JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 4, pp. 1-9, April 2026
https://doi.org/10.9708/jksci.2026.31.04.001

Low-Cost Edge AI Accelerators for Real-Time Object Detection: A Comparative Analysis of Inference Performance and Cost Efficiency

Pil-Seong Jeong*

*Professor, Dept. of Information and Communication Engineering, Myongji College, Seoul, Korea

[Abstract]

As the edge AI accelerator market rapidly expands, selecting optimal hardware from diverse heterogeneous platforms for real-time object detection has become a critical challenge. However, TOPS (Tera Operations Per Second) figures provided by manufacturers represent only theoretical maximums and fail to reflect real-world performance. This study empirically compares inference performance and cost efficiency for four low-cost heterogeneous edge AI accelerators: Jetson Orin Nano Super, Jetson Orin NX, Hailo-8 M.2, and Rockchip RK3588, using YOLO-series object detection models. Experimental results demonstrate that Hailo-8 achieved 101.2 FPS for YOLOv8s, approximately 4.6 times faster than Jetson platforms. RK3588 achieved 34.6 FPS for YOLOv8n, outperforming Orin NX and Orin Nano Super. In cost efficiency (measured as FPS per USD), RK3588 and Hailo-8 showed 2.2-7.8 times better performance than Jetson platforms. Notably, YOLOv10n with NMS-free architecture exhibited poor performance on RK3588 NPU, highlighting the importance of model-accelerator compatibility. This study provides a cost-efficiency metric framework for edge AI platform selection.

▸ **Key words**: Edge AI, Heterogeneous Accelerator, Object Detection, YOLO, Benchmarking, Cost Efficiency

[요 약]

엣지 AI 가속기 시장이 급성장함에 따라 실시간 객체탐지를 위한 최적의 이기종 플랫폼 선택이 중요한 과제가 되었다. 그러나 제조사가 제공하는 TOPS 수치는 이론적 최대치로, 실제 성능을 반영하지 못한다. 본 연구에서는 저비용 이기종 엣지 AI 가속기 4종(Jetson Orin Nano Super, Jetson Orin NX, Hailo-8 M.2, Rockchip RK3588)을 대상으로 YOLO 계열 객체탐지 모델의 추론 성능과 비용 효율성을 비교 분석하였다. 실험 결과, YOLOv8s 기준 Hailo-8이 101.2 FPS로 Jetson 대비 약 4.6배 빠른 성능을 보였으며, YOLOv8n 기준 RK3588이 34.6 FPS로 Orin NX와 Orin Nano Super를 능가하였다. 비용 효율성(달러당 FPS)에서 RK3588과 Hailo-8이 Jetson 대비 2.2-7.8배 우수하였다. 특히 NMS-free 아키텍처의 YOLOv10n이 RK3588 NPU에서 13.0 FPS의 저조한 성능을 보여, 모델-가속기 호환성의 중요성을 확인하였다. 본 연구는 엣지 AI 플랫폼 선택을 위한 비용 효율성 지표 프레임워크를 제시한다.

▸ **주제어**: 엣지 AI, 이기종 가속기, 객체탐지, YOLO, 벤치마킹, 비용 효율성

• First Author: Pil-Seong Jeong, Corresponding Author: Pil-Seong Jeong
*Pil-Seong Jeong (ibetter.kr@gmail.com), Dept. of Information and Communication Engineering, Myongji College
• Received: 2026. 02. 02, Revised: 2026. 03. 18, Accepted: 2026. 03. 23.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

딥러닝 기술의 비약적인 발전과 함께 실시간 추론이 필요한 애플리케이션이 급증하면서, 클라우드 기반 처리의 한계가 부각되고 있다[1][2]. 자율주행, 스마트 팩토리, 지능형 영상 분석 등의 응용 분야에서는 네트워크 지연, 대역폭 제약, 데이터 프라이버시 문제를 해결하기 위해 엣지 단에서의 AI 추론이 필수적으로 요구된다[3][4]. 특히 YOLO 계열 객체탐지 모델은 실시간 영상 분석의 핵심 기술로 자리 잡았으며, 이를 엣지 디바이스에서 효율적으로 실행하기 위한 하드웨어 플랫폼 선택이 중요한 과제로 대두되고 있다.

이러한 수요에 대응하여 NVIDIA, Rockchip, Hailo 등 주요 반도체 기업들은 저전력 환경에서 높은 연산 성능을 제공하는 이기종 엣지 AI 가속기를 경쟁적으로 출시하고 있다[5]. 특히 2024년에는 NVIDIA Jetson Orin Nano Super($249)와 같은 고성능 저가 제품이 출시되어 엣지 AI의 진입 장벽이 크게 낮아졌다. 그러나 이기종 엣지 AI 가속기의 급증은 연구자와 개발자들에게 플랫폼 선택의 어려움을 가중시키고 있다.

현재 엣지 AI 플랫폼 선택에는 다음과 같은 문제점이 존재한다. 첫째, 제조사가 제시하는 TOPS(Tera Operations Per Second) 수치는 이론적 최대치에 불과하며, 실제 워크로드 성능과 상관관계가 낮다[6][7]. 예를 들어, 67 TOPS를 표방하는 가속기와 26 TOPS 가속기가 실제 객체탐지 작업에서는 역전된 성능을 보일 수 있다. 둘째, GPU와 NPU는 각기 다른 아키텍처와 소프트웨어 스택을 사용하므로, 이기종 플랫폼 간 공정한 비교 기준이 부재하다[8]. 셋째, 하드웨어 도입 비용 대비 성능에 대한 체계적 분석이 부족하다[9].

기존 연구들은 엣지 디바이스 벤치마킹에 기여했으나, 몇 가지 한계가 있다. Abdulkadhim과 Repas[6]는 자동화된 벤치마킹 프레임워크를 제안했으나 비용 효율성 분석이 포함되지 않았다. YOLOBench[7]는 550개 이상의 YOLO 모델을 평가했으나 NPU 가속기를 포함하지 않았다. Alqahtani 등[8]은 Raspberry Pi와 Jetson을 비교했으나 최신 NPU 플랫폼(Hailo, RK3588)은 제외되었다. 즉, 저비용 GPU(Jetson)와 NPU(Hailo, RK3588)를 동일 조건에서 비교한 비용 효율성 연구가 부족한 실정이다.

본 연구에서는 이러한 연구 갭을 해결하기 위해 저비용 이기종 엣지 AI 가속기 4종을 대상으로 추론 성능과 비용 효율성을 실증적으로 분석한다. 본 연구의 주요 기여는 다음과 같다.

- 저비용 이기종 가속기 4종(Jetson Orin NX, Orin Nano Super, Hailo-8, RK3588)의 YOLO 기반 객체탐지 성능을 동일 조건에서 비교 분석
- 비용 효율성 지표(달러당 FPS, TOPS당 FPS)를 활용한 플랫폼 선택 프레임워크 제시
- TOPS 수치와 실제 성능 간 괴리를 통계적으로 검증
- NMS-free 구조 등 최신 모델 아키텍처의 가속기별 호환성 분석

본 논문의 나머지 구성은 다음과 같다. 2장에서는 엣지 AI 가속기 아키텍처와 관련 연구를 검토한다. 3장에서는 실험 환경 및 결과를 제시하고, 4장에서는 결론을 맺는다.

II. Related Works

1. Edge AI Accelerator Architectures

엣지 AI 가속기는 아키텍처에 따라 크게 GPU 기반과 NPU 기반으로 분류된다. 각 아키텍처는 서로 다른 설계 철학을 기반으로 하며, 이로 인해 동일한 워크로드에서도 상이한 성능 특성을 보인다.

1.1 GPU-based Accelerator

NVIDIA Jetson 시리즈는 엣지 AI 시장의 사실상 표준으로 자리 잡았다[1]. CUDA 코어와 Tensor 코어를 결합하여 다양한 정밀도(FP32/FP16/INT8)의 연산을 지원하며, TensorRT 추론 엔진을 통해 모델 최적화가 가능하다. 특히 풍부한 개발 생태계와 문서화로 인해 개발 생산성이 높다는 장점이 있다. 본 연구에서는 Jetson Orin Nano Super(67 TOPS)와 Orin NX(100 TOPS)를 비교 대상으로 선정하였다.

1.2 NPU-based Accelerator

NPU(Neural Processing Unit)는 신경망 연산에 특화된 전용 가속기로, GPU 대비 높은 전력 효율을 특징으로 한다[5].

Hailo-8은 데이터플로우 아키텍처 기반으로 26 TOPS를 2.5W 이하에서 달성하여 뛰어난 전력 효율(TOPS/W)을 보인다. M.2 폼팩터로 Raspberry Pi 5에 장착 가능하나, 전용 컴파일러(Dataflow Compiler)를 통해 생성된 HEF 파일만 실행 가능하다는 제약이 있다.

Rockchip RK3588은 ARM big.LITTLE 아키텍처 기반 SoC로, 내장 NPU가 6 TOPS를 제공한다. $139의 저렴한 가격이 장점이나, RKNN SDK의 모델 호환성이 Jetson TensorRT 대비 제한적이다[10].

2. Comparison with Prior Studies

엣지 디바이스 벤치마킹에 관한 연구는 지속적으로 진화해 왔다. 초기 연구들은 주로 단일 플랫폼 또는 동일 아키텍처 내에서의 성능 비교에 집중하였으나, 최근에는 이기종 플랫폼 간 비교로 연구 범위가 확장되고 있다.

Abdulkadhim과 Repas[6]는 이기종 엣지 디바이스를 위한 자동화된 벤치마킹 프레임워크 SHEAB를 제안하였다. 이 연구는 Jetson과 Raspberry Pi에서 YOLO, ResNet 등의 모델을 자동으로 평가할 수 있는 파이프라인을 구축했으나, 비용 효율성 분석은 포함되지 않았으며 NPU 기반 가속기는 평가 대상에서 제외되었다. YOLOBench[7]는 550개 이상의 YOLO 기반 모델을 4개의 임베디드 하드웨어 플랫폼에서 체계적으로 평가한 대규모 벤치마킹 연구이다. 다양한 YOLO 변형 모델의 성능을 비교했다는 점에서 의의가 있으나, NPU 플랫폼은 Google Coral TPU에 한정되었고 Hailo나 Rockchip과 같은 최신 NPU는 포함되지 않았다. Morales-García 등[8]은 YOLOv8의 다양한 변형(n, s, m, l, x)을 NVIDIA Jetson Orin NX에서 TensorRT를 활용하여 체계적으로 평가하였다. FP16/FP32 정밀도에서의 추론 속도와 정확도를 분석했으나, 단일 GPU 플랫폼에 한정되었으며 NPU 가속기와의 비교나 비용 효율성 분석은 포함되지 않았다. Ogden 등[9]은 비용 효율성 관점에서 이기종 플랫폼을 분석한 선구적 연구로, "Bang for the Buck" 개념을 도입하여 달러당 성능을 평가하였다. 그러나 이 연구 역시 Hailo-8, RK3588과 같은 최신 NPU 플랫폼은 포함되지 않았으며, 모델-가속기 호환성에 대한 분석은 수행되지 않았다.

이처럼 기존 연구들은 비용 효율성 분석의 부재, 최신 NPU 플랫폼 미포함, 모델-가속기 호환성 분석 부족이라는 한계가 있다. 본 연구는 저비용 GPU(Jetson)와 NPU(Hailo-8, RK3588)를 동일 조건에서 비교하고, 비용 효율성 지표를 제시함으로써 이러한 연구 갭을 해소하고자 한다.

III. Experiments

1. Experimental Setup

1.1 Hardware Configuration

본 연구에서는 저비용 엣지 AI 가속기 4종을 비교 대상으로 선정하였다. 플랫폼 선정 기준은 다음 네 가지이다. (1)가속기 아키텍처 다양성: GPU(NVIDIA Ampere) 2종과 NPU(Hailo, Rockchip) 2종을 포함하여 이기종 비교가 가능하도록 구성하였다. (2)가격대: 엣지 배포의 비용 민감성을 고려하여 $100~$650 범위의 플랫폼으로 한정하였다. (3)시장 가용성 및 커뮤니티 지원: 2025년 기준 개발자 생태계가 활발하고 구매가 용이한 플랫폼을 선정하였다. (4)TOPS 대비 실성능 검증: 공칭 TOPS가 6~100 범위에 걸쳐 있어 TOPS와 실제 추론 성능 간 관계를 분석할 수 있다. Table 1은 실험에 사용된 플랫폼 사양을 보여준다.

Table 1. Specifications of Benchmark Edge Platforms

Platform	AI Accelerator	TOPS	Memory	Price
Orin NX	Ampere GPU	100	16GB LPDDR5	$599
Orin Nano Super	Ampere GPU	67	8GB LPDDR5	$249
RPi 5 + Hailo-8 M.2	Hailo NPU	26	8GB LPDDR4x	~$350
Orange Pi 5 Plus	Rockchip NPU	6	16GB LPDDR4x	$139

Jetson 플랫폼은 Orin NX 16GB(reComputer J4012)와 Jetson Orin Nano Super Developer Kit을 사용하였으며, 두 플랫폼 모두 NVIDIA Ampere 아키텍처 기반의 GPU를 탑재하고 있다. NPU 플랫폼으로는 Raspberry Pi 5에 Hailo-8 M.2 모듈을 장착한 구성과 Rockchip RK3588 SoC를 탑재한 Orange Pi 5 Plus를 선정하였다.

1.2 Software Configuration

각 플랫폼의 소프트웨어 환경은 Table 2와 같다. Jetson 플랫폼은 JetPack 6.2.2와 TensorRT 10.3을 사용하였으며, 기본 비교는 각 플랫폼의 권장 정밀도(Jetson FP16, NPU INT8)로 수행하였다. 추가로 Jetson 플랫폼에서 INT8 정밀도 실험을 수행하여 정밀도 차이에 따른 성능 변화를 분석하였다.

플랫폼 간 기본 비교는 각 플랫폼의 실무 배포 권장 정밀도를 기준으로 수행하였다(Jetson: FP16, RK3588/Hailo-8: INT8). 이는 실제 배포 시나리오를 반영한 것이다. 정밀도 차이가 성능 비교에 미치는 영향을 분석하기 위해, Jetson 플랫폼에서 INT8 캘리브레이션(COCO128 데이터셋, 128 이미지)을 수행하고 INT8 추론 성능과 정확도를 추가 측정하였다.

Table 2. Software Environment

Platform	SDK/Runtime	Precision	Model Format
Orin NX, Nano Super	JetPack 6.2.2, TensorRT 10.3	FP16, INT8	.engine
RPi 5 + Hailo-8 M.2	HailoRT 4.19	INT8	.hef
Orange Pi 5 Plus	RKNN-Toolkit-Lite2 2.3.2	INT8	.rknn

1.3 Benchmark Models and Measurement

벤치마크 모델로 YOLOv5n/s, YOLOv8n/s, YOLOv10n, YOLO11n을 선정하였다. YOLO 계열 모델을 선정한 이유는 (1) 엣지 환경 실시간 객체탐지의 사실상 표준이며, (2) nano/small 크기가 엣지 배포에 적합하고, (3) 주요 아키텍처 전환점을 포괄하기 때문이다. 구체적으로 YOLOv5는 anchor-based 아키텍처의 대표, YOLOv8은 anchor-free 전환의 대표, YOLOv10은 NMS-free 아키텍처의 대표, YOLO11은 최신 세대를 대표한다. YOLOv8을 중점 비교 모델로 활용한 이유는 4종 플랫폼 중 유일하게 전 플랫폼에서 테스트 가능한 모델이 YOLOv8s이며(Table 3 참조), nano와 small 두 가지 크기를 지원하여 모델 규모에 따른 성능 변화를 분석할 수 있기 때문이다. 입력 크기는 640×640, Batch Size=1 조건에서 측정하였으며, 워밍업 10회 후 100회 추론을 3회 반복하여 총 300회의 측정값을 수집하였다. 측정 환경은 실내 25±2°C에서 수행하였다.

Table 3. Model Availability by Platform

Model	Orin NX	Orin Nano Super	RPi 5 + Hailo-8 M.2	Orange Pi 5 Plus
YOLOv5n	Available	Available	Unavailable	Available
YOLOv5s	Available	Available	Unavailable	Available
YOLOv8n	Available	Available	Unavailable	Available
YOLOv8s	Available	Available	Available	Available
YOLOv10n	Available	Available	Unavailable	Available
YOLO11n	Available	Available	Unavailable	Available

단, Hailo-8 플랫폼은 전용 컴파일러(Dataflow Compiler)로 생성된 HEF 파일만 실행 가능하여 제한된 모델만 테스트 가능하였다. Table 3은 각 플랫폼에서 테스트 가능한 모델을 정리한 것이다.

Hailo-8은 Hailo Model Zoo에서 제공하는 사전 컴파일된 HEF 파일만 사용 가능하여 YOLOv8s 외 모델 테스트가 제한되었다. 따라서 본 연구의 성능 비교는 두 가지 축으로 수행하였다. (1)4종 플랫폼 직접 비교는 YOLOv8s 기준으로 수행하여 Hailo-8을 포함한 전체 플랫폼을 평가하였다. (2)3종 플랫폼 비교(Hailo-8 제외)는 YOLOv8n을 포함한 전 모델로 수행하여 모델 아키텍처별 성능 특성과 플랫폼 호환성을 분석하였다. YOLOv8n 기준 3종 비교는 연구 목표인 4종 비교를 보완하는 분석으로, nano 크기 모델의 플랫폼별 확장성과 TOPS 대비 실성능 검증에 활용하였다.

2. Inference Performance Results

2.1 Performance Comparison

Table 4와 Table 5는 각 플랫폼의 권장 정밀도 기준 YOLO 모델 추론 성능을 보여준다. YOLOv8n 기준 3종 플랫폼 비교에서 RK3588이 34.6 FPS로 가장 높은 성능을 기록하였으며, Orin NX(31.4 FPS)와 Orin Nano Super(28.0 FPS)를 능가하였다. 이는 6 TOPS의 RK3588이 100 TOPS의 Orin NX보다 높은 추론 성능을 보인 것으로, TOPS 수치와 실제 성능 간의 괴리를 보여준다.

Table 4. Jetson GPU Inference Performance (FP16, FPS)

Platform	YOLO				
	v5n	v8n	v8s	v10n	v11n
Orin NX	32.1	31.4	22.1	30.7	30.6
Orin Nano Super	28.9	28.0	19.1	27.2	27.2

Table 5. NPU Inference Performance (INT8, FPS)

Platform	YOLO				
	v5n	v8n	v8s	v10n	v11n
RPi 5 + Hailo-8 M.2	-	-	101.2	-	-
Orange Pi 5 Plus	28.1	34.6	21.0	13.0	30.6

YOLOv8s 기준 4종 플랫폼 비교에서는 Hailo-8이 101.2 FPS로 압도적인 성능을 보였다. 이는 Jetson 플랫폼(19-21 FPS) 대비 약 4.9배 빠른 수치로, Hailo-8의 데이터플로우 아키텍처가 YOLOv8s 모델에 매우 효율적으로 최적화되어 있음을 보여준다.

Fig. 1은 3종 플랫폼(Hailo-8 제외)의 모델별 FPS를 비교한 것이다. RK3588이 YOLOv8n에서 최고 성능을 기록한 반면, YOLOv10n에서는 NMS-free 구조의 NPU 비호환으로 현저히 낮은 성능을 보였다. Hailo-8은 YOLOv8s 단일 모델만 지원하여 본 그래프에서 제외하였으며, 해당 성능(101.2 FPS)은 Table 5에 제시하였다.

2.2 FP16 vs INT8 Precision Comparison

정밀도 차이가 성능에 미치는 영향을 분석하기 위해

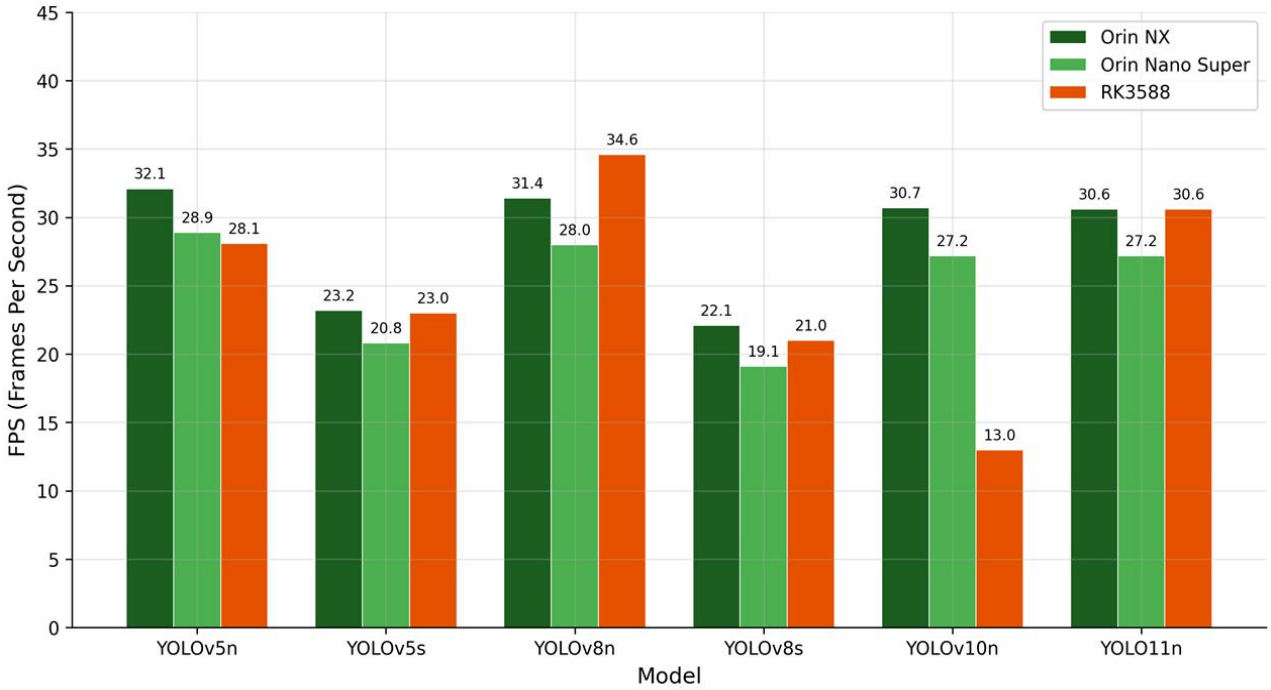

Fig. 1. YOLO Model Inference Performance by Platform (FP16)

Jetson 플랫폼에서 INT8 추론을 추가 수행하였다. Table 6과 Table 7은 각각 Orin NX와 Orin Nano Super의 FP16 대비 INT8 성능 향상을 보여준다. YOLOv10n은 NMS-free 아키텍처의 TensorRT INT8 그래프 최적화 비호환으로 INT8 엔진 빌드에 실패하여 측정하지 못하였다.

INT8 양자화를 통해 nano 모델은 5-11%, small 모델은 23-28%의 FPS 향상을 달성하였다. small 모델의 향상이 더 큰 것은 연산량이 많은 모델에서 INT8의 연산 절감 효과가 크기 때문이다. YOLOv10n은 NMS-free 아키텍처의 TensorRT INT8 그래프 최적화 비호환으로 INT8 변환에 실패하였다. Jetson INT8 적용 후에도 RK3588 대비 성능 우위는 제한적이다. YOLOv8n 기준 Orin NX INT8(33.5 FPS)은 RK3588 INT8(34.6 FPS)에 여전히 미달하였다. 이는 NPU의 비용 효율성 우위가 단순 정밀도 차이만으로는 설명되지 않으며, 아키텍처 수준의 최적화 차이에 기인함을 시사한다.

Table 6. Orin NX FP16 vs INT8 Performance (FPS)

Model	FP16	INT8	Gain
YOLOv5n	32.1	33.8	+5.3%
YOLOv5s	23.2	29.3	+26.3%
YOLOv8n	31.4	33.5	+6.7%
YOLOv8s	22.1	27.7	+25.3%
YOLOv10n	30.7	-	-
YOLO11n	30.6	33.6	+9.8%

Table 7. Orin Nano Super FP16 vs INT8 Performance (FPS)

Model	FP16	INT8	Gain
YOLOv5n	28.9	31.0	+7.3%
YOLOv5s	20.8	25.6	+23.1%
YOLOv8n	28.0	30.7	+9.6%
YOLOv8s	19.1	24.5	+28.3%
YOLOv10n	27.2	-	-
YOLO11n	27.2	30.2	+11.0%

Fig. 2는 Orin NX에서 FP16과 INT8 정밀도별 FPS를 비교한 것이다. small 모델(YOLOv5s, YOLOv8s)에서 25-26%의 큰 성능 향상이 관찰된 반면, nano 모델은 5-10% 수준의 향상에 그쳤다.

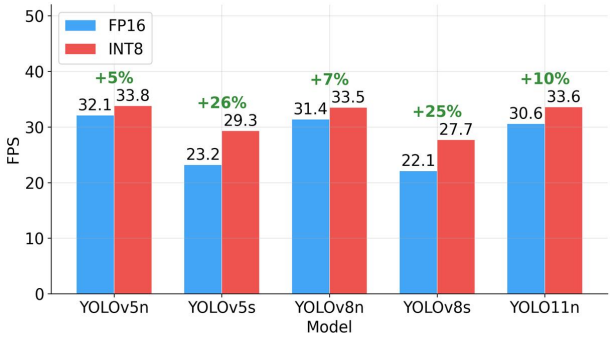

Fig. 2. FP16 vs INT8 Performance on Orin NX

2.3 Detection Accuracy (mAP)

양자화에 따른 정확도 변화를 분석하기 위해 COCO val2017 데이터셋(5,000 이미지)에서 mAP50-95를 측정하였다. Table 8은 Jetson 플랫폼의 정밀도별 mAP 결과를 보여준다.

Table 8. Detection Accuracy on COCO val2017 (mAP50-95)

Model	Orin NX		Drop	Orin Nano Super	
	FP16	INT8		FP16	INT8
YOLOv5n	28.0	25.1	-2.9	34.0	-
YOLOv5s	37.1	35.6	-1.5	42.7	-
YOLOv8n	37.3	36.0	-1.3	37.1	36.2
YOLOv8s	44.9	43.4	-1.5	44.8	-
YOLOv10n	38.5	-	-	38.3	-
YOLO11n	39.5	38.1	-1.4	39.1	-

INT8 양자화에 따른 정확도 하락은 1.3-2.9 mAP 포인트로, 일반적으로 허용 가능한 수준이다. 특히 YOLOv5n의 정확도 하락(-2.9pp)이 가장 컸는데, 이는 nano 모델의 파라미터 수가 적어 양자화에 상대적으로 민감하기 때문으로 분석된다. 같은 플랫폼(Orin NX vs Nano Super) 간 FP16 mAP는 거의 동일하여 하드웨어 차이가 정확도에 영향을 미치지 않음을 확인하였다.

Fig. 3은 Orin NX에서 FP16과 INT8 정밀도별 COCO val2017 mAP50-95를 비교한 것이다. INT8 양자화에 따른 정확도 하락은 모든 모델에서 1.3-2.9pp 범위로, 실용적으로 허용 가능한 수준임을 확인할 수 있다.

2.4 Measurement Reliability

본 실험의 반복 측정(워밍업 10회 후 100회 추론 × 3회 반복, 총 300회)은 동일 장비에서의 기술적 반복(technical repetitions)으로, 생물학적 반복(biological replicates)과는 구별된다. 따라서 결과 해석은 평균±표준편차 및 백분위수 기반의 기술통계를 중심으로 수행하며, 통계 검정은 측정의 안정성을 보조적으로 확인하는 용도로 제시한다.

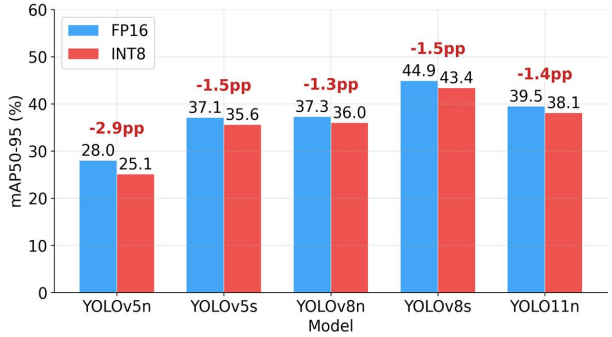

Fig. 3. FP16 vs INT8 Detection Accuracy (COCO val2017, Orin NX)

YOLOv8n 3종 플랫폼 비교에서 RK3588(34.6±1.2 FPS)은 Orin NX(31.4±0.8 FPS) 대비 3.2 FPS 높은 평균 성능을 보였다. 표준편차가 평균 차이 대비 충분히 작아 플랫폼 간 성능 차이는 측정 변동성을 상회한다(보조 검정: Welch's ANOVA, $F(2, 547.3) = 892.41$, $p < 0.001$).

YOLOv8s 4종 플랫폼 비교에서 Hailo-8(101.2±0.5 FPS)은 Jetson 플랫폼(19-21 FPS) 대비 약 80 FPS의 절대적 차이를 보였으며, 이는 측정 변동성과 무관하게 명확히 구분되는 수준이다

2.5 Latency Distribution Analysis

추론 지연시간의 분포 특성을 분석하기 위해 각 플랫폼에서 상세 지연시간 측정을 수행하였다. Table 9와 Table 10은 각각 Orin NX와 RK3588의 모델별 지연시간 분포를 비교한 것이다.

플랫폼 간 지연시간 분포에서 뚜렷한 차이가 관찰되었다. Orin NX(TensorRT)는 매우 안정적인 지연시간 분포를 보여 P95와 P50의 차이가 1-2ms에 불과한 반면, RK3588(RKNN)은 P95-P50 차이가 8-13ms로 상대적으로 넓은 분포를 보였다. 이는 TensorRT의 GPU 스케줄링이 RKNN NPU 대비 더 일관된 추론 시간을 제공함을 나타낸다.

RK3588에서 YOLOv8n이 평균 28.88ms로 가장 낮은 지연시간을 보였으나, P95 지연시간이 36.73ms로 실시간 처리 기준(30 FPS, 33.3ms)을 초과하였다. 반면 Orin NX에서는 YOLOv8n INT8의 P95가 30.62ms로 실시간 기준 내에 안정적으로 수렴하였다. YOLOv10n은 RK3588에서 평균 76.98ms로 NMS-free 구조의 NPU 비호환성이 지연시간 관점에서도 확인되었다.

Table 9. Orin NX Latency Distribution (INT8, N=300)

Model	AVG (ms)	P50 (ms)	P95 (ms)	P99 (ms)
YOLOv5n	29.55	29.44	30.31	30.37
YOLOv8n	29.85	29.78	30.62	30.91
YOLOv8s	36.13	36.30	37.19	37.53
YOLO11n	29.76	29.54	31.04	32.94

Table 10. RK3588 Latency Distribution (INT8, N=300)

Model	AVG (ms)	P50 (ms)	P95 (ms)	P99 (ms)
YOLOv5n	35.56	34.60	44.09	47.46
YOLOv8n	28.88	27.93	36.73	40.46
YOLOv8s	47.58	46.21	59.69	64.88
YOLOv10n	76.98	76.86	80.70	82.93
YOLO11n	32.67	31.57	41.22	44.81

3. Cost Efficiency Analysis

3.1 Cost Efficiency Metrics

비용 효율성 분석을 위해 두 가지 지표를 정의하였다. 첫째, 달러당 FPS(FPS per USD)는 하드웨어 도입 비용 대비 추론 성능을 나타낸다. 둘째, TOPS당 FPS(FPS per TOPS)는 공칭 연산 성능 대비 실제 활용률을 나타낸다.

Table 11은 YOLOv8s 기준 4종 플랫폼의 비용 효율성 분석 결과를 보여준다.

Table 11. Cost Efficiency Analysis (YOLOv8s, Default Precision)

Platform	Price (USD)	FPS	FPS/USD	FPS/TOPS
RPi 5 + Hailo-8	~350	101.2	0.289	3.89
Orange Pi 5 Plus	139	21.0	0.151	3.50
Orin Nano Super	249	19.1	0.077	0.29
Orin NX	599	22.1	0.037	0.22

Fig. 4는 YOLOv8s 기준 4종 플랫폼의 FPS/USD를 시각화한 것이다. NPU 기반 플랫폼(Hailo-8, RK3588)이 GPU 기반 Jetson 플랫폼 대비 현저히 높은 비용 효율성을 보이며, 특히 Hailo-8은 0.289 FPS/USD로 최고 효율을 기록하였다.

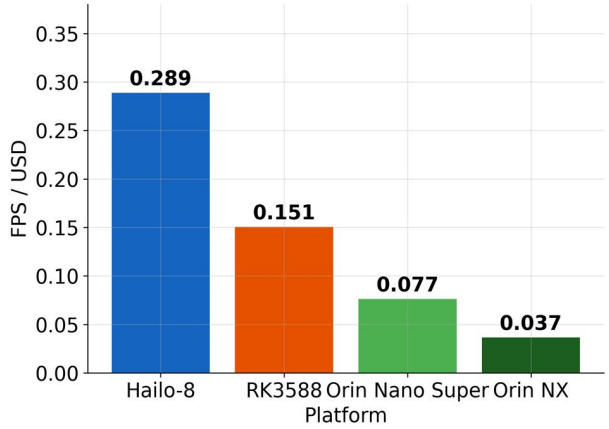

Fig. 4. Cost Efficiency Comparison – YOLOv8s (FPS per USD)

YOLOv8n 기준 3종 플랫폼(Hailo-8 제외) 비교 결과는 Table 12와 같다. Fig. 5는 YOLOv8n 기준 FPS/USD를 보여주며, RK3588이 0.249 FPS/USD로 가장 높은 비용 효율성을 보였다.

Table 12. Cost Efficiency Analysis (YOLOv8n, 3 Platforms)

Platform	Price (USD)	FPS	FPS/USD	vs Jetson
Orange Pi 5 Plus	~139	34.6	0.249	2.2-4.7x
Orin Nano Super	249	28.0	0.112	baseline
Orin NX	599	31.4	0.052	0.47x

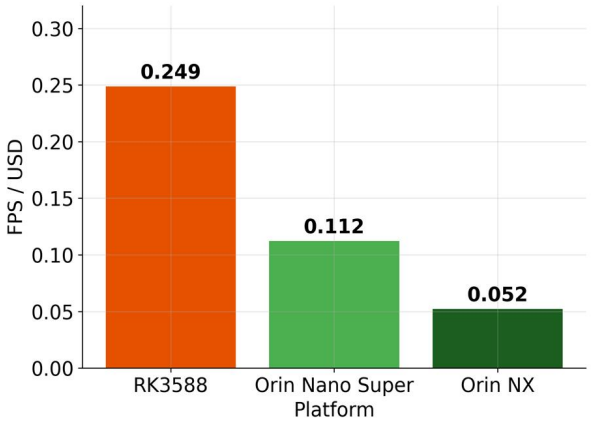

Fig. 5. Cost Efficiency Comparison – YOLOv8n (FPS per USD)

YOLOv8n에서도 RK3588이 0.249 FPS/USD로 가장 높은 비용 효율성을 보였으며, Orin Nano Super(0.112 FPS/USD) 대비 2.2배, Orin NX(0.052 FPS/USD) 대비 4.7배 우수하였다.

3.2 Analysis Results

NPU 플랫폼(Hailo-8, RK3588)이 GPU 플랫폼(Jetson) 대비 비용 효율성에서 현저히 우수한 성능을 보였다. Hailo-8은 0.289 FPS/USD로 Orin NX(0.037 FPS/USD) 대비 7.8배, RK3588은 0.151 FPS/USD로 4.1배 높은 비용 효율성을 기록하였다.

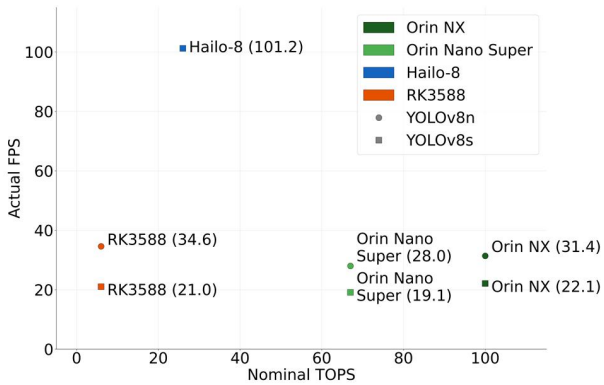

Fig. 6. Nominal TOPS vs. Actual Inference Performance

Fig. 6은 공칭 TOPS와 실제 추론 성능(FPS) 간의 관계를 산점도로 나타낸 것이다. 공칭 연산 성능이 높을수록 실제 FPS가 비례하여 증가하지 않으며, 오히려 TOPS 대비 실제 활용률(FPS/TOPS)은 저사양 NPU 플랫폼에서 더 높게 나타나는 역상관 경향을 보여준다.

특히 주목할 점은 TOPS와 실제 성능 간의 역상관 관계이다. 공칭 TOPS가 가장 높은 Orin NX(100 TOPS)가 FPS/TOPS 지표에서 가장 낮은 0.22를 기록한 반면, 가장 낮은 RK3588(6 TOPS)이 3.50으로 가장 높았다. 이는 TOPS가 이론적 최대치에 불과하며, 실제 워크로드에서는 메모리 대역폭, 소프트웨어 최적화 수준, 모델-가속기 호환성의 영향이 크기 때문으로 분석되었다.

4. Model-Accelerator Compatibility

Table 4에서 YOLOv10n의 RK3588 성능(13.0 FPS)이 다른 모델(28-35 FPS) 대비 62.4% 낮은 것을 확인할 수 있다. YOLOv10은 기존 YOLO 모델과 달리 NMS(Non-Maximum Suppression) 후처리를 제거한 NMS-free 아키텍처를 채택하고 있다. 이 구조는 end-to-end 학습이 가능하다는 장점이 있으나, RK3588 NPU의 RKNN 런타임에서 효율적으로 가속되지 않는 것으로 나타났다.

반면 Jetson 플랫폼에서는 YOLOv10n이 27-31 FPS로 다른 모델과 유사한 성능을 보였다. 이는 TensorRT가 NMS-free 구조를 효과적으로 최적화할 수 있기 때문으로 해석된다. 이러한 결과는 최신 모델 아키텍처 도입 시 가속기별 호환성 검증이 필수적임을 시사한다.

5. Platform Selection Guidelines

본 연구 결과를 바탕으로 응용 도메인별 플랫폼 선택 가이드라인을 Table 13에 제시한다.

본 가이드라인은 추론 속도 및 하드웨어 가격을 기준으로 한 1차 선별 지표이며, 실제 배포 시에는 전력 소비, 양자화에 따른 정확도 변화, 개발 생태계 성숙도, 모델 변환 제약 등을 종합적으로 고려해야 한다. 고속 처리가 필요한 경우 Hailo-8이 유력한 후보이나(YOLOv8s 기준), HEF 파일 의존성으로 모델 선택이 제한될 수 있다. 비용이 최우선인 경우 RK3588이 139로 높은 FPS/$를 보였으나, YOLOv10과 같은 최신 아키텍처와의 호환성은 사전 검증이 필요하다. 개발 유연성과 생태계가 중요한 경우 Jetson 플랫폼이 TensorRT의 광범위한 모델 지원과 풍부한 문서화로 적합하다.

IV. Conclusions

본 연구에서는 저비용 이기종 엣지 AI 가속기 4종(Jetson Orin NX, Orin Nano Super, Hailo-8, RK3588)을 대상으로 YOLO 기반 객체탐지 워크로드의 추론 성능과 비용 효율성을 비교 분석하였다. 각 플랫폼에서 300회 반복 측정(기술적 반복)을 수행하였으며, 주요 연구 결과는 다음과 같다.

첫째, 추론 성능에서 NPU 플랫폼이 특정 조건에서 GPU를 능가하였다. Hailo-8은 YOLOv8s(INT8) 단일 모델 기준으로 101.2 FPS를 기록하여 Jetson 대비 4.6배 빠른 성능을 보였다. 단, 이 결과는 Hailo Model Zoo의 사전 최적화된 HEF 파일에 기반한 것으로, 다른 모델로의 일반화에는 추가 검증이 필요하다. RK3588은 YOLOv8n에서 34.6 FPS로 Orin NX(31.4 FPS)와 Orin Nano Super(28.0 FPS)를 능가하였다.

둘째, 비용 효율성(달러당 FPS)에서 NPU 플랫폼이 GPU 대비 우수하였다. YOLOv8s 기준 4종 비교에서 Hailo-8은 0.289 FPS/USD, RK3588은 0.151 FPS/USD를 기록하여 Jetson Orin NX(0.037 FPS/USD)를 큰 폭으

Table 13. Platform Selection Guidelines by Application Domain (Speed/Cost Criteria)

Application Domain	Priority	Candidate Platform	Rationale	Additional Considerations
High-Speed Processing (100+ FPS)	Performance First	RPi 5 + Hailo-8	Highest FPS based on YOLOv8s	HEF model contraints, power verification required
Low-Budget Project	Cost First	Orange Pi 5 Plus	$139, Highest FPS/$	Model-accelerator compatibility pre-verification required
General Purpose / Development & Research	Flexiblilty	Orin Nano Super	TensorRT ecosystem	FP16/INT8 precision selectable
Industrial Use	Stability	Orin NX	16GB, industrial-grade case	High initial cost ($599)
Lastest Model Experimenation	Compatibility	Jetson Series	Full architecture support	NMS-free and other lastest architecture compatibility

로 상회하였다. YOLOv8n 기준 3종 비교(Hailo-8 제외)에서도 RK3588이 0.249 FPS/USD로 Orin Nano Super(0.112) 대비 2.2배 우수하였다.

셋째, TOPS 수치와 실제 성능 간 괴리를 실증하였다. 6 TOPS의 RK3588이 100 TOPS의 Orin NX보다 높은 FPS를 기록하였으며, TOPS당 FPS 지표에서 RK3588(3.50)이 Orin NX(0.22)보다 약 16배 높아 TOPS가 플랫폼 선택 기준으로 적절하지 않음을 확인하였다.

넷째, 모델-가속기 호환성의 중요성을 확인하였다. YOLOv10n의 NMS-free 구조가 RK3588 NPU에서 13.0 FPS로 다른 모델(28-35 FPS) 대비 62.4% 낮은 성능을 보였으며, TensorRT INT8 변환에서도 실패하여 최신 아키텍처 도입 시 가속기별 호환성 검증이 필수적임을 시사하였다.

다섯째, INT8 양자화 시 정확도 손실은 1.3-2.9 mAP 포인트로 허용 가능한 수준이며, small 모델에서 23-28%의 유의미한 성능 향상을 달성할 수 있음을 확인하였다.

본 연구의 한계점은 다음과 같다. 첫째, Hailo-8은 사전 컴파일된 HEF 파일에 의존하여 YOLOv8s 단일 모델로만 테스트되었으므로 해당 플랫폼의 성능 결론은 이 모델에 한정된다. 둘째, 전력 측정과 배치 처리 시나리오가 미포함되었다. 향후 FPS/Watt 효율성 분석, Transformer/LLM 워크로드 확장 연구가 필요하다.

본 연구에서 제시한 비용 효율성(달러당 FPS, TOPS당 FPS) 분석 프레임워크가 엣지 AI 플랫폼 선택 의사결정에 기여하기를 기대한다.

REFERENCES

[1] X. Wang, Y. Han, V. Leung, D. Niyato, X. Yan, and X. Chen, "Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models," ACM Computing Surveys, Vol. 57, No. 6, pp. 1-49, 2025.

[2] S. Bharadwaj et al., "Object Detection on Low-Compute Edge SoCs: A Reproducible Benchmark and Deployment Guidelines," Scientific Reports, Vol. 16, Article 36862, 2026, doi: 10.1038/s41598-026-36862-y.

[3] Y. Mao, X. Yu, K. Huang, Y.-J. A. Zhang, and J. Zhang, "Green Edge AI: A Contemporary Survey," Proceedings of the IEEE, Vol. 112, No. 7, pp. 879-911, 2024.

[4] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," IEEE Internet of Things Journal, Vol. 7, No. 8, pp. 7457-7469, 2020.

[5] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of Deep Learning Accelerators for Edge and Emerging AI Applications," Electronics, Vol. 13, No. 15, pp. 2988, 2024.

[6] M. Abdulkadhim and S. R. Repas, "SHEAB: A Novel Automated Benchmarking Framework for Edge AI," Technologies, Vol. 13, No. 11, pp. 515, 2025.

[7] Deeplite, "YOLOBench: Benchmarking Efficient Object Detectors on Embedded Systems," Proceedings of ICCV RCV Workshop, 2023.

[8] A. Morales-García, A. F. Skarmeta, and J. Santa, "Benchmarking YOLOv8 Variants for Object Detection Efficiency on NVIDIA Jetson Orin NX," Computers, Vol. 15, No. 2, Article 74, 2025, doi: 10.3390/computers15020074.

[9] S. Ogden, L. Xu, and T. Guo, "Bang for the Buck: Evaluating the Cost-Effectiveness of Heterogeneous Edge Platforms for Neural Network Workloads," Proceedings of IEEE/ACM Symposium on Edge Computing (SEC), pp. 51-63, 2023.

[10] M. Qian, Y. Li, X. Zhao, and H. Zhang, "Real-Time Wire Rope Detection Method Based on Rockchip RK3588," Scientific Reports, Vol. 15, Article 1, 2025.

[11] A. Tschand et al., "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI," in Proc. IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 1-14, 2025, doi: 10.1109/HPCA61900.2025.00069.

Authors

Pil-Seong Jeong received the B.S. degree in Electronic Engineering from Seoul National University of Science and Technology, Korea, in 2004, and the M.S. and Ph.D. degrees in Electronic and Communication Engineering from Kwangwoon University, Korea, in 2008 and 2013, respectively. Prof. Jeong is currently a Professor with the Department of Information and Communication Engineering, Myongji College, Seoul, Korea. His research interests include edge AI, AIoT, embedded systems, and real-time object detection on heterogeneous edge accelerators.