Evaluation Metrics for Image Segmentation

Semantic Segmentation (Binary)

Metric	Technical Name	Meaning	Example	Analogy	Use Case
Accuracy	Acc	Fraction of correctly labeled pixels	Model labels most cat and sofa pixels correctly	Student answers most questions correctly	Simple and intuitive overall measure
`Acc = (TP + TN) / (TP + TN + FP + FN)`					
Precision	P	Of predicted “cat” pixels, how many are truly cat	Model wrongly labels sofa as cat	Doctor gives wrong diagnosis	Good when false positives are costly
`P = TP / (TP + FP)`					
Recall	R	Of all true cat pixels, how many predicted	Model misses cat’s tail	Metal detector misses some coins	Good when false negatives are costly
`R = TP / (TP + FN)`					
F1 Score / Dice	F1 / DSC	Balance of precision & recall	Model finds cat but includes sofa pixels	Balancing accuracy and completeness	Useful for imbalanced classes
`F1 = 2TP / (2TP + FP + FN)`					
Intersection over Union	IoU / Jaccard	Overlap ÷ total area	Predicted cat overlaps 80% with true cat	Two circles overlap 80%	Standard benchmark for segmentation
`IoU = TP / (TP + FP + FN)`					

Semantic Segmentation (Multi-Class)

Metric	Technical Name	Meaning	Example	Analogy	Use Case
Pixel Accuracy	PA	Overall % pixels correct	City scene: model labels sky correctly, dominates result	Exam score inflated by easy questions	Quick overall measure but biased

$PA = \Sigma\ TP_i\ /\ \Sigma\ (TP_i + FP_i + FN_i)$					
Mean Pixel Accuracy	mPA	Accuracy averaged per class	Small traffic sign accuracy matters equally	Teacher averages grades across subjects	Reduces bias from large classes
$mPA = (1/N)\ \Sigma\ (TP_i\ /\ (TP_i + FN_i))$					
Mean IoU	mIoU	IoU averaged across classes	Average IoU of road, car, sky	Report card with subject averages	Most widely used segmentation metric
$mIoU = (1/N)\ \Sigma\ (TP_i\ /\ (TP_i + FP_i + FN_i))$					
Mean Dice Score	mDice	Dice averaged across classes	Dice score for road, car, sky averaged	Balanced grading across subjects	Balances recall and precision across classes
$mDice = (1/N)\ \Sigma\ (2TP_i\ /\ (2TP_i + FP_i + FN_i))$					
Frequency Weighted IoU	FWIoU	IoU weighted by class size	Road pixels dominate image so weighted more	Major subjects count more in GPA	Reflects impact of frequent classes
$FWIoU = (1/\Sigma\ n_k)\ \Sigma\ n_i\ \cdot\ (TP_i\ /\ (TP_i + FP_i + FN_i))$					

Instance Segmentation

Metric	Technical Name	Meaning	Example	Analogy	Use Case
Average Precision	AP	Detecting + segmenting objects	Street scene: model finds only 2 of 3 cars	Counting apples but missing one	Balances precision & recall for objects
$AP = \int_0^1 P(R)\ dR$					
mean Average Precision	mAP	AP across IoU thresholds	Car passes loose overlap but fails strict	Graded both for correctness and neatness	Comprehensive object-level evaluation
$mAP = (1/T)\ \Sigma\ AP(IoU=t)$					
AP_small / AP_medium / AP_large	AP_S / AP_M / AP_L	Performance by object size	Detects big bus but misses tiny sign	Reads big letters but not small ones	Reveals sensitivity to object scale
AP formula restricted by object size					

Mask IoU	Mask IoU	IoU for each matched mask	Dog mask overlaps 70% with ground truth	Two shapes overlap by 70%	Direct overlap measure for masks
$IoU = \|M_{pred} \cap M_{gt}\| / \|M_{pred} \cup M_{gt}\|$					
Boundary F-measure	BF	Quality of edges	Tumor boundary predicted with blurred edges	Coloring outside the lines	Captures contour accuracy
$BF = 2 \cdot (P_b \cdot R_b) / (P_b + R_b)$					

Panoptic Segmentation

Metric	Technical Name	Meaning	Example	Analogy	Use Case
Panoptic Quality	PQ	Overall quality (shape × detection)	Model finds road and sky but misses a car	Cooked tasty dish but forgot side item	Single metric combining segmentation & detection
$PQ = \Sigma\ IoU(p,g) / (\|TP\| + 0.5\|FP\| + 0.5\|FN\|)$					
Segmentation Quality	SQ	Accuracy of object shapes	Car mask slightly smaller than ground truth	Cookie cutter shape slightly off	Focuses on mask quality
$SQ = \Sigma\ IoU(p,g) / \|TP\|$					
Recognition Quality	RQ	Correctness of object detection	Model detects 2 of 3 cars	2 students present, 1 absent	Focuses on object detection quality
$RQ = \|TP\| / (\|TP\| + 0.5\|FP\| + 0.5\|FN\|)$					

Specialized Metrics (Medical & 3D)

Metric	Technical Name	Meaning	Example	Analogy	Use Case
Hausdorff Distance	HD	Worst boundary error	Tumor boundary missed by 5 mm	Measuring largest gap with ruler	Highlights worst-case error
$HD = \max\ \{\ \sup_{x \in A} \inf_{y \in B} d(x,y),\ \sup_{y \in B} \inf_{x \in A} d(x,y)\ \}$					

Average Symmetric Surface Distance	ASSD	Average boundary error	Tumor edges off by ~1 mm	Average gap between outlines	Smoother measure than HD
$ASSD = (1/\|A\|+1/\|B\|) (\Sigma\ d(a,B) + \Sigma\ d(b,A))$					
Volumetric Overlap Error	VOE	Overlap error in 3D	Liver segmentation overlaps 85% with truth	Two Lego models overlap only partly	Good for volumetric structures
$VOE = 1 - (\|A \cap B\| / \|A \cup B\|)$					