Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things

Jing Zhang, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—In the Internet-of-Things (IoT) era, billions of sensors and devices collect and process data from the environment, transmit them to cloud centers, and receive feedback via the Internet for connectivity and perception. However, transmitting massive amounts of heterogeneous data, perceiving complex environments from these data, and then making smart decisions in a timely manner are difficult. Artificial intelligence (AI), especially deep learning, is now a proven success in various areas, including computer vision, speech recognition, and natural language processing. AI introduced into the IoT heralds the era of AI of things (AIoT). This article presents a comprehensive survey on AIoT to show how AI can empower the IoT to make it faster, smarter, greener, and safer. Specifically, we briefly present the AIoT architecture in the context of cloud computing, fog computing, and edge computing. Then, we present progress in AI research for IoT from four perspectives: 1) perceiving; 2) learning; 3) reasoning; and 4) behaving. Next, we summarize some promising applications of AIoT that are likely to profoundly reshape our world. Finally, we highlight the challenges facing AIoT and some potential research opportunities.**

***Index Terms*—3-D, aged care, artificial intelligence (AI), biometric recognition, causal reasoning, cloud/fog/edge computing, deep learning, human–machine interaction, Internet of Things (IoT), machine translation (MT), privacy, security, sensors, smart agriculture, smart city, smart grids, speech recognition.**

I. Introduction

THE Internet of Things (IoT), a term originally coined by Kevin Ashton at MIT's Auto-ID Center [1], refers to a global intelligent network that enables cyber–physical interactions by connecting numerous things with the capacity to perceive, compute, execute, and communicate with the Internet; process and exchange information between things, data centers, and users; and deliver various smart services [2], [3]. From the radio-frequency identification (RFID) devices developed in the late 1990s to modern smart things, including cameras, lights, bicycles, electricity meters, and wearable devices, the IoT has developed rapidly over the last twenty years in parallel with advances in networking technologies, including Bluetooth, Wi-Fi, and long-term evolution (LTE). The IoT represents a key infrastructure for supporting various applications [4], e.g., smart homes [5], [6], smart transportation [7], [8], smart grids [9], and smart healthcare [10], [11]. According to McKinsey's report [12], the IoT sector will contribute $2.7 to $6.2 trillion to the global economy by 2025.

A typical IoT architecture has three layers [13]: 1) a perception layer; 2) a network layer; and 3) an application layer. The perception layer lies at the bottom of the IoT architecture and consists of various sensors, actuators, and devices that function to collect data and transmit them to the upper layers. The network layer lies at the center of the IoT architecture and comprises different networks [e.g., local area networks (LANs), cellular networks, and the Internet] and devices (e.g., hubs, routers, and gateways) enabled by various communication technologies, such as Bluetooth, Wi-Fi, LTE, and fifth-generation mobile networks (5G). The application layer is the top IoT layer and it is powered by cloud computing platforms, offering customized services to users, e.g., data storage and analysis. In conventional IoT solutions, data collected from sensors are transmitted to the cloud computing platform through the networks for further processing and analysis before delivering the results/commands to end devices/actuators.

However, this centralized architecture faces significant challenges in the context of the massive numbers of sensors used across various applications. Based on reports from Cisco [14] and IDC [15], 50 billion devices will be IoT connected by 2025, generating 79.4 ZB of data. Transmitting this huge amount of data requires massive bandwidth, and cloud processing and sending the results back to end devices lead to high latency. To address this issue, "fog computing," coined by Cisco [16], aims to bring storage, computation, and networking capacity to the edge of the network (e.g., to distributed fog nodes, such as routers) in proximity to the devices. Fog computing offers the advantages of low latency and high computational capacity for IoT applications [17], [18]. "Edge computing" has also recently been proposed by further deploying computing capacity on edge devices in proximity to sensors and actuators [19], [20]. Note that the terms fog computing and edge computing are interchangeable in some literature [19], [21] or the fog is treated as a part of the broader concept of edge computing [22]. For clarity, here we treat them as different concepts, i.e., fog computing at the network side and edge computing at the thing side. Edge

Manuscript received September 16, 2020; revised October 24, 2020; accepted November 12, 2020. Date of publication November 19, 2020; date of current version May 7, 2021. This work was supported by the Australian Research Council Projects under Grant FL-170100117, Grant DP-180103424, and Grant IH-180100002. *(Corresponding author: Dacheng Tao.)*

The authors are with the School of Computer Science, Faculty of Engineering, University of Sydney, Darlington, NSW 2008, Australia (e-mail: jing.zhang1@sydney.edu.au; dacheng.tao@sydney.edu.au).

Digital Object Identifier 10.1109/JIOT.2020.3039359

2327-4662 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF EXEMPLAR AIOT SENSORS. A: AGRICULTURE, C: CITIES/HOMES/BUILDINGS, E: EDUCATION, G: GRIDS, H: HEALTHCARE, I: INDUSTRY, S: SECURITY, AND T: TRANSPORTATION

Sensor Type	Scalar	Vector	Multimedia
Sensors	altimeter, ammeter, hygrometer, light meter, manometers, ohmmeter, tachometer, thermometer, voltmeter, wattmeter	anemometer, accelerometer, gyroscope	microphone, camera, lidar, CT/MRI/ultrasound scanner
Data Type	scalar	vector	2/3/4D tensor
Applications	A,C,E,G,H,I,T	A,C,G,I,T	A,C,E,G,H,I,S,T

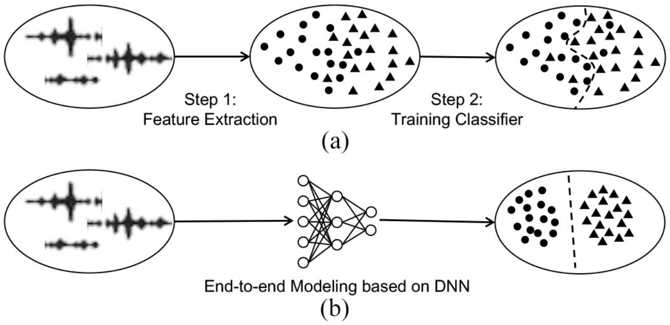

Fig. 1. Schematic paradigm of (a) classical machine learning methods and (b) deep learning.

computing can process and analyze data on premises and make decisions instantly, thereby benefitting latency-sensitive IoT applications. The processed data from different devices can then be aggregated at the fog node or cloud center for further analysis to enable various services.

In addition to these challenges created by massive numbers of sensors, another challenge arises through their heterogeneous nature [23], including scalar sensors, vector sensors, and multimedia sensors as summarized in Table I. Perceiving and understanding dynamic and complex environments from sensor data are fundamental to IoT applications providing useful services to users. As a result, various intelligent algorithms have been proposed for certain applications with scalar and vector sensors, e.g., decision rules-based methods and data-driven methods. Typically, these methods use handcrafted features extracted from data for further prediction, classification, or decision [Fig. 1(a)]. However, this paradigm of using handcrafted features and shallow models is unsuited to modern IoT applications with multimedia sensors. First, multimedia sensor data are high dimensional and unstructured (semantics are unavailable without additional processing), so it is difficult to design handcrafted features for them without domain knowledge. Second, handcrafted features are usually vulnerable to noise and different types of variance (e.g., illumination and viewpoint) in data, limiting their representation and discrimination capacity. Third, feature design and model training are separate, without joint optimization.

The last few years has witnessed a renaissance in artificial intelligence (AI) assisted by deep learning. Deep neural networks (DNNs) have been widely used in many areas and have achieved excellent performance in many applications, including speech recognition [24], face recognition [25], image classification [26], object detection [27], semantic segmentation [28], and natural language processing [29], benefitting from their powerful capacity to feature learn and end-to-end model [Fig. 1(b)]. Moreover, with modern computational devices, e.g., graphics processing units (GPUs) and tensor processing units (TPUs), DNNs can efficiently and automatically discover discriminative feature representations from large-scale labeled or unlabeled data sets in a supervised or unsupervised manner [30]. Deploying DNNs into cloud platforms, fog nodes, and edge devices in IoT systems enables the construction of an intelligent hybrid computing architecture capable of leveraging the power of deep learning to process massive quantities of data and extract structured semantic information with low latency. Therefore, advances in deep learning have paved a clear way for improving the perceiving ability of IoT systems with large numbers of heterogeneous sensors.

Although an IoT's perception system is a critical component of the architecture, simply adapting to and interacting with the dynamic and complex world are insufficient. For example, edge cases exist in the real world that may not be seen in the training set nor defined in the label set, resulting in the degeneration of a pretrained model. Another example is in industry, where the operating modes of machines may drift or change due to fatigue or wear and tear. Consequently, models trained for the initial mode cannot adapt to this variation, leading to a performance loss. These issues are related to some well-known machine learning research topics, including few-shot learning (FSL) [31], zero-shot learning (ZSL) [32], metalearning [33], unsupervised learning (USL) [34], semisupervised learning (SSL) [35], transfer learning (TL) [36], and domain adaptation (DA) [37], [38]. Deep learning has facilitated progress in these areas, suggesting that deep learning can be similarly leveraged to improve IoT system learning. Furthermore, to interact with the environment and humans, an IoT system should be able to reason and behave. For example, a man parks his car in a parking lot every morning and leaves regularly on these days. Therefore, a smart parking system may infer that he probably works nearby. Then, it can recommend and introduce some parking offers, car maintenance, and nearby restaurants to him via an AI chatbot. These application scenarios could benefit from recent advances in causal inference and discovery [39], graph-based reasoning [40], reinforcement learning (RL) [41], and speech recognition and synthesis [24], [42].

According to Cisco's white paper [43], 99.4% of physical objects are still unconnected. Advanced communication technologies, such as Wi-Fi 6 (IEEE 802.11ax standard) and 5G and AI technologies will enable mass connection. This heralds the era of the AI of things (AIoT), where AI encounters IoT. Both academia and industry have invested heavily in AIoT, and various AIoT applications have now been developed, providing services and creating value. Therefore, here we performed a survey of this emerging area to demonstrate how AI technologies empower things with intelligence and enhance applications.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

A. Contributions of This Survey

There are several excellent existing surveys on IoT covering different perspectives, a detailed discussion and comparison of which is provided below. Here, we specifically focus on AIoT and provide an overview of research advances, potential challenges, and future research directions through a comprehensive literature review and detailed discussion. The contributions of this survey can be summarized as follows.

1) We discuss AIoT system architecture in the context of cloud computing, fog computing, and edge computing.
2) We present progress in AI research for IoT, applying a new taxonomy: perceiving, learning, reasoning, and behaving.
3) We summarize some promising applications of AIoT and discuss enabling AI technologies.
4) We highlight challenges in AIoT and some potential research opportunities.

B. Relationship to Related Surveys

We first review existing surveys related to IoT and contrast them with our work. Since the IoT is related to many topics, such as computing architectures, networking technologies, applications, security, and privacy, surveys have tended to focus on one or some of these topics. For example, Atzori *et al.* [44] described the IoT paradigm from three perspectives: 1) "things" oriented; 2) "Internet" oriented; and 3) "semantic" oriented, corresponding to sensors and devices, networks, and data processing and analysis, respectively. They reviewed enabling technologies and IoT applications in different domains and also analyzed some remaining challenges with respect to security and privacy. Whitmore *et al.* [45] presented a comprehensive survey on IoT and identified recent trends and challenges. We review the other surveys according to the specific topic covered.

1) Architecture: In [46], several typical IoT architectures were reviewed, including software-defined network-based architectures, the MobilityFirst architecture, and the CloudThings architecture. They argued that future IoT architectures should be scalable, flexible, interoperable, energy efficient, and secure such that the IoT system can integrate and handle huge numbers of connected devices. Lin *et al.* [13] discussed two typical architectures: 1) the three-layer architecture (i.e., with a perception layer, network layer, and application layer) and 2) the service-oriented architecture. For the IoT computing architecture, integrating cloud computing [47] with fog/edge computing [13] has attracted increasing attention. Chiang and Zhang [17] and Pan and McElhannon [20] provided a detailed review of fog computing and edge computing for IoT. Since we focus on AI-empowered IoT, we are also interested in the cloud/fog/edge computing architectures of IoT systems, especially those tailored for deep learning. More detail will be presented in Section II.

2) Networking Technologies: Connecting massive numbers of things to data centers and transmitting data at scale relies on various networking technologies. Verma *et al.* [48] presented a comprehensive survey of network methodologies, including data center networks, hyperconverged networks, massively parallel mining networks, and edge analytics networks, which support real-time analytics of massive IoT data. Wireless sensor networks have also been widely used in IoT to monitor physical or environmental conditions [49]. The recently developed 5G mobile networks can provide very high data rates at extremely low latency and a manifold increase in base station capacity. 5G is expected to boost the number of connected things and drive the growth of IoT applications [50]. Due to the massive numbers of sensors and network traffic, resource management in IoT networks has become a topic of interest, with advanced deep learning technologies showing promising results [51]. Although we also focus on deep learning for IoT, we are more interested in its role in IoT data processing rather than networking, which is, therefore, beyond the scope of this survey.

3) Data Processing: Massive sensor data must be processed to extract useful information before being used for further analysis and decision making. Data mining and machine learning approaches have been used for IoT data processing and analysis [52], [53]. Moreover, the context of IoT sensors can provide auxiliary information to help understand sensor data. Therefore, various context-aware computing methods have been proposed for IoT [54]. There has recently been rapid progress in deep learning, with these positive effects also impacting IoT data processing, e.g., streaming data analysis [55], mobile multimedia processing [56], manufacturing inspection [57], and health monitoring. In contrast, we conduct this survey on deep learning for IoT data processing using a new taxonomy, i.e., how deep learning improves the ability of IoT systems to perceive, learn, reason, and behave. Since deep learning is itself a rapidly developing area, our survey covers the latest progress in deep learning in various IoT application domains.

4) Security and Privacy: Massive user data are collected via ubiquitous connected sensors, which may be transmitted and stored in the cloud through IoT networks. These data may contain some biometric information, such as faces, voice, or fingerprints. Cyberattacks on IoT systems may result in data leakage, so data security and privacy have become a critical concern in IoT applications [18]. Recently, access control [58] and trust management [59] approaches have been reviewed to protect the security and privacy of IoT. We also analyze this issue and review progress advanced by AI, such as federated learning (FL) [60].

5) Applications: Almost all surveys refer to various IoT application domains, including smart cities [61], smart homes [6], smart healthcare [62], smart agriculture [63], and smart industry [4]. Furthermore, IoT applications based on specific things, e.g., the Internet of Vehicles (IoV) [7] and Internet of Video Things (IoVT) [23], have also been rapidly developed. We also summarize some promising applications of AIoT and demonstrate how AI enables them to be faster, smarter, greener, and safer.

C. Organization

The organization of this article is shown in Fig. 2. We first discuss AIoT computing architecture in Section II. Then, we

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

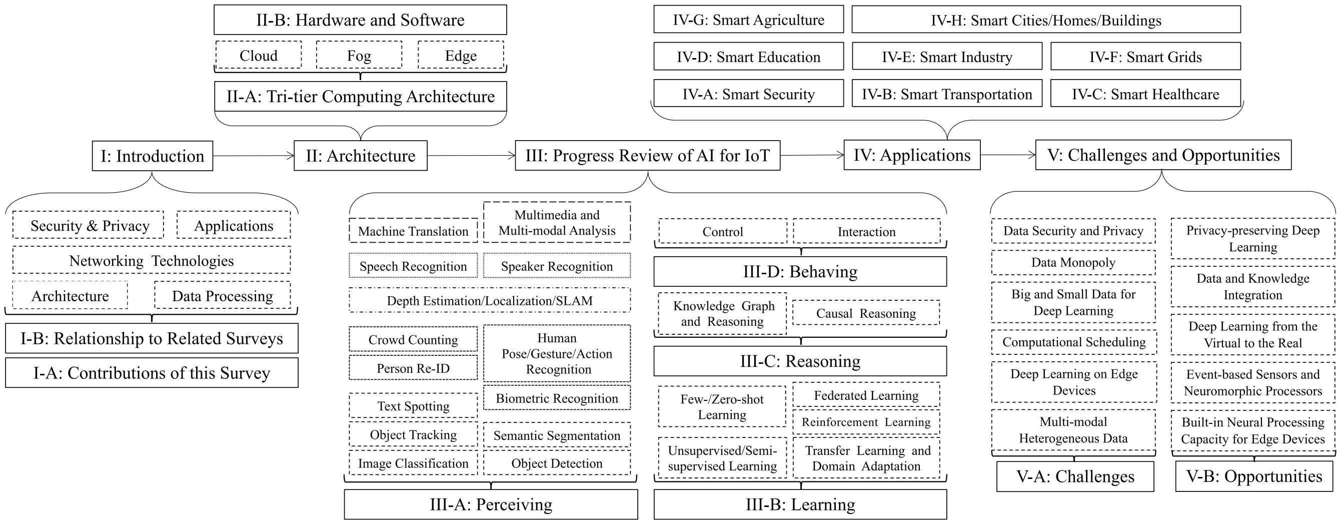
II-B: Hardware and Software
Cloud
Fog
Edge
II-A: Tri-tier Computing Architecture
IV-G: Smart Agriculture
IV-H: Smart Cities/Homes/Buildings
IV-D: Smart Education
IV-E: Smart Industry
IV-F: Smart Grids
IV-A: Smart Security
IV-B: Smart Transportation
IV-C: Smart Healthcare
I: Introduction
II: Architecture
III: Progress Review of AI for IoT
IV: Applications
V: Challenges and Opportunities
Security & Privacy
Applications
Networking Technologies
Architecture
Data Processing
I-B: Relationship to Related Surveys
I-A: Contributions of this Survey
Machine Translation
Multimedia and Multi-modal Analysis
Speech Recognition
Speaker Recognition
Depth Estimation/Localization/SLAM
Crowd Counting
Person Re-ID
Human Pose/Gesture/Action Recognition
Text Spotting
Biometric Recognition
Object Tracking
Semantic Segmentation
Image Classification
Object Detection
III-A: Perceiving
Control
Interaction
III-D: Behaving
Knowledge Graph and Reasoning
Causal Reasoning
III-C: Reasoning
Few-/Zero-shot Learning
Federated Learning
Reinforcement Learning
Unsupervised/Semi-supervised Learning
Transfer Learning and Domain Adaptation
III-B: Learning
Data Security and Privacy
Data Monopoly
Big and Small Data for Deep Learning
Computational Scheduling
Deep Learning on Edge Devices
Multi-modal Heterogeneous Data
V-A: Challenges
Privacy-preserving Deep Learning
Data and Knowledge Integration
Deep Learning from the Virtual to the Real
Event-based Sensors and Neuromorphic Processors
Built-in Neural Processing Capacity for Edge Devices
V-B: Opportunities

Fig. 2. Diagram of the organization of this article.

present a comprehensive survey of enabling AI technologies for AIoT in Section III, followed by a summary of AIoT applications in Section IV. The challenges faced by AIoT and research opportunities are discussed in Section V, followed by conclusions in Section VI.

II. Architecture

In this section, we discuss the architecture for AIoT applications. Similar to [13] and [23], we also adopt a tri-tier architecture but from the perspective of computing. For simplicity, we term the three layers as the cloud/fog/edge computing layer, as shown in Fig. 3. The edge computing layer may function like the perception layer in [13] and smart visual sensing block in [23]. It also supports control and execution over sensors and actuators. Thereby, this layer aims to empower AIoT systems with the ability to perceive and behave. The fog computing layer is embodied in the fog nodes within the networks, such as hubs, routers, and gateways. The cloud computing layer supports various application services, functioning similarly to the application layer [13] and intelligent integration block in [23]. The fog and cloud computing layers mainly aim to empower AIoT systems with the ability of learning and reasoning since they can access massive amounts of data and have vast computation resources. It is noteworthy that the edge things and fog nodes are always distributed while the cloud is centralized in the AIoT network topology.

A. Tri-Tier Computing Architecture

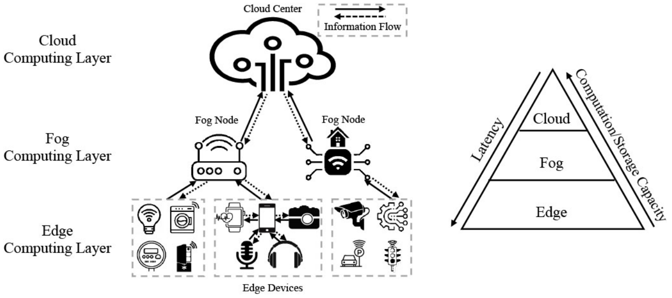

Fig. 3. Diagram of the tri-tier computing architecture of AIoT.

1) Cloud Computing Layer: The cloud enables AIoT enterprises to use computing resources virtually via the Internet instead of building their physical infrastructure on premises. It can provide flexible, scalable, and reliable resources, including computation, storage, and network for enabling various AIoT applications. Typically, real-time data streams from massive distributed sensors and devices are transmitted to the remote cloud center through the Internet, where they are further integrated, processed, and stored. With the off-the-shelf deep learning tools and scalable computing hardware, it is easy to set up the production environment on the cloud, where DNNs are trained and deployed to process the massive amounts of data. An important feature of cloud computing is that it provides elastic computing resources in the pay-as-you-go way, which is useful for the AIoT services with fluctuant traffic loads. Another feature is that it can leverage all the data from the registered devices in an AIoT application, which is useful for training deep models with better representation and generalization ability.

2) Fog Computing Layer: Fog computing brings storage, computation, and networking capacity to the edge of the network that is in the proximity of devices. The facilities or infrastructures that provide fog computing service are called fog nodes, e.g., routers, switches, gateways, and wireless access points. Although functioning similarly to cloud computing, fog computing offers a key advantage, i.e., low latency, since it is closer to devices. Besides, fog computing can provide continuity of service without the need for the Internet, which is important for specific AIoT applications with an unstable Internet connection, e.g., in agriculture, mining, and shipping domains. The other advantage of fog computing is the protection of data security and privacy since data can be held within the LAN. Fog nodes are better suited for deploying DNNs rather than training since they are designed to store data from local devices, which are incomplete compared with

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

those on the cloud. Nevertheless, model training can still be scheduled on fog nodes by leveraging FL [60].

3) Edge Computing Layer: The term of edge computing is interchangeable with fog computing in some literature [19], [21] or denotes a broader concept that the fog can be treated as a part of it [22]. Nevertheless, we treat them as different concepts for clarity in this article. Specifically, we distinguish them based on their locations within the LAN, i.e., fog computing at the network side and edge computing at the thing side. In this sense, edge computing refers to deploying computing capacity on edge devices in proximity to sensors and actuators. A great advantage of edge computing over fog and cloud computing is the reduction of latency and network bandwidth since it can process data into compact structured information on-site before transmission, which is especially useful for AIoT applications using multimedia sensors. However, due to its limited computation capacity, only lightweight DNNs can run on edge devices. Therefore, research topics, including neural network architecture design or search for mobile setting and network pruning/compression/quantization, have attracted increasing attention recently.

In practice, it is common to deploy multiple different models into cloud platforms, fog nodes, and edge devices in an AIoT system to build an intelligent hybrid computing architecture. By intelligently offloading part of the computation workload from edge devices to the fog nodes and cloud, it is expected to achieve low latency while leveraging deep learning capacities for processing massive amounts of data. For example, a lightweight model can be deployed on edge devices to detect cars in a video stream. It can act as a trigger to transmit keyframes to fog nodes or the cloud for further processing.

B. Hardware and Software

1) Hardware: While GPU is initially developed for accelerating image rendering on display devices, the general-purpose GPU turns the massive computational power of its shader pipeline into general-purpose computing power (e.g., for massive vector operations), which has sparked the deep learning revolution along with DNN and big data. Lots of operations in the neural network, such as convolution can be computed in parallel on GPU, significantly reducing the training and inference time. Recently, an application-specific integrated circuit (ASIC) named TPU is designed by Google specifically for neural network machine learning. Besides, field-programmable gate arrays (FPGAs) have also been used for DNN acceleration due to their low-power consumption and high throughput. Several machine learning processors have also been developed for fog and edge computing, e.g., Google Edge TPU and NVIDIA Jetson Nano.

2) Software: Researchers and engineers must design, implement, train, and deploy DNNs easily and quickly. To this end, different open-source deep learning frameworks have been developed, from the beginners, such as Caffe[1] and MatConvNet[2] to the popular TensorFlow[3] and PyTorch.[4] MatConvNet is a MATLAB toolbox for implementing convolutional neural networks (CNNs). Caffe is implemented in C++ with Python and MATLAB interfaces and well known for its speed but does not support distributed computation and mobile deployment. Caffe2 improves it accordingly, which has been later merged into PyTorch. The features, such as dynamic computation graphs and automatic computation of gradients in TensorFlow and PyTorch, have made them easy to use and popular. They also support for deploying models into mobile devices by enabling model compression/quantization and hardware acceleration. Porting models among different frameworks is necessary and useful. The open neural network exchange (ONNX)[5] offers this feature by defining an open format built to represent machine learning models, which has been supported by TensorFlow and Pytorch. There are other deep learning frameworks, such as MXNet,[6] Theano,[7] PaddlePaddle,[8] and neural network inference computing framework for mobile devices such as ncnn.[9]

III. Progress Review of AI for IoT

In this section, we comprehensively review the progress of enabling AI technologies for AIoT applications, especially deep learning. We conduct the survey by applying a new taxonomy, i.e., how deep learning improves the ability of AIoT systems for perceiving, learning, reasoning, and behaving. To prevent it from being a survey on deep learning, we carefully select the topics and technologies that are closely related to and useful for various AIoT applications. Moreover, we only outline the trend of the research progress and highlight state-of-the-art technologies rather than diving into the details. We specifically discuss their potentials for AIoT applications. We hope this survey can draw an overall picture of AI technologies for AIoT and provide insights into their utility.

A. Perceiving

Empowering things with the perceiving ability, i.e., understanding the environment using various sensors, is fundamental for AIoT systems. In this part, we will focus on several related topics as diagrammed in Fig. 4.

First, we present a review of the progress in generic scene understanding, including image classification, object detection, and tracking, semantic segmentation, and text spotting.

1) Image Classification: Image classification refers to recognizing the category of an image. Classical machine learning methods based on hand-crafted features have been surpassed by DNNs [26] on large-scale benchmark data sets such as ImageNet [30], sparking a wave of research on the architecture of DNNs. From AlexNet [26] to ResNet [64], more

[1] https://github.com/BVLC/caffe

[2] https://github.com/vlfeat/matconvnet

[3] https://github.com/tensorflow/tensorflow

[4] https://github.com/pytorch/pytorch

[5] https://github.com/onnx/onnx

[6] https://github.com/apache/incubator-mxnet

[7] https://github.com/Theano/Theano

[8] https://github.com/PaddlePaddle

[9] https://github.com/Tencent/ncnn

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

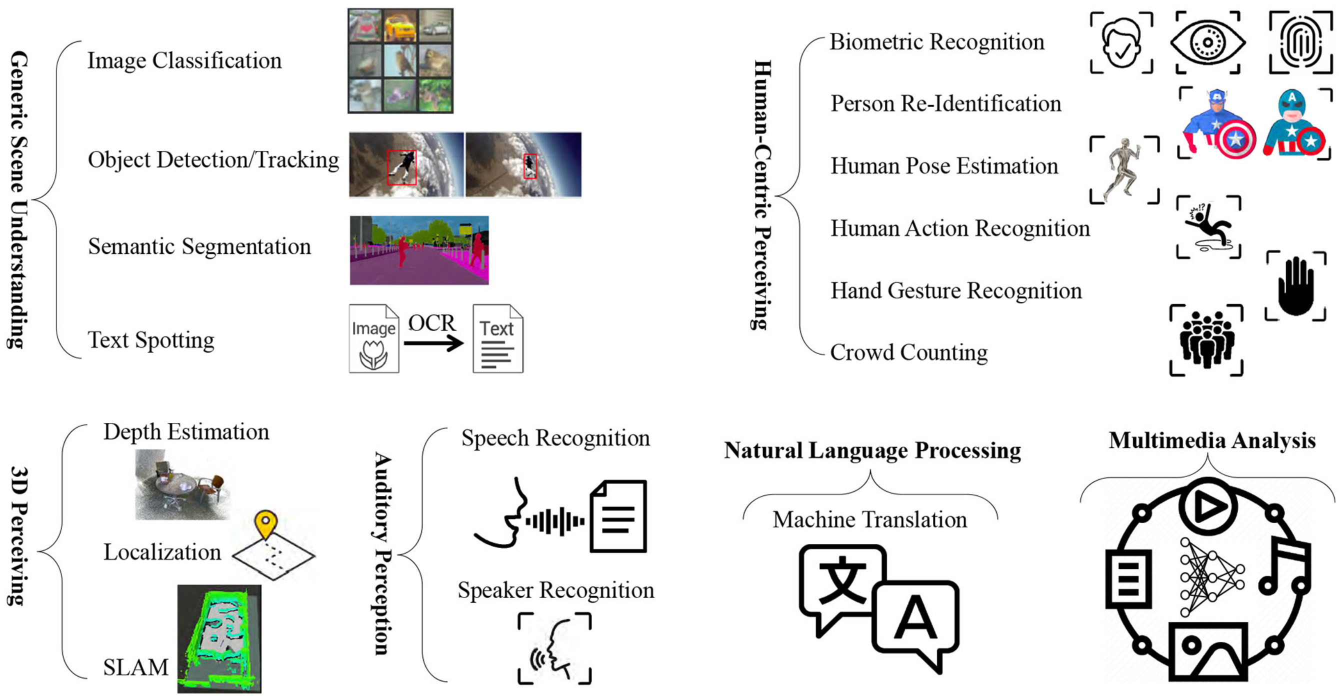

Fig. 4. Diagram of the perceiving-related topics in AIoT.

and more advanced network architectures have been devised by leveraging stacked 3 × 3 convolutional layer for reducing network parameters and increasing network depth, 1 × 1 convolutional layer for feature dimension reduction, residual connections for preventing gradient vanishing and increasing network capacity, and dense connections for reusing features from previous layers as shown in Fig. 5. A brief summary of representative deep CNNs is listed in Table II. As can be seen, with the increase of network depth and the number of parameters, the representation capacity also increases, leading to lower top1 classification error on the ImageNet data set. Besides, the architecture of the network matters. Even with fewer model parameters and computational complexity, the recently proposed networks, such as ResNet and DenseNet, outperform previous ones such as VGGNet. Lightweight networks are appealing to AIoT applications where DNNs are deployed on edge devices. Recently, some computationally efficient networks such as MobileNet have been proposed by leveraging depthwise convolutions [65], pointwise convolutions [66], or binary operations [67]. Besides, network compression, such as pruning and quantization can be used to obtain lightweight models from heavy models, which will be reviewed in Section III-A17. Image recognition can be very useful in many AIoT applications, such as smart education tools or toys, which help and teach children to explore the world with cameras. Besides, some popular applications in smartphones also benefit from the advances in this area for recognizing flowers and birds, and food items and calories.

2) Object Detection: Generic object detection refers to recognizing the category and location of an object, which is used as a prepositive step for many downstream tasks, including face recognition, person reidentification, pose estimation, behavior analysis, and human–machine interaction (HMI). The methods for object detection from images have been revolutionized by DNNs. State-of-the-art methods can be categorized into two groups: 1) two-stage methods and 2) one-stage methods. The former follows a typical "proposal→detection" paradigm [27], while the latter directly evaluates all the potential object candidates and outputs the detection results [68]. Recently, one-stage anchor-free detectors have been proposed by representing object location using points or regions rather than anchors [69], achieving a better tradeoff between speed and accuracy, which is appealing to AIoT applications that require onboard detection. Detection of specific category of objects, such as pedestrian, car, traffic sign, and the license plate has been widely studied, which are useful for improving the perceiving ability of AIoT systems for traffic and public safety surveillance and autonomous driving [70]. Besides, object detection is a crucial technique for video data structuring in many AIoT systems using visual sensors, which aims to extract and organize compact structured semantic information from video data for further retrieval, verification, statistics, and analysis at low transmission, storage, and computation cost.

3) Object Tracking: Classical object tracking methods include generative and discriminative methods, where the former ones try to search the most similar regions to the target and the latter ones leverage both foreground target and background context information to train an online discriminative classifier [71]. Later, different deep learning methods have been proposed to improve the classical methods by learning multiresolution deep features [72], end-to-end representation learning [73], and leveraging siamese networks [74]. Object trackers usually run much faster than object detectors, which can be deployed on edge devices of AIoT applications, such as video surveillance and autonomous driving for object trajectory generation and motion prediction. One

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

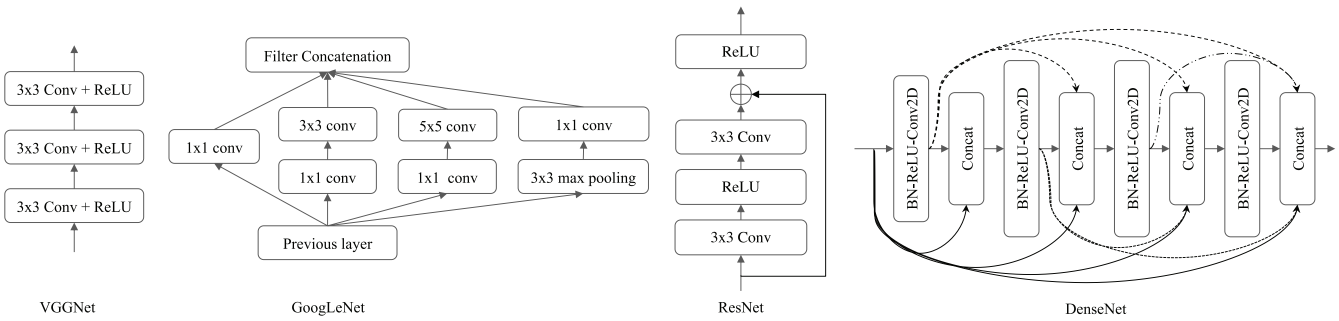

Fig. 5. Basic blocks of representative deep CNNs.

TABLE II
SUMMARY OF REPRESENTATIVE DEEP CNNS. PARAM.: NUMBER OF PARAMETERS; COMP.: COMPUTATIONAL COMPLEXITY (MACS)

Network	Year	Depth	Param. (M)	Comp. (G)	top1-err
AlexNet	2012	8	61.10	0.72	43.45
VGGNet	2014	11	132.86	7.63	30.98
		13	133.05	11.34	30.07
		16	138.36	15.5	28.41
		19	143.67	19.67	27.62
GoogLeNet	2014	22	6.62	1.51	30.22
Inception v3	2015	48	27.16	2.85	22.55
ResNet	2015	18	11.69	1.82	30.24
		34	21.80	3.68	26.70
		50	25.56	4.12	23.85
		101	44.55	7.85	22.63
		152	60.19	11.58	21.69
DenseNet	2016	121	7.98	2.88	25.35
		169	14.15	3.42	24.00
		201	20.01	4.37	22.80
		161	28.68	7.82	22.35

possible solution is to leverage the hybrid computation architecture (see Section II-A) by deploying object tracker on edge devices while deploying object detectors on the fog nodes or cloud, i.e., tracking across all frames while detecting only on keyframes. In this way, only keyframes and the compact structured detection results should be transmitted via the network, thereby reducing the network bandwidth and processing latency.

4) Semantic Segmentation: Semantic segmentation refers to predicting pixel-level category label for an image. The fully CNN with an encoder–decoder structure has become the *de facto* paradigm for semantic segmentation [75], [76], since it can learn discriminative and multiresolution features through cascaded convolution blocks while preserving spatial correspondence. Many deep models have been proposed to improve the representation capacity and prediction accuracy from the following three aspects: 1) context embedding; 2) resolution enlarging; and 3) boundary refinement. Efficient modules are proposed to exploit context information and learn more representative feature representations, such as global context pooling module in the ParseNet [77], atrous spatial pyramid pooling in the DeepLab models [28], and the pyramid pooling module in the PSPNet [78]. Enlarging the resolution of feature maps is beneficial for improving prediction accuracy, especially for small objects. Typical techniques include using the deconvolutional layer, unpooling layer, and dilated convolutional layer. Boundary refinement aims to obtain sharp boundaries between different categories in the segmentation map, which can be achieved by using conditional random field as the postprocessing technique on the predicted probability maps [28].

There are two research topics related to semantic segmentation, i.e., instance segmentation and panoptic segmentation. Instance segmentation refers to detecting foreground objects as well as obtaining their masks. A well-known baseline model is Mask R-CNN, which adopts an extra branch for object mask prediction in parallel with the existing one for bounding box regression [79]. Its performance can be improved further by exploiting and enhancing the feature hierarchy of deep convolutional networks [80], employing nonlocal attention [81], and leveraging the reciprocal relationship between detection and segmentation via hybrid task cascade [82]. Panoptic segmentation refers to simultaneously segmenting the masks of foreground objects as well as background stuff [83], i.e., unifying both the semantic segmentation and instance segmentation tasks. A simple but strong baseline model is proposed in [84], which adds a semantic segmentation branch into the Mask R-CNN framework and uses a shared feature pyramid network backbone [80]. Semantic segmentation in many subareas, such as medical image segmentation [75], road detection [85], and human parsing [86], is useful in various AIoT applications. For example, it can be used to recognize the dense pixel-level drivable area and traffic participants, such as cars and pedestrians, which can be further combined with 3-D measure information to get a comprehensive understanding of the driving context and make smart driving decisions accordingly. Moreover, obtaining the foreground mask or body parts matters for many AIoT applications, e.g., video editing for entertainment and computational advertising, virtual try-on, and augmented/virtual reality (AR/VR). Besides, the structured semantic mask is also useful for semantic-aware efficient and adaptive video coding.

5) Text Spotting: Text spotting is a composite task, including text detection and recognition. Although text detection is related to generic object detection, it is a different and challenging problem: 1) while generic objects have regular shapes, text may be in variable length and shape depending on the number of characters and their orientation and 2) the appearance of the same text may change significantly due to fonts, styles, as well as background context. Deep learning

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE III
REPRESENTATIVE BENCHMARK DATA SETS IN GENERIC SCENE UNDERSTANDING. BBOX: BOUNDING BOX; MASK: PIXEL-LEVEL SEMANTIC MASK

Area	Dataset	Link	Volume	Label Type
Image Classification	ImageNet	http://www.image-net.org/	1.2M	Category, BBox
	CIFAR-10/-100	https://www.cs.utoronto.ca/~kriz/cifar.html	6k	Category
	Caltech-UCSD Birds	http://www.vision.caltech.edu/visipedia/CUB-200-2011.html	11,788	Category, Attributes, BBox
	Caltech 256	http://www.vision.caltech.edu/Image_Datasets/Caltech256/	30,607	Category
Object Detection	COCO	https://cocodataset.org/#detection-2020	200k	Category, BBox, Mask
	Pascal VOC	http://host.robots.ox.ac.uk/pascal/VOC/	10k	Category, BBox, Mask
Object Tracking	MOT	https://motchallenge.net/	22	BBox
	KITTI-Tracking	http://www.cvlibs.net/datasets/kitti/eval_tracking.php	50	3D BBox
	UA-DETRAC	http://detrac-db.rit.albany.edu/	140k	BBox
Semantic Segmentation	Cityscape	https://www.cityscapes-dataset.com/	25k	Mask
	ADE20K	https://groups.csail.mit.edu/vision/datasets/ADE20K/	22,210	Category, Attributes, Mask
	PASCAL-Context	https://cs.stanford.edu/~roozbeh/pascal-context/	19,740	Mask
Text Spotting	Total-text	https://github.com/cs-chan/Total-Text-Dataset	1,555	Polygon Box, Text
	SCUT-CTW1500	https://github.com/Yuliang-Liu/Curve-Text-Detector	1,500	BBox, Text
	LSVT	https://ai.baidu.com/broad/introduction?dataset=lsvt	450k	Binary Mask, Text

has advanced this area by learning more representative feature [87], devising better representation of text proposals [88], and using large-scale synthetic data set [89]. Recently, end-to-end modeling of text detection and recognition has achieved impressive performance [90], [91]. Each subtask can benefit from the other by leveraging more supervisory signals and learning a shared feature representation. Moreover, rather than recognizing text at the character level, recognizing text at the word or sentence level can benefit from the word dictionary and language model. Specifically, the idea of sequence-to-sequence modeling and connectionist temporal classification (CTC) [92] from the areas of speech recognition and machine translation (MT) has also been explored. Since the text is very common in real-world scenes, e.g., traffic sign, nameplate, and information board, text spotting can serve as a useful tool in many AIoT applications for "reading" text information from scene images, e.g., live camera translator for education, reading assistant for the visually impaired [93], optical character recognition (OCR) for automatic document analysis, and store nameplate recognition for self-localization and navigation. The representative benchmark data sets in the areas related to generic scene understanding are summarized in Table III.

Next, we present a review of the progress in human-centric perceiving, including biometric recognition, such as face/fingerprint/iris recognition, person reidentification, pose/gesture/action estimation, and crowd density estimation.

6) Biometric Recognition: Biometric recognition, based on face, fingerprint, or iris, is a long-standing research topic. We first review the progress in face recognition. There are usually four key stages in a face recognition system, i.e., face detection, face alignment, face representation, and face classification/verification. Face detection, as a specific sub-area of object detection, benefits from the recent success of deep learning in generic object detection. Nevertheless, special effort should be made to address the following challenges, i.e., vast scale variance, severe imbalance of positive and negative proposals, profile and front face, occlusion, and motion blur. One of the most famous classical methods is the Viola–Jones algorithm, which sets up the fundamental face detection framework [94]. The idea of using cascade classifiers inspires many deep learning methods, such as cascade CNN [95]. Recently, jointly modeling face detection with other auxiliary tasks, including face alignment, pose estimation, and gender classification, can achieve improved performance, owing to the extra abundant supervisory signals for learning a shared discriminative feature representation [96], [97]. Note that the all-in-one model is appealing to some AIoT application where multiple structured facial information could be extracted.

Face alignment, also known as (a.k.a.), facial landmark detection aims to detect facial landmarks from a face image, which is useful for front face alignment and face recognition. Typically, face facial landmark detectors are trained and deployed in a cascade manner that a shape increment is learned and used to update the current estimate at each level [98]. Face landmark detectors are usually lightweight and run very fast, which are very useful for latency-sensitive AIoT applications.

For face recognition, significant progress has been achieved in the last decade, mainly owing to deep representation learning and metric learning. The milestone work in [25] proposes to learn discriminative deep bottleneck features using classification and verification losses. Nevertheless, they face a challenge to scale to orders of magnitude larger data sets with more identities. To address this issue, a representation learning method using triplet loss is proposed to directly learn discriminative and compact face embedding [99]. Face recognition is one of the most widely used perceiving techniques for identity verification and access control in various AIoT applications, e.g., smart cities and smart homes. Associating the facial identity with one's accounts can create vast business value, e.g., mobile payment, membership development and promotion, and fast track in smart retail. Regarding the number of people to be recognized and privacy concerns, either offline or online solutions can be used, where models are deployed on edge devices, fog nodes, or cloud centers [100], [101]. A research topic related to practical face recognition applications is liveness detection and spoof detection. Different methods have been proposed based on action imitation, speech collaboration, and multimodal sensors [102].

In addition to face recognition, iris, fingerprint, and palmprint recognition have also been studied for a long period and are widely used in practical AIoT applications. Compared with fingerprint, palmprint has abundant features and can be

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

captured using common built-in cameras of mobile phones rather than sensitive sensors. Typically, a palmprint recognition system is composed of a palmprint image acquisition system, a palmprint region of interest (ROI) extraction module, a feature extraction module, and a feature matching module for recognition or verification. Both hand-crafted features, such as line features, orientation-based features, and the orthogonal line ordinal feature and deep learning-based feature representation have been studied in the literature [103], [104]. For example, Zhang *et al.* [105] proposed a novel device to capture palmprint images in a contactless way, which can be used for access control, aviation security, and e-banking. It uses blockwise statistics of competitive code as features and the collaborative representation-based framework for classification. Besides, a DCNN-based palmprint verification system named DeepMPV is proposed for mobile payment in [104]. It first extracts the palmprint ROI by using pretrained detectors and then trains a siamese network to match palmprints. Recently, Amazon has announced its new payment system called Amazon One, which is a fast, convenient, and contactless way for people to use their palms to make payments based on palmprint recognition. In practice, the choice of a specific biometric recognition solution depends on sensors, usage scenarios, latency, and power consumption. Although biometric recognition offers great utility, the concerns about data security and privacy have to be carefully addressed in practical AIoT systems.

7) Person Reidentification: Person reidentification, as a subarea of image retrieval, refers to recognizing an individual captured in disjoint camera views. In contrast to face recognition in a controlled environment, person reidentification is more challenging due to the variations in the uncontrolled environment, e.g., viewpoint, resolution, clothing, and background context. To address these challenges, different methods have been proposed [110], including deep metric learning based on various losses, integration of local features and context, multitask learning based on extra attribute annotations, and using human pose and parsing mask as guidance. Recently, generative adversarial networks (GANs) have been used to generate style-transferred images for bridging the domain gap between different data sets [111]. Person reidentification has vast potential for AIoT applications, such as smart security in an uncontrolled and noncontact environment, where other biometric recognition techniques are not applicable. Although extra efforts are needed to build practical person reidentification systems, one can leverage the idea of human-in-the-loop AI to achieve high performance with low labor effort. For example, the person reidentification model can be used for initial proposal ranking and filtering, then human experts are involved to make final decisions.

8) Human Pose Estimation and Gesture/Action Recognition: Human pose estimation, a.k.a. human keypoint detection refers to detecting body joints from a single image. There are two groups of human pose estimation methods, i.e., top-down methods and bottom-up methods. The former consists of two stages, including person detection and keypoint detection, while the latter directly detects all keypoints from the image and associates them with corresponding person instances. Although top-down methods still dominate the leaderboard of public benchmark data sets such as MS COCO,[10] they are usually slower than bottom-up methods [112]. Recent progress in this area can be summarized in the following aspects.

1) Learning better feature representation from stronger backbone network, multiscale feature fusion, or context modeling [113].
2) Effective training strategy, including online hard keypoint mining, hard negative person detection mining, and harvesting extra data [107].
3) Subpixel representation or postprocessing techniques [107], [114]. Recently, dealing with pose estimation in crowd scenes with severe occlusions also attracts much attention. The other related topic is 3-D human pose estimation from a single image or multiview images [115], aiming to estimate the 3-D coordinate of each keypoint rather than the 2-D coordinate on the image plane.

Once we detect the human keypoints for each frame given a video clip, the skeleton sequence for each person instance can be obtained, from which we can recognize the action. This process is known as skeleton-based action recognition. To model the long-term temporal dependencies and dynamics, as well as spatial structures within the skeleton sequence, different neural networks have been exploited for action recognition, such as the deep recurrent neural network (RNN) [116], CNN [117], and deep graph convolutional networks (GCNs) [118]. Besides, since some joints may be more relevant to specific actions than others, attention mechanism has been used to automatically discover informative joints and emphasize their importance for action recognition [119]. Estimation of human pose and recognition of action can be very useful in many real-world AIoT scenarios, such as rehabilitation exercises monitoring and assessment [120], dangerous behavior monitoring [121], and HMI.

Hand gesture recognition is also a hot research topic and has many practical applications, such as HMI and sign language recognition. Different sensors can be used in AIoT systems for gesture recognition, such as millimeter-wave radar and visual sensors, such as RGB camera, depth camera, and event camera [122]–[124]. Nevertheless, due to the prevalence of cameras and great progress in deep learning and computer vision, visual hand gesture recognition has the vast potential, which can be categorized into two groups, i.e., static ones and dynamic ones. The former aims to match the gesture in a single image to some predefined gestures, while the latter tries to recognize the dynamic gesture from an image sequence, which is more useful. Usually, there are three phases in dynamic hand gesture recognition, i.e., hand detection, hand tracking, and gesture recognition. While hand detection and tracking can benefit from recent progress in generic object detection and tracking as described in Sections III-A2 and III-A3, hand gesture recognition can also borrow useful ideas from the area of action recognition, e.g., exploiting RNN and 3-D CNN to capture the gesture dynamics from image sequences. Hand gesture

[10] http://cocodataset.org/index.htm#keypoints-leaderboard

[11] https://www.youtube.com/watch?v=__eLCXUKtec

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

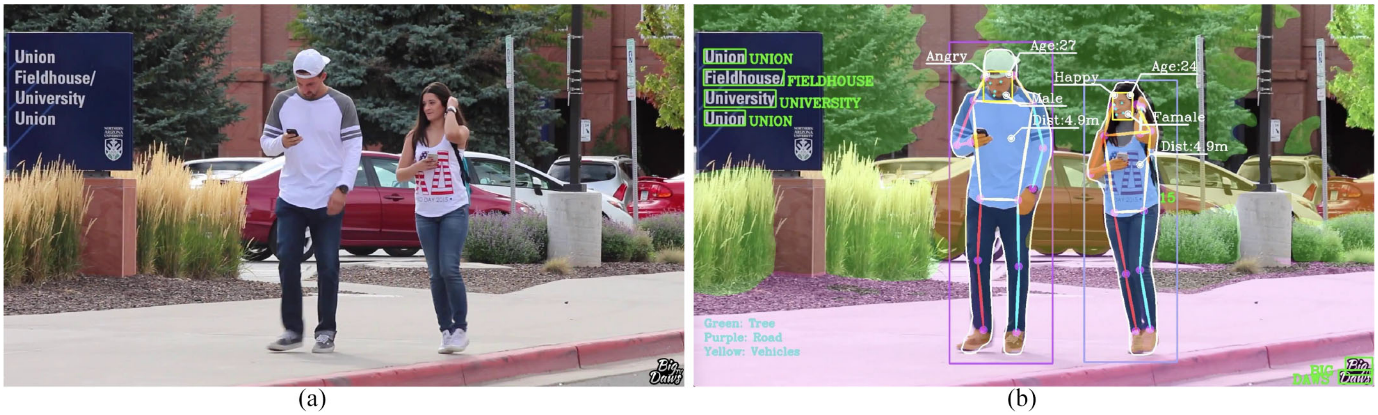

Fig. 6. Demonstration of AI techniques for generic scene understanding, human-centric perceiving, and 3-D perceiving. (a) Frame from the video “Walking Next to People”[11]. (b) Processed result by using different perceiving methods, i.e., semantic segmentation [85]; object detection [106]; text spotting [91]; human parsing [86]; human pose estimation [107]; face detection, alignment, and facial attribute analysis [96], [108]; and depth estimation [109].

TABLE IV
REPRESENTATIVE BENCHMARK DATA SETS IN HUMAN-CENTRIC PERCEIVING. ID: IDENTITY; BBOX: BOUNDING BOX

Area	Dataset	Link	Volume	Label Type
Face Recognition	FFHQ	https://github.com/NVlabs/ffhq-dataset	70k	ID
	FDDB	http://vis-www.cs.umass.edu/fddb/	2,845	ID, BBox
	YouTube Faces DB	https://www.cs.tau.ac.il/~wolf/ytfaces/	3,425	ID, BBox
Fingerprint Recognition	FVC2000	http://bias.csr.unibo.it/fvc2000/databases.asp	3,520	ID
	LivDet Databases	http://livdet.org/registration.php	11k	ID
Iris Recognition	LivDet Databases	http://livdet.org/registration.php	7,223	ID
	IrisDisease	http://zbum.ia.pw.edu.pl/AGREEMENTS/IrisDisease-v2_1.pdf	2,996	ID
Person Re-ID	Market-1501	http://zheng-lab.cecs.anu.edu.au/Project/project_reid.html	1,501	ID, BBox
	DukeMTMC-ReID	https://github.com/sxzrt/DukeMTMC-reID_evaluation	1,404	ID, BBox
	CUHK03	https://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html	1,360	ID, BBox
Pose Estimation	COCO	https://cocodataset.org/#keypoints-2020	200k	Keypoints
	MPII	http://human-pose.mpi-inf.mpg.de/	25k	Keypoints
	DensePose-COCO	http://densepose.org/	50k	Keypoints
Gesture Recognition	DVS128	https://www.research.ibm.com/dvsgesture/	1,342	Category
	MS-ASL	https://www.microsoft.com/en-us/research/project/ms-asl/	25k	Category
Action Recognition	UCF101	https://www.crcv.ucf.edu/data/UCF101.php	13,320	Category
	ActivityNet	http://activity-net.org/	19,994	Category
Crowd Counting	NWPU-Crowd	https://gjy3035.github.io/NWPU-Crowd-Sample-Code/	5,109	Dots, BBox
	JHU-CROWD++	http://www.crowd-counting.com/	4,372	Dots, BBox
	UCF-QNRF	https://www.crcv.ucf.edu/data/ucf-qnrf/	1,535	Dots

recognition can be very useful for interactions with things in AIoT systems, e.g., noncontact control of television and car infotainment system, and communication with the speech and hearing impaired [125].

9) Crowd Counting: In the video surveillance scenario, it is necessary to count the crowd in both indoor and outdoor areas and prevent crowd congestion and accident. For practical AIoT applications with crowd counting ability, WI-FI, Bluetooth, and camera-based solutions have been proposed by estimating the connections between smartphones and WI-FI access points or Bluetooth beacons [126] or estimating the crowd density of a crowd image [127]. Although counting the detected faces or heads in a crowd image can be used for crowd counting intuitively, the person instance in a crowd image is always in relatively low resolution and blurry, which limits the performance of the detection model. Besides, detecting a vast amount of persons in a single shot is computationally inefficient. Therefore, most CNN-based methods directly regress the crowd density map, in which the ground truth is constructed by placing Gaussian density maps at the head regions. Since it is costly to collect and annotate crowd images, synthetic data sets can be used and have demonstrated its value for this task, i.e., either being used in the pretraining-finetuning scheme or by DA [128]. Despite the progress in this area, more efforts are needed to address real-world challenges for practical AIoT applications, e.g., designing lightweight and computational efficient crowd counting models, simultaneous crowd counting and crowd flow estimation, and integration of multimodal sensors for more accurate crowd counting. The representative benchmark data sets in the aforementioned research areas related to human-centric perceiving are summarized in Table IV.

In the following, we review several topics related to 3-D perceiving, including depth estimation, localization, and simultaneous localization and mapping (SLAM).

10) Depth Estimation/Localization/SLAM: Estimating depth using cameras is a long-standing research topic [109], [129]–[131]. In real-world AIoT applications, there can be several configurations, such as the monocular camera, stereo camera, and multiview camera system. Recently, depth estimation from monocular video together with camera pose

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE V
Representative Benchmark Data Sets in 3-D Perceiving

Area	Dataset	Link	Volume	Label Type
Depth Estimation	KITTI	http://www.cvlibs.net/datasets/kitti/eval_depth.php	93k	Depth Maps
	NYU Depth Dataset V2	https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html	1,449	Depth Maps
	Make3D	http://make3d.cs.cornell.edu/data.html#object	534	Depth Maps
SLAM	KITTI	http://www.cvlibs.net/datasets/kitti/eval_odometry.php	22	Poses
	EUROC MAV Dataset	https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets	12	Poses
	TUM Visual-Inertial	https://vision.in.tum.de/data/datasets/visual-inertial-dataset	43	Poses

estimation has attracted a lot of attention. In contrast to traditional matching and optimization-based methods, current research on this topic mainly focuses on deep learning in an unsupervised or self-supervised way [132]. Nevertheless, they construct the self-supervisory signals based on the reprojection photometric loss with respect to depth and camera pose derived from the well-defined multiview geometry, which is similar to the matching error or photometric error terms in the traditional optimization objective. Although CNN has powerful representation capacity, special effort has to be made to address the challenges, including occlusions and dynamic objects, as well as the scale issue (per-frame ambiguity and temporally inconsistent).

The aforementioned camera pose estimation is also related to visual odometry (VO) and visual-inertial odometry (VIO) [133], [134], which aim to calculate sequential camera poses of an agent based on the camera and inertial measurement unit (IMU) sensors. VO and VIO are always used at the front end in a SLAM system, where the back end refers to the nonlinear optimization of the pose graph, aiming to obtain globally consistent and drift-free pose estimation results. In traditional methods such as ORB-SLAM [135], the front end and back end are two separate modules. Recently, a differentiable architecture named neural graph optimizer is proposed for global pose graph optimization [136]. Together with a local pose estimation model, it achieves a complete end-to-end neural network solution for SLAM.

Depth estimation, pose estimation, VO/VIO, and SLAM constitute the important 3-D perceiving ability of AIoT, which could be very useful in smart transportation [137], smart industry [138], smart agriculture [139]–[141], smart cities, and homes [142]–[144]. For example, deploying multiple cameras at different viewpoints, one can construct a multiview visual system for depth estimation and object or scene 3-D reconstruction. In the autonomous driving scenario, depth estimation can be integrated into the object detection module and road detection module for forward collision warning. Besides, SLAM can be used for lane departure warning, lane keeping, high-precision map construction, and update [137]. Other use cases may include self-localization and navigation for the agricultural robot, sweeper robot, service robot, and unmanned aerial vehicle (UAV) [138]–[140]. The representative benchmark data sets in the areas of 3-D perceiving are summarized in Table V. Fig. 6 presents an example of using AI techniques for generic scene understanding, human-centric perceiving, and 3-D perceiving.

Due to sensor quality and imaging conditions, the captured image may need to be preprocessed to enhance illumination, increase contrast, and rectify distortions before being used in the aforementioned visual perception tasks. In the following, we briefly review the recent progress in the area of image enhancement as well as image rectification and stitching.

11) Image Enhancement: Image enhancement is a task-oriented task that refers to enhancing the specific property of a given image, such as illumination, contrast, and sharpness. Images captured in a low-light environment are in low visibility and difficult to see details due to insufficient incident light or underexposure. An image can be decomposed into the reflectance map and illumination map based on the Retinex theory [145]. Then, the illumination map can be enhanced, thereby balancing the overall illumination of the original low-light image. However, it is a typical ill-posed problem to obtain the reflectance and illumination from a single image. To address this issue, different prior-based or learning-based low-light enhancement methods have been proposed in recent literature. For example, LIME leverages a structure prior of the illumination map to refine the initial estimation [146] while a piecewise smoothness constraint is used in [147]. Since low-light images usually contain noises that will be amplified after enhancement, some robust Retinex models have been proposed to account for noise and estimate reflectance, illumination, and noise simultaneously [147], [148].

Images captured in a haze environment are in low contrast due to the haze attenuation and scattering effects. Recovering the clear image from a single hazy input is also an ill-posed problem, which can be addressed by both prior-based and learning-based methods [149]–[151]. For example, He *et al.* propose a dark channel prior to estimate the haze transmission efficiently [149]. Cai *et al.* [150] proposed the first deep CNN model for image dehazing, which outperforms traditional prior-based methods by leveraging the powerful representation capacity of CNNs. Recently, Zhao *et al.* [152] proposed a real-world benchmark to evaluate dehazing methods according to visibility and realness. When images are captured in the low-light and haze environment, it comes to the more challenging case, i.e., nighttime image dehazing. Similarly, some methods have been proposed based on either statistical priors or deep learning, e.g., maximum reflectance prior [153], glow separation [154], and ND-Net [155].

12) Image Rectification and Stitching: Wide field-of-view (FOV) cameras, such as fisheye cameras, have been widely used in different AIoT applications, e.g., video surveillance and autonomous driving since they can capture a larger scene area than narrow FOV cameras. However, the captured images contain distortions since they break the perspective transformation assumption. To facilitate downstream tasks, the distorted image should be rectified beforehand. The rectification

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
REPRESENTATIVE BENCHMARK DATA SETS IN AUDITORY PERCEPTION. ID: IDENTITY; BBOX: BOUNDING BOX OF SPEAKERS

Area	Dataset	Link	Volume	Label Type
Speech Recognition	CHiME-3	http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/	14,658 utterances	Text, ID
	VoxCeleb	http://www.robots.ox.ac.uk/~vgg/data/voxceleb/	1M utterances	Text, BBox, Attributes
	TIMIT APCSC	https://catalog.ldc.upenn.edu/LDC93S1	630 speakers	Text, ID
Speaker Verification	VoxCeleb	http://www.robots.ox.ac.uk/~vgg/data/voxceleb/	1M utterances	Text, BBox, Attributes
	TIMIT APCSC	https://catalog.ldc.upenn.edu/LDC93S1	630 speakers	Text, ID
	Common Voice	https://commonvoice.mozilla.org/en/datasets	61,528 voices	ID, Attributes

methods can be categorized as camera calibration-based methods and distortion model-based methods. The former calibrate the intrinsic and extrinsic parameters of cameras and then rectify the distorted image by following perspective transformation. The widely used calibration method is proposed by Zhang [156] based on planar patterns at a few different orientations, where radial lens distortion is modeled. The latter directly estimate the distortion parameters of a distortion model and map the distorted image to the rectified image based on it accordingly. Different geometric cues have been exploited for formulating the optimization constraints in optimization-based methods or loss functions in learning-based methods [157], such as lines and vanishing points. Given two or more fisheye cameras with calibrated parameters, a panorama image can be obtained from their images by image stitching. For example, Liu *et al.* [158] proposed an online camera pose optimization method for the surround view system, which is composed of several fisheye cameras around the vehicle. The surround view system can capture a 360° view around the vehicle, which is useful in IoV for the advanced driver assistant system and crowd-sourcing high-precision map update.

In addition to the above reviewed visual perception methods, we then present a brief review of auditory perception, specifically speech perception. We include two topics in the following part, i.e., speech recognition and speaker verification.

13) Speech Recognition: Speech recognition, a.k.a. automatic speech recognition (ASR), is a subfield of computational linguistics that aims to recognizing and translating spoken language into text automatically. Traditional ASR models are based on hand-crafted features like cepstral coefficient and hidden Markov model (HMM) [159], which have been revolutionized by the DNN for end-to-end modeling without the need of domain knowledge for feature engineering, HMM design, as well as explicit dependency assumption. For example, RNN, especially long short-term memory (LSTM), is used to model the long-range dependencies in the speech sequence and decode the text sequentially [24]. However, for one thing, extra effort is needed to presegment training sequences so that the classification loss can be calculated at each point in the sequence independently, for another, RNN processes data in a sequential manner, which is parallel-unfriendly. To address the first issue, the CTC is proposed by directly maximizing the probabilities of the correct label sequence in a differentiable way [92]. To mitigate the other issue, the transformer architecture is devised using scaled dot-product attention and multihead attention [160].

Recently, real-time ASR systems have been developed either using on-device computing or in a cloud-assisted manner [161], [162]. ASR is very useful in many AIoT applications since speech is one of the most important noncontact interaction modes. For example, ASR can be used in the smart input system [163], automatic transcription system, smart voice assistant [164], [165], and computer-assisted speech rehabilitation and language teaching [166]. The computing paradigm could be on-device edge computing (e.g., offline mode of smart voice assistant), fog computing with a powerful computing device and sound pickup system (e.g., automatic transcription system for conferences), as well as the cloud computing with acceptable latency (e.g., online mode of smart voice assistant). Besides, some related techniques for music and humming recognition and birdsong recognition could be useful to empower AIoT systems for music retrieval and recommendation and wild bird conservation.

14) Speaker Recognition: While face recognition aims to recognize an individual through one's unique facial patterns, speaker recognition achieves the same goal using one's voice characteristics. A speaker recognition system is composed of three modules, i.e., speech acquisition and production, feature representation and selection, and pattern matching and classification [167]. Previously, speaker recognition methods are dominated by the *i*-vector representation and probabilistic linear discriminant analysis framework [168], where *i*-vector refers to extracting low-dimensional speaker embeddings from sufficient statistics. Recently, several end-to-end deep speaker recognition models have been devised [169], achieving better performance than *i*-vector baselines. Similar to the techniques in face recognition, speaker recognition also benefits from the advances in deep metric learning, i.e., leveraging the contrastive loss or triplet loss to learn discriminative speaker embeddings from large-scale data sets. Speaker recognition is one of the important means for identity identification, which has many applications in various AIoT domains, for example, automatic transcription system for multiperson meetings, personalized recommendation by smart voice assistants [170], and audio forensics [171]. Besides, speaker recognition can be integrated with face recognition for access control. The representative benchmark data sets in the areas related to auditory perception are summarized in Table VI.

Next, we present a review of the progress in natural language processing (taking MT as an example) and multimedia and multimodal analysis.

15) Machine Translation: MT is also a subfield of computational linguistics that aims to translate text from one language to another automatically. Neural MT (NMT) based

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
REPRESENTATIVE BENCHMARK DATA SETS IN NATURAL IMAGE PROCESSING AND MULTIMEDIA ANALYSIS. I: IMAGE; T: TEXT; V: VIDEO; AND A: AUDIO

Area	Dataset	Link	Volume	Label Type
Machine Translation	WMT	http://www.statmt.org/wmt14/translation-task.html	50M words	Text
	NIST 2008	https://catalog.ldc.upenn.edu/LDC2011S08	942 hours	Text, Attributes
	TedTalks	http://opus.nlpl.eu/TedTalks.php	2.81M tokens	Text
Text-to-Image	Caltech-UCSD Birds	http://www.vision.caltech.edu/visipedia/CUB-200-2011.html	11,788	Category, Text
	COCO	https://cocodataset.org/#captions-2015	123,287	Caption
	Oxford-102 Flowers	http://www.robots.ox.ac.uk/~vgg/data/flowers/102/	8,189	Category, Text
Image Captioning	COCO	https://cocodataset.org/#captions-2015	123,287	Caption
	nocaps	https://nocaps.org/	15,100	Caption, Category
	Flickr30k	http://shannon.cs.illinois.edu/DenotationGraph/	31,783	Caption
Coss-Media Retrieval	Wikipedia	https://en.wikipedia.org/wiki/Wikipedia:Featured_articles	2,866	I-T pairs
	PKU XMediaNet	http://59.108.48.34/tiki/XMediaNet/	40k	I-T-V-A pairs

on deep learning has made rapid progress in recent years, outperforming the traditional statistical MT methods or example-based MT methods by leveraging the powerful representation capacity and large-scale training data. The prevalent architecture for NMT is the encoder–decoder [172]. Later, attention mechanism is used to attend to all source words (i.e., global attention) or only part of them (i.e., local attention) when decoding at each step of RNN [173]–[175]. Attention can be useful for learning context features related to the target and achieve joint alignment and translation, showing better performance for long sentences. Unsupervised representation learning has shown promising performance for many downstream language tasks by learning context-aware and informative embeddings, e.g., BERT [29]. Recently, unsupervised NMT has also been studied, which could be trained on monolingual corpora. For example, leveraging BERT as contextual embedding has been proved useful for NMT by borrowing informative context from the pretrained model [176]. Together with speech recognition and speech synthesis, MT can be extended to translation speech from one language to another, which is very useful in many AIoT applications, such as language education [166], automatic translation and transcription, and multilingual customer service (e.g., subway broadcast).

16) Multimedia and Multimodal Analysis: With the rapid growth of multimedia content (e.g., text, audio, image, and video) created in various Internet platforms, understanding the content becomes a hot research topic. Recent studies on cross-media matching and retrieval try to align both domains semantically by leveraging deep learning, especially adversarial learning [177]. However, the modality-exclusive information impedes representation learning. To address this issue, disentangled representation learning has been proposed [178], which tries to maximize the mutual information between feature embeddings from different modalities and separate modality-exclusive features from them. Image/video captioning and text-to-image generation are two generative tasks related to cross-modal matching, where captioning refers to generating a piece of text description for a given image or video [179] while text-to-image generation aims to generate a realistic image that matches the given text description [180].

In addition to the aforementioned multimedia content, there are other modalities of data that are also useful for scene understanding, e.g., depth image, Lidar point cloud, thermal infrared image. By using them with RGB images as input, cross-modal perceiving has attracted increasing attention in real-world applications, e.g., scene parsing for autonomous driving [85], [181], object detection and tracking in low-light scenarios [182], [183], and action recognition [184]. There are three ways of fusing multimodal data, i.e., at the input level [181], at the feature level [85], [182], [183], [185], [186], and at the output level [184], respectively. Among them, fusing multimodal data at the feature level is most prevalent, which can be further categorized into three groups, i.e., early fusion [186], late fusion [185], and fusion at multiple levels [85], [182]. For example, a multibranch group fusion module is proposed to fuse features from RGB and thermal infrared images at different levels in [182], since the semantic information and visual details differ at different levels. Besides, Chen *et al.* [85] leveraged the residual learning idea to fuse the multilevel RGB image features and Lidar features via a residual structure in a cascaded manner.

Multimedia generation and cross-modal analysis are useful in some AIoT applications, e.g., television program retrieval/recommendation based on speech description [165], automatic (personalized) item description generation in e-commerce, a teaching assistant in education, multimedia content understanding and responding in a chatbot, nighttime object detection and tracking for smart security, and action recognition for rehabilitation monitoring and assessment. Another research topic that is close to AIoT is multimedia coding, which has also been advanced by deep learning [187]. It is noteworthy that a novel idea named video coding for machines is proposed recently [188], which attempts to bridge the gap between feature coding for machine vision and video coding for human vision. It can facilitate downstream tasks given the compact coded features as well as support human-in-the-loop inspection and intervention, therefore, having vast potential for supporting many AIoT applications. The representative benchmark data sets in the areas related to natural language processing and multimedia analysis are summarized in Table VII.

In the end, we briefly review the progress in network compression and neural architecture search (NAS).

17) Network Compression and NAS: Network compression is an effective technique to improve the efficiency of

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

DNNs for AIoT applications with limited computational budgets. It mainly included four kinds of techniques, i.e., network pruning, network quantization, low-rank factorization, and knowledge distillation. Typically, network pruning consists of three stages: 1) training a large network; 2) pruning the network according to a certain criterion; and 3) retraining the pruned network. Network pruning can be carried out at different levels of granularity, e.g., weight pruning, neuron pruning, filter pruning, and channel pruning based on the magnitude of weights or responses calculated by L_1/L_2 norm [189]. Network quantization compresses the original network by reducing the number of bits required for each weight, which significantly reduces memory use and float point operations with a slight loss of accuracy. Usually, uniform precision quantization is adopted inside the whole network, where all layers share the same bit width. Recently, a mixed-precision model quantization method has been proposed by leveraging the power of NAS [190], where different bit widths are assigned to different layers/channels. For other techniques, we recommend the comprehensive review in [191].

Instead of manually designing the network, NAS aims to automatically search the architecture from a predefined search space [192]. Most NAS methods fall into three categories, i.e., evolutionary methods, RL-based methods, and gradient-based methods. Evolutionary methods need to train a population of neural network architectures, which are then evolved with recombination and mutation operations. RL-based methods model the architecture generation process as a Markov decision process, treat the validation accuracy of the sampled network architecture as the reward, and update the architecture generation model (e.g., RNN controller) via RL algorithms. The above two kinds of methods require rewards/fitness from the sampled neural architecture, which usually leads to a prohibitive computational cost. In contrast, gradient-based methods adopt a continuous relaxation of the architecture representation. Therefore, the optimization of neural architecture can be conducted in a continuous space using gradient descent, which is orders of magnitude faster.

B. Learning

Since the real world is dynamic and complex, using a fixed model in AIoT systems cannot adapt to the variations, probably leading to a performance loss. Thereby, empowering things with learning ability is important for AIoT so that it can update and evolve in response to the variations. Here, we briefly review the progress in several subareas of machine learning as diagrammed in Fig. 7.

First, we review some research topics in machine learning, where none or few data/annotations from the target task are available, i.e., USL, SSL, TL, DA, FSL, and ZSL.

1) Unsupervised/Semisupervised Learning: Deep USL refers to learning from data without annotations based on DNNs, e.g., deep autoencoders, deep belief networks, and GAN, which can model the probability distribution of data. Recently, various GAN models have been proposed, which can generate high resolution and visually realistic images from random vectors. Accordingly, the models are expected to have learned a high-level understanding of the semantics of training data. For example, the recent BigBiGAN model can learn discriminative visual representation with good transferring performance on downstream tasks, by devising an encoder to learn an inverse mapping from data to the latent space [193]. Another hot research subarea is self-supervised learning, which learns discriminative visual representation by solving predefined pretext tasks [194]. For example, the recently proposed SimCLR method defines a context-based contrasting task for self-supervised learning [34], obtaining comparable performance as fully supervised models.

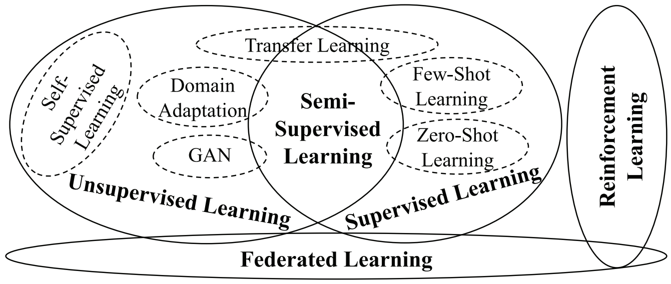

Fig. 7. Diagram of the learning-related topics in AIoT.

SSL refers to learning from both labeled and unlabeled data [195]. Usually, the amount of unlabeled data is much larger than that of labeled data. Recent studies adopt a teacher–student training paradigm, i.e., pseudolabels are generated by the teacher model on the unlabeled data set, which is then combined with the labeled data and used to train or fine-tune the student model. For example, an iterative training scheme is proposed in [35], where the trained student model is used as the teacher model at the subsequent training round. The method outperforms the fully supervised counterpart on ImageNet by a large margin.

Since annotating large-scale data can be prohibitively expensive and time consuming, USL and SSL can be useful for continually improving models in AIoT systems by harvesting the large-scale unlabeled data collected by massive numbers of sensors [196]. Besides, the multimodal data from heterogeneous sensors (e.g., RGB/infrared/depth camera, IMU, Lidar, and microphone) can be used to design cross modal-based pretext tasks (e.g., by leveraging audio–visual correspondence and ego-motion) and free semantic label-based pretext task (e.g., by leveraging depth estimation and semantic segmentation) for self-supervised learning [197].

2) Transfer Learning and Domain Adaptation: TL is a subfield of machine learning, aiming to address the learning problem of a target task without sufficient training data by transferring the learned knowledge from a source-related task [198]. Note that different from the aforementioned SSL where labeled and unlabeled data are usually drawn from the same distribution, TL does not require the data distributions of the source and the target domains to be identical. For example, it has been almost the *de facto* practice to fine-tune the models pretrained on ImageNet in different downstream tasks, e.g., object detection and semantic segmentation, for faster convergence and better generalization. In a recent study [36], a computational taxonomic map is discovered for TL between 26

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

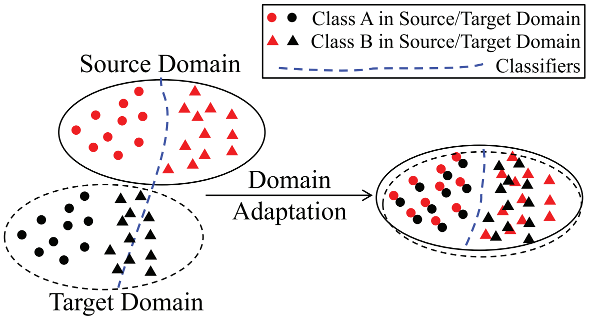
● ● Class A in Source/Target Domain
▲ ▲ Class B in Source/Target Domain
Classifiers
Source Domain
Domain Adaptation
Target Domain

Fig. 8. Illustration of DA.

visual tasks, providing valuable empirical insights, e.g., what tasks transfer well to other target tasks, and how to reuse supervision among related tasks to reduce the demand for labeled data while achieving same performance.

DA is also a long-standing research topic related to TL, which aims to learn a model from one or multiple source domains that performs well on the target domain for the same task (Fig. 8). When there are no annotations available in the target domain, this problem is a.k.a. unsupervised DA (UDA). Visual DA methods try to learn domain-invariant representations by matching the distributions between source and target domains at the appearance level, feature level, or output level, thereby reducing the domain shift. DA has been used in many computer vision tasks, including classification, object detection, and especially semantic segmentation [37], where obtaining the dense pixel-level annotations in the target domain is costly and time consuming. Recently, a mobile DA framework is proposed for edge computing in AIoT [38] by knowledge distillation from the teacher model on the server to the student model on the edge device.

In real-world AIoT systems, there are always many related tasks involved, e.g., object detection and tracking, and semantic segmentation in video surveillance. Therefore, finding the TL dependencies across these tasks and leveraging such prior knowledge to learn better models are of practical value for AIoT [199]–[202]. DA could be useful for AIoT applications when deploying models to new scenarios or new working modes of machines [128], [203]–[206], e.g., "synthetic→real," "daytime→nighttime," or "clear→rainy."

3) Few-/Zero-Shot Learning: FSL, as an application of meta-learning (i.e., learning to learn), aims to learn from only a few samples with annotations [31]. Prior knowledge can be leveraged to facilitate addressing the unreliable empirical risk minimizer issue in FSL due to the small few-shot training set. For example, prior knowledge can be used to augment training data by transforming samples from the training set, or an extra weakly labeled/unlabeled data set, or extra similar data sets. Besides, it can also be used to constrain hypothesis space and alter the search strategy in hypothesis space. In real-world AIoT applications, there are always some rare cases that need to be recognized by AI models, e.g., a car collision, cyber attack, and machine fault. However, the collection and annotation of such large-scale cases are usually very difficult. Thereby, FSL can be used to learn suitable models in these scenarios [207].

ZSL refers to learning a model with good generalization ability that can recognize unseen samples, whose classes have not been seen previously. Usually, auxiliary semantic information is provided to describe both seen and unseen classes, e.g., attributes-based description and text-based description. Thereby, each category can be represented as a feature vector in the attribute space or lexical space (a.k.a. semantic space). In some cases, the semantic space is a given learned semantic space, e.g., label embedding space or text-embedding space, where semantically similar classes are embedded as nearby vectors. Therefore, ZSL can be formulated as learning a mapping from the data space to the semantic space [32]. ZSL can be useful in some AIoT application scenarios. For example, in fault diagnosis [208], some specific types of faults may occur in the future and should be recognized, but no training instances belonging to them have ever been collected previously. In other cases, classes may change over time or new classes may emerge, e.g., electric bicycle and tricar, where ZSL can be leveraged to adapt to the variations.

Then, we review another two learning topics related to AIoT, i.e., RL and FL.

4) Reinforcement Learning: As one of the three basic learning paradigms along with supervised learning and USL, RL aims to learn a policy model for an agent through interactions with the environment such that it can maximize the cumulative reward. Recently, rapid progress has been made by incorporating deep learning in RL (i.e., DRL) [209]. DNN has a strong representation capacity for learning compact and discriminative feature representation from the high-dimensional image and video data, which enables RL to deal with previously intractable problems by learning better policy and value function. There are two main groups of DRL methods, including deep Q-network [210] and policy gradient-based methods, such as asynchronous advantage actor–critic [211]. While actor–critic-based methods directly optimize the cumulative reward, a surrogate objective function can be used to address the distribution shift problem in policy gradient-based RL methods.

DRL can empower things with the ability to interact with and adapt to the dynamic world, making it useful in many AIoT applications (Fig. 9), such as autonomous driving in smart transportation [199], [203], 3-D-landmark detection of CT scans [212] and robot control [213] in smart healthcare, course recommendation in smart education [41], real-time scheduling (RTS) for smart factory [214], load scheduling in smart grids [215], [216], plant growth control in smart agriculture [217], [218], and network management in smart cities [51], [219]. Moreover, DRL can be integrated into the FL framework for privacy-preserving learning [199], [216].

5) Federated Learning: FL was initially proposed to address the learning problem in which data sets are distributed across multiple devices and not allowed to be leaked to others [220]. Different data owners collaboratively train a model, whose performance is expected to match the performance of the model directly trained on the union of all data [60]. The architecture of FL usually consists of a central server (or collaborator) and many distributed client devices. Gradients are computed in each client using their own data and aggregated

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

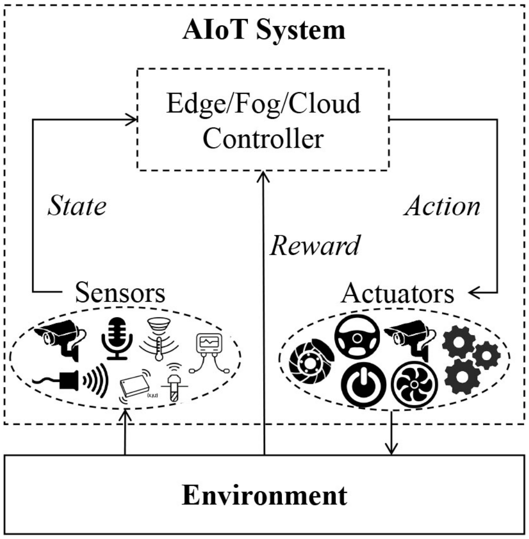

Fig. 9. Illustration of RL in AIoT.

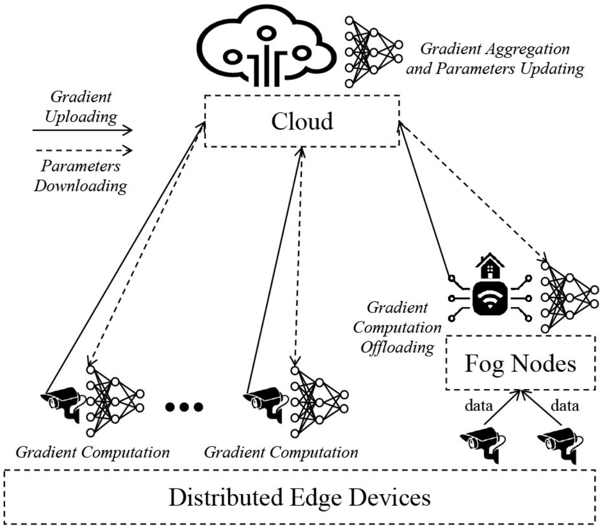

Fig. 10. Illustration of FL in AIoT. Some edge devices offload gradient computation to fog nodes.

(or concatenated) in the server, and then sent back to clients to update their models (Fig. 10). FL offers a general learning framework to use massive distributed and isolated data while preserving data security and privacy, which is particularly appealing to AIoT applications in the context of edge computing [197], [221]. For example, FL can be used to improve the perceiving and learning ability of AIoT devices, such as connected vehicles for autonomous driving [199] and wearable devices for health monitoring [197], [222].

C. Reasoning

In our real world, there is much knowledge carried by Web information, medical records, financial transactions, etc., which can be used to reason the answer to a question or infer patient cohorts. We humans have the ability of causal reasoning, such as causal inference and causal discovery. Empowering AIoT with such reasoning abilities is important for making smart and explainable decisions. In this part, we present the review on two related topics, i.e., knowledge graph (KG) and reasoning, as well as causal reasoning.

1) Knowledge Graph and Reasoning: KG is an efficient structured way to represent knowledge in a graph, where nodes represent entities, and edges represent relations (a.k.a. facts) in the forms of triples, i.e., (head entity, relation, and tail entity). Some well-known KGs, such as WordNet, Freebase, YAGO, and NELL have been constructed and used in many applications via knowledge reasoning. Knowledge reasoning refers to inferring new knowledge based on the existing knowledge, e.g., identification and removal of erroneous knowledge, adding missing knowledge, answering questions, and drawing conclusions. Classical knowledge reasoning methods are based on rules, including first-order predicate logic rules, probability rules, ontology languages, and path rules. Recently, KG embedding-based methods have attracted significant attention, aiming to embed the entities and relations in a KG into continuous vector spaces such that the reasoning can be done by leveraging translational distance models and semantic matching models [223].

Knowledge reasoning is useful for many downstream tasks, which can be categorized into in-KG and out-of-KG. In-KG applications include graph refinement (e.g., completion and error detection), triple/entity classification, and entity resolution. Out-of-KG applications include relation extraction, question answering, and recommendation, which are related to specific AIoT scenarios, such as network forensics analysis in smart security [224], smart assistants in smart healthcare and agriculture [39], [225], explainable recommendation in smart education and e-commerce [40], [226], digital twins in smart factory [163], [227], and fault diagnosis in smart grids [228]. Note that one crucial feature of knowledge reasoning is explainable, which has the potential to be incorporated with deep learning and mitigate the interpretability issue in AIoT applications.

2) Causal Reasoning: Causality refers to the generic relationship between an effect and the cause, where the cause partly gives rise to the effect and the effect partly depends on the cause. Causal reasoning includes the topics of causal inference that aims to estimate the causal effect and causal discovery that aims to find causal relations. One classical way of reasoning causality is via randomized controlled trial by evaluating the outcomes from the treatment group and control group, which is, however, costly and time consuming. Recently, learning causality from observational data has attracted much attention [229]. There are two well-known causal models used for learning causality, i.e., the structural causal models and the potential outcome framework (a.k.a. Rubin Causal Model). Different representation learning, multitask learning, and metalearning methods have been proposed for causal inference under the potential outcome framework.

Causal reasoning is useful in many AIoT applications, such as online recommendation in smart e-learning [226], fault analysis in smart grids [230], and driving safety in smart transportation [231]. Moreover, causal inference can be used to unfold the "black-box" decision process of DNNs to address the interpretability issue, including model-based interpretation methods [232] and example-based interpretation

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

methods [233]. They are crucial for building explainable AIoT systems-based DNNs, which have been widely used in many perceiving tasks as described in Section III-A.

D. Behaving

The behaving ability is also very important for AIoT systems passively responding to environment variations and raised requests, or actively exploring the unknown. Thereby, in this part, we present a brief review of two topics related to behaving in AIoT, i.e., control and interaction.

1) Control: The term of control here refers to controlling sensors and actuators in an AIoT system to transform the current system state to the target. The system state can be measured by sensors or calculated based on the perceiving methods described in Section III-A. The target state can be calculated based on predefined rules or determined by a decision model. The control algorithms are task oriented and specifically designed in different AIoT systems regarding their physical structures and sensors. For example, there are two kinds of control in autonomous driving systems, i.e., lateral control (steering) for autonomic turning and lane keeping, and longitudinal control (brake and throttle) for autonomic braking and forward collision avoidance. Multirobot systems have been used in many AIoT applications, such as smart logistics and precision agriculture, where decentralized control methods and coordination strategies are proposed for achieving and maintaining formations. Recently, deep RL has been used for autonomous robot control, e.g., self-driving vehicle [199], [203], medical robot [213], and mobile service robot. In the public safety or traffic video surveillance scenarios, the active cameras can change its orientation and focal length by pan-tilt-zoom control to track and focus on specific targets [234], [235].

2) Interaction: Real-world AIoT systems may interact with humans and the environment in different ways, which can be categorized into three groups: 1) using input and output devices for communication, e.g., keyboard, mouse, touch screen, microphone, and headset; 2) using mechanical arms for gripping and moving items, e.g., humanoid robots and industrial robots; and 3) using gearing for moving and transportation, e.g., wheels of mobile robots. In the first category, AIoT systems may communicate with humans via multimedia, such as text, speech, image, and video. For example, AIoT systems can send a message to user end or edge devices and display it on the screen as a reminder or a response to the user query. The message may contain text, audio, and video, which can be prerecorded or searched from the database. Besides, the content can also be automatically created according to the user request. For example, in the dialogue system of a smart home assistant, the user may send: 1) a picture or video to the assistant for generating a text description about it, which is related to image and video captioning [179]; 2) a keyword to the assistant for writing a poem [236]; 3) a piece of text description or a sketch to the assistant for creating an image [180], [237]; and 4) lyrics and score to the assistant for composing a piece of music [238]. Speech is another way for AIoT systems communicating with humans, which is enabled by speech synthesis [42], e.g., in question-answering system [163] and multilingual translator [166]. Moreover, AIoT systems can create virtual 3-D objects or scenes and render them on the AR/VR glasses or headsets. Human can interact with the virtual objects in the virtual environment or real objects in the augmented real environment.

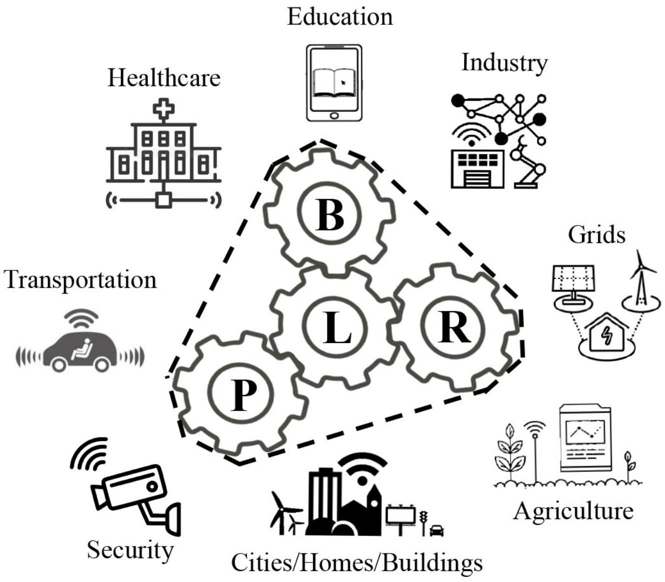

Fig. 11. AI empowers things with the ability of perceiving (P), learning (L), reasoning (R), and behaving (B) in many AIoT domains.

In the second group, a service robot in the smart home scenario may use its arms with flexible joints to grip a cup and fill it with water for elderly people. Besides, robot arms in the scenario of the Industrial IoT (IIoT) can be used for assembling in industrial product lines [138]. Note that the human–robot interaction is a long-standing research topic that special efforts have been made in task planning and programming [239], safe interaction [138], and imitation learning [240]. Recently, the idea of a digital twin has been proposed for the smart factory, various human interactions are enabled, such as question-answering and voice control [163], [227]. In the third group, due to the simple kinematics, wheel-based mobile robots are widely used for SLAM and navigation to explore and interact with the unknown environment [137], [139]. Note that there may be several interactive ways in an AIoT system upon the usage scenarios and devices.

IV. AIoT Applications

As reviewed in Section III, the progress of AI shows a great potential to empower the connected things in AIoT systems with the ability for perceiving, learning, reasoning, and behaving. The resulting AIoT systems will have a huge impact on the economic sectors and our living environments, such as security, transportation, healthcare, education, industry, energy, agriculture, as well as our homes and cities. In this part, we showcase some promising applications of AIoT in these areas (Fig. 11) and demonstrate how will AI enable the AIoT systems to be faster, smarter, greener, and safer.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
AIoT Applications Empowered by Different AI Technologies

	Perceiving	Learning	Reasoning	Behaving
Smart Security	object detection/tracking [234], [235], [242], text-spotting [242], biometric recognition [100], [170], [243], [241], [171], person re-identification [206], pose/gesture/action recognition [221], crow counting [126], [128], speech/speaker recognition [170], [171]	TL [128], DA [128], [206], FL [221]	KG reasoning [224]	control [234], [235], interaction [243]
Smart Transportation	pose/gesture/action recognition [244], [245], [121], object detection/tracking [245], [121], [137], semantic segmentation [203], [85], 3D [245], [121], [85], [137], text-spotting [242], speech/speaker recognition [164], [165], [121], multimedia [165]	TL [203], [199], DA [203], RL [203], [199], FL [199]	causal reasoning [231]	control [203], [199], interaction [231], [165], [164], [244], [245], [121], [137]
Smart Healthcare	object detection [120], [212], [213], pose/gesture/action recognition [120], 3D [213], multimedia [39]	SSL [197], RL [212], [213], FL [197], [222]	KG reasoning [39]	control [213], interaction [39]
Smart Education	image classification [207], pose/gesture/action recognition [125], [246], text-spotting [246], speech/speaker recognition [166], machine translation [166]	FSL [207], RL [41]	KG reasoning [226]	interaction [166], [247], [41], [246]
Smart Industry	classification [208], object detection/tracking [138], 3D [138], speech/speaker recognition [163]	DA [204], ZSL [208], RL [214]	KG reasoning [227], [163]	control [138], interaction [138], [227], [204], [163]
Smart Grids	image classification [248], object detection/tracking [248], 3D [249]	USL [250], [249], SSL [249], [196], TL [200], RL [215], [216]	KG reasoning [228], causal reasoning [230]	control [248], interaction [249]
Smart Agriculture	object detection/trakcing [139], [201], [140], [141], counting [202], [141], [205], semantic segmentation [202], [141], 3D [139], [140], [141], [251]	TL [201], [202], DA [205], RL [217], [218]	KG reasoning [225]	control [217], interaction [139], [140], [225]
Smart Cities / Homes / Buildings	object detection/tracking [5], text spotting [93], pose/gesture/action recognition [124], biometric recognition [170], speech/speaker recognition [170], [161], [162], 3D [252], [142], [143], [144]	RL [253], [219]	KG reasoning [254]	control [255], interaction [252], [254], [142], [124], [143], [144]

A. Smart Security

The goal of smart security is to ensure the security of our physical world and cyberspace, which can be achieved with the help of various AIoT systems. One of the most important features of them is *human-centric perceiving*, which can recognize the identities of individuals and analyze their behaviors to prevent illegal activities. For example, face recognition systems have been deployed in building entrance, railway station, and airport, enabled by cloud/fog computing [100] or edge computing [101]. Despite their utility, one major concern is data security and privacy preservation. Recently, a lightweight solution is proposed in [241] by block-based logic transformation, which not only reduces feature size but also preserves the original feature, therefore appealing to resource-limited AIoT devices and resistant to potential attacks. Beyond the biometric features from the face, fingerprint, and iris for recognition, spatial and temporal human body features (e.g., shape and gait) are leveraged for person reidentification, which aims to recognize individuals and trace their trajectories in multiple cameras. However, deploying the technique in a real-world scenario faces the domain shift challenge arisen from camera view differences (e.g., viewpoint, illumination, and resolution). To address it, Zhang *et al.* [206] proposed a style translation-based method for cross-domain person reidentification in camera sensor networks, which can reduce domain shift and learn domain-invariant features. Besides, recognizing identities of human-related objects, e.g., vehicle license plate recognition [242], is also useful for tracing human trajectory. Moreover, to capture salient and high-resolution humans and vehicles, active cameras can be used in AIoT systems by adjusting the orientation and focal length according to target location [234], [235]. Beyond the aforementioned techniques for individual analysis, estimating the crowd density and monitoring the crowd flow is also very important for public safety [128], e.g., avoid deadly accidents of trampling and crushing of pedestrians. Some AIoT applications in different domains are summarized in Table VIII according to the enabling AI technologies.

B. Smart Transportation

Smart transportation enabled by AIoT covers traffic participants (e.g., smart IoV [256]), traffic infrastructures [257], and industry applications (e.g., smart connected logistics [258]). Among them, the self-driving car is a typical example empowered by AI, which integrates various perceiving, learning, reasoning, and behaving abilities together. The self-driving system should perceive the driving environment, such as detecting road [85], traffic sign [259], pedestrian [260], and car [261], estimating the intention of cars and pedestrians and predicting their trajectories [244], [245]. Besides, it should also measure the pose and location of landmarks (e.g., traffic-signs) for SLAM [137]. Based on them, the self-driving system can determine its driving policy and interact with other traffic participants. Recently, deep RL is leveraged to learn driving policy directly from visual input (e.g., front-view images). However, carrying out the training in the real world is unaffordable. To address the issue, both DA [203] and TL

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

methods [199] are proposed by leveraging virtual 3-D game engines.

AIoT can also enable in-car driver monitoring and interactions with the infotainment system. Monitoring dangerous driver behaviors is crucial for preventing traffic accidents. To this end, a mobile application is developed in [121] for detecting dangerous driving behaviors, e.g., distraction and drowsiness. It leverages multimodel data for behavior detection, including images for face state detection (e.g., eyes openness and head yaw angles), motion data from IMU sensors for estimating vehicle location and speed, light level for estimating lighting conditions, and speech data for speech state detection (e.g., speech rate and loudness). Moreover, the research in [231] shows that identifying the causal factors of driving fatigue can be used for triggering corresponding countermeasures. For example, sleep-related fatigue is resistant to most interventions while task-related fatigue (e.g., distraction) can be counteracted effectively. Recently, AI interaction techniques, such as gesture and speech recognition, have been used for noncontact control in the in-car infotainment system [165]. The study in [164] shows that voice control allows drivers to keep watching out-car environment and demands lower mental load. In contrast, handheld controls may distract driver's attention and require higher mental demand, probably leading to dangerous driving behaviors.

C. Smart Healthcare

AIoT systems for smart healthcare cover several phases, including monitoring, examination, surgery, and rehabilitation. For monitoring, both wearable devices with motion sensors [262] and cameras [263] can be used for human activity recognition. Using motion sensors such as accelerometer, human activities can be recognized from time-series motion data based on a CNN model. Recently, a mobile robot with a camera is used for human activity recognition, where the control policy is obtained using deep RL by maximizing the recognition accuracy while minimizing its energy consumption. Note that the recognition module can be deployed in the edge devices or in the fog node depending on the model size and computational demand, which can be connected to the smart healthcare systems in community healthcare centers or hospitals. For examination, deep learning has been used for medical image understanding, such as 3-D-landmark detection in CT scans [212] and semantic segmentation [75]. Usually, these models are deployed on the private cloud of the hospital due to their high computational cost and the privacy concern. Recently, deep RL has been used to control the surgical robot [213] for multilateral cutting in 2D orthotropic gauze. Furthermore, AIoT systems can be useful for various rehabilitation monitoring and assessment [120], e.g., stroke rehabilitation and ankles rehabilitation. Via the connected 3-D AR/VR devices, therapists can assess the rehabilitation and make better treatments accordingly, which is helpful for patients in rural or distant areas. Recently, online healthcare services or medical assistant robots can offer convenient information-query and auxiliary diagnosis services. For example, a hierarchical attention network is proposed to make explainable and accurate answers by exploiting the structural, linguistics, and visual information within a multimodal medical KG [39].

D. Smart Education

AI technologies can empower AIoT things to help children and students recognize new species, learn native or foreign languages, choose personalized learning resources, and help those with visual impairments learn through interactions. For example, a weight imprinting-based FSL method is proposed in [207], which can be used in edge devices to recognize new bird and animal species with only a few labeled samples. In [246], Raspberry Pi is used to build a smart classroom by leveraging hand gesture recognition and text recognition technologies. This AIoT system enables to control Raspberry Pi to capture the lecture notes in blackboard/whiteboard via static hand gesture, recognize the text in the image (e.g., numbers, characters, and symbols), and convert them into an editable format, which is then saved in private cloud and shared to students for further editing or collaboration via the desktop application. A speech-to-speech multilingual translation system on mobile devices is proposed in [166], which includes three modules, i.e., speech recognition, language translation, and text-to-speech synthesis. It can work in offline mode and offers grammatical information that is useful for language learners. Recently, many mobile translator products have been released in CES 2020, which can translate tens of languages, benefiting from the advances of AI technologies such as deep learning. As a complement of on-campus learning, online learning via massive open online course (MOOC) platforms (e.g., Coursera) has become very popular. Learners watch the courses via their end devices connected to the cloud. Recommending courses from massive numbers of candidates could be helpful for personalized learning. To this end, deep RL-based [41] and rule-based [226] recommendation methods have been proposed. Besides, helping children with visual impairments in education also matters. For example, multisensory interactive maps are designed [247], which can help children to acquire skills via interactions, e.g., touching, listening, tasting, and scenting.

E. Smart Industry

Digital twin, i.e., a mirror digital representation of a physical system, has demonstrated great value for smart factories in Industry 4.0, e.g., monitoring the manufacturing process, diagnosing the fault, and preventing downtime. AIoT can be a critical part of implementing digital twins where the connected sensors and actuators can collect real-time data from production lines and send them to the digital twin running in the cloud (Fig. 12). Moreover, AI technologies can enable an intelligent analysis of data and help to make smart decisions. Recently, a service-oriented digital twin model is proposed in [163], which uses an ontology-oriented knowledge structure to represent the knowledge about the manufacturing system from the sensing data. It also designs a vocal interaction system for knowledge retrieval based on speech recognition and text-to-speech synthesis. In [227], a KG-based

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

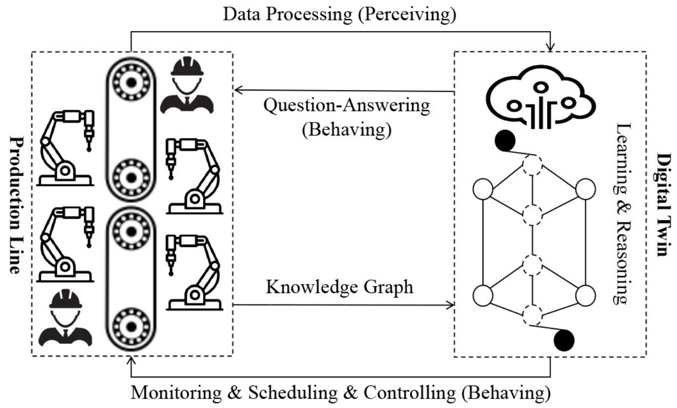

Fig. 12. Diagram of the digital twin system in a smart factory.

digital twin model is introduced which is composed of four parts, i.e., feature extraction, ontology creation, KG generation, and semantic relation extraction. It can extract and infer knowledge from large-scale production line data and enhance manufacturing process management via semantic relation reasoning. RTS in the smart factory is another hot research topic. In [214], a RL-based RTS model is proposed, which can incrementally update and maintain the knowledge base in RTS during operations to respond to shop floor environment change.

A typical example of AIoT application in the smart industry is the printed circuit board (PCB) manufacturing. There are three scenarios that are related to AIoT systems with different sensors and devices, i.e., manufacturing, visual defect inspection, and machine fault diagnosis. First, industrial robots have been widely used in the production line of smart factories, e.g., for drilling and grasping. AI technologies can be used to improve their functionalities. For example, Bousmalis *et al.* [204] proposed a deep robotic grasping model named GraspGAN, which bridges the domain gap between synthetic images and real-world ones via the pixel-level image translation and a feature-level domain classifier. To increase the safety, speed, and accuracy of autonomous picking and palletizing, Krug *et al.* [138] proposed a novel grasp representation scheme allowing redundancy in the gripper pose placement. Second, PCB defect inspection carried out by workers manually is laborious and time consuming. Recently, deep learning-based methods have been proposed for automatic real-time visual defect inspection [264]. Third, it is important to predict and diagnose machine faults from sensor data to reduce PCB defects, thereby increasing production efficiency and reducing losses. Although the digital twin system provides a useful mirror virtual environment for creating and testing new equipment and models, it is still challenging to fast adapt the trained model or control policy to the physical world. Thereby, more efforts should be made in the areas of DA, TL, and metalearning. Besides, since it is difficult to collect and annotate edge samples in the industrial context, zero-/FSL is also worth further study. In addition, causal analysis of the product defects based on data and knowledge is also of practical importance.

F. Smart Grids

AIoT in the smart grids can be used for grid fault diagnosis, load monitoring and scheduling, and cyber attack detection. For example, UAVs connected to the control center via the cellular network are used for damage classification and estimation of power distribution poles [248]. The captured image by UAVs is sent back to the cloud and processed by a CNN model to predict falling damage extent and fire damage extent. Besides, some "industrial stethoscopes" are designed to recognize and localize fault sound sources in visual scenes via cameras and multiple microphones, where the algorithm can run on edge devices for real-time monitoring. Recently, a two-stream network with an attention mechanism is proposed to directly localize a sound source in images [249]. AI technologies have also advanced fault diagnosis. For example, a convolutional sparse autoencoder-based USL method has been proposed for power transmission line fault diagnosis based on voltage and current signals [250]. It can detect faults within 7-ms latency, which is appealing to real-world applications. Besides, knowledge representation [228] and causal relationship discovery [230] are also explored for power grid fault diagnosis and impact causal analysis. For load monitoring and electric vehicle charging scheduling, TL [200] and deep RL [215] methods are proposed. Furthermore, to preserve the privacy of the households while intelligently managing their load scheduling, a distributed deep RL method is proposed inspired by the idea of FL [216], where the action networks are located at distributed households and the critic network is located at an aggregator from a trusted third party. The security concern about the smart grids under cyber attacks has attracted much attention. Recently, a semisupervised deep learning method is proposed based on autoencoder and GANs, which can effectively detect false data injection attacks in smart grids.

G. Smart Agriculture

Recently, the concept of precision agriculture becomes popular, referring to observing, measuring, and responding to crop variability through sensors, autonomous agricultural machines, and geographic information systems, which can be achieved by AIoT in smart agriculture. For example, crop counting and yield estimation is an important topic of precision agriculture [141], [202]. UVAs are used for capturing images of crops and fruits and sending them to the cloud for further counting [202]. Note that accumulating the counting results across image frames does not lead to the total yield since fruits are double-counted in adjacent frames. To address this issue, a detection-tracking-counting-based method is proposed [141], which can reject outliers and double-counted fruits. UAVs can also be used for continuous crop monitoring by capturing crop field images over time and temporally aligning them [251]. The path planning [140] and self-localization and navigation [139] abilities of UAVs are critical to complete the above tasks, which have been studied in the agricultural scenario. For plant grow control, an AIoT system is set up in a tomato greenhouse [217], which is composed of a wireless sensor network and cloud

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

computing center empowered by AI technologies. Deep RL is used for obtaining the optimal control policy on the illumination, nutrition, and ventilation conditions in the greenhouse. For crop management and pest/disease control, agricultural knowledge retrieval and answering system will be helpful. To this end, an agricultural KG, namely, AgriKG [225], is automatically constructed by leveraging NLP and deep learning techniques.

H. Smart Cities/Homes/Buildings

Smart cities/homes/buildings AIoT is related to those in the aforementioned smart sectors and can be empowered by similar AI technologies. The examples are as follows.

1) Continuous speaker authentication [170] in smart home voice assistants by integrating body-surface vibration signals with speech signal is related to smart security (Section IV-A).
2) HMI systems (e.g., control television) based on hand gesture recognition [122], [124] in smart homes share the similar techniques for traffic sign language recognition in smart transportation (Section IV-B) and sign language recognition in smart education (Section IV-D). Recently, an event-based gesture recognition system is proposed by using dynamic vision sensors and a neurosynaptic event-based processor [124], which can identify gestures with a low latency energy consumption.
3) Visual-navigation systems for the visually impaired [142]–[144], share some basic features, such as localization, obstacle recognition, and path planning of SLAM for autonomous driving in smart transportation and smart agriculture (Sections IV-B and IV-G), but also have some differences. For example, these systems usually contain some feedback modules (e.g., via vibration or speech) specifically designed for the visually impaired.
4) Sound visualization systems for the hearing impaired are related to the industrial stethoscopes for localizing fault sound sources in the grid (Section IV-F), but also offer different features. For example, the gender of speakers, sound types, loudness, and sound directions is displayed as indicator icons [252], specifically designed for the hearing impaired.
5) The energy management and optimization systems in smart buildings [253], [255] can be regarded as the extension of smart grids AIoT (Section IV-F).

V. Challenges and Opportunities

A. Challenges

1) Multimodal Heterogeneous Data Processing, Transmission, and Storage: AIoT systems contain massive numbers of heterogeneous sensors that generate a large data stream of different formats, sizes, and timestamps, thereby significantly challenging further processing, transmission, and storage. An efficient coding scheme can be used to reduce network bandwidth and transmission latency. For example, the video coding scheme for machines [188] is promising for facilitating downstream computer vision tasks and requires further study. The AI perceiving technologies mentioned in Section III-A could be used to extract compact and structural representations from the data, which would be transmission and storage friendly. However, the structural representation is task oriented and should be calculated at the network edge to minimize bandwidth and latency, which represents another challenge.

2) Deep Learning on Edge Devices: Deploying deep CNN models on edge devices is crucial for real-time data stream processing and low latency in AIoT systems. However, edge devices are limited by their limited computational and storage resources. Thereby, how to design or automatically search lightweight, computationally efficient, and hardware-friendly DNN architectures is of practical value but still remains challenging. Moreover, network pruning, compression, and quantization are also worth further exploring.

3) Computational Scheduling in the AIoT Architecture: As described in Section II-A, a typical AIoT architecture contains heterogeneous computing resources, including cloud centers, fog nodes, and edge devices. In real-world AIoT systems, some intense computation may be needed to offload to the fog node or cloud center from the edge devices, thereby creating a computational scheduling challenge. Specifically, when scheduling computation across different resources, the following factors should be taken into account: data type and volume, network bandwidth, processing latency, performance accuracy, energy consumption, and data security and privacy in each specific application scenario. Moreover, a dynamic adaptive scheduling strategy would deal with unbalanced data flow and user demands over time.

4) Big and Small Data for Deep Learning in AIoT: Big data generated from massive numbers of sensors are ubiquitous in AIoT systems, with huge potential for deep learning. As reviewed in Section III-A, deep supervised learning methods have achieved remarkable success for perceiving in different areas due to large-scale labeled data. However, most AIoT data are unlabeled, and labeling them would be both time and financially expensive. Although there has been rapid progress in USL, especially self-supervised learning [34], [194], [265], future efforts are expected to further leverage AIoT data, especially multimodal data. Furthermore, given the small scale of labeled data, TL, SSL, and FSL in the AIoT context might provide solutions to challenges from new classes, rare cases, and state drifting of devices.

5) Data Monopoly: In the AI era, data provide a valuable resource for creating new products and improving services. AIoT companies collect and exploit massive data, thereby leading to new opportunities for data collection and exploitation. This positive loop could lead to a data monopoly, i.e., vast proprietary data protected by established interests that cannot be accessed by other entities. Consequently, new competitors face a *de facto* barrier to market entry, and a data monopoly becomes a real threat to free-market competition.

6) Data Security and Privacy: Due to sensors being ubiquitous in smart homes, hospitals, and cities, vast biometric data (e.g., face image, voice, action, pulse, imaging data, etc.) of AIoT users or informed and uninformed participants may be collected. This raises important concerns with regards

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
LIST OF ABBREVIATIONS

Abbreviation	Description
5G	5th Generation Mobile Networks
AI	Artificial Intelligence
AIoT	Artificial Intelligence of Things
AR/VR	Augmented Reality / Virtual Reality
ASIC	Application-Specific Integrated Circuit
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Networks
CTC	Connectionist Temporal Classification
DA	Domain Adaptation
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
FL	Federated Learning
FOV	Field-of-View
FPGA	Field-Programmable Gate Arrays
FSL	Few-Shot Learning
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HMI	Human-Machine Interaction
HMM	Hidden Markov Model
IIoT	Industrial Internet of Things
IMU	Inertial Measurement Unit
IoT	Internet of Things
IoV	Internet of Vehicles
IoVT	Internet of Video Things
KG	Knowledge Graph
LAN	Local Area Network
LSTM	Long Short-Term Memory
LTE	Long Term Evolution
MOOC	Massive Open Online Course
MT	Machine Translation
NAS	Neural Architecture Search
NMT	Neural Machine Translation
OCR	Optical Character Recognition
ONNX	Open Neural Network Exchange
PCB	Printed Circuit Board
RFID	Radio-Frequency Identification
RL	Reinforcement Learning
RNN	Recurrent Neural Network
ROI	Region of Interest
RTS	Real-Time Scheduling
SLAM	Simultaneous Localization and Mapping
SSL	Semi-Supervised Learning
TL	Transfer Learning
TPU	Tensor Processing Unit
UAV	Unmanned Aerial Vehicle
UDA	Unsupervised Domain Adaptation
USL	Unsupervised Learning
VIO	Visual Inertial Odometry
VO	Visual Odometry
ZSL	Zero-Shot Learning

to data security and privacy. Who owns these data? How long will these data be retained? How will these data be used? Legislation is important in response to these concerns, a good example being the general data protection regulation (GDPR)[12] enforced by the European Union, which gives individuals control over their personal data. Controllers and processors of personal data must take appropriate measures to protect data security and privacy.

7) Growing Energy Consumption in Data Centers: According to [266], electricity use by communication technologies is expected to account for 21% of global total usage, with data centers contributing more than 1/3 to this. Therefore, enhancing energy efficiency in data centers is required for a sustainable future. For example, some data centers are located in cold climates to take advantage of air cooling. Other solutions include water cooling and immersing servers in a nonconductive oil or mineral bath. Workload analysis, task scheduling, and virtual machine consolidation have also been studied to improve power efficiency in data centers [267], [268]. The rapidly growing number of cloud centers mirrors the rapid growth in AIoT applications. Consequently, more efforts should continue to be made to address energy consumption in data centers.

B. Opportunities

1) Built-in Neural Processing Capacity for Edge Devices: Many edge devices are equipped with specialized chips (e.g., GPUs in smartphones and intelligent cameras) to accelerate neural network processing. Consequently, building neural processing capacity into edge devices is very useful for AIoT applications. First, it reduces processing latency and network bandwidth consumption. Since the sensing data can be processed on-site, only a small amount of processed data need to be transmitted. Second, it can protect data security and privacy. For example, for biometric verification, registered user biometric data could be stored with encryption on local hardware, with only the built-in verification capacity on the edge devices exposed to the applications, thereby reducing the risk of data leakage. Third, it enables distributed and asymmetric model training. An FL framework can be used to train models on distributed edge devices by leveraging their local sensor data. Moreover, some groups of devices may choose different model updating policies than others depending on their usage scenarios.

2) Event-Based Sensors and Neuromorphic Processors: Traditional camera sensors continually generate dense data once opened, which are further fed into deep CNNs for further GPU processing. Generally, all the pixels are used in the calculation, resulting in high computational costs. Recently, event-based sensors and neuromorphic processors have been proposed [124]. For example, event-based cameras only record pixels that change in brightness, thereby reducing redundant data generation and transmission. Event-based neuromorphic processors can operate on sparse and asynchronous event streams directly, avoiding dense and redundant computations on regular sensing data such as GPUs. These can be used in many AIoT applications, such as gesture/action recognition with low-power consumption and latency.

3) Deep Learning From the Virtual to the Real: In embodied AI, it is difficult and/or costly to train models in the real world, e.g., autonomous driving, robot arm control, and robot navigation. 3-D virtual platforms that mimic real-world scenarios have been proposed, such as Voyage Deepdrive,[13] OpenAI gym,[14] and Habitat,[15] which are very useful for cost effectively training deep learning models, especially for deep

[12] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[13] https://deepdrive.voyage.auto/

[14] https://gym.openai.com/

[15] https://aihabitat.org/

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

RL. Nevertheless, the critical issue of domain shift between the virtual and physical environments must be addressed before deploying the trained model into real-world scenarios. Recently, TL and DA have attracted significant attention to address this issue in the setting of both USL and RL.

4) Data and Knowledge Integration for Perceiving, Learning, Reasoning, and Behaving: Deep learning model performance is largely determined by large-scale training data. However, humans learn new concepts-based not only on data but on prior knowledge. Likewise, prior knowledge can be very useful for training deep learning models in a data-efficient way. For example, attribute-based class description enables ZSL for new concepts via attribute transfer. Another example is KGs, which represent structural relationships between entities. Knowledge can be extracted from unstructured data to build KGs, learn knowledge-embedding representations, and for reasoning. Integrated with deep learning (e.g., graph neural networks), this is a useful approach in many areas, such as question and answer systems and fault/disease diagnosis, opening up promising research avenues toward human-level cognition intelligence. Therefore, data and knowledge integration are important for improving the perceiving, learning, reasoning, and behavior of AIoT.

5) Privacy-Preserving Deep Learning: Deep learning requires large-scale data generated by different things from different users in the context of AIoT. Individuals may worry about data security and privacy if data are transmitted to and stored in the cloud. To alleviate these concerns, privacy-preserving deep learning has attracted attention from both the deep learning and information security communities. The recently proposed FL framework (refer to Section III-B5) is a representative and promising solution that allows data to be stored locally in distributed devices. Homomorphic encryption [269] has been used in FL to prevent data leakage to the server. All abbreviations in this article are listed in Table IX.

VI. Conclusion

Here, we present a comprehensive survey of AIoT, covering AIoT computing architectures; AI technologies for empowering IoT with perceiving, learning, reasoning, and behaving abilities; promising AIoT applications; and the challenges and opportunities facing AIoT research. The three-tier computing architecture of AIoT provides different computing resources for deep learning whilst also posing new challenges, e.g., in the design and search of lightweight models and computation scheduling within the three-tier architecture. Deep learning has rapidly progressed in many perceiving areas and enables many AIoT applications. Nevertheless, more effort should be made to improve edge intelligence. In the context of USL and other machine learning topics such as RL, deep learning has attracted an increasing amount of attention and is useful for further improving the intelligence of AIoT systems to handle dynamic and complex environments. Moreover, reasoning based on KGs and causal analysis is a challenging but active research area, having the potential to enable AIoT systems to approach human-level cognitive intelligence. To respond to the dynamic environment, AIoT behaves via control and interaction, where deep learning has demonstrated its value in improving control accuracy and enabling multimodal interactions. In the future, empowered by rapidly developing AI technologies, many fast, smart, green, and safe AIoT applications are expected to deeply reshape our world.

References

[1] K. Ashton, "That 'Internet of Things' thing," *RFID J.*, vol. 22, no. 7, pp. 97–114, 2009.

[2] *Terms of the Ubiquitous Network*, CCSA Standard YDB 062-2011, 2011.

[3] *Overview of IoT*, ITU-T Standard Y.2060, 2012.

[4] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[5] X. Li, R. Lu, X. Liang, X. Shen, J. Chen, and X. Lin, "Smart community: An Internet of Things application," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 68–75, Nov. 2011.

[6] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: Challenges and solutions," *J. Clean. Prod.*, vol. 140, pp. 1454–1464, Jan. 2017.

[7] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibañez, "Internet of Vehicles: Architecture, protocols, and security," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3701–3709, Oct. 2018.

[8] S. Aslam, M. P. Michaelides, and H. Herodotou, "Internet of ships: A survey on architectures, emerging applications, and challenges," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9714–9727, Oct. 2020.

[9] Z. Guan, J. Li, L. Wu, Y. Zhang, J. Wu, and X. Du, "Achieving efficient and secure data acquisition for cloud-supported Internet of Things in smart grid," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1934–1944, Dec. 2017.

[10] L. Catarinucci *et al.*, "An IoT-aware architecture for smart healthcare systems," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 515–526, Dec. 2015.

[11] A. M. Rahmani *et al.*, "Exploiting smart e-health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," *Future Gene. Comput. Syst.*, vol. 78, pp. 641–658, Jan. 2018.

[12] E. Lamarre and B. May, *Ten Trends Shaping the Internet of Things Business Landscape*. New York, NY, USA: McKinsey Digit., 2019.

[13] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[14] J. Macaulay, L. Buckalew, and G. Chung, *Internet of Things in Logistics: A Collaborative Report by DHL and Cisco on Implications and Use Cases for the Logistics Industry*, DHL Trend Res., Bonn, Germany and Cisco Consulting Services, San Jose, CA, USA, 2015.

[15] IDC. (2019). *The Growth in Connected IoT Devices is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast*. Accessed: Jun. 18, 2019. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS45213219

[16] F. Bonomi, "Connected vehicles, the Internet of Things, and fog computing," in *Proc. 8th ACM Int. Workshop Veh. Inter. Netw.*, 2011, pp. 13–15.

[17] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[18] J. Ni, K. Zhang, X. Lin, and X. S. Shen, "Securing fog computing for Internet of Things applications: Challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 601–628, 1st Quart., 2018.

[19] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[20] J. Pan and J. McElhannon, "Future edge cloud and edge computing for Internet of Things applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018.

[21] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[22] E. M. Tordera *et al.*, "What is a fog node a tutorial on current concepts towards a common definition," 2016. [Online]. Available: arXiv:1611.09193.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[23] C. W. Chen, "Internet of video things: Next-generation IoT with visual sensors," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6676–6685, Aug. 2020.

[24] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 338–342.

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1701–1708.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Toward real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*. New York, NY, USA: Curran Assoc., 2015, pp. 91–99.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[31] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[32] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, 2019.

[33] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: arXiv:2002.05709.

[35] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 10687–10698.

[36] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 3712–3722.

[37] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 435–445.

[38] J. Yang, H. Zou, S. Cao, Z. Chen, and L. Xie, "MobileDA: Toward edge-domain adaptation," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6909–6918, Aug. 2020.

[39] Y. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1089–1097.

[40] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5329–5336.

[41] J. Zhang, B. Hao, B. Chen, C. Li, H. Chen, and J. Sun, "Hierarchical reinforcement learning for course recommendation in MOOCs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 435–442.

[42] S. Ö. Arık *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 195–204.

[43] J. Bradley, J. Barbier, and D. Handler, "Embracing the Internet of everything to capture your share of $14.4 trillion," Cisco, San Jose, CA, USA, White Paper, vol. 318, 2013.

[44] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, 2010.

[45] A. Whitmore, A. Agarwal, and L. Da Xu, "The Internet of Things—A survey of topics and trends," *Inf. Syst. Front.*, vol. 17, no. 2, pp. 261–274, 2015.

[46] I. Yaqoob *et al.*, "Internet of Things architecture: Recent advances, taxonomy, requirements, and open challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 10–16, Jun. 2017.

[47] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.

[48] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, and N. Kato, "A survey on network methodologies for real-time analytics of massive IoT data and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1457–1477, 3rd Quart., 2017.

[49] L. Mainetti, L. Patrono, and A. Vilei, "Evolution of wireless sensor networks towards the Internet of Things: A survey," in *Proc. 19th Int. Conf. Softw. Telecommun. Comput. Netw.*, Split, Croatia, 2011, pp. 1–6.

[50] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[51] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2nd Quart., 2020.

[52] C.-W. Tsai, C.-F. Lai, M.-C. Chiang, and L. T. Yang, "Data mining for Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 77–97, 1st Quart., 2014.

[53] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of Things data analysis: A survey," *Digit. Commun. Netw.*, vol. 4, no. 3, pp. 161–175, 2018.

[54] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 414–454, 1st Quart., 2014.

[55] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.

[56] K. Ota, M. S. Dao, V. Mezaris, and F. G. D. Natale, "Deep learning for mobile multimedia: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3s, pp. 1–22, 2017.

[57] L. Li, K. Ota, and M. Dong, "Deep learning for smart industry: Efficient manufacture inspection system with fog computing," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4665–4673, Oct. 2018.

[58] J. Qiu, Z. Tian, C. Du, Q. Zuo, S. Su, and B. Fang, "A survey on access control in the age of Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 4682–4696, Jun. 2020.

[59] Z. Yan, P. Zhang, and A. V. Vasilakos, "A survey on trust management for Internet of Things," *J. Netw. Comput. Appl.*, vol. 42, pp. 120–134, Jun. 2014.

[60] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[61] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[62] C. A. Tokognon, B. Gao, G. Y. Tian, and Y. Yan, "Structural health monitoring framework based on Internet of Things: A survey," *IEEE Internet Things J.*, vol. 4, no. 3, pp. 619–635, Jun. 2017.

[63] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. N. Hindia, "An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3758–3773, Oct. 2018.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[65] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: arXiv:1704.04861.

[66] J. Zhang, Y. Cao, Y. Wang, C. Wen, and C. W. Chen, "Fully point-wise convolutional neural network for modeling statistical regularities in natural images," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 984–992.

[67] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 525–542.

[68] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788.

[69] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[70] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 2722–2730.

[71] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, "A survey of appearance models in visual object tracking," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, pp. 1–48, 2013.

[72] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.

[73] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2805–2813.

[74] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 4282–4291.

[75] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[76] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[77] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–11.

[78] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2881–2890.

[79] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 2961–2969.

[80] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[81] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.

[82] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 4974–4983.

[83] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 9404–9413.

[84] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 6399–6408.

[85] Z. Chen, J. Zhang, and D. Tao, "Progressive LiDAR adaptation for road detection," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 693–702, May 2019.

[86] H. He, J. Zhang, Q. Zhang, and D. Tao, "Grapy-ML: Graph pyramid mutual learning for cross-dataset human parsing," in *Proc. 34th AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10949–10956.

[87] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 512–528.

[88] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5676–5685.

[89] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 2315–2324.

[90] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 5238–5246.

[91] J. Liu, Z. Chen, B. Du, and D. Tao, "ASTS: A unified framework for arbitrary shape text spotting," *IEEE Trans. Image Process.*, vol. 29, pp. 5924–5936, 2020.

[92] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[93] C. Mancas-Thillou, S. Ferreira, J. Demeyer, C. Minetti, and B. Gosselin, "A multifunctional reading assistant for the visually impaired," *EURASIP J. Image Video Process.*, vol. 2007, no. 1, 2007, Art. no. 064295.

[94] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Kauai, HI, USA, 2001, pp. 511–518.

[95] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 5325–5334.

[96] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[97] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[98] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1685–1692.

[99] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 815–823.

[100] T. Soyata, R. Muraleedharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture," in *Proc. IEEE Symp. Comput. Commun.*, Cappadocia, Turkey, 2012, pp. 59–66.

[101] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Dept. School Comput. Sci., CMU, Pittsburgh, PA, USA, Rep. CMU-CS-16-118, 2016.

[102] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surveys*, vol. 50, no. 1, pp. 1–37, 2017.

[103] L. Fei, G. Lu, W. Jia, S. Teng, and D. Zhang, "Feature extraction methods for palmprint recognition: A survey and evaluation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 2, pp. 346–363, Feb. 2019.

[104] Y. Zhang, L. Zhang, X. Liu, S. Zhao, Y. Shen, and Y. Yang, "Pay by showing your palm: A study of palmprint verification on mobile platforms," in *Proc. IEEE Int. Conf. Multimedia Expo*, Shanghai, China, 2019, pp. 862–867.

[105] L. Zhang, L. Li, A. Yang, Y. Shen, and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," *Pattern Recognit.*, vol. 69, pp. 199–212, Sep. 2017.

[106] Z. Chen, J. Zhang, and D. Tao, "Recursive context routing for object detection," *Int. J. Comput. Vis.*, pp. 1–19, Aug. 2020, doi: 10.1007/s11263-020-01370-7.

[107] J. Zhang, Z. Chen, and D. Tao, "Toward high performance human keypoint detection," 2020. [Online]. Available: arXiv:2002.00537.

[108] X. Zeng, C. Ding, Y. Wen, and D. Tao, "Soft-ranking label encoding for robust facial age estimation," 2019. [Online]. Available: arXiv:1906.03625.

[109] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 2002–2011.

[110] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020. [Online]. Available: arXiv:2001.04193.

[111] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 79–88.

[112] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7291–7299.

[113] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5693–5703.

[114] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 7093–7102.

[115] J. Wang, S. Huang, X. Wang, and D. Tao, "Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 7771–7780.

[116] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1110–1118.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[117] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 3288–3297.

[118] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," 2018. [Online]. Available: arXiv:1801.07455.

[119] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[120] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, "A kinect-based system for cognitive rehabilitation exercises monitoring," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 620–631, 2014.

[121] A. Kashevnik, I. Lashkov, and A. Gurtov, "Methodology and mobile application for driver behavior analysis and accident prevention," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2427–2436, Jun. 2020.

[122] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, 2015.

[123] J. Lien *et al.*, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–19, 2016.

[124] A. Amir *et al.*, "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7243–7252.

[125] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[126] L. Schauer, M. Werner, and P. Marcus, "Estimating crowd densities and pedestrian flows using Wi-Fi and Bluetooth," in *Proc. 11th Int. Conf. Mobile Ubiquitous Syst. Comput. Netw. Serv.*, 2014, pp. 171–177.

[127] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 30, 2020, doi: 10.1109/TPAMI.2020.3013269.

[128] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 8198–8207.

[129] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[130] Z.-F. Wang and Z.-G. Zheng, "A region based stereo matching algorithm using cooperative optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.

[131] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5667–5675.

[132] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 8977–8986.

[133] N. Yang, L. V. Stumberg, R. Wang, and D. Cremers, "D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 1281–1292.

[134] L. Han, Y. Lin, G. Du, and S. Lian, "DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Macau, China, 2019, pp. 6906–6913.

[135] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[136] E. Parisotto, D. Singh Chaplot, J. Zhang, and R. Salakhutdinov, "Global pose estimation with an attention-based recurrent network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, 2018, pp. 237–246.

[137] X. Shao, L. Zhang, T. Zhang, Y. Shen, H. Li, and Y. Zhou, "A tightly-coupled semantic SLAM system with visual, inertial and surround-view sensors for autonomous indoor parking," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2691–2699.

[138] R. Krug *et al.*, "The next step in robot commissioning: Autonomous picking and palletizing," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 546–553, Jan. 2016.

[139] F. A. Cheein, G. Steiner, G. P. Paina, and R. Carelli, "Optimized EIF-SLAM algorithm for precision agriculture mapping based on stems detection," *Comput. Electron. Agr.*, vol. 78, no. 2, pp. 195–207, 2011.

[140] B. H. Y. Alsalam, K. Morton, D. Campbell, and F. Gonzalez, "Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, 2017, pp. 1–12.

[141] X. Liu *et al.*, "Robust fruit counting: Combining deep learning, tracking, and structure from motion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1045–1052.

[142] Y. H. Lee and G. Medioni, "RGB-D camera based wearable navigation system for the visually impaired," *Comput. Vis. Image Understand.*, vol. 149, pp. 3–20, Aug. 2016.

[143] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *Proc. IEEE Int. Conf. Robot. Autom.*, Singapore, 2017, pp. 6533–6540.

[144] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 583–593, Mar. 2018.

[145] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

[146] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, pp. 982–993, 2017.

[147] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, pp. 2828–2841, 2018.

[148] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, and Y. Zhou, "Zero-shot restoration of underexposed images via robust retinex decomposition," in *Proc. IEEE Int. Conf. Multimedia Expo*, London, U.K., 2020, pp. 1–6.

[149] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[150] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, pp. 5187–5198, 2016.

[151] J. Zhang and D. Tao, "FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2019.

[152] S. Zhao, L. Zhang, S. Huang, Y. Shen, and S. Zhao, "Dehazing evaluation: Real-world benchmark datasets, criteria, and baselines," *IEEE Trans. Image Process.*, vol. 29, pp. 6947–6962, 2020.

[153] J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. Wen Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7418–7426.

[154] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 226–234.

[155] J. Zhang, Y. Cao, Z.-J. Zha, and D. Tao, "Nighttime dehazing with a synthetic benchmark," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2355–2363.

[156] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[157] Z. Xue, N. Xue, G.-S. Xia, and W. Shen, "Learning to calibrate straight lines for fisheye image rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 1643–1651.

[158] X. Liu, L. Zhang, Y. Shen, S. Zhang, and S. Zhao, "Online camera pose optimization for the surround-view system," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 383–391.

[159] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 11. Tokyo, Japan, 1986, pp. 49–52.

[160] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, AB, Canada, 2018, pp. 5884–5888.

[161] Y.-S. Chang, S.-H. Hung, N. J. Wang, and B.-S. Lin, "CSR: A cloud-assisted speech recognition service for personal mobile device," in *Proc. Int. Conf. Parallel Process.*, Taipei City, Taiwan, 2011, pp. 305–314.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[162] Y. He *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., 2019, pp. 6381–6385.

[163] F. Longo, L. Nicoletti, and A. Padovano, "Ubiquitous knowledge empowers the smart factory: The impacts of a service-oriented digital twin on enterprises' performance," *Annu. Rev. Control*, vol. 47, pp. 221–236, Jan. 2019.

[164] J. M. Owens, S. B. McLaughlin, and J. Sudweeks, "On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players," *SAE Int. J. Passenger Cars Mech. Syst.*, vol. 3, no. 1, pp. 734–743, 2010.

[165] I. Tashev, M. Seltzer, Y.-C. Ju, Y.-Y. Wang, and A. Acero, "Commute UX: Voice enabled in-car infotainment system," in *Proc. Workshop Speech Mobile Pervasive Environ. (SiMPE Mobile HCI)*, Sept. 2009.

[166] K. Angelov, B. Bringert, and A. Ranta, "Speech-enabled hybrid multilingual translation for mobile devices," in *Proc. Demonstrations 14th Conf. Eur. Ch. Assoc. Comput. Linguist.*, 2014, pp. 41–44.

[167] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[168] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[169] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2018, pp. 1086–1090.

[170] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 343–355.

[171] A. Ross, S. Banerjee, and A. Chowdhury, "Security in smart cities: A brief review of digital forensic schemes for biometric data," *Pattern Recognit. Lett.*, vol. 138, pp. 346–354, Oct. 2020.

[172] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop Syntax Semantic Struct. Stat. Transl.*, 2014, pp. 103–111.

[173] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 3104–3112.

[174] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. [Online]. Available: arXiv:1609.08144.

[175] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2017, pp. 5998–6008.

[176] J. Zhu *et al.*, "Incorporating BERT into neural machine translation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.

[177] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[178] W. Guo, H. Huang, X. Kong, and R. He, "Learning disentangled representation for cross-modal retrieval with deep mutual information estimation," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1712–1720.

[179] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surveys*, vol. 51, no. 6, pp. 1–36, 2019.

[180] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 1505–1514.

[181] L. Li, B. Qian, J. Lian, W. Zheng, and Y. Zhou, "Traffic scene segmentation based on RGB-D image and deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1664–1669, May 2018.

[182] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multilevel CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2019.

[183] C. Wang *et al.*, "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 7064–7073.

[184] J. Liu, Y. Li, S. Song, J. Xing, C. Lan, and W. Zeng, "Multi-modality multi-task recurrent neural network for online action detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2667–2682, Sep. 2019.

[185] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, "Large-margin multi-modal deep learning for RGB-D object recognition," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.

[186] Z. Xu, S. Liu, J. Shi, and C. Lu, "Outdoor RGBD instance segmentation with residual regretting learning," *IEEE Trans. Image Process.*, vol. 29, pp. 5301–5309, 2020.

[187] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding: A review and a case study," *ACM Comput. Surveys*, vol. 53, no. 1, pp. 1–35, 2020.

[188] L.-Y. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," 2020. [Online]. Available: arXiv:2001.03569.

[189] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 1389–1397.

[190] Z. Cai and N. Vasconcelos, "Rethinking differentiable search for mixed-precision neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 2349–2358.

[191] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.

[192] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, pp. 1–21, Mar. 2019.

[193] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 10542–10552.

[194] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 4, 2020, doi: 10.1109/TPAMI.2020.2992393.

[195] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[196] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Trans. Smart Grid*, early access, Jul. 20, 2020, doi: 10.1109/TSG.2020.3010510.

[197] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien, "Federated self-supervised learning of multi-sensor representations for embedded intelligence," *IEEE Internet Things J.*, early access, Jul. 15, 2020, doi: 10.1109/JIOT.2020.3009358.

[198] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proc. IEEE*, early access, Jul. 7, 2020, doi: 10.1109/JPROC.2020.3004555.

[199] X. Liang, Y. Liu, T. Chen, M. Liu, and Q. Yang, "Federated transfer reinforcement learning for autonomous driving," 2019. [Online]. Available: arXiv:1910.06001.

[200] M. D'Incecco, S. Squartini, and M. Zhong, "Transfer learning for non-intrusive load monitoring," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1419–1429, Mar. 2020.

[201] S. Bargoti and J. Underwood, "Deep fruit detection in orchards," in *Proc. IEEE Int. Conf. Robot. Autom.*, Singapore, 2017, pp. 3626–3633.

[202] S. W. Chen *et al.*, "Counting apples and oranges with deep learning: A data-driven approach," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 781–788, Apr. 2017.

[203] X. Pan, Y. You, Z. Wang, and C. Lu, "Virtual to real reinforcement learning for autonomous driving," 2017. [Online]. Available: arXiv:1704.03952.

[204] K. Bousmalis *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *Proc. IEEE Int. Conf. Robot. Autom.*, Brisbane, QLD, Australia, 2018, pp. 4243–4250.

[205] M. Valerio Giuffrida, A. Dobrescu, P. Doerner, and S. A. Tsaftaris, "Leaf counting without annotations using adversarial unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Long Beach, CA, USA, 2019, pp. 2590–2599.

[206] Z. Zhang, Y. Wang, and S. Liu, "Cross-domain person re-identification using dual generation learning in camera sensor networks," *Ad Hoc Netw.*, vol. 97, Feb. 2020, Art. no. 102019.

[207] N. Passalis, A. Iosifidis, M. Gabbouj, and A. Tefas, "Hypersphere-based weight imprinting for few-shot learning on embedded devices," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 3, 2020, doi: 10.1109/TNNLS.2020.2979745.

[208] L. Feng and C. Zhao, "Fault description based attribute transfer for zero-sample industrial fault diagnosis," *IEEE Trans. Ind. Informat.*, early access, Apr. 20, 2020, doi: 10.1109/TII.2020.2988208.

[209] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[210] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[211] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[212] F.-C. Ghesu *et al.*, "Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 176–189, Jan. 2019.

[213] B. Thananjeyan, A. Garg, S. Krishnan, C. Chen, L. Miller, and K. Goldberg, "Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning," in *Proc. IEEE Int. Conf. Robot. Autom.*, Singapore, 2017, pp. 2371–2378.

[214] Y.-R. Shiue, K.-C. Lee, and C.-T. Su, "Real-time scheduling for a smart factory using a reinforcement learning approach," *Comput. Ind. Eng.*, vol. 125, pp. 604–614, Nov. 2018.

[215] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[216] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grid," *IEEE Trans. Ind. Informat.*, early access, Jul. 6, 2020, doi: 10.1109/TII.2020.3007167.

[217] A. Somov *et al.*, "Pervasive agriculture: IoT-enabled greenhouse for plant growth control," *IEEE Pervasive Comput.*, vol. 17, no. 4, pp. 65–75, Oct./Dec. 2018.

[218] J. Binas, L. Luginbuehl, and Y. Bengio, "Reinforcement learning for sustainable agriculture," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2019, pp. 1–3.

[219] L. Zhao, J. Wang, J. Liu, and N. Kato, "Routing for crowd management in smart cities: A deep reinforcement learning perspective," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 88–93, Apr. 2019.

[220] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016. [Online]. Available: arXiv:1610.05492.

[221] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 35–44, 2020.

[222] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *Int. J. Med. Informat.*, vol. 112, pp. 59–67, Apr. 2018.

[223] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[224] W. Wang and T. E. Daniels, "Building evidence graphs for network forensics analysis," in *Proc. 21st Annu. Comput. Security Appl. Conf.*, Tucson, AZ, USA, 2005, p. 11.

[225] Y. Chen, J. Kuang, D. Cheng, J. Zheng, M. Gao, and A. Zhou, "AgriKG: An agricultural knowledge graph and its applications," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2019, pp. 533–537.

[226] N. Bendakir and E. Aïmeur, "Using association rules for course recommendation," in *Proc. AAAI Workshop Educ. Data Min.*, vol. 3, 2006, pp. 1–10.

[227] A. Banerjee, R. Dalal, S. Mittal, and K. P. Joshi, *Generating Digital Twin Models Using Knowledge Graphs for Industrial Production Lines*, UMBC Inf. Syst. Dept., Baltimore, MD, USA, 2017.

[228] L. Wang *et al.*, "Knowledge representation and general Petri net models for power grid fault diagnosis," *IET Gener. Transm. Distrib.*, vol. 9, no. 9, pp. 866–873, Jun. 2015.

[229] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surveys*, vol. 53, no. 4, pp. 1–37, 2020.

[230] H. Jiang, X. Dai, D. W. Gao, J. J. Zhang, Y. Zhang, and E. Muljadi, "Spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification, and impact causal analysis," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2525–2536, Apr. 2016.

[231] J. F. May and C. L. Baldwin, "Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 12, no. 3, pp. 218–224, 2009.

[232] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 981–990.

[233] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2376–2384.

[234] J. Denzler, M. Zobel, and H. Niemann, "Information theoretic focal length selection for real-time active 3-D object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, 2003, pp. 400–407.

[235] E. Sommerlade and I. Reid, "Information-theoretic active scene exploration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–7.

[236] X. Zhang and M. Lapata, "Chinese poetry generation with recurrent neural networks," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2014, pp. 670–680.

[237] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *Advances in Neural Information Processing Systems*, 2019, pp. 887–897.

[238] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation—A survey," 2017. [Online]. Available: arXiv:1709.01620.

[239] P. Tsarouchi, S. Makris, and G. Chryssolouris, "Human–robot interaction review and challenges on task planning and programming," *Int. J. Comput. Integr. Manuf.*, vol. 29, no. 8, pp. 916–931, 2016.

[240] T. Zhang *et al.*, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Brisbane, QLD, Australia, 2018, pp. 1–8.

[241] W. Yang, S. Wang, G. Zheng, J. Yang, and C. Valli, "A privacy-preserving lightweight biometric system for Internet of Things security," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 84–89, Mar. 2019.

[242] O. Bulan, V. Kozitsky, P. Ramesh, and M. Shreve, "Segmentation- and annotation-free license plate recognition with deep localization and failure identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2351–2363, Sep. 2017.

[243] P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, R. Jidhesh, and T. Vadivel, "Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things," *Clust. Comput.*, vol. 22, no. 4, pp. 7733–7744, 2019.

[244] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6120–6127.

[245] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 6262–6271.

[246] S. He, M. Han, N. Patel, and Z. Li, "Converting handwritten text to editable format via gesture recognition for education," in *Proc. 51st ACM Techn. Symp. Comput. Sci. Educ.*, 2020, p. 1369.

[247] E. Brule, G. Bailly, A. Brock, F. Valentin, G. Denis, and C. Jouffrais, "MapSense: Multi-sensory interactive maps for children living with visual impairments," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 445–457.

[248] M. M. Hosseini, A. Umunnakwe, M. Parvania, and T. Tasdizen, "Intelligent damage classification and estimation in power distribution poles using unmanned aerial vehicles and convolutional neural networks," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3325–3333, Jul. 2020.

[249] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4358–4366.

[250] K. Chen, J. Hu, and J. He, "Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1748–1758, May 2018.

[251] N. Chebrolu, T. Läbe, and C. Stachniss, "Robust long-term registration of UAV images of crop fields for precision agriculture," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3097–3104, Oct. 2018.

[252] D. Jain *et al.*, "Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 241–250.

[253] E. Mocanu *et al.*, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

[254] D. Le-Phuoc, H. N. M. Quoc, H. N. Quoc, T. T. Nhat, and M. Hauswirth, "The graph of things: A step towards the live knowledge graph of connected things," *J. Web Semantics*, vols. 37–38, pp. 25–35, Mar. 2016.

[255] J. Pan, R. Jain, S. Paul, T. Vu, A. Saifullah, and M. Sha, "An Internet of Things framework for smart energy in buildings: Designs, prototype, and experiments," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 527–537, Dec. 2015.

[256] H. Jiang, Z. Zhang, J. Dang, and L. Wu, "A novel 3-D massive MIMO channel model for vehicle-to-vehicle communication environments," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 79–90, Jan. 2018.

[257] D. Jiang, L. Huo, Z. Lv, H. Song, and W. Qin, "A joint multi-criteria utility-based network selection approach for vehicle-to-infrastructure networking," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3305–3319, Oct. 2018.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.

[258] C. Lee, Y. Lv, K. Ng, W. Ho, and K. Choy, "Design and application of Internet of Things-based warehouse management system for smart logistics," *Int. J. Prod. Res.*, vol. 56, no. 8, pp. 2753–2768, 2018.

[259] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 2110–2118.

[260] Z. Chen, W. Ouyang, T. Liu, and D. Tao, "A shape transformation-based dataset augmentation framework for pedestrian detection," 2019. [Online]. Available: arXiv:1912.07010.

[261] H.-N. Hu *et al.*, "Joint monocular 3D vehicle detection and tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 5390–5399.

[262] T. Q. Trung and N.-E. Lee, "Flexible and stretchable physical sensor integrated platforms for wearable human-activity monitoring and personal healthcare," *Adv. Mater.*, vol. 28, no. 22, pp. 4338–4372, 2016.

[263] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.

[264] Y. Gao, J. Lin, J. Xie, and Z. Ning, "A real-time defect detection method for digital signal processing of industrial inspection applications," *IEEE Trans. Ind. Informat.*, early access, Jul. 31, 2020, doi: 10.1109/TII.2020.3013277.

[265] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[266] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[267] A. Verma, G. Dasgupta, T. K. Nayak, P. De, and R. Kothari, "Server workload analysis for power minimization using consolidation," in *Proc. Conf. USENIX Annu. Techn. Conf.*, 2009, p. 28.

[268] H. Yuan, J. Bi, M. Zhou, Q. Liu, and A. C. Ammari, "Biobjective task scheduling for distributed green data centers," *IEEE Trans. Autom. Sci. Eng.*, early access, Jan. 7, 2020, doi: 10.1109/TASE.2019.2958979.

[269] Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 1333–1345, 2018.

Jing Zhang (Member, IEEE) received the B.Sci. degree from Henan University, Kaifeng, China, in 2010, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2015.

He is a Research Fellow of Computer Vision with the School of Computer Science and the Faculty of Engineering, University of Sydney, Darlington, NSW, Australia. He has published several papers on top tier conferences and journals, including CVPR, ICCV, NeurIPS, AAAI, Multimedia, SIGIR, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His research interests include computer vision and deep learning.

Dr. Zhang serves as a Reviewer for a number of journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *International Journal of Computer Vision*, CVPR, ECCV, NeurIPS, ICLR, AAAI, IJCAI, and ACM Multimedia.

Dacheng Tao (Fellow, IEEE) is currently a Professor of Computer Science and an ARC Laureate Fellow with the School of Computer Science and the Faculty of Engineering, University of Sydney, Darlington, NSW, Australia. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences.

Prof. Tao received the 2015 Australian Scopus-Eureka Prize and the 2018 IEEE ICDM Research Contributions Award. He is a Fellow of the Australian Academy of Science, AAAS, and ACM.

Authorized licensed use limited to: National Central University. Downloaded on May 05,2024 at 10:17:34 UTC from IEEE Xplore. Restrictions apply.