Artificial Intelligence of Things: A Survey

SHAKHRUL IMAN SIAM, The Ohio State University, USA
HYUNHO AHN, The Ohio State University, USA
LI LIU, Michigan State University, USA
SAMIUL ALAM, The Ohio State University, USA
HUI SHEN, The Ohio State University, USA
ZHICHAO CAO, Michigan State University, USA
NESS SHROFF, The Ohio State University, USA
BHASKAR KRISHNAMACHARI, University of Southern California, USA
MANI SRIVASTAVA, University of California, Los Angeles, USA
MI ZHANG, The Ohio State University, USA

The integration of the Internet of Things (IoT) and modern Artificial Intelligence (AI) has given rise to a new paradigm known as the Artificial Intelligence of Things (AIoT). In this survey, we provide a systematic and comprehensive review of AIoT research. We examine AIoT literature related to sensing, computing, and networking & communication, which form the three key components of AIoT. In addition to advancements in these areas, we review domain-specific AIoT systems that are designed for various important application domains. We have also created an accompanying GitHub repository, where we compile the papers included in this survey: https://github.com/AIoT-MLSys-Lab/AIoT-Survey. This repository will be actively maintained and updated with new research as it becomes available. As both IoT and AI become increasingly critical to our society, we believe AIoT is emerging as an essential research field at the intersection of IoT and modern AI. We hope this survey will serve as a valuable resource for those engaged in AIoT research and act as a catalyst for future explorations to bridge gaps and drive advancements in this exciting field.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Computing methodologies → Artificial intelligence**; • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Artificial Intelligence of Things, AIoT, Edge AI

ACM Reference Format:
Shakhrul Iman Siam, Hyunho Ahn, Li Liu, Samiul Alam, Hui Shen, Zhichao Cao, Ness Shroff, Bhaskar Krishnamachari, Mani Srivastava, and Mi Zhang. 2024. Artificial Intelligence of Things: A Survey. *ACM Trans. Sensor Netw.* , (August 2024), 70 pages. https://doi.org/10.1145/3690639

Authors' addresses: Shakhrul Iman Siam, The Ohio State University, USA, siam.5@osu.edu; Hyunho Ahn, The Ohio State University, USA, ahn.377@osu.edu; Li Liu, Michigan State University, USA, liuli9@msu.edu; Samiul Alam, The Ohio State University, USA, alam.140@osu.edu; Hui Shen, The Ohio State University, USA, shen.1780@osu.edu; Zhichao Cao, Michigan State University, USA, caozc@msu.edu; Ness Shroff, The Ohio State University, USA, shroff.11@osu.edu; Bhaskar Krishnamachari, University of Southern California, USA, bkrishna@usc.edu; Mani Srivastava, University of California, Los Angeles, USA, mbs@ucla.edu; Mi Zhang, The Ohio State University, USA, mizhang.1@osu.edu.

1550-4859/2024/August-ART $15.00
https://doi.org/10.1145/3690639

1 INTRODUCTION

The proliferation of the Internet of Things (IoT) such as smartphones, wearables, drones, and smart speakers, as well as the gigantic amount of data they capture, have revolutionized the way we work, live, and interact with the world. Equipped with sensing, computing, networking, and communication capabilities, these devices are able to collect, analyze and transmit a wide range of data including images, videos, audio, texts, wireless signals, physiological signals from individuals and the physical world. In recent years, advancements in Artificial Intelligence (AI), particularly in deep learning (DL)/deep neural network (DNN), foundation models, and Generative AI, have propelled the integration of AI with IoT, making the concept of **Artificial Intelligence of Things (AIoT)** a reality. The synergy between IoT and modern AI enhances decision making, improves human-machine interactions, and facilitates more efficient operations, making AIoT one of the most exciting and promising areas that have the potential to fundamentally transform how people perceive and interact with the world.

As illustrated in Figure 1, at its core, AIoT is grounded on three key components: sensing, computing, and networking & communication. Specifically, AIoT utilizes a variety of onboard sensors such as cameras, microphones, motion and physiological sensors to collect data from individuals and the physical world. The collected sensor data are processed by modern AI algorithms for a variety of tasks such as classification, localization, anomaly detection, and many others. Lastly, the networking & communication component of AIoT ensures the reliable transmission of the sensor data and/or the computed outcomes to the cloud, edges or other nearby AIoT devices. Compared to conventional IoT, the computing component of AIoT is concentrated on AI-oriented compute tasks. Moreover, the sensing and networking & communication components of AIoT are AI empowered. It is these two key distinctions that allow AIoT to empower billions of everyday devices with breakthroughs brought by modern AI.

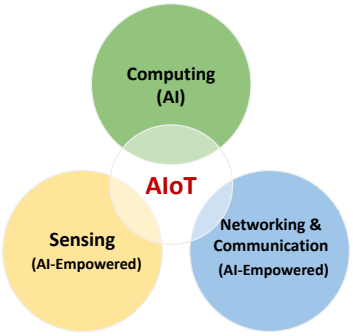

Fig. 1. Overview of AIoT.

Besides advancements in the three key components, domain-specific AIoT systems have been proposed and developed across a wide range of application domains. For example, in the domain of healthcare, AIoT systems enable remote patient monitoring, facilitate disease diagnosis on site, and act in the form of assistive technology that helps people with disabilities. In the domain of Augmented, Virtual, and Mixed Reality, AIoT systems enable 3D tracking to provide immersive user experiences. In the domain of video streaming and analytics, AIoT systems have been developed to enhance video quality and optimize video processing efficiency. All these developed domain-specific systems demonstrate the potential of AIoT on revolutionizing a wide range of industries.

The overarching goal of this survey is to provide a systematic and comprehensive review of AIoT research. As shown in Figure 2, we organize the literature of AIoT in a taxonomy consisting of four main categories: **sensing**, **computing**, **networking & communication**, and **domain-specific AIoT systems**. Specifically,

- **Sensing:** Sensing serves as the foundation of AIoT. In §2, we survey AI-empowered sensing mechanisms and techniques in AIoT that cover research directions related to motion sensing, wireless sensing, vision sensing, acoustic sensing, multi-modal sensing, earable sensing, and Generative AI for sensing.
- **Computing:** Computing is the brain of AIoT. In §3, we survey fundamental compute tasks that lie at the core of AIoT, covering topics related to on-device inference, offloading, on-device training, federated learning, and AI agents for AIoT.
- **Networking & Communication**: Networking and communication serve as the backbone of AIoT. In §4, we survey AI-empowered networking and communication techniques related to a variety of networks including cellular/mobile networks, Wi-Fi networks, visible light communication, and LoRa/LoRaWAN.
- **Domain-specific AIoT Systems**: The advancements in sensing, computing, networking & communication lay the foundation for the development of AIoT systems designed for specific application domains. In §5,

we survey these AIoT systems in important application domains including healthcare and well-being, video streaming and analytics, autonomous driving, as well as augmented, virtual, and mixed reality.

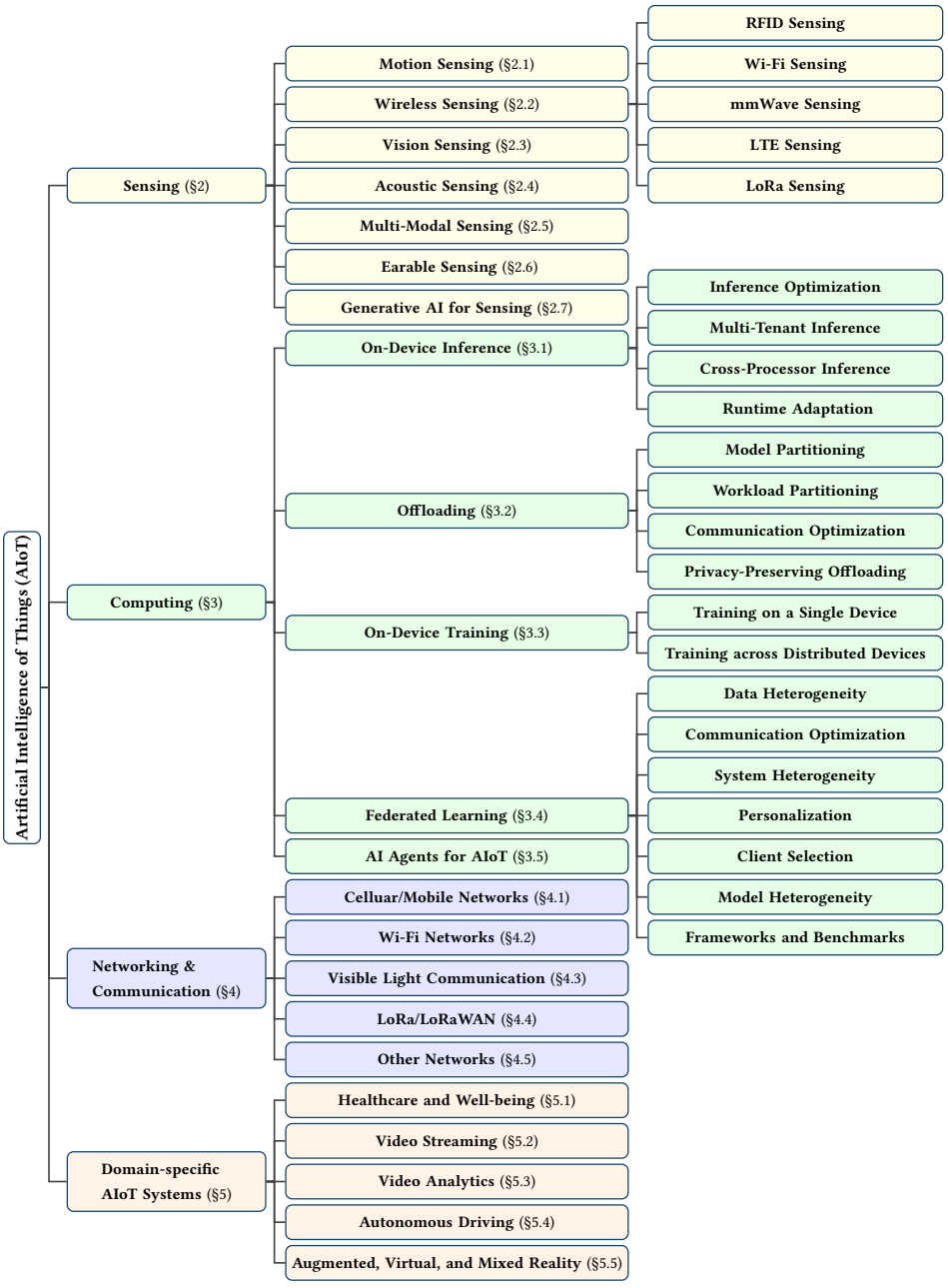

Fig. 2. Taxonomy of Artificial Intelligence of Things (AIoT) literature.

Sensing Type	Contact Type	Computation Requirement	Privacy-Intrusive	Range	Advantages	Disadvantages
Motion Sensing	Contact	Lightweight	Low	Short	Low power, cost-effective	Require body contact, sensitive to location, limited information
Wireless Sensing	Contactless	Heavy	Low	Medium to Long	Can penetrate walls, large coverage area	High computational cost, signal interference from other devices
Vision Sensing	Contactless	Heavy	High	Medium to Long	Rich information, versatile applications	Privacy concerns, high computational cost, sensitive to lighting conditions and occlusions
Acoustic Sensing	Contactless	Lightweight	High	Short to Medium	Rich information, versatile applications	Privacy concerns, affected by background noises
Multi-Modal Sensing	Both	Heavy	Variable	Variable	Combine strengths of multiple sensors, robust	Complex integration, high computational cost
Earable Sensing	Contact	Lightweight	Variable	Short	Close proximity to signal sources, versatile applications	Limited to what can be measured at the ear, sensitive to ambient noises and ear positioning

Table 1. Comparison of different sensing modalities.

We have established a **GitHub repository** to organize the papers featured in the survey at https://github.com/AIoT-MLSys-Lab/AIoT-Survey. We will actively maintain the repository and incorporate new research as it emerges.

Although there are several surveys on topics relevant to AIoT [21, 30, 89, 94, 168, 196, 233, 324, 329], they focus on some specific aspects of AIoT. In contrast, this survey provides a holistic view of AIoT research. More importantly, *we primarily focus on literature on sensing, computing, networking & communication, and domain-specific AIoT systems that are built upon modern AI techniques such as DL, foundation models, and Generative AI.* We hope this survey along with the GitHub repository could serve as valuable resources to help researchers and practitioners gain a comprehensive understanding of AIoT research and inspire them to contribute to this important and exciting field.

2 SENSING

2.1 Motion Sensing

Motion sensing involves the use of motion sensors such as Inertial Measurement Unit (IMU) sensors (i.e., accelerometers, gyroscopes, and magnetometers) attached to the individuals to capture various types of motions such as arm postures, body movements, and physical activities. As summarized in Figure 3, depending on the sensing tasks, existing works on AI-empowered motion sensing can be grouped into two categories: human activity recognition, and arm tracking.

Human Activity Recognition. One of the most important tasks of motion sensing is human activity recognition (HAR). Most existing HAR frameworks are limited to a few predefined activities and require prior knowledge or labeled data for supervised training. To address this limitation, Liu et al. [160] introduce Lasagana, an unsupervised learning-based HAR framework that extracts common bases of human motions in an unsupervised manner, creating a universal multi-resolution representation for common human activities. Their prototype system achieves 98.9% precision in activity classification and nearly 100% recall with about 90% precision in activity indexing. Another major limitation of existing motion sensing-based HAR frameworks is that machine learning

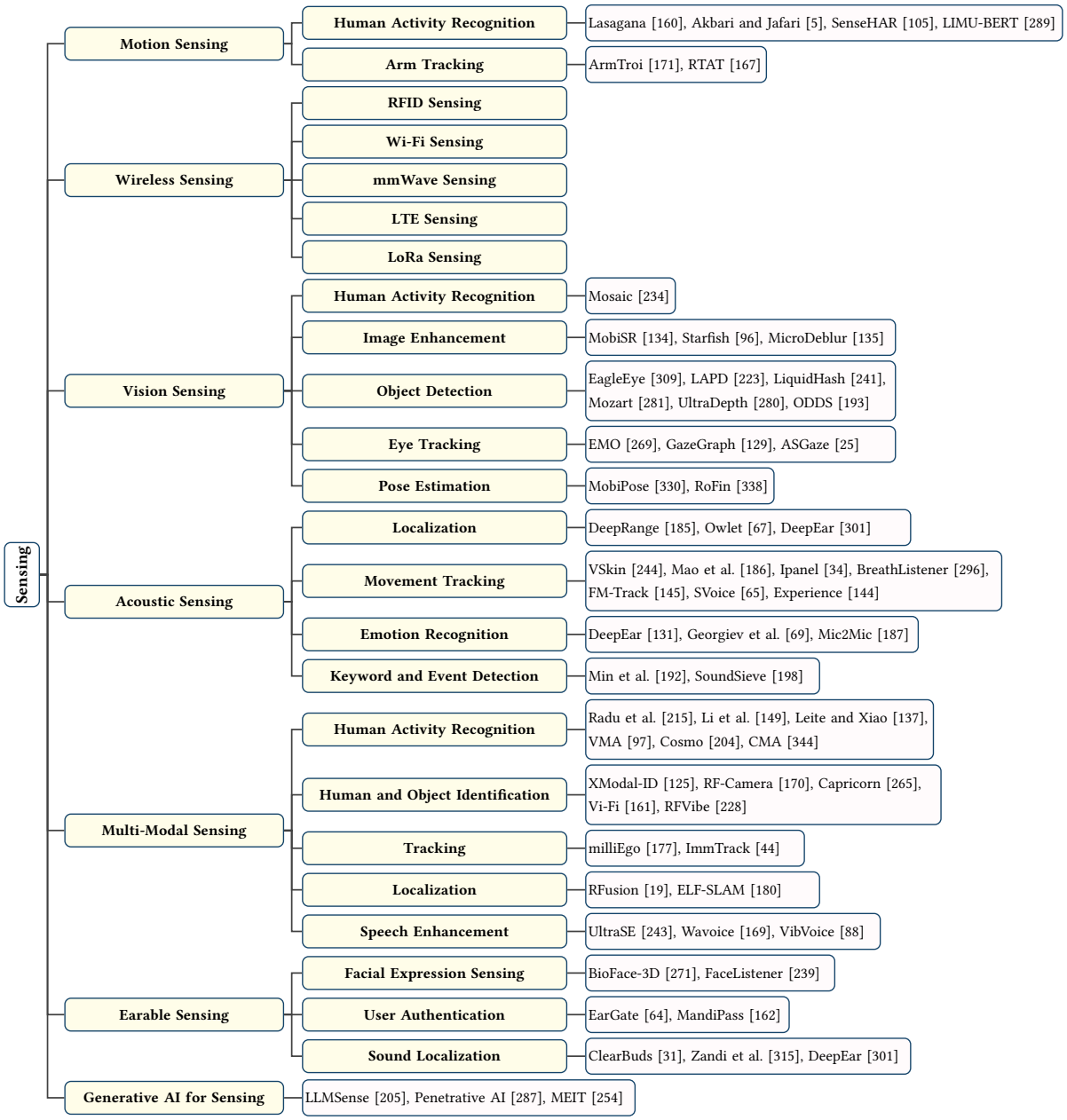

Fig. 3. Summary of topics related to sensing.

(ML) algorithms trained on specific sensors require retraining upon any system configuration changes, such as adding a new sensor. To address this limitation, Akbari and Jafari [5] propose a training scheme for the newly added sensors to identify human activities that were previously detected by existing sensors. As another line of

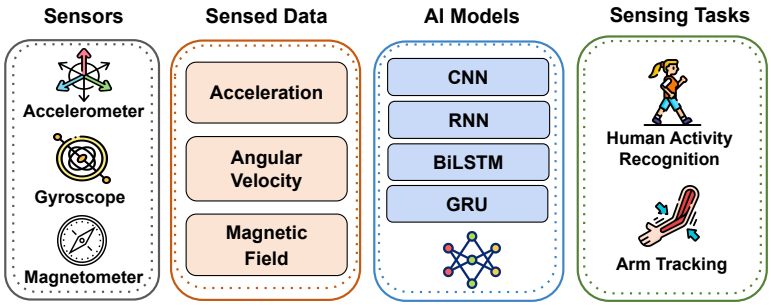

Fig. 4. Illustration of AI-empowered motion sensing pipeline.

research, Jeyakumar et al. [105] target the device heterogeneity problem, which encompasses variations in sensor types, data formats, and sampling rates, leading to lower activity recognition performance in real-life scenarios. To address this issue, they propose a DL-based HAR framework named SenseHAR, which allows sensor fusion while being robust to device heterogeneity. SenseHAR offers easy calibration for new devices, allowing seamless integration and utilization of different devices with varying sampling frequencies, sensors, and applications. Xu et al. [289] tackle the challenges of limited labeled data and device placement diversity in HAR. They propose LIMU-BERT, a lightweight DL-based HAR framework that employs self-supervised learning to extract general features from unlabeled sensor data. It adopts the key principles of the BERT framework for motion sensing and a classifier consisting of three stacked Gated Recurrent Units (GRU). The model's efficiency and ability to learn robust features make it suitable for real-time applications on mobile devices.

Arm Tracking. Another important task of motion sensing is arm tracking, which uses motion sensors to track the movements, positions, and posture of an individual's arm. Most arm tracking systems require attaching multiple sensors to an individual's arm, which can limit flexibility and have a negative impact on the overall user experience. To address this issue, Liu et al. [171] propose ArmTroi, a real-time 3D arm skeleton tracking system that uses a single motion sensor worn on the wrist. ArmTroi adopts an attention and recurrent neural network (RNN)-based network, which is lightweight and suitable for mobile and real-time applications. The authors also prototype the system on LG smartwatches, Google Glass, and Samsung Galaxy S7. ArmTroi achieves real-time arm tracking with 92.7% gesture recognition precision, and demonstrates its efficacy through fitness and gesture-based control applications. As another line of research, the differences among accelerometers, gyroscopes, and magnetometers of the IMU sensors as well as the heavy computation costs incurred by DL models make it challenging to leverage all of these sensors for accurate and real-time arm tracking. To address this issue, Liu et al. [167] propose RTAT, a real-time arm tracking system that utilizes a Bidirectional Long Short-Term Memory (BiLSTM)-based multitask neural network to track both the orientation and location of an arm simultaneously. RTAT also incorporates an attention mechanism to dynamically learn the importance of different IMU sensor streams to achieve high accuracy and low latency.

2.2 Wireless Sensing

Wireless sensing uses wireless signals to sense individuals and objects in the environment in a contact-free manner. As summarized in Figure 5, based on the frequency bands wireless signals belong to, existing works on AI-empowered wireless sensing can be grouped into five categories: RFID sensing, Wi-Fi sensing, mmWave sensing, LTE sensing, and LoRa sensing.

2.2.1 RFID Sensing. Radio Frequency Identification (RFID) is a technology that employs an RFID tag and reader, enabling the retrieval of information from the tag using radio frequency (RF) signals emitted by the reader. By

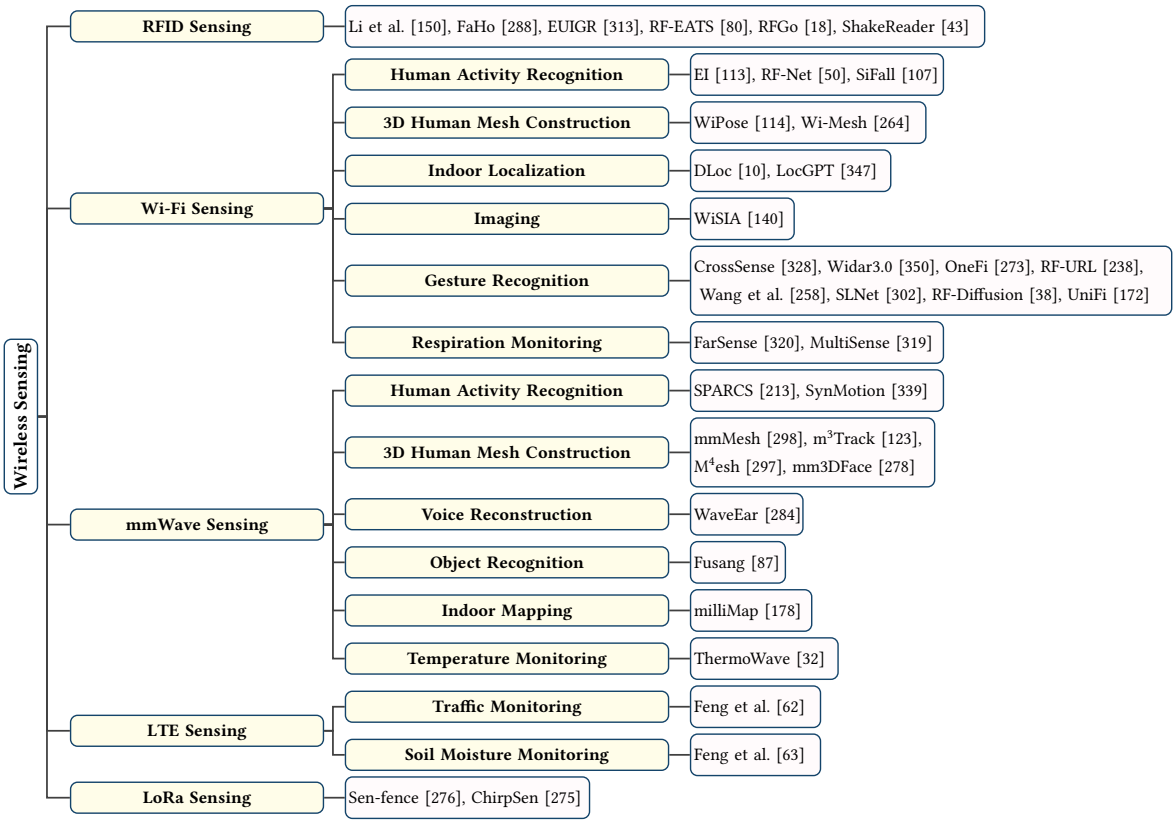

Fig. 5. Summary of topics related to wireless sensing.

attaching RFID tags to individuals or objects, RFID can be deployed for sensing tasks such as localization, object tracking, and classification. Li et al. [150] utilize an RFID sensing system for activity recognition in the medical environment by attaching RFID tags to objects in clinical settings and recording the Received Signal Strength (RSS) from these tags. These collected data subsequently serve as the input for a convolutional neural network (CNN), enabling the recognition of activities that involve the usage of certain objects. While this approach effectively identifies activities using RFID, the received signal comprises both the Line of Sight (LOS) signal and multiple reflections from obstacles. This complicates the localization task, making it challenging to determine which signal accurately represents the RFID tag's location In response to this issue, Xu et al. [288] introduce an algorithm that transforms the RFID signal into a hologram that encapsulates the probable location of the tag. A CNN is then employed to accurately identify the tag's actual position within this hologram. The accuracy of existing RFID systems is significantly impacted by the subjects and the surrounding environmental conditions. In the context of gesture recognition, some studies consider environmental variations but often neglect the impact on the user. To address this issue, Yu et al. [313] develop a discriminator DNN, which identifies the user and its environment in the data. Simultaneously, it also has gesture labeling DNN, which predicts the probability of gestures. Through adversarial training of both DNNs, the gesture labeling DNN learns to create representations that are indistinguishable from the domain discriminator, resulting in a gesture recognizer that is independent of the user and environment. Ha et al. [80] introduces RF-EATS, a system designed to noninvasively sense food

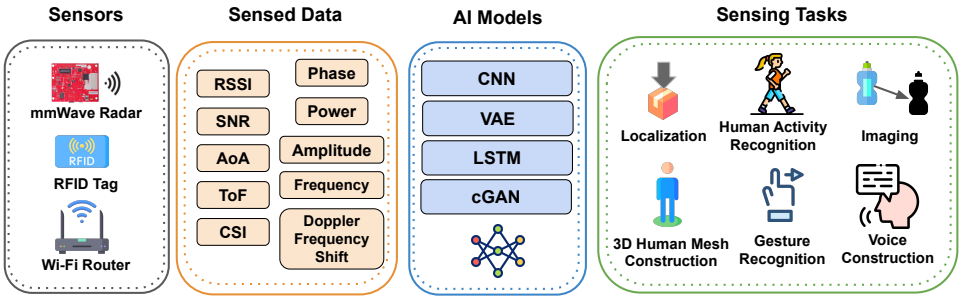

Fig. 6. Illustration of AI-empowered wireless sensing pipeline.

and liquids within closed containers using passive RFID tags. The authors attach the RFID tag to the liquids and detect whether this liquid is fake or not. To manage the diversity in environmental conditions, the study employs Variational Autoencoders (VAE) to synthesize multiple samples. A classifier is then trained to distinguish counterfeit liquids using these augmented datasets. Bocanegra et al. [18] design an RFID reader system capable of simultaneous multi-tag reading via an array of deployed antennas. To determine whether an RFID tag is within the checkout area, they also utilize a neural network, training it in a supervised manner using data captured from the reader. Ultra-high-frequency (UHF) RFID is more appealing to retailers because it can rapidly scan multiple RFID-tagged items, substantially increasing operational efficiency; however, smartphones currently lack direct communication capabilities with UHF RFID tags. To bridge the gap, Cui et al. [43] introduce ShakeReader, a system designed to enhance interaction between smartphones and UHF RFID-tagged items without requiring hardware modifications to existing RFID systems or smartphones. ShakeReader enables users to obtain item-specific information by performing predefined gestures, such as shaking the smartphone near the RFID tag. The system utilizes a reflector polarization model to analyze the backscattered signal from the tag, which is affected by the smartphone's gestures. This model accounts for both the signal propagation and the polarization changes caused by the reflection from the smartphone, enabling the detection of specific gestures using the RFID reader even with a single tag.

2.2.2 Wi-Fi Sensing. Wi-Fi sensing takes advantage of the ubiquitous Wi-Fi signals and their associated hardware to detect and interpret human movements or changes in the environment. Depending on the sensing tasks, existing works on AI-empowered Wi-Fi sensing can be grouped into the following categories.

Human Activity Recognition. One important task of Wi-Fi sensing is human activity recognition (HAR). The major challenge in device-free human activity recognition is that wireless signals are highly influenced by the specific environment and individual characteristics of the human subject, leading to poor generalization of models across different subjects and environments. To address this challenge, Jiang et al. [113] propose EI for HAR that learns domain-independent features from activity data collected in different domains. EI accepts multiple types of input signals, including Wi-Fi Channel State Information (CSI). The DL model of EI incorporates an adversarial network, including a CNN-based feature extractor, an FC-layer-based activity recognizer that predicts activity type from extracted features, and a domain discriminator that predicts the domain. Ding et al. [50] present RF-Net, a metric-based meta-learning approach for one-shot human activity recognition using Wi-Fi that can perform the recognition in a new environment with only one observation per label. RF-Net classifies a new observation in new environments by calculating a weighted sum of all the labels in the training dataset. The weights are given by the similarity between the query observation and all the data in the support dataset of the new environment. Lastly, Ji et al. [107] propose SiFall to formulate the fall detection problem as adaptive anomaly detection out of normal repeatable human activities instead of seeking features to characterize fall activity.

3D Human Mesh Construction. 3D human mesh construction in Wi-Fi sensing refers to the creation of three-dimensional representations of the human body using Wi-Fi signals. Jiang et al. [114] present WiPose, a 3D human pose skeleton construction framework that recovers human joints on both limbs and torso of the human body using commercial Wi-Fi devices. WiPose records CSI using a single antenna transmitter with multiple distributed receivers and designs an LSTM-based deep learning model that accepts the sequence of Doppler Frequency Shift (DFS) profile transformed from non-overlapping CSI segments and outputs a series of features. The learned features from LSTM are regarded as the rotation of human body joints and fed to the forward kinematics layers to calculate the actual joint locations based on a given skeletal structure. Wang et al. [264] present Wi-Mesh, which further improves the 3D human mesh construction task with DNN based on GRU and self-attention. Wi-Mesh leverages a commodity 3-antenna transmitter and two receivers with 9 antennas in an L shape to record CSI. Received signals at the specific antenna array can be used to calculate the 2D AoA of the signal reflections based on phase shift, providing spatial information about the objects and environment. Wi-Mesh generates thirty 2D AoA spectrums per second and extracts only human images by subtracting the static components in consecutive images since the human body is moving. Wi-Mesh tracks way more body locations than WiPose and also outperforms WiPose with an average joint location error of 2.4cm and body vertices location error of 2.81cm, though using more complicated antenna arrays.

Indoor Localization. Indoor localization in Wi-Fi sensing refers to the process of using Wi-Fi signals to determine the position of objects or individuals within indoor environments. Unlike outdoor localization, which commonly relies on GPS (Global Positioning System), indoor localization requires a different set of technologies and methodologies due to challenges such as the unavailability of GPS signals indoors, multi-path reflection, and interference from walls and other structures. Thus, indoor localization remains a "last-mile" problem when forming a positioning system without blind spots. Wi-Fi has been broadly utilized to address the indoor localization problem. Ayyalasomayajula et al. [10] present DLoc, a DL-based wireless localization algorithm and an automated mapping platform MapFind, which altogether forms a positioning system with a map inspired by outdoor localization services. MapFind constructs location-tagged maps of the environment and generates training data for DLoc. Together, they solve the active indoor localization scenario in which off-the-shelf Wi-Fi devices like smartphones can access a map of the environment and estimate their position by sending packets to surrounding Wi-Fi access points with respect to that map. While deep learning approaches for indoor localization rely on high-quality training samples and are hard to adapt to varied scenarios, Zhao et al. [347] propose LocGPT, which is a specialized Generative Pre-training Transformer (GPT) variant that excels in generating profound contextual insights, to explore the underlying principles of indoor localization. The model is configured with 36 million parameters tailored for transfer learning. To facilitate the benchmarking, training, and transfer learning in indoor localization, they have established Ray, the first 3D indoor localization dataset on a scale of millions, including RFID, Wi-Fi, and BLE samples. LocGPT achieves near-par accuracy when fine-tuned with merely half the conventional dataset, which shows its superiority in transfer learning within the indoor localization domain.

Imaging. Wi-Fi imaging exploits the capabilities of Wi-Fi signals to create images of objects or humans in the environment. Li et al. [140] present a Wi-Fi imaging system WiSIA that is capable of simultaneously detecting and segmenting objects and humans within the imaging plane using commodity Wi-Fi devices. WiSIA leverages two receivers with three orthogonal antennas sharing the same transmitter antenna as the imaging model on the object side to record CSI that contains the changes in the Wi-Fi signal of both amplitude and phase. WiSIA incorporates a conditional Generative Adversarial Network (cGAN) to refine the boundaries in an image-to-image translation fashion. WiSIA achieves 0.9 in similarity and tagging accuracy for all five tested objects which is comparable to the state-of-the-art computer vision and acoustics imaging while outperforming the state-of-the-art vision-based method in conditions with darkness or obstructions.

Gesture Recognition. Wi-Fi signals can be used in the gesture recognition task by analyzing the variations in the Wi-Fi signal caused by human body movements. Zhang et al. [328] propose CrossSense, a system designed to improve the scalability and efficiency of WiFi-based gesture recognition. The primary challenge addressed is the need for extensive, site-specific training data collection, which is labor-intensive and impractical for large-scale deployments. CrossSense tackles this by using machine learning to generate synthetic training samples from existing measurements, allowing these samples to be effectively used across different environments. Zheng et al. [350] propose Widar3.0, a Wi-Fi-based zero-effort cross-domain gesture recognition system. Widar3.0 calculates the body-coordinate velocity profile (BVP) of gestures from CSI at the lower signal level, which represents power distribution over different velocities and is unique from gesture to gesture while independent from the domain. On this basis, Widar3.0 adopts a one-fits-all model based on CNN, GRU, and dense layers that requires only one-time training but can adapt to different data domains. Similar to Widar3.0, OneFi [273] proposes to use velocity distribution which can be derived from DFS as the unique feature that describes a gesture. It adopts a backbone based on self-attention, noted as Wi-Fi Transformer, as the gesture recognition framework. To avoid model re-training, OneFi adopts a lightweight one-shot learning framework based on transductive fine-tuning and opens up a new direction for one-shot (or few-shot) learning in Wi-Fi-based gesture recognition. Song et al. [238] present RF-URL, an unsupervised representation learning framework for human gesture recognition tasks. RF-URL combines signal-processing-based RF sensing with learning-based RF sensing by using a contrastive framework. Experimental results indicate that RF-URL pre-training model is capable of extracting general information for gesture recognition and applying it effectively across different datasets. Wang et al. [258] carry out an in-depth study on the domain variation problem in Wi-Fi-based gesture recognition task, which can alter multi-path effects and introduce noise into wireless signals. These variations, including changes in the environment, can lead to significant performance degradation in Wi-Fi sensing applications due to the resulting fluctuations in wireless signal patterns. To mitigate these effects, the authors propose a robust framework based on conformal prediction, which quantifies the similarity between testing and training data without the need for retraining or generating new features. Yang et al. [302] propose SLNet, an architecture for enhancing wireless sensing applications through the integration of deep learning and spectrogram analysis. SLNet utilizes neural networks to generate super-resolution spectrograms, addressing the limitations of traditional time-frequency uncertainty. This design improves the accuracy of Wi-Fi-based gesture recognition, human identification, fall detection, and breathing estimation tasks. Experiments demonstrate that SLNet achieves superior performance with reduced computational demands, making it suitable for practical deployment on edge devices. Chi et al. [38] introduce RF-Diffusion, a novel approach to generating high-quality, time-series radio frequency (RF) data using diffusion models. The proposed methodology involves training RF-Diffusion with a real-world dataset to generate synthetic RF signals of the designated type. These synthetic samples are then integrated with the original dataset, and collectively employed to train the wireless sensing model. The authors highlight that RF-Diffusion when used as a data augmentation tool, leads to substantial improvements in Wi-Fi-based gesture recognition accuracy. This enhancement is attributed to the model's ability to produce diverse and high-quality RF data that enriches the training datasets of existing systems.

Respiration Monitoring. Wi-Fi signals can be used for respiration monitoring by analyzing the subtle variations in wireless signals caused by the movement of a person's chest during breathing. Existing methods of respiration monitoring are limited by short sensing ranges, susceptibility to noise, and issues with phase offset stability. To overcome these limitations, Zeng et al. [320] introduce FarSense, a system for enhancing Wi-Fi-based respiration sensing. FarSense leverages the CSI ratio from two antennas to overcome the limitations of existing methods that rely on individual CSI readings. By using the CSI ratio, FarSense cancels out most of the noise and phase offset issues, significantly extending the sensing range. The system combines the amplitude and phase information of the CSI ratio to address the "blind spots" problem and improves the sensitivity of detecting subtle respiration

signals. Zeng et al. [319] present MultiSense, a system for accurately monitoring the respiration patterns of multiple individuals simultaneously using commodity Wi-Fi devices. MultiSense overcomes the challenges faced in existing methods by leveraging multiple antennas on Wi-Fi devices and modeling the multi-person respiration sensing problem as a Blind Source Separation (BSS) problem. MultiSense cancels out time-varying phase offsets and removes background static signals, allowing for robust separation and continuous monitoring of detailed respiration patterns.

2.2.3 mmWave Sensing. Millimeter Wave (mmWave) sensing refers to the use of electromagnetic waves with wavelengths in the millimeter range, typically between 30 GHz and 300 GHz frequency band, for a variety of sensing tasks. The high frequency, short wavelength, and broadband capacity make mmWave more sensitive to minor reflection distance variations, and thus can provide finer sensing resolution. At the same time, mmWave has limited penetration capabilities so it can easily be attenuated or blocked by obstacles. As such, mmWave sensing often requires a direct line-of-sight between the transceivers and the sensing target. Depending on the sensing tasks, existing works on AI-empowered mmWave sensing can be grouped into the following categories.

Human Activity Recognition. The capability of mmWave signals to capture micro-motions and micro-vibrations of different human body parts makes it feasible for the task of human activity recognition (HAR). Pegoraro et al. [213] introduce SPARCS for mmWave-based HAR. It focuses on extracting micro-Doppler signatures of human movement from irregular and sparse Channel Impulse Response (CIR) samples. This approach leverages the inherent sparsity of the mmWave channel to reduce sensing overhead drastically while integrating seamlessly with existing communication protocols. By formulating micro-Doppler extraction as a sparse recovery problem, SPARCS achieves high-quality human activity recognition with significantly lower overhead compared to existing methods, demonstrating its applicability and efficiency in real-world scenarios. While research on introducing DL to mmWave-based human activity recognition achieves promising performance, collecting and labeling mmWave datasets for such tasks is difficult and expensive. To close the gap, Zhang et al. [339] present SynMotion which synthesizes mmWave signals at high quality using widely available vision-based human motion datasets with the coordinates of body skeletal points and designs a few/zero-shot synthetic-to-real transfer learning framework for downstream human activity recognition.

3D Human Mesh Construction. mmwave signals can also be used for 3D human mesh construction by providing detailed information about the human body contours and structure. Xue et al. [298] present mmMesh, a DL-based real-time 3D human mesh construction framework to model the moving subject with commercial portable mmWave devices. mmMesh utilizes range and angle information to remove noisy reflections from static objects in the IF signals collected by commercial devices and generate the 3D point clouds as input to the DL model. Kong et al. [123] propose m^3Track to enable simultaneous tracking of the 3D postures of multiple users leveraging a single commercial mmWave device. m^3Track obtains the Range-Doppler-Profile of the IF signals by range-FFT and doppler-FFT that contains information on the users and background objects. It distinguishes multiple users and backgrounds by sliding a convolutional kernel along the range bins of the Range-Doppler-Profile and performing convolution operations to detect the ranges that contain users. Xue et al. [297] develop M^4esh for multi-subject 3D human mesh reconstruction. The tracking scheme of M^4esh integrates techniques adopted by mmMesh and m^3Track, including subject detection, 3D point cloud generation for each subject, and per-subject mesh reconstruction. Similarly, Xie et al. [278] propose mm3dFace to move towards the reconstruction of human face. It proposes to leverage commercial mmWave radar to reconstruct 3D human faces that continuously express facial expressions in a passive manner. mm3dFace captures human face information from the recorded IF signal. By applying range-FFT to the IF signal and AoA calculation, it obtains the range profile, azimuth profile, and elevation profile, which together form a Range-Angle-Profile in the three-dimensional space. The three-dimensional profile captures the side view and frontal view of human faces.

Voice Reconstruction. Voice Reconstruction refers to the process of capturing and reconstructing the human voice by detecting subtle vibrations with millimeter-wave signals. Xu et al. [284] propose WaveEar, which leverages mmWave devices to enable noise-resilient speech sensing for voice-user interface (VUI) in environments with audible and inaudible interference. The authors conducted an in-depth study of human voice generation to obtain insights into voice vibration caused by the integrated effort of three physiological organs, e.g., lungs, vocal cords, and articulators. WaveEar designs a low-cost mmWave probe that employs a phased directional array to locate the speaker by throat vibration and then transmits mmWave signals towards the near-throat region of the speaker and processes the reflected signal for voice reconstruction.

Object Recognition. The broadband nature of mmWave makes it also suitable for object recognition. He et al. [87] present Fusang, a system that adopts commercial off-the-shelf mmWave devices for accurate and robust 3D object recognition. Fusang leverages the large bandwidth of mmWave radars to capture a unique set of fine-grained responses reflected by objects with different shapes. It generates the High-Resolution Range Profile (HRRP) from the IF signal and constructs two novel graph-structured features, as the HRRP data of different objects in the spectrum is not always distinguishable. Fusang extracts the set of formants that denotes the peaks in the HRRP envelope and iteratively bisects the frequency bands to a point when there is no more than one formant falling in each subband to build a binary tree with subbands that contain formants as leaf nodes.

Indoor Mapping. Indoor mapping using mmWave involves creating detailed maps or spatial representations of environments using the data obtained from mmWave radar sensors. State-of-the-art mapping approaches are mainly based on optical sensors, such as lidar and cameras. One of the advantages of mmWave over optical sensors is its ability to penetrate through certain materials and resilience to poor illumination. Lu et al. [178] present milliMap, which adopts a single-chip mmWave radar for dense indoor map generation and simple object annotation in low-visibility environments under emergency situations. milliMap adopts conditional GAN supervised by a co-located liar to generate dense patches similar to lidar ground truth from mmWave scans. In this way, milliMap overcomes the sparsity and multi-path noise of mmWave signals. It also identifies different objects from the spectral response of mmWave reflections by a CNN-based semantic recognizer.

Temperature Sensing. Temperature sensing refers to the continuous or periodic process of measuring and recording temperature levels in a given environment, object, or individual. While most wireless temperature monitoring solutions are not cost-effective and generate electronic wastes, ThermoWave [32] enables ecological, battery-less, and ultra-low-cost wireless temperature monitoring using mmWave signals. Specifically, ThermoWave is designed based on the principle of thermal scattering effect of mmWave. Specifically, it attaches ThermoTags made of cholesteryl material inked film or paper which aligns the molecular patterns at different temperatures and senses the temperature-induced pattern change by scattered mmWave signals. The ThermoTags are of low cost (less than 0.01 dollars per tag). ThermoWave adopts a mmWave-radar-based ThermoScanner to receive the temperature-modulated mmWave scattering and extract thermal features from it.

2.2.4 LTE Sensing. Long-Term Evolution (LTE) sensing leverages the capabilities of LTE wireless broadband communication technology for the task of sensing. Feng et al. [62] explore the use of LTE signals for pervasive sensing applications both indoors and outdoors. It aims to address the limitations of existing wireless sensing technologies, such as Wi-Fi, which are constrained by coverage and performance issues. Specifically, the authors propose to leverage the widespread and diverse LTE infrastructure to achieve comprehensive and reliable sensing without affecting LTE data communication. Through advanced techniques to mitigate interference and noise, the authors demonstrate the effectiveness of LTE sensing in two key applications: indoor respiration monitoring and outdoor traffic monitoring. In [63], the authors leverage the infrastructure of LTE base stations to provide a cost-effective and energy-efficient solution for the application of soil moisture monitoring. By utilizing commercial off-the-shelf hardware, including software-defined radios and a Raspberry Pi, the proposed system achieves high

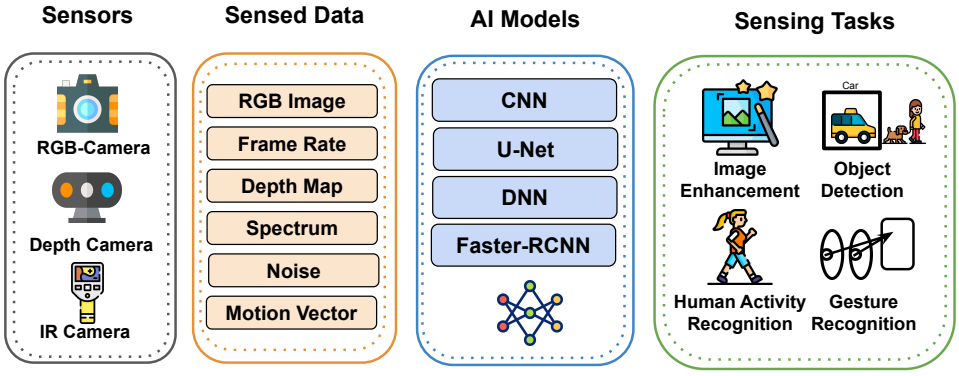

Fig. 7. Illustration of AI-empowered vision sensing pipeline.

accuracy comparable to high-end sensors but at a fraction of the cost. They have deployed their prototype system and examined its robustness across various soil types and conditions, demonstrating its potential for applications in precision agriculture and environmental monitoring.

2.2.5 LoRa Sensing. The long-range, low-power characteristics of LoRa networks make it popular among large-scale remote-area IoT applications. However, the use of LoRa for sensing tasks is yet to be explored due to challenges related to interference, sensing range, and many more. To address these challenges, Xie and Xiong [276] introduce Sen-fence, which explores advanced signal processing techniques that maximize movement-induced signal variations, thereby increasing the sensing range. Additionally, the authors introduce a novel "virtual fence" method, which confines sensing activities to a specific area of interest, thereby reducing the impact of environmental noise and interference. Sen-fence achieves a 50-meter range for fine-grained human respiration detection while effectively managing interference for practical LoRa sensing applications. Though the proposed method in Sen-fence is effective for detecting tiny movements like respiration but struggles with larger movements such as human walking. To address this issue, the authors in [275] introduce ChirpSen, a system designed to enhance the sensing range of LoRa-based localization by fully exploiting the properties of chirp signals. ChirpSen employs a chirp concentration scheme that concentrates the power of all signal samples in a LoRa chirp at one timestamp, thus increasing the signal power as well as the sensing range. Real-world experiments demonstrate that ChirpSen significantly enhances detection capabilities, extending the range for monitoring human respiration at a distance of 138 meters and tracking a walking human at up to 210 meters.

2.3 Vision Sensing

Vision sensing involves the use of vision sensors such as RGB cameras, depth cameras, and near-infrared (NIR) image sensors to capture and analyze visual information for various sensing tasks. As summarized in Figure 3, depending on the sensing tasks, existing works on AI-empowered vision sensing can be grouped into five categories: human activity recognition, image enhancement, object detection, eye tracking, and pose estimation.

Human Activity Recognition. DL-based models used in vision sensing for HAR can be computationally demanding, posing a significant challenge when it comes to execution on mobile and IoT devices. Moreover, vision systems that rely on RGB cameras are intrinsically susceptible to privacy leakage by hacking. To tackle this problem, Shim et al. [234] choose to use a Near-Infrared (NIR) image sensor to monitor human activities that inherently does not contain enough data to reveal personal identity. Although the NIR sensor loses a lot of

spatial information, the authors have demonstrated that the temporal information and pixel-wise computation over DNN are enough to recognize the performed activities.

Image Enhancement. Image enhancement involves manipulating the image itself to improve its quality. A key technique within this area is super resolution, which aims to increase the resolution of the image. However, executing this task on-device poses significant challenges due to the immense computational complexity and substantial storage requirements. To mitigate these issues, Lee et al. [134] employ two distinct compressed DNNs and schedule their operations across CPU, GPU, and DSP. Captured images by vision sensors are often transmitted over low-power, unreliable IoT networks. However, traditional methods such as JPEG, designed for use on reliable networks, are still commonly employed for image transmission. To efficiently transmit and receive high-quality image data over this unstable network, Hu et al. [96] find the optimal encoder and decoder pair of DNN by employing neural architecture search methods. Motion blurs on IoT devices are a severe problem while capturing the image. Existing solutions to this problem often necessitate additional hardware or have high computational demands that are ill-suited to microcontrollers. To solve this problem, Lee [135] adopt depth-independent convolution operations on DNN to estimate the blur kernel. This predicted kernel is then applied to the blurred image to recover the original, clear image. Additionally, the algorithm employs a matrix transformation, converting it to a Toeplitz Matrix. This transformation yields computational advantages, making it particularly efficient for deployment in extremely resource-constrained microcontroller environments.

Object Detection. Object detection is one of the most fundamental and important tasks in vision sensing. Recognizing faces in crowded environments is a critical challenge, particularly in applications like finding missing children. Existing DNN methods suffer from the low-resolution problem of the detected face. To solve the problem, Yi et al. [309] design a three-step mult-DNN pipeline consisting of detection, clarification, and recognition. During the clarification phase, the system recovers missing elements of the low-resolution image by fine-tuning it with the target's face. The research by Sami et al. [223] leverage a Time of Flight (ToF) sensor embedded in mobile phones to locate and identify concealed spy cameras. Conventional methods typically necessitate manual interpretation to discern these hidden devices. However, the incorporation of a ToF sensor enables the system to detect distinctive reflections emitted by spy cameras. Following this, deep learning techniques are deployed to filter out false positives from the detected images and effectively pinpoint the hidden cameras in an automated manner. Sun et al. [241] have shown the use of a smartphone camera to detect counterfeit liquid products, eliminating the need for additional hardware. The method tracks the movement of bubbles in the liquid using Faster-RCNN and U-Net and verifies the product's authenticity using the AdaBoost algorithm. Depth-contained images acquired from depth sensors can be employed in the detection and classification tasks of DNN [193, 280, 281]. These methods are effective compared to RGB cameras in low-light environments. Both Xie et al. [280, 281] employ the indirect Time-of-Flight (iToF) depth camera to capture the high-resolution texture depth map while Mithun et al. [193] use Kinect for XBOX One to achieve it. In particular, in the construction of the depth map, Xie et al. [281] employ an autoencoder that exploits the phase components of the iToF camera. On the other hand, Xie et al. [280] employs an additional distorting IR source and uses the energy difference of the signal depending on the texture.

Eye Tracking. Exploiting eyewear devices for eye tracking presents unique challenges due to their limited computational resources and the variability in eye characteristics across different individuals. To address this issue, Wu et al. [269] design EMO, a personalized DNN classifier that classifies emotions using images captured from a single eye by the eyewear. Likewise, Lan et al. [129] employ eyewear devices for extracting gaze data and aims to use it for cognitive context sensing. However, this approach also suffers from the diversity of people. To address this issue, this research adopts the few-shot learning method. These allow for rapid adjustment to new environments when operating a spatial-temporal graph-based DNN system, which is used to classify activities from gaze information. Tracking the gaze from the eye is highly demanding because of the small size of the

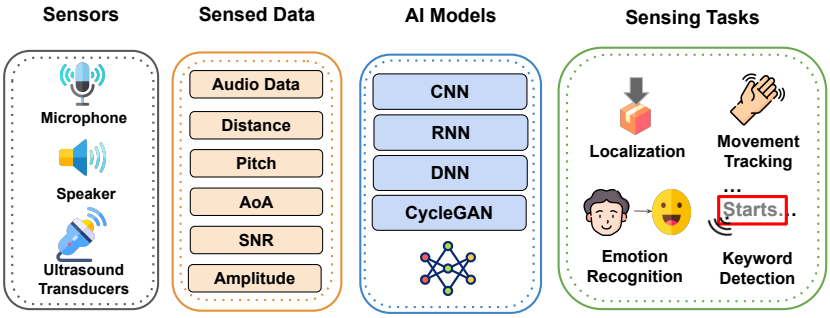

Fig. 8. Illustration of AI-empowered acoustic sensing pipeline.

iris and subtle hints concerning the directions. Existing commercial systems are expansive, while low-cost RGB camera approaches suffer from the insufficiency of datasets. To effectively track eye gaze, Cao et al. [25] have developed a geographical gaze model that maps the relationship between the smartphone screen and the iris boundary, which contains the gaze directions. To accurately extract the iris boundary over the eye, the authors employ U-Net and further refine the resulting pixels to enhance the accuracy of eye tracking.

Pose Estimation. Pose estimation is the process of determining the position and orientation of the human body, in a 3D space using visual inputs. Zhang et al. [330] introduce MobiPose, a system designed to achieve efficient and accurate real-time multi-person pose estimation on mobile devices. MobiPose introduces a motion-vector-based approach that tracks human proposals across consecutive frames to eliminate the need for repeated human detection. It also introduces a mobile-friendly model employing lightweight, multi-stage feature extractions utilizing heterogeneous computing resources (CPU and GPU) to perform pose estimation in parallel, thereby minimizing latency. Traditional 60Hz cameras have limited capabilities when it comes to tracking delicate finger movements due to their low sampling rate. Consequently, the performance of 3D hand pose reconstruction displays restricted accuracy. To address this issue, Zhang et al. [338] have developed a 3D hand pose reconstruction method that utilizes the camera and wearable gloves embedded with LEDs on the fingertips and wrist. The camera captures the strip effect of the rolling shutter from the LEDs on the gloves, and a CNN identifies the location and bounding box of these strips. This information is then used to construct a 3D representation of the hand posture.

2.4 Acoustic Sensing

Acoustic sensing involves utilizing acoustic sensors to capture, measure, and analyze acoustic signals for sensing purposes. As summarized in Figure 3, depending on the sensing tasks, existing works on AI-empowered acoustic sensing can be grouped into four categories: localization, movement tracking, emotion recognition, and keyword and event detection.

Localization. Localization using acoustic sensing refers to the process of determining the position or location of objects or sources of sound using sound waves. Mao et al. [185] introduce DeepRange, which investigates the limitations of traditional signal processing methods in localization tasks utilizing aquatic signals, particularly in scenarios with a low SNR environment. They pose the question of whether DNNs can automatically learn features from received acoustic signals to estimate distance, potentially surpassing the performance of conventional signal processing algorithms devised by domain experts. The study introduces a DNN-based ranging system, which directly employs raw acoustic signals without feature extraction and indicates superior performance compared to established signal processing approaches. Conventional methodologies for sound source localization require multiple microphone arrays, which is impractical for tiny devices. Addressing this, Owlet [67] place

a microphone inside the stencil with sound holes. The incoming sound through these apertures indicates the direction based on hole patterns. Nevertheless, the approach remains susceptible to environmental factors, such as reflective wall signals. To mitigate this, the authors employed a CNN to estimate the Direction of Arrival (DoA), trained on a synthetic dataset representative of various environments. Yang and Zheng [301] introduce DeepEar, a DL-based framework to improve sound localization using only two microphones, particularly in scenarios with multiple sound sources. Drawing inspiration from the biological function of human ears, which shape sound waves to provide more spatial information, the authors design a neural network architecture that simulates human auditory processing. This includes a gammatone filterbank mimicking the cochlea's role by transforming audio into the time-frequency domain, followed by an autoencoder that extracts high-level sound representations. These features are then utilized by a deep neural network to pinpoint sound locations accurately.

Movement Tracking. Movement tracking using acoustic sensing involves detecting and monitoring the movement of objects or individuals through the analysis of sound waves. Acoustic signals, as they propagate through the human body, undergo a range of transformations. By performing a comprehensive analysis of these signal changes, we can effectively track the movements. Sun et al. [244] introduce VSkin, a system that can detect finger movement on mobile devices using acoustic signals. VSkin utilizes both structure-borne and air-borne sounds to detect touch and measure finger movements on all surfaces of a device, not only limited to the touchscreen. The existing method of movement tracking frequently encounters challenges such as low SNR, interference, and mobility, which may affect the accuracy of the tracking. To overcome these issues, Mao et al. [186] employ the 2D MUSIC algorithm [266] to produce joint of distance and Angle of Arrival (AoA) profiles derived from hand motion. Leveraging this profile, an RNN is utilized to precisely track both the distance and AoA of hand movements on a room-scale. Chen et al. [34] present Ipanel, a system that uses acoustic signals created by finger movements on a hardwood tabletop to extend mobile device interactions beyond the small screen and onto surrounding surfaces. Unlike traditional finger tracking systems that use a fixed frequency acoustic signal, Ipanel tracks the dynamically changing frequencies of acoustic signals produced when fingers slide on a surface. IPanel extracts distinctive features from both the spatio-temporal and frequency domain characteristics of the acoustic signals, converting them into images, which are then processed by a CNN for finger movement recognition. The system supports recognition of common gestures like clicks, flips, scrolls, and zooms, as well as handwriting recognition of numbers and alphabets with high accuracy. Acoustic signals can capture human breathing patterns, with a key advantage being the elimination of specialized wearable sensors. Leveraging this, Xu et al. [296] designs a model to monitor the drivers through accurate breathing pattern extraction. The initial stage involves the isolation of environmental driving noise, which is then followed by the reconstruction of detailed breathing waveforms via the application of GAN. Li et al. [145] present FM-Track, a system for tracking multiple moving targets using acoustic signals without physical contact. The authors propose a chirp-based signal model that integrates range, velocity, and angle information from the reflected signals to accurately determine the position and movement of each target. FM-Track can track up to four targets simultaneously within a 3-meter range, demonstrating its efficacy through experiments on both smartphones and smart speakers. Fu et al. [65] employ ultrasound signals emitted from a smartphone to detect the articulatory movements of the mouth. Using the reflected signals from these movements, the study successfully reconstructs audible speech with a DNN named SiVoNet by training the network supervised way using paired audible speech. They implemented a prototype for a comprehensive evaluation, using a Samsung Galaxy S8 to validate performance on a commercial smartphone platform. The evaluation results show that SiVoNet can reconstruct speech with a Character Error Rate (CER) as low as 7.62%, outperforming state-of-the-art acoustic-based approaches. Experience [144] investigate the challenges and solutions related to the deployment of acoustic sensing system-based movement tracking in real-world scenarios. The authors identify several critical issues, such as audible sound leakage, high power consumption, and performance

degradation due to device mobility. Li et al. [144] propose a power control mechanism by dynamically adjusting the transmission power and switching between idle and active states based on detected activity to reduce power consumption. They built a prototype of their proposed power control schemes for hand tracking on a Samsung S9+ smartphone, reducing average power consumption from 22% to 10% over two hours.

Emotion Recognition. Emotion recognition through acoustic sensing involves analyzing voice and sound patterns to determine the emotional state of a speaker. Lane et al. [131] present DeepEar, a mobile audio sensing framework to perform audio inference tasks such as ambient scene analysis, emotion recognition, and stress detection. DeepEar is designed to address the challenge of diverse and noisy acoustic environments that mobile users encounter. The framework consists of multiple DNNs, each specialized in a specific audio sensing task, and employs advanced DL techniques for pre-training and fine-tuning. Georgiev et al. [69] address the challenge of performing multiple audio analysis tasks, including emotion recognition, on resource-constrained mobile and embedded devices. Existing solutions for audio sensing focus exclusively on the operation of a single DNN. However, Georgiev et al. [69] have shown that by sharing layers among different audio task DNN models, it can reduce its computation cost while achieving comparable accuracy. Microphone variability, which refers to differences in audio data quality and characteristics recorded by different microphones, can significantly impact the robustness and accuracy of audio-sensing tasks. To address this challenge, Mathur et al. [187] design Mic2Mic, which leverages Cycle-Consistent Generative Adversarial Networks (CycleGANs) to ensure that emotion recognition and other audio sensing tasks can be performed accurately across different devices. Mic2Mic learns a translation function between audio data recorded from different microphones, effectively reducing the domain shift caused by microphone variability.

Keyword and Event Detection. Keyword detection in acoustic sensing involves the identification and recognition of specific words or phrases from audio signals. Selecting the device with the best audio quality leads to clearer and more distinguishable audio features, which are critical for accurate keyword recognition. Min et al. [192] introduce a real-time assessment framework to determine the optimal audio input from various devices. This model routinely evaluates potential devices and selects the most suitable one for operation within the execution duty cycle. They introduced two models for this assessment: probability-based and data-driven DNN models. It demonstrates that it achieves higher accuracy while consuming less energy than its baseline counterparts in keyword detection tasks. Event detection refers to the process of identifying and recognizing specific events or activities based on sound signals captured by acoustic sensors. Traditional systems often miss parts of longer-duration events due to intermittent power, resulting in incomplete audio data. To mitigate these challenges, SoundSieve [198] employ a regression neural network to predict the importance of upcoming audio segments and captures only the most relevant segments of an audio clip. With this predictive capability, the device can decide whether to enter a sleep cycle or remain awake to capture the signal.

2.5 Multi-Modal Sensing

Multi-modal sensing involves the use of more than one sensing modality where the key advantage is its ability to combine distinct information provided by each of the included sensing modalities. At the same time, determining which sensing modalities to include, and how to combine them effectively, are highly dependent on the specific application. As summarized in Figure 3, depending on the sensing tasks, existing works on AI-empowered multi-modal sensing can be grouped into five categories: human activity recognition, human and object identification, tracking, localization, and speech enhancement.

Human Activity Recognition. Human activity recognition (HAR) using multi-modal sensing integrates data from different sensory modalities to detect and identify human activities. Traditional ML strategies typically

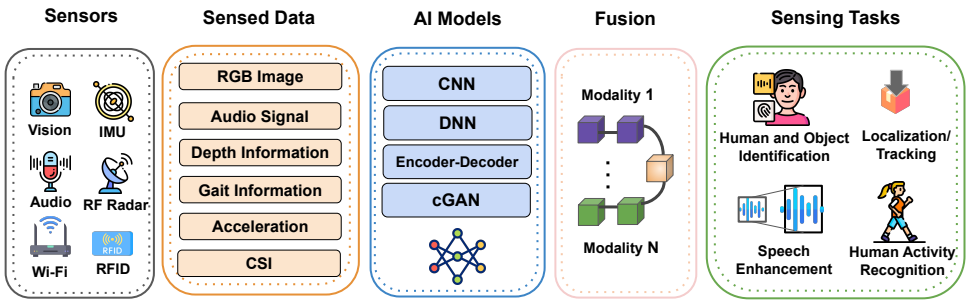

Fig. 9. Illustration of AI-empowered multi-modal sensing pipeline.

employ one of two methods for sensor fusion: feature concatenation and Ensemble classifiers. Feature concatenation merges modalities but neglects inter-sensor correlation. Ensemble classifiers, on the other hand, uses separate classifiers but compromise intra-sensor correlation by fusing outputs later. Radu et al. [215] propose Modality-Specific Architecture that can learn both inter and intra-sensor correlation for the task of HAR. The network comprises multiple distinct branches, each dedicated to a specific modality. The outputs from these branches are then combined using fully connected layers. The task of HAR requires high accuracy with minimal inference latency. In multi-modal environments where sensors transmit data to a computing device, network fluctuations can cause asynchronous arrival of modalities. Straightforward approaches, such as waiting for delayed modalities or ignoring them, compromise both latency and accuracy. To address this challenge, Li et al. [149] introduce speculative inference Instead of waiting for delayed sensor data, it imputes the missing values and utilizes this generated data for subsequent inferences. If the accuracy falls below acceptable levels, the system executes a rollback of its results and re-initiates the inference process. Leite and Xiao [137] propose reducing the number of sensors used to lower computational demands, although this can potentially degrade accuracy. To mitigate this, the authors introduce a pipeline that prioritizes sensors based on their impact on accuracy. During the model training phase, sensors that have minimal or negative effects on accuracy are excluded. This approach significantly reduces memory usage and inference time while maintaining high accuracy in HAR. In HAR, when a model trained in one domain is deployed in another domain, a degradation in performance occurs due to differences between the two domains. In multi-modal environments, these challenges are amplified due to the presence of additional variable factors. To tackle this issue, Hu et al. [97] propose VMA, which transfers the DNN from one domain to another in the presence of multiple domains and modalities. The key idea is that changing one factor would have higher accuracy than changing multiple factors. Thus, VMA identifies pairs of domains wherein only one factor differs between them. Leveraging these pairs, it finds a path to effectively transition from one domain to the desired target domain by sequentially modifying only one factor at a time. Ouyang et al. [204] propose Cosmo, a two-stage fusion learning system for enhancing HAR using multimodal data when labeled data are limited. In the initial stage, Cosmo leverages unlabeled data to discern consistent information, which denotes shared information that is uniformly present across different modalities. During the subsequent stage, Cosmo focuses on capturing complementary information, identifying the distinct and unique characteristics inherent to each modality, and leveraging the labeled data. As such, Cosmo achieve 26.73% accuracy compared to the supervised fusion learning baseline. Lastly, Zhang et al. [344] introduce CMA, a method for HAR by associating wearable IMU sensors with structural vibration signals. In CMA, all data is initially aligned and segmented according to the timestamp. Then, each data segment detects the activity in their data using a threshold. Lastly, the system utilizes a Temporal Convolutional Network to determine if the data segment sourced from distinct modalities points to an identical activity and individual.

Human and Object Identification. Human and object identification involves the ability to detect, recognize and categorize of individuals, objects or both. One effective method for identifying individuals is analyzing their gait, as it constitutes a unique characteristic for each person. While some studies utilize camera-based techniques for this purpose, they often struggle in low-light conditions and require the subject to be within the camera's field of view. Also, RF signals offer advantages like penetration through obstacles and not being affected by lighting conditions, but their accuracy may decrease when there is a significant difference between the training and testing environments. To tackle this problem, Korany et al. [125] introduce XModal-ID, a gait-based identifying system using Wi-Fi signal and video footage. It determines whether a person within a Wi-Fi area is the same as the individual captured in the video footage. From the video, it creates the 3D mesh of a human and simulates how Wi-Fi signal would be after the signal is reflected from a 3D mesh human. This Wi-Fi signal implicitly contains gait information since it is reflected by the human body joint while moving. Thus, by comparing the simulated Wi-Fi data with the real-world Wi-Fi signal captured in the Wi-Fi area, it can identify whether the two sets of data correspond to the same individual or not. Liu et al. [170] present an innovative system called RF-Camera, which combines RFID and computer vision techniques to recognize human interactions with physical objects in environments involving multiple subjects and objects. To achieve this goal, RF-Camera uses the Kinect DK system which is equipped with an RGB camera and depth camera to detect the human and its relevant hand trajectory. At the same time, an RFID system is used to identify the items and track their movement. Current vision-based methods for video scene analysis excel at recognizing and identifying objects and people, i.e., extrinsic details. Nevertheless, they cannot be used when it comes to capturing intrinsic details, such as discerning the state of a washing machine. To bridge this gap, Capricorn [265] integrates both RF and vision sensors, aiming to understand a scene's external and internal details comprehensively. Specifically, the camera provides data about types of objects and their respective bounding boxes. Concurrently, UWB radar detects object vibrations, leveraging this data to infer the internal states of these objects. In [161], the authors propose Vi-Fi, which utilizes an RGB-D camera and smartphones to associate multiple individuals with their respective smartphone identifiers. It accomplishes this by capturing bounding box information and depth data from the RGB-D camera, as well as IMU sensor data and Wi-Fi Fine Timing Measurements (FTM) from smartphones. Subsequently, this diverse data is fed into LSTM models, and the output features are combined to generate an association score between the smartphones and their bounding boxes. Vi-Fi achieves an association accuracy of 81% in real-time and 91% in offline processing, demonstrating its effectiveness in identifying humans and objects in complex environments. When it comes to object identification, capturing the material and shape information is of vital importance. mmWave signals can obtain rich information from the reflecting surfaces thanks to its broadband signals. However, the reflections from stationary objects contain less information than vibrating objects. To exploit its capability to the fullest, RFVibe [228] fuses mmWave signals with acoustic signals for contactless material and object identification. Particularly, it plays an audio sound towards the object to generate micro-vibrations in the object and shines a millimeter wave radar signal on the object at the same time. By analyzing the physical properties of the reflected wireless signal, these micro-vibrations can be captured. RFVibe extracts several features, including frequency features, power features, and damping features. RFvibe adopts a CNN-based neural network to enable accurate identification of these features under different setups and locations. The neural network consists of three feature heads that transform features from different sources into a common latent space and a classification head that takes in the intermediate feature maps and outputs the probability distribution of possible classes.

Tracking. Tracking humans or objects has been explored using various modalities. One method is utilizing the mmWave radar because it offers spatial information and the ability to construct data points in space. However, this sensor struggles in scenarios involving rapid movements. To address this limitation, Lu et al. [177] introduce milliEgo, a robust egomotion estimation system that combines the capabilities of the IMU sensor and mmWave

radar. To integrate the information from these two sensors, the authors proposed a two-stage intra- and inter-sensor cross-self attention mechanism, which interchangeably learns how to compensate for one another sensors during each step. Consequently, this approach outperforms its counterparts, which solely rely on the IMU sensor, combining RGB with IMU and integrating depth information with IMU. Another combination of mmWave radar and IMU sensor has also been explored for tracking interpersonal distances by ImmTrack [44]. Since it requires tracking multiple individuals, IMU data from multiple individuals' smartphones and the corresponding mmWave data are generated. To associate them, cosine similarity metrics are employed. Once associated, the IMU data, initially in its local coordinate system, is translated to the mmWave's global coordinate system, making it suitable for monitoring interpersonal distances.

Localization. Multi-modal sensing can also enhance the performance of localization. For example, Boroushaki et al. [19] introduce RFusion, a multi-modal localization system that utilizes both RF and vision sensing modalities. When estimating a location using a single RF antenna, there's a broad potential location area. Introducing an RGB-D camera can narrow down this area by leveraging depth information. Nevertheless, even with this refinement, multiple candidate locations remain, necessitating measurements from various positions. By optimizing this measurement trajectory through reinforcement learning, RFusion achieves centimeter-level accuracy, improving travel distance efficiency by twice as much compared to its baseline. As another example, ELF-SLAM [180] propose to combine both motion sensing and acoustic sensing for localization. IMU sensor inherently is susceptible to noise and biases that can accumulate over time. The authors propose to leverage the acoustic information emitted and captured by smartphones. As this acoustic data is reflected by surfaces, the captured echoes carry distinct spatial information based on their location. This enables precise indoor location alignment by compensating the inaccurate misaligned parts of the IMU sensor with spatial information in the acoustic data.

Speech Enhancement. Speech enhancement refers to the process of improving the quality and intelligibility of speech signals, typically in the presence of noise or other degrading factors. Traditional research relying solely on audio data often requires multiple microphone arrays and is significantly influenced by the environment in which the data is captured. While multi-modal solutions exist that combine camera-captured lip movements with audio, their accuracy degrades in low-light conditions. Consequently, Sun and Zhang [243] introduce UltraSE, a system that combines ultrasound signals with audible sound to enhance the user's speech. The user holds the phone near the mouth, and it emits ultrasound. Since this signal is reflected by the lip, it contains the articulation gestures that do not contain the noise of the audible sound. By fusing this noise-free ultrasound data with audio data in a cGan-based DNN, it produces de-noised audio output. However, this method has a limitation in terms of short working distances. Also, it has to hold the phone to capture the data. To address this issue, Liu et al. [169] design Wavoice, which aims to remove noise from the audio signal using mmWave that can operate long distances. They discovered a strong correlation between mmWave and audio signals, as both carry information about vocal fold vibrations, making them suitable for fusion. By integrating these two, the audio data offsets the motion interference inherent in mmWave signals, while the mmWave data counteracts the noise limitations of the audio signal. As a result, Wavoice surpasses its audio-only speech recognition baseline more than 20 times. Although Wavoice successfully separates the clear speech, it has the constraint that it needs the mmWave radar device. To address this issue, by narrowing the focus to scenarios involving head-mounted wearables like wireless earbuds or VR/AR headsets, VibVoice [88] leverage the IMU sensors that most of these devices are equipped with. IMU accelerometer attached to the head is capable of detecting vibrations generated by the speaker's voice via bone conduction through the skull, devoid of any external environmental noises. To integrate the IMU and audio modalities, they employ encoder-decoder architectures. The encoder extracts essential features from each modality and merges them while the decoder subsequently reconstructs human speech.

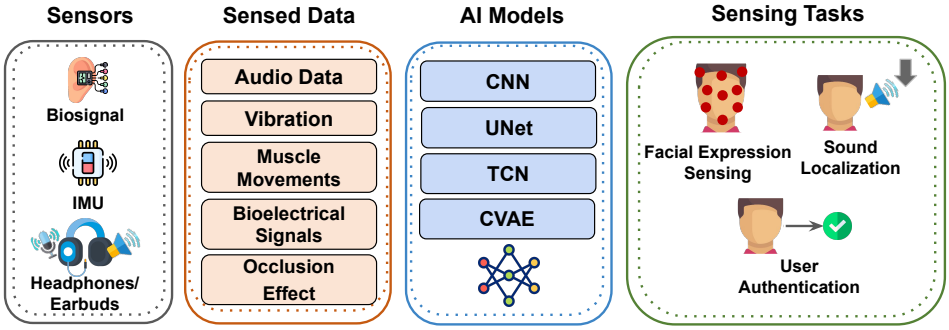

Fig. 10. Illustration of AI-empowered Earable sensing pipeline.

2.6 Earable Sensing

Earables are wearable devices attached to ears in the form of headphones or wireless earbuds. As summarized in Figure 3, depending on the sensing tasks, existing works on AI-empowered earable sensing can be grouped into three categories: facial expression sensing, user authentication, and sound localization.

Facial Expression Sensing. Conventional methods for capturing facial expressions are primarily counted on video cameras. However, video cameras are limited in low-light environments and pose substantial risks of privacy infringement. In contrast, earables avoid such limitations and have demonstrated significant promise for a variety of facial expression sensing tasks. For example, Wu et al. [271] propose BioFace-3D, which leverages EMG and EOG signals captured by earables to detect the facial muscle activities, track 2D landmarks, and perform continuous 3D facial reconstruction using a CNN. As another example, Song et al. [239] propose FaceListener, which uses the commodity headphone to recognize a user's facial expressions. FaceListener emits ultrasound signals to detect face movements and uses this information to create a facial landmark model and recognize facial expressions based on an LSTM model.

User Authentication. Earables have also been utilized to identify unique individual characteristics, such as a person's gait for the purpose of user authentication. For gait-based user authentication, traditional methods often require special equipment, which is cost-prohibitive and limited in range. In contrast, Ferlini et al. [64] propose EarGate, which employs an in-ear microphone to capture bone-conducted sounds induced by walking to detect the user's gait for user identification. Furthermore, they demonstrate that classification performance can be notably improved through transfer learning. Liu et al. [162] introduce MandiPass, a biometric-based authentication system that utilizes intracorporal biometric called MandiblePrint, derived from the vibrations of human mandibles. It uses an Inertial Measurement Unit (IMU) embedded in an earphone to capture the MandiblePrint when a user voices some specific sound. This sound generates vibrations in the throat that propagate through the mandible to the ear, where they are sensed by the IMU. MandiPass validates the feasibility of MandiblePrint through theoretical modeling and experimental vibration propagation, demonstrating its potential as a user authentication method.

Sound Localization. Sound localization in earable sensing refers to the ability of ear-worn devices to determine the direction of incoming sound sources. It is essential for enhancing spatial awareness and improving user experience in hearing aids, augmented reality, and personal assistants. Chatterjee et al. [31] emphasize the importance of sound localization in enhancing user experience, particularly in distinguishing between the target speaker and background noise. The authors use binaural wireless earbuds and dual-channel neural networks to separate the target voice from the noises. These networks consist of a time domain network called CB-Conv-TasNet and a frequency-based network called CB-UNet to exploit both spatial and acoustic information. As a result, it achieves

a better scale-invariant signal-to-distortion ratio (SI-SDR) than AirPods Pro, which is based on beamforming. Another critical task in sound localization is individualizing the Head-Related Transfer Function (HRTF). This individualization typically demands extensive and cost-intensive measurements in an anechoic chamber. To address this issue, Zandi et al. [315] introduce a simplified approach for conducting these measurements and propose to use a conditional variational autoencoder to achieve HRTF individualization. Lastly, Yang and Zheng [301] introduce DeepEar to address the issue of sound localization with two microphones. Unlike traditional methods that rely on extensive microphone arrays, DeepEar employs binaural microphones, which are more compact and thus more suitable for integration into devices like hearing aids. DeepEar leverages a multisector-based neural network that divides space into sectors for detecting multiple sound sources simultaneously.

2.7 Generative AI for Sensing

Advancements in Generative AI have provided AIoT with opportunities to leverage state-of-the-art generative models such as Large Language Models (LLMs) to perceive, interpret, and present IoT sensor data in ways that are not attainable before [263]. Generative AI can correlate sensor readings with relevant contextual information, such as historical data, environmental conditions, and operational status so as to provide deeper insights into the sensor data and make decisions; it can improve user experiences by allowing non-technical users to interact with sensor systems and perform data querying using natural language; it can also help translate raw sensor data into human-understandable reports and summaries, making it easier for users to understand key information contained inside sensor data.

Some efforts have been made to leverage such unique capabilities of Generative AI for sensing. For example, Ouyang and Srivastava [205] propose LLMSense, a prompting framework for LLMs to make sense of raw sensor data and low-level perception results. This framework can be implemented in an edge-cloud system, with small LLMs running on edge devices to summarize sensor data and high-level reasoning performed on the cloud to ensure data privacy. Two approaches are proposed to improve the performance of LLMSense: summarizing sensor data before reasoning and selectively including historical sensor data. Results show that LLMSense achieves high accuracy in tasks such as dementia diagnosis using behavior data and occupancy tracking with environmental sensor data. In [287], the authors propose Penetrative AI to explore how LLMs can be extended to interact with the physical world using IoT sensors and actuators. As a prompting framework, Penetrative AI shows how carefully constructed prompts can harness LLMs' embedded world knowledge for tasks such as user activity sensing and heartbeat detection. Specifically, Penetrative AI operates on two levels: textualized signal processing, where sensor data are converted into text for LLM analysis; and digitized signal processing, where LLMs directly interpret sensor data. Using heartbeat detection as an example, Penetrative AI demonstrates that LLMs can effectively analyze real-world sensor data with proper guidance, illustrating the potential of integrating LLMs into cyber-physical systems to enhance their intelligence and functionality. Lastly, Wan et al. [254] go one step further beyond prompting and propose a multimodal LLM named MEIT that translates raw ECG sensor data into human-understandable reports. For cardiologists, the task of interpreting ECG data and writing reports can be both intricate and time-consuming. MEIT aims to fill this gap by automating the ECG report generation task. Specifically, MEIT involves instruction tuning a multimodal LLM to integrate raw ECG data with corresponding textual instructions, ensuring that the generated reports are clinically relevant and accurate. Experimental results demonstrate the superior performance of MEIT in generating accurate and professional ECG reports, underscoring its potential for real-world clinical applications.

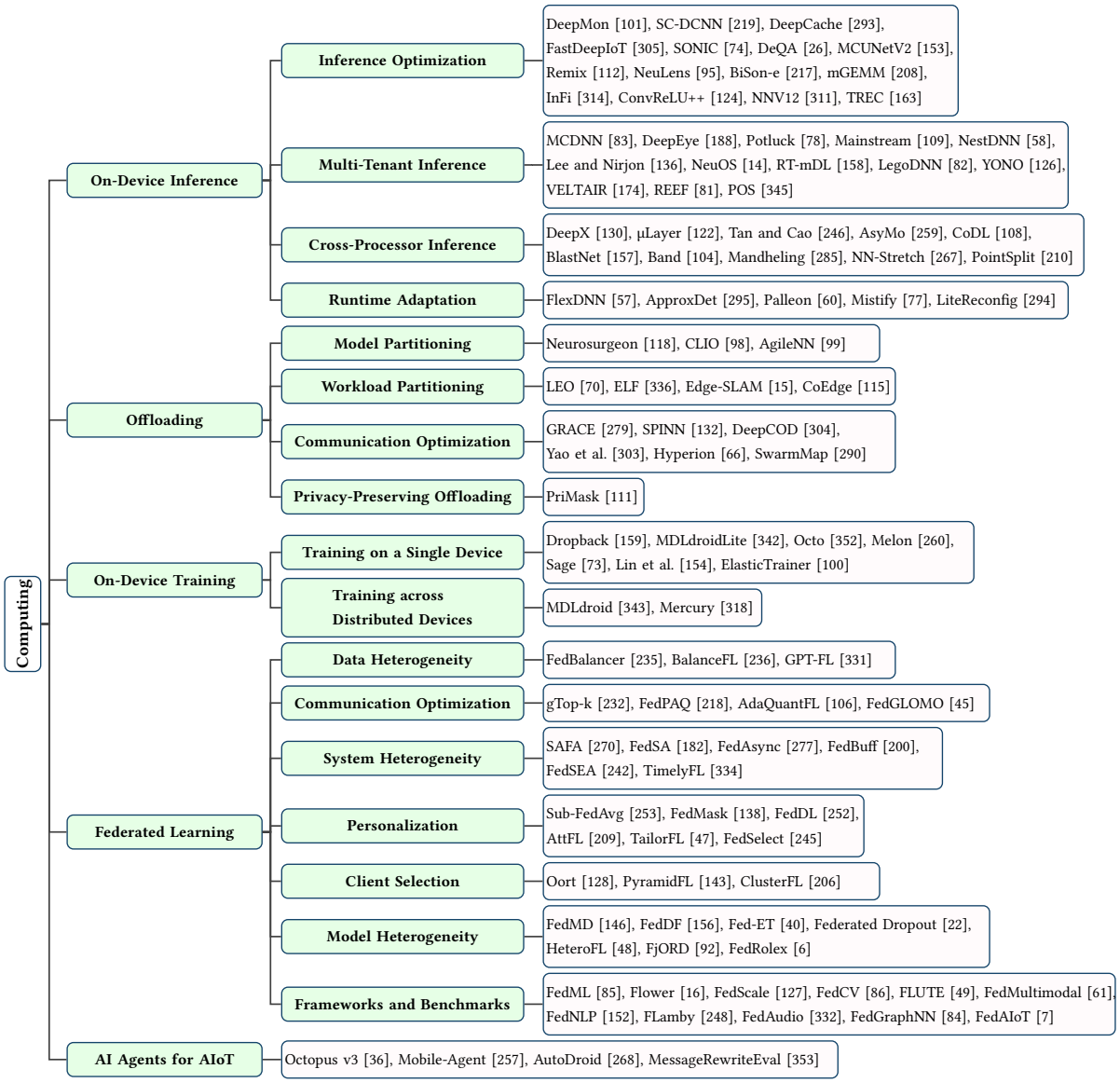

Fig. 11. Summary of topics related to computing.

3 COMPUTING

3.1 On-Device Inference

One of the most fundamental and essential compute tasks of AIoT is to perform inferences on the device. On-device inference is particularly critical for latency-sensitive applications or scenarios where cloud connectivity is not available. As summarized in Figure 11, existing works on on-device inference can be grouped into four categories: inference optimization, multi-tenant inference, cross-processor inference, and runtime adaptation.

3.1.1 Inference Optimization. IoT devices are constrained in their onboard computing power, memory resources, and battery life. The objective of inference optimization is to enhance the computational and energy efficiency as well as to reduce memory demands and efficiently utilize memory resources during the inference process. For example, Huynh et al. [101] propose DeepMon, an on-device inference framework that allows large DNNs to run on mobile devices at low latency for continuous vision applications. They propose a caching mechanism that exploits the similarities between consecutive images to cache intermediate processed data within CNN, which allows DeepMon to execute very deep models such as VGG-16 in near real-time. Ren et al. [219] propose SC-DCNN, an optimization framework of stochastic computing (SC) for CNNs. They propose to apply SC to CNNs by designing function blocks and implementing hardware-oriented max-pooling in the SC domain. In addition, they propose to perform holistic optimizations for feature extraction blocks and weight storage schemes. By calculating multiplications and additions with AND gates and multiplexers in SC, SC-DCNN achieves a significant reduction in energy consumption. Xu et al. [293] propose DeepCache, which adopts proven video compression techniques to systematically search for neighboring image blocks with similarities, rather than restricting matching solely to blocks in the same positions. They propose dividing video frames into regions, searching for similar regions in cached frames using a specialized matcher, and dynamically merging adjacent regions to maintain cache effectiveness. In [305], the authors propose FastDeepIoT, which incorporates a profiling module and a compression steering module to optimize execution time and reduce energy consumption. The profiling module generates diverse training structures and builds an interpretable model for predicting the execution time, while the compression steering module enables existing DL compression algorithms to collaboratively minimize both execution time and energy consumption. In SONIC [74], the authors explore the opportunity of DNN inference intermittently on energy-harvesting systems. They propose loop continuation that significantly reduces the cost of ensuring accurate intermittent execution for DNN inference by modifying loop control variables within a loop nest, as opposed to dividing an extended loop into multiple tasks. Cao et al. [26] propose DeQA, a set of optimization techniques designed to enable Question Answering (QA) systems to run on mobile devices. DeQA reduces memory demands by loading partial indexes, dividing data into smaller units, and replacing in-memory lookups with a key-value database, altogether reducing the memory requirements of QA systems to just a few hundred megabytes. Lin et al. [153] propose MCUNetV2, a scheduling technique in a patch-based manner to minimize memory usage for tiny DL. They propose initially executing the model on a limited spatial region, followed by the remainder of the network operating with a smaller peak memory consumption in the usual manner. Additionally, they propose to redistribute the receptive field to reduce the computation overhead caused by the patch-based initial stage. Jiang et al. [112] propose Remix, an adaptive image partitioning and selective execution strategy that involves the execution of existing DNNs on non-uniformly partitioned image blocks. They propose to leverage historical frames to learn the distribution of target objects and achieve higher detection accuracy with a given latency budget or higher inference speedup without accuracy deduction. Hou et al. [95] propose a dynamic inference mechanism known as the Assemble Region-Aware Convolution (ARAC) supernet, which removes redundant operations within CNN models by leveraging spatial redundancy and channel slicing. They propose to split the CNN inference flow into multiple micro-flows and load them into GPU as single models. In this way, NeuLens outperforms baseline methods in terms of latency reduction (up to 58%) while achieving accuracy improvement (up to 67.9%) within the same latency and memory constraints. Reggiani et al. [217] propose BiSon-e, a RISC-V-based architecture that features a binary segmentation to enhance the CPU pipeline. They propose to perform Single Instruction Multiple Data (SIMD) operations on existing scalar Functional Units (FUs) to increase the performance of narrow integer applications on resource-constrained edge devices. In this way, BiSon-e achieves significant energy efficiency and execution time deduction. To address the overload caused by the convolution layer, Park et al. [208] propose mGEMM, which expands the structure of the GEMM and eliminates the problems of memory overhead and low data reuse rate of the GEMM. They propose a reusable block of highly optimized computation on the inner computation kernel and partitioned the

computation for the loops outside of the inner kernel. In [314], the authors propose a learnable input filtering framework named InFi that unifies both approaches. They propose treating skip as a special case of reuse and designing a filter that supports both skip and reuse functions, requiring only maintaining an additional key-value table for reuse in the inference phase. In this way, InFi achieves lower energy consumption and latency. Kong et al. [124] propose a lossless acceleration method ConvReLU++, which achieves early negative result detection by employing reference-based upper-bound calculations. This approach guarantees that once intermediate results turn negative, the final results will be negative. When negative results are detected, the remaining computations can be skipped, leading to a significant latency reduction in ConvReLU++. Yi et al. [311] propose NNV12, an on-device framework that optimizes cold inference. They propose three optimization techniques encompassing kernel selection, weight transformation caching, and pipelined inference, to effectively reduce the latency of cold inference. In addition, they propose a heuristic-based kernel scheduling scheme, which fully harnessed three optimization techniques and led to substantial enhancements in the latency of cold inference. Lastly, Liu et al. [163] propose a set of optimization techniques for the Transient Redundancy Elimination-based Convolution (TREC), which recognizes and prevents redundant computations present in the form of identical tiles within input data or activation maps. They propose to repurpose parts of a matrix used in DNN computations as hashing vectors and embed a two-step stack for storing clustering IDs in TREC, aided by a reversed index for efficient entry location, which collaboratively eliminates significant memory overhead.

3.1.2 Multi-Tenant Inference. Multi-tenant inference refers to the simultaneous execution of multiple distinct AI models, often originating from multiple concurrently running applications. The key to multi-tenant inference is to efficiently manage and process inference requests from multiple tenants with limited resources on the device. Han et al. [83] propose MCDNN, a framework for executing DNNs in video stream analytics using an approximation-based approach. They propose a heuristic scheduling algorithm designed to address approximate model scheduling, which allocates resources based on their usage frequency and utilizes a catalog to choose the most accurate model variant. Mathur et al. [188] propose DeepEye, a small wearable camera running multiple models locally, enabling near real-time image analysis. They propose an inference pipeline that increased processor utilization by scheduling the execution of computation-heavy layers and the loading of memory-heavy layers across multiple models. They also built prototype hardware powered by a quad-core Qualcomm Snapdragon 410 processor on a custom integrated carrier board to demonstrate the feasibility of their design. Guo and Hu [78] propose Potluck, which caches the previously computed results to provide cross-applications approximate deduplication. They propose a set of algorithms tuning the similarity threshold that regulates the degree to which various raw inputs are considered to be "the same", which makes Potluck decreases processing latency for vision workloads. Jiang et al. [109] propose Mainstream, a video processing system that addresses resource contention by sharing the same portion of DNN when inference is taken, which avoids redundant work. Additionally, they use an analytical model to estimate the effects of DNNs for an event and give the optimal model and sample rate option, resulting in significant overall event F1-score improvement. In [58], the authors propose NestDNN, a framework that enables resource-aware on-device DL in multi-tenant settings. The key idea of NestDNN is to transform a DNN model into a multi-capacity model, where sub-models with smaller capacity are nested inside sub-models with larger capacity through shared parameters. At runtime, NestDNN incorporates a resource-aware scheduler which selects the optimal sub-model for each DNN model and allocates it the optimal amount of runtime resources so as to jointly maximize the overall performance of all the concurrently running applications. Lee and Nirjon [136] propose a concept of neural weight virtualization. Having each block of memory represent a block of weights for one or more DNNs makes it possible for multiple DNNs to be put into the main memory which has a smaller capacity than the total size of the DNNs. In this way, weight virtualization achieves significant improvement in execution time and energy efficiency. Bateni and Liu [14] propose NeuOS, a latency-predictable framework for DNN-driven autonomous systems. They introduced the notion of a cohort, which represents a group of DNN

instances capable of communication via a shared channel. They also propose a technique to predict the best system-level power configuration for each DNN of the cohort to meet the deadline for processing. Ling et al. [158] propose RT-mDL, a framework that enables heterogeneous DL tasks to execute on edge devices by concurrently optimizing DNN model scaling and real-time scheduling. They propose a model scaling algorithm constrained by storage limitations that generates a range of model variants and overall optimizes the DL execution by identifying the optimal combination of task priorities and scaling levels of DL tasks. Han et al. [82] explore a block-level scaling of DNNs, which only extracts and re-training descendant blocks from a complete DNN. Additionally, they employ a runtime scalar to determine the most effective combination of blocks to maximize accuracy. In this way, LegoDNN offers a wider range of model sizes without increasing time cost, resulting in significant improvement in accuracy and energy consumption reduction. In [126], the authors propose YONO based on product quantization to compress heterogeneous models into two codebooks. Additionally, they enable in-memory model execution and support model switching for dissimilar multi-task learning on microcontrollers, achieving significant latency and energy consumption reduction. Liu et al. [174] introduce VELTAIR, a scheduling approach that adapts its granularity to efficiently minimize scheduling conflicts. Additionally, they propose an adaptive compilation strategy that enables dynamic selection of programs with appropriate exclusive and shared resource usage patterns, aimed at mitigating overall performance degradation caused by interference. In REEF [81], the authors explore preemptive scheduling support for inference tasks on GPU. They propose a reset-based preemption mechanism that initiates a real-time kernel on the GPU through proactive termination and subsequent restoration of best-effort kernels. Zhang et al. [345] propose POS, an operator-level scheduling framework combined with four operator-scheduling strategies. They propose abstracting the multi-model inference into a computation graph-based unified intermediate representation and finding optimal scheduling strategies for operators in the computation graph automatically with a learning-based operator-scheduling algorithm.

3.1.3 Cross-Processor Inference. Cross-Processor Inference refers to the ability of a model to perform inference across different types of processors (i.e., CPUs, GPUs, TPUs) within a device. Modern IoT devices are often equipped with multiple heterogeneous processors, each of which is optimized for certain computing tasks. This provides a great opportunity to leverage these heterogeneous processing units to collaboratively perform inference in a cross-processor manner. The realization of cross-processor inference involves a pivotal strategy: model partitioning. This technique capitalizes on the multiple processors to optimize inference tasks by partitioning the models and executing individual partitions on different processors. For example, Lane et al. [130] propose DeepX: a software accelerator for DL execution that allows any developer to use DL methods and automatically lowers resource usage. They propose a deep architecture decomposition algorithm that can decompose models into unit blocks for heterogeneous local device processors, maximizing resource utilization. In [122], the authors propose μLayer, a low latency on-device inference runtime that accelerates each layer by utilizing the onboard CPU and GPU simultaneously. They propose channel-wise workload distribution to distribute the output channels of an NN layer to both CPU and GPU to fully utilize the resources, achieving a significant reduction in latency. Tan and Cao [246] explore model partitioning between CPU and Neural Processing Units (NPUs). NPUs run DNN models faster but with less accuracy. Consequently, they propose heuristic-based algorithms and Machine Learning based Model Partition, which can explore a range of layer combinations to determine the part for CPU and NPU separately with optimal time-accuracy trade-off. Wang et al. [259] propose AsyMo, which focuses on partitioning the matrix multiplication blocks of DL models on asymmetric multiprocessors. They propose cost-model-directed block partitioning and asymmetry-aware scheduling to balance the tasks. Additionally, they propose to set the frequency by offline profiling energy curves, which achieve more energy efficiency than baselines. Jia et al. [108] propose CoDL, a concurrent DL inference framework that makes optimal use of diverse processors to expedite the execution at the operator level. They propose to use hybrid-dimensional partitioning and operator chaining to reduce sharing-related overhead, and an accurate, lightweight method to predict latency by considering

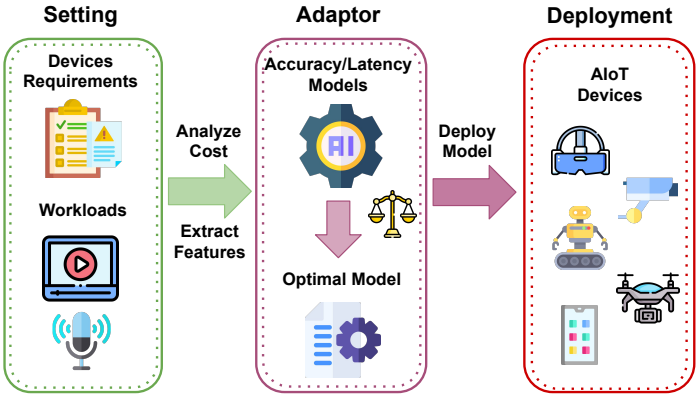

Fig. 12. Illustration of runtime adaptation pipeline.

non-linearity and concurrency. In this way, CoDL achieves higher speedup and more energy saving compared with other methods. Ling et al. [157] propose a model inference abstraction duo-block consisting of a CPU block and a GPU block. Such a duo-block is generated based on neural architecture search (NAS) techniques. They also propose a dynamic cross-processor scheduler that enhances the concurrent real-time DNN inference by optimizing CPU/GPU utilization. Current mobile inference frameworks struggle to efficiently utilize diverse processors for multi-DNN workloads in applications due to a focus on a single DNN per processor, hampering performance and posing a challenge to serving multi-DNN tasks. To address this issue, Jeong et al. [104] propose Band, a mobile DNN runtime for scheduling multi-DNN requests based on a central component. They propose using a model analyzer for model partitioning into subgraphs. A scheduler assigns subgraph-worker pairs, followed by execution of subgraphs on relevant processors by workers. In this way, Band outperforms TensorFlow Lite in terms of end-to-end performance. Xu et al. [285] propose Mandheling, a system that leverages the benefits of Digital Signal Processors (DSP) in integer-based numerical computations during mixed precision training. They propose a co-scheduling technique between CPU and DSP to mitigate the overhead caused by DSP-unfriendly operators, which achieves latency improvement. In addition, they propose incorporating DSP compute subgraph reuse, self-adaptive rescaling, and batch splitting to collaboratively eliminate the preparation overhead on DSP. Wei et al. [267] propose NN-Stretch, an automated model adaptation strategy that splits the DL model based on processor architecture traits. They propose structure-preserved meeting point identification and capacity-guaranteed depth-width scaling. They also propose a sub-graph-based spatial scheduler for parallel inference across heterogeneous processors. Another crucial component of cross-processor inference is distributing the workload across various processing units to minimize idle time. Park et al. [210] propose PointSplit, a 3D object detection framework for multi-accelerator edge devices. They propose a 2D semantics-aware biased sampling method to sample two complementary point sets and schedule them to be processed on GPU and NPU separately.

3.1.4 Runtime Adaptation. Runtime adaptation in on-device inference refers to the ability of AI models to adjust and tailor their runtime behaviors in response to the changing available resources of the devices and evolving data inputs over time to deliver optimized system performance. For example, input images with contents that are easy to recognize do not need a large DNN model to process. Given that, in [57], the authors propose FlexDNN, an input-adaptive framework which leverages the early exit mechanism to construct a single DNN model but dynamically adapts its model capacity to matching the difficulty levels of the input images at runtime. In this way, FlexDNN is able to achieve a significant reduction in frame drop rate and energy consumption while maintaining accuracy. Xu et al. [295] propose ApproxDet, a multi-branch framework employed to identify the

optimal configuration branch for adaptive video object detection based on the characteristics of video content and available resources at runtime. They propose an accuracy and latency-driven scheduler to select the optimal execution branch for the specific user requirement, which achieves 52.9% latency reduction with higher accuracy over YOLOv3 and lower switching overhead compared to other baselines. Feng et al. [60] propose Palleon, which dynamically selects an optimal DNN model by automatically detecting class skews. They propose a class-skew detector to generate precise class skew profiles and catch class skew switches. In addition, they propose Bayesian filter and separability-aware model selection techniques to improve accuracy and overall energy consumption. Guo et al. [77] propose Mistify, an intermediate layer that automates the process of porting a cloud-based model to a range of models optimized for edge devices across different points in the design space. They propose an architecture adaptor and a parameter-tuning coordinator, which collaboratively selects the optimal model that adapts to users' hardware profiles and performance targets. Lastly, LiteReconfig proposed in [294] consists of two components that collaborate as a scheduler to determine the execution branch to activate at runtime. The first component analyzes the cost and benefits associated with all potential features, and the scheduler selects which features to utilize for selecting the execution branch. The second component chooses the optimal execution branch within the execution kernel to adapt to different video contents and available resources.

3.2 Offloading

Given the limited memory and computing capacities of IoT devices, some of them may not be able to run the most efficient AI models by just using their own onboard resources. In such scenarios, it is necessary to offload the execution of part or even the whole model to nearby resourceful edges or the cloud. As summarized in Figure 11, existing works on offloading can be grouped into four categories: model partitioning, workload partitioning, communication optimization, and privacy-preserving offloading.

Model Partitioning. Model partitioning refers to the task of partitioning the AI model between the IoT devices and the nearby resourceful edge or cloud server such that different parts of the AI model are executed in a distributed manner. For example, Kang et al. [118] propose Neurosurgeon, a framework that automatically partitions the DNN computation at the layer level. Neurosurgeon partitions the DNN into two parts for computation on mobile devices and the cloud, respectively, and trains a predictive model during the deployment phase to identify the optimal partition point of the model. In this way, Neurosurgeon achieves significant end-to-end inference latency and energy consumption reduction over cloud-only methods. Huang et al. [98] propose CLIO, a framework enabling model compilation for extremely resource-constrained devices. They propose a novel technique for progressively partitioning models between the cloud and an end device, offering a variety of accuracy-bandwidth tradeoffs. This technique can be integrated with existing model compression and adaptive model partitioning techniques to achieve enhanced performance. In [99], the authors propose AgileNN, an offloading technique that minimizes online computation and communication costs by putting a few valuable features computed locally and thus reducing the size of the local model. They propose using eXplainable AI to estimate the most important features in the top k and retained by the local network to make a part prediction combined with the prediction by the remote network from other less important features for the final result.

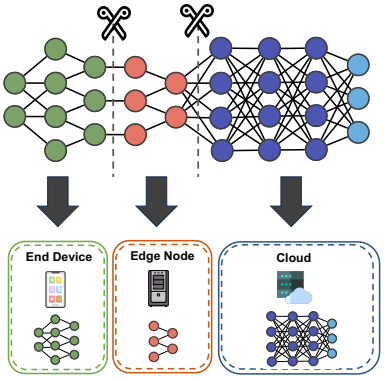

Fig. 13. Illustration of model partitioning.

Workload Partitioning. Workload partitioning refers to the distribution of workloads such as input data (e.g., images, SLAM map) and different DL models within the same processing pipeline across various edge devices and cloud servers to optimize performance, reduce latency, and improve resource utilization. In [70], the

authors propose a sensing algorithm scheduler LEO that specializes in offloading workloads generated by sensor applications to heterogeneous processors. They propose to bring together critical ideas scattered in existing offloading solutions to maximize the performance without changing accuracy, and LEO runs as a service on LPU to perform both frequent and joint schedule optimization for concurrent pipelines, which also makes LEO more energy efficient compared with other baseline methods. Current offloading solutions always assume the presence of a dedicated and robust server to which all inferences can be offloaded. However, it's possible not to be able to find such a server in reality. To address this issue, Zhang et al. [336] propose ELF, a framework that accelerates mobile deep vision applications through parallel offloading, without being restricted to specific server provisioning. They propose a recurrent region proposal algorithm by predicting a new video frame's region proposals based on the ones detected in previous frames, which achieve less latency compared with other baseline methods. Then, these predicted RPs are partitioned into "RP boxes" and offloaded to multiple servers, both partitioning and parallel processing make ELF achieve less resource demands. Ben et al. [15] propose Edge-SLAM, a system that leverages edge computational resources to offload parts of Visual-SLAM. They propose to run the tracking module of VIsual-SLAM on mobile devices and move the left to nearby edge devices, which makes Edge-SLAM achieve significantly reduced latency. Additionally, they propose adding a partial global map as a fixed-size local map on the mobile device to achieve constant memory usage with minimal loss of accuracy in the final map. Another line of research in workload partitioning involves dividing different DL models within the same processing pipeline across edge devices and cloud server. In [115], the authors propose CoEdge, a cooperative edge system for distributed real-time deep learning tasks. They propose a hierarchical DL task scheduling framework integrated with global task dispatching and local batched real-time DL execution to maximize the utilization of edge resources. Additionally, a GPU-aware concurrent DL containerization method is proposed to furnish an isolated execution environment for every task. These techniques make CoEdge achieve less deadline missing rate and less end-to-end latency compared with other baseline methods.

Communication Optimization. Communication between IoT devices and the cloud is often conducted through wireless channels in which the bandwidth can be quite limited. To ensure a timely exchange of migrated workloads between IoT devices and the cloud while minimizing bandwidth usage and power consumption, efficient communication is crucial. Xie and Kim [279] explore a DNN-aware compression algorithm measuring the perception model of a DNN to compress the input while maintaining inference accuracy. They propose to use the gradient concerning the input to characterize the DNN's perception. Using this estimated perceptual model, GRACE addresses a series of optimization challenges to ascertain the optimal codec parameters within the existing codec framework. In this way, GRACE achieves considerable compression ratio gains with little loss of accuracy. In [132], the authors propose SPINN, a synergistic progressive inference system that simultaneously employs an early-exit policy both in the cloud and locally. They propose an early-exit-aware cancellation mechanism that allows the interruption of the inference when having a confident early prediction evaluated by the wrapper of an intermediate classifier to provide robust operation under uncertain connectivity. Additionally, they propose a CNN-specific packing mechanism and an SLA- and condition-aware scheduler that make SPINN achieve higher throughput, higher accuracy, and less energy cost compared with other baseline methods. Yao et al. [304] propose Deep Compressive Offloading, an asymmetric encoder/decoder framework that uses an efficient encoder on a local device while utilizing a relatively complex decoder on a server. In this way, most of the processing burden can be put on the server side and achieve a significant latency improvement. Additionally, they propose an effective system DeepCOD which incorporates a performance predictor and a runtime partition decision maker, which achieves higher speedup for inference. In [303], the authors explored an edge-cloud training pipeline by harnessing parallel processing capabilities spanning both edge and cloud environments. They propose to apply scheduled feature replay and error-feedback compression, which fully utilize the computing capabilities available at the edge. Additionally, they offered a context-aware decision engine to adaptively organize parallel execution

and compression, which keeps the overall latency low. Fu et al. [66] propose Hyperion, a distributed mobile offloading framework that supports various applications and heterogeneous hardware. They propose a regularity-aware kernel analyzer to break down the tasks into smaller parts while ensuring that only the necessary data is transmitted, which makes Hyperion more efficient. Before scheduling, they propose a context-aware computing time predictor to predict the runtime duration of a given slice and a pipeline-enabled and network-adaptive scheduler to determine the optimal number of slices to be offloaded for each computational unit, both achieve superior speedup compared with the baseline. However, as the number of agents increases, the operational overhead, which relies on a central node, also increases. To address this issue, Xu et al. [290] propose SwarmMap, a framework that scales up collaborative edge-based Visual-SLAM service. They propose a change log-based map information tracker to achieve the minimum bandwidth consumption for map synchronization. Additionally, they propose a SLAM-specific task-aware scheduler that makes decisions based on the status of agents to minimize the procession time. Further, they propose a map backbone profiling technique to mitigate storage overhead without reducing accuracy.

Privacy-Preserving Offloading. IoT devices often collect personal data that may contain privacy-sensitive information. In scenarios where data are also needed to offload along with the workloads to edge or cloud servers, it is imperative to ensure that this data is handled in a way that preserves the privacy of users. PriMask [111] introduce a small-scale neural network – named MaskNet – to mask the data before its transmission to the cloud. The data masked by MaskNet cannot be recovered by the cloud, thus preserving the privacy after offloading. Moreover, each mobile device has its own unique MaskNet, which ensures that a privacy breach affecting the MaskNet of one device does not compromise the privacy of data on other devices.

3.3 On-Device Training

Besides on-device inference, another fundamental and essential compute task of AIoT is on-device training. As summarized in Figure 11, existing works on on-device training can be grouped into two categories: training on a single device, and training across distributed devices.

Training on a Single Device. In the case of single-device training, the entire training process takes place on a single device. To achieve effective training on a single device, existing efforts have mainly focused on the exploration of memory optimization. For example, Lis et al. [159] propose Dropback, which only trains a fraction of the weights who have the highest accumulated gradients while keeping the remaining weights not stored in memory, which significantly reduces the memory access cost. Zhang et al. [342] propose MDLdroidLite, a learning framework that transforms regular DNNs into resource-efficient models for on-device learning. They propose a Release-and-Inhibit Control (RIC) technique to wisely grow each layer independently from tiny to backbone, which avoids redundant resource overhead. In addition, they propose a RIC-adaption pipeline that transfers existing parameters to new-born parameters during growth. In this way, MDLdroidLite achieves 28X to 50X fewer model parameters compared with other baselines. In [352], the authors propose Octo, a cross-platform system designed for lightweight on-device learning that leverages the fixed-point computational capabilities of embedded processors. They propose an INT8 training technique with loss-aware compensation and parameterized range clipping methods to efficiently apply quantization in forward pass and backward pass, respectively. In this way, Octo achieves higher training efficiency compared with other baselines. Wang et al. [260] propose Melon, a memory-optimized on-device training framework that retrofits established recomputation and micro-batch techniques to fit into resource-constrained devices. They further propose a lifetime-aware memory pool to optimize memory utilization based on the characteristics of DNN training. In addition, they propose an on-the-fly memory adapting technique to quickly adjust to changes in the memory budget and resume execution using the partial results. In this way, Melon achieves higher training throughput with the same batch size. In [73], the authors propose Sage, an on-device training framework that incorporates memory-optimized techniques. They

propose to separate differentiable operations from computable operations by employing a two-layer abstraction to represent a node in the computational graph, then Sage applies operator fusion and subgraph reduction to minimize the graph size. Additionally, they propose to dynamically adapt to the memory budgets by using gradient accumulation and checkpointing. Lin et al. [154] propose an on-device training framework with algorithm-system co-design. They propose a quantization-aware scaling technique to align the accuracy with the floating-point counterpart by automatically scaling the gradient with varying bit-precision. To save memory during backward computation, they propose a sparse update technique to skip the computation of less important layers and sub-tensors. In [100], the authors propose ElasticTrainer, a technique that can dynamically select the optimal trainable network portion at training time. They propose a tensor importance evaluator by leveraging the XAI technique to define the importance of a tensor in a specific epoch. On the other hand, they propose a tensor timing profiler to compute the backward pass timing of each tensor. Based on importance and time, they propose a tensor selector to select the optimal trainable network portion, which makes ElasticTrainer achieve higher training speedup with less energy consumption compared with baselines.

Training across Distributed Devices. In the case of training across distributed devices, DL models are trained collaboratively across a network of IoT devices where data on each device can be exchanged with other devices. In doing so, the collective computational power and data across the multiple devices can be leveraged to jointly train and update the DL models. For example, Zhang et al. [343] propose MDLdroid, a decentralized mobile DL training framework for mobile sensing applications. They propose a chain-directed synchronous stochastic gradient descent algorithm that dynamically aggregates and manages the model with one of the neighbors based on runtime resource status. Additionally, they propose a chain-scheduler, an agent-based multi-goal reinforcement learning technique, incorporating an accelerated reward function to effectively and equitably manage and allocate resources. In this way, MDLdroid achieves high training accuracy with low overhead. As another example, in [318], the authors propose Mercury, an importance sampling-based on-device distributed training framework. The key principle behind the design of Mercury is that not all the data samples contribute equally to model training. Given that, in each training iteration, Mercury identifies and selects data samples that provide more important information. By focusing on those more important data samples, Mercury considerably enhances the training efficiency of each iteration. As a result, the total number of iterations and total training time is reduced.

3.4 Federated Learning

As data collected by IoT devices often contain privacy-sensitive information, federated learning (FL) emerges as a privacy-preserving approach that can train models across decentralized devices while keeping data on each device to preserve data privacy [117, 256, 333]. Unlike fully on-device training, FL has the advantage of allowing information to be shared among devices, making it suitable for more complex applications that require more data volume. Instead of gathering data from different devices into a central server for training, the model is disseminated to the participating devices in FL. These devices then conduct local training for a number of rounds and communicate only their model updates or gradients back to the central server for aggregation. The updated global model is subsequently broadcasted to the next set of participating devices for further training rounds [189]. As summarized in Figure 11, existing works on FL for IoT can be grouped into seven categories: data heterogeneity, communication optimization, system heterogeneity, personalization, client Selection, model heterogeneity, as well as frameworks and benchmarks.

Data Heterogeneity. Unlike centralized training, data distributed across the devices participating in the FL process is generally non-IID (non-independent and identically distributed). Such data heterogeneity could make the local models overfit to local data, and aggregating these models could lead to convergence issues. Shuai et al. [236] propose BalanceFL, which scales the model weights making it behave as if it was trained on uniform distributed data. As such, it allows the global model to effectively learn both common and rare classes from a

long-tailed real-world dataset, and thus mitigates the bias caused by data heterogeneity. Shin et al. [235] propose FedBalancer, which uses a data selection strategy to select informative samples with adaptive deadline control. In doing so, the global model avoids overfitting caused due to data heterogeneity and makes convergence more stable. Lastly, Zhang et al. [331] propose GPT-FL, which pre-trains the global model using synthetic data generated by generative models before fine-tuning with federated training. This makes the global model start from a more stable point instead of starting from scratch such that data heterogeneity does not strongly affect convergence.

Communication Optimization. Communication between client devices and the central server in FL is often conducted through bandwidth-limited wireless networks. Therefore, reducing bandwidth usage between client devices and the central server can significantly enhance FL efficiency. Shi et al. [232] introduce gTop-k. Instead of accumulating the local top-k gradients from all the clients to update the model in each iteration, gTop-k chooses the global top-k gradients from a subset of clients, which considerably reduces the amount of gradients to communicate. Reisizadeh et al. [218] propose FedPAQ, which quantizes model updates to reduce their sizes before uploading to the server while the server only periodically averages the updates. The quantized updates and the periodic averaging on the server lead to lower communication costs. Similarly, Jhunjhunwala et al. [106] propose an adaptive quantization scheme called AdaQuantFL, which achieves communication efficiency through quantization while maintaining a low error floor by changing the number of quantization levels during training. Lastly, Das et al. [45] propose FedGLOMO to reduce the variance of local updates by global aggregation with momentum. This results in faster convergence and an overall lower number of communication rounds.

System Heterogeneity. The participating devices in FL can be heterogeneous in their available on-device computing resources and network bandwidths. Such system heterogeneity would inevitably cause different participating devices to complete their local training at different times. Consequently, the slowest clients become the bottlenecks in the FL process. One key technique to address system heterogeneity is the design of semi-asynchronous or asynchronous communication protocols. For example, Wu et al. [270] introduce SAFA, which uses a lag-tolerant model distribution algorithm and version-aware aggregation method based on a cache system. This decouples the global model broadcast and gradient upload process, making the system more tolerant of lagging clients. Ma et al. [182] propose FedSA, which is a semi-asynchronous mechanism where the server aggregates a subset of local models by their arrival order in each round. The authors show that this approach improves convergence both theoretically and experimentally. Xie et al. [277] propose FedAsync, where the updates to the server and the broadcast to the clients are done asynchronously with a buffer. The updates from clients that are far behind the server schedule are deprioritized or excluded entirely. This avoids the destabilizing effects of stragglers and increases the number of communication rounds the system can complete within a time frame. Nguyen et al. [200] propose FedBuff, which also uses a buffered asynchronous aggregation scheme sending updates asynchronously but aggregating and broadcasting updates synchronously. This not only makes the system lag-tolerant, but also makes it compatible with Secure Aggregation and Differential Privacy. Sun et al. [242] introduce FedSEA in which the authors design a scheduler that can efficiently predict the arriving time of local updates from devices and adjust the synchronization time point according to the devices' predicted arriving time. In doing so, it reduces the total number of straggling clients. Zhang et al. [334] propose an asynchronous FL framework named TimelyFL. The key idea of TimelyFL is adaptive partial training, which allows each client to train part of the model based on the available resources of each client at runtime. In doing so, more clients are able to join in the global update without staleness.

Personalization. Besides training a global model, another use case of FL is to personalize the global model for participating clients such that the personalized model can better fit the needs of the end user. For example, Sub-FedAvg [253] creates a personalized sub-network for each client from the global model by applying structured pruning on convolutional filters and unstructured pruning on fully connected layers. Li et al. [138] propose

FedMask where each device learns a sparse binary mask and applies the learned sparse binary mask to local models to create personalized and sparse local models for each client. Instead of creating a personalized model for each user, Tu et al. [252] propose FedDL, a clustering approach in which the client pool is grouped into several clusters, and one personalized model is assigned to each cluster. Similarly, AttFL [209], designed for time series mobile and embedded sensor data, groups clients with similar contextual goals using cosine similarity, and redistributes updated personalized model parameters for improved inference performance at each local device. Deng et al. [47] propose TailorFL, a resource-aware and data-directed pruning strategy that makes each device's sub-model structure match its available resource and correlate with its local data distribution. Lastly, FedSelect [245] incrementally expands sub-networks to personalize client parameters, concurrently conducting global aggregations on the remaining parameters. This enables the personalization of both client parameters and sub-network structure during the training process.

Client Selection. In each round of FL, the central server selects a subset of clients to participate in the federated training process. The client selection strategy to determine which subset of clients to be included in each round plays a significant role in FL. For example, Lai et al. [128] introduce Oort, a utility-based client selection scheme that takes both data and system utilities into account, where data utility is measured by the importance of model update and system utility is measured by the local training speed and the available network bandwidth for communication. By selecting clients with the highest utilities, Oort enhances both data and system efficiency and outperforms random client selection in terms of time-to-accuracy performance. PyramidFL [143] moves one step further and proposes to exploit data and system utilities within the selected clients to further enhance the time-to-accuracy performance of federated training. Lastly, Ouyang et al. [206] introduce ClusterFL that minimizes the empirical training loss of multiple learned models while automatically capturing the intrinsic clustering relationship among the clients. This helps select and drop the clients with little correlation with others in each cluster, which speeds up the federated training process.

Model Heterogeneity. In standard FL, the participating clients and the central server collaboratively train the same model. However, imposing the same model on all devices would exclude low-end devices that do not have the enough memory. Moreover, state-of-the-art AI is increasingly reliant on large models, such as LLMs. Requiring the server and client models to be identical makes it impossible for standard FL to train such large models due to the resource limitations of client devices. Given that, model-heterogeneous FL was introduced to address this issue, allowing for the training of models with varying capacities across the server and clients. One primary approach for model-heterogeneous FL is based on knowledge distillation (KD). For example, Li and Wang [146] propose FedMD, where clients train their own local models on a public dataset and upload their logit vectors to the server for KD. Since only logits are sent, clients' local models can have different architecture and sizes. Lin et al. [156] propose FedDF to train the global model through ensemble distillation in which client models with different sizes and architectures are used as teachers. An unlabeled dataset is used and the predictions of the teacher models on that dataset are used to distill the global model. Similarly, Cho et al. [40] propose Fed-ET in which models of different architectures and sizes are trained on clients' private data and then used to train a larger model at the server. However, Fed-ET uses weighted consensus distillation where the client updates are weighed based on a consensus function. This deprioritizes underperforming clients, resulting in higher accuracy. The other primary approach for model-heterogeneous FL is based on partial training where different parts of the global model are extracted and disseminated to different clients for local training. For instance, Federated Dropout [22] propose to randomly extract sub-models of different sizes from the global model. Given the random nature, the sub-models extracted from the global model in each round can be different. During the update step, the server aggregates the sampled client updates with weighted averaging based on how many updates each part of the global model receives. Different from Federated Dropout [22], HeteroFL [48] and FjORD [92] propose static sub-model extraction schemes where the sub-models extracted from the global model in each round are

always the same. However, the key issue of static sub-model extraction schemes is that part of the global model cannot be trained on data across all the clients. This inevitably biases the global model training, especially data heterogeneity across the clients is high. To address this key issue, Alam et al. [6] propose FedRolex, which is a rolling sub-model extraction scheme that allocates sub-models of different sizes to clients and progressively rolls the sub-model extraction window across the entire global model. In doing so, all parts of the global model are evenly trained on the entire client data.

Frameworks and Benchmarks. Frameworks and benchmarks play important roles in enabling FL on IoT devices. Popular FL frameworks include FedML [85], which implements a wide range of FL algorithms and datasets to facilitate developing and evaluating FL algorithms for a wide range of applications. Flower [16] is another FL framework that is built on top of Ray [199] and is heavily customizable to different FL algorithms. FedScale [127] and FLUTE [49] provide high-level APIs to implement, deploy, and evaluate FL algorithms at scale. These frameworks are, however, not specifically geared towards IoT devices. In terms of benchmarks, existing FL benchmarks are predominantly conducted on datasets in domains of computer vision (FedCV [86]), natural language processing (FedNLP [152]), medical imaging (FLamby [248]), speech and audio (FedAudio [332]), multimodal (FedMultimodal [61]), and graph neural networks (FedGraphNN [84]). These datasets, however, do not come from genuine IoT devices and therefore do not accurately reflect the distinctive characteristics of IoT data. In contrast, in [7], the authors propose FedAIoT, an FL benchmark designed for IoT devices. FedAIoT includes eight datasets collected from IoT devices such as smartphones, smartwatches, Wi-Fi routers, drones, and smart home sensors. It also includes an FL framework customized for IoT, which supports IoT-friendly models and facilitates non-IID data partitioning, IoT-specific data preprocessing, quantized training, and noisy label emulation.

3.5 AI Agents for AIoT

Traditional machine learning approaches focus on low-level basic recognition tasks. However, real-world applications can be complicated and require not only basic perception but also performing more complicated tasks such as making higher-level plans and decisions based on reasoning. AI agents, powered by advanced generative AI models such as LLMs, can autonomously perform such complicated tasks, thereby significantly enhancing the capabilities of AIoT. Some efforts have been made to build AI agents for AIoT. For example, multimodal input is crucial for developing AI agents for AIoT as IoT devices in general collect data from multiple sensing modalities such as language, vision, and audio. Chen and Li [36] introduce Octopus v3, a multimodal model with functional tokens tailored for AI agents, which supports both English and Chinese and operates efficiently on various edge devices such as Raspberry Pi. In [257], the authors introduce Mobile-Agent, an AI agent designed for mobile devices. Mobile-Agent can interpret user instructions to identify and locate elements on the mobile app's interface. It then autonomously plans and executes tasks, navigating apps step-by-step without requiring system-specific customization. Wen et al. [268] introduce AutoDroid, a mobile task automation framework designed to handle arbitrary tasks on an Android application without manual intervention. AutoDroid combines the capabilities of LLMs with dynamic app analysis to manage unseen tasks. During the offline stage, it gathers app-specific knowledge by exploring UI relationships and creating simulated tasks. In the online stage, AutoDroid uses memory-augmented LLMs to guide the next actions and complete tasks based on these suggestions. Experimental results demonstrate that AutoDroid effectively automates tasks and outperforms existing training-based and LLM-based methods. Lastly, as another line of research, text rewriting is a crucial feature of AI agents, as it can enhance communication by transforming informal or incorrect text into well-structured content. Despite advancements in LLMs for text summarization and rewriting, their large size and computation time make them challenging to use on mobile devices. Developing a smaller model with similar capabilities is also challenging due to the need to balance size and performance and the requirement for expensive data labeling. Zhu et al.

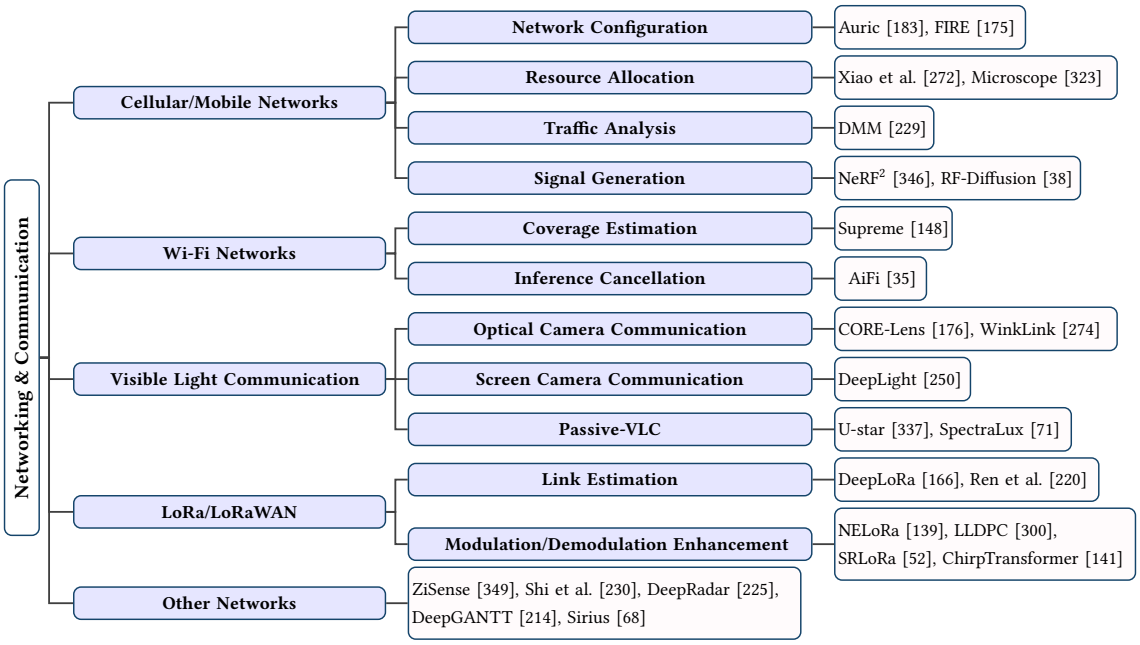

Fig. 14. Summary of topics related to networking & communication.

[353] present MessageRewriteEval, a compact yet powerful language model for text rewriting tasks that operate efficiently on mobile devices. They present an innovative method for fine-tuning instructions for a mobile-centric text rewriting model, enabling high-quality training data generation without human labeling.

4 NETWORKING & COMMUNICATION

4.1 Cellular/Mobile Networks

As cellular networks evolve over many generations, they play an increasingly important role in providing mobile, reliable, and evolving communication. As summarized in Figure 14, existing works on AI-empowered cellular/mobile networks can be grouped into four categories: network configuration, resource allocation, traffic analysis, and signal generation.

Network Configuration. Cellular/mobile network parameters are typically manually configured based on rulebooks. Unfortunately, this process is time-consuming, error-prone, and difficult to maintain. AI-guided network configuration has been explored to provide a data-driven approach to improve network performance and service robustness. For example, adding new carriers to accommodate increasing voice and data traffic can make cellular network configuration tasks very challenging. To address this issue, Mahimkar et al. [183] propose Auric, which uses a series of carrier attributes as inputs to train a DL model and outputs network configuration based on geographical proximity. Experimental results show that Auric leads to 96% accuracy across a large number of carriers and configuration parameters when evaluated on real-world LTE network data. As another line of research, Liu et al. [175] introduce FIRE, a system that employs a variant of variational autoencoders (VAE) for downlink channel estimation. In doing so, it eliminates the overhead of requesting feedback from client devices and improves the quality of FDD (Frequency Domain Duplexing) MIMO systems. Moreover, FIRE effectively

supports MIMO transmissions in real-world settings, achieving an SNR enhancement of over 10 dB compared to the state of the arts.

Resource Allocation. AI techniques can also enhance the performance of cellular/mobile networks by managing and distributing critical network resources based on various factors such as demand, network conditions, and service requirements in a data-driven manner. For example, Xiao et al. [272] conduct an extensive measurement study on the ecosystem of mobile virtual network operators (MVNO). Based on the findings, the authors propose to leverage big data analytics and ML-based techniques to optimize an MVNO's service such as predicting monthly data usage to optimize data plan reselling, customer churn profiling, and mitigation. As another example, Zhang et al. [323] propose Microscope, a DL-based framework that decomposes per-service level resource demand based on spatio-temporal features hidden in traffic aggregates. In doing so, Microscope reduces relative demand estimation error to below 1.2%, allowing cellular operators to allocate network resources more accurately.

Traffic Analysis. Traffic analysis refers to the techniques to monitor, analyze, and optimize the flow of data across the network. AI-based traffic analysis can help in forecasting future traffic demands and making adjustments to enhance the overall network efficiency. Shen et al. [229] propose a fast map-matching system named DMM for cellular data. DMM utilizes a recurrent neural network (RNN) to determine the most probable road trajectory given a series of cell tower locations. To make DMM practically useful in real-world scenarios, DMM also incorporates a number of techniques such as spatial-aware representation of cell tower sequences, an encoder-decoder structure for variable input and output lengths, and a reinforcement learning-based model to optimize the matched results.

Signal Generation. Lastly, the success of Generative AI in natural language processing and computer vision has sparked interests in using Generative AI in the domain of cellular/mobile networks. For example, NeRF^2 [346] introduces a radio-frequency (RF) radiance field that uses a Neural Radiance Network to model a continuous volumetric scene function, which captures the propagation of RF signals in complex environments. The model trained with signal measurements and a physical model of ray tracing can generate synthetic RF datasets that can be adopted to boost the training of application-layer artificial neural networks (ANNs). Experimental results demonstrate the effectiveness of NeRF^2 in the fields of indoor localization and 5G MIMO. As another example, Chi et al. [38] present RF-Diffusion, a novel approach for generating high-quality time-series RF signals using a generative model. The method involves using time-frequency diffusion theory and a hierarchical diffusion transformer to generate high-quality synthetic RF signals by leveraging the unique characteristics of RF signals in both time and frequency domains. RF-Diffusion demonstrates superior performance compared to other generative models including DDPM, DCGAN, and CVAE, achieving higher structural similarity and better SNRs.

4.2 Wi-Fi Networks

As summarized in Figure 14, existing works on AI-empowered Wi-Fi networks can be grouped into two categories: coverage estimation and interference cancellation.

Coverage Estimation. AI-empowered Wi-Fi coverage estimation aims to leverage AI algorithms to obtain the distribution and strength of Wi-Fi signals in a specific area with higher resolution. For example, inspired by advancements in image super-resolution, Li et al. [148] propose Supreme, which constructs fine-grained radio maps based on coarse-grained radio maps crowd-sourced across sites with a deep spatial-temporal reconstruction network consisting of 3D convolutions, spatial-temporal residual blocks, and reconstruction subnets. The authors have conducted experiments on a dataset consisting of six months of data collected from two university campuses. Experimental results show that Supreme outperforms state-of-the-art baselines based on coarse-grained radio maps and achieves lower localization error in a Wi-Fi fingerprint-based localization case study.

Interference Cancellation. As the number of wireless devices increases, multiple devices may simultaneously transmit data within the same unlicensed Wi-Fi band. This can cause severe performance degradation. To ensure

reliable communication, advanced interference cancellation techniques are needed. Chen et al. [35] introduce AiFi, an AI-empowered interference cancellation method for commodity Wi-Fi devices to estimate interference using knowledge gathered from the Wi-Fi receiver's physical layer without extra RF hardware. AiFi leverages the domain knowledge of Wi-Fi physical layer information including pilot information (PI) and channel state information (CSI) to guide the DL model design. Specifically, AiFi first extracts the interference features from Wi-Fi physical layer, estimates interference via an attention network using these features, and finally removes those interference from the received signal using a fully-connected network and an LSTM. Experiments show that AiFi effectively boosts the MAC frame reception rate by 18× with a cancellation delay under 1ms per frame.

4.3 Visible Light Communication

Visible light communication (VLC) uses visible light as a data transmission medium to connect devices and communicate. VLC requires bit encoding using visible light sources, and light-sensitive sensors as receivers. As summarized in Figure 14, in VLC, AI has been used to improve the performance of optical-camera communication, screen-camera communication, and passive-VLC.

Optical-Camera Communication. Optical-camera communication (OCC) relies on LED lighting infrastructures as transmitters and cameras as receivers. The coded information is either transmitted directly from LED lights or reflected from the illuminated objects and is received by the camera. Liu et al. [176] introduce CORE-Lens, which addresses the challenges posed by the mutual interference between OCC and object recognition (OR) in indoor environments. Traditional OCC systems often suffer from the entanglement of light patterns used for communication with the background, which degrades both OR accuracy and OCC decoding performance. CORE-Lens addresses these challenges by employing a disentangled representation learning (DRL) approach combined with GAN-based image reconstruction. Experimental results show that CORE-Lens achieves superiority in both visible light sensing and communications compared to conventional approaches. Xiao et al. [274] propose WinkLink, an OCC system that enables robust transmission behind complex backgrounds even under low SNR conditions. They design a two-stage DNN and a context-aware demodulation protocol to extract subtle signals in the lossy OCC channel. WinkLink is trained solely on a synthesized dataset yet generalizes well to unseen real-world backgrounds.

Screen-Camera Communication. Screen-camera communication (SCC) encodes video content in a human-imperceptible manner on a screen as the light source, and uses cameras capturing images of such screen content work as receivers. Existing techniques on SCC often suffer from high decoding errors due to screen extraction inaccuracies and perceptible flickers on common refresh rate screens. To address this issue, Tran et al. [250] present DeepLight, an innovative approach for SCC that addresses the challenges of decoding inaccuracies and perceptible screen flickers. For the bit encoder, DeepLight applies a Manchester coding strategy. For the decoder, Deeplight adopts the state-of-the-art deep object detection pipeline to extract the screen from a camera frame and then adopts a DNN-based model to decode spatially encoded bits in the frame simultaneously. Experimental results show that DeepLight is able to achieve high decoding accuracy (frame error rate < 0.2) and moderately high data throughput (≥ 0.95Kbps) at extended distances.

Passive-VLC. Instead of relying on active light sources for data transmission, passive-VLC uses ambient light which can be modulated and then detected by a receiver to decode the transmitted information. Essentially, passive-VLC systems leverage changes in light intensity or other properties of ambient light to convey information. Zhang et al. [337] design U-star, a system consisting of passive Underwater Optical Identification (UOID) tags and DL-enabled camera-based tag readers, providing objects/human identification and location-based services as underwater navigation assistance in scenarios such as dive and rescue. U-star employs a three-dimensional multi-color cube-shaped design for the UOID tags and adopts the CycleGAN-based underwater denoising model

that converts underwater UOID images into clear ones. Experiments under different underwater scenarios show that U-star achieves a bit error rate of 0.003 at 1m and less than 0.05 at up to 3m, which is sufficient for guiding underwater navigation. Ghiasi et al. [71] present SpectraLux, an approach to transmit and decode data using low-power liquid crystal (LC) cells. It utilizes the physical characteristics of LC shutters toggling between being translucent and opaque when switching the voltage from 0V to 5V, emitting different spectrums of the incident light. SpectraLux adopts a spectrometer that captures 256 bands of incoming light and achieves multi-symbol decoding by feeding PCA-reduced spectrum features to CNNs for classification. SpectraLux shows the potential of utilizing the wide spectrum of ambient light in passive-VLC.

4.4 LoRa/LoRaWAN

LoRa (Long Range) is a rising low-power wide-area communication technology. LoRa's physical layer adopts the chirp spread spectrum (CSS) modulation which is known for its resistance to interference and capacity to travel long distances, making it particularly suitable for various IoT applications. LoRaWAN (Long Range Wide Area Network) refers to the protocol and system architecture for networks of LoRa nodes which is an open standard that ensures interoperability among different manufacturers and developers. As summarized in Figure 14, existing works on AI-empowered LoRa/LoRaWAN can be grouped into two categories: link estimation and modulation/demodulation enhancement.

Link Estimation. To study LoRa link coverage in the wild in supporting smarter LoRa deployments, Liu et al. [166] propose DeepLoRa, a DL-based framework for LoRa path loss estimation of long-distance links in real-world environments. To do so, DeepLoRa extracts land-cover types along a LoRa link from multi-spectral remote sensing images, and exploits the order dependency of the land-cover sequence by utilizing Bi-LSTM (Bidirectional Long Short Term Memory) for path loss estimation. Experimental results on a real LoRaWAN dataset show that DeepLoRa is able to achieve less than 4dBm estimation error, which is 2× smaller than state-of-the-art approaches. Moreover, the study conducted in [220] further corroborates that DeepLoRa outperforms other link estimation approaches in terms of LoRa localization accuracy.

Modulation/Demodulation Enhancement. Enhancements in LoRa modulation and demodulation are essential for improving the performance, efficiency, and reliability of data transmission in LoRa systems. For example, Li et al. [139] present NELoRa, a neural-enhanced LoRa demodulation framework that takes advantage of the powerful feature learning capability of DL to enable LoRa communication under ultra-low SNR. The key idea of NELoRa is the dual-DNN design: the first DNN is used as a noise filter to extract clean chirp symbols from the noisy LoRa packets, and the second DNN is used as a decoder that decodes the extracted clean chirp symbols. Experimental results show that NELoRa outperforms the standard LoRa demodulation method under a wide range of LoRa configurations in both indoor and outdoor deployments. Yang and Du [300] propose LLDPC, which enables low-density parity-check (LDPC) coding in LoRa networks under the inspiration of the wide usage of LDPC coding in other wireless networks. LDPC requires the Log-likelihood Ratio (LLR) for decoding which is not applicable to the CSS modulation adopted by LoRa. Moreover, the mainstream decoding algorithms for LDPC need multiple iterations to achieve effective error correction, resulting in long decoding latency that exceeds the maximum ACK time of the LoRa gateway. To tackle these challenges, LLDPC extracts LLR by treating CSS demodulation as a classification task and outputs the probability of all possible decoding results. It further utilizes a Graph Neural Networks (GNN) for fast belief propagation to achieve efficient LDPC decoding. Du et al. [52] propose SRLoRa, which decodes LoRa signals by leveraging spatial diversity from multiple gateways. Specifically, SRLoRa employs CNN-based interleaving denoising layers to extract features under ultra-low SNR and consolidates features from different gateways in the merging layers. The merged signals with accumulated energy are then fed to a CNN decoder for decoding. Lastly, Li et al. [141] further establish an encoding framework, providing four features including on-air time, selective initial frequency, chirp repeating, and symbol hopping, to

combat various challenges of weak signals, signal collisions, and environment dynamics. On the decoder side, the neural-enhanced decoder is adopted and optimized for decoding the symbols with symbol hopping based encoding in terms of input and parameter sizes.

4.5 Other Networks

Besides the wireless networks mentioned above, AI has also been applied to various other types of networks for diverse objectives. For instance, ZiSense [349] is proposed to enhance the energy efficiency of sensor nodes in co-existence environments by using a sequence of received signal strength (RSS) values to predict the presence of ZigBee signals through a decision tree model. Shi et al. [230] propose to improve the configuration of wireless mesh networks (WMN) by DL-based domain adaption that adapts models for network configuration prediction trained on simulation to its corresponding physical network. In particular, the authors develop a teacher-student neural network that learns robust configuration prediction models from large-scale inexpensive simulation data with minor physical measurements to close the simulation-to-reality gap. Perez-Ramirez et al. [214] present DeepGANTT, a DL-based scheduler that leverages GNN to provide a near-optimal solution for the NP-hard carrier scheduling problem in RFID backscatter networks. In those networks, battery-free RFID tags harvest energy from excitation in the environment, and IoT devices equipped with RFID readers provide them with the carrier for communication. To avoid collisions, DeepGANTT trains a carrier scheduler based on GNN to handle and capture the interdependence of nodes in the irregular network topology graphs. DeepGANTT breaks the scalability constraints of the optimal scheduler used for training and can generalize to networks 6× larger in the number of nodes and 10× larger in the number of tags. Sarkar et al. [225] propose DeepRadar that utilizes DL to detect radar signals and estimate their spectral occupancy for incumbent protection and efficient spectrum sharing. This approach involves spectrogram image learning (SIL) based on YOLO (You Only Look Once) model that learns an object detection model using spectrograms, including both radar and non-radar data. Lastly, Garg and Roy [68] design Sirius, a self-localization system, where the node computes its own location onboard, using a single receiver for low-power IoT nodes to close the gap between the needs for accurate and robust localization and the lack of efficient solutions in the low-power scenario. Instead of relying on strictly synchronized antenna arrays to estimate angle-of-arrival (AoA) and time-of-flight (ToF) which requires resources low-power nodes do not possess, Sirius uses antennas whose gain pattern can be reconfigured by the on/off of controllable switches in real-time to embed direction specific encoding to the received signal. The gain patterns are passed to AI models to estimate the angle in degrees. Experimental results show that Sirius is able to obtain competitive performance compared to state-of-the-art antenna array-based systems, achieving 7-degree median error in AoA estimation and 2.5-meter median error in localization.

5 DOMAIN-SPECIFIC AIOT SYSTEMS

5.1 Healthcare and Well-being

One important application domain of AIoT systems is healthcare and well-being. As summarized in Figure 15, existing works on AIoT systems for healthcare and well-being can be grouped into four categories: vital sign monitoring, in-situ illness detection and monitoring, assistive technology, and personal health insight generation.

Vital Sign Monitoring. One of the primary use cases of AIoT systems developed for healthcare and well-being is monitoring an individual's vital signs such as cardiac signals, breathing, and blood pressure. For instance, one of the key challenges of vital sign monitoring is motion artifacts caused by body movements. In [37], the authors introduce MoVi-Fi to monitor breathing and heartbeats in a contactless way using RF signals under the existence of body movements. MoVi-Fi utilizes deep contrastive learning to separate vital signs from the body movements and further uses an encoder-decoder model to refine and recreate the vital sign waveforms. In [335], the authors observe that vital signs including breathing and heartbeats cause subtle facial vibrations.

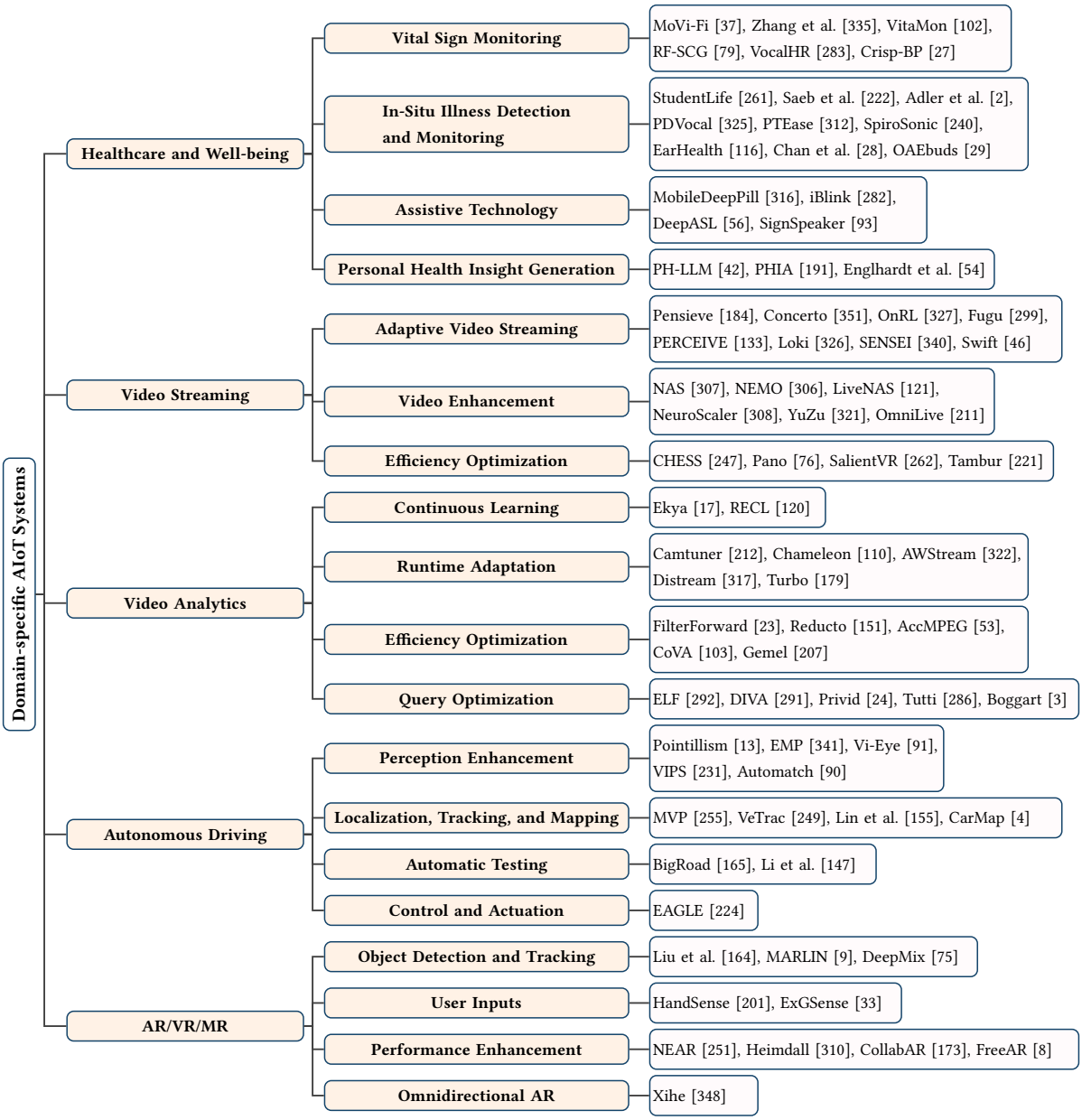

Fig. 15. Summary of topics related to domain-specific AIoT systems.

They propose to use the motion sensors inside the commodity AR/VR headsets to capture those subtle facial vibrations and employ an LSTM-based model to reconstruct the vital sign waveforms. VitaMon [102], RF-SCG [79], and VocalHR [283] focus on monitoring cardiac signals. Specifically, VitaMon [102] proposes to use video to measure the inter-heartbeat interval (IBI). Since blood absorbs more light than other tissues, video can effectively

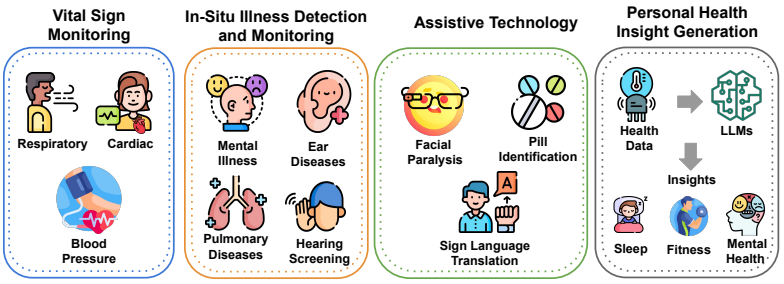

Fig. 16. Summary of AIoT systems for healthcare and well-being.

detect the changes in blood vessel volume that occur with each heartbeat. Based on this principle, VitaMon employs a CNN to identify and reconstruct the peak of each heartbeat across consecutive facial image frames. RF-SCG [79], on the other hand, uses mmWave to reconstruct the seismocardiogram (SCG) waveforms that detect fine-grained cardiovascular events. RF-SCG emits mmWave radar signals and captures the reflections from the human body, and proposes a CNN-based model to translate the mmWave reflections to SCG waveforms. VocalHR [283] proposes to infer cardiac activities from human voice production. It extracts phonation and articulatory features from human voice that are related to cardiac activities, and transforms these vocal features into cardiac activities through an LSTM-based model. Lastly, Cao et al. [27] introduce Crisp-BP, a blood pressure monitoring system that leverages wrist-worn devices equipped with PPG sensors. The sensors emit green and infrared light, which measures volume changes in blood vessels, which are processed by a BLSTM-based model to estimate both diastolic and systolic blood pressure.

In-Situ Illness Detection and Monitoring. Another important use case of AIoT systems developed for healthcare and well-being is to detect or monitor the progress of illnesses such as mental illnesses, lung and ear diseases in non-clinical settings. Mental illnesses are a leading cause of disability worldwide [197]. One pioneering work for mental illnesses is StudentLife [261], where the study identifies relationships between smartphone sensor data and students' mental health and academic performance. As another pioneering work in this domain, Saeb et al. [222] propose to use smartphone GPS data and phone usage data to capture and detect various daily-life behavioral markers from individuals with depression and utilize AI models to analyze the collected sensor data to infer depressive symptom severity. Adler et al. [2] focus on leveraging smartphone sensor data to predict early warning signs of psychotic relapse in patients with schizophrenia spectrum disorders. They develop encoder-decoder neural network models that could identify behavioral anomalies occurring within 30 days before a relapse. Parkinson's disease is another use case. It is observed that non-speech body sounds [216], such as breathing and throat-clearing sounds, are highly correlated to Parkinson's disease. Zhang et al. [325] propose PDVocal, which leverages everyday smartphone voice activities such as calls and chats to capture these sounds and employs a ResNet-based DL model to assess the presence probability of Parkinson's disease. In terms of lung diseases, Yin et al. [312] introduce PTEase, a 3D-printed mouthpiece that attaches to a smartphone for pulmonary disease detection. The smartphone emits acoustic waves via its speaker. These waves travel through the airway and are captured by the smartphone's microphone, providing detailed information about the user's airway conditions which is crucial for pulmonary disease detection and lung function assessment. Similarly, Song et al. [240] introduce SpiroSonic, a smartphone-based system for conducting spirometry tests by monitoring the motion of the chest wall during breathing. SpiroSonic emits an ultrasound wave from the smartphone speaker and captures the reflected wave from the chest wall. It extracts specific features such as the maximum velocity of chest wall motion, the chest wall displacement during the first second of exhalation, and the peak chest wall displacement. These features are then used as inputs to a regression neural network, which

provides an assessment of the user's lung function. In terms of ear diseases, Jin et al. [116] propose EarHealth, an earphone-based system that detects three ear diseases: otitis media, ruptured eardrums, and earwax blockages. By emitting sound waves into the ear and capturing the echoes using its integrated microphone, EarHealth analyzes the captured data that contains crucial information about the ear through a multi-view DL model to detect and monitor these ear diseases. Chan et al. [28] present the development and clinical evaluation of a low-cost otoacoustic emissions (OAE) probe designed to facilitate early hearing screening. Conventional OAE tests require highly sensitive and expensive acoustic hardware, making it inaccessible to low and middle-income countries. To fill this gap, the authors propose to develop a low-cost probe using off-the-shelf microphones and earphones connected to a smartphone. The probe functions by emitting two pure tones through the earphones, prompting the cochlea to generate distortion-product OAEs, which are then captured by a microphone. In [29], they further improve the design of the hearing screening probe using wireless earbuds, and propose OAEbuds, which employs a two-step protocol combining frequency-modulated continuous wave (FMCW) signals and wideband pulses to separate OAEs from in-ear reflections. The clinical study shows that OAEbuds achieves sensitivity and specificity comparable to commercial medical devices, demonstrating its potential to make hearing screening more affordable and accessible.

Assistive Technology. AIoT systems for healthcare and well-being has also been developed as assistive technologies, which help individuals with disabilities perform tasks that might otherwise be difficult or impossible. For instance, Zeng et al. [316] introduce MobileDeepPill, a mobile assistive technology that automatically identifies prescription pills in real-world settings using smartphone cameras. MobileDeepPill identifies pills by employing a multi-CNN model to extract a pill's three distinctive characteristics including color, shape, and imprints. It also adopts knowledge distillation to reduce the size of the multi-CNN model for on-device inference. Xiong et al. [282] propose iBlink, a smart glasses-based assistive technology for individuals with facial paralysis. Most individuals with facial paralysis are not able to blink on one side of the face, which could lead to blindness. iBlink aids individuals with facial paralysis to blink by detecting the non-paralyzed side's blinking using a camera and CNN and applying electrical stimulation to trigger blinking on the paralyzed side. As another line of research, DeepASL [56] and SignSpeaker [93] focus on developing sign language translation systems that bridge the communication gap between deaf people and people with normal hearing ability. Specifically, DeepASL uses infrared light-based sensing to capture and extract skeleton joint information of fingers, palms, and forearms when the user performs sign language. On the other hand, SignSpeaker derives sign-related information using motion sensors from a smartwatch. Both systems utilize the Connectionist Temporal Classification (CTC) technique to construct the sentence-level translation from the word-level translation.

Personal Health Insight Generation. The emergence of LLMs opens up a wide range of possibilities in the application domain of healthcare and well-being. One of the most promising capabilities is to generate personal health insights based on data collected from health-related sensors inside an individual's mobile and wearable devices. For example, Cosentino et al. [42] introduce PH-LLM, a Personal Health Large Language Model based on a fine-tuned version of Gemini designed to generate insights and recommendations for improving sleep and fitness behaviors. PH-LLM collects data from multiple sources, including medical records, wearable sensor data, and self-reported health data from each individual. By integrating this information, it seeks to understand each individual's unique health profile and to provide tailored health recommendations and predictions. Similarly, Merrill et al. [191] present PHIA, a Personal Health Insights Agent to analyze behavioral health data from wearable sensors using LLMs. PHIA can address both factual and open-ended health queries, and generate personalized, actionable health insights with high accuracy. Lastly, Englhardt et al. [54] explore the potential of using LLMs to derive clinically relevant insights from multi-sensor data collected from mobile and wearable devices. The authors develop chain-of-thought prompting methods to facilitate LLMs in reasoning about activity, sleep, and social interaction data, and their relation to mental health conditions such as depression and anxiety. While the

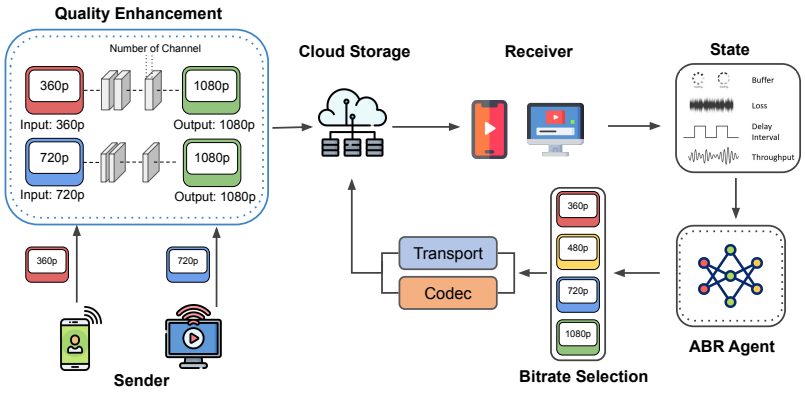

Fig. 17. Illustration of the architecture of AIoT systems for video streaming.

authors initially focused on using LLMs for diagnostic task, they found greater potential in generating detailed, natural language summaries that integrate multiple data streams, offering a more comprehensive understanding of a patient's health condition.

5.2 Video Streaming

Video streaming involves the continuous and seamless transmission of video and audio content from a server to a client over a network. It has become one of the most widely used technologies that enables services such as live streaming and video conferencing, which are integral to people's daily lives. As summarized in Figure 15, existing works on AIoT systems for video streaming can be grouped into three categories: adaptive video streaming, video enhancement, and efficiency optimization.

Adaptive Video Streaming. One key challenge of video streaming is to maintain a consistent high-quality viewing experience and uninterrupted playback when network bandwidth fluctuates due to factors such as congestion, interference, and user mobility. Adaptive video streaming addresses this challenge by dynamically adjusting the video quality in real-time based on the available network bandwidth, ensuring a smoother viewing experience. For example, Mao et al. [184] propose Pensieve, an adaptive video streaming framework that employs reinforcement learning (RL) to autonomously learn adaptive bitrate (ABR) algorithms to eliminate the need for pre-programmed control rules. Zhou et al. [351] present Concerto, which identifies an important factor of poor quality of experience (QoE): the lack of coordination between application-layer video codecs and the transport-layer protocols. To address this issue, Concerto introduces a video bitrate adaptation strategy based on deep imitation learning, which is abel to identify the most suitable bitrate for codec and transport layers and successfully boosts the QoE. While both Pensieve and Concerto exhibit their potential, a notable challenge arises from the fact that the learning models are commonly trained within simulators or emulators. Unfortunately, this can result in poor performance when applied in real-world scenarios. Zhang et al. [327] present OnRL, which effectively bridges the gap between simulation and real-world scenarios by introducing an online RL framework designed for real-time mobile video telephony applications. One challenge with RL is that the algorithm might make incorrect exploitation decisions. OnRL addresses this issue by introducing a hybrid learning approach: if the RL model performance deviates from the expected, the system switches to a rule-based ABR algorithm; otherwise, it continues to follow the RL strategy. Another challenge with RL-based approaches is that acquiring suitable training data and creating a suitable environment is not trivial. Yan et al. [299] address this challenge by developing an ABR algorithm and training it directly within the real deployment environment using in-situ data. As another line of research, Lee et al. [133] introduce PERCEIVE, which utilizes a 2-stage LSTM model for

cellular uplink channel throughput prediction and adapts the video encoding bitrate based on the prediction results to improve user experience in mobile live streaming applications. Zhang et al. [326] show that attempts to combine such hybrid approaches do not effectively utilize the combined strengths of both methods, often resulting in suboptimal performance. Therefore, they propose Loki, which strives for a more profound collaboration of rule-based methods with learning-based methods. This is achieved by converting a "white-box" rule-based approach into a similar "black-box" neural network model using a customized imitation learning model. Zhang et al. [340] introduce SENSEI, a streaming optimization scheme that capitalizes on users' varying sensitivity levels to different segments of a video. This approach is rooted in the understanding that users are more attuned to crucial moments (e.g., goal-scoring moments in a sports video) and are more displeased by buffering interruptions during these instances compared to less critical parts. Given that, SENSEI reduces the current video quality to conserve bandwidth, which can later be allocated to deliver higher quality during moments of heightened user sensitivity. This strategy enhances the QoE within the same bandwidth constraints by efficiently adapting video quality to users' sensitivity patterns. Lastly, Dasari et al. [46] introduce Swift, an adaptive video streaming system featuring a layered encoder. Instead of encoding video segments separately in various qualities, Swift encodes the video segments into layers. In doing so, it significantly reduces bandwidth usage and achieves a quicker response time to fluctuations in network conditions.

Video Enhancement. Another key challenge of video streaming is the inherent limitation in the resolution of the original source video, which can affect the viewing experience on high-definition displays. Video enhancement techniques address this challenge by enhancing the quality of videos by upgrading their resolution beyond the resolution of the original source video, thereby providing a better viewing experience for users with high-definition displays. Yeo et al. [307] introduce NAS, a super resolution-based video delivery framework that leverages client-side computation and DNNs to enhance user QoE. Their approach involves combining scalable DNNs with adaptive predictions that can adjust their processing requirements dynamically in response to the available resources. Reinforcement learning is used to determine the best time to download a DNN model and choose the appropriate video bitrate for each video segment. However, one key limitation of NAS [307] is its high computational demand and power consumption, making it less competitive to be deployed on mobile devices. To make video enhancement feasible for mobile devices, Yeo et al. [306] propose NEMO, which capitalizes on the inherent temporal redundancy in videos by applying super-resolution to only some specific frames while reusing the super-resolution results to enhance the entire video. However, due to the involvement of resource-intensive offline computation, NEMO is not ideal for live video streaming. In contrast, Kim et al. [121] design LiveNAS specifically for live video streaming scenarios. LiveNAS utilizes real-time online training and incorporates recently trained outcomes for super-resolution within the context of live video. Similarly, Yeo et al. [308] present NeuroScaler, a streamlined and scalable neural-enhancing framework for live video streaming. NeuroScaler focuses on reducing the costs of live video streaming and includes cost-reducing algorithms for video super-resolution and a specialized hybrid video codec that drastically cuts encoding expenses for selective super-resolution outputs. Zhang et al. [321] move one step further and propose YuZu, a super-resolution-based video streaming system for 3D video streaming. This system addresses key limitations of existing 3D video streaming methods, such as high bandwidth consumption and the ineffectiveness of viewport-based streaming when the entire scene is within the view. Lastly, Park et al. [211] explore omnidirectional video (i.e., 360° video) streaming, and develop OmniLive, a super-resolution-based omnidirectional video streaming system that utilizes GPU to sustain a high super-resolution quality at 30 frames per second across a range of mobile devices.

Efficiency Optimization. Enhancing the efficiency of video streaming services is also important. Tang et al. [247] present CHESS, a video popularity prediction scheme designed to forecast the future popularity of videos. Since only a small fraction of videos gain significant popularity and contribute to the majority of watch time, by prioritizing these popular videos rather than processing all videos uniformly, CHESS effectively allocates

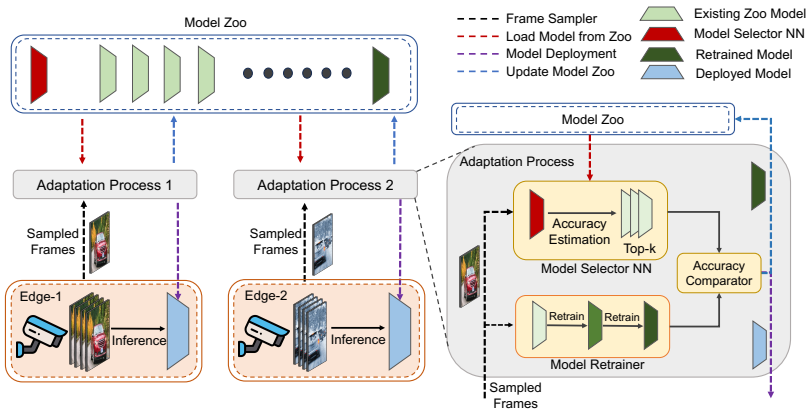

Fig. 18. Illustration of continuous learning for video analytics.

processing resources to optimize the user experience. Omnidirectional video streaming consumes more bandwidth compared to standard video streaming. One potential solution for bandwidth optimization is the viewport-driven approach, which focuses on streaming only the region that the viewer is watching (viewport) in high quality. However, this approach comes with constraints as it requires predicting the viewer's future gaze direction, and any prediction errors can lead to rebuffering or drops in quality. To address this challenge, Guan et al. [76] develop Pano, a method that leverages the sensitivity of users to variations in quality distortion, which effectively balances the trade-off between quality and bandwidth allocation. This approach allows for increasing quality to the highest noticeable extent when there is surplus bandwidth and decreasing quality to an almost unnoticeable degree when bandwidth is limited. Improving QoE for mobile omnidirectional video streaming is crucial, particularly in bandwidth-limited wireless networks. Previous research on omnidirectional video streaming has attempted optimization based on head movement trajectory (HMT) but often falls short in achieving precise HMT predictions. To address this challenge, Wang et al. [262] introduce SalientVR, a framework that integrates gaze information into a saliency-driven mobile 360° video streaming system. SalientVR holds the potential to elevate the QoE by utilizing user gaze patterns to deliver content more accurately and effectively for mobile VR devices. Lastly, Rudow et al. [221] introduce Tambur, a scheme designed to address bandwidth-efficient loss recovery for video conferencing. Existing streaming codes fall short of meeting the specific demands of video conferencing due to the frequent loss of packets, often occurring in bursts, which can impede the rendering of video frames. Tambur introduce a learning-based predictive model for effectively configuring bandwidth overhead and achieves a noteworthy reduction in both the frequency and cumulative duration of freezes.

5.3 Video Analytics

Video cameras have been deployed at scale at places such as streets and intersections, stores and shopping malls, as well as homes and office buildings. Analyzing video streams collected from these distributed cameras enables many applications such as security and surveillance, traffic management, and customer behavior analysis. As summarized in Figure 15, existing works on AIoT systems for video analytics can be grouped into four categories: continuous learning, runtime adaptation, efficiency optimization, and query optimization.

Continuous Learning. In video analytics, it is inevitable that new video data emerge. Therefore, it is critical for video analytics systems to adapt to such data drift. Although continuous learning can effectively tackle data drift by periodically retraining models on new data, supporting continuous learning on video analytics systems is not trivial. Bhardwaj et al. [17] propose Ekya, a video analytics system that addresses the challenge of jointly

supporting inference and continuous learning on edge servers. The key idea of Ekya is to identify models that need retraining the most while balancing the resources for joint retraining and inference. Ekya can enhance the performance of video analytics, particularly in dynamically changing environments where data drift is a significant factor in performance. However, since the retraining process consumes the majority of the time, relying solely on model retraining may not be resource-efficient for real-time video analytics tasks. Khani et al. [120] propose RECL, an end-to-end system that integrates model reusing with model retraining to overcome this problem. RECL performs continuous model retraining as well as leverages historical specialized DNNs and shares this model zoo across various edge devices. Additionally, RECL efficiently allocates GPU resources by utilizing an iterative training scheduler, which prioritizes retraining jobs based on their progression rate. RECL shows remarkable improvement in both accuracy and mAP for object detection and image classification tasks, outperforming all baseline models, including Ekya.

Runtime Adaptation. Another critical capability of video analytics systems is runtime adaptation. Adapting camera parameters and settings has a significant impact on video analytics performance, particularly due to weather and lighting conditions. To maintain high accuracy, it becomes essential to adapt camera parameters in response to these conditions. However, the task of identifying the optimal camera settings for specific scenes is challenging. To mitigate the impact of environmental condition changes on video analytics performance, Paul et al. [212] propose Camtuner, a reinforcement learning-based approach to dynamically adapt non-automated camera parameters. Apart from camera parameters, various other factors within a video analytics pipeline can impact its performance, including frame resolution, frame sampling rate, and the choice of DNN models. Collectively, these components can be referred to as the overall configuration. Choosing a suitable configuration can impact both the resource utilization and accuracy of a video analytics application. Although adapting model configurations frequently can optimize resource usage, it incurs high costs due to the large number of possible configurations. To address this issue, Jiang et al. [110] introduce Chameleon, a technique for achieving a balance between resource allocation and accuracy by choosing the appropriate neural network configuration. Zhang et al. [322] propose AWStream, an adaptive stream processing system with low latency and high accuracy. AWStream's main contribution is its runtime system that consistently monitors and adjusts to network conditions. It optimizes streaming data rate based on available bandwidth and employs learned Pareto-optimal configurations to maintain high accuracy. Zeng et al. [317] propose Distream, which focuses on runtime adaptation to the dynamic workloads generated by distributed video cameras. Depending on the deployment location, the number of objects captured by each camera and its corresponding workload is different and varies throughout the day. The key idea of Distream is to adaptively balance the workloads across the cameras and also partition the workloads between cameras and the edge server. As such, Distream fully utilizes the compute resources at cameras and the edge server to enhance system performance. Lastly, Lu et al. [179] propose Turbo, which capitalizes on managing latent computing resources to enhance overall performance, particularly in object detection tasks. The proposed approach revolves around a multi-exit GAN structure, which is paired with an adaptive scheduler that dynamically determines the optimal enhancement level for each incoming frame, thereby maximizing object detection accuracy in real-time. The adaptive scheduler makes on-the-fly decisions about the most appropriate enhancement levels based on the current resource availability. In terms of results, Turbo presents remarkable improvements in absolute mAP.

Efficiency Optimization. Enhancing the efficiency of video analytics systems is also important. The widespread deployment of video cameras, numbering in the thousands and operating continuously, leads to a massive amount of data that needs to be transmitted and processed. Transmitting and processing all video frames from the edge to the server can be extremely expensive due to the bandwidth constraints and computational resources required. Canel et al. [23] propose FilterForward, which only selects and transmits the relevant video frames to save bandwidth. Similarly, Li et al. [151] introduce Reducto, another filtering-based technique which implements on-camera filtering and dynamically adjusts filtering decisions to cater to live video analytics requirements.

Experimental results show that Reducto outperforms FilterForward by 93% in terms of frame filtering efficiency. As another line of research, Du et al. [53] introduce AccMPEG. The proposed key techniques involve the design of a cheap camera-side model to efficiently decide which regions of the frames should be encoded high-quality and which regions should be subjected to lower-quality encoding. Additionally, AccMPEG allows for quick customization to different DNNs, with training times reduced to mere minutes, further demonstrating its efficiency. Hwang et al. [103] introduce CoVA, a cascade architecture that reduces the need for full video decoding. By leveraging compressed-domain analysis, CoVA efficiently detects and tracks objects across frames, only decoding a minimal subset of frames necessary for DNN processing. CoVA's design not only optimizes computational efficiency but also supports both temporal and spatial queries, broadening its applicability in video analytics. Lastly, video analytics systems often host multiple tasks like object detection, face recognition, and semantic segmentation, where different models can be used for different tasks. Given the limited GPU resources of edge devices, attempting to load all models can exceed GPU memory limit. Padmanabhan et al. [207] introduce Gemel, a model merging technique that can efficiently merge and share layers from models with the same architectures. In doing so, Gemel effectively reduces the number of swaps required and the amount of data loaded into GPU memory, resulting in fewer frame drops and improved accuracy.

Query Optimization. A video analytics query is an inquiry submitted to a video analytics system to retrieve useful information and insights from video data. Summarizing a video scene with object count is a common query type to get insight from a video stream. Due to the energy constraint of edge devices, continuously transmitting video streams is a challenging task. ELF, presented by Xu et al. [292], is a framework designed to continually summarize video scenes through the aggregation of object counts, all while operating within the confines of limited energy resources. Rather than transmitting raw video data, the approach involves sending only numerical data, such as count numbers or other relevant query-related values. In many camera setups, a significant portion of cameras often remain inactive and unqueried. This scenario can be referred to as "zero-streaming" where the inactive cameras store video data in their local storage and communicate with the server only when a specific query is requested. Xu et al. [291] propose DIVA, an approach to effectively query video analytics on zero streaming cameras. When it comes to tasks like retrieval, tagging, and counting, DIVA consistently demonstrates superior performance compared to other baseline methods. As zero streaming cameras primarily store data on their local storage, a drawback of DIVA is its susceptibility to video data loss in the event of camera storage failures. Video analytics queries can sometimes raise concerns about privacy violations, as users may request sensitive information about others, potentially infringing on their privacy. Cangialosi et al. [24] introduce Privid, a method aimed at extracting valuable information from video data without compromising privacy. Privid's approach involves breaking the video into smaller segments and executing processing code on each segment individually rather than processing the entire video at once. A Privid query comprises a set of statements in a PrividQL language, similar to SQL, along with executable video processing components. They run an experiment on video data collected from three cameras and apply Detectronv2 for object detection and DeepSORT for object tracking. The rise of 5G technology has propelled the expansion of ultra-fast video analytics, largely due to the growing need for low-latency processing capabilities. Tutti, developed by Xu et al. [286], combines the 5G radio access network and edge computing at the user level to ensure optimal performance for video analytics tasks with low latency. Tutti achieves a remarkable reduction in response latency and demonstrates substantial progress in enhancing QoE for video analytics applications. In response to diverse applications, video analytics platforms have gradually moved away from providing pre-defined video processing results. Instead, they now enable users to utilize their customized models, all while ensuring a consistent commitment to specified accuracy standards. Recent optimization efforts involve preprocessing video data in advance to construct indices that can expedite subsequent queries. However, these optimizations were tailored for scenarios where models were predefined and

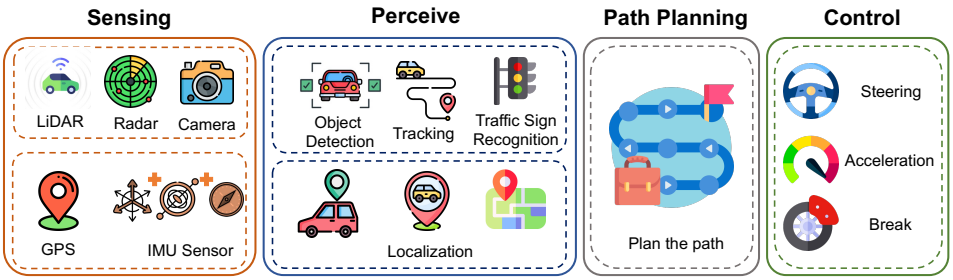

Fig. 19. Illustration of the architecture of AIoT systems for autonomous driving.

not user-provided. Agarwal and Netravali [3] introduce Boggart, a comprehensive pipeline for a video analytics platform that can function as a versatile accelerator using the model provided by the users.

5.4 Autonomous Driving

Autonomous driving enables a vehicle to navigate and operate partially with or fully without human intervention. By fusing real-time data collected through IoT sensors with AI-driven perception and decision-making algorithms, AIoT systems are contributing to making autonomous vehicles safer, more efficient, and adaptable to changing road conditions. As summarized in Figure 15, existing works on AIoT systems for autonomous driving can be grouped into four categories: perception enhancement, localization, tracking, and mapping, automatic testing, and control and actuation.

Perception Enhancement. Perception enhancement involves the use of AI to process data from various sensors such as cameras, LiDAR, and radar more accurately, allowing an autonomous vehicle to have a more comprehensive understanding of its surroundings. While LiDAR-based systems offer detailed spatial mapping, they fail in adverse weather conditions because LiDAR beams struggle to penetrate through elements like fog. To address this limitation, Bansal et al. [13] introduce Pointillism, an innovative concept called cross-potential point clouds, which leverages the spatial diversity generated by utilizing multiple radar systems to effectively address the issues related to noise and sparsity in radar-based point clouds. Single-vehicle 3D sensors have two primary limitations: susceptibility to occlusion by non-transparent objects and reduced detail perception at greater distances. Zhang et al. [341] introduce EMP, a collaborative approach where all nearby connected autonomous vehicles (CAVs) share sensor data with each other. This sharing allows each vehicle to create a more comprehensive and higher-resolution perception compared to relying solely on its own sensors. The emerging paradigm of infrastructure-assisted autonomous driving leverages infrastructure elements like smart lampposts to assist autonomous vehicles. However, a challenge arises when the vehicle and infrastructure point clouds do not hold a significant overlap or similarity, resulting in a drop in accuracy and delays. Vi-Eye, presented by He et al. [91] is a pioneering system capable of aligning vehicle-infrastructure point clouds with centimeter-level accuracy in real time. VIPS, developed by Shi et al. [231], takes this capability a step further, achieving decimeter-level accuracy while still maintaining real-time performance. VIPS distinguishes itself from Vi-Eye by adopting an alternative strategy. While Vi-Eye relies on highly accurate point cloud transmission between infrastructure and vehicles, VIPS focuses on aligning two graphs generated from simplified and diverse representations of objects detected by the vehicle and infrastructure. He et al. [90] present Automatch, an innovative solution by utilizing traffic cameras to enhance the perception and localization capabilities of autonomous vehicles, particularly at intersections. The pioneering aspect of the system is that it enables vehicles to expand their range of perception by correlating images taken by both traffic cameras and on-vehicle cameras.

Localization, Tracking, and Mapping. Localization is the process of determining the precise position of a vehicle within a known environment by comparing sensor data to pre-existing maps or reference points. Accurate localization in environments like tunnels and underpasses, where Global Navigation Satellite Systems (GNSS) signals are unavailable, can be a challenging task in autonomous driving. MVP [255] address this challenge by extracting magnetic fingerprints from anomalies in the geomagnetic field. These magnetic fingerprints are then compared to a magnetic map, allowing for precise positioning of vehicles without relying on GNSS signals. In the context of autonomous driving, tracking refers to the continuous monitoring and prediction of the movements and trajectories of vehicles on a roadway. VeTrac [249] employs traffic cameras as a sensing network to reconstruct large-scale vehicle trajectories, addressing the limitations of GPS-dependent solutions. It achieves this through a vision-based vehicle detection and tracking algorithm applied to video frames collected from the traffic cameras. Lin et al. [155] identifies three primary computational bottlenecks in autonomous driving systems: object detection, object tracking, and localization, which collectively consume over 94% of computational resources in the system. In response to these challenges, the authors have developed an end-to-end autonomous driving system that draws from the most cutting-edge system designs found in both academic research and industry practices. Mapping is a continuous process that involves creating and continually updating a detailed map of the surroundings of a vehicle through the use of various sensors such as LiDAR, cameras, and radar. Maps used in autonomous driving systems require continuous updates to account for significant changes in the environment, which can affect the features visible to a vehicle. CarMap, developed by Ahmad et al. [4] offer an innovative solution by collecting 3D maps from vehicles equipped with LiDAR and advanced cameras, ensuring near real-time map updates. As each vehicle travels through a road segment, it uploads map updates to a cloud service over a cellular network, making these updates accessible to other vehicles.

Automatic Testing. Automatic testing involves identifying and analyzing various events or scenarios that autonomous vehicles may encounter on the road and testing the vehicle's AI-driven systems to ensure they respond appropriately to these events. BigRoad [165] provides a cost-effective and dependable solution for collecting extensive driving data by utilizing a smartphone and an Inertial Measurement Unit (IMU) installed within the vehicle. This system extracts internal driver inputs, such as steering wheel angles, driving speed, and acceleration, and also discerns external perceptions of road conditions, including the distinction between wet and dry surfaces. This information can be highly valuable for various purposes, including autonomous vehicle testing and evaluation. Automatic testing of autonomous driving technology is a complicated process due to the necessity of addressing unusual events and corner cases like road obstacles, pedestrians on highways, or wildlife encounters. To address this challenge, Li et al. [147] introduce an automatic system that utilizes an algorithm to identify and respond to unusual driving events effectively. The results of detecting unusual events can be valuable for retraining and enhancing a self-steering algorithm, particularly in more complex driving scenarios.

Control and Actuation. Autonomous control systems manage components that interact with their environments while making decisions independently, without human intervention. Prior works in autonomous AIoT control systems involve multiple stages, including data acquisition from sensors, processing with deep neural networks, and control of configuration parameters to interact with the external environment. The multiple stages suffer from performance bottlenecks due to the difficulty in tuning each step. For instance, even lightweight deep neural networks for object detection have millions of parameters and are too complex for embedded platforms. This complexity makes it infeasible to run multi-stage AIoT control algorithms in real-time on platforms with memory and computation constraints. Sandha et al. [224] present EAGLE, an end-to-end deep reinforcement learning (RL) solution that trains a neural network policy to directly use images as input for controlling the PTZ camera. The proposed system bypasses the conventional multi-stage process of object identification, tracking, and control by directly mapping raw photos to control actions using a neural network policy. The paper demonstrates Eagle's

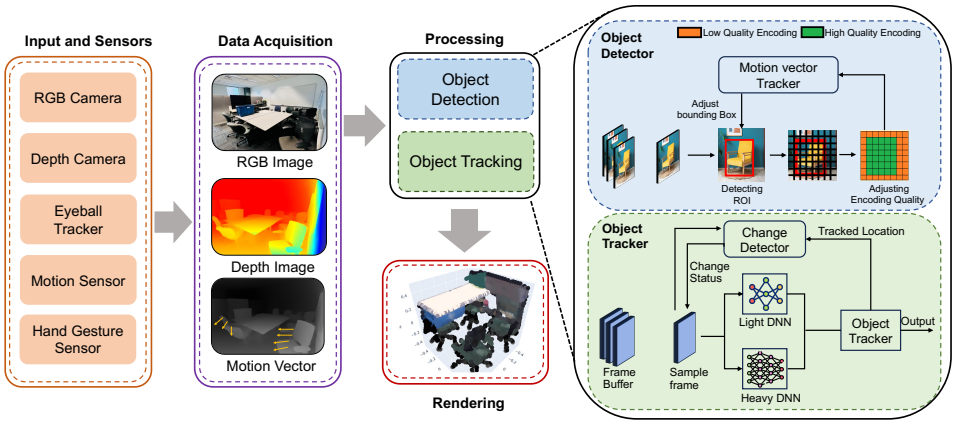

Fig. 20. Illustration of the architecture of AIoT systems for AR/VR/MR.

effectiveness in various scenarios and its successful transfer from simulation to real-world applications, making significant contributions to the fields of edge AI and autonomous camera control.

5.5 Augmented, Virtual, and Mixed Reality

Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR) redefine our perception of the world. Specifically, AR enriches reality by overlaying digital content on our surroundings; VR immerses us in entirely digital environments; and MR provides an interactive experience between the virtual and real worlds. As summarized in Figure 15, existing works on AIoT systems for AR/VR/MR can be grouped into four categories: object detection and tracking, user inputs, performance enhancement, and omnidirectional AR.

Object Detection and Tracking. Object detection and tracking is one of the most fundamental tasks in AR/MR. Liu et al. [164] present an efficient offloading-based object detection and tracking system for AR/MR, which offloads the object detection task to the cloud while conducting tracking on AR devices. The key technique incorporated in the system is a dynamic region of interest (RoI) encoding technique that encodes regions where objects are not likely to be detected in lower quality. As such, the proposed system reduces both offloading latency and bandwidth consumption while maintaining object detection accuracy. Apicharttrisorn et al. [9] propose MARLIN, a lightweight object detection and tracking framework for AR. Instead of running computationally expensive DNN on each frame, it initiates the DNN execution on the initial frame and then assesses if there are significant scene changes using a change detector specifically designed to identify alterations in the background. If such frame changes are not detected, MARLIN opts for a more lightweight tracking scheme, conserving computational resources while maintaining tracking accuracy. Guan et al. [75] move one step further and propose DeepMix that focuses on 3D object detection for AR/MR, aiming to provide an efficient solution in this computationally demanding domain. Instead of relying on computationally intensive DNN-based 3D object detection models for bounding box inference, DeepMix offloads 2D RGB images to the edge for 2D object detection and then utilizes the returned 2D bounding boxes in conjunction with depth data captured by headsets to estimate 3D bounding boxes. DeepMix was prototyped on a Microsoft HoloLens 2. Evaluation results show that compared to existing methods based on 3D object detection, DeepMix not only enhances detection accuracy but also considerably decreases end-to-end latency.

User Inputs. Capturing user inputs in an accurate, intuitive and user-friendly manner is another important task in AR/VR/MR. Existing systems face challenges in capturing user-friendly inputs, particularly in detecting

subtle and low-effort finger gestures, which are more suitable for head-mounted devices (HMD) controllers. Nguyen et al. [201] introduce HandSense, a system using capacitively coupled electrodes to precisely capture and recognize micro-finger gestures for interaction with HMD. They develop an electrode placement configuration on fingertips that minimizes the need for extensive hand movements and utilize several DNN-based methods to recognize the gestures. Experimental results show that HandSense is able to achieve a 97% accuracy in recognizing 14 gestures performed by 10 subjects. As another line of research, in interactive VR applications, conventional techniques have limitations as they cannot capture the upper face of users, which is mostly occluded by the head-mounted display. To address this limitation, Chen et al. [33] propose ExGSense, which detects and recognizes eye and mouth gestures as VR inputs. This capability is made possible through the utilization of sparse near-eye biopotential signal measurements combined with a DNN-based classifier. They evaluated their prototype with 42 facial gestures, achieving 93% accuracy for user-specific and 77% for user-independent evaluation.

Performance Enhancement. Performance enhancement involves optimizing software and hardware to reduce latency, increase processing speed, and improve resource management. Trinelli et al. [251] present NEAR, a transparent AR processing system designed to reduce latency and enhance performance when integrating AR features into streaming videos from lightweight IoT devices. NEAR introduces a simplified SOCKS 5 proxy, a video decoder, and an encoder for the extraction and re-injection of video streams into network flows. This setup enables offloading heavy computational tasks, like object detection, to edge devices, reducing the processing load on both source and consumer devices. NEAR operates without requiring modifications to the IoT streaming devices or client-side applications, ensuring a seamless integration of AR and other computationally intensive functions directly within the network. Mobile DL frameworks often encounter limitations, particularly related to multi-DNN GPU contention, which can significantly increase inference latency. Unlike desktop GPUs, mobile GPUs cannot effectively implement multi-tasking approaches due to their constraints. Heimdall, introduced by [310], can efficiently manage the demands of multiple DNN rendering tasks on mobile devices, ensuring minimal latency and optimal performance in emerging AR applications. Heimdall introduces an innovative GPU coordinator that effectively handles multiple DNN rendering tasks on both GPU and CPU by decomposing complex DNNs into smaller units and adopting flexible scheduling techniques. The approach significantly reduces the contention between DNNs and rendering tasks, which typically degrades performance on mobile devices, thus enhancing overall system performance. Heimdall was prototyped on various mobile GPUs and AR applications, showing it boosts frame rates from 12 to 30 fps and reduces worst-case DNN inference latency by up to 15 times compared to the baseline multi-threading approach. Liu et al. [173] introduce CollabAR, the concept of collaborative image recognition into its design, capitalizing on users' temporally and spatially correlated images to enhance image recognition accuracy. The edge-assisted design of the system significantly reduces end-to-end latency, ensuring seamless and efficient performance on commodity mobile devices. CollabAR attains a recognition accuracy rate exceeding 96% even for images with substantial distortions. FreeAR, presented by [8], enhances the performance of mobile AR by introducing infrastructure-free AR experiences through collaborative time slicing and efficiently distributing compute-intensive tasks across multiple user devices. In FreeAR, all devices unite under a common coordinate system. The chosen primary device takes charge by executing DNNs, enabling it to update the device pose, physical object locations, and 3D virtual overlays, much like traditional AR systems. Meanwhile, secondary devices shift into a low-power mode, where they track their locations within the converged coordinate system using lightweight methods. With this approach, FreeAR can establish a low-power framework, enabling users to seamlessly engage in AR experiences without relying on infrastructure support.

Omnidirectional AR. Lastly, omnidirectional AR refers to AR experiences that provide a 360-degree view of the environment, allowing users to interact with and view augmented content from any direction. In mobile AR applications, achieving accurate omnidirectional lighting is crucial to avoid undesirable visual effects. However, accurately estimating omnidirectional lighting in practical scenarios can be challenging, primarily due to the

influence of environmental lighting conditions and the dynamic nature of mobile users. Zhao and Guo [348] introduce Xihe, a mobile AR application capable of real-time and precise omnidirectional lighting estimation by employing a sphere-based point cloud sampling technique. Combined with 3D vision-based lighting estimation pipeline, this sampling technique delivers significantly improved results over farthest point sampling techniques.

6 DISCUSSIONS

Lastly, in the field of AIoT, addressing issues such as bias and fairness, security and privacy, as well as legal and ethical concerns is as crucial as tackling the technical challenges in sensing, computing, and networking & communication, and domain-specific AIoT systems as we have covered in previous sections. In this section, we provide a brief discussion on these issues.

Bias and Fairness in AIoT. The integration of IoT and AI significantly broadens the functional capabilities of AIoT. As a result, it raises the need for fairness considerations of AIoT since its extended capabilities allow it to be widely deployed in daily life. Bias is referred to the systematic deviation in data or algorithms used by AIoT that leads to unfair or discriminatory outcomes. Balasingam et al. [12] address the challenge of balancing throughput and fairness in mobility platforms that allocate tasks to vehicles for applications such as food delivery and ridesharing. They show that current ridesharing platforms often fail to ensure that riders from different neighborhoods receive equitable service. This issue arises when the algorithm prioritizes ride requests from certain neighborhoods over others, typically favoring areas with higher demand or shorter and more profitable trips. Given that, they introduce Mobius, a system engineered to balance high throughput and fairness among customers by effectively managing the inherent trade-offs in shared mobility, which enhances the overall performance and fairness of mobility platforms. Bias in AIoT may arise through federated learning (FL), where models are trained across multiple edge devices, influenced by the heterogeneous resources and data distributions of these devices. Selialia et al. [226] observe that sample feature heterogeneity, resulting from different feature representations at various devices, is a major contributor to bias in FL. Their results show that existing bias mitigation techniques, such as normalization do not fully eliminate bias, with bias levels being proportional to the degree of heterogeneity in sensor sampling features. Lastly, Bae and Xu [11] focus on biases in pedestrian trajectory prediction models used in autonomous vehicles. They highlight that many DL models trained on pedestrian data are biased, particularly against vulnerable demographics like children and the elderly, who exhibit different walking patterns compared to the general adult population. This bias can lead to higher prediction errors for these groups and increasing their risk of involvement in vehicle crashes.

Security in AIoT. The vulnerabilities inherent in AIoT pose critical security concerns. One of the root causes is the limited resources of AIoT devices, which makes it challenging to implement robust security measures. For example, to make an effective balance between security needs with resource limitations, Luo et al. [181] propose ShieldScatter, a lightweight solution to enhance IoT security by utilizing battery-free backscatter tags. These tags create fine-grained multi-path propagation signatures, allowing for the identification of legitimate users and the detection of attackers. ShieldScatter provides a cost-effective method that does not require expensive hardware modifications, offering a practical security solution for resource-constrained IoT devices. As another example, in contact-free smart sensing devices, limited storage resources necessitate the use of cloud storage. However, data stored in the cloud is particularly vulnerable due to the open nature of cloud environments, making it susceptible to potential third-party attacks. To mitigate these risks, Mei et al. [190] introduce a novel Cloud-Edge-End cooperative storage scheme that leverages the distinct characteristics of the cloud, edge, and endpoint layers. This scheme employs a strategically designed data partitioning strategy, which involves storing sensory data across the three layers separately. By doing so, it increases the difficulty of potential security breaches while offering robust protection against both internal and external attacks. To protect from malicious attacks in IoT environment, several DL-based detection mechanisms are proposed [59, 119, 142]. Khan et al. [119] investigate

the robustness of SplitFed Learning – a hybrid of split learning and federated learning (FL) – against model poisoning attacks, where attackers deliberately inject fake data into the network. SplitFed combines the parallel computation efficiency of FL with the resource efficiency and improved privacy of split learning. The study shows that SplitFed, due to its smaller client-side model portions, is inherently more robust to model poisoning attacks compared to FL. Li et al. [142] focus on physical adversarial attacks on DL-based Wi-Fi sensing systems. This attack manipulates Wi-Fi packet preambles to subtly alter the Channel State Information, thereby influencing the DL models that rely on this data, without interrupting normal communication. Demonstrating high success rates of attack in activity recognition and user authentication, this study exposes significant security vulnerabilities in current Wi-Fi sensing systems. Lastly, Dong et al. [51] explore a critical security vulnerability in modern mobile devices that utilize dynamic refresh rate switching to optimize power consumption. The authors present an innovative attack vector named RefreshChannels, where two colluding apps modulate the display's refresh rate to covertly transmit sensitive information, bypassing the operating system's sandboxing and isolation measures. They also propose countermeasures to mitigate the RefreshChannels attack such as restricting refresh rate API access, limiting refresh rate change frequency, introducing delays and randomization, and detecting abnormal refresh rate patterns.

Privacy in AIoT. Since AIoT could gather a diverse array of data such as an individual's location, personal healthcare record, behavior patterns, and biometric information that is rich in personal information, the collection and processing of such personal data can raise significant privacy concerns. To protect the privacy of individuals, various regulations have been implemented. The European Union (EU)'s General Data Protection Regulation (GDPR) offers comprehensive data protection rules for handling EU citizens' personal data [55]. In the U.S., the California Consumer Privacy Act (CCPA) outlines consumer rights regarding personal information collected by businesses, while the Health Insurance Portability and Accountability Act (HIPAA) stringently controls the handling of personal healthcare data [202, 203]. Alongside these legal frameworks, numerous research efforts are underway to tackle privacy-related challenges. Abadi et al. [1] introduce Differential Privacy (DP), a technique that injects noise into data to preserve sensitive personal information. They introduce DP into DL model training with their proposed DP-SGD method, which has proven to maintain high accuracy while effectively preserving privacy. Fully Homomorphic Encryption (FHE) is another privacy-preserving mechanism which enables computation to be performed over encrypted data. FHE ensures that original data remains hidden and is not decrypted during processing. However, due to its significant computational demands, AIoT is exploring alternatives like Partially Homomorphic Encryption (PHE) and Somewhat Homomorphic Encryption (SHE) to reduce computational overhead. Shafagh et al. [227] propose Pilatus, a PHE scheme for IoT while sharing the data with the cloud. Pilatus protects data privacy by ensuring that the cloud stores only encrypted data while still enabling operations like summation. Mo et al. [195] introduce PPFL, a framework that leverages Trusted Execution Environments (TEEs) to prevent private information leakage in federated learning scenarios. Though federated learning enables decentralized training across multiple devices without aggregating user data, model updates can still leak sensitive information, posing significant privacy risks. To address this, PPFL employs TEEs to securely process model updates, ensuring that both local training on clients and secure aggregation on servers are protected from potential adversaries. Singh et al. [237] introduce SnoopDog, a framework designed to address the privacy issues arising from hidden wireless sensors, such as secret cameras and microphones. SnoopDog identifies Wi-Fi-based sensors monitoring users by detecting causal patterns between trusted sensor data like IMU readings and Wi-Fi traffic. Although the current implementation of SnoopDog is limited to Wi-Fi-connected devices, future enhancements could extend its capabilities to other wireless communication standards like Zigbee or Bluetooth. Conventional privacy-preserving machine learning (PPML) methods often face significant latency issues due to computation overhead of encryption processes. To address this issue, Chien et al. [39] introduce Enc^2, a hybrid method that combines encoding and homomorphic encryption to enhance PPML for resource-constrained IoT

devices. The proposed method performs most of the computations on plaintext, thus reducing latency and shifting the encoding burden from the IoT device to the cloud. Lastly, Corbett et al. [41] introduce BystandAR, which addresses the privacy concerns posed by Augmented Reality (AR) devices that unintentionally capture the visual data of bystanders. BystandAR leverages eye gaze and voice indicators to differentiate between subjects and bystanders, protecting the bystander's privacy in real-time without offloading data to external servers.

Legal and Ethical Concerns in AIoT. Finally, AIoT must adhere to ethical norms and legal obligations. Mittelstadt [194] discusses the intersection of ethical issues and the deployment of health-related IoT technologies, emphasizing the importance of designing these technologies in ways that are both ethically responsible and legally compliant. It also underscores the need for responsible design and deployment of IoT technologies, ensuring they are trustworthy, respect user rights, and enhance healthcare delivery without compromising ethical standards. Gill [72] highlights the importance of addressing ethical dilemmas in the adoption of autonomous vehicles. The study focuses on the ethical dilemma of programming autonomous vehicles to make decisions in situations where harm is unavoidable, such as whether to protect passengers or pedestrians. Despite industry and policymakers' tendencies to downplay these ethical issues, the findings underscore the necessity of addressing these dilemmas to ensure the successful deployment and acceptance of autonomous vehicles. Bouderhem [20] proposes a comprehensive ethical framework to govern the use of AI in healthcare. This framework is based on values such as human dignity, fairness, transparency, accountability, and inclusivity. Bouderhem [20] also discusses the role of the European Union's General Data Protection Regulation (GDPR) and the AI act as models for creating robust regulatory frameworks.

7 CONCLUDING REMARKS

In this survey, we present a comprehensive review of AIoT research. We organize the AIoT literature into a taxonomy that includes four categories: sensing, computing, networking & communication, and domain-specific AIoT systems, and review key topics within each category. We hope our survey serves as a foundational reference, enabling researchers and practitioners to gain a comprehensive understanding of AIoT and inspiring further contributions to this exciting and important field.

8 ACKNOWLEDGEMENT

We would like to thank the editorial board and the anonymous reviewers of ACM Transactions on Sensor Networks (TOSN) for their helpful and constructive comments. We would also like to thank Christopher Ellis, Vishnu Chhabra, Imran Kibria, and Anwesha Roy for their help. Ness Shroff has been supported in part by National Science Foundation (NSF) grants NSF AI Institute (AI-EDGE) CNS-2112471, CNS-2312836, CNS-2106933, CNS-2106932, CNS-2312836, CNS-1955535, and CNS-1901057, by Army Research Office under Grant W911NF-21-1-0244 and was sponsored by the Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-23-2-0225. Bhaskar Krishnamachari is supported in part by Defense Advanced Research Projects Agency (DARPA) under Contract Number HR001120C0160 and in part by ARL under Cooperative Agreement W911NF-17-2-0196. Mani Srivastava is supported in part by Air Force Office of Scientific Research (AFOSR) under Cooperative Agreement FA95502210193, DEVCOM ARL under Cooperative Agreement W911NF-17-2-0196, and National Institutes of Health (NIH) mDOT Center under Award 1P41EB028242. Zhichao Cao and Mi Zhang are supported in part by NSF under award NeTS-2312675. Icons inside Figure 4, 6, 7, 8, 9, 10, 12, 13, 16, 17, 18, 19, 20 are made by Freepik from https://www.flaticon.com. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the ARL, the Department of Defense or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*.

[2] Daniel A Adler, Dror Ben-Zeev, Vincent WS Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. 2020. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR mHealth and uHealth* 8, 8 (2020).

[3] Neil Agarwal and Ravi Netravali. 2023. Boggart: Towards General-Purpose Acceleration of Retrospective Video Analytics. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA.

[4] Fawad Ahmad, Hang Qiu, Ray Eells, Fan Bai, and Ramesh Govindan. 2020. CarMap: Fast 3d feature map updates for automobiles. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*.

[5] Ali Akbari and Roozbeh Jafari. 2019. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*.

[6] Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. 2022. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *Advances in Neural Information Processing Systems* 35 (2022).

[7] Samiul Alam, Tuo Zhang, Tiantian Feng, Hui Shen, Zhichao Cao, Dong Zhao, JeongGil Ko, Kiran Somasundaram, Shrikanth S Narayanan, Salman Avestimehr, et al. 2023. FedAIoT: A Federated Learning Benchmark for Artificial Intelligence of Things. *arXiv preprint arXiv:2310.00109* (2023).

[8] Kittipat Apicharttrisorn, Jiasi Chen, Vyas Sekar, Anthony Rowe, and Srikanth Krishnamurthy. 2022. Breaking Edge Shackles: Infrastructure-Free Collaborative Mobile Augmented Reality. In *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*. Boston Massachusetts.

[9] Kittipat Apicharttrisorn, Xukan Ran, Jiasi Chen, Srikanth Krishnamurthy, and Amit Roy-Chowdhury. 2019. Frugal following: power thrifty object detection and tracking for mobile augmented reality. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. New York New York.

[10] Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Sethi, Deepak Vasisht, and Dinesh Bharadia. 2020. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[11] Andrew Bae and Susu Xu. 2023. Discovering and Understanding Algorithmic Biases in Autonomous Pedestrian Trajectory Predictions *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA.

[12] Arjun Balasingam, Karthik Gopalakrishnan, Radhika Mittal, Venkat Arun, Ahmed Saeed, Mohammad Alizadeh, Hamsa Balakrishnan, and Hari Balakrishnan. 2021. Throughput-fairness tradeoffs in mobility platforms *(MobiSys '21)*. Association for Computing Machinery, New York, NY, USA.

[13] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia. 2020. Pointillism: accurate 3D bounding box estimation with multi-radars. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. Virtual Event Japan.

[14] Soroush Bateni and Cong Liu. 2020. {NeuOS}: A {Latency-Predictable}{Multi-Dimensional} Optimization Framework for {DNN-driven} Autonomous Systems. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*.

[15] Ali Ali Ben, Marziye Kouroshli, Sofiya Semenova, Zakieh Hashemifar, Steven Ko, and Karthik Dantu. 2022. Edge-SLAM: Edge-assisted visual simultaneous localization and mapping. *ACM Transactions on Embedded Computing Systems* 22, 1 (2022).

[16] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[17] Romil Bhardwaj, Zhengxu Xia, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, Nikolaos Karianakis, Kevin Hsieh, Paramvir Bahl, and Ion Stoica. 2022. Ekya: Continuous Learning of Video Analytics Models on Edge Compute Servers. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. Renton, WA.

[18] Carlos Bocanegra, Mohammad Khojastepour, Mustafa Arslan, Eugene Chai, Sampath Rangarajan, and Kaushik Chowdhury. 2020. RFGo: a seamless self-checkout system for apparel stores using RFID. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[19] Tara Boroushaki, Isaac Perper, Mergen Nachin, Alberto Rodriguez, and Fadel Adib. 2021. Rfusion: Robotic grasping via rf-visual sensing and learning. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[20] Rabaï Bouderhem. 2024. Shaping the future of AI in healthcare through ethics and governance. *Humanit. Soc. Sci. Commun.* 11, 416 (March 2024).

[21] Han Cai, Ji Lin, Yujun Lin, Zhijian Liu, Haotian Tang, Hanrui Wang, Ligeng Zhu, and Song Han. 2022. Enable Deep Learning on Mobile Devices: Methods, Systems, and Applications. *ACM Transactions on Design Automation of Electronic Systems* 27, 3 (May 2022).

[22] Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. 2018. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210* (2018).

[23] Christopher Canel, Thomas Kim, Giulio Zhou, Conglong Li, Hyeontaek Lim, David Andersen, Michael Kaminsky, and Subramanya Dulloor. 2019. Scaling Video Analytics on Constrained Edge Nodes. 1905.13536 [cs.CV]

[24] Frank Cangialosi, Neil Agarwal, Venkat Arun, Srinivas Narayana, Anand Sarwate, and Ravi Netravali. 2022. Privid: Practical, Privacy-Preserving Video Analytics Queries. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. Renton, WA.

[25] Jiani Cao, Chengdong Lin, Yang Liu, and Zhenjiang Li. 2023. Gaze Tracking on Any Surface with Your Phone. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (Boston, Massachusetts) *(SenSys '22)*. New York, NY, USA.

[26] Qingqing Cao, Noah Weber, Niranjan Balasubramanian, and Aruna Balasubramanian. 2019. Deqa: On-device question answering. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*.

[27] Yetong Cao, Huijie Chen, Fan Li, and Yu Wang. 2021. Crisp-BP: Continuous wrist PPG-based blood pressure measurement. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[28] Justin Chan, Nada Ali, Ali Najafi, Anna Meehan, Lisa R Mancl, Emily Gallagher, Randall Bly, and Shyamnath Gollakota. 2022. An off-the-shelf otoacoustic-emission probe for hearing screening via a smartphone. *Nature biomedical engineering* 6, 11 (2022).

[29] Justin Chan, Antonio Glenn, Malek Itani, Lisa R Mancl, Emily Gallagher, Randall Bly, Shwetak Patel, and Shyamnath Gollakota. 2023. Wireless earbuds for low-cost hearing screening. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[30] Zhuoqing Chang, Shubo Liu, Xingxing Xiong, Zhaohui Cai, and Guoqing Tu. 2021. A Survey of Recent Advances in Edge-Computing-Powered Artificial Intelligence of Things. *IEEE Internet of Things Journal* 8, 18 (September 2021).

[31] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[32] Baicheng Chen, Huining Li, Zhengxiong Li, Xingyu Chen, Chenhan Xu, and Wenyao Xu. 2020. ThermoWave: a new paradigm of wireless passive temperature monitoring via mmWave sensing. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[33] Chen Chen, Ke Sun, and Xinyu Zhang. 2021. ExGSense: Toward Facial Gesture Sensing with a Sparse Near-Eye Sensor Array. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*. Nashville TN USA.

[34] Mingshi Chen, Panlong Yang, Jie Xiong, Maotian Zhang, Youngki Lee, Chaocan Xiang, and Chang Tian. 2019. Your Table Can Be an Input Panel: Acoustic-based Device-Free Interaction Recognition. 3, 1, Article 3 (mar 2019).

[35] Ruirong Chen, Kai Huang, and Wei Gao. 2022. AiFi: AI-Enabled WiFi Interference Cancellation with Commodity PHY-Layer Information. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[36] Wei Chen and Zhiyuan Li. 2024. Octopus v3: Technical Report for On-device Sub-billion Multimodal AI Agent. *arXiv preprint arXiv:2404.11459* (2024).

[37] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th annual international conference on mobile computing and networking*.

[38] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*.

[39] Hao-Jen Chien, Hossein Khalili, Amin Hass, and Nader Sehatbakhsh. 2023. Enc2: Privacy-Preserving Inference for Tiny IoTs via Encoding and Encryption. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (Madrid, Spain) *(ACM MobiCom '23)*. Association for Computing Machinery, New York, NY, USA, Article 35.

[40] Yae Jee Cho, Andre Manoel, Gauri Joshi, Robert Sim, and Dimitrios Dimitriadis. 2022. Heterogeneous Ensemble Knowledge Transfer for Training Large Models in Federated Learning. *International Joint Conference on Artificial Intelligence (IJCAI)* (2022).

[41] Matthew Corbett, Brendan David-John, Jiacheng Shang, Y. Charlie Hu, and Bo Ji. 2023. BystandAR: Protecting Bystander Visual Data in Augmented Reality Systems. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services* (Helsinki, Finland) *(MobiSys '23)*. Association for Computing Machinery, New York, NY, USA.

[42] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A. Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, Megan Walker, Jeffrey Yu, Tim Althoff, Conor Heneghan, John Hernandez, Mark Malhotra, Leor Stern, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Jiening Zhan, Shruthi Prabhakara, Daniel McDuff, and Cory Y. McLean. 2024. Towards a Personal Health Large Language Model. *arXiv* (June 2024). 2406.06474

[43] Kaiyan Cui, Yanwen Wang, Yuanqing Zheng, and Jinsong Han. 2021. ShakeReader: 'Read' UHF RFID using Smartphone. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*.

[44] Yimin Dai, Xian Shuai, Rui Tan, and Guoliang Xing. 2023. Interpersonal distance tracking with mmWave radar and IMUs. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[45] Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. 2022. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*. PMLR.

[46] Mallesham Dasari, Kumara Kahatapitiya, Samir Das, Aruna Balasubramanian, and Dimitris Samaras. 2022. Swift: Adaptive Video Streaming with Layered Neural Codecs. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. Renton, WA.

[47] Yongheng Deng, Weining Chen, Ju Ren, Feng Lyu, Yang Liu, Yunxin Liu, and Yaoxue Zhang. 2022. TailorFL: Dual-Personalized Federated Learning under System and Data Heterogeneity. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[48] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264* (2020).

[49] Dimitrios Dimitriadis, Mirian Hipolito Garcia, Daniel Diaz, Andre Manoel, and Robert Sim. 2022. FLUTE: A Scalable, Extensible Framework for High-Performance Federated Learning Simulations. *ArXiv* abs/2203.13789 (2022).

[50] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[51] Gaofeng Dong, Jason Wu, Julian De Gortari Briseno, Akash Deep Singh, Justin Feng, Ankur Sarker, Nader Sehatbakhsh, and Mani Srivastava. 2024. RefreshChannels: Exploiting Dynamic Refresh Rate Switching for Mobile Device Attacks. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services* (Minato-ku, Tokyo, Japan) *(MOBISYS '24)*. Association for Computing Machinery, New York, NY, USA.

[52] Jialuo Du, Yidong Ren, Zhui Zhu, Chenning Li, Zhichao Cao, Qiang Ma, and Yunhao Liu. 2023. SRLoRa: Neural-enhanced LoRa Weak Signal Decoding with Multi-gateway Super Resolution. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*.

[53] Kuntai Du, Qizheng Zhang, Anton Arapin, Haodong Wang, Zhengxu Xia, and Junchen Jiang. 2022. AccMPEG: Optimizing Video Encoding for Video Analytics. *arXiv* (April 2022). 2204.12534

[54] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 56 (may 2024).

[55] European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[56] Biyi Fang, Jillian Co, and Mi Zhang. 2017. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*.

[57] Biyi Fang, Xiao Zeng, Faen Zhang, Hui Xu, and Mi Zhang. 2020. Flexdnn: Input-adaptive on-device deep learning for efficient mobile vision. In *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE.

[58] Biyi Fang, Xiao Zeng, and Mi Zhang. 2018. Nestdnn: Resource-aware multi-tenant on-device deep learning for continuous mobile vision. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*.

[59] Habiba Farrukh, Reham Aburas, Siyuan Cao, and He Wang. 2020. FaceRevelio: a face liveness detection system for smartphones with a single front camera. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[60] Boyuan Feng, Yuke Wang, Gushu Li, Yuan Xie, and Yufei Ding. 2021. Palleon: A runtime system for efficient video processing toward dynamic class skew. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*.

[61] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. 2023. FedMultimodal: A Benchmark For Multimodal Federated Learning. 2306.09486 [cs.DC] https://arxiv.org/abs/2306.09486

[62] Yuda Feng, Yaxiong Xie, Deepak Ganesan, and Jie Xiong. 2021. Lte-based pervasive sensing across indoor and outdoor. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[63] Yuda Feng, Yaxiong Xie, Deepak Ganesan, and Jie Xiong. 2022. Lte-based low-cost and low-power soil moisture sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[64] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[65] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[66] Ziyan Fu, Ju Ren, Yunxin Liu, Ting Cao, Deyu Zhang, Yuezhi Zhou, and Yaoxue Zhang. 2022. Hyperion: A Generic and Distributed Mobile Offloading Framework on OpenCL. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[67] Nakul Garg, Yang Bai, and Nirupam Roy. 2021. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*.

[68] Nakul Garg and Nirupam Roy. 2023. Sirius: A self-localization system for resource-constrained iot sensors. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[69] Petko Georgiev, Sourav Bhattacharya, Nicholas Lane, and Cecilia Mascolo. 2017. Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017).

[70] Petko Georgiev, Nicholas Lane, Kiran Rachuri, and Cecilia Mascolo. 2016. LEO: Scheduling sensor inference algorithms across heterogeneous mobile processors and network resources. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*.

[71] Seyed Keyarash Ghiasi, Vivian Dsouza, Koen Langendoen, and Marco Zuniga. 2023. SpectraLux: Towards Exploiting the Full Spectrum with Passive VLC. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[72] Tripat Gill. 2021. Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics Inf. Technol.* 23, 4 (Dec. 2021).

[73] In Gim and JeongGil Ko. 2022. Memory-efficient DNN training on mobile devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[74] Graham Gobieski, Brandon Lucia, and Nathan Beckmann. 2019. Intelligence beyond the edge: Inference on intermittent embedded systems. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*.

[75] Yongjie Guan, Xueyu Hou, Nan Wu, Bo Han, and Tao Han. 2022. DeepMix: Mobility-aware, Lightweight, and Hybrid 3D Object Detection for Headsets. http://arxiv.org/abs/2201.08812 arXiv:2201.08812 [cs].

[76] Yu Guan, Chengyuan Zheng, Xinggong Zhang, Zongming Guo, and Junchen Jiang. 2019. Pano: Optimizing 360° Video Streaming with a Better Understanding of Quality Perception. In *Proceedings of the ACM Special Interest Group on Data Communication* (Beijing, China) *(SIGCOMM '19)*. New York, NY, USA.

[77] Peizhen Guo, Bo Hu, and Wenjun Hu. 2021. Mistify: Automating {DNN} Model Porting for {On-Device} Inference at the Edge. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*.

[78] Peizhen Guo and Wenjun Hu. 2018. Potluck: Cross-application approximate deduplication for computation-intensive mobile applications. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*.

[79] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless seismocardiography via deep learning radars. In *Proceedings of the 26th annual international conference on mobile computing and networking*.

[80] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. 2020. Food and liquid sensing in practical environments using {RFIDs}. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*.

[81] Mingcong Han, Hanze Zhang, Rong Chen, and Haibo Chen. 2022. Microsecond-scale preemption for concurrent {GPU-accelerated}{DNN} inferences. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*.

[82] Rui Han, Qinglong Zhang, Chi Liu, Guoren Wang, Jian Tang, and Lydia Chen. 2021. Legodnn: block-grained scaling of deep neural networks for mobile vision. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[83] Seungyeop Han, Haichen Shen, Matthai Philipose, Sharad Agarwal, Alec Wolman, and Arvind Krishnamurthy. 2016. Mcdnn: An approximation-based execution framework for deep stream processing under resource constraints. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*.

[84] Chaoyang He, Keshav Balasubramanian, Emir Ceyani, Carl Yang, Han Xie, Lichao Sun, Lifang He, Liangwei Yang, Philip S Yu, Yu Rong, et al. 2021. Fedgraphnn: A federated learning system and benchmark for graph neural networks. *arXiv preprint arXiv:2104.07145* (2021).

[85] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. 2020. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518* (2020).

[86] Chaoyang He, Alay Dilipbhai Shah, Zhenheng Tang, Di Fan1Adarshan Naiynar Sivashunmugam, Keerti Bhogaraju, Mita Shimpi, Li Shen, Xiaowen Chu, Mahdi Soltanolkotabi, and Salman Avestimehr. 2021. Fedcv: a federated learning framework for diverse computer vision tasks. *arXiv preprint arXiv:2111.11066* (2021).

[87] Guorong He, Shaojie Chen, Dan Xu, Xiaojiang Chen, Yaxiong Xie, Xinhuai Wang, and Dingyi Fang. 2023. Fusang: Graph-inspired Robust and Accurate Object Recognition on Commodity mmWave Devices. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[88] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[89] Shibo He, Kun Shi, Chen Liu, Bicheng Guo, Jiming Chen, and Zhiguo Shi. 2022. Collaborative Sensing in Internet of Things: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials* 24, 3 (2022).

[90] Yuze He, Li Ma, Jiahe Cui, Zhenyu Yan, Guoliang Xing, Sen Wang, Qintao Hu, and Chen Pan. 2022. AutoMatch: Leveraging Traffic Camera to Improve Perception and Localization of Autonomous Vehicles. In *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*. Boston Massachusetts.

[91] Yuze He, Li Ma, Zhehao Jiang, Yi Tang, and Guoliang Xing. 2021. VI-eye: semantic-based 3D point cloud registration for infrastructure-assisted autonomous driving. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. New Orleans Louisiana.

[92] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. 2021. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems* 34 (2021).

[93] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. Signspeaker: A real-time, high-precision smartwatch-based sign language translator. In *The 25th Annual International Conference on Mobile Computing and Networking*.

[94] Kun Mean Hou, Xunxing Diao, Hongling Shi, Hao Ding, Haiying Zhou, and Christophe de Vaulx. 2023. Trends and Challenges in AIoT/IIoT/IoT Implementation. *Sensors* 23, 11 (January 2023). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[95] Xueyu Hou, Yongjie Guan, and Tao Han. 2022. NeuLens: spatial-based dynamic acceleration of convolutional neural networks on edge. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[96] Pan Hu, Junha Im, Zain Asgar, and Sachin Katti. 2020. Starfish: Resilient image compression for AIoT cameras. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[97] Zhizhang Hu, Yue Zhang, Tong Yu, and Shijia Pan. 2022. VMA: Domain Variance-and Modality-Aware Model Transfer for Fine-Grained Occupant Activity Recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[98] Jin Huang, Colin Samplawski, Deepak Ganesan, Benjamin Marlin, and Heesung Kwon. 2020. Clio: Enabling automatic compilation of deep learning pipelines across iot and cloud. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[99] Kai Huang and Wei Gao. 2022. Real-time neural network inference on extremely weak devices: agile offloading with explainable AI. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[100] Kai Huang, Boyuan Yang, and Wei Gao. 2023. ElasticTrainer: Speeding Up On-Device Training with Runtime Elastic Tensor Selection. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[101] Loc Huynh, Youngki Lee, and Rajesh Balan. 2017. Deepmon: Mobile gpu-based deep learning framework for continuous vision applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*.

[102] Sinh Huynh, Rajesh Balan, JeongGil Ko, and Youngki Lee. 2019. VitaMon: measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*.

[103] Jinwoo Hwang, Minsu Kim, Daeun Kim, Seungho Nam, Yoonsung Kim, Dohee Kim, Hardik Sharma, and Jongse Park. 2022. CoVA: Exploiting Compressed-Domain Analysis to Accelerate Video Analytics. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*. Carlsbad, CA.

[104] Joo Jeong, Jingyu Lee, Donghyun Kim, Changmin Jeon, Changjin Jeong, Youngki Lee, and Byung-Gon Chun. 2022. Band: coordinated multi-dnn inference on heterogeneous mobile processors. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[105] Jeya Jeyakumar, Liangzhen Lai, Naveen Suda, and Mani Srivastava. 2019. SenseHAR: a robust virtual activity sensor for smartphones and wearables. In *SenSys '19: Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. New York, NY, USA.

[106] Divyansh Jhunjhunwala, Advait Gadhikar, Gauri Joshi, and Yonina C Eldar. 2021. Adaptive quantization of model updates for communication-efficient federated learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[107] Sijie Ji, Yaxiong Xie, and Mo Li. 2022. SiFall: Practical Online Fall Detection with RF Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[108] Fucheng Jia, Deyu Zhang, Ting Cao, Shiqi Jiang, Yunxin Liu, Ju Ren, and Yaoxue Zhang. 2022. Codl: efficient cpu-gpu co-execution for deep learning inference on mobile devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. Association for Computing Machinery New York, NY, USA.

[109] Angela Jiang, Daniel Wong, Christopher Canel, Lilia Tang, Ishan Misra, Michael Kaminsky, Michael Kozuch, Padmanabhan Pillai, David Andersen, and Gregory Ganger. 2018. Mainstream: Dynamic Stem-Sharing$for Multi-Tenant Video Processing. In 2018 USENIX Annual Technical Conference (USENIX ATC 18)$.

[110] Junchen Jiang, Ganesh Ananthanarayanan, Peter Bodik, Siddhartha Sen, and Ion Stoica. 2018. Chameleon: Scalable Adaptation of Video Analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. New York, NY, USA.

[111] Linshan Jiang, Qun Song, Rui Tan, and Mo Li. 2022. PriMask: Cascadable and Collusion-Resilient Data Masking for Mobile Cloud Inference. In *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*. http://arxiv.org/abs/2211.06716 arXiv:2211.06716 [cs].

[112] Shiqi Jiang, Zhiqi Lin, Yuanchun Li, Yuanchao Shu, and Yunxin Liu. 2021. Flexible high-resolution object detection on edge devices with tunable latency. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[113] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*.

[114] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[115] Zhehao Jiang, Neiwen Ling, Xuan Huang, Shuyao Shi, Chenhao Wu, Xiaoguang Zhao, Zhenyu Yan, and Guoliang Xing. 2023. CoEdge: A Cooperative Edge System for Distributed Real-Time Deep Learning Tasks. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[116] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. EarHealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[117] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning* 14, 1–2 (2021).

[118] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News* 45, 1 (2017).

[119] Momin Ahmad Khan, Virat Shejwalkar, Amir Houmansadr, and Fatima M. Anwar. 2023. Security Analysis of SplitFed Learning *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA.

[120] Mehrdad Khani, Ganesh Ananthanarayanan, Kevin Hsieh, Junchen Jiang, Ravi Netravali, Yuanchao Shu, Mohammad Alizadeh, and Victor Bahl. 2023. RECL: Responsive Resource-Efficient Continuous Learning for Video Analytics. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA.

[121] Jaehong Kim, Youngmok Jung, Hyunho Yeo, Juncheol Ye, and Dongsu Han. 2020. Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication* (Virtual Event, USA) *(SIGCOMM '20)*. New York, NY, USA.

[122] Youngsok Kim, Joonsung Kim, Dongju Chae, Daehyun Kim, and Jangwoo Kim. 2019. μlayer: Low latency on-device inference using cooperative single-layer acceleration and processor-friendly quantization. In *Proceedings of the Fourteenth EuroSys Conference 2019*.

[123] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[124] Rui Kong, Yuanchun Li, Yizhen Yuan, and Linghe Kong. 2023. ConvReLU++: Reference-based Lossless Acceleration of Conv-ReLU Operations on Mobile CPU. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[125] Belal Korany, Chitra Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *The 25th Annual International Conference on Mobile Computing and Networking*.

[126] Young Kwon, Jagmohan Chauhan, and Cecilia Mascolo. 2022. Yono: Modeling multiple heterogeneous neural networks on microcontrollers. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[127] Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. 2022. Fedscale: Benchmarking model and system performance of federated learning at scale. In *International Conference on Machine Learning*. PMLR.

[128] Fan Lai, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. 2021. Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*.

[129] Guohao Lan, Bailey Heit, Tim Scargill, and Maria Gorlatova. 2020. GazeGraph: Graph-based few-shot cognitive context sensing from human visual behavior. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[130] Nicholas Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, Lei Jiao, Lorena Qendro, and Fahim Kawsar. 2016. Deepx: A software accelerator for low-power deep learning inference on mobile devices. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[131] Nicholas Lane, Petko Georgiev, and Lorena Qendro. 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*.

[132] Stefanos Laskaridis, Stylianos Venieris, Mario Almeida, Ilias Leontiadis, and Nicholas Lane. 2020. SPINN: synergistic progressive inference of neural networks over device and cloud. In *Proceedings of the 26th annual international conference on mobile computing and networking*.

[133] Jinsung Lee, Sungyong Lee, Jongyun Lee, Sandesh Sathyanarayana, Hyoyoung Lim, Jihoon Lee, Xiaoqing Zhu, Sangeeta Ramakrishnan, Dirk Grunwald, Kyunghan Lee, et al. 2020. PERCEIVE: deep learning-based cellular uplink prediction using real-time scheduling

patterns. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*.
[134] Royson Lee, Stylianos Venieris, Lukasz Dudziak, Sourav Bhattacharya, and Nicholas Lane. 2019. Mobisr: Efficient on-device super-resolution through heterogeneous mobile processors. In *The 25th annual international conference on mobile computing and networking*.
[135] Seulki Lee. 2023. MicroDeblur: Image Motion Deblurring on Microcontroller-based Vision Systems. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.
[136] Seulki Lee and Shahriar Nirjon. 2020. Fast and scalable in-memory deep multitask learning via neural weight virtualization. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*.
[137] Clayton Leite and Yu Xiao. 2021. Optimal sensor channel selection for resource-efficient deep activity recognition. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*.
[138] Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. 2021. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.
[139] Chenning Li, Hanqing Guo, Shuai Tong, Xiao Zeng, Zhichao Cao, Mi Zhang, Qiben Yan, Li Xiao, Jiliang Wang, and Yunhao Liu. 2021. NELoRa: Towards ultra-low SNR LoRa communication with neural-enhanced demodulation. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.
[140] Chenning Li, Zheng Liu, Yuguang Yao, Zhichao Cao, Mi Zhang, and Yunhao Liu. 2020. Wi-fi see it all: generative adversarial network-augmented versatile wi-fi imaging. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.
[141] Chenning Li, Yidong Ren, Shuai Tong, Shakhrul Iman Siam, Mi Zhang, Jiliang Wang, Yunhao Liu, and Zhichao Cao. 2024. ChirpTransformer: Versatile LoRa Encoding for Low-power Wide-area IoT. In *Proceedings of ACM MobiSys*.
[142] Changming Li, Mingjing Xu, Yicong Du, Limin Liu, Cong Shi, Yan Wang, Hongbo Liu, and Yingying Chen. 2024. Practical Adversarial Attack on WiFi Sensing Through Unnoticeable Communication Packet Perturbation *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA.
[143] Chenning Li, Xiao Zeng, Mi Zhang, and Zhichao Cao. 2022. PyramidFL: A fine-grained client selection framework for efficient federated learning. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.
[144] Dong Li, Shirui Cao, Sunghoon Ivan Lee, and Jie Xiong. 2022. Experience: practical problems for acoustic sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.
[145] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.
[146] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581* (2019).
[147] Hongyu Li, Hairong Wang, Luyang Liu, and Marco Gruteser. 2018. Automatic Unusual Driving Event Identification for Dependable Self-Driving. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. Shenzhen China.
[148] Kehan Li, Jiming Chen, Baosheng Yu, Zhangchong Shen, Chao Li, and Shibo He. 2020. Supreme: Fine-grained radio map reconstruction via spatial-temporal fusion network. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.
[149] Tianxing Li, Jin Huang, Erik Risinger, and Deepak Ganesan. 2021. Low-latency speculative inference on distributed multi-modal data streams. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*.
[150] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall Burd. 2016. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*.
[151] Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Xu, and Ravi Netravali. 2020. Reducto: On-Camera Filtering for Resource-Efficient Real-Time Video Analytics. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication* (Virtual Event, USA) *(SIGCOMM '20)*. New York, NY, USA.
[152] Bill Yuchen Lin, Chaoyang He, Zihang Zeng, Hulin Wang, Yufen Huang, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2021. Fednlp: A research platform for federated learning in natural language processing. *arXiv preprint arXiv:2104.08815* (2021).
[153] Ji Lin, Wei-Ming Chen, Han Cai, Chuang Gan, and Song Han. 2021. Mcunetv2: Memory-efficient patch-based inference for tiny deep learning. *arXiv preprint arXiv:2110.15352* (2021).
[154] Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. 2022. On-device training under 256kb memory. *Advances in Neural Information Processing Systems* 35 (2022).
[155] Shih-Chieh Lin, Yunqi Zhang, Chang-Hong Hsu, Matt Skach, Md Haque, Lingjia Tang, and Jason Mars. 2018. The Architectural Implications of Autonomous Driving: Constraints and Acceleration. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. Williamsburg VA USA.
[156] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems* 33 (2020).

[157] Neiwen Ling, Xuan Huang, Zhihe Zhao, Nan Guan, Zhenyu Yan, and Guoliang Xing. 2022. BlastNet: Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[158] Neiwen Ling, Kai Wang, Yuze He, Guoliang Xing, and Daqi Xie. 2021. Rt-mdl: Supporting real-time mixed deep learning tasks on edge platforms. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[159] Mieszko Lis, Maximilian Golub, and Guy Lemieux. 2019. Full deep neural network training on a pruned weight budget. *Proceedings of Machine Learning and Systems* 1 (2019).

[160] Cihang Liu, Lan Zhang, Zongqian Liu, Kebin Liu, Xiangyang Li, and Yunhao Liu. 2016. Lasagna: towards deep hierarchical understanding and searching over mobile sensing data. In *MobiCom '16: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. New York, NY, USA.

[161] Hansi Liu, Abrar Alali, Mohamed Ibrahim, Bryan Cao, Nicholas Meegan, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, et al. 2022. Vi-Fi: Associating Moving Subjects across Vision and Wireless Sensors. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[162] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, Xian Xu, and Kui Ren. 2021. MandiPass: Secure and Usable User Authentication via Earphone IMU. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*.

[163] Jiesong Liu, Feng Zhang, Jiawei Guan, Hsin-Hsuan Sung, Xiaoguang Guo, Xiaoyong Du, and Xipeng Shen. 2023. Space-Efficient TREC for Enabling Deep Learning on Microcontrollers. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*.

[164] Luyang Liu, Hongyu Li, and Marco Gruteser. 2019. Edge Assisted Real-time Object Detection for Mobile Augmented Reality. In *The 25th Annual International Conference on Mobile Computing and Networking*. Los Cabos Mexico.

[165] Luyang Liu, Hongyu Li, Jian Liu, Cagdas Karatas, Yan Wang, Marco Gruteser, Yingying Chen, and Richard Martin. 2017. BigRoad: Scaling Road Data Acquisition for Dependable Self-Driving. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. Niagara Falls New York USA.

[166] Li Liu, Yuguang Yao, Zhichao Cao, and Mi Zhang. 2021. DeepLoRa: Learning accurate path loss model for long distance links in LPWAN. In *INFOCOM*.

[167] Miaomiao Liu, Sikai Yang, Wyssanie Chomsin, and Wan Du. 2022. Real-Time Tracking of Smartwatch Orientation and Location by Multitask Learning. In *SenSys '22: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. New York, NY, USA.

[168] Sicong Liu, Bin Guo, Cheng Fang, Ziqi Wang, Shiyan Luo, Zimu Zhou, and Zhiwen Yu. 2023. Enabling Resource-Efficient AIoT System With Cross-Level Optimization: A Survey. *IEEE Communications Surveys & Tutorials* (2023).

[169] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[170] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[171] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time Arm Skeleton Tracking and Gesture Inference Tolerant to Missing Wearable Sensors. In *MobiSys '19: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA.

[172] Yan Liu, Anlan Yu, Leye Wang, Bin Guo, Yang Li, Enze Yi, and Daqing Zhang. 2024. UniFi: A Unified Framework for Generalizable Gesture Recognition with Wi-Fi Signals Using Consistency-guided Multi-View Networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024).

[173] Zida Liu, Guohao Lan, Jovan Stojkovic, Yunfan Zhang, Carlee Joe-Wong, and Maria Gorlatova. 2020. CollabAR: Edge-assisted Collaborative Image Recognition for Mobile Augmented Reality. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. Sydney, NSW, Australia.

[174] Zihan Liu, Jingwen Leng, Zhihui Zhang, Quan Chen, Chao Li, and Minyi Guo. 2022. VELTAIR: towards high-performance multi-tenant deep learning services via adaptive compilation and scheduling. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.

[175] Zikun Liu, Gagandeep Singh, Chenren Xu, and Deepak Vasisht. 2021. FIRE: enabling reciprocity for FDD MIMO systems. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[176] Ziwei Liu, Tianyue Zheng, Chao Hu, Yanbing Yang, Yimao Sun, Yi Zhang, Zhe Chen, Liangyin Chen, and Jun Luo. 2022. CORE-lens: Simultaneous communication and object recognition with disentangled-GAN cameras. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[177] Chris Lu, Muhamad Saputra, Peijun Zhao, Yasin Almalioglu, Gusmao Pedro De, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[178] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on*

Mobile Systems, Applications, and Services.
[179] Yan Lu, Shiqi Jiang, Ting Cao, and Yuanchao Shu. 2023. Turbo: Opportunistic Enhancement for Edge Video Analytics. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (Boston, Massachusetts) *(SenSys '22)*. New York, NY, USA.
[180] Wenjie Luo, Qun Song, Zhenyu Yan, Rui Tan, and Guosheng Lin. 2022. Indoor Smartphone SLAM with Learned Echoic Location Features. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.
[181] Zhiqing Luo, Wei Wang, Jun Qu, Tao Jiang, and Qian Zhang. 2018. ShieldScatter: Improving IoT Security with Backscatter Assistance *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA.
[182] Qianpiao Ma, Yang Xu, Hongli Xu, Zhida Jiang, Liusheng Huang, and He Huang. 2021. FedSA: A semi-asynchronous federated learning mechanism in heterogeneous edge computing. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021).
[183] Ajay Mahimkar, Ashiwan Sivakumar, Zihui Ge, Shomik Pathak, and Karunasish Biswas. 2021. Auric: using data-driven recommendation to automatically generate cellular configuration. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*.
[184] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (Los Angeles, CA, USA) *(SIGCOMM '17)*. New York, NY, USA.
[185] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic Ranging via Deep Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 143 (dec 2020).
[186] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*.
[187] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas Lane. 2019. Mic2mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems. In *Proceedings of the 18th international conference on information processing in sensor networks*.
[188] Akhil Mathur, Nicholas Lane, Sourav Bhattacharya, Aidan Boran, Claudio Forlivesi, and Fahim Kawsar. 2017. Deepeye: Resource efficient local execution of multiple deep vision models using wearable commodity hardware. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*.
[189] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR.
[190] Yaxin Mei, Wenhua Wang, Yuzhu Liang, Qin Liu, Shuhong Chen, and Tian Wang. 2023. Privacy-Enhanced Cooperative Storage Scheme for Contact-Free Sensory Data in AIoT with Efficient Synchronization. *ACM Trans. Sen. Netw.* (sep 2023). Just Accepted.
[191] Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. 2024. Transforming Wearable Data into Health Insights using Large Language Model Agents. *arXiv* (June 2024). 2406.06464
[192] Chulhong Min, Alessandro Montanari, Akhil Mathur, and Fahim Kawsar. 2019. A closer look at quality-aware runtime assessment of sensing models in multi-device environments. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*.
[193] Niluthpol Mithun, Sirajum Munir, Karen Guo, and Charles Shelton. 2018. ODDS: real-time object detection using depth sensors on embedded GPUs. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.
[194] Brent Mittelstadt. 2017. Ethics of the health-related internet of things: a narrative review. *Ethics Inf. Technol.* 19, 3 (Sept. 2017).
[195] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. 2021. PPFL: privacy-preserving federated learning with trusted execution environments. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* (Virtual Event, Wisconsin) *(MobiSys '21)*. Association for Computing Machinery, New York, NY, USA.
[196] Mehdi Mohammadi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani. 2018. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. *IEEE Communications Surveys & Tutorials* 20, 4 (2018).
[197] David C. Mohr, Mi Zhang, and Stephen Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology (ARCP)* 13, 1 (2017). http://www.annualreviews.org/doi/pdf/10.1146/annurev-clinpsy-032816-044949
[198] Mahathir Monjur, Yubo Luo, Zhenyu Wang, and Shahriar Nirjon. 2023. SoundSieve: Seconds-Long Audio Event Recognition on Intermittently-Powered Systems. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[199] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael Jordan, et al. 2018. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*.
[200] John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. 2022. Federated learning with buffered asynchronous aggregation. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
[201] Viet Nguyen, Siddharth Rupavatharam, Luyang Liu, Richard Howard, and Marco Gruteser. 2019. HandSense: capacitive coupling-based dynamic, micro finger gesture recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. New York New York.
[202] State of California. 2018. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov.

[203] U.S. Department of Health and Human Services. 1996. Health Insurance Portability and Accountability Act of 1996. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf.

[204] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[205] Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. 2403.19857 [cs.AI] https://arxiv.org/abs/2403.19857

[206] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*.

[207] Arthi Padmanabhan, Neil Agarwal, Anand Iyer, Ganesh Ananthanarayanan, Yuanchao Shu, Nikolaos Karianakis, Guoqing Xu, and Ravi Netravali. 2023. Gemel: Model Merging for Memory-Efficient, Real-Time Video Analytics at the Edge. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA.

[208] Jongseok Park, Kyungmin Bin, and Kyunghan Lee. 2022. mGEMM: low-latency convolution with minimal memory overhead optimized for mobile devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[209] JaeYeon Park, Kichang Lee, Sungmin Lee, Mi Zhang, and JeongGil Ko. 2023. AttFL: A Personalized Federated Learning Framework for Time-series Mobile and Embedded Sensor Data Processing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 116 (sep 2023).

[210] Keondo Park, You Choi, Inhoe Lee, and Hyung-Sin Kim. 2023. PointSplit: Towards On-device 3D Object Detection with Heterogeneous Low-power Accelerators. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[211] Seonghoon Park, Yeonwoo Cho, Hyungchol Jun, Jeho Lee, and Hojung Cha. 2023. OmniLive: Super-Resolution Enhanced 360° Video Live Streaming for Mobile Devices. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services* (Helsinki, Finland) *(MobiSys '23)*. New York, NY, USA.

[212] Sibendu Paul, Kunal Rao, Giuseppe Coviello, Murugan Sankaradas, Oliver Po, Y. Hu, and Srimat Chakradhar. 2023. Enhancing Video Analytics Accuracy via Real-Time Automated Camera Parameter Tuning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (Boston, Massachusetts) *(SenSys '22)*. New York, NY, USA.

[213] Jacopo Pegoraro, Jesus O Lacruz, Michele Rossi, and Joerg Widmer. 2022. SPARCS: A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[214] Daniel F Perez-Ramirez, Carlos Pérez-Penichet, Nicolas Tsiftes, Thiemo Voigt, Dejan Kostić, and Magnus Boman. 2023. DeepGANTT: A Scalable Deep Learning Scheduler for Backscatter Networks. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[215] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas Lane, Cecilia Mascolo, Mahesh Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018).

[216] Tauhidur Rahman, Alexander T Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. Bodybeat: A Mobile System for Sensing Non-Speech Body Sounds. In *The 12th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

[217] Enrico Reggiani, Cristóbal Lazo, Roger Bagué, Adrián Cristal, Mauro Olivieri, and Osman Unsal. 2022. Bison-e: A lightweight and high-performance accelerator for narrow integer linear algebra computing on the edge. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.

[218] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International conference on artificial intelligence and statistics*. PMLR.

[219] Ao Ren, Zhe Li, Caiwen Ding, Qinru Qiu, Yanzhi Wang, Ji Li, Xuehai Qian, and Bo Yuan. 2017. Sc-dcnn: Highly-scalable deep convolutional neural network using stochastic computing. *ACM SIGPLAN Notices* 52, 4 (2017).

[220] Yidong Ren, Li Liu, Chenning Li, Zhichao Cao, and Shigang Chen. 2022. Is lorawan really wide? fine-grained lora link-level measurement in an urban environment. In *2022 IEEE 30th International Conference on Network Protocols (ICNP)*. IEEE.

[221] Michael Rudow, Francis Yan, Abhishek Kumar, Ganesh Ananthanarayanan, Martin Ellis, and K.V. Rashmi. 2023. Tambur: Efficient loss recovery for videoconferencing via streaming codes. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA.

[222] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research (JMIR)* 17, 7 (2015).

[223] Sriram Sami, Sean Tan, Bangjie Sun, and Jun Han. 2021. LAPD: Hidden Spy Camera Detection using Smartphone Time-of-Flight Sensors. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[224] Sandeep Singh Sandha, Bharathan Balaji, Luis Garcia, and Mani Srivastava. 2023. Eagle: End-to-end Deep Reinforcement Learning based Autonomous Control of PTZ Cameras *(IoTDI '23)*. New York, NY, USA.

[225] Shamik Sarkar, Milind Buddhikot, Aniqua Baset, and Sneha Kumar Kasera. 2021. DeepRadar: A deep-learning-based environmental sensing capability sensor design for CBRS. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[226] Khotso Selialia, Yasra Chandio, and Fatima M. Anwar. 2023. Federated Learning Biases in Heterogeneous Edge-Devices: A Case-Study *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA.

[227] Hossein Shafagh, Anwar Hithnawi, Lukas Burkhalter, Pascal Fischli, and Simon Duquennoy. 2017. Secure sharing of partially homomorphic encrypted iot data. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*.

[228] Hailan Shanbhag, Sohrab Madani, Akhil Isanaka, Deepak Nair, Saurabh Gupta, and Haitham Hassanieh. 2023. Contactless Material Identification with Millimeter Wave Vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[229] Zhihao Shen, Wan Du, Xi Zhao, and Jianhua Zou. 2020. DMM: Fast map matching for cellular data. In *Proceedings of the 26th annual international conference on mobile computing and networking*.

[230] Junyang Shi, Mo Sha, and Xi Peng. 2021. Adapting wireless mesh network configuration from simulation to reality via deep learning based domain adaptation. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*.

[231] Shuyao Shi, Jiahe Cui, Zhehao Jiang, Zhenyu Yan, Guoliang Xing, Jianwei Niu, and Zhenchao Ouyang. 2022. VIPS: real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. Sydney NSW Australia.

[232] Shaohuai Shi, Qiang Wang, Kaiyong Zhao, Zhenheng Tang, Yuxin Wang, Xiang Huang, and Xiaowen Chu. 2019. A distributed synchronous SGD algorithm with global top-k sparsification for low bandwidth networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE.

[233] Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, and Khaled B. Letaief. 2020. Communication-Efficient Edge AI: Algorithms and Systems. http://arxiv.org/abs/2002.09668 arXiv:2002.09668 [cs, eess, math].

[234] Seungwoo Shim, Hyeonho Shin, Myeongkyun Cho, Youngki Lee, Jinwoo Shin, and Song Kim. 2023. Mosaic: Extremely Low-resolution RFID Vision for Visually-anonymized Action Recognition. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.

[235] Jaemin Shin, Yuanchun Li, Yunxin Liu, and Sung-Ju Lee. 2022. FedBalancer: data and pace control for efficient federated learning on heterogeneous clients. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[236] Xian Shuai, Yulin Shen, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. 2022. BalanceFL: Addressing class imbalance in long-tail federated learning. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[237] Akash Deep Singh, Luis Garcia, Joseph Noor, and Mani Srivastava. 2021. I Always Feel Like Somebody's Sensing Me! A Framework to Detect, Identify, and Localize Clandestine Wireless Sensors. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association.

[238] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. 2022. Rf-url: unsupervised representation learning for rf sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[239] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[240] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[241] Bangjie Sun, Sean Tan, Zhiwei Ren, Mun Chan, and Jun Han. 2022. Detecting counterfeit liquid food products in a sealed bottle using a smartphone camera. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[242] Jingwei Sun, Ang Li, Lin Duan, Samiul Alam, Xuliang Deng, Xin Guo, Haiming Wang, Maria Gorlatova, Mi Zhang, Hai Li, et al. 2022. FedSEA: A Semi-Asynchronous Federated Learning Framework for Extremely Heterogeneous Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[243] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th annual international conference on mobile computing and networking*.

[244] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*.

[245] Rishub Tamirisa, John Won, Chengjun Lu, Ron Arel, and Andy Zhou. 2024. FedSelect: Customized Selection of Parameters for Fine-Tuning during Personalized Federated Learning. *2024 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[246] Tianxiang Tan and Guohong Cao. 2021. Efficient execution of deep neural networks on mobile devices with npu. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)*.

[247] Linpeng Tang, Qi Huang, Amit Puntambekar, Ymir Vigfusson, Wyatt Lloyd, and Kai Li. 2017. Popularity Prediction of Facebook Videos for Higher Quality Streaming. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*. Santa Clara, CA.

[248] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. 2022. FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. *arXiv preprint arXiv:2210.04620* (2022).

[249] Panrong Tong, Mingqian Li, Mo Li, Jianqiang Huang, and Xiansheng Hua. 2021. Large-scale vehicle trajectory reconstruction with camera sensing network. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. New Orleans Louisiana.

[250] Vu Tran, Gihan Jayatilaka, Ashwin Ashok, and Archan Misra. 2021. Deeplight: Robust & unobtrusive real-time screen-camera communication for real-world displays. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*.

[251] Marco Trinelli, Massimo Gallo, Myriana Rifai, and Fabio Pianese. 2019. Transparent AR Processing Acceleration at the Edge. In *Proceedings of the 2nd International Workshop on Edge Systems, Analytics and Networking*. Dresden Germany.

[252] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[253] Saeed Vahidian, Mahdi Morafah, and Bill Lin. 2021. Personalized federated learning by structured and unstructured pruning under data heterogeneity. In *2021 IEEE 41st international conference on distributed computing systems workshops (ICDCSW)*. IEEE.

[254] Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. 2024. MEIT: Multi-Modal Electrocardiogram Instruction Tuning on Large Language Models for Report Generation. *arXiv* (March 2024). 2403.04945

[255] Chia-Cheng Wang, Jyh-Cheng Chen, Yi Chen, Rui-Heng Tu, Jia-Jiun Lee, Yu-Xin Xiao, and Shan-Yu Cai. 2021. MVP: magnetic vehicular positioning system for GNSS-denied environments. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. New Orleans Louisiana.

[256] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. 2021. A Field Guide to Federated Optimization. 2107.06917 [cs.LG]

[257] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-Agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158* (2024).

[258] Kailong Wang, Cong Shi, Jerry Cheng, Yan Wang, Minge Xie, and Yingying Chen. 2022. Solving the WiFi Sensing Dilemma in Reality Leveraging Conformal Prediction. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[259] Manni Wang, Shaohua Ding, Ting Cao, Yunxin Liu, and Fengyuan Xu. 2021. Asymo: scalable and efficient deep-learning inference on asymmetric mobile cpus. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[260] Qipeng Wang, Mengwei Xu, Chao Jin, Xinran Dong, Jinliang Yuan, Xin Jin, Gang Huang, Yunxin Liu, and Xuanzhe Liu. 2022. Melon: Breaking the memory wall for resource-efficient on-device machine learning. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*.

[261] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*.

[262] Shibo Wang, Shusen Yang, Hailiang Li, Xiaodan Zhang, Chen Zhou, Chenren Xu, Feng Qian, Nanbin Wang, and Zongben Xu. 2022. SalientVR: saliency-driven mobile 360-degree video streaming with gaze information. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. Sydney NSW Australia.

[263] Xin Wang, Zhongwei Wan, Arvin Hekmati, Mingyu Zong, Samiul Alam, Mi Zhang, and Bhaskar Krishnamachari. 2024. IoT in the Era of Generative AI: Vision and Challenges. *arXiv preprint arXiv:2401.01923* (2024).

[264] Yichao Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Wi-Mesh: A WiFi Vision-based Approach for 3D Human Mesh Construction. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[265] Ziqi Wang, Ankur Sarker, Jason Wu, Derek Hua, Gaofeng Dong, Akash Singh, and Mani Srivastava. 2022. Capricorn: Towards Real-time Rich Scene Analysis Using RF-Vision Sensor Fusion. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[266] Mati Wax, Tie-Jun Shan, and Thomas Kailath. 1984. Spatio-temporal spectral analysis by eigenstructure methods. *IEEE transactions on acoustics, speech, and signal processing* 32, 4 (1984).
[267] Jianyu Wei, Ting Cao, Shijie Cao, Shiqi Jiang, Shaowei Fu, Mao Yang, Yanyong Zhang, and Yunxin Liu. 2023. NN-Stretch: Automatic Neural Network Branching for Parallel Inference on Heterogeneous Multi-Processors. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[268] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. AutoDroid: LLM-powered Task Automation in Android *(ACM MobiCom '24)*. New York, NY, USA.
[269] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 2020. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*.
[270] Wentai Wu, Ligang He, Weiwei Lin, Rui Mao, Carsten Maple, and Stephen Jarvis. 2020. SAFA: A semi-asynchronous protocol for fast federated learning with low overhead. *IEEE Trans. Comput.* 70, 5 (2020).
[271] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: Continuous 3D facial reconstruction through lightweight single-ear biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.
[272] Ao Xiao, Yunhao Liu, Yang Li, Feng Qian, Zhenhua Li, Sen Bai, Yao Liu, Tianyin Xu, and Xianlong Xin. 2019. An in-depth study of commercial MVNO: Measurement and optimization. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*.
[273] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.
[274] Rui Xiao, Leqi Zhao, Feng Qian, Lei Yang, and Jinsong Han. 2024. Practical Optical Camera Communication Behind Unseen and Complex Backgrounds. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*.
[275] Binbin Xie, Minhao Cui, Deepak Ganesan, Xiangru Chen, and Jie Xiong. 2023. Boosting the long range sensing potential of lora. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[276] Binbin Xie and Jie Xiong. 2020. Combating interference for long range LoRa sensing. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.
[277] Cong Xie, Sanmi Koyejo, and Indranil Gupta. 2020. Asynchronous Federated Optimization. 1903.03934 [cs.DC] https://arxiv.org/abs/1903.03934
[278] Jiahong Xie, Hao Kong, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Feilong Tang. 2023. mm3DFace: Nonintrusive 3D Facial Reconstruction Leveraging mmWave Signals. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[279] Xiufeng Xie and Kyu-Han Kim. 2019. Source compression with bounded dnn perception loss for iot edge computer vision. In *The 25th Annual International Conference on Mobile Computing and Networking*.
[280] Zhiyuan Xie, Xiaomin Ouyang, Xiaoming Liu, and Guoliang Xing. 2021. UltraDepth: Exposing high-resolution texture from depth cameras. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.
[281] Zhiyuan Xie, Xiaomin Ouyang, Li Pan, Wenrui Lu, Guoliang Xing, and Xiaoming Liu. 2023. Mozart: A Mobile ToF System for Sensing in the Dark through Phase Manipulation. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[282] Sijie Xiong, Sujie Zhu, Yisheng Ji, Binyao Jiang, Xiaohua Tian, Xuesheng Zheng, and Xinbing Wang. 2017. iBlink: Smart glasses for facial paralysis patients. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*.
[283] Chenhan Xu, Tianyu Chen, Huining Li, Alexander Gherardi, Michelle Weng, Zhengxiong Li, and Wenyao Xu. 2022. Hearing Heartbeat from Voice: Towards Next Generation Voice-User Interfaces with Cardiac Sensing Functions. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.
[284] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*.
[285] Daliang Xu, Mengwei Xu, Qipeng Wang, Shangguang Wang, Yun Ma, Kang Huang, Gang Huang, Xin Jin, and Xuanzhe Liu. 2022. Mandheling: Mixed-precision on-device dnn training with dsp offloading. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.
[286] Dongzhu Xu, Anfu Zhou, Guixian Wang, Huanhuan Zhang, Xiangyu Li, Jialiang Pei, and Huadong Ma. 2022. Tutti: Coupling 5G RAN and Mobile Edge Computing for Latency-Critical Video Analytics. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking* (Sydney, NSW, Australia) *(MobiCom '22)*. New York, NY, USA.
[287] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative AI: Making LLMs Comprehend the Physical World *(HOTMOBILE '24)*. New York, NY, USA.

[288] Huatao Xu, Dong Wang, Run Zhao, and Qian Zhang. 2019. FaHo: Deep learning enhanced holographic localization for RFID tags. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*.

[289] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *SenSys '21: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. New York, NY, USA.

[290] Jingao Xu, Hao Cao, Zheng Yang, Longfei Shangguan, Jialin Zhang, Xiaowu He, and Yunhao Liu. 2022. {SwarmMap}: Scaling up real-time collaborative visual {SLAM} at the edge. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*.

[291] Mengwei Xu, Tiantu Xu, Yunxin Liu, and Felix Lin. 2021. Video Analytics with Zero-streaming Cameras. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*.

[292] Mengwei Xu, Xiwen Zhang, Yunxin Liu, Gang Huang, Xuanzhe Liu, and Felix Lin. 2020. Approximate Query Service on Autonomous IoT Cameras. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services* (Toronto, Ontario, Canada) *(MobiSys '20)*. New York, NY, USA.

[293] Mengwei Xu, Mengze Zhu, Yunxin Liu, Felix Lin, and Xuanzhe Liu. 2018. Deepcache: Principled cache for mobile deep vision. In *Proceedings of the 24th annual international conference on mobile computing and networking*.

[294] Ran Xu, Jayoung Lee, Pengcheng Wang, Saurabh Bagchi, Yin Li, and Somali Chaterji. 2022. LiteReconfig: Cost and content aware reconfiguration of video object detection systems for mobile GPUs. In *Proceedings of the Seventeenth European Conference on Computer Systems*.

[295] Ran Xu, Chen lin Zhang, Pengcheng Wang, Jayoung Lee, Subrata Mitra, Somali Chaterji, Yin Li, and Saurabh Bagchi. 2020. ApproxDet: content and contention-aware approximate object detection for mobiles. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[296] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*.

[297] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. 2022. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[298] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*.

[299] Francis Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in situ: a randomized experiment in video streaming. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. Santa Clara, CA.

[300] Kang Yang and Wan Du. 2022. LLDPC: A low-density parity-check coding scheme for LoRa networks. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[301] Qiang Yang and Yuanqing Zheng. 2024. DeepEar: Sound Localization With Binaural Microphones. *IEEE Transactions on Mobile Computing* 23, 1 (2024).

[302] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. 2023. {SLNet}: A Spectrogram Learning Neural Network for Deep Wireless Sensing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*.

[303] Dixi Yao, Liyao Xiang, Zifan Wang, Jiayu Xu, Chao Li, and Xinbing Wang. 2021. Context-aware compilation of dnn training pipelines across edge and cloud. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021).

[304] Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Shengzhong Liu, Huajie Shao, and Tarek Abdelzaher. 2020. Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency. In *Proceedings of the 18th conference on embedded networked sensor systems*.

[305] Shuochao Yao, Yiran Zhao, Huajie Shao, ShengZhong Liu, Dongxin Liu, Lu Su, and Tarek Abdelzaher. 2018. Fastdeepiot: Towards understanding and optimizing neural network execution time on mobile and embedded devices. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*.

[306] Hyunho Yeo, Chan Chong, Youngmok Jung, Juncheol Ye, and Dongsu Han. 2020. NEMO: Enabling Neural-Enhanced Video Streaming on Commodity Mobile Devices. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. New York, NY, USA, Article 28.

[307] Hyunho Yeo, Youngmok Jung, Jaehong Kim, Jinwoo Shin, and Dongsu Han. 2018. Neural Adaptive Content-aware Internet Video Delivery. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*. Carlsbad, CA.

[308] Hyunho Yeo, Hwijoon Lim, Jaehong Kim, Youngmok Jung, Juncheol Ye, and Dongsu Han. 2022. NeuroScaler: Neural Video Enhancement at Scale. In *Proceedings of the ACM SIGCOMM 2022 Conference* (Amsterdam, Netherlands) *(SIGCOMM '22)*. New York, NY, USA.

[309] Juheon Yi, Sunghyun Choi, and Youngki Lee. 2020. EagleEye: Wearable camera-based person identification in crowded urban spaces. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[310] Juheon Yi and Youngki Lee. 2020. Heimdall: mobile GPU coordination platform for augmented reality applications. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. London United Kingdom.

[311] Rongjie Yi, Ting Cao, Ao Zhou, Xiao Ma, Shangguang Wang, and Mengwei Xu. 2023. Boosting DNN Cold Inference on Edge Devices. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services* (Helsinki, Finland) *(MobiSys '23)*. New York, NY, USA.

[312] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, and Wei Gao. 2023. PTEase: Objective Airway Examination for Pulmonary Telemedicine using Commodity Smartphones. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.

[313] Yinggang Yu, Dong Wang, Run Zhao, and Qian Zhang. 2019. RFID based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*.

[314] Mu Yuan, Lan Zhang, Fengxiang He, Xueting Tong, and Xiang-Yang Li. 2022. Infi: end-to-end learnable input filter for resource-efficient mobile-centric inference. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.

[315] Navid Zandi, Awny El-Mohandes, and Rong Zheng. 2022. Individualizing Head-Related Transfer Functions for Binaural Acoustic Applications. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.

[316] Xiao Zeng, Kai Cao, and Mi Zhang. 2017. MobileDeepPill: A small-footprint mobile deep learning system for recognizing unconstrained pill images. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*.

[317] Xiao Zeng, Biyi Fang, Haichen Shen, and Mi Zhang. 2020. Distream: Scaling Live Video Analytics with Workload-Adaptive Distributed Edge Intelligence. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20)*. New York, NY, USA.

[318] Xiao Zeng, Ming Yan, and Mi Zhang. 2021. Mercury: Efficient on-device distributed dnn training via stochastic importance sampling. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[319] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. 2020. MultiSense: Enabling multi-person respiration sensing with commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020).

[320] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019).

[321] Anlan Zhang, Chendong Wang, Bo Han, and Feng Qian. 2022. YuZu: Neural-Enhanced Volumetric Video Streaming. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. Renton, WA.

[322] Ben Zhang, Xin Jin, Sylvia Ratnasamy, John Wawrzynek, and Edward Lee. 2018. AWStream: Adaptive Wide-Area Streaming Analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. New York, NY, USA.

[323] Chaoyun Zhang, Marco Fiore, Cezary Ziemlicki, and Paul Patras. 2020. Microscope: mobile service traffic decomposition for network slicing as a service. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[324] Chaoyun Zhang, Paul Patras, and Hamed Haddadi. 2019. Deep Learning in Mobile and Wireless Networking: A Survey. *IEEE Communications Surveys & Tutorials* 21, 3 (2019).

[325] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. 2019. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *The 25th annual International Conference on Mobile Computing and Networking*.

[326] Huanhuan Zhang, Anfu Zhou, Yuhan Hu, Chaoyue Li, Guangping Wang, Xinyu Zhang, Huadong Ma, Leilei Wu, Aiyun Chen, and Changhui Wu. 2021. Loki: Improving Long Tail Performance of Learning-Based Real-Time Video Adaptation by Fusing Rule-Based Models. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) *(MobiCom '21)*. New York, NY, USA.

[327] Huanhuan Zhang, Anfu Zhou, Jiamin Lu, Ruoxuan Ma, Yuhan Hu, Cong Li, Xinyu Zhang, Huadong Ma, and Xiaojiang Chen. 2020. OnRL: Improving Mobile Video Telephony via Online Reinforcement Learning. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. New York, NY, USA, Article 29.

[328] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th annual international conference on mobile computing and networking*.

[329] Jing Zhang and Dacheng Tao. 2021. Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet of Things Journal* 8, 10 (May 2021).

[330] Jinrui Zhang, Deyu Zhang, Xiaohui Xu, Fucheng Jia, Yunxin Liu, Xuanzhe Liu, Ju Ren, and Yaoxue Zhang. 2020. MobiPose: Real-time multi-person pose estimation on mobile devices. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[331] Tuo Zhang, Tiantian Feng, Samiul Alam, Dimitrios Dimitriadis, Mi Zhang, Shrikanth S Narayanan, and Salman Avestimehr. 2023. Gpt-fl: Generative pre-trained model-assisted federated learning. *arXiv preprint arXiv:2306.02210* (2023).

[332] Tuo Zhang, Tiantian Feng, Samiul Alam, Sunwoo Lee, Mi Zhang, Shrikanth S Narayanan, and Salman Avestimehr. 2023. Fedaudio: A federated learning benchmark for audio tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[333] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and Salman Avestimehr. 2022. Federated Learning for Internet of Things: Applications, Challenges, and Opportunities. *IEEE Internet of Things Magazine (IEEE IoTM)* (2022).
[334] Tuo Zhang, Lei Gao, Sunwoo Lee, Mi Zhang, and Salman Avestimehr. 2023. TimelyFL: Heterogeneity-aware Asynchronous Federated Learning with Adaptive Partial Training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
[335] Tianfang Zhang, Cong Shi, Payton Walker, Zhengkun Ye, Yan Wang, Nitesh Saxena, and Yingying Chen. 2023. Passive Vital Sign Monitoring via Facial Vibrations Leveraging AR/VR Headsets. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[336] Wuyang Zhang, Zhezhi He, Luyang Liu, Zhenhua Jia, Yunxin Liu, Marco Gruteser, Dipankar Raychaudhuri, and Yanyong Zhang. 2021. Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.
[337] Xiao Zhang, Hanqing Guo, James Mariani, and Li Xiao. 2022. U-star: An underwater navigation system based on passive 3d optical identification tags. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*.
[338] Xiao Zhang, Griffin Klevering, Juexing Wang, Li Xiao, and Tianxing Li. 2023. RoFin: 3D Hand Pose Reconstructing via 2D Rolling Fingertips. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*.
[339] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. 2022. Synthesized Millimeter-Waves for Human Motion Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.
[340] Xu Zhang, Yiyang Ou, Siddhartha Sen, and Junchen Jiang. 2021. SENSEI: Aligning Video Streaming Quality with Dynamic User Sensitivity. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*.
[341] Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y. Guo, Feng Qian, and Z. Mao. 2021. EMP: edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. New Orleans Louisiana.
[342] Yu Zhang, Tao Gu, and Xi Zhang. 2020. MDLdroidLite: A release-and-inhibit control approach to resource-efficient deep neural networks on mobile devices. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.
[343] Yu Zhang, Tao Gu, and Xi Zhang. 2021. MDLdroid: A ChainSGD-reduce approach to mobile deep learning for personal mobile sensing. *IEEE/ACM Transactions on Networking* 30, 1 (2021).
[344] Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. 2023. CMA: Cross-Modal Association Between Wearable and Structural Vibration Signal Segments for Indoor Occupant Sensing. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.
[345] Ziyang Zhang, Huan Li, Yang Zhao, Changyao Lin, and Jie Liu. 2023. POS: An Operator Scheduling Framework for Multi-model Inference on Edge Intelligent Computing. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*.
[346] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. 2023. Nerf2: Neural radio-frequency radiance fields. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*.
[347] Xiaopeng Zhao, Guosheng Wang, Zhenlin An, Qingrui Pan, and Lei Yang. 2024. Understanding Localization by a Tailored GPT. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*.
[348] Yiqin Zhao and Tian Guo. 2021. Xihe: A 3D Vision-based Lighting Estimation Framework for Mobile Augmented Reality. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*.
[349] Xiaolong Zheng, Zhichao Cao, Jiliang Wang, Yuan He, and Yunhao Liu. 2014. Zisense: towards interference resilient duty cycling in wireless sensor networks. In *Proceedings of ACM SenSys*.
[350] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*.
[351] Anfu Zhou, Huanhuan Zhang, Guangyuan Su, Leilei Wu, Ruoxuan Ma, Zhen Meng, Xinyu Zhang, Xiufeng Xie, Huadong Ma, and Xiaojiang Chen. 2019. Learning to Coordinate Video Codec with Transport Protocol for Mobile Video Telephony. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. New York, NY, USA, Article 29.
[352] Qihua Zhou, Song Guo, Zhihao Qu, Jingcai Guo, Zhenda Xu, Jiewei Zhang, Tao Guo, Boyuan Luo, and Jingren Zhou. 2021. Octo:{INT8} training with loss-aware compensation and backward quantization for tiny on-device learning. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*.
[353] Yun Zhu, Yinxiao Liu, Felix Stahlberg, Shankar Kumar, Yu hui Chen, Liangchen Luo, Lei Shu, Renjie Liu, Jindong Chen, and Lei Meng. 2023. Towards an on-device agent for text rewriting. *arXiv preprint arXiv:2308.11807* (2023).