



신경망 자동 탐색 기술

성명 이재성

소속 중앙대

인공지능 기술의 대중화
(AI Democratization)를 위한
제2회 탱고 커뮤니티 컨퍼런스





목 차

1

기술 개요

03

1. 관련 기술 및 기술 개발 범위
2. 신경망 자동 탐색 기술 프로세스

2

개발 내용

07

1. YOLO 기반 신경망 자동 탐색 기술
2. 모바일 특화 신경망 탐색 기술

3

프레임워크

14

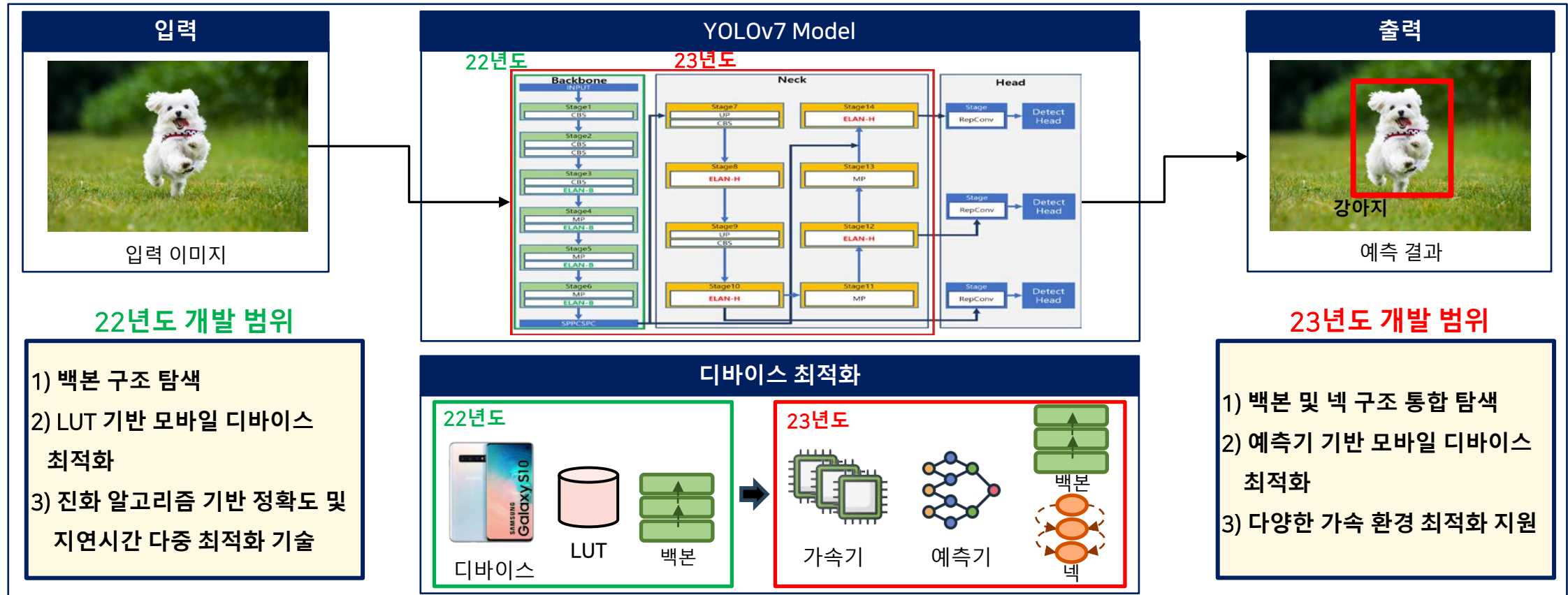
1. 신경망 자동 탐색 모듈 시뮬레이션



관련 기술 및 기술 개발 범위

객체 검출 및 기술 개발 범위

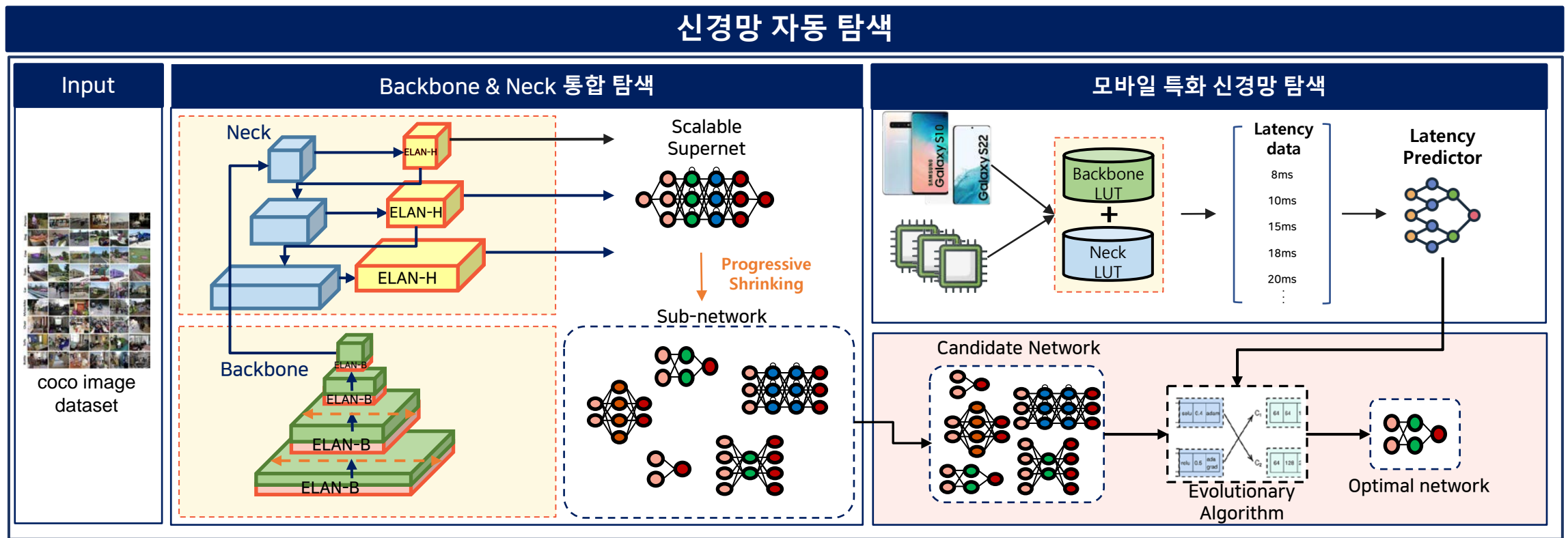
- 최신 객체 검출 YOLOv7 기반 신경망 자동 탐색 기술 개발



신경망 자동 탐색 기술 프로세스

백본과 넥 신경망 탐색을 통합한 신경망 자동 생성 기술

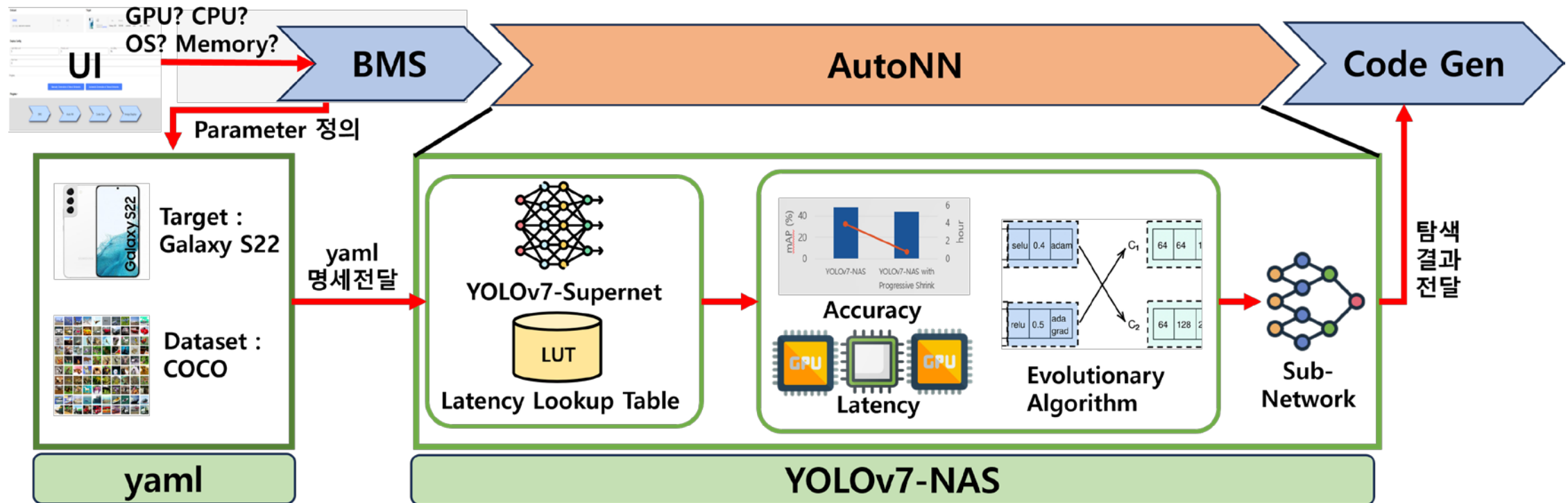
- 백본과 넥의 동시 탐색을 통해 객체 탐지 신경망의 구성 요소와 연결 구조를 고려하여 최적의 신경망 구조 탐색 가능



신경망 자동 탐색 기술 프로세스

TANGO 프레임워크 내 AutoNN (신경망 자동 탐색) 단계 개요

- 진화 알고리즘 기반 다중 최적화 기술을 통해 사용자 요구사항에 맞는 최적의 Sub-network를 탐색하는 단계
- TANGO UI로부터 입력된 요구사항과 BMS 단계에서 전달된 yaml 파일을 활용

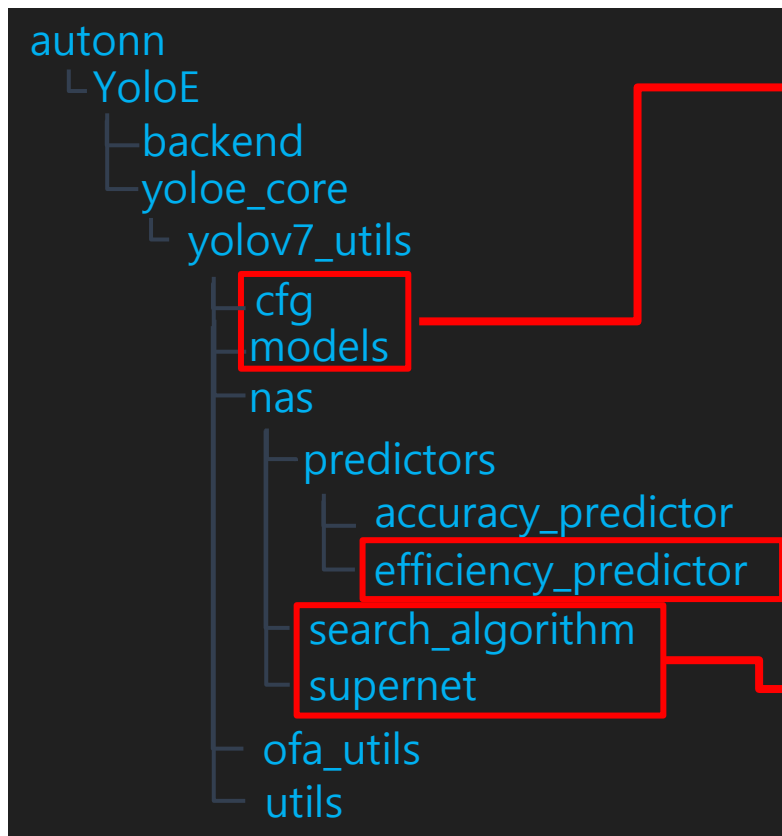




신경망 자동 탐색 기술 프로세스

TANGO 프레임워크 내 디렉토리 트리

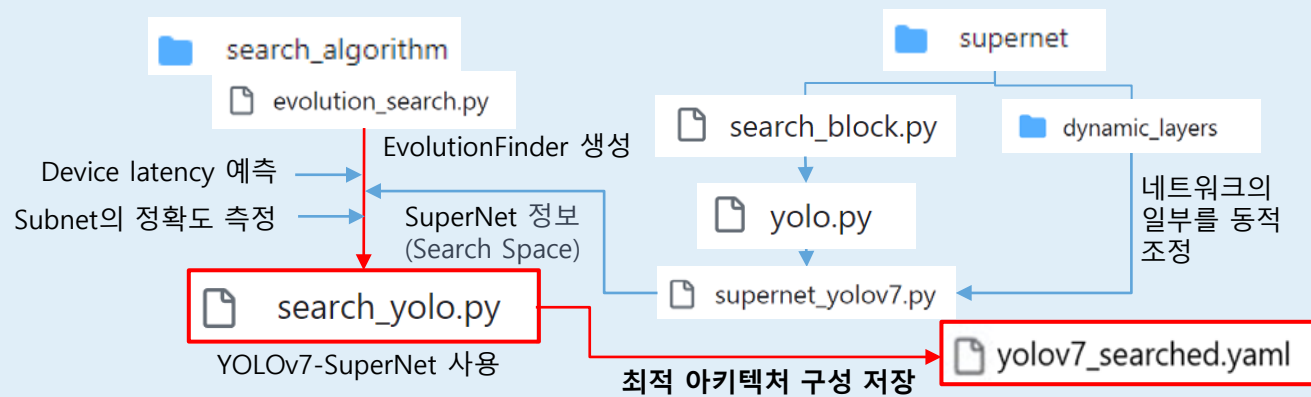
- TANGO 프레임워크 활용을 위한 통합된 디렉토리 트리



- Base model : YOLOv7-Supernet, YOLOv7 저장
- Hyperparameter : 노코드 지원을 위해 default로 저장된 값 활용
- 필요시 해당 디렉토리내 yaml 파일 수정을 통해 변경 가능

- Device의 LUT를 기반으로 학습한 pt파일 및 efficiency network 구조 저장
- 다른 device의 LUT가 있다면 해당 디렉토리 내 파일을 활용해 재학습 가능

- Base model을 기반으로 backbone 및 neck 구조 탐색 수행

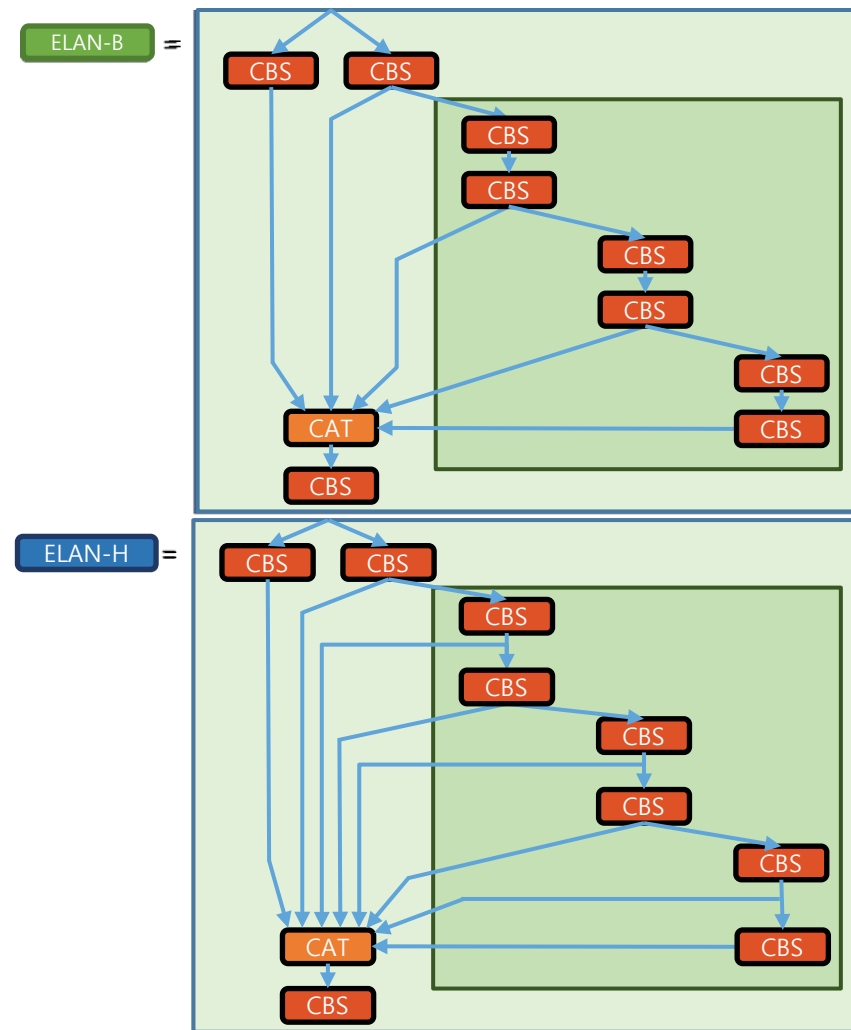
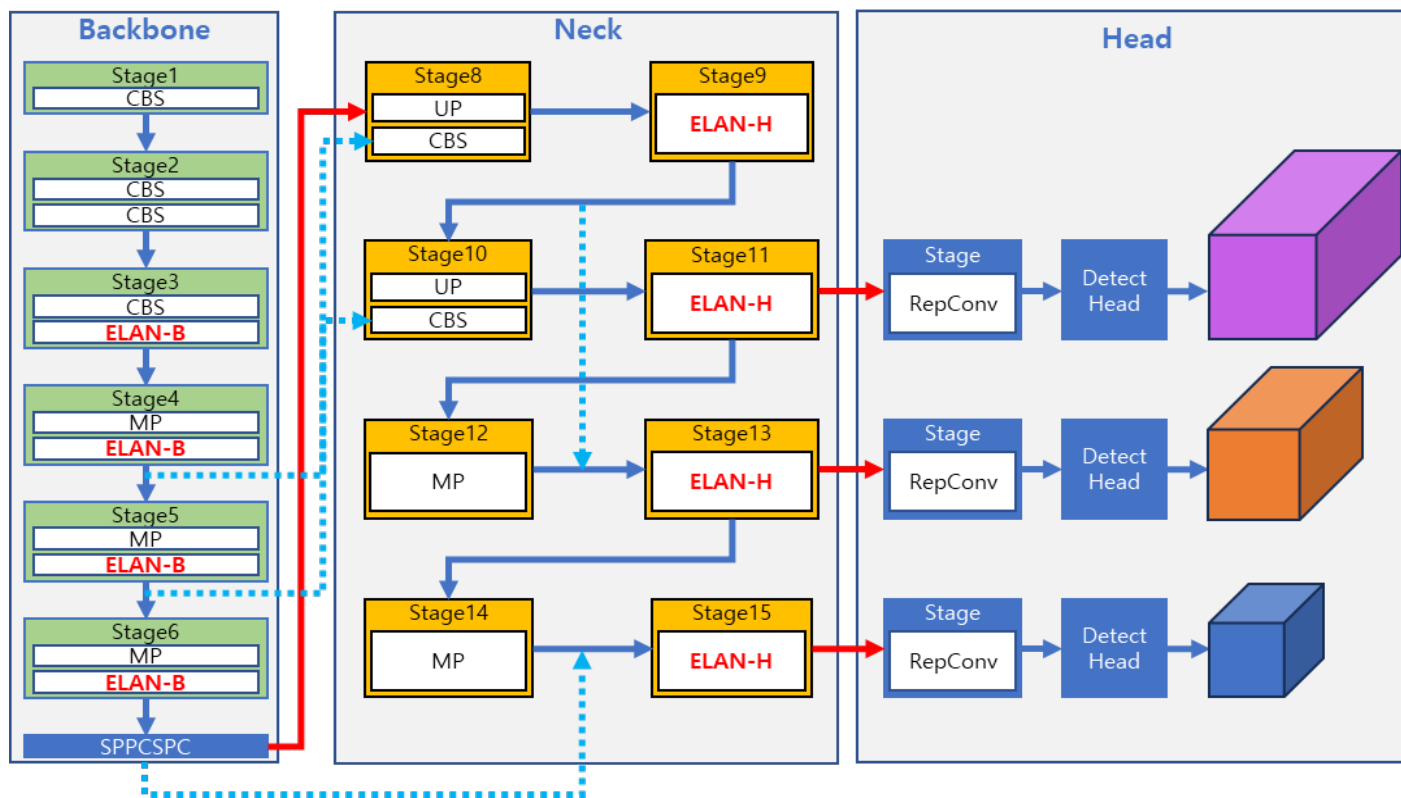




YOLO 기반 신경망 자동 탐색 기술

탐색공간(Search Space) 정의

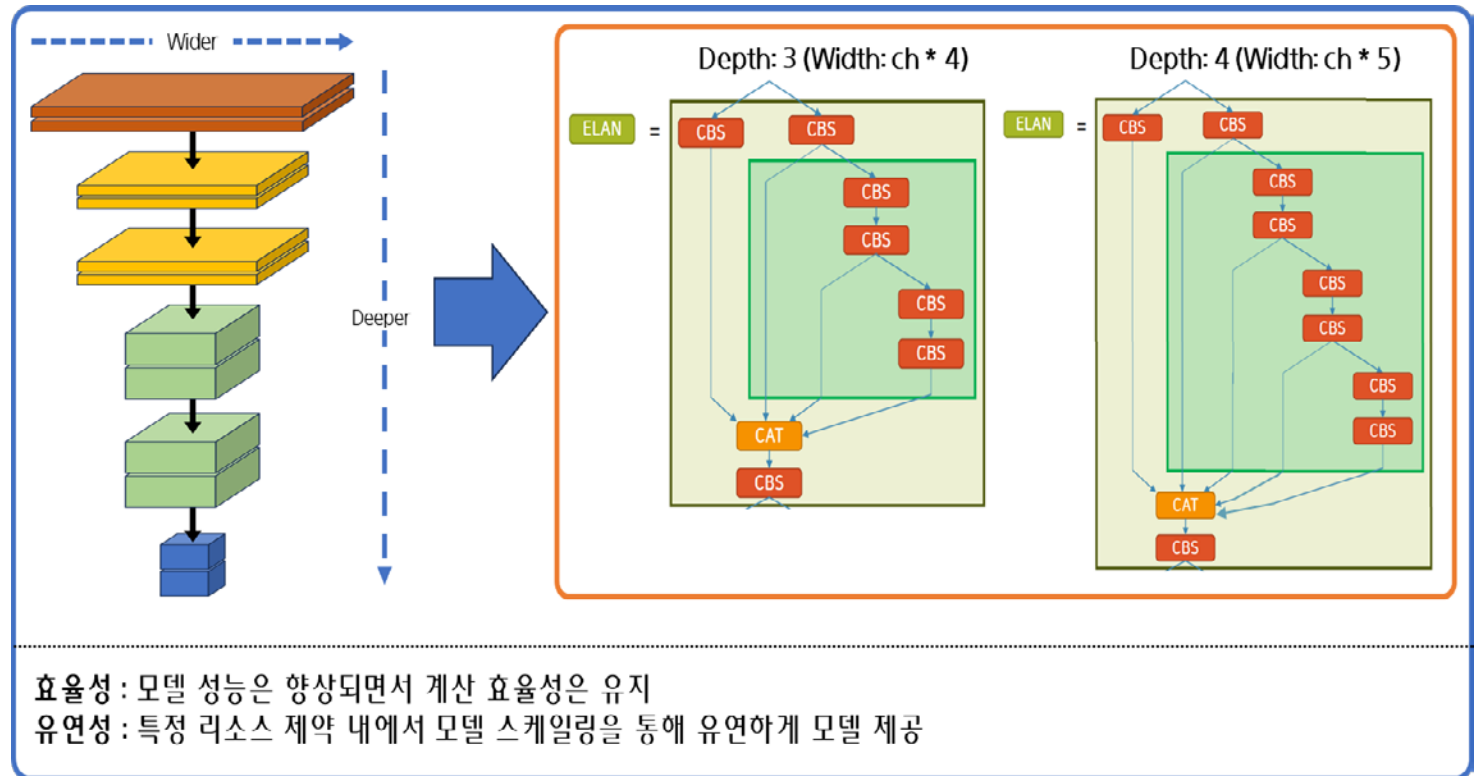
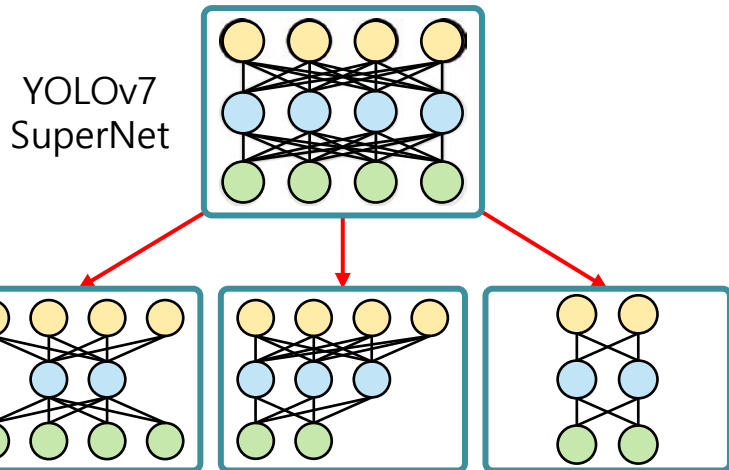
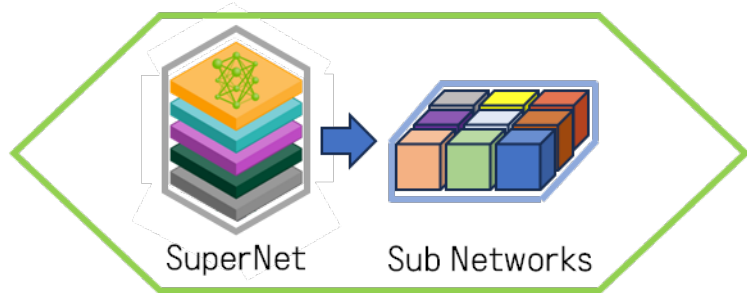
- YOLOv7의 연산자 블록과 전체 신경망 구조 고려
- 백본과 넥 신경망 크기 조절이 용이한 탐색공간 정의



YOLO 기반 신경망 자동 탐색 기술

백본과 넥 구조 통합 탐색

- YOLOv7 기반 백본과 넥 동시에 탐색 가능한 YOLOv7-SuperNet 구성 (YOLOv7 & YOLOv7-tiny 지원)
- ELAN 모듈 기반 Compound Scaling 조합 탐색 방법 적용 (Depth를 조절하면 자동적으로 width 조절)

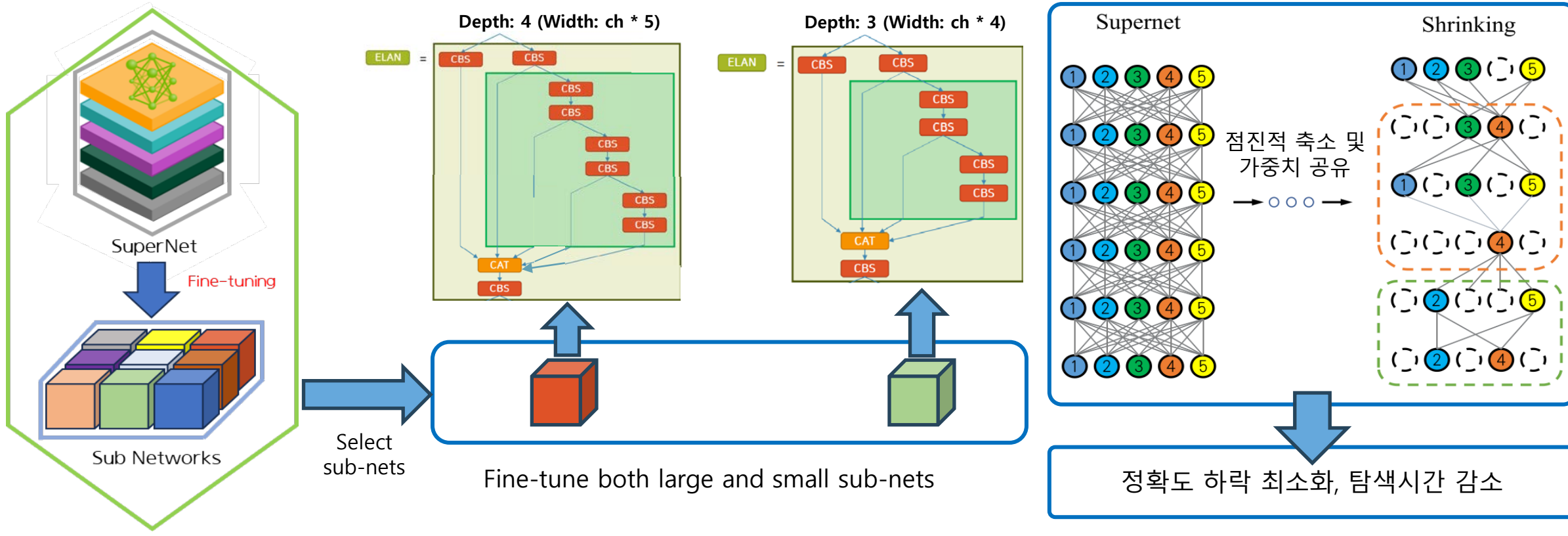




YOLO 기반 신경망 자동 탐색 기술

Progressive Shrinking 학습 기술을 활용한 Fine-tuning 간소화

- 정밀한 Pre-training이 없는 경우, Sub-network는 Scratch부터 학습해야 하므로 탐색시간 증가
- 탐색시간 감소를 위해 점진적 Sub-network 축소 및 가중치 공유

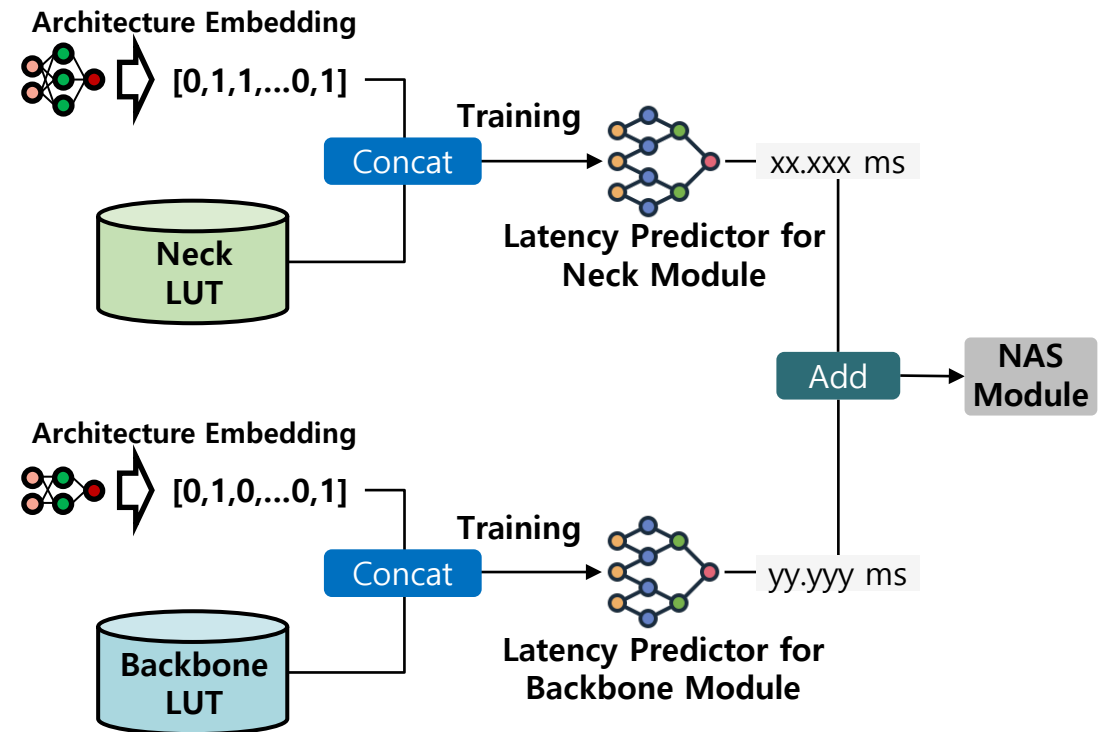
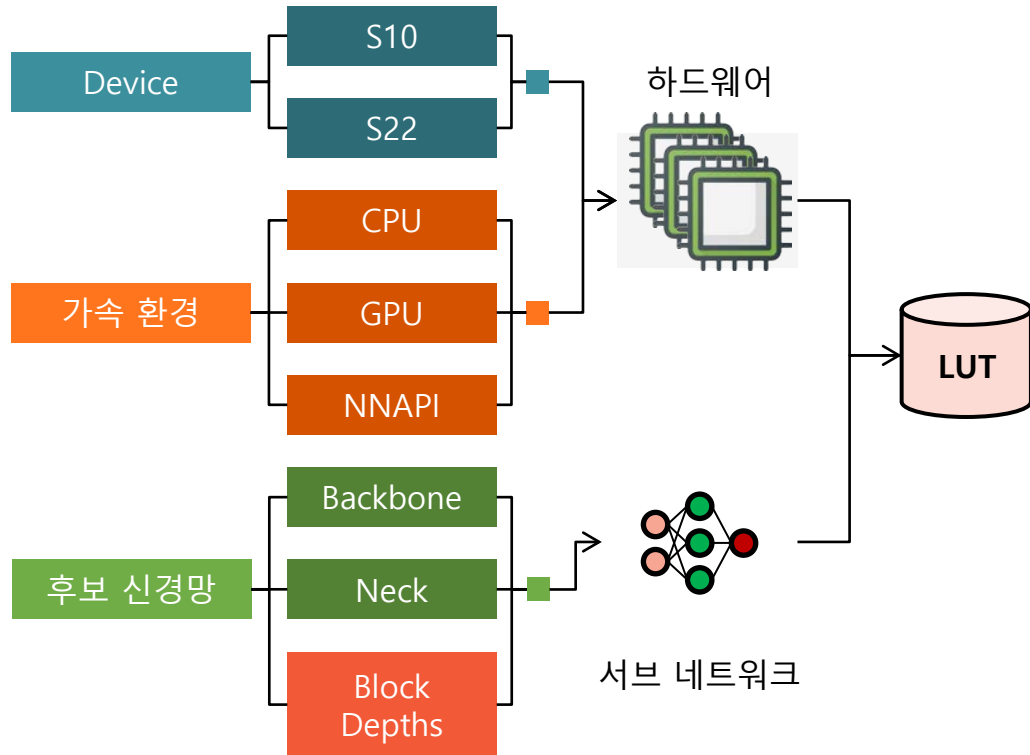




모바일 특화 신경망 탐색 기술

YOLOv7 Backbone / Neck 신경망 지연시간 LUT 구성

- Galaxy S10 및 S22에 대해 YOLOv7 모델의 Backbone / Neck 신경망 지연시간 LUT 구성
- 신경망 오퍼레이션 각각의 지연시간을 측정 후 더한 것과 전체 신경망의 총 지연시간이 불일치하는 문제 해결



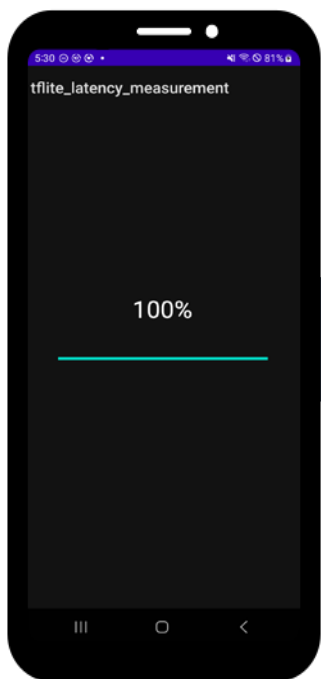


모바일 특화 신경망 탐색 기술

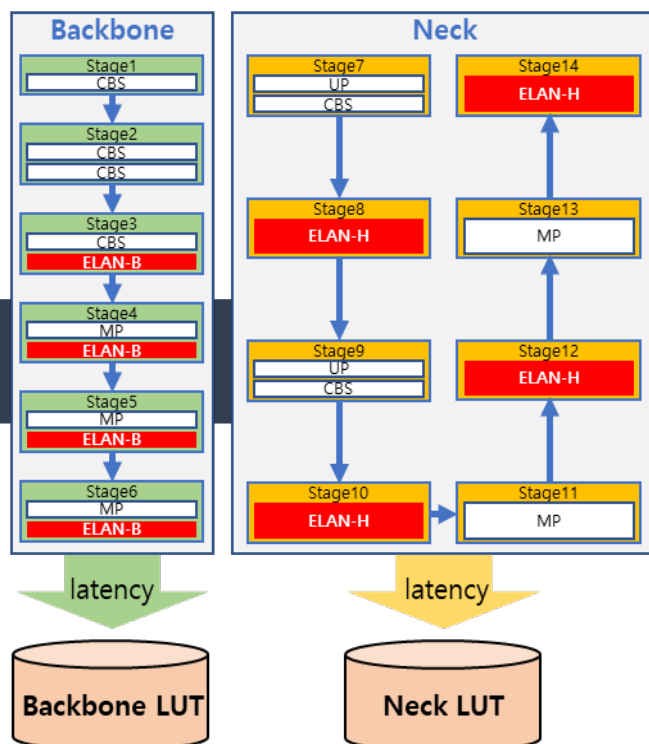
YOLOv7 Backbone / Neck 신경망 지연시간 측정

- 지연시간 측정에 필요한 안드로이드 응용프로그램 개발
- JSON 형식으로 아키텍처-지연시간 쌍으로 저장

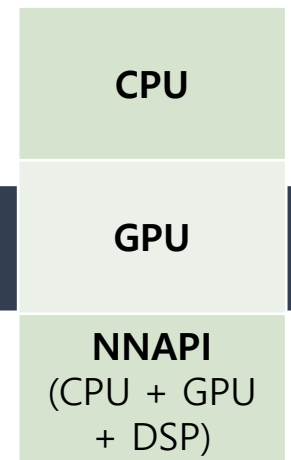
LUT 측정 화면 (S22)



YOLOv7 Sub-network 구성



하드웨어 가속기 선택



LUT 측정 예시 (S22,

ELAN Block Depths	Latency (ms)
1,1,1,1	132.3286
1,1,1,2	135.1316
1,1,1,3	136.6132
⋮	⋮
5,5,5,5	354.0421

LUT 측정 JSON 파일

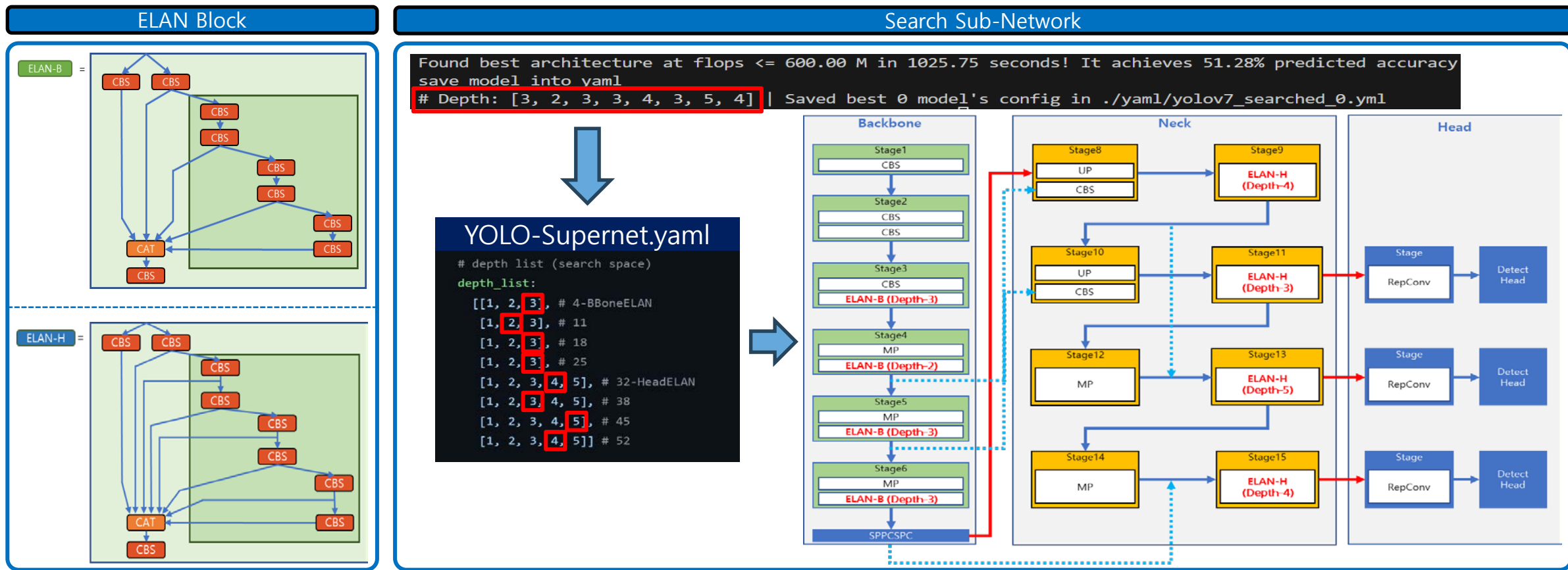




모바일 특화 신경망 탐색 기술

신경망 자동 생성 결과

- 타겟 디바이스 제약 조건을 만족시키는 신경망 구조 생성





AutoNN 모듈 시뮬레이션

AutoNN 수행 과정

- 전달된 yaml 명세를 기준으로 base 모델을 선택한 다음 그에 맞는 AutoNN 단계를 수행
- ex) 하드웨어 설정이 Galaxy S22인 경우, YOLOv7-NAS 실행

yaml





Target : Galaxy S22



Dataset : COCO

Dataset
coco
폴더 생성일 2023-09-14 10:03:10
FILES SIZE
- -

Target
 s22
2023-10-18 created by automltest
Info Galaxy_S22 Memory 128 MB OS android CPU arm Accelerator cpu Engine tf-lite

Progress -


```
0/0 4.836 0.07755 0.03972 0.102 0.2193 1715 640: 0%| |
0/0 4.836 0.07755 0.03972 0.102 0.2193 1715 640: 100%| |
0/0 4.836 0.07755 0.03972 0.102 0.2193 1715 640: 100%| |

-----GET /status_request-----
nasinfo.status: completed
Auto NN 완료
```

T

A

N

감사합니다.

