موضوع

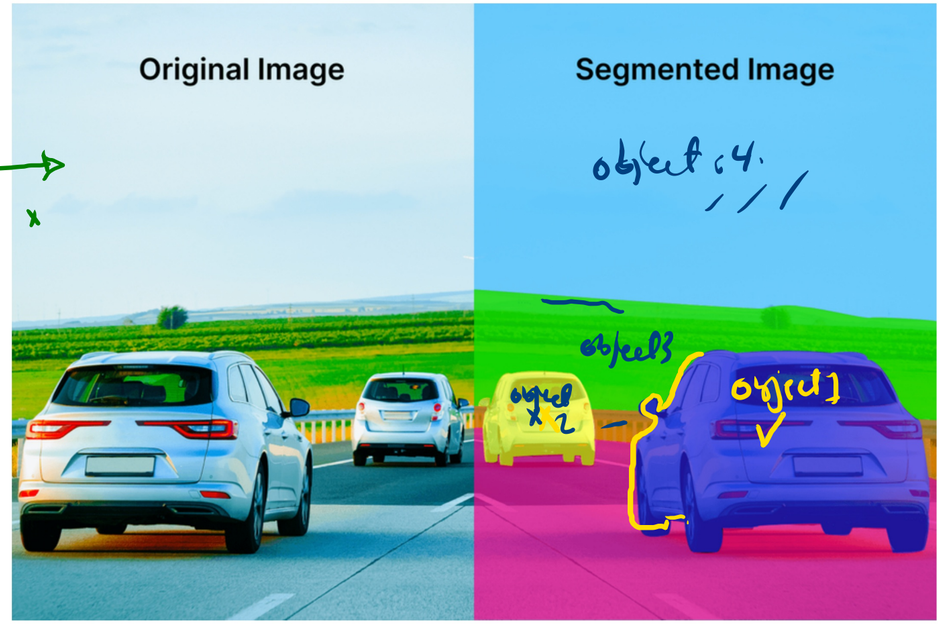
Image Segmentation

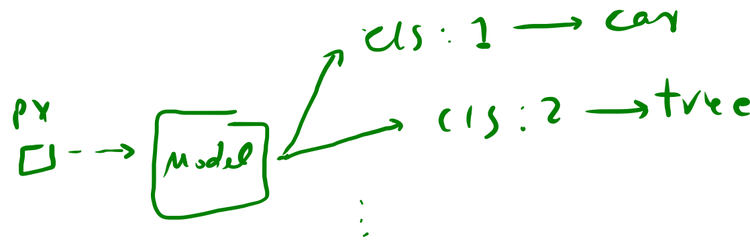
ناحیه بندی تصاویری

[pixel classification]

ناحیه بندی تصویر یک تکنیک بینایی کامپیوتر است که یک تصویر دیجیتال را به گروه‌های مجزا از پیکسل‌ها – بخش‌های تصویر – تقسیم می‌کند تا تشخیص شیء و وظایف مرتبط را ممکن سازد.

https://www.ibm.com/think/topics/image-segmentation



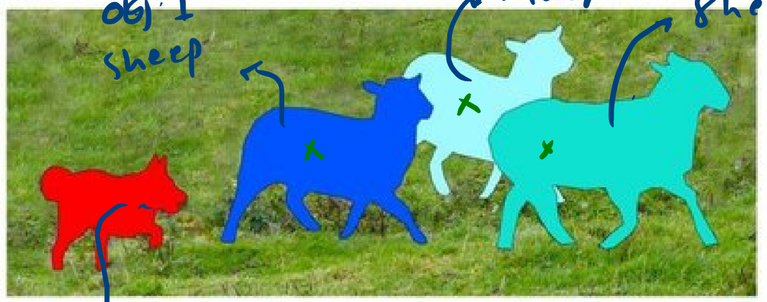


object Recog.

original

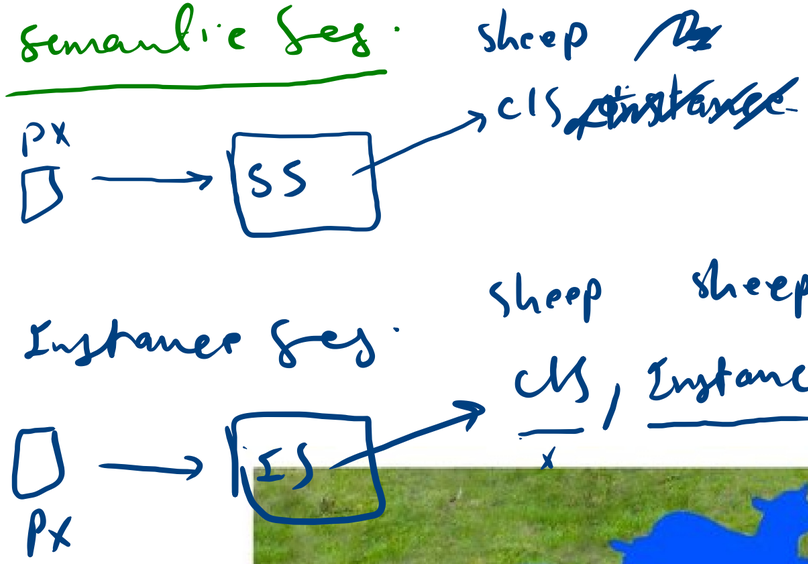
تفاوت سگمنتیشن در این دو تصویر چیست؟

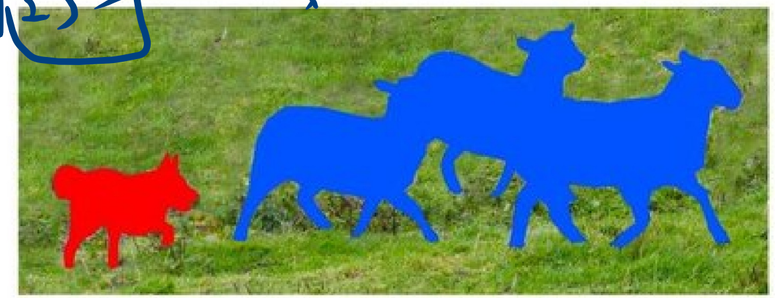
1 seg.



cls
object

2 seg.





Semantic Segmentation

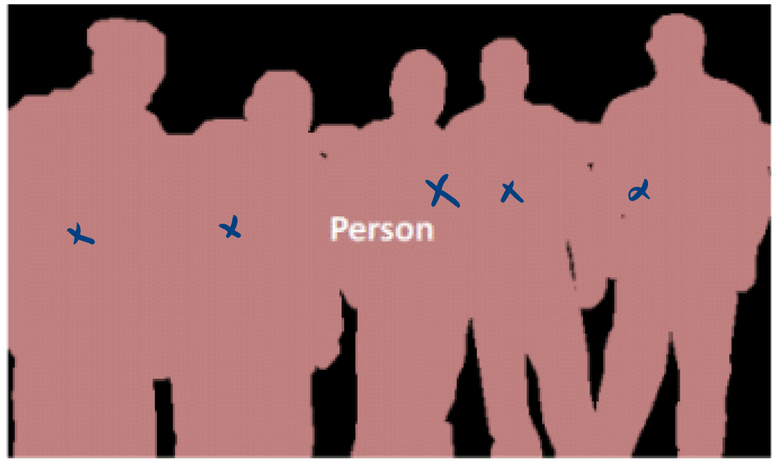
هر پیکسل تصویر را به یک کلاس معنایی اختصاص می‌دهد!

Instance Segmentation

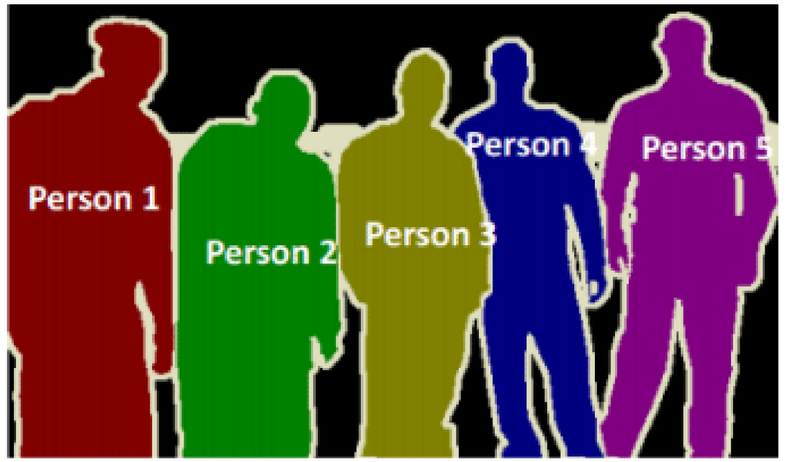
بین نمونه‌های مختلف از یک کلاس تمایز قائل می‌شود!

نوع سگمنتیشن دو تصویر زیر ؟

Semantic



Instance Seg.



Instance Seg.

Panoptic Segmentation

Semantic Seg. + Instance Seg.

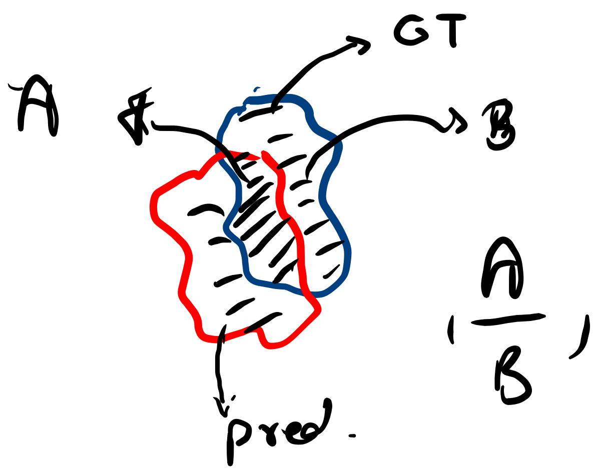
cls

Instance

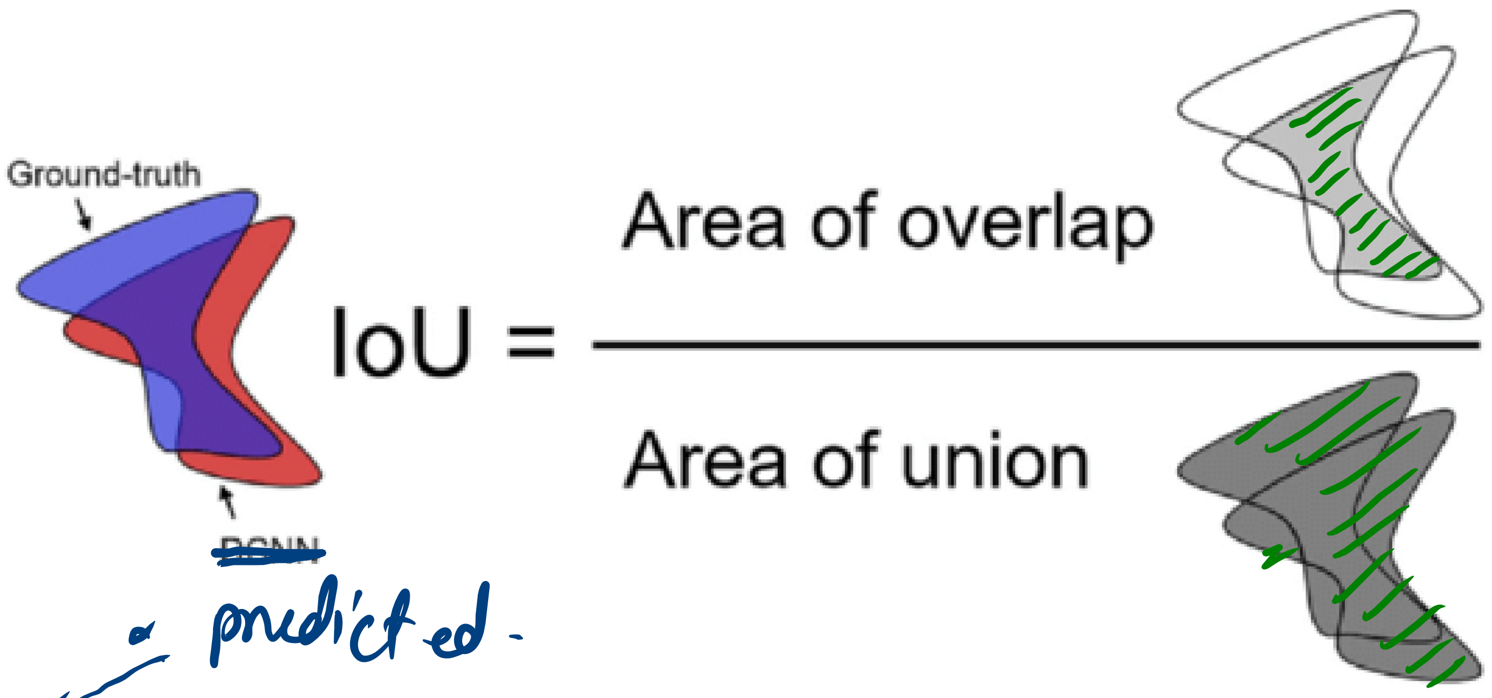
موضوع

Semantic Segmentation

ناحیه بندی معنایی



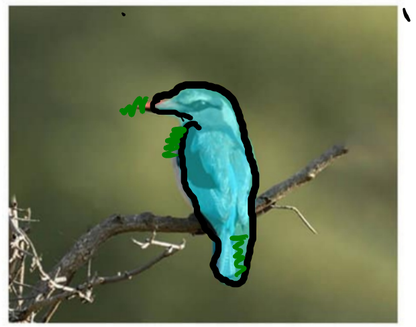
IoU (Intersection over Union)



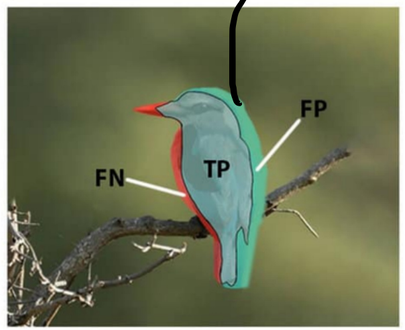
IoU (Intersection over Union)

$$IoU = \frac{TP}{(TP + FP + FN)}$$

Ground Truth Mask



Predicted Mask



mIoU (mean Intersection over Union)

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c},$$

$$\mathrm{mean\_IoU} = \frac{1}{C}\sum_c \mathrm{IoU}_c.$$

IoU_person → person : 0.98

IoU_ball → ball : 0.7

IoU_car → car : 0.82

$$mIOU = \frac{0.98 + 0.7 + 0.82}{3} = \frac{2.5}{3} \approx 0.83$$

Pixel Accuracy (PA)

$$PA = \frac{\text{تعداد پیکسل هایی که درست کلاسیفای شده اند!}}{\text{تعداد کل پیسکل ها}}$$

mean Pixel Accuracy (mPA)

$$mPA = \frac{\text{مجموع تعداد پیکسل هایی که درست کلاسیفای شده اند! (روی همه کلاس ها)}}{\text{تعداد کل پیسکل های همه ی کلاس ها}}$$

TABLE 1. Popular semantic segmentation datasets.

Dataset	Training	Validation	Testing	Classes	Resolution
MS COCO [207]	118K	5000	-	80	vary in size
Pascal VOC 2012 [9]	1464	1449	-	20	vary in size
Pascal Context [10]	4998	5105	9637	59	vary in size
Pascal-Part [211]	1716	-	1817	6	vary in size
Cityscapes [11]	2975	500	1525	19	2048 × 1024
GTA5 [52]	-	-	-	19	1914 × 1052
KITTI [208]	-	-	-	3	vary in size
CamVid [12]	367	101	233	11	960 × 720
ADE20K [13]	20K	2000	-	150	vary in size
SYNTHIA [55]	-	-	-	13	1280 ×960
Stanford Background [213]	-	-	-	8	320 × 240

Heavy and Lightweight Deep Learning Models for Semantic Segmentation: A Survey

CRISTINA CĂRUNTA[1], ALINA CĂRUNTA[2], AND CĂLIN-ADRIAN POPA[1], (Member, IEEE)
[1]Department of Computers and Information Technology, Politehnica University of Timişoara, 300223 Timişoara, Romania
[2]Department of Computer Science, West University of Timişoara, 300223 Timişoara, Romania
Corresponding author: Călin-Adrian Popa (calin.popa@cs.upt.ro)

14 January 2025

Heavy and Lightweight Deep Learning Models for Semantic Segmentation: A Survey

CRISTINA CĂRUNTA[1], ALINA CĂRUNTA[2], AND CĂLIN-ADRIAN POPA[1], (Member, IEEE)
[1]Department of Computers and Information Technology, Politehnica University of Timişoara, 300223 Timişoara, Romania
[2]Department of Computer Science, West University of Timişoara, 300223 Timişoara, Romania
Corresponding author: Călin-Adrian Popa (calin.popa@cs.upt.ro)

14 January 2025

TABLE 2. Performance of complex semantic segmentation models ('𝕋/𝕍' represents test/validation set).

Method	Year	Backbone	VOC 2012	mIoU Cityscapes	ADE20K
FCN [8]	2015	VGG-16	62.2 𝕋	-	-
DeconvNet [16]	2015	VGG-16	72.5 𝕋	-	-
ParseNet [15]	2016	VGG-16	69.8 𝕋	-	-
RefineNet [90]	2017	ResNet-152	83.4 𝕋	-	40.7 𝕍
		ResNet-101	-	73.6 𝕋	-
PSPNet [30]	2017	ResNet-101	85.4 𝕋	80.2 𝕋	-
		ResNet-269	-	-	44.94 𝕍
DeepLabV3+ [84]	2018	Xception	89.0 𝕋	82.1 𝕋	-
DenseASPP [96]	2018	DenseNet-161	-	80.6 𝕋	-
EncNet [92]	2018	ResNet-101	85.9 𝕋	-	44.65 𝕍
DANet [1]	2019	ResNet-101	82.6 𝕋	81.5 𝕋	-
FDNet [193]	2019	-	84.2 𝕋	-	-
CFNet [205]	2019	ResNet-101	87.2 𝕋	79.60 𝕋	44.89 𝕍
CCNet [87]	2020	ResNet-101	-	81.9 𝕋	45.76 𝕍
SANet [88]	2020	ResNet-101	83.2 𝕋	-	-
CDN [191]	2020	-	85.7 𝕋	80.5 𝕋	55.27 𝕋
HRNet [24]	2021	HRNetV2-W48	-	82.5 𝕋	-
SETR [98]	2021	ViT-L	-	81.64 𝕋 82.15 𝕍	50.28 𝕍
TrSeg [99]	2021	DRN [202]	-	79.9 𝕋	-
Segmenter [100]	2021	ViT-L	-	81.3 𝕍	53.63 𝕍
PVT [102]	2021	PVT-Large	-	-	44.8 𝕍
Swin [35]	2021	Swin-L	-	-	53.5 𝕍 62.8 𝕋
FLANet [89]	2022	HRNet-W48	-	83.6 𝕋	46.99 𝕍
		ResNet-150	88.5 𝕋	-	-
SegNeXt [97]	2022	MSCAN-L	90.6 𝕋	-	-
ConvNeXt [42]	2022	ConvNeXt-XL	-	-	54.0 𝕍
PVT V2 [103]	2022	PVT V2-B5	-	-	48.7 𝕍
HRViT [104]	2022	HRViT-b3	-	83.16 𝕍	50.20 𝕍
P2T [105]	2022	P2T-Large	-	-	49.4 𝕍
FsaNet [22]	2023	ResNet-101	-	83.05 𝕋	44.10 𝕍
InternImage [94]	2023	InternImage-H	-	86.1 𝕋 87.0 𝕍	62.9 𝕍
ONE-PEACE [38]	2023	-	-	-	63.0 𝕍
SERNet-Former [222]	2024	Efficient-ResNet	-	87.35 𝕍	-

Heavy and Lightweight Deep Learning Models for Semantic Segmentation: A Survey

CRISTINA CĂRUNTA[1], ALINA CĂRUNTA[2], AND CĂLIN-ADRIAN POPA[1], (Member, IEEE)
[1]Department of Computers and Information Technology, Politehnica University of Timişoara, 300223 Timişoara, Romania
[2]Department of Computer Science, West University of Timişoara, 300223 Timişoara, Romania
Corresponding author: Călin-Adrian Popa (calin.popa@cs.upt.ro)

14 January 2025

TABLE 2. Performance of complex semantic segmentation models. ('𝕋/𝕍' represents test/validation set).

Method	Year	Backbone	mIoU VOC 2012	mIoU Cityscapes	mIoU ADE20K
FCN [8]	2015	VGG-16	62.2 𝕋	-	-
DeconvNet [16]	2015	VGG-16	72.5 𝕋	-	-
ParseNet [15]	2016	VGG-16	69.8 𝕋	-	-
RefineNet [90]	2017	ResNet-152	83.4 𝕋	-	40.7 𝕍
		ResNet-101	-	73.6 𝕋	-
PSPNet [30]	2017	ResNet-101	85.4 𝕋	80.2 𝕋	-
		ResNet-269	-	-	44.94 𝕍
DeepLabV3+ [84]	2018	Xception	89.0 𝕋	82.1 𝕋	-
DenseASPP [96]	2018	DenseNet-161	-	80.6 𝕋	-
EncNet [92]	2018	ResNet-101	85.9 𝕋	-	44.65 𝕍
DANet [1]	2019	ResNet-101	82.6 𝕋	81.5 𝕋	-
FDNet [193]	2019	-	84.2 𝕋	-	-
CFNet [205]	2019	ResNet-101	87.2 𝕋	79.60 𝕋	44.89 𝕍
CCNet [87]	2020	ResNet-101	-	81.9 𝕋	45.76 𝕍
SANet [88]	2020	ResNet-101	83.2 𝕋	-	-
CDN [191]	2020	-	85.7 𝕋	80.5 𝕋	55.27 𝕋
HRNet [24]	2021	HRNetV2-W48	-	82.5 𝕋	-
SETR [98]	2021	ViT-L	-	81.64 𝕋 82.15 𝕍	50.28 𝕍
TrSeg [99]	2021	DRN [202]	-	79.9 𝕋	-
Segmenter [100]	2021	ViT-L	-	81.3 𝕍	53.63 𝕍
PVT [102]	2021	PVT-Large	-	-	44.8 𝕍
Swin [35]	2021	Swin-L	-	-	53.5 𝕍 62.8 𝕋
FLANet [89]	2022	HRNet-W48	-	83.6 𝕋	46.99 𝕍
		ResNet-150	88.5 𝕋	-	-
SegNeXt [97]	2022	MSCAN-L	90.6 𝕋	-	-
ConvNeXt [42]	2022	ConvNeXt-XL	-	-	54.0 𝕍
PVT V2 [103]	2022	PVT V2-B5	-	-	48.7 𝕍
HRViT [104]	2022	HRViT-b3	-	83.16 𝕍	50.20 𝕍
P2T [105]	2022	P2T-Large	-	-	49.4 𝕍
FsaNet [22]	2023	ResNet-101	-	83.05 𝕋	44.10 𝕍
InternImage [94]	2023	InternImage-H	-	86.1 𝕋 87.0 𝕍	62.9 𝕍
ONE-PEACE [38]	2023	-	-	-	63.0 𝕍
SERNet-Former [222]	2024	Efficient-ResNet	-	87.35 𝕍	-

Heavy and Lightweight Deep Learning Models for Semantic Segmentation: A Survey

CRISTINA CĂRUNTA[1], ALINA CĂRUNTA[2], AND CĂLIN-ADRIAN POPA[1], (Member, IEEE)
[1]Department of Computers and Information Technology, Politehnica University of Timişoara, 300223 Timişoara, Romania
[2]Department of Computer Science, West University of Timişoara, 300223 Timişoara, Romania
Corresponding author: Călin-Adrian Popa (calin.popa@cs.upt.ro)

14 January 2025

Method	Year	Backbone	GPU	Resolution	FPS	Params (M)	mIoU
ENet [113]	2016	-	Titan X	640 × 360 1280 × 720 1920 × 1080	135.4 46.8 21.6	0.37	58.3 𝕋
LinkNet [115]	2017	ResNet-18	Titan X	640 × 360 1280 × 720 1920 × 1080	65.8 18.7 8.5	11.5	76.4 𝕍
ERFNet [116]	2018	-	Titan X	640 × 360 1280 × 720 1920 × 1080	83.3 24.4 11.4	-	69.7 𝕋
ContextNet [125]	2018	-	Titan X	2048 × 1024	18.3	0.85	65.9 𝕍
BiSeNet [128]	2018	Xception-39 ResNet-18	Titan XP	2048 × 1024	105.8 65.5	5.8 49.0	68.4 𝕋 74.7 𝕋
ShelfNet [118]	2019	ResNet-18	GTX 1080Ti	2048 × 1024	36.9	-	74.8 𝕋
SwiftNetRN-18 [126]	2019	ResNet-18	GTX 1080Ti	2048 × 1024	39.9	11.8	75.5 𝕋
LEDNet [143]	2019	ResNet	GTX 1080Ti	1024 × 512	71	0.94	70.6 𝕋
FDDWNet [120]	2020	-	RTX 2080Ti	1024 × 512	60	0.80	71.5 𝕋
LDPNet [160]	2020	-	GTX 1080Ti	1024 × 512	87	0.8	71.1 𝕋
CARNet [144]	2020	ResNet-152	Titan X	1024 × 512	29.4	63.3	75.2 𝕋
ERPNet [121]	2021	ResNet-18 MobileNet V3	GTX 2080	896 × 896 780 × 780	66.1 109.5	10.98 1.46	75.3 𝕋 73.6 𝕋
DDRNet [139]	2021	ResNet	GTX 2080Ti	2048 × 1024	101.6	5.7	77.4 𝕋
BiSeNet V2 [129]	2021	-	GTX 1080Ti	2048 × 1024	156	-	72.6 𝕋
SegFormer [39]	2021	MiT-B5	-	1024 × 1024	2.5	84.7	84.0 𝕍
RTFormer [164]	2022	-	RTX 2080Ti	2048 × 1024	110	4.8	76.3 𝕍
RELAXNet [153]	2022	-	GTX 2080Ti	1024 × 512	64	1.9	74.8 𝕋
JPANet [152]	2022	-	GTX 1080Ti	1024 × 512	109.9	3.49	71.62 𝕋
PP-LiteSeg [203]	2022	STDC1 [131]	GTX 1080Ti	1024 × 512	273.6	-	72.0 𝕋 73.1 𝕍
		STDC2 [131]		1536 × 768	102.6		77.5 𝕋 78.2 𝕍
BFMNet [140]	2023	ResNet-18	RTX 2080Ti	1536 × 768 2048 × 1024	63.7 31.4	22.3	77.7 𝕋 78.9 𝕋
MLFNet [141]	2023	ResNet-18 ResNet-34	Titan XP	1024 × 512	95.1 72.2	9.90 13.03	71.0 𝕋 72.1 𝕋
BiSeNet V3 [135]	2023	STDC2	GTX 1080Ti	1536 × 768	93.8	-	79.0 𝕋
LETNet [166]	2023	-	RTX 3090	1024 × 512	150	0.95	72.8 𝕋
AFFormer [168]	2023	-	V100 NVIDIA	2048 × 1024	22	3.0	78.7 𝕍
PIDNet [33]	2023	ResNet	RTX 3090	2048 × 1024	93.2 31.1	7.6 36.9	78.6 𝕋 80.6 𝕋
DSNet [221]	2024	-	RTX 3090	2048 × 1024	37.6	6.8	80.4 𝕍

TABLE 3. Performance of real-time models on Cityscapes dataset. ('𝕋/𝕍' represents test/validation set; 'FPS' denotes frames per second; 'Params (M)' is the number of learnable parameters).

سگمنتیشن معنایی : مقایسه مدل ها

Heavy and Lightweight Deep Learning Models for Semantic Segmentation: A Survey

CRISTINA CĂRUNTA[1], ALINA CĂRUNTA[2], AND CĂLIN-ADRIAN POPA[1], (Member, IEEE)
[1]Department of Computers and Information Technology, Politehnica University of Timişoara, 300223 Timişoara, Romania
[2]Department of Computer Science, West University of Timişoara, 300223 Timişoara, Romania
Corresponding author: Călin-Adrian Popa (calin.popa@cs.upt.ro)

14 January 2025

mIOU > 70

FPS > 100

Method	Year	Backbone	GPU	Resolution	FPS	Params (M)	mIoU
ENet [113]	2016	-	Titan X	640 × 360 1280 × 720 1920 × 1080	135.4 46.8 21.6	0.37	58.3 𝕋
LinkNet [115]	2017	ResNet-18	Titan X	640 × 360 1280 × 720 1920 × 1080	65.8 18.7 8.5	11.5	76.4 𝕍
ERFNet [116]	2018	-	Titan X	640 × 360 1280 × 720 1920 × 1080	83.3 24.4 11.4	-	69.7 𝕋
ContextNet [125]	2018	-	Titan X	2048 × 1024	18.3	0.85	65.9 𝕍
BiSeNet [128]	2018	Xception-39 ResNet-18	Titan XP	2048 × 1024	105.8 65.5	5.8 49.0	68.4 𝕋 74.7 𝕋
ShelfNet [118]	2019	ResNet-18	GTX 1080Ti	2048 × 1024	36.9	-	74.8 𝕋
SwiftNetRN-18 [126]	2019	ResNet-18	GTX 1080Ti	2048 × 1024	39.9	11.8	75.5 𝕋
LEDNet [143]	2019	ResNet	GTX 1080Ti	1024 × 512	71	0.94	70.6 𝕋
FDDWNet [120]	2020	-	RTX 2080Ti	1024 × 512	60	0.80	71.5 𝕋
LDPNet [160]	2020	-	GTX 1080Ti	1024 × 512	87	0.8	71.1 𝕋
CARNet [144]	2020	ResNet-152	Titan X	1024 × 512	29.4	63.3	75.2 𝕋
ERPNet [121]	2021	ResNet-18 MobileNet V3	GTX 2080	896 × 896 780 × 780	66.1 109.5	10.98 1.46	75.3 𝕋 73.6 𝕋
DDRNet [139]	2021	ResNet	GTX 2080Ti	2048 × 1024	101.6	5.7	77.4 𝕋
BiSeNet V2 [129]	2021	-	GTX 1080Ti	2048 × 1024	156	-	72.6 𝕋
SegFormer [39]	2021	MiT-B5	-	1024 × 1024	2.5	84.7	84.0 𝕍
RTFormer [164]	2022	-	RTX 2080Ti	2048 × 1024	110	4.8	76.3 𝕍
RELAXNet [153]	2022	-	GTX 2080Ti	1024 × 512	64	1.9	74.8 𝕋
JPANet [152]	2022	-	GTX 1080Ti	1024 × 512	109.9	3.49	71.62 𝕋
PP-LiteSeg [203]	2022	STDC1 [131] STDC2 [131]	GTX 1080Ti	1024 × 512 1536 × 768	273.6 102.6	-	72.0 𝕋 73.1 𝕍 77.5 𝕋 78.2 𝕍
BFMNet [140]	2023	ResNet-18	RTX 2080Ti	1536 × 768 2048 × 1024	63.7 31.4	22.3	77.7 𝕋 78.9 𝕋
MLFNet [141]	2023	ResNet-18 ResNet-34	Titan XP	1024 × 512	95.1 72.2	9.90 13.03	71.0 𝕋 72.1 𝕋
BiSeNet V3 [135]	2023	STDC2	GTX 1080Ti	1536 × 768	93.8	-	79.0 𝕋
LETNet [166]	2023	-	RTX 3090	1024 × 512	150	0.95	72.8 𝕋
AFFormer [168]	2023	-	V100 NVIDIA	2048 × 1024	22	3.0	78.7 𝕍
PIDNet [33]	2023	ResNet	RTX 3090	2048 × 1024	93.2 31.1	7.6 36.9	78.6 𝕋 80.6 𝕋
DSNet [221]	2024	-	RTX 3090	2048 × 1024	37.6	6.8	80.4 𝕍

TABLE 3. Performance of real-time models on Cityscapes dataset. ('𝕋/𝕍' represents test/validation set; 'FPS' denotes frames per second; 'Params (M)' is the number of learnable parameters).

https://paperswithcode.com/sota/semantic-segmentation-on-ade20k

https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012

https://paperswithcode.com/sota/semantic-segmentation-on-cityscapes

Semantic seg.

InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions

Wenhai Wang[1*], Jifeng Dai[2,1*], Zhe Chen[3,1*], Zhenhang Huang[1*], Zhiqi Li[3,1*], Xizhou Zhu[4*],
Xiaowei Hu[1], Tong Lu[3], Lewei Lu[4], Hongsheng Li[5], Xiaogang Wang[4,5], Yu Qiao[1✉]
[1]Shanghai AI Laboratory [2]Tsinghua University
[3]Nanjing University [4]SenseTime Research [5]The Chinese University of Hong Kong

https://github.com/OpenGVLab/InternImage

The End