Instance Level Detection and Beyond

The University of Hong Kong
Xie Enze
Supervisor：Dr. Luo Ping

Introduction

- My research is computer vision， especially detection and segmentation with self／semi／weak supervision．

- My work PolarMask was selected as CVPR 2020 Top－10 Influential Papers and with 730＋ github star．

- I co－developed OpenSelfSup (1．2k github star)， which is a popular self－supervised learning toolbox．

- Publication： 7 first／co－first papers (TPAMI， CVPR， ICCV， ECCV)． Citation： 443

  - Google scholar： https：／／scholar．google．com／citations？user＝42MVVPgAAAAJ&hl＝zh－CN

Outline

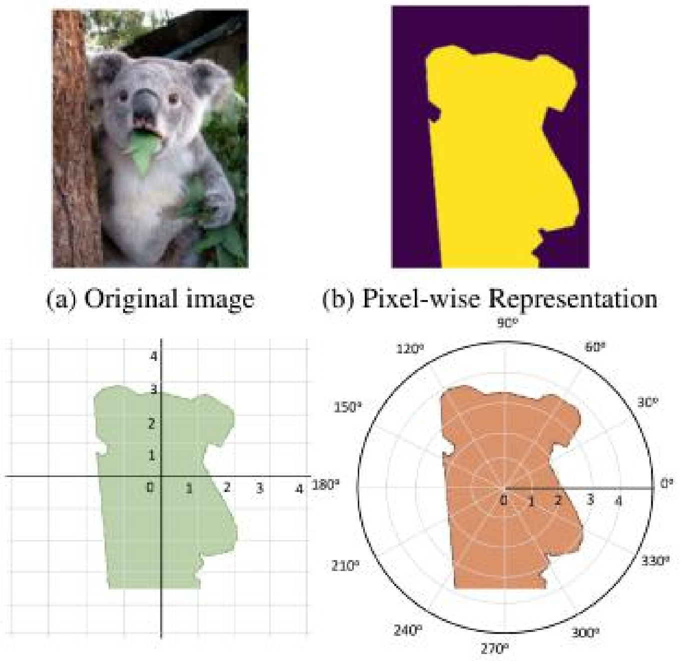
- 1. PolarMask： Single Shot Instance Segmentation with Polar Representation (CPVR20 Oral & TPAMI21, Top—10 influence paper)

- 2. DetCo： Unsupervised Contrastive Learning for Object Detection(a new pretrain friendly for object detection, work in Huawei)

- Pyramid Vision Transformer： A Versatile Backbone for Dense Prediction without Convolutions (say goodbye to CNNs)

PolarMask

Summary：We introduced a Polar Representation to reformulate the instance segmentation problem.



(a) Original image (b) Pixel-wise Representation

(c) Cartesian Representation (d) Polar Representation

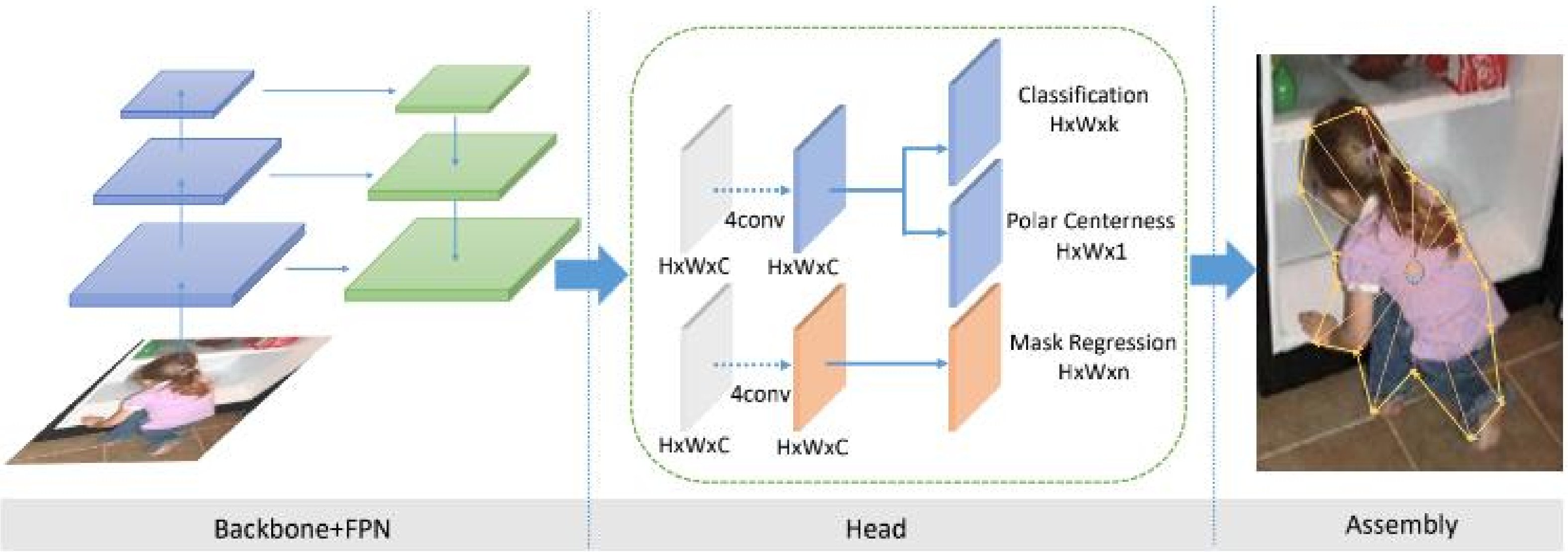
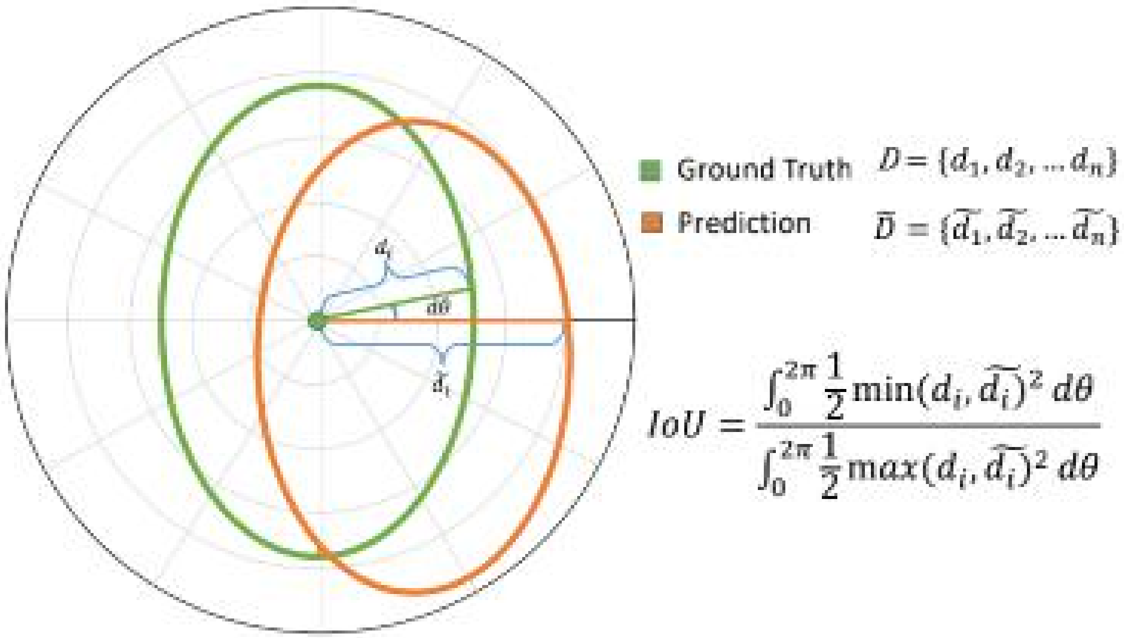


Figure 2 – The overall pipeline of PolarMask. The left part contains the backbone and feature pyramid to extract features of different levels. The middle part is the two heads for classification and polar mask regression. H, W, C are the height, width, channels of feature maps, respectively, and k is the number of categories (e.g., $k = 80$ on the COCO dataset), n is the number of rays (e.g., $n = 36$)

PolarMask


🟩 Ground Truth $D = \{d_1, d_2, \dots d_n\}$
🟧 Prediction $\tilde{D} = \{\tilde{d_1}, \tilde{d_2}, \dots \tilde{d_n}\}$
d_i
$d\theta$
$\tilde{d_i}$

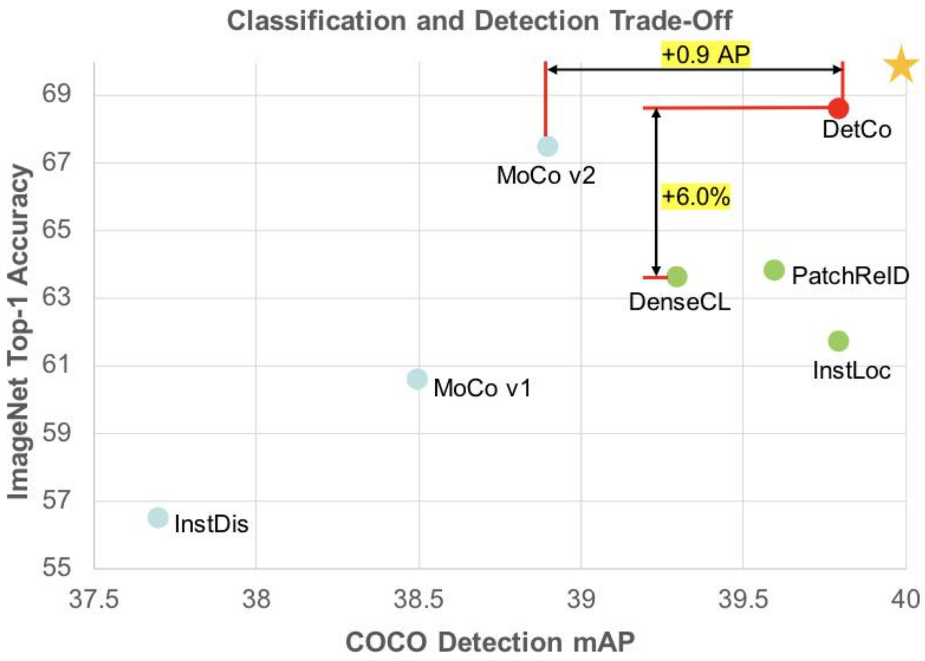
$$IoU = \frac{\int_0^{2\pi} \frac{1}{2} \min(d_i, \tilde{d_i})^2 \, d\theta}{\int_0^{2\pi} \frac{1}{2} max(d_i, \tilde{d_i})^2 \, d\theta}$$

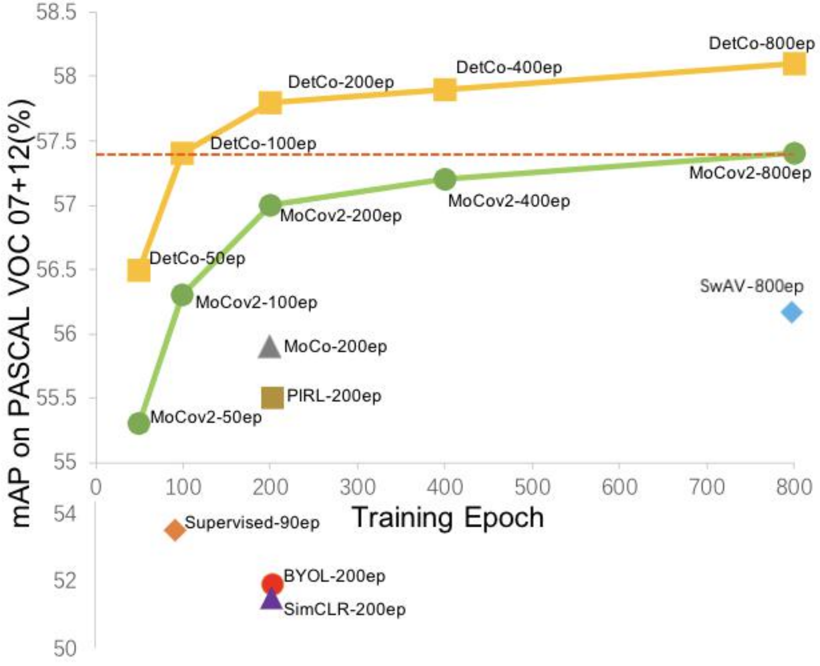
Figure 5 – Mask IoU in Polar Representation. Mask IoU (interaction area over union area) in the polar coordinate can be calculated by integrating the differential IoU area in terms of differential angles.

Mask IoU Loss

Use calculus to design loss

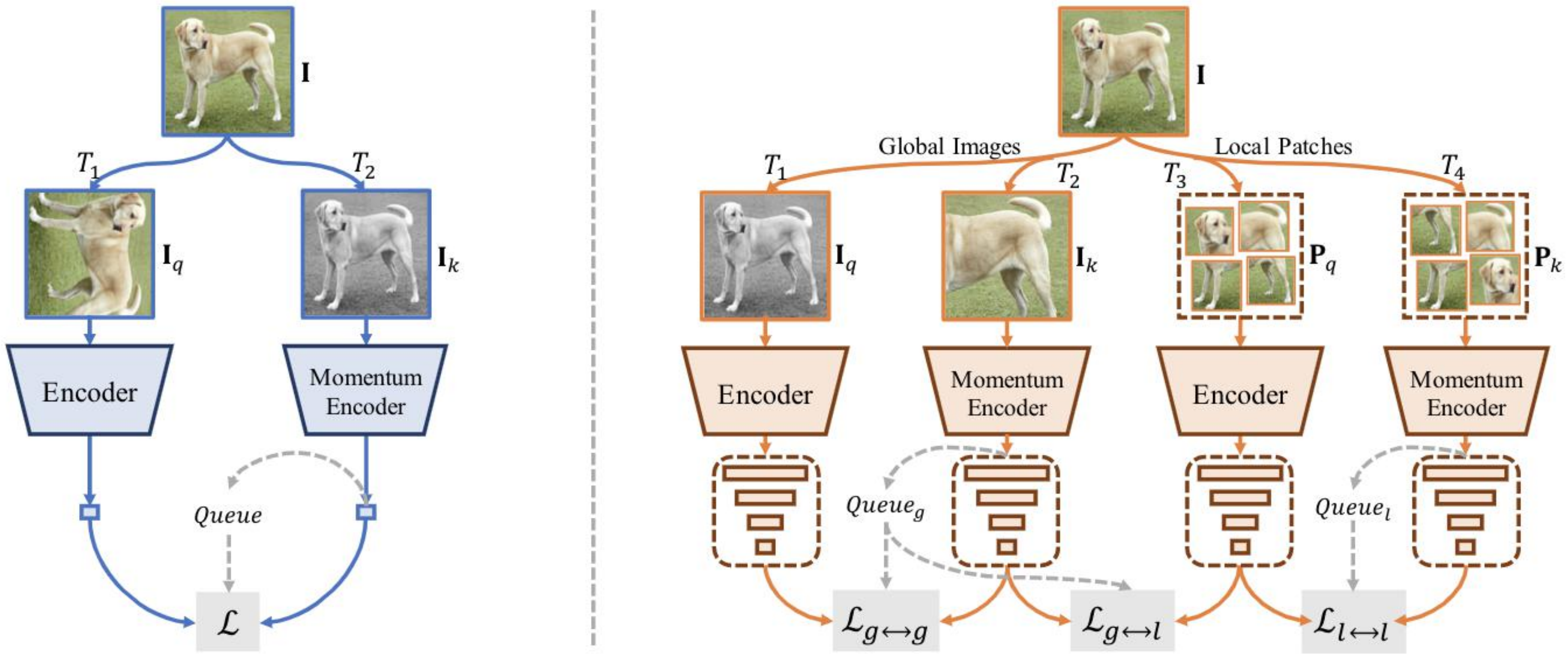
DetCo


Classification and Detection Trade-Off
+0.9 AP
+6.0%
DetCo
MoCo v2
PatchReID
DenseCL
InstLoc
MoCo v1
InstDis
ImageNet Top-1 Accuracy
69
67
65
63
61
59
57
55
37.5
38
38.5
39
39.5
40
COCO Detection mAP


mAP on PASCAL VOC 07+12(%)
58.5
58
57.5
57
56.5
56
55.5
55
54
52
50
0
100
200
300
400
500
600
700
800
Training Epoch
DetCo-800ep
DetCo-400ep
DetCo-200ep
DetCo-100ep
DetCo-50ep
MoCov2-800ep
MoCov2-400ep
MoCov2-200ep
MoCov2-100ep
MoCov2-50ep
SwAV-800ep
MoCo-200ep
PIRL-200ep
Supervised-90ep
BYOL-200ep
SimCLR-200ep

DetCo

Summary： We successfully use large—scale unlabeled data to help downstream object detection tasks.



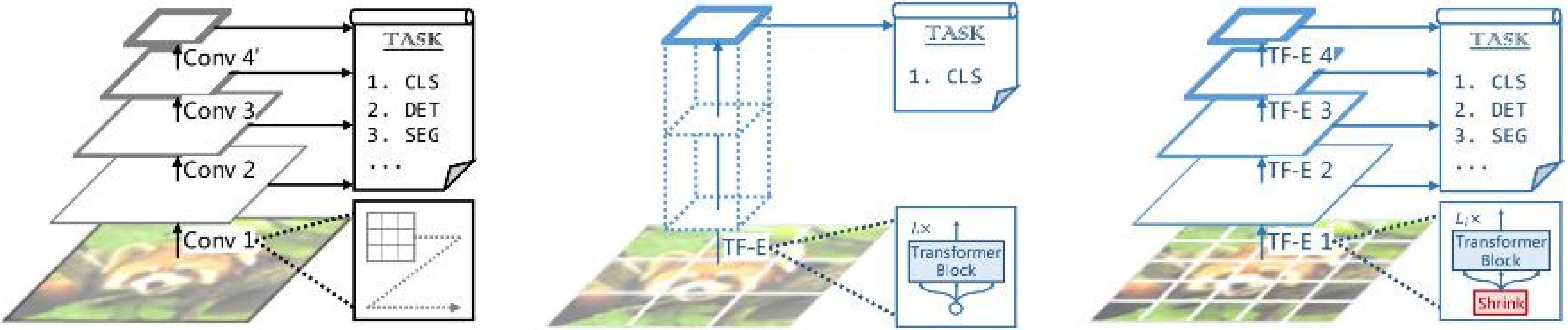
(a) MoCo

(b) DetCo (ours)

Figure 2. **The overall pipeline of DetCo compared with MoCo [19].** (a) is MoCo's framework, which only considers the single high-level feature and learning contrast from a global perspective. (b) is our DetCo, which learns representation with multi-level supervision and adds two additional local patch sets for input, building contrastive loss cross the global and local views. Note that “T” means image transforms. “$Queue_{g/l}$” means different memory banks [40] for global/local features.

PVT：next generation versatile vision backbone

Summary：A pure Transformer backbone for dense prediction, *e.g.* object detection and semantic segmentation.

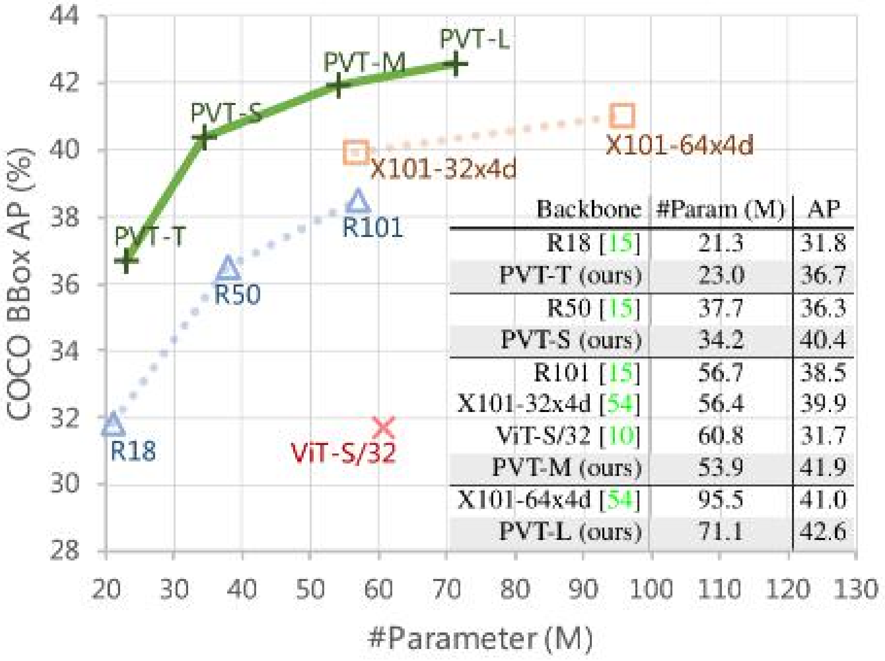


(a) CNNs: VGG [41], ResNet [15], *etc.*

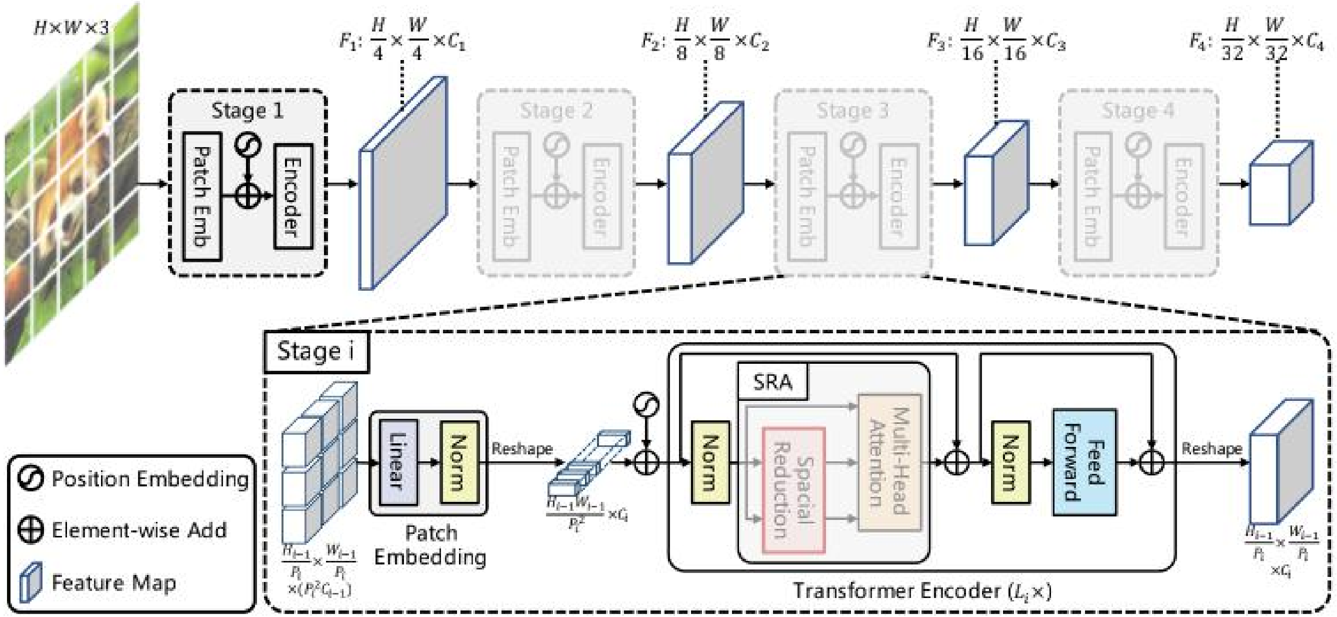
(b) Vision Transformer [10]

(c) Pyramid Vision Transformer (ours)

PVT：next generation versatile vision backbone



Backbone	#Param (M)	AP
R18 [15]	21.3	31.8
PVT-T (ours)	23.0	36.7
R50 [15]	37.7	36.3
PVT-S (ours)	34.2	40.4
R101 [15]	56.7	38.5
X101-32x4d [54]	56.4	39.9
ViT-S/32 [10]	60.8	31.7
PVT-M (ours)	53.9	41.9
X101-64x4d [54]	95.5	41.0
PVT-L (ours)	71.1	42.6



$H\times W\times 3$
$F_1: \frac{H}{4}\times\frac{W}{4}\times C_1$
$F_2: \frac{H}{8}\times\frac{W}{8}\times C_2$
$F_3: \frac{H}{16}\times\frac{W}{16}\times C_3$
$F_4: \frac{H}{32}\times\frac{W}{32}\times C_4$
Stage 1
Patch Emb
Encoder
Stage 2
Patch Emb
Encoder
Stage 3
Patch Emb
Encoder
Stage 4
Patch Emb
Encoder
Stage i
Linear
Norm
Patch Embedding
Reshape
$\frac{H_{i-1}}{P_i}\times\frac{W_{i-1}}{P_i}\times(P_i^2C_{i-1})$
$\frac{H_{i-1}W_{i-1}}{P_i^2}\times C_i$
SRA
Norm
Spacial Reduction
Multi-Head Attention
Norm
Feed Forward
Transformer Encoder ($L_i\times$)
Reshape
$\frac{H_{i-1}}{P_i}\times\frac{W_{i-1}}{P_i}\times C_i$
Position Embedding
Element-wise Add
Feature Map

PVT：next generation versatile vision backbone

- 1. Combine DETR and Trans2Seg, we first build a **pure Transformer** Detection and Segmentation pipeline.

- 2. We show that vision tasks can work better without CNNs.(say goodbye to CNNs)

- 3. improve PVT, making it faster and stronger.