BLANKSKIP: Early-exit Object Detection onboard Nano-drones

Carlo Marra[†*], Beatrice Alessandra Motetti[†], Alessio Burrello[†], Enrico Macii[†], Massimo Poncino[†], Daniele Jahier Pagliari[†]

[†]Politecnico di Torino, Turin, Italy

[*]Corresponding author, email: carlo.marra@polito.it

Abstract

Deploying tiny computer vision Deep Neural Networks (DNNs) on-board nano-sized drones is key for achieving autonomy, but is complicated by the extremely tight constrained of their computational platforms ($\approx 10\,\mathrm{MiB}$ memory, $\approx 1\,\mathrm{W}$ power budget). Early-exit adaptive DNNs that dial down the computational effort for "easy-to-process" input frames represent a promising way to reduce the average inference latency. However, while this approach is extensively studied for classification, its application to dense tasks like object detection (OD) is not straight-forward. In this paper, we propose BLANKSKIP*, an adaptive network for on-device OD that leverages a simple auxiliary classification task for early-exit, i.e., identifying frames with no objects of interest. With experiments using a real-world nano-drone platform, the Bitcraze Crazyflie 2.1, we achieve up to 24% average throughput improvement with a limited 0.015 mean Average Precision (mAP) drop compared to a static MobileNet-SSD detector, on a state-of-the-art nano-drones OD dataset.*

1. Introduction

Nano-drones are small Unmanned Aerial Vehicles (UAVs) that enable unique applications such as indoor exploration, infrastructure inspection, and search-and-rescue operations, where their agility and small form factor (diameter smaller than 10 cm) open scenarios inaccessible to larger systems [22]. Operating autonomously in these settings requires onboard perception: for instance the drone must be able to detect and localize objects in real time, without relying on remote computation offloading, as the latency and communication overhead introduced would be incompatible with fast, reactive flight. However, the hardware aboard nano-drones is severely constrained by strict power and weight budgets. Platforms such as the Bitcraze Crazyflie 2.1 [3], equipped with the GAP8 System-on-Chip (SoC) Artificial Intelligence (AI) deck [7], offer only a few hundred MHz of clock frequency, less than 10 MiB of total onboard RAM, and operate under a total power envelope of a few hundred milliwatts for the compute part [17]. Under these constraints, even compact Deep Neural Network (DNN) architectures designed for embedded devices deliver low frame rates, that leave little margin for the multitasking required by autonomous flight (e.g., simultaneous navigation, obstacle avoidance, and target detection). Existing approaches to onboard Object Detection (OD), in particular, have focused on compressing DNNs through quantization, pruning, and lightweight architectural design [11, 26, 29], yielding models that fit within the memory and compute budget of embedded processors. Yet these solutions share a fundamental limitation: they apply a fixed computational graph to every input frame, regardless of its content. This is particularly wasteful in nano-drone exploration scenarios, where a large fraction of frames may contain little to no information. Adaptive inference methods [27, 31] address this inefficiency by adjusting compute at runtime in an input-dependent way. Among them, Early Exit (EE) DNNs have shown promising results in reducing average latency for image classification [24]. However, their application to OD, and to severely constrained platforms such as nano-drones, remains largely unexplored.

In this work, we propose BLANKSKIP, an EE mechanism for OD on nano-drones that exploits the sparsity of object presence across frames. Rather than applying the full detection pipeline to every input, we decompose the problem into two sequential sub-tasks: a lightweight binary classifier, attached at an intermediate layer of the backbone, first determines whether a frame contains any object of interest; only if so, inference continues through the full detection head. This yields a reduction in average latency proportional to the fraction of empty frames in the scene, directly translating into higher effective frame rates and freed computational budget for concurrent tasks.

More in detail, the following are our main contributions:

- We train BLANKSKIP end-to-end with a composite loss that jointly optimizes detection quality and early-exit reliability, with class weights enabling explicit control over

the missed-detection vs. computational savings trade-off.
- We formulate the joint optimization of branch placement, training hyperparameters, and confidence threshold as a Bayesian Hyper-Parameter Optimization (HPO) problem.
- We evaluate BLANKSKIP on two datasets against a static (non-adaptive) baseline: on Himax EE, a nano-drone-specific dataset from [17], we achieve a mAP of 0.593, matching the result of a static baseline using the same backbone from the original paper (0.591) while skipping 39.8% of frames and reducing the average Multiply-And-Accumulate (MAC) operations per frame by 22.5%. On the more challenging Cityscapes dataset [6], we maintain competitive mAP (0.204 vs. 0.229, -11%) while skipping 26.7% of frames and reducing the average MACs by 12.1%.
- We deploy an 8-bit quantized version of our model on the Bitcraze Crazyflie 2.1 GAP8 SoC, achieving a mAP of 0.588 at 1.86 average Frames Per Second (FPS), a 24% throughput improvement over the baseline (1.50 FPS), with 0.015 mAP drop.

2. Background and Related Works

2.1. Nano-drone computational platforms

As our target for deployment, we consider the Bitcraze Crazyflie 2.1, a Commercial-Off-The-Shelf (COTS) quadrotor with a diameter of 10 cm and a weight of 27 g [3]. The drone is operated by an STM32F405 Microcontroller Unit (MCU) dedicated to state estimation and actuation control. The platform can be extended with companion boards, most notably the AI-deck, which embeds a GAP8 SoC from GreenWaves Technologies [7]. GAP8 is a parallel ultra-low-power processor based on the RISC-V architecture, designed for Digital Signal Processing (DSP) and AI applications, featuring one fabric controller core and a cluster of 8 additional cores clocked up to 175 MHz, designed for parallel workloads acceleration. The on-chip memory hierarchy consists of 64 kB of L1 scratchpad shared among the cluster cores and 512 kB of L2 SRAM on the fabric controller side. The AI-deck further provides a Himax HM01B0 ultra-low-power grayscale camera and off-chip L3 memories (8 MB RAM and 64 MB Flash) accessible via HyperBus.

2.2. Object detection on constrained devices

Deploying OD on resource-constrained devices has attracted growing research interest, driven by the demand for real-time onboard perception in mobile and embedded systems. The Single Shot Detector (SSD) [21] has become a widely adopted framework for this setting, consisting of a feature extraction backbone and multiple convolutional prediction heads. MobileNets [11, 26] are commonly used as low-cost backbones for SSD in embedded contexts. Compact end-to-end detectors such as You Only Look Once (YOLO) Nano [29] have also been explored, leveraging Neural Architecture Search (NAS) driven design to minimize both model size and inference cost; while SSD-based models offer a more compelling speed advantage, YOLO-based alternatives usually trade some throughput for improved detection accuracy and robustness [14].

These architectures are typically further compressed for deployment via post-training quantization [12] and structured pruning [8]. Nevertheless, deploying such models on severely constrained platforms remains challenging: the work of [28] reports only 0.71 FPS running SSDLite-MobileNetV2 on a Raspberry Pi B3+ mounted on a standard-sized drone, already exceeding the power and size budget of nano-drone systems. The authors of [16] demonstrate a more power-efficient deployment on a GAP8 MCU, achieving 1.6 FPS at 117 mW, yet relying on a fixed computational graph regardless of input content.

Adaptive inference. The goal of adaptive inference is to dynamically adjust the computational cost at runtime based on the complexity of each input [9]. Several directions have been explored: Slimmable Networks [31] allow a single model to operate at different channel widths, while cascade approaches route easy inputs through lightweight models and hard inputs through heavier ones [20]. In EE models, side-branch classifiers are inserted at intermediate layers, and inference is terminated as soon as a confidence criterion, typically entropy-based, is met.

BranchyNet [27] introduced the EE paradigm, demonstrating that a large fraction of inputs can be correctly classified at shallow layers with significant latency savings. A key advantage of EE, which makes this solution deployment-friendly, is that input-adaptive compute is achieved within a single model, without the memory overhead of multiple networks, except for the extra branch, which is usually negligible. While this is also true for Slimmable networks [31], the latter introduce additional complexity during inference (extra routing logic, partial loading of each layer's parameters, which influences optimal computation scheduling, etc).

However, while EE has been extensively studied for image classification [24, 27, 31], its application to dense tasks like OD remains largely underexplored. This is mainly because entropy-based EE criteria designed for classification do not naturally extend to tasks such as OD, which produce dense outputs (grids of bounding box locations and class probabilities), each with its own score distribution.

AdaDet [30] addresses this gap by integrating detection heads at intermediate backbone layers and introducing a bounding-box uncertainty criterion to assess exit confidence, while DynamicDet [20] adopts a cascade of two detectors with a learned difficulty router. These works, however, target high-power GPU platforms and are incompatible with the tight computational constraints of nano-drone

deployment. While they do not report exact model sizes, their simpler models require ≈5-10 GFLOPs per inference, roughly one order of magnitude more than the DNNs we deploy onboard the Crazyflie in this paper.

Sabet et al. [25] propose temporal early-exits for video object detection, where changes between consecutive frames trigger full computation while unchanged frames reuse previous detections.

This approach relies on fixed camera placement (the authors target surveillance scenarios), therefore, it does not apply to drone use cases. Hector et al. [10] introduce elastic neural networks for edge deployment, incorporating multiple detection heads at different depths to enable runtime adaptation based on external resource constraints (e.g., battery levels). However, this method relies on an exogenous signal rather than per-sample content-based decisions to tune the computational effort, which limits applicability to scenarios with highly variable frame content such as nano-drone exploration.

3. Methodology

In nano-drone applications, OD networks must process video streams in real-time under severe power, memory, and compute constraints. However, a significant fraction of frames captured during exploration may contain no objects of interest. Therefore, processing these "empty" frames unnecessarily wastes compute and, therefore, latency. Based on this intuition, we propose BLANKSKIP, a simple yet effective EE mechanism that classifies frames as empty before completing the full detection pipeline.

BLANKSKIP reduces the average OD latency, enabling higher average frame rates. This, in turn, can improve detection rates during fast flight maneuvers and, consequently, enable more efficient exploration of the environment. For example, [17] has analyzed the non-trivial dependencies between flight speed, inference latency and overall exploration time for nano-drones, showing that faster-moving UAVs often "miss" objects due to the low OD rates. Additionally, in multi-task deployments, freed computational resources thanks to EE can be allocated to other, non-critical, concurrent tasks.

3.1. Problem formulation

We reformulate OD as two sequential sub-problems:

1. *Binary Classification*: Determine whether an image contains objects of interest (non-empty) or not (empty);
2. *Object Detection*: Execute full detection only when the image is confidently classified as non-empty.

A straightforward approach would involve two separate models, one per stage. However, this would greatly increase the memory footprint and total computational cost of the system. Instead, we share the same backbone network for both tasks. To achieve this, we adapt the EE paradigm, originally proposed in [27]: an intermediate branch performs binary empty/non-empty classification to gate the OD pipeline. When an image is confidently classified as empty, the system bypasses the remaining backbone layers and detection stages entirely, yielding the computational savings described above while maintaining a single unified model.

Let $I \in \mathbb{R}^{H\times W\times C}$ denote an input image and $\mathcal{D}(\cdot)$ a conventional OD pipeline composed of a feature extraction backbone and detection heads. At intermediate layer ℓ of the backbone, we extract feature maps $\mathbf{x}_\ell \in \mathbb{R}^{H'\times W'\times C'}$ and pass them to a lightweight EE branch $f_{\text{EE}}(\cdot)$ that produces binary classification probabilities:

$$P(c \mid \mathbf{x}_\ell) = \text{softmax}(f_{\text{EE}}(\mathbf{x}_\ell)), \quad c \in \{0, 1\} \tag{1}$$

where $c = 1$ denotes "empty" and $c = 0$ denotes "non-empty". The inference decision mechanism is:

$$\mathcal{O}(I) = \begin{cases} \emptyset & \text{if } P(1 \mid \mathbf{x}_\ell) \geq \tau \\ \mathcal{D}(I) & \text{otherwise} \end{cases} \tag{2}$$

where $\tau \in [0.5, 1]$ is a confidence threshold controlling the trade-off between computational savings and detection reliability. High-confidence gating (high τ) minimizes false positives, i.e. incorrectly skipping images containing objects, which is critical as it would result in missed detections. Conversely, false negatives (empty images classified as non-empty), incur a computational overhead but do not necessarily affect detection accuracy, as $\mathcal{D}(\cdot)$ can still produce empty outputs when no objects are present.

3.2. Network architecture

Our OD pipeline is based on a SSD algorithm composed of a MobileNetV2 feature extractor and multiple detection heads [21, 26]. As illustrated in Fig. 1, the EE branch attaches at an intermediate layer ℓ of the backbone, which shall balance sufficient semantic abstraction for reliable empty/non-empty discrimination with the computational savings obtained when early-exit triggers.

The EE branch $f_{\text{EE}}(\cdot)$ is deliberately lightweight to minimize overhead. Its architecture consists of:

1. Two 3×3 convolutional layers with batch normalization and ReLU6 activation, where the second uses stride 2;
2. Global Average Pooling (GAP) to collapse spatial dimensions into a compact feature vector;
3. Two fully-connected layers.

Formally, the EE branch computes:

$$\mathbf{h} = \text{GAP}(\sigma(\text{BN}(W_2 * \sigma(\text{BN}(W_1 * \mathbf{x}_\ell))))) \tag{3}$$

$$f_{\text{EE}}(\mathbf{x}_\ell) = W_4(\sigma(W_3\mathbf{h})) \tag{4}$$

where $*$ denotes convolution and σ is ReLU6. The branch dimensions are designed such that the first convolutional

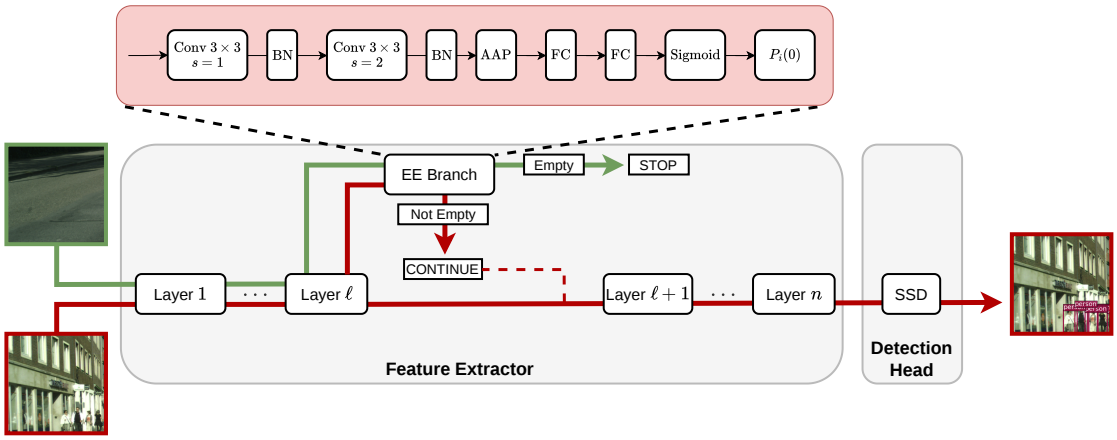

Figure 1. Overview of the proposed BLANKSKIP architecture. The backbone processes the input frame through layers 1 to ℓ. Then, the EE branch predicts the "emptyness" probability $P_i(1)$. If the scene is classified as *empty*, inference stops early (green path), avoiding the remaining backbone layers and the SSD detection head. Otherwise, the full pipeline runs to produce bounding box predictions (red path).

layer maintains the channel dimension of $\mathbf{x}_\ell$, while the second reduces it to 64 channels. The two fully-connected layers produce 64-dimensional intermediate feature vectors, before the final 2-class output. The EE branch introduces 50K–138K additional parameters depending on the channel dimension at layer ℓ, increasing the static MobileNetV2 SSD parameter count by only 1.1–3.0% over the original 4.67M. Its computational cost in MAC operations varies with branch placement and the spatial resolution of the feature map, as detailed below.

3.3. Training strategy

The system is trained end-to-end with a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{EE}} \tag{5}$$

where $\mathcal{L}_{\text{loc}}$ is Smooth-L1 localization loss, $\mathcal{L}_{\text{cls}}$ is classification loss [21], and $\mathcal{L}_{\text{EE}}$ is an additional weighted binary cross-entropy term penalizing empty/non-empty misclassifications by the EE branch. Formally:

$$\mathcal{L}_{\text{EE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_0 \cdot (1-y_i)\log P_i(0) + w_1 \cdot y_i \log P_i(1)\right] \tag{6}$$

where $P_i(c) = P(c \mid \mathbf{x}_\ell^{(i)})$ and $y_i \in \{0,1\}$ indicates whether image i is empty. The class weights w_0 and w_1 enable tuning the trade-off between detection reliability and computational efficiency. Higher w_0 penalizes false positives (non-empty images incorrectly classified as empty), prioritizing detection completeness, which is critical when missing objects are costly. Conversely, a higher w_1 penalizes false negatives (empty images unnecessarily processed), maximizing energy savings in resource-constrained scenarios. The weighting parameter λ in Eq. 5 balances the EE objective with detection accuracy. Ground truth labels are derived automatically from detection annotations, requiring no additional annotation effort.

3.4. Architectural exploration

The design of an EE network involves several interdependent decisions that cannot be optimized in isolation. The placement of the EE at layer ℓ affects both the semantic richness of features available for classification and the potential computational savings [27, 30]. The relationship between branch placement and compute cost is non-trivial: placing it earlier enables larger potential backbone savings, but the branch must process high-resolution feature maps, thus being more expensive in terms of MACs; conversely, deeper layers have smaller spatial dimensions, making the branch usually lighter in terms of computation, but leave less backbone computation to skip. Additionally, later layers encode semantically richer representations, potentially improving the branch's ability to discriminate between empty and non-empty frames. This creates the potential for a sweet spot within the network where branch cost, compute savings, and classification accuracy are jointly optimized. Moreover, the optimal branch placement influences training dynamics, and thus the optimal learning rate, loss weights, etc.

This interdependency extends to the confidence threshold τ, another critical design parameter. While the threshold can be adjusted at inference time to dynamically tune the accuracy/compute trade-off, its optimal value is coupled with training hyperparameters. We address this multi-dimensional problem through a Bayesian HPO stage [23]. We select a Bayesian Optimization method because our search space contains mixed parameter types, including categorical branch placement, continuous learning rates, and

loss weights. Moreover, each objective evaluation requires full network training, making exhaustive search computationally prohibitive.

We employ the Tree-structured Parzen Estimator (TPE) [2] implemented in Optuna [1] to search over candidate EE branch placements, learning rates, batch sizes, loss weights, and class weights. The composite HPO objective function balances three competing goals: preserving detection quality (measured by mAP without early-exit to isolate detection performance), maximizing computational savings, and achieving high EE classification accuracy. Formally, we maximize:

$$\mathcal{J} = \text{mAP}_{\text{baseline}} \times \mathcal{S}(\ell) \times A_{\text{EE}}(\ell) \tag{7}$$

where $\text{mAP}_{\text{baseline}}$ is the detection performance without early-exit. $\mathcal{S}(\ell) = 1 - \mathcal{F}_{\text{EE}}(\ell)/\mathcal{F}_{\text{full}}$ represents the static Floating-point Operations (FLOPs) saving achievable by exiting at layer ℓ, with $\mathcal{F}_{\text{EE}}(\ell)$ and $\mathcal{F}_{\text{full}}$ denoting the computational cost of early-exit and full forward pass, respectively. Lastly, $A_{\text{EE}}(\ell)$ is the EE branch classification accuracy at the optimal threshold. For each candidate configuration, we perform full network training followed by post-training EE threshold (τ) optimization via ternary search.

3.5. Deployment details

To deploy baseline (static) MobileNet-SSD detectors and BLANKSKIP EE solutions on the Floating Point Unit-less GAP8 SoC, we quantize all DNNs to 8-bit integer format with Quantization-Aware Training (QAT), using the open-source PLiNIO library [13]. We apply affine quantization, with simple min-max quantizers for weights and PaCT [5] for activations. We deploy full-integer models, folding normalization layers parameters in the preceding linear/convolutional layers. We use the open-source DORY compiler [4] to generate low-level C code for GAP8, accelerating all supported layers on the 8-core parallel cluster.

4. Experimental Results

4.1. Experimental setup & datasets

We compare BLANKSKIP with the work of Lamberti et al. [17], which trained static MobileNet-SSDs to perform object detection while navigating an environment through different bio-inspired algorithms, targeting the same nano-UAV platform considered in our work. In addition, to show that our method generalizes to larger and more complex datasets, we also evaluate it on a version of Cityscapes [6] adapted for OD and for our device class.

Himax. The Himax dataset [17] contains 321 training and 279 test images at 240×320 resolution, collected with the on-board camera of the Crazyflie 2.1 UAV, and annotated for two object categories: "Bottle" and "Tin can". Since all original images contain at least one object, the dataset does not reflect the high proportion of empty frames typical of real-world deployment scenarios. To address this, we augment the dataset with synthetically generated negative samples. Specifically, for each image we extract random object-free regions (40–70% of the image area), verified by zero intersection with all bounding boxes; positive examples are generated analogously by generating a randomly centered crop around one of the objects in the original dataset. All crops are resized to 240×320 via letterbox scaling. The augmented training set comprises 496 images (321 positive, 175 negative), and the test set 462 images (279 positive, 183 negative). We refer to this augmented dataset as Himax EE, to distinguish it from the original Himax dataset, on which we also report results for direct comparison with [17].

Cityscapes. Cityscapes is an urban scene understanding dataset originally designed for semantic and instance segmentation, covering eight object categories: "Person", "Rider", "Car", "Truck", "Bus", "Train", "Motorcycle", and "Bicycle". To adapt it for OD, we derive bounding box annotations directly from the instance masks. The original images are at 2048×1024 resolution, which makes one-shot processing unfeasible on nano-UAV class devices. In fact, the peak intermediate activations of a static MobileNet-SSD would reach 62.88 MiB, which is almost 1 order of magnitude more than the total RAM available in GAP8. A common solution to enable inference with such high-resolution images on memory-limited targets is sequential patch-based processing [18], at the cost of increased latency and of a potential mAP drop for objects crossing the patch boundaries.

We follow this approach, partitioning each image into eight non-overlapping 512×512 tiles, thus reducing the peak memory to 7.86 MiB, which fits the GAP8 available RAM. The resulting dataset comprises 27,800 tiles (19,040 train / 4,760 validation / 4,000 test). Notably, approximately 36%, 32%, and 24% of the train, validation, and test tiles are empty, not requiring extra negative sampling for EE training. This suggests that EE can be beneficial not only when processed frames are frequently empty, but also in scenarios with denser object distributions, if images are processed in patches.

Implementation details. All models are implemented in PyTorch 2.10. We generate all BLANKSKIP models attaching EE branches to a MobileNet-SSD network with a width multiplier $\alpha = 1.0\times$. For Himax and Himax EE experiments, the weights of the backbone are initialized from a detector pre-trained on an Open Images [15] subset covering the Bottle and Tin can categories, as proposed by the dataset authors [17]. For Cityscapes experiments, weights are initialized from a COCO [19] pre-trained detector.

Table 1. Effect of EE branch positioning on Himax EE (float32).

Model	τ	mAP	$\text{mAP}_{\text{no-EE}}$	Acc [%]	Skip [%]	MAC_{avg} [M]	MAC_{EE} [M]
Static 0.5×	–	0.412	–	–	–	193	–
Static 0.75×	–	0.479	–	–	–	358	–
Static 1.0×	–	0.591	–	–	–	534	–
EE, Stage 1 ($\ell = 1$)	0.541	0.543	0.603	85.3	34.2	708	375
EE, Stage 2 ($\ell = 2$)	0.717	0.533	0.587	89.6	39.2	456	178
EE, Stage 3 ($\ell = 4$)	0.742	0.555	0.596	89.6	38.7	406	165
EE, Stage 4 ($\ell = 9$)	**0.685**	**0.593**	**0.616**	**94.4**	**39.8**	**414**	**230**
EE, Stage 5 ($\ell = 11$)	0.904	0.592	0.620	95.7	41.1	430	266

MAC_{avg}: dataset average. MAC_{EE}: early-exit path cost. $\text{mAP}_{\text{no-EE}}$: mAP with early-exit disabled.

4.2. Himax EE results

We compare our EE solution with three static MobileNetV2-SSD configurations with width multipliers $\alpha \in \{0.5, 0.75, 1.0\}$. The full-width model (SSD-MbV2-1.0) has 4.67 M parameters and requires 534 MMACs per forward pass, while SSD-MbV2-0.75 and SSD-MbV2-0.5 have 2.68 M and 1.34 M parameters and require 358 MMACs and 193 MMACs, respectively.

All baseline models are trained following the recommended hyperparameters configuration from [17]: 100 epochs, RMSprop optimizer with an initial learning rate of 1×10^{-4} and a batch size of 24. The learning rate is decayed by a factor of 0.95 every 25 epochs. Data augmentation includes random horizontal mirroring, random crop sampling, and random brightness adjustment, individually applied with a probability of 0.5. We first verify our training setup on the original Himax dataset (without empty frames), reproducing their baseline results: our static MobileNetV2-SSDs achieve a mAP of 0.571, 0.495, and 0.429 at width multipliers $1.0\times$, $0.75\times$, and $0.50\times$, respectively, similar or better than the 0.550, 0.460, and 0.430 reported in the original work. On Himax EE, which includes empty frames, the static MobileNetV2-SSDs achieve a mAP of 0.591, 0.479 and 0.412 respectively, showing that the augmented dataset exhibits similar complexity to the original one.

For BLANKSKIP models, in order to analyze in detail the impact of EE branch placement on computational cost and accuracy, we perform 5 separate Bayesian HPO runs, one for each of the first 5 stages of the MobileNetV2 $1.0\times$ backbone. Each stage comprises one or more inverted residual block layers, sharing the same output resolution and channel count. Namely, stages $(1, 2, 3, 4, 5)$ include $(1, 2, 3, 4, 3)$ layers respectively. Each run forces the EE branch to be attached to one of the layers belonging to that stage, and optimizes five hyperparameters: the specific EE layer within the stage, the branch learning rate, the batch size, the loss weight λ (Eq. 5), and the class weight w_1 for non-empty samples. Each optimization run performs 100 trials, maximizing the composite objective presented in Eq. 7. For each candidate configuration, the confidence threshold τ is selected via ternary search on validation data to maximize overall classification accuracy.

Table 1 reports detailed metrics for the five EE configurations discovered through Bayesian optimization, in float32 precision. The table reports mAP and $\text{mAP}_{\text{no-EE}}$, representing detection accuracy with and without early-exit enabled, respectively. Acc denotes the EE branch classification accuracy, while Skip indicates the percentage of images triggering early-exit. MAC_{avg} shows the average computational cost across the dataset with EE active, and MAC_{EE} represents the cost of the early-exit path when taken.

Notably, all EE configurations achieve a $\text{mAP}_{\text{no-EE}}$ higher than the static full-width model (0.591). We attribute this improvement to the fact that the EE branch likely provides beneficial gradient signals to intermediate layers during training, acting as a regularizer similarly to what observed in BranchyNet [27].

Stage 4 ($\ell = 9$) emerges as the most accurate configuration, achieving a mAP of 0.593 with an average cost of 414 MMACs per image. The 94.4% EE branch accuracy enables skipping full detection for 39.8% of the images. Compared to a static full-width baseline (534 MMACs), Stage 4 reduces average inference cost by 22.5% while maintaining competitive detection accuracy. Crucially, when early-exit is taken, the cost drops to only 230 MMACs, a 57.9% reduction compared to the baseline.

Interestingly, the results show that placing the EE branch in shallow layers (stage 1 or 2) is significantly less accurate (85.3-89.6%). This demonstrates that identifying frames with no objects of interest, without relying on priors on the stillness of the frame (as in [25]) is a non-trivial task that requires latent representations with sufficient semantic abstraction, thus justifying an EE approach versus other forms of skipping. Additionally, Stage 1 placement also suffers from prohibitively high early-exit costs due to processing large feature maps at shallow depths. Despite a 34.2% skip rate, the average cost exceeds the static baseline (708 MMACs), resulting in a net computational overhead rather

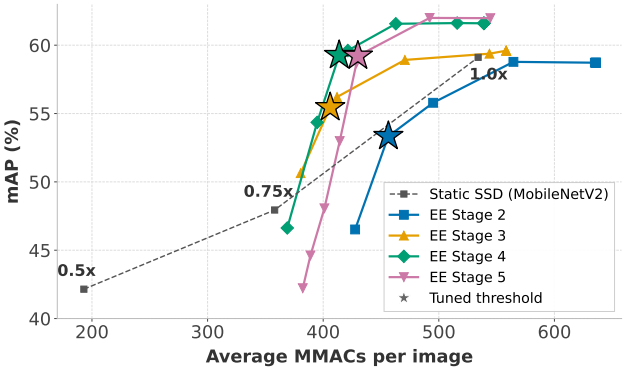

Figure 2. Accuracy-efficiency trade-off for with threshold τ sweep (0.5-0.99) on the Himax EE dataset. Stars mark optimal thresholds according to Eq. 7. Dashed line: static SSD at different widths. Stage 1 is omitted due to prohibitively high exit costs from large feature maps (120×160).

than savings. Moreover, the low EE classification accuracy at earlier stages also further degrades the overall mAP.

The limited skip rates (35–41%) across all configurations reflect the dataset composition: with approximately 40% negative samples in the test set, most images require full detector processing. Nevertheless, the results validate BLANKSKIP's effectiveness in preserving detection quality while reducing inference cost, especially considering that, in real-world settings, the share of empty images could be substantially higher.

Figure 2 visualizes the accuracy-efficiency Pareto frontier obtained across stages 2-5 by sweeping the EE decision threshold τ of Eq. 2 in the [0.5,0.99] range, with stars marking the optimal operating points selected by Bayesian Optimization to maximize Eq. 7 and reported in Table 1. Points to the right of each curve correspond to *larger* τ values, that let the system skip detection only if the EE confidence is very high, thus yielding higher-cost, higher-mAP results. Conversely, points to the left are obtained with *small* τ, thus achieving lower-mAP and lower-cost. The dashed line represents the static SSD Pareto frontier at different width multipliers. For better visualization, we omit stage 1 entirely due to the sub-optimal results discussed above, and we do not show points reaching a mAP smaller than the static 0.5× architecture.

The threshold sweep reveals further insights into the behavior of each configuration. The mAP obtained placing the branch at Stage 5, despite being competitive at the optimum according to Eq. 7, degrades more rapidly as τ decreases. Stage 2, while offering the lowest early-exit path cost (178 MMACs), is unable to match the accuracy of deeper configurations even at high thresholds, confirming that shallow representations are insufficient for reliable empty-frame detection. Based on this analysis, we select Stage 4 for QAT and deployment.

Note that, once the EE branch placement is selected, different operating modes (e.g., different points from the green curve in Fig. 2) can still be switched while the system is operating, by changing a single scalar configuration (τ), thus enabling runtime tuning of the latency vs mAP trade-off.

4.2.1. Deployment results

To validate the practical feasibility of our approach on nano-drone hardware, we deploy the optimized Stage 4 BLANKSKIP model alongside static baseline configurations (1.0×, 0.75×, and 0.5× width multipliers) on the GAP8 microcontroller using the DORY framework [4]. We configure the GAP8 multicore cluster to operate at 1.2 V supply voltage with a clock frequency of 160 MHz as in the reference paper [17].

Prior to deployment, we apply QAT to all models for 100 epochs, starting from their float32 checkpoints, maintaining the same optimizer (RMSprop, batch size 64) and learning rate schedule (decay factor 0.95 every 25 epochs). We empirically find that a higher initial learning rate of 1×10^{-2} is necessary to facilitate adaptation to the quantized domain. To verify this training setup, we also apply QAT to the 1.0× static model on the original Himax dataset, obtaining a mAP of 0.547, higher than the Int8 result reported in [17] of 0.500.

Table 2 reports Int8 performance on Himax EE, and deployment metrics across all configurations, including mAP, computational cost, inference efficiency in MAC operations per cycle, latency, throughput in FPS, and model size in MiB. Note that our results for static models differ slightly from those in [17] because we use the open-source DORY compiler [4], rather than the proprietary toolchain of GAP8.

The static 1.0×-width baseline (no EE) processes frames at 1.50 FPS with an efficiency of 4.96 MAC/cycle and 0.555 mAP. Adding the EE branch introduces minimal computational overhead: the full forward pass requires only 539 MMACs (0.9% increase) due to the lightweight branch architecture that adds only 0.078M parameters, maintaining nearly identical throughput at 1.48 FPS.

When early-exit is triggered at layer 9, computational cost drops dramatically to 230 MMACs, a 57% reduction compared to the static baseline. On the Himax EE dataset, where 39.6% of frames exit early, the average throughput therefore reaches 1.86 FPS (24% improvement with respect to the 1.0× static model, and comparable to 0.75×) while achieving 0.540 mAP, with only a 2.7% drop compared to the static 1.0× DNN. The EE branch accuracy when quantized remains high (90.9%), with a false positive rate of 7.5%. This demonstrates that content-based early-exit enables significant on-device acceleration on resource-constrained microcontrollers with minimal accuracy degradation for autonomous nano-drone deployment.

Table 2. Comparison of static MobileNetV2-SSD configurations and BLANKSKIP on the Himax EE dataset. Int8 quantization is applied to all models. Deployment metrics are measured on GAP8.

Model	mAP	Operations (MMACs)	Efficiency (MAC/cycle)	Latency (ms)	Throughput (FPS)	Model Size (MB)
Static MbV2-SSD 0.5×	0.415	193	4.14	285.3	3.50	1.16
Static MbV2-SSD 0.75×	0.498	358	4.22	523.6	1.91	2.53
Static MbV2-SSD 1.0×	0.555	534	4.96	666.7	1.50	4.41
EE, Stage 4 ($\ell = 9$)	0.540	414	4.87	537.1	1.86	4.49

Table 3. Cityscapes results (float32).

Model	mAP	Acc	Skip	MAC_{avg}	MAC_{EE}
Static 1.0×	0.229	-	-	1807	-
EE, Stage 4 ($\ell = 10$)	0.204	82.3	26.7	1588	855

Acc: EE branch accuracy, Skip: early-exited images [Unit = %]; MAC_{avg}: dataset average. MAC_{EE}: early-exit path cost (backbone + EE branch) [Unit = M (10^6)]

Finally, we also simulated the execution of our BLANKSKIP on the in-field data collected in [17]. Namely, we asked the authors to provide us with the raw video from the drone camera in one of their exploration experiments[1] and manually annotated all frames with bounding boxes for bottles and tin cans. We filtered out all boxes whose area is smaller than the smallest object in the Himax training set, as we cannot expect the DNN to predict out of distribution; hence, those frames are effectively considered empty from the model's perspective. Lastly, we ran our EE model on the resulting data. The results show that 71.7% of all frames are effectively empty. Our EE model identifies them with a precision of 85.48%. As a result, the EE branch exits early on 54.83% of all frames (1288 out of 2349), yielding an average throughput of 2.07 FPS, a 37.7% improvement over a standard static MobileNetV2-SSD (1.50 FPS).

4.3. Cityscapes results

As a last experiment, to evaluate the generalizability of BLANKSKIP to larger and more complex datasets, we applied it to the patched version of Cityscapes described in Sec. 4.1. The HPO identifies layer 10 as the optimal EE branch positioning, showing that the optimal depth range where EE should be inserted remains similar across datasets.

Table 3 reports the results in terms of mAP and total MAC operations, comparing BLANKSKIP with a static 1.0× model. Our model achieves a 12% average computational reduction (1588 vs 1807 MMACs) by exiting early on 26.7% of tiles. When EE triggers, inference costs only 855 MMACs, a 53% reduction compared to the full pass (1855 MMACs), which itself incurs a modest 48 MMACs overhead over the static baseline due to the additional EE branch. The higher absolute costs compared to Himax EE reflect the larger input resolution (512×512 vs 240×320), resulting in approximately $3.5\times$ more MMACs overall.

Unlike on Himax EE, the mAP of the EE model (0.204) is slightly lower than the static baseline (0.229). This difference likely reflects the increased complexity of Cityscapes: with 8 object categories across diverse urban scenes and a higher object density per image, the empty-frame detection task is more challenging, as evidenced by the lower EE branch accuracy. The lower skip rate (26.7% vs 39.8% on Himax EE) further limits computational savings, although the latter could increase substantially in real-world deployments, where empty frames are more frequent. Nonetheless, these results demonstrate that BLANKSKIP successfully generalizes to larger, multi-class detection scenarios, offering viable new trade-offs between accuracy and efficiency in more demanding settings.

5. Conclusions

We have presented BLANKSKIP, a simple yet effective method to improve the average OD throughput in constrained embedded devices such as nano-drones, with minimal accuracy drops. With experiments on two different datasets, we have shown that our solution provides multiple Pareto-optimal, runtime-switchable operating points in terms of mAP vs complexity, while incurring minimal memory overheads. Our future works will encompass ways to improve the EE accuracy, e.g., by tuning the threshold τ dynamically based on previously processed frames.

Acknowledgments

We thank the authors of [17] for providing us with pretrained model checkpoints for comparisons, and raw drone videos.

This publication is part of the project PNRR-NGEU which has received funding from the MUR – DM 118/2023.

[1] Drone flight video with onboard camera feed and detection overlay available at this link: www.youtube.com/watch?v=BTin8g0nyko.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. 5

[2] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, page 2546–2554, Red Hook, NY, USA, 2011. Curran Associates Inc. 5

[3] Bitcraze AB. Crazyflie 2.1 Datasheet, Rev. 3. https://www.bitcraze.io/documentation/hardware/crazyflie_2_1/crazyflie_2_1-datasheet.pdf, 2019. Accessed: 2026. 1, 2

[4] Alessio Burrello, Angelo Garofalo, Nazareno Bruschi, Giuseppe Tagliavini, Davide Rossi, and Francesco Conti. Dory: Automatic end-to-end deployment of real-world dnns on low-cost iot mcus. *IEEE Transactions on Computers*, 70 (8):1253–1268, 2021. 5, 7

[5] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 5

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 2, 5

[7] Eric Flamand, Davide Rossi, Francesco Conti, Igor Loi, Antonio Pullini, Florent Rotenberg, and Luca Benini. 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)*, pages 1–4, 2018. 1, 2

[8] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015. 2

[9] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic Neural Networks: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(11):7436–7456, 2022. 2

[10] Rory Hector, Mohammad Umer, Asif Mehmood, Zhu Li, and Shuvra Bhattacharyya. Scalable object detection for edge cloud environments. *Frontiers in Sustainable Cities*, 3:675889, 2021. 3

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1, 2

[12] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018. 2

[13] Daniele Jahier Pagliari, Matteo Risso, Beatrice Alessandra Motetti, and Alessio Burrello. Plinio: A user-friendly library of gradient-based methods for complexity-aware dnn optimization. In *2023 Forum on Specification & Design Languages (FDL)*, pages 1–8, 2023. 5

[14] Jeong-ah Kim, Ju-Yeong Sung, and Se-ho Park. Comparison of faster-rcnn, yolo, and ssd for real-time vehicle type recognition. In *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pages 1–4, 2020. 2

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 5

[16] Lorenzo Lamberti, Manuele Rusci, Marco Fariselli, Francesco Paci, and Luca Benini. Low-power license plate detection and recognition on a risc-v multi-core mcu-based vision system. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021. 2

[17] Lorenzo Lamberti, Luca Bompani, Victor Javier Kartsch, Manuele Rusci, Daniele Palossi, and Luca Benini. Bio-inspired autonomous exploration policies with cnn-based object detection on nano-drones. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6, 2023. 1, 2, 3, 5, 6, 7, 8

[18] Ji Lin, Wei-Ming Chen, Han Cai, Chuang Gan, and Song Han. Mcunetv2: memory-efficient patch-based inference for tiny deep learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. 5

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[20] Zhihao Lin, Yongtao Wang, Jinhe Zhang, and Xiaojie Chu. Dynamicdet: A unified dynamic architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2, 3, 4

[22] Daniele Palossi, Andres Gomez, Stefan Draskovic, Andrea Marongiu, Lothar Thiele, and Luca Benini. Extending the lifetime of nano-blimps via dynamic motor control. *J. Signal Process. Syst.*, 91(3–4):339–361, 2019. 1

[23] Andrew Quitadadmo, James Johnson, and Xinghua Shi. Bayesian hyperparameter optimization for machine learning based eqtl analysis. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, page 98–106, New York, NY, USA, 2017. Association for Computing Machinery. 4

[24] Haseena Rahmath P, Vishal Srivastava, Kuldeep Chaurasia, Roberto G. Pacheco, and Rodrigo S. Couto. Early-exit deep

neural network - a comprehensive survey. *ACM Comput. Surv.*, 57(3), 2024. 1, 2

[25] Amin Sabet, Jonathon Hare, Bashir Al-Hashimi, and Geoff V. Merrett. Temporal early exits for efficient video object detection. *SSRN Electronic Journal*, 2022. 3, 6

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1, 2, 3

[27] Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469, 2016. 1, 2, 3, 4, 6

[28] Quang Khôi Trân, Nguyen Quang, and Khanh Ngo. Object detection for drones on raspberry pi potentials and challenges. *IOP Conference Series: Materials Science and Engineering*, 1109:012033, 2021. 2

[29] Alexander Wong, Mahmoud Famuori, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl, and Jonathan Chung. Yolo nano: A highly compact you only look once convolutional neural network for object detection. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 22–25. IEEE, 2019. 1, 2

[30] Le Yang, Ziwei Zheng, Jian Wang, Shiji Song, Gao Huang, and Fan Li. Adadet: An adaptive object detection system based on early-exit neural networks. *IEEE Transactions on Cognitive and Developmental Systems*, 16(1):332–345, 2024. 2, 4

[31] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas S. Huang. Slimmable neural networks. *CoRR*, abs/1812.08928, 2018. 1, 2