JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 5, pp. 1335-1343, May. 2026

Check for updates

화재 및 연기 탐지를 위한 커스텀 YOLOv8의 엣지 디바이스 환경 성능 분석

임 창 건[1] · 최 재 명[2*]
[1]목원대학교 컴퓨터융합학부 학석사과정
[2]목원대학교 컴퓨터융합학부 부교수

Performance Analysis of Custom YOLOv8 for Fire and Smoke Detection in Edge Device Environments

Chang Geon Lim[1] · Jae Myeong Choi[2*]

[1]Bachelor's and Master's Program Student, Division of Computer Convergence, Mokwon University, Daejeon 35349, Korea
[2]Associate Professor, Division of Computer Convergence, Mokwon University, Daejeon 35349, Korea

[요 약]

현대 사회의 화재는 인명과 재산에 막대한 피해를 주는 치명적인 재난으로 건축물의 고층화 및 복합 용도화에 따른 조기 감지의 중요성이 증대되고 있다. 기존 센서 기반 감지 방식의 물리적 한계를 극복하기 위해 본 논문에서는 엣지 디바이스인 NVIDIA Jetson Orin Nano를 기반으로 실시간 화재 및 연기 탐지를 수행하는 최적화된 YOLOv8 커스텀 모델을 제안하였다. 실험 결과, 제안된 모델(YOLOv8-RE2)은 mAP@0.5:0.95 기준 48.06%의 정확도를 기록하여 베이스라인인 YOLOv8-M(48.49%)과 대등한 성능을 보였으나, 구조적 복잡성으로 인해 추론 속도는 약 27.5% 낮은 13.76 FPS를 나타냈다. 결론적으로 본 연구는 엣지 환경에서 커스텀 모델의 실용 가능성을 확인하였으며, 향후 어텐션 메커니즘 고도화 등을 통한 경량화 및 속도 최적화 연구의 필요성을 시사한다.

[Abstract]

Fire is a fatal disaster that causes immense damage to life and property; therefore, its early detection has become increasingly important. To overcome the limitations of conventional sensor-based fire detection, this paper proposes an optimized custom YOLOv8(You Only Look Once) model for real-time detection of smoke and fire on the NVIDIA Jetson Orin Nano edge device. The experimental results indicated that the proposed YOLOv8-RE2 model achieved an accuracy of 48.06% in terms of mAP(Mean Average Precision)@0.5:0.95, thus performing comparably to the YOLOv8-M baseline. However, owing to structural complexity, the inference speed was 13.76 FPS, which is ~27.5% lower than the baseline. This study confirms that custom models can be applied in edge environments and suggests the necessity of investigating lightweight design and speed optimization via advanced attention mechanisms.

색인어 : 화재 탐지, 엣지 컴퓨팅, 객체 탐지, 딥러닝 최적화, 젯슨 오린 나노
Keyword : Fire Detection, Edge Computing, Object Detection, Deep Learning Optimization, Jetson Orin Nano

http://dx.doi.org/10.9728/dcs.2026.27.5.1335

Received 04 March 2026; **Revised** 17 March 2026
Accepted 30 March 2026

***Corresponding Author; Jae Myeong Choi**
Tel: ※ 개인정보 표시제한
E-mail: jmchoi@mokwon.ac.kr

Ⅰ. 서 론

현대의 화재는 여전히 가장 치명적인 재난 중 하나로서 인명과 재산에 막대한 피해를 입히고 있다. 소방청의 2024년도 화재 통계연감에 따르면, 2023년 국내에서 발생한 화재는 총 38,857건이며 이에 따라 2,204명이 부상, 284명이 사망하는 인명피해를 발생시켰으며, 약 1조 2천억의 재산 피해를 입힌 것으로 집계되었다[1]. 특히 최근 발생하는 화재는 건축물의 고층화와 복합 용도화로 인해 연기 확산 속도가 매우 빨라, 조기 감지에 실패할 경우 대규모 인명피해로 이어질 가능성이 높다. 대표적으로 2024년 8월 부천시 호텔에서 발생한 화재는 전기적 요인으로 시작되어 급격한 연기 확산으로 인해 7명이 사망하고 12명이 부상을 입는 참사로 이어졌다[1].

전통적인 열 감지기나 연기 센서는 화재가 일정 규모 이상으로 성장하거나 연기가 센서 위치에 도달해야만 반응한다는 물리적 한계가 존재한다. 이는 골든타임 확보를 어렵게 하며, 천장이 높은 대형 공간이나 개방된 실외 환경에서는 감지 효율이 저하된다는 한계점이 존재한다. 이를 개선하기 위해 최근 기존에 설치된 CCTV나 무인 항공기의 영상을 실시간으로 분석하여 화염이나 연기의 시각적 특징을 포착하는 컴퓨터 비전 기반의 지능형 화재 탐지 시스템이 주목받고 있다[2],[3].

2019년 대구 사우나 화재 사례로 알 수 있듯, 화재 징후를 초기에 시각적으로 인지하고 신속하게 대응하는 체계의 구축은 인명 피해를 최소화하는 데 있어 필수적인 요소로 간주된다[4].

또한 합성곱 신경망(CNN, Convolutional Neural Network)의 발전으로 영상 기반 화재 탐지의 정확도가 향상되었다. 그중 YOLO(You Only Look Once)는 1단계 탐지 방식(One-Stage Detection)을 채택하여 높은 정확도와 빠른 처리 속도를 동시에 제공함으로써 실시간 탐지 분야의 표준으로 자리 잡았다. 최근에는 RGB 채널 영상을 활용한 YOLOv8 경량 모델 기반의 실시간 화재 탐지 성능에 관한 연구가 진행되어 딥러닝 기반 감지 시스템의 가능성을 입증한 바 있다[5]. 그러나 모델의 성능이 고도화될수록 네트워크의 깊이가 깊어지고 파라미터 수가 증가해 연산 복잡도 또한 높아지는 경향이 있다. 이는 고성능 GPU 서버가 아닌 전력 공급과 연산 자원이 한정되어 있는 엣지 디바이스 환경에서의 지연시간 문제를 유발할 위험성이 존재한다[6]. 따라서 본 논문에서는 엣지 디바이스인 NVIDIA Jetson Orin Nano를 기반으로 높은 신뢰성과 실시간성을 동시에 확보할 수 있는 최적화된 화재 탐지 모델을 제안하고 그 성능을 검증한다.

Ⅱ. 관련 연구

2-1 YOLOv8

YOLOv8은 이전 버전인 YOLOv5와 YOLOv7의 장점을 유지하면서, Anchor-Free 탐지 방식과 새로운 Backbone 구조를 도입한 모델이다[7]. YOLOv8은 CSP(Cross Stage Partial) 구조를 개선한 C2f 모듈을 백본으로 사용하며, 경량화된 연산에서도 특징 정보의 추출을 효과적으로 수행한다. 또한 기존의 Anchor-Free 방식을 채택한 헤드 구조를 통해 앵커 박스 설정에 대한 의존성을 제거하고, 다양한 크기의 객체에 유연하게 대응할 수 있다.

손실함수 측면에서는 분류 손실(VFL)과 박스 회귀 손실(DFL + CIoU)을 결합하여 학습의 안정성을 향상시켰다. 이러한 구조적 특징은 비정형적이고 불규칙한 화재 및 연기 객체를 탐지하는 데 유리하게 작용할 것으로 기대된다.

2-2 TensorRT

엣지 디바이스는 현장에서 데이터를 직접 처리하며 통신 지연을 줄이고 데이터 보안을 강화하는 핵심 기술이다. 하지만 엣지 디바이스는 컴퓨팅 리소스가 제한적이라는 특성이 존재한다. Polenakis et al.[8]은 엣지 디바이스에서 YOLO 알고리즘의 경량화 버전들을 비교하며 그 가능성을 검토하였으나, 복잡한 환경에서의 정확도와 속도 균형에 대한 과제를 남겼다. 이러한 하드웨어의 제약을 극복하기 위해 NVIDIA TensorRT와 같은 하드웨어 가속 SDK를 활용한 연구가 필수적이다[3].

TensorRT는 딥러닝 추론을 위한 고성능 최적화 라이브러리로, Precision Calibration을 통해 기본적인 FP32(32-bit Floating Point) 연산을 FP16(16-bit)으로 변환하여 메모리 사용량을 절반 수준으로 줄이고 연산 처리량을 극대화한다. Layer & Tensor Fusion 기법을 사용하여 서로 다른 레이어를 하나의 커널로 병합하여 GPU의 메모리 접근 횟수를 줄이고 지연시간을 단축시킨다. Kernel Auto-Tuning을 통해 Jetson Orin Nano에 가장 최적화된 알고리즘과 커널을 자동으로 선택하여 하드웨어 성능을 최대한 활용한다[9].

Ⅲ. 제안 모델

표 1. 제안 모델의 구조적 특징 비교
Table 1. Comparison of structural features of the proposed model

Category	YOLOv8-M	YOLOv8-RE2
Target	80 Class	2 Class
Backbone	Standard C2f	Deep C2f, SPPF
Attention	X	CBAM, SimAM, ECA
Neck	Concat	BiFPN_Add
Head	3-Scale	4-Scale
Parameters	25.9M	63.4M

본 논문에서 제안하는 YOLOv8-RE2 모델의 구조적 변화

점은 다음 표 1과 같다. 표 1에서 알 수 있듯이, 제안 모델은 탐지 정확도 향상을 위해 백본의 깊이를 확장하고 다중 어텐션 메커니즘을 결합함으로써 파라미터 수가 베이스라인인 YOLOv8-M 모델의 25.9M 대비 약 2.4배 증가한 63.4M으로 설계되었다. 이는 엣지 디바이스의 제한된 자원 내에서 연산 부담을 초래할 수 있으나, 미세 특징 추출 성능을 극대화하기 위한 구조적 선택이다.

3-1 네트워크 아키텍처

본 논문에서 제안하는 모델의 흐름은 다음 그림 1과 같다. 이미지로부터 특징을 추출하는 특징 추출부(Backbone)에서는 핵심 블록인 C2f 모듈의 반복 연산 횟수를 기존 모델 대비 150~300% 확장하였다. 화재와 연기는 일반적인 사물(자동차, 사람 등)과 달리 정형화된 외곽선이 없고 질감이 매우 복잡하기 때문에, 네트워크 깊이 확장을 통해 보다 심층적이고 추상적인 특징 표현을 학습할 수 있도록 유도하였다. 또한 추출된 정보를 융합하는 Neck과 최종 바운딩 박스를 예측하는 탐지부(Head)를 소형 객체 특화형 다중 융합 아키텍처로 변경하였다.

3-2 소형 객체 특화 4-Scale 헤드

기존의 YOLOv8 아키텍처는 Backbone을 통과하며 다운샘플링된 특징 맵을 기반으로 P3, P4, P5의 3가지 크기에서만 객체를 탐지한다. 그러나 초기 화재 불씨나 원거리 카메라에서 포착되는 발생 초기의 연기는 10×10 픽셀 이하의 매우 작은 크기를 갖는 경우가 많으며, 기존의 최소 해상도인 P3에서는 다운샘플링 과정 중 객체의 픽셀 정보가 소실되어 탐지율이 저하될 우려가 있다.

이를 보완하기 위해, 본 논문에서 제안하는 모델은 다운샘플링 횟수를 줄여 원본 이미지 대비 해상도 감소가 비교적 적은 1/4 스케일의 고해상도 P2 레이어 연산 경로를 추출하여 네트워크의 Neck과 Head에 새롭게 추가 통합하였다. 결과적으로 그림 1의 최하단에 나타난 바와 같이 4-Scale Detection Head(P2, P3, P4, P5) 구조를 완성하였다. 연산량의 증가가 수반되나 고해상도의 공간적 세부 정보를 유지함으로써 픽셀 단위가 극히 작은 초기 발화점과 연기의 경계면을 보다 정밀하게 식별할 수 있는 구조적 기반을 마련하였다.

3-3 다중 어텐션 메커니즘 기반 특징 강조

화재 관제 환경은 조명의 급격한 변화를 비롯해 구름, 안개, 야간의 빛 번짐 등 화재 및 연기와 시각적 특징이 유사한 배경 노이즈가 무수히 산재해 있다. 이러한 배경 노이즈를 필터링하고 타겟 객체에 적절한 가중치를 부여하기 위해, 네트워크의 각기 다른 위치에 복합 어텐션 메커니즘을 통합하였다. 그림 1의 Backbone 네트워크를 구성하는 각 C2f 모듈 후단에 CBAM(Convolutional Block Attention Module)을 결합하였다. CBAM은 특징 맵에서 채널 어텐션과 공간 어텐션을 순차적으로 적용하여, 화염의 색상 정보나 연기의 불규칙한 질감을 전경으로 분리해 낸다.

또한 Backbone의 마지막 특징 추출 계층에는 SPPF 모듈과 함께 SimAM(Simple, Parameter-Free Attention Module)을 적용하였다. SimAM은 에너지 함수를 기반으로 각 뉴런의 3D 어텐션 가중치를 계산하여, 엣지 디바이스에

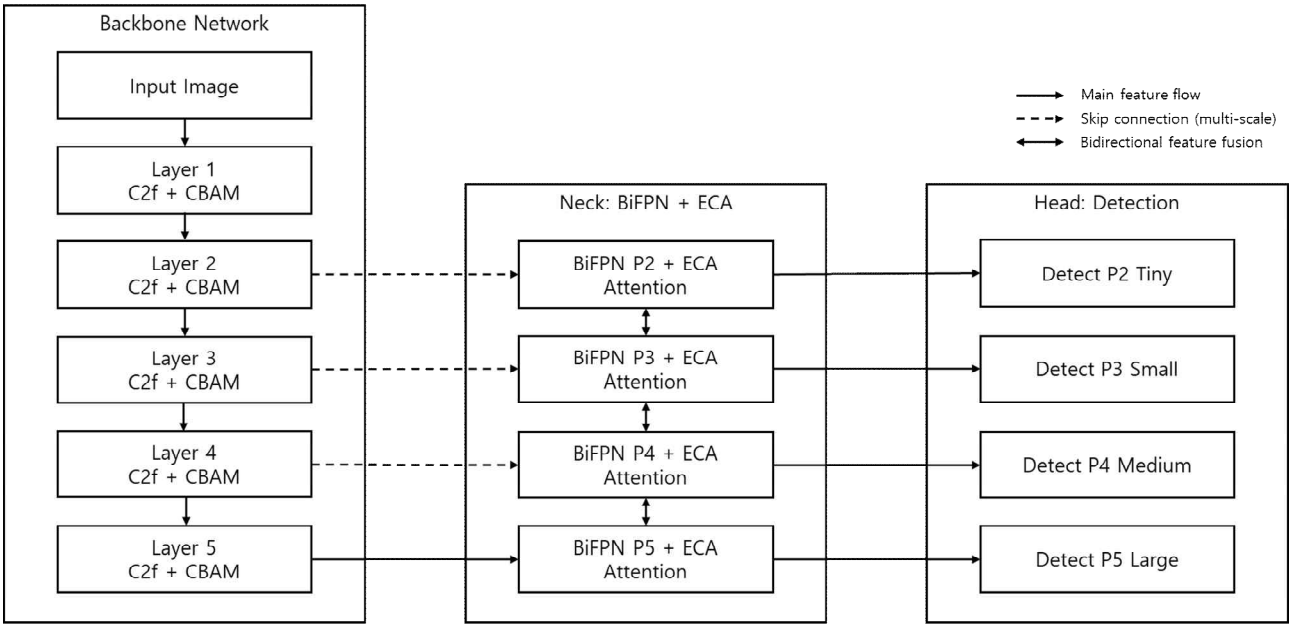

그림 1. 제안하는 모델의 전체 아키텍처
Fig. 1. Full architecture of the proposed model

파라미터 연산 부담을 가중하지 않으면서 전역적인 문맥 정보를 보전하는 역할을 수행한다.

이와 함께 특징이 융합되는 Neck의 다중 경로 전반에는 ECA(Efficient Channel Attention)를 결합하였다. 차원 축소 없이 1D 합성곱을 통해 국소적인 채널 간 상호작용을 계산하는 ECA는 연산 비용이 비교적 적으며, 특징 융합 시 발생할 수 있는 노이즈 증폭을 효과적으로 억제한다.

3-4 BiFPN 기반 양방향 특징 융합

일반적인 객체 탐지 모델의 Neck 구조는 서로 다른 해상도의 특징 맵을 결합할 때 단순한 Concatenation 또는 Element-wise Addition 연산을 수행한다. 그러나 이러한 방식은 의미론적 정보가 풍부한 깊은 레이어와 공간적 정보가 풍부한 얕은 레이어의 기여도를 동등하게 취급하므로, 다중 스케일 탐지 시 정보 불균형을 초래할 수 있다. 이러한 문제를 개선하고자 제안 모델의 Neck 구조에 BiFPN (Bidirectional Feature Pyramid Network)을 채택하였다. 그림 1의 중앙부에 도식화된 것처럼, BiFPN은 상위 레이어에서 하위 레이어로 내려가는 Top-Down 경로와 다시 올라가는 Bottom-Up 경로, 그리고 동일한 스케일의 입력을 직접 연결하는 우회 경로를 구성한다. 특히 서로 다른 해상도의 노드를 병합할 때, 학습 가능한 정규화된 가중치를 사용하여 네트워크가 P2부터 P5까지로 확장된 특징 맵들의 융합 비율을 데이터에 맞게 조정하도록 함으로써 높은 객체 탐지 강건성을 확보할 수 있도록 설계하였다.

Ⅳ. 실험 및 결과 분석

4-1 데이터셋

본 논문에서는 화재 탐지 모델의 성능 검증을 위해 D-fire 데이터셋을 활용하였다[10]. D-fire 데이터셋은 화재 및 연기 탐지 알고리즘의 학습을 위해 설계된 21,527장의 이미지로 구성되어 있다. 데이터셋은 화재와 연기, 두 가지 클래스로 구성되어 있으며, 전체 데이터는 학습, 검증, 테스트 데이터셋으로 70:20:10의 비율로 분할하였으며, 데이터 증강을 적용하지 않고 데이터셋 원본을 그대로 활용하였다.

4-2 실험 환경 및 실험 설정

모델을 학습 후 실제 화재 탐지 성능 검증을 위한 추론은 자원이 제한된 엣지 디바이스 환경에서 실시되었다.

모델의 학습에서는 NVIDIA A100 GPU가 탑재된 고성능 워크스테이션을 사용하였다. NVIDIA A100 GPU는 대용량 메모리와 Tensor Core를 통해 복잡한 YOLO 모델의 학습 시간을 단축하고 안정적인 수렴을 달성할 수 있다. 학습에 적용된 하이퍼파라미터 설정은 표 2와 같다.

또한 학습이 완료된 모델은 ONNX 포맷으로 변환한 후, 엣지 디바이스인 NVIDIA Jetson Orin Nano에서 추론을 실시하였다. Jetson Orin Nano는 1024개의 CUDA 코어와 32개의 Tensor 코어를 갖춘 엣지용 AI 모듈로, 본 연구에서는 TensorRT 8.5를 활용하여 FP16 정밀도로 모델을 최적화한 후 실시간 추론 속도를 측정하였다.

표 2. 하이퍼파라미터 설정
Table 2. Hyperparameter settings

Parameter	Value
Image Size	640 × 640
Optimizer	SGD
Epochs	300
Patience	50
Batch Size	32
Learning Rate	0.01

이미지 사이즈는 정확도와 연산 효율 간의 균형을 고려하여 640×640으로 설정하였다. 옵티마이저는 SGD를 채택하였으며, 전체 학습 과정은 300에폭으로 설정하였고, 과적합을 방지하기 위해 Early Stopping Patience를 50으로 설정하였다.

4-3 정량적 정확도 분석

최종 학습된 모델들의 성능 지표는 다음 표 3과 같다.

표 3. YOLO 모델의 정량적 성능 비교
Table 3. Performance comparison of YOLO models

Model	Precision	Recall	mAP @0.5	mAP @0.5:0.95
YOLOv8-L	80.32%	73.95%	79.63%	48.30%
YOLOv8-M	81.03%	74.84%	80.13%	48.49%
YOLOv8-RE2	80.62%	72.62%	79.10%	48.06%

실험 결과, YOLOv8-M 모델이 mAP@0.5:0.95 기준 48.49%로 가장 높은 정확도를 기록하였다. 통상적으로 모델의 크기가 클수록 정확도가 높을 것으로 예상되나, 본 실험에서는 YOLOv8-M 모델이 과적합 없이 최적의 일반화 성능을 보였다.

본 연구에서 제안한 모델인 YOLOv8-RE2는 48.06%의 정확도를 기록하였다. 이는 베이스라인인 YOLOv8-M 모델 대비 0.42%p의 차이를 기록하며, 성능 저하 없이 대등한 수준의 탐지 능력을 확보했음을 나타낸다. 특히 정밀도 측면에서는 80.62%를 기록하며 대형 모델인 YOLOv8-L 모델 대

비 오탐지 억제 능력이 우수함을 확인하였다.

그림 2에서 그림 5까지는 각 모델의 성능 차이를 막대그래프로 시각화한 것이다.

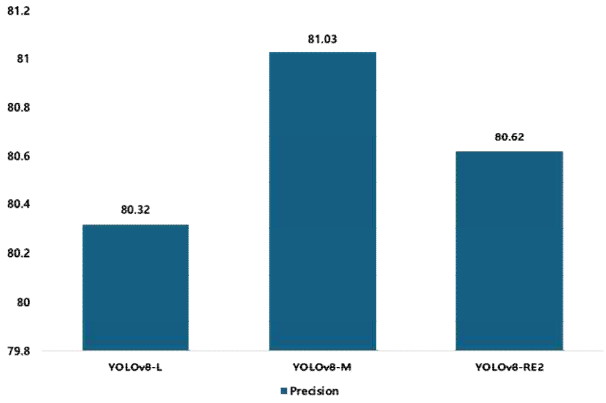

그림 2. 각 모델의 정밀도 성능 비교
Fig. 2. Compare precision performance for each model

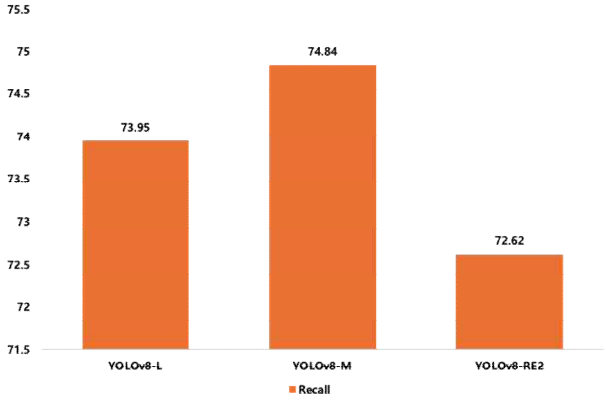

그림 3. 각 모델의 Recall 성능 비교
Fig. 3. Compare Recall performance for each model

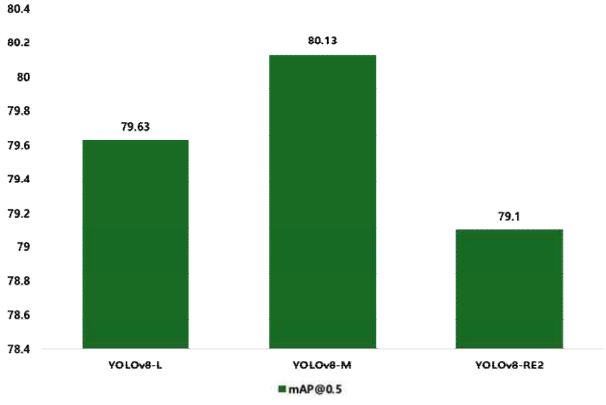

그림 4. 각 모델의 mAP@0.5 성능 비교
Fig. 4. Compare mAP@0.5 performance for each model

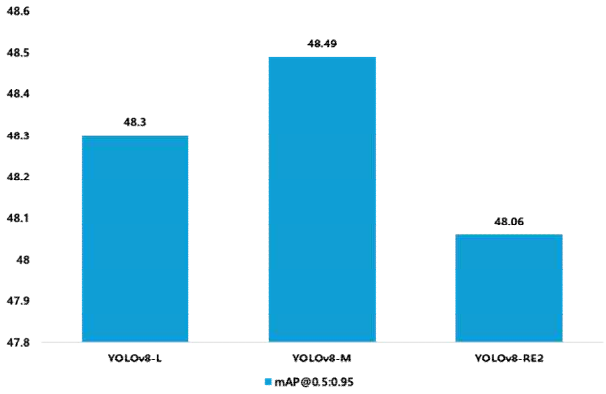

그림 5. 각 모델의 mAP@0.5:0.95 성능 비교
Fig. 5. Compare mAP@0.5:0.95 performance for each model

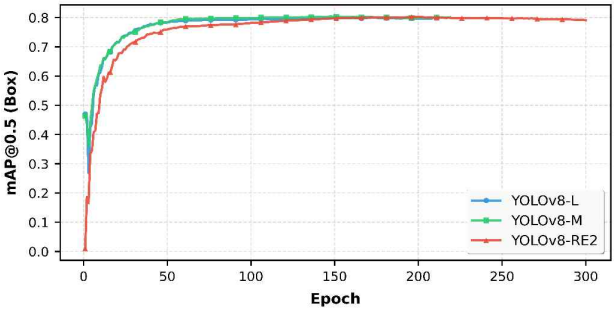

그림 6. 각 모델의 mAP@0.5 학습곡선 그래프
Fig. 6. mAP@0.5 learning curve graph for each model

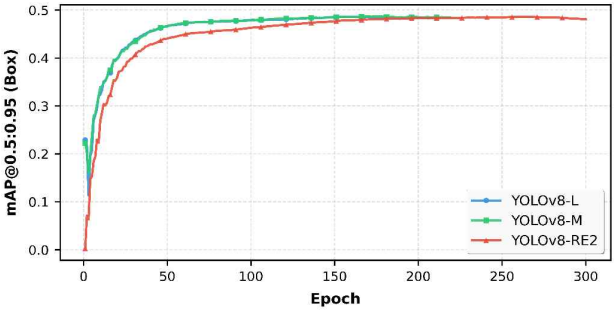

그림 7. 각 모델의 mAP@0.5:0.95 학습곡선 그래프
Fig. 7. mAP@0.5:0.95 learning curve graph for each model

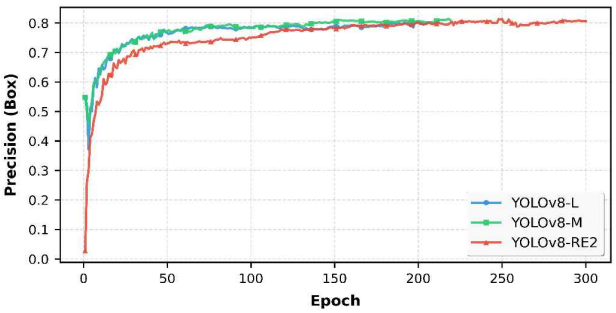

그림 8. 각 모델의 Precision 학습곡선 그래프
Fig. 8. Precision learning curve graph for each model

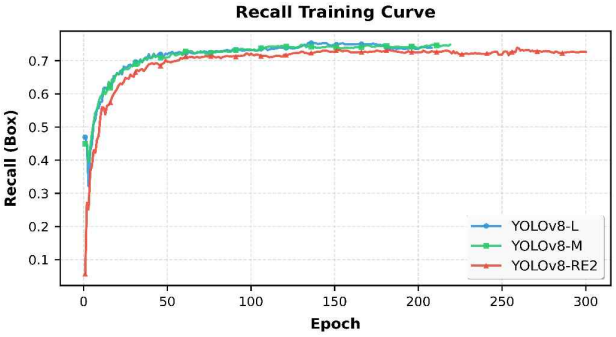

그림 9. 각 모델의 Recall 학습곡선 그래프
Fig. 9. Recall learning curve graph for each model

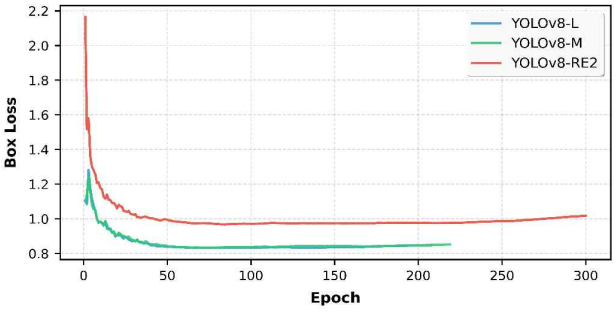

그림 10. 각 모델의 Box Loss 학습곡선 그래프
Fig. 10. Box Loss learning curve graph for each model

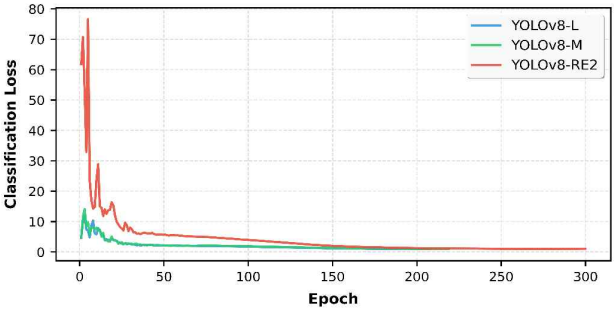

그림 11. 각 모델의 Classification Loss 학습곡선 그래프
Fig. 11. Classification Loss learning curve graph for each model

그림 6에서 그림 11까지는 학습 진행에 따른 성능 지표 변화를 시각화한 것이다. 모든 모델이 안정적으로 수렴되었으나, YOLOv8-M 모델이 전반적인 지표에서 우위를 나타낸다. mAP 및 Precision 지표에서는 학습 초기 약 50 에폭 구간에서 세 모델 모두 급격한 성능 향상을 보이며, 이후 완만한 수렴 양상을 보인다. 특히 YOLOv8-M 모델은 수렴 속도가 가장 빠르고 최종 수렴 값도 높은 반면, YOLOv8-RE2 모델은 구조적 복잡성으로 인해 상대적으로 완만한 상승 곡선을 보이며 수렴 시점이 다소 지연되는 경향을 확인할 수 있다. Box Loss 및 Classification Loss에서는 YOLOv8-RE2 모델이 학습 후반부에도 타 모델 대비 높은 손실 값을 유지하는 경향이 나타나, 데이터 증강 미적용에 따른 일반화 성능 저하가 손실 수렴에도 영향을 미친 것으로 판단된다.

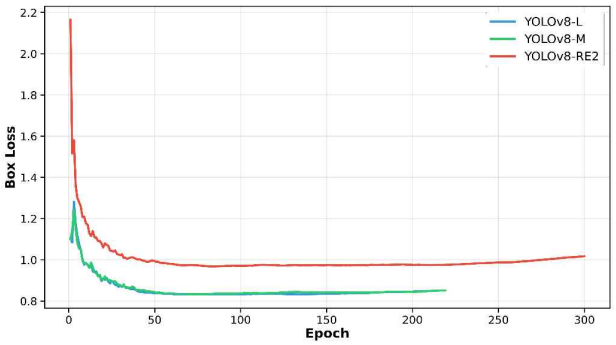

그림 12. 각 모델의 Box Loss 비교 그래프
Fig. 12. Box Loss comparison graph for each model

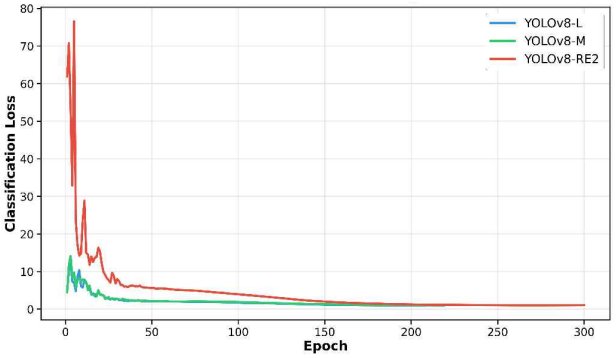

그림 13. 각 모델의 Classification Loss 비교 그래프
Fig. 13. Classification Loss comparison graph for each model

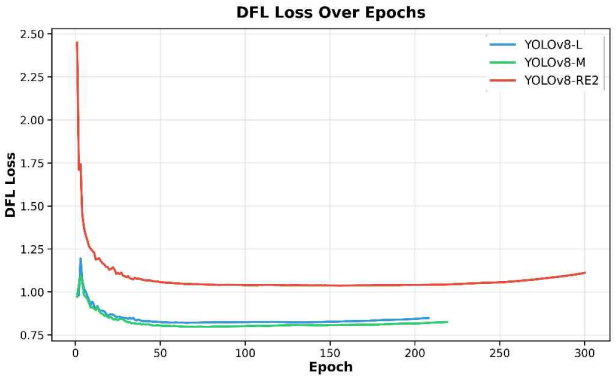

그림 14. 각 모델의 DFL Loss 비교 그래프
Fig. 14. DFL Loss comparison graph for each model

그림 12에서 그림 14까지는 학습 과정 중 각 모델의 손실 함수를 시각화한 것이다. 각 모델을 비교해 보면 YOLOv8-M 모델이 Box Loss와 Classification Loss 모두에서 가장 낮은 값을 나타내었다. 또한 DFL Loss의 경우 세 모델이 유사한 수렴 양상을 보이나, YOLOv8-RE2 모델은 학습 후반부에서도 타 모델 대비 높은 손실 값을 유지하는 경향이 나타나 모델 복잡도 증가에 따른 학습 수렴의 어려움을 반영하는 것으로 판단된다.

4-4 추론 속도 및 효율성 분석

실시간 화재 감시 시스템 구축을 위해 가장 중요한 지표인 추론 속도(FPS)와 효율성을 측정한 결과는 다음 표 4와 같다. 모든 속도 측정은 엣지 디바이스인 Jetson Orin Nano에서 TensorRT FP16으로 최적화하여 수행하였다.

표 4. 각 모델의 추론 속도와 효율성 비교
Table 4. Efficiency comparison

Model	mAP @0.5:0.95	Inference Time	FPS	Efficiency Score
YOLOv8-L	48.30%	52.69ms	16.81	8.12
YOLOv8-M	48.49%	59.49ms	18.98	9.20
YOLOv8-RE2	48.06%	72.67ms	13.76	6.61

표 4에서 효율성 점수는 모델의 탐지 성능을 나타내는 mAP와 실시간 처리 능력을 나타내는 FPS를 종합적으로 평가하기 위해 고안된 지표이다. 이 지표는 다음과 같은 수식으로 산출된다.

$$\text{Efficiency Score} = \text{mAP}_{0.5:0.95} \times \text{FPS} \quad (1)$$

여기서 mAP0.5:0.95는 IoU(Intersection over Union) 임계값을 0.5에서 0.95까지 0.05 간격으로 변화시키며 측정한 평균 정확도의 평균값으로서 모델의 전반적인 탐지 정확도를 나타낸다. 또한 FPS는 초당 처리 가능한 프레임 수로, 모델의 추론 속도를 나타낸다. 따라서 효율성 점수가 높다는 것은 모델이 높은 정확도를 유지하면서 빠른 속도로 동작함을 의미하며, 이는 엣지 디바이스와 같은 실시간 응용 분야에서의 실용성을 평가하는 척도가 된다.

FPS는 YOLOv8-M 모델이 18.98 FPS로 가장 빠른 성능을 보여주었으며, YOLOv8-L 모델이 16.81 FPS, YOLOv8-RE2 모델이 13.76 FPS를 기록하였다. 효율성 점수 산출 결과 YOLOv8-M 모델이 가장 높은 효율성 점수를 기록하였다.

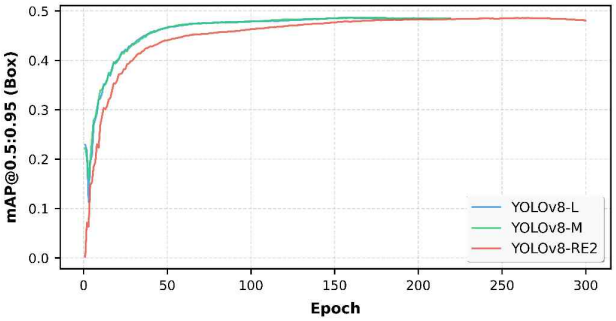

그림 15. 각 모델의 수렴 속도 분석 그래프
Fig. 15. Rate of convergence in the learning process

그림 15는 학습 과정에서의 수렴 속도를 시각화한 것이다. YOLOv8-M 모델은 학습 초기부터 가파른 성능 향상을 보이며 빠르게 안정화되지만, YOLOv8-RE2 모델은 상대적으로 완만한 상승 양상을 보인다. 이는 모델 구조의 복잡성으로 인해 학습 난이도가 높았을 가능성을 시사한다.

4-5 제안 모델의 한계점 및 성능 저하 원인 분석

실험 결과, 소형 객체 탐지에 특화되도록 설계된 제안 모델이 베이스라인인 YOLOv8-M 모델 대비 정확도와 추론 속도 측면에서 다소 낮은 성능을 기록하였다. 특히 재현율이 하락한 점과, 대형 모델인 YOLOv8-L 모델보다도 낮은 추론 속도를 기록한 원인을 분석하면 다음과 같다.

가장 핵심적인 성능 저하 요인은 고해상도 특징 맵 연산 및 파라미터 급증에 따른 하드웨어 병목 현상이다. 실제 모델의 네트워크 복잡도를 분석한 결과, 베이스라인인 YOLOv8-M 모델의 파라미터 수가 약 25.9M인 반면, 제안 모델은 심층 C2f 모듈 확장과 다중 어텐션 추가로 인해 파라미터 수가 대형 모델인 YOLOv8-L을 상회하는 63.4M으로 2배 이상 급증하였다. 또한 기존 YOLOv8이 P3 크기부터 탐지를 수행하는 반면, 제안 모델은 더욱 고해상도인 P2 레이어에서 탐지를 수행한다. 이와 같이 C2f 모듈 확장과 P2 고해상도 레이어 추가로 인해 파라미터 연산량이 크게 증가하였다. 이는 Jetson Orin Nano의 제한된 메모리 대역폭과 충돌하여 병목 현상을 유발한 것으로 판단된다.

이와 더불어, 다중 어텐션 메커니즘 도입에 따른 TensorRT 최적화 효율 저하 역시 추론 속도 하락을 가중시켰다. 제안 모델은 배경 노이즈 필터링을 위해 네트워크 전반에 CBAM, SimAM, ECA를 결합하였다. 이러한 어텐션 모듈들은 글로벌 평균 풀링과 시그모이드 활성화 함수 등 메모리 접근이 빈번한 연산을 포함한다. 이는 NVIDIA TensorRT의 주요 가속 기법인 레이어 및 텐서 퓨전이 연속적으로 병합되는 것을 방해하며, 궁극적으로 GPU 커널 호출 횟수를 증가시켜 실제 추론 속도를 대폭 하락시키는 요인으로 작용하였다.

나아가, 모델 복잡도 증가에 비해 훈련 데이터의 다양성이 부족했던 점은 재현율 하락의 주된 원인으로 확인되었다. 그림 2 및 그림 3의 결과에서 제안 모델의 Precision은 80.62%로 높은 수준을 유지했으나, Recall은 72.62%로 가장 낮았다. 이는 P2 레이어와 다중 어텐션이 훈련 데이터의 미세한 특징에는 강하게 반응하도록 학습되었으나, 데이터 증강 기법을 배제함에 따라 깊고 복잡해진 네트워크가 충분한 일반화 성능을 확보하지 못했기 때문이다. 즉, 모델이 부분적인 과적합 현상을 겪게 되었으며, 이는 그림 15에 나타난 학습 수렴 지연을 유발한 근본적인 이유로 분석된다.

결론적으로 제안 모델은 연산량이 극도로 제한된 엣지 환경에서 미세 특징 추출이라는 목적을 달성하기 위해 추론 속도와 일반화 성능 간의 Trade-off가 발생한 한계를 나타낸다.

Ⅴ. 결론 및 향후 연구

5-1 결론

본 논문에서는 연산 자원이 제한된 엣지 디바이스 환경인 NVIDIA Jetson Orin Nano에서 초기 화재 및 연기를 실시간으로 정밀하게 탐지하기 위해, 기존 YOLOv8 아키텍처를 재설계한 커스텀 모델인 YOLOv8-RE2를 제안하고 그 실효성을 검증하였다. 초기 화재의 발화점 및 발생 초기 연기와 같이 픽셀 단위의 미세한 특징을 포착하기 위해 네트워크 하단에 고해상도 P2 레이어를 추가하고, 복잡한 배경 노이즈를 필터링하고자 다중 어텐션 메커니즘을 통합 적용하였다.

실험 결과, 제안 모델은 정밀도 측면에서 80.62%를 기록하며 대형 모델인 YOLOv8-L 모델의 80.32%, YOLOv8-M 모델의 81.03%와 대등한 수준의 우수한 오탐지 억제 능력을 입증하였다. 이는 실제 화재 관제 시스템에 도입 시, 오경보로 인한 소방력 낭비를 줄이는 데 중요한 지표로 작용할 수 있음을 시사한다.

그러나 실시간성을 평가하는 추론 속도에서는 13.76 FPS를 기록하여 베이스라인 모델인 YOLOv8-M 모델의 18.98 FPS 대비 약 27.5% 하락하였으며, mAP@0.5:0.95 기준 탐지 정확도 역시 48.06%로 다소 저조한 수치를 기록하였다. 이러한 결과가 도출된 핵심적인 원인은 미세 특징 추출 성능을 극대화하기 위해 도입한 구조적 복잡성이 엣지 디바이스의 하드웨어적 한계 및 최적화 엔진의 특성과 충돌했기 때문이다.

이는 P2 고해상도 레이어 추가에 따른 연산량 급증 및 다중 어텐션 모듈로 인한 TensorRT 최적화 효율 저하가 복합적으로 작용한 결과로, 엣지 환경에서의 구조적 복잡성과 하드웨어 제약 간의 Trade-off를 명확히 보여주는 사례로 볼 수 있다.

결론적으로 본 연구를 통해, 연산량이 극도로 제한된 엣지 환경에서는 단순히 해상도를 높이고 어텐션 모듈을 추가하는 소프트웨어적 접근만으로는 실시간성과 정확도의 완벽한 균형을 맞추기 어렵다는 것을 확인하였다. 따라서 엣지 컴퓨팅 기반의 지능형 영상 분석 시스템을 구축하기 위해서는 알고리즘의 고도화뿐만 아니라, 타겟 하드웨어의 특성과 최적화 컴파일러의 동작 방식을 종합적으로 고려한 설계가 필수적이라는 점을 본 연구를 통해 확인하였다.

5-2 향후 연구

본 논문에서 도출된 한계점과 하드웨어 병목 현상의 근본적인 원인을 바탕으로, 향후 연구에서는 엣지 디바이스의 아키텍처 특성을 충분히 고려한 커스텀 모델의 고도화를 진행할 계획이다. 또한 본 연구에서는 추론 속도 분석이 단일 측정 기반의 FPS 비교에 그쳐 통계적 유의성 검증이 이루어지지 못한 한계가 존재한다. 향후 연구에서는 동일 조건 하에서 반복 측정을 통한 평균 및 표준편차 산출, 그리고 모델 간 추론 시간 차이에 대한 통계적 유의성 검증을 추가로 수행하여 실험 결과의 신뢰도를 높이고자 한다. 또한 엣지 환경에 최적화된 경량 어텐션 메커니즘을 개발하여 TensorRT의 레이어 퓨전이 단절되는 현상을 방지하며, 기존의 메모리 집약적인 어텐션 모듈을 배제하는 대신, 학습 단계에서 복잡한 구조를 가지지만 추론 단계에서 단일 합성곱 연산으로 병합되는 재매개변수화 기법을 활용하거나, GPU 커널 친화적인 연산만으로 구성된 엣지 전용 어텐션 모듈을 새롭게 설계하여 탐지 정확도를 유지하면서 추론 지연 시간을 베이스라인 수준으로 회복하고자 한다.

또한, 네트워크 압축 및 지식 증류 기술을 도입하여 P2 레이어의 막대한 연산 부하 문제를 해결할 계획이다. 구조 압축을 통해 불필요한 채널과 노드를 제거하고, 고해상도 특징을 잘 추출하는 무거운 교사 모델의 지식을 가벼운 모델에 전달하는 지식 증류 기법을 적용함으로써, 추론 시에는 무거운 P2 레이어 없이도 초기 발화점을 정밀하게 탐지할 수 있는 경량화 아키텍처를 구축하고자 한다. 최종적으로 이러한 구조적 경량화와 데이터 고도화 과정을 통합함으로써, 엣지 환경의 엄격한 자원 제약을 극복하고 실제 복합 건물이나 산림 화재 관제 시스템에 즉시 투입할 수 있는 빠르고 신뢰성 높은 지능형 재난 감시 시스템을 완성하고자 한다.

감사의 글

본 연구는 2026년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0031136, 2026년 지역혁신클러스터육성).

참고문헌

[1] National Fire Agency. 2024 Fire Statistical Yearbook [Internet]. Available: https://www.nfa.go.kr.

[2] C.-S. Yun and Y.-H. Park, “Lightweight Shuffling Attention for Real-Time Smoke and Fire Detection,” *Journal of the Institute of Electronics and Information Engineers*, Vol. 61, No. 8, pp. 59-71, August 2025. https://doi.org/10.7471/ieie.2025.62.8.059

[3] H.-S. Jo, H.-J. Kwon, H.-Y. Jung, and S.-A. Lee, “Fire Detection with YOLOv8 and YOLOv9 Models Based on NVIDIA Jetson AGX Orin,” in *Proceeding of the KIISE Korea Computer Congress 2024*, Jeju, pp. 1764-1766, June 2024.

[4] N. Choi. Daegu Bathhouse Fire... Concluded as Total Negligence [Internet]. Available: https://www.fpn119.co.kr/

112715.

[5] C. G. Lim and J. M. Choi, “Real-Time Fire and Smoke Detection Performance Evaluation Based on YOLOv8n Using RGB Images,” *Journal of Platform Technology*, Vol. 13, No. 6, pp. 82-90, 2025. https://doi.org/10.23023/JPT.2025.13.6.082

[6] L. T. Ramos, E. Casas, C. Romero, F. Rivas-Echeverría, and E. Bendek, “A Study of YOLO Architectures for Wildfire and Smoke Detection in Ground and Aerial Imagery,” *Results in Engineering*, Vol. 26, 104869, June 2025. https://doi.org/10.1016/j.rineng.2025.101123

[7] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLOv8 [Internet]. Available: https://github.com/ultralytics/ultralytics.

[8] I. Polenakis, C. Sarantidis, I. Karydis, and M. Avlonitis, “Smoke Detection on the Edge: A Comparative Study of YOLO Algorithm Variants,” *Signals*, Vol. 6, No. 4, pp. 936-953, 2025. https://doi.org/10.3390/signals6040060

[9] M. T. Mohammed and M. S. Shubber, “Impact of Colour Space Transformation on Smoke Detection Accuracy Using RESNET50,” *Iraqi Journal of Data Science*, Vol. 2, No. 1, pp. 37-49, 2025. https://doi.org/10.51173/ijds.v2i1.15

[10] P. V. A. B. de Venâncio, A. C. Lisboa, and A. V. Barbosa, “An Automatic Fire Detection System Based on Deep Convolutional Neural Networks for Low-Power, Resource-Constrained Devices,” *Neural Computing and Applications*, Vol. 34, No. 18, pp. 15349-15368, 2022. https://doi.org/10.1007/s00521-022-07467-z

임창건(Chang Geon Lim)

2020년～현 재: 목원대학교 컴퓨터융합학부 학석사과정
※관심분야 : 영상처리, 컴퓨터비전, 딥러닝 등

최재명(Jae Myeong Choi)

2014년 : 목원대학교 일반대학원 IT 공학과 박사

2014년 3월～현 재: 목원대학교 컴퓨터융합학부 부교수
※관심분야 : 무선통신시스템, 통신재난, 스마트재난시스템, IoT, 재난안전통신, 사회재난(통신)정책, 빅데이터, 재난예측, 재난관리, 디지털콘텐츠 등