The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025

YOLOv8 Real-Time Object Detection on Jetson Orin Nano

Chaitanya Krishna Sai Jonnadula[1] , Muthukumaran Thulasingam[2],
[1,2]National Small Industries Corporation – Technical Services Centre, Hyderabad, Telangana-India
Corresponding Author : chaitanyajonnadula21@gmail.com

జెట్సన్ ఓరిన్ నానోలో YOLOv8 రియల్-టైమ్ ఆబ్జెక్ట్ డిటెక్షన్

చైతన్య కృష్ణ సాయి జొన్నాదుల [1] , ముత్తుకుమారన్ తులసింగం [2] ,
[1,2] నేషనల్ స్మాల్ ఇండస్ట్రీస్ కార్పొరేషన్ – టెక్నికల్ సర్వీసెస్ సెంటర్, హైదరాబాద్, తెలంగాణ-భారతదేశం
సంబంధిత రచయిత : chaitanyajonnadula21@gmail.com

వియుక్త

రియల్-టైమ్ ఆబ్జెక్ట్ డిటెక్షన్‌లో ఇటీవలి పురోగతులు ఎంబెడెడ్ అప్లికేషన్‌లకు అనువైన ఎండ్-టు-ఎండ్, సింగిల్-షాట్ ఇన్ఫెరెన్స్ పైప్‌లైన్‌లను ప్రారంభించడం ద్వారా కంప్యూటర్ విజన్ రంగాన్ని మార్చాయి. అల్ట్రాలిటిక్స్ YOLOv8 మోడల్ YOLO కుటుంబంలో తాజా పునరుక్తిని సూచిస్తుంది, యాంకర్-ఫ్రీ డిటెక్షన్ హెడ్, CSP-ఆధారిత బ్యాక్‌బోన్‌లు మరియు అధునాతన శిక్షణ ఆప్టిమైజేషన్‌లను పరిచయం చేస్తుంది. ఈ పని NVIDIA జెట్సన్ ఓరిన్ నానోపై YOLOv8 యొక్క పూర్తి విస్తరణ పైప్‌లైన్‌ను అందిస్తుంది, ఇది ఆంపియర్ GPUతో కూడిన 40 TOPS ఎడ్జ్ AI పరికరం, లాజిటెక్ USB కెమెరాతో ఇన్‌పుట్ సోర్స్‌గా కలిపి ఉంటుంది. అమలు అధిక-త్రూపుట్, తక్కువ-లేటెన్సీ ఇన్ఫెరెన్స్ కోసం PyTorch శిక్షణ, ONNX ఎగుమతి మరియు TensorRT ఆప్టిమైజేషన్‌ను అనుసంధానిస్తుంది. ప్రయోగాత్మక ఫలితాలు VGA రిజల్యూషన్ కోసం సెకనుకు సుమారు 30 ఫ్రేమ్‌ల వద్ద వ్యక్తులు మరియు వస్తువులను నమ్మదగిన గుర్తింపును ప్రదర్శిస్తాయి, ఇది పరిమిత పవర్ ఎన్వలప్‌లలో ఎంబెడెడ్ రియల్-టైమ్ విజన్ యొక్క సాధ్యాసాధ్యాలను హైలైట్ చేస్తుంది. క్వాంటైజేషన్, పవర్-పెర్ఫార్మెన్స్ ట్రేడ్-ఆఫ్‌లు మరియు రోబోటిక్స్ మరియు స్మార్ట్ సర్వైలెన్స్ కోసం అప్లికేషన్ చిక్కులతో సహా సిస్టమ్ యొక్క విస్తరణ పరిగణనలు చర్చించబడ్డాయి .

కీలకపదాలు:
YOLOv8, ఆబ్జెక్ట్ డిటెక్షన్, జెట్సన్ ఓరిన్ నానో, ఎడ్జ్ AI, TensorRT, రియల్-టైమ్ ఇన్ఫరెన్స్

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025

1. పరిచయం

ఆబ్జెక్ట్ డిస్కవరీ అనేది కంప్యూటర్ దృష్టికి పునాది, ఇది స్వతంత్ర నావిగేషన్ మరియు నిఘా నుండి ఉత్తేజిత వాస్తవికత వరకు పనులను అనుమతిస్తుంది. R-CNN, ఫాస్ట్ R-CNN మరియు ఫాస్టర్ వంటి సాంప్రదాయ బహుళ-దశ సెన్సార్లు అత్యాధునిక డెలికేసీని సాధించాయి కానీ అధిక కంప్యూటేషనల్ అవుట్‌ఫ్లోతో బాధపడ్డాయి, ఎంబెడెడ్ బయాస్‌పై రియల్-టైమ్ డిప్లాయ్‌మెంట్ కోసం వాటి మైలేజీని పరిమితం చేశాయి. విరుద్ధంగా, YOLO (యు ఓన్లీ లుక్ ఓర్లీ లుక్) కుటుంబం సింగిల్-స్టేజ్ డిస్కవరీని ఆవిష్కరించింది, పనిని తిరోగమన సమస్యగా రూపొందించింది మరియు తద్వారా ముగింపు వేగంలో ఆర్డర్స్-ఆఫ్-మాగ్నిట్యూడ్ పురోగతిని సాధించింది. 2023లో అల్ట్రాలిటిక్స్ ప్రవేశపెట్టిన YOLOv8, డిస్కవరీ డెలికేసీ మరియు కంప్యూటేషనల్ ఎఫెక్టివ్‌నెస్ మధ్య అనుకూలమైన సమతుల్యతను అందించడానికి యాంకర్-ఫ్రీ డిస్కవరీ హెడ్‌లు, ప్రభావవంతమైన CSP చైన్‌లు మరియు స్ట్రీమ్‌లైన్డ్ ట్రైనింగ్ ఛానెల్‌లను అనుసంధానిస్తుంది. ఇంతలో, NVIDIA జెట్సన్ ఓరిన్ నానో ఎడ్జ్ AI కోసం సరసమైన కానీ ముఖ్యమైన ప్లాట్‌ఫామ్‌ను అందిస్తుంది, దాని 1024 CUDA కోర్లు, 32 టెన్సర్ కోర్లు మరియు LPDDR5 మెమరీ సబ్‌సిస్టమ్‌ను ఉపయోగించి 7 - 15 W పవర్ వద్ద 40 కవర్ల AI పనితీరును అందిస్తుంది. ఈ పరికరంలో YOLOv8 ని నాటడం వలన కాంపాక్ట్ ఎంబెడెడ్ టాకిల్‌తో అత్యాధునిక డిస్కవరీ అల్గోరిథంల వివాహంలో ఆచరణాత్మక జ్ఞానాన్ని అందిస్తుంది .

పరిశోధన లక్ష్యాలు మరియు పద్ధతి

USB కెమెరా ఇన్‌పుట్ ఉపయోగించి రియల్-టైమ్ విజన్ పనుల కోసం NVIDIA జెట్సన్ ఓరిన్ నానోపై YOLOv8 ఆబ్జెక్ట్ డిటెక్షన్ ఫ్రేమ్‌వర్క్‌ను అమలు చేయడం యొక్క ప్రభావాన్ని అంచనా వేయడం ఈ అధ్యయనం లక్ష్యం. పరిశోధన లక్ష్యాలు:

1. రియల్-టైమ్ డిటెక్షన్ కోసం ఎంబెడెడ్ GPU ప్లాట్‌ఫారమ్ (జెట్సన్ ఓరిన్ నానో)పై YOLOv8 ని అమలు చేయడం యొక్క సాధ్యాసాధ్యాలను పరిశోధించడానికి.
2. ONNX మార్పిడి మరియు TensorRT త్వరణం (FP 16 మరియు INT 8) ఉపయోగించి అంచు విస్తరణ కోసం YOLOv8 మోడల్‌ను ఆప్టిమైజ్ చేయడానికి.
3. లైవ్ వీడియో పరిస్థితుల్లో ఫ్రేమ్స్ పర్ సెకండ్ (FPS), జాప్యం మరియు GPU వినియోగం పరంగా సిస్టమ్ పనితీరును కొలవడానికి.

4. వివిధ రకాల లైటింగ్ మరియు దృశ్య సంక్లిష్టతలలో బహుళ వస్తువుల (ఉదాహరణకు, వ్యక్తులు, సెల్ ఫోన్లు) గుర్తింపు ఖచ్చితత్వం మరియు దృఢత్వాన్ని అంచనా వేయడానికి.
5. ఎడ్జ్ AI విస్తరణలలో మోడల్ పరిమాణం, ఖచ్చితత్వం మరియు వేగం మధ్య ట్రేడ్-ఆఫ్‌ల గురించి అంతర్దృష్టులను అందించడానికి.

2. సాహిత్య సర్వే

ఆబ్జెక్ట్ డిటెక్షన్ రెండు పరిపూరక అక్షాలతో అభివృద్ధి చెందింది: ప్రతి-చిత్ర ఖచ్చితత్వాన్ని మెరుగుపరచడం మరియు నిర్బంధ హార్డ్‌వేర్‌పై నిజ-సమయ ఆపరేషన్‌ను ప్రారంభించడానికి అనుమితి జాప్యాన్ని తగ్గించడం. ప్రారంభ అధిక-ఖచ్చితత్వ విధానాలు రెండు-దశల నమూనాను అనుసరించాయి, ఇక్కడ ప్రాంత ప్రతిపాదనలు ఉత్పత్తి చేయబడతాయి మరియు వర్గీకరించబడతాయి. R-CNN మరియు దాని ఉత్పన్నాలు (ఫాస్ట్ R-CNN , ఫాస్టర్ R-CNN) ఈ వర్క్‌ఫ్లోను అధికారికీకరించాయి మరియు సెలెక్టివ్ సెర్చ్ / రీజియన్ ప్రతిపాదన నెట్‌వర్క్‌లను లోతైన వర్గీకరణదారులతో కలపడం ద్వారా బలమైన గుర్తింపు పనితీరును సాధించాయి, కానీ అధిక గణన ఓవర్‌హెడ్ మరియు రియల్-టైమ్ అన్వయతను పరిమితం చేసే జాప్యం యొక్క ఖర్చుతో [14]. పెద్ద లేబుల్ చేయబడిన కార్పోరాపై (ఉదా., ఇమేజ్‌నెట్) ప్రీట్రైన్ చేయబడిన డీప్ కన్వల్యూషనల్ బ్యాక్‌బోన్‌ల విజయం తదుపరి డిటెక్టర్‌లకు ప్రాతినిధ్య పునాదిని అందించింది [10, 8].

జాప్య పరిమితులను తీర్చడానికి, సింగిల్-స్టేజ్ డిటెక్టర్లు డిటెక్షన్‌ను దట్టమైన, పర్-లొకేషన్ ప్రిడిక్షన్ సమస్యగా రీఫ్రేమ్ చేశాయి. సింగిల్ షాట్ డిటెక్టర్ (SSD) బహుళ-స్కేల్ ఫీచర్ మ్యాప్‌లు మరియు డిఫాల్ట్ బాక్స్‌లను ఒకే ఫార్వర్డ్ పాస్‌లో ఆబ్జెక్ట్ క్లాస్‌లు మరియు లొకేషన్‌లను అంచనా వేయడానికి ప్రవేశపెట్టింది, ఇది వినియోగదారు GPU లు మరియు ఎంబెడెడ్ బోర్డులపై ఆచరణాత్మక రియల్-టైమ్ డిటెక్షన్‌ను అనుమతిస్తుంది [9]. అదే సమయంలో, YOLO కుటుంబం గ్రిడ్ కణాల నుండి బౌండింగ్ బాక్స్‌లు మరియు క్లాస్ సంభావ్యతలను నేరుగా అంచనా వేసే ఎండ్-టు-ఎండ్ రిగ్రెషన్ ఫార్ములేషన్‌ను ప్రాచుర్యం పొందింది, పోటీ ఖచ్చితత్వాన్ని కొనసాగిస్తూ ఏకీకృత సింగిల్-నెట్‌వర్క్ విధానం అధిక నిర్గమాంశను సాధించగలదని నిరూపించింది [1]. తదుపరి YOLO పునర్విమర్శలు నిరూపితమైన డీప్-లెర్నింగ్ పద్ధతులను క్రమంగా చేర్చాయి—యాంకర్ బాక్స్‌లు మరియు బ్యాచ్ నార్మలైజేషన్ (YOLOv2), లోతైన అవశేష వెన్నెముకలు మరియు మల్టీ-స్కేల్ ప్రిడిక్షన్ (YOLOv3), మరియు క్రాస్-స్టేజ్ పాక్షిక కనెక్షన్‌లు ప్లస్ బలమైన ఆగ్మెంటేషన్ మరియు లాస్ ఫంక్షన్‌లు (YOLOv4)—వేగం/ఖచ్చితత్వ సరిహద్దును నిరంతరం మెరుగుపరచడానికి [2–4].

YOLO వంశం వెలుపల ఉన్న ఆర్కిటెక్చరల్ ఆవిష్కరణలు కూడా సామర్థ్య సరిహద్దును అభివృద్ధి చేశాయి. అవశేష కనెక్షన్లు (ResNet) ప్రవణతలు అదృశ్యం కాకుండా చాలా లోతైన వెన్నెముకలను ఎనేబుల్ చేశాయి, డిటెక్షన్ హెడ్‌ల కోసం ఫీచర్ వెలికితీతను మెరుగుపరిచాయి [8]. మెరుగైన FLOP ల-నుండి-ఖచ్చితత్వ ట్రేడ్-ఆఫ్‌ల కోసం రిజల్యూషన్, లోతు మరియు వెడల్పును సంయుక్తంగా స్కేల్ చేయడానికి EfficientDet మోడల్ స్కేలింగ్ మరియు వెయిటెడ్ బై-డైరెక్షనల్ ఫీచర్ పిరమిడ్ (Bi FPN)తో కాంపౌండ్ స్కేలింగ్ నియమాన్ని ప్రవేశపెట్టింది [11]. YOLOv7 మరియు YOLOv8 వంటి ఇటీవలి ఆబ్జెక్ట్ డిటెక్టర్లు సమర్థవంతమైన ఫీచర్ అగ్రిగేషన్ మాడ్యూల్‌లను (E-ELAN, CSP వేరియంట్‌లు), యాంకర్-ఫ్రీ హెడ్‌లను నొక్కి చెబుతున్నాయి మరియు అనుమితి ఓవర్‌హెడ్‌ను తగ్గించేటప్పుడు చిన్న-వస్తువు గుర్తింపు మరియు కన్వర్జెన్స్‌ను మెరుగుపరిచే "బ్యాగ్-ఆఫ్-ట్రిక్స్" శిక్షణను అందిస్తున్నాయి [5, 6]. YOLOv8 యొక్క యాంకర్-ఫ్రీ హెడ్ బాక్స్ రిగ్రెషన్‌ను సులభతరం చేస్తుంది మరియు తరచుగా యాంకర్-ఆధారిత వేరియంట్‌లకు సంబంధించి అనుమితి ఖర్చు మరియు హైపర్‌పారామీటర్ సెన్సిటివిటీని తగ్గిస్తుంది [6].

ముడి మోడల్ డిజైన్‌కు మించి, ఆచరణాత్మక అంచు విస్తరణ మోడల్ స్కేలింగ్ మరియు హార్డ్‌వేర్-అవేర్ ఆప్టిమైజేషన్‌లపై ఆధారపడి ఉంటుంది. తేలికైన నిర్మాణాలు (ఉదా. మొబైల్‌నెట్ వేరియంట్‌లు) మరియు స్కేల్డ్ YOLO "నానో/స్మాల్" వేరియంట్‌లు పారామితి మరియు FLOPల తగ్గింపులను అందిస్తాయి, ఇవి ఎంబెడెడ్ SoC లపై రియల్-టైమ్ అనుమితిని సాధ్యం చేస్తాయి [16]. సాఫ్ట్‌వేర్ వైపు, NVIDIA యొక్క CUDA మరియు TensorRT వంటి విక్రేత స్టాక్‌లు శిక్షణ పొందిన నెట్‌వర్క్‌లను ఆప్టిమైజ్ చేసిన రన్‌టైమ్ ఇంజిన్‌లుగా మారుస్తాయి, ఇవి లేయర్ ఫ్యూజన్, కెర్నల్ ఆటోట్యూనింగ్ మరియు మిక్స్‌డ్-ప్రెసిషన్ ఎగ్జిక్యూషన్ (FP 16/INT 8) నిర్వహిస్తాయి, ఇవి అమాయక ఫ్రేమ్‌వర్క్ అనుమితిపై బహుళ-ఫోల్డ్ స్పీడ్‌అప్‌లను అందిస్తాయి [12, 13]. క్వాంటైజేషన్ (పోస్ట్-ట్రైనింగ్ మరియు క్వాంటైజేషన్-అవేర్ శిక్షణ) ఒక కీలకమైన లివర్: INT8 అమలు ప్రధాన త్రూపుట్ లాభాలను ఇస్తుంది కానీ ఖచ్చితత్వ క్షీణతను పరిమితం చేయడానికి జాగ్రత్తగా క్రమాంకనం లేదా తిరిగి శిక్షణ అవసరం.

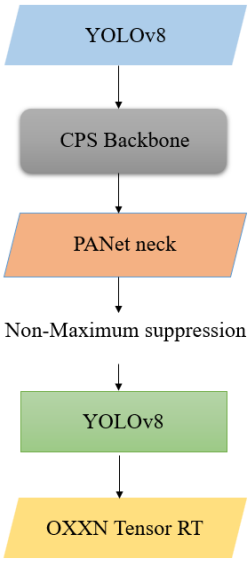
ఎంబెడెడ్ NVIDIA జెట్సన్ ప్లాట్‌ఫారమ్‌లు ఆన్-డివైస్ విజన్ కోసం వాస్తవ పరీక్షా కేంద్రంగా మారాయి, అధ్యయనాలు జెట్సన్ నానో, జేవియర్ మరియు AGX ఓరిన్ బోర్డులపై టెన్సార్‌ఆర్‌టి ఆప్టిమైజేషన్ మరియు మోడల్ ప్రూనింగ్/క్వాంటైజేషన్ తర్వాత రియల్ టైమ్‌లో YOLO వేరియంట్‌లను అమలు చేస్తున్నట్లు చూపిస్తున్నాయి [7]. ఈ రచనలు (1) మోడల్ ఎంపిక (వేరియంట్ పరిమాణం) కఠినమైన రియల్-టైమ్ పరిమితులను తీర్చగలదా అని నియంత్రిస్తుందని, (2) టెన్సర్‌ఆర్‌టి FP16 అమలు తరచుగా అనుకూలమైన ఖచ్చితత్వం/జాప్య బిందువుకు సరిపోతుందని మరియు (3) క్రమాంకనం లేదా పరిమాణీకరణ-అవేర్ శిక్షణను ఉపయోగించినప్పుడు INT8 నిరాడంబరమైన ఖచ్చితత్వ ఖర్చుతో త్రూపుట్‌ను మరింత పెంచుతుందని స్థిరంగా నివేదిస్తున్నాయి.

ఈ పురోగతులు ఉన్నప్పటికీ, అత్యాధునిక డిటెక్టర్‌లను శక్తి-నిర్బంధిత అంచు పరికరాలకు అనువదించేటప్పుడు అంతరాలు ఉంటాయి. మొదటిది, ఆధునిక యాంకర్-రహిత హెడ్‌లు మరియు హార్డ్‌వేర్ యాక్సిలరేటర్‌ల మధ్య పరస్పర చర్య తక్కువగా అన్వేషించబడింది: యాంకర్-రహిత డిజైన్‌లు అల్గారిథమిక్ సంక్లిష్టతను తగ్గిస్తాయి కానీ ఆప్టిమైజ్ చేయబడిన కెర్నల్‌లకు సంబంధించి వాటి మెమరీ/యాక్సెస్ నమూనాలు నిజమైన నిర్గమాంశను ప్రభావితం చేస్తాయి. రెండవది, క్వాంటైజేషన్ దృఢత్వం - ముఖ్యంగా వైవిధ్యమైన లైటింగ్ మరియు మూసివేత కింద చిన్న-వస్తువు గుర్తింపు కోసం - ప్లాట్‌ఫారమ్‌లు మరియు పనిభారాలలో మరింత అనుభావిక మూల్యాంకనం అవసరం. మూడవది, జెట్సన్ టైర్‌లలో (ఓరిన్ నానో, జేవియర్ NX , AGX ఓరిన్) గుర్తింపు ఖచ్చితత్వం, జాప్యం, శక్తి వినియోగం మరియు ఉష్ణ ప్రవర్తనను సంయుక్తంగా నివేదించే సమగ్ర మూల్యాంకనాలు పరిమితం.

ఈ అధ్యయనం జెట్సన్ ఓరిన్ నానోలో తాజా YOLOv8 ఆర్కిటెక్చర్‌ను అమలు చేయడం ద్వారా, మోడల్ స్కేలింగ్, ONNX ఎగుమతి మరియు TensorRT ఆప్టిమైజేషన్ (FP16 /INT8 క్రమాంకనం) లను కలిపి, రియల్ కెమెరా ఫీడ్‌ల కింద త్రూపుట్ (FPS), పర్-ఫ్రేమ్ లేటెన్సీ, GPU వినియోగం మరియు డిటెక్షన్ రోబస్ట్‌నెస్ మధ్య ఆచరణాత్మక ట్రేడ్-ఆఫ్‌లను అంచనా వేయడం ద్వారా ఆ అంతరాలను పరిష్కరిస్తుంది. [1–14] లో నమోదు చేయబడిన పురోగతిలో మా ప్రయోగాత్మక ఫలితాలను గుర్తించడం ద్వారా, మేము రియల్-వరల్డ్ ఎడ్జ్- AI అప్లికేషన్‌ల కోసం ఆధునిక సింగిల్-స్టేజ్ డిటెక్టర్‌ల సాధ్యత యొక్క సమగ్ర అంచనాను అందిస్తాము.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025

3. పద్ధతి



చిత్రం 1. రియల్-టైమ్ YOLOv8 ఆబ్జెక్ట్ డిటెక్షన్ కోసం ప్రతిపాదిత పద్ధతి

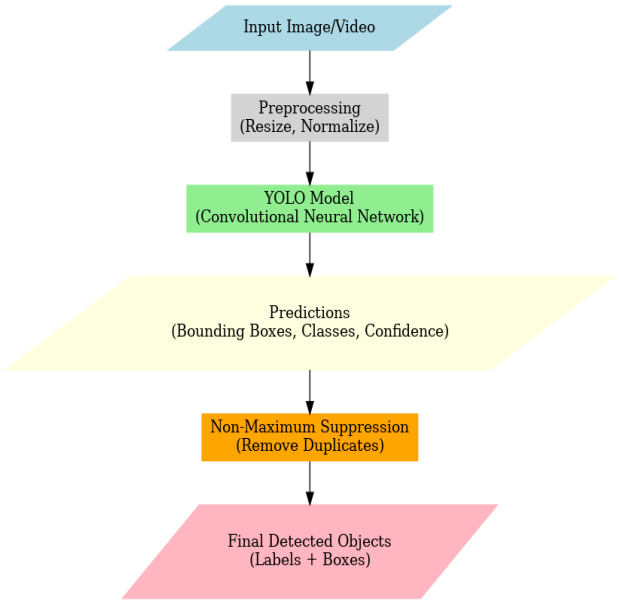
జెట్సన్ ఓరిన్ నానోపై YOLOv8 ని అమలు చేయడానికి ప్రతిపాదిత పద్ధతిలో Fig . 1 లో చూపిన విధంగా నిర్మాణాత్మక పైప్‌లైన్ ఉంటుంది.

1. **మోడల్ ఇనిషియలైజేషన్ (YOLOv8):** ముందస్తు శిక్షణ పొందిన YOLOv8 మోడల్ దాని యాంకర్-ఫ్రీ డిటెక్షన్ హెడ్ మరియు గణన సామర్థ్యం కారణంగా ఎంపిక చేయబడింది, ఇది ఎడ్జ్ AI కి అనుకూలంగా ఉంటుంది.
2. **ఫీచర్ ఎక్స్‌ట్రాక్షన్ (CSP బ్యాక్‌బోన్):** ఇన్‌పుట్ ఫ్రేమ్‌లు క్రాస్-స్టేజ్ పార్షియల్ (CSP) బ్యాక్‌బోన్ ద్వారా పంపబడతాయి, ఇది గ్రేడియంట్ ప్రవాహాన్ని పెంచుతుంది మరియు బహుళ-స్థాయి లక్షణాలను సంగ్రహించేటప్పుడు మెమరీ వినియోగాన్ని తగ్గిస్తుంది.
3. **ఫీచర్ అగ్రిగేషన్ (PANet నెక్):** పాత్ అగ్రిగేషన్ నెట్‌వర్క్ (PANet) వివిధ పరిమాణాల వస్తువులలో గుర్తింపును మెరుగుపరచడానికి లోతులేని మరియు లోతైన ఫీచర్ మ్యాప్‌లను అనుసంధానిస్తుంది.
4. **డిటెక్షన్ హెడ్ & నాన్-మాగ్జిమమ్ సప్రెషన్:** డిటెక్షన్ హెడ్ బౌండింగ్ బాక్స్‌లు, క్లాస్ స్కోర్‌లు మరియు కాన్ఫిడెన్స్ విలువలను అంచనా వేస్తుంది. ఫిల్టర్ రిడెండెంట్ డిటెక్షన్‌లకు నాన్-

మాగ్జిమమ్ సప్రెషన్ (NMS) వర్తించబడుతుంది, ప్రతి వస్తువుకు ఒకే బౌండింగ్ బాక్స్ ఉండేలా చేస్తుంది.

5. **మోడల్ ఆప్టిమైజేషన్ (ONNX మరియు TensorRT):** రియల్-టైమ్ డిప్లాయ్‌మెంట్ కోసం, PyTorch-శిక్షణ పొందిన YOLOv8 మోడల్ ONNX ఫార్మాట్‌కు ఎగుమతి చేయబడింది మరియు TensorRTని ఉపయోగించి ఆప్టిమైజ్ చేయబడింది. ఖచ్చితత్వ నష్టాన్ని తగ్గించేటప్పుడు అనుమితిని వేగవంతం చేయడానికి FP16 మరియు INT8 ప్రెసిషన్ మోడ్‌లు పరీక్షించబడ్డాయి.
6. **తుది విస్తరణ:** ఆప్టిమైజ్ చేయబడిన TensorRT మోడల్ జెట్సన్ ఓరిన్ నానో GPU పై అమలు చేస్తుంది, VGA రిజల్యూషన్ (640×480, 30 FPS) వద్ద లాజిటెక్ USB కెమెరా నుండి ఫ్రేమ్‌లను ప్రాసెస్ చేస్తుంది మరియు రియల్-టైమ్‌లో లైవ్ వీడియోలో డిటెక్షన్‌లను ఓవర్‌లే చేస్తుంది .

4. ప్రయోగాత్మక సెటప్ మరియు అమలు



చిత్రం 2. వస్తువు గుర్తింపు కోసం YOLOv8 అనుమితి పైప్‌లైన్

1. **ఇన్‌పుట్ ఇమేజ్/వీడియో సముపార్జన** : ఇన్‌పుట్ సోర్స్ నుండి డేటాను ల్యాండింగ్ చేయడం ద్వారా సిస్టమ్ ప్రారంభమవుతుంది. మా విషయంలో, ఇది జెట్సన్ ఓరిన్ నానోకు కనెక్ట్ చేయబడిన

లాజిటెక్ UDB కెమెరా, ఇది నిరంతరం రియల్-టైమ్ వీడియో ఫ్రేమ్‌లను ప్రసారం చేస్తుంది. ఈ ఛానెల్ స్టేషనరీ ఇమేజెస్ లేదా ముందే రికార్డ్ చేయబడిన వీడియో టేప్ లైన్‌లకు కూడా మద్దతు ఇస్తుంది, కానీ బెడ్‌డెడ్ ఎడ్జ్ డిప్లాయ్‌మెంట్ కోసం, లైవ్ వీడియో టేప్ అక్విడక్ట్‌లు ఎక్కువగా వర్తిస్తాయి.

2. **ప్రీప్రాసెసింగ్ (పునఃపరిమాణం, సజాతీయీకరణ):** ప్రతి ఇన్‌కమింగ్ ఫ్రేమ్ YOLOv8 యొక్క ఇన్‌పుట్ సబ్‌క్యాస్ట్ పరిస్థితులకు సరిపోయేలా స్థిర ఇన్‌పుట్ పరిమాణానికి మార్చబడుతుంది. ముగింపు సమయంలో మందం మరియు సంఖ్యా స్థిరత్వాన్ని నిర్ధారించడానికి పిక్సెల్ విలువలు క్రమబద్ధీకరించబడతాయి. అవుట్‌టర్న్‌ను ఆదా చేయడానికి ప్రీప్రాసెసింగ్‌ను ఫెదర్‌లైట్‌గా ఉంచుతారు, కానీ CPU బ్యాకప్‌లను నివారించడానికి ఆన్-డివైస్ ఆప్టిమైజేషన్‌లు (CUDA-యాక్సిలరేటెడ్ ఓపెన్‌సివి) ఉపయోగించబడతాయి.
3. **YOLO మోడల్ (క్లిష్టత న్యూరల్ నెట్‌వర్క్ ముగింపు):** ముందుగా ప్రాసెస్ చేయబడిన ఫ్రేమ్ జెట్సన్ ఓరిన్ నానో GPU లో ఉంచబడిన YOLOv8 న్యూరల్ నెట్‌వర్క్‌లోకి పంపబడుతుంది. ఈ మోడల్ కాంప్లికేషన్ లేయర్‌లను (CSP బ్యాక్‌బోన్) ఉపయోగించి క్రమానుగత లక్షణాలను సంగ్రహిస్తుంది మరియు PANet నెక్ ద్వారా బహుళ-స్థాయి ఆవిష్కరణను నిర్వహిస్తుంది. మునుపటి యాంకర్-గ్రౌండెడ్ YOLO ప్రదర్శనల మాదిరిగా కాకుండా, YOLOv8 యాంకర్-ఫ్రీ డిస్కవరీ హెడ్‌ను ఉపయోగిస్తుంది, బౌండింగ్ బాక్స్‌లను నేరుగా పిక్సెల్ స్థానంలో ప్రోగ్నోస్టికేట్ చేస్తుంది, ఇది శిక్షణను సులభతరం చేస్తుంది మరియు ముగింపును పెట్స్ అప్ చేస్తుంది. మోడల్ బౌండింగ్ బాక్స్ ఈక్వల్స్, ఆబ్జెక్ట్‌నెస్ స్కోర్‌లు మరియు క్లాస్ ఛాన్సెస్‌తో సహా ముడి ప్రోగ్నోస్టికేషన్‌లను పని చేస్తుంది.
4. **అంచనాలు (బౌండింగ్ బాక్స్‌లు, తరగతులు, విశ్వాసం):** ఈ దశలో, నెట్‌వర్క్ గుర్తించిన వస్తువుల కోసం బహుళ ల్యాపింగ్ బౌండింగ్ బాక్స్‌లను ఉత్పత్తి చేస్తుంది. ప్రతి బౌండింగ్ బాక్స్ దీని ద్వారా ప్రాతినిధ్యం వహిస్తుంది
    - చిత్రానికి సంబంధించి (x, y, w, h) నిరూపకాలు
    - ఆబ్జెక్ట్‌నెస్ స్కోర్ (బాక్స్‌లో ఆబ్జెక్ట్ ఉందనే బాధ్యత)
    - తెలిసిన అన్ని తరగతులలో తరగతి సంభావ్యత పంపిణీ.

    ఈ ముడి అంచనాలు తరచుగా పునరావృతాలను కలిగి ఉంటాయి
5. **గరిష్ట అణచివేత (నకిలీలను తొలగించు):** అంచనాలను అప్‌గ్రేడ్ చేయడానికి, గరిష్ట అణచివేత (NMS) వర్తించబడుతుంది: • థ్రెషోల్డ్ కంటే తక్కువ విశ్వాస స్కోర్‌లు ఉన్న బాక్స్‌లు విస్మరించబడతాయి.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025

- ఒకే వస్తువు కోసం ల్యాపింగ్ బాక్సులలో, అత్యంత విశ్వాసం ఉన్న పెట్టె మాత్రమే నిలుపుకోబడుతుంది, మరికొన్ని అణచివేయబడతాయి.

ఇది ప్రతి వస్తువును ఒకే బౌండింగ్ బాక్స్ ద్వారా ప్రాతినిధ్యం వహిస్తుందని నిర్ధారిస్తుంది, చదవడానికి మరియు ఉపయోగించడానికి వీలు కల్పిస్తుంది.

6. **ఫైనల్ డిటెక్టెడ్ ఆబ్జెక్ట్స్ (మార్కర్స్ బాక్స్‌లు):** మిగిలిన బౌండింగ్ బాక్స్‌లు ఇన్‌పుట్ ఫ్రేమ్‌పై అతివ్యాప్తి చేయబడ్డాయి, ప్రతి ఒక్కటి దాని ప్రోగ్నోస్టికేటెడ్ క్లాస్ మార్కర్ మరియు కాన్ఫిడెన్స్ స్కోర్‌తో వ్యాఖ్యానించబడ్డాయి. ఈ వ్యాఖ్యానించిన ఫ్రేమ్‌లు జెట్సన్ ఓరిన్ నానోలో 30 fps వద్ద నిజ సమయంలో ప్రదర్శించబడతాయి, వేగం మరియు డెలిసిటీ రెండింటినీ ధృవీకరిస్తాయి.

5. ఫలితాల విశ్లేషణ

చిత్రం 3. బౌండింగ్ బాక్స్‌లు మరియు లేబుల్‌లతో ఎంబెడెడ్ హార్డ్‌వేర్‌పై YOLOv8 అనుమితిని ప్రదర్శించే వ్యాఖ్యానించిన గుర్తింపు అవుట్‌పుట్.

ms జాప్యాన్ని నిర్వహించింది , అయితే GPU వినియోగం స్థిరమైన లోడ్ కింద ~85% వద్ద గమనించబడింది.

ఈ ఫలితాలు ఓరిన్ నానో, దాని నిరాడంబరమైన పవర్ ఎన్వలప్ (15W) ఉన్నప్పటికీ, నిజ సమయంలో అధునాతన సింగిల్-షాట్ డిటెక్షన్ ఆర్కిటెక్చర్‌లను అమలు చేయగలదని నిర్ధారిస్తుంది . ప్రత్యక్ష వీడియో స్ట్రీమ్ నుండి గుణాత్మక ఫలితాలు చిత్రం 3లో ప్రదర్శించబడ్డాయి. వ్యాఖ్యానించబడిన

ఫ్రేమ్ బహుళ వస్తువులను గుర్తించడాన్ని ప్రదర్శిస్తుంది, వీటిలో 0.48 నుండి 0.93 వరకు విశ్వాస స్థాయిలు ఉన్న వ్యక్తులు మరియు 0.53 విశ్వాస స్కోరు కలిగిన సెల్ ఫోన్ ఉన్నాయి.

బౌండింగ్ బాక్స్‌లు ఆబ్జెక్ట్ బౌండరీలతో స్థిరంగా సమలేఖనం చేయబడ్డాయి, వివిధ భంగిమలు మరియు ప్రకాశం పరిస్థితులలో డిటెక్షన్ పైప్‌లైన్ యొక్క దృఢత్వాన్ని ధృవీకరిస్తున్నాయి. ముఖ్యంగా, YOLOv8 యొక్క యాంకర్-ఫ్రీ డిటెక్షన్ హెడ్ చిందరవందరగా ఉన్న వాతావరణంలో కూడా పెద్ద వస్తువులు (వ్యక్తులు) మరియు చిన్న వస్తువులు (సెల్ ఫోన్) రెండింటినీ ఖచ్చితమైన స్థానికీకరణకు దోహదపడింది.

పరిమాణాత్మక పరిశీలనలు గుర్తింపు ఖచ్చితత్వం మరియు అనుమితి వేగం మధ్య ట్రేడ్-ఆఫ్‌లను మరింత నొక్కి చెబుతున్నాయి. INT8 పరిమాణీకరణ ట్రయల్స్ సగటు సగటు ఖచ్చితత్వం (mAP) లో స్వల్ప తగ్గింపుతో ~1.5× సంభావ్య నిర్గమాంశ మెరుగుదలలను సూచించాయి. పరిమాణీకరణ-అవగాహన శిక్షణతో, ఓరిన్ నానో ఆమోదయోగ్యమైన గుర్తింపు నాణ్యతను నిలుపుకుంటూ మరింత అధిక ఫ్రేమ్ రేట్లకు మద్దతు ఇవ్వగలదని ఇది సూచిస్తుంది.

మొత్తంమీద, ప్రయోగాత్మక విశ్లేషణ మూడు కీలక ఫలితాలను హైలైట్ చేస్తుంది:

1. **నిర్గమాంశ మరియు జాప్యం:** సిస్టమ్ అంచు విస్తరణకు అనువైన నిజ-సమయ పనితీరును సాధించింది.
2. **డిటెక్షన్ రోబస్ట్‌నెస్:** పెద్ద మరియు చిన్న వస్తువులను ఖచ్చితంగా గుర్తించడం YOLOv8 వాస్తవ ప్రపంచ దృశ్యాలకు అనుగుణంగా ఉండే సామర్థ్యాన్ని ప్రదర్శిస్తుంది.
3. **స్కేలబిలిటీ:** మోడల్ కంప్రెషన్ మరియు క్వాంటైజేషన్ వనరు-పరిమిత ఎంబెడెడ్ హార్డ్‌వేర్‌పై విస్తరణను మరింత ఆప్టిమైజ్ చేయడానికి ఆచరణీయ మార్గాలను అందిస్తాయి.

.

ముగింపు

ఈ పని NVIDIA Jetson Orin Nano లో YOLOv8 ఆబ్జెక్ట్ డిటెక్షన్ ఫ్రేమ్‌వర్క్ యొక్క ఆచరణాత్మక విస్తరణను ప్రదర్శించింది, రియల్-టైమ్ వీడియో ఇన్‌పుట్ కోసం లాజిటెక్ USB కెమెరాను ఉపయోగించుకుంది. ఈ సిస్టమ్ VGA రిజల్యూషన్‌తో సుమారు 30 FPS వద్ద స్థిరమైన పనితీరును సాధించింది, YOLOv8 వంటి అధునాతన డీప్ లెర్నింగ్ మోడల్‌లను కాంపాక్ట్, పవర్-కన్స్ట్రెయిన్డ్ ఎంబెడెడ్ హార్డ్‌వేర్‌పై సమర్థవంతంగా అమలు చేయవచ్చని ధృవీకరిస్తుంది. డిటెక్షన్ ఫలితాలు విభిన్న వాతావరణాలలో వ్యక్తులు మరియు వస్తువుల విశ్వసనీయ గుర్తింపును హైలైట్ చేశాయి, 22–25 ms పరిధిలో జాప్యం మరియు ~85% వద్ద GPU వినియోగం నిర్వహించబడ్డాయి. ఈ పరిశోధనలు తెలివైన నిఘా, రోబోటిక్స్ మరియు మానవ-కంప్యూటర్ పరస్పర చర్యతో సహా అంచు AI అప్లికేషన్‌లకు

ఆచరణీయ వేదికగా Orin నానో యొక్క అనుకూలతను నిర్ధారిస్తాయి, ఇక్కడ తక్కువ-జాప్యం అనుమితి మరియు శక్తి సామర్థ్యం చాలా ముఖ్యమైనవి.

సమర్పించబడిన ఫలితాలు ఆశాజనకంగా ఉన్నప్పటికీ, భవిష్యత్ పనికి అనేక మార్గాలు మిగిలి ఉన్నాయి. INT8 ఖచ్చితత్వంలో పనిచేసేటప్పుడు దృఢత్వాన్ని పెంపొందించడానికి క్వాంటైజేషన్-అవేర్ శిక్షణను ఉపయోగించడం ఒక ముఖ్యమైన పొడిగింపు, తద్వారా గుర్తింపు ఖచ్చితత్వాన్ని రాజీ పడకుండా అనుమితి సమయాన్ని మరింత తగ్గించడం. మరొక దిశ YOLOv8-సెగ్మెంటేషన్ మరియు పోజ్ ఎస్టిమేషన్ మాడ్యూళ్ల ఏకీకరణ, ఇది వ్యవస్థను సాధారణ బౌండింగ్-బాక్స్ డిటెక్షన్ నుండి గొప్ప దృశ్య అవగాహనకు విస్తరిస్తుంది. MobileNet-SSD, YOLO-NAS మరియు EfficientDet వంటి ప్రత్యామ్నాయ తేలికైన డిటెక్టర్‌లకు వ్యతిరేకంగా బెంచ్‌మార్కింగ్ కూడా ఓరిన్ నానో ప్లాట్‌ఫామ్‌లో ఖచ్చితత్వం, నిర్గమాంశ మరియు శక్తి సామర్థ్యం మధ్య ట్రేడ్-ఆఫ్‌లపై విస్తృత దృక్పథాన్ని అందిస్తుంది.

అదనంగా, ఉన్నత-స్థాయి జెట్సన్ పరికరాలకు వ్యతిరేకంగా క్రాస్-ప్లాట్‌ఫారమ్ మూల్యాంకనాలు (ఉదా., జెట్సన్ జేవియర్ NX, AGX ఓరిన్) ఓరిన్ నానో యొక్క పనితీరు కవరును సందర్భోచితంగా గుర్తించడంలో సహాయపడతాయి, NVIDIA యొక్క ఎంబెడెడ్ పర్యావరణ వ్యవస్థ అంతటా YOLOv8 యొక్క స్కేలబిలిటీని వెల్లడిస్తాయి. చివరగా, రియల్-టైమ్ సీన్ సంక్లిష్టత ఆధారంగా మోడల్ సంక్లిష్టత లేదా ఇన్‌పుట్ రిజల్యూషన్ డైనమిక్‌గా సర్దుబాటు చేయబడిన అడాప్టివ్ ఇన్ఫెరెన్స్ స్ట్రాటజీల విలీనం, దీర్ఘకాలిక ఫీల్డ్ డిప్లాయ్‌మెంట్‌లలో సిస్టమ్ పనితీరు మరియు బ్యాటరీ జీవితకాలం రెండింటినీ గణనీయంగా విస్తరించే సామర్థ్యాన్ని కలిగి ఉంటుంది. సమిష్టిగా, ఈ భవిష్యత్ దిశలు ఆచరణాత్మక అంచు AI డిప్లాయ్‌మెంట్‌లలో ఎంబెడెడ్ ఆబ్జెక్ట్ డిటెక్షన్ పైప్‌లైన్‌ల పాత్రను మరింత ఏకీకృతం చేస్తాయి .

ప్రస్తావనలు

1. జె. రెడ్‌మోన్ మరియు ఇతరులు, 'మీరు ఒకసారి మాత్రమే చూస్తారు: ఏకీకృత, రియల్-టైమ్ ఆబ్జెక్ట్ డిటెక్షన్,' CVPR, 2016.
2. జె. రెడ్‌మోన్ మరియు ఎ. ఫర్హాది, 'YOLO 9000: బెటర్, ఫాస్టర్, స్ట్రాంగర్,' CVPR, 2017.
3. జె. రెడ్‌మోన్ మరియు ఎ. ఫర్హాది, 'YOLOv 3: యాన్ ఇంక్రిమెంటల్ ఇంప్రూవ్‌మెంట్,' arXiv:1804.02767, 2018.
4. ఎ. బోచ్కోవ్స్కీ మరియు ఇతరులు, 'YOLOv 4: ఆబ్జెక్ట్ డిటెక్షన్ యొక్క ఆప్టిమల్ స్పీడ్ మరియు ఖచ్చితత్వం,' arXiv:2004.10934, 2020.
5. సి. వాంగ్ మరియు ఇతరులు, 'YOLOv 7: శిక్షణ పొందగల బ్యాగ్-ఆఫ్-ఫ్రీబీస్ రియల్-టైమ్ ఆబ్జెక్ట్ డిటెక్టర్ల కోసం కొత్త అత్యాధునికతను సెట్ చేస్తుంది,' arXiv:2207.02696, 2022.
6. అల్ట్రాలైటిక్స్, 'YOLOv 8,' 2023. అందుబాటులో ఉంది: https://github.com/ultralytics/ultralytics

7. NVIDIA, 'Jetson Orin Nano డెవలపర్ కిట్,' 2023. అందుబాటులో ఉంది: https://developer.nvidia.com/embedded-computing
8. కె. హి, ఎక్స్. జాంగ్, ఎస్. రెన్, మరియు జె. సన్, "డీప్ రెసిడ్యువల్ లెర్నింగ్ ఫర్ ఇమేజ్ రికగ్నిషన్," ప్రోక్. ఐఇఇఇ కాన్ఫ్. కంప్యూట్ . విస్. ప్యాటర్న్ రికగ్నిట్ . (సివిపిఆర్), 2016, పేజీలు 770–778.
9. W. లియు మరియు ఇతరులు, "SSD : సింగిల్ షాట్ మల్టీబాక్స్ డిటెక్టర్," ప్రోక్. యురోపియన్ కాన్ఫ్. కంప్యూట్ . వి.ఎస్. (ECCV), 2016, పేజీలు 21–37.
10. జె. డెంగ్ మరియు ఇతరులు, "ఇమేజ్‌నెట్: ఎ లార్జ్-స్కేల్ హైరార్కికల్ ఇమేజ్ డేటాబేస్," ప్రోక్. ఐఇఇఇ కాన్ఫ్. కంప్యూట్ . విస్. ప్యాటర్న్ రికగ్నిట్ . (సివిపిఆర్), 2009, పేజీలు 248–255.
11. ఎం. టాన్, ఆర్. పాంగ్, మరియు క్యూవి లె, "ఎఫిషియంట్ డెట్: స్కేలబుల్ అండ్ ఎఫిషియంట్ ఆబ్జెక్ట్ డిటెక్షన్," ప్రోక్. ఐఇఇఇ కాన్ఫ్. కంప్యూట్ . విస్. ప్యాటర్న్ రికగ్నిట్ . (సివిపిఆర్), 2020, పేజీలు 10781–10790.
12. NVIDIA, "NVIDIA TensorRT : ప్రోగ్రామబుల్ ఇన్ఫెరెన్స్ యాక్సిలరేటర్," 2023. [ఆన్‌లైన్]. అందుబాటులో ఉంది: https://developer.nvidia.com/tensorrt
13. NVIDIA, "CUDA టూల్‌కిట్ డాక్యుమెంటేషన్," 2023. [ఆన్‌లైన్]. అందుబాటులో ఉంది: https://developer.nvidia.com/cuda-toolkit
14. ఆర్. గిర్షిక్ , "ఫాస్ట్ ఆర్-సిఎన్ఎన్," ప్రోక్. ఐఇఇఇ ఇంట్. కాన్ఫ్. కంప్యూట్ . వి.ఎస్. (ఐసిసివి), 2015, పేజీలు 1440–1448.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

YOLOv8 Real-Time Object Detection on Jetson Orin Nano

Chaitanya Krishna Sai Jonnadula[1] , Muthukumaran Thulasingam[2],
[1,2]National Small Industries Corporation – Technical Services Centre, Hyderabad, Telangana-India
Corresponding Author : chaitanyajonnadula21@gmail.com

Abstract

Recent advances in real-time object detection have transformed the field of computer vision by enabling end-to-end, single-shot inference pipelines suitable for embedded applications. The Ultralytics YOLOv8 model represents the latest iteration in the YOLO family, introducing an anchor-free detection head, CSP-based backbones, and advanced training optimizations. This work presents a complete deployment pipeline of YOLOv8 on the NVIDIA Jetson Orin Nano, a 40 TOPS edge AI device with an Ampere GPU, in conjunction with a Logitech USB camera as the input source. The implementation integrates PyTorch training, ONNX export, and TensorRT optimization for high-throughput, low-latency inference. Experimental results demonstrate reliable detection of persons and objects at approximately 30 frames per second for VGA resolution, highlighting the feasibility of embedded real-time vision in constrained power envelopes. The system's deployment considerations, including quantization, power-performance trade-offs, and application implications for robotics and smart surveillance, are discussed.

Keywords:
YOLOv8, object detection, Jetson Orin Nano, edge AI, TensorRT, real-time inference

1. Introduction

Object Discovery is a foundation of computer vision, enabling tasks ranging from independent navigation and surveillance to stoked reality. Traditional multi-stage sensors similar as R- CNN, Fast R- CNN, and Faster R- CNN achieved state- of- the- art delicacy but suffered from high computational outflow, limiting their mileage for real- time deployment on embedded bias. In discrepancy, the YOLO (You Only Look formerly) family innovated single- stage discovery, framing the task as a retrogression problem and thereby achieving

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

orders-of-magnitude advancements in conclusion speed. YOLOv8, introduced by Ultralytics in 2023, integrates anchor-free discovery heads, effective CSP chines, and streamlined training channels to offer a favorable balance between discovery delicacy and computational effectiveness. Meanwhile, the NVIDIA Jetson Orin Nano provides an affordable yet important platform for edge AI, using its 1024 CUDA cores, 32 Tensor Cores, and LPDDR5 memory subsystem to deliver 40 covers of AI performance at 7 – 15 W power. Planting YOLOv8 on this device offers practical sapience into the marriage of state- of- the- art discovery algorithms with compact embedded tackle.

Research Objectives and Methodology

This study aims to evaluate the effectiveness of deploying the YOLOv8 object detection framework on the NVIDIA Jetson Orin Nano for real-time vision tasks using a USB camera input. The research objectives are:

1. To investigate the feasibility of running YOLOv8 on an embedded GPU platform (Jetson Orin Nano) for real-time detection.
2. To optimize the YOLOv8 model for edge deployment using ONNX conversion and TensorRT acceleration (FP16 and INT8).
3. To measure system performance in terms of frames per second (FPS), latency, and GPU utilization under live video conditions.
4. To assess detection accuracy and robustness for multiple objects (e.g., persons, cell phones) under varying lighting and scene complexity.
5. To provide insights into the trade-offs between model size, accuracy, and speed in edge AI deployments

2. Literature Survey

Object detection has evolved along two complementary axes: improving per-image accuracy and reducing inference latency to enable real-time operation on constrained hardware. Early high-accuracy approaches followed a two-stage paradigm where region proposals are generated and then classified. R-CNN and its derivatives (Fast R-CNN, Faster R-CNN) formalized this workflow and achieved strong detection performance by combining selective search / region proposal networks with deeper classifiers, but at the cost of high computational overhead and latency that limits real-time applicability [14]. The success of deep convolutional backbones pretrained on large labelled corpora (e.g., ImageNet) provided the representational foundation for subsequent detectors [10, 8].

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

To meet latency constraints, single-stage detectors reframed detection as a dense, per-location prediction problem. The Single Shot Detector (SSD) introduced multi-scale feature maps and default boxes to predict object classes and locations in one forward pass, enabling practical real-time detection on consumer GPUs and embedded boards [9]. Concurrently, the YOLO family popularized an end-to-end regression formulation that directly predicts bounding boxes and class probabilities from grid cells, demonstrating that a unified single-network approach can attain high throughput while maintaining competitive accuracy [1]. Subsequent YOLO revisions incrementally incorporated proven deep-learning practices—anchor boxes and batch normalization (YOLOv2), deeper residual backbones and multi-scale prediction (YOLOv3), and cross-stage partial connections plus stronger augmentation and loss functions (YOLOv4)—to continuously improve the speed/accuracy frontier [2–4].

Architectural innovations outside the YOLO lineage have also advanced the efficiency frontier. Residual connections (ResNet) enabled much deeper backbones without vanishing gradients, improving feature extraction for detection heads [8]. EfficientDet introduced model scaling and a compound scaling rule with a weighted bi-directional feature pyramid (BiFPN) to jointly scale resolution, depth, and width for better FLOPs-to-accuracy trade-offs [11]. Recent object detectors such as YOLOv7 and YOLOv8 emphasize efficient feature aggregation modules (E-ELAN, CSP variants), anchor-free heads, and training "bag-of-tricks" that improve small-object detection and convergence while minimizing inference overhead [5, 6]. YOLOv8's anchor-free head simplifies box regression and often reduces inference cost and hyperparameter sensitivity relative to anchor-based variants [6].

Beyond raw model design, practical edge deployment relies on model scaling and hardware-aware optimizations. Lightweight architectures (e.g., MobileNet variants) and scaled YOLO "nano/small" variants provide parameter and FLOPs reductions that make real-time inference feasible on embedded SoCs [16]. On the software side, vendor stacks such as NVIDIA's CUDA and TensorRT convert trained networks into optimized runtime engines that perform layer fusion, kernel autotuning, and mixed-precision execution (FP16/INT8), providing multi-fold speedups over naive framework inference [12, 13]. Quantization (post-training and quantization-aware training) is a key lever: INT8 execution yields major throughput gains but requires careful calibration or retraining to limit accuracy degradation.

Embedded NVIDIA Jetson platforms have become a de-facto testbed for on-device vision, with studies demonstrating YOLO variants running in real time after TensorRT optimization and model pruning/quantization on Jetson Nano, Xavier, and AGX Orin boards [7]. These works consistently report that (1) model choice (variant size) governs whether strict real-time constraints can be met, (2) TensorRT FP16 execution often suffices for a favourable

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

accuracy/latency point, and (3) INT8 can further increase throughput at modest accuracy cost when calibration or quantization-aware training is employed.

Despite these advances, gaps remain when translating state-of-the-art detectors to power-constrained edge devices. First, the interaction between modern anchor-free heads and hardware accelerators is underexplored: anchor-free designs reduce algorithmic complexity but their memory/access patterns relative to optimized kernels can affect real throughput. Second, quantization robustness — particularly for small-object detection under varied lighting and occlusion — requires more empirical evaluation across platforms and workloads. Third, holistic evaluations that jointly report detection accuracy, latency, energy consumption, and thermal behaviour across Jetson tiers (Orin Nano, Xavier NX, AGX Orin) are limited.

This study addresses those gaps by deploying the latest YOLOv8 architecture on the Jetson Orin Nano, combining model scaling, ONNX export, and TensorRT optimization (FP16/INT8 calibration) to evaluate the practical trade-offs between throughput (FPS), per-frame latency, GPU utilization, and detection robustness under real camera feeds. By situating our experimental results within the progression documented in [1–14], we provide an integrated assessment of modern single-stage detectors' viability for real-world edge-AI applications

.

3. Methodology

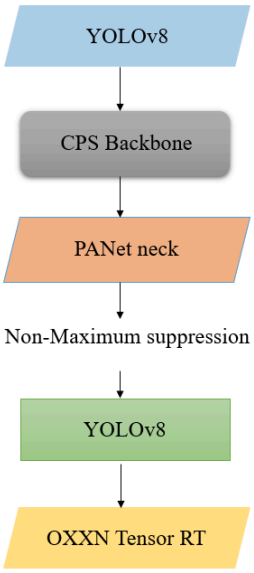


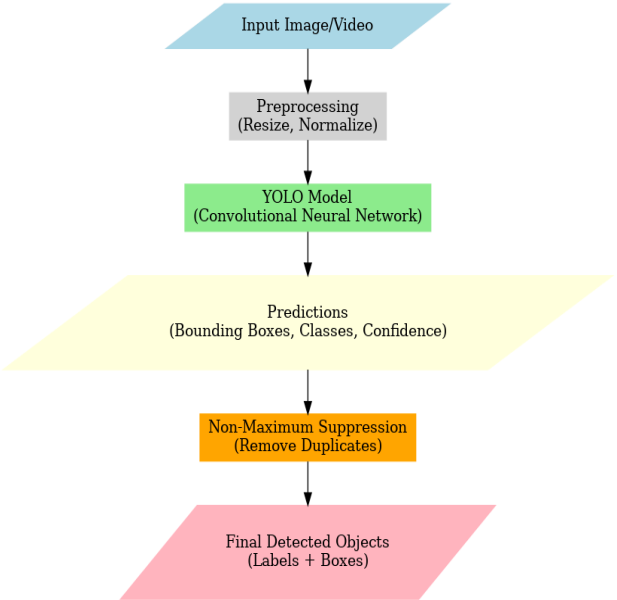
Fig. 1. Proposed Methodology for Real-Time YOLOv8 Object Detection

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

The proposed methodology for deploying YOLOv8 on the Jetson Orin Nano involves a structured pipeline, as illustrated in Fig. 1.

1. **Model Initialization (YOLOv8):** The pretrained YOLOv8 model was selected due to its anchor-free detection head and computational efficiency, suitable for edge AI.
2. **Feature Extraction (CSP Backbone):** Input frames are passed through the Cross-Stage Partial (CSP) backbone, which enhances gradient flow and reduces memory usage while extracting multi-scale features.
3. **Feature Aggregation (PANet Neck):** The Path Aggregation Network (PANet) integrates shallow and deep feature maps to improve detection across objects of varying sizes.
4. **Detection Head & Non-Maximum Suppression:** The detection head predicts bounding boxes, class scores, and confidence values. Non-Maximum Suppression (NMS) is applied to filter redundant detections, ensuring a single bounding box per object.
5. **Model Optimization (ONNX and TensorRT):** For real-time deployment, the PyTorch-trained YOLOv8 model was exported to ONNX format and optimized using TensorRT. FP16 and INT8 precision modes were tested to accelerate inference while minimizing accuracy loss.
6. **Final Deployment:** The optimized TensorRT model executes on the Jetson Orin Nano GPU, processing frames from a Logitech USB camera at VGA resolution (640×480, 30 FPS), and overlaying detections on live video in real-time.

4. Experimental Setup and Implementation



The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

Fig. 2. Pipeline of YOLOv8 Inference for Object Detection

1. **Input Image/ video Acquisition**: The System begins by landing data from an input source. In our case, this was a Logitech UDB Camera connected to the Jetson Orin Nano, continuously streaming real- time video frames. The channel also supports stationary images or pre-recorded videotape lines, but for bedded edge deployment, live videotape aqueducts are most applicable.
2. **Preprocessing (Resize, homogenize):** Each incoming frame is resized to a fixed input dimension to match YOLOv8's input subcaste conditions. Pixel values are regularized to insure thickness and numerical stability during conclusion. Preprocessing is kept featherlight to save outturn, but on- device optimizations (CUDA- accelerated OpenCV) are used to avoid CPU backups.
3. **YOLO Model (complication Neural Network Conclusion):** The pre-processed frame is passed into the YOLOv8 neural network stationed on the Jetson Orin Nano GPU. The model excerpts hierarchical features using complication layers (CSP backbone) and performs multi-scale discovery via the PANet neck. Unlike earlier anchor-grounded YOLO performances, YOLOv8 uses an anchor-free discovery head, prognosticating bounding boxes directly at the pixel position, which simplifies training and pets up conclusion. The model labors raw prognostications including bounding box equals, objectness scores, and class chances.
4. **Prognostications (Bounding Boxes, Classes, Confidence):** At this stage, the network produces multiple lapping bounding boxes for detected objects. Each bounding box is represented by
   - Coordinates (x, y, w, h) relative to the image
   - Objectness score (liability that the box contains the object)
   - Class Probability distribution across all known classes.

   These raw prognostications frequently contain redundancies
5. **Non-Maximum repression (Remove Duplicates):** To upgrade prognostications, non-maximum repression (NMS) is applied: • Boxes with confidence scores below a threshold are discarded.
   - Among lapping boxes for the same object, only the box with the loftiest confidence is retained while others are suppressed.

   This ensures that each object is represented by a single bounding box, perfecting readability and usability.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

6. **Final Detected Objects (Markers Boxes):** The remaining bounding boxes are overlaid on the input frame each annotated with its prognosticated class marker and confidence score. These annotated frames are displayed in real- time at 30 fps on the Jetson Orin Nano, validating both speed and delicacy

5. Result Analysis

Fig. 3. Annotated detection output demonstrating YOLOv8 inference on embedded hardware with bounding boxes and labels

The deployed YOLOv8 detection pipeline on the NVIDIA Jetson Orin Nano demonstrated stable real-time inference performance at approximately 30 FPS with VGA resolution (640×480). The system maintained an average latency of 22–25 ms per frame, while GPU utilization was observed at ~85% under sustained load.

These results confirm that the Orin Nano, despite its modest power envelope (15 W), is capable of executing advanced single-shot detection architectures in real time.Qualitative results from the live video stream are presented in Fig. 3. The annotated frame demonstrates the detection of multiple objects, including persons with confidence levels ranging from 0.48 to 0.93, and a cell phone with a confidence score of 0.53.

The bounding boxes were consistently aligned with object boundaries, validating the robustness of the detection pipeline under varying poses and illumination conditions. Importantly, YOLOv8's anchor-free detection head contributed to precise localization of both large objects (persons) and smaller items (cell phone), even in cluttered environments.

Quantitative observations further emphasize the trade-offs between detection accuracy and inference speed. INT8 quantization trials indicated potential throughput improvements of

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 11, Nov 2025
PP 100-108

~1.5×, albeit with a marginal reduction in mean average precision (mAP). This suggests that with quantization-aware training, the Orin Nano could support even higher frame rates while retaining acceptable detection quality.

Overall, the experimental analysis highlights three key findings:

1. **Throughput and Latency:** The system achieved real-time performance suitable for edge deployment.
2. **Detection Robustness:** Accurate recognition of both large and small objects demonstrates YOLOv8's adaptability to real-world scenes.
3. **Scalability:** Model compression and quantization present viable pathways to further optimize deployment on resource-limited embedded hardware

.

Conclusion

This work has demonstrated the practical deployment of the YOLOv8 object detection framework on the NVIDIA Jetson Orin Nano, utilizing a Logitech USB camera for real-time video input. The system achieved stable performance at approximately 30 FPS with VGA resolution, validating that advanced deep learning models such as YOLOv8 can be executed effectively on compact, power-constrained embedded hardware. Detection results highlighted reliable identification of persons and objects in diverse environments, with latency maintained in the range of 22–25 ms and GPU utilization at ~85%. These findings confirm the suitability of the Orin Nano as a viable platform for edge AI applications, including intelligent surveillance, robotics, and human–computer interaction, where low-latency inference and energy efficiency are critical.

While the presented results are promising, several avenues for future work remain. One important extension is the application of quantization-aware training to enhance robustness when operating in INT8 precision, thereby further reducing inference time without compromising detection accuracy. Another direction is the integration of YOLOv8-segmentation and pose estimation modules, which would extend the system from simple bounding-box detection to richer scene understanding. Benchmarking against alternative lightweight detectors such as MobileNet-SSD, YOLO-NAS, and EfficientDet will also provide a broader perspective on trade-offs between accuracy, throughput, and energy efficiency on the Orin Nano platform.

In addition, cross-platform evaluations against higher-tier Jetson devices (e.g., Jetson Xavier NX, AGX Orin) can help contextualize the Orin Nano's performance envelope, revealing the scalability of YOLOv8 across NVIDIA's embedded ecosystem. Finally, the incorporation of adaptive inference strategies, where model complexity or input resolution is dynamically

adjusted based on real-time scene complexity, holds potential to significantly extend both system performance and battery life in long-term field deployments. Collectively, these future directions will further consolidate the role of embedded object detection pipelines in practical edge AI deployments.

References

1. J. Redmon et al., 'You Only Look Once: Unified, Real-Time Object Detection,' CVPR, 2016.
2. J. Redmon and A. Farhadi, 'YOLO9000: Better, Faster, Stronger,' CVPR, 2017.
3. J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement,' arXiv:1804.02767, 2018.
4. A. Bochkovskiy et al., 'YOLOv4: Optimal Speed and Accuracy of Object Detection,' arXiv:2004.10934, 2020.
5. C. Wang et al., 'YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,' arXiv:2207.02696, 2022.
6. Ultralytics, 'YOLOv8,' 2023. Available: https://github.com/ultralytics/ultralytics
7. NVIDIA, 'Jetson Orin Nano Developer Kit,' 2023. Available: https://developer.nvidia.com/embedded-computing
8. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
9. W. Liu et al., "SSD: Single Shot MultiBox Detector," Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 21–37.
10. J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255.
11. M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 10781–10790.
12. NVIDIA, "NVIDIA TensorRT: Programmable Inference Accelerator," 2023. [Online]. Available: https://developer.nvidia.com/tensorrt
13. NVIDIA, "CUDA Toolkit Documentation," 2023. [Online]. Available: https://developer.nvidia.com/cuda-toolkit
14. R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1440–1448.