●YOLOX

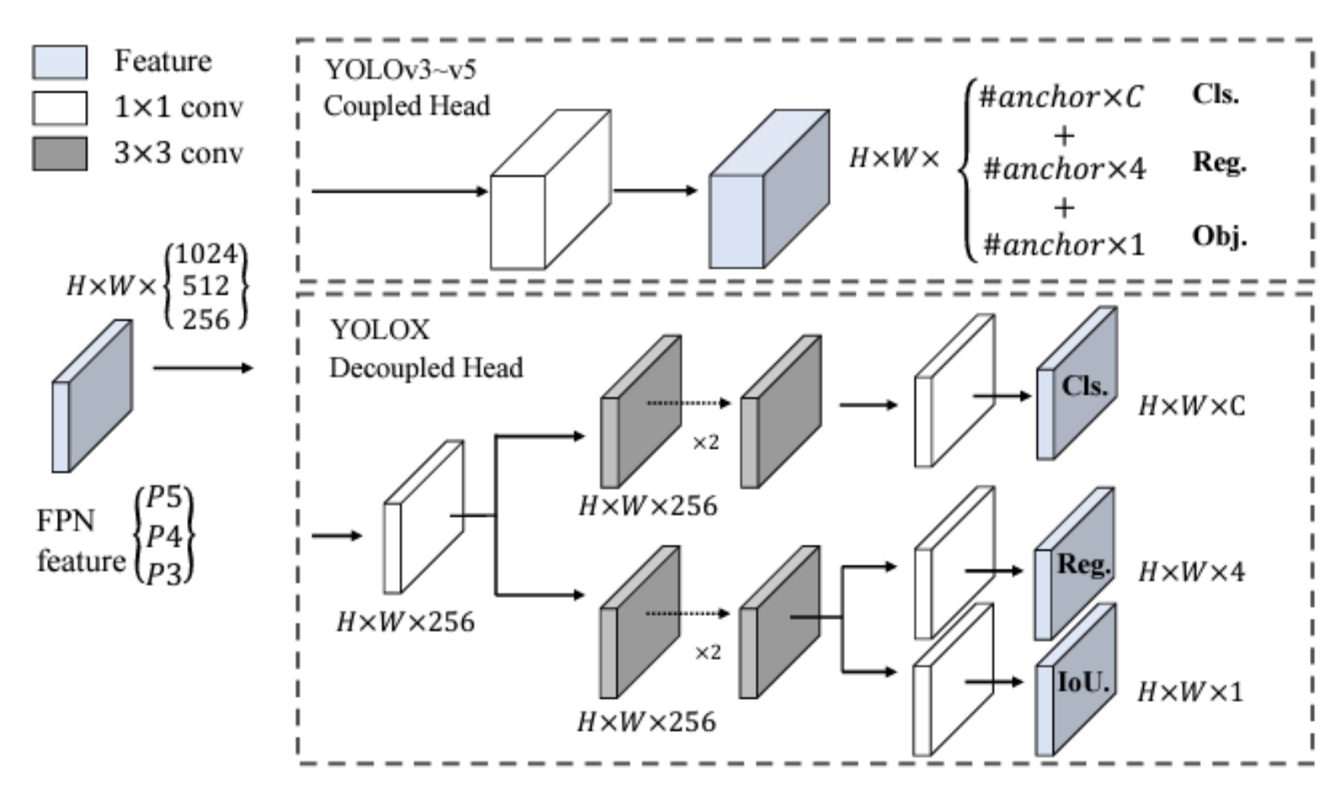
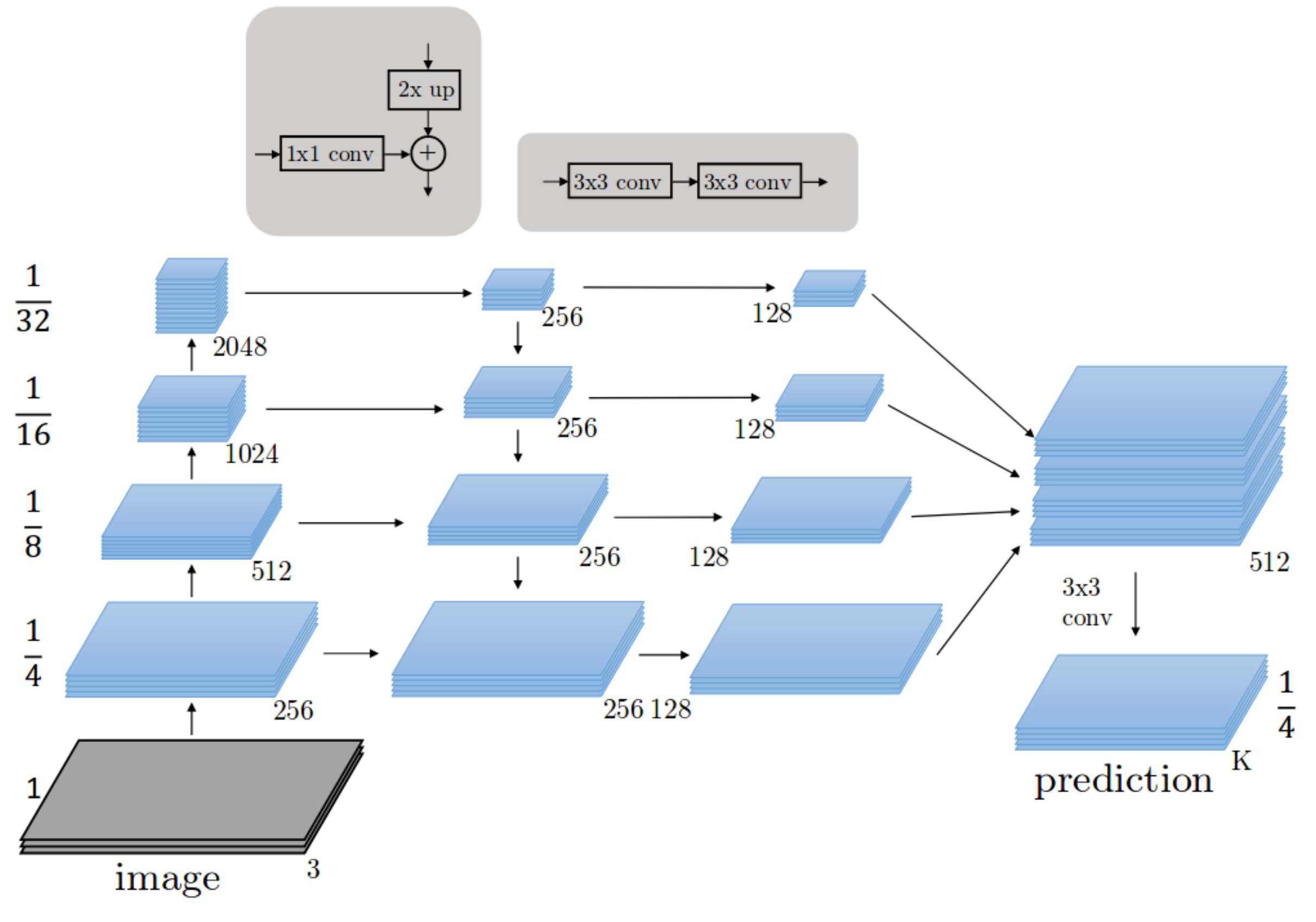


Figure 2: Illustration of the difference between YOLOv3 head and the proposed decoupled head. For each level of FPN feature, we first adopt a 1 × 1 conv layer to reduce the feature channel to 256 and then add two parallel branches with two 3 × 3 conv layers each for classification and regression tasks respectively. IoU branch is added on the regression branch.

- YOLOX는 기본적으로 1 Stage Detector로 Input - Backbone - Neck - Dense Prediction의 구조를 가진다.
- YOLOX는 Darknet53의 Backbone을 통해 Feature Map을 추출하며, SPP Layer를 통해 성능을 개선한다.
- YOLOv4와 YOLOv5의 파이프라인은 Anchor Based 위주로 최적화가 진행되어있기 때문에, General 한 성능이 오히려 떨어질 수 있다고 생각하여 본 논문의 저자들은 YOLOv3-SPP를 기본 베이스 모델로 삼았다.
- FPN을 통해 Multi-Scale Feature Map을 얻고 이를 통해 작은 해상도의 Feature Map에서는 큰 Object를 추출하고 큰 해상도의 Feature Map에서는 작은 Object를 추출하게끔 한 Neck 구조를 차용하였다.
- 최종적으로 Head 부분에서는 기존 YOLOv3와 달리 Decoupled Head를 사용하였다.
- YOLOX는 위와 같은 네트워크 위에 크게 4가지 방법(Decoupled head, Strong data augmentation, Anchor-free, Multi positives)을 가지고 성능 향상을 이끌어냈다.

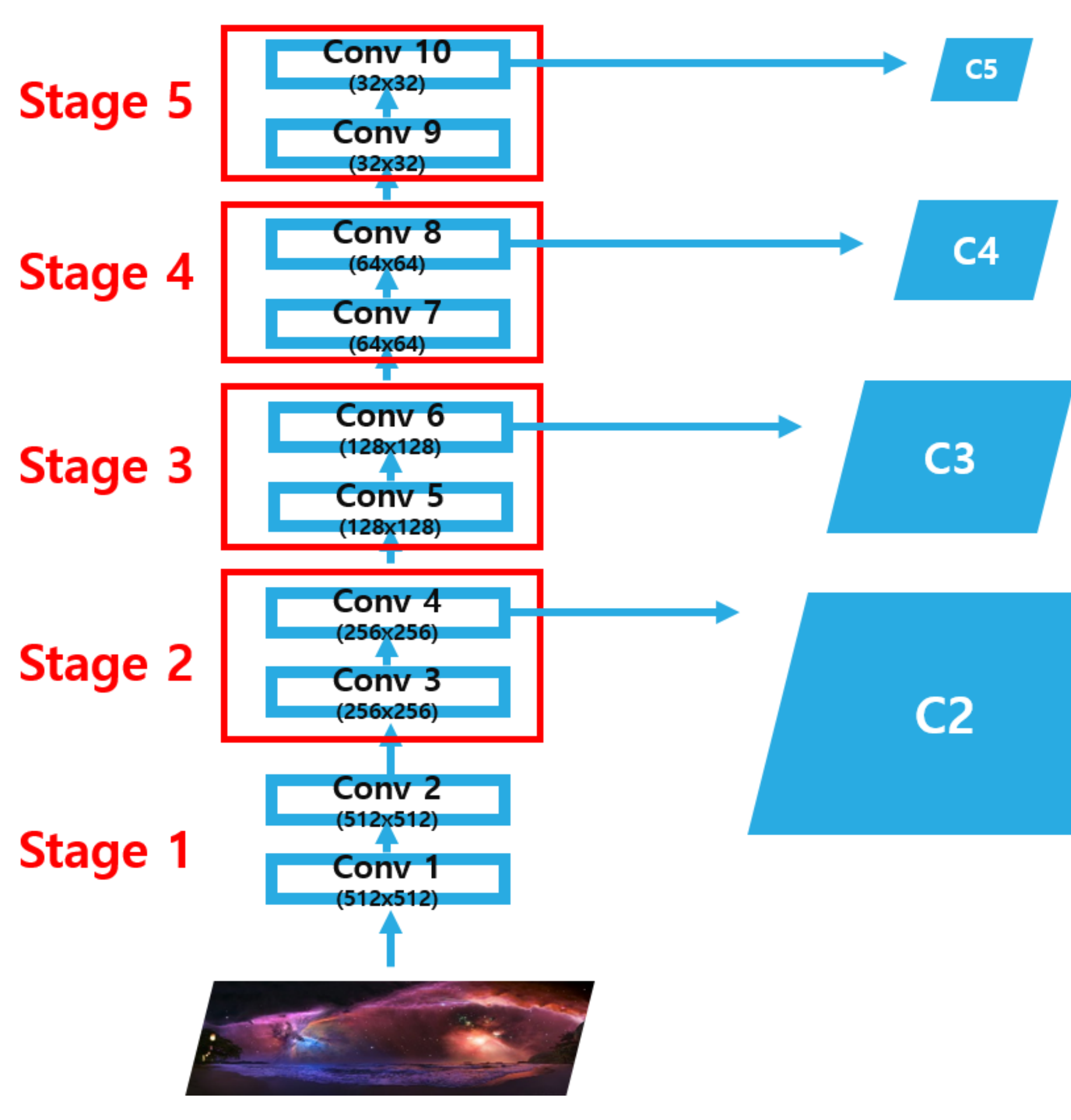
●FPN (Feature Pyramid Network)

- 기존의 object를 detect하는 방법은 모델의 추론속도가 느리고, 메모리 누수가 커서, 컴퓨팅 자원을 적게 차지하면서 다양한 크기의 객체를 인식하는 새로운 방법인 FPN이 제시됨
- Anchor free의 단점으로 멀티 스케일을 고려하지 못해서 작은 오브젝트 등을 잘 찾기가 힘든 부분이 있는데, 저자는 이런 약점들을 FPN를 통해서 멀티스케일을 고려해서 극복하고자 했다.



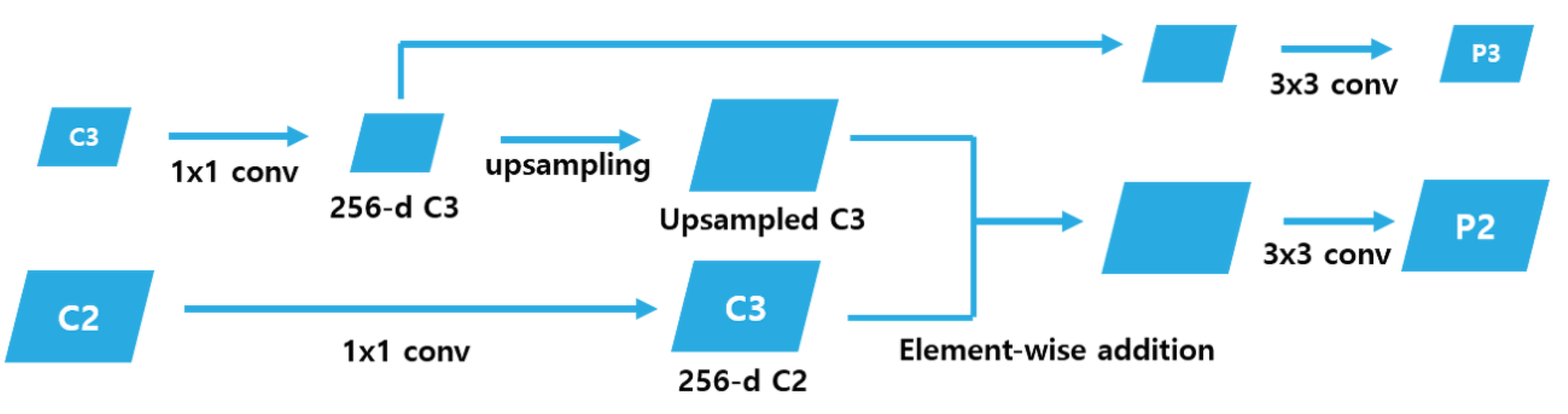
- FPN은 임의의 크기의 single-scale 이미지를 convolutional network에 입력하여 다양한 scale의 feature map을 출력하는 네트워크이다. -FPN이 feature map을 추출하여 피라미드를 건축하는 과정은 bottom-up pathway, top-down pathway, lateral connections에 따라 진행된다.

Bottom-up pathway



- Bottom-up pathway 과정은 이미지를 convolutional network에 입력하여 forward pass하여 2배씩 작아지는 feature map을 추출하는 과정이다.
- 이 때 각 stage의 마지막 layer의 output feature map을 추출한다. 깊은 layer 수록 강력한 feature를 보유하고있기 때문이다.
- 위 c2, c3, c4, c5는 원본 이미지의 1/4 1/8 1/16 1/32 크기를 가진 feature이다.
- Conv1의 feature map은 너무 많은 메모리를 차지하여 제외됨.

Top-down pathway

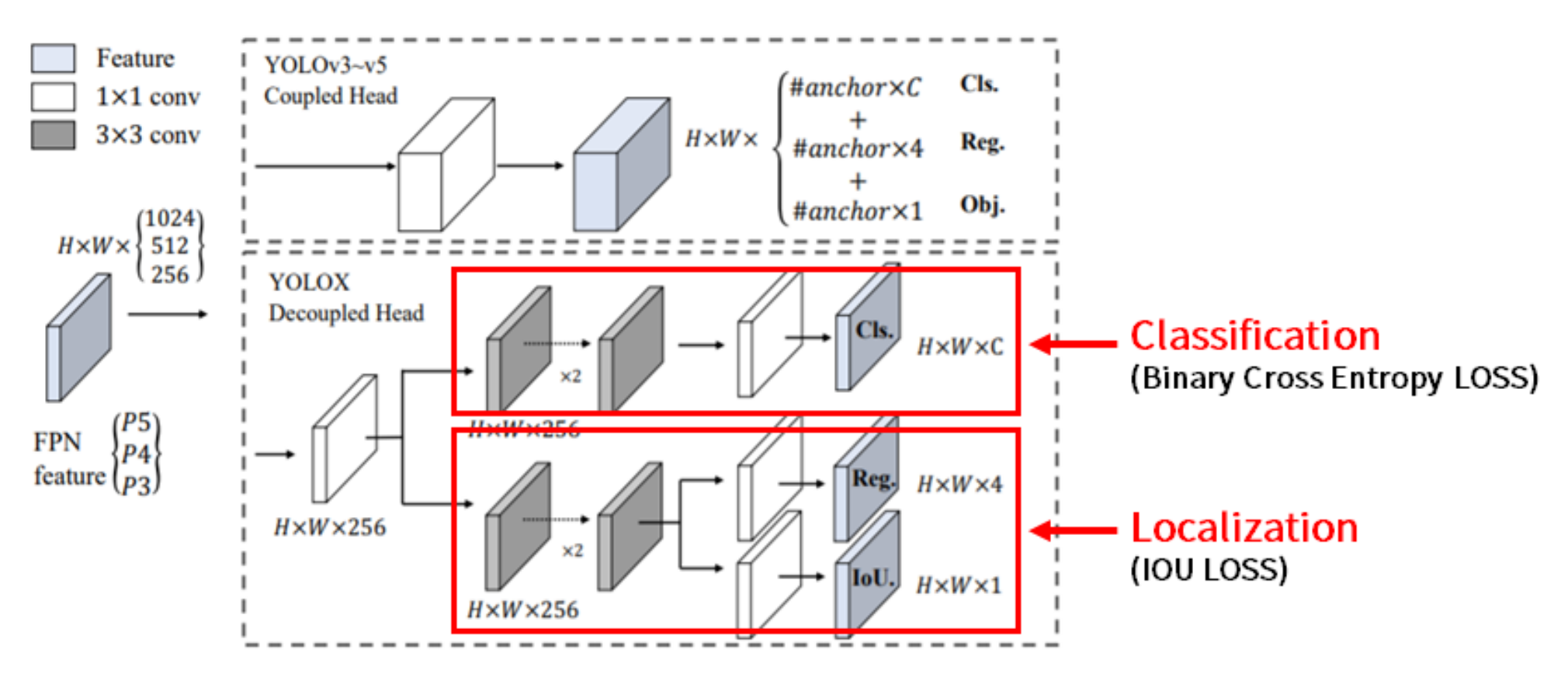


- Top-down Pathway는 각 pyramid level에 있는 feature map을 2배로 upsampling하고 channel 수를 동일하게 맞춰주는 과정이다.
- 각 pyramid level의 feature map을 2배로 upsampling해주면 바로 아래 level의 feature map와 크기가 같아진다.
- 모든 pyramid level의 feature map에 1*1 conv를 적용하여 channel을 256으로 수정해줌.
- Upsampling된 feature map과 바로 아래 feature map과 elementwise addition 연산을 하는 Lateral connections 과정을 수행.
- 이후 각각의 feature map에 3*3 conv 연산을 적용함.
- 적용된 p2, p3,…들은 c2, c3,…들과 크기가 동일
- 가장 높은 레벨에 있는 C2는 1*1 conv를 통해 256 channels로 늘 려준 후 그대로 출력되어 p2가 됨.

Conclusion

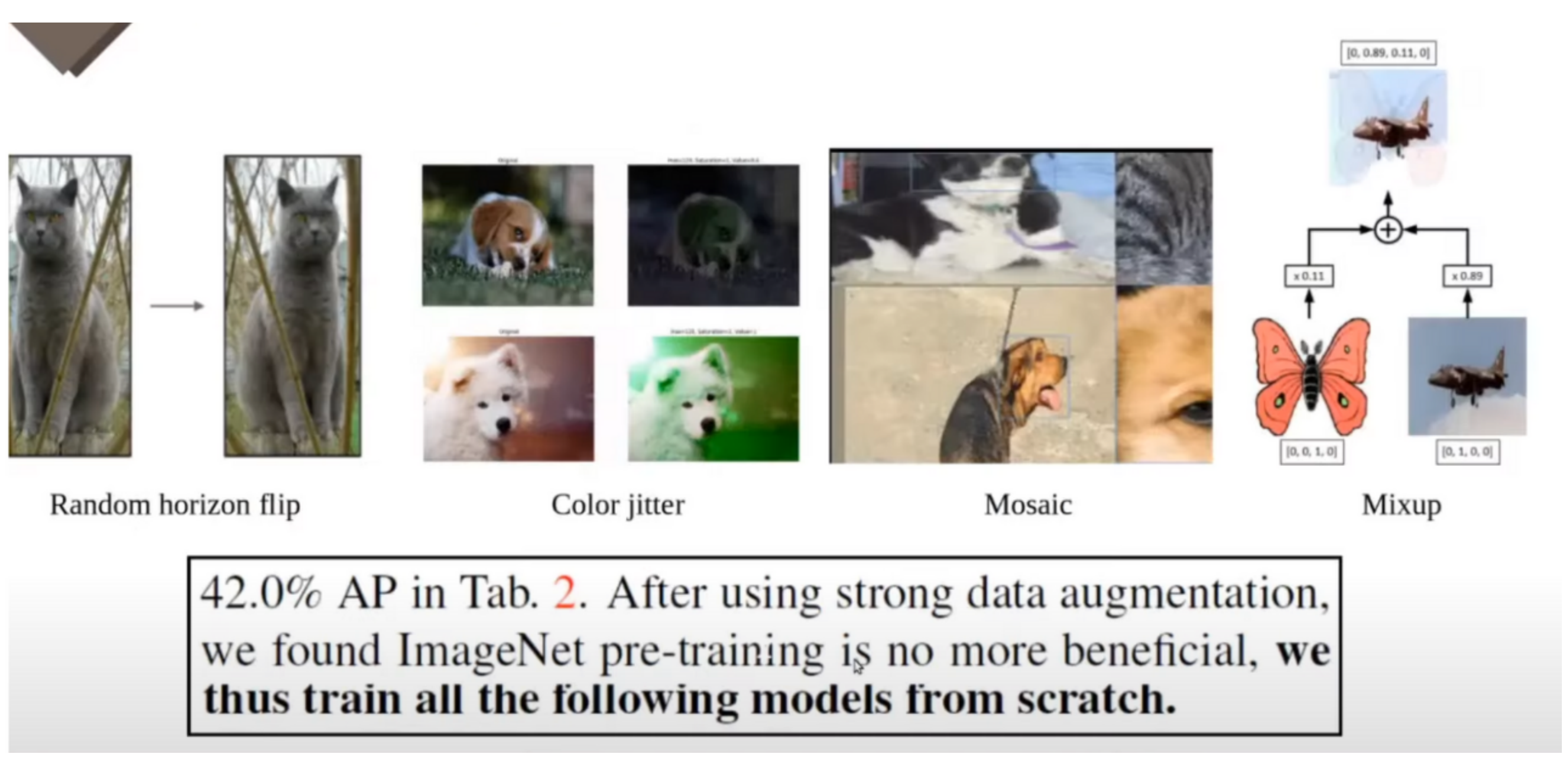
- 결과 feature maps의 고해상도 feature map은 저수준 특징을 가지지 만 객체의 위치에 대한 정보를 상대적으로 정확하게 보존함.
- 이런 고해상도 feature map의 특징을 element-wise addition을 통해 저해상도 feature map에 전달하기 때문에 p가 c에 비해 작은 객체를 더 잘 detect한다.

●Decoupled head



- YOLOV3부터 V5까지는 하나의 헤드만 사용했다. 기존에는 하나의 Head에서 Classification과 Localization을 함께 진행했다. 그러나 Classification에는 Fully Connected Layer가 효과적이지만, 반면에 Localization에는 Convolution Head가 보다 적절한데 이 두가지가 서로 상충된다.
- 따라서 Decoupled head를 사용하여 Classification엔 FC Head를, Localization에는 Convolution Head를 적용했다.Classification 에는 BCE Loss를 사용하고 Localization에는 IoU Loss를 사용하여 학습을 진행한다. 실험 결과, 해당 방식이 기존 Coupled Head 방식보다 Convergence 속도가 빠르고 AP가 향상되었다고 한다.

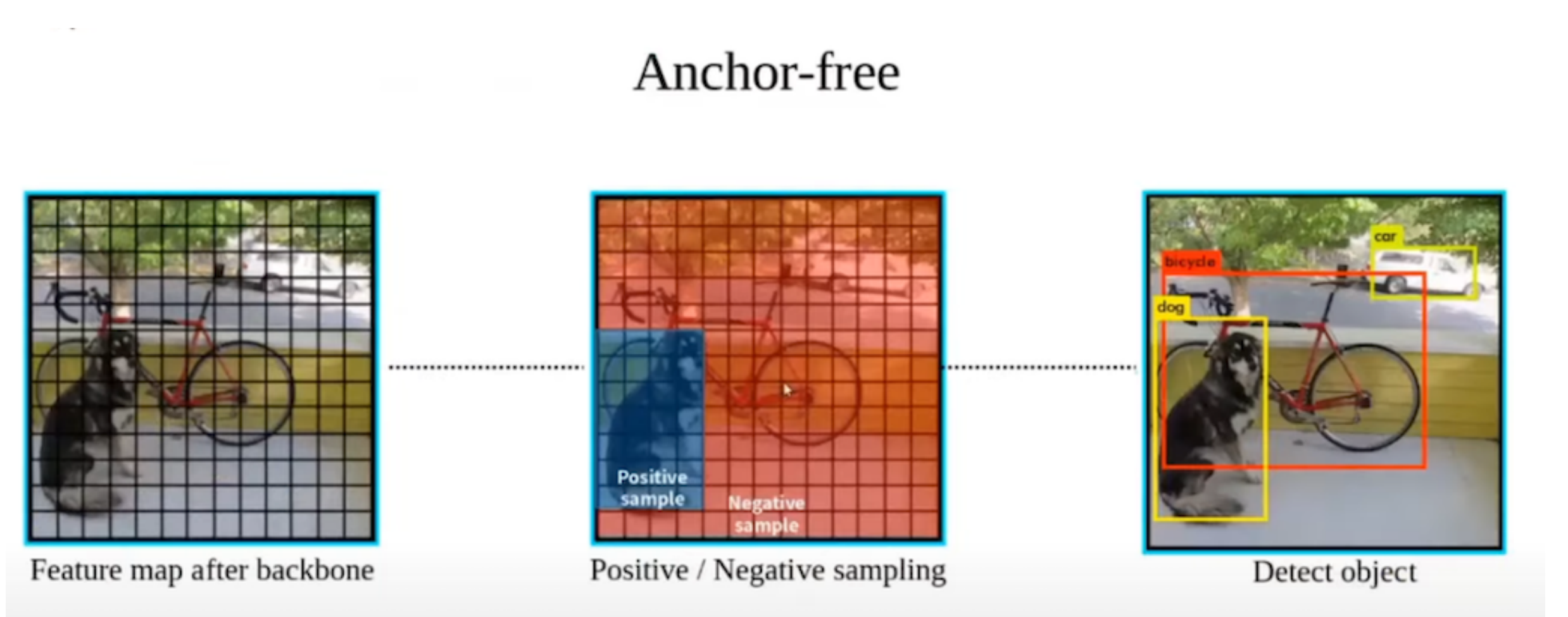
●Strong data augmentation



42.0% AP in Tab. 2. After using strong data augmentation, we found ImageNet pre-training is no more beneficial, **we thus train all the following models from scratch.**

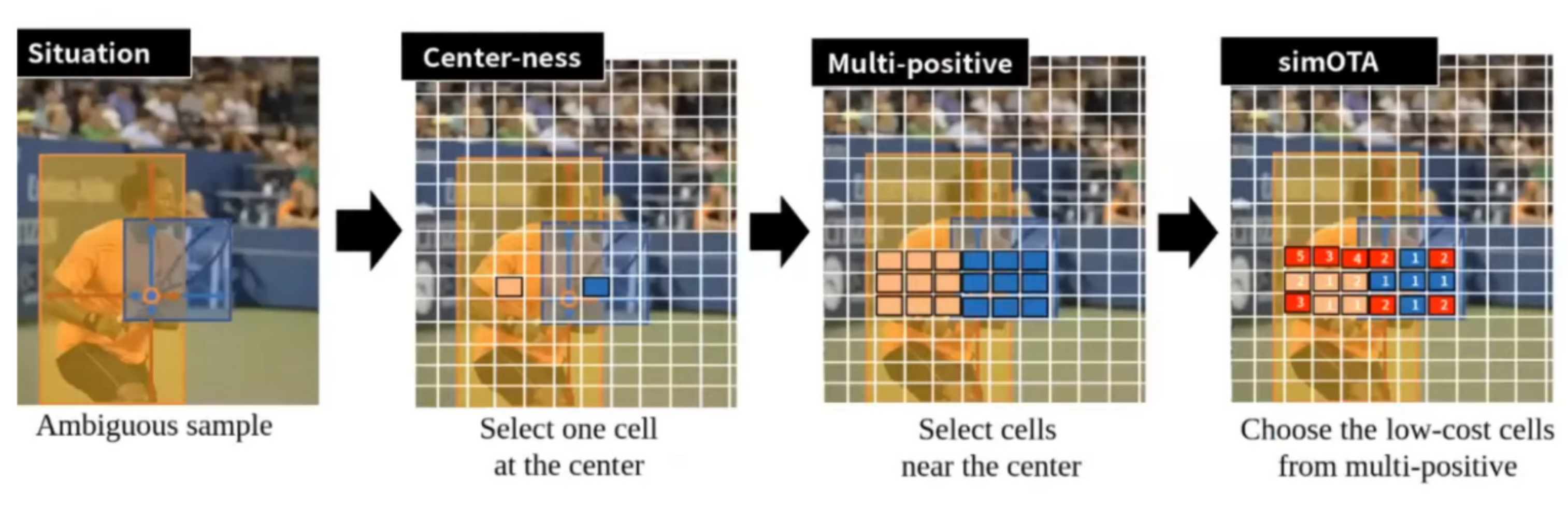
- 이 논문에서는 총 4가지의 Data augmentation 방법을 적용했다.
- random horizon flip, 원본 영상의 hsl을 변경하여 증강시키는 클러스터라는 방법, 원본 이미지 외에 3개의 추가적인 사진을 섞는 Mosaic이라는 방법(ultralytics의 YOLOv3 에서 적용됨), 마지막으로 해당 이미지와 레이블의 다른 것을 조금씩 섞는 Object Detection에서 사용되는 Mixup이라는 방법을 사용했다.

●Anchor-free



- YOLOX에서는 Anchor를 사용하지 않는 Anchor free 방법을 사용한다.
- 기존 Anchor 기반의 Detector들은 개발자들이 직접 Heuristic 하게 Tuning을 진행해주어야 하는 불편함이 존재했다. 또한 그렇게 Tuning된 Anchor Size 또한 특정 Task에 종속적이므로 General한 성능은 떨어지는 이슈가 존재했다.
- Anchor Free 방식은 학습을 보다 간편하고 편하게 해주고, 다양한 Hyperparameter들을 Tuning해야 하는 필요성이 없으며, 그로 인해 다양한 분야에 General 하게 일정한 성능을 보장한다.
- Anchor free 메서드는 feature extraction 과정까지는 같지만, Anchor free라는 이름처럼 각 Cell마다 Anchor를 사용하지 않고, 바로 bounding box나 class를 classification 하는 과정을 거친다.
- 위 그림에서 그라운드 박스가 파란색 Cell이라고 하면, GT bounding box에 속하는 Cell들은 Positive 샘플로 할당하고, 속하지 않는 것은 Negative 샘플로 할당한다.

●Multi positive



- 기본적인 Anchor free의 방법만으로는 Anchor base방식의 accuracy를 따라잡을 수가 없다. 따라서 본 논문에서는 accuracy를 향상하기 위해서 다른 방법을 적용 했다.
- 그 중 하나가 CenterYOLOness이다. 좌측 사진과 같이 예측하고자 하는 Cell이 있다고 할 때, 각 Cell 중에 원래 Positive가 있는데, 오브젝트에 중심에 해당하는 Cell들만 Positive로 할당하는 CenterYOLOness 방법을 사용을 해서 정확도를 더 향상했다.
- 저자는 가운데에 있는 Cell 말고 옆에 주위에 있는 Cell들도 더 충분히 좋은 prediction을 할 수 있다고 보았고, 오브젝트 중심에 있으면서도 Loss가 낮은 k개의 샘플을 Positive 샘플로 적용하는 Multi positive를 사용했다.
- 이렇게 positive Sample을 증강해줌으로써, 심각한 class imbalance도 어느정도 상쇄시킬 수 있다.

●성능 비교

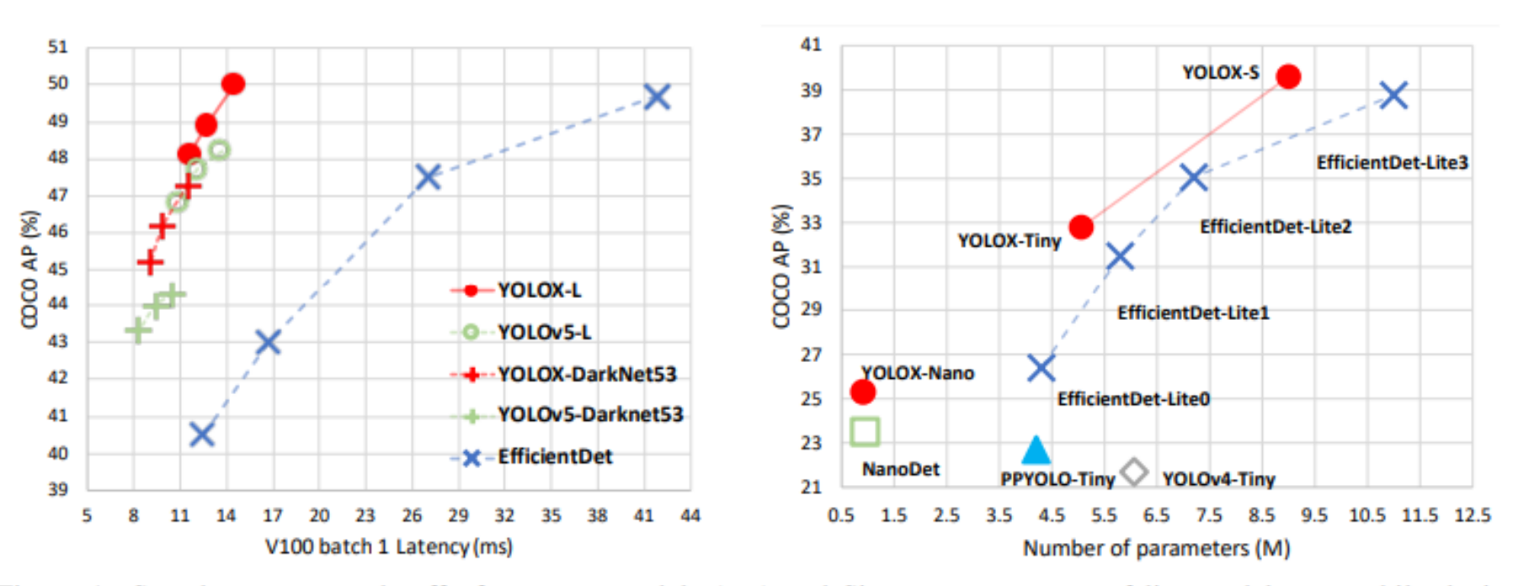
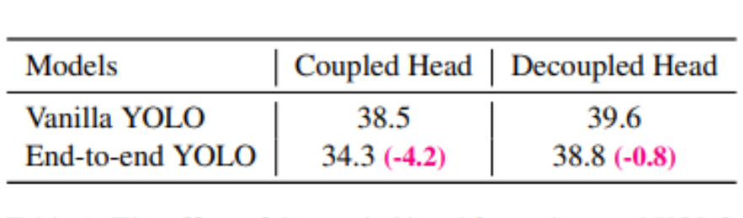


Figure 1: Speed-accuracy trade-off of accurate models (top) and Size-accuracy curve of lite models on mobile devices (bottom) for YOLOX and other state-of-the-art object detectors.

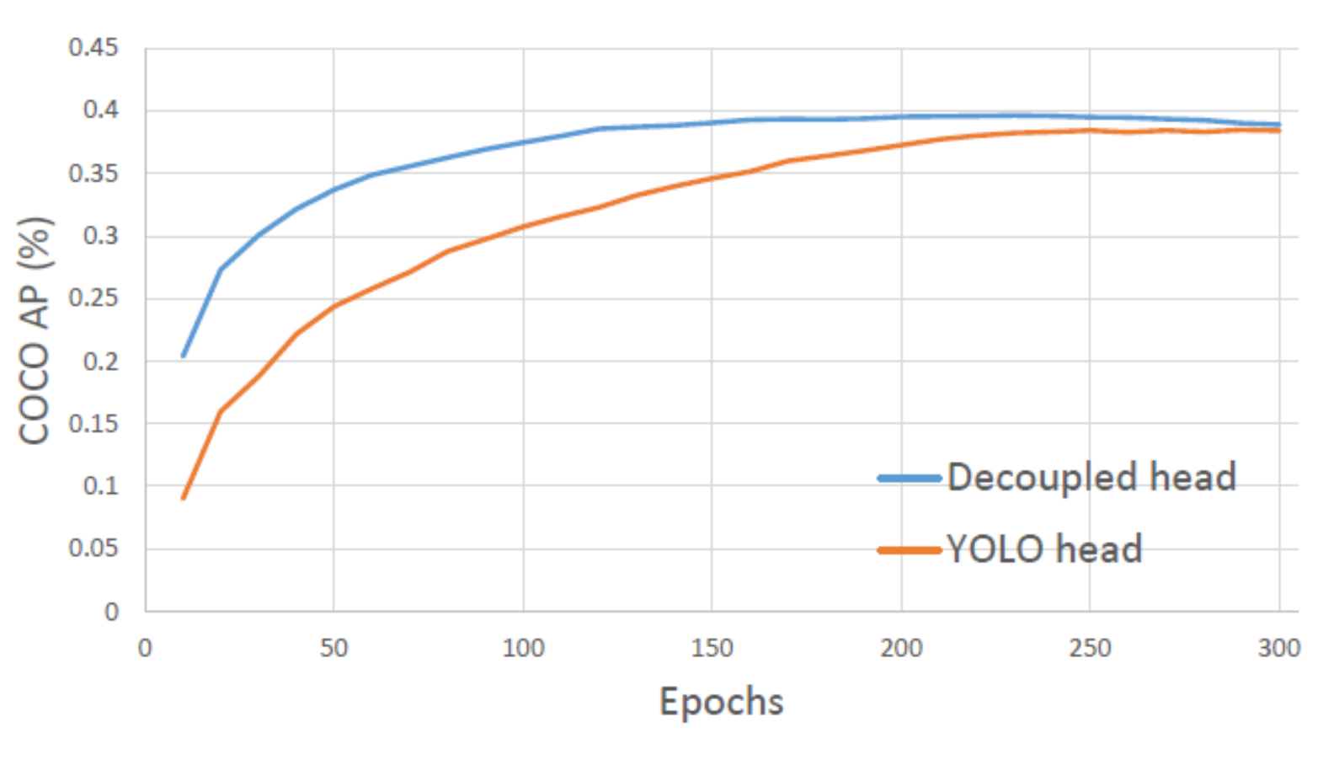
- 기존 모델들에 Anchor free를 적용한 결과, 기존 모델보다 높은 성능을 보여주었다.



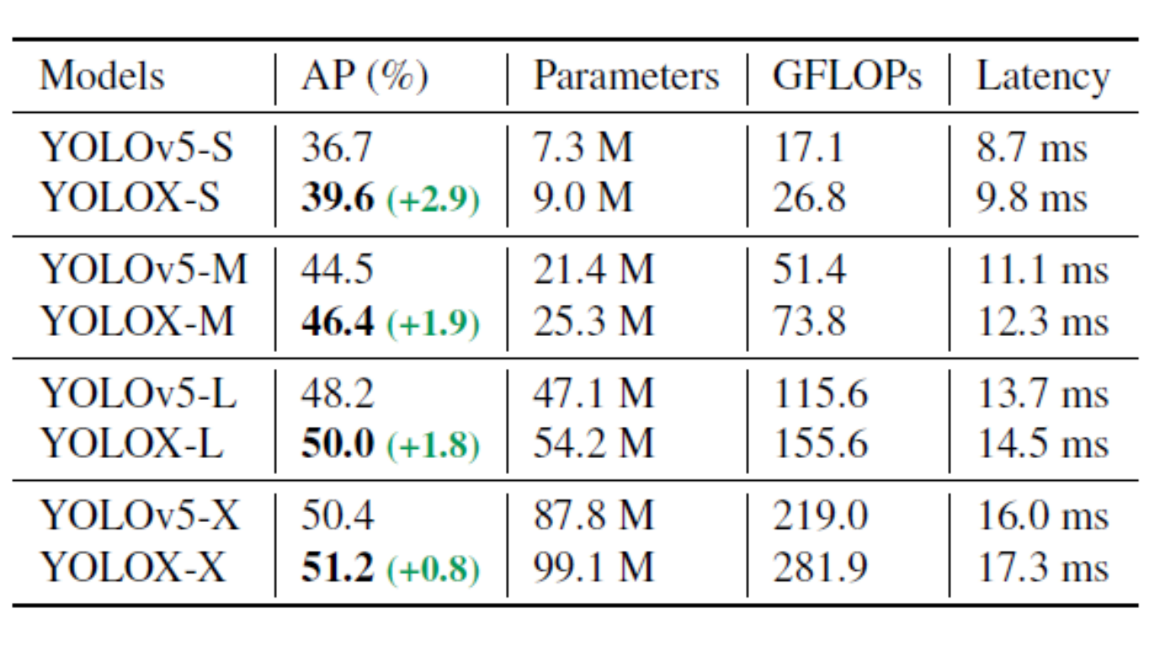
Models	Coupled Head	Decoupled Head
Vanilla YOLO	38.5	39.6
End-to-end YOLO	34.3 (-4.2)	38.8 (-0.8)

Table 1: The effect of decoupled head for end-to-end YOLO in terms of AP (%) on COCO.

- Coupled Head에 비해 Decoupled Head를 사용했을 때의 성능이 더 좋은것을 확인할 수 있다.

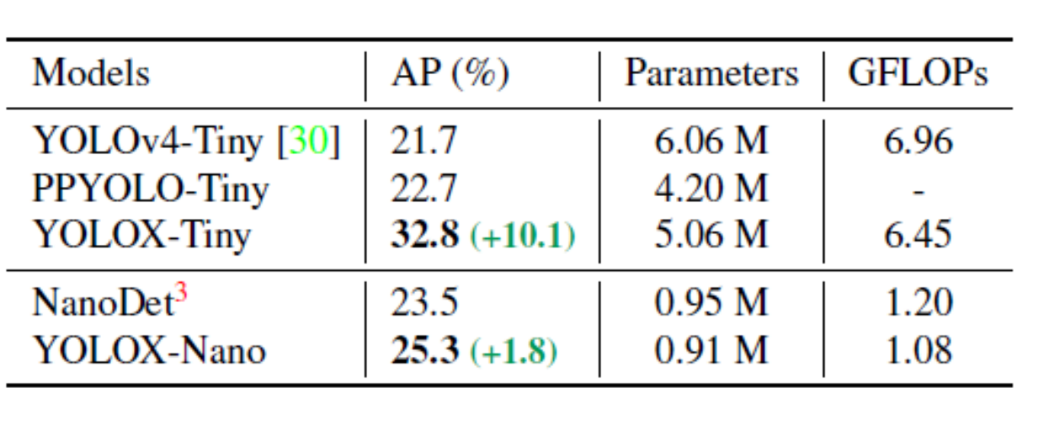


- Decoupled head와 YOLO head의 epoch당 AP값으로 표현된 학습 속도. Decoupled head가 조금 더 converging 속도가 빠른 것을 확인할 수 있다.



Models	AP (%)	Parameters	GFLOPs	Latency
YOLOv5-S	36.7	7.3 M	17.1	8.7 ms
YOLOX-S	**39.6** (+2.9)	9.0 M	26.8	9.8 ms
YOLOv5-M	44.5	21.4 M	51.4	11.1 ms
YOLOX-M	**46.4** (+1.9)	25.3 M	73.8	12.3 ms
YOLOv5-L	48.2	47.1 M	115.6	13.7 ms
YOLOX-L	**50.0** (+1.8)	54.2 M	155.6	14.5 ms
YOLOv5-X	50.4	87.8 M	219.0	16.0 ms
YOLOX-X	**51.2** (+0.8)	99.1 M	281.9	17.3 ms

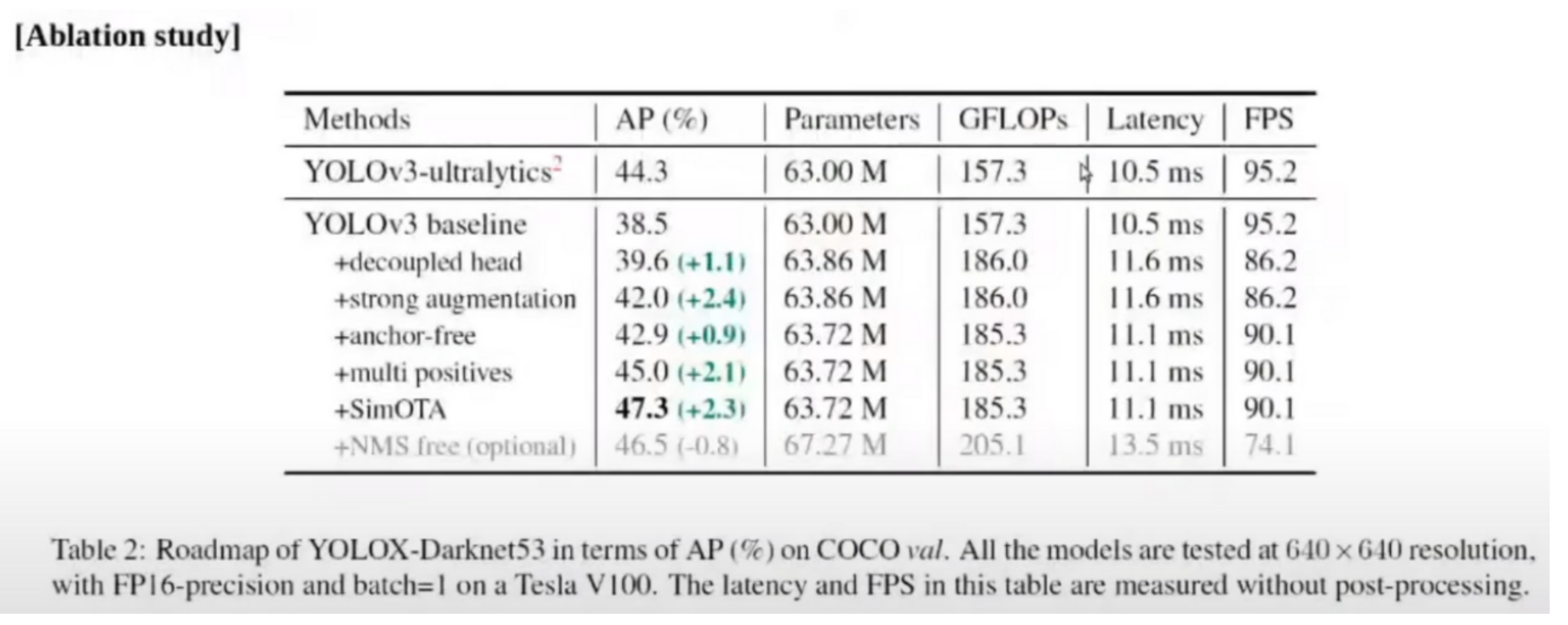
Table 3: Comparison of YOLOX and YOLOv5 in terms of AP (%) on COCO. All the models are tested at 640 × 640 resolution, with FP16-precision and batch=1 on a Tesla V100.



Models	AP (%)	Parameters	GFLOPs
YOLOv4-Tiny [30]	21.7	6.06 M	6.96
PPYOLO-Tiny	22.7	4.20 M	-
YOLOX-Tiny	**32.8** (+10.1)	5.06 M	6.45
NanoDet[3]	23.5	0.95 M	1.20
YOLOX-Nano	**25.3** (+1.8)	0.91 M	1.08

Table 4: Comparison of YOLOX-Tiny and YOLOX-Nano and the counterparts in terms of AP (%) on COCO *val*. All the models are tested at 416 × 416 resolution.

- 다양한 backbone과 다양한 사이즈, 그리고 경량화된 모델에서도 성능 개선이 된 결과를 확인할 수 있다.

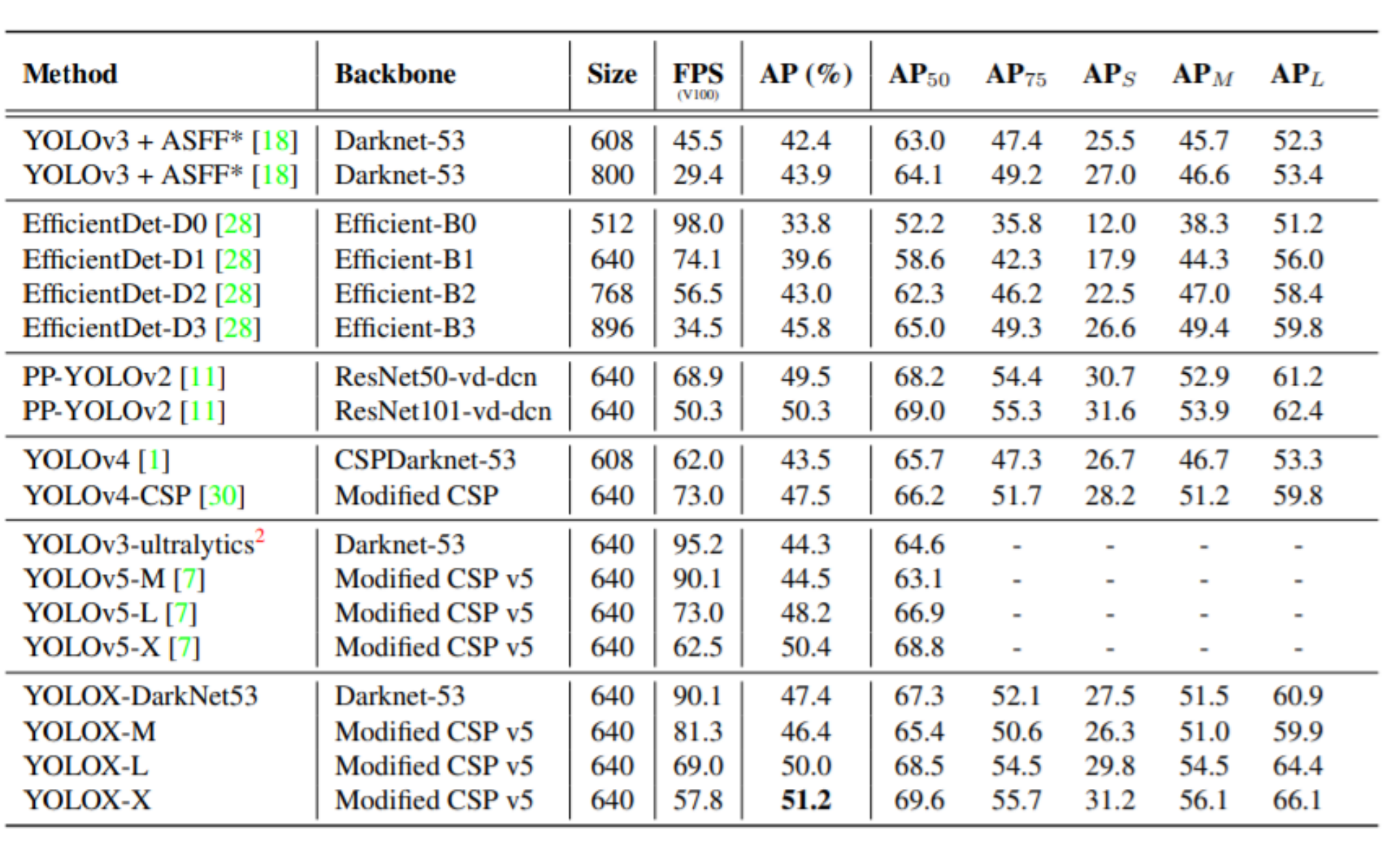


[Ablation study]

Methods	AP (%)	Parameters	GFLOPs	Latency	FPS
YOLOv3-ultralytics[2]	44.3	63.00 M	157.3	10.5 ms	95.2
YOLOv3 baseline	38.5	63.00 M	157.3	10.5 ms	95.2
+decoupled head	39.6 (+1.1)	63.86 M	186.0	11.6 ms	86.2
+strong augmentation	42.0 (+2.4)	63.86 M	186.0	11.6 ms	86.2
+anchor-free	42.9 (+0.9)	63.72 M	185.3	11.1 ms	90.1
+multi positives	45.0 (+2.1)	63.72 M	185.3	11.1 ms	90.1
+SimOTA	**47.3** (+2.3)	63.72 M	185.3	11.1 ms	90.1
+NMS free (optional)	46.5 (-0.8)	67.27 M	205.1	13.5 ms	74.1

Table 2: Roadmap of YOLOX-Darknet53 in terms of AP (%) on COCO *val*. All the models are tested at 640×640 resolution, with FP16-precision and batch=1 on a Tesla V100. The latency and FPS in this table are measured without post-processing.

- YOLO V3 base모델에 비해서, 앞에서 소개한 방법들을 적용하는 것이, 성능 향상에 도움이 되었다는 것을 알 수 있다.
- NMS-free는 두개의 conv layer를 추가해 one-to-one 라벨할당과 gradient를 멈춘다. 이것이 성능과 inference 속도를 저하하므로, 옵셔널하게 사용가능하게 하고, 최종모델엔 반영하지 않았다.



Method	Backbone	Size	FPS (V100)	AP (%)	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv3 + ASFF* [18]	Darknet-53	608	45.5	42.4	63.0	47.4	25.5	45.7	52.3
YOLOv3 + ASFF* [18]	Darknet-53	800	29.4	43.9	64.1	49.2	27.0	46.6	53.4
EfficientDet-D0 [28]	Efficient-B0	512	98.0	33.8	52.2	35.8	12.0	38.3	51.2
EfficientDet-D1 [28]	Efficient-B1	640	74.1	39.6	58.6	42.3	17.9	44.3	56.0
EfficientDet-D2 [28]	Efficient-B2	768	56.5	43.0	62.3	46.2	22.5	47.0	58.4
EfficientDet-D3 [28]	Efficient-B3	896	34.5	45.8	65.0	49.3	26.6	49.4	59.8
PP-YOLOv2 [11]	ResNet50-vd-dcn	640	68.9	49.5	68.2	54.4	30.7	52.9	61.2
PP-YOLOv2 [11]	ResNet101-vd-dcn	640	50.3	50.3	69.0	55.3	31.6	53.9	62.4
YOLOv4 [1]	CSPDarknet-53	608	62.0	43.5	65.7	47.3	26.7	46.7	53.3
YOLOv4-CSP [30]	Modified CSP	640	73.0	47.5	66.2	51.7	28.2	51.2	59.8
YOLOv3-ultralytics[2]	Darknet-53	640	95.2	44.3	64.6	-	-	-	-
YOLOv5-M [7]	Modified CSP v5	640	90.1	44.5	63.1	-	-	-	-
YOLOv5-L [7]	Modified CSP v5	640	73.0	48.2	66.9	-	-	-	-
YOLOv5-X [7]	Modified CSP v5	640	62.5	50.4	68.8	-	-	-	-
YOLOX-DarkNet53	Darknet-53	640	90.1	47.4	67.3	52.1	27.5	51.5	60.9
YOLOX-M	Modified CSP v5	640	81.3	46.4	65.4	50.6	26.3	51.0	59.9
YOLOX-L	Modified CSP v5	640	69.0	50.0	68.5	54.5	29.8	54.5	64.4
YOLOX-X	Modified CSP v5	640	57.8	**51.2**	69.6	55.7	31.2	56.1	66.1

- COCO dataset에 적용한 SOTA 모델과 비교한 성능. YOLOX-X의 AP가 51.2%로 가장 높다.