

Copyright © 2022 Association for Computing Machinery.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

This is the peer-reviewed version of the following original article:

Title: PhiNets: a scalable backbone for low-power AI at the edge

The final version of this paper appears in ACM Transactions on Embedded Computing Systems, Published by ACM, <https://doi.org/10.1145/3510832>

PhiNets: a scalable backbone for low-power AI at the edge

FRANCESCO PAISSAN*, ALBERTO ANCILOTTO*, and ELISABETTA FARELLA, E3DA Unit, Digital Society Center - Fondazione Bruno Kessler (FBK)

In the Internet of Things era, where we see many interconnected and heterogeneous mobile and fixed smart devices, distributing the intelligence from the cloud to the edge has become a necessity. Due to limited computational and communication capabilities, low memory and limited energy budget, bringing artificial intelligence algorithms to peripheral devices, such as end-nodes of a sensor network, is a challenging task and requires the design of innovative solutions. In this work, we present *PhiNets*, a new scalable backbone optimized for deep-learning-based image processing on resource-constrained platforms. *PhiNets* are based on inverted residual blocks specifically designed to decouple the computational cost, working memory, and parameter memory, thus exploiting all available resources for a given platform. With a YoloV2 detection head and Simple Online and Realtime Tracking, the proposed architecture achieves state-of-the-art results in (i) detection on the COCO and VOC2012 benchmarks, and (ii) tracking on the MOT15 benchmark. *PhiNets* obtain a reduction in parameter count of around 90% with respect to previous state-of-the-art models (EfficientNetv1, MobileNetv2) and achieve better performance with lower computational cost. Moreover, we demonstrate our approach on a prototype node based on an STM32H743 microcontroller (MCU) with 2MB of internal Flash and 1MB of RAM and achieve power requirements in the order of 10 mW. The code for the *PhiNets* is publicly available on GitHub¹.

Additional Key Words and Phrases: Multi-Object Tracking, Neural Networks, Edge AI, Tiny ML

ACM Reference Format:

Francesco Paissan, Alberto Ancilotto, and Elisabetta Farella. 2022. PhiNets: a scalable backbone for low-power AI at the edge. *ACM Trans. Embedd. Comput. Syst.* 1, 1, Article 1 (January 2022), 19 pages. <https://doi.org/10.1145/3510832>

1 INTRODUCTION

Over the past decade, we have witnessed two parallel trends. On one side, the increasing popularity of the internet of things, i.e., intelligent networked things everywhere, is a consequence of the growing capabilities of the embedded systems, enhanced with capable processing units working at always increasing frequencies and offering attractive low-power modes [11, 12, 27]. On the other side, with the advent of deep learning techniques, machine learning algorithms' size grows exponentially, thanks to the improvements in processor speeds and the availability of large training data. However, embedded systems cannot sustain the resource requirements of standard deep learning techniques, adequate for GP-GPUs [6, 14, 33].

How then to compose the need for exploiting the opportunity to bring intelligence at the edge with the complexity of deep learning? In this context, the junction point is TinyML [39, 41], a

*Both authors contributed equally to this research.

¹https://github.com/fpaissan/phinets_pl

Authors' address: Francesco Paissan, fpaissan@fbk.eu; Alberto Ancilotto, aancilotto@fbk.eu; Elisabetta Farella, efarella@fbk.eu, E3DA Unit, Digital Society Center - Fondazione Bruno Kessler (FBK), Via Sommarive, 18, Povo, Trento, Italy, 38123.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

1539-9087/2022/1-ART1 \$15.00

<https://doi.org/10.1145/3510832>

cutting-edge field that brings the machine learning (ML) transformative power to the performance- and power-constrained domain of tiny devices and embedded systems.

TinyML has been applied to several different classes of problems and devices, such as audio processing and sound event detection [6] [38], biosignals processing [21], gesture recognition [43] and general time series data [15]. Among the several application domains where to explore this novel trend, computer vision is one of the most popular. In this domain, object detection and tracking should be real-time, reliable, and accurate. Current best-performing pipelines for multi-object detection and tracking imply using many computational resources, thus substantially limiting the application scenarios in which such techniques can be exploited. The current limitations mainly depend on the high computational cost of state-of-the-art methods, which require GPUs for real-time inference and thus are not a good candidate for resource-constrained devices.

The most efficient solutions for Multi-Object Tracking (MOT) follow the tracking-by-detection paradigm. The tracking algorithm consists of an association algorithm based on the detected bounding boxes. Many pipelines are available for object detection. In particular, the main difference for what concerns object detection is the number of stages required to detect objects in one frame. The detection pipeline can be one-stage (single-shot bounding box regression) [2, 26] or two-stage (region proposal, object identification) [17, 31]. The one-stage detectors are the most efficient from the computational complexity perspective; thus, they are the go-to solution for lightweight, real-time object detection.

The most popular one-stage detectors are YOLO and SSD. The core of the detection pipeline is the convolutional backbone for latent space representation of the images, which accounts for most of the Multiply-Accumulate operations (MAC). After that, the detection head is applied to the output of the backbone and outputs the bounding box regression. Since the most expensive part of the pipeline is the backbone, many works have proposed scalable architectures to improve the efficiency of image processing [19, 37]. The ability of the networks to change computational requirements is an asset for embedded inferencing since the embedded platforms are diverse in hardware constraints. For example, typical IoT-oriented MCUs, such as the STM32F7 MCU or STM32L4, only have 320kB SRAM/1MB Flash and 32KB SRAM/256KB Flash, respectively, thus they cannot run the same networks. Also, developing a CNN architecture to tackle vision problems efficiently is not a new problem. Networks using minimal resources, such as MobileNets [19] and EfficientNets [36], have already been optimized to allow state-of-the-art performance with less than 1B Multiply and Accumulate operations (MACs), but with low performance on MCU scale computational resources.

Our work contributes to the state-of-the-art by proposing a novel scalable backbone, *PhiNets*, for detection and multi-object tracking on resource-constrained platforms. We prove the efficiency of *PhiNets* by comparing them with existing lightweight backbones within a YOLOv2 [30] detection head and Simple Online Real-time Tracking (SORT) tracker [1]. Furthermore, we will argue why our convolutional block is computationally cheaper with respect to current state-of-the-art solutions and how, given the architecture of *PhiNets*, we can do a one-shot search of the optimal parameters for every computational constraint set (i.e., every target platform). Moreover, we implemented the tracking inference on off-the-shelf MCUs with state-of-the-art consumption (1.3mJ per frame).

The novelty behind *PhiNets* is the convolutional block, which decouples the computational constraints and allows targeting specific platforms by optimizing the architecture hyper-parameters. In particular, our approach can be coupled with specific Neural Architecture Search (NAS) algorithms to boost the performance on specific tasks, without compromising the computational efficiency. Thus, our work has a meaningful impact in the fields of:

- embedded vision processing by proposing a new architecture family, *PhiNets*, which pushes forward the state-of-the-art in object detection on tiny devices;
- low-power image processing, since our pipeline requires only 1.3mJ per frame or 13mW at 10 fps;

2 RELATED WORKS

2.1 Scalable backbones

In this work, we present an MCU-friendly MOT pipeline based on deep learning. We propose an optimised backbone between the stages of a detection pipeline. The backbone is the sub-network in charge of compressing the input in a latent representation useful to localize, classify, detect, and track objects. We focused on backbone optimisation since it accounts for the most considerable computational cost in network inference. Thus, a reduction in the backbone complexity has a significant impact on the computational cost of the network. Moreover, by means of transfer learning, the same backbone can be applied to other input domains (e.g., image classification and sound event detection), as shown in [5] [35, 36]. Scalable backbones are neural networks which size can be varied by tuning different hyperparameters. These are an excellent solution for embedded processing since MCUs vary in resource availability, and diverse network architectures can be implemented based on the specific application domains. In the MobileNets paper, [19, 35], the authors presented a scalable backbone with good performance in vision benchmarks. In the work presented by Howard et al. [19], the architecture is based on depth-wise separable convolutions, and the authors present width and resolution multipliers. Both parameters are constrained in $[0, 1]$ (where 1 represents the standard architecture) and reduce parameter count and computational complexity quadratically. Instead, in the work of Sandler et al. [35] inverted residual structures and linear bottlenecks represent the building blocks of the architecture. Moreover, the same scaling model presented by Howard et al. [19] is applied. Despite the model's ability to scale, the authors did not study any particular scaling model and its relationships with the model's performance in different vision tasks. The first EfficientNet paper [36] proposed a Neural Architecture Search (NAS) based approach for vision tasks that is also capable of scaling in depth. Moreover, it proved that scaling the neural network architecture one dimension at a time is less efficient than scaling all three dimensions (resolution, width, depth) simultaneously. The well-described compound scaling model, which consists of scaling all three dimensions by a power of the initial coefficients, proved to give a state-of-the-art performance in neural network scaling. For EfficientNetV2 [37], the authors improved the model proposed by Tan et al. [36] with NAS to optimize the model's parameter efficiency. The work of Xiaoliang Dai et al. [7] presents a scaling framework for adapting networks to different hardware architectures efficiently, without relying on costly architecture search. It uses an accuracy predictor based on the Gaussian Process with Bayesian optimization instead of sampling the accuracy by training multiple networks. This work presents a general framework that can be used to adapt any network architecture to a target platform, achieving state-of-the-art performance for a given latency or energy budget. Although this scaling technique only optimizes a single metric (latency), it can be used with our proposed approach as the first step of the network adaptation procedure presented later. The main advantage of our approach over this technique is that, thanks to our convolutional block and network architecture scaling, we can disjointly optimize the required parameter memory and working memory. This is done by analytically computing the last two hyperparameters needed, as opposed to other approaches for porting neural networks to edge devices that only modify some parts of the network, limiting flexibility [16] [5]. Similarly, Han Cai et al. [4] propose a different approach for network adaptation, consisting in training a big network from which different sub-networks can be obtained by activating and de-activating

paths. As per ChamNet [7], this approach could be used together with our proposed methodology by formalizing the best sub-network fitting predefined latency requirements; then, by computing the t_0 and β hyperparameters to tailor the obtained architecture for the available flash and RAM of the target platform (as shown in Sec. 3.3).

2.2 Detection methods

Object detection is a field of computer vision dealing with the detection of semantic objects in images. Different techniques have been developed in the past decades and will be hereafter compared based on the hardware implementation feasibility and computational complexity. We can split the main detectors based on how many stages are required to extract the bounding boxes. The two-stage detectors, as for example Faster R-CNN [31] or Mask R-CNN [17], are based on region proposal techniques that are then processed (i) to generate a region of interest or (ii) to perform object classification and bounding box regression. On the other hand, single-stage detectors such as YOLO (You Only Look Once) [30] or SSD (Single Shot Multibox Detectors) [26] solve detection as a regression problem by learning to infer class probability and bounding box coordinates from input images. While two-stage detectors usually have higher accuracy scores with respect to single-stage detectors, they require more computational power and energy to infer a frame. Thus, to address MCU-friendly MOT applications, single-stage detectors are preferable.

2.3 Tracking methods

Many object association techniques can be implemented for MOT, ranging from Intersection over Union (IoU) comparison to unified detection and tracking techniques [42]. As it was for detection, there is a trade-off between computational complexity and performance, compared in terms of ID switches and MOTA score. There are two categories of tracking algorithms: algorithms perform tracking after a detection pipeline (e.g., SORT [1], DeepSORT [40], IoU) and the other category composed of algorithms that perform detection and tracking together (e.g., FairMOT [42] and FairMOT Lite).

	IDs	MOTA %	MOTP	Hz (fps)
IoU	287	61.6	0.116	5.33
SORT	48	54.3	0.172	5.31
DeepSORT	47	55.9	0.175	3.64
FairMOT	49	48.9	0.192	0.2
FairMOT Lite	60	46.9	0.196	3.85

Table 1. Results of trackers on a sequence from MOT significant to smart cities environment. IoU, SORT, DeepSORT, and FairMOT lite use YOLOv5S as object detector. The benchmarking is performed on a Intel(R) Core(TM) i9-10900KF CPU @ 3.70GHz

Simple Online and Realtime Tracking (SORT) exploits the Hungarian association algorithm to associate bounding boxes from consecutive frames by maximizing the IoU score. The same approach is performed by Deep SORT, in which the score to be minimized is the distance (e.g., Euclidean, cosine, correlation, etc.) between the latent representation of the content of the bounding boxes. This enables the algorithm to be more robust with respect to IoU because the visual attributes of the images are also taken into account.

FairMOT instead is based on CenterNet [8] and performs the detection and tracking together. This approach does not require anchors for detection. In fact, the bounding box shape is inferred

from the center of the blob. Some modification of this algorithm in which YOLOv5S is implemented in place of CenterNet is referred to as FairMOT Light here.

In Table 1, we quantitatively compared the presented trackers to help understanding trackers' performance and computational cost (expressed in terms of frames per second). A more detailed discussion on the trackers is performed in Section 3.4.

2.4 Vision-based MCU applications

Although tiny vision is an emerging technology, the main focus of the literature is on detection and classification tasks, thus a subset of the tracking pipeline. Some works explore both neural architectural optimisation and inference optimisation employing custom compilers and operations. In [24], an MCU-oriented NAS (TinyNAS) is combined with a lightweight inference engine (TinyEngine), enabling ImageNet-scale inference on MCU. Industry-oriented tools as STM X-Cube-AI can be exploited to implement artificial neural networks on MCUs [34], though compromising the inference performance with respect to manual network implementation via CMSIS-NN [22]. On the other end, some works [32] exploit extreme parameter reduction using XNOR Networks. However, the performance of those approaches, and in general of Binary Neural Networks (BNNs), is notably lower than the one achieved by classical CNNs; [3] thus, many application scenarios are not addressable with BNNs. Another way to implement vision intelligence at the edge is by exploiting custom hardware architectures, which use parallel computing to speed up computation [10, 13].

In this context, our work is towards the design of a novel scalable backbone, which maximises resource usage by decoupling the computational requirements of the neural network, allowing to design networks that take advantage of the different computational capabilities of different platforms. Since the proposed work applies at the architecture definition stage, i.e. before any hardware-specific implementation, it is suitable for various platforms as it does not rely on any specific custom hardware, compiler, or runtime.

In the following sections, we describe the proposed network architecture family and how it achieves good performance in object detection and tracking while being suitable for microcontroller-scale execution.

3 PHINETS ARCHITECTURE

When constraining computational cost and memory usage to fit neural networks on an MCU, scaling approaches like the ones presented in EfficientNet highlight the inefficiency of current state-of-the-art architectures. This is proved by the significant performance drop-offs on computer vision tasks when these networks are constrained to lower powered devices, as demonstrated by the authors in [19] and also confirmed empirically in Fig. 7.

In this work, we present *PhiNets*, an efficient neural network family developed for MCU inference. *PhiNets* aim at solving the main drawbacks of current state-of-the-art scalable backbones for image processing at the edge. They are a family of networks optimized for inference within the one to ten million MACs range, tuned to minimize the performance's drop while scaling with the architecture specifications.

In the following subsections, we will introduce the main network building blocks (3.1), the main constraints when it comes to MCU inference (3.2) and how *PhiNets* solve this problem (3.3). In the end, we will present the detection and tracking pipelines selected for the benchmarking (3.4).

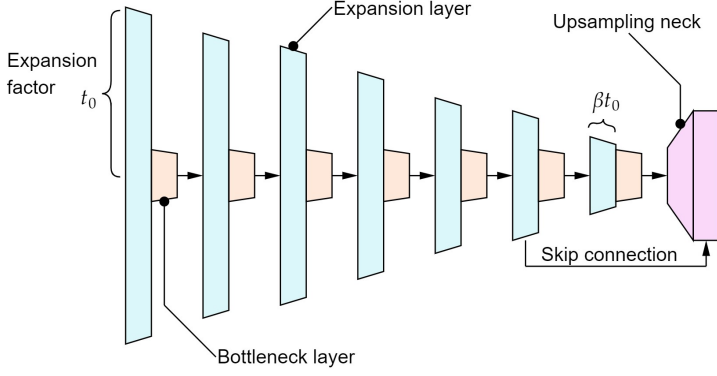


Fig. 1. An overview of the *PhiNets* family network architecture. The proposed architecture scales with respect to the expansion factor t_0 , number of convolutional blocks, shape factor β and width multiplier as explained in Sec 3.3.

3.1 Network building blocks

As it is common in parameter-efficient architectures, such as [18, 35, 37], the network is composed of a sequence of inverted residual blocks (by default, there are seven blocks for our architecture), each followed by a swish activation function. Fig. 1 shows the network architecture overview.

As illustrated in Fig. 2, the number of filters in the first bottleneck layer is 24α , where α is a hyperparameter that works similarly to how it does in the MobilenetV2 architecture, while the multiplication factor gets doubled every time the feature map is downsampled in the network. Squeeze-and-Excitation blocks [20] are inserted after each convolutional block and skip connections are used between the bottleneck layers that have the same feature map resolution, where no downsampling of the feature map takes place, as in MobileNetV2 [35]. The expansion factor used in the inverted residual block of index N , where N equals 0 for the first layer, is determined by two hyperparameters, i.e. t_0 (*base expansion factor*) and β (*shape hyperparameter*), as

$$t = t_0 \left(\frac{(\beta - 1)N + B}{B} \right) \quad (1)$$

Five strided convolutions are used for down-sampling the feature maps through the network with a spatial resolution reduction of a factor of $32\times$ between input and output tensors.

As the network has been primarily developed for object detection tasks, we maximized the receptive field of each element in the output grid, also considering the resolution of the output tensor. Reducing the latter too much would affect the spatial information flow through the network and significantly lower the performance [2]. To tackle this issue, we placed a neck composed of a $2\times$ up-sampling layer and a skip connection to the latest convolutional block of the exact resolution after the sequence of convolutional blocks. This allows the proposed architecture to outperform models like MobileNet and EfficientNet when performing object detection tasks at low input resolutions. While MobileNets and EfficientNets need to compromise between output tensor resolution (e.g., a 128×128 input would be represented as only 4×4) and receptive field, *PhiNets* allow for higher receptive fields while maintaining fine granularity in the output grid, increasing detector performance.

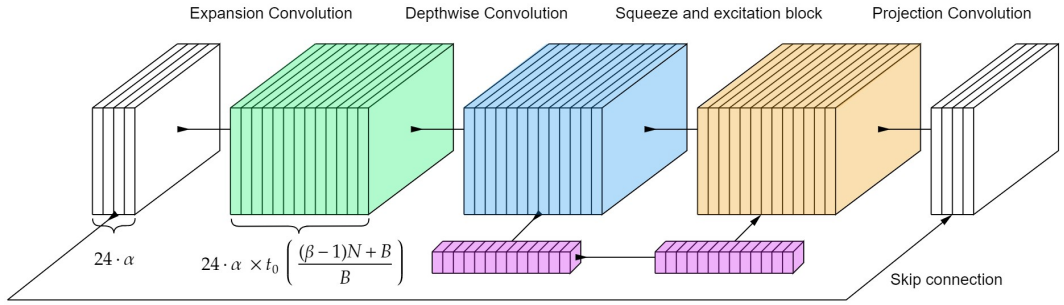


Fig. 2. An overview of the *PhiNets* convolutional block structure. First, the number of channels is increased with a pointwise convolution, followed by a depthwise convolution and SE block. Finally, a second pointwise convolution connects to the low dimensionality bottleneck block.

3.2 Hardware-constrained and hardware-aware scaling

When bringing CNNs to edge devices, the general approach consists of (i) architectural space exploration to identify the best performing networks for a given task (ii) selection of the best network fitting in the most restrictive resource constraints. This approach is known as hardware-constrained network architecture search and is the most often used technique for architecture search in constrained devices. While this helps optimise a network for a very resource-limited device such as a microcontroller, this approach results in a network that provides sub-optimal resource utilization of the platform, as only the most stringent requirement is met. But, when working with a resource-constrained platform, we usually need to face three different constraints:

- the number of operations (MAC) required for network inference. High-performance microcontrollers, coupled with efficient inferencing frameworks, can achieve tens to hundreds of MMAC per second. For real-time video applications, this means that a top-of-the-line MCU can run a 10MMAC detection network at $\approx 10\text{Hz}$;
- the dynamic memory (working memory - WM). When executing a network's computational graph, at each point, the CPU must compute a matrix multiplication between the output of the previous layer or the input of the network and the following channel filter matrix. The RAM usage is determined by the size of these tensors, plus the memory required to keep the tensors for skip connections for later usage;
- the static memory (parameter memory - PM). As FLASH memory is the most expensive part of the microcontroller's die both in terms of cost and area, this can usually contain 100KB to 1MB of data. Assuming that all network parameters get quantized to 8bit integers, a maximum of 100K to 1M parameters can be used, based on the selected platform.

Our work proposes an optimized architecture family, *PhiNets*, which inverts the hardware network architecture search paradigm, using a *hardware-aware* network scaling pipeline. Resource constraints can be met with minimal performance loss by varying different sets of hyperparameters. Moreover, eventual performance bottlenecks can be easily identified thanks to the structure of the network.

3.3 Meeting the requirements: decoupling MAC and memory usage efficiently

Resource usage for the three main hardware constraints of the network can be optimized in a decoupled way, i.e., using different hyperparameter combinations to meet different resource constraints and achieve optimal use of the available hardware. This optimization will allow for superior

performance with respect to networks generated using hardware-constrained scaling techniques, as network architectures are not designed considering the strictest hardware requirements, but instead tailored to a specific platform. The following sections will highlight how the network parameters are connected to the resource requirements for *PhiNets*.

3.3.1 Number of operations. The number of operations (MAC) for the baseline network depends on network input resolution $w \times h$, on the network width, which scales quadratically with the parameter α , and on the network depth (determined by the number of blocks B). These parameters can be tuned using a compound scaling methodology as proposed in [37] to obtain the best performing network for a given operation count or can be determined by other implementation factors (e.g. the resolution of the camera used can set a fixed $w \times h$).

Parameters $w \times h$, α and B determine the number of operations for the base network. This can be defined by real-time requirements for the system, power consumption targets, or accuracy requirements. Fig. 3 shows the effects of the three parameters on the complexity of the networks. Different networks have been generated varying the $w \times h$, α and B hyperparameters, and the Tensorflow profiler has been used in order to obtain the number of operations for each network.

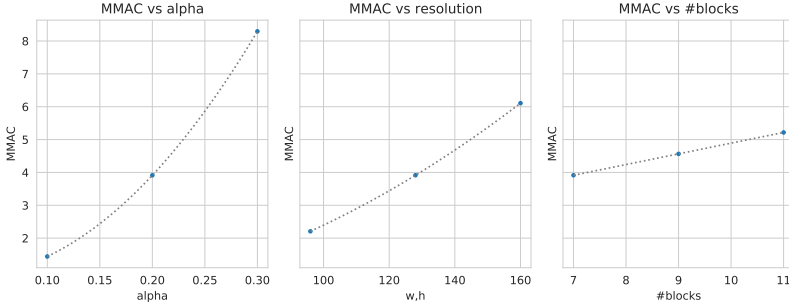


Fig. 3. *PhiNets* computational complexity with respect to α , $w \times h$ and B . The plots highlight an exponential increase in MAC count with the number of filters and input resolution, while a linear trend with respect to the number of convolutional blocks.

3.3.2 Dynamic memory. The dynamic memory will be determined by the size of the tensors in the expansion layers of the first convolutional block. The tensor size in the later layers increases linearly with the network's depth and decreases quadratically with resolution. Moreover, no tensors need to be kept in memory for the first block as there are no previous layers with residual connections. Varying the base expansion factor t_0 scales the dynamic memory required by the network linearly. Note that it is recommended to keep this parameter between 2 and 8, using by default 6 for networks larger than 5MMAC and 5 for networks smaller than that. Fig. 4 (left) shows the linear effects of the expansion factor on the working memory requirements of the network. Different networks have been generated varying the t_0 hyperparameter, and STMicroelectronics-Cube-Ai has been used to obtain the working memory requirements for each network.

3.3.3 Parameter memory. The parameter memory is determined by the convolutional kernels in the network. In particular, it is determined by the size of the kernels used in the expansion layers of the later network blocks, where the tensors usually have a low spatial resolution, but a high number of filters is used. Given how the expansion factor of later layers is related to β , the parameter memory required by the network varies with this parameter. In particular, the relationship between the number of parameters in the network and the shape hyperparameter is

modelled as $\#Params \approx C \cdot \frac{1}{2}(1 + \beta)$ with C number of parameters of the base network ($\beta = 1$). Fig. 4 (right) shows the effects of the shape factor on the number of parameters of the network.

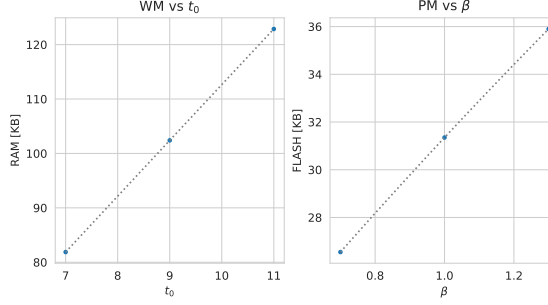


Fig. 4. Left: *PhiNets* RAM requirements from a default network by varying t_0 . Right: *PhiNets* FLASH requirements from a default network by varying β . As shown, the trend is linear in both cases.

3.3.4 Network tuning strategy. While tuning the network, it is recommended to optimize the number of operations of the base network firstly, then RAM and FLASH usage. In particular, a general network-tuning approach can be as follows:

- (1) Estimation of the available time for an inference operation and estimate of the corresponding MAC count given the platform performance;
- (2) Selection of the hyperparameters $w \times h$, α and B , based either on a compound scaling technique, from a default architecture, or with a network architecture search algorithm (such as the scaling presented in ChamNet [7] or Once For All [4]), to achieve the correct number of operations;
- (3) Tuning of the t_0 hyperparameter knowing input resolution and available WM from the size of input and output tensors for the first convolutional block (assuming network weights and activations get quantized to 8bit integers)

$$WM \approx \left(\frac{w}{2} \times \frac{h}{2} \cdot 24\alpha + \frac{w}{4} \times \frac{h}{4} \cdot 24\alpha \right) t_0$$

For example, going from $t_0 = 5$ to $t_0 = 4$ decreases the needed RAM for the network by 20%;

- (4) Tuning of the β hyperparameter to achieve the desired number of parameters. Knowing the number of parameters of the starting architecture P_0 and the target P_{target} , we can obtain β from:

$$\beta \approx 2 \frac{P_{target}}{P_0} - 1$$

For example, going from $\beta = 1$ to $\beta = 0.6$ decreases the number of parameters by 20%;

The optimization procedure can be repeated multiple times if higher precision is required, as varying t_0 and β has second-order effects on the number of operations.

3.4 Detection and tracking

Nowadays, there are many alternatives for both detection [2, 17, 26] and tracking [1, 40, 42]. We took into account the algorithms considering the trade-off between computational cost and performance. We used the YOLOv2 [30] detection head working at a single scale for the object detection task, as this requires only a single convolutional layer for bounding box prediction and class identification of all objects in the frame, leading to minimal operation count networks. Choosing YoloV2 also

helps in embedded processing since the computational complexity is affected mainly by the *PhiNet* architecture and does not directly depend on the number of objects in the frame.

For the tracker, we used a tracking-by-detection pipeline based on the proposed object detector. Different alternatives like SORT, DeepSORT, and IoU association can be considered to work with our architecture. They all have a computational cost that scales linearly with the number of objects to be tracked, thus implying a limit in the number of possibly tracked objects in resource-constrained platforms targeting real-time applications. Moreover, while IoU is the shallowest concerning the operation count, it has many ID switches, thus performing worse than SORT and DeepSORT, considering this critical metric in low-fps applications, where an object might be in the scene a total of 10-20 frames. Between DeepSORT and SORT, we selected SORT since having the embedding extractor as in the DeepSORT architecture implies using a lower complexity object detector. In tracking-by-detection pipelines, it is crucial to have good detection performance since it directly impacts tracking performance. Thus, we are interested in having the best performing detector possible since it will help the shallower tracker achieve the same MOTA score as the more complex one, working with a worse detector.

4 RESULTS

Since, as already described, we used a tracking-by-detection pipeline, and the tracking performance is highly dependent on the detections, we decided to split the benchmarking in detection performance, tracking performance, and power consumption to show how each component of our pipeline was performing.

4.1 Baseline architectures

PhiNets baseline architectures were tested, with the parameters summarized in the Table 2.

Resolution	α	B	β	t_0	MAC	Parameters	Task
128×128	0.35	7	1	6	9.85 M	61.2 K	Detection
128×128	0.25	7	1	6	6.08 M	37.9 K	Detection
96×96	0.25	7	1	5	3.01 M	31.8 K	Detection
96×96	0.15	7	1	5	1.23 M	14.3 K	Detection
160×160	0.3	7	1	5	10.42 M	39.9 K	Tracking
160×160	0.2	7	1	5	4.96 M	21.6 K	Tracking
128×128	0.2	7	1	5	3.18 M	21.6 K	Tracking

Table 2. Parameters for generating the benchmarked *PhiNets*

Networks have been grouped by task, as we found that a slightly lower resolution but a higher number of filters benefited, at the same MAC count, more the detection task than the tracking one, for which a higher resolution proved to be the better choice.

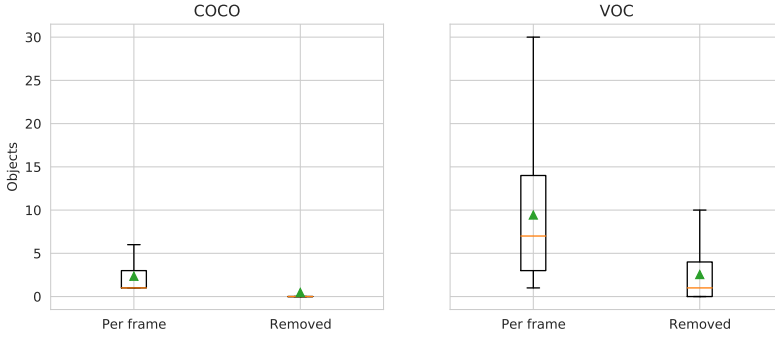


Fig. 5. Number of objects removed per frame from COCO and VOC2012 datasets based on area constraint. The removal prevents having objects which are represented by only a couple of pixels on the down-sampled image (input of the backbone).

4.2 Detection

To evaluate object detection performance towards tiny multi-object tracking, we trained EfficientNets, MobileNets and *PhiNets* sized between 1M and 10M MAC on a subset of the MS COCO [25] and VOC2012 [9] object detection benchmarks.

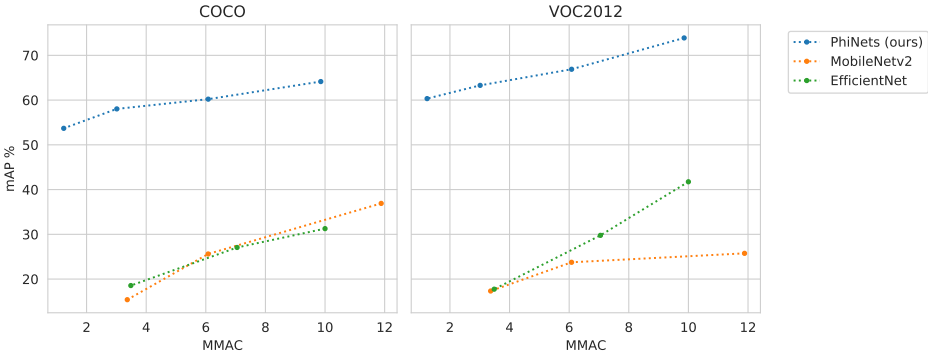


Fig. 6. Comparison of scalable backbones applied before a YOLOv2 detection head for MCU-scale object detection considered in terms of mean Average Precision (mAP) vs MAC count. The best fitting and performing models for MCU inference are the ones in the top-left area of the plot (requiring less operations with better performance).

Given the application constraints of tiny vision, mainly regarding the input resolution, we reduced the training set by considering only the "person" class and by using only targets whose bounding box was more extensive in area than $1/64$ of the original image size. This pre-processing of the dataset is visualised in Fig. 5, where we investigated the number of objects per frame in both datasets and the number of removed objects based on the above constraints.

The networks were trained for 60 (COCO) / 120 (VOC) epochs, using the Adam optimizer. The first three epochs for both datasets are used for warm-up, with a 5×10^{-3} learning rate, while for the remaining epochs, we used cosine decaying on the learning rate, starting from 1×10^{-2} .

As shown in Fig. 6, all the *PhiNets* perform better in object detection on the 1-10 MMAC range, given their advanced scalability features. In fact, *PhiNet*'s performance is almost constant (within the same benchmarking dataset) with respect to the number of MMAC, while EfficientNets and MobileNets have a performance drop greater than 15% mAP in the depicted MAC range.

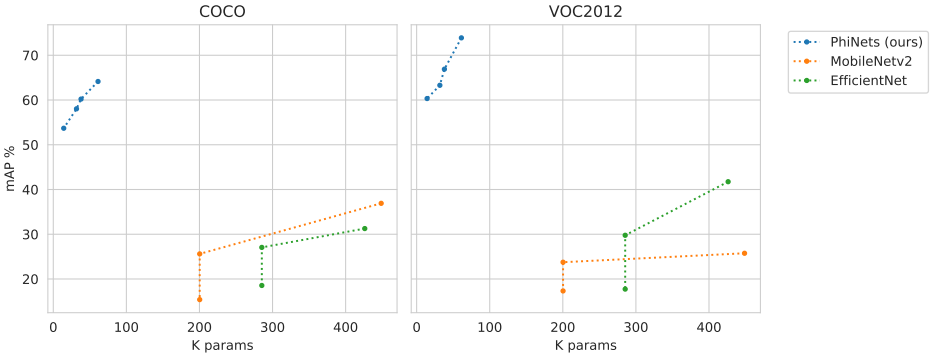


Fig. 7. Comparison of scalable backbones applied before a YOLOv2 detection head for MCU-scale object detection. The plots show a vertical section when we downsized the network by lowering the input resolution, keeping the same number of filters (and thus parameters), in order to avoid drastic performance drop in EfficientNets and MobileNets. The best fitting and performing models for MCU inference are the ones in the top-left area of the plot (requiring less parameters with better performance).

We achieved the same behaviour also in the analysis with respect to the number of parameters, depicted in Fig. 7, proving that *PhiNets* are more efficient in parameter count than previous state-of-the-art architectures. In conclusion, *PhiNets* set a new standard for embedded object detection on MCUs by achieving higher performance with less operations and parameters.

4.3 Multi-Object Tracking

After the detection, we perform multi-object tracking using SORT [1]. We benchmarked the proposed backbones performance on the MOT15 dataset [23] after augmenting the training of the detectors with 360 epochs on the benchmark data.

The relationship between detection and tracking performance in tracking-by-detection pipelines is intuitively linear. In fact, since the tracking is based on the detection IoU, a bad detector implies low tracking performance. Thus, as expected by analysing the detection results, and as we empirically proved in Fig. 8, *PhiNets* are the best performing backbone for tracking in the 1-10MMAC range.

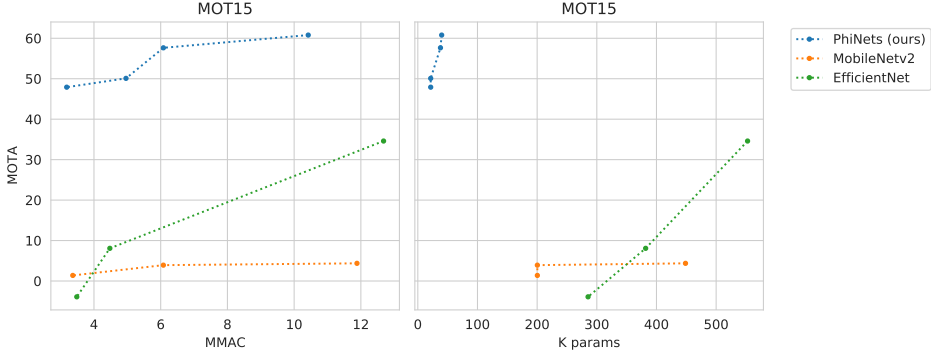


Fig. 8. Results for MOT15 tracking showing the relation between Multi Object Tracking Accuracy (MOTA) score and network complexity (in terms of MAC count - left - and parameters count - right). The plots show a vertical section when we downsized the network by lowering the input resolution, keeping the same number of filters (and thus parameters), in order to avoid drastic performance drop in EfficientNets and MobileNets. The best fitting and performing models for MCU inference are the ones in the top-left area of the plot (requiring less parameters and operations with better performance).

4.4 Power consumption

In order to test the power consumption of the network on a off-the-shelf MCU, a prototype hardware board (shown in Figure 9) was developed, based around an STM32H743 microcontroller, with 2MB of internal Flash and 1MB of RAM. The proposed architecture is, however, not limited or targeted towards any particular hardware platform, rather it is developed to be scalable and portable to a variety of MCUs. The hardware prototype has been developed in order to interface with a prototype low power vision sensor [28] [29], but the architecture can be used with most commercial cameras depending on the use case. Since the specific camera used in our prototype and the other devices on the endnode (e.g. the RF transceiver) are not the target of our approach, in this section we will analyze only the power consumption of the microcontroller.

The MCU was powered at $V_{dd} = 1.8V$ from a switch-mode power supply, and the internal LDO powering the core was set to output 1.1V. The microcontroller was run at a constant frequency of 300MHz, as this allowed the best efficiency in terms of energy requirements per network inference run. This can be seen from the graphs in Fig. 10, where we analyze the effects of different running frequencies and MCU core voltages on the required current (and, consequently, energy consumption). Tests were run using ST's proprietary STMicroelectronics-Cube-Ai runtime, and code was compiled using arm-eabi-none-gcc with -O2 optimization level.

The energy required by the MCU for an inference pass was estimated by sampling the current through a shunt resistor at the input of the MCU power bus. Empirically, we sampled the current $I(\tau)$ every $t_s = 10\mu s$ for the duration of a single inference pass (which, depending on the number of operations of the model, takes from 15ms to 120ms at the chosen clock frequency) and computed the energy required for inference using the relationship $E = V_{dd} \sum_{\tau} I(\tau) t_s$.

We investigated the relationship between computational complexity and energy required per inference in Fig. 11. As we can see, energy (and average current) required by the MCU for an inference pass is, in a first approximation, linearly dependent on the number of operations of the network, with a coefficient of around 1.3 mJ/MMAC. This allows us to estimate the possible working points of the microcontroller, by using this data to approximate the energy per frame required for all networks in the range of 1-10 MMAC. We suppose to run the network at a fixed

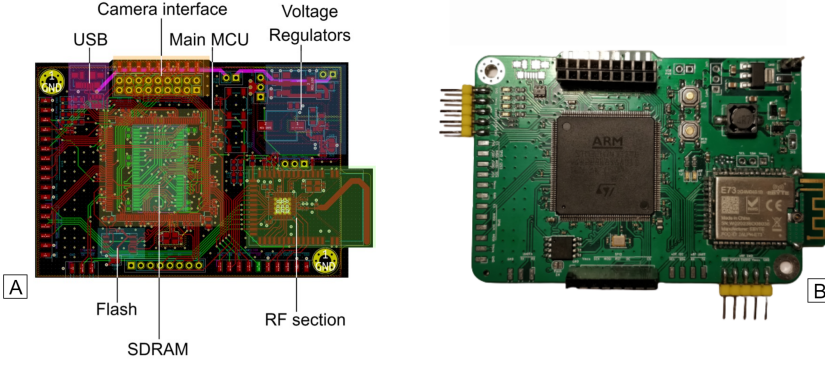


Fig. 9. The system hardware prototype realized. The board includes an STM32H743 MCU, switch mode power regulator, a DCMI camera interface and display connection, external FLASH and DRAM and a bluetooth section (FLASH, DRAM and BLE are not needed for this application). A: Layout of board and functional blocks. B: Final prototype picture, with soldered components

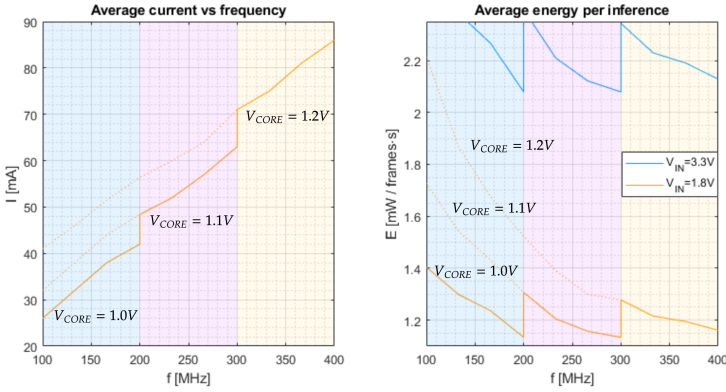


Fig. 10. Trade-offs between core voltage, running frequency and power consumption for the target MCU. Left: current absorbed by the MCU at different running speeds when computing inference. Right: estimated energy per inference with the reference 1.23 MMAC network. The platform shows the minimum energy consumption running at 200 or 300MHz when powered at 1.8V. It is possible to use a higher core voltage at lower running frequencies (dotted line), but this only increases the MCU power consumption without bringing additional benefits. Blue, red and yellow areas show the optimal (most efficient) frequency ranges for different core voltages.

number of iterations per second, then having the MCU enter a low power mode (for which current consumption is approximated using data from the datasheet) for the remaining time before the next frame. For example, we can run the 6.1MMAC network at 100% duty cycle, achieving around 14fps while using 162mW, for a fast and accurate detector. Suppose very low power consumption is the target. In that case, instead, the 1.23MMAC network can run at 10fps by waking up the MCU 10 times per second, performing the computation (which takes approximately 15ms), and sleeping for the remaining 85ms, with a total power consumption of 13mW. The relationship between power consumption, inference speed and accuracy is plotted in Fig. 13.

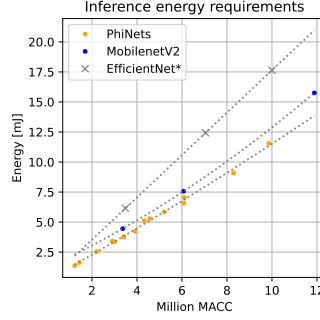


Fig. 11. Linear relationship between MAC and energy requirements for the realized hardware prototype. Energy was measured with the MCU running at 300MHz and powered at 1.8V . As energy required is, in a first approximation, linearly dependent on network complexity (thus execution time), we can interpolate data points for different networks to obtain an estimation for all possible working points of the platform (Fig. 13). As expected, the relationship between energy per frame and number of operations for our model and the MobileNetV2 architecture are similar, as both networks are based on similar convolutional blocks. The performance of our model is, as demonstrated before, much higher than competing approaches for an equal number of operations. EfficientNet values are only estimates extrapolated from smaller models, as the high parameter count of the architecture made it impossible to run on our platform.

This hardware - software combination allowed for a state of the art power consumption of under 1.3mJ for the 1.2MMAC *PhiNet* ($53.7 / 60.3$ mAP on a subset of the COCO/VOC2012 datasets) and 11.8mJ for the 9.8MMAC *PhiNet* ($64.1 / 73.9$ mAP on COCO/VOC2012, 60.8 MOTA on MOT15).

Figure 13 shows the working points that can be reached with the proposed hardware and software with respect to performance, power consumption, and inference speed. The platform is capable of running object detection at over 50fps with the proposed hardware, at a power consumption from $1.3\text{mW}/\text{fps}$ (for networks achieving $53.7 / 60.3$ mAP on the selected subsets of COCO/VOC datasets) to $11.8\text{mW}/\text{fps}$ ($64.1 / 73.9$ mAP on COCO/VOC), or, in other words, 10fps tracking at 13mW to 118mW , depending on the performance required by the specific application.

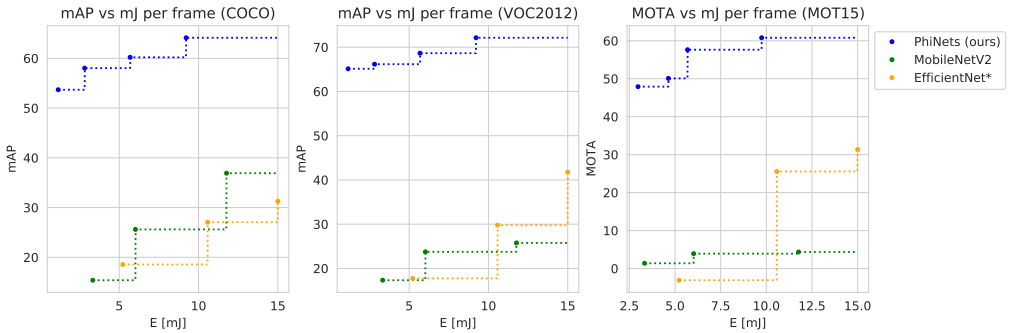


Fig. 12. Comparison between the pareto curve of the performance over energy per frame curve for our approach, MobileNetV2 and EfficientNet. Energy per frame values are obtained as described before, with the MCU running at 300MHz and powered at 1.8V . Values for EfficientNet are extrapolated from smaller networks, as the networks shown required more parameter memory than what is available on our prototype.

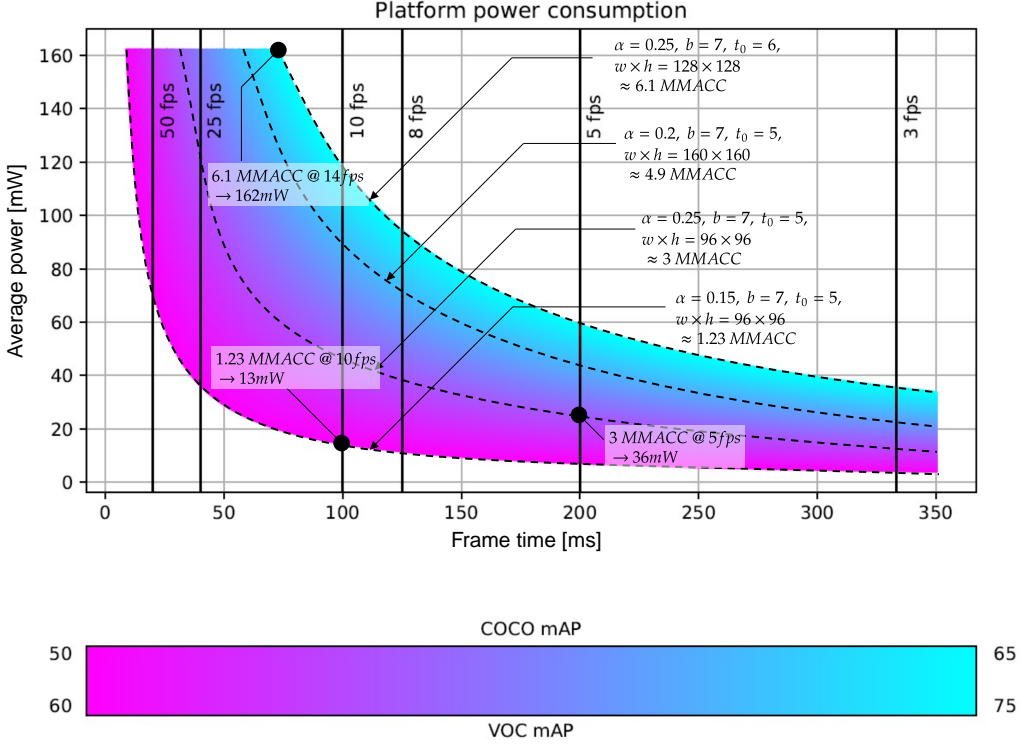


Fig. 13. Possible working points of the platform, with respect to latency, performance and power consumption. Data has been recorded running the platform at 1.8V, at a frequency of 300MHz. Networks with higher precision are in cyan, networks with higher performance and lower power requirements in purple. Smaller networks can achieve either very low power consumption figures, or very high speeds, with, in specific fields, small accuracy losses with respect to larger ones. A usage example will be presented in section 4.5

4.5 Case study: monitoring application

A concise case study is here presented considering a scenario akin to the ones shown in the sequences TownCentre (top-left in Fig 14) and PETS09 (bottom-left in Fig 14), where a camera is mounted on an elevated position, overseeing a walkway.

Since the camera is mounted on an elevated point, there are no significant variations in the size and pose of the targets, thus allowing good tracking performance using smaller networks, which require lower power consumption. For example, the resulting MOTA score on PETS09 is 62.86 for the smallest *PhiNet* and 68.51 for the largest one tested. Given the large area seen by the camera, we can use low frame rates since targets will require multiple seconds to cross from one side of the image to the other. We can estimate the required energy usage for always-on tracking from Fig 13.

Given the target experimental results, we can use the smallest tracking network presented in Table 2, at 3.18MMAC complexity. Fig. 11 shows the measured energy consumption per inference at 3.21mJ, or 16mW at the chosen framerate. A similar result can be extrapolated from the plot in Fig. 13, knowing the target frame rates and noting that the network we are using is on the left side of the mAP bar. In particular, knowing the target mAP of the network (we are in the purple area of the mAP bar), we know that, on the top plot, we are moving on the dashed line corresponding to the 3 MMAC network. On this line, we intercept the vertical bar for 3fps operation



Fig. 14. Visual example of detection and tracking pipeline's output.

at a power consumption near $15mW$. Such low energy requirements are ideal for always-on IoT nodes, allowing for long lifetime of the devices' batteries thanks to lower energy absorption.

5 CONCLUSION

In this paper we presented a new scalable backbone for low computational complexity image processing based on inverted residual blocks. The backbone was benchmarked on detection and tracking tasks with state-of-the-art results. We achieved 20% higher mAP with respect to EfficientNets and MobileNets on the COCO and VOC2012 detection benchmarks for the same computational complexity and 80% less parameters. Moreover, we achieved significantly higher MOTA score with the same compression factor. The main asset of *PhiNets* is that our model outperforms the hardware-constrained scaling by disjointly optimizing MAC, working memory and parameter memory as per the hardware-aware architecture scaling paradigm.

In conclusion, we proved that our approach successfully runs on a IoT endnode with a power consumption that allows the node to be powered using a solar panel.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 957337 (MARVEL project). This paper reflects only the authors' views and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- [1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.
- [2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).
- [3] Adrian Bulat and Georgios Tzimiropoulos. 2019. Xnor-net++: Improved binary neural networks. *arXiv preprint arXiv:1909.13863* (2019).
- [4] Han Cai, Chuang Gan, and Song Han. 2019. Once for All: Train One Network and Specialize it for Efficient Deployment. *CoRR abs/1908.09791* (2019).
- [5] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. 2020. Tiny Transfer Learning: Towards Memory-Efficient On-Device Learning. *CoRR abs/2007.11622* (2020).
- [6] Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, and Elisabetta Farella. 2019. Neural Network Distillation on IoT Platforms for Sound Event Detection. In *Proc. Interspeech 2019*. 3609–3613. <https://doi.org/10.21437/Interspeech.2019-2394>

- [7] Xiaoliang Dai, Peizhao Zhang, Bichen Wu, Hongxu Yin, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu, Yiming Wu, Yangqing Jia, Peter Vajda, Matt Uyttendaele, and Niraj K. Jha. 2018. ChamNet: Towards Efficient Network Design through Platform-Aware Model Adaptation. *CoRR* abs/1812.08934 (2018).
- [8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6569–6578.
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n. d.]. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [10] Eric Flamand, Davide Rossi, Francesco Conti, Igor Loi, Antonio Pullini, Florent Rotenberg, and Luca Benini. 2018. GAP-8: A RISC-V SoC for AI at the Edge of the IoT. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 1–4.
- [11] Raspberry Pi Foundation. [n. d.]. Raspberry Pi hardware. <https://www.raspberrypi.org/documentation/hardware/raspberrypi/>
- [12] FriendlyARM. [n. d.]. NanoPi NEO-LTS. <https://www.friendlyarm.com>
- [13] Angelo Garofalo, Manuele Rusci, Francesco Conti, Davide Rossi, and Luca Benini. 2020. PULP-NN: accelerating quantized neural networks on parallel ultra-low-power RISC-V processors. *Philosophical Transactions of the Royal Society A* 378, 2164 (2020), 20190155.
- [14] Cong Hao, Jordan Dotzel, Jinjun Xiong, Luca Benini, Zhiru Zhang, and Deming Chen. 2021. Enabling Design Methodologies and Future Trends for Edge AI: Specialization and Co-design. *IEEE Design & Test* (2021).
- [15] Shayan Hassantabar, Prerit Terway, and Niraj K. Jha. 2020. TUTOR: Training Neural Networks Using Decision Rules as Model Priors. *CoRR* abs/2010.05429 (2020).
- [16] Shayan Hassantabar, Zeyu Wang, and Niraj K. Jha. 2019. SCANN: Synthesis of Compact and Accurate Neural Networks. *CoRR* abs/1904.09090 (2019).
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
- [20] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [21] Arlene John, Barry Cardiff, and Deepu John. 2021. A 1D-CNN Based Deep Learning Technique for Sleep Apnea Detection in IoT Sensors. *CoRR* abs/2105.00528 (2021).
- [22] Liangzhen Lai, Naveen Suda, and Vikas Chandra. 2018. Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus. *arXiv preprint arXiv:1801.06601* (2018).
- [23] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942* (2015).
- [24] Ji Lin, Wei-Ming Chen, Yujun Lin, John Cohn, Chuang Gan, and Song Han. 2020. Mcunet: Tiny deep learning on iot devices. *arXiv preprint arXiv:2007.10319* (2020).
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.
- [27] Arm Ltd. [n. d.]. Cortex-M. <https://developer.arm.com/ip-products/processors/cortex-m>
- [28] Francesco Paissan, Gianmarco Cerutti, Massimo Gottardi, and Elisabetta Farella. 2019. People/car classification using an ultra-low-power smart vision sensor. In *2019 IEEE 8th International Workshop on Advances in Sensors and Interfaces (IWASI)*. IEEE, 91–96.
- [29] Francesco Paissan, Massimo Gottardi, and Elisabetta Farella. 2021. Enabling energy efficient machine learning on a Ultra-Low-Power vision sensor for IoT. *arXiv preprint arXiv:2102.01340* (2021).
- [30] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.
- [31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).
- [32] Wojciech Romaszkan, Tianmu Li, and Puneet Gupta. 2020. 3PXNet: Pruned-Permuted-Packed XNOR Networks for Edge Machine Learning. *ACM Trans. Embed. Comput. Syst.* 19, 1, Article 5 (Feb. 2020), 23 pages. <https://doi.org/10.1145/3371157>

- [33] Manuele Rusci, Davide Rossi, Eric Flamand, Massimo Gottardi, Elisabetta Farella, and Luca Benini. 2018. Always-ON Visual Node with a Hardware-Software Event-Based Binarized Neural Network Inference Engine. In *Proceedings of the 15th ACM International Conference on Computing Frontiers (CF '18)*. Association for Computing Machinery, New York, NY, USA, 314–319. <https://doi.org/10.1145/3203217.3204463>
- [34] Fouad Sakr, Francesco Bellotti, Riccardo Berta, and Alessandro De Gloria. 2020. Machine Learning on Mainstream Microcontrollers. *Sensors* 20, 9 (2020). <https://doi.org/10.3390/s20092638>
- [35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.
- [36] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.
- [37] Mingxing Tan and Quoc V Le. 2021. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298* (2021).
- [38] Ganesh Venkatesh, Alagappan Valliappan, Jay Mahadeokar, Yuan Shangguan, Christian Fuegen, Michael L. Seltzer, and Vikas Chandra. 2021. Memory-efficient Speech Recognition on Smart Devices. *CoRR* abs/2102.11531 (2021).
- [39] Xiaofei Wang, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. 2020. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 22, 2 (2020), 869–904.
- [40] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.
- [41] Dianlei Xu, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. 2020. Edge Intelligence: Architectures, Challenges, and Applications. *arXiv preprint arXiv:2003.12172* (2020).
- [42] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2020. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *arXiv preprint arXiv:2004.01888* (2020).
- [43] Andy Zhou, Rikky Muller, and Jan M. Rabaey. 2021. Memory-Efficient, Limb Position-Aware Hand Gesture Recognition using Hyperdimensional Computing. *CoRR* abs/2103.05267 (2021).