Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Empowering Retail Visitors Through Edge AI: A User-Centered Case Study of Cooperative Totem Systems

Gianluca Brilli[1], Federica Caruso[2], Giacomo Valente[2], Alberto Carlevaro[3], Chiara Garibotto[4], Jacopo Motta[3], Vittoriano Muttillo[5], Damiano Vallocchia[6], Paolo Burgio[1]

[1]Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Italy (e-mail: name.surname@unimore.it)
[2]Department of Engineering, Information Science and Mathematics, University of L'Aquila, Italy (e-mail: name.surname@univaq.it)
[3]Aitek SPA, Italy (e-mail: name.surname@aitek.it)
[4]Department of Naval, Electrical, Electronic and Telecommunications Engineering, University of Genoa, Italy (e-mail: chiara.garibotto@unige.it)
[5]University of Teramo, Italy (e-mail: vmuttillo@unite.it)
[6]Ro Technology SRL, Italy (e-mail: damiano.vallocchia@rotechnology.it)

Corresponding author: Gianluca Brilli (e-mail: gianluca.brilli@unimore.it).

This project has received funding from the Key Digital Technologies Joint Undertaking (KDT JU) under grant agreement No 877056. The JU receives support from the European Union Horizon 2020 research and innovation programme and Spain, Italy, Austria, Germany, Finland, Switzerland.

ABSTRACT Retail is one of the most significant and competitive economic sectors, and interactive systems such as informative totems play an increasingly important role in enhancing the visitor experience within shopping mall environments by offering value-added services. Over the years, these systems have evolved toward AI-driven context awareness to personalize content based on the people interacting with them. To maintain soft real-time responsiveness, however, existing systems typically rely on cloud-based processing of AI workloads. This raises significant privacy concerns, as sensitive visitor data must be transmitted to external systems or third-party services. Although prior literature demonstrates notable progress in integrating AI-driven context awareness into interactive systems for shopping mall environments, current solutions do not enable fully edge-confined execution of AI workloads and therefore cannot guarantee that sensitive data remain local. To address this gap, this paper presents an industrial case study aimed at developing a Cooperative Intelligent Totem System, in which all AI tasks are executed locally within an edge infrastructure composed of a cooperating totem and roof nodes, without relying on any external systems or third-party services. Experimental results show that the system achieves accurate AI-driven perception, consistently satisfies the one-second responsiveness requirement, and scales up to 18 simultaneous users when supported by multiple cooperating roof nodes, all while keeping sensitive data strictly confined to the edge.

INDEX TERMS Context-aware interactive totems, cooperative edge intelligence, workload sharing, edge-processing.

I. INTRODUCTION

Nowadays, retail represents one of the most relevant and competitive economic sectors [1], [2]. In 2024, U.S. retail sales reached $5.29 trillion, with forecasts projecting growth to $5.42 trillion in 2025 [3]; In Europe, retail sales rose 1.5% year-over-year through October 2025, while in the first six months of 2025 retail sales grew +2% driven by tourism and experiential formats in shopping malls [4]. In fact, enhancing the visitor experience within a shopping mall is recognised as one of the key factors contributing to increased sales [5], [6]. Therefore, to enhance the visitor experience, modern shopping mall managers must deliver shopping experiences that meet visitor expectations and strengthen engagement [7].

Among the technologies driving this evolution, digital signage has emerged as an effective enabler for creating appealing store layouts and delivering engaging communication [8]. In particular, **interactive systems**, such as interactive totems, enhance the visitor experience within the shopping mall by offering value-added services such as checking product availability, navigating shopping mall layouts, and displaying advertising content [2], [9], [10].

In order to continuously adapt to the surrounding envi-

ronment, interactive totems are evolving toward **AI-driven context awareness** [11], namely the capacity to understand who is in front of the totem and tailor the interaction through AI-based perception. Here, context-awareness is achieved via AI-driven analysis of visitors (e.g., age, gender, spoken language), enabling the interactive totem to tailor content to the detected audience. AI-algorithms have proven highly effective for this purpose [12], [13], enabling tailored recommendations and improved management of shopping malls [5], [9], based on sensitive perceptual data (e.g., facial images and utterances).

A critical requirement for interactive totems is the ability to provide **soft real-time responsiveness** (hereinafter referred to as real-time responsiveness), as delays in interaction may cause visitors to disengage [5]. To achieve this real-time responsiveness, many existing interactive systems rely on cloud-based execution of computationally heavy AI algorithms [5], where higher processing capacity is available. However, sending sensitive visitor data to the cloud raises significant privacy concerns, as personal data may be exposed to third-party infrastructures [14], [15]. For this reason, there is a growing consensus, in both research and industry, that keeping AI-based processing of sensitive data **confined to the edge** (i.e., a processing with no data shared with external systems or third-party services) is a robust strategy for reducing unnecessary exposure [16], [17]. This architectural choice aligns with widely accepted data-protection principles such as data minimization and purpose limitation [18]. As a result, edge-confined processing has become a preferred technical solution in domains involving sensitive perceptual data, including smart cameras, mobile devices, and ambient intelligence. Since interactive totems rely on similar classes of perceptual data, adopting an edge-confined processing strategy is a natural and technically justified design choice, enabling responsible data handling while still supporting real-time, personalized interaction.

While the literature shows significant progress in integrating AI-driven context awareness within interactive systems [19]–[25], existing solutions do not provide a fully edge-confined AI-based processing of sensitive data. Specifically, no AI-driven interactive system deployed in shopping malls provides simultaneously: *(i) AI-driven context-awareness, (ii) real-time responsiveness, and (iii) edge-confined processing*.

To address this gap, this paper presents an industrial case study aimed at developing a *Cooperative Context-Aware Totem System*, in which all AI tasks are executed locally within an edge infrastructure, without relying on external systems or third-party services (e.g., cloud platforms). Specifically, the main results of this case study consist of:

- The design of the proposed Cooperative Context-Aware Totem System, composed of cooperating totem and roof nodes, developed to satisfy the user requirements elicited through a user-centered analysis of the context of use, while ensuring (i) AI-driven context awareness, (ii) real-time responsiveness, and (iii) edge-confined processing.
- The implementation of a prototype of the Cooperative Context-Aware Totem System on two AMD Zynq UltraScale+ platforms.
- The evaluation of the implemented prototype in terms of the accuracy of its AI tasks, its real-time responsiveness, and its scalability, conducted in a laboratory setting that emulates a real shopping mall environment.

The remainder of this paper is structured as follows. Section II reviews the related work. Section III presents the case study, detailing the definition of user requirements through a user-centred design methodology. Section IV describes the proposed cooperative intelligent totem system. Section V reports the experimental results and provides a discussion. Finally, Section VI concludes the paper and outlines future work.

II. RELATED WORK

A growing body of work has investigated AI-based interactive systems for shopping mall environments, spanning both academic research and commercial products. These systems typically combine sensing modules with AI inference to analyse the presence and behaviour of visitors and to generate real-time responses. However, existing systems differ widely in how sensing, inference, and data handling are organised, which results in different degrees of support for *(i) AI-driven Context-Awareness*, (*ii) real-time responsiveness*, and *(iii) edge-confined processing*. The following analysis reviews existing systems with respect to these three capabilities, with a comparative summary provided in Table 1.

Academic research. Sung et al. [19] propose a mixed-reality AI-based interactive system deployed in a shopping mall environment, where visitors interact with an augmented virtual guide using real-time speech recognition and synthesis. Specifically, the system was designed to enable immediate vocal interaction and to enhance visitors engagement. However, the authors do not describe how sensitive data (i.e., spoken utterances) are processed or stored, nor whether they are handled locally or shared with external systems or third parties (e.g., cloud services).

Bhuvana et al. [20] propose an emotion-aware analytics approach applicable to shopping mall environments, based on facial expression recognition and posture analysis to estimate visitor engagement and support AI-based interactions with service robots. The authors describe how multi-modal vision techniques could provide insights into the emotional states and behavioral cues of visitors in real-time. However, the proposed approach remains conceptual and does not present a deployed system or a full processing pipeline. In addition, the authors do not specify how sensitive data (i.e., facial images) are processed or stored, nor whether such data are handled locally or shared with external systems or third parties (e.g., cloud services).

Beem et al. [21] present a cloud-based personalisation interactive system for e-commerce settings, described through a case study of a fashion retailer that combines *Amazon*

TABLE 1. Comparison of existing academic and commercial AI-based interactive systems

Type	Work	(i) AI-driven Context-Awareness	(ii) Real-time Responsiveness	(iii) Edge-confined Processing
Academic Research	**Sung et al. [19]**	✓	✓	✗
	Bhuvana et al. [20]	✓	✓	✗
	Beem et al. [21]	✓	✓	✗
Commercial Product	**Hikvision [22]**	✓	✗	✗
	Victek [23]	✓	✓	✗
	Tacteasy [24]	✓	✓	✗
	IntelligentKiosk [25]	✓	✓	✗

Personalize [26] with a fine-tuned GPT model for dynamic content generation. The authors emphasize the significance of AI-driven personalisation and outline the technical challenges associated with real-time recommendation pipelines. However, the system is limited to digital retail environments and relies entirely on cloud-based services, with no information provided on the handling of sensitive data.

Taken together, existing academic research works show a range of approaches for enabling *(i) AI-driven context-awareness* and *(ii) real-time responsiveness*, but provide limited detail on how sensitive data (e.g., facial images, spoken utterances) are processed in practice, and none explicitly addresses *(iii) edge-confined processing*.

Commercial products. Several commercial vendors offer AI-based interactive systems for shopping mall environments, providing functionalities such as visitor presence sensing, face detection, demographic estimation, and content adaptation in real-time (e.g., Victek [23], Tacteasy [24], IntelligentKiosk [25]). These systems typically combine embedded camera modules with cloud-connected management platforms, supporting context-aware behavior and immediate responsiveness to visitor presence or movement. However, publicly available documentation describes only high-level capabilities and does not specify how captured sensitive data are processed, stored, or transmitted, nor whether any component can operate without reliance on external systems or third parties (e.g., cloud services). As a result, their data-handling architectures remain unclear, particularly with respect to the possibility of *(iii) edge-confined processing*.

Overall, the reviewed academic and commercial AI-based interactive systems consistently support capabilities (i) and (ii), but none provide evidence of capability (iii). This limitation is relevant because recent studies indicate that edge-confined processing can reduce privacy risks by limiting the exposure of sensitive data during transmission and storage [12], [15], [27], [28]. The system proposed in this paper overcomes this limitation by ensuring that sensitive data and inference processing remain confined to the local edge deployment environment.

III. CASE STUDY

This work builds on an industrial case study conducted in collaboration with *Aitek*[1], an Italian company specialized in video analytics and digital signage, within Use Case 6 "*Intelligent Totems*" of the *Fractal* European project [29]. The case study examines the design of a *Cooperative Intelligent Totem System* (hereinafter *System*) deployed in shopping malls, a context of use defined by concrete physical, organizational, and user-related constraints that must be explicitly addressed during system design and evaluation [30].

The goal of this case study is to develop a *System* prototype capable of *(i) AI-driven Context-Awareness*, *(ii) real-time responsiveness*, and *(iii) edge-confined processing*. Achieving this goal requires a systematic understanding of the *context of use*, encompassing the shopping mall environment, the characteristics of the visitors who interact with the *System*, and the interaction conditions under which the *System* is expected to operate [31]. To this end, a User-Centered Design (UCD) process [32] is adopted, applying its first two stages, devoted to understanding the context of use (Section III-A) and to specifying the user requirements (Section III-B), whose outcomes informed the design of the *System* presented in Section IV.

A. UNDERSTANDING THE CONTEXT OF USE

Within the context of use considered in this case study, the **environment** is a shopping mall, an enclosed indoor place that aggregates multiple stores and shared public spaces and is characterized by continuous pedestrian circulation and complex spatial layouts that may extend across several floors connected by escalators or elevators [33], [34].

The physical characteristics of the shopping mall, including strong artificial lighting and pervasive background music, result in an environment that is generally bright and noisy. Human dynamics within this environment are continuous, and crowd conditions are highly dynamic, with density levels fluctuating over time and peaking during weekends, sales periods, and seasonal events [35], [36].

The technological infrastructure of the shopping mall includes video surveillance systems used for safety and security monitoring [37], [38], as well as informative totems typically

[1]https://www.aitek.it/

located at points of high pedestrian circulation, such as entrances or corridor junctions, which provide mall directories and advertising [2], [9], [10].

The **users** involved in this context of use include shopping mall visitors, store personnel, and shopping mall staff.

Visitors represent the primary user group, comprising a heterogeneous population with balanced proportions of men and women, differing in age, nationality, and motivations for visiting the shopping mall [39]–[41]. These motivations may be *hedonic*, such as leisure and social engagement, or *utilitarian*, such as targeted shopping or quick access to specific services [40], [42], [43]. Such demographic and motivational differences affect visitors needs and their preferred modalities for receiving informative and advertising content, influencing both what content they consume and how it is most effectively presented within the shopping mall [44].

Store personnel constitute a secondary user group whose needs are shaped by marketing-related activities aimed at promoting products, attracting visitors, and evaluating the effectiveness of promotional strategies within the shopping mall environment [45]. These needs encompass the creation and delivery of tailored content used to engage visitors within the shopping mall.

Shopping mall staff, including security, administrative, and customer service personnel, represent responsible for the overall daily functioning of shopping mall environments. These employees are integral to operational continuity, visitor assistance, and the provision of a safe and secure environment. Their effectiveness relies heavily on situational awareness of activities throughout the shopping mall, timely access to relevant information for visitor support, and continuous monitoring of security-related incidents [46], [47].

This analysis of the environment and users completes the first stage of the UCD process (*understanding the context of use*), and provides the basis for the second stage, detailed in the next subsection.

B. SPECIFY USER REQUIREMENTS

In the second stage of the UCD process, Personas and Scenarios [48] were modeled to represent users from the three user groups identified in the analysis of the context of use and their typical interactions with the system. Specifically, Personas provide concise characterizations of these users, and Scenarios describe typical situations in which they engage with the *System* in the shopping mall environment. These artifacts support the subsequent specification of user requirements.

1) Personas and Scenarios

The Personas and Scenarios developed correspond to the three user groups identified in the analysis of the context of use: (1) Visitors, (2) Store personnel, and (3) Shopping mall staff.

1) **Elena Lopez - Visitor**: Elena López is a Spanish-speaking senior visitor traveling in Berlin who is unfamiliar with the layout of the shopping mall. Her goal is to locate shops quickly and access information that is readable and linguistically accessible. When interacting with the *System* to search for a specific shop, she relies on the interface to present content in her preferred language with adequate text readability, enabling her to identify relevant shop categories and obtain clear directions.
2) **Michael Kant - Store Personnel**: Michael Kant works as part of the store personnel of a clothing shop located along one of the shopping mall main corridors. His goal is to promote current offers and assess whether they attract the attention of passing visitors. Over time, he observes how visitors behave in the vicinity of his store and how the *System* adapts the advertising content shown to approaching visitors. When certain promotional messages appear to be associated with increased visitor interest around the shop area, he keeps them active; when interest remains limited, he updates the advertising content or reorganizes the shop-front layout to improve visibility.
3) **Katrina Weber - Shopping Mall Staff**: Katrina Weber is part of the shopping mall staff and is responsible for monitoring daily operations and ensuring that visitor circulation remains smooth. Her goal is to identify situations that may require an operational intervention. While overseeing the shopping mall, she reviews information provided by the *System* about localized changes in visitor presence within specific areas of the shopping mall. In fact, when the *System* indicates that an unusual queue is forming near the restrooms due to increased crowd density, she immediately dispatches maintenance staff to verify the cause and restore normal conditions.

These Personas and Scenarios helped clarify user needs and interaction conditions, which in turn informed the specification of the User Requirements.

2) User Requirements

The specified user requirements consolidate the needs identified through Personas and Scenarios, formalizing both the *System* functions that must be supported (i.e., *functional requirements*) and the conditions under which these functions must be delivered (i.e, *non-functional requirements*). Their specification adheres to the overall goal of the case study, namely developing a *System* prototype capable of *(i) AI-driven context-awareness*, *(ii) real-time responsiveness*, and *(iii) edge-confined processing.*

a: Functional Requirements

- The *System* shall perform **people detection** through AI-based visual sensing analysis to identify the presence and number of visitors within the area sensed by the *System*.
- The *System* shall perform **crowd density estimation** through AI-based visual sensing analysis to identify abnormal or critical crowding conditions in its surroundings.

- The *System* shall perform **face detection** through AI-based visual sensing analysis to localize visitors and enable subsequent tailored interactions.
- The *System* shall estimate the visitors **age** and expressed **gender** through AI-based visual sensing analysis.
- The *System* shall detect the visitors **spoken language** through AI-based audio sensing analysis.
- The *System* shall deliver **tailored content** based on the detected demographic characteristics of the visitors (i.e., age, gender, and language).

Please note that crowd density estimation is specified as a separate functional requirement because it involves distinct processing and supports different management actions than those associated with the people detection function.

The reliance on AI-based sensing and data-driven inference to implement these functions is essential for achieving *(i) AI-driven Context-Awareness*.

b: Non-Functional Requirements

- The *System* shall deliver the tailored content within **one second** from the initial face detection of the visitor.
- The *System* shall store all **sensitive data locally**, and any cooperation shall occur exclusively within the **local edge infrastructure** of the *System* (edge-confined), **without the involvement of external systems or third parties (e.g., cloud services)**.

Please note that the one-second latency requirement follows established usability guidelines [49], which indicate that system responses occurring within approximately one second are perceived by visitors as instantaneous.

Meeting these two non-functional requirements supports the goal of the case study of achieving *(ii) real-time responsiveness* and *(iii) edge-confined processing*.

The following section details how these user requirements were translated into the development of the *System* prototype.

IV. THE PROPOSED COOPERATIVE INTELLIGENT TOTEM SYSTEM

The design of the proposed *System* follows the methodological framework developed within the European project *Fractal*[2], whose main objective is to introduce a novel approach to edge computing based on adaptive and context-aware computing nodes, referred to as *fractal nodes*. The proposed *System* is implemented as a network of cooperative fractal nodes forming a local edge infrastructure. The primary rationale behind this architectural choice is the adoption of an edge-confined processing strategy: by deploying a fractal node surrounded by interconnected edge nodes with which it can share computation, all sensitive data remains strictly within the local edge infrastructure [50].

When the fractal node must execute computationally demanding tasks involving sensitive data, the system enables cooperation among fractal nodes in order to distribute a portion of the workload to nearby fractal nodes within the same local edge infrastructure (a mechanism referred to as *workload sharing*).

[2]https://fractal-project.eu/

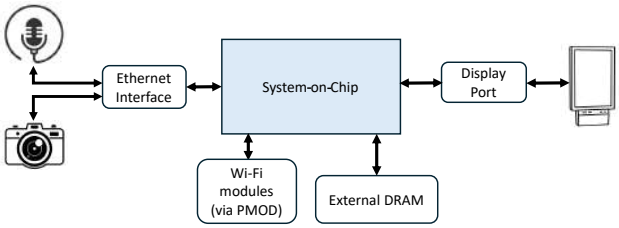

FIGURE 1. Hardware Infrastructure of the Totem Node.

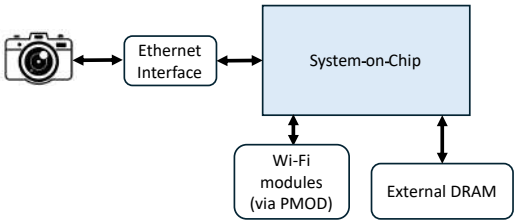

FIGURE 2. Hardware Infrastructure of the Roof Node.

Workload sharing is the key feature that makes the proposed *System* cooperative: as detailed in Section IV-A, a node in the proposed *System* can offload part of its computation to another node while keeping all processing confined within the edge domain.

This design preserves real-time responsiveness while ensuring that no sensitive data leaves the local network, avoiding any interaction with external systems or third parties (e.g., cloud services).

Within the proposed *System*, two types of fractal nodes are defined: the *totem node* and the *roof node*.

The totem node, whose hardware infrastructure is shown in Fig. 1, integrates a camera and a microphone as sensing devices, both connected through an Ethernet interface, and a display as its actuator, connected via DisplayPort. Two PMOD Wi-Fi connectors are also included, enabling the attachment of two Wi-Fi antennas that support the cooperation infrastructure (described in Section IV-A3). At the core of the totem node, a System-on-Chip (SoC) executes all tasks required to satisfy the user requirements defined in the previous section.

Given its intended deployment in shopping malls, the totem node is approximately *180 cm* tall, with the camera and microphone positioned at *150–160 cm* to ensure ergonomic and effective interaction with visitors.

The totem node is responsible for direct human–system interaction and for generating tailored content in real-time. Its tasks include: *Face Detection* (FD), which takes as input frames from the camera and outputs one cropped face image per detected visitor; *Age Estimation* (AE), which takes as input a face image and outputs the estimated age; *Gender Classification* (GC), which takes as input a face image and identifies the expressed gender; *Idiom Recognition* (IR),

which takes as input audio from the microphone and outputs the recognized spoken language; *Rule-Based Recommendation* (RBR), which receives age, gender, and spoken language data and selects the tailored content; *Content Selection* (CS), which displays the selected advertisement on the display; and *Load Balancing* (LB), which determines whether, in case of need, there are available nodes for workload sharing.

The roof node, whose hardware infrastructure is shown in Fig. 2, is installed above the totem node (typically on the ceiling of the shopping mall) and is equipped with a camera connected through an Ethernet interface as its primary sensor. Two PMOD Wi-Fi connectors are also included, enabling the attachment of two Wi-Fi antennas that support the cooperation infrastructure described later in this section. At the core of the roof node, a SoC executes all tasks required to satisfy the user requirements defined in the previous section.

The roof node supervises the area surrounding the totem node, monitors visitor presence, and assists the totem node by executing offloaded tasks when necessary. Its tasks include: *People Detection* (PD), which takes as input frames from its camera and identifies the presence of visitors; and *Density Estimation* (DE), which processes camera frames to produce an estimate of the number of visitors in the monitored area. Depending on the detected situation, the roof node also generates event signals (Alarms 1–5) to trigger the specific reactions (through task executions) on the totem node.

The next subsection details the operational flow through which the totem node and the roof node coordinate to satisfy the user requirements.

A. OPERATIONAL FLOW

The totem node and roof node respond to external stimuli and interact to provide the required AI-driven context awareness in real-time and in an edge-confined manner. The operational flow of these two nodes within the proposed *System* is formalized through finite state machines, shown in Fig. 3 by their associated state–transition diagrams. The left side of Fig. 3 shows the state–transition diagram of the totem node, whereas the right side presents the corresponding diagram for the roof node.

Each state–transition diagram is composed of circles representing the states of the node and directed edges representing the transitions between states. For readability, states are named as T_i and R_i (i = 1, 2, ...) for totem node and roof node, respectively; state names appear in bold in the diagrams. Inside each state, the name of the tasks executed in that state (e.g., *PD*, *DE*) may be shown. In such cases, the indicated tasks are executed in parallel on the corresponding node. Alternatively, a state may contain an action (e.g., *Send Alarm 1*), meaning that the corresponding action is performed in that state. Upon entering a state, the tasks or actions indicated in that state are started, and the node remains in that state until their execution completes. Transitions between states are annotated with events or guard conditions (e.g., detection results, resource availability, or alarm signals) that trigger the state change and govern the progression of each node through its operational workflow.

The *System* operates in two possible modes: *no workload sharing* and *workload sharing*. The active mode is determined by the value of a configurable parameter, *MAXF*, which depends on the hardware characteristics of the totem node (details provided in the Section IV-B). When the number of visitors approaching the totem node is smaller than *MAXF*, the totem node can complete all required tasks within the one-second deadline without assistance from roof nodes, and thus operates in *no workload sharing* mode. Conversely, when the number of detected faces exceeds *MAXF*, the totem node requires computational support from neighboring roof nodes and switches to *workload sharing* mode.

1) Operational Flow without Workload Sharing

In this subsection, the operational flow of the *System* in the case where the number of visitors approaching the totem node is smaller than *MAXF* is described, meaning that the totem node is capable of completing all required tasks within the one-second deadline without performing workload sharing.

The totem node starts in state (**T0**), with its display turned off. The roof node starts in state **R0**, where it announces its availability to assist the totem node through workload sharing (the cooperation infrastructure is described in Section IV-A3).

The roof node then transitions to state **R1**, where it monitors the area in front of the totem node using its dedicated camera. In this state, it continuously executes the PD task to detect the presence of visitors and, in parallel, runs the DE task to estimate crowd density.

From state **R1**, when one or more visitors enter the totem node area, the roof node transitions to state **R3** and sends *Alarm 1* to the totem node. Upon receiving *Alarm 1*, the totem node transitions to state **T1**, powers on the display corresponding to the activated side, and shows a welcome message through the CS task.

The roof node then moves to state **R4**, where it continues monitoring the scene. When a visitor approaches the totem node and enters the proximity area, the roof node enters state **R6** and triggers *Alarm 2*, and then transitions to state **R7**. If instead the visitor moves away from the totem node area, the roof node transitions to state **R5**, sends *Alarm 4* to the totem node (indicating that the display can be turned off), and then returns to state **R1**.

In state **R7**, the roof node continues executing PD and DE. If the visitor moves out of the totem node proximity area but remains within the totem node area, the roof node transitions to state **R8**, sends *Alarm 3*, and then returns to state **R4**.

In states **R1**, **R4**, and **R7**, if the DE task reports a number of visitors greater than a predefined *threshold* (representing an abnormal crowding situation typical of shopping mall environments), the roof node transitions to state **R2**, triggers *Alarm 5*, and then returns to state **R1**. *Alarm 5* is propagated to notify the shopping mall security system about the detected abnormal crowd condition.

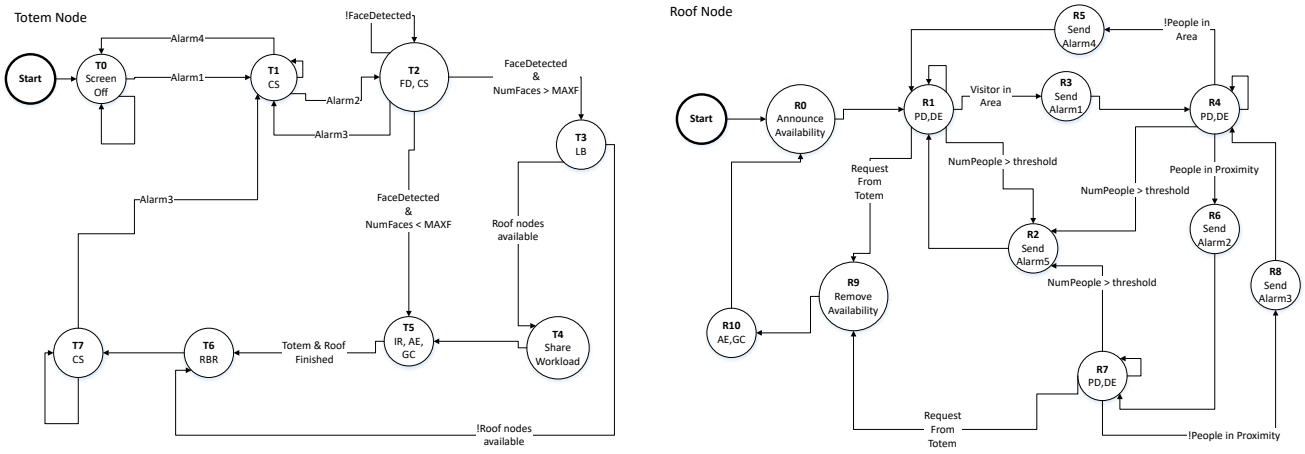

FIGURE 3. Operational Flow among totem and roof nodes.

Upon receiving *Alarm 2*, the totem node transitions to state **T2** and executes the FD task to acquire facial images using its camera, while continuing to display the welcome message through the CS task. If the number of detected faces is smaller than *MAXF*, the totem node transitions to state **T5**, where it executes the IR, AE, and GC tasks. The IR task is executed once per interaction, whereas AE and GC are executed for each detected face.

Once these tasks are complete, the totem node (for completeness) verifies that any potential roof node side computations have also finished, although in this scenario no workload sharing occurs, and transitions to state **T6**. The outputs produced by IR, AE, and GC are aggregated and passed to the RBR task, which selects the tailored content that best matches the inferred visitor profile.

Finally, in state **T7**, the CS task renders the selected tailored content on the totem node display.

It is assumed that once the visitor receives the tailored content, they eventually leave the totem node proximity area. When this occurs, the roof node transitions to state **R8**, generates *Alarm 3*, and then returns to state **R4**; the totem node receives *Alarm 3* and returns to state **T1**. If the same visitor approaches the totem node again, the operational flow is repeated.

If visitors leave the entire totem node area, the roof node transitions to state **R5** and issues *Alarm 4*, after which it returns to state **R1**. This causes the totem node to switch off its display and return to state **T0**, thus reducing energy consumption when no visitors are present.

This concludes the description of the operational flow for the case in which the totem node operates entirely autonomously, without requiring workload sharing from the roof nodes.

2) Operational Flow with Workload Sharing

In this subsection, the operational flow of the *System* in the case where the number of visitors approaching the totem node exceeds *MAXF* is described. In this condition, the totem node is not able to complete all required computations within the one-second deadline on its own and therefore requires computational support from roof nodes through workload sharing.

As described in the previous subsection, the roof node starts in state **R0**, where it announces its availability to support the totem node, and then moves to state **R1**. When visitors approach the totem node, the roof node eventually reaches state **R7**, while the totem node reaches state **T2** and completes the FD task.

At this point, the totem node determines that the number of detected faces is greater than *MAXF*. It therefore transitions to state **T3**, where it executes the LB task to identify whether neighboring roof nodes are available for workload sharing. If one or more roof nodes are available, the totem node prepares an offloading request and transitions to state **T4**, where it sends the input data required for the AE and GC tasks (i.e., the cropped facial images to be processed) to the selected roof node.

After this transmission, the totem node moves to state **T5**, where it performs its own instances of the AE and GC tasks on the remaining subset of images (i.e., those not offloaded to the roof node).

Upon receiving the offloading request, the roof node transitions to state **R9** and withdraws its availability announcement. This is done to ensure that once a roof node has been allocated to support workload sharing, it cannot be selected again until the current offloading session is completed. The roof node then moves to state **R10**, where it executes its own AE and GC tasks on the offloaded images.

States **T5** and **R10** represent the core of the workload-sharing mechanism: the totem node and the roof node execute AE and GC computations in parallel, thereby reducing the overall processing time and ensuring compliance with the one-second deadline.

Once the roof node completes its assigned computations,

it sends the results back to the totem node and transitions to state **R0**, where it re-announces its availability. Meanwhile, the totem node waits for the roof node results and for its own local computations to finish. After receiving and aggregating all outputs, it transitions to state **T6**, where the RBR task selects the tailored content to display. Finally, in state **T7**, the CS task projects the selected tailored content on the totem node display.

It is worth noting that if, in state **T3**, the LB task determines that no roof nodes are available, the totem node becomes aware that it cannot complete the required computations within the one-second deadline. As a consequence, it cannot produce a fully AI-driven context-aware advertisement in time. In this case, the finite state machine transitions directly from state **T3** to state **T6**. The RBR task, informed that the system reached state **T6** via the LB path (i.e., without performing AE and GC on all detected faces), selects a fallback content strategy. Specifically, instead of delivering a tailored advertisement, the RBR task chooses a general, non-tailored message designed to mitigate the lack of timely AI-driven context-aware processing (e.g., a generic promotional banner).

3) The Cooperation Infrastructure Mechanism

As described in the previous paragraph, when the number of detected faces exceeds *MAXF*, the totem node verifies whether roof nodes have available computational resources (state **T3**). This state is also referred to as *neighbor discovery*. If one or more roof nodes are available, the totem node initiates a workload–sharing procedure to offload part of the computation (state **T4**), a state also referred to as *task offloading*.

Neighbor discovery is performed using the standard Wi-Fi infrastructure available in the shopping mall, which allows nodes to publish and subscribe to availability information over the existing network. At the transport layer, the Transmission Control Protocol (TCP) is used, while at the application layer, the Message Queuing Telemetry Transport (MQTT) protocol is adopted [51]. MQTT is a publish/subscribe protocol that enables roof nodes to broadcast their computational availability by publishing messages to predefined topics. Each node acts as an MQTT client and subscribes to these topics to receive updates from other nodes within Wi-Fi range.

In the proposed cooperative infrastructure, a dedicated MQTT broker runs on a separate device, managing all published messages and maintaining the current availability status of neighboring nodes. Importantly, the messages exchanged over MQTT contain only resource–availability information and no sensitive visitor data. Thus, the involvement of an external device running the MQTT broker does not jeopardize the edge-confined processing strategy and does not expose sensitive data to third parties.

Furthermore, the Wi-Fi/MQTT-based infrastructure is also employed by roof nodes to communicate *Alarm 1*, *Alarm 2*, *Alarm 3*, and *Alarm 4* to the totem node, as well as *Alarm 5* to the shopping mall security system.

Once a suitable neighbor node is identified, task offloading is carried out using Wi-Fi Direct, which enables direct peer-to-peer communication between nodes without the involvement of intermediate access points. This mechanism preserves edge-confined processing, as all exchanged data remains within the local network domain, and reduces latency. At the transport layer, TCP is again employed, while at the application layer, the Hypertext Transfer Protocol (HTTP) is used to transmit task data and to receive the corresponding results.

In the proposed *System*, each roof node connects to the MQTT broker to announce its resource availability (state **R0**) and to receive potential offloading requests (transition "Request from Totem" in Fig. 3). When the totem node determines that offloading is required, it queries the available nodes via MQTT (state **T3**) and initiates workload sharing with those offering free resources (state **T4**).

During this process, the totem node transmits the cropped facial images to the assisting roof nodes, which execute the AE and GC tasks on behalf of the totem node. Once the computations are completed, each assisting roof node returns the results to the totem node, which aggregates all received outputs and forwards them to the RBR task (state **T6**). The RBR task then selects the tailored content, which is finally displayed through the CS task (state **T7**).

Through this cooperative workflow, the proposed *System* guarantees the execution of AI workloads within the one-second real-time deadline, while maintaining complete edge-confined data handling and avoiding any involvement of external systems or third parties (e.g., cloud services).

B. TASK IMPLEMENTATION

This section describes the implementation details of the tasks executed by the totem node and the roof node. Furthermore, it presents the hardware infrastructure within the SoC embedded in both the totem and roof nodes, as used in the prototype, and defines the value of *MAXF*, which specifies the maximum number of faces that the totem node can process locally within the one-second deadline.

1) Totem Node

The totem node executes the tasks *Face Detection* (FD), *Age Estimation* (AE), *Gender Classification* (GC), *Idiom Recognition* (IR), *Rule-Based Recommendation* (RBR), *Content Selection* (CS), and *Load Balancing* (LB).

a: Face Detection (FD) Task

This task processes a continuous video stream from the totem node camera to determine whether one or more individuals are standing in front of the totem node and to locate their faces.

It takes as input RGB images of size 320×320 pixels captured by the totem node camera and produces cropped outputs of size 240×200 pixels, each corresponding to a

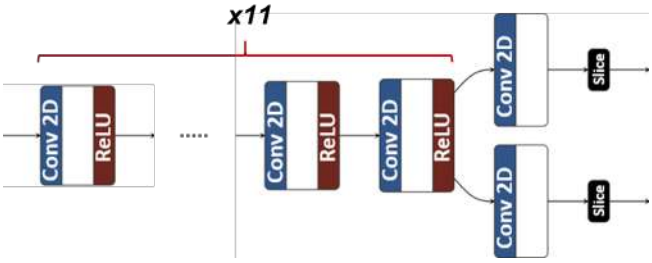

FIGURE 4. Neural network model architecture of the FD task.

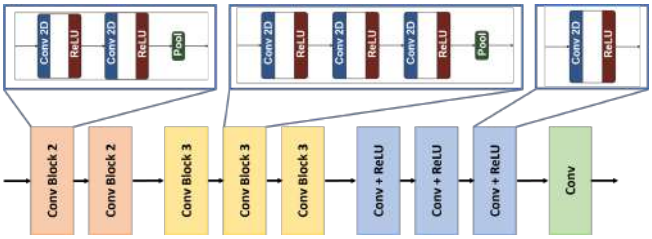

FIGURE 5. Neural network model architecture of AE and GC tasks.

detected facial region. These cropped images are forwarded to the subsequent age and gender analysis tasks (AE and GC). The FD task is triggered upon reception of *Alarm 2* from the roof node.

FD relies on AI-based algorithms, specifically on DenseBox [52], a fully convolutional neural network designed for real-time face detection, an essential requirement for the totem node. DenseBox eliminates the need for region–proposal stages and directly detects faces with high efficiency, making it well-suited for edge devices operating under strict timing constraints. The network architecture is shown in Fig. 4.

The neural network is developed using the Caffe framework [53] and the task is implemented in C++. The neural network is trained on the WIDER Face dataset [54], which contains a large variety of facial images exhibiting diverse poses, scales, lighting conditions, and occlusions. This variability makes it an ideal dataset for developing face detectors that must operate reliably in shopping malls.

b: Age Estimation (AE) and Gender Classification (GC) Tasks

These two tasks estimate the visitor *age* and *gender* (interpreted as the expressed gender) from the cropped facial images produced by the FD task. Both tasks take as input images of size 240×200 pixels and produce textual numerical outputs, which are subsequently forwarded to the RBR task.

The AE task is formulated as a regression problem, whereas the GC task is addressed as a binary classification problem. Both models are implemented as convolutional neural networks based on the VGG16 architecture [55], specifically the VGG-Face variant [56], selected for its proven transferability and high accuracy in face-related recognition tasks. The network architecture is shown in Fig. 5. The neural networks used for AE and GC share the same architecture, differing only in the output layer.

The tasks are implemented in Python using TensorFlow and Keras, and trained offline using the MORPH II dataset [57], which provides 55135 facial images annotated with age and gender. This dataset is widely considered one of the most comprehensive benchmarks available for age and gender estimation tasks (see [44] for further discussion).

To reflect the insights derived from the analysis of the context of use (Section III-A), which indicates that men and women visit shopping malls in approximately equal proportions [58], the original MORPH dataset (originally imbalanced with about 85% male and 15% female images) was rebalanced [59]. Data normalization and random sampling techniques were applied to obtain a gender-balanced subset. The resulting dataset, corresponding to approximately 30% of the original, contains 16978 images, equally divided between male and female subjects.

c: Idiom Recognition (IR) Task

The IR task processes audio data acquired from a microphone integrated into the totem node camera to identify the spoken language of the visitor.

It takes audio files in WAV format as input and produces a string corresponding to the detected spoken language as output. The identified language is then provided to the RBR task, which uses it to adapt the tailored content displayed on the totem node display.

This task combines deep learning and statistical approaches to ensure robust accuracy even in noisy environments and across diverse accents, as emerged from the analysis of the context of use. For speech-to-text (STT) approaches, neural network-based automatic speech recognition models (Whisper [60] and Wav2Vec 2.0 [61]) are used to transcribe speech before linguistic pattern analysis. In non-STT approaches, acoustic features are extracted using Gaussian mixture models, hidden Markov models, and support vector machines to classify languages based on phonotactic characteristics. This hybrid strategy enables accurate spoken language detection and stable performance across different visitor profiles. The task is implemented in Python.

d: Rule-Based Recommendation (RBR) and Content Selection (CS) Tasks

The RBR task aggregates the outputs of AE, GC, and IR (i.e., age, gender, and spoken language) and determines which tailored content should be shown on the display of the totem node.

It takes as input a JSON-formatted file containing the inferred visitor profile and applies a rule-based decision model to select the tailored content.

The result of this decision process is forwarded to the CS task, which renders the selected tailored content on the corresponding display.

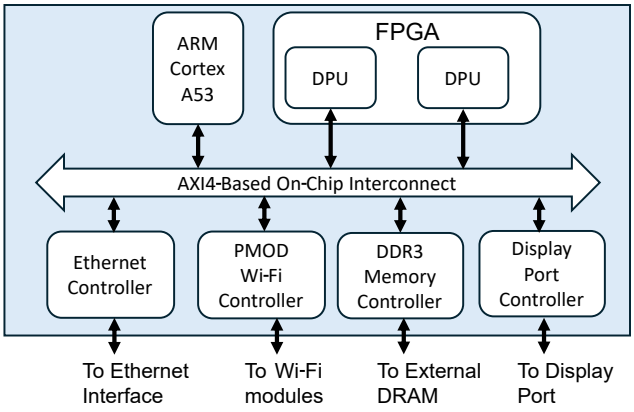

FIGURE 6. SoC architecture implemented on both the totem node and the roof node. The SoC integrates a quad-core ARM Cortex-A53 processor cluster and an FPGA fabric hosting two *Deep Learning Processing Unit* (DPU) cores, interconnected through an AXI4-based on-chip interconnect. Two PMOD Wi-Fi controllers (only one is depicted in the figure for readability) enable both standard Wi-Fi and Wi-Fi Direct communication with other nodes. An external DDR3 DRAM module is shared through the DDR3 memory controller, providing storage for all processing tasks. The Ethernet controller interfaces with the cameras (and the microphone in the totem node), while the DisplayPort controller enables video streaming to the totem node display.

e: Load Balancing (LB) Task

The LB task is responsible for determining whether there are roof nodes available to support the totem node in completing all computations within the one-second deadline. It does not perform the full workload-sharing procedure; rather, it constitutes the decision-making step that checks for the presence of helper nodes and selects which of them should be involved in the subsequent offloading phase.

The LB task is designed to be *reactive* rather than *proactive*. In the context of shopping malls, visitor arrivals are highly dynamic and unpredictable, making it inefficient to reserve resources in advance or maintain pre-allocated computation slots on the roof nodes. Proactive reservation would lead to unnecessary resource locking and reduced overall system efficiency, as roof nodes might remain idle while being marked as unavailable. A reactive strategy, instead, evaluates resource availability only when the totem node detects a workload that exceeds its local processing threshold (*MAXF*).

A detailed description of the full workload-sharing procedure and the algorithm governing the LB task is provided in [28].

Hardware Infrastructure within the System-on-Chip

All the tasks on the totem node are deployed on an *AMD Xilinx Zynq UltraScale+* SoC device [62]. This heterogeneous platform integrates a quad-core ARM Cortex-A53 processor cluster together with a programmable logic fabric that hosts two *Deep Learning Processing Unit* (DPU) cores [63], which are hardware accelerators optimized for neural-network inference. The SoC architecture implemented on the totem node is shown in Fig. 6.

The SoC is soldered on a *ZCU102* development board. The SoC corresponds to the light-blue block highlighted in Fig. 1, while the ZCU102 board provides the external interfaces required by the prototype, including the Ethernet interface, the PMOD connectors for the Wi-Fi modules [64], the external DRAM, and the DisplayPort output, as shown in Fig. 1.

The quad-core ARM Cortex-A53 runs a Linux operating system.

Task allocation across the hardware resources is as follows: the RBR, IR, CS, and LB tasks execute on the ARM cores, while the FD, AE, and GC tasks are accelerated on the two DPU cores. Specifically, the two DPU cores are contented by the execution of FD, AE, and GC. The IR task uses all four ARM cores to efficiently handle audio input and AI inference. The scheduling of tasks and the implementation of the finite state machine shown in Fig. 3 are carried out by a dedicated runtime manager, developed on top of the Linux operating system.

With this hardware configuration, the value of *MAXF* is set to 2. This value was determined empirically by measuring the response time of the individual tasks on the prototype platform. Further details on this evaluation are provided in Section V.

2) Roof Node

The roof node executes two tasks: *People Detection* (PD) and *Density Estimation* (DE).

a: People Detection (PD) Task

The PD task analyzes the video stream captured by the roof node camera to identify and track the presence of visitors within the totem node area.

Each video frame, originally acquired in Full HD (1920×1080 pixels), is cropped to a square region and resized to 300×300 pixels. This pre-processing step preserves the spatial context surrounding the totem node while reducing the computational load, enabling real-time inference on an edge-computing platform.

Based on the detected pedestrian positions, the PD task generates the alarms discussed in Section IV-A (Alarm 1 to Alarm 4), which are forwarded to the totem node.

The PD task is based on a customized YOLOv5n model [65], retrained through transfer learning to specialize from the original 80 COCO object classes to a single *pedestrian* class. This specialization is motivated by the fact that the standard YOLOv5n detector includes classes such as bicycle, car, dog, and chair, which are irrelevant for the shopping mall environment. Restricting the model to a single class reduces the output dimensionality and improves both inference speed and accuracy in crowded indoor environments (as it can be in a shopping mall environment).

The customized training dataset combines approximately 40000 pedestrian images from COCO [66], 40000 from the Open Images Dataset [67], and 10000 images captured directly by the roof node cameras in the target shopping mall environment. All local images were manually annotated using

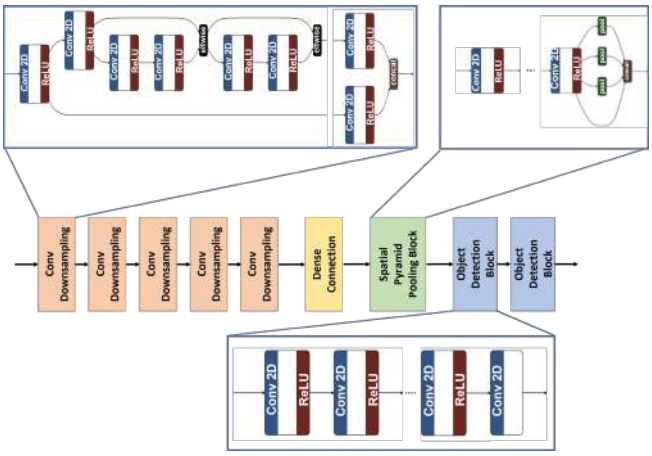

FIGURE 7. YOLO-based model used by DE and PD tasks.

the VGG Image Annotator tool. The PD task is implemented in C++.

b: Density Estimation (DE) Task

The DE task receives a video stream from the roof node camera and estimates the density of visitors within the monitored area.

It takes as input images of size 416×416 pixels, obtained by cropping and resizing the original Full HD (1920×1080) frames. The DE task outputs a numerical value indicating the estimated number of visitors in the scene. This value is compared against a threshold (as discussed in Section IV-A); if the threshold is exceeded, the roof node triggers *Alarm 5* to notify the mall security system of a potential crowding condition.

DE is implemented using the YOLOv4 convolutional neural network [68], a state-of-the-art model for real-time object detection whose architecture is similar to that shown in Fig. 7. YOLOv4 was selected due to its favorable trade-off between accuracy and computational load, making it suitable for continuous operation on edge devices. The model was developed using the Darknet framework [69], it was trained on the COCO dataset [66], and deployed in C++.

Hardware Infrastructure within the System-on-Chip

All the tasks on the roof node are deployed on an *AMD Xilinx Zynq UltraScale+* SoC device [62]. The SoC architecture implemented on the roof node is similar to the one used for the totem node and shown in Fig. 6, with the only difference that the roof node does not include a DisplayPort controller. As in the totem node, the Zynq UltraScale+ device is mounted on a *ZCU102* development board, which provides all external interfaces required by the prototype.

The quad-core ARM Cortex-A53 runs a Linux operating system identical to that deployed on the totem node.

The PD task executes directly on the camera unit, enabling continuous operation with minimal latency. The DE task runs on one of the DPU cores, benefiting from hardware acceleration for convolutional neural networks. The scheduling of tasks and the implementation of the finite state machine shown in Fig. 3 are carried out by a dedicated runtime manager developed on top of the Linux operating system.

V. EXPERIMENTAL RESULTS

The experimental activities aim to verify whether the proposed *System* satisfies the functional and non-functional requirements reported in Section III-B2.

To this end, a prototype composed of one totem node and one roof node was implemented and evaluated in a laboratory environment emulating a real shopping mall. The hardware platform consists of two *ZCU102* development boards, each hosting an *AMD Xilinx Zynq UltraScale+* SoC. For both boards, a Linux-based operating system was built using PetaLinux 2021.2 [70]. The system software was developed with the Vitis 2021.2 toolchain [71], and the deep-learning workloads were deployed on the DPU cores using Vitis AI 2.5 [72]. The overall prototype architecture is shown in Fig. 8, which shows the interconnection between the two nodes and the MQTT broker running on a Raspberry Pi board [73].

All experiments were conducted within the facilities of *Aitek*. A snapshot of the experimental setup is shown in Fig. 9, where the green rectangle outlines the total area observed by the roof node, the red polygon marks the totem node area, and the yellow bounding box highlights the detected visitor.

The objectives of the experimental evaluation are threefold:

(1) **Verification of functional requirements.** This objective consists of two parts. First, the correct execution of all tasks is verified, both in isolation and when invoked according to the sequence defined by the finite state machine in Fig. 3. Ensuring that tasks do not block or hang is essential, particularly because the proposed implementation executes multiple AI workloads on an edge device with limited computational resources, where susceptibility to overload is inherently higher [74]. Second, the quality of the tasks FD, PD, DE, AE, GC, and IR is evaluated by measuring their inference accuracy on dedicated test sets, as detailed in Section V-A.

(2) **Verification of non-functional requirements.** This objective involves assessing the real-time responsiveness of the system and confirming that the overall processing pipeline remains below the one-second limit, as required by the real-time constraint. Section V-B reports the associated results.

(3) **Scalability evaluation.** This objective involves assessing how the proposed *System* behaves as the number of visitors increases, and how the response time behaves when additional roof nodes are introduced to support workload sharing. Section V-C reports the associated results.

The experimental campaign was conducted under both operational flows described earlier: (i) the operational flow *without workload sharing* (see Section IV-A1), and (ii) the operational flow *with workload sharing* (see Section IV-A2).

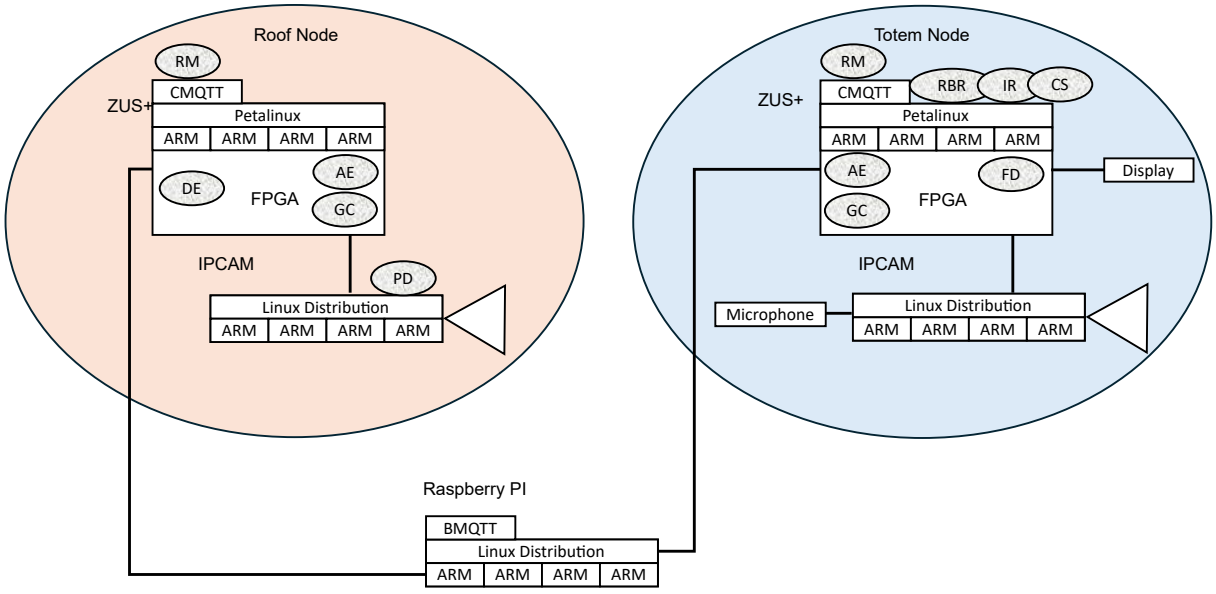

FIGURE 8. Prototype architecture. ZUS+ refers to the Zynq UltraScale+ SoC, while *BMQTT* and *CMQTT* denote the MQTT broker and MQTT client, respectively.

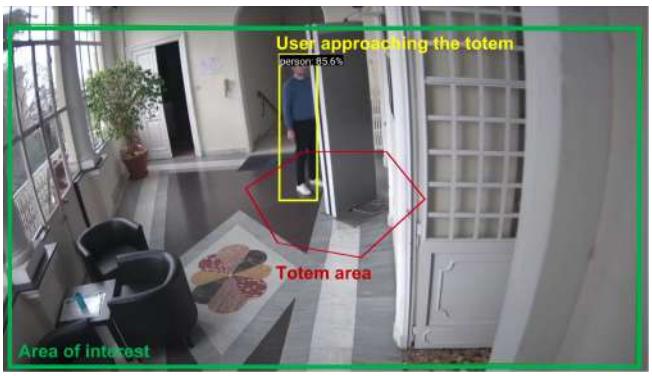

FIGURE 9. Snapshot of the prototype installed at *Aitek* facilities. The view is captured from the roof node camera. The green rectangle outlines the total area of interest, the red polygon defines the totem node area, and the yellow box marks the detected visitor.

It is worth noting that operating without workload sharing requires that no more than *MAXF* visitors (with $MAXF = 2$ in our prototype, see Section IV-B) enter the totem node proximity area. Conversely, the operational flow with workload sharing is triggered whenever more than *MAXF* visitors approach the totem node.

In both operational modes, the test setup reflected a realistic crowded scenario: several individuals were present within the broader totem node area, as commonly observed in shopping mall environments.

A. VERIFICATION OF FUNCTIONAL REQUIREMENTS

1) Correctness of Task Execution

In order to verify the correctness of task execution, each task was deployed on its corresponding hardware component as described in Section IV-B, and the correct execution flow was validated in both operational flows. By correct execution flow, it is referred to as the ability of each task to receive its expected inputs, produce an output, and operate in a stable manner over repeated executions [75]. This phase did not involve any evaluation of inference accuracy (which is addressed in the next subsection); rather, it focused on ensuring that the system components were properly scheduled and coordinated by the runtime manager (namely, the software component responsible for coordinating tasks and executing the finite state machines of Fig. 3), and that the corresponding finite state machines could execute without blocking or instability.

The experiments revealed that the runtime manager needed to be explicitly pinned to one core of the ARM Cortex-A53 processor to ensure stable and predictable timing performance. As shown in Fig. 8, the runtime manager is another task on both the totem node and the roof node. Without this configuration, the Linux scheduler occasionally descheduled the runtime manager, causing temporary stalls or latency fluctuations in the overall execution pipeline.

In addition, bandwidth-regulation mechanisms [76] were introduced on the target architecture to control the memory-access rate of bandwidth-intensive workloads, such as the FD task. This step proved essential: without bandwidth regulation, preliminary tests showed that the CS component experienced intermittent display refresh issues due to memory-bus contention, negatively affecting the visitors experience. After applying bandwidth regulation, all components executed correctly, and the system maintained a stable and continuous behavior in both operational flows.

2) Task Accuracy

The accuracy was evaluated separately for the FD, PD, DE, IR, GC, and AE tasks. Since all these tasks rely on AI-based

TABLE 2. Accuracy and required operations (OPS) per image for the FD task using two network configurations distinguished by input size.

Input Size	OPS (per image)	Float Acc.	Quant. Acc.
320×320	0.49G	0.8833	0.8783
360×640	1.11G	0.8931	0.8922

models (see Section IV-B), the reported accuracy corresponds to the inference performance of the underlying neural networks used by each task. For each task, the test sets were constructed from a combination of samples from the corresponding public datasets (as detailed in Section IV-B) and samples collected directly from the prototype environment. More specifically, these samples are constituted of images (for FD, AE, GC, DE, and PD) and audio recordings (for IR) acquired using the setup shown in Fig. 9. Furthermore, the response time of each task in isolation was measured to characterize its standalone timing performance.

a: Face Detection (FD) Task

The accuracy of the FD model (part of the FD task) was evaluated using the FDDB dataset [77], which was selected for its rich, diverse set of annotated faces captured under unconstrained real-world conditions. Unlike more controlled benchmarks, FDDB includes faces with substantial variation in pose, illumination, scale, and occlusion, conditions that closely reflect the visual challenges encountered in shopping malls. For these reasons, FDDB represents an appropriate benchmark for assessing the robustness of the FD model in the target context of use.

Two network configurations were evaluated, differing only in input resolution: one processes 320×320 pixel images, while the other uses a larger 360×640 pixel input. Table 2 reports the accuracy obtained by both models when deployed on the totem node. The columns labeled *Float Acc.* and *Quant. Acc.* refer respectively to the accuracy achieved with the floating-point model and with its quantized counterpart executed on the DPU accelerator. Quantization was applied to improve inference speed and reduce memory footprint, while preserving accuracy as much as possible.

The results show that the accuracy difference between the two input resolutions is negligible, whereas the computational cost (measured in operations per image) nearly doubles for the higher-resolution model. For this reason, the 320×320 configuration was adopted in the final implementation.

Regarding the response time, the FD task is executed on the DPU accelerator and achieves an average response time in isolation of 0.2 ms per image.

b: People Detection (PD) Task

For object detection tasks such as PD, traditional accuracy metrics are not meaningful; for this reason, standard detection metrics are reported, namely precision, recall, and mean Average Precision (mAP), which are the accepted indicators of detection performance [78].

TABLE 3. Precision and Recall metrics related to the PD model (part of the PD task).

Metric	Description	Value
Precision (BoxP)	Correct detections / total detections	0.877
Recall (R)	Correct detections / total ground truths	0.777
mAP@50	Mean Average Precision at IoU 0.5	0.869
mAP@50–95	Mean Average Precision at IoU 0.5–0.95	0.595

The precision, recall, and mAP values for the PD model (part of the PD task) were evaluated on a dataset of 5486 images containing a total of 11696 annotated visitor instances. Table 3 reports the corresponding performance metrics. The model achieves a precision (BoxP) of 0.877 and a recall (R) of 0.777, indicating a good balance between correctly detected instances and missed detections. The mAP at an IoU threshold of 0.5 is 0.869, while the stricter mAP averaged across IoU thresholds from 0.5 to 0.95 is 0.595.

Regarding response time, the PD task is executed on the hardware of the roof node camera and exhibits an average pre-processing time of 0.1 ms, inference time of 3.2 ms, and post-processing time of 1.2 ms per image, for a total average response time of approximately 4.5 ms per frame. These results confirm that the model achieves high detection performance while maintaining very low latency.

c: Density Estimation (DE) Task

The accuracy of the DE model (part of the DE task) was evaluated using the COCO2014-5k subset [66]. This dataset was selected because it provides a large number of diverse indoor and outdoor scenes with significant variations in crowd density, scale, illumination, and occlusion. Such variability makes COCO2014-5k a widely accepted benchmark for evaluating density-related tasks and ensures that the DE model is tested under conditions representative of real shopping mall environments.

Two network configurations were evaluated, both using 416×416 input images but differing in the required computational complexity. The first configuration corresponds to the original quantized model, which requires 60.1 GOPS per image, while the second configuration applies pruning techniques to reduce complexity to 38.2 GOPS. Table 4 reports the quantitative results.

In terms of accuracy, the quantized version achieves a mAP@50–95 of 0.3730, slightly outperforming the pruned configuration, which reaches 0.3590. These results indicate that pruning minimally affects detection accuracy while significantly reducing computational demand, making the pruned model the preferred choice for deployment in the prototype.

Regarding the response time, the DE task is executed on the DPU accelerator and achieves an average response time in isolation of 200 ms per image.

TABLE 4. Accuracy and required operations (OPS) per image for the DE model (part of the DE task).

Input Size	OPS (per image)	Float Acc. (mAP@50–95)	Quant. Acc. (mAP@50–95)
416×416	60.1G	0.3950	0.3730
416×416	38.2G	0.3810	0.3590

TABLE 5. Recognition accuracy of the IR model (part of the IR task) for different hot words.

Hot Word	Language	Recognition Accuracy [%]
Buongiorno	Italian	87.5
Ciao	Italian	62.5
Hello	English	100.0
Hi	English	87.5
Average		**84.4**

d: Idiom Recognition (IR) Task

The accuracy of the IR model (part of the IR task) was evaluated using two target languages: English and Italian. These languages were chosen because they represent the most common idioms spoken by visitors in the commercial malls where the proposed *System* is intended to be deployed. Italian is the primary local language, while English is widely used as the default international communication language in public environments, tourism, and customer-facing services. Focusing on these two spoken languages provides a realistic and representative assessment of the behavior of the IR model in the target environment.

For the test set, utterances were recorded from eight individuals, including both male and female speakers, each greeting the totem node in different ways. Specifically, the test set consisted of 32 WAV audio files containing the hot words *"Buongiorno"* and *"Ciao"* for Italian, and *"Hello"* and *"Hi"* for English. The per-class recognition accuracy is summarized in Table 5. Incorrect classifications were not due to language misidentification but rather to the inability to detect any of the predefined hot words, typically caused by noisy audio or unclear pronunciation.

The results show that the highest accuracy was obtained for the hot word *"Hello"*, whereas the lowest was observed for *"Ciao"*. Overall, the IR model correctly identified the spoken language in 84% of cases.

Regarding the response time, the IR task is executed on the ARM Cortex A53 processor and achieves an average response time in isolation of 100 ms per utterance.

e: Gender Classification (GC) Task

In evaluating the accuracy of the GC model (part of the GC task), it is important to assess potential sources of bias in model accuracy, particularly with respect to ethnicity and age. In a shopping mall environment, visitors are expected to represent diverse ethnic and age groups; therefore, any imbalance in the training dataset can lead to systematic bias in the model inference. If certain demographic categories

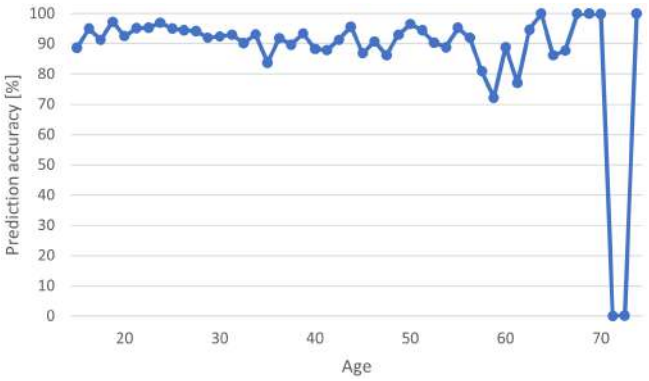

FIGURE 10. GC accuracy (%) across age groups ranging from 10 to 70 years.

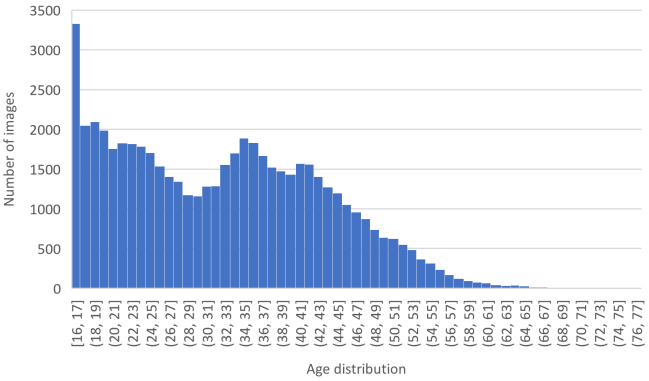

FIGURE 11. Training set distribution for AE and GC. The dataset is notably underrepresented beyond 60 years.

are underrepresented during training, the model is likely to exhibit reduced accuracy for those groups.

Fig. 10 shows the gender classification accuracy of the AI model across different age ranges. The model shows consistently high accuracy for individuals aged between 18 and 55, achieving an accuracy exceeding 85%, which confirms its ability to generalize effectively across younger and middle-aged visitors. However, a marked decline in accuracy is observed in the 60–65 age range, where accuracy drops to 0%. This decrease can be attributed to the limited number of training samples for this age group. Fig. 11 shows the distribution of training samples by age and gender, highlighting underrepresented age groups, particularly beyond 60 years.

Fig. 12 reports the gender classification accuracy across different ethnic groups, together with their respective representation in the dataset. The results reveal a correlation between accuracy and the proportion of samples per group. Higher accuracy is achieved for well-represented groups (White (94.86%), Hispanic (92.20%), and Black (91.76%)), while lower accuracy is observed for underrepresented categories such as Asian (77.78%) and Other (75%). This evidence confirms that dataset imbalance affects generalization and suggests that expanding the diversity of the training data would improve fairness and robustness in real-world shopping mall environments.

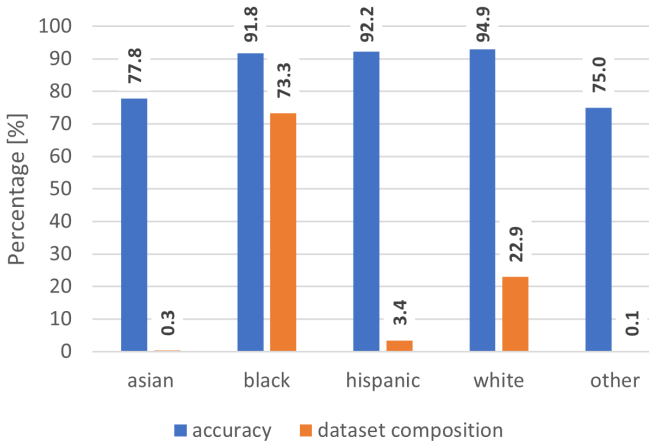

FIGURE 12. GC accuracy across different ethnic groups, alongside the corresponding dataset composition.

Regarding the response time, the GC task is executed on the DPU accelerator and achieves an average response time in isolation of 100 ms per image.

f: Age Estimation (AE) Task

For the AE model (part of the AE task), as for the GC model, it is important to analyze potential sources of bias in model performance.

For the AE model, accuracy is not a meaningful performance metric, since age estimation is formulated as a regression problem in which the output is a continuous numerical value (see Section IV-B). Therefore, the objective is not to predict an exact class label, but to minimize the deviation between the estimated age and the ground-truth age. For this reason, and in accordance with standard practice in the age estimation literature [79], performance was evaluated using the prediction error rather than accuracy. Fig. 13 reports the per–age-group error rate, which provides a more informative view of the model behaviour across the age spectrum. This representation highlights age-dependent performance variations and reveals the impact of dataset imbalance. In particular, Fig. 13 shows the percentage error rate of the age estimation model across different age groups. The results show a clear U-shaped trend, with higher classification errors occurring at both extremes of the age spectrum. Specifically, the model shows elevated error rates for younger individuals (under 20) and older adults (over 60), with error exceeding 30%. Conversely, the model achieves its best performance among middle-aged individuals (35–45 years), with error rates below 15%, demonstrating strong generalization in this demographic range.

As shown in Fig. 11, these findings indicate that the model struggles particularly with age groups that are underrepresented in the training dataset, emphasizing the importance of balanced data distribution for accurate age estimation.

Regarding the response time, the AE task is executed on the DPU accelerator and achieves an average response time in isolation of 100 ms per image.

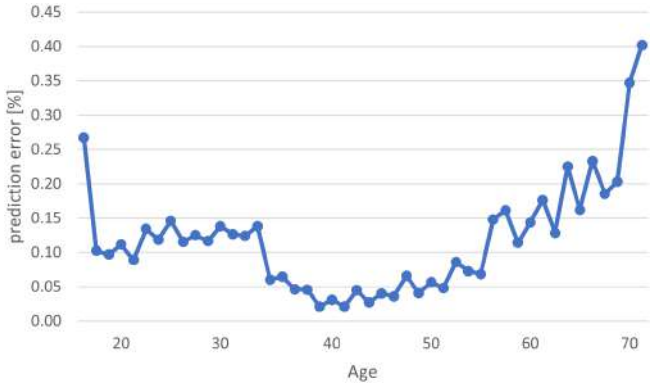

FIGURE 13. Percentage error rate of the AE task across different age groups.

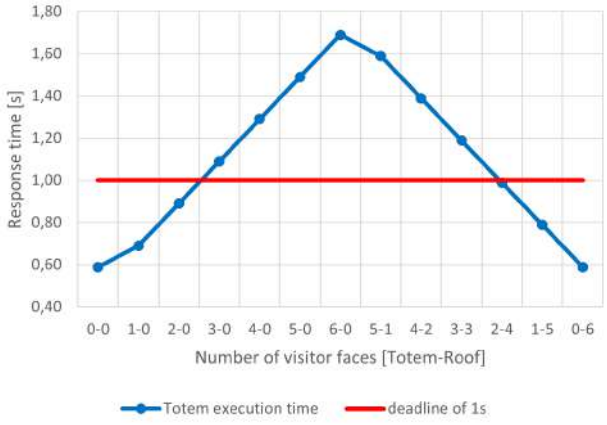

FIGURE 14. System response time in seconds, varying the number of visitors in front of the totem node with a different load distribution between the totem node and the roof node.

B. VERIFICATION OF NON-FUNCTIONAL REQUIREMENTS

Fig. 14 reports the response time of the task execution pipeline on the totem node, measured from the completion of the FD task (state **T2** in Fig. 3) to the projection of the tailored content by the CS task (state **T6**). The X-axis represents the workload distribution using the notation $(t–r)$, where t denotes the number of visitor faces processed locally on the totem node and r denotes those processed remotely on the roof node.

For example, the configuration *2–0* indicates that two visitors are present in the totem node proximity area and the totem node processes both faces locally. As expected, the response time remains below the one-second deadline, confirming that the *System* correctly handles up to two faces without requiring workload sharing.

The configuration *3–0* is included to show that, with three visitors, local processing alone is no longer sufficient to meet the one-second deadline. This empirical result motivates the choice of $MAXF = 2$ in the proposed *System*, as discussed in Section IV-B.

The remaining data points in Fig. 14 show the results obtained when varying both the distribution of computation

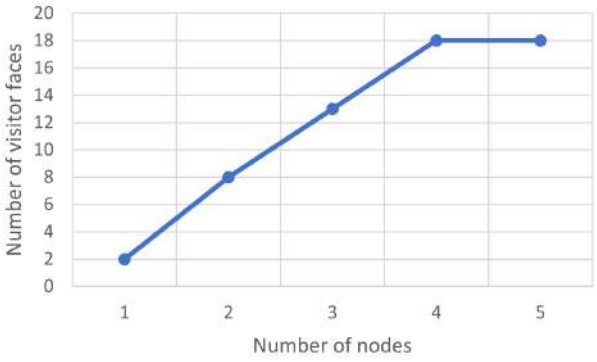

FIGURE 15. Number of visitor faces that can be processed in 1 second, varying the number of roof nodes.

between the totem and the roof node and the number of visitors in the totem node proximity area. It is worth noting that one facial image corresponds to each detected visitor.

Although the state–transition diagram in Fig. 3 represents only one offloading request for readability, the prototype runtime manager deployed on the roof node was configured to handle up to six simultaneous requests from the totem node. With this configuration, the pair composed of one totem node and one roof node can process up to eight visitors in the totem node proximity area while still meeting the one-second deadline, thus enabling tailored content generation for groups of up to eight visitors.

Overall, the cooperative behavior between nodes is clearly visible in the results: as soon as the local workload exceeds the capability of the totem node (*MAXF*), the roof node absorbs the additional computational demand, keeping the end-to-end response time within the required real-time bound.

C. SCALABILITY

Scalability tests were conducted to evaluate how the proposed *System* behaves as the number of visitors in the totem node proximity area increases beyond eight. Indeed, up to eight visitors can be handled by the pair composed of the totem node and a single roof node. The goal of this evaluation is to determine how many visitors can be managed simultaneously within the totem node proximity area while keeping the overall response time below the one-second deadline, assuming additional roof nodes can be added to assist with workload sharing.

To this end, the number of cooperating roof nodes was progressively increased, enabling the *System* to distribute computational workloads dynamically across additional nodes when required.

Fig. 15 reports the scalability results. The Y-axis shows the maximum number of visitor faces that can be processed while meeting the one-second deadline, whereas the X-axis indicates the total number of active nodes in the system (the sum of the totem node and roof nodes). For example, when more than eight faces must be processed, the workload needs to be distributed across more than two nodes (specifically, one totem node and two roof nodes) to remain within the real-time constraint. With a total of four nodes (one totem node and three roof nodes), the *System* can deliver tailored content to up to 18 individuals simultaneously.

D. DISCUSSION

This section discusses the experimental results from multiple perspectives, analyzing perception accuracy, real-time responsiveness, scalability limits, privacy implications, and the applicability of the proposed architecture to other deployment scenarios.

a: Perception accuracy

The experimental results show that the proposed *System* achieves high accuracy across all perception tasks (FD, PD, DE, AE, GC, and IR), enabling robust AI-driven context awareness in realistic shopping mall conditions. The FD task reliably acquires facial images even in the presence of clutter and background motion; AE and GC exhibit strong performance across age and gender ranges well represented in the training dataset; and IR achieves high reliability for the selected idioms. A known limitation arises for underrepresented age groups (e.g., elderly individuals aged 60+), where the AE task shows reduced accuracy due to dataset imbalance. This behavior is consistent with the literature on deep learning models, which shows that they are strongly affected by the distribution of the training data.

b: Real-time responsiveness

Regarding real-time responsiveness, the prototype consistently satisfies the one-second deadline established by usability requirements, as long as the number of visitors in proximity does not exceed *MAXF*. When the workload overcomes this limit, the cooperative workload-sharing mechanism allows the system to offload AE and GC to roof nodes while remaining within the time constraint. This confirms that the combination of hardware acceleration (via DPU cores), lightweight runtime management, and edge cooperation is effective in sustaining real-time operation under increasing load.

c: Scalability and deployment challenges

The scalability experiments clarify the limits of the proposed architecture when considering larger deployments. Although serving tailored content to 18 simultaneous visitors is not a realistic scenario for a single totem in a shopping mall, the result is meaningful for two reasons.

First, beyond this point, the one-second deadline cannot be guaranteed even when additional roof nodes are available. The limiting factor is not the computational capacity of the nodes, but the communication layer: the Wi-Fi Direct network becomes saturated due to increased peer-to-peer traffic and connection management overhead. As the number of simultaneous offloading requests grows, medium contention

and connection coordination delays dominate the overall latency, preventing further scaling. In a real-world shopping mall with multiple active totems, additional contention for shared wireless resources would further amplify this effect. Therefore, scaling to a full mall deployment would require localized coordination strategies or a hierarchical communication structure to mitigate Wi-Fi Direct bottlenecks.

Second, the results suggest that the same hardware platform could support multiple displays. For example, a totem equipped with three displays could simultaneously deliver tailored content to approximately six visitors per display while still preserving real-time responsiveness within the one-second deadline.

d: Privacy-Preserving Edge Infrastructure

When compared with state-of-the-art solutions, a distinctive feature of the proposed prototype is its strict adoption of *edge-confined processing*. All computation, including the tasks managing the sensitive data such as AE and GC, remains fully within the local edge infrastructure. Once the nodes are deployed and verified, they can be considered part of a trusted execution environment in which no sensitive data are transmitted to external systems or third parties (e.g., cloud services). If additional computational power is needed, more roof nodes can be added without weakening this trust boundary, since all cooperative communication occurs inside the edge domain. This architectural property ensures that the system simultaneously supports real-time tailored interactions while maintaining a confined processing strategy aligned with the responsible handling of sensitive perception data.

e: Applicability to other environments

The proposed cooperative intelligent totem architecture is not limited to shopping mall deployments but can be applied to any distributed edge scenario that requires soft real-time guarantees and dynamic workload redistribution among neighboring nodes. The key enabler of this generalization is the offloading mechanism combined with peer-to-peer coordination, which allows a local node to maintain deadline compliance by selectively delegating computationally intensive tasks when local demand increases.

In transportation hubs (e.g., airports or train stations), the same architecture could support adaptive passenger services by distributing perception and analytics workloads across multiple kiosks or edge stations. Similarly, in healthcare facilities such as hospitals, the approach could enable personalized information and assistance services by dynamically sharing processing tasks among nearby terminals, ensuring real-time responsiveness even during peak usage conditions. More generally, any smart environment characterized by spatially distributed interactive nodes and variable user density can benefit from the proposed cooperative edge architecture without modifications to its core design principles.

VI. CONCLUSIONS AND FUTURE WORK

This paper presented an AI-driven context-aware interactive totem system built as a local edge infrastructure with cooperative nodes. The *System* integrates a set of AI-based tasks together with a rule-based recommendation task able to provide tailored content to visitors. A key characteristic of the design is the adoption of an edge-confined processing strategy, ensuring that all sensitive data remain within the local edge infrastructure while still enabling real-time tailored interactions. A complete functional prototype was implemented on AMD Xilinx Zynq UltraScale+ devices and evaluated in a laboratory environment emulating real shopping mall conditions.

Experimental results show that the *System* meets its functional requirements, offering robust inference accuracy across perception tasks and stable end-to-end execution without process blocking or pipeline stalls. The *System* satisfies the one-second responsiveness requirement for up to two simultaneous visitors without cooperation and up to eight visitors through workload sharing with a single roof node. Scalability tests further show that increasing the number of roof nodes extends the number of visitors supported within the one-second deadline, up to a practical limit imposed by communication overhead.

The results confirm that cooperative computation at the edge is a viable and effective approach for real-time AI-driven context-aware services in shopping mall environments. The ability to maintain real-time responsiveness while avoiding any interaction with external systems or third parties (e.g., cloud services) positions the proposed architecture as a strong candidate for privacy-aware interactive retail systems and similar deployments.

Future developments will address three main directions. First, mechanisms for reducing model bias will be explored, particularly for underrepresented age and ethnic groups, through dataset augmentation, domain adaptation, and more balanced training strategies. Second, the *System* will be extended to support multi-display totem configurations, enabling simultaneous tailored content delivery to different visitor groups. Finally, federated learning will be integrated within the local edge infrastructure. By enabling participating nodes to collaboratively update shared AI models without exchanging raw data, federated learning can enhance model accuracy over time while preserving data locality and further strengthening the edge-confined processing paradigm.

REFERENCES

[1] K. Amadeo, "What is retailing, and why it's important to the economy," https://www.thebalancemoney.com/what-is-retailing-why-it-s-important-to-the-economy-3305718, Jan. 2023, the Balance. Accessed: Oct. 31, 2025.

[2] MarketsandMarkets, "Artificial Intelligence in Retail Market by Type & Technology - Global Forecast to 2022," https://rb.gy/s9mjhk, 2022, accessed: 11-July-2022.

[3] N. R. Federation. (2025, April) Nrf forecasts 2025 retail sales to hit $5.42 trillion, despite economic uncertainty. Accessed: 2025-12-04. [Online]. Available: https://nrf.com/media-center/press-releases/nrf-forecasts-2025-retail-sales-to-hit-5-42-trillion-despite-economic-uncertainty

[4] Trading Economics. (2025, October) Euro area retail sales yoy. Accessed: 2025-12-04. [Online]. Available: https://it.tradingeconomics.com/euro-area/retail-sales-annual
[5] V. Shankar, "How artificial intelligence (ai) is reshaping retailing," *Journal of Retailing*, vol. 94, no. 4, pp. vi–xi, 2018.
[6] M. Avello, D. Gavilán, C. Abril, and R. Manzano, "Experiential shopping at the mall: influence on consumer behaviour," *China-USA Business Review*, vol. 10, no. 1, 2011.
[7] D. Grewal, A. L. Roggeveen, and J. Nordfält, "The future of retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 1–6, 2017, the Future of Retailing.
[8] PlaySignage. (2024) Digital signage: Impacting store layout & customer flow. Accessed: 2025-06-13. [Online]. Available: https://playsignage.com/blog/Digital-Signage-Impacting-Store-Layout-and-Customer-Flow/
[9] Y. Ma, S. Wang, R. Sun *et al.*, "ISA: An intelligent shopping assistant," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, 2020, pp. 5–9. [Online]. Available: https://aclanthology.org/2020.aacl-demo.3/
[10] L. Andrade, J. Quintero, E. Gamess, and A. Russoniello, "A proposal for a technological solution to improve user experience in a shopping center based on indoor geolocation services," *International Journal of Advanced Computer Science and Applications*, vol. 9, 01 2018.
[11] S. Peretti, F. Caruso, G. Valente, L. Pomante, and T. Di Mascio, "Educating artificial intelligence following the child learning development trajectories," *Behaviour & Information Technology*, vol. 0, no. 0, pp. 1–17, 2025.
[12] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.
[13] S. Huang, H. Yang, Y. Yao, X. Lin, and Y. Tu, "Deep adaptive interest network: Personalized recommendation with context-aware learning," *arXiv preprint arXiv:2409.02425*, 2024. [Online]. Available: https://arxiv.org/abs/2409.02425
[14] University of Delaware. (2024) Managing data confidentiality. Accessed: 2025-06-13. [Online]. Available: https://www1.udel.edu/security/data/confidentiality.html
[15] A. Guha, D. Grewal, P. K. Kopalle, M. Haenlein, M. J. Schneider, H. Jung, R. Moustafa, D. R. Hegde, and G. Hawkins, "How artificial intelligence will affect the future of retailing," *Journal of Retailing*, vol. 97, no. 1, pp. 28–41, 2021.
[16] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629v1*, 2016. [Online]. Available: https://arxiv.org/pdf/1602.05629v1
[17] O. Rudovic, A. Bindal, V. Garg, P. Simha, P. Dighe, and S. Kajarekar, "Streaming on-device detection of device directed speech from voice and touch-based invocation," in *ICASSP*, 2022. [Online]. Available: https://arxiv.org/abs/2110.04656
[18] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016, articles 5(1)(b) and 5(1)(c): Purpose Limitation and Data Minimization.
[19] E. C. Sung, S. Bae, D.-I. D. Han, and O. Kwon, "Consumer engagement via interactive artificial intelligence and mixed reality," *International Journal of Information Management*, vol. 60, 2021.
[20] B. K. V, A. N. Rao, and N. Yedukondalu, "Emotrustbot: Affective intelligence and trust-aware adaptation for emotionally aligned human-robot interaction," in *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2025, pp. 1602–1609.
[21] V. R. Beem, "Ai-driven personalization in retail: Transforming customer experience through intelligent product recommendations," *European Journal of Computer Science and Information Technology*, vol. 13, no. 38, pp. 117–131, 2025.
[22] Hikvision, "Hikvision 4k ai digital signage totem," Hikvision Digital Technology Co., Ltd., 2025, accessed: 29 October 2025. [Online]. Available: https://display.hikvision.com/en/products/interactive-flat-panel-display/digital-signage/
[23] Victek LED, "Ai camera digital signage kiosk with age and gender recognition," Victek LED Technology Co., Ltd., 2025, accessed: 29 October 2025. [Online]. Available: https://victekled.en.made-in-china.com/
[24] Tacteasy, "The evolution of self-service kiosks: Leading the way with ai integration," Tacteasy Inc., 2025, accessed: 29 October 2025. [Online]. Available: https://www.tacteasy.com/the-evolution-of-self-service-kiosks-tacteasy-leading-the-way-with-ai-integration/
[25] Intelligent Kiosk, "Targeted advertising digital signage technology," Intelligent Kiosk Ltd., 2025. [Online]. Available: https://www.intelligent-kiosk.com/targeted-advertising-digital-signage-technology.html
[26] Amazon, "Amazon personalize," https://aws.amazon.com/pm/personalize/, 2025.
[27] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.
[28] G. Valente, F. Caruso, L. Pomante, and T. Di Mascio, "Reactive load balancing for sentient spaces in absence of cloud and fog," *Electronics*, vol. 14, no. 17, 2025.
[29] A. Lojo, L. Rubio, J. M. Ruano, T. Di Mascio, L. Pomante, E. Ferrari, I. García Vega, F. K. Gürkaynak, M. Labayen Esnaola, V. Orani, and J. Abella, "The ecsel fractal project: A cognitive fractal and secure edge based on a unique open-safe-reliable-low power hardware platform," in *2020 23rd Euromicro Conference on Digital System Design (DSD)*, 2020, pp. 393–400.
[30] R. K. Yin, *Case study research and applications*. Sage Thousand Oaks, CA, 2018.
[31] International Organization for Standardization, *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts*, ISO Std., 2018. [Online]. Available: https://www.iso.org/standard/63500.html
[32] D. A. Norman and S. W. Draper, Eds., *User-Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986.
[33] P. Van den Berg, H. Larosi, S. Maussen, and T. Arentze, "Sense of place, shopping area evaluation, and shopping behaviour," *Geographical Research*, vol. 59, no. 4, pp. 584–598, 2021.
[34] L. Ortegón-Cortázar and M. Royo-Vela, "Attraction factors of shopping centers: Effects of design and eco-natural environment on intention to visit," *European Journal of Management and Business Economics*, vol. 26, no. 2, pp. 199–219, 2017.
[35] T. Pei, Y. Liu, H. Shu, Y. Ou, M. Wang, and L. Xu, "What influences customer flows in shopping malls: Perspective from indoor positioning data," *ISPRS International Journal of Geo-Information*, vol. 9, no. 11, p. 629, 2020.
[36] C. Chu, H. Zhang, P. Wang, and F. Lu, "Deepindoorcrowd: Predicting crowd flow in indoor shopping malls with an interpretable transformer network," *Transactions in GIS*, vol. 27, no. 6, pp. 1699–1723, 2023.
[37] M. Sajid, A. H. Khan, K. R. Malik, J. A. Khan, and A. Alwadain, "A new approach of anomaly detection in shopping center surveillance videos for theft prevention based on rlcnn model," *PeerJ Computer Science*, vol. 11, p. e2944, 2025.
[38] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," *Expert systems with Applications*, vol. 42, no. 21, pp. 7991–8005, 2015.
[39] V. Jackson, L. Stoel, and A. Brantley, "Mall attributes and shopping value: Differences by gender and generational cohort," *Journal of retailing and consumer services*, vol. 18, no. 1, pp. 1–9, 2011.
[40] S. Makgopa, "Determining consumers' reasons for visiting shopping malls," *Innovative Marketing*, vol. 12, no. 2, pp. 22–27, 2016.
[41] Data Insights Market, "Shopping mall visitor counting system market's evolution: Key growth drivers 2025-2033," Data Insights Market, Tech. Rep., oct 2025. [Online]. Available: https://www.datainsightsmarket.com/reports/shopping-mall-visitor-counting-system-1391726
[42] H. Koksal, "Shopping motives, mall attractiveness, and visiting patterns in shopping malls in the middle east: a segmentation approach," *Contemporary Management Research*, vol. 15, no. 1, pp. 1–23, 2019.
[43] B. J. Babin, W. R. Darden, and M. Griffin, "Work and/or fun: measuring hedonic and utilitarian shopping value," *Journal of consumer research*, vol. 20, no. 4, pp. 644–656, 1994.
[44] T. Di Mascio, S. Peretti, F. Caruso, and D. Cassioli, "The "great beauty" of diversity: Smart totems to promote gender uniqueness," in *2022 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, 2022, pp. 28–33.
[45] A. Zairis, "The retail store managers' role: Evidence from greece," *International Journal of Business Science & Applied Management (IJBSAM)*, vol. 8, no. 1, pp. 28–40, 2013.
[46] P. Prihandoko, N. Wulandari, and J. Eska, "Implementation of convolutional neural networks (cnn) for crowd counting in shopping mall envi-

ronments," *IJISTECH (International Journal of Information System and Technology)*, vol. 8, no. 4, pp. 267–274, 2024.
[47] H. Kim, M. Button, and J. Lee, "Public perceptions of private security in shopping malls: A comparison of the united kingdom and south korea," *International journal of law, crime and justice*, vol. 53, pp. 89–100, 2018.
[48] J. Pruitt and J. Grudin, "Personas: practice and theory," in *Proceedings of the 2003 Conference on Designing for User Experiences*, ser. DUX '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 1–15.
[49] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994.
[50] X. Wang, B. Wang, Y. Wu, Z. Ning, S. Guo, and F. R. Yu, "A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1729–1757, 2025.
[51] *MQTT: Message Queuing Telemetry Transport*, MQTT.org, 2025, accessed: October 2025. [Online]. Available: https://mqtt.org/
[52] S. Huang, B. Yang, F. Wang, and W. Liu, "Densebox: Unifying landmark localization with end to end object detection," in *IEEE International Conference on Computer Vision (ICCV) Workshops*. IEEE, 2015, pp. 1–9.
[53] Berkeley AI Research, *Caffe Deep Learning Framework*, 2025, available at: https://caffe.berkeleyvision.org/.
[54] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[56] G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are cnn-based features for age and gender classification?" in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–6.
[57] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 341–345, 2006.
[58] H. Hu and C. R. Jasper, "Men and women: A comparison of shopping mall behavior," *Journal of Shopping Center Research*, vol. 11, no. 1, pp. 113–131, 2004.
[59] T. Di Mascio, F. Caruso, and S. Peretti, "How to make an artificial intelligence algorithm "ecological"? insights from a holistic perspective," in *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, ser. CHItaly '23. New York, NY, USA: Association for Computing Machinery, 2023.
[60] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356
[61] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477
[62] AMD, "Zynq ultrascale+ mpsoc," 2024, accessed: 2025-03-10. [Online]. Available: https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html
[63] Xilinx, Inc., *Xilinx Deep Learning Processing Unit (DPU)*, Xilinx, Inc., 2024, available at: https://www.xilinx.com/products/intellectual-property/dpu.html.
[64] Digilent, *Pmod ESP32 Reference Manual*, 2022. [Online]. Available: https://digilent.com/reference/pmod/pmodesp32/reference-manual?redirect=1
[65] G. Jocher *et al.*, "Yolov5," https://github.com/ultralytics/yolov5, 2020.
[66] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European conference on computer vision (ECCV)*, pp. 740–755, 2014.
[67] Google, *Open Images Dataset*, 2025, available at: https://storage.googleapis.com/openimages/web/index.html.
[68] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[69] Joseph Redmon, *Darknet: Open Source Neural Networks in C*, 2025, available at: https://pjreddie.com/darknet/.
[70] *PetaLinux Tools Documentation: Reference Guide (Version 2021.2)*, Xilinx, Inc., 2021, accessed: October 2025. [Online]. Available: https://www.xilinx.com/support/download/index.html/content/xilinx/en/downloadNav/embedded-design-tools.html
[71] *Vitis Unified Software Platform (Version 2021.2)*, Xilinx, Inc., 2021, accessed: October 2025. [Online]. Available: https://www.xilinx.com/support/download/index.html/content/xilinx/en/downloadNav/vitis.html
[72] *Vitis AI User Guide (UG1414) Version 2.5*, Advanced Micro Devices, Inc. (AMD), 2024, accessed: October 2025. [Online]. Available: https://docs.amd.com/r/2.5-English/ug1414-vitis-ai
[73] *Raspberry Pi: Official Website and Documentation*, Raspberry Pi Ltd., 2025, accessed: October 2025. [Online]. Available: https://www.raspberrypi.com/
[74] F. Vahid and T. D. Givargis, *Embedded System Design: A Unified Hardware/Software Introduction*. Hoboken, NJ, USA: Wiley, 2001.
[75] D. Ciambrone, V. Muttillo, L. Pomante, and G. Valente, "Hepsim: An esl hw/sw co-simulator/analysis tool for heterogeneous parallel embedded systems," in *2018 7th Mediterranean Conference on Embedded Computing (MECO)*, 2018, pp. 1–6.
[76] G. Valente, G. Brilli, T. D. Mascio, A. Capotondi, P. Burgio, P. Valente, and A. Marongiu, "Fine-grained qos control via tightly-coupled bandwidth monitoring and regulation for fpga-based heterogeneous socs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 36, no. 2, pp. 326–340, 2025.
[77] V. Jain and E. G. Learned-Miller. (2010) Fddb: A benchmark for face detection in unconstrained settings. [Online]. Available: https://api.semanticscholar.org/CorpusID:8299268
[78] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539
[79] T. Di Mascio, P. Fantozzi, L. Laura, and V. Rughetti, "Age and gender (face) recognition: A brief survey," in *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference*. Cham: Springer International Publishing, 2022, pp. 105–113.

GIANLUCA BRILLI is an assistant professor in computer engineering at the University of Modena and Reggio Emilia, within the High-Performance Real-Time Laboratory (HiPeRT-Lab) which is located in Modena, Italy. His main expertise lies in the field of software and hardware acceleration using reconfigurable embedded systems, and his main research interests are main memory QoS regulation and memory interference mitigation on FPGA-based heterogeneous systems. He graduated in computer engineering in 2018 and he received his Ph.D. in mathematics in February 2022 at the University of Modena and Reggio Emilia.

FEDERICA CARUSO received the M.S. degree in Computer Engineering in 2018 and the Ph.D. degree in Information and Communication Technologies in 2022 from the University of L'Aquila, L'Aquila (Italy). Her research interests include human-computer interaction, assistive technologies, technology-enhanced learning, serious games, gamification, and immersive virtual reality technologies. In particular, she is working on methodologies for designing serious games and immersive virtual reality-based systems in different application domains. From 2025, she is an Assistant Professor in Human-Computer Interaction at the Department of Information Engineering, Computer Science and Mathematics (DISIM) at the University of L'Aquila. She is the author and coauthor of 35 research articles in peer-reviewed journals and national and international conference proceedings.

GIACOMO VALENTE received the M.S. degree in Electronic Engineering in 2014 and the Ph.D. degree in Information and Communication Technology in 2018 from the University of L'Aquila. His primary research activities are in electronic design automation, reconfigurable computer architectures, and real-time systems. Since 2022, he has been an Assistant Professor in Computer Architecture at the Department of Information Engineering, Computer Science, and Mathematics of the University of L'Aquila. He is the author or co-author of more than 30 research articles in peer-reviewed journals and international conference proceedings. He has been also a reviewer and member of several TPCs related to his research topics.

ALBERTO CARLEVARO (Student Member, IEEE) received the M.S. degree in applied mathematics from the University of Genoa, Genoa, Italy, in 2020, with a thesis in mathematical physics, and the Ph.D. degree in electronics from the Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture (DITEN), University of Genoa, in collaboration with the Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT), National Council of Research of Italy (CNR), Rome, Italy, in 2024, with a thesis on mathematical methods for trustworthy artificial intelligence. He was a Visiting Research Scholar with the EECS Department, University of California at Berkeley, Berkeley, CA, USA, working on Physics-Informed Machine Learning. His current research interests include conformal prediction for trustworthy artificial intelligence.

CHIARA GARIBOTTO [SM] was born in Chiavari, Italy, in 1985. She received the Bachelor's degree in Telecommunication Engineering in 2012, the Master's degree in Multimedia Signal Processing and Telecommunication Networks in 2015, and the Ph.D. degree in Science and Technology for Electronic and Telecommunication Engineering in 2019, from the University of Genoa, Italy. She is currently an Assistant Professor at the University of Genoa. Her research interests include context awareness and intelligent sensing in the Internet of Things, e-health, and wireless positioning. She is the Secretary of the IEEE Communication Society eHealth Technical Committee.

JACOPO MOTTA is a Software Developer in AI and Computer Vision at Aitek S.p.A. (M.S., Electronic Engineering, University of Genoa, 2021), focusing on deep-learning video analytics, real-time inference, and production deployment, as well as applied machine learning beyond vision.

VITTORIANO MUTTILLO received his master's degree in computer science engineering in 2015 and his Ph.D. in information and communication technologies in 2019, both from the University of L'Aquila. He is currently a research fellow at the Centre of Excellence DEWS, University of L'Aquila. His research interests focus on embedded systems, with a particular emphasis on Electronic Design Automation, model-based system-level HW/SW co-design, mixed-criticality and cyber-physical systems on heterogeneous multi/many-core platforms, as well as automated software engineering and AI/ML-based digital twin modeling and simulation. He is also an active member of the HiPEAC network, contributing to several cutting-edge European project research in system-level design and high-performance computing.

DAMIANO VALLOCCHIA received the Bachelor's degree in Information Engineering in 2015 and the Master's degree in Telecommunications Engineering in 2019 at the University of L'Aquila, Italy. He currently works as a Telecommunication Engineer in the Research and Development area in Ro Technology, dealing with EU and National funded research projects, focusing on providing cyber and network security solutions.

PAOLO BURGIO got a Ph.D in Electronics Engineering jointly between the University of Bologna and the University of Southern-Brittany, in 2013. His research topics are next-generation predictable systems based on heterogeneous many-cores and GP-GPUs, with an eye on compilers and parallel programming models. Since 2014, he has been a member of HiPeRT Lab at Univ. of Modena, where he currently coordinates the activities on autonomous vehicles and smart cities. He is the co-founder of the HiPeRT srl startup. Like most of the italians, he has a special interest in good food and football.