

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Fully Onboard AI-powered Human-Drone Pose Estimation on Ultra-low Power Autonomous Flying Nano-UAVs

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Palossi D., Zimmerman N., Burrello A., Conti F., Muller H., Gambardella L.M., et al. (2022). Fully Onboard AI-powered Human-Drone Pose Estimation on Ultra-low Power Autonomous Flying Nano-UAVs. IEEE INTERNET OF THINGS JOURNAL, 9(3), 1913-1929 [10.1109/JIOT.2021.3091643].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/847001> since: 2022-01-23

*Published:*

DOI: <http://doi.org/10.1109/JIOT.2021.3091643>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Burrello A., Garofalo A., Bruschi N., Tagliavini G., Rossi D., Conti F., “DORY: Automatic End-to-End Deployment of Real-World DNNs on Low-Cost IoT MCUs”, in IEEE Transactions on Computers, vol. 70, no. 8, pp. 1253-1268, 1 Aug. 2021, doi: 10.1109/TC.2021.3066883.**

The final published version is available online at:

<https://ieeexplore.ieee.org/document/9381618>

#### Rights/License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it>)  
When citing, please refer to the published version.*

# Fully Onboard AI-powered Human-Drone Pose Estimation on Ultra-low Power Autonomous Flying Nano-UAVs

Daniele Palossi, Nicky Zimmerman, Alessio Burrello, Francesco Conti *Member, IEEE*, Hanna Müller, Luca Maria Gambardella, Luca Benini *Fellow, IEEE*, Alessandro Giusti, Jérôme Guzzi *Member, IEEE*

**Abstract**—Many emerging applications of nano-sized unmanned aerial vehicles (UAVs), with a few  $\text{cm}^2$  form-factor, revolve around safely interacting with humans in complex scenarios, for example, monitoring their activities or looking after people needing care. Such sophisticated autonomous functionality must be achieved while dealing with severe constraints in payload, battery, and power budget ( $\sim 100 \text{ mW}$ ). In this work, we attack a complex task going from perception to control: to estimate and maintain the nano-UAV's relative 3D pose with respect to a person while they freely move in the environment – a task that, to the best of our knowledge, has never previously been targeted with fully onboard computation on a nano-sized UAV. Our approach is centered around a novel vision-based deep neural network (DNN), called PULP-Frontnet, designed for deployment on top of a parallel ultra-low-power (PULP) processor aboard a nano-UAV. We present a vertically integrated approach starting from the DNN model design, training, and dataset augmentation down to 8-bit quantization and deployment in-field. PULP-Frontnet can operate in real-time (up to 135 frame/s), consuming less than 87 mW for processing at peak throughput and down to 0.43 mJ/frame in the most energy-efficient operating point. Field experiments demonstrate a closed-loop top-notch autonomous navigation capability, with a tiny 27-grams Crazyflie 2.1 nano-UAV. Compared against an ideal sensing setup, onboard pose inference yields excellent drone behavior in terms of median absolute errors, such as positional (onboard: 41 cm, ideal: 26 cm) and angular (onboard:  $3.7^\circ$ , ideal:  $4.1^\circ$ ). We publicly release videos and the source code of our work.

**Index Terms**—Autonomous UAV, Convolutional Neural Networks, Ultra-low-power, Nano-UAV, Artificial Intelligence.

## SUPPLEMENTARY MATERIAL

Videos, project's code, and dataset are available at: <https://github.com/idsia-robotics/pulp-frontnet>.

This work has been partially funded by the Swiss National Science Foundation (SNSF) Spark (grant no. 190880), by the Swiss National Centre of Competence in Research (NCCR) Robotics, and by the EU H2020 project ALOHA (grant no. 780788) and I-SWARM (grant no. 871743).

D. Palossi, L. M. Gambardella, A. Giusti, and J. Guzzi are with the Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Via La Santa 1, 6900 Lugano, Switzerland (e-mail: name.surname@idsia.ch). N. Zimmerman is with the Institute of Geodesy and Geoinformation (IGG), University of Bonn, Nussallee 15, 53115 Bonn, Germany (e-mail: name.surname@igg.uni-bonn.de). A. Burrello, F. Conti, and L. Benini are with the Department of Electrical, Electronic and Information Engineering (DEI) of University of Bologna, Viale del Risorgimento 2, 40136 Bologna, Italy (e-mail: name.surname@unibo.it). D. Palossi, H. Müller, and L. Benini are with the Integrated Systems Laboratory (IIS) of ETH Zürich, ETZ, Gloriastrasse 35, 8092 Zürich, Switzerland (e-mail: name.surname@iis.ee.ethz.ch).

Copyright (c) 2021 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org).

Manuscript received March 08, 2021; revised May 04, 2021.



Fig. 1. Our prototype based on the COTS Crazyflie 2.1 nano-quadrotor performing the HDI task, with only onboard computational resources.

## I. INTRODUCTION

Unmanned aerial vehicles (UAVs) equipped with artificial intelligence (AI) algorithms are an important recent development in the Internet of Things (IoT) domain. Applications have been explored in a wide variety of fields, including search and rescue missions, human-drone interaction, precision agriculture [1], [2], monitoring and transportation tasks, providing network and cellular coverage [3]. Most such systems [4], [5] are based on standard- or micro-sized UAVs, as defined in Table I, presenting a taxonomy of the most popular size classes [6]. These relatively large robots can afford for powerful, but bulky, onboard *embedded computers*: for example, the widely-adopted NVIDIA Tegra series [4], [5], [7] is capable of tens of TOP/s within tens of Watts. These computers can run computation- and memory-intensive perception workloads, such as object detection and simultaneous localization and mapping algorithms [7], [8], [9], that are required for their autonomous operation.

Nano-UAVs are an emerging class of aircraft with a much smaller size (sub-ten centimeters), weight (a few tens of grams), and electronic power envelope (sub-Watt). These versatile aircraft enable exciting use cases [10], out of reach for bulkier aircraft. For example, they can be used as unobtrusive and mobile “smart sensors” autonomously flying where their presence is more valuable, such as in cluttered indoor

environments and near humans [11]. They can also perform onboard data analytics to preselect relevant information to be transmitted to the IoT backbone [12].

These applications rely on sensors, such as cameras [9], that yield information-rich but difficult-to-interpret data; therefore, the aircraft need to solve challenging perception tasks aboard to achieve autonomous operation. Unfortunately, the onboard computing capability of nano-aircraft has been traditionally limited to simple microcontroller units (MCUs) [13]. This class of devices can deliver up to a few hundreds of MOp/s, which is insufficient to meet the real-time requirements of state-of-the-art (SoA) perception and navigation algorithms.

In this work, we tackle the problem of endowing a nano-UAV (Bitcraze Crazyflie 2.1 quadrotor) with the ability to solve a challenging visual perception task only with onboard resources. Our task is estimating the relative pose of a free-moving human subject from low-resolution images acquired by the robot's front-looking camera, as illustrated in Figure 1. This perception ability is then used to control the drone to stay at a constant distance in front of the subject, following their movement – a key capability for many human-drone interaction (HDI) tasks.

Our work leverages on the parallel ultra-low-power (PULP) paradigm [14], [15], employing a commercial off-the-shelf (COTS) printed circuit board (PCB) from Bitcraze, called *AI-deck*. This pluggable PCB is compatible with the Crazyflie 2.1 and features a PULP-based *GreenWaves Technologies* GAP8 System-on-Chip (SoC) coupled with an ULP QVGA gray-scale image sensor.

Relying on this PULP-based SoC, we solve the problem of mapping one low-resolution image to the relative pose of the subject ( $x$ ,  $y$ ,  $z$ , and orientation components). For this purpose, we employ our novel streamlined CNN called *PULP-Frontnet*, designed to take advantage of the GAP8 SoC architecture to achieve energy-efficient calculation and precise pose prediction. Our CNN takes inspiration from the *Proximity* network [11] which addresses the same task but exploiting a power-unconstrained remote computer fed with high-resolution images, radio-streamed from a standard-size quadrotor.

A recent research trend demonstrates how addressing the problem at both hardware and algorithmic levels can enable autonomous navigation for the nano-sized class of vehicles [14], [16]. On the one hand, the combination of the PULP computing paradigm [17] with the heterogeneous architectural model [15] enables flexible and energy-efficient computation within the limited power envelope of a nano-UAV. On the other hand, a new class of algorithms based on deep convolutional neural networks (CNNs) represents a lightweight alternative [11], [4], [5], [18] to traditional perception approaches.

Our work provides the following contributions beyond the SoA:

- we introduce PULP-Frontnet, a novel CNN for pose estimation, which we explore in three variants with different computational, performance, and memory trade-offs on the GAP8 SoC. The CNN topologies are designed to meet

TABLE I  
UAVS TAXONOMY BY VEHICLE CLASS-SIZE [6].

Vehicle class	$\varnothing$ : Weight [cm:kg]	Power [W]	Onboard device
<i>standard-size</i> [4]	$\sim 50 : \geq 1$	$\geq 100$	Desktop
<i>micro-size</i> [5]	$\sim 25 : \sim 0.5$	$\sim 50$	Embedded
<i>nano-size</i> [14]	$\sim 10 : \sim 0.01$	$\sim 5$	MCU
<i>pico-size</i> [13]	$\sim 2 : \leq 0.001$	$\sim 0.1$	ULP

both the strict power budget of IoT MCUs and the real-time requirement of autonomous nano-drones;

- we present our dataset augmentation methodology, which maximizes the model's generalization capability with synthetic pitch, photometric, optical, and geometric enhancements;
- using open-source tools [19], [20], we demonstrate our methodology from perception to control (including training, aggressive 8-bit quantization, CNN deployment, and low-level controller), with no drop in regression performance, even compared to the full precision (float 32-bit) Proximity CNN. We achieve an onboard peak inference performance of 135 frame/s within 86 mW and a top energy efficiency of  $\sim 0.43$  mJ/frame;
- we experimentally evaluate how the CNN design impacts on *i*) regression performance, *ii*) power consumption, *iii*) inference rate, and *iv*) closed-loop control accuracy;
- we prove our methodology in the field presenting a closed-loop, fully working demonstration of PULP-Frontnet on a 27-grams nano-UAVs, achieving 100% success-rate on all tests (18 runs on never-seen-before subjects), with behavior comparable with an ideal motion-capture system (median absolute angular error below  $5^\circ$ );
- we publicly open-source our novel dataset, training pipeline, and deployable implementations.

Our work demonstrates that deep learning models for robotic perception, trained and deployed with the proposed methodology, can afford extreme complexity reduction (up to  $24\times$  fewer operations and  $33\times$  less memory, vs. the Proximity NN). Then, offloading our models on an energy-efficient PULP processor, we can achieve a real-time execution even aboard a resource-constrained nano-drone, with no compromise in the quality-of-results, as shown in the supplementary video material.

The rest of the paper is organized as follows: Section II provides the SoA overview of deep learning-based nano-UAVs. Section III introduces the hardware background of our work. Section IV presents in detail our *i*) PULP-Frontnet CNN, *ii*) our dataset augmentation methodology, *iii*) the employed training, quantization, and deployment policies, and *iv*) the proposed onboard control. Section V shows the experimental evaluation of the work, considering *i*) the PULP-Frontnet regression performance, *ii*) the onboard power consumption, inference rate, energy efficiency, and memory use, and *iii*) the final control accuracy with in-field experiments. Finally, Section VI concludes the paper.

## II. RELATED WORK

For a palm-size “flying IoT node”, HDI is a first-class use case that can enhance the user experience (e.g., increased safety) thanks to the small size of the node [21]. In this context, our work addresses the pose estimation of a user from low-resolution images, enabling effective HDI. The *Proximity* NN [11] represents our application baseline, as the same vision-based task is addressed. This NN is based on ResNet [22] and has been demonstrated with a remote commodity desktop computer’s GPU. The *Proximity* NN is coupled with a Parrot Bebop 2 quadrotor flying near the user and streaming front-looking high-resolution images to the remote computer. This allows the model to estimate the subject’s pose relative to the drone, determine the appropriate control input, and send it back to the drone, achieving its control task, i.e., staying in front of the user. Our PULP-Frontnet NN solves the same visual perception task and yields an equivalent quality of robot behavior (see Section V-A), but employs a novel streamlined DL model, e.g., without residual shortcuts (up to  $\sim 24\times$  and  $\sim 33\times$  fewer operations and memory, respectively). Ultimately, our model runs entirely aboard a Crazyflie 2.1 nano-drone (i.e., around  $15\times$  lighter than a Parrot Bebop 2) with no need of any external computer/infrastructure.

Moving into nano-scale UAVs, we focus on those which employ novel deep learning-based algorithms [14], [23], [24], [25]. These approaches are rapidly gaining attention, as they are lightweight compared to the more traditional ones based on the *localization-mapping-planning* cycle [8], and represent a natural fit within resource-constrained nano-drones. Among these, we can distinguish two primary “flavors”: *autonomous* systems that rely only on onboard sensing and computational resources and *automatic* systems that need some off-board aid.

To date, the vast majority of works that combine nano-aircrafts with onboard computation are severely limited on their task’s complexity and applicability. In [23], the DL paradigm applied to nano-UAVs is further streamlined into the swarm scenario. The authors introduce a simple DL model composed of two NNs, running in an *inner-outer* fashion, where the inner NN runs as many times as the number of drones in the swarm. The model predicts the target  $z$  component of the drone’s pose to keep the fleet’s formation during group maneuvers. It uses a 6-element input vector – representing relative position and velocity – per neighboring drone (up to five). Therefore, their biggest input is between  $128 - 512\times$  smaller than ours, and their peak number of operations to predict  $z$  is  $\sim 27\text{ kMAC}$ , i.e., three orders of magnitude less than PULP-Frontnet.

Moving to nano-sized *automatic* systems, i.e., requiring off-board resources, the SoA is characterized by a broad spectrum of use-cases and solutions. For example, vision-based DL/RL algorithms for obstacle avoidance [24], [25], [26] being computed on external computers, and nano-drones localization systems with additional ad-hoc infrastructure (e.g., ultra-wideband anchors, motion-capture cameras, etc.) [24], [27]. Offloading computational-intense workloads to remote base-stations can enable complex algorithms fed with abundant sensory data streams from the aircraft. Nevertheless, this

approach has several drawbacks [28], such as: *i)* latency, *ii)* limited operations distance, *iii)* reliability and security issues on the communication channels, and *iv)* onboard power-consumption overhead due to the high-frequency streaming (e.g., video).

As an alternative to both previous classes of solutions, we are witnessing the advent of ultra-low-power application-specific integrated circuits (ASICs) designed to enable complex functionalities (e.g., visual odometry, simultaneously localization and mapping) aboard small-size robotic platforms [29], [30], [31]. With a power consumption between ten to a few hundreds of mW, these systems have been proven compatible with the power envelope of a small-sized UAV. However, to date, these approaches *i)* have not yet been demonstrated on a real-life flying nano/pico-UAV and *ii)* they only account for one among other fundamental functionalities. Therefore, ASICs increase the system’s complexity because they need co-processors for both basic flying functionalities and to micro-manage control and data transfers. On a severely constrained system (i.e., weight and size), this is a big downside and a push towards more integrated solutions (e.g., SoCs).

Lastly, the *PULP-Dronet* project [14], [16] presents an “halfway” architectural improvement in the SoA. Unlike the aforementioned ASIC designs, PULP-Dronet proposes a novel PULP-based hardware design, called *PULP-Shield*, extending the computational capabilities aboard the Bitcraze Crazyflie 2.1 nano-UAV, with a multi-core general-purpose SoC. This design has been recently launched as a COTS pluggable PCB by Bitcraze, under the commercial name of *AI-deck* – both the robotic platform and the additional processor are the same adopted in our work. Additionally, PULP-Dronet presents a vision-based DL algorithm for autonomous driving of a nano-UAV, tackling lane detection and obstacle avoidance tasks. From a methodology perspective, our work enhances the PULP-Dronet approach in multiple ways:

- addressing a 4-output pose estimation task for HDI;
- making use of open-source quantization/deployment tools [19], [20], as well as employing a  $2\times$  more aggressive quantization scheme (i.e., 8-bits vs. 16-bits);
- including the development flow for ad-hoc dataset collection and its augmentation;
- proposing a novel streamlined DL model (up to  $10\times$  and  $8\times$  fewer operations and memory, respectively);
- introducing a thorough model-size analysis to study the relation between power consumption, memory constraints, regression performance, and control accuracy.

Ultimately, our models push further the onboard NN’s inference performance with a peak throughput of  $135\text{ frame/s}$  @  $86\text{ mW}$  – PULP-Dronet peaked at  $18\text{ frame/s}$  @  $272\text{ mW}$ .

## III. BACKGROUND

### A. PULP paradigm & GAP8 System-on-Chip

Several recently emerged application scenarios (including autonomous flying on nano-UAVs) require local processing capabilities in the order of billions of operations per second on top of devices with power budgets limited to  $1\text{--}100\text{ mW}$  –  $10\text{--}100\times$  more than an off-the-shelf microcontroller unit (MCU).



Parallel ultra-low power (PULP) processing is a recently proposed paradigm that is getting industrial and academic traction to respond to this heightened need of performance and energy efficiency for low-power edge devices. PULP computers couple inside the same System-on-Chip (SoC) a traditional MCU, meant to manage I/O-oriented tasks, with a programmable general-purpose accelerator that is dedicated to the execution of data-parallel computational kernels [17], such as the linear algebra at the heart of artificial intelligence.

In this work, we focus on a commercial embodiment of the PULP paradigm proposed by GreenWaves Technologies: the GAP8<sup>1</sup> SoC, shown in Figure 2. GAP8 employs nine identical RISC-V cores: one, called FABRIC CONTROLLER (FC), is used as the main core in the MCU; eight are used to build up the parallel general-purpose programmable accelerator, i.e., the CLUSTER (CL). All cores are based on the open-source RI5CY design [32] and use a relatively simple four-stage in-order single-issue pipeline to implement the *RV32IMC* instruction set. To improve energy efficiency on integer linear algebra and digital signal processing, RI5CY implements the *XpulpV2* instruction set extension, which includes hardware loops, address post-increment for load/store operations, single instruction multiple data (SIMD) vectorial arithmetic, and dot product instructions for 8- and 16-bit data.

GAP8 includes a full fledged MCU system, organized around the FC. Data and code are stored on a relatively large (512 kB) L2 memory, accessible from both the FC and the CL; the FC has access to a further 16 kB of private L1 memory (uncached) for data and a 1 kB instruction cache. I/O is performed by means of a programmable controller called micro-DMA [33] ( $\mu$ DMA), with support for common serial interfaces (e.g., UART, I2C, SPI, I2S) as well as for 8-bit camera parallel interface (CPI). Additionally, the SoC supports the 8-bit HyperBus protocol by Cypress Semiconductor<sup>2</sup>, enabling external L3 DRAMs and Flash memory, with a bandwidth up to 200 Mbit s<sup>-1</sup>. Each interface is implemented as a separate  $\mu$ DMA channel; the  $\mu$ DMA autonomously moves data between the supported interfaces and the L2 memory, with minimal software intervention. The channels operate independently one from another, hence different transfers can be partially overlapped in time.

The main MCU system is accelerated with a CLUSTER organized around a 64 kB 16-banks L1 scratchpad memory, with word interleaving. The L1 is shared by the eight cluster cores through a high-bandwidth interconnect, with zero wait states from all cores in absence of bank collisions. Program code for both the FC and the CLUSTER resides in L2 memory. The 8 cores share a single 4 kB instruction cache, optimized for a single-program multiple data stream (SPMD) programming model [34], and use a dedicated hardware block to enable low-latency synchronization. Data transfers between L2 and cluster L1 are explicit and performed through a programmable DMA controller within the CL domain.

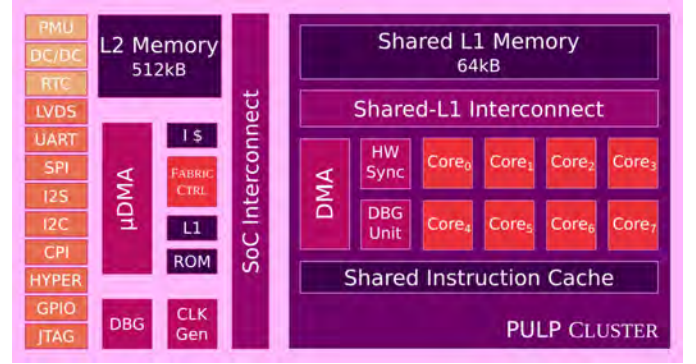


Fig. 2. GAP8 System-on-Chip architecture.

### B. Robotic platform & heterogeneous model

Our robotic platform is represented by the commercial off-the-shelf (COTS) *Bitcraze Crazyflie 2.1*<sup>3</sup> quadrotor, an open-source and open-hardware nano-drone with a weight of 27 g and a diameter of 10 cm. The main processor aboard the drone is the STM32F405 microcontroller unit (MCU) that, together with a combined – i.e., accelerometer/gyroscope – Bosch BMI088 inertial measurement unit (IMU) lays a reliable ground of basic control functionalities. The STM32 MCU runs up to 168 MHz and features 192 kbit SRAM and 1 Mbit flash, on-chip memories, allowing for onboard *inertial state estimation* and *actuation control* tasks. The former utilizes the IMU's input data to feed an extended Kalman filter (eKF) performing the state estimation at 100 Hz; meanwhile, the latter is embodied by a proportional-integral-derivative (PID) control loop cascade. The cascade is composed of two control loops, one controlling the attitude at 500 Hz, and a second one updating the position at 100 Hz.

In our configuration, the robotic platform is extended with two COTS pluggable printed circuit board (PCB) from Bitcraze, the *Flow-deck*<sup>4</sup> and the *AI-deck*<sup>5</sup>, extending the onboard capabilities. The former weighs 3.5 g and features the PMW3901 *optical flow* (OF) visual sensor and the VL53L1x time-of-flight (ToF) ranging module. The OF camera enables the drone to detect its motions in any direction; meanwhile, the ToF sensor provides a distance measurement from the ground. This two sensory information is forwarded to the onboard state estimation increasing its accuracy and reliability, for example, reducing long-term drift. The second expansion board, the AI-deck, represents the high-level onboard computing device in charge of executing complex – otherwise non-addressable – navigational algorithms such as the proposed CNNs.

The AI-deck weighs 4.4 g and is the first commercial embodiment of the *visual navigation engine* – called PULP-Shield – introduced in [14], [16]. Like its predecessor, the AI-deck features a general-purpose GAP8 SoC, additional off-chip memories as big as 512 Mbit HyperFlash and 64 Mbit HyperRAM, and a Himax HM01B0 ULP monochrome QVGA camera. The only exception w.r.t. the first PULP-Shield pro-

<sup>1</sup>[https://greenwaves-technologies.com/gap8\\_gap9](https://greenwaves-technologies.com/gap8_gap9)

<sup>2</sup><https://www.cypress.com/products/hyperbus-memory>

<sup>3</sup><https://www.bitcraze.io/products/crazyflie-2-1>

<sup>4</sup><https://www.bitcraze.io/products/flow-deck-v2>

<sup>5</sup><https://store.bitcraze.io/products/ai-deck>

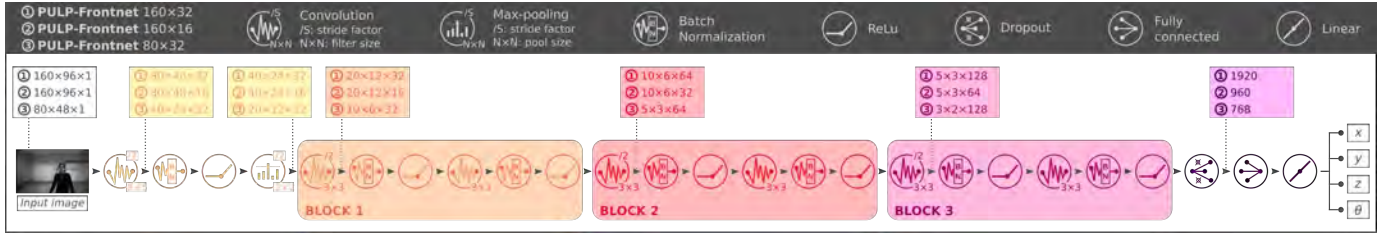


Fig. 3. PULP-Frontnet neural network, exploring three model sizes, varying memory and computational requirements.

prototype is represented by the additional ESP32-based WiFi<sup>6</sup> transceiver and a UART communication channel between the STM32 and the GAP8, instead of the SPI one initially proposed. Even if the availability of the WiFi module eases remote visual-based computation, in the rest of this work, we will refer to a configuration where the expensive WiFi (up to multiples order of magnitude higher power consumption than the SoC) is always turned off. Our primary mission is to develop a fully autonomous system where the whole navigation intelligence runs aboard the nano-drone without any external communication/computation. In our case, the only exception of active WiFi transmission is for dataset collection and showcasing purposes.

The combination of the STM32 MCU with the GAP8 embodies the *host-accelerator* heterogeneous model at the ULP-scale [15]. The *host* (i.e., STM32) is devoted to control-oriented tasks and part of the sensor interfacing. Instead, the computational intensive navigation workloads are offloaded to the general-purpose *accelerator* (i.e., GAP8). Due to the AI-deck's multi-level (L1-L2-L3) memory hierarchy and the pairing of the low-resolution camera with the SoC, the system minimizes communication overhead (e.g., input images are directly fed to the GAP8 without any need to pass through the host) and exploits the locality of data. Even if these basic concepts and functionalities are designed for nano-drones autonomous driving scenario, they are general and applicable to any IoT node requiring visual processing capabilities.

#### IV. METHODS

##### A. PULP-Frontnet neural network

To solve our pose estimation task, we present the *PULP-Frontnet* CNN, shown in Figure 3. The proposed neural network is inspired by the original *Proximity* network [11], where the same task was addressed with different ResNet-based topology and robotic platform. Our model takes as input a front-looking gray-scale image from the low-resolution camera aboard the nano-drone and outputs four independent variables, defining the target pose.

In our design, we employ a classical pattern where each convolution is followed, in order, by a batch normalization and activation (i.e., ReLU) stage, stabilizing the learning process with a per-layer scaling effect [35]. The model is characterized by a first  $5 \times 5$  convolutional layer, followed by a  $2 \times 2$  max-pooling. Each of them reduces by  $4\times$  the output feature map size due to a striding factor of two (both horizontal and

vertical). Then a *block pattern* of two  $3 \times 3$  convolutional layers is repeated three times, where each block doubles the number of output channels and reduces by  $4\times$  the output feature map size. The last part of the model presents a dropout stage followed by a fully connected layer that outputs the pose as a point in the 3-dimensional space ( $x$ ,  $y$ ,  $z$ ), and a rotation angle w.r.t. the gravity  $z$ -axis ( $\theta$ ).

To successfully deploy PULP-Frontnet on top of a resource-constrained MCU, such as the GAP8 SoC, the NN's execution must comply with the strict real-time constraints dictated by the application scenario while respecting the bounds imposed by the on-chip and onboard resources. The main constraints can be summarized as follow: *i) throughput or minimum frame-rate*, *ii) quality-of-result or regression performance*, and *iii) onboard/on-chip memory limits*. In this light, it is clear we need a reliable methodology and strategy to reduce the memory and computational loads, to ease the deployment on the available resources while exploiting the hardware architecture at best to meet the real-time constraint.

Therefore, to comply with the given architectural constraints, we introduce fixed-point arithmetic and 8-bit integer data (see Section IV-C) instead of floating-point calculation on a 32-bit data type – as per the original proximity network. This transformation represents an industry-standard with many advantages, such as  $4\times$  reduction in the memory need; fast and efficient execution on devices without hardware support for floating-point calculation, such as the GAP8 SoC and many other commercial MCUs; and enabling for optimized signal processing instructions (e.g., packed-SIMD). The price for these enhancements at both memory and computational level is a minimal drop in the CNN's accuracy [36].

For this reason, in this work, we aim at investigating the relationship between memory footprint and computational requirements (i.e., number of operations) of the proposed model w.r.t. its regression performance and closed-loop in-field control accuracy (see Section V-A and V-C). Therefore, we play with the memory/operations knobs by varying input image size and the number of channels between different NN's blocks, affecting both parameters (i.e., weights) and intermediate feature map sizes. This process results in three PULP-Frontnet NN variants, as shown in Figure 3 and detailed in Table II. Here, the number of operations accounts only for convolutional and fully connected layers; instead, the memory requirements consider the input image, all weights, and all intermediate buffers to store the feature maps – i.e., what we would obtain from a straightforward implementation.

The first version, named  $160 \times 32$ , is characterized by the

<sup>6</sup><https://www.u-blox.com/en/product/nina-w10-series-open-cpu>

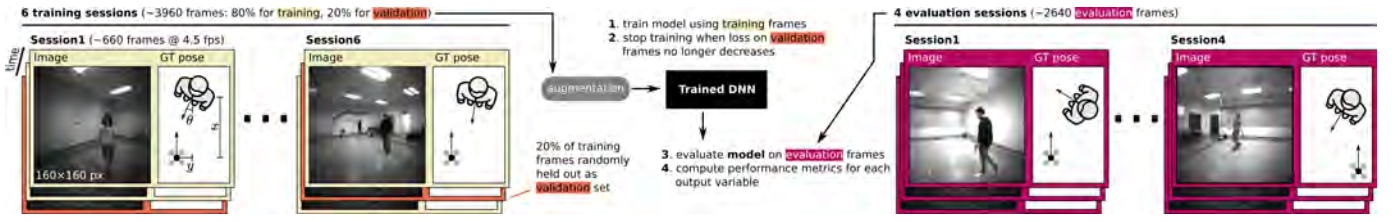


Fig. 4. Organization and splitting of the collected data sequences. Data from 6 training sessions (left) is used, with random augmentations, to learn a model (center), which is then evaluated with data from 4 different sessions (right).



Fig. 5. The original dataset image (left) is cropped at a random height to simulate pitch variations; a random subset of photometric, optical and geometric augmentations (top) are then applied. Bottom: ten random augmentations originating from the same source image.

biggest memory footprint and the highest number of multiply-and-accumulate (MAC) operations needed to perform one input image inference. The  $160 \times 16$  NN represents the extreme edge on the memory minimization exploration axis instead. Lastly, the NN called  $80 \times 32$  uses a smaller input image, w.r.t. the previous two NNs, showing the minimum computational requirements. Table II also compares the proposed PULP-Frontnet models with the Proximity NN. With  $7\text{--}24\times$  fewer operations and  $12\text{--}33\times$  less memory need, if compared to the Proximity NN, our model makes it possible to envision an outstanding performance on our deployment robotic platform.

TABLE II  
MAC OPERATIONS AND MEMORY FOOTPRINT FOR ONE FRAME INFERENCE OF THE PULP-FRONTNET MODELS AND THE PROXIMITY NN [11].

PULP-Frontnet	$160 \times 32$	$160 \times 16$	$80 \times 32$	NN [11]
Operations [MMAC]	14.1	4.3	4.0	96.5
Memory [kB]	499	184	348	6116
# Parameters	$3.03 \times 10^5$	$7.80 \times 10^4$	$2.99 \times 10^5$	$1.26 \times 10^6$

### B. Dataset collection & augmentation

**Dataset collection.** The dataset used to train, validate, and test the PULP-Frontnet models is collected in a  $10 \times 10$  m room equipped with a motion capture system (mocap), composed of 12 Optitrack PM13 cameras. The dataset is acquired using the same deployment robotic platform, introduced in Section III-B, and therefore using the onboard QVGA, gray-scale, Himax camera. In our dataset collection setup, the quadrotor is equipped with a mocap target (i.e., reflective marker) and affixed with a horizontal attitude (zero pitch and roll) on a wheeled cart with adjustable height.

During the dataset acquisition, an operator moves the cart around the room, continuously changing its position, heading, and height (in a range between 1.20 m to 1.45 m). Simultaneously, the recorded subject moves freely in the environment, wearing either a baseball cap or an almost-invisible headband

with a mocap target affixed. The operator and the subject move so that the latter is visible in most frames while capturing a wide distribution of camera-subject distances and headings. Because the camera is moving and the room's setup (e.g., lighting) is purposefully changed between different recording sessions, the images contain varied backgrounds, sometimes cluttered with objects including furniture, lab equipment, and people other than the subject.

Thanks to the markers applied, on both quadrotor and subject, the mocap system can track their pose at 200 Hz, resulting in a precise ground truth labeling information. These poses are recorded using the ROS [37] framework and synchronized with the  $160 \times 160$  pixels video frames (streamed to a host computer via WiFi from the quadrotor). Once the quadrotor and subject poses are known in the room reference frame, the relative pose of the subject with respect to the drone is computed and decomposed in its  $x$ ,  $y$ ,  $z$  and  $\theta$  components. We recorded ten distinct sessions featuring different subjects of various ages, height, clothing, and hairstyles. Furthermore, some subjects wear face masks and eyeglasses only for part of the session to enhance data variety. In total, we collected about 6600 frames representing 25 minutes of acquisition time (150 s per subject at  $\sim 4.5$  frame/s), as described in Figure 4.

Among the ten recorded sessions, we use six for training and the remaining four for evaluation; this ensures that the same subject does not appear both in training and evaluation sets. Therefore, our evaluation metrics quantify the model's ability to generalize to unseen subjects and do not reward overfitting on the training ones. The training dataset is further split by holding out a random 20% of its frames for validation, for which the loss is monitored as training progresses. The model that yields the lowest validation loss is selected following a standard *early-stopping* pattern. The organization of the dataset, and the workflow for training and evaluating models, is illustrated in Figure 4.

**Dataset augmentation.** During a mission, the quadrotor pitch changes in a range of approximately  $\pm 15^\circ$  to generate forward/backward accelerations: this heavily affects the image



acquired, which is not stabilized optically nor electronically. Therefore, a CNN trained only on images acquired with a flat attitude returns increasingly inaccurate results as the in-field acquisition pitch deviates from  $\pm 0^\circ$ .

Our mitigation strategy to this problematic effect is to apply, during training, a synthetic pitch augmentation technique (see Figure 5) to ensure that our model is robust to the actual pitch of the quadrotor during the mission. In particular, we observe that we can generate  $160 \times 96$  pixel images with an approximated pitch in the range  $\pm 14^\circ$  by cropping a subset of the rows from a  $160 \times 160$  pixel image acquired with a horizontal and constant attitude. For example, the top 96 rows of a dataset image, compared to its middle 96 rows, depict the same scene with an approximated pitch of  $+14^\circ$  (see Figure 5 on the left). This approximation disregards photometric effects due to vignetting in the full-frame image and ignores perspective distortion; however, the approximation is acceptable for our purposes and can be implemented as an inexpensive augmentation strategy within the training pipeline. In particular, for a given  $160 \times 160$  pixel dataset image, we crop  $160 \times 96$  training samples at random vertical positions; this yields more than 66'000 training instances.

To further promote generalization, we apply additional data augmentation techniques, as shown in Figure 5. First, we apply the following photometric and optical augmentations, each independently sampled with a 50% probability:

- *contrast* jittering (multiplicative factor uniformly sampled in the range  $[0.7, 2.0]$ ), accounting for erratic auto exposure behavior in our camera model;
- *brightness* jittering in the range  $[-0.2, 0.2]$ ;
- *gamma* correction (exponent uniformly sampled in the range  $[0.4, 2.0]$ );
- synthetic *vignetting* effect with random radius and strength, which accounts for the strong vignetting present in the Himax images, and its likely variability in different cameras and illumination conditions;
- smoothing using a gaussian kernel with  $\sigma = 3$  pixels, accounting for *blurring* due to camera motion, vibration, and lens defocus.

Finally, with a 50% probability, we horizontally *flip* the image while correspondingly altering the ground truth, i.e., negate the  $y$  and  $\theta$  variables. Note that this automatically ensures that the distribution of these variables in the datasets is symmetric.

### C. Training, quantization, and deployment.

**Network training and quantization.** We divide the procedure to produce a deployable PULP-Frontnet in several steps, implemented using the PyTorch framework<sup>7</sup> and the open-source NEMO library [19]. First, we train a *full-precision* floating-point PULP-Frontnet on the dataset described in Section IV-B, minimizing the L1 loss for the pose vector  $(x, y, z, \theta)$ . We use the Adam optimizer with learning rate  $10^{-4}$  and early stopping over 100 epochs; we then select the model with the lowest validation loss, which is obtained after on average 80 epochs.

After full-precision training has ended, we perform a *fake-quantized fine-tuning* step. We use linear uniform per-layer quantization and a variant of PACT [36] for training, with quantization functions of the form:

$$Q(\mathbf{t}) = \varepsilon_t \cdot \left\lfloor \frac{\mathbf{t}}{\varepsilon_t} \right\rfloor \quad (1)$$

where  $\mathbf{t}$  is the tensor,  $Q(\mathbf{t})$  its quantized representation, and  $\varepsilon_t$  is a scalar representing the difference between two consecutive fixed-point values. We chose to use 256 levels (8 bits) for activations and 128 levels for weights; given the weights' asymmetric distribution around 0, these 128 levels can be represented using an 8-bit signed integer. The network is manipulated so that convolutional and fully connected layers use weights that have passed through the quantization function of Equation 1, with  $\varepsilon_{\mathbf{W}} = (\mathbf{W}_{\max} - \mathbf{W}_{\min}) / (2^7 - 1)$ .  $\mathbf{W}_{\max}$  and  $\mathbf{W}_{\min}$  are fixed to the layer-wise maximum and minimum values of weights, respectively. Batch normalization and pooling layers are left untouched at this stage.

To quantize activations, we replace all ReLUs in the network with the quantization function of Equation 1, choosing  $\varepsilon_{\mathbf{x}} = \alpha / (2^8 - 1)$ .  $\alpha$  is initialized to the maximum value reached by the output of each ReLU over the validation set and is then trained by back-propagation. To fine-tune the fake-quantized network, we initialize it with the equivalent quantized value of the full-precision weights, then we calibrate the  $\alpha$  parameters using the validation set, and finally, we perform 100 epochs using the Adam optimizer to minimize the L1 loss. We use an initial learning rate of  $10^{-4}$  with 0.95 decay and weight decay of  $10^{-6}$ . After the fine-tuning procedure has completed, the network is transformed into an *integer deployable* one [19]. Weights  $\mathbf{W}$  are approximated without network performance loss as:

$$\mathbf{W} \approx \varepsilon_{\mathbf{W}} \cdot \mathbf{W}_{\min}^* + \varepsilon_{\mathbf{W}} \cdot \mathbf{W}^*, \quad (2)$$

where  $\mathbf{W}_{\min}^* = Q(\mathbf{W}_{\min}) / \varepsilon_{\mathbf{W}}$  and  $\mathbf{W}^*$  is an integer tensor with values in the range  $[-64, +63]$ . Notice that  $\mathbf{W}_{\min}^*$  is also defined in the same  $[-64, +63]$  region; therefore,  $\mathbf{W}$  as a whole can be accurately represented with an 8-bit signed integer even if weights are distributed asymmetrically around 0. We replace floating-point batch normalization layers with integer ones using 32-bit parameters. Finally, tensors outcoming activation layers are represented using 8-bit unsigned integers, while intermediate outputs use a 32 bits data type. Therefore, the entire network can run entirely in the integer domain and produces a vector of four 32-bit fixed-point values that approximate  $(x, y, z, \theta)$ .

**Deployment strategy.** The quantized PULP-Frontnet models' deployment is based on the PULP-NN library [20] for optimized 8-bits fixed-point arithmetic. PULP-NN exploits the eight cores general-purpose CLUSTER of the GAP8 SoC to parallelize kernels' execution on the spatial dimensions (i.e., *width*  $\times$  *height*) and the SIMD and bit-manipulation ISA extensions to achieve the best performance and energy efficiency, peaking at 15.6 MAC/cycle for squared-size images. However, the available kernels operate on the shared L1 64 kB memory, constraining their applicability to small layers; therefore, these kernels are not suitable for deploying our network without any additional intermediate manipulation.

<sup>7</sup><https://github.com/idsia-robotics/pulp-frontnet>

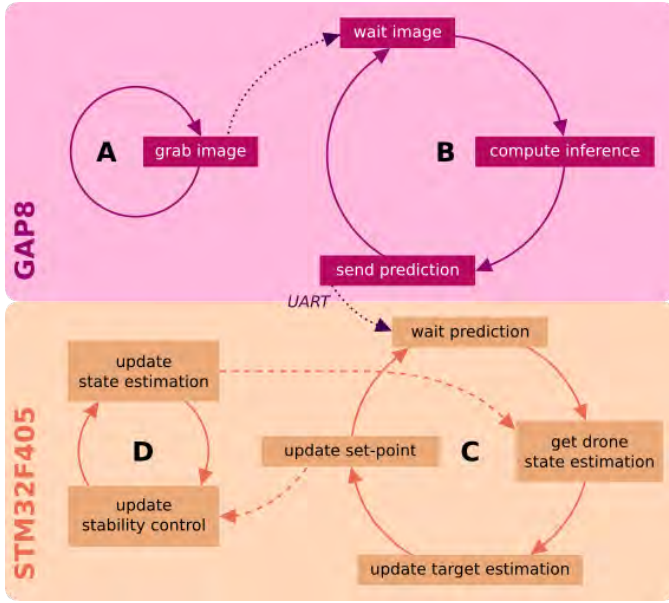


Fig. 6. The four main loops that define the drone behavior. (A) camera loop and (B) inference loop run on the GAP8 SoC, while (C) high-level control loop and (D) low-level control loop run on the STM32F405. Dark violet dotted arrows mean synchronization.

For this purpose, we employ the DORY tool [20] that automatically produces C “wrapping” code to manage the two levels of on-chip memory (i.e., L1, L2) and the external RAM, orchestrating weights and activation movements to maximize PULP-NN kernels performance. Thanks to general templates and tensor tiling, DORY divides the layers in nodes, which are executed in L1 by inserting *i*) double-buffered L2-L1 DMA calls, and *ii*) calls to basic kernels in the PULP-NN library, which performs computation on local L1 data. The data movements are always overlapped with computation due to asynchronous and non-blocking DMA calls. Further, DORY always operates storing the network’s weights in the external RAM, loading them into the L2 during the previous layer’s execution by employing the  $\mu$ DMA – i.e., during the execution of layer  $i$ , the weights of layer  $i+1$  are transferred –, realizing a two-level double buffering between RAM, L2, and L1.

This policy enables the execution of NNs whose weights would not fit in the available L2 memory constraint. However, it prevents the possibility of “ahead-of-time” weights pre-loading into the L2 for small networks that would not benefit from this continuous data movement. Therefore, for our exploration on the relation between memory, power consumption, and the CNN regression performance, we manually modify the C code produced by DORY to investigate our smallest PULP-Frontnet model (i.e.,  $160 \times 16$ ) pre-loading all the weights in L2. In this way, we restrict the RAM utilization only to an initialization stage before the mission starts.

#### D. Onboard closed-loop control

Figure 6 illustrates the control flow that makes the drone hover in front of a person. Pose estimations are computed on the GAP8 SoC by the PULP-Frontnet CNN and then sent through the UART interface to the Crazyflie’s flight controller

(STM32F405 MCU), where they are filtered and then used to compute low-level control set-points. This flow is organized in four loops:

- (A) **camera loop**: the camera regularly grabs  $162 \times 162$  gray-scale images;
- (B) **inference loop**: once a new image is available, it is cropped to the target size (e.g.,  $160 \times 96$ ) and pushed to the inference loop. The predicted pose in the drone-relative frame is sent through UART from the GAP8 to the STM32;
- (C) **high-level control loop**: the STM32 receives the pose and transforms it to the drone odometry frame, fused with the previous target state estimation, and uses it to update a low level set-point;
- (D) **low-level control loop**: the high-rate stability loop regularly updates the drone’s state estimation and applies a cascade of PID controllers to reach the set-point decided by the high-level loop.

The only synchronization points between different loops are *i*) on the wait for a new image in the inference loop, and *ii*) on the wait for a new pose in the high-level control loop. Therefore, the high-level control’s frequency is bounded by the maximum inference loop (i.e., up to 135 Hz for the fastest  $80 \times 32$  NN), which in turn also depends on the maximum image grabbing rate (i.e., up to  $\sim 160$  Hz).

**Notation and state space.** As presented in Section IV-B, PULP-Frontnet outputs a pose estimation that does not depend on the pitch and roll components of the orientations of the drone and the human subject. Therefore, we only consider poses that belong to the Euclidean subgroup  $SE(3)$  generated by translations and rotations around the common z-axis aligned with gravity. We denote the pose of object  $\mathcal{A}$  – drone or subject – with respect to frame  $\mathcal{B}$  as  $p_{\mathcal{A}}^{\mathcal{B}} = (\vec{p}, \theta) \in \mathbb{R}^3 \times S^1$ , where  $\vec{p}$  represents the position and  $\theta$  the rotation around the common z-axis. To better react to the subject’s movements, the drone also keeps track of their linear and angular velocity  $v$  as part of the subject’s state. We denote the state of object  $\mathcal{A}$  with respect to frame  $\mathcal{B}$  as  $\xi_{\mathcal{A}}^{\mathcal{B}} = (p_{\mathcal{A}}^{\mathcal{B}}, v_{\mathcal{A}}^{\mathcal{B}}) \in \mathbb{R}^3 \times S^1 \times \mathbb{R}^4$ .

We introduce three frames, which all share the same z-axis orientation:  $\mathcal{D}$  attached to the drone,  $\mathcal{H}$  attached to the subject, and  $\mathcal{O}$  as the world-fixed drone odometry frame. In Figure 7, we depict the top-down view of the human subject  $\mathcal{H}$  walking sideways to their right and the drone  $\mathcal{D}$  (violet) trying to stay in front at constant distance  $\Delta$ . It does so by moving towards target pose  $\mathcal{D}'$  (red) while rotating towards the violet line. Frames are drawn with a solid x-axis (with a unit vector  $\vec{e}$  that points away from the front of the object), a dashed y-axis, and share the same z-axis exiting the drawing. All computations, except inference, are done in the odometry frame  $\mathcal{O}$ ; therefore, we later drop the related index to simplify the notation. In the following, differences between angles in  $S^1$  are meant as real values in  $[-\pi, \pi]$ .

**High-level control loop.** As illustrated in loop C of Figure 6, the high-level control loop, which is in charge of updating the low-level set-point, comprises four steps:

- *wait prediction*: the loop waits until a new prediction  $\tilde{p}_{\mathcal{H}}^{\mathcal{D}}$  is computed from the inference loop, based on the current camera image;

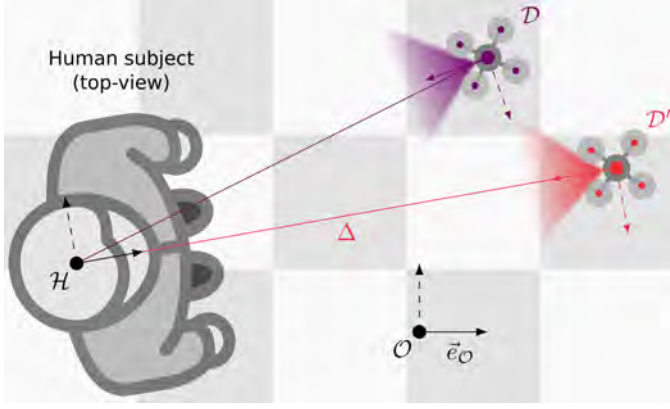


Fig. 7. Our three reference frames:  $\mathcal{D}$ ,  $\mathcal{H}$ , and  $\mathcal{O}$ . Top-down view of the human subject  $\mathcal{H}$  walking sideways to their right and the drone  $\mathcal{D}$  (violet) trying to stay in front at distance  $\Delta$  by moving towards target pose  $\mathcal{D}'$  (red).

- *get drone state estimate*: the loop reads and stores the current drone state estimation  $\xi_{\mathcal{D}}^{\mathcal{O}}$  from the low-level control;
- *update target estimation*: using the current transformation  $\xi_{\mathcal{D}}^{\mathcal{O}}$ , the loop computes the prediction to the odometry frame  $\tilde{p}_{\mathcal{H}}^{\mathcal{O}}$ , which it uses to update the subject state estimation  $\xi_{\mathcal{H}}^{\mathcal{O}}$  in a Kalman filter;
- *update set-point*: finally, the loop updates the desired velocity  $v_{\mathcal{D}}^{\mathcal{O}}$  to make the drone staying in front of the subject, the desired distance  $\Delta$ .

**Kalman filter.** We model the drone dynamics and inference prediction as a stochastic linear process with normally-distributed and zero-mean noise. More precisely, we let  $(p_n, v_n)$  be the values of  $\xi_{\mathcal{H}}^{\mathcal{O}}$  and  $o_n$  those of  $\tilde{p}_{\mathcal{H}}^{\mathcal{O}}$  at time  $t_n$ , and assume that:

$$p_{n+1} = p_n + v_n(t_{n+1} - t_n) \quad (3)$$

$$v_{n+1} = v_n + a(t_{n+1} - t_n) \quad (4)$$

$$o_n = p_n + \epsilon, \quad (5)$$

where the acceleration  $a \in \mathbb{R}^4$  has covariance  $Q$  and zero mean, and the observation error  $\epsilon \in \mathbb{R}^3 \times S^1$  has covariance  $R$  and zero mean. We make two further assumptions: the processes are isotropic, and invariants, i.e., covariances  $Q$  and  $R$  are constant and diagonal. We can then decouple the Kalman filters for each component at the cost of neglecting that predictions, w.r.t. the drone longitudinal and lateral axis, have slightly different MSE and that errors depend on the human subject relative position.

**Velocity control.** We define simple, uncoupled linear and angular controls to let the drone hover in front of the human subject. The target linear velocity  $\vec{v}_{\mathcal{D}}'$  is computed to allow the drone to reach the target position  $\tilde{p}_{\mathcal{D}'}$  at distance  $\Delta$  in front of the subject in time  $\tau$ . Assuming the subject is smoothly moving at an almost constant speed, we add to  $\tau$  the subject's estimated velocity  $\vec{v}_{\mathcal{H}}$ , where:

$$\tilde{p}_{\mathcal{D}'} = \tilde{p}_{\mathcal{H}} + \vec{e}_{\mathcal{H}}\Delta \quad (6)$$

$$\vec{v}_{\mathcal{D}}' = \frac{\tilde{p}_{\mathcal{D}'} - \tilde{p}_{\mathcal{D}}}{\tau} + \vec{v}_{\mathcal{H}} \Big|_{[-\vec{v}_{\max}, \vec{v}_{\max}]} \quad (7)$$

The goal of the angular control is to keep the subject centered in the image frame. Therefore, we first compute a target

orientation  $\theta_{\mathcal{D}}'$  as the current orientation to face the person and then the angular speed to reach it over time  $\tau$ :

$$\theta_{\mathcal{D}}' = \angle(\vec{e}_{\mathcal{D}}, \tilde{p}_{\mathcal{H}} - \tilde{p}_{\mathcal{D}}) \quad (8)$$

$$\omega_{\mathcal{D}}' = \frac{\theta_{\mathcal{D}}' - \theta_{\mathcal{D}}}{\tau} \Big|_{[-\omega_{\max}, \omega_{\max}]} \quad (9)$$

All velocities are clamped within the maximal ranges of  $v_{\max} = 1 \text{ m s}^{-1}$  for the linear speed, and the angular speed below  $\omega_{\max} = 0.8 \text{ rad s}^{-1}$ .

**Low-level control loop.** The lowest level of our control is based on the open-source `Controller_PID` offered by the Crazyflie 2.1 firmware<sup>8</sup>. In particular, the drone applies a cascade of PID controllers (with no synchronization with the high-level controller) to update *i*) the target attitude from the current speed and target velocity (@100 Hz), *ii*) the target attitude rate from the current attitude and the target attitude (@500 Hz), and *iii*) motor commands from the current attitude rate and the target attitude rate (@500 Hz). We limit the absolute value of the target pitch to  $12^\circ$  so to respect the pitch limit used for the dataset augmentation. The low-level control loop is also in charge of updating the drone state estimation  $\xi_{\mathcal{D}}^{\mathcal{O}}$ , fusing the onboard measurements from the IMU, OF camera, and ToF distance sensor (z-direction) using an extended Kalman filter.

## V. RESULTS

In Section V-A, we evaluate the regression performance of different models with offline experiments on the testing set. In particular, we compare full precision and quantized variants of the same network, different network architectures, and quantitative performance across the output variables. Section V-B evaluates energy efficiency, power consumption, and computational/memory requirements of the proposed PULP-Frontnet variants. Finally, Section V-C analyzes the quadrotor behavior on in-field person-tracking experiments, using only onboard sensing and computation.

### A. Regression performance

In this section, we report the regression performance metrics for our three PULP-Frontnet networks (both full-precision and quantized) and for the Proximity NN [11] (only full-precision). All models are trained on the same training set and augmented as defined in IV-B. Inputs are scaled to the appropriate input resolution for each model ( $160 \times 96$  and  $80 \times 48$  for the PULP-Frontnet, and  $108 \times 60$  for Proximity NN) using bilinear interpolation. All models are then evaluated on the same testing set, sampled without augmentation ( $\sim 4'000$  images).

Table III reports, for each output variable, the mean absolute error (MAE), expressed in meters for  $x, y, z$  and radians for  $\theta$ , and the mean squared error (MSE), highlighting the best score (in bold) for both metrics across all the networks. The MAE on  $x$  and  $y$  variables is  $\sim 0.2 \text{ m}$  for all models, which indicates a good ability to localize the subject on the horizontal plane. For the  $z$ , the error is lower w.r.t. other variables due to

<sup>8</sup><https://github.com/bitcraze/crazyflie-firmware>



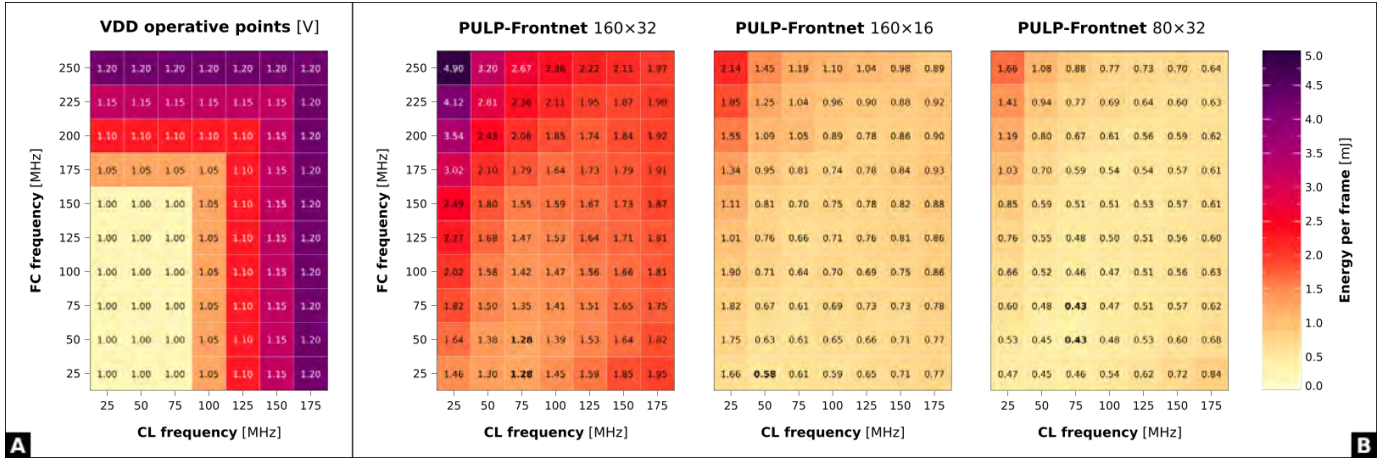


Fig. 8. A) VDD operative points used for the frequencies sweeping in B. B) Energy efficiency analysis, as the energy required to run the inference on a single frame, for all three NNs, sweeping both FC and CL frequency. The most energy-efficient operative points are reported in **bold** (per NN).

TABLE III  
REGRESSION PERFORMANCE FOR OUR NETWORKS (QUANTIZED) AND PROXIMITY NN [11] (FULL PRECISION).

Network	MAE [ $\cdot 10^{-3}$ ]				MSE [ $\cdot 10^{-3}$ ]			
	$x$	$y$	$z$	$\theta$	$x$	$y$	$z$	$\theta$
$160 \times 32$	<b>195</b>	192	101	482	<b>66</b>	<b>78</b>	<b>20</b>	386
$160 \times 16$	203	191	110	492	74	83	25	412
$80 \times 32$	226	<b>178</b>	106	556	88	84	29	504
NN [11]	210	219	<b>96</b>	<b>473</b>	79	91	21	<b>385</b>

its reduced variance. Compared to MAE, MSE values further penalize large errors: on this metric, the  $160 \times 32$  model consistently outperforms smaller networks, which is explained by its utilization in training. Additionally, Table III shows how the proposed PULP-Frontnet models have comparable performance with the original Proximity network, which has many more parameters and higher computational/memory requirements (see Section IV-A and Table II).

Figure 9 presents a comparison based on the coefficient of determination  $R^2$ , a standard adimensional metric for regression performance.  $R^2$  represents the fraction of the variance in the target variable explained by the model (higher is better). A trivial model that predicts the average of the target variable for the whole testing dataset yields  $R^2 = 0$ ; in this case, the MSE is the same as the target variable's variance – i.e., the model captures no useful information other than the average. At the other extreme, an ideal model yields  $R^2 = 1.0$ , and lastly, if  $0 < R^2 < 1$ , the model can account for only part of the variance in the data. A model might have a negative  $R^2$  in case its MSE exceeds the variance of the data, which frequently occurs with models operating on high-dimensional inputs. Note that comparing MAE and MSE metrics across different output variables can be misleading. For example, on a variable with very low variance (such as  $z$ ), even a trivial model that returns the variable's average would yield very low MAE and MSE. Therefore, when comparing across variables, the  $R^2$  metric is a better indication of regression performance, which does not depend on the output variable's variance.

Figure 9 shows a consistent pattern among all networks: the prediction is best for  $x$  and  $y$  variables ( $R^2 > 0.5$ ).  $z$ , which encodes height, proves harder, and  $\theta$ , representing the subject's head orientation, results being the most complex variable to estimate. The estimation of  $\theta$  is more sensitive to challenging testing images, such as those where the subject is very far, or looking away from the drone, or only partly visible. Additionally, the limited image resolution has a higher impact on  $z$  and  $\theta$  than on  $x$  and  $y$ , as confirmed by the systematic loss in accuracy reducing the input image's size from the  $160 \times 16$  to the  $80 \times 32$  NN. Lastly, considering the effect of quantization vs. full precision, it introduces an approximation of the original network, but it also has a beneficial regularization effect, reducing the parameter space. In the case of the experiment in Figure 9, these effects are well balanced, and the differences between quantized and full precision models are well within noise margins.

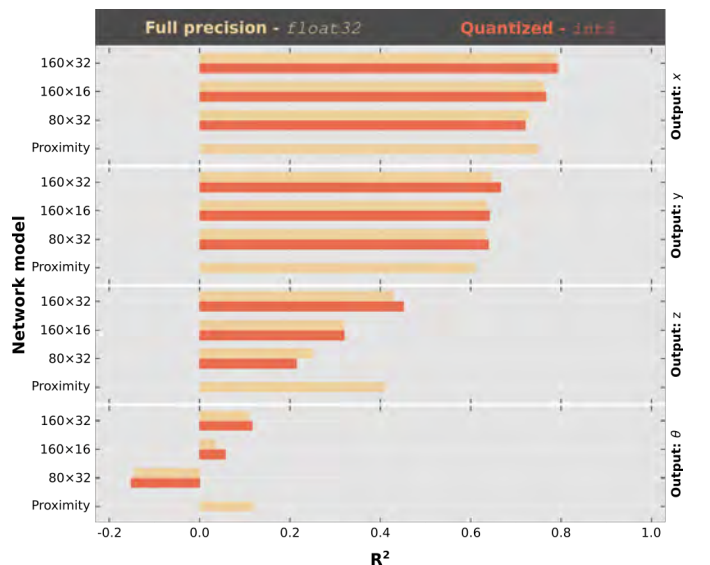


Fig. 9. Regression performance ( $R^2$ ) on the testing dataset.



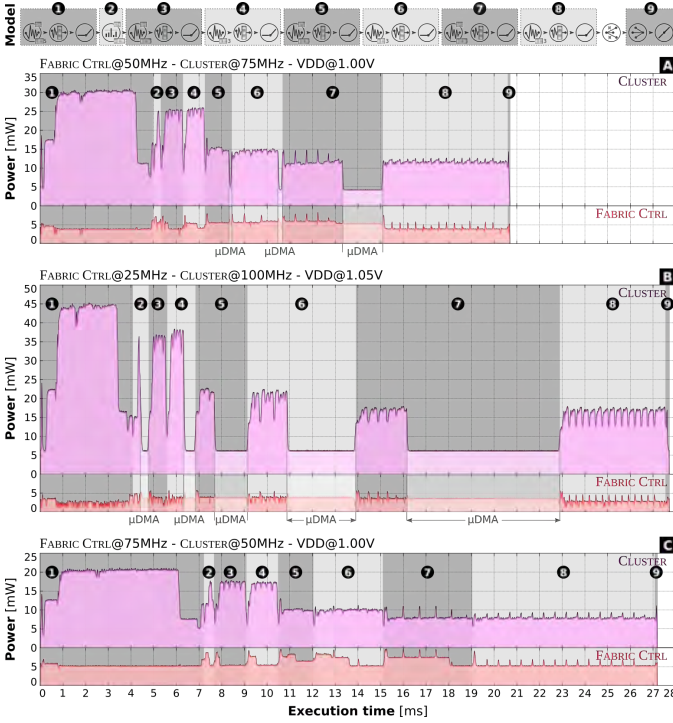


Fig. 10. Power traces of PULP-Frontnet  $80 \times 32$  for one-frame inference in three operative points: A) FC@50 MHz, CL@75 MHz, B) FC@25 MHz, CL@100 MHz, and C) FC@75 MHz, CL@50 MHz. Every trace highlights all the computational stages reported in the **Model**. Measurements are taken after the internal DC/DC converter (i.e., accounting for both FABRIC CTRL and CLUSTER separately).

### B. Onboard performance

Our performance investigation starts from Figure 8, where we show the PULP-Frontnet energy efficiency sweeping all the operative points of the GAP8 SoC. FC and CL frequency are explored with a growing step of 25 MHz, affecting also the minimum voltage required to enable the desired frequencies (VDD growing step 0.05 V), as shown in Figure 8-A. The maximum frequencies, and the required VDD, are selected according to the GAP8 SoC datasheet<sup>9</sup>. For each network, i.e., PULP-Frontnet  $160 \times 32$ ,  $160 \times 16$ , and  $80 \times 32$ , we show the energy efficiency heat-map as the required energy to perform the inference on one frame. The model  $160 \times 32$  is the least efficient of the three, as it requires 14.0 MMAC operations per frame, resulting in 1.28 mJ/frame running at its best configuration of FC@25-50 MHz-CL@75 MHz. Both remaining models, i.e.,  $160 \times 16$  and  $80 \times 32$ , show a higher – and similar – energy efficiency, due to their reduced number of operations, as much as 4.3 MMAC and 4.0 MMAC per frame, respectively. In Figure 8-B, we highlight the most energy efficient operative points: FC@25 MHz-CL@50 MHz for the model  $160 \times 16$ , and FC@50-75 MHz-CL@75 MHz for the  $80 \times 32$  one, consuming 0.58 mJ/frame and 0.43 mJ/frame, respectively. The rest of this section will refer to these most energy-efficient configurations to evaluate both power consumption and inference performance.

Figure 10 shows the power traces for one-frame inference, reporting both FABRIC CTRL (FC) and CLUSTER (CL) power domains separately. Figure 10-A refers to the most energy-efficient operative point highlighted in Figure 8-B, i.e., FC@50 MHz, CL@75 MHz ( $80 \times 32$ ). This configuration exhibits some cluster idleness, up to almost 2 ms within layer 7. As reported on the x-axis, the  $\mu$ DMA activity (data transfers from the off-chip DRAM to the on-chip L2 memory) does not entirely overlap with the cluster's computation. As the  $\mu$ DMA bandwidth depends on the FC's frequency, it is possible to end up in an operative scenario for which the  $\mu$ DMA can not satisfy the data's demand of a faster cluster. This situation can be even more exacerbated, reducing on one side the FC's frequency and at the same time increasing the CL's one, e.g., FC@25 MHz, CL@100 MHz as reported in Figure 10-B. In this case, the CL's idleness increases being well visible for layers 2, 4, 5, 6, and 7. On the contrary, by selecting an operative point such as FC@75 MHz, CL@50 MHz (Figure 10-C), we can hide all the  $\mu$ DMA latencies, obtaining a perfect pipelining between the cluster computation and  $\mu$ DMA data transfers, therefore, avoiding any CL's idleness.

Despite the CL's idle time of the first configuration, i.e., FC@50 MHz, CL@75 MHz, it results being the most energy-efficient as minimizing the  $\mu$ DMA overhead (i.e., the configuration in Figure 10-C brings, as a consequence, a higher mean power consumption of the FC, i.e., 5.9 mW instead of 4.7 mW. From the energy point of view, the two configurations (Figure 10-A vs. C) have a similar energy cost for the CL, i.e., 320 mJ and 327 mJ, respectively, but a very different one for the FC's domain, as 98 mJ vs. 161 mJ. This extra cost ( $\sim 63$  mJ) is more than  $3\times$  the energy overhead for the idleness in the first configuration, turning in a less efficient configuration. In all power traces, the first computational stage 1 is the most power-hungry due to better utilization of the eight general-purpose cores within the cluster. Feeding in input the full image allows for an almost-ideal spatial ( $width \times height$ ) parallelization that is not always the case for the remaining convolutional layers, which operate on very small spatial dimensions ( $3 \times 2$ ) but very deep input tensor, as in layer 8 with 128 channels, given the PULP-NN spatial parallelization scheme (each core operates on different chunks of spatial data with all input and output channels).

Figure 11 assesses the inference throughput (frame/s) vs. power consumption of the three NN models. Each model is evaluated in three different configurations: i) the one that generates the highest throughput, ii) a second one referring to the most energy-efficient operative point identified in Figure 8, and iii) a last one for the minimum power consumption. Each configuration reports both mean and peak power, the latter as a marker on top of each icon. All three NNs are almost iso-power for a given configuration, showing how the difference in their respective overall energy efficiency comes from a reduced execution time, i.e., increased throughput. The model PULP-Frontnet  $80 \times 32$  shows the minimum mean power of 8.6 mW paired with a performance of 18.5 frame/s. The most power-hungry configuration is represented by the model  $160 \times 16$  running at maximum frequency (FC@250 MHz, CL@175 MHz)

<sup>9</sup>[https://gwt-website-files.s3.amazonaws.com/gap8\\_datasheet.pdf](https://gwt-website-files.s3.amazonaws.com/gap8_datasheet.pdf)

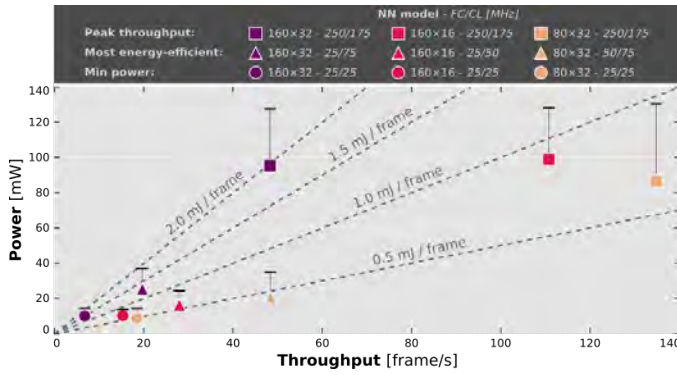


Fig. 11. Throughput vs. power consumption for all three NNs, each in three different operative points: *i*) peak throughput, *ii*) the most energy-efficient point highlighted in Figure 8, and *iii*) minimum power consumption. Dashed gray lines show the levels of energy efficiency in mJ/frame.

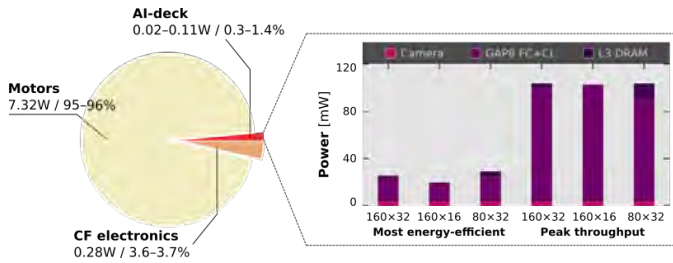


Fig. 12. The nano-drone's power envelope break-down, with AI-deck zoom-in. In the right plot the GAP8 SoC is shown for both the most energy-efficient operative points and the peak throughput ones.

achieving 110.7 frame/s within 99 mW. On the contrary, the overall peak throughput is given by the model 80x32, reaching 134.7 frame/s, with a total SoC power consumption of 86.6 mW. Lastly, considering the most energy-efficient configuration for all three NNs, the model 80x32 results 2.5x and 1.7x faster than the models 160x32 and 160x16, respectively, with very similar power consumption (~20 mW).

In Figure 12 is reported the power brake-down for the entire nano-UAV. As shown in the plot on the left, the vast majority of the total power consumption is represented, as expected, by the Crazyflie (CF) electric motors, accounting for 95–96% of the total, depending on the specific GAP8's configuration. The CF electronics (i.e., LEDs, STM32 MCU, etc.) consumes a small fraction of the total, such as 3.6–3.7%, leaving for the AI-deck the smallest part, i.e., between 0.3 and 1.4%, depending on the SoC's operative point, and shown on the plot on the right. In this zoom-in plot, we consider the two SoC's configurations previously introduced, namely the *most energy-efficient* (different for each model, see Figure 8) and the *peak throughput* one (FC@250 MHz, CL@175 MHz). Additionally, we show the constant power consumption for the Himax camera (i.e., 4 mW) and the power required to perform all the L3-L2 weight transfers (not present in the model 160x16 as all the weights are pre-loaded in the GAP8's L2 memory). From this analysis, we see how the AI-deck power consumption is *i*) always dominated by the GAP8 component and *ii*) it never exceeds 110 mW when running at the maximum frequencies. Such a small impact on the whole

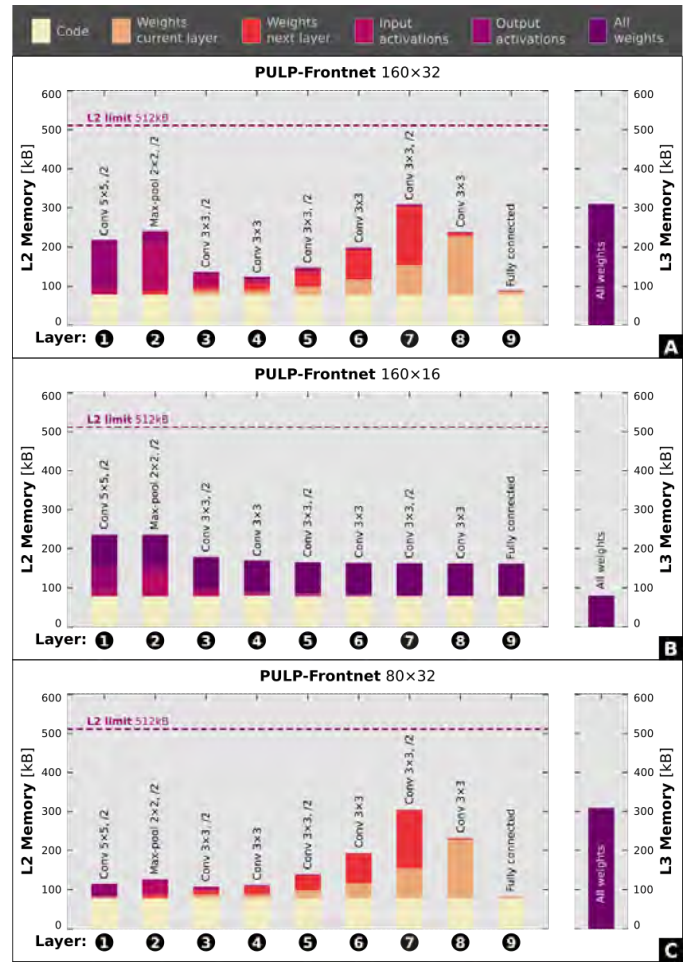


Fig. 13. Memory pressure on both L2 and L3 memory for each NN's layer. PULP-Frontnet A) 160x32, B) 160x16, and C) 80x32.

system's power budget demonstrates running PULP-Frontnet at the highest performance point and enables the possibility to further extend the onboard intelligence with additional tasks, aiming at more complex mission objectives.

Figure 13 reports the memory pressure on both L2 and L3, for each NN model, namely PULP-Frontnet 160x32, 160x16, and 80x32. Every sub-figure (A-B-C) shows the amount of L2 memory required at every layer for: *i*) code allocation, *ii*) parameters (i.e., weights) occupancy for the current layer, *iii*) weights allocation for the next layer computation, *iv*) input, and *v*) output buffers (i.e., activations). The only exception to this representation is given in Figure 13-B where we keep all the NN's weights always in L2 – due to their small footprint – avoiding the indication of *current layer* and *next layer* weights allocation. Each subplot also shows the L2 memory upper-bound of the GAP8 SoC, as a dashed line at 512 kB. On the right area of each sub-figure, we also report the total parameters memory footprint located in L3 (i.e., DRAM) that the  $\mu$ DMA needs to transfer in L2. In Figure 13-A/C, during each layer's computation, the  $\mu$ DMA transfers the weights required for the next layer, overlapping it with the current computation (i.e., double buffering scheme). In Figure 13-B all the weights are transferred before the inference starts and



are kept available in L2 for the entire application's lifetime, without any additional need to move data from L3.

For all three proposed NNs, the code footprint is almost constant  $f$ , i.e.,  $\sim 80$  kB and the first layer is the one which requires the largest *output activations* buffer that serves as the input buffer for the second layer. All three models show a slight pressure on both L2 and L3 memory, as the only potential violation of the L2 upper-bound is represented by the second layer of PULP-Frontnet  $160 \times 32$ , that would require 541 kB to allocate all weights and intermediate buffers simultaneously in L2. As already introduced in Section III, the missed opportunity for better exploitation of the L2 memory is a consequence of the deployment tool, i.e., DORY, that is designed to work with NN models characterized by a higher volume of L3 data. This situation highlights the possibility of the proposed PULP-Frontnet to run together with additional workloads, paving the ground for advanced multi-tasks execution of autonomous navigation algorithms on the GAP8 SoC.

### C. In-field control accuracy

We conclude the experimental analysis by putting the whole system to the test, i.e., pose estimation task and autonomous navigation, only with the sensory information and computational resources aboard our nano-drone prototype. We design an experiment where a subject follows a predefined path, while the drone's task is to stay in front of them, at a fixed distance  $\Delta = 1.3$  m. We assess the task's quality, using the networks and controller presented in Section IV. We repeat the experiment multiple times using the three proposed NN topologies, and compare how well the drone tracks the target pose in each run. We foster our key findings with a video of the proposed setup/experiment, showing the whole closed-loop system's capability at the following link: <https://github.com/idsia-robotics/pulp-frontnet>.

**Experimental setup.** When testing the system, we noticed that subjects tend to adapt their motion to the drone behavior, e.g., a drone reacting slowly and erratically leads them to move slower. For our experiment, this would make different runs not comparable to each other. To ensure objective measurements across runs, we ask subjects to completely ignore the drone behavior as they move: this is possible since the drone is very small and subjects don't feel threatened by potential collisions. To control the subject's motion, we add markers to the floor for each step to be taken: subjects are instructed to step every beat of a metronome, and therefore move at the same speed in every run, independently on the drone behavior. Figure 14 illustrates the setup and the path that our subjects are instructed to follow. The entire pattern takes 50 s among eight phases (0-7), and no pause is taken between them.

**Init:** The run starts with the subject standing at pose  $\mathcal{H}_0$ , facing towards the top side of the map. The drone is initially hovering at pose  $\mathcal{D}_0$ , 3.6 m in front of the subject. The drone points  $30^\circ$  to the left of the subject; the camera can therefore see the subject, facing towards it. The camera field-of-view is highlighted in Figure 14.

**Phase 0 (5 s):** The subject stands still for 5 s; during this time, the drone is expected to rotate left by  $30^\circ$  and move forwards to go at 1.3 m in front of the subject.

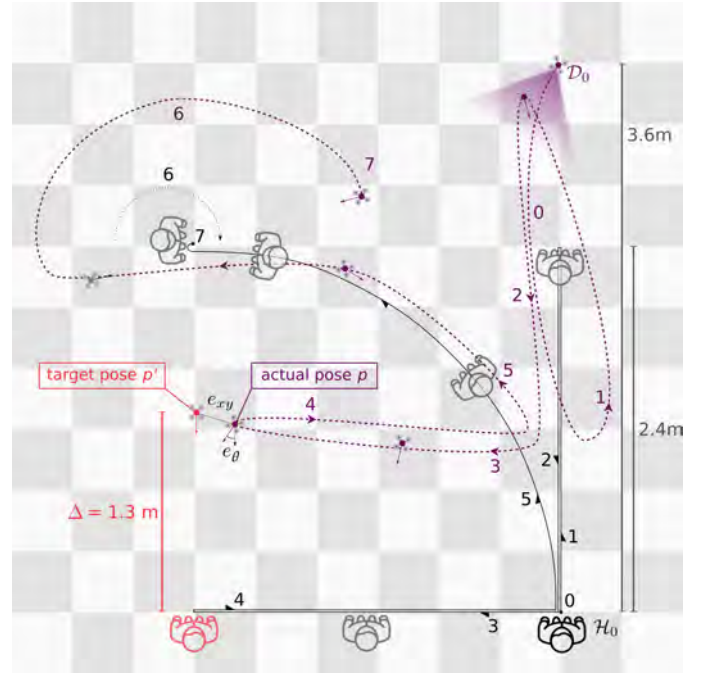


Fig. 14. Top-view of the in-field experimental setup (description in text).

**Phases 1/2 (12 s):** The subject walks forward covering 2.4 m in 6 s, and then backward for the same time and distance. The drone is expected to track the user by moving backward and then forward.

**Phases 3/4 (14 s):** Without changing their orientation (i.e., still facing towards the top side of the map), the subject moves sideways towards their left for 2.4 m and 7 s, and then towards their right.

**Phase 5 (6 s):** The subject walks along a quarter of a circle, with radius 2.4 m, counterclockwise, facing the direction of the path, in 6 s. At the end of this phase, the subject faces towards the left of the map.

**Phase 6 (8 s):** The subject rotates in place, clockwise, by  $180^\circ$ , in 8 s; at the end of this phase, the subject faces towards the right of the map. The drone is expected to perform a half circle with 1.3 m radius while always pointing at its center.

**Phase 7 (5 s):** The subject stands in place for 5 s (this gives the drone enough time to complete its motion).

Note that the experiment challenges the drone with increasingly difficult tasks: reaching a standing target in phase 0; following a target that moves without changes in orientation in phases 1-4; keeping track of a target that moves and rotates in phase 5; staying in front of a target that spins in place in phase 6. Note in particular that phases 5-7 test the drone's ability to predict the orientation of the subject's head, which is the most challenging component of the pose to predict (Section V-A).

For every experimental run, we record the output of inference, and the true poses of both subjects  $p_H^W$  and drone  $p_D^W$ , captured in the world-fixed motion capture frame  $\mathcal{W}$ . From these data, in post-processing, we build a dataset consisting of a list of predicted  $\hat{p}_H^D$  and ground truth  $p_H^D$  poses of the subject relative to the drone. We run the experiment with two different

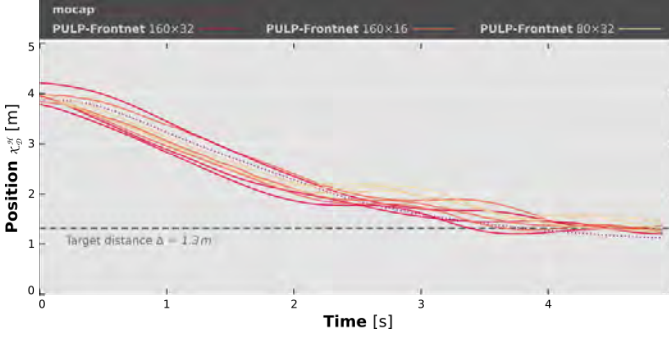


Fig. 15. Ground truth  $x$  component of the drone pose with respect to the subject during phase 0, for all 10 runs (3 for each network, plus 1 using ground truth relative poses). The dashed gray line represents the target distance  $\Delta = 1.3$  m.

subjects, neither of whom is part of the training dataset. Before the experiment, each subject practices the timed movements a few times, as they require some coordination, especially in sudden changes of direction. For each subject, we perform ten runs: three runs for each of the three networks described in Section IV-A, in which the model is run at its maximal throughput operative point; and one run, acting as an upper bound of the achievable performance, where the controller uses as input the ground truth relative pose (measured by the motion tracking system), yielding to a total of twenty runs.

**Metrics.** For each run, we measure the following metrics. To measure the models' ability to interpret the images, we compute the  $R^2$  of the prediction with respect to the ground truth for each model's output variable; this metric is comparable with those reported in Section V-A on the testing set. In this case, however, the images are acquired during flight, with continuously variable pitch and roll, and with a different distribution of the target variable values, as the drone is actively tracking the subject. To measure how well the drone is performing the task, we compute statistics on the difference between the drone target pose  $p'$  (calculated from the ground truth pose of the subject) and the drone's actual pose  $p$  during the entire run. An example of the two poses at a specific point in time (the end of phase 3) is represented in red and violet, respectively, in Figure 14. In particular, we separately quantify:

- $e_{xy}$ : the horizontal component of the distance between  $p$  and  $p'$  (absolute position error);
- $e_\theta$ : the difference in orientation between  $p$  and  $p'$  (absolute angular error).

Note that we ignore the  $z$  component of the position error, because the target height is approximately constant in our task.

**Results.** Figure 15 illustrates how the drone approaches the subject (who is standing still) during phase 0 of each run. In this phase, the drone is expected to rotate  $30^\circ$  to its left and reach a distance of 1.3 m from the subject standing at an initial pose approximately 3.6 m away. When provided pose estimations from inference, the trajectory converges marginally slower to the target pose than the trajectory observed with perfect estimations. In particular, minor oscillations in the final part of the trajectory are caused by errors in the distance prediction, depending on the drone pitch. As the drone gets closer to the target pose, the model detects a decreasing

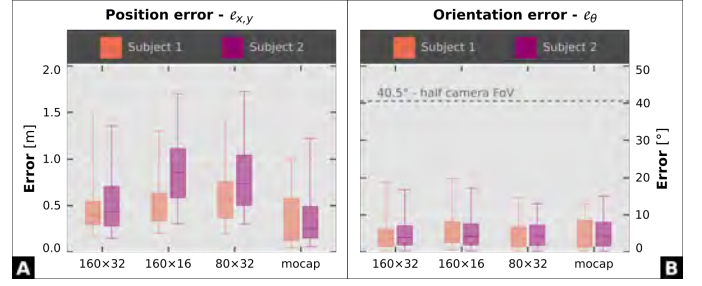


Fig. 16. Distribution of control errors for the two subjects (color), using the proposed models, A) position error and B) orientation error. Boxplot whiskers mark  $5^{th}$  and  $95^{th}$  percentile of data. The dashed line on the right plot marks the half-field of view of the camera ( $40.5^\circ$ ): when the angular error is smaller, the subject is visible in the frame.

distance, which yields the controller to pitch up to decelerate: the changes in the image due to the different camera pitch lead the model to estimate a slightly higher distance of the subject than previously thought, which pushes the drone to pitch down again to get a little closer; due to the synthetic pitch augmentation technique described in Section IV-B, these pitch-dependent errors are small enough that the control is stable; in contrast, experiments on models trained without pitch augmentation show unstable behavior<sup>10</sup>.

TABLE IV  
IN-FIELD EXPERIMENT RESULTS.

Network	Rate [Hz]	Regression $R^2$			Median pose error	
		$x$	$y$	$\theta$	$e_{xy}$ [m]	$e_\theta$ [ $^\circ$ ]
$160 \times 32$	48	0.87	<b>0.87</b>	<b>0.71</b>	<b>0.41</b>	<b>3.7</b>
$160 \times 16$	111	<b>0.93</b>	0.82	0.56	0.61	4.7
$80 \times 32$	<b>135</b>	0.88	0.75	0.54	0.63	4.0
mocap	30	1.00	1.00	1.00	0.26	4.1

Table IV reports the rate of the high-level controller on the STM32 (see Section IV-D) and summarizes the metrics over all runs. The  $R^2$  scores indicate good regression performance; in comparison with scores obtained on the testing set (Figure 9), all variables ( $\theta$  in particular) are estimated significantly better, for all networks. For example, network  $160 \times 32$  improves its  $R^2$  from (0.70, 0.67, 0.12) to (0.87, 0.87, 0.71) on variables ( $x, y, \theta$ ). This improved performance is a consequence of the closed-loop system under test actively following the subject. In fact, in the in-field tests, the images are acquired mainly from a frontal pose and a close distance, creating a virtuous circle where the better the drone follows the subject, the easier it is to predict the correct pose. In general, performance trends across various networks and different variables match those observed on the more challenging testing set.

Table IV also reports the median values for metrics  $e_{xy}$  and  $e_\theta$ , and Figure 16 illustrates their distribution, separately on the

<sup>10</sup>We have also tested the system while subjects move freely. While we do not report quantitative results for these runs, we have observed that: *i*) the drone's motion is stable against small movements of the subjects; *ii*) the drone is generally able to track free-to-move subjects; *iii*) tracking failures are primarily due to the drone's angular speed being limited: fast-moving subjects may exit the field of view.



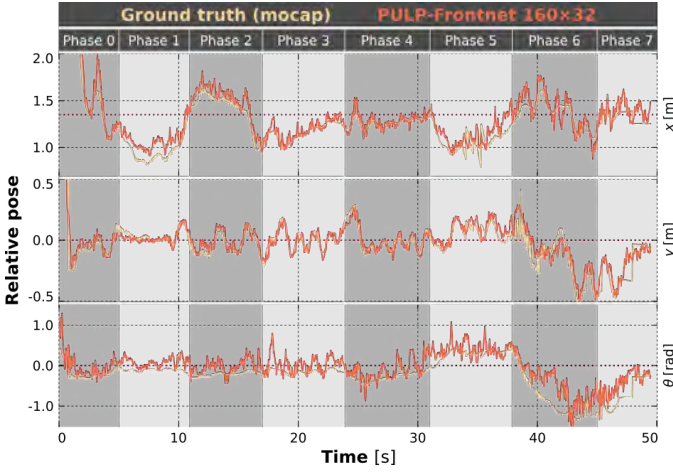


Fig. 17. Ground truth (yellow) and network prediction (orange) versus time, during a single experimental run using the  $160 \times 32$  network, reporting  $x$ ,  $y$ , and  $\theta$  components of the relative pose. The horizontal dashed line (violet) corresponds to the relative pose that the controller attempts to keep. We highlight the time-intervals associated with the different phases of the run.

two subjects. For all runs, the angular error  $e_\theta$  is consistently lower than  $20^\circ$ , with a median value below  $5^\circ$ . This shows that, for both subjects, all neural networks effectively estimate the subject's horizontal position to adjust the drone yaw to keep the user in the center of the frame. In contrast, the position error  $e_{xy}$ , while generally below 1.0 m, changes depending on the network and the subject. The lower bound for this error is given by the mocap runs, in which the controller is fed with the ground truth relative pose of the subject w.r.t. the drone. We observe that the position error obtained using our best model is less than two times larger than the lower bound. Lower position errors correspond to networks with higher  $R^2$  values, with the  $160 \times 32$  network performing best on both metrics. It is important to note that mistakes in estimating the user's head orientation (i.e., on the  $\theta$  output variable) are reflected as *position* errors in this experiment.

Figure 17 compares the network prediction to the ground truth during a single run. We observe that, for all pose components, the prediction tracks the ground truth without systematic bias and with good accuracy. During the first few seconds (phase 0),  $x$  is correctly estimated much above the desired value, which leads the drone to come closer to the user quickly. For example, during phases 1 and 2, when the subject walks forward and then backward, the robot moves to keep the distance to the desired value, making the  $x$  component reflects the user's movements. For  $\theta$ , predictions are noisier than  $x$  and  $y$ , but still manage to capture large-scale patterns in the target variable. This effect is well visible in the last 10 seconds, i.e., phases 6 and 7, where the ground truth of  $\theta$  is consistently negative as the drone tracks the user's in-place rotation; the prediction properly also captures this pattern.

## VI. CONCLUSION

In this work, we presented PULP-Frontnet, a novel CNN that visually estimates the pose of a freely-moving human subject, controlling the robot to stay at a constant distance in front of them. Solving this HDI problem on an autonomous

nano-drone is a challenging and valuable task in the IoT domain. These robotic helpers can be envisioned as the next-generation ubiquitous IoT devices, ideal for indoor operations near humans. We propose a general methodology for CNNs' architecture design, dataset collection and augmentation strategies, 8-bit quantization, and deployment on a PULP-based multi-core SoC (i.e., the GWT GAP8). We consider three CNN variants with different trade-offs on computation, performance, power envelope, and memory needs, running on the GAP8 SoC aboard a COTS Crazyflie 2.1 nano-quadrotor (i.e., 27 grams). Our results show a remarkable peak performance of 135 frame/s onboard inference rate within only 86 mW power consumption. Our CNN shows the same regression performance of the resource-unconstrained full-precision baseline, even involving subjects never seen during training. In-field experiments, made with a fully integrated demonstrator, exhibit excellent control performance (median absolute angular error below  $5^\circ$ ) with minimal resource use (down to 4.3 MMAC/frame operations, and 184 kB memory footprint). With a peak energy efficiency of 0.43 mJ/frame, we leave more than enough computational power for additional data analytics aboard our nano-UAVs, paving the way to the ultimate mobile IoT edge-node.

## REFERENCES

- [1] H. Shakhathreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019.
- [2] N. H. Motlagh, T. Taleb, and O. Arouk, "Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives," *IEEE Internet of Things Journal*, vol. 3, no. 6, Dec 2016.
- [3] B. Alzahrani, O. S. Oubbati, A. Barnawi, M. Atiquzzaman, and D. Alghazzawi, "Uav assistance paradigm: State-of-the-art in applications and challenges," *Journal of Network and Computer Applications*, vol. 166, p. 102706, 2020.
- [4] N. Smolyanskiy, A. Kamenev, J. Smith, and S. Birchfield, "Toward low-flying autonomous mav trail navigation using deep neural networks for environmental awareness," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017.
- [5] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebsch, J. Nieto, and R. Siegwart, "weednet: Dense semantic weed classification using multispectral images and mav for smart farming," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 588–595, 2018.
- [6] D. Palossi, A. Marongiu, and L. Benini, "Ultra low-power visual odometry for nano-scale unmanned aerial vehicles," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017. IEEE, 2017, pp. 1647–1650.
- [7] B. Bodin, H. Wagstaff, S. Saecdi, L. Nardi, E. Vespa, J. Mawer, A. Nisbet, M. Lujan, S. Furber, A. J. Davison, P. H. J. Kelly, and M. F. P. O'Boyle, "Slambench2: Multi-objective head-to-head benchmarking for visual slam," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3637–3644.
- [8] Y. Li, N. Brasch, Y. Wang, N. Navab, and F. Tombari, "Structure-slam: Low-drift monocular slam in indoor environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6583–6590, 2020.
- [9] Y. Lu, Z. Xue, G.-S. Xia, and L. Zhang, "A survey on vision-based uav navigation," *Geo-spatial information science*, vol. 21, no. 1, pp. 21–32, 2018.
- [10] D. Floreano and R. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, pp. 460–6, 05 2015.
- [11] D. Mantegazza, J. Guzzi, L. M. Gambardella, and A. Giusti, "Vision-based control of a quadrotor in user proximity: Mediated vs end-to-end learning approaches," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6489–6495.
- [12] C. Avasalcai, I. Murturi, and S. Dustdar, "Edge and fog: A survey, use cases, and future challenges," *Fog Computing: Theory and Practice*, pp. 43–65, 2020.

- [13] R. J. Wood, B. Finio, M. Karpelson, K. Ma, N. O. Pérez-Arancibia, P. S. Sreetharan, H. Tanaka, and J. P. Whitney, *Progress on "Pico" Air Vehicles*. Cham: Springer International Publishing, 2017.
- [14] D. Palossi, A. Loquercio, F. Conti, E. Flamand, D. Scaramuzza, and L. Benini, "A 64-mw dnn-based visual navigation engine for autonomous nano-drones," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8357–8371, 2019.
- [15] F. Conti, D. Palossi, A. Marongiu, D. Rossi, and L. Benini, "Enabling the heterogeneous accelerator model on ultra-low power microcontroller platforms," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 1201–1206.
- [16] D. Palossi, F. Conti, and L. Benini, "An open source and open hardware deep learning-powered visual navigation engine for autonomous nano-uavs," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 604–611.
- [17] F. Conti, R. Schilling, P. D. Schiavone, A. Pullini, D. Rossi, F. K. Gürkaynak, M. Muehlberghuber, M. Gautschi, I. Loi, G. Haugou, S. Mangard, and L. Benini, "An iot endpoint system-on-chip for secure and energy-efficient near-sensor analytics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2481–2494, 2017.
- [18] D. Gandhi, L. Pinto, and A. Gupta, "Learning to fly by crashing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, sep 2017.
- [19] F. Conti, "Technical report: Nemo dnn quantization for deployment model," *arXiv preprint arXiv:2004.05930*, 2020.
- [20] A. Burrello, A. Garofalo, N. Bruschi, G. Tagliavini, D. Rossi, and F. Conti, "Dory: Automatic end-to-end deployment of real-world dnns on low-cost iot mcus," *IEEE Transactions on Computers*, pp. 1–1, 2021.
- [21] Y. Hiroi and A. Ito, "Influence of the size factor of a mobile robot moving toward a human on subjective acceptable distance," *Mobile Robots-Current Trends*, pp. 177–190, 2011.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.
- [23] G. Shi, W. Hönig, Y. Yue, and S. J. Chung, "Neural-swarm: Decentralized close-proximity multirotor control using learned interactions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3241–3247.
- [24] B. Broecker, K. Tuyls, and J. Butterworth, "Distance-based multi-robot coordination on pocket drones," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6389–6394.
- [25] K. Kang, S. Belkhale, G. Kahn, P. Abbeel, and S. Levine, "Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6008–6014.
- [26] O. Andersson, M. Wzorek, and P. Doherty, "Deep learning quadcopter control via risk-aware active learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 3812–3818.
- [27] W. Zhao, A. Goudar, J. Panerati, and A. P. Schoellig, "Learning-based bias correction for ultra-wideband localization of resource-constrained mobile robots," *arXiv preprint arXiv:2003.09371*, 2020.
- [28] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, "Iot: Internet of threats? a survey of practical security vulnerabilities in real iot devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8182–8201, 2019.
- [29] A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, and V. Sze, "Navion: A 2-mw fully integrated real-time visual-inertial odometry accelerator for autonomous navigation of nano drones," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, 2019.
- [30] Z. Li, Y. Chen, L. Gong, L. Liu, D. Sylvester, D. Blaauw, and H. Kim, "An 879gops 243mw 80fps vga fully visual cnn-slam processor for wide-range autonomous exploration," in *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2019, pp. 134–136.
- [31] J.-H. Yoon and A. Raychowdhury, "31.1 a 65nm 8.79 tops/w 23.82 mw mixed-signal oscillator-based neuroslam accelerator for applications in edge robotics," in *2020 IEEE International Solid-State Circuits Conference-ISSCC*. IEEE, 2020, pp. 478–480.
- [32] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gürkaynak, and L. Benini, "Near-threshold risc-v core with dsp extensions for scalable iot endpoint devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, Oct 2017.
- [33] A. Pullini, D. Rossi, G. Haugou, and L. Benini, "μdma: An autonomous i/o subsystem for iot end-nodes," in *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, 2017, pp. 1–8.
- [34] I. Loi, A. Capotondi, D. Rossi, A. Marongiu, and L. Benini, "The quest for energy-efficient i\$ design in ultra-low-power clustered many-cores," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, no. 2, pp. 99–112, 2018.
- [35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 07–09 Jul 2015, pp. 448–456.
- [36] J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang, "Accurate and efficient 2-bit quantized neural networks," in *Proceedings of the 2nd SysML Conference*, vol. 2019, 2019.
- [37] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.



**Daniele Palossi** received the Ph.D. in Information Technology and Electrical Engineering from the ETH Zürich in 2019 and is currently a postdoctoral researcher at the Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland, and the Integrated Systems Laboratory, ETH Zürich, Switzerland. His research focuses on the embedded domain with special emphasis on energy-efficient ultra-low-power platforms, algorithms for autonomous navigation, and resource-constrained small-size cyber-physical systems. He has been the recipient of the Swiss National Science Foundation (SNSF) Spark Grant and the 2<sup>nd</sup> prize at the Design Contest held at the ACM/IEEE ISLPED'19.



**Nicky Zimmerman** is a Ph.D. student at the Photogrammetry Lab at the Rheinische Friedrich-Wilhelms-Universität Bonn since March 2021. She received her M.Sc. in Informatics from Università della Svizzera italiana with a thesis on Embedded Implementation of Vision-Based Navigation for Nano-drones. She previously worked for General Motors and Intel.



**Alessio Burrello** received his B.Sc. and M.Sc. degree in Electronic Engineering at the Politecnico of Turin, Italy, in 2016 and 2018. He is currently working toward his Ph.D. degree at the Department of Electrical, Electronic and Information Technologies Engineering (DEI) of the University of Bologna, Italy. His research interests include parallel programming models for embedded systems, machine and deep learning, hardware oriented deep learning, and code optimization for multi-core systems.



**Francesco Conti** received the Ph.D. in Electronic Engineering from the University of Bologna, Italy, in 2016. He is currently an Assistant Professor in the DEI Department of the University of Bologna. From 2016 to 2020, he held a position as postdoctoral researcher at the Integrated Systems Laboratory of ETH Zürich in the Digital Systems group, and a research grant from the University of Bologna. His research focuses on the development of deep learning based intelligence on top of ultra-low power, ultra-energy efficient programmable Systems-on-Chip – from both the hardware and software perspective. His work has resulted in 50+ publications in international conferences and journals and has been awarded several times, including the 2020 IEEE TCAS-I Darlington Best Paper Award, the 2018 Hipec Tech Transfer Award, the 2018 ESWEEK Best Paper Award, and the 2014 ASAP Best Paper Award.



**Alessandro Giusti** received the Ph.D. in Computer Science from Politecnico di Milano, Italy, in 2009. He is currently Professor in Artificial Intelligence for Autonomous Robotics at the Dalle Molle Institute for Artificial Intelligence (IDSIA) in Lugano (Switzerland), affiliated with the University of Applied Sciences and Arts of Southern Switzerland (SUPSI) and Università della Svizzera italiana (USI). His research focuses on human-robot interaction and self-supervised deep learning for perception tasks in aerial, ground and industrial robotics. His work has resulted in 100+ publications in international conferences and journals and has been awarded several times, including the 2020 IEEE Communications Society Charles Kao Award, Best Demo and Best Video Awards at HRI 2019, the shortlist for the AAAI 2016 Video Competition, the 1<sup>st</sup> place in three international competitions on biomedical image analysis and four additional best paper awards.



**Jérôme Guzzi** holds a master's degree in physics from ETH Zürich and a Ph.D. in computational science from USI Lugano. He focuses his research on mobile robots, multi-robot systems, and human-robot interaction.



**Hanna Müller** received her B.Sc. and M.Sc. in Electrical Engineering and Information Technologies from ETH Zürich, Switzerland, in 2017 and 2020. She is currently pursuing a Ph.D. degree with the Integrated Systems Laboratory at ETH Zürich. Her research interests include low-power systems, nano-drones, autonomous navigation, mapping and drone swarms.



**Luca Maria Gambardella** is currently an Internationally Recognized AI Expert, with 35 years of experience in the field. He is also a Full Professor with the Faculty of Informatics, USI, Lugano, and a Professor with the IDSIA USI-SUPSI, which he directed for 25 years. He is the Co-Founder, CTO, and the Head of applied AI at Artificialy in Lugano, Switzerland. Scientifically, he is one of the pioneers of the ant colonies optimization metaheuristics. In his career, he has successfully negotiated, conducted, and delivered many AI and ML projects with Swiss and international research agencies and companies. He has published more than 300 scientific articles, with H-index 74, and more than 61500 citations.



**Luca Benini** is the Chair of Digital Circuits and Systems at ETH Zürich and a Full Professor at the University of Bologna. He has served as Chief Architect for the Platform2012 in STMicroelectronics, Grenoble. Dr. Benini's research interests are in energy-efficient systems and multi-core SoC design. He is also active in the area of energy-efficient smart sensors and sensor networks. He has published more than 1'000 papers in peer-reviewed international journals and conferences, five books and several book chapters. He is a Fellow of the ACM and a member of the Academia Europaea.