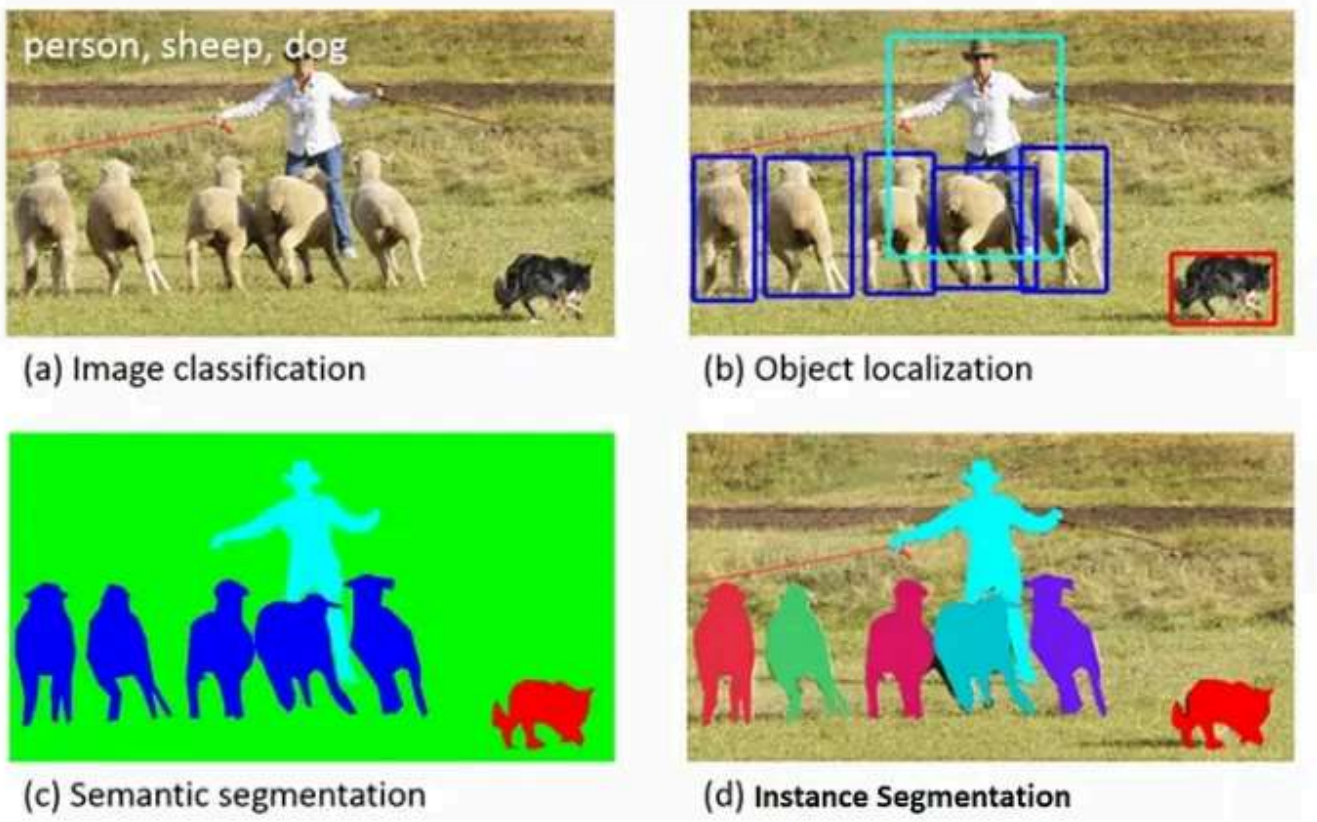
Semantic Segmentation語意分割

分類

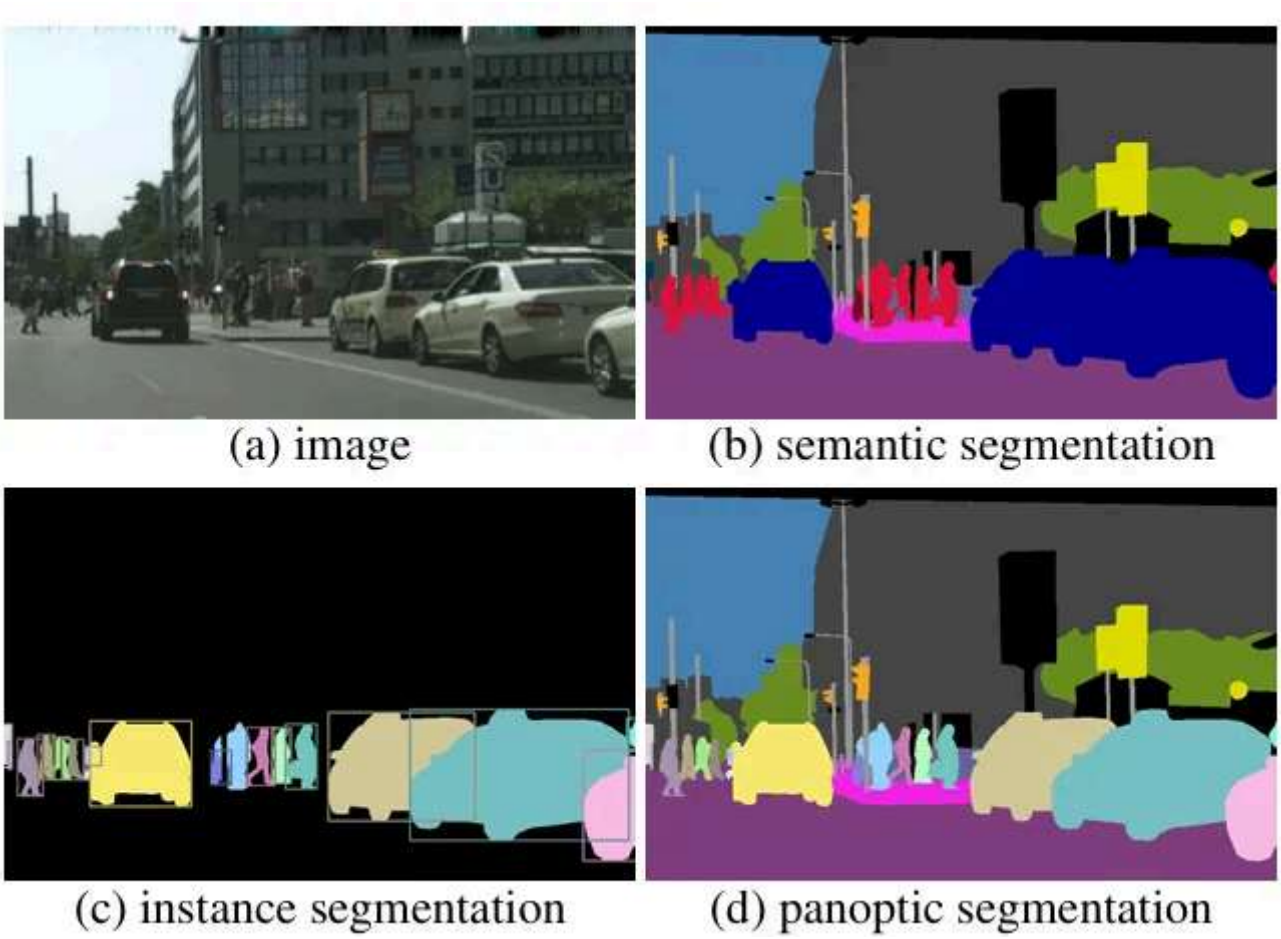
WZX

6 min read · Just now

- Semantic Segmentation (語義分割) 是將圖像中的所有像素點進行分類。
- Instance Segmentation (實例分割) 是物件偵測和語義分割的結合，任務相對較難。針對感興趣的像素點進行分類，並且將各個物件定位，即使是相同類別也會分割成不同物件。
- Panoptic Segmentation (全景分割) 則是更進一步結合了語義分割和實例分割，對各像素進行檢測與分割，同時也將背景考慮進去。



來源



來源

需求

針對風景、街景、房子、道路這類場景的語意切割

選擇:Panoptic Segmentation (全景分割)

- 對於 “Thing” **類別** (可數的、有明確個體的物件，如汽車、行人、動物、房子)，需要區分出 **不同的實例**。
- 對於 “Stuff” **類別** (不可數的、無定形的背景區域，如天空、道路、草地、道路、背景)，只需要分配類別標籤，不需要實例ID。

評估指標

- **PQ (Panoptic Quality):** 全景分割最重要的綜合指標，越高越好。
- **PQ_th (PQ for Things):** 代表模型在 “Thing” 類別上的表現。
- **PQ_st (PQ for Stuff):** 代表模型在 “Stuff” 類別上的表現。
- **mIoU:** 傳統的語意分割指標。
- **AP:** 通常與實例分割的平均精度相關。

Filter: ResNet | ResNet-50 | multi-scale | FPN | Transformer | Neighborhood Attention | single scale | ResNeXt | Swin-Transformer | untagged

Edit Leaderboard

Rank	Model	PQ	PQst	PQth	mIoU	AP	Extra Training Data	Paper	Code	Result	Year	Tags
1	**OneFormer** (ConvNeXt-L, single-scale, 512x1024, Mapillary Vistas-pretrained)	70.1	74.1	64.6	84.6	48.7	✓	OneFormer: One Transformer to Rule Universal Image Segmentation			2022	
2	**Panoptic-DeepLab** (SWideRNet [1, 1, 4.5], Mapillary Vistas, multi-scale)	69.6			85.3	46.8	✓	Scaling Wide Residual Networks for Panoptic Segmentation			2020	
3	**OneFormer** (ConvNeXt-L, single-scale)	68.51			83.0	46.5	×	OneFormer: One Transformer to Rule Universal Image Segmentation			2022	

paperwithcode

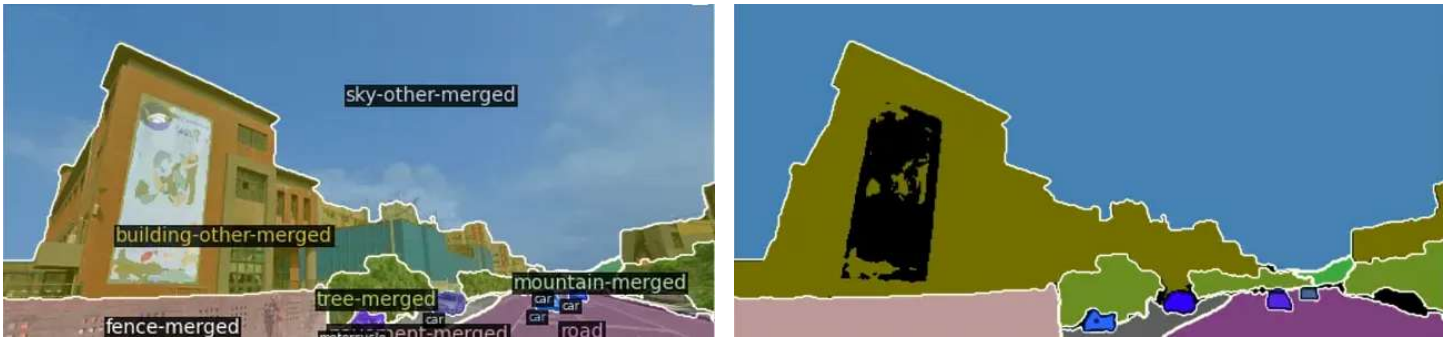
12	**OneFormer** (Swin-L, single-scale)	67.2			83.0	45.6	×	OneFormer: One Transformer to Rule Universal Image Segmentation			2022	

訓練資料集

特性	COCO (133 類)	Cityscapes (19 類)	ADE20K (150 類)
主要應用領域	通用物件，日常多樣場景	都市街景，自動駕駛	極度多樣的室內外場景解析

原圖

預訓練



COCO (133 classes)

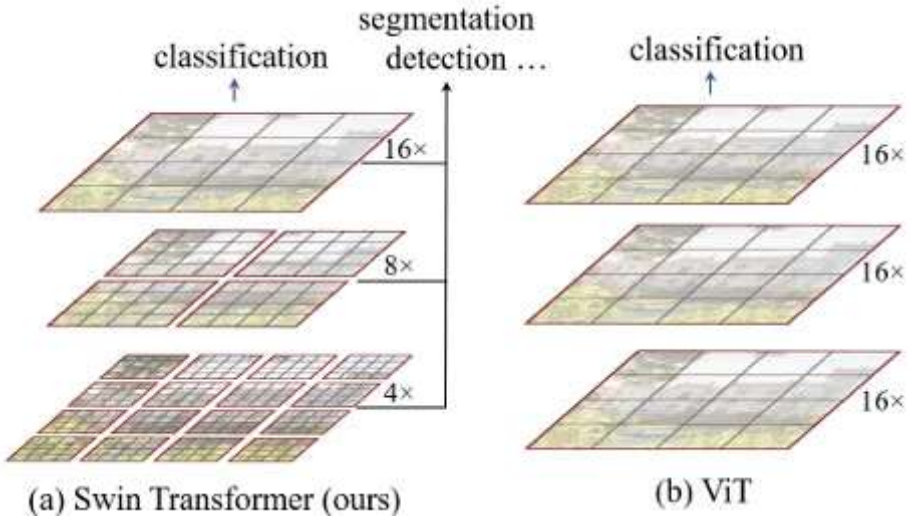
Cityscapes (19 classes)

ADE20K (150 classes)

Backbone

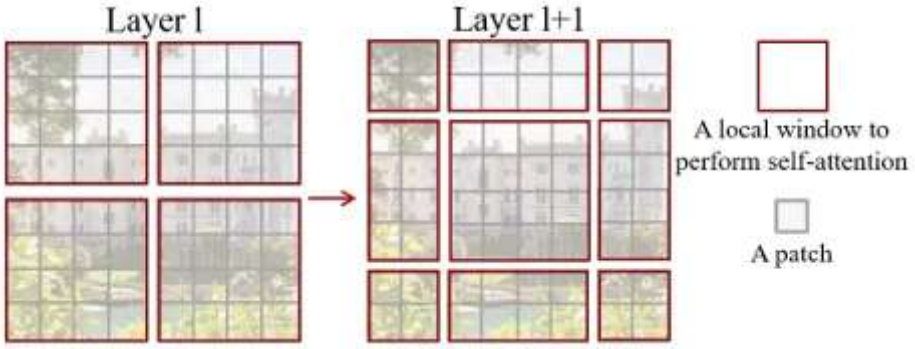
Swin Transformer

- **分層結構 (Hierarchical Structure)：** Swin Transformer 模擬了傳統 CNN 的分層特徵提取方式。在不同階段逐漸減小特徵圖的解析度並增加通道數，從而能夠捕捉不同尺度的視覺信息。



來源

- **滑動窗口注意力 (Shifted Window based Self-Attention)：** 傳統的 Vision Transformer (ViT) 在計算自注意力時，每個 token 都需要與所有其他 token 進行交互，這導致了平方級別的計算複雜度，對於高解析度圖像處理成本很高。Swin Transformer 引入 **基於窗口的自注意力 (Window based Self-Attention)**，將自注意力計算限制在不重疊的局部窗口內。為了實現跨窗口的信息交互，進一步提出了 **滑動窗口 (Shifted Window)** 機制，在連續的 Transformer Block 中交替使用常規窗口和移位後的窗口。

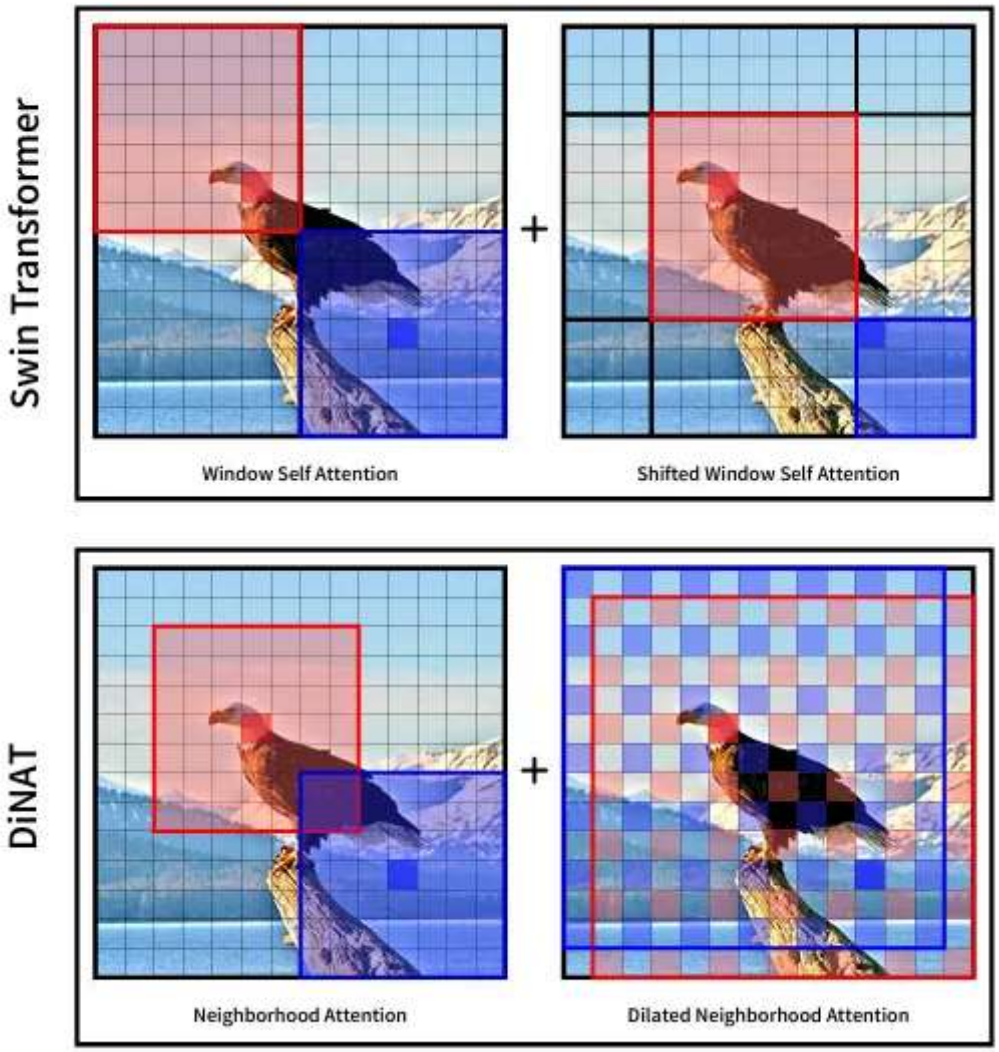


來源

- **相對位置編碼 (Relative Position Bias)：** 在計算自注意力時，Swin Transformer 加入了相對位置偏置，而不是像 ViT 使用絕對位置編碼。這有助於模型更好地理解 token 之間的相對空間關係。

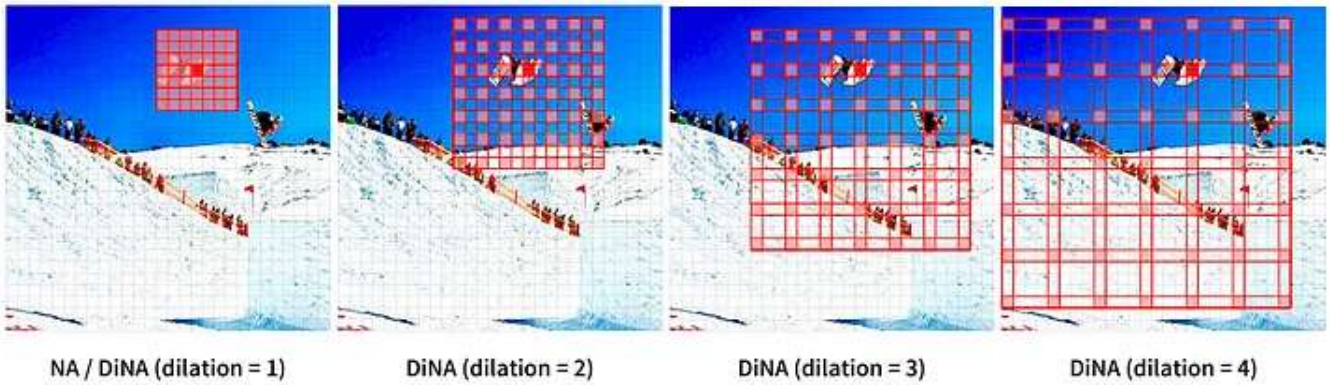
DiNAT (Dilated Neighborhood Attention Transformer)

- **擴展鄰域注意力 (Dilated Neighborhood Attention, DiNA)：** DiNAT 的核心創新在於其擴展鄰域注意力機制。



來源

- 傳統的局部注意力 (如 Swin Transformer 中的窗口注意力) 僅關注一個固定大小的鄰域。
- DiNA 允許注意力機制以不同的 **擴展率 (dilation rate)** 來關注更廣泛的上下文區域，而不需要增加參數數量或計算密度。代表一個 token 可以關注到其周圍更遠的 token，從而獲得更大的感受野和更豐富的上下文信息。

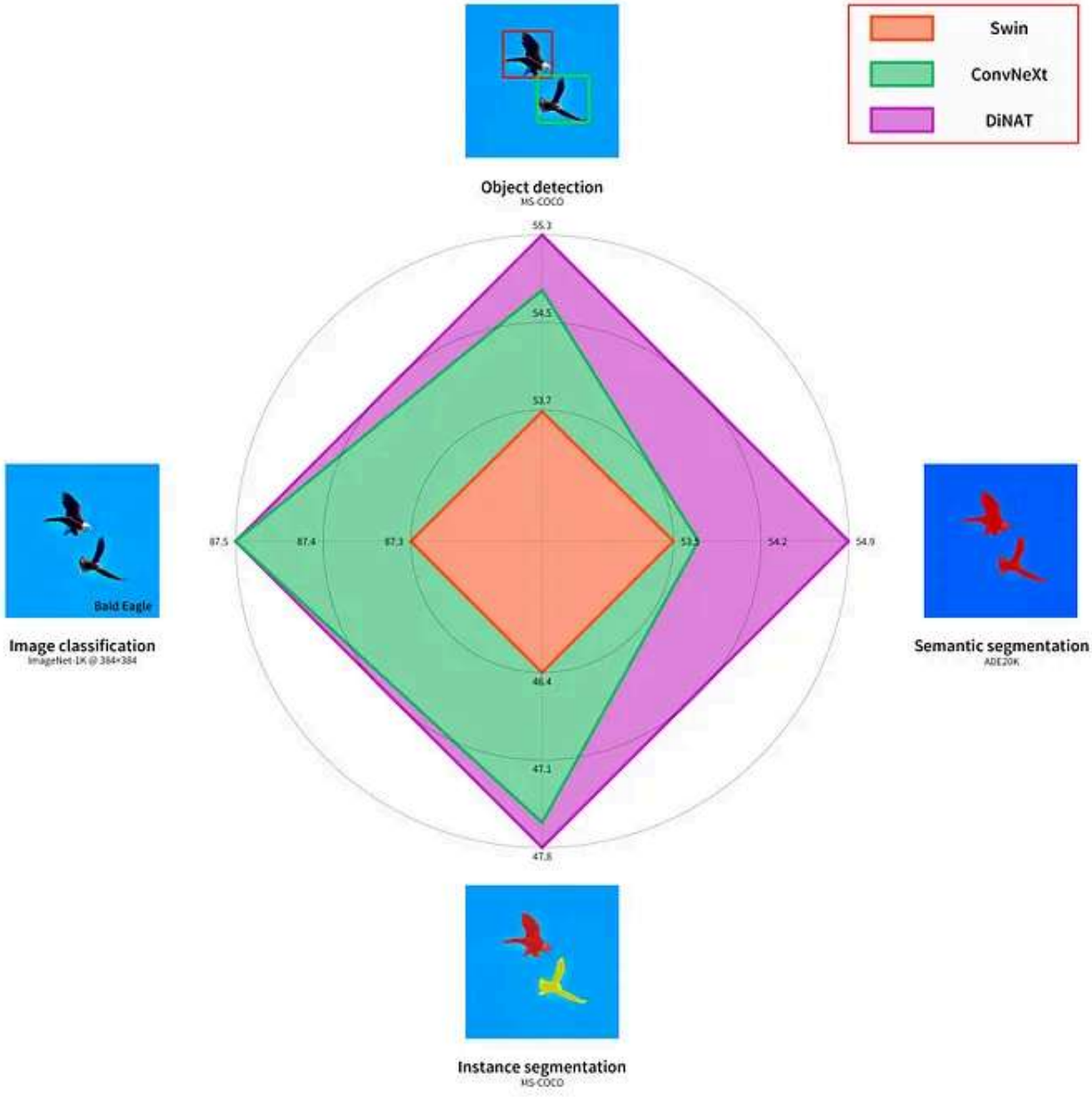


來源

- **多尺度感受野：** 通過在不同層或不同注意力中使用不同的擴展率，DiNAT 可以有效地聚合多尺度的上下文信息。
- **保持局部計算效率：** 儘管引入了擴展，DiNA 仍然保持了局部計算的特性，避免了全域自注意力的高計算成本。

總結比較：

DiNAT-L 的主要創新在於其擴展鄰域注意力，提供比 Swin 的滑動窗口注意力更大的有效感受視野能力，從而帶來性能提升。



來源

論文

Panoptic Segmentation

[1801.00868] Panoptic Segmentation

Swin Transformer

[2103.14030] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

DiNAT

[2209.15001] Dilated Neighborhood Attention Transformer

Semantic Segmentation | Swin Transformer | Dinat