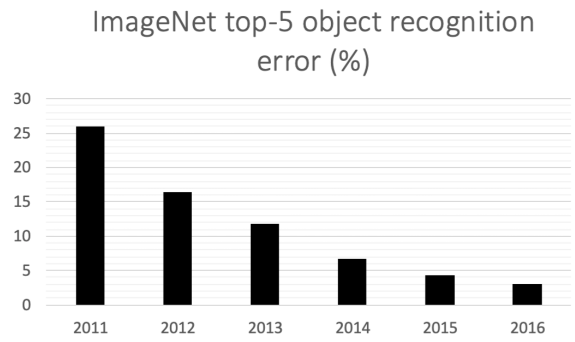
"Deep Dive": Beyond ImageNet, what's other benchmark and metrics commonly used in CV?

In the class, ImageNet is mentioned to be the booster of modern deep learning algorithms, so I want to deep dive into how benchmark and metrics evolve. This report mainly focus on Computer Vision (CV).

Impact of Deep Learning in Computer Vision

ImageNet 1000-object category recognition challenge



But the neural networks you have seen so far won't work well on images!

1. Critical Role Benchmarks Play
2. Mathematical Definition of Core Metrics

Critical Role Benchmarks Play

Introduction to Machine Learning Evaluation

A misconception in machine learning is the conflation of **loss functions** with **evaluation metrics**. While both provide scalar signals regarding model performance, their utility diverges significantly in practice and intent.

Loss Function is a differentiable function used during training to optimize model parameters via gradient descent. It provides a smooth optimization landscape (e.g., Cross-Entropy, MSE).

Evaluation Metric measures real-world performance and is often non-differentiable. Unlike loss functions that penalize confidence levels, metrics like Accuracy treat predictions categorically.

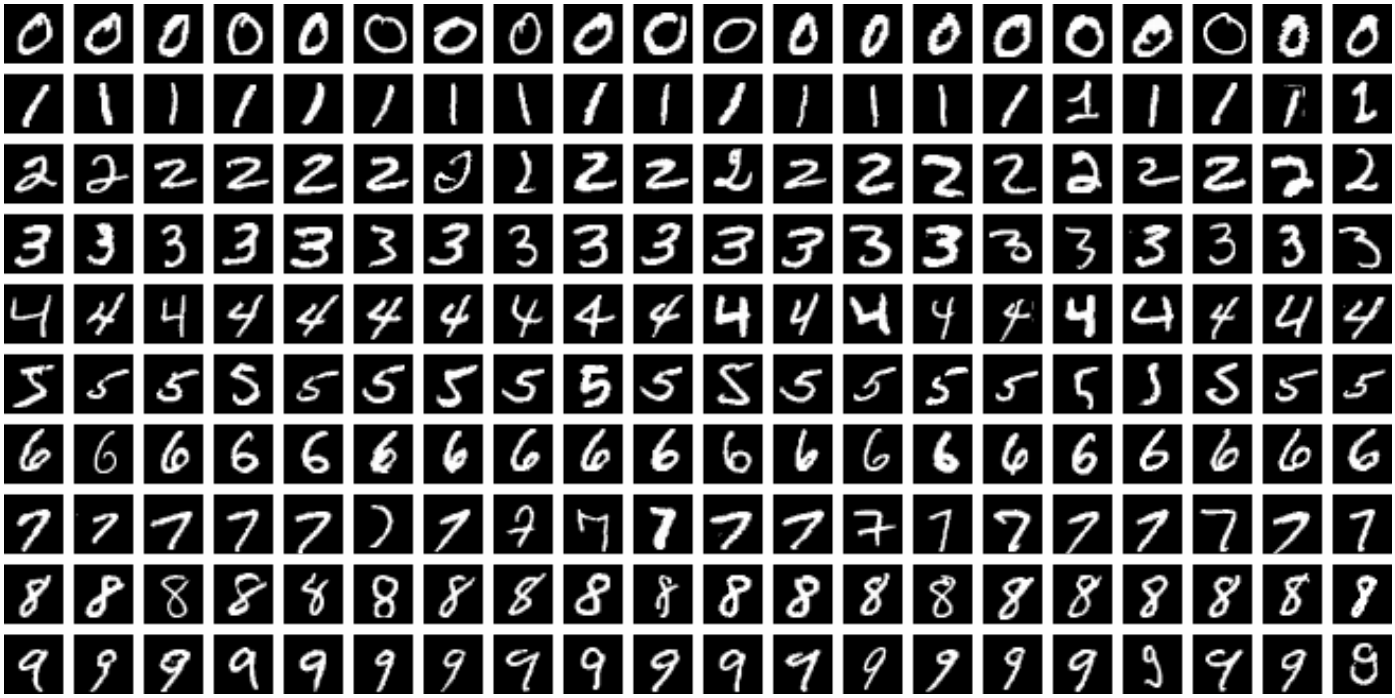
A benchmark is a standardized experimental framework that enables fair comparison of different models by holding evaluation conditions constant. It consists of three components:

- **Dataset**: Proper train/validation/test splits to prevent data leakage
- **Task Definition**: Precise specification of inputs and expected outputs
- **Evaluation Protocol**: Standardized metrics and scoring methods

Benchmarks evolve over time—when models saturate a benchmark by exceeding human performance, new, more challenging benchmarks emerge (e.g., GLUE to SuperGLUE in NLP) to continue driving progress.

Computer Vision: From Pixels to Perception

In the late 1980s and 1990s, computer vision was largely defined by the struggle to perform basic pattern recognition. The release of the **MNIST (Modified National Institute of Standards and Technology)** database in 1998 by Yann LeCun and colleagues marked a watershed moment. MNIST consists of 60,000 training images and 10,000 test images of grayscale digits (28×28 pixels).

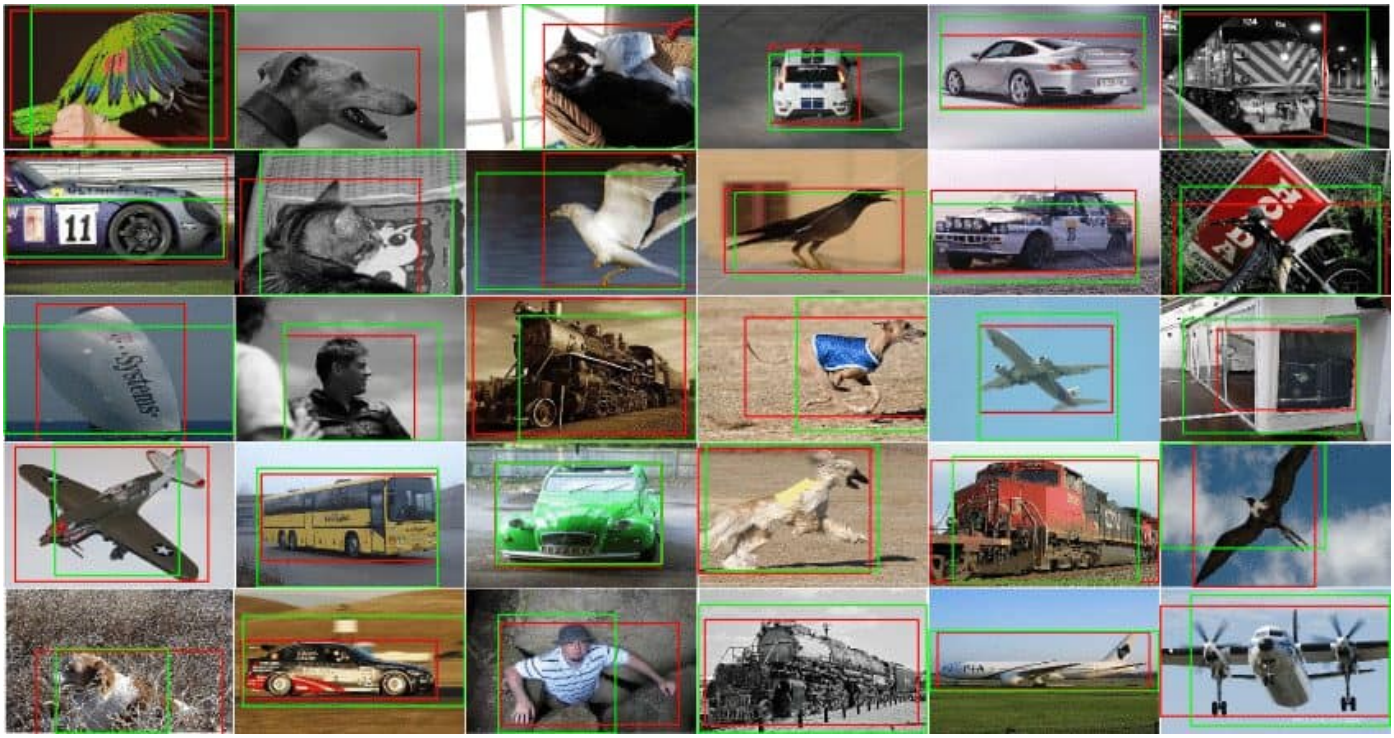


If MNIST was the cradle of computer vision, **ImageNet** was its proving ground.

ImageNet, conceived by Fei-Fei Li in 2006, consists of 14 million images across 20,000 categories using crowdsourced annotation. The pivotal moment came in 2012 when AlexNet reduced error rates from 26% to 15.3%, marking a paradigm shift from handcrafted features to learned representations. Beyond the competition itself, ImageNet established the standard workflow of pre-training on large-scale data followed by fine-tuning on task-specific datasets, creating the foundation for transfer learning that powers modern applied AI.

As ImageNet classification saturated, focus shifted to object detection and segmentation.

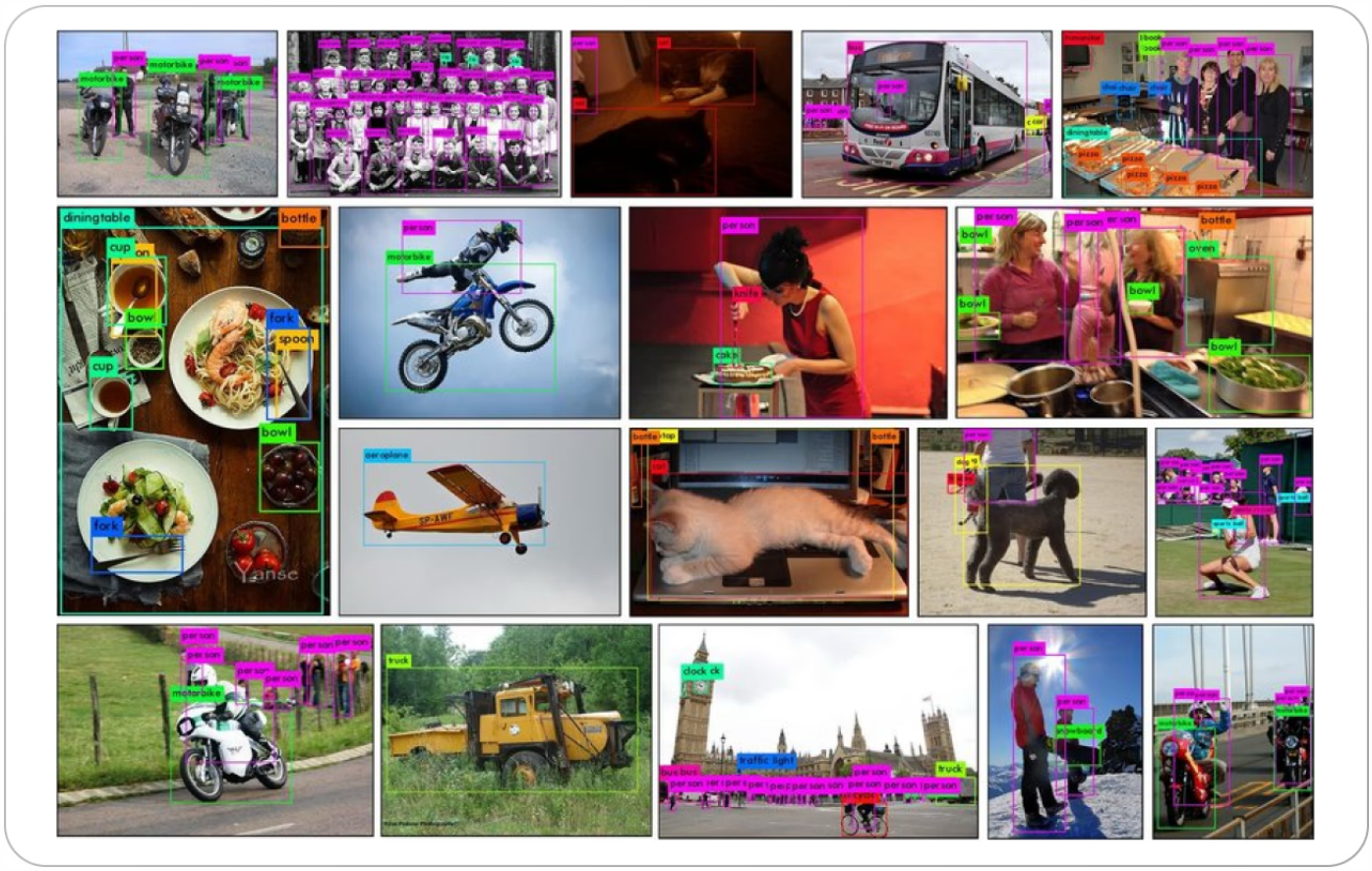
PASCAL VOC (2005-2012) introduced 20 object classes and IoU-based bounding box evaluation, but featured "iconic" images with centered, well-lit objects that models began saturating by 2012.



MS COCO (2014) pushed toward non-iconic scene understanding with three key differences:

- **Contextual Complexity**: Objects are often small, occluded, or in backgrounds, requiring scene understanding
- **Instance Density**: Averages 7.7 objects per image vs. 3 in PASCAL VOC
- **Task Diversity**: Supports detection, instance segmentation, keypoint detection, and captioning

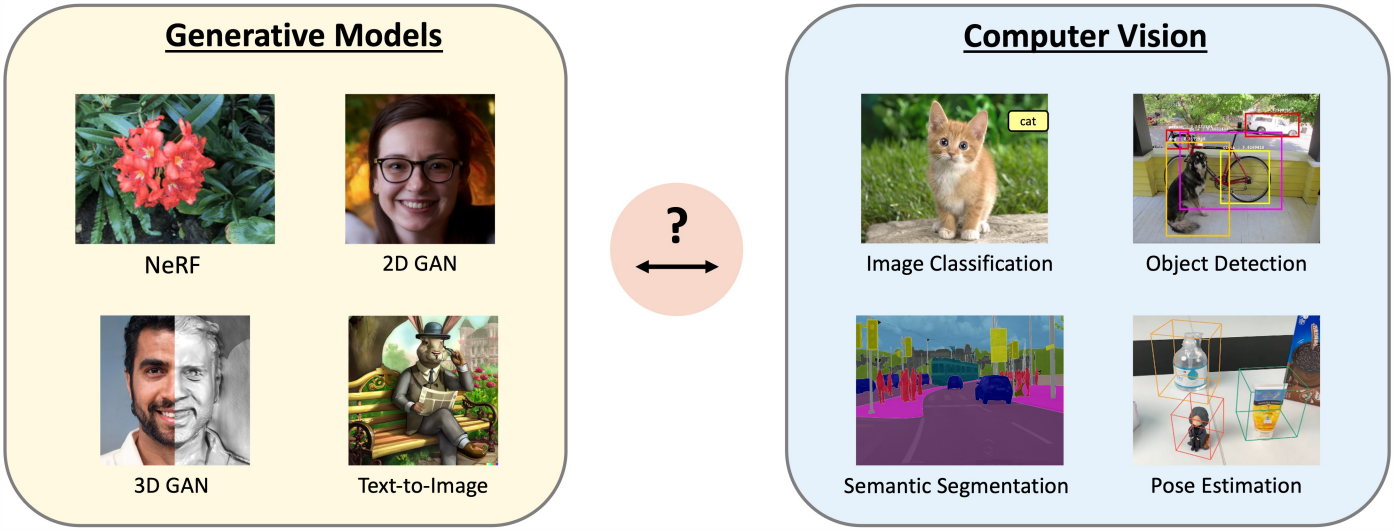
COCO's difficulty is reflected in its stricter metrics—mAP@[.50:.05:.95] averages precision across ten IoU thresholds (0.50 to 0.95), demanding spatial precision rather than rough correctness.



Computer Vision Metrics: A Story of Broken Assumptions

Generative Vision: The Evaluation Crisis

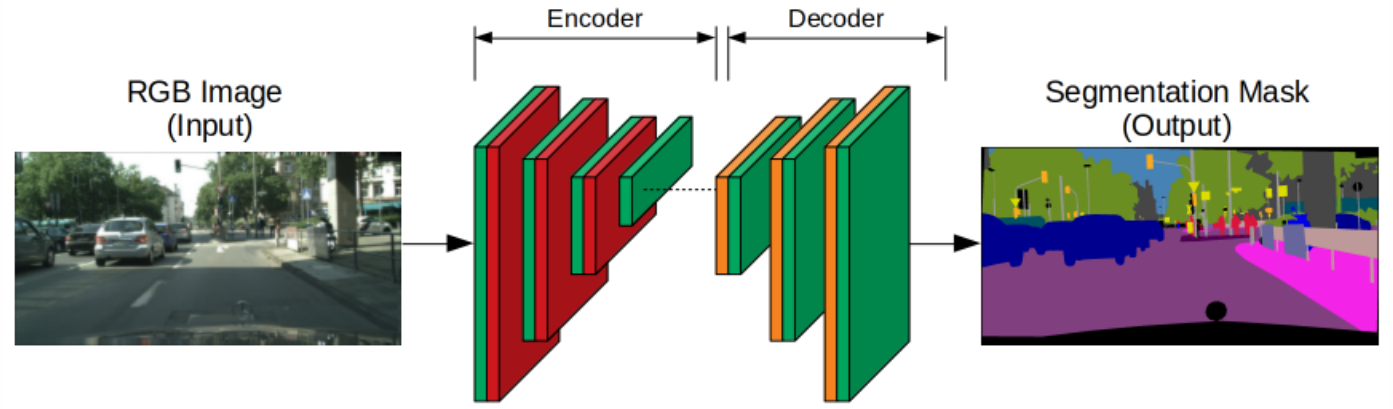
GANs and diffusion models broke the rules—no single ground truth exists. This forced the field to invent distributional metrics: **Inception Score** for clarity and diversity, **FID** for detecting mode collapse, and text-to-image benchmarks like **DrawBench** testing compositional reasoning with adversarial prompts ("a horse riding an astronaut").



But before we could evaluate what models create, we had to fix how we measure what they see.

Segmentation: When 90% Accuracy Means Nothing

Your self-driving car reports 90% pedestrian detection accuracy. Great! Except it labels everything as "road" and misses all the people.



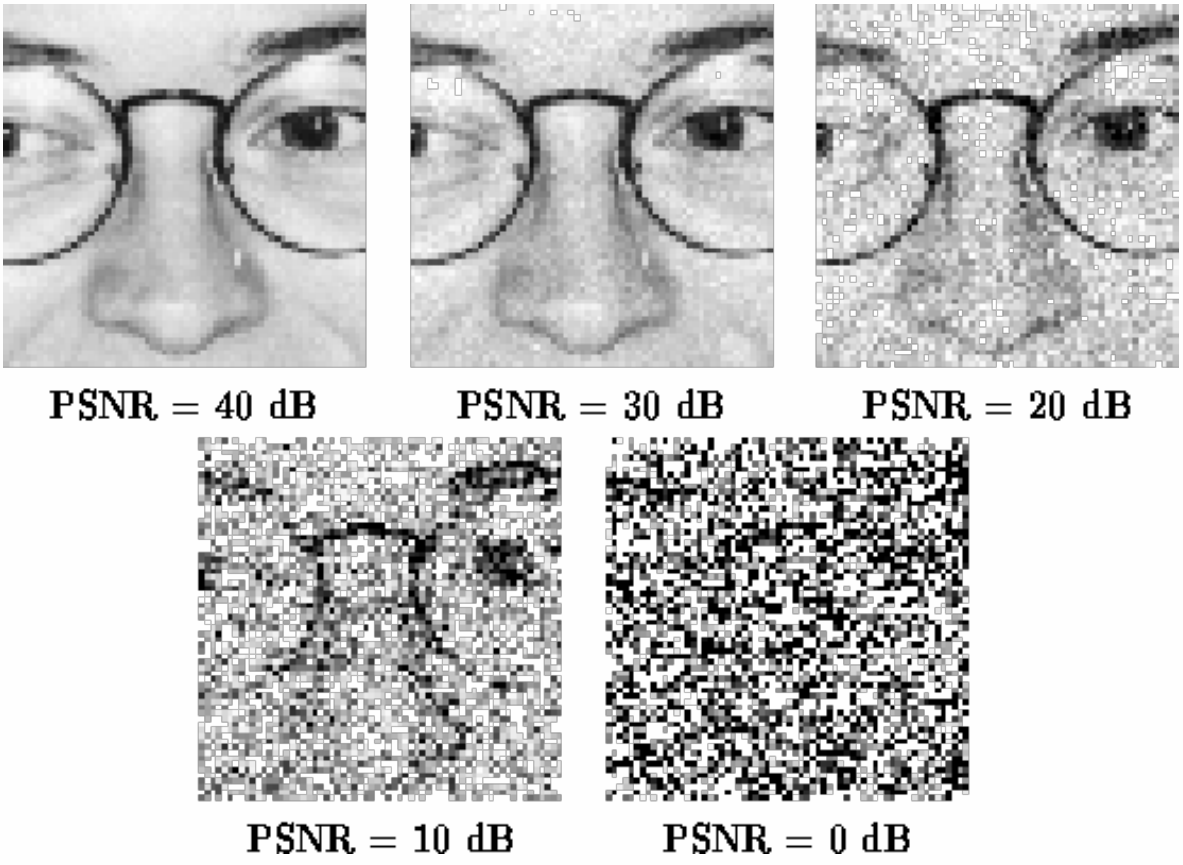
The problem: Pixel accuracy gets dominated by backgrounds. A lazy model ignoring people entirely still gets an A.

The fix: **IoU** penalizes both missing objects AND oversized boundaries. **Dice Coefficient** prioritizes foreground—crucial when brain tumors occupy 0.1% of an MRI.

This exposed a deeper issue: not all pixels matter equally. But what about image quality itself?

Image Quality: PSNR's 30-Year Lie

Engineers trusted **PSNR** for decades: higher = better. Then someone looked at the images.



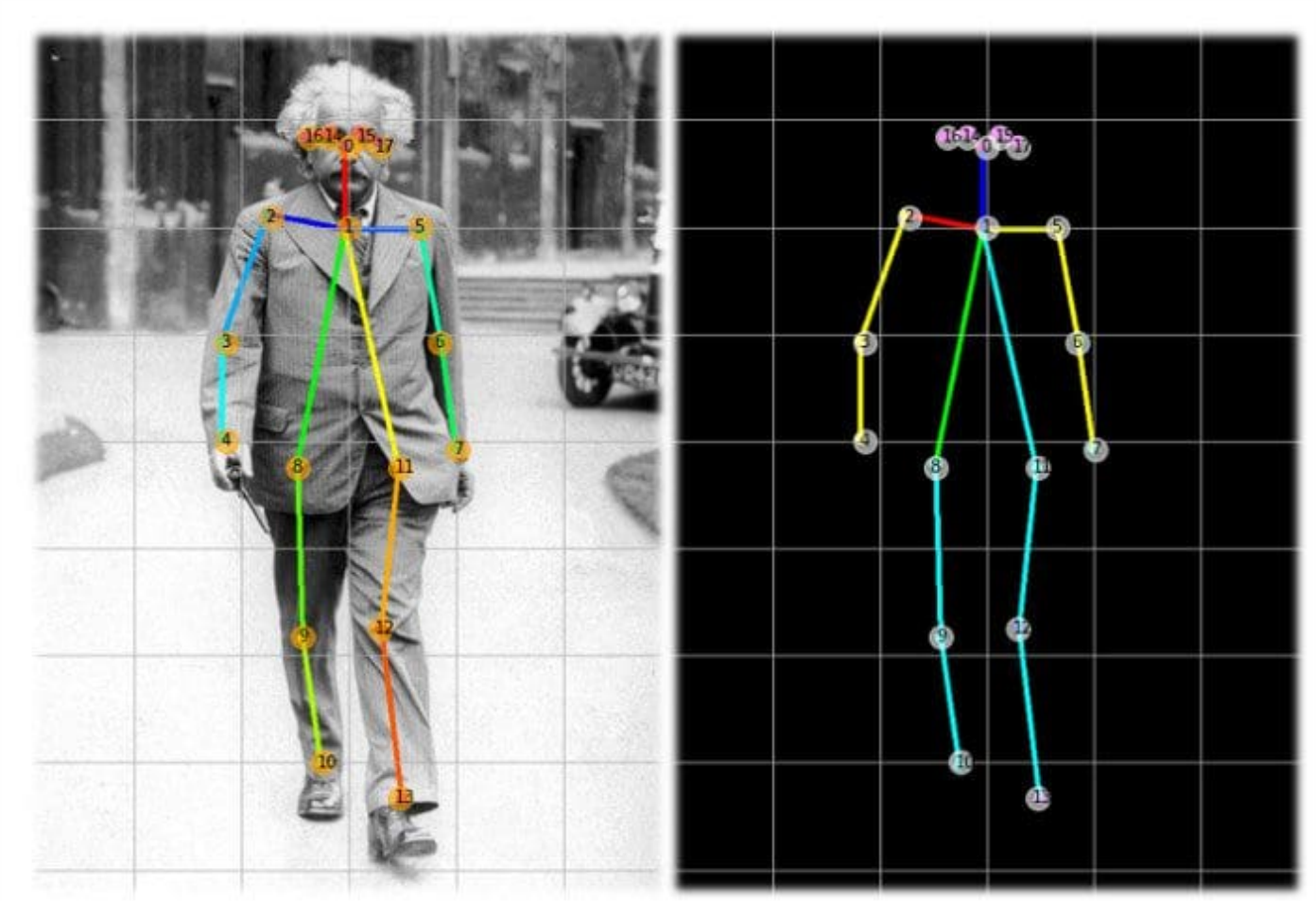
Two compressed photos: blurry (35dB) vs. sharp but noisy (32dB). PSNR picks blurry. Humans unanimously choose sharp.

The revolution: **SSIM** (2004) measures structural changes humans notice. **LPIPS** (2018) goes nuclear—compares images in neural feature space where two golden retrievers on grass share zero pixels but identical representations.

Quality metrics evolved to match human perception. Spatial tasks needed the same treatment.

Pose Estimation: The 5-Pixel Paradox

Is 5-pixel error good or bad? Person fills the frame? Terrible. Tiny distant figure? Perfect.

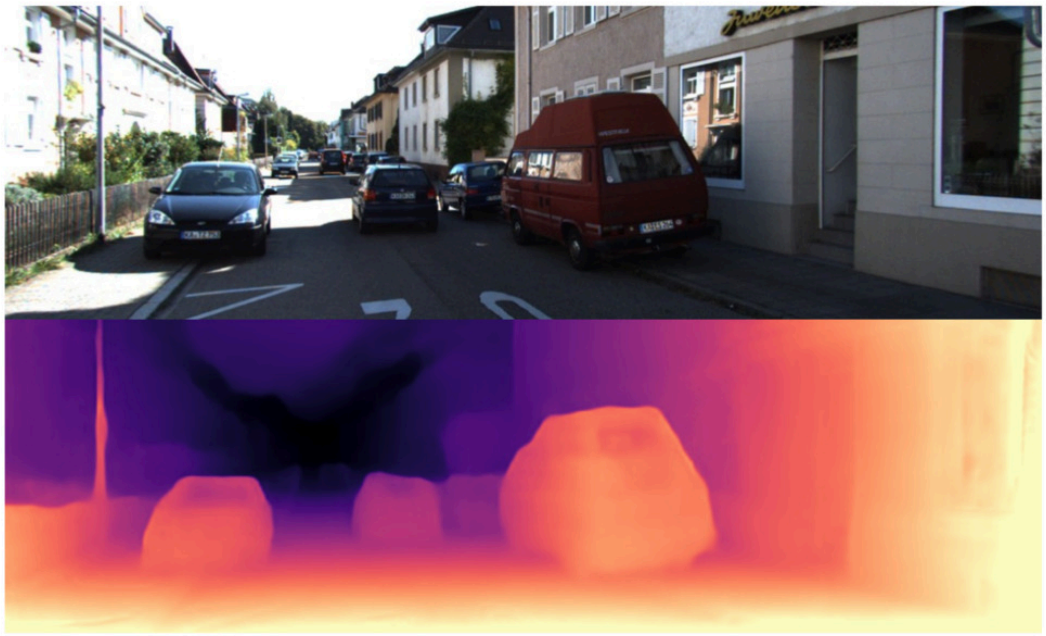


OKS normalized by object size and baked in per-keypoint tolerance—shoulders need precision, hips get slack (anatomically ambiguous). **PCK** simplified it: "Within 50% of torso? Yes/no."

Localizing 2D keypoints was solved, but understanding 3D space remained.

Depth Estimation: When 10cm ≠ 10cm

Wall: 9.9m vs 10.0m = 10cm error. Book: 0.9m vs 1.0m = 10cm error.



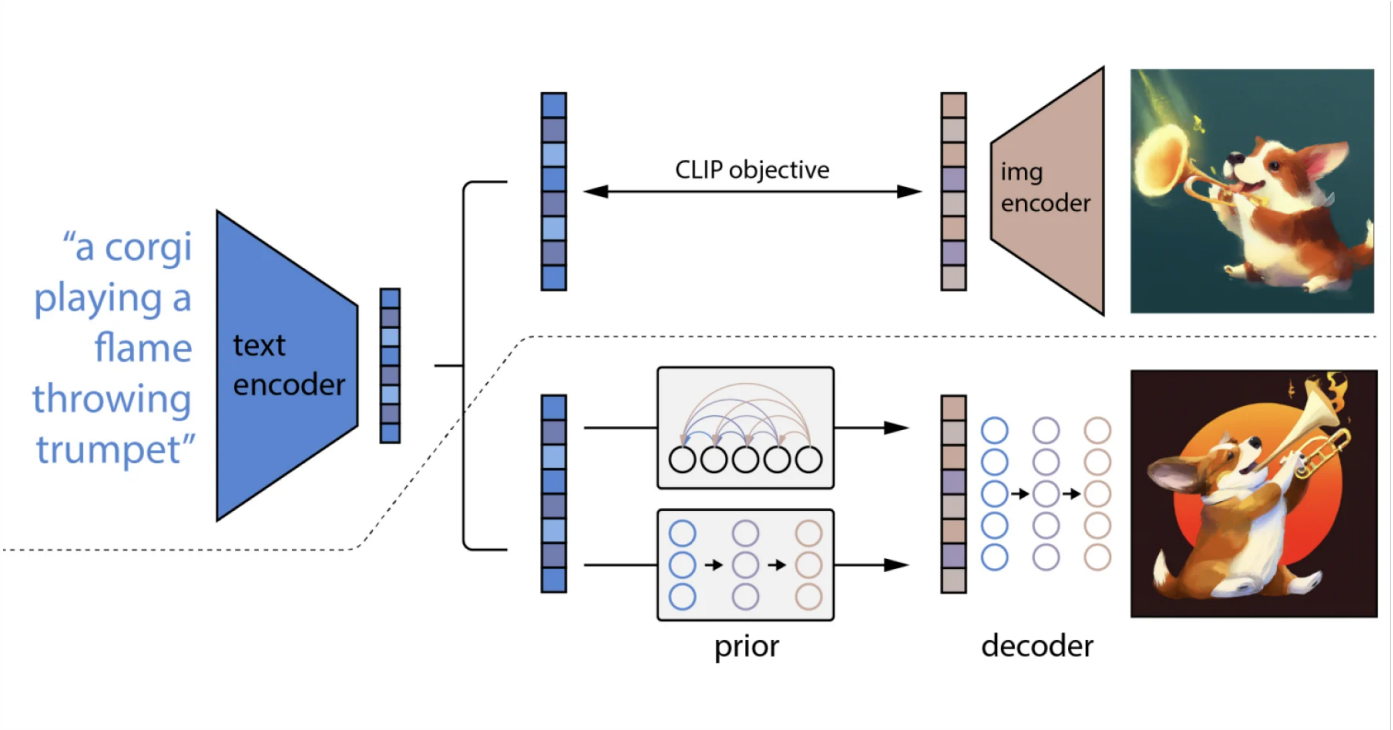
RMSE says equal. Your brain says one's perfect, one's wrong.

AbsRel measures proportions (1% vs 10%). **$\delta < 1.25$** asks: "Within 25%?" Self-driving cars don't need millimeters—just "roughly 10m" vs "roughly 3m."

These metrics worked for perceiving the world. Then generative AI demanded we measure creation itself.

Text-to-Image: Beautiful Chaos Returns

2021: DALL-E generates photorealistic images. Magic!
2022: "Wait, is the horse riding the astronaut?" Not so magic.



The crisis: FID measures realism, not instruction-following. Gorgeous random sunsets score perfectly while ignoring "robot playing chess."

CLIP Score (2021): Trained on 400M pairs, finally answers "Did you follow instructions?"

This exposed systematic failures:

- "Three cats" → five cats (can't count)
- "Box left of sphere" → backwards (spatial reasoning broken)
- "Horse riding astronaut" → reversed roles (compositional failure)

Current reality: No single metric captures "good generation." FID (realism) + CLIP Score (alignment) + compositional benchmarks (stress tests).

The evaluation crisis came full circle—messy, multi-metric, but finally honest about what we can and can't measure.

Mathematical Definition of Core Metrics

Intersection over Union (IoU)

IoU measures how well a predicted bounding box overlaps with the ground truth:

$$IoU = \frac{\text{Area}(B_p \cap B_{gt})}{\text{Area}(B_p \cup B_{gt})} \tag{1}$$

It's scale-invariant—a 5-pixel error on a small object hurts IoU more than the same error on a large object, matching human perception of accuracy.

Mean Average Precision (mAP)

mAP averages precision across all classes along the Precision-Recall curve:

- **Pascal VOC**: Uses mAP@50 (predictions with IoU ≥ 0.50 count as correct)
- **MS COCO**: Uses mAP@[.50:.05:.95] (averages AP across IoU thresholds from 0.50 to 0.95), rewarding precise localization

Fréchet Inception Distance (FID)

FID measures the similarity between real and generated image distributions using features from an Inception network:

$$FID = || \mu_r - \mu_g ||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \tag{2}$$

- **First term** ($|| \mu_r - \mu_g ||^2$): Measures if generated images are centered in the same feature space as real images
- **Second term** (Trace): Captures diversity—if the model has mode collapse, the covariance structures will differ, increasing FID

Lower FID indicates generated images are more realistic and diverse.

Dice Coefficient

The Dice Coefficient measures pixel-level overlap between predicted mask AA A and ground truth mask BB B:

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{3}$$

This metric is the harmonic mean of precision and recall at the pixel level. It ranges from 0 to 1, where 1 indicates perfect overlap. Dice is equivalent to the F1-score and relates to IoU as: $Dice = \frac{2 \times IoU}{1 + IoU}$. It's particularly popular in medical image segmentation due to its sensitivity to small structures.

Mean Intersection over Union (mIoU)

mIoU extends IoU to multi-class segmentation by computing IoU for each class independently and averaging:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{4}$$

where C is the number of classes, TP_c is true positive pixels, FP_c is false positives, and FN_c is false negatives for class c. Unlike Pixel Accuracy, mIoU treats all classes equally, preventing large classes (like background) from dominating the metric. This makes it the standard benchmark metric for semantic segmentation tasks.

Peak Signal-to-Noise Ratio (PSNR)

PSNR measures pixel-level reconstruction quality between images:

$$PSNR = 10 \log_{10} \frac{MAX^2}{MSE} \tag{5}$$

where MAX is the maximum possible pixel value (255 for 8-bit images) and MSE is the mean squared error. Higher values indicate better quality. However, PSNR is purely pixel-based and doesn't account for perceptual factors—images with high PSNR can still appear blurry or unnatural to humans.

Structural Similarity Index (SSIM)

SSIM evaluates image similarity through three components: luminance, contrast, and structure. For local windows x and y:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

- μ_x, μ_y: Mean intensities (luminance comparison)
- σ_x, σ_y: Standard deviations (contrast comparison)
- σ_{xy}: Covariance (structure comparison)
- C_1, C_2: Stability constants

SSIM ranges from -1 to 1, with 1 indicating identical images. Unlike PSNR, SSIM correlates better with human perceptual judgment by considering structural information rather than raw pixel errors.

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS measures perceptual distance using deep features from pretrained networks (typically VGG):

$$LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (F_l^{(x)}(h,w) - F_l^{(y)}(h,w))||^2 \quad (7)$$

where F_l denotes features at layer l, and w_l are learned weights. By operating in learned feature space rather than pixel space, LPIPS captures perceptual similarity—images that humans perceive as similar will have low LPIPS distance even if pixel values differ significantly. Lower values indicate greater similarity.

Object Keypoint Similarity (OKS)

OKS adapts the IoU concept for keypoint detection:

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (8)$$

- d_i: Euclidean distance between predicted and ground truth keypoint
- s: Object scale (square root of object area)
- k_i: Per-keypoint constant controlling falloff (reflects annotation variability)
- $delta(v_i > 0)$: Indicator for visible keypoints onl

The exponential decay means nearby predictions score higher, with the scale normalization ensuring fairness across object sizes. OKS ranges from 0 to 1, where 1 indicates perfect alignment. This is the standard metric for COCO keypoint evaluation.

Percentage of Correct Keypoints (PCK)

PCK measures the fraction of keypoints predicted within a threshold distance from ground truth:

$$PCK@\alpha = \frac{\#\text{ keypoints with } d_i < \alpha \cdot d_{ref}}{\text{Total keypoints}} \tag{9}$$

where α is the threshold (commonly 0.5 or 0.2) and d_{ref} is a reference distance, typically torso diameter or head size. PCK@0.5 means the predicted keypoint must be within 0.5× the reference distance to count as correct. This binary metric is simpler than OKS but less nuanced.

Absolute Relative Error (AbsRel)

AbsRel normalizes depth error by ground truth depth:

$$AbsRel = \frac{1}{N}\sum_{i}\frac{|d_i - d_i^*|}{d_i^*} \tag{10}$$

where d_i is predicted depth and d_i^* is ground truth. This relative error ensures that a 10cm error at 1m distance is penalized more heavily than a 10cm error at 10m distance, aligning with human perception that relative accuracy matters more than absolute accuracy in depth estimation.

Threshold Accuracy (δ < 1.25)

This metric measures the percentage of pixels where the depth prediction is within a multiplicative factor:

$$\delta_t = \%\text{ of pixels where } \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) < t \tag{11}$$

Common thresholds are $t \in \{1.25, 1.25^2, 1.25^3\}$. A prediction passes if it's within 25% (for t=1.25) of the ground truth in either direction. Higher values indicate more accurate depth predictions, with typical benchmarks reporting all three thresholds.

CLIP Score

CLIP Score leverages the CLIP model's joint vision-language embedding to measure alignment:

$$\text{CLIP-S} = \cos(\text{CLIP}_{\text{image}}(I), \text{CLIP}_{\text{text}}(T)) \tag{12}$$

where the cosine similarity measures how well generated image I matches text prompt T in CLIP's learned embedding space. Higher scores indicate better semantic alignment. Unlike FID (which measures realism), CLIP Score specifically evaluates whether the image content matches the text description.

Inception Score (IS)

IS evaluates both image quality and diversity using a classifier's predictions:

$$IS = \exp\left(\mathbb{E}_x[KL(p(y|x) \parallel p(y))]\right) \tag{13}$$

- $p(y|x)$: Conditional class distribution for image x (should be peaked for clear images)
- $p(y)$: Marginal class distribution (should be uniform for diverse generation)
- KL: Kullback-Leibler divergence

High IS requires both confident predictions (low conditional entropy) and diverse outputs (high marginal entropy). However, IS has limitations—it can't detect mode collapse if the model generates diverse but all high-quality images from a subset of classes. FID has largely superseded IS for realism evaluation.

Reference

https://en.wikipedia.org/wiki/MNIST_database

https://cv.gluon.ai/build/examples_datasets/imagenet.html

https://viso.ai/deep-learning/pascal-voc-dataset/

https://www.v7labs.com/blog/coco-dataset-guide

https://generative-vision.github.io/workshop-CVPR-23/

https://encord.com/blog/image-segmentation-for-computer-vision-best-practice-guide/

https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/VELDHUIZEN/node18.html

https://mobidev.biz/blog/human-pose-estimation-technology-guide

https://github.com/topics/depth-estimation

https://www.edge-ai-vision.com/2023/01/from-dall%C2%B7e-to-stable-diffusion-how-do-text-to-image-generation-models-work/