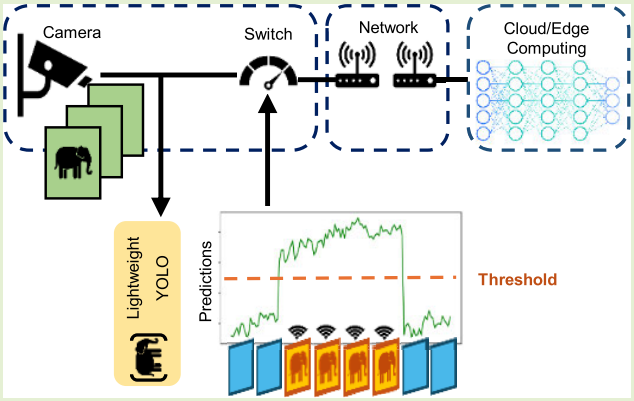
Intelligent Sensing Framework: Near-Sensor Machine Learning for Efficient Data Transmission

Wenjun Huang, Arghavan Rezvani, Hanning Chen, *Graduate Student Member, IEEE*, Yang Ni, Sanggeon Yun, Sungheon Jeong, Guangyi Zhang, and Mohsen Imani

***Abstract*—Applications in the Internet of Things (IoT) utilize machine learning (ML) to analyze sensor-generated data. However, a major challenge lies in the lack of targeted intelligence in current sensing systems, leading to vast data generation and increased computational and communication costs. To address this challenge, we propose a novel sensing framework to equip sensing systems with intelligent data transmission capabilities by integrating a highly efficient ML model placed near the sensor. This model provides prompt feedback for the sensing system to transmit only valuable data while discarding irrelevant information by regulating the frequency of data transmission. The near-sensor model is quantized and optimized for real-time sensor control. To enhance the framework's performance, the training process is customized, and a "lazy" sensor deactivation strategy utilizing temporal information is introduced. The suggested framework is orthogonal to other IoT frameworks and can be considered as a plug-in for selective data transmission. The framework is implemented, encompassing both software and hardware components. The experiments demonstrate that the framework utilizing the suggested module achieves over 85% system efficiency in terms of energy consumption and storage, with negligible impact on performance. This framework has the potential to significantly reduce data output from sensors, benefiting a wide range of IoT applications.**



***Index Terms*—Energy efficiency, intelligent sensing, Internet of Things (IoT), machine learning (ML), near-sensor computing.**

I. Introduction

THE prevalence of ubiquitous sensors is currently experiencing an exponential surge, both in terms of their quantity and the vast amount of data they generate. Despite the rapid growth, existing approaches to sensor data processing and transmission cannot keep pace due to their algorithmic and architectural limitations [1]. In numerous Internet of Things (IoT) applications, data collected by sensors are analyzed using machine learning (ML) models [2], [3], [4], [5]. As the volume of data continues to grow, many applications opt to send the data to more computationally powerful nodes, such as edge or cloud computing nodes, to execute the learning algorithms. In either scenario, a large volume of data is transmitted at a high rate to ensure that all necessary information is captured and processed for various tasks. The significant amount of data conveyed in both scenarios places high demands on energy and storage resources, resulting in considerable resource pressure and wastage [6]. This is especially problematic for applications that require a relatively complex and expensive ML model. Fig. 1 depicts a typical IoT system for video monitoring systems, where dense data generated by the camera is continuously analyzed using complex ML models. In the system, visual signals captured by surveillance cameras are transmitted continuously to a costly ML model, which may be hosted on a central server, such as a cloud or edge computing node. Depending on the intended purposes, the ML

Manuscript received 5 July 2024; accepted 3 August 2024. Date of publication 15 August 2024; date of current version 31 October 2024. This work was supported in part by DARPA Young Faculty Award, National Science Foundation (NSF) under Grant 2127780, Grant 2319198, Grant 2321840, Grant 2312517, and Grant 2235472; in part by the Semiconductor Research Corporation (SRC), Office of Naval Research through the Young Investigator Program under Award N00014-21-1-2225 and Award N00014-22-1-2067; and in part by the Air Force Office of Scientific Research under Award FA9550-22-1-0253, along with generous gifts from Xilinx and Cisco. The associate editor coordinating the review of this article and approving it for publication was Dr. Te Han. *(Wenjun Huang and Arghavan Rezvani contributed equally to this work.) (Corresponding author: Mohsen Imani.)*

The authors are with the Department of Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: m.imani@uci.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSEN.2024.3440988, provided by the authors.

Digital Object Identifier 10.1109/JSEN.2024.3440988

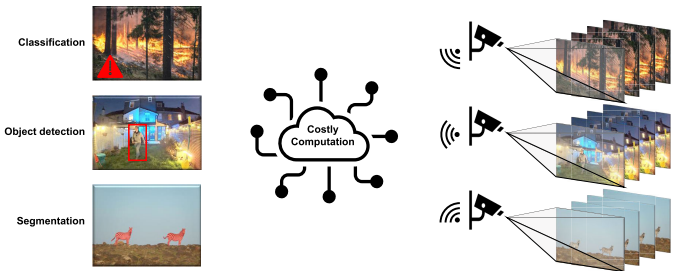


Fig. 1. Application scenarios of an intelligent system.

model performs various tasks, including but not limited to classification, object detection, and segmentation [7], [8].

Many studies attempted to alleviate the energy and storage pressures in IoT applications from multiple perspectives, e.g., computing offloading, resource allocation, and so on. Traditional methods have shown substantial progress in tackling these issues. Certain research efforts leveraged the Lyapunov optimization algorithm [9] to identify the optimal decision [10]. Others framed resource allocation and computing offloading as optimization challenges [11], [12], [13], [14], [15]. However, these approaches exhibit certain limitations. First, they require knowledge of the underlying model, which proves challenging due to the intricate and dynamic nature of IoT systems. Second, they are vulnerable to being stuck at local optima. Some research [16], [17], [18], [19], [20] have introduced intelligent offloading strategies grounded in deep learning (DL). Furthermore, some research have placed emphasis on the optimization of hardware structures, thereby enhancing the efficiency of edge computing [21], [22], [23].

Different from the work above, which uses ML/DL algorithms to automate offloading and resource allocation, some research proposes solutions to reduce data generated by the sensor. For example, in the realm of computer vision, analyze-then-compress (ATC) approaches present an alternative strategy in which front-end devices extract and transmit features to a central server. Depending on the specific scenario in which it is being applied, ATC approaches utilize a variety of traditional feature extraction algorithms, ranging from handcrafted methods (e.g., [24], [25], [26]) to information theory-based methods [27], [28]. In recent years, more advanced deep learning-based methods have garnered significant attention. Several early layers of DNN are deployed on the front-end devices for extracting highly compact and representative features. In the face recognition task, for example, the face of an individual can be represented by features with several hundred dimensions [29], [30], [31]. By representing data in such features, the amount of data that needs to be transmitted can be significantly decreased. In addition, only a few lightweight operations are required to be performed on the central server.

However, a notable limitation of DNN-based ATC methods is their restricted capacity for generalization. Given the meticulous design of DNN architectures, the features they extract and transmit to the central server are often highly abstract and tailored specifically to the intended task. However, visual signal carrying pertinent information typically undergoes a sequence of downstream tasks for comprehensive analysis. Consequently, the inherent challenge arises from the deficiency in generalization, rendering it difficult to design a backbone network capable of extracting features suitable for all such tasks. Moreover, in numerous scenarios, it becomes useful to retain visual signals for subsequent analysis or future reference. The transmission of excessively abstract features significantly complicates the process of reconstructing the original visual signal on the server side. Although front-end devices possess the capability to store the original signals, their constrained storage capacity poses a challenge.

In addition, all the efforts mentioned above, whether from an IoT or ML perspective, still need to process all the data generated from the sensor, neglecting the fact that in many IoT applications (e.g., fire alarm, wildlife monitoring, crime surveillance [32], and healthcare [33]), only a small fraction of sensor activity typically contains valuable information. Hence, it is unnecessary to run a costly service, such as a large-scale DNN model, that handles a continuous and complete stream of sensor data, whether on the edge or in the cloud. This is because the service specifically targets only that small fraction of valuable data, yet it still requires processing substantial amounts of irrelevant information.

Spiking neural networks (SNNs) and event cameras, on the other hand, generate data only when there is a change in the scene, reducing the amount of data needed for transmission. However, in a static scene, an event camera would barely generate any data, effectively rendering it blind to stationary information. This limitation restricts its applicability, particularly for tasks involving slowly moving objects. Moreover, the spatial resolution of event cameras is generally lower compared to high-resolution frame-based cameras, which can be a limiting factor for applications that require detailed spatial information. Event cameras can also be sensitive to noise, especially in low-light conditions, resulting in spurious events that add complexity to the data processing. Last but not least, the price of event cameras is generally higher than that of traditional RGB cameras, which can limit their applicability for widespread deployment.

Observing the limitations of the approaches previously discussed, in this article, we rethink and redesign the sensing system, proposing a new framework that is orthogonal to previous research directions. Rather than reducing the data representation or determining where and how data should be relocated for service execution, our framework focuses on reducing the amount of data sent out from the sensor side by identifying valuable information. Our framework, acting as a "filter," can be applied before any aforementioned approaches, and easily be integrated into any system as a plug-in.

Our proposed framework consists of a few components. First, we deploy a lightweight model near the sensor to detect whether a frame contains useful information, which we refer to as a frame of interest (FOI) and only send out those FOIs. The model helps mitigate the huge amount of unnecessary analysis of costly ML models over the central server. Although this process can also be deployed before the costly ML models at the same place, our near-sensor model offers substantial savings in transmission costs, encompassing energy, bandwidth, and more. To enable intelligent sensing, the near-sensor

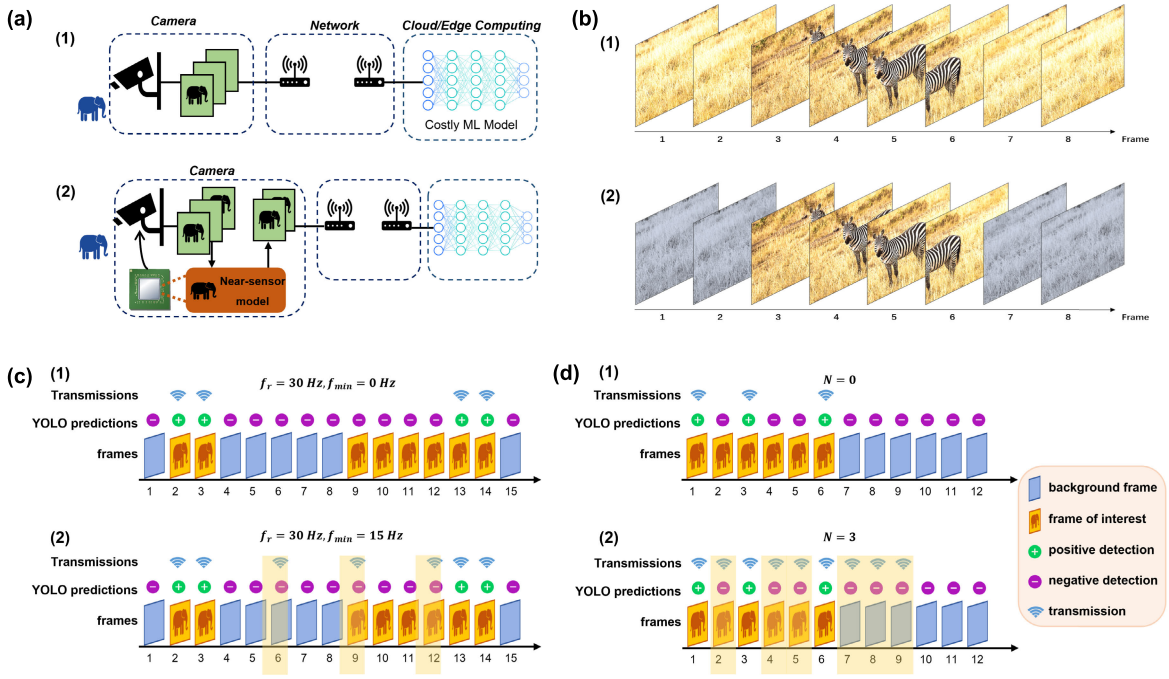


Fig. 2. Motivation and design of our proposed intelligent sensing module. (a) General system framework of conventional systems and our system. (b) Visualization of the data transmission in our system. (c) Illustration of minimum data transmission frequency (denoted by f_{min}) in our system. f_r denotes the camera's refresh rate. (d) Illustration of lazy sensor deactivation scheme in our system, N is the number for deactivation count.

model should be fast enough to process frames in real-time and provide feedback. With the help of this feedback, our framework produces selective and sparse data. Furthermore, we enhance the overall performance of the framework by introducing several effective schemes to mitigate potential misdetections of the lightweight model, which we explore in Section II.

In this work, we describe the following contributions.

1) We propose a new framework that improves IoT system energy and storage efficiency orthogonal to the previous approaches. It can be readily inserted into any existing system, serving as an intelligent data generation "filter." We call the sensor exploiting this framework an "intelligent sensor" in the rest of the article.
2) To illustrate our framework, we design a modified DNN model tailored to near-sensor computing.
3) We introduce schemes for alleviating possible misdetections of the near-sensor model, including nonzero minimum transmission frequency and lazy deactivation. We also conduct a thorough investigation into their impact on overall system performance.
4) We implement the framework encompassing both software and hardware components. Our experiments demonstrate that utilizing our intelligent framework leads to a substantial reduction in energy and storage consumption in sensing systems.

II. Methods

A. Framework Overview

Fig. 2(a) illustrates the framework of a conventional system and our framework. In Fig. 2(a-1), the conventional sensor captures and transmits all the frames to the costly models, regardless of the presence of useful information in the frames. On the contrary, the intelligent sensor equipped with our framework, as shown in Fig. 2(a-2), utilizes a lightweight model near the sensor to detect and control the FOI transmission. The model is deployed on an edge computing device integrated into the camera, connecting to the image sensor. Specifically, the camera captures a continuous stream of frames, which are then fed to the lightweight model for real-time predictions. With the presence of FOI, the camera raises the data transmission frequency, and the frames are transmitted to the central server for more sophisticated operations; if the frame is detected as background (with no interest), the camera will turn off the data transmission. Fig. 2(b) provides a visualized example, where the transmitted frames are presented in color while the discarded frames are shaded in gray. The system adopting our framework, as demonstrated in Fig. 2(b-2), outperforms conventional systems depicted in Fig. 2(b-1) by exclusively transmitting frames containing a zebra, resulting in a reduction of storage and energy consumption by half in this particular instance.

This is because transmitting only the necessary FOIs to the central server reduces the number of inferences needed by the complex ML model on the central server, which is the primary source of energy consumption. This reduction is achieved while introducing only a negligible energy overhead associated with the near-sensor model. This is in contrast to previous approaches that would transmit all frames to the server based on the camera's refresh rate, resulting in significant energy waste due to performing inference on numerous unnecessary frames.

In this work, we concentrate on the effect of our proposed framework on energy consumption reduction. Each

element in the framework is elaborated on in the following sections.

B. Near-Sensor Model

The near-sensor model is tasked with distinguishing FOIs from all other frames. One way to tackle this problem is by using a classifier. However, the frames captured by a sensor may contain multiple objects of interest with varying scales and positions, while classifiers are typically trained on images that contain a single, centered object (such as those found in CIFAR-10, CIFAR-100 [34], and ImageNet [35]). These classifiers have limitations in detecting multiple objects with varying scales and positions. As a result, a deep object detection model is often employed instead. Among different object detection models, YOLO [36], a single-stage detector, is selected. Compared with two-shot detectors (e.g., R-CNN, Fast R-CNN, Faster R-CNN, and R-FCN [37], [38], [39], [40]), YOLO is lightweight, faster, and with comparable accuracy in a suitable scenario [41], [42]. These features make YOLO a good candidate for being embedded into the sensor.

The output layer of YOLO contains bounding box predictions concatenated to the class prediction and objectness confidence. However, the goal of our intelligent sensor is to detect the existence of objects of interest, regardless of their position in the frame. Therefore, we can only keep the objectness confidence in the output, which can be used further to determine FOI. We set a threshold for the objectness confidence, and only the frames with confidence exceeding this threshold are transmitted. As increasing the threshold, the detection becomes stricter, resulting in fewer frames being considered FOI. Our framework's definition allows us to customize the YOLO model in the following ways.

1) Model Optimization: The architectures of the YOLO series contain several repetitive blocks. Although these blocks contribute to the model capacity, they make the model power-hungry and slower during inference. For example, the YOLOv5 model family has five variations: x-large, large, medium, small, and nano. While each model shares the same structure, they differ in the network's depth and the number of filters in different layers (width). Since our model does not predict bounding boxes, we can modify its depth and width to create a more lightweight model that still achieves comparable performance on our task. In other words, in contrast to the YOLO model, which predicts both the class and location of an object, our proposed near-sensor model requires only class prediction. This simplification allows us to reduce the number of model parameters without compromising the accuracy of the class prediction task.

In our experiments, we utilized three YOLO-based models, namely, YOLOv5n, YOLOv5nm, and YOLOv5ns. YOLOv5n stands for Yolov5 nano, which is the smallest introduced YOLOv5 model. By modifying the depth and width of the YOLOv5n, we achieved more lightweight models, which we called nano-medium (YOLOv5nm) and nano-small (YOLOv5ns). In YOLOv5nm, the depth and width are half of the depth and width of the YOLOv5n, and in YOLOv5ns, this ratio is one-third for depth and one-fourth for width.

As mentioned earlier, the output of the YOLO model not only contains confidence but also concatenates the bounding box information, which is not required in our proposed module. Consequently, we can remove the part of the model associated with the bounding box during the inference to reduce the model size.

2) Inference Simplification: YOLO utilizes a nonmax suppression (NMS) algorithm as the final step to pick the most appropriate bounding box for the object among all of the predicted boxes for that specific object. The NMS algorithm starts with selecting the box with the highest objectness score among all, removing all the boxes with high overlap with the selected box, and repeating these steps iteratively. However, since the sensing scenario does not require bounding boxes, we can simplify this step. Instead of running the NMS algorithm, we only keep the highest objectness confidence. If there is one confidence value greater than the threshold, it indicates the presence of at least one object in the prediction. Therefore, by solely comparing the highest confidence value with the threshold, we can achieve comparable performance, resulting in reduced inference time.

3) Model Quantization and Loss Function Customization: Model quantization is another well-known approach to accelerating model inference. It involves using fewer bits to store model parameters while maintaining nearly the same level of accuracy [43], [44], [45], [46]. Aggressive quantization leads to a highly lightweight model, but at the cost of reduced accuracy compared to the original model. On the other hand, less aggressive quantized models experience minimal accuracy loss, but they are not as lightweight as aggressively quantized models [47]. The amount of tolerable accuracy loss varies across different tasks. For this work, we utilized the **kmeans-lut** quantization which is a Look-up-table (LUT) based quantization [48], where LUT is generated by k-means clustering.

Moreover, refining the loss function can enhance the performance of the model when subjected to intensive quantization. The conventional YOLOv5 has three loss terms

$$L = l_{\text{obj}} + l_{\text{cls}} + l_{\text{bbox}} \tag{1}$$

where l_{obj}, l_{cls}, and l_{bbox} are objectness confidence loss, classification loss, and bounding box loss, respectively. Among the loss terms, reducing the l_{obj} and l_{cls} loss terms contributes to accurate object detection and classification, resulting in improved performance of our model. In contrast, the l_{bbox} loss term, which corresponds to the precise bounding box position, has a negative impact on our model. This is because it forces the model to make a compromise during the gradient descent search, making it more difficult for the model to converge to the optimal. By removing the l_{bbox} term, our near-sensor model can prioritize the detection of FOIs without considering the bounding box generation, enabling the model to achieve a higher degree of quantization while maintaining a comparable level of accuracy. Therefore, the following loss function was adapted for training the near-sensor model with a faster convergence to improve accuracy in our task

$$L = l_{\text{obj}} + l_{\text{cls}}. \tag{2}$$

C. Data Transmission Frequency

The prediction of the near-sensor model regulates the frequency of data transmission, thereby reducing the volume of data transmitted to the central server. If the camera records FOIs, it should be configured to transmit all FOIs to the server, with a frequency equivalent to the camera's refresh rate. Conversely, when the camera captures background frames, it should lower the data transmission rate to save energy. This reduced frequency is referred to as the minimum transmission frequency. The minimum transmission frequency can vary between zero and the camera's refresh rate. If the minimum transmission frequency is set to match the camera's refresh rate, all frames captured by the camera are forwarded to the server, indicating that the transmission is unaffected by the predictions of the lightweight model. In this scenario, the volume of data transmitted to the server is identical to that of conventional systems. Conversely, when the minimum transmission frequency is set to zero, any frames identified as background frames would not be transmitted to the server. Fig. 2(c-1) demonstrates the data transmission of our framework. The blue frames represent the background and the yellow frames with an elephant depict FOIs. A positive or negative sign is used to present the near-sensor model predictions. The frames that are marked with a positive sign represent the prediction of FOIs. The frames being transmitted to the server are indicated by the Wi-Fi icon. Only the frames that are recognized as FOI (frames 2, 3, 13, and 14), are transmitted.

However, even though the lightweight model displays a high level of accuracy, it is still inevitable to misdetect some FOIs as background frames, and these misdetected frames are all discarded when the minimum transmission frequency is zero since the data transmission is completely halted. This wrong discard can be alleviated by increasing the minimum transmission frequency, which means that even if the frames are detected as background frames, they are still transmitted to the server regularly at a lower nonzero frequency. An example demonstrating the effect of increasing the minimum transmission frequency is shown in Fig. 2(c-2). When the minimum transmission frequency equals zero [see Fig. 2(c-1)], all the frames detected as background are discarded by the intelligent sensor. This significantly reduces the amount of data transmitted while also losing some useful information (e.g., frames 9–12). To reduce the number of missing FOIs, we increased the transmission frequency in Fig. 2(c-2). In the figure, the camera's refresh rate $f_r = 30$ Hz, and the minimum transmission frequency $f_{\min}$ is set to $f_r/2$ (i.e., 15 Hz). Under this setting, even if the transmission frequency is tuned down, the sensor would also send one frame every two frames. From the figure, we can observe that although the prediction of the lightweight model maintains the same, we transmit more FOIs to the server (frames 9 and 12).

D. Lazy Sensor Deactivation

Since FOIs contain valuable information, in this work, the priority is given to transmitting all FOIs rather than mistakenly transmitting a background frame. Therefore, we define misdetections as the FOIs which are not transmitted. Considering the fact that the frames in a video have temporal correlation, we assume that if the camera captures an FOI, the following frame is likely to be an FOI as well. Thus, in order to reduce the misdetection of FOIs, inspired by [49], we proposed a scheme for lazy sensor deactivation, which considers the detection results of neighboring frames. However, unlike the work in [49] which schedules observation points over the target execution, our scheme entails monitoring the number of consecutive background frames detected by the near-sensor model. The camera maintains a high transmission frequency until the count (C_1) of consecutive background detection reaches a predefined number (N). Once the number is met, the camera tunes down the transmission frequency and resets the count. The count is reset to zero whenever an FOI is identified. The adoption of our lazy sensor deactivation scheme enables the detector to rectify the misdetection of a single frame by utilizing the adjacent frame's information. In comparison to the detector without the lazy sensor deactivation scheme, utilizing our approach preserves more FOIs since an occasional misdetection cannot affect the transmission frequency. Decisions for tuning the transmission frequency are made based on a few adjacent frames. Fig. 2(d) provides an example that demonstrates the advantages of implementing lazy sensor deactivation. In this example, the value of N is set to 3. When compared to the system that does not utilize lazy deactivation [shown in Fig. 2(d-1)], the implementation of lazy deactivation [see Fig. 2(d-2)] also transmits the FOIs that are misdetected by the near-sensor model, as demonstrated by the transmission of frames 2, 4, and 5.

The utilization of the lazy sensor deactivation scheme incurs two costs from a storage perspective. The first cost arises when a negative sample is mistakenly identified as an FOI, leading to the reset and restart of the count. In the worst case scenario, a single negative sample misdetection results in storing $2N$ additional frames. Nonetheless, this cost is acceptable as our primary concern lies in preserving the completeness of FOIs. The inclusion of a few extra negative samples following FOIs does not influence the pertinent information we aim to retain. Moreover, given that N is not an excessively large value, our storage capacity can handle these rare occurrences.

The second cost inherent in the lazy sensor deactivation scheme manifests in the recovery of the transmission frequency to a high level and the subsequent repetition of counting following each period of frequency decrease. Given that, a large proportion of frames comprise background and such frames often appear in the form of segments, the frames following a low transmission frequency period are more likely to be background as well. As a result, in a long sequence of background frames, the detector stores N more frames after each low transmission frequency period. To mitigate the redundancy following each period, we introduced one more count (C_2) to monitor the number of consecutive low transmission frequency periods. This count is used to calculate N_{new} for consecutive background frames

$$N_{\text{new}} = \max\left(1, \frac{N}{2^{C_2}}\right). \tag{3}$$

Upon tuning down the transmission frequency, C_2 increments by 1. However, the detection of an FOI interrupts the consecutive low transmission frequency periods, resetting the count C_2 to zero. At the start of each period, (3) determines the number (N_{new}) for that particular period. Using the count C_2 for consecutive low transmission frequency periods gradually decreases the threshold from N to 1 in the long run, leading to greater storage and energy savings than the vanilla scheme.

Algorithm 1 outlines the pseudocode of the proposed scheme, which incorporates minimum transmission frequency and lazy sensor deactivation. The pseudocodes 6–12 indicate the code for lazy sensor deactivation and 14–20 indicate the code for minimum transmission frequency, where f_r is the camera's refresh rate, f_{min} is the minimum transmission frequency, and C_3 is a count used to determine whether a frame should be transmitted in a minimum transmission frequency period.

E. Dual-Camera Collaboration

Recording and analyzing valuable information necessitate the utilization of high-resolution images, thereby engendering a predilection for high-resolution cameras. Nevertheless, the sustained operation of such cameras for near-sensor computing proves to be energy burdensome, given their elevated power consumption. Our objective is to furnish dependable performance while concurrently minimizing energy consumption. Therefore, we integrated an additional low-resolution, and power-efficient, camera into the sensor configuration.

During periods devoid of FOIs, the high-resolution camera remains inactive to conserve power, while the low-resolution camera is engaged in executing the near-sensor model, as elucidated in Section II-B. When an FOI is detected by the near-sensor model utilizing the low-resolution camera, the high-resolution camera is activated, capturing and subsequently transmitting the pertinent frames. During this phase, the low-resolution camera is deactivated, given the superior quality of the frames obtained by the high-resolution camera.

The power consumption breakdown of using dual-camera collaboration on the sensor side is depicted in Fig. 3. The incorporation of a power-efficient low-resolution camera results in significant power savings (highlighted in shadow), even compared to using our module with a single high-resolution camera. Given the infrequent occurrence of FOIs, our dual-camera collaboration scheme proves to be highly effective in mitigating energy consumption at the sensor side over an extended duration.

Note that our dual-camera collaboration requires an additional low-resolution camera, which increases the cost per device. The price of an image sensor depends on several factors, such as resolution, sensor size, technology, dynamic range, and noise suppression. While a basic CMOS sensor might cost a few dollars, high-end specialized sensors can be significantly more expensive, potentially reaching thousands of dollars. In our system, a basic sensor is sufficient for FoI detection, resulting in an increase of only 1/1000 to 1/100 over the original expenses. On the other hand, these basic sensors support at least 30 frames per second (frames/s), which is sufficient for our task and does not impact efficiency.

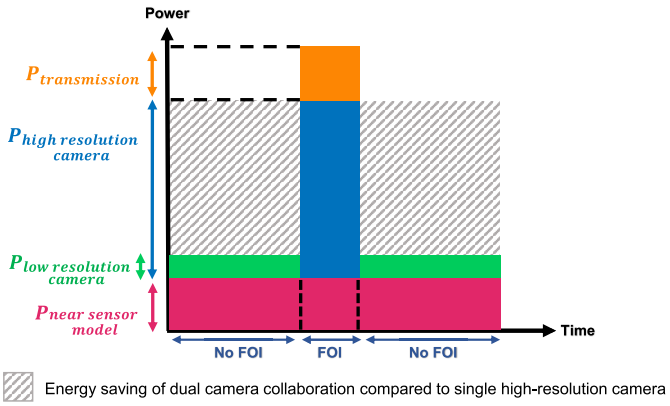


Fig. 3. Energy consumption breakdown on the sensor for the framework utilizing dual-camera collaboration.

III. Experiments

A. Experimental Setup

In this work, we trained and evaluated our framework in the context of animal detection using the Microsoft Common Objects in Context (MS COCO) dataset [50], which is widely used for object detection tasks. In this context, the images in the dataset were selected and relabeled. The images containing at least one object belonging to the animal category are considered FOI and are labeled 1. The remaining frames are considered background and labeled as 0. The near-sensor lightweight model detects and transmits FOIs while filtering out the background frames. The detected frames are transmitted to a more sophisticated model, in our case a well-trained Fast R-CNN model, to perform advanced operations. The framework is implemented using PyTorch [51]. In accordance with the scenario, we ordered the data in the testset with a specific logic: FOIs and background frames are presented in a fragmented manner, appearing consecutively and alternating with each other. The frames in fragments are ordered randomly.

B. Parameter Evaluation Metrics

The goal of the framework is to detect the animals in all FOIs while minimizing the system's energy consumption and occupying minimal storage. In essence, our module aims to minimize the misdetection rate, defined as the fraction of FOIs that are not detected by our near-sensor model

$$P_{\text{miss}} = \frac{n_{\text{miss}}}{n_{\text{FOI}}} \tag{4}$$

where n_{miss} is the number of missed FOIs by the near-sensor model and n_{FOI} is the total number of FOIs in the stream.

In addition, we prioritized the percentage of transmission reduction achieved by our module in comparison to sending all frames captured by the camera, which is defined as

$$P_{\text{trans}} = \frac{n_{\text{trans}}}{n_{\text{frames}}} \tag{5}$$

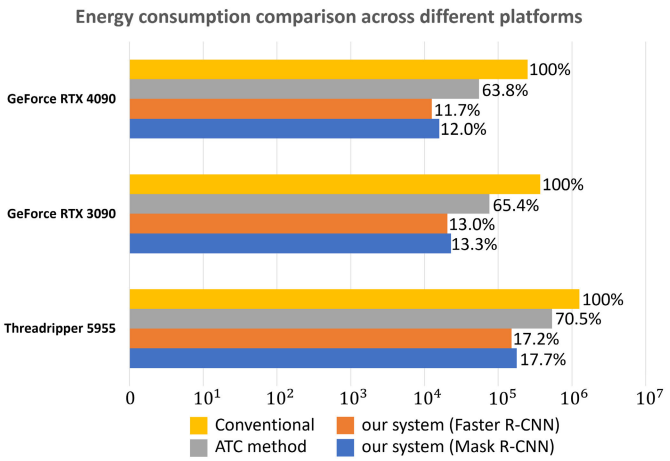


Fig. 4. Energy consumption of the baselines and the system adopting our framework. The experiments are conducted with $M = 20$ (the total number of the frames $n_{\text{total}} = 21336$, $P_{\text{miss}} = 3\% \pm 0.6\%$). The conventional system implements Fast R-CNN on the server.

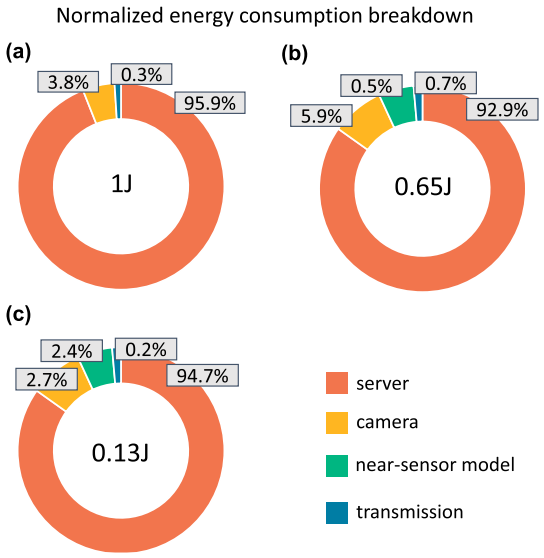


Fig. 5. Normalized energy consumption breakdown. (a) Conventional system. (b) ATC method. (c) System with our framework.

where n_{trans} is the number of transmitted frames and n_{frames} is the total number of frames in the testset. The energy consumption of the system, including transmission and inference energy of the Fast R-CNN model, is closely related to the number of transmitted FOIs; thus the percentage of transmission reduction serves as a key indicator of the effectiveness of our framework. In addition to impacting energy consumption, P_{trans} also reflects the amount of storage that can be saved on the server.

It is worth emphasizing that our framework involves tradeoffs among its various parameters. Altering the values of these parameters can lead to different performances with respect to missed detection frames and the percentage of transmission reduction. For instance, the most extreme scenario is to maintain the transmission frequency equal to the camera's refresh rate and transmit all frames to the server. Although this setting would result in zero missed detection frames, it would also lead to the highest possible energy consumption. In the following, we analyze the influence of each parameter on the performance of our framework utilizing the proposed module and discuss the tradeoffs between these parameters.

This study explores the impact of four key parameters on our system's performance.

1) The confidence threshold (T) of YOLO.
2) The ratio (M) of the number of background frames to the number of FOIs.
3) Minimum transmission frequency (f_{min}).
4) The count (N) at which the sensor deactivates.

C. Results

The comparison of the conventional system, the ATC method, and the system with our framework are shown in Fig. 4 (the x-axis is illustrated in log scale). For the system with our framework, we implemented two models (i.e., Mask R-CNN [52] and Faster R-CNN) on the server, sorted in descending order of model complexity. For the conventional system, we used the least complex model, i.e., Fast R-CNN. For the ATC method, we implemented Mask R-CNN by adopting the approach described in [53]. This is because, in ATC methods, all downstream tasks rely on the same abstract features, requiring the complex server-side model to perform inference for all tasks. The energy consumptions are measured on three platforms: GeForce RTX 4090, GeForce RTX 3090, and AMD Threadripper 5955.

The energy consumption of conventional systems comprises three components: energy consumption by the sensor, transmission energy, and inference energy consumed at the server. Both the ATC method and the system with our framework introduce an additional component: near-sensor model energy consumption. Despite this additional energy consumption, both the ATC method and our framework reduce the overall system energy consumption. This reduction is achieved because both methods perform preliminary processing on low-power devices, which facilitates subsequent inference. As shown in the figure, our proposed framework helps consume less than 18% energy in all settings compared to the conventional system, and less than 25% energy compared to the ATC method.

Although the ATC method also employs a near-sensor model as a feature extractor, it transmits the features of all frames to the server for inference. In contrast to the ATC approaches, our framework does not transmit abstract features of frames but only transmits the original frames containing useful information, therefore, significantly reducing the number of frames sent to the server. While the size of data transmitted per frame may be larger compared to the ATC approaches, our framework involves the transmission of fewer frames, ultimately resulting in a reduction in the overall amount of data transmitted. In addition, the complex ML model on the server in the ATC method only has access to abstract features, therefore, the downstream tasks, such as object detection and segmentation, rely on a single model. However, our near-sensor model sends the original FoIs to the server, allowing for the deployment of complex models

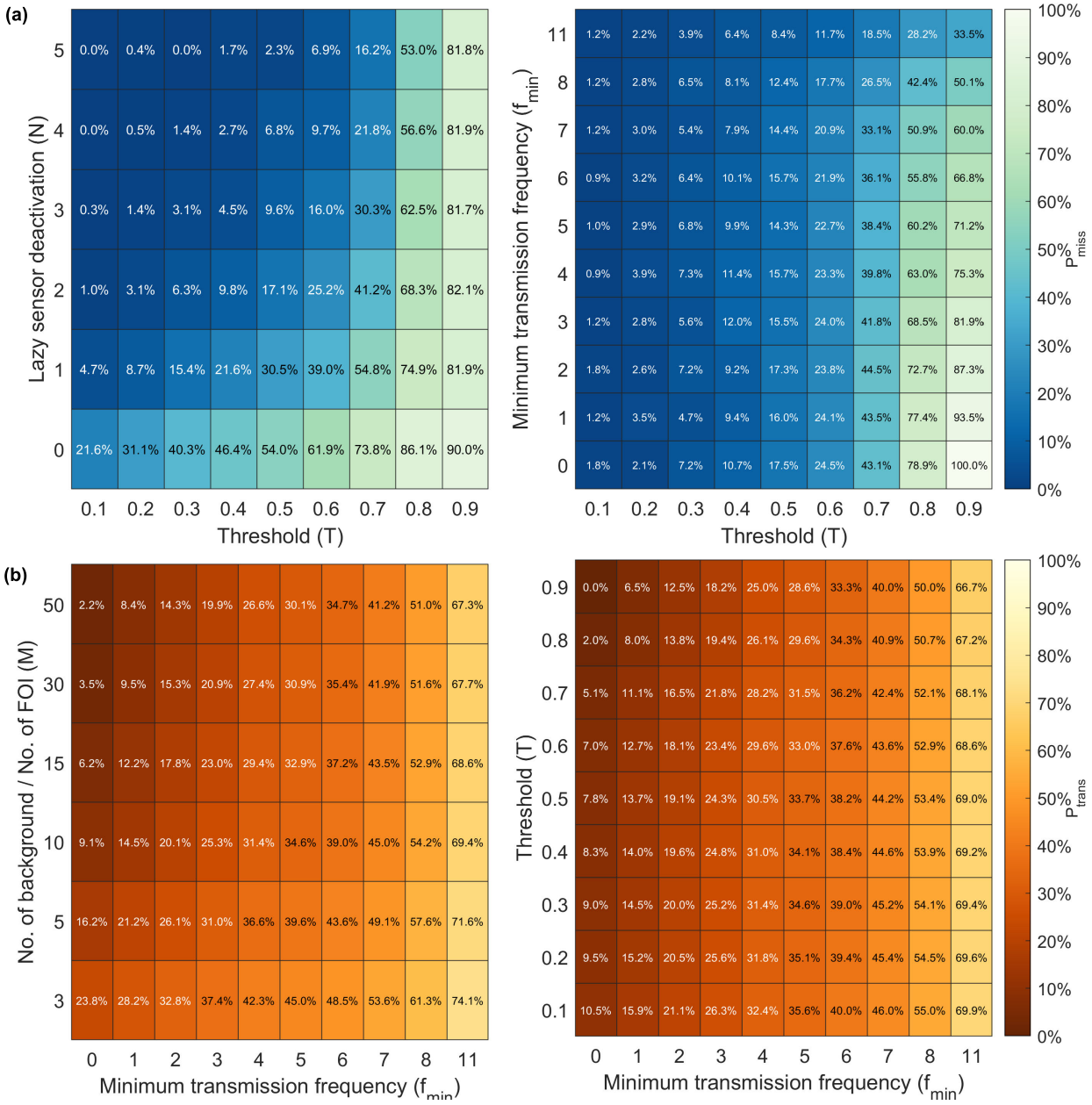


Fig. 6. Performance evaluation. (a) Heatmaps that display the miss rate P_{miss} with different parameter combinations (threshold (T), the ratio (M) of the number of background frames to the number of FOIs, minimum transmission frequency (f_{min}), and the count (N) at which the sensor deactivates). (b) Heatmaps that display the percentage of transmission P_{trans} with different parameter combinations.

specifically tailored to each task. This results in superior performance for each individual task.

The normalized energy consumption breakdown is depicted in Fig. 5. The conventional system, which consumes the most energy, is normalized as 1, with the others adjusted accordingly. Despite our framework introducing a negligible energy portion to the system (near-sensor model), it reduces the need for server-side inferences, resulting in a substantial decrease in overall system energy consumption compared to both the conventional system and the ATC method. It is capable of saving 87% on the energy consumption of the conventional system and keeps valuable information. While the ATC method eases server-side inferences, the number of inferences remains high, leading to a higher server energy consumption compared with our system. On the other hand, our dual-camera collaboration reduces camera energy consumption compared to both the conventional system and the ATC method.

D. Parameter Impact Analysis

Fig. 6(a) and (b) presents heatmaps illustrating the impact of the key parameters on P_{miss} and P_{trans}. For both metrics, a lower value indicates better performance. Regarding P_{miss}, a higher value of T leads to a notable increase in the number of missed FOIs. However, incorporating a lazier deactivation scheme and increasing the minimum transmission frequency can mitigate the adverse effects associated with a higher value of T. When examining the left panel in Fig. 6, N exerts a dominant influence on P_{miss}. As N increases, accuracy

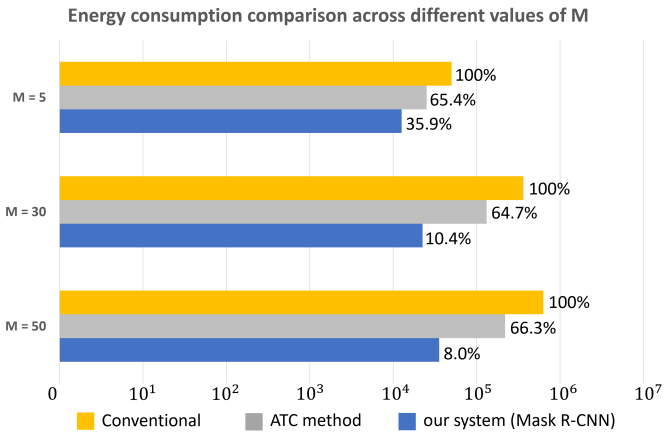


Fig. 7. Energy consumption comparison across different values of *M*. All servers are equipped with GeForce RTX 3090.

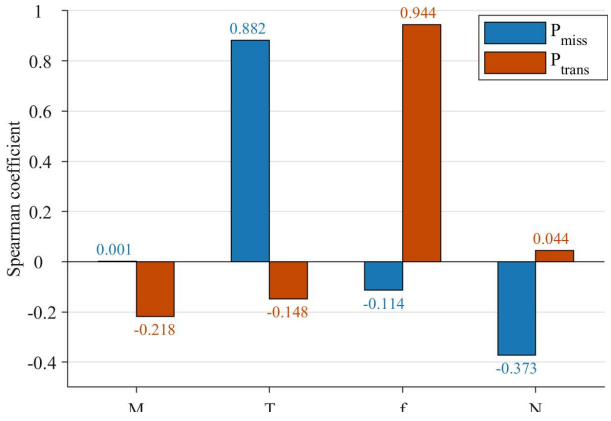


Fig. 8. Spearman coefficient of the parameters. The magnitude of the coefficient reflects the strength of the association between P_{miss}, P_{trans} and the variables (**T:** the confidence threshold of the model, **M:** the ratio of the number of background frames to the number of FOIs, **f:** minimum transmission frequency, and **N:** The count at which sensor deactivates).

improves significantly, leading to a decrease in the percentage of misdetections. Conversely, raising the minimum transmission frequency has a predominantly negative impact on P_{trans}. However, adopting a high value of T can reduce the amount of data transmission.

We also examined the relationship between M and P_{trans} and found that our framework exhibits a clear advantage as M increases, in Fig. 7. Specifically, as M increases from 5 to 50, our approach can save energy ranging from 65% to 92% compared to the conventional system. In contrast, the energy consumption of both the conventional method and the ATC method increases proportionally as the amount of data grows.

The Spearman correlation coefficients [54] of the parameters are presented in Fig. 8. In the figure, a positive coefficient indicates a positive correlation between two variables, while a negative coefficient indicates a negative correlation. The magnitude of the coefficient reflects the strength of the association between the variables. Specifically, T shows a significant positive correlation with P_{miss}, indicating that as T increases, P_{miss} also increases. Conversely, N exhibits a significant negative correlation with P_{miss}, indicating that higher values of N are associated with lower values of P_{miss}. On the other hand, the coefficient between N and P_{trans} is only 0.044, suggesting that changes in N have little influence on the percentage of data transmission. This observation aligns with our analysis in Section II-D. Moreover, as the near-sensor model operates continuously, the energy reduction in our framework primarily stems from the decrease in data transmission and the resulting reduction in server-side inference. Given that N has a negligible effect on P_{trans}, it also has minimal impact on overall energy consumption. In addition, an increase in f corresponds to an increase in P_{trans}, demonstrating a strong direct positive correlation.

TABLE I
MODEL PARAMETERS

Model Name	No. of Parameters	GFLOPS
YOLOv5 n	1,765,270 (100%)	4.2
YOLOv5 nm	433,190 (24.5%)	1.1
YOLOv5 ns	108,806 (6.2%)	0.4

E. Model Customization Analysis

We compared the performance of various lightweight near-sensor models mentioned in Section II-B1. We evaluated the tradeoff between the sensitivity and specificity of these models using receiver operating characteristic (ROC) curves and area under the curve (AUC), as illustrated in Fig. 9(a). In addition, Table I displays the number of parameters and GFLOPS of each model. While the AUC of YOLOv5ns is only slightly lower than that of YOLOv5n, the reduction in model size is significant, with the number of parameters decreasing to only 6.2% of the latter. Furthermore, the detrimental effect resulting from the reduction in model size can be alleviated by incorporating the lazy sensor deactivation scheme and elevating the minimum transmission frequency.

We also investigated the influence of quantization on the model performance. YOLOv5n trained on the original loss is quantized into different bit precisions, i.e., 16-bit float point (*fp16*), 8-bit integer (*int8*), 5-bit integer (*int5*), and 4-bit integer (*int4*). The performance of both the *fp16* and *int8* quantized models remains unaffected. However, as illustrated by Fig. 9(b), when we further reduce bit precision to *int5*, a slight degradation in AUC is observed (from 0.97 to 0.96), and a degradation in performance is noticeable when the model is quantized to *int4* (from 0.97 to 0.93).

As discussed in Section II-B3, the simplified loss function reduces task difficulty, allowing the model to become more lightweight or be more aggressively quantized. Fig. 9(c) illustrates the impact assessment of our tailored loss function, demonstrating that with the tailored loss, the model can achieve intensive quantization while maintaining comparable performance levels. The model trained on the adapted loss achieves a higher AUC score under the same level of aggressive quantization (*int4* quantization) compared to the model trained on the original loss. It achieves the same AUC as the model training on the original loss with the precision float

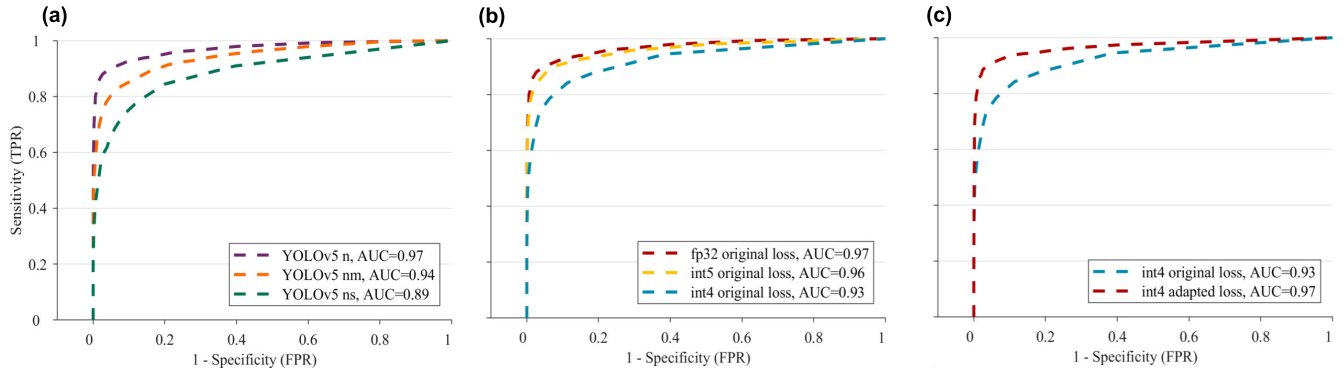


Fig. 9. Model comparison. (a) ROC curves of three lightweight models. (b) ROC curves of the models with different quantization trained by original loss. (c) ROC curves of the model subjected to *int4* quantization, trained with our adapted loss function and the original loss function.

TABLE II
DESIGN ACCELERATION ON AMD-XILINX ZCU104

	LUT	FF	BRAM	URAM	DSP
Total	84.9K	146.5K	224	40	844
Available	230.4K	460.8K	312	96	1728
Utilization	36.87%	31.80%	71.79%	41.67%	48.84%

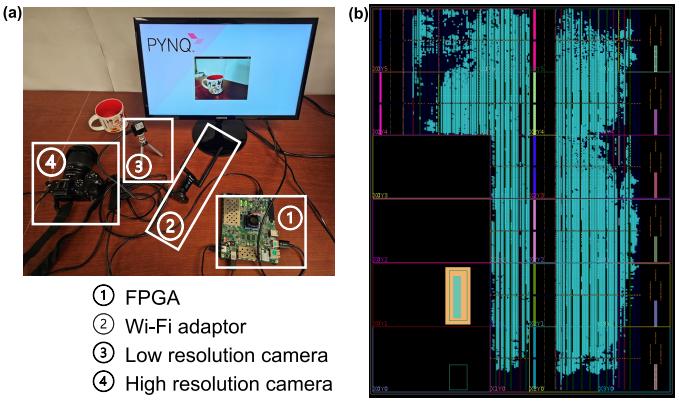


Fig. 10. Experiment setup. (a) Sensor side setup. (b) Accelerator placement layout on AMD Xilinx ZCU104 FPGA.

fp32. Since *fp32* requires 32 bits (4 bytes) per parameter, while *int4* only requires 4 bits (0.5 bytes) per parameter, jointly adopting the quantization and the customized loss can make the model 8× smaller without losing performance.

F. Hardware Implementation

The setup on the sensor side is depicted in Fig. 10(a). A high-resolution camera (④), a low-resolution camera (③), and a Wi-Fi adaptor (②) are connected to the FPGA board (①) via cable to capture and transmit the frames to the server. In addition, a screen is utilized for visualizing the information captured by the camera.

To meet the requirements of the proposed scenario, the near-sensor model is deployed on a resource-limited low-power edge-level FPGA: AMD-Xilinx Zynq UltraScale+ MPSoC ZCU104 (ZCU104) [55]. FPGAs are semiconductor devices that are based on a matrix of configurable logic blocks (CLBs) connected via programmable interconnects. Through hardware programming (such as Verilog or HLS), we can implement an ML accelerator on FPGA. The host program, executed on the ARM Cortex-A53 processor on the ZCU104's processing system (PS), was developed in Python. The communication between the PS and the programmable logic (PL) is established through the AMBA Advanced eXtensible Interface (AXI). Here, the PS side is a host ARM processor and the PL side is a reconfigurable logic. Our architecture design is implemented on the top of PL (reconfigurable logic).

To leverage hardware acceleration, we utilized the AMD-Xilinx deep learning unit (DPU) intellectual property (IP) as our hardware accelerator on the ZCU104's PL side. Our model was integrated into the DPU using the Vitis AI framework [56]. Vitis AI is an ML compiler framework developed by AMD-Xilinx that automatically maps ML operations (such as convolution and fully connected layers) into Xilinx hardware IP. The Vitis AI version that we choose is 2.0. Furthermore, the cameras are connected to the host ARM CPU, which facilitates communication with the cloud server. TCP protocol is used as the communication protocol. Table II, we present the FPGA resource utilization result. In Fig. 10(b), we present the accelerator placement layout on AMD Xilinx ZCU104 FPGA. The overview of our hardware platform is shown in Fig. 11.

Considering the constraints of resources such as power and space, we sometimes need to reduce the acceleration performance of deep processing units (DPU) [57], [58]. For instance, in Table II, we select the parallelism for input, output, and pixel processing of convolution operations to be 16, 16, and 8, respectively. If the goal is to reduce power consumption and resource utilization, one strategy is to decrease computation parallelism. Another strategy involves employing knowledge distillation and quantization to minimize model size, thereby reducing the computational overhead of edge hardware accelerators [5], [59]. In this work, we concentrate on accelerating the near-sensor framework on edge FPGAs. However, we may also consider other AI computing platforms such as Google Edge TPU and NVIDIA Jetson Nano [60]. These chips facilitate easier programming of ML models but compromise the capability for hardware resource reconfiguration.

G. Other Applications

In addition to visual monitoring, our proposed framework can be readily applied for multiple other tasks, such as audio processing and radar monitoring.

For the audio processing task, we used the UrbanSound8K dataset [61], a public audio dataset for urban sound classification applications. It contains ten classes, including car horn, gunshot, and dog bark. We focused on the gunshot and siren classes as the audio of interest. The dataset was reorganized and relabeled following the strategy outlined in Section III-A. The near-sensor model detects the audio of interest, while the server model classifies the specific class of that audio segment. A frequency-domain filter bank is applied to the audio signals, which are windowed in the time domain, to generate Mel spectrograms. These spectrograms are then fed into a CNN for classification. Compared to conventional methods, the system adopting our framework maintains comparable accuracy while consuming only 25% of the energy.

For radar monitoring, we evaluated the framework using the CRUW dataset [62], a public camera-radar dataset designed for autonomous vehicle applications. The radar images in this dataset are captured by the TI AWR1843, which operates at approximately 30 W [63]. The dataset was processed following the procedure outlined in Section III-A. Under the same deployment settings, the system using our framework achieved comparable performance while consuming only 18% of the energy required by the conventional system.

H. Security and Privacy Analysis

While our framework offers advantages such as reduced energy consumption and minimized bandwidth requirements, it necessitates an investigation of its security and privacy implications. Security considerations encompass data encryption both in transit and at rest. On the other hand, privacy concerns entail data minimization through near-sensor processing and anonymization techniques, user consent, and transparency regarding data usage. Compared with ATC methods that transmit abstract features, our framework transmits the original frames, thus sacrificing data encryption during transmission. However, it's worth noting that the conventional system also lacks data encryption during data transmission. This weakness can be mitigated by employing encryption techniques tailored specifically to image data. In addition, both ATC methods and our framework introduce an extra near-sensor model component. Since the model is situated near the sensor and only processes incoming data locally, it does not leak any information, ensuring data safety.

IV. Limitations

While our framework is highly effective, there are a few limitations to consider for further improvements in future works.

1) *Initial Training and Labeled Data Requirement:* Deploying the framework necessitates training the near-sensor model with labeled data. In some scenarios, obtaining labeled data may be challenging, or it might be difficult to fully cover the data distribution of sensor data. This introduces an initial cost associated with the deployment of the near-sensor module. However, after this initial cost, the framework can save substantial energy, especially on the server side.
2) *Accuracy Tradeoffs:* Due to the lightweight nature of the near-sensor model, its accuracy may be lower than that of the original, more complex model. Although we have proposed schemes such as lazy deactivation and maintaining a nonzero transmission frequency to mitigate possible misdetections, the near-sensor model's accuracy is still slightly lower than the server-side model. While this loss of accuracy is negligible in many scenarios, it becomes critical in applications where high accuracy is paramount. In such cases, the near-sensor model may need to be less lightweight, which would reduce the energy savings.
3) *Environmental Constraints:* The performance of our framework is dependent on the ratio of background frames to FOIs. In environments where this ratio is lower, the energy savings and efficiency improvements may not be as significant.

In future work, we aim to enhance the performance of the near-sensor model while maintaining its low energy consumption and reducing the initial deployment cost of the framework.

V. Conclusion

In this article, we introduce a novel framework for intelligent sensing that addresses some of the challenges associated with analyzing large-scale sensor data using complex ML models. Our framework is designed based on the observation that in many IoT applications, only a small proportion of sensor data conveys information of interest. Therefore, our framework intelligently selects the data generated by the sensors and only transmits and analyzes the data with useful information. It employs a near-sensor model to detect information of interest and control the data transmission, and a complex model located in the server to implement more sophisticated inference. The near-sensor model and the minimum transmission frequency are beneficial to reduce the energy and storage requirements, with a focus on decreasing the transmission frequency when no useful information is detected.

We set up the system with our framework on a low-power FPGA and evaluated the performance. The experimental results demonstrate that our framework significantly reduces total energy consumption and storage usage to less than 10% of that of conventional systems while retaining over 95% of useful information. Furthermore, we customized the model architecture and the loss function to suit our specific scenario and implemented quantization to achieve additional model compression. By jointly applying the customized loss function and quantization, the near-sensor model achieves an $8\times$ reduction in size without any loss in performance.

We also investigated the key factors influencing the framework's effectiveness. Instead of completely halting data transmission when no FoI is detected, we maintain a nonzero minimum transmission frequency. This ensures regular, low-frequency transmission to the server even in the absence of

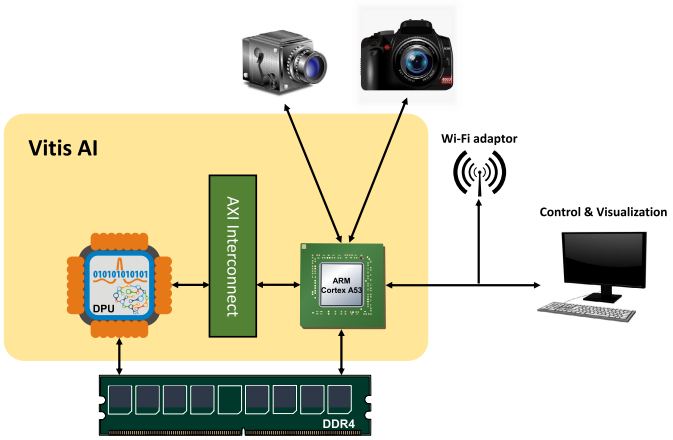


Fig. 11. Hardware platform overview.

Algorithm 1 Intelligent Data Transmission

Require: YOLO prediction(y), Lazy sensor deactivation count(N), Camera refresh rate(f_r), Minimum transmission frequency(f_{min}), $C_1, C_2, C_3 = 0, 0, 0$

Ensure: Transmission decision(D)

```
if y == 1 then
    C1, C2, C3 = 0, 0, 0
    return D = 1
else
    if C3 == 0 then
        C1 = C1 + 1
        if C1 ≤ max(1, N/2^C2) then
            return D = 1
        else
            C1, C2, C3 = 0, C2 + 1, C3 + 1
            return D = 0
        end if
    else
        C3 = C3 + 1
        if C3 == fr/fmin then
            C1, C3 = C1 + 1, 0
            return D = 1
        else
            return D = 0
        end if
    end if
end if
```

FoIs, thereby benefiting the integrity of the useful information. In addition, our lazy sensor deactivation scheme leverages the temporal correlation between adjacent frames, achieving a balance between resource consumption and accuracy. Furthermore, our proposed framework demonstrates greater effectiveness as the ratio of background frames to FoIs increases.

In addition to the evaluations under the visual monitoring scenario, we extended the framework to other tasks, e.g., audio processing and radar monitoring. The results show the versatility and applicability of our framework under different scenarios.

We also discussed the limitations of our current framework and outlined potential directions for future improvements.

Data Availability Statement

The dataset Microsoft COCO object detection for this study can be found in [50]. The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Appendix
Supplementary Material

Video Demo

Our research includes a video demonstration showcasing the results. In the demo, our model detects the animals appearing in the frames. The video can be accessed at the following link: https://drive.google.com/file/d/1-IpRLfd8Ym38p8APCJgxNq5igiK5ARa5/view?usp=sharing.

Acknowledgment

The authors express their gratitude to Andrew Ding for the discussion and meticulous proofreading to improve the earlier version of the article.

References

[1] M. M. Sadeeq, N. M. Abdulkareem, S. R. M. Zeebaree, D. M. Ahmed, A. S. Sami, and R. R. Zebari, "IoT and cloud computing issues, challenges and opportunities: A review," *Qubahan Academic J.*, vol. 1, no. 2, pp. 1–7, Mar. 2021.

[2] L. Yang and A. Shami, "IoT data analytics in dynamic environments: From an automated machine learning perspective," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105366.

[3] Y. Djenouri, A. Belhadi, G. Srivastava, E. H. Houssein, and J. C. Lin, "Sensor data fusion for the industrial artificial intelligence of things," *Expert Syst.*, vol. 39, no. 5, Jun. 2022, Art. no. e12875.

[4] S. Yun et al., "HyperSense: Hyperdimensional intelligent sensing for energy-efficient sparse data processing," *Adv. Intell. Syst.*, vol. 1, Jun. 2024, Art. no. 2400228.

[5] Y. Ni et al., "HEAL: Brain-inspired hyperdimensional efficient active learning," 2024, *arXiv:2402.11223*.

[6] V. Tsakanikas, T. Dagiuklas, M. Iqbal, X. Wang, and S. Mumtaz, "An intelligent model for supporting edge migration for virtual function chains in next generation Internet of Things," *Sci. Rep.*, vol. 13, no. 1, p. 1063, Jan. 2023.

[7] M. Kumari and A. Kaul, "Deep learning techniques for remote sensing image scene classification: A comprehensive review, current challenges, and future directions," *Concurrency Comput., Pract. Exper.*, vol. 35, no. 22, Oct. 2023, Art. no. e7733.

[8] H. Chen et al., "TaskCLIP: Extend large vision-language model for task oriented object detection," 2024, *arXiv:2403.08108*.

[9] G. Wang, X. Yang, W. Cai, and Y. Zhang, "Event-triggered online energy flow control strategy for regional integrated energy system using Lyapunov optimization," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106451.

[10] C.-F. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132–4150, Aug. 2019.

[11] Z. Sun et al., "Cloud-edge collaboration in industrial Internet of Things: A joint offloading scheme based on resource prediction," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17014–17025, Sep. 2022.

[12] Y.-H. Chiang, T. Zhang, and Y. Ji, "Joint cotask-aware offloading and scheduling in mobile edge computing systems," *IEEE Access*, vol. 7, pp. 105008–105018, 2019.

[13] M. Chen and Y. Hao, "Task offloading for mobile edge computing in software defined ultra-dense network," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 587–597, Mar. 2018.

[14] Y. Wang, X. Tao, X. Zhang, P. Zhang, and Y. T. Hou, "Cooperative task offloading in three-tier mobile computing networks: An ADMM framework," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2763–2776, Mar. 2019.

[15] Z. Zheng, L. Song, Z. Han, G. Y. Li, and H. V. Poor, "A stackelberg game approach to proactive caching in large-scale mobile edge networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5198–5211, Aug. 2018.

[16] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1960–1971, Apr. 2019.

[17] S. Yu, X. Wang, and R. Langar, "Computation offloading for mobile edge computing: A deep learning approach," in *Proc. IEEE 28th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Aug. 2017, pp. 1–6.

[18] Z. Ali, L. Jiao, T. Baker, G. Abbas, Z. H. Abbas, and S. Khaf, "A deep learning approach for energy efficient computational offloading in mobile edge computing," *IEEE Access*, vol. 7, pp. 149623–149633, 2019.

[19] M. Issa et al., "Hyperdimensional hybrid learning on end-edge-cloud networks," in *Proc. IEEE 40th Int. Conf. Comput. Design (ICCD)*, Oct. 2022, pp. 652–655.

[20] H. Yang et al., "BrainIoT: Brain-like productive services provisioning with federated learning in industrial IoT," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2014–2024, Feb. 2022.

[21] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[22] C. Lammie, A. Olsen, T. Carrick, and M. Rahimi Azghadi, "Low-power and high-speed deep FPGA inference engines for weed classification at the edge," *IEEE Access*, vol. 7, pp. 51171–51184, 2019.

[23] H. Chen, Y. Ni, W. Huang, and M. Imani, "Scalable and interpretable brain-inspired hyper-dimensional computing intelligence with hardware–software co-design," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2024, pp. 1–8.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, vol. 3951. May 2006, pp. 404–417.

[26] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.

[27] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *physics/0004057*.

[28] L. Xiang, J. Luo, and A. Vasilakos, "Compressed data aggregation for energy efficient wireless sensor networks," in *Proc. Annu. IEEE Commun. Soc. Conf. Sens., Mesh Ad Hoc Commun. Netw.*, Jun. 2011, pp. 46–54.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[30] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*.

[31] A. M. Ghosh and K. Grolinger, "Edge-cloud computing for Internet of Things data analytics: Embedding intelligence in the edge with deep learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2191–2200, Mar. 2021.

[32] B. Yogameena, K. Menaka, and S. Saravana Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intell. Transp. Syst.*, vol. 13, no. 7, pp. 1190–1198, Jul. 2019.

[33] W. Huang et al., "Exploration of using a pressure sensitive mat for respiration rate and heart rate estimation," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 298–301.

[34] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.

[35] J. Deng, "A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit., 2009*, Oct. 2009, pp. 1–26.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[38] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–29.

[40] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–24.

[41] L. Deng et al., "Lightweight aerial image object detection algorithm based on improved YOLOv5s," *Sci. Rep.*, vol. 13, no. 1, p. 7817, May 2023.

[42] M. Zahrawi and K. Shaalan, "Improving video surveillance systems in banks using deep learning techniques," *Sci. Rep.*, vol. 13, no. 1, p. 7911, May 2023.

[43] A. Alqahtani, X. Xie, and M. W. Jones, "Literature review of deep network compression," *Informatics*, vol. 8, no. 4, p. 77, Nov. 2021.

[44] A. Goel, C. Tung, Y.-H. Lu, and G. K. Thiruvathukal, "A survey of methods for low-power deep learning and computer vision," in *Proc. IEEE 6th World Forum Internet Things (WF-IoT)*, Jun. 2020, pp. 1–6.

[45] C. N. Coelho et al., "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Mach. Intell.*, vol. 3, no. 8, pp. 675–686, Jun. 2021.

[46] I. Chakraborty, D. Roy, I. Garg, A. Ankit, and K. Roy, "Constructing energy-efficient mixed-precision neural networks through principal component analysis for edge intelligence," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 43–55, Jan. 2020.

[47] H. Wu, P. Judd, X. Zhang, M. Isaev, and P. Micikevicius, "Integer quantization for deep learning inference: Principles and empirical evaluation," 2020, *arXiv:2004.09602*.

[48] F. Cardinaux et al., "Iteratively training look-up tables for network quantization," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 860–870, May 2020.

[49] G. Caravagna, G. Costa, and G. Pardini, "Lazy security controllers," in *Proc. Int. Workshop Secur. Trust Manage.*, 2012, pp. 33–48.

[50] T.-Y. Lin, "Microsoft COCO: Common objects in context," in *Computer Vision ECCV 2014* (Lecture Notes in Computer Science), vol. 8693. Cham, Switzerland: Springer, 2014, pp. 740–755.

[51] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–19.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[53] Y. Matsubara and M. Levorato, "Split computing for complex object detectors: Challenges and preliminary results," in *Proc. 4th Int. Workshop Embedded Mobile Deep Learn.*, Sep. 2020, pp. 7–12.

[54] C. Spearman, "The proof and measurement of association between two things," *Int. J. Epidemiology*, vol. 39, no. 5, pp. 1137–1150, Oct. 2010.

[55] A. Xilinx. (2023). *Zynq Ultrascale*. Accessed: Dec. 12, 2023. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/zcu104.html

[56] V. Kathail, "Xilinx Vitis unified software platform," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, Feb. 2020, pp. 173–174.

[57] H. Chen, A. Zakeri, F. Wen, H. E. Barkam, and M. Imani, "HyperGRAF: Hyperdimensional graph-based reasoning acceleration on FPGA," in *Proc. 33rd Int. Conf. Field-Programmable Log. Appl. (FPL)*, Sep. 2023, pp. 34–41.

[58] H. Lee et al., "Comprehensive integration of hyperdimensional computing with deep learning towards neuro-symbolic AI," in *Proc. 60th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2023, pp. 1–6.

[59] H. E. Barkam et al., "Reliable hyperdimensional reasoning on unreliable emerging technologies," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2023, pp. 1–9.

[60] W. Huang et al., "EcoSense: Energy-efficient intelligent sensing for in-shore ship detection through edge-cloud collaboration," 2024, *arXiv:2403.14027*.

[61] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[62] Y. Wang, G. Wang, H.-M. Hsu, H. Liu, and J.-N. Hwang, "Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2809–2818.

[63] W. Li et al., "Real-time fall detection using mmWave radar," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 16–20.

Wenjun Huang received the B.Sc. degree in information science and technology from Southeast University, Nanjing, China, in 2019, and the M.Sc. degree in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 2021. He is currently pursuing the Ph.D. degree with the University of California at Irvine, Irvine, CA, USA.

He was an Associate Research Scientist at GN ReSound in the Netherlands. His research interests include embedded systems, computer vision, and machine learning for healthcare.

Arghavan Rezvani received the B.S. degree in computer engineering from Sharif University of Technology, Tehran, Iran, in 2022. She is now pursuing the Ph.D. degree in deep learning and computer vision with the University of California at Irvine, Irvine, CA, USA.

Hanning Chen (Graduate Student Member, IEEE) received the B.S. degree in microelectronic engineering from Nanjing University, Nanjing, China, in 2019, and the M.S. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2021. He is now pursuing the Ph.D. degree in computer science with the University of California at Irvine, Irvine, CA, USA.

He is a member of the Bio-Inspired Architecture and Systems Laboratory, University of California at Irvine. His research interests include computer architecture and machine learning.

Yang Ni received the bachelor's degree from the University of Glasgow, Glasgow, Scotland, in 2019, and the master's degree from the University of California at San Diego, San Diego, CA, USA, in 2021. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of California at Irvine, Irvine, CA, USA.

He is a member of the Bio-Inspired Architecture and Systems Laboratory (BIASLab), University of California at Irvine. His research interests include efficient machine learning, vector symbolic architecture, and reinforcement learning.

Mr. Ni received the Best Paper Award at the Design Automation and Test in Europe (DATE) Conference in 2022.

Sanggeon Yun received the B.S. degree in computer science from Kookmin University, Seoul, South Korea, in 2023. He is now pursuing the Ph.D. degree with the Bio-Inspired Architecture and Systems (BIASLab), University of California at Irvine, Irvine, CA, USA, under the supervision of Prof. Mohsen Imani.

His research interests include hyperdimensional computing, machine learning, natural language processing, human–computer interaction, and information visualization.

Sungheon Jeong received the B.S. degree in IT engineering from Pusan National University, Busan, South Korea, in 2022. He is now pursuing the Ph.D. degree in computer science with the University of California at Irvine, Irvine, CA, USA.

His current research interests include cover computer vision and reinforcement learning and also linear algebra and probabilistic approaches to OOD, black box interpretation, and latent representation.

Guangyi Zhang received the B.Eng. degree in electrical engineering from McGill University, Montreal, QC, Canada, in 2020. He is currently pursuing the master's degree in computer science with the University of California at Irvine, Irvine, CA, USA.

His research interests include macromodeling, machine learning, federated learning, theoretical machine learning, and generative models.

Mohsen Imani received the Ph.D. degree from the Department of Computer Science and Engineering, University of California at San Diego, San Diego, CA, USA, in 2020.

He is one of the pioneers in hyperdimensional computing (HDC) and its applications in the cognitive learning domain. He boasts an impressive publication record with over 170 papers in top conferences and journals and holds 20 U.S. patents. His contributions have paved a new path in brain-inspired hyperdimensional computing, enabling ultraefficient and real-time cognitive learning. His research has been a key factor in initiating multiple programs at the Semiconductor Research Corporation (SRC), the Defense Advanced Research Projects Agency (DARPA), Intel, IBM, and CISCO.

Dr. Imani's research has earned him several prestigious awards, including the DARPA Young Faculty Award 2023, the SRC Young Faculty Award 2023, the ONR Young Investigator Program Award 2023, the DARPA Riser 2022, the Bernard Gordon Engineering Leadership Award, and the Outstanding Researcher Award. He has also received six Best Paper Awards and nominations at top conferences. He has a long history of successful technology transfers to multiple companies and governmental agencies.