiscale cross-modal fusion of RGB-T to fully extract high-level semantic features

extracts features directly from the 2D view as the output of a decision tree-based classifier model. The difference is that the authors propose a more lightweight approach compared to neural network architectures, the 2D Orthogonal Locally Persistent Projection (2D-OLPP).

Although multimodal data provide data describing the environment from different perspectives, redundant data may contain many features of low relevance, which may reduce the ability of the model to fit and predict the environment. The authors in [109] introduce decision tree to identify the features that significantly differentiate between high and low SA levels in the process of SA assessment of drivers of self-driving cars using eye movement signals and EEG. The features with highest importance are concatenated and classified using ANN. The performance of EEG only, eye movement signals only, and the combination of the two was compared with accuracy of 0.76, 0.77, and 0.81, respectively. This indicates the advantages of multimodal fusion.

Late fusion, where multiple models are applied to the corresponding modalities separately and fused at the output layer. Models that are independent of each other can avoid misinterpretation of the data by the models, but they may also lose relevant information between modalities and have higher computational complexity. The use of different models for different modes provides a certain degree of convenience in design, and thus the late fusion method is widely used in work related to multimodal SA systems. Similar to the previously mentioned [100], [101], different modalities are designed separately for feature extraction and classification models, which are then fused using multiple classification results. When using deep features from multiple modalities for a downstream task, the attention mechanism can learn relationships between modalities to learn more representative features. In [112], the authors use cross-attention for visual and textual deep features to influence the final confidence level based on the effectiveness of different modalities at different disaster times. The modality with more confident representations can reduce the impact of the modality with misleading representations through cross-attentive linking. Experiments show that this approach has better performance as well as higher robustness compared to directly splicing the features of multiple modalities. Similarly, the authors in [116] investigated the use of cross-attention to learn an efficient representation of long sequences of multimodal information.

The depth representation of three temporal modalities (EEG signals, eye movement signals, and vehicle sensor signals) is first learned using LSTM, and then a decision-level fusion of the feature representations for each modality is performed using the cross-attention mechanism. Employing the graph-based clustering, [111] also completes the fusion of location information and text classification. In [115], two independent XGBoost are used to classify EEG and eye movement data, and the output probabilities of the two models are combined into an ensemble and fused end-to-end using the proposed Random Vector Function Link Stacking (RVFL-S). Viewing the fusion of results from multiple models as a process of logical reasoning suggests a systematic approach to combining information for decision-making. Corresponding rules have been developed in [113] for different SA objectives. Once multi-sensor information has been extracted, classified, and processed, the corresponding rules guide the system in interpreting the classification results and formulating predictions and plans accordingly. Similarly to [32], the authors in [114] adopted the digital twin approach and used visualization to fuse the multi-sensor data. In addition, ontology-based approaches can provide a flexible and interpretable logic for the later fusion of systems with a larger number of modalities (e.g., [102], [103], [110]).

In model-based approaches, how the modalities are fused depends on the structural design of the model. Quan Guo et al. [117] adopted an unsupervised neural network-based method, proposed to process incomplete multimodalities from social media in two steps. In [122], the authors propose a multi-interaction and multi-level neural network by choosing four modalities: text, image, text in image, and image illustration as signals for perceiving tweets with sarcastic sentiments in social media. By mimicking the brain's process of perceiving sarcasm, the proposed network has four components: extraction, interaction, integration, and cognition. Images and texts are acquired modal representations using CNN-based and BiLSTM-based networks, respectively. Based on the gate and attention mechanisms, the four modal representations are subsequently used to compute interactions ranging from unidirectional to quadratic for multimodal integration. The integrated feature vectors are activated, and classification is done by a linear layer. In addition, the authors in [121] employed RNN in order to introduce temporal dependency for human emotion recognition. Visual, audio and text are represented using GRUs and then fused using the attention mechanism, and feature representations from different time points are further fused using RNN.

Based on Transformer and Attention mechanism, ViLBERT [125] processes image and text using two encoder sets, then the features of different modalities can interact with each other in parallel space. Applied in [119] to enhance the system's understanding of the correlation between text and images in social media. Similarly, Vison Transformer (ViT) is used as the image encoder in ALBEF [126], and the first six layers of the BERT model are used as text encoders, while the last six layers are used as multimodal encoders for multimodal fusion.

$$MMEncoder_{ALBEF} = ViT\left(X_{img}\right) \otimes BERT\left(X_{txt}\right) \quad (64)$$

where X_{img} and X_{txt} denotes the image and txt labeled multimodal data.

When the data to be processed is unstructured (e.g., images, text, and files in various formats), the absence of modality is common. In [118], the authors use separate modal representation methods for image and text modalities, and then use Variational Autoencoders (VAE) separately in order to learn the common features between the different modalities. In [120], multivariate time series are used as inputs to the system, and VAE is used to learn normal and abnormal patterns in the data as well as the differences between them to output anomaly monitoring results, enabling an end-to-end SA system.

In vision-based SA systems, in addition to the common RGB images, thermal cameras monitor the infrared radiation emitted by an object to provide thermal imaging images. Zhou et al. [124] proposed a Multi-Task Awareness Network (MTANet) with hierarchical multimodal fusion for urban scene understanding based on RGB-T signals. The network contains two encoders using ResNet152 as a backbone for extracting feature maps from RGB images and thermal imaging images, respectively. The feature maps extracted by the two encoders are selected and fused with complementary information from the RGB and thermal features through spatial and channel attention modules, and residual concatenation is performed when the decoder reconstructs the image to ensure that the features at different scales have been correctly reconstructed. The network outperforms other baseline models by obtaining an average IoU of 56.1% and 78.6% in the RGB-T semantic segmentation datasets MFNet [127] and PST900 [128], respectively. In a subsequent work, dynamic bilateral cross-fusion network (DBCNet) [123] utilized the cross-fusion of features from multi-scale RGB and thermal images to further improve the model's understanding of the urban environment.

C. PHASE SUMMARY OF MULTIMODAL SYSTEMS IN SITUATION AWARENESS

The survey work in this section reveals the following key points and insights for multimodal systems in SA:

- The representation methods of modalities in SA systems is closely related to factors such as the information carried by the modalities themselves and their format. From conventional statistical features to representation learning based on AI techniques, the obtained representations are becoming increasingly complex but converging to semantics. This is helpful in providing more advanced SAs, but at the same time may create problems of interpretability.
- Multi-modal or multi-sensor SA systems do increase the perceived range of the environment, but they do not imply an increase in the SA level, as in the earlier

work [91]. It is worth mentioning that modal representations based on AI technology may help in the comprehension session of SA, but more complex logic is also bound to cause an increase in computational load.

- Most of the related work on multimodal SA systems has employed late fusion, i.e., the use of multiple models for different modalities, followed by fusion of the outputs of the individual models at the decision level. This fusion approach can be modular and may help in the expansion of the system. Moving into the AI era, modality-based approaches allow for better learning of correlations between modalities and also blur the boundaries between modal representation and fusion, allowing them to be implemented end-to-end in a single model. However, it has only been applied in a few works.

IV. SITUATION AWARENESS WITH 3D TECHNOLOGIES

Perception of space is a common task in situation awareness, which often determines the ability of complex systems to perform their tasks efficiently and safely. Simple spatial perception is low-dimensional, e.g., using human presence sensors to determine whether a space is occupied. Combined with a 3D model of the environment, it can provide users with services such as localization and navigation. The integration of visual sensors, such as cameras, has enhanced spatial awareness in situation awareness systems, making it more akin to human visual perception. Some work has used thermal infrared sensors, LIDAR, depth sensors, etc. in some work individually to provide sensing of different usage scenarios or combining them to provide more comprehensive information, which may involve multimodal fusion. Situation awareness systems usually need to translate and understand the raw data from sensors with the help of algorithms such as detection, segmentation, tracking, scanning, etc., and AI-based approaches have advantages. This section explores how a situation awareness system can utilize the raw data provided by various sensors in order to achieve spatial awareness from the perspective of 3D technology. Table 4 shows a summary of 3D technologies for situation awareness.

A. 3D LOCATING AND NAVIGATION

3D modeling of the environment is the basis for 3D localization and navigation, while the focus is on how to transform sensor signals into semantic information with real-world meaning and link them to the 3D model of the environment. The authors in [129] designed a model for indoor and outdoor urban environments that combines 3D indoor and outdoor GIS signals with a 3D model to enable localization in both areas and seamless switching of navigation between the two areas. In the model, objects such as elevators, staircases, doors and windows inside the building as well as water sources and roads outside the building are endowed with semantic information. The authors define more than ten indoor emergency events and use smoke sensors, heat sensors, etc. for their detection. The 3D indoor-outdoor modeling makes the environment intuitively perceivable, and for the various semantic objects allows for timely and accurate representations of the detected events, providing critical information to the people inside.

The easy to deploy and distributed nature of the Internet of Things (IoT) facilitates 3D localization and navigation. The authors in [130] built a wireless sensor network (WSN) to provide users with situation awareness and emergency localization and navigation. The authors model a dense and discrete sensor network as an undirected graph, where each vertex in the undirected graph represents a sensor node and the raw data from each sensor is converted into a semantic hazard index based on its relevance. Nodes or clusters of nodes with a high hazard index are designated as hazard fields by the authors, signaling potential danger zones within the network, which becomes the condition of the route planning of people in the undirected graph to a defined evacuation exit. In a multi-story building, the stacks of WSNs deployed for each floor form a three-dimensional network. The authors conducted experiments using a three-story building where the three levels of the WSNs were connected at the location of the staircase and three exits with different evacuation capacities were set up. When there are sensors in the WSNs that detect danger, the escape routes can avoid the danger field and tend to leave through the exit with high evacuation capacity.

Route planning may need to consider different goals of situation awareness and corresponding environments. For indoor evacuation in case of fire, the authors in [131] used a cellular automata model, a typical microscopic evacuation model, to model the environment. In addition, the authors utilized flame and fire dynamics models (FDS) for predicting the fire's course and extent. Multi-sensor data streams constructed from temperature sensors, smoke sensors, gas-phase sensors, and cameras were converted into fire indices and crowdedness indices and assigned values to corresponding locations in the cellular automata-based 3D model. Cells containing hazard information were defined as obstacle cells, implying escape routes to be avoided. In simulation experiments, using the proposed method possesses shorter escape time and probability compared to the baseline method with the same number of escapees.

The spatial perception provided by 3D maps for disaster and accident scenes helps a lot, but what the maps show can be inaccurate. With the help of UAVs, the authors in [132] fused images from a bird's-eye view with 3D maps to achieve a more accurate and time-sensitive situation awareness. Position sensors on board the UAVs can add position, orientation, and attitude information to the acquired images, which are references for the alignment of the images with the 3D maps.

B. 3D OBJECT DETECTION

In SA systems based on visual perception, 3D object detection is one of the most fundamental tasks and plays a decisive role. Based on multimodal fusion, RGB images together with

TABLE 3. 3D technologies for situation awareness.

Distribution	Year	study	Modality
3D location and navigation	2015	[129]	3D model, GSI data
	2017	[130]	3D model, WSN data
	2020	[131]	3D model, multi-sensor data
	2021	[132]	3D map, image, GSI data
3D object detection	2017	[133]	image (RGB) and LIDAR (BEV + FV)
	2017	[134]	image (RGB) and LIDAR (BEV)
	2018	[135]	image (RGB) and LIDAR (BEV)
	2018	[136]	image with depth information (RGB-D)
	2019	[137]	image with depth information (RGB-D)
	2021	[138]	image (RGB) and LIDAR (BEV)
3D semantic segmentation	2017	[139]	LIDAR (point cloud)
	2018	[140], [141], [142]	LIDAR (point cloud)
	2020	[143]	LIDAR (point cloud)

depth information or LIDAR data can construct a 3D model for the environment.

Typically, in [133], Chen et al. proposed a 3D object detection network for autonomous driving scenarios. The bird's eye view (BEV) and front viewpoint (FV) cloud images generated by LIDAR and the RGB images generated by the camera are used as data sources. The bird's eye view is represented by height, intensity, and density while the front view provides supplementary information. Based on the Regional Proposal Network (RPN), the authors proposed to use a deep fusion approach, which combines early and late fusion, to hierarchically fuse multi-view features, which improved the performance by 0.5% to 1%. The proposed method achieved average precision (AP) of 71.29% to 87.65% at 3D Intersection over Union (IoU) of 0.75 to 0.25.

Similarly, the work in [135] applies RPN as a backbone network for 3D object detection, but the authors discard the LIDAR front view, which carries less information, and use only BEV and RGB images as inputs to the network. The use of a 3D Anchor grid is proposed for multi-view alignment and cropping for multimodal fusion. The proposed method achieves up to 88.53%, 58.75% and 68.06% AP for 3D target recognition for cars, pedestrians and bicycles, respectively. Recently, the authors in [138] introduced an Attention Generation Module (AGM) to improve the detection performance of occluded objects. The authors routinely used a network based on CNN architecture for feature extraction of LIDAR BEV and RGB images, and modal alignment using 3D anchors. The AGM is introduced in the subsequent backpropagation operation of the network that generates 3D ROIs to enhance the network's attention to the objects. The proposed network improves 1.3% in AP compared to [135]. Besides, The authors in [134] use HOG descriptors and LBP descriptors to feature representations of RGB images and LIDAR point cloud data and connect them for multimodal early fusion. The authors trained a pedestrian detector based on a multi-view approach and used Random Forest (RF) to classify the detection window. The proposed method achieves an AP of up to 82% for 3D target detection of cars.

Depth image (RGB-D) sensors are a well-established 3D sensor that produces RGB images with depth maps. The authors in [136] propose a 3D object detection method for RGB-D images. For RGB images, the authors use a well-established 2D CNN object detector to propose and categorize the content in the picture. The authors propose to project a 2D bounding box at an angle according to the camera viewpoint onto a 3D depth map as the search space for objects. For depth maps, similar to LIDAR-generated FVs, points belonging to the same object are always relatively close to each other and are therefore more easily segmented within the specified search space. The proposed method achieves up to 83.76% AP for 3D inspection of cars.

Beyond detection, pose estimation of detected objects is a further understanding of the 3D environment. Based on RGB-D images, the authors in [137] define a Normalized Object Coordinate Space (NOCS) to include all the detected 3D objects in it. Utilizing the Mask R-CNN architecture, the authors employ separate head structures for the x, y, and z axes to process proposed regions. The depth map is subsequently aligned to the RGB image to output the proposed regions in 3D. The advantage of this approach is that the CNN-based network does not need to involve 3D data, but rather the model can be trained using regular datasets to achieve more detection objects and higher detection performance.

C. 3D SEMANTIC SEGMENTATION

3D semantic segmentation provides a higher level of understanding for situation awareness systems. From object detection to semantic segmentation, the system's ability to

understand visual data is refined from the region level to the pixel level. Charles et al. [139] proposed PointNet for classification, segmentation, and semantic interpretation of 3D objects, pioneering the use of deep learning architectures to directly process point cloud data. Technically, PointNet is divided into two networks for classification and segmentation, extracting respectively a global feature and point-wise feature, respectively, on the point cloud data to maintain the representation performance across different sensory fields. As we may have a point cloud data input represent as

$$\{P_1, P_2, \ldots, P_n\} \tag{65}$$

where P_i denotes one point in 3D space, and

$$Pi = (x_{Pi}, y_{Pi}, z_{Pi}, F_{Pi}) \tag{66}$$

where x, y, z are the coordinates and F_{Pi} could be the optional features such as color. A transformation matrix T is learned by a shadow MLP to represent the point cloud data:

$$T = MLP_{Trans}(x_1, x_2, \ldots, x_n) \tag{67}$$

The point-wise feature f_i and global feature f_G extraction can be described as below:

$$f_i = MLP_{Feature}(T \cdot x_i) \tag{68}$$

$$f_G = Pooling(f_1, f_2, \ldots, f_n) \tag{69}$$

After training with a 3D object dataset with 40 classes, it achieved a mIoU metric of 83.7%. Additionally, DOPS [143] of the same directly applies the CNN-based network directly to the point cloud data to ensure the integrity of the spatial information.

The authors in [140] propose the SqueezeSeg algorithm based on SqueezeNet, which uses spherical projection to convert point cloud data from data structures unsuitable for 2D CNNs into a front view. Based on a classical encoder-decoder architecture, SqueezeSeg learns the shapes of various objects in the 3D dataset as a priori knowledge, making it possible to predict the 3D shapes of the detected objects, even though they may be incomplete. Thanks to SqueezeNet's streamlined network architecture, SqueezeSeg has an impressive inference speed. Building on the descripted foundations, the subsequent SqueezeSeg V2 [141] and SqueezeSeg V3 [142] maintain excellent performance.

D. PHASE SUMMARY OF 3D TECHNOLOGY IN SITUATION AWARENESS

The survey work in this section reveals the following key points and insights for 3D technology in SA:

- Situation awareness heavily relies on spatial perception, which is evolving with the integration of diverse sensor technologies. From traditional human presence sensors to sophisticated visual sensors like cameras, the shift towards 3D models enhances spatial understanding and facilitates tasks such as localization and navigation.
- Combining various sensor modalities such as thermal infrared sensors, LIDAR, and depth sensors allows for comprehensive spatial awareness. Fusion of multi-view, multi-sensor and multi-modal enable the translation of raw sensor data into meaningful representations, enhancing the system's ability to detect, segment, track, and scan the environment.
- Existing methods have gained promising results in 3D target detection and semantic segmentation, but they are still limited in their ability to interpret the contextual relationships of the detected objects. This may need to be addressed by introducing graph-based approaches [130].

V. SITUATION AWARENESS WITH INTERPRETABLE TECHONOLOGIES

Although technologies such as AI and multimodality have greatly improved the ability of SA systems to perceive, understand, predict and even make decisions, complex models and changing real-world environments affect their reliability and robustness [144]. In particular, the introduction of deep learning has made the complex, multi-layered model a 'black box'. This can lead to a lack of understanding and trust in SA systems by users.

Commonly used interpretability methods in building SA systems are: Fuzzy classifier, Local Interpretable Model-agnostic Explanations (LIME) [145], SHapely Additive exPlanations (SHAP) [146], Class Activation Mapping (CAM) [147], etc.

The concept of fuzzy classifier was introduced in [84], [91], and [92] to increase the interpretability of SA. Derived from fuzzy logic, its variables can be any real number between zero and one, which mimics human reasoning patterns and distinguishes it from clear logic (Boolean logic) binarization. Suppose there is a set of input features $X = \{x_1, x_2, \ldots, x_i\}$, then there is a corresponding set of fuzzy sets $A = \{A_1, A_2, \ldots, A_i\}$. The affiliation function $\mu_A(x)$ is used to represent the degree of affiliation of x_i to A_i. A set of manually formulated if-then rules reason the input fuzzy set A_i to the output fuzzy set $\hat{A}_i$ and output clear and interpretable predictive values.

LIME is a local surrogate model, i.e., it focuses on interpreting individual predictions. A new dataset is obtained as a set of input samples x_i' that are randomly perturbed according to a specified distribution, and the corresponding predictions generated by the original (black box) model $f(\cdot)$. LIME trains an interpretable model $g(\cdot)$ (e.g., a decision tree-based model) on this new dataset. The weight w_i of a perturbed sample is determined by its similarity to the original sample, and the closer it is to the original sample, the higher weight it gets. Then an interpretability model can be represented as:

$$g(x_i) = argmin_g \sum_{i=1}^{N} w_i \cdot Loss\left(f(x_i), g\left(x_i'\right)\right) + \Omega(g) \tag{70}$$

where $f(x_i)$ and $g(x_i')$ denotes the prediction of the original model and the new interpretable model for the i^{th} original sample and perturbed sample, respectively. $\Omega(g)$ is a regularization term.

Besides, SHAP is used to calculate the significance value of each feature to improve the interpretability of the predicted values, independent of the model. This approach was introduced in [21] and [48]. Specifically, the Shapley value ϕ_i for feature i can be computed as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f\left(S \cup \{i\}\right) - f\left(S\right)\right] \tag{71}$$

where N is the set of all features and S is a subset of features except the i^{th} feature. The contribution of each feature to the model's prediction results is indicated by the calculated Shapley value. In combination with LIME, Kernel SHAP can also train a linear local surrogate model that provides explanations for localized (single-sampled) prediction results.

CAM provides CNN models with interpretable visualizations. Zhou et al. [147] retrained the model after adding global average pooling (GAP) in front of the final classifier (linear layer) of the CNN one to obtain CAM. the obtained CAM can be represented as:

$$L_{CAM}^{c} = \sum_{i=1}^{n} w_i^c \cdot M_i \tag{72}$$

where M_i is the i^{th} feature map generated by the last convolution layer. w_i^c is the weight of the cth neuron of the linear layer in the i^{th} feature map, which corresponds to the c^{th} classification category. Subsequently, Grad-CAM [148] combines the gradient information to compute importance weights for the feature map, which allows the model to extract any layer of CAM without including the GAP layer and without retraining.

A. INTERPRETABLE FORECASTING AND DECISION-MAKING

The authors in [149] introduce a knowledge graph to increase the interpretability of AI models for demand forecasting. Entities and concepts from data sources are identified and a knowledge graph is constructed. Predictions from trained AI models are queried in the constructed knowledge graph to determine their reliability, which relies on a rule-based process. In [150], the authors investigate the interpretable classification and prediction of bearing faults. The input signal is the acceleration signal of a bearing, which is converted into a time-frequency map using Short-Time Fourier Transform (STFT) and feature maps are extracted using CNN. Grad-CAM is used to compute the attention of the model to show the frequency bands of signals that are of most interest for different fault types. The authors used decision trees and fuzzy classifiers to validate and interpret the predictions of the CNN, respectively.

Based on reinforcement learning, the authors in [151] introduce deep Q-learning to enable epidemic diagnosis. Based on the observable Markov decision process, the prediction process of the model becomes relatively transparent. The authors point out that utilizing deep Q learning to aid diagnosis can effectively improve the accuracy and timeliness of epidemic diagnosis.

B. INTERPRETABLE VISUAL-BASED SYSTEM

Segmentation of tasks to provide better interpretability. In [152], the authors investigate the use of AI models for interpretable SA assessment of drivers in an autonomous driving environment. A DNN is first used to predict the driver's gaze angle, after which a Gaussian mixture model is used to map its features to the same space to predict the driver's region of attention.

In the field of medical diagnosis, Allahabadi et al. [153] utilized ResNet and U-Net as backbone networks for image feature extraction medical image segmentation respectively to assist doctors in diagnosing whether a patient is infected with Covid-19. Based on LIME, the authors delineated the image into hyper pixels and evaluated the significance of each hyper pixel by inputting only a portion of the hyper pixels. In addition, a physician-oriented human-computer interface visualizes the segmentation and diagnosis results. The authors in [154] use Grad-CAM to explain the results of a deep learning model for medical image classification. The authors point out that their heatmaps generated by the trained model effectively point out regions of interest for diagnosis, and their visualization results are consistent with the expert's diagnosis.

Besides, Zhang et al. [155] proposed an interpretable CNN for object classification task. The authors designed a loss function that passes the loss gradients of different classes to the corresponding filters, i.e., different filters will be responsible for different object classes. The interpretable convolutional layer is used to replace the last conventional convolutional layer of the CNN. The accuracy of image classification is improved while providing interpretability.

C. INTERPRETABLE MULTIMODAL SYSTEM

As mentioned in previous sections, the introduction of multimodal data is very effective for SA. However, complex modal representations and modal fusion methods make the multimodal system difficult to interpret. In addition to interpreting unimodal modalities in multimodal systems, such as the use of Grad-CAM in vision to achieve attention-based interpretation, interpreting the complementarities between different modalities can help to understand the mechanisms of multimodal enhancement of system performance [156].

The authors in [157] provides an attention-based interpretation of the image along with an attribute-based interpretation of the image using textual modalities, i.e., adding an attribute prediction network output describing the attributes of the image. The most salient attributes of each image can be obtained using backpropagation and they are interpreted in the form of text for the image.

TABLE 4. Summary of applications using SA based on AI and multimodal technologies.

Focus Area	Year	study
Military Defence	2015	[161]
	2017	[162]
	2020	[163]
	2023	[164]
Disaster and Risk Management	2015	[165]
	2018	[166]
	2019	[167]
	2019	[168]
Medical and Healthcare	2014	[169]
	2020	[170]
	2021	[171]
Smart Home	2016	[172]
	2017	[173]
	2019	[174]
Vehicle Assistance	2015	[175]
	2020	[176]
	2022	[177]
Power and Energy	2016	[85]
	2020	[178]
	2023	[179]
Industry and Equipment	2024	[180]
	2020	[181]
	2023	[182]

In addition, a graph-based approach, [158] proposes to construct images into scene graphs using annotations of the images. This facilitates vision-based SA systems to understand the environment from a semantic perspective. In [159], the authors investigate the perception of human movements. Each frame in the video is utilized to generate a corresponding scene graph, which is merged into a spatio-temporal graph in the order of frames, aiming to explore the process of model inference while monitoring human actions.

VI. APPLICATIONS USING SITUATION AWARENESS BASED ON AI AND MULTIMODAL TECHOLOGIES

This section will discuss five applications of situation awareness that use both AI and multimodal techniques: (1) Military Defense; (2) Disaster and Risk Management; (3) Medical and Healthcare; (4) Smart City and Life; (5) Vehicle Assistance; (6) Power and Energy; (7) Industry and Equipment. Table 5 shows summary of applications using situation awareness based on AI and multimodal technologies.

A. MILITARY DEFENSE

The concept of situation awareness began in the military and is now a necessity to provide and enhance military defense. The essence of military defense is the monitoring, identification and tracking of enemies or unknown targets and the perception of the battlefield environment. Earlier works of SA for military defense used vision-based methods such as visible light images and infrared images [160]. However, the dynamic and unpredictable nature of the battlefield environment necessitates comprehensive situational awareness capabilities for effective military defense, so environmental factors such as GPS, radar and radio signals are also attempted to be considered [65].

Using perception based on infrared and visible images, Gundogdu et al. [161] introduced a visual tracking method to provide military defense SA. The authors pointed out that visible images are easier to achieve segmentation of the target contours as they are more textured and have less edge noise compared to infrared images. Their aim is to fully utilize images from both modalities for environmentally adaptive target tracking. Using a MOSSE-based filter, a template-based tracking method, the authors propose an algorithm that switches between image modality and monitoring template when monitoring the tracker's inability to adapt to the target's appearance. Experiments yielded that the algorithm maintains performance when switching from visible to infrared sequences and obtained AUC indices of 0.327 and 0.36 for visible and infrared sequences, respectively.

Liu and Liu [162] considering the limited computational and memory resources of military embedded platforms, an early fusion method is proposed for visible images, infrared images, and motion images, where the motion images are derived from the difference between visible images of different frames. The authors fused the three images to become a color image through RGB channel, and then output the target frame through CNN extracted features and ROI classification and regression. The proposed multimodal target detector achieves top1 accuracy and average precision both higher than 98% while maintaining computational efficiency and low computational resource requirements. The authors improved the methodology for target region proposals in their subsequent work [163] by using a more advanced fully convolutional network based RPN, which achieved a better average accuracy of higher than 99.8% while using less time.

Lee et al. [164] designed a deep learning-based decision-making system for military defense that incorporates a GNN-based robust tactical map fusion technique (RTMF) and a spatio-temporal multilayer model (STBR) for ontology-based battlefield recognition. The RTMF first performs target monitoring using a single-stage target detector with image data from different agents (e.g., soldiers, vehicles, and UAVs), and then projects the multi-agent target monitoring results and textual descriptions onto the hypergraph plane using a GNN to project the multi-agent target monitoring results and textual descriptions onto the hypergraph plane for multimodal fusion. The STBR provides well-established ontological rules and uses ANN to provide combat decisions and recommendations.

B. DISASTER AND RISK MANAGEMENT

Monitoring and forecasting of events and their locations is a major part of disaster and risk management [165]. Information collection is a key factor, and multimodal techniques enhance the range and reliability of predictions. Commonly, multimodal data acquisition is realized through multi-sensors or multi-agents, which are deployed in the environment for sensing various aspects of the environment. Nowadays, the growth of the Internet has made data access easier. On social media, users may be able to become a multimodal agent for disaster and risk management, which enhanced the flexibility of the SA system.

Using multi-sensor IoT-based sensing, Gu et al. [167] designed a multimodal system for open pit mine disaster SA. Multimodal information from various sources, including sensors and cameras, is integrated at the data level using an SVM-based model and outputs the site safety level represented by the data. The fused data will be used as input to the ELM model for the prediction of mining area disasters.

Wang and Gerber [165] considered combining textual information from tweets with structured mobile network data to enable prediction of future crime locations. The authors categorized the system into three models: (1) a SVM-based text-rich classification model for extracting location-related features from tweet information and text information and predicting the user's nearest next location type; (2) a text retrieval model based on a vector space model, which instances a document for each location and utilizes nearby tweets as search keywords to obtain the most probable destinations; and (3) a text-retrieval model based on a vector space model, which uses nearby tweets as search keywords for the most probable destinations; and (4) a text retrieval model based on the vector space model. (3) a text enrichment regression model for determining the distance of features captured from text content to each site type and feature. The proposed algorithm is able to distinguish crime locations containing 25 crime types with 35% accuracy. Vomfell et al. [166] proposed the use of tweets and cab data to predict the number of crimes. The correlation between the number of crimes and location information is learned using models based on machine learning architectures such as RF, GBM and ANN on textual information from tweets and cab traffic data.

Based on the textual information, Rizk et al. [168] introduced the image information that tweets also contain in order to classify social media species disaster-related tweets. The method consists of two components corresponding to the first level (perception) and second level (comprehension) of SA. The model at the first level uses metrics such as RGB histogram, HSI/V histogram, and gradient direction histogram to represent the image information, and the BoW method is used to describe the semantic information. Separately trained classifiers classify the two modalities. SVM based classifiers are applied with the second quarter of the model to fuse and classify the data representations and features generated by the first level model at the decision level. The method obtained 92.24% accuracy in the test.

C. MEDICAL AND HEALTHCARE

Compared to SAs in other domains, SAs in medical and healthcare usually focus more on sensing the user's physical conditions. With the addition of multimodality, devices such as portable and wearable physiological signal sensors, sensor-integrated smart phone and person-facing cameras, etc. can be deployed to provide sensing.

Haghighi et al. [169] designed a health monitoring system based on wireless sensors and wearable sensors. The acceleration sensor data from the mobile phone, consisting of x-, y-, and z-axis values, is preprocessed into a vector format and inputted into a K-NN classifier. The action-related classifications output by the classifier will be fused with low-level sensory data such as heart rate, blood pressure and body temperature. Based on fuzzy inference, the authors have created a rule base and corresponding inference rules for multimodal fusion at the decision level. Weights are assigned to each semantic variable (e.g., heart rate) based on its contextual significance. The weighted average of all the variables will be used as the confidence level for the system to judge whether the participant is healthy or not. The proposed system considers being run on a removable device and adopts an intelligent adaptation strategy that reduces the sampling accuracy of the sensors to save power when the participant's health monitoring results are positive, which increases the lifetime of application by over 38%.

Henaien et al. [170] proposed a rule-based medical monitoring system. The authors divided the system into several layers, most notably (1) an active and assisted living sensor layer that includes all wearable sensors and nearby sensors that can represent information about the patient's health status, location, movement, etc.; (2) an ontological model data layer that includes ontological rules based on expert recommendations; and (3) a machine-learning based inference layer that is used to learn new predictive and defensive healthcare or technology rules, where decision tree learning algorithms are applied.

Saad et al. [171] proposed a recurrent neural network-based SA recommendation system. Body temperature, blood pressure, heart rate and respiratory rate data from wearable sensors are used as multimodal inputs to the system. The RNN-based network learns physiological signals acquired at different times and relevant medical recommendations suggested by experts. It outperforms traditional baseline methods in disease matching proportion, time complexity, and system latency due to the efficiency of RNN inference. Similar features are recognized and used to provide learned counterparts and medical advice based on classification. Symptoms not previously encountered are used for training the model. The authors compare traditional baseline methods, and the proposed method is superior in terms of the proportion of

disease matches, time complexity, and system latency benefiting from the inference efficiency of RNNs.

D. SMART HOME

A key element of the smart home is the perception of the environment, which is where situation awareness applications come in. In smart homes, the ability to accurately detect the activities of home users and respond to their needs is the focus, with more emphasis on the interaction between the user and the environment than applications in medical and healthcare.

For smart life, Lee and Lin [172] proposed a smart home system based on multi-sensor and AI technology. A wearable device with integrated accelerometers and gyroscopes and location beacons arranged throughout the room constitute a sensor network. The system can deduce the user's position based on the Received Signal Strength Indicator (RSSI) of multiple received beacons. Accelerometers and gyroscopes are used to determine the user's state, and three decision trees are used to determine the user's posture (sitting, lying, or standing), activity (chatting, watching, and reading), and whether or not the user is sleeping, respectively. The Hidden Markov Model enhances the confidence of state classification by calculating the likelihood of hidden states. When the system is aware of the user's activity, it can change the light color and play specific types of music for specific activities to achieve a livable smart home environment. Zhang et al. [173] Inferring the user's posture and position using multiple sensors integrated in smartphones and smartwatches. SVM was used for posture sensing, and in combination with heartbeat data and position data, the system was able to determine if the user had fallen and deliver an alert.

Wang et al. [174] considered the impact of alarm sounds on user's emotions and used multimodal data from smartphones and contextual information such as weather and social information to sense the situation in order to provide appropriate alarm sounds. Acoustic features such as over-zero rate, pitch type, and speed were used to evaluate the Arousal-Valence value of the alarm sound, and a K-NN algorithm was used to assign a reasonable Arousal-Valence value to the new alarm sound. The alarm sound will be changed according to the extracted vectors including feature vectors indicating user's sleep, context vectors indicating user's mood and outdoor weather and social vectors indicating social relationships. Experiments have shown that the application of the system has improved the emotional state of the user by about 11%.

E. VEHICLE ASSISTANCE

Vehicle assistance is provided to ground, air and sea vehicles. The assistance consists of sensing the state of the vehicle itself and its driver and performing the corresponding actions, which can improve driving safety and even enable autonomous driving. Many of the methods mentioned in the previous sections are also in the context of vehicle assistance, e.g., [18], [20], [21], [47], [56], [59], etc.

When an automated vehicle encounters an unmanageable situation, it needs to hand over control to the driver, which requires ensuring that the driver has sufficient situation awareness. Hofbauer et al. [176] introduced eye-movement data sensors to enable the assessment of the driver's situation awareness. The multimodal data consists of eye movement data and images captured by the vehicle camera. The system first senses the image content using a machine learning based region of interest (ROI) prediction network. Comparing the regions covered by the eye-movement data, the authors define four scenarios: (1) undetected, i.e., the eye-movement data does not overlap with the ROI region; (2) detected, i.e., the eye-movement data overlaps with the ROI region; (3) comprehended, i.e., the regions overlap for a period of time that exceeds a formulated threshold; and (4) distraction, i.e., the eye-movement data overlaps with the non-ROI region. Hu et al. [175] suggest that playing appropriate music in a vehicle can help to improve driver fatigue and negative emotions, thus enhancing the driver's situation awareness. The proposed system was experimentally demonstrated to reduce the fatigue level and negative mood level by about 49% and 36%, respectively.

Automated driving requires continuous decision support and state updates. A reinforcement learning approach is introduced in [177] to deal with the control transition problem of self-driving cars in specific traffic situations. In the event that an autonomous vehicle encounters a special situation that requires a transfer of driving authority, the safety state of the vehicle (e.g., speed, road congestion, etc.) can be modelled using RL and an optimal takeover strategy can be fitted to avoid the vehicle from performing a minimal-risk operation (i.e., stopping as safely as possible). The RL-based SA system treats the area around a single vehicle as a cell and senses the average speed of all vehicles in the cell as well as the number of automated and manually driven vehicles to serve as the environment state. Also, a successful takeover is performed as a condition for model reward. Based on learning and updating the Q-functions using a dual-delay deep deterministic policy gradient training strategy (TD3), longer autopilot times and distances are achieved compared to conventional takeover schemes.

F. POWER AND ENERGY

Energy and power grids continue to increase in size and complexity as technology advances. At the same time, the reliability and security of the grid is a continuous concern. The introduction of SA is essential to enhance the security of the grid and avoid attacks against energy systems [183].

As reviewed in the previous work on smart grids, PMUs are important sensors that sense the conditions of the grid system in real time. Using only the data measured by PMUs, the authors in [85] employed random matrix theory (RMT) to model the data and provide a more rudimentary SA for data analysis as well as visualization of the grid system from a statistical perspective. In [178], the authors introduced deep

reinforcement learning to simulate and interact with the grid through a data-driven agent. Their designed reinforcement learning agent can receive signals collected from a variety of sensors in the system, including PMUs, in order to sense and understand the state of the system at the current moment, and iterate through the state by updating the Q-value in the reinforcement learning in order to formulate a strategy for autonomous voltage control.

AI-based SA of batteries can not only sense the state of the battery, but also optimize its performance and service life. In order to better predict the service life of batteries, the authors in [179] used state-space modeling of the degradation process of lithium-ion batteries, estimated the implicit parameters of the state model using particle filtering and expectation-maximization methods, and corrected the estimated parameters using support vector regression. This effectively models the battery degradation process with stochasticity and achieves superior performance in lifetime prediction.

G. INDUSTRY AND EQUIPMENT

The traditional labor-intensive model of industrial production and equipment maintenance is becoming obsolete, and AI-based SA may update this model. Based on a deep learning approach, the authors in [180] predicted the lifetime of the equipment by collecting sensor data from the equipment to extract features of interest and training a time-series prediction data. In [181], the authors enhance the prediction ability for long-term time series data using LSTM-RNN approach. Their model was trained and tested using sensor data from electrical equipment in railroads as samples to predict the future maintenance time of the equipment.

In a production environment, the efficiency optimization of mining with complex working conditions using a vision-based SA approach is proposed in [182]. The improved YOLOv5 can accurately monitor the position of different minerals on the production equipment for mineral separation and resource recovery. This significantly optimizes the efficiency of industrial production while saving human resources.

H. PHASE SUMMARY OF APPLICATIONS USING AI-EMPOWERED SITUATION AWARENESS

The survey work in this section reveals the following key points and insights for applications in SA:

- A key challenge in applications is the integration of heterogeneous data from multiple sensors and modalities, such as in [162], [165], [172], and [176]. Ensuring the quality, accuracy and consistency of data from various sources can be challenging, especially in dynamic and unpredictable environments.
- In applications such as smart homes and vehicle assistance, the collection and analysis of multimodal data raises concerns about privacy, data security, and ethical considerations. Balancing the benefits of enhanced situation awareness with the protection of individual privacy rights is a critical challenge that needs to be addressed, federated learning might be a potential method [184].
- Integrating AI-based SA systems into human-centered workflows and decision-making processes requires careful consideration of the dynamics of human-AI interaction [170], [176]. Designing interfaces and interaction mechanisms that enable humans and AI systems to collaborate and communicate effectively is an important challenge.

VII. SITUATION AWARENESS-RELATED DATASETS

This section discusses the current status and development of SA-related datasets. Datasets are often closely related to practical applications. Therefore, this section reviews related work from healthcare, autonomous systems, and disaster management perspectives. These datasets play a crucial role in realizing SA by providing the necessary information for AI models to accurately sense, understand, and predict situations. It is important to note that only a few datasets declare themselves to be SA-relevant, and thus the work in this section relies on the paradigms, tasks, and applications discussed in the previous section. Table 6 shows the datasets discussed in this section.

A. OVERVIEW OF DATASETS

1) AUTONOMOUS SYSTEM DATASETS

Karlsruhe Institute of Technology and Toyota Technological Institute dataset (KITTI) [185] is one of the largest international dataset of computer vision algorithm reviews for autonomous driving scenarios. It was recorded in a real traffic environment over a period of up to six hours, and the dataset consists of multimodal information such as corrected and synchronized imagery, radar, GPS, and IMU speed information. However, its limitations may have included a relatively small size compared to real-world variability, focusing mainly on urban scenes, which might not adequately represent diverse driving conditions (e.g., rural, adverse weather). Although the KITTI dataset does not directly focus on SA, it has become a benchmark for many studies [133], [134], [135], [136], [138], [140], [142], [195], indirectly developing vision-based SA systems for autonomous vehicles.

Cityscapes [186] dataset provides 25K images of driving scenes in 50 cities. 5K of these images have pixel-level annotations with up to 97% coverage for various categories such as cars, pedestrians, roads, and buildings; whereas the other 20K images have coarse-grained annotations. Its limitations include an exclusive focus on the urban environment, which may lack diversity in road types, landscapes and geographic areas. This leads to its under-representation of transportation patterns and infrastructure in cities with rural or other cultural backgrounds. Focusing on semantic segmentation, the Cityscapes dataset provides significant contributions in paved road and urban traffic environments. Vision-based

TABLE 5. Summary of SA-related datasets.

Focus Area	Datasets	Modality(ies)	data volume
Autonomous System	KITTI [185]	RGB, LIDAR, Location, Vehicle motion	15K images with LiDAR data
	Cityscapes [186]	RGB, Location, Vehicle motion	25K images (5K fine annotation + 20K weakly annotation)
	Waymo Open [187]	RGB, LIDAR, Location, Vehicle motion	over 10 hours video with LiDAR data
	ROAD [188]	RGB	122K frames from approx. 3 hours video
Disaster Management	CrisisLex [189]	Text	7 datasets, over 300K tweet texts
	xBD [190]	Image	22K images
	CrisisMMD [191]	Image, Text	16K tweet texts + 18K images
	VIDI [192]	Video	4534 video clips
Healthcare	SEED [193]	EEG, eye movement signal	Over 10 hours EEG signals
	PhysioNet [194]	Image, Physiological signals, clinical diagnosis	Over 200K hospital admissions, over 370K images

autonomous vehicles were able to clearly perceive road conditions at a semantic level.

Waymo Open [187] dataset is a dataset that focuses on vision-based perception and motion prediction in autonomous driving. The Perception Dataset includes 1950 high-resolution (1920*1280px) videos of 20 seconds duration and corresponding point cloud data labeled with the four categories of 'Vehicles', 'Pedestrians', 'Cyclists', and 'Signs', whereas the Motion Dataset has 100K segments of over 200M frames of data, and the label lacks the 'Signs' category than the former. Although the images in this dataset have a higher resolution than the other similar datasets, its data volume is significantly lower than that of others. This may raise its limitation that the data is underrepresented. Besides, Access to this dataset is protected by a protocol, resulting in the possibility of receiving restrictions or limitations on content access. We mentioned the importance of 3D visual perception in Section IV, and the Waymo Open dataset elevates the visual perception of autonomous vehicles SA systems from 2D images to 3D point cloud data. Compared to the KITTI dataset, it has a higher data volume and a more advanced sensor configuration.

ROAD [188] dataset, as recent work, specifically focuses on situational awareness in autonomous driving, utilizing a purely visual approach. The dataset comprises 22 segments, each approximately 8 minutes long, containing 560,000 entity labels, 640,000 action labels, and 499,000 location labels. The limitation of the dataset is that it does not focus on the actions of pedestrians. Pedestrians, as important participants in traffic, may have an impact on driver behavior, resulting in behaviors that cannot be explained or categorized by action labels. In contrast to the previously discussed datasets, the ROAD dataset declares that it focuses on the situational awareness task and attends in more detail to the situational awareness of pedestrians by autonomous vehicles.

When applied to SA, the discussed datasets may contribute to the following:

- Perception and object detection: The annotation data provided by the dataset for various objects, such as cars, pedestrians, and cyclists, contributes to training artificial intelligence models for situational awareness in detecting and tracking surrounding entities in autonomous driving scenarios.
- Semantic Segmentation: Pixel-level annotation enables artificial intelligence models to comprehend and classify each pixel in an image, thereby enhancing the understanding of scenes in urban environments and improving situational awareness capabilities.
- Multi-sensor Fusion: The dataset's provision of multimodal sensor data aids in the fusion of information from LiDAR, cameras, and radar, enabling a comprehensive understanding of the environment and enhancing situational awareness by providing multiple perspectives.
- Environment comprehension: The multimodal dataset's diverse data types (stereo images, LiDAR point clouds) enable algorithms to understand the environment's geometry and semantics, contributing to a comprehensive understanding of the driving scene for better situational awareness.

2) DISASTER MANAGEMENT DATASETS

CrisisLex [189] is a series of datasets comprising multiple text-based collections focused on disasters. The collection includes datasets on diverse topics such as political events (BlackLivesMatterU/T1), natural disasters (CrisisLexT6, CrisisLexT26, ChileEarthquakeT1, etc.), climate-related issues (ClimateConE350), and more. Specifically, the CrisisLexT26 dataset comprises 250,000 tweets categorized into 26 types of disasters that occurred between 2012 and 2013. On the other hand, a portion of the dataset was collected decades ago, which may result in insufficient timeliness and comprehensiveness. However, the CrisisLex series of datasets provides a rich corpus for disaster management orientation, which directly enhances the performance of text-based SA systems.

xView Building Damage (xBD) [190] dataset provides high-resolution satellite images capturing building damages resulting from natural disasters. The dataset recorded 22,068

images, including pre-disaster and post-disaster images; and 19 different event markers, including earthquakes, floods, wildfires, and volcanic eruptions. They can be used for damage localization and assessment. Since the data is satellite imagery, this makes the image quality heavily dependent on the weather conditions at the location where it was taken. Observation of the dataset revealed a number of post-disaster features and buildings that were obscured by large areas of cloud. As an image dataset, the xBD dataset provides different perspectives for SA systems to perceive disasters in a remote sensing manner. Different from the disaster pictures taken from the ground perspective, remote sensing images can improve the accuracy and recall of SA due to its single perspective making the features more regular, but obtaining real-time remote sensing images may be an obstacle when deploying the system.

CrisisMMD [191] dataset is similar to CrisisLex, as both collect disaster-related data from social media. However, CrisisMMD focuses more on image information within the dataset's content. The multimodal data of images and texts contribute to providing more comprehensive information for disaster monitoring. The dataset contains three types of labels:(1) whether the content is related to a disaster or not; (2) eight types of disaster events (including car accidents, building damage, casualties, missing persons, etc.); (3) three levels of damage severity. In social media, the posting of messages is usually multimodal. The CrisisMMD dataset simulates a more realistic environment for SA systems and provides richer labels as well.

VIDI [192] dataset is a up-to-date video dataset for disaster event classification. It contains a total of 4534 video clips from 43 event categories collected from social media. Unlike perceiving events with words, the information density of visuals is more sparse and closer to the way humans perceive events. In contrast to other vision-based datasets (images), the VIDI dataset annotates the videos, allowing AI-based models to model the temporal correlation of the visuals, which improves the perception of the SA system.

When applied to SA, the discussed datasets may contribute to the following:

- Disaster Monitoring and Assessment: For instance, using satellite images from the xBD dataset enables AI models to monitor, analyze building damages, and evaluate the extent of the damage. This contributes to managing disaster-stricken areas.
- Detecting Disasters from the Internet: Learning from multimodal data disseminated on social media within the dataset assists situational awareness systems in perceiving patterns of disasters from various modes such as images, texts, and others.

3) HEALTHCARE DATASETS

The SJTU Emotion EEG Dataset (SEED) [193] contains EEG and eye-tracking signals obtained from 15 subjects while they watched 15 movie segments. The original dataset included three types of movie segments categorized as positive, negative, and neutral. Each segment lasted around four minutes, resulting in a collective signal recording of over 10 hours. Subsequent extensions, such as SEED-IV and SEED-V, introduced movie segments inducing emotions like 'happiness,' 'sadness,' 'fear,' and 'disgust,' aiming to broaden the range of captured EEG signals. The limitation of this dataset is that it involves a small number of subjects, which may have limitations in data representativeness. Secondly, the movie clips used for data collection were all Chinese movies, which may lead to errors in the dataset in terms of culture, age, and gender. However, the contribution of SEED is its use of a multimodal approach to provide individual emotion perception for SA systems.

PhysioNet [194] stands as an expansive repository housing a wide array of freely accessible physiological signal datasets and associated resources. One of the prominent medical datasets, MIMIC-III, houses clinical records of 53,423 patients admitted to intensive care units between 2001 and 2012. It encompasses diverse patient information, vital sign measurements, medication records, and additional data. In contrast, the visually centered dataset, MIMIC-CXR, predominantly comprises 377,110 annotated chest X-ray images, featuring details on anatomical locations and image orientations. However, most of the datasets are unimodal. This results in datasets that may lack sufficient contextual information for their use as medical contexts, making it challenging to interpret physiological signals into real-world scenarios.

When applied to SA, the discussed datasets may contribute to the following:

- Physiological signal comprehension: Establishing connections between an extensive range of emotional labels, disease classifications, and various biological signals enhances the situational awareness system's capacity to comprehend biological signals comprehensively. This augmentation significantly aids in elevating the system's perceptual capabilities in tasks such as emotion detection and disease monitoring.
- Disease prediction: Temporal patient case data can serve as a valuable source of historical information for AI models, augmenting their ability to predict potential diseases. Aligning and fusing multimodal physiological signals with time-series-based medical records may substantially enhance the model's robustness and performance.

B. POTENTIAL BIASES

As with several of the previously described datasets, potential biases may be introduced in the experimental design, data collection, and labeling. Researchers should be cognizant enough of their limitations and potential biases when constructing or using datasets.

Bias due to data collection methods typically occurs in datasets where the design of the experiment was used for

collection. For example, the authors of the SEED dataset used a device with integrated eye movement monitoring and EEG collection to collect data. Compared to a similar EEG dataset, the circuit design, sampling location, sampling frequency, and other parameters may not be the same. Second, the measurement of physiological signals, such as ECG and EEG, is a discrete analog of electrical signals to the human body, and thus the acquired signals are inherently subject to error [196]. Additionally, sampling bias is a pervasive factor that stems from non-random sampling. The presence of sampling bias can lead to a less representative dataset. Using the Cityscapes dataset as an example, although it contains data from up to 50 cities, all of these cities are located in Europe. This may make the models trained using this dataset less resilient to different transportation environments in other regions.

VIII. OPEN CHALLENGES AND FUTURE WORK

This section presents various open challenges encountered in situation awareness empowered by AI and multimodal technologies. The challenges captured in this work include further projections, federated learning and distributed computing for SA systems, Explainable AI for SA, SA of autonomous vehicles and mixed reality for SA.

A. FURTHER PROJECTIONS

Munir et al. [7] indicate that the deployment of AI is particularly useful for the prediction phase of SA, but only a few of the existing work based on AL techniques actually achieve prediction of the environment, e.g., [47], [50], [104], [197], etc., and most of these are predicting a single or few variables in the environment.

B. REAL-TIME SA WITH UPDATES

Increasing computing power enables AI model-based SA methods to provide or evaluate SA at near real-time speeds, but most of the applied AI models are based on supervised learning, which may lead to obsolescence of the trained models in the face of ever-changing environments, so the introduction of the concepts of on-line, incremental, or reinforcement learning is worth considering.

C. FEDERATED LEARNING AND DISTRIBUTED COMPUTATING FOR SITUATION AWARENESS SYSTEMS

The trend of multiplying data volumes and model parameters has led to the gradual dominance of big data and big models, which usually require the support of huge computational resources. Situation awareness systems will probably be hindered by environmental factors when using these technologies in the future. New challenges may lie in realizing distributed computing for situation awareness systems. Federated learning is a machine learning technique which trains models through multiple decentralized devices or servers for distributed deployment. These devices or servers hold local data samples but do not exchange data for better privacy. Federated learning for situation awareness is a new research area. The feasibility of using federated learning to achieve distributed computing for situation awareness systems [184] and its improvement in the accuracy of situation awareness systems has been demonstrated in recent work.

D. EXPLAINABLE AI FOR SITUATION AWARENESS

The domains where situation awareness is applied, such as military, medical, etc., may be very sensitive and less fault tolerant. However, when introducing AI-based approaches in situation awareness systems, factors such as false patterns, missing and noise in the training data may cause the predictions made by the model to lack in correctness and robustness, resulting in them not being trusted. Explainable Artificial Intelligence (XAI) refers to the ability of an intelligent being to communicate clearly and effectively with users, affected persons, decision makers, developers, etc., of an AI system in an explainable, understandable, and human-computer interactive manner in order to gain the trust of human beings while meeting regulatory requirements. In [198], the authors discuss a useful taxonomy for XAI approaches from three perspectives: (1) simplifying interpretation, (2) adding visual interpretation, and (3) measuring feature relevance.

E. SITUATION AWARENESS AND EXTENDED REALITY

Interaction and feedback is an important part of the situation awareness system. Extended Reality (XR), as the sum of Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), provides more intuitive and immersive methods of interaction, enabling users to interact with digital content in innovative ways. In [199], the authors point out the focus of realizing XR: (1) display devices, which provide the hardware foundation for realizing XR; (2) tracking algorithms, which track the user's body parts, gestures, orientations, and entities in the environment and establish correspondences in space; and (3) AI automation, which includes 3D scanning, image segmentation, and spatial computation (e.g., interpreting the physical scene). With the support of IoT technology, multimodal technology, and 3D situation awareness, work has been done to introduce XR technology into situation awareness systems, such as [23], which combines the results of the perception of the environment with real-time navigational footage using AR technology in order to provide more integrated information to the user.

IX. CONCLUSION

This paper provided a comprehensive overview of context-aware perception instances based on artificial intelligence technology and multimodal systems. The review encompassed various models and architectures in artificial intelligence, including machine learning, deep learning, and reinforcement learning, as well as the situational awareness tasks targeted by these methods. Findings from the survey revealed that as models become more complex, there is indeed a significant enhancement in their performance; however, interpretability has emerged as a new issue.

Of particular note is the reinforcement learning methods, due to their unique characteristics, provide sustainable

predictive layers and additional decision-making layers to SA systems. Furthermore, multimodal data offers more comprehensive environmental information to SA systems, demanding refined representation and fusion methods for multimodal data, which this paper deeply delves into.

Moreover, this paper extensively explored 3D technologies associated with situational awareness and use cases of artificial intelligence and multimodal approaches in various fields of situational awareness. Concerning data sources, only a relatively small number of datasets are constructed with the direct goal of providing situational awareness, but diverse datasets can provide situational awareness capabilities to the system indirectly. This paper summarized them from various fields, modalities, volumes and limitations.

The survey results have important implications for a variety of practical applications. For example, in military defense, the integration of reinforcement learning methods into SA systems can enable a more robust prediction and decision-making layer that improves the ability to monitor, identify, and track enemy or unknown targets in dynamic battlefield environments. Similarly, in disaster and risk management, improvements in the representation and fusion of multimodal data can enable more accurate event monitoring and prediction, enhancing preparedness and response.

The review also highlighted unresolved challenges and suggests future work in the following areas: (1) enhancing SA capabilities to offer deeper and more comprehensive environmental predictions; (2) developing real-time SA updated with AI technology; (3) employing distributed computing and federated learning for SA; (4) creating interpretable AI models for SA systems; (5) utilizing augmented reality for more comprehensive interaction.

REFERENCES

[1] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 37, no. 1, pp. 32–64, Mar. 1995, doi: 10.1518/001872095779049543.

[2] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task load Index): Results of empirical and theoretical research," in *Advances in Psychology*. Amsterdam, The Netherlands: North Holland, 1988, pp. 139–183.

[3] M. R. Endsley, "Situation awareness global assessment technique (SAGAT)," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, May 1988, pp. 789–795, doi: 10.1109/NAECON.1988.195097.

[4] R. M. Taylor, "Situational awareness rating technique (Sart): The development of a tool for aircrew systems design," in *Situational Awareness*. Evanston, IL, USA: Routledge, 2017, pp. 111–128.

[5] D. O'HARE, M. Wiggins, A. Williams, and W. Wong, "Cognitive task analyses for decision centred design and training," *Ergonomics*, vol. 41, no. 11, pp. 1698–1718, Nov. 1998, doi: 10.1080/001401398186144.

[6] P. Salmon, N. Stanton, G. Walker, and D. Green, "Situation awareness measurement: A review of applicability for C4i environments," *Appl. Ergonom.*, vol. 37, no. 2, pp. 225–238, Mar. 2006, doi: 10.1016/j.apergo.2005.02.001.

[7] A. Munir, A. Aved, and E. Blasch, "Situational awareness: Techniques, challenges, and prospects," *AI*, vol. 3, no. 1, pp. 55–77, Jan. 2022, doi: 10.3390/ai3010005.

[8] U. Ju, L. L. Chuang, and C. Wallraven, "Acoustic cues increase situational awareness in accident situations: A VR car-driving study," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3281–3291, Apr. 2022, doi: 10.1109/TITS.2020.3035374.

[9] M. Capallera, L. Angelini, Q. Meteier, O. A. Khaled, and E. Mugellini, "Human-vehicle interaction to support driver's situation awareness in automated vehicles: A systematic review," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2551–2567, Mar. 2023, doi: 10.1109/TIV.2022.3200826.

[10] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[11] G. D'Aniello, R. Gravina, M. Gaeta, and G. Fortino, "Situation-aware sensor-based wearable computing systems: A reference architecture-driven review," *IEEE Sensors J.*, vol. 22, no. 14, pp. 13853–13863, Jul. 2022, doi: 10.1109/JSEN.2022.3180902.

[12] Q. Li, K. K. H. Ng, S. C. M. Yu, C. Y. Yiu, and M. Lyu, "Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110179, doi: 10.1016/j.knosys.2022.110179.

[13] B. Lu, M. Coombes, B. Li, and W.-H. Chen, "Improved situation awareness for autonomous taxiing through self-learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3553–3564, Dec. 2016, doi: 10.1109/TITS.2016.2557588.

[14] P. Campigotto, C. Rudloff, M. Leodolter, and D. Bauer, "Personalized and situation-aware multimodal route recommendations: The favour algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 92–102, Jan. 2017, doi: 10.1109/TITS.2016.2565643.

[15] D. Thekke Kanapram, F. Patrone, P. Marin-Plaza, M. Marchese, E. L. Bodanese, L. Marcenaro, D. Martín Gómez, and C. Regazzoni, "Collective awareness for abnormality detection in connected autonomous vehicles," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3774–3789, May 2020, doi: 10.1109/JIOT.2020.2974680.

[16] C. Andrieu, M. Davy, and A. Doucet, "Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1762–1770, Jul. 2003, doi: 10.1109/tsp.2003.810284.

[17] A. Krayani, A. S. Alam, L. Marcenaro, A. Nallanathan, and C. Regazzoni, "An emergent self-awareness module for physical layer security in cognitive UAV radios," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 2, pp. 888–906, Jun. 2022, doi: 10.1109/TCCN.2022.3161937.

[18] M. Capallera, Q. Meteier, E. De Salis, M. Widmer, L. Angelini, S. Carrino, A. Sonderegger, O. A. Khaled, and E. Mugellini, "A contextual multimodal system for increasing situation awareness and takeover quality in conditionally automated driving," *IEEE Access*, vol. 11, pp. 5746–5771, 2023, doi: 10.1109/ACCESS.2023.3236814.

[19] M. Tajdinian, M. Mohammadpourfard, Y. Weng, and I. Genc, "Preserving microgrid sustainability through robust islanding detection scheme ensuring cyber-situational awareness," *Sustain. Cities Soc.*, vol. 96, Sep. 2023, Art. no. 104592, doi: 10.1016/j.scs.2023.104592.

[20] Y. Wang, A. Klautau, M. Ribero, A. C. K. Soong, and R. W. Heath, "mmWave vehicular beam selection with situational awareness using machine learning," *IEEE Access*, vol. 7, pp. 87479–87493, 2019, doi: 10.1109/ACCESS.2019.2922064.

[21] F. Zhou, X. J. Yang, and J. C. F. de Winter, "Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2284–2295, Mar. 2022, doi: 10.1109/tits.2021.3069776.

[22] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, no. 4, pp. 588–597, Oct. 2009, doi: 10.1109/tase.2008.2004965.

[23] R. W. Liu, Y. Guo, J. Nie, Q. Hu, Z. Xiong, H. Yu, and M. Guizani, "Intelligent edge-enabled efficient multi-source data fusion for autonomous surface vehicles in maritime Internet of Things," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 3, pp. 1574–1587, Sep. 2022, doi: 10.1109/tgcn.2022.3158004.

[24] N. A. Stanton, P. M. Salmon, G. H. Walker, E. Salas, and P. A. Hancock, "State-of-science: Situation awareness in individuals, teams and systems," *Ergonomics*, vol. 60, no. 4, pp. 449–466, Apr. 2017, doi: 10.1080/00140139.2017.1278796.

[25] M. R. Endsley, "Supporting human-AI teams: Transparency, explainability, and situation awareness," *Comput. Hum. Behav.*, vol. 140, Mar. 2023, Art. no. 107574, doi: 10.1016/j.chb.2022.107574.

[26] J. Jiang, A. J. Karran, C. K. Coursaris, P.-M. Léger, and J. Beringer, "A situation awareness perspective on human-AI interaction: Tensions and opportunities," *Int. J. Hum.-Comput. Interact.*, vol. 39, no. 9, pp. 1789–1806, May 2023, doi: 10.1080/10447318.2022.2093863.

[27] T. Nguyen, C. P. Lim, N. D. Nguyen, L. Gordon-Brown, and S. Nahavandi, "A review of situation awareness assessment approaches in aviation environments," *IEEE Syst. J.*, vol. 13, no. 3, pp. 3590–3603, Sep. 2019, doi: 10.1109/JSYST.2019.2918283.

[28] S. Thombre, Z. Zhao, H. Ramm-Schmidt, J. M. Vallet García, T. Malkamäki, S. Nikolskiy, T. Hammarberg, H. Nuortie, M. Z. H. Bhuiyan, S. Särkkä, and V. V. Lehtola, "Sensors and AI techniques for situational awareness in autonomous ships: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 64–83, Jan. 2022, doi: 10.1109/TITS.2020.3023957.

[29] D. Qiao, G. Liu, T. Lv, W. Li, and J. Zhang, "Marine vision-based situational awareness using discriminative deep learning: A survey," *J. Mar. Sci. Eng.*, vol. 9, no. 4, p. 397, Apr. 2021, doi: 10.3390/jmse9040397.

[30] Z. Dong, T. Xu, Y. Li, P. Feng, X. Gao, and X. Zhang, "Review and application of situation awareness key technologies for smart grid," in *Proc. IEEE Conf. Energy Internet Energy Syst. Integr.*, Nov. 2017, pp. 1–6, doi: 10.1109/EI2.2017.8245450.

[31] M. Panteli and D. S. Kirschen, "Situation awareness in power systems: Theory, challenges and applications," *Electric Power Syst. Res.*, vol. 122, pp. 140–151, May 2015, doi: 10.1016/j.epsr.2015.01.008.

[32] X. He, Q. Ai, J. Wang, F. Tao, B. Pan, R. Qiu, and B. Yang, "Situation awareness of energy Internet of Things in smart city based on digital twin: From digitization to informatization," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 7439–7458, May 2023, doi: 10.1109/JIOT.2022.3203823.

[33] H. Bavle, J. L. Sanchez-Lopez, C. Cimarelli, A. Tourani, and H. Voos, "From SLAM to situational awareness: Challenges and survey," *Sensors*, vol. 23, no. 10, p. 4849, 2110.

[34] G. D'Aniello, R. Gravina, M. Gaeta, and G. Fortino, "Situation awareness in multi-user wearable computing systems," in *Proc. IEEE Conf. Cognit. Comput. Aspects Situation Manage. (CogSIMA)*, Jun. 2022, pp. 133–138.

[35] R. Power, B. Robinson, J. Colton, and M. Cameron, "Emergency situation awareness: Twitter case studies," in *Information Systems for Crisis Response and Management in Mediterranean Countries*. Cham, Switzerland: Springer, 2014, pp. 218–231.

[36] R. Lamsal, A. Harwood, and M. R. Read, "Socially enhanced situation awareness from microblogs using artificial intelligence: A survey," *ACM Comput. Surveys*, vol. 55, no. 4, pp. 1–38, Apr. 2023, doi: 10.1145/3524498.

[37] M. R. Carlos, L. C. González, J. Wahlström, G. Ramírez, F. Martínez, and G. Runger, "How smartphone accelerometers reveal aggressive driving behavior?—The key is the representation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3377–3387, Aug. 2020, doi: 10.1109/TITS.2019.2926639.

[38] Y. Liu, P. Zhou, L. Yang, Y. Wu, Z. Xu, K. Liu, and X. Wang, "Privacy-preserving context-based electric vehicle dispatching for energy scheduling in microgrids: An online learning approach," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 462–478, Jun. 2022, doi: 10.1109/TETCI.2021.3085964.

[39] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and H. Shakeel, "Deep learning for multi-class identification from domestic violence online posts," *IEEE Access*, vol. 7, pp. 46210–46224, 2019, doi: 10.1109/ACCESS.2019.2908827.

[40] Q. Luo, Y. Chen, L. Chen, X. Luo, H. Xia, Y. Zhang, and L. Chen, "Research on situation awareness of airport operation based on Petri nets," *IEEE Access*, vol. 7, pp. 25438–25451, 2019, doi: 10.1109/ACCESS.2019.2900988.

[41] D. M. Beskow, S. Kumar, and K. M. Carley, "The evolution of political memes: Detecting and characterizing Internet memes with multi-modal deep learning," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102170, doi: 10.1016/j.ipm.2019.102170.

[42] M.-S. Baek, W. Park, J. Park, K.-H. Jang, and Y.-T. Lee, "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation," *IEEE Access*, vol. 9, pp. 131906–131915, 2021, doi: 10.1109/ACCESS.2021.3112682.

[43] L. Yang and Q. Zhao, "A BiLSTM based pipeline leak detection and disturbance assisted localization method," *IEEE Sensors J.*, vol. 22, no. 1, pp. 611–620, Jan. 2022, doi: 10.1109/JSEN.2021.3128816.

[44] A. Sepehr, O. Gomis-Bellmunt, and E. Pouresmaeil, "Employing machine learning for enhancing transient stability of power synchronization control during fault conditions in weak grids," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 2121–2131, May 2022, doi: 10.1109/TSG.2022.3148590.

[45] K. Yin, Y. Yang, C. Yao, and J. Yang, "Long-term prediction of network security situation through the use of the transformer-based model," *IEEE Access*, vol. 10, pp. 56145–56157, 2022, doi: 10.1109/ACCESS.2022.3175516.

[46] J. P. A. Dantas, M. R. O. A. Maximo, A. N. Costa, D. Geraldo, and T. Yoneyama, "Machine learning to improve situational awareness in beyond visual range air combat," *IEEE Latin Amer. Trans.*, vol. 20, no. 8, pp. 2039–2045, Aug. 2022, doi: 10.1109/TLA.2022.9853232.

[47] R. W. Liu, M. Liang, J. Nie, Y. Yuan, Z. Xiong, H. Yu, and N. Guizani, "STMGCN: Mobile edge computing-empowered vessel trajectory prediction using spatio-temporal multigraph convolutional network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7977–7987, Nov. 2022, doi: 10.1109/TII.2022.3165886.

[48] C. Y. Yiu, K. K. H. Ng, X. Li, X. Zhang, Q. Li, H. S. Lam, and M. H. Chong, "Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks," *Adv. Eng. Informat.*, vol. 53, Aug. 2022, Art. no. 101698, doi: 10.1016/j.aei.2022.101698.

[49] Z. Zhong, G. Kaiser, and B. Ray, "Neural network guided evolutionary fuzzing for finding traffic violations of autonomous vehicles," *IEEE Trans. Softw. Eng.*, vol. 49, no. 4, pp. 1860–1875, Apr. 2023, doi: 10.1109/TSE.2022.3195640.

[50] L. Sui, X. Guan, C. Cui, H. Jiang, H. Pan, and T. Ohtsuki, "Graph learning empowered situation awareness in Internet of Energy with graph digital twin," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 7268–7277, May 2023, doi: 10.1109/TII.2022.3227641.

[51] I. Bisio, C. Garibotto, F. Lavagetto, and A. Sciarrone, "Performance analysis of VCA-based target detection system for maritime surveillance," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5010–5020, Apr. 2023, doi: 10.1109/TVT.2022.3223250.

[52] Y. Wang, J. Wang, J. Jiang, S. Xu, and J. Wang, "SA-LSTM: A trajectory prediction model for complex off-road multi-agent systems considering situation awareness based on risk field," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 1–12, Nov. 2023, doi: 10.1109/TVT.2023.3287227.

[53] Y. Guo, R. W. Liu, Y. Lu, J. Nie, L. Lyu, Z. Xiong, J. Kang, H. Yu, and D. Niyato, "Haze visibility enhancement for promoting traffic situational awareness in vision-enabled intelligent transportation," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 15421–15435, Dec. 2023, doi: 10.1109/tvt.2023.3298041.

[54] M. Cafaro, I. Epicoco, M. Pulimeno, and E. Sansebastiano, "Toward enhanced support for ship sailing," *IEEE Access*, vol. 11, pp. 87047–87061, 2023, doi: 10.1109/access.2023.3303808.

[55] J. Wu, K. Ota, M. Dong, J. Li, and H. Wang, "Big data analysis-based security situational awareness for smart grid," *IEEE Trans. Big Data*, vol. 4, no. 3, pp. 408–417, Sep. 2018, doi: 10.1109/TBDATA.2016.2616146.

[56] J. Guo, X. Li, Z. Liu, J. Ma, C. Yang, J. Zhang, and D. Wu, "TROVE: A context-awareness trust model for VANETs using reinforcement learning," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6647–6662, Jul. 2020, doi: 10.1109/JIOT.2020.2975084.

[57] J. Liao, T. Liu, X. Tang, X. Mu, B. Huang, and D. Cao, "Decision-making strategy on highway for autonomous vehicles using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 177804–177814, 2020, doi: 10.1109/ACCESS.2020.3022755.

[58] C. Xu, T. Zhang, X. Kuang, Z. Zhou, and S. Yu, "Context-aware adaptive route mutation scheme: A reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13528–13541, Sep. 2021, doi: 10.1109/jiot.2021.3065680.

[59] L. Zhong, L. Zhao, C. Ding, X. Ge, J. Chen, Y. Zhang, and L. Zhang, "Vision-based 3D aerial target detection and tracking for maneuver decision in close-range air combat," *IEEE Access*, vol. 10, pp. 4157–4168, 2022, doi: 10.1109/ACCESS.2022.3140331.

[60] C. Wu, X. Cai, J. Sheng, Z. Tang, B. Ai, and Y. Wang, "Parameter adaptation and situation awareness of LTE-R handover for high-speed railway communication," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1767–1781, Mar. 2022, doi: 10.1109/TITS.2020.3026195.

[61] H. Shu, T. Liu, X. Mu, and D. Cao, "Driving tasks transfer using deep reinforcement learning for decision-making of autonomous vehicles in unsignalized intersection," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 41–52, Jan. 2022, doi: 10.1109/TVT.2021.3121985.

[62] L. Chen, Y. He, Q. Wang, W. Pan, and Z. Ming, "Joint optimization of sensing, decision-making and motion-controlling for autonomous vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 4642–4654, May 2022, doi: 10.1109/TVT.2022.3150793.

[63] R. D. Costa, C. M. Hirata, and V. U. Pugliese, "A comparative study of situation awareness-based decision-making model reinforcement learning adaptive automation in evolving conditions," *IEEE Access*, vol. 11, pp. 16166–16182, 2023, doi: 10.1109/ACCESS.2023.3245055.

[64] X. Liu, M. Derakhshani, L. Mihaylova, and S. Lambotharan, "Risk-aware contextual learning for edge-assisted crowdsourced live streaming," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 740–754, Mar. 2023, doi: 10.1109/JSAC.2022.3229423.

[65] Y. Xu, M. Zhang, and B. Jin, "Pursuing benefits or avoiding threats: Realizing regional multi-target electronic reconnaissance with deep reinforcement learning," *IEEE Access*, vol. 11, pp. 63972–63984, 2023, doi: 10.1109/ACCESS.2023.3289077.

[66] X. Zhang, H. Zhang, H. Zhou, C. Huang, D. Zhang, C. Ye, and J. Zhao, "Safe reinforcement learning with dead-ends avoidance and recovery," *IEEE Robot. Autom. Lett.*, vol. 9, no. 1, pp. 491–498, Jan. 2024, doi: 10.1109/lra.2023.3333248.

[67] Y. Li, "Deep reinforcement learning: An overview," 2018, *arXiv:1701.07274*.

[68] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[69] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/bf00116251.

[70] S. L. Salzberg, "C4.5: Programs for machine learning," *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/bf00993309.

[71] L. Breiman, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 2017, doi: 10.1201/9781315139470.

[72] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.

[73] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[74] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[75] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3149–3157.

[76] H. Wu, R. Srikant, X. Liu, and C. Jiang, "Algorithms with logarithmic or sublinear regret for constrained contextual bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–19.

[77] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2015, *arXiv:1409.2329*.

[78] Z. Wu, C. Liu, J. Wen, Y. Xu, J. Yang, and X. Li, "Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss," *IEEE Trans. Image Process.*, vol. 32, pp. 682–693, 2023, doi: 10.1109/TIP.2022.3231744.

[79] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[80] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4439–4449, Sep. 2021, doi: 10.1109/TCYB.2020.3035613.

[81] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3688–3704, Jul. 2022, doi: 10.1109/TPAMI.2021.3053577.

[82] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances neural information processing systems*, 2017, pp. 1–11.

[83] A. Coronato, G. De Pietro, and G. Paragliola, "A situation-aware system for the detection of motion disorders of patients with autism spectrum disorders," *Expert Syst. Appl.*, vol. 41, no. 17, pp. 7868–7877, Dec. 2014, doi: 10.1016/j.eswa.2014.05.011.

[84] D. Campo, M. Baydoun, P. Marin, D. Martin, L. Marcenaro, A. de la Escalera, and C. Regazzoni, "Learning probabilistic awareness models for detecting abnormalities in vehicle motions," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1308–1320, Mar. 2020, doi: 10.1109/TITS.2019.2909980.

[85] X. He, R. C. Qiu, Q. Ai, L. Chu, X. Xu, and Z. Ling, "Designing for situation awareness of future power grids: An indicator system based on linear eigenvalue statistics of large random matrices," *IEEE Access*, vol. 4, pp. 3557–3568, 2016, doi: 10.1109/ACCESS.2016.2581838.

[86] Q. Wang, S. Bu, Z. He, and Z. Y. Dong, "Toward the prediction level of situation awareness for electric power systems using CNN-LSTM network," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6951–6961, Oct. 2021, doi: 10.1109/TII.2020.3047607.

[87] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012, doi: 10.1109/TPAMI.2011.176.

[88] D. Cavaliere, J. A. Morente-Molinera, V. Loia, S. Senatore, and E. Herrera-Viedma, "Collective scenario understanding in a multivehicle system by consensus decision making," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 9, pp. 1984–1995, Sep. 2020, doi: 10.1109/TFUZZ.2019.2928787.

[89] C. V. L. Pennington, R. Bossu, F. Ofli, M. Imran, U. Qazi, J. Roch, and V. J. Banks, "A near-real-time global landslide incident reporting tool demonstrator using social media and artificial intelligence," *Int. J. Disaster Risk Reduction*, vol. 77, Jul. 2022, Art. no. 103089, doi: 10.1016/j.ijdrr.2022.103089.

[90] Y. Tian, H. Meng, F. Yuan, Y. Ling, and N. Yuan, "Vision transformer with enhanced self-attention for few-shot ship target recognition in complex environments," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3268455.

[91] J. Byun, I. Hong, B. Lee, and S. Park, "Intelligent household LED lighting system considering energy efficiency and user satisfaction," *IEEE Trans. Consum. Electron.*, vol. 59, no. 1, pp. 70–76, Feb. 2013, doi: 10.1109/TCE.2013.6490243.

[92] A. A. P. Wai, W. Huang, V. F. S. Fook, J. Biswas, H. Chi-Chun, and L. Koujuch, "Situation-aware patient monitoring in and around the bed using multimodal sensing intelligence," in *Proc. 6th Int. Conf. Intell. Environments*, Jul. 2010, pp. 128–133, doi: 10.1109/IE.2010.31.

[93] M. N. Alkhomsan, M. A. Hossain, S. M. M. Rahman, and M. Masud, "Situation awareness in ambient assisted living for smart healthcare," *IEEE Access*, vol. 5, pp. 20716–20725, 2017, doi: 10.1109/ACCESS.2017.2731363.

[94] F. Dell'Agnola, U. Pale, R. Marino, A. Arza, and D. Atienza, "MBioTracker: Multimodal self-aware bio-monitoring wearable system for online workload detection," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 5, pp. 994–1007, Oct. 2021, doi: 10.1109/TBCAS.2021.3110317.

[95] J. Chen, X. Gao, J. Rong, and X. Gao, "A situation awareness assessment method based on fuzzy cognitive maps," *J. Syst. Eng. Electron.*, vol. 33, no. 5, pp. 1108–1122, Oct. 2022, doi: 10.23919/JSEE.2022.000108.

[96] F. A. Ghaleb, B. A. S. Al-Rimy, A. Almalawi, A. M. Ali, A. Zainal, M. A. Rassam, S. Z. M. Shaid, and M. A. Maarof, "Deep Kalman neuro fuzzy-based adaptive broadcasting scheme for vehicular ad hoc network: A context-aware approach," *IEEE Access*, vol. 8, pp. 217744–217761, 2020, doi: 10.1109/ACCESS.2020.3040903.

[97] R. Böck, S. Glüge, A. Wendemuth, K. Limbrecht, S. Walter, D. Hrabal, and H. C. Traue, "Intraindividual and interindividual multimodal emotion analyses in human-machine-interaction," in *Proc. IEEE Int. Multi-Disciplinary Conf. Cognit. Methods Situation Awareness Decis. Support*, Mar. 2012, pp. 59–64, doi: 10.1109/COGSIMA.2012.6188409.

[98] H.-G. Kim and G. Y. Kim, "Deep neural network-based indoor emergency awareness using contextual information from sound, human activity, and indoor position on mobile device," *IEEE Trans. Consum. Electron.*, vol. 66, no. 4, pp. 271–278, Nov. 2020, doi: 10.1109/TCE.2020.3015197.

[99] W. Wang, H. Yin, C. Chen, A. Till, W. Yao, X. Deng, and Y. Liu, "Frequency disturbance event detection based on synchrophasors and deep learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3593–3605, Jul. 2020, doi: 10.1109/TSG.2020.2971909.

[100] Y. Gao, H. Zhang, X. Zhao, and S. Yan, "Event classification in microblogs via social tracking," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–14, May 2017, doi: 10.1145/2967502.

[101] R.-Q. Wang, Y. Hu, Z. Zhou, and K. Yang, "Tracking flooding phase transitions and establishing a passive hotline with AI-enabled social media data," *IEEE Access*, vol. 8, pp. 103395–103404, 2020, doi: 10.1109/ACCESS.2020.2994187.

[102] F. Paganelli and D. Giuli, "An ontology-based system for context-aware and configurable services to support home-based continuous care," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 2, pp. 324–333, Mar. 2011, doi: 10.1109/TITB.2010.2091649.

[103] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "Semantically enhanced UAVs to increase the aerial scene understanding," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 49, no. 3, pp. 555–567, Mar. 2019, doi: 10.1109/TSMC.2017.2757462.

[104] K. Saleem, M. Saleem, R. Z. Ahmad, A. R. Javed, M. Alazab, T. R. Gadekallu, and A. Suleman, "Situation-aware BDI reasoning to detect early symptoms of COVID-19 using smartwatch," *IEEE Sensors J.*, vol. 23, no. 2, pp. 898–905, Jan. 2023, doi: 10.1109/JSEN.2022.3156819.

[105] T. Rehder, A. Koenig, M. Goehl, L. Louis, and D. Schramm, "Lane change intention awareness for assisted and automated driving on highways," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 2, pp. 265–276, Jun. 2019, doi: 10.1109/TIV.2019.2904386.

[106] P. P. Paul, M. Gavrilova, and S. Klimenko, "Situation awareness through multimodal biometric template security in real-time environments," in *Proc. Int. Conf. Cyberworlds*, Oct. 2013, pp. 82–88, doi: 10.1109/CW.2013.80.

[107] R. M. Hegde, J. Kurniawan, and B. D. Rao, "On the design and prototype implementation of a multimodal situation aware system," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 645–657, Jun. 2009, doi: 10.1109/tmm.2009.2017631.

[108] S. Liu, S. You, Z. Lin, C. Zeng, H. Li, W. Wang, X. Hu, and Y. Liu, "Data-driven event identification in the U.S. power systems based on 2D-OLPP and RUSBoosted trees," *IEEE Trans. Power Syst.*, vol. 37, no. 1, pp. 94–105, Jan. 2022, doi: 10.1109/TPWRS.2021.3092037.

[109] J. Yang, N. Liang, B. J. Pitts, K. O. Prakah-Asante, R. Curry, M. Blommer, R. Swaminathan, and D. Yu, "Multimodal sensing and computational intelligence for situation awareness classification in autonomous driving," *IEEE Trans. Hum.-Mach. Syst.*, vol. 53, no. 2, pp. 270–281, Apr. 2023, doi: 10.1109/THMS.2023.3234429.

[110] G. Xu, Y. Cao, Y. Ren, X. Li, and Z. Feng, "Network security situation awareness based on semantic ontology and user-defined rules for Internet of Things," *IEEE Access*, vol. 5, pp. 21046–21056, 2017, doi: 10.1109/ACCESS.2017.2734681.

[111] C. Fan, F. Wu, and A. Mostafavi, "A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters," *IEEE Access*, vol. 8, pp. 10478–10490, 2020, doi: 10.1109/ACCESS.2020.2965550.

[112] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, "Multimodal categorization of crisis events in social media," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14667–14677, doi: 10.1109/CVPR42600.2020.01469.

[113] S. M. Khan and M. Chowdhury, "Situation-aware left-turning connected and automated vehicle operation at signalized intersections," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 13077–13094, Aug. 2021, doi: 10.1109/JIOT.2021.3064041.

[114] Y. Chen, Q. Jing, L. Xiao, Y. Ding, M. Hu, W. Che, and H. Lin, "A multi-level situational awareness method with dynamic multimodal data visualization for air pollution monitoring," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 489–492, doi: 10.1109/IGARSS46834.2022.9883066.

[115] R. Li, J. Cui, R. Gao, P. N. Suganthan, O. Sourina, L. Wang, and C.-H. Chen, "Situation awareness recognition using EEG and eye-tracking data: A pilot study," in *Proc. Int. Conf. Cyberworlds*, Sep. 2022, pp. 209–212, doi: 10.1109/CW55638.2022.00049.

[116] H. Yang, J. Wu, Z. Hu, and C. Lv, "Real-time driver cognitive workload recognition: Attention-enabled learning with multimodal information fusion," *IEEE Trans. Ind. Electron.*, vol. 71, no. 5, pp. 1–11, May 2023, doi: 10.1109/TIE.2023.3288182.

[117] Q. Guo, J. Jia, G. Shen, L. Zhang, L. Cai, and Z. Yi, "Learning robust uniform features for cross-media social data by using cross autoencoders," *Knowledge-Based Syst.*, vol. 102, pp. 64–75, Jun. 2016, doi: 10.1016/j.knosys.2016.03.028.

[118] M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, and Z. Huang, "Incomplete cross-modal retrieval with dual-aligned variational autoencoders," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 3283–3291, doi: 10.1145/3394171.3413676.

[119] T. Sosea, I. Sirbu, C. Caragea, D. Caragea, and T. Rebedea, "Using the image-text relationship to improve multimodal disaster tweet classication," in *Proc. 18th Int. Conf. Inf. Syst. Crisis Response Manag.*, 2021, pp. 1–14.

[120] Z. Li, Y. Zhao, Y. Geng, Z. Zhao, H. Wang, W. Chen, H. Jiang, A. Vaidya, L. Su, and D. Pei, "Situation-aware multivariate time series anomaly detection through active learning and contrast VAE-based models in large distributed systems," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2746–2765, Sep. 2022, doi: 10.1109/JSAC.2022.3191341.

[121] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 320–334, Jan. 2022, doi: 10.1109/TAFFC.2020.3000510.

[122] F. Yao, X. Sun, H. Yu, W. Zhang, W. Liang, and K. Fu, "Mimicking the Brain's cognition of sarcasm from multidisciplines for Twitter sarcasm detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 228–242, Jan. 2023, doi: 10.1109/TNNLS.2021.3093416.

[123] W. Zhou, T. Gong, J. Lei, and L. Yu, "DBCNet: Dynamic bilateral cross-fusion network for RGB-T urban scene understanding in intelligent vehicles," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 53, no. 12, pp. 7631–7641, Dec. 2023, doi: 10.1109/tsmc.2023.3298921.

[124] W. Zhou, S. Dong, J. Lei, and L. Yu, "MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 48–58, Jan. 2023, doi: 10.1109/TIV.2022.3164899.

[125] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*.

[126] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," 2021, *arXiv:2107.07651*.

[127] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115, doi: 10.1109/IROS.2017.8206396.

[128] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9441–9447, doi: 10.1109/ICRA40945.2020.9196831.

[129] H. Tashakkori, A. Rajabifard, M. Kalantari, and M. Aleksandrov, "Indoor incident situation awareness using a 3D indoor/outdoor spatial city model," in *Proc. 2nd Int. Conf. Inf. Commun. Technol. Disaster Manage. (ICT-DM)*, Nov. 2015, pp. 240–245, doi: 10.1109/ICT-DM.2015.7402050.

[130] C. Wang, H. Lin, R. Zhang, and H. Jiang, "SEND: A situation-aware emergency navigation algorithm with sensor networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 4, pp. 1149–1162, Apr. 2017, doi: 10.1109/TMC.2016.2582172.

[131] C. Wang, J. Luo, C. Zhang, and X. Liu, "A dynamic escape route planning method for indoor multi-floor buildings based on real-time fire situation awareness," in *Proc. IEEE 26th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2020, pp. 222–229, doi: 10.1109/ICPADS51040.2020.00039.

[132] T. Yamanouchi, G. Urakawa, and S. Kashihara, "UAV 3D-draping system for sharing situational awareness from aerial imagery data," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Oct. 2021, pp. 229–232, doi: 10.1109/GHTC53159.2021.9612450.

[133] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534, doi: 10.1109/CVPR.2017.691.

[134] A. González, D. Vázquez, A. M. López, and J. Amores, "On-board object detection: Multicue, multimodal, and multiview random forest of local experts," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3980–3990, Nov. 2017, doi: 10.1109/TCYB.2016.2593940.

[135] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8, doi: 10.1109/IROS.2018.8594049.

[136] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927, doi: 10.1109/CVPR.2018.00102.

[137] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2637–2646, doi: 10.1109/CVPR.2019.00275.

[138] R. Khamsehashari and K. Schill, "Improving deep multi-modal 3D object detection for autonomous driving," in *Proc. 7th Int. Conf. Autom., Robot. Appl. (ICARA)*, Feb. 2021, pp. 263–267, doi: 10.1109/ICARA51699.2021.9376453.

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

[140] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893, doi: 10.1109/ICRA.2018.8462926.

[141] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4376–4382, doi: 10.1109/ICRA.2019.8793495.

[142] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2020, pp. 1–19.

[143] M. Najibi, G. Lai, A. Kundu, Z. Lu, V. Rathod, T. Funkhouser, C. Pantofaru, D. Ross, L. S. Davis, and A. Fathi, "DOPS: Learning to detect 3D objects and predict their 3D shapes," 2020, *arXiv:2004.01170*.

[144] S. A. and S. R., "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decis. Analytics J.*, vol. 7, Jun. 2023, Art. no. 100230, doi: 10.1016/j.dajour.2023.100230.

[145] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" 2016, *arxiv:1602.04938*.

[146] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arxiv:1705.07874*.

[147] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.

[148] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[149] J. M. Rožanec, B. Fortuna, and D. Mladenić, "Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI)," *Inf. Fusion*, vol. 81, pp. 91–102, May 2022, doi: 10.1016/j.inffus.2021.11.015.

[150] H.-Y. Chen and C.-H. Lee, "Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis," *IEEE Access*, vol. 8, pp. 134246–134256, 2020, doi: 10.1109/ACCESS.2020.3006491.

[151] K. Lin, J. Liu, and J. Gao, "AI-driven decision making for auxiliary diagnosis of epidemic diseases," *IEEE Trans. Mol., Biol. Multi-Scale Commun.*, vol. 8, no. 1, pp. 9–16, Mar. 2022, doi: 10.1109/TMBMC.2021.3120646.

[152] R. Yahyaabadi and S. Nikan, "An explainable attention zone estimation for level 3 autonomous driving," *IEEE Access*, vol. 11, pp. 93098–93110, 2023, doi: 10.1109/access.2023.3309810.

[153] H. Allahabadi et al., "Assessing trustworthy AI in times of COVID-19: Deep learning for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients," *IEEE Trans. Technol. Soc.*, vol. 3, no. 4, pp. 272–289, Dec. 2022, doi: 10.1109/TTS.2022.3195114.

[154] N. I. Papandrianos, A. Feleki, S. Moustakidis, E. I. Papageorgiou, I. D. Apostolopoulos, and D. J. Apostolopoulos, "An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and grad-CAM," *Appl. Sci.*, vol. 12, no. 15, p. 7592, Jul. 2022, doi: 10.3390/app12157592.

[155] Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, and S.-C. Zhu, "Interpretable CNNs for object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3416–3431, Oct. 2021, doi: 10.1109/TPAMI.2020.2982882.

[156] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021, doi: 10.1109/ACCESS.2021.3070212.

[157] W. Xu, J. Wang, Y. Wang, G. Xu, D. Lin, W. Dai, and Y. Wu, "Where is the model looking at?—Concentrate and explain the network attention," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 506–516, Mar. 2020, doi: 10.1109/JSTSP.2020.2987729.

[158] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017, doi: 10.1007/s11263-016-0981-7.

[159] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Explainable video action reasoning via prior knowledge and state transitions," in *Proc. 27th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 521–529, doi: 10.1145/3343031.3351040.

[160] J. Gong, G. Fan, L. Yu, J. P. Havlicek, D. Chen, and N. Fan, "Joint view-identity manifold for infrared target tracking and recognition," *Comput. Vis. Image Understand.*, vol. 118, pp. 211–224, Jan. 2014, doi: 10.1016/j.cviu.2013.10.002.

[161] E. Gundogdu, H. Ozkan, H. S. Demir, H. Ergezer, E. Akagündüz, and S. K. Pakin, "Comparison of infrared and visible imagery for object tracking: Toward trackers with superior IR performance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPRW.2015.7301290.

[162] S. Liu and Z. Liu, "Multi-channel CNN-based object detection for enhanced situation awareness," 2017, *arXiv:1712.00075*.

[163] S. Liu, H. Liu, V. John, Z. Liu, and E. Blasch, "Enhanced situation awareness through CNN-based deep multimodal image fusion," *Opt. Eng.*, vol. 59, no. 5, p. 1, May 2020, doi: 10.1117/1.oe.59.5.053103.

[164] C.-E. Lee, J. Baek, J. Son, and Y.-G. Ha, "Deep AI military staff: Cooperative battlefield situation awareness for commander's decision making," *J. Supercomput.*, vol. 79, no. 6, pp. 6040–6069, Apr. 2023, doi: 10.1007/s11227-022-04882-w.

[165] M. Wang and M. S. Gerber, "Using Twitter for next-place prediction, with an application to crime prediction," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 941–948, doi: 10.1109/SSCI.2015.138.

[166] L. Vomfell, W. K. Härdle, and S. Lessmann, "Improving crime count forecasts using Twitter and taxi data," *Decis. Support Syst.*, vol. 113, pp. 73–85, Sep. 2018, doi: 10.1016/j.dss.2018.07.003.

[167] Q. Gu, S. Jiang, M. Lian, and C. Lu, "Health and safety situation awareness model and emergency management based on multi-sensor signal fusion," *IEEE Access*, vol. 7, pp. 958–968, 2019, doi: 10.1109/ACCESS.2018.2886061.

[168] Y. Rizk, H. S. Jomaa, M. Awad, and C. Castillo, "A computationally efficient multi-modal classification approach of disaster-related Twitter images," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 2050–2059, doi: 10.1145/3297280.3297481.

[169] P. Delir Haghighi, A. Perera, M. Indrawan-Santiago, and T. Minh Huynh, "Situation-aware mobile health monitoring," presented at the *Proc. 11th Int. Conf. Mobile Ubiquitous Syst. Comput., Netw. Services*, Jul. 2014.

[170] A. Henaien, H. Ben Elhadj, and L. Chaari Fourati, "Combined machine learning and semantic modelling for situation awareness and healthcare decision support," in *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*. Cham, Switzerland: Springer, 2020, pp. 197–209.

[171] A. Saad, H. Fouad, and A. A. Mohamed, "Situation-aware recommendation system for personalized healthcare applications," *J. Ambient Intell. Humanized Comput.*, Feb. 2021, doi: 10.1007/s12652-021-02927-1.

[172] S.-Y. Lee and F. J. Lin, "Situation awareness in a smart home environment," in *Proc. IEEE 3rd World Forum Internet Things*, Dec. 2016, pp. 678–683.

[173] S. Zhang, P. McCullagh, H. Zheng, and C. Nugent, "Situation awareness inferred from posture transition and location: Derived from smartphone and smart home sensors," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 6, pp. 814–821, Dec. 2017, doi: 10.1109/THMS.2017.2693238.

[174] J. Wang, Y. Guo, W. Han, J. Zheng, H. Peng, X. Hu, and J. Cheng, "Mobile crowdsourcing based context-aware smart alarm sound for smart living," *Pervas. Mobile Comput.*, vol. 55, pp. 32–44, Apr. 2019, doi: 10.1016/j.pmcj.2019.02.003.

[175] X. Hu, J. Deng, J. Zhao, W. Hu, E. C.-H. Ngai, R. Wang, J. Shen, M. Liang, X. Li, V. C. M. Leung, and Y.-K. Kwok, "SAfeDJ: A crowd-cloud codesign approach to situation-aware music delivery for drivers," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 1, pp. 1–24, Oct. 2015, doi: 10.1145/2808201.

[176] M. Hofbauer, C. B. Kuhn, L. Püttner, G. Petrovic, and E. Steinbach, "Measuring driver situation awareness using region-of-interest prediction and eye tracking," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2020, pp. 91–95, doi: 10.1109/ISM.2020.00022.

[177] R. Alms, A. Noulis, E. Mintsis, L. Lücken, and P. Wagner, "Reinforcement learning-based traffic control: Mitigating the adverse impacts of control transitions," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 187–198, 2022, doi: 10.1109/OJITS.2022.3158688.
[178] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020, doi: 10.1109/TPWRS.2019.2941134.
[179] Y. Keshun, Q. Guangqi, and G. Yingkui, "Remaining useful life prediction of lithium-ion batteries using EM-PF-SSA-SVR with gamma stochastic process," *Meas. Sci. Technol.*, vol. 35, no. 1, Oct. 2023, Art. no. 015015, doi: 10.1088/1361-6501/acfbef.
[180] Y. Keshun, Q. Guangqi, and G. Yingkui, "Optimizing prior distribution parameters for probabilistic prediction of remaining useful life using deep learning," *Rel. Eng. Syst. Saf.*, vol. 242, Feb. 2024, Art. no. 109793, doi: 10.1016/j.ress.2023.109793.
[181] Q. Wang, S. Bu, and Z. He, "Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6509–6517, Oct. 2020, doi: 10.1109/TII.2020.2966033.
[182] Y. Keshun and L. Huizhong, "Feature detection of mineral zoning in spiral slope flow under complex conditions based on improved YOLOv5 algorithm," *Phys. Scripta*, vol. 99, no. 1, Dec. 2023, Art. no. 016001, doi: 10.1088/1402-4896/ad0f7d.
[183] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2218–2234, May 2020, doi: 10.1109/TSG.2019.2949998.
[184] S. Seo, J. Lee, H. Ko, and S. Pack, "Situation-aware cluster and quantization level selection algorithm for fast federated learning," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13292–13302, Aug. 2023, doi: 10.1109/JIOT.2023.3262582.
[185] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.
[186] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.
[187] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," presented at the *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.
[188] G. Singh, S. Akrigg, M. D. Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Grazioso, A. Bradley, G. D. Gironimo, and F. Cuzzolin, "ROAD: The road event awareness dataset for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1036–1054, Jan. 2023, doi: 10.1109/TPAMI.2022.3150906.
[189] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "CrisisLex: A lexicon for collecting and filtering microblogged communications in crises," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, vol. 8, no. 1, pp. 1–12, doi: 10.1609/icwsm.v8i1.14538.
[190] R. Gupta, "XBD: A dataset for assessing building damage from satellite imagery," 2019, *arXiv:1911.09296*.
[191] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," 2018, *arXiv:1805.00713*.
[192] D. Sesver, A. E. Gençoglu, Ç. E. Yildiz, Z. Günindi, F. Habibi, Z. A. Yazici, and H. K. Ekenel, "VIDI: A video dataset of incidents," in *Proc. IEEE 14th Image, Video, Multidimensional Signal Process. Workshop*, Jun. 2022, pp. 1–5, doi: 10.1109/IVMSP54334.2022.9816319.
[193] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015, doi: 10.1109/TAMD.2015.2431497.
[194] G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: A web-based resource for the study of physiologic signals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 70–75, May 2001, doi: 10.1109/51.932728.
[195] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253, doi: 10.1109/CVPR.2018.00033.
[196] L. Yan and J. Bae, "Challenges of physiological signal measurements using electrodes: Fudamentals to understand the instrumentation," *IEEE Solid StateCircuits Mag.*, vol. 9, no. 4, pp. 90–97, Jul. 2017, doi: 10.1109/MSSC.2017.2745860.
[197] N. Gamal, S. Ghoniemy, H. M. Faheem, and N. A. Seada, "Sentiment-based spatiotemporal prediction framework for pandemic outbreaks awareness using social networks data classification," *IEEE Access*, vol. 10, pp. 76434–76469, 2022, doi: 10.1109/ACCESS.2022.3192417.
[198] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers Big Data*, vol. 4, pp. 1–17, Jul. 2021.
[199] A. Çöltekin, I. Lochhead, M. Madden, S. Christophe, A. Devaux, C. Pettit, O. Lock, S. Shukla, L. Herman, Z. Stacho, P. Kubíček, D. Snopková, S. Bernardes, and N. Hedley, "Extended reality in spatial sciences: A review of research challenges and future directions," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 7, p. 439, Jul. 2020, doi: 10.3390/ijgi9070439.

JIELI CHEN received the B.Eng. degree in communication engineering from Dongguan University of Technology, China, and the M.Sc. degree in communication and signal processing from Newcastle University, U.K. He is currently pursuing the Ph.D. degree with Xi'an Jiaotong-Liverpool University. His research interests include situation and data analytics, machine learning, and multimedia signal processing.

KAH PHOOI SENG (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from the University of Tasmania, Australia. She was a Professor at the Department Head of Computer Science and Networked System, Sunway University. Before joining Sunway University, she was an Associate Professor at the School of Electrical and Electronic Engineering, Nottingham University. She has worked or attached to Australian-based and U.K.-based universities, including Monash University, Griffith University, the University of Tasmania, the University of Nottingham, Sunway University, Edith Cowan University, and Charles Sturt University. Prior to joining the Queensland University of Technology (QUT), she was an Adjunct Professor at the School of Engineering and Information Technology, UNSW. She is currently a Professor of artificial intelligence with Xian Jiaotong-Liverpool University and an Adjunct Professor with the School of Computer Science, QUT. She has a strong record of publications and has published more than 250 papers in journals and internationally refereed conferences. She has participated in more than U.S. $1.8 million research grant projects from the government and industry in Australia and overseas. She has supervised or co-supervised 15 Ph.D. students to completion and more than 25 higher-degree research students. Her research interests include computer science and engineering, including artificial intelligence (AI), data science and machine learning, big data, multimodal information processing, intelligent systems, the Internet of Things (IoT), embedded systems, mobile software development, affective computing, computer vision, and the development of innovative technologies for real-world applications. She is an Associate Editor of IEEE Access. She also serves on the editorial board or committees of several journals and international conferences.

JEREMY SMITH (Member, IEEE) received the degree in engineering science from the University of Liverpool in 1984. He then undertook Ph.D. research at the Automated Welding Group, under the leadership of Professor Lucas. He is currently a Professor with the Department of Electrical and Electronics, UoL. His research interests include vision-based sensors and control systems, advanced digital and parallel processing systems, embedded systems, neural network and fuzzy logic, robotic and navigation.

LI-MINN ANG (Senior Member, IEEE) received the B.Eng. (Hons.) and Ph.D. degrees from Edith Cowan University (ECU), Australia. He was an Associate Professor of networked and computer systems at the School of Information and Communication Technology (ICT), Griffith University. He has worked at Australian and U.K. Universities, including Monash University, the University of Nottingham, ECU, California State University (CSU), and Griffith University. He is currently a Professor of electrical and computer engineering with the School of Science, Technology, and Engineering, University of the Sunshine Coast (USC), Australia. His research interests include computer, electrical, and systems engineering, including the Internet of Things, intelligent systems and data analytics, machine learning, visual information processing, embedded systems, wireless multimedia sensor systems, reconfigurable computing (FPGA), and the development of innovative technologies for real-world systems, including smart cities, engineering, agriculture, environment, health, and defense. He is a Senior Fellow of the Higher Education Academy, U.K.