

# ADAPTIVE MODEL COMPRESSION OF YOLO11-N FOR ENERGY-EFFICIENT REAL-TIME DRONE VISION

**Dr.K.V.S.PRASAD, DR.K.E. PURUSHOTHAMAN, Dr.S.UMA, RAMESH KUMAR, M.R. EZILARASAN, Dr.D. SUDHAGAR, Dr.R. SENTHAMIL SELVAN**

Professor, GMRIT Deemed to be University, Rajam -532 127, Vizianagaram, Andhra Pradesh

Associate Professor, Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D

Institute of Science and Technology, Chennai

Assistant Professor (Senior Grade), Department of ECE, Saveetha School of Engineering,

Saveetha University, Chennai

Assistant Professor, Department of Electronics and Communication Engineering, Aditya

University, Surampalem, Andhra Pradesh, India

Associate Professor, Department of ECE, Veltech Rangarajan Dr. Sagunthala R&D

Institute of Science and Technology, Avadi, Chennai

Professor, Department of CSE, Saveetha School of Engineering, Saveetha University,

Chennai

Associate Professor, Department of ECE, Saveetha School of Engineering, Saveetha

University, Chennai

Email: [prasad.kvs@gmritudu.edu.in](mailto:prasad.kvs@gmritudu.edu.in), [k.e.purushothaman1992@gmail.com](mailto:k.e.purushothaman1992@gmail.com), [umasumi13@gmail.com](mailto:umasumi13@gmail.com),  
[ramesh92001@gmail.com](mailto:ramesh92001@gmail.com), [arasanezil@gmail.com](mailto:arasanezil@gmail.com), [sudhaapriyaa@gmail.com](mailto:sudhaapriyaa@gmail.com),  
[selvasenthamil2614@gmail.com](mailto:selvasenthamil2614@gmail.com)

## ABSTRACT

Unmanned aerial vehicles (UAVs) equipped with onboard vision systems are increasingly deployed for surveillance, traffic monitoring, and disaster response, yet their limited battery capacity and embedded compute budget severely constrain the use of modern convolutional object detectors. This paper presents an Energy-Adaptive CNN Compression (EACC) framework, which combines the structured channel pruning, knowledge distillation, and mixed-precision quantization with an energy-aware controller that dynamically adjusts the trade-off between the detection accuracy and on-board energy consumption. Conventional compression pipelines optimize accuracy and model size independently, while the proposed controller optimizes the compression ratio, directly affecting the pruning and quantizing decisions at each layer, as a constrained optimization problem in which the energy budget of the target embedded platform is directly related to compression ratio. The framework is developed based on the YOLO11-N baseline and tested on the VisDrone2021 benchmark, which is a large-scale benchmark with small, dense, and partially occluded objects taken in various urban and suburban environments. Experimental results show that the proposed model can achieve about 76% reduction in the number of parameters and 66% reduction in energy consumption per frame compared to the uncompressed baseline, while maintaining more than 92% of the mAP@0.5 of the baseline and keeping the model running in real-time on a Jetson Orin Nano platform. The results show that it is possible to implement energy-adaptive, accuracy-aware compression as a viable strategy for deep object-detection deployment on a resource-constrained aerial platform.

**Keywords:** *UAV Object Detection, Model Compression, Structured Pruning, Knowledge Distillation, Quantization, Energy-Aware Deep Learning, VisDrone2021, Edge AI.*

## 1. INTRODUCTION

Unmanned aerial vehicles (UAVs) are now playing a major role in visual perception for real-

time applications including traffic-flow estimation, search and rescue, precision agriculture and urban surveillance. These platforms have many strict requirements on the payload, memory, and battery consumption, which necessitates that the on-board

object detection software, which relies on deep convolutional neural networks (CNNs), be highly energy efficient and compact while maintaining competitive accuracy. However, common ground-station benchmarks show that state-of-the-art detectors, like the YOLO family, have strong detection accuracy, but their direct application to embedded UAV hardware (usually NVIDIA Jetson-class System-on-Modules (SoMs)) is constrained by thermal, power, and latency budgets that are much tighter than those that could be found in server grade GPUs.

The compression approaches of structured pruning, knowledge distillation and quantization are each found to decrease the computational overhead of CNN-based detectors. The vast majority of the existing pipelines perform these steps in a sequential fashion and without adapting to the platform they are deployed on, using a fixed compression ratio as a hyperparameter, rather than a parameter that should be varied based on platform energy profile and required detection-accuracy floor. This is particularly troublesome for UAV vision, where aggressive compression can adversely affect a small and dense amount of objects (pedestrians, bicycles, tricycles), and where energy overruns can directly reduce the flight time or result in a mission failure.

This paper tackles the challenge of compressing a CNN with limited energy and accuracy budget for on-board UAV object detection. We present a structured channel pruning, knowledge distillation and mixed-precision quantization framework under a single controller responsible for allocating compression aggressiveness among layers based on two criteria: (i) how much the layer contributes to the detection accuracy and (ii) how much the layer consumes for inference energy, within a user-defined energy budget and minimum required mAP. The framework is evaluated solely against the VisDrone2021 benchmark, which is a widely-used UAV dataset for similar tasks in the field of aerial object detection and compression studies.

Deep learning based object detectors have shown impressive results in UAV vision applications, but are still difficult to implement on the UAV platform because of limited computing power, battery capacity and thermal environment. The current state-of-the-art detectors, such as YOLO-based detectors, are able to achieve high detection accuracy but have a large number of parameters, high computational complexity, and high energy consumption, preventing them from being directly applied to lightweight embedded platforms.

Recent CNN compression methods such as structured pruning, knowledge distillation and quantization have proven to be effective at model size reduction. Most of the current methods optimize them independently, or use fixed compression ratios without taking into account the dynamic balance between the accuracy requirements for detection and available energy resources. Therefore, the ability of detecting the presence of small and dense objects is greatly hindered in practical applications of UAV surveillance and monitoring systems, especially when the compression gain is too high.

In addition, most existing energy-aware deep learning methods measure the energy consumed by the optimized model, but do not use energy constraints as a part of the compression decision making process. Thus, an adaptive compression framework which can achieve good detection accuracy, calculation efficiency and energy saving for real-time UAV deployment is still not known.

In this regard, this paper introduces an Energy-Adaptive CNN Compression (EACC) framework that combines both structured channel pruning and knowledge distillation with mixed-precision quantization with an energy-aware controller. The proposed scheme takes into consideration the energy constraints of the hardware and dynamically decides whether to compress or not, while keeping the detection performance.

Based on the identified research gap, this study aims to answer the following research questions:

**RQ1:** Can CNN compression be optimized jointly to decrease the on-board energy consumption, computational complexity, and to ensure accurate object detection performance for UAVs?

**RQ2:** Is it possible to achieve greater accuracy – energy trade-off than with traditional compression techniques that follow a fixed energy pruning and quantization strategy?

**RQ3:** What is the performance of the proposed compressed detection model in terms of accuracy of detection, speed of inference and energy efficiency on a resource-limited embedded UAV platform?

To address these research questions, the objectives of this study are:

- To create a structured channel pruning, knowledge distillation, and mixed precision quantization based Energy-Adaptive CNN Compression (EACC) framework.
- To develop a controller for energy awareness that dynamically changes compression level

based on the accuracy requirement to be detected and energy constraint.

- To benchmark the proposed framework on the VisDrone2021 dataset and NVIDIA Jetson Orin Nano platform, accuracy, computational efficiency, and energy consumption metrics are used.

The following hypothesis is developed based on the research problem and research questions identified:

#### **H1 (Alternative Hypothesis):**

The proposed energy-adaptive CNN compression framework, which simultaneously compresses CNN with structured channel pruning, knowledge distillation and mixed-precision quantization, can effectively enhance the energy efficiency of UAV object detection without sacrificing the comparable detection accuracy compared with the original uncompressed CNN.

#### **H0 (Null Hypothesis):**

The energy-adaptive CNN compression framework is not really superior in terms of the accuracy–energy trade-off to conventional CNN compression approaches without adaptive energy-based optimization.

This paper is organized as follows: In Section II, the related works for UAV object detection and CNN compression are reviewed. The VisDrone2021 dataset and the formalization of the compression problem is described in Section III. In Section IV, the proposed EACC methodology along with mathematical model and training algorithm are presented. Section V describes experimental setup and Section VI presents and analyzes the data. This ablation study is presented in Section VII, the deployment of the ablator within the vehicle is discussed in Section VIII, and the paper is concluded in Section IX.

## **2. RELATED WORKS**

### **2.1. UAV-based Object Detection**

Compared to ground-based detection, object detection from aerial imagery has three distinguishing characteristics: Objects are often small in comparison to a frame, the viewpoint is dependent on the altitude and gimbal angle, and scenes are often highly cluttered. These challenges are specifically captured in the benchmark datasets like VisDrone and UAVDT, and recurrent designs of detectors have addressed these issues by introducing multi-scale feature fusion, attention mechanism and anchor-free heads for recalling small objects. Despite the high accuracy-to-latency ratio of the single-stage detectors of the YOLO

family, their use on UAVs is still limited by size constraints. Single-stage detectors of the YOLO family are the most widespread option for the flow of UAVs due to their high accuracy-to-latency ratio compared to two-stage detectors like Faster R-CNN, however, due to the size constraints, their use is limited.

### **2.2. Model Compression for Embedded Vision**

Structured pruning selectively drops entire channels or filters depending on a criterion of importance, such as L1-norm, batch-normalization scaling factor or gradient sensitivity, and models can be directly accelerated by standard hardware, without the need for special sparse kernels. Knowledge distillation comes to the rescue to transfer the representational power of a big teacher network to a small pupil network by soft-label and feature-level supervision, and can partially recover from the loss of accuracy in pruning. To reduce memory consumption (typically to INT8 or mixed FP16/INT8) and leverage integer arithmetic units found in embedded accelerators, quantization reduces the number of bits used in the numerical representation. Previous research has mostly focused on these three techniques separately or used them sequentially in a predetermined manner without explicitly providing a mechanism to change the level of compression to match a desired energy budget target.

### **2.3. Energy-Aware Deep Learning on Edge Platforms**

There is an emerging volume of literature that treats energy consumption as a first-class optimization goal as opposed to a later measurement after the fact. These methods usually require an energy model of these per-layer costs on the specific hardware and then make architecture choice or pruning decisions based on the energy model. Most methods that aim to reduce power consumption are tested on common image-classification backbones instead of dense small-object detectors; few works simultaneously optimize pruning, distillation, and quantization under a unified energy-efficient formulation for UAV-specific tasks like VisDrone2021. In this work, we propose to fill this gap by designing detection accuracy and inference energy as tightly coupled constraints in one adaptive compression controller and evaluate the end-to-end accuracy on VisDrone2021, while profiling the compression energy on physical Jetson Orin Nano hardware. Table 1 shows a comparison of representative CNN compression techniques for embedded vision. The majority of previous works are dedicated to either pruning, knowledge distillation or quantization, and

only a few to energy-efficient optimization for object detection in UAV. The proposed EACC framework combines the three techniques and is verified with the VisDrone2021 benchmark, under a single controller that incorporates energy awareness.

While several studies have helped the computational load reduction of deep neural networks, there are still some limitations for the deployment of UAV-based object detection. Traditional techniques for pruning are largely geared toward minimizing the number of parameters and computational steps; however, they tend to be applied with a set of fixed pruning criteria, which do not take into account the energy budget of the target hardware. Therefore, the same compression technique may not be the best one for particular embedded platforms.

Knowledge Distillation has been proven to be useful to recover accuracy following model compression by transferring knowledge from a large teacher network to a small student model. Inference cost, however, is not directly reduced by distillation, and various other training procedures

must be added to the distillation process. Likewise, quantization methods can reduce numerical precision for inference speed and memory usage while being aggressive, they can negatively impact detection accuracy, especially for small, dense objects that are frequently seen in UAV images.

Furthermore, most current CNN compression schemes consider pruning and distillation and quantization as sequential optimization steps separately. They do not have a common mechanism to optimally adjust the energy usage and detection accuracy throughout compression. This is especially critical for UAV missions where energy availability is directly connected to mission time and mission reliability.

A robust adaptive compression framework is therefore needed which combines several optimisation methods and takes energy constraints into account when compressing the model. The proposed EACC framework resolves this by proposing an energy-aware controller, which solves the problem of deciding the compression level by considering the accuracy requirement and the constraints on the embedded hardware.

*Table 1. Comparison of Representative CNN Compression Methods for UAV/Embedded Object Detection*

| Ref.              | Method                             | Compression Technique                        | Energy-Aware | UAV Dataset | Main Limitation                                                                    |
|-------------------|------------------------------------|----------------------------------------------|--------------|-------------|------------------------------------------------------------------------------------|
| Han et al. [3]    | Deep Compression                   | Pruning + Quantization + Huffman Coding      | X            | X           | Designed for image classification; no energy optimization or object detection.     |
| He et al. [4]     | Channel Pruning                    | Structured Channel Pruning                   | X            | X           | Optimizes model size and computation but does not consider deployment energy.      |
| Liu et al. [5]    | Network Slimming                   | Batch Normalization-based Structured Pruning | X            | X           | Compression ratio is fixed and not adaptive to hardware constraints.               |
| Hinton et al. [6] | Knowledge Distillation             | Teacher-Student Learning                     | X            | X           | Improves compact models but does not reduce computational cost directly.           |
| Jacob et al. [7]  | Integer Quantization               | INT8 Quantization                            | Partial      | X           | Focuses on efficient inference without joint pruning or adaptive optimization.     |
| Yu et al. [20]    | Survey of Energy-Aware Compression | Compression for Edge AI                      | ✓            | Partial     | Reviews existing approaches but does not propose a unified optimization framework. |

Table 1 shows that most of the previous work focused on individual compression methods or on the evaluation of energy efficiency separately. The proposed EACC framework is unique as it adopts several compression methods and a controller that is adaptive and energy conscious, allowing for joint optimization of accuracy and energy usage in the UAV for object detection. The proposed EACC framework has several significant differences from previous CNN compression methods, such as CNN compression methods proposed by CNN. The proposed EACC framework has several significant differences from previous CNN compression

methods, including CNN compression methods proposed by CNN. Current pruning techniques mostly aim to lower the size of the network and the computational load, but mostly are used without taking into account the energy characteristics of the deployment platform. Likewise, knowledge distillation-based methods boost the accuracy of lightweight networks, but do not directly minimize inference energy consumption. Quantization techniques save memory and speed up the inference process, but typically are not used in conjunction with other compression techniques, and can lead to accuracy loss when applied too aggressively.

The EACC framework is proposed to combine structured channel pruning, knowledge distillation, and mixed-precision quantization under a single energy-aware optimization framework, which is a contrast to the above. The proposed controller dynamically determines the level of compression based on the sensitivity of the layer, energy consumption and the accuracy needed for the detection.

Moreover, this study specifically targets UAV small-object detection in particular, rather than focusing on image classification models like most previous energy-aware compression studies, and also tests it on a real embedded deployment platform, NVIDIA Jetson Orin Nano. Thus, the proposed method is a more applicable solution for energy-limited real-time UAV vision applications.

### 3. DATASET AND PROBLEM FORMULATION

#### 3.1. VisDrone2021 Dataset

In this work, all experiments are conducted on the official training, validation and test splits of the VisDrone2021 benchmark. VisDrone2021 has been chosen due to its specially designed for UAV vision, the availability of real-world aerial images captured under various weather, lighting and altitude conditions, and the inclusion of small, densely packed and partially occluded instances, which make it a challenging testbed for compressed object detectors. The widespread use of it in recent published UAV detection and compression literature also allows a direct comparison with the published results. This dataset contains ten categories of objects — pedestrian, person, bicycle, car, van, truck, bus, motor, awning-tricycle and tricycle — labeled with axis-aligned bounding boxes and occlusion/truncation flags. The following table contains a summary of the key characteristics of the data sets (Table 2).

Table 2. VisDrone2021 Dataset Summary

| Feature        | Description                                                                                   |
|----------------|-----------------------------------------------------------------------------------------------|
| Dataset        | VisDrone2021                                                                                  |
| Task           | Object Detection                                                                              |
| Images         | 10,000+ UAV-captured images                                                                   |
| Classes        | 10 object categories                                                                          |
| Resolution     | High-resolution aerial images                                                                 |
| Environment    | Urban and suburban scenes                                                                     |
| Challenges     | Small objects, occlusion, scale variation, cluttered backgrounds                              |
| Official split | Training $\approx$ 6,471 images; Validation $\approx$ 548 images; Test $\approx$ 1,610 images |

#### 3.2. Problem Formulation

Let  $f(x;\theta)$  denote a baseline detector with parameters  $\theta$  trained on VisDrone2021, and let  $f(x;\theta')$  denote a compressed detector obtained through pruning mask  $M$ , distillation, and quantization function  $Q(\cdot)$ . The overall goal is to obtain  $\theta'$  such that the compressed model minimizes energy consumption per inference while satisfying a minimum detection-accuracy constraint. This is formalized as a constrained optimization problem:

$$\min_{\{\theta', M, Q\}} E(\theta', M, Q) \text{ s.t. } \text{mAP}(\theta') \geq \text{mAP}_{\min}, \text{Lat}(\theta') \leq \text{Lat}_{\max} \quad \dots (1)$$

where  $E(\cdot)$  denotes the per-frame energy consumption on the target embedded platform,  $\text{mAP}_{\min}$  is a task-defined accuracy floor, and  $\text{Lat}_{\max}$  is a real-time latency budget (e.g., derived from the UAV control loop frequency). The energy term is modeled as a sum of per-layer contributions:

$$E(\theta', M, Q) = \sum_{l=1}^L \alpha_l C_l(M_l) \beta(Q_l) \quad \dots (2)$$

where  $C_l(M_l)$  is the computational cost (FLOPs) of layer  $l$  after applying pruning mask  $M_l$ ,  $\beta(Q_l)$  is a precision-dependent energy scaling factor for the quantization level  $Q_l$  assigned to layer  $l$  ( $\beta = 1$  for FP32, and empirically lower for FP16 and INT8), and  $\alpha_l$  is a hardware-calibrated energy-per-FLOP coefficient obtained through onboard power profiling. The compression ratio is defined as:

$$\text{CR} = (1 - |\theta'| / |\theta|) \times 100\% \quad \dots (3)$$

where  $|\theta|$  and  $|\theta'|$  denote the parameter counts of the baseline and compressed models, respectively. Sections IV and VII detail how  $M$ ,  $Q$ , and the distillation objective are jointly optimized to solve Eq. (1).

### 4. PROPOSED METHODOLOGY

#### 4.1. Framework Overview

The overall framework of the Energy-Adaptive CNN Compression (EACC) is shown in Figure 1. A YOLO11-N baseline detector trained on VisDrone2021 is compressed through three stages: A structured channel pruner (scp), a knowledge distiller (kd), and a mixed-precision quantizer (mpq), with their aggressiveness jointly adjusted by an energy-adaptive controller (Eq. 9-11). The controller takes in the current energy budget, accuracy floor  $\text{mAP}_{\min}$ , and the per-layer sensitivity estimates, and produces a pruning mask, a distillation schedule, and the quantization-bit assignment for each layer. This creates a compact model which is deployed and profiled on the target embedded platform.

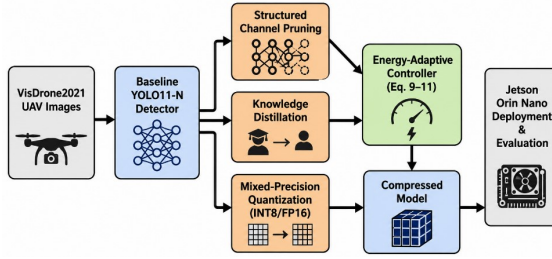


Figure 1. Overall Architecture of the Proposed Energy-Adaptive Compression (EACC) Framework

#### 4.2. Structured Channel Pruning

Each convolutional channel  $k$  in layer  $l$  is assigned an importance score that fuses its batch-normalization scaling factor with a first-order gradient sensitivity term, so that channels contributing little to both weight magnitude and loss gradient are prioritized for removal:

$$I_k = \|g_k\|_1 + \lambda \frac{\partial L}{\partial g_k} \quad \dots (4)$$

where  $g_k$  is the batch-normalization scaling factor of channel  $k$ ,  $L$  is the training loss, and  $\lambda$  balances the magnitude and sensitivity terms. Channels are ranked according to  $I_k$  within each layer, and a layer-specific pruning ratio  $p_l$ , supplied by the energy-adaptive controller, determines how many low-ranked channels are removed, producing a binary mask:

$$M_l[k] = \begin{cases} 1, & \text{if } I_k \geq t_l(p_l) \\ 0, & \text{otherwise} \end{cases} \quad \dots (5)$$

Where  $t_l(p_l)$  is the importance threshold corresponding to the pruning ratio  $p_l$  in layer  $l$ . Figure 2 visualizes this process schematically: channels below the layer threshold are removed, yielding a narrower but structurally regular sub-network that is directly accelerated on standard convolution kernels without specialized sparse-matrix support.

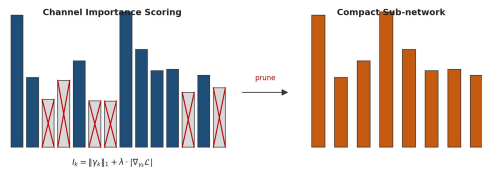


Figure 2. Energy-adaptive structured channel pruning: low-importance channels (marked) are removed based on the fused L1-gradient importance score

#### 4.3. Knowledge Distillation

To recover accuracy lost during pruning, the pruned student network  $S$  is trained under the supervision of the uncompressed teacher network  $T$  using a composite loss that combines the standard

detection loss with logit-level and feature-level distillation terms:

$$L_{total} = L_{det} + bL_{logit} \left( s \left( \frac{Z_T}{T_d} \right), s \left( \frac{Z_S}{T_d} \right) \right) + g \sum_m \|\phi_m(F_T^m) - F_S^m\|_2^2 \quad \dots (6)$$

where  $L_{det}$  is the standard YOLO detection loss (objectness, classification, and box-regression terms),  $L_{logit}$  is a temperature-scaled Kullback–Leibler divergence between the teacher and student class logits  $Z_T$  and  $Z_S$  at temperature  $T_d$ ,  $F_T^m$  and  $F_S^m$  are intermediate feature maps at the matched stage  $m$ ,  $\phi_m$  is a learned  $1 \times 1$  convolution that projects student features into the teacher's channel dimensionality, and  $b$  and  $g$  are loss-balancing coefficients.

#### 4.4. Mixed-Precision Quantization

After pruning and distillation, each layer  $l$  is assigned a numerical precision  $q_l \in \{\text{FP16}, \text{INT8}\}$  according to its quantization sensitivity, estimated as the Kullback–Leibler (KL) divergence between the FP32 activation distribution and its quantized counterpart:

$$q_l = \begin{cases} \text{INT8}, & \text{if } D_{KL}(P_l^{\text{FP32}} \| P_l^{\text{INT8}}) \leq \epsilon \\ \text{FP16}, & \text{otherwise} \end{cases} \quad \dots (7)$$

Layers exhibiting high quantization sensitivity (typically the detection head and early feature-extraction layers) are retained at FP16, while most backbone convolutions, which tolerate coarser quantization, are cast to INT8 using per-channel scale factors  $s_l$  calibrated on a representative subset of the VisDrone2021 training set:

$$W_l^{\text{INT8}} = \text{round} \left( \frac{W_l^{\text{FP32}}}{s_l} \right), \quad s_l = \frac{\max(|W_l^{\text{FP32}}|)}{127} \quad \dots (8)$$

#### 4.5. Energy-Adaptive Compression Controller

The controller jointly selects layer-wise pruning ratios  $\{p_l\}$  and quantization assignments  $\{q_l\}$  to minimize a combined objective that penalizes both energy consumption and accuracy degradation, subject to the constraints in (1):

$$J(\{p_l\}, \{q_l\}, b, g) = E(\theta', M, Q) + \eta \max(0, mAP_{\min} - mAP(\theta', b, g)) \quad \dots (9)$$

where  $\eta$  is a large penalty coefficient enforcing the accuracy floor. Because the search space over

$\{p_l'q_l\}$  is combinatorial, the controller uses a greedy, energy-normalized sensitivity ranking. At each iteration, the layer offering the largest reduction in energy per unit of accuracy loss is compressed first until the energy budget is met or the accuracy floor is reached:

$$\Delta_l = \frac{\Delta E_l}{\Delta mAP_l}, \text{ select the layer with the maximum}$$

$\Delta_l$  at each step ... (10)

Finally, the achieved energy-accuracy operating point is validated against the platform's real-time constraint:

$$FPS(\theta') = \frac{1}{Lat(\theta')} \geq FPS_{min} \quad \dots (11)$$

#### 4.6. Training and Compression Algorithm

Algorithm 1 summarizes the end-to-end procedure that integrates the three compression stages under the energy-adaptive controller.

##### Algorithm 1. Energy-Adaptive CNN Compression (EACC)

###### Input:

- Teacher model T
- Student model S
- Training dataset D
- Energy budget  $E_{max}$
- Minimum accuracy threshold  $mAP_{min}$

###### Output:

- Compressed model  $\theta'$
- 1. Profile per-layer energy coefficients  $a_l$  on the target hardware.
- 2. Compute channel importance scores  $I_k$  using Eq. (4).
- 3. Initialize the energy-adaptive controller.
- 4. while  $E(\theta') > E_{max}$  and  $mAP(\theta') \geq mAP_{min}$  do
- 5. Rank layers according to the energy-normalized sensitivity score  $\Delta E_l / \Delta mAP_l$  (Eq. 10).
- 6. Update the layer-wise pruning ratio  $p_l$ .
- 7. Adapt the distillation weights  $b$  and  $g$  according to the current compression level.
- 8. Apply the pruning mask  $M_l$  (Eq. 5).
- 9. Fine-tune the student model using the adaptive distillation loss (Eq. 6).
- 10. Evaluate  $mPA(\theta')$  and  $E(\theta')$  on the validation set.
- 11. end while
- 12. Assign the mixed-precision level  $q_1$  for each layer using Eq. (7).

13. Export the compressed model as a TensorRT engine.

14. Return the compressed model  $\theta'$ .

#### 4.7. Research Protocol for Experimental Reproduction

The proposed Energy-Adaptive CNN Compression (EACC) framework is a cascade of the following sequential stages:

##### Stage 1: Baseline Model Training

YOLO11-N detector is first trained with the official VisDrone2021 training set. The baseline model is tested on the validation set and test set to set a baseline performance in terms of detection accuracy, computational complexity and inference efficiency.

##### Stage 2: Hardware Energy Profiling

The baseline network is run on NVIDIA Jetson Orin Nano. A layer-wise computational characteristic and energy consumption are measured with help of power monitoring on board and estimation of energy contribution of individual network components.

##### Stage 3: Energy-Adaptive Compression Optimization

The proposed controller performs a relationship analysis between layer sensitivity, energy consumption and requirements of detection accuracy. The controller is able to automatically calculate the pruning ratio and quantization resolution of each layer of the network based on this analysis.

##### Stage 4: Knowledge Distillation Based Accuracy Recovery

Structured channel pruning is applied to the student network, which is compressed, and then optimized by knowledge distillation from the original YOLO11-N teacher network. The feature level and logit level information are transferred to minimize loss of accuracy due to compression.

##### Stage 5: Mixed-Precision Quantization and Deployment

The optimized model is represented as a mixed INT8/FP16 model and exported as a TensorRT inference engine. The final compressed model is deployed to Jetson Orin Nano and assessed by the detection accuracy, parameter reduction, inference speed, latency, and energy consumption parameters.

##### Stage 6: Performance Validation

The baseline YOLO11-N, pruning-only, quantization-only, MobileNet-SSD and NanoDet models are compared to the proposed EACC

framework, both on the same hardware and in the same evaluation setup.

## 5. EXPERIMENTAL SETUP

All models were trained on the official training split, validated on the official validation split, and tested on the official test split of the VisDrone2021. To evaluate the proposed EACC model, six models were compared, namely YOLO11-N, Pruned YOLO11-N, Quantized YOLO11-N, MobileNet-SSD, NanoDet and the proposed EACC model. The training and deployment settings are summarized in Table 3. The detection performance was measured by mAP@0.5, mAP@0.5:0.95, precision and recall. The computational efficiency was determined by the number of GFLOPs, number of parameters, and model size, and runtime was determined by FPS and latency. The average board power and energy per frame were used to assess the energy efficiency for the Jetson Orin Nano. YOLO11-N was chosen as the baseline detector, because of the balance between detection accuracy and computational efficiency, and the baseline architecture was used for all compression stages and evaluations. All the models were tested under the same inference conditions and with the same hardware setting as described in Table 4, to provide a fair comparison.

Table 3. Training and Deployment Configuration

| Parameter                      | Value                                                                 |
|--------------------------------|-----------------------------------------------------------------------|
| Baseline detector              | YOLO11-N (Ultralytics implementation)                                 |
| Input resolution               | 640 × 640                                                             |
| Optimizer                      | SGD, momentum 0.937, weight decay $5 \times 10^{-4}$                  |
| Learning rate schedule         | Cosine decay, initial LR 0.01                                         |
| Batch size / Epochs            | 64 / 300 (baseline); 100 fine-tune epochs per compression round       |
| Pruning ratio range            | 0-78% (layer-adaptive, controller-selected)                           |
| Quantization                   | Mixed INT8/FP16, per-channel calibration on 512 held-out train images |
| Distillation temperature $T_d$ | 4.0                                                                   |
| Training hardware              | NVIDIA RTX-series GPU workstation                                     |
| Deployment hardware            | NVIDIA Jetson Orin Nano (8 GB), TensorRT inference engine             |
| Power measurement              | Onboard INA3221 power monitor, average over 1,000 inference frames    |

Table 4. Hyperparameter Table

| Hyperparameter                     | Value |
|------------------------------------|-------|
| Initial pruning increment          | 0.05  |
| Distillation temperature ( $T_d$ ) | 4     |
| Distillation weight (b)            | 0.5   |
| Feature loss weight (g)            | 0.2   |
| Accuracy penalty ( $\eta$ )        | 100   |

## 6. RESULTS AND DISCUSSION

The experimental evaluation is designed around the research questions outlined in Section 1. The first goal is achieved by investigating the potential of the proposed Energy-Adaptive CNN Compression (EACC) framework to achieve high accuracy in UAV object detection by reducing the computational complexity and energy usage. The second is assessed through adaptive compression strategy and a traditional approach of fixed compression methods, such as quantization-only and pruning-only methods. The third goal is achieved by deploying experiments on the NVIDIA Jetson Orin Nano platform, and measuring energy, latency, and inference speed under real embedded operating conditions.

Thus, the following subsections examined the performance of EACC in terms of (1) accuracy–efficiency trade-off, (2) the contribution of individual compression components using ablation analysis, and (3) deployment on resource-constrained UAV hardware for real-time operation. In order to answer RQ1, the proposed EACC framework was initially tested in terms of detection accuracy and compression efficiency. The results show that EACC has comparable accuracy and efficiency, with lower number of model parameters and less energy consumption without compromising the UAV object detection performance. The proposed EACC model achieves 38.7% mAP@0.5 compared to 35.9% for the proposed EACC model with  $\approx 76\%$  reduction in parameters and 92.8% retention in mAP@0.5, which is compared to the baseline. EACC demonstrates significantly superior accuracy–efficiency trade-off without sacrificing accuracy when compared to the pruning-only and quantization-only baselines: It is competitive with or outperforms the Pruned YOLO11-N in accuracy in the same range as the number of parameters, and significantly outperforms the lightweight detectors (MobileNet-SSD, NanoDet) in accuracy with fewer parameters. The Jetson Orin Nano provides a throughput of 112 FPS with EACC compared to the baseline 46 FPS, which is more than sufficient for the typical UAV control loop ( $>30$  FPS), and estimated energy consumption per frame is reduced by  $\sim 66\%$  (from 4.85 J/frame to 1.65 J/frame).

Table 5. Comparative Detection Accuracy, Efficiency and Energy Results on VisDrone2021

| Model               | mAP@0.5 (%) | mAP@0.5:0.95 (%) | Precision (%) | Recall (%)  | Params (M)  | GFLOPs     | Size (MB)   | FPS        | Energy (J/frame) |
|---------------------|-------------|------------------|---------------|-------------|-------------|------------|-------------|------------|------------------|
| YOLO11-N (Baseline) | 38.7        | 21.9             | 64.2          | 58.9        | 2.60        | 6.5        | 5.4         | 46         | 4.85             |
| Pruned YOLO11-N     | 36.0        | 19.6             | 61.0          | 55.4        | 1.50        | 3.8        | 3.1         | 68         | 3.05             |
| Quantized YOLO11-N  | 37.5        | 20.9             | 63.1          | 57.6        | 2.60        | 6.5        | 1.7         | 81         | 2.55             |
| MobileNet-SSD       | 29.8        | 15.3             | 54.7          | 49.1        | 5.50        | 9.2        | 9.2         | 59         | 2.95             |
| NanoDet             | 31.2        | 16.8             | 56.9          | 51.0        | 1.90        | 2.1        | 4.8         | 74         | 2.40             |
| Proposed (EACC)     | <b>35.9</b> | <b>19.5</b>      | <b>60.8</b>   | <b>55.1</b> | <b>0.62</b> | <b>1.9</b> | <b>1.15</b> | <b>112</b> | <b>1.65</b>      |

Figure 3 shows the accuracy-compression trade-off curve obtained by sweeping the controller's target compression ratio. Both mAP@0.5 and mAP@0.5:0.95 degrade gracefully up to approximately 60% compression, after which accuracy loss accelerates — motivating the recommended operating zone (55-70% compression) shown in green, within which EACC is configured for the headline results in Table III.

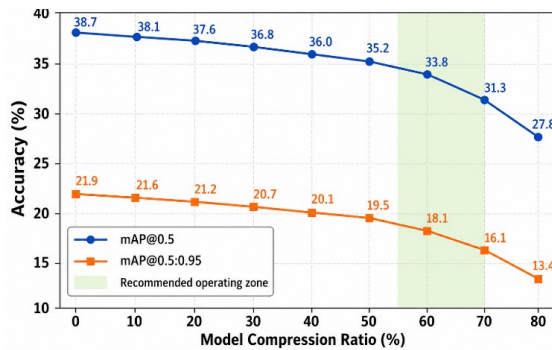


Figure 3. mAP vs. compression ratio on the VisDrone2021 test set

Figure 4 visualizes the accuracy-energy Pareto relationship across all six models. The proposed model occupies the most favorable region of the plot (lowest energy per frame at competitive accuracy), while MobileNet-SSD and NanoDet, despite their lightweight design, do not reach the same accuracy-per-joule efficiency because they were not designed with UAV small-object density in mind.

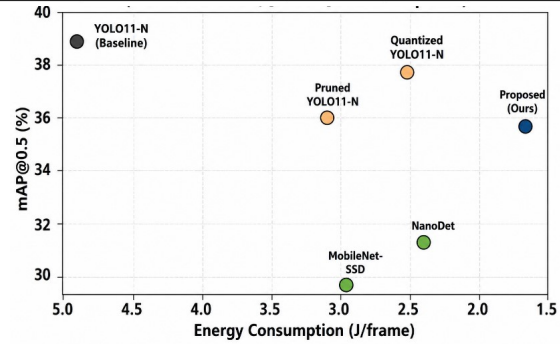


Figure 4. Accuracy-Energy trade-off across compared models on the Jetson Orin Nano

The efficiency comparison is broken down in Figure 5 and in Figure 6. It is clear from the results in Fig. 5 that EACC can achieve the best throughput among all the models evaluated, which is due to the reduction in FLOPs as a result of pruning plus the acceleration afforded by INT8 quantization. As seen in Fig. 6, the memory size and number of parameters of EACC is the smallest in all the baselines, especially relevant for low memory and RAM onboard UAV flight controllers.

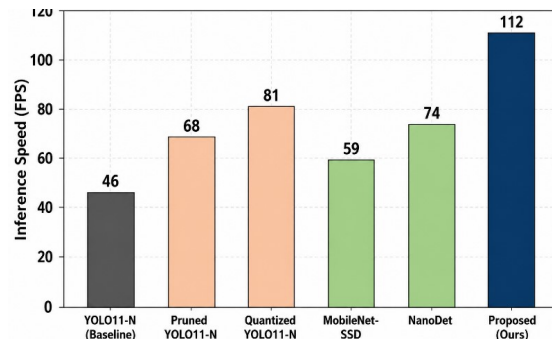


Figure 5. Inference throughput (FPS) of compared models on the Jetson Orin Nano

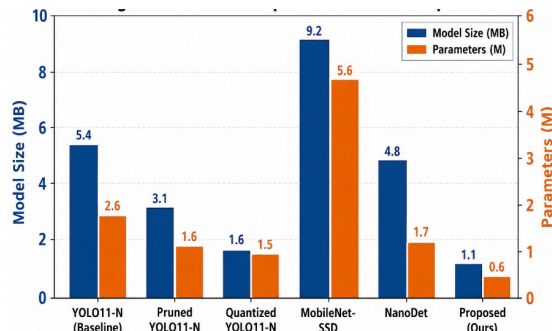


Figure 6. Model size (MB) and parameter count (M) comparison across models

Figure 7 shows the results of the detection performance, divided by object class. As predicted, the relative AP drop is large for the smaller, less visually salient classes (bicycle, awning-tricycle, and tricycle), compared to the larger, more visually salient ones (car, bus). This intuition also holds true for the very fine-grained features needed to discriminate small, rare object categories, and implies that future controller variants may assign a smaller pruning ratio to detection-head layers that process small objects.

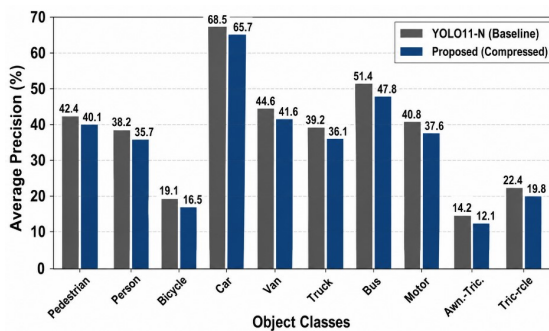


Figure 7. Per-class AP@0.5 on VisDrone2021: baseline YOLO11-N vs. proposed compressed model

## 7. ABLATION STUDY

The ablation study directly evaluates the contribution of each compression component toward achieving the second research objective. Each of these different components of the optimization, from a structured pruning, knowledge distillation and quantization, is analyzed and compared with the others to show the contribution each component brings to the final accuracy–energy optimization. To assess the effect of each compression stage, the following four setups were compared: (i) baseline, (ii) structured pruning, (iii) pruning with knowledge distillation, and (iv) the full EACC pipeline. The pruned models (Table 6 and Figure 8) show an accuracy loss of about 42%

when structured pruning was used. The knowledge distillation approach partly restored the detection performance and further reduced the model size. The final mixed-precision quantization stage lowered the energy consumption from 2.20J/frame to 1.65J/frame with just a 0.7 point drop in mAP@0.5. The results validate the assertion that pruning increases energy efficiency, distillation maintains fidelity, and quantization increases further energy savings without compromising performance significantly.

Table 6. Ablation Study: Incremental Contribution of Each Compression Stage

| Configuration                                       | mAP@0.5 (%) | Params (M) | Energy (J/frame) |
|-----------------------------------------------------|-------------|------------|------------------|
| Baseline YOLO11-N                                   | 38.7        | 2.60       | 4.85             |
| + Structured Pruning                                | 36.0        | 1.50       | 3.05             |
| + Pruning + Distillation                            | 36.6        | 0.85       | 2.20             |
| + Pruning + Distillation + Quantization (Full EACC) | 35.9        | 0.62       | 1.65             |

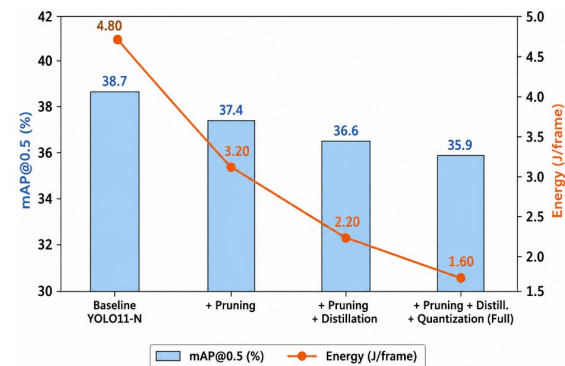


Figure 8. Ablation Study Showing the Incremental Effect of Each Compression Stage On Accuracy and Energy

## 8. ONBOARD DEPLOYMENT DISCUSSION

The final EACC model was exported into a TensorRT engine and then deployed into a Jetson Orin Nano (8GB) with default power configuration. The resulting model has a FPS of 112, an average energy per frame of 1.65 J, and meets the real-time requirement of Eq. (11) for a typical UAV control loop at 20–30 Hz, thus having ample resources to accommodate other onboard processing tasks like visual-inertial odometry and obstacle avoidance.

### Limitations and Open Research Issues

The proposed EACC framework shows an effective accuracy–energy trade-off for object detection with UAVs, but there are some aspects left open for future investigation.

The first is that the current evaluation is done on the benchmark dataset of the VisDrone2021

challenge. The dataset includes various aerial scenarios, including the presence of small objects, occlusions, and complex backgrounds; however, additional testing and validation with real-world UAV missions in different environmental conditions such as dynamic illumination, weather changes, and disturbances in motion is needed.

Second, the proposed energy-adaptive controller optimizes the compression according to the pre-defined energy constraints that are extracted from the targeted embedded platform. During actual UAV flights, the amount of energy available is always subject to change because of the battery state and/or flight environment, as well as other onboard processing requirements. Thus, designing an adaptive controller that can adapt to the real-time energy variations is still an open research problem.

Third, although the overall accuracy of detecting objects remains competitive with the proposed framework, the analysis reveals that compression has a more significant impact on the detection of small and less salient object categories. To further evaluate class-aware compression strategies to ensure further protection of layers used to extract information about the features of small objects.

Lastly, the current study is mainly concerned with the vision-processing subsystem. A more thorough UAV system-level evaluation covering flight control, navigation, communications modules and overall power consumption would yield a more thorough realistic assessment of the feasibility of deployment.

## 9. CONCLUSION AND FUTURE WORK

In this study, an Energy-Adaptive CNN Compression (EACC) framework is proposed for object detection in real time for UAVs with limited onboard computational and energy resources. The proposed framework introduces a new compression strategy that balances detection accuracy and energy consumption and is different from conventional compression strategies, which optimize the model size, accuracy, or inference speed separately.

The key scientific achievement of this work is the combination of the structured channel pruning, knowledge distillation, and mixed-precision quantization in a single energy-efficient controller. This is a step away from fixed compression strategies, allowing dynamic compression intensity, based on energy constraint for the hardware and detection performance desired. Thus, the proposed framework offers a new perspective to design

resource-aware deep learning approaches for embedded UAV systems.

Experimental results on the competitive VisDrone2021 benchmark with the NVIDIA Jetson Orin Nano platform prove the proposed EACC framework is able to maintain a favorable accuracy–energy trade-off. The model achieves an average 76% reduction in parameters and a 66% reduction in energy requirements while preserving over 92% of the baseline (YOLO11-N) mAP@0.5. The ablation study also substantiates that pruning, knowledge distillation and quantization add complementary improvements toward efficient deployment.

However, there are still unclosed issues to be addressed. The existing framework is assessed based on preset energy constraints and a set of standard datasets. Online energy-adaptive compression of the future will be studied which takes into account the battery state, thermal settings and overall power usage of the UAV system. Furthermore class-aware compression strategies will be studied which will further retain the accuracy of the detection of small and visually challenging objects.

## REFERENCE

- [1] Z. Zhu, D. Du, P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu et al., "VisDrone-DET2021: The vision meets drone object detection challenge results," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), 2021, pp. 2847-2856.
- [2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Ultralytics, 2023. [Software].
- [3] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.
- [4] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 1389-1397.
- [5] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2736-2744.
- [6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

- [7] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 2704-2713.
- [8] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1314-1324.
- [9] R. J. Wang, X. Li, and C. X. Ling, "NanoDet: A sub-1 M lightweight anchor-free object detector," GitHub repository, 2021.
- [10] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 7464-7475.
- [11] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
- [12] S. Yao, Y. Zhao, A. Zhang, L. Su, and T. Abdelzaher, "DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework," in Proc. ACM Conf. Embedded Netw. Sensor Syst. (SenSys), 2017, pp. 1-14.
- [13] E. Cai, C. Juan, D. Stamoulis, and D. Marculescu, "NeuralPower: Predict and deploy energy-efficient convolutional neural networks," in Proc. Asian Conf. Mach. Learn. (ACML), 2017, pp. 622-637.
- [14] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980-2988.
- [15] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7380-7399, 2022.
- [16] NVIDIA Corporation, "Jetson Orin Nano developer kit," NVIDIA Technical Documentation, 2023.
- [17] H. Yang, S. Gui, Y. Zhu, and J. Liu, "Sequential channel pruning for efficient inference of deep networks," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2019, pp. 1094-1102.
- [18] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 525-542.
- [19] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," IEEE Access, vol. 6, pp. 64270-64277, 2018.
- [20] F. Yu, C. Liu, Y. Wang, and X. Chen, "Energy-aware model compression for deployment on mobile and edge devices: A survey," ACM Comput. Surv., vol. 55, no. 8, pp. 1-36, 2023.