A TinyML Framework for Quantifying Artifacts' Holding Power in Smart Museums

Rafiq Ul Islam
DIMES Dept., University of Calabria
Rende, Italy
rafiqul.islam@dimes.unical.it

Claudio Savaglio
DIMES Dept., University of Calabria
Rende, Italy
csavaglio@dimes.unical.it

Giancarlo Fortino
DIMES Dept., University of Calabria
Rende, Italy
g.fortino@unical.it

Pietro Manzoni
Universitat Politècnica de València - DISCA
Valencia, Spain
pmanzoni@disca.upv.es

***Abstract*—TinyML enables efficient on-device machine learning for resource-constrained edge devices, addressing privacy and computational challenges in real-world applications. This paper presents a TinyML-based Edge-Cloud system using the FOMO (Faster Objects, More Objects) model deployed on a constrained edge device to analyze visitor engagement in Smart Museums. Unlike conventional camera-based tracking, our approach ensures privacy-preserving analytics by processing visual data locally, extracting only anonymized metrics (e.g., dwell time) to allow for holding power (HP) calculations. We optimize FOMO via data augmentation (DA) and quantization, significantly improving performance; with the optimal configuration, we achieved 86.5% F1-score at 155ms inference latency (a 9.2× speedup over the baseline configuration) while maintaining 88% precision. Our results demonstrate that TinyML-enhanced edge deployment offers a robust, low-latency, and privacy-compliant solution for museum analytics, outperforming traditional cloud and wearable-based approaches.**

***Index Terms*—TinyML, Smart Museum, Smart Environment, Edge AI, Embedded Learning.**

I. Introduction

Tiny Machine Learning (TinyML) is a rapidly growing subfield of Machine Learning (ML) that focuses on developing optimized algorithms, software frameworks, and hardware architectures to enable on-device, low-power ML inference. It bridges the gap between traditional ML/deep learning (DL) and embedded systems, allowing advanced data processing directly on microcontrollers (MCUs) and resource-constrained edge devices [1]. As a cornerstone of Edge Intelligence [2], TinyML enables distributed smart systems to process data locally while minimizing latency, bandwidth, and energy consumption, critical for applications that demand efficiency and privacy preservation.

Smart Museums, as an emerging domain within smart systems, require solutions that balance analytical capabilities with ethical constraints, a challenge uniquely addressed by TinyML's edge intelligence paradigm. In Smart Museums, conventional camera-based tracking systems raise significant privacy concerns due to collecting and storing visitor images or video footage. TinyML addresses these concerns [3] by enabling on-device processing, where visual data is analyzed locally without transmitting raw images to the cloud or storing identifiable information in databases. Instead, the embedded TinyML model extracts only quantitative metrics (e.g., attention time, visitor count) at the edge, forwarding anonymized results to cloud systems for further analysis. This approach promises compliance with data protection regulations while maintaining the functionality required, particularly for computing Holding Power (HP), a metric commonly used to quantify how effectively a display or exhibit captures and retains a visitor's attention. It measures how long people dwell in front of a specific exhibit or point of interest.

Various platforms have emerged to streamline the development and deployment of TinyML solutions on edge devices. EdgeImpulse [4] stands out for its comprehensive toolchain and integration with efficient object detection models. EdgeImpulse provides an end-to-end cloud-based solution to implement TinyML models for edge devices [5]. Starting from the data collection, it provides the data preprocessing, data augmentation (DA), and fine-tuning of pretrained models according to the edge device's requirements, and finally deploys the ML model for a variety of targeted environments, e.g., OpenMV and Arduino. In EdgeImpulse, FOMO ("Faster Objects, More Objects") represents a significant advancement in ML algorithms tailored for resource-constrained devices, facilitating real-time object counting and localization while consuming 30 times less power and memory compared to MobileNet SSD or YOLOv5 [6], providing a better performance factor as compared to other models and uniquely suitable for our privacy-preserving edge device processing platform. The fully convolutional architecture of FOMO accommodates a range of input resolutions (for instance, $96x96$ to $1024x1024$) by generating proportional heatmaps, which are particularly advantageous for identifying minute defects within extensive images. In addition, it integrates effortlessly with MobileNetV2, allowing modifications to the complexity of the model through adjustable alpha values and facilitating transfer learning, empowering users to modify pre-existing models to improve FOMO localization efficiency [7], [8].

In this paper, we present a TinyML-based solution deployed on the edge device and using FOMO model to *(i)* detect the **number of visitors** focusing on a particular Smart Museum artifact, and to *(ii)* calculate its the **Holding Power (HP)** by leveraging the resource-contrained device connected using a LoRaWAN channel which enables efficient long-range communication and ultra-low power consumption while ensuring reliable data delivery even in radio frequency (RF)-challenged indoor environments (which is ideal in Smart Museum scenarios).

The remainder of the paper is organized as follows. Section II described the related work, while Section III provides insight into the proposed system. In Section IV we have discussed the results and, finally, we conclude our work with some future research directions in Section V.

II. Related Work

Museums reflect the historical and cultural significance of past events. The attraction of each artifact in the museum depends on the visitor's interest and how efficiently the artifact is placed there. To optimize visitor engagement, curators must see the visitor's response toward the artifacts to make better decisions about the visibility, placement, spatial allocation, and environmental factors. For this purpose, the attractiveness and HP of artifacts are two primary considerations to evaluate. Artifact attractiveness is defined as the ratio of people who stopped at a particular artifact to the total number of people who visited the museum [9], [10] and HP, defined as the ratio of average time spent in front of the artifact to the time 'necessary' to observe an artifact [11]–[14] given by (Equation 1).

$$\text{Holding Power} = \frac{Estimated\,Occupancy\,Duration}{Time\,to\,observe\,the\,artifact} \quad (1)$$

Measuring visitor interaction with the artifacts provides valuable insights about museum environments and the popularity of artifacts inside the museum, which could be used for further decision-making processes. The authors of [12] share this idea and used RFID tags to trace visitor positions near artifacts and relate them to the blinds provided to visitors at the museum entrance. In [15], different communication protocols, including LoRaWAN, to track visitors' movements inside the museum are compared, while the authors of [16] analyze visitor attention by tracking movement patterns and time spent in specific areas, offering valuable insights into exhibition engagement and the effective utilization of space. Similarly, the attraction power of artifacts is calculated using the LoRaWAN Edge-Cloud architecture in [17], where only sensor-based tracking is implemented without any ML model implementation. Although these approaches partially overlap with our proposal for the technical and technology framework, our method introduces a more robust and resource-constrained solution, using ML to improve accuracy and efficiency.

Other works, such as [18], exploit a multilayer perceptron (MLP) that leverages user contextual data, combining physical and digital interaction behaviors to model user preferences. This hybrid approach, which comprises both behavioral patterns, showed improved point of interest (POI) recommendation performance across all metrics compared to baselines. This method requires substantial Cloud processing, negating the benefits of Edge Computing. Similarly, the approach used in [19] leverages the multisensor of smartphones and smart watches to track human activity recognition. The study explored how Naive Bayes (NB) and K-Nearest Neighbor (KNN) outperform in applications like e-health and remote monitoring for activity recognition. A similar approach could be used in a smart museum scenario to track people who focus on objects, but an edge-optimized approach can better serve this cause. The use of additional beacons and smart devices is not a minimal solution for a smart museum application; on the other hand, automatic tracking can be imposed using the YOLO and Openpose ML models, but deploying them over Edge devices is difficult due to their high computational costs.

The calculation of the HP for museum artifacts has been based on manual observations or sensor techniques. With the advancement in AI and ML domains, systems have been developed for object detection and automated tracking, perfect for resource-abundant systems. With the rise of TinyML [20] and efficient object detection models like FOMO, work done in [6], [21] has shown the implementation for different object detection applications that allow a real-time implementation for low-power devices such as Nicla-vision. However, object detection and tracking-based Edge AI have not been applied in prior work for HP estimation in museum scenarios. Unlike prior work that relies on bulky ML models or limited sensor networks, our solution uniquely combines TinyML-optimized edge vision with privacy-preserving LoRaWAN analytics to enable real-time artifact engagement tracking.

III. Proposed Solution

The proposed system integrates an embedded edge device with a camera and microcontroller as the core component for real-time data capture, on-device inference, and local dwell metric computation. The overall architecture is organized across three hierarchical layers: Edge, Communication, and Cloud, shown in Figure 1. At the edge layer, a camera-based edge node performs lightweight inference to assess visitor engagement through focus duration. This information is transmitted via an LPWAN protocol in the communication layer, supported by a gateway and an application layer protocol to ensure efficient and structured data delivery. Finally, the Cloud layer comprises a message broker for data ingestion, a time-series database for structured storage and

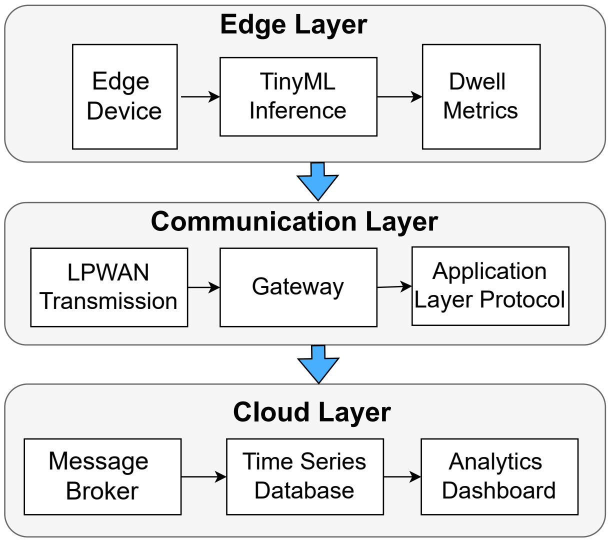

Fig. 1. High-Level System Architecture

processing, and a visualization dashboard for analytics and Holding Power KPI monitoring.

In our implementation (see Figure 2), we used an Arduino Nicla Vision[1] board, which features an onboard camera, to capture visual data and run inference using a FOMO-based object detection model. The computed dwell metrics were transmitted over LoRaWAN to The Things Network (TTN) server, where the data was collected via the MQTT protocol. On the Cloud side, a Dockerized Telegraf container acts as an MQTT client, subscribing to TTN topics and processing incoming payloads. Telegraf then forwards the data to an InfluxDB instance—a time-series database used for storing and analyzing engagement metrics. These metrics are visualized in real time using Grafana, enabling both live monitoring and artifact-level analysis. By combining TinyML, LoRaWAN, and Docker environment, the architecture provides an efficient and scalable solution to provide insights for museum curators.

Our work began with the collection of an image data set to train the TinyML model. We conducted a two-fold data collection strategy. Initially, we collected data from Mohadikar'a Kaggle repository [22], which consists of 2000 images of faces from different angles. We further narrowed down the data selection from these data by considering faces looking into the camera as *Focused* and the faces looking in other directions as *Non-focused* categories. We conducted a controlled laboratory data collection using the same edge device's camera for both categories.

Using Edge Impulse, we uploaded both datasets to label regions of interest (ROIs), such as the visitor's focus on the artifact. For the custom dataset, we annotated ROIs frame-by-frame, labeling *Focused* bounding boxes around faces with a direct gaze toward the artifact (pitch/yaw within

[1]https://docs.arduino.cc/hardware/nicla-vision/

TABLE I
MODEL TRAINING CONFIGURATION

Parameter	Configuration
Model Architecture	FOMO (Faster Objects, More Objects)
Input resolution	96×96 pixels
Data Augmentation	Random flipping, Rotation (±20%), Zoom (±20%), Brightness (±0.2), Contrast, Saturation (80-120%), and Hue (±0.1) adjustments.
Dataset Split	80% training, 20% testing
Optimizer	Adam (LR: 0.005)
Training Cycles	100 epochs (early stopping, patience=5)

±15°) as a focused tolerance threshold and *Non-focused* faces turned away (pitch/yaw more than ±15°) or distracted by other objects. For model training, we utilized FOMO, a lightweight object detection framework optimized for edge devices. The model was trained and validated, optimizing the hyperparameters summarized in TableI to ensure accurate object detection and tracking. The input resolution was set to $96 * 96$ pixels to balance speed and accuracy. To enhance the robustness of the model, we applied a comprehensive DA pipeline that included random flips, rotations, zoom, and adjustments in contrast, brightness, saturation, and hue. These augmentations simulate real-world variations, diversifying the training dataset. The data set was partitioned into 80% training and 20% testing subsets. Training employed the Adam optimizer (learning rate: 0.005) over 100 epochs with early stopping (patience=5) to prevent overfitting. This configuration achieved optimal performance while respecting the computational constraints of the embedded edge device, i.e., Nicla Vision, in our prototype.

After confirming its performance on test data discussed in Section IV, the model was exported in the OpenMV library (.tfLite format) for deployment on the embedded edge device. The deployed model operates on the edge device, detecting and calculating the duration of time each visitor focuses on an artifact shown in Algorithm 1 (It continuously processes camera frames using a FOMO model to detect and track each visitor, updating their dwell times and unique counts in real time. It periodically packages and transmits visitor metrics **active count**, **total dwell time**, and **IDs** via LoRaWAN at fixed transmission intervals. InfluxDB provides a task creation facility to use the incoming data for any KPI calculations. Finally, we calculate the HP KPI using an InfluxDB task and write this in the bucket as a separate field, shown in Algorithm 2(It queries time-series data for total visit time and visitor count within a defined time window, joins them, and computes holding power for each record. It then stores the resulting holding-power metrics in the target data bucket. The last node in this data pipeline is Grafana, which provides visualization of this KPI.

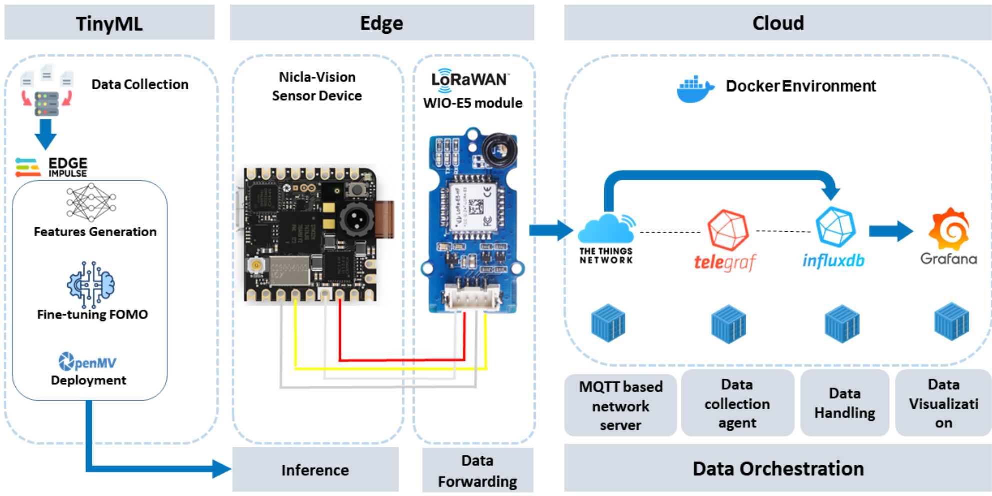

Fig. 2. Hardware/Software Architecture

IV. Results Analysis

A real-life artifact's HP is calculated by combining edge computing with Cloud-based data processing. It is worth noting that, instead of using the estimated occupancy of (Eq. 1), which resorts to the estimated occupancy time, we now have precise estimation about the time visitors spend observing a particular artifact, thus enabling a more accurate calculation of the artifact's HP. Moreover, we have also modified (Eq. 1) to get optimized results between a range of (0-1) in order to compare the results of different artifacts.

The initial formulation of HP quantified engagement as the ratio of the total observed occupancy time to the maximum possible observation time see. (Equation 2)

$$Holding\ Power = \frac{\sum t_i}{n \times T_{\max}} \tag{2}$$

where t_i is the observation time of visitor i, n is the total number of visitors, and $T_{\max} = 40$ s(for our artifact).

While this provided a baseline metric, it introduced two critical limitations:

1) **Cumulative dilution:** HP becomes negligible as n grows over time.
2) **Lack of normalization:** Prevents cross-artifact comparisons due to dependence on visitors' volume.

To address these limitations, we propose a daily HP metric, resetting calculations each exhibition day (9 AM–5 PM) during the execution of the InfluxDB task, to avoid temporal bias and enforce a bounded range $HP_{\text{daily}} \in [0, 1]$. This normalized HP metric enables consistent cross-artifact comparison. Although validation against curator assessments remains future work, it already serves as a reliable automated KPI.

Initially, the model was trained without data augmentation, resulting in poor accuracy due to limited image variations. By incorporating the discussed augmentation techniques on int8 quantization, we significantly improved model

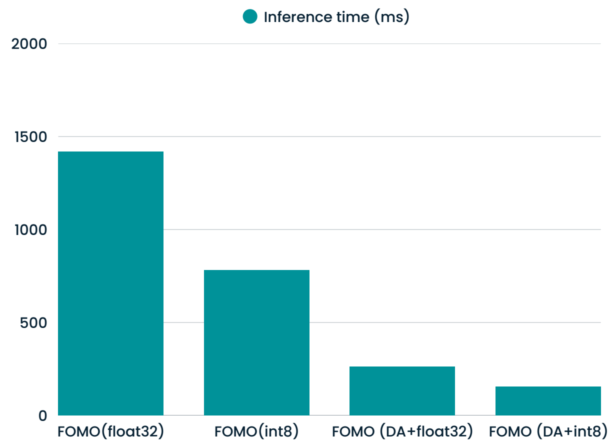

Fig. 3. Inference Latency Comparison with DA, float32 and int8 quantization

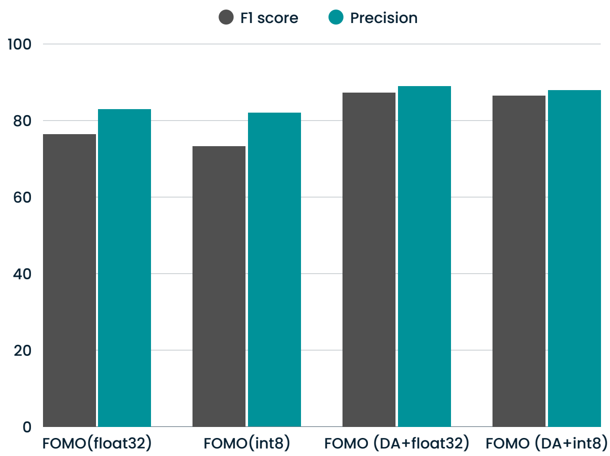

Fig. 4. Model Performance comparison with DA, float32 and int8 quantization

Algorithm 1 Edge-Optimized Visitor's time counting

Require: Camera stream I_t, FOMO model M, LoRaWAN interface L
Ensure: Visitor metrics $\mathcal{M}$, Transmission packets P

Initialize:
Parameters: $\tau_{\text{match}} = 80\text{px}, \tau_{\text{exclusion}} = 50\text{px}, T_{\text{tx}} = 30\text{s}$
$\mathcal{O} \leftarrow \emptyset$ {Active visitor set}
$\mathcal{T} \leftarrow \emptyset$ {Dwell time dictionary}
$N \leftarrow 0$ {Unique visitor counter}
M.load("trained.tflite")
L.join("OTAA")
for each frame I_t at time t **do**
  $\mathcal{D} \leftarrow M.\text{predict}(I_t)$ $\{\mathcal{D} = \{(x_i, y_i, w_i, h_i, s_i)\}\}$
  for each detection $d_i \in \mathcal{D}$ **do**
    $c_i \leftarrow (x_i + w_i/2, y_i + h_i/2)$
    $o_j \leftarrow \arg\min_{o \in \mathcal{O}_{\text{inactive}}} \|c_i - o.\text{pos}\|$
    if $\|c_i - o_j.\text{pos}\| < \tau_{\text{match}}$ **then**
      $o_j.\text{update}(c_i, \text{active} \leftarrow \text{True})$
    else if $\min_{o_k \in \mathcal{O}_{\text{active}}} \|c_i - o_k.\text{pos}\| > \tau_{\text{exclusion}}$ **then**
      $\mathcal{O} \leftarrow \mathcal{O} \cup \{\text{new\_track}(c_i)\}$
      $N \leftarrow N + 1$
      $\mathcal{T}[o_j.\text{id}] \leftarrow 0$
    end if
  end for
  $\Delta t \leftarrow t - t_{\text{prev}}$
  for each $o_j \in \mathcal{O}_{\text{active}}$ **do**
    $\mathcal{T}[o_j.\text{id}] \leftarrow \mathcal{T}[o_j.\text{id}] + \Delta t$
  end for
  if $t \bmod T_{\text{tx}} = 0$ **then**
    $P \leftarrow (\text{ID}, |\mathcal{O}_{\text{active}}|, \sum \mathcal{T}, N) \oplus \{\text{id} : \mathcal{T}[\text{id}]\}$
    $L.\text{transmit}(\text{hex\_encode}(P))$
  end if
end for
return SUCCESS =0

performance, showcasing the effectiveness of data augmentation (DA) in enhancing inference time and accuracy. Quantitative evaluations demonstrate the impact of quantization and DA on our FOMO-based artifact engagement detection system. The baseline `float32` model achieved 76.4% F1-score (1419ms inference), while `int8` quantization reduced latency by 45% (782ms) with minimal accuracy drop (73.3% F1). Though this introduced a marginal 3.1-point accuracy drop, the latency reduction proved critical for real-time deployment on the Nicla Vision's 32MHz Cortex-M7 MCU. Comparative analysis showed that quantization errors primarily affected low-confidence detections (less than 0.6) at artifact edges, but preserved robustness for high-probability center-region engagements. Significantly, DA-enhanced models showed 14% superior performance:the `float32`+DA variant attained 87.3% F1-score at 261ms (5.4× faster than baseline), and its `int8`-quantized counterpart delivered the optimal balance, 86.5% F1-score at just

Algorithm 2 Holding Power Calculation

Require: Time series data from "source_bucket"
Ensure: Holding power metric in "target_bucket"

$t_{start} \leftarrow \text{date.truncate}(\text{now}(), \text{day}) + \text{TimeWindow}(h)$
$t_{end} \leftarrow t_{start} + \text{Exhibition\_Duration}(h)$
$\mathcal{T} \leftarrow \text{query}(t_{start}, t_{end}, \text{"total\_time"})$
$\mathcal{N} \leftarrow \text{query}(t_{start}, t_{end}, \text{"total\_count"})$
$\mathcal{J} \leftarrow \text{join}(\mathcal{T}, \mathcal{N})$
$\mathcal{H} \leftarrow \text{map}(\mathcal{J}, \text{function calculate\_hp}(r)$
if $r.\text{people} > 0$ **then**
  return $\frac{r.\text{time}}{r.\text{people} \times 40}$
else
  return 0
end if
end function)
write($\mathcal{H}$, "target_bucket")
return SUCCESS =0

155ms inference (9.2× speedup vs baseline) while maintaining 88% precision. See Fig. 4 and Fig.3. This establishes FOMO+DA+`int8` as our preferred configuration, which combines near-state-of-the-art accuracy with real-time edge deployment capabilities crucial for museum environments. After inference on Nicla Vision, the data were sent to InfluxDB, where a task was processed for HP calculation, and the results were written in a separate Influx bucket after every 30 seconds. This data is used for visual representation of HP in Grafana shown in Fig. 5.

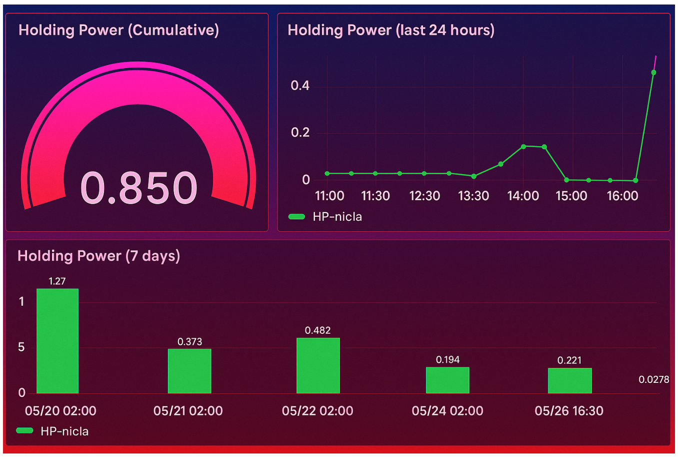

Fig. 5. Grafana Dashboard

The consolidated results show that the synergistic integration of int8 quantization and DA within the FOMO model yields optimal performance for HP estimation, achieving an F1 score of 86.5%, minimal inference latency, and efficient deployment on resource-constrained edge devices.

V. Conclusion and Future Work

Smart museums stand at the intersection of cultural heritage preservation and technological innovation, taking advantage of edge intelligence for real-time analytics while addressing ethical concerns. In this context, TinyML is a key enabler for next-generation cultural analytics.

This work presented a hybrid edge-cloud system that uses the FOMO model deployed on an embedded edge device with a camera and microcontroller for local data capture and processing. The system offers an efficient, privacy-preserving solution for on-device tracking visitors' focus time, ensuring compliance with regulations like GDPR and mitigating re-identification risks, and combining low power consumption with high performance. It achieves an F1-score of 86.5% and an inference time of just 155 ms(over 9× faster than traditional `float32` baselines), enabled by `int8` quantization and data augmentation techniques. Privacy is preserved through on-device processing, which avoids transmitting visual data to the cloud. Robustness is enhanced through training with augmented data, ensuring reliable measurement of visitor engagement. Periodic uploads of head pose metrics to the cloud support curatorial decision-making. Compared to traditional camera-based systems, this approach significantly reduces cost and complexity.

Future work includes extending the FOMO model to support multi-artifact attention tracking using attention heatmaps and deployment in multiple museum environments to validate generalization and robustness under real-world conditions (e.g., crowding, occlusion). Voice-based sentiment analysis, via keyword detection and emotional tone recognition, could further assist curators in evaluating artifact appeal. The exploitation of non IP-based network is another interesting research direction to explore [23]. Finally, developing energy-harvesting modules would enable deployment in heritage sites with limited power infrastructure, advancing sustainable cultural analytics.

ACKNOWLEDGMENT

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union–NextGenerationEU–Project "Entrust: usEr ceNtric plaTform foR continoUS healThcare"–CUP H53D23008110001 - Grant Assignment Decree No. 1382 adopted on 01-09-2023 by the Italian MUR. This work was also partially supported by the research project MORELLINO (CIPROM/2023/29), funded by "Direcció General de Ciència i Investigació" Generalitat Valenciana - SPAIN and by the European Union's Horizon Europe project REMARKABLE under the MSCA-SE grant ID: 101086387.

REFERENCES

[1] P. Warden and D. Situnayake, *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media, 2020. [Online]. Available: https://www.oreilly.com/library/view/tinyml/9781492052036/

[2] V. Barbuto, C. Savaglio, M. Chen, and G. Fortino, "Disclosing edge intelligence: A systematic meta-survey," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 44, 2023.

[3] H. Han and J. Siebert, "Tinyml: A systematic review and synthesis of existing research," 02 2022, pp. 269–274.

[4] S. Hymel, C. Banbury, D. Situnayake, A. Elium, C. Ward, M. Kelcey, M. Baaijens, M. Majchrzycki, J. Plunkett, D. Tischler, A. Grande, L. Moreau, D. Maslov, A. Beavis, J. Jongboom, and V. J. Reddi, "Edge impulse: An mlops platform for tiny machine learning," 2023. [Online]. Available: https://arxiv.org/abs/2212.03332

[5] J. R. et al., "Edge impulse: An mlops platform for tiny machine learning," *Proceedings of Machine Learning and Systems*, vol. 5, pp. 254–268, 2023.

[6] J. Da Silva, T. Flores, S. Júnior, and I. Silva, "Tinyml-based pothole detection: A comparative analysis of yolo and fomo model performance," in *2023 IEEE Latin American conference on computational intelligence (LA-CCI)*. IEEE, 2023, pp. 1–6.

[7] M. N. A. B. Hisham, R. Seman, Z. Saad, and Z. H. C. Soh, "Solar panel defect detection using fomo on edge impulse," *Environment-Behaviour Proceedings Journal*, vol. 10, no. SI27, 2025.

[8] F. Abedrabbo, V. Iglesias, V. Moya, and A. Pilco, "Fomo model based logo detection in embroidery process," 01 2024.

[9] B. Serrell, "Paying attention: The duration and allocation of visitors' time in museum exhibitions," *Curator: The museum journal*, vol. 40, no. 2, pp. 108–125, 1997.

[10] ——, *Paying attention: Visitors and museum exhibitions*. American Association of Museums, 1998.

[11] J. K. Smith and L. F. Smith, "Spending time on art," *Empirical studies of the arts*, vol. 19, no. 2, pp. 229–236, 2001.

[12] J. Lanir, T. Kuflik, J. Sheidin, N. Yavin, K. Leiderman, and M. Segal, "Visualizing museum visitors' behavior: Where do they go and what do they do there?" *Personal and Ubiquitous Computing*, vol. 21, pp. 313–326, 4 2017.

[13] S. Macdonald, *A companion to museum studies*. John Wiley & Sons, 2011.

[14] A. Bollo and L. Dal Pozzolo, "Analysis of visitor behaviour inside the museum: An empirical study," in *Proceedings of the 8th international conference on arts and cultural management*, vol. 2. Citeseer, 2005, pp. 1–13.

[15] D. R. Gawade, R. B. Simorangkir, D. Zorbas, S. Kumar, S. Ziemann, D. Iacopino, J. Barton, K. Schuhmann, M. Anders, B. O'Flynn, and J. L. Buckley, "A museum artefact monitoring testbed using lorawan," in *Proc. - IEEE Symposium on Computers and Communications*, vol. 2021-September. IEEE, 2021.

[16] R. Pierdicca, M. Marques-Pita, M. Paolanti, and E. S. Malinverni, "Iot and engagement in the ubiquitous museum," *Sensors (Switzerland)*, vol. 19, 3 2019.

[17] R. U. Islam, M. Viviani, S. Colace, S. Laurita, C. Savaglio, and G. Fortino, "Improving visitors experience in smart museums," in *Proceedings of The First Latin American Conference on Internet of Things*. IEEE, Apr. 2025, to appear.

[18] S. H. Hashemi and J. Kamps, "Exploiting behavioral user models for point of interest recommendation in smart museums," *New Review of Hypermedia and Multimedia*, vol. 24, no. 3, pp. 228–261, 2018.

[19] M. A. Hanif, T. Akram, A. Shahzad, M. A. Khan, U. Tariq, J. I. Choi, Y. Nam, and Z. Zulfiqar, "Smart devices based multisensory approach for complex human activity recognition," *Computers, Materials and Continua*, vol. 70, pp. 3221–3234, 2022.

[20] A. Elhanashi, P. Dini, S. Saponara, and Q. Zheng, "Advancements in tinyml: Applications, limitations, and impact on iot devices," *Electronics*, vol. 13, no. 17, p. 3562, 2024.

[21] R. Alaa, H. Al-Libawy, and E. A. Hussein, "Low-cost blind spot detection system based on lite object detection algorithm and limited resources hardware," in *2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*. IEEE, 2024, pp. 469–474.

[22] M. Mohadikar, "AFLW 2000 headpose estimation dataset," https://www.kaggle.com/datasets/androsstrk/aflw-2000-headpose-estimation-dataset, 2025, accessed: 2025-04-10.

[23] R. U. Islam, C. Savaglio, and G. Fortino, "Leading smart environments towards the future internet through name data networking: A survey," *Future Generation Computer Systems*, p. 107754, 2025.