Localization and Object Detection

A fixed dense tensor becomes a scored set of boxes

Prof. Nipun Batra

ES 667 · Deep Learning · IIT Gandhinagar

L8B kept spatial features; detection must spend them

Two real images turn “where” into supervised geometry

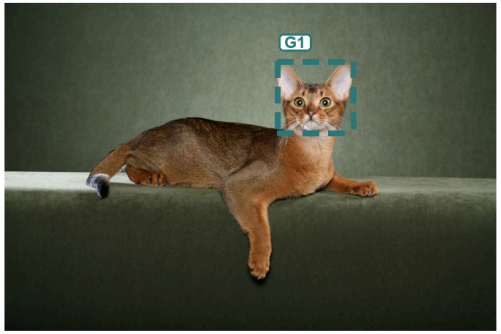
V VISUAL

A REAL TWO-IMAGE MINI-BATCH

Two photographs · two official XML head boxes

I1 · Abyssinian_1

OBSERVED PHOTO + XML GT



VOC XML (333, 72)–(425, 158)

course xyxy [332, 71, 425, 158]

93 × 87 source pixels · ground truth, not model output

I2 · Bengal_10

OBSERVED PHOTO + XML GT

VOC XML (315, 60)–(446, 219)

course xyxy [314, 59, 446, 219]

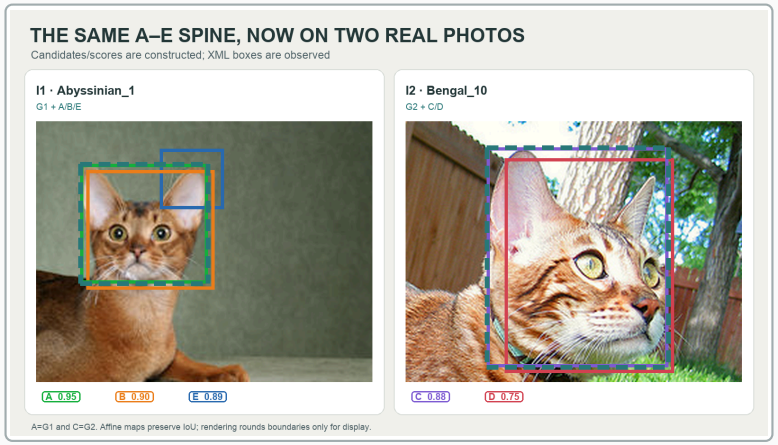
132 × 160 source pixels · ground truth, not model output

> OBSERVED pixels + official XML head boxes. Detection reuses the L8B backbone, but changes the head, targets, geometry, and evaluation.

Commit: which operations turn five candidates into trustworthy detections?

Q QUESTION

HELD CASE · OPENING MINI-BATCH CONSTRUCTED, not model output: A/B/E belong to I1; C/D belong to I2.



A · THRESHOLD

score threshold alone

B · GEOMETRY

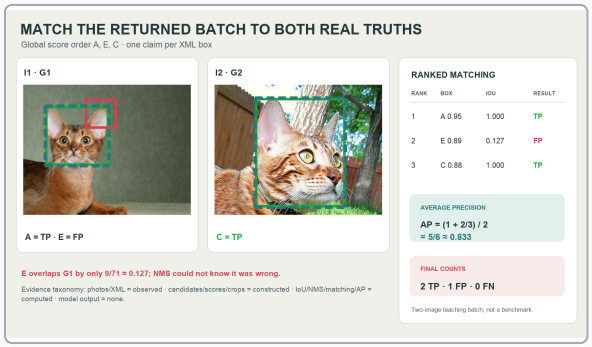
decode → score → NMS → match

C · LOSS

rerun loss at inference

Outline

V VISUAL



OBSERVED pixels + GT · CONSTRUCTED candidates · COMPUTED matches

01 · BOX CONTRACT

Choose coordinates, decode rules, and class scores.

02 · OVERLAP

Use IoU to compare boxes under an explicit convention.

03 · TRAIN

Assign candidates; combine classification and localization losses.

04 · INFER

Decode, threshold, and suppress duplicate predictions.

05 · EVALUATE

Match once, rank by score, and accumulate AP.

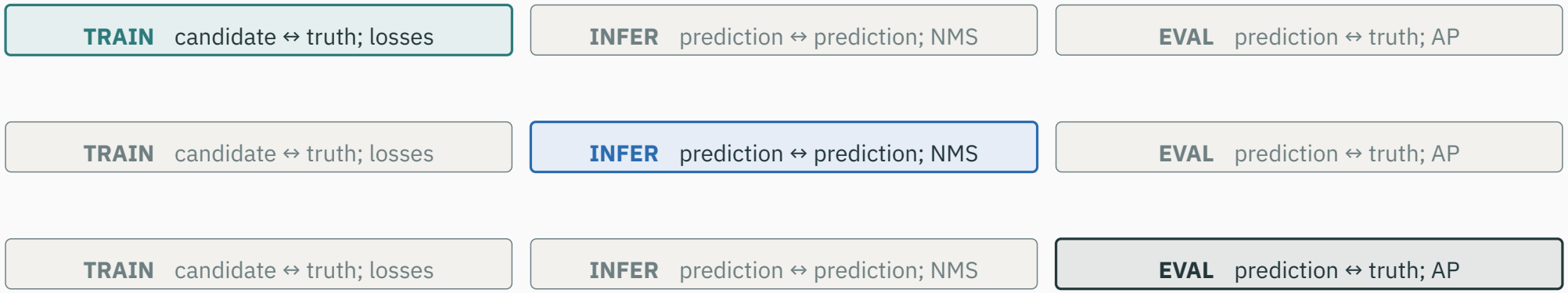
06 · ARCHITECTURES

Relate two-stage, one-stage, FPN, and anchor-free heads.

given / eval · truth / train · raw candidate · infer / control · survivor / TP · suppressed / FP

Three clocks reuse boxes—but never the same comparison

V VISUAL



TRAIN

assign candidate ↔ truth

INFER

compare prediction ↔ prediction

EVAL

match prediction ↔ truth

The clock strip will remain visible whenever “IoU threshold” could otherwise mean three different things.

A box is four numbers with an explicit convention

Detection predicts a set, not a longer class vector

V VISUAL

classifier: $x \to \hat{p} \in \Delta^{K-1}$

detector: $x \to \left\{\left(\hat{y}_i, \hat{b}_i, s_i\right)\right\}_{i=1}^{N}$

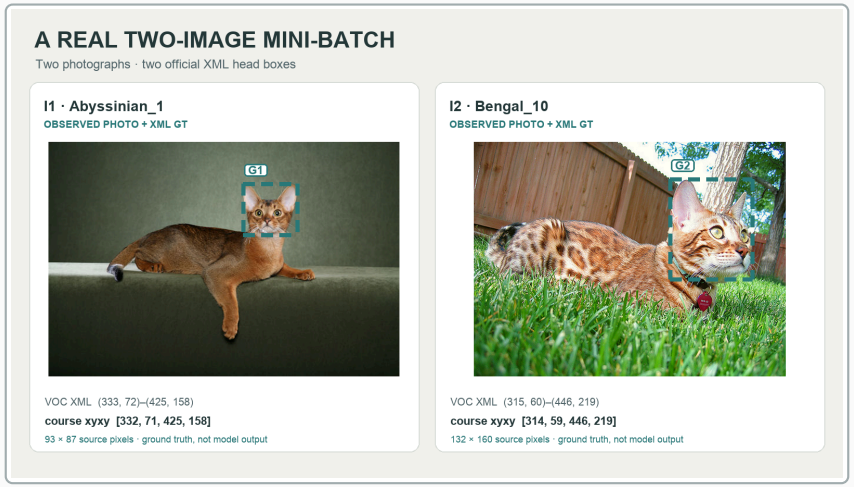
CLASS	BOX	SCORE
$\hat{y}_i$: what is here?	$\hat{b}_i \in \mathbb{R}^4$: where is it?	s_i: how confident?

A fixed network tensor may contain many candidate triples; thresholding and NMS make the returned set variable-length.

Our coordinate convention prevents the silent “+1” bug

D DERIVATION

HELD CASE · REAL ANNOTATION CONTRACT VOC XML uses one-based inclusive corners; the course converts once to zero-based half-open $[x_0, y_0, x_1, y_1)$.



Example I1: VOC $(333, 72)$–$(425, 158)$ becomes $[332, 71, 425, 158)$, so area is $93 \cdot 87 = 8,091$. No later component adds $+1$.

The canonical truths fix every later area calculation

D DERIVATION

HELD CASE · CONSTRUCTED NUMERIC SPINE $G_1 = [10, 10, 50, 50]$ in I1; $G_2 = [60, 15, 90, 55]$ in I2. Separate positive affine maps preserve IoU.

G_1

$w = 50 - 10 = 40$

$h = 50 - 10 = 40$

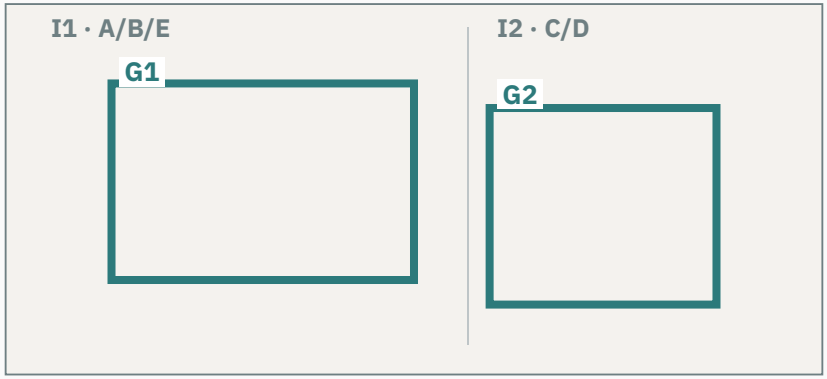
$|G_1| = 1,600$

G_2

$w = 90 - 60 = 30$

$h = 55 - 15 = 40$

$|G_2| = 1,200$



A localization head shares evidence, then asks two questions

V VISUAL

SHARED LOCAL FEATURE

$$h = f_{\theta(x)}$$

→

CLASS HEAD

$$z = W_c h + b_c$$

BOX HEAD

$$t = W_b h + b_b \in \mathbb{R}^4$$

what is here?

how should the reference move?

> Detection is multi-task learning repeated at spatial locations—not a classifier with four unexplained extra outputs.

Commit: what offsets move this anchor to its matched truth?

Q QUESTION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

HELD CASE · POSITIVE ANCHOR Reference $a = (50, 50, 40, 40)$; matched target $b = (58, 46, 80, 40)$, both in center-size form.

TRANSLATE

$$t_x = \frac{x - x_a}{w_a}$$
$$t_y = \frac{y - y_a}{h_a}$$

RESIZE

$$t_w = \log\left(\frac{w}{w_a}\right)$$
$$t_h = \log\left(\frac{h}{h_a}\right)$$

Calculate all four before the answer. Which offset must be zero? Which must be $\log 2$?

The anchor target is a scaled correction, not an absolute box

A ANSWER

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

$$t_x = \frac{58-50}{40} = 0.20 \qquad t_w = \log\left(\frac{80}{40}\right) = \log 2 \approx 0.693$$
$$t_y = \frac{46-50}{40} = -0.10 \qquad t_h = \log\left(\frac{40}{40}\right) = 0$$

$$t = (0.20, -0.10, 0.693, 0)$$

Translations are measured in anchor sizes; log-scale changes keep decoded widths positive and make zero mean “unchanged.”

Decoding must recover the same geometry

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

$$\hat{x} = x_a + t_x w_a = 50 + 0.20(40) = 58$$
$$\hat{y} = y_a + t_y h_a = 50 - 0.10(40) = 46$$

$$\hat{w} = w_a \exp(t_w) = 40 \exp(0.693) \approx 80$$
$$\hat{h} = h_a \exp(t_h) = 40 \exp(0) = 40$$

Encode and decode are inverse contracts; a unit test should recover b before any model is trained.

A YOLO-style anchor head emits nine scalars per candidate

V VISUAL

TRAIN candidate ↔ truth; losses

INFER prediction ↔ prediction; NMS

EVAL prediction ↔ truth; AP

HELD CASE · HEAD CONVENTION For this teaching ledger only: one objectness logit + $K = 4$ class logits + four box offsets.

OBJECTNESS	CLASSES	OFFSETS
one logit is anything here?	$K = 4$ logits which class?	four numbers how should the anchor move?

$$1 + K + 4 = 1 + 4 + 4 = 9$$

Teaching convention only: other heads may fold background into classes or predict box sides.

The dense tensor fixes capacity before the object count is known

D DERIVATION

TRAIN candidate ↔ truth; losses

INFER prediction ↔ prediction; NMS

EVAL prediction ↔ truth; AP

HELD CASE · TENSOR LEDGER 20×20 locations, $A = 3$ anchors per location, $K = 4$ classes, nine outputs per anchor.

quantity	calculation	count
candidate boxes	$20 \cdot 20 \cdot 3$	$1,200$
output scalars	$20 \cdot 20 \cdot 3 \cdot 9$	$10,800$

The head always emits $10,800$ scalars. Labels and post-processing decide which candidate records matter.

Toy assignment says which candidates receive which losses

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

For one truth, three anchors have IoUs 0.76, 0.48, and 0.12. Use the declared **toy** policy $\tau_+ = 0.70$, $\tau_- = 0.30$.

IoU	assignment	loss terms
0.76	positive	objectness + class + box
0.48	ignore	no gradient in this toy policy
0.12	background	objectness only

> Assignment is a training rule, not AP matching and not NMS. Real detectors use architecture-specific thresholds or dynamic matchers.

One positive candidate produces a fully auditable loss

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

Target $t = (.20, -.10, .693, 0)$; prediction $\hat{t} = (.25, -.05, .60, .05)$. Let $p_{\text{obj}} = .90$, $p_{\text{cup}} = .80$ and $\lambda = 2$.

OBJECTNESS	CLASS	BOX L_1
$-\log .90 \approx .105$	$-\log .80 \approx .223$	$\lvert.05\rvert + \lvert.05\rvert + \lvert-.093\rvert + \lvert.05\rvert = .243$

$$\mathcal{L} = .105 + .223 + 2(.243) \approx .815$$

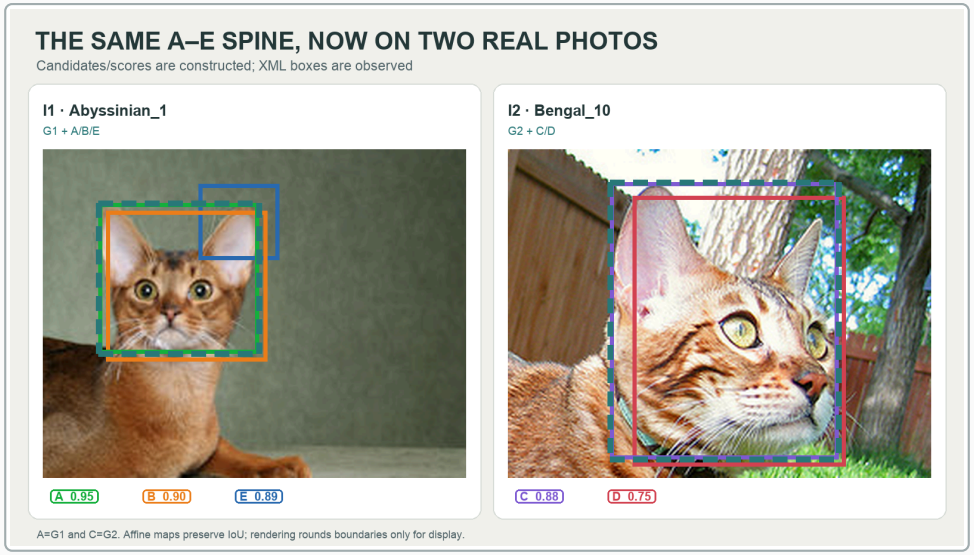
Write the target, parameterization, reduction, and λ. "Box loss" without those contracts is not reproducible.

IoU makes box quality geometric

IoU asks how much two regions agree

V VISUAL

HELD CASE · SAME I1 PAIR $A = G_1 = [10, 10, 50, 50]$ and $B = [12, 12, 52, 52]$ are affinely placed on the real Abyssinian head ROI.



$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \in [0, 1]$$

0: disjoint 1: identical invariant to scaling the whole scene

The held overlap is 0.822, not a visual guess

D DERIVATION

HELD CASE · SAME PAIR $A = [10, 10, 50, 50]$, $B = [12, 12, 52, 52]$; both have area $1,600$.

$$w_I = \min(50, 52) - \max(10, 12) = 38 \qquad A_I = 38 \cdot 38 = 1,444$$

$$h_I = \min(50, 52) - \max(10, 12) = 38 \qquad A_U = 1,600 + 1,600 - 1,444 = 1,756$$

$$\mathrm{IoU}(A, B) = \frac{1444}{1756} \approx 0.822$$

The same 0.822 will later justify suppressing B around kept box A.

Safe IoU clamps disjoint intersections to zero

D DERIVATION

For each axis, $w_I = \max\big(0, \min\big(x_1^A, x_1^B\big) - \max\big(x_0^A, x_0^B\big)\big)$ $h_I = \max\big(0, \min\big(y_1^A, y_1^B\big) - \max\big(y_0^A, y_0^B\big)\big)$

Then, $A_I = w_I h_I$ $A_U = A_A + A_B - A_I$ $\mathrm{IoU} = \frac{A_I}{A_U}$

Without $\max(0, \cdot)$, two disjoint boxes can produce two negative side lengths—and a meaningless positive "intersection."

Also assert valid boxes: $x_1 \geq x_0$, $y_1 \geq y_0$, and define the zero-union case explicitly.

Equal coordinate error can mean unequal overlap

Q QUESTION

Two predictions move one boundary by 10 pixels.

Large object: 200×200
a 10-pixel error is 5% of its width.

Small object: 20×20
the same error is 50% of its width.

Coordinate losses see four residuals. IoU sees the region those residuals create.

Use coordinate losses for a stable regression signal; report overlap-aware metrics for the geometry the task actually values.

Smooth L_1 and GIoU repair different failures

D DERIVATION

SMOOTH $L_1 \cdot \beta = 1$

quadratic near 0; linear for large residuals

$0.2 \to 0.02$; $3 \to 2.5$

GIOU

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|C \setminus (A \cup B)|}{|C|}$$

C: smallest enclosing box

> Plain $1 - \mathrm{IoU}$ is flat while boxes are disjoint. GIoU changes with the enclosing gap; Smooth L_1 instead controls coordinate-residual robustness.

Rezatofighi et al. · Generalized IoU

TRAIN: dense candidates need selective supervision

Dense heads create a foreground–background imbalance

V VISUAL

TRAIN candidate ↔ truth; losses

INFER prediction ↔ prediction; NMS

EVAL prediction ↔ truth; AP

If thousands of easy background candidates contribute ordinary cross-entropy, their count can drown out the rare positives and hard negatives.

Focal loss turns confidence into a gradient-volume control

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

$$\mathrm{FL}(p_t) = -\alpha_{t(1-p_t)}^{\gamma} \log p_t$$

p_t	γ	α_t
probability assigned to the true binary label	how strongly easy examples are down-weighted	optional class-balance weight

> The next numeric sets $\alpha_t = 1$ and $\gamma = 2$ to isolate the focusing factor.

With $\gamma = 2$, the hard example is about $978\times$ louder

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

Easy: $p_t = .9$

$\mathrm{CE} = -\log .9 \approx .10536$

$(1 - p_t)^2 = .01$

$\mathrm{FL} \approx .0010536$

Hard: $p_t = .2$

$\mathrm{CE} = -\log .2 \approx 1.60944$

$(1 - p_t)^2 = .64$

$\mathrm{FL} \approx 1.03004$

$$\frac{1.03004}{.0010536} \approx 977.6 \approx 978\times$$

Focal loss does not delete easy examples; it makes already-correct examples quiet enough that rare mistakes remain audible.

Diagnose training failures as symptom → suspect → test

V VISUAL

TRAIN candidate ↔ truth; losses

INFER prediction ↔ prediction; NMS

EVAL prediction ↔ truth; AP

symptom	suspect	smallest discriminating test
box loss falls; IoU stays near 0	box format / decode mismatch	round-trip one known anchor and draw decoded boxes
almost no positive anchors	assignment geometry too strict	histogram best IoU per truth; inspect τ_+
objectness predicts background	easy negatives dominate	log pos/neg loss mass; compare focal or sampling

Inspect geometry and assignment before changing the backbone: a model cannot learn targets that the pipeline encoded incorrectly.

Follow along: boxes, IoU, matching, and anchor round-trips

I INTERACTIVE

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

INTERACTIVE

↗ colab.research.google.com/github/nipunbatra/dl-teaching/blob/master/notebooks/L09/01_iou_and_bbox_losses.ipynb

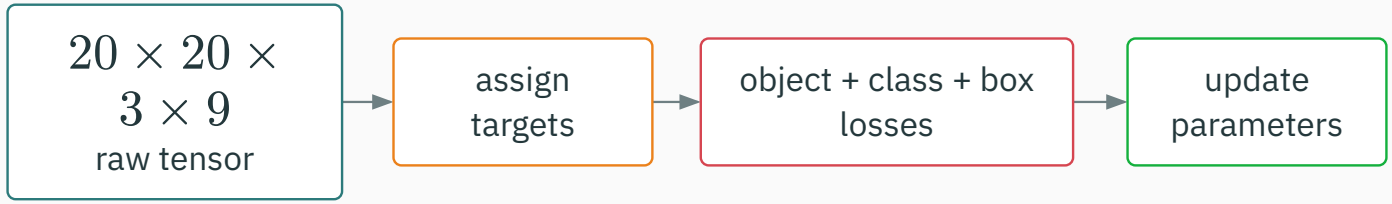
Open the exact Colab. Predict the $\frac{1444}{1756}$ IoU, implement the clamp, verify the anchor encode/decode round-trip, and build the held matching ledger before reading its outputs.

BOX	OVERLAP	ANCHOR
$[x_0, y_0, x_1, y_1]$; no $+1$	$\mathrm{IoU}(A, B) \approx .822$	$t = (.20, -.10, .693, 0)$

TRAIN closes with targets; inference opens without them

V VISUAL

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

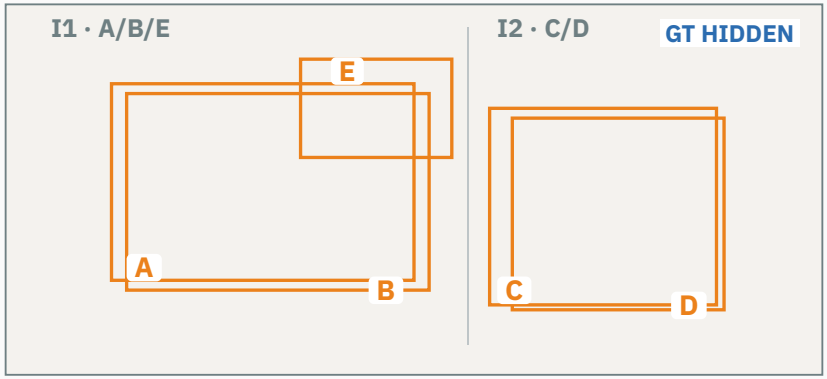
At inference there is no truth, no assignment, and no loss—only decoded boxes, scores, thresholds, and prediction-to-prediction suppression.

INFER: score first, then suppress duplicates

The constructed raw ledger isolates the inference rule

V VISUAL

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



A .95 B .90 E .89 C .88 D .75

A/B/E share (I1, head); C/D share (I2, head). These scores are pedagogical constructions, not model output. INFER sees neither XML truth nor TP/FP labels.

Score thresholding removes weak evidence—not duplicates

Q QUESTION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

Use a deliberately permissive score threshold $s \geq .70$.

A $.95$	B $.90$	E $.89$	C $.88$	D $.75$

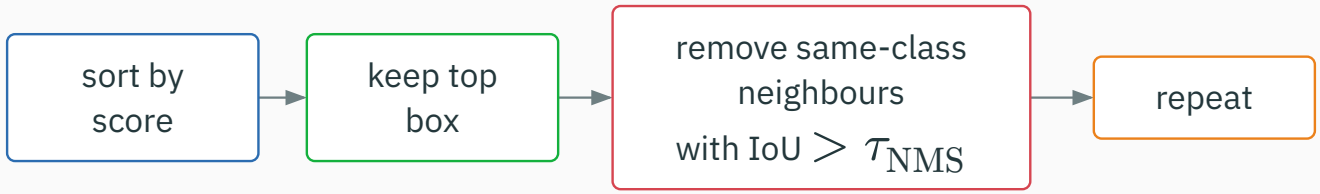
All five survive.

Thresholding can drop low scores, but it cannot tell that two high-scoring boxes describe the same object.

NMS is a greedy prediction-to-prediction comparison

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

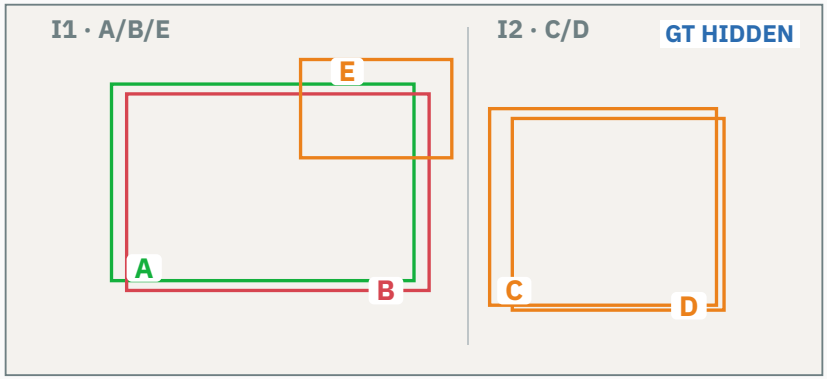


No ground truth participates. Standard NMS has no learned parameter and no gradient; here $\tau_{\text{NMS}} = .50$.

Step 1: A claims its high-overlap neighbourhood

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



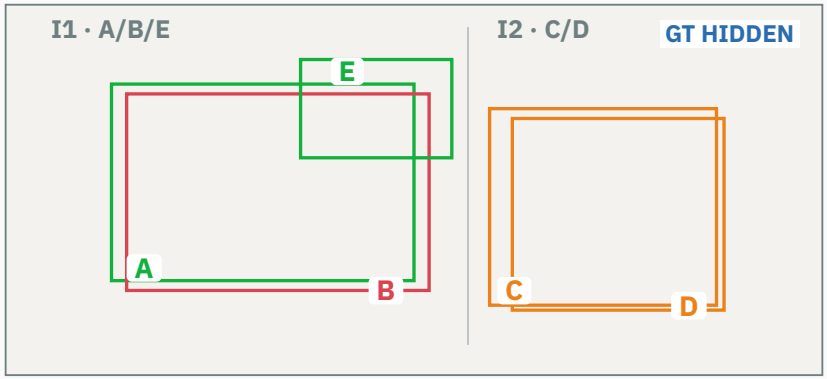
$$\mathrm{IoU}(A, B) = .822 > .50 \rightarrow \text{keep A, suppress B}$$

E, C, and D remain because none overlaps A above the NMS threshold.

Step 2: E survives because NMS cannot test correctness

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



$E = .89$ is now the highest remaining score; its IoU with every survivor is below $.50$.

Keep E. NMS removes duplicates around a stronger prediction; it does not remove isolated false alarms.

Step 3: C claims the I2 head neighbourhood

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | EVAL prediction ↔ truth; AP

HELD CASE · LAST PAIR $C = [60, 15, 90, 55]$ and $D = [63, 17, 91, 56]$.

$$A_I = (90 - 63)(55 - 17) = 27 \cdot 38 = 1,026 \; A_U = 1,200 + 1,092 - 1,026 = 1,266$$

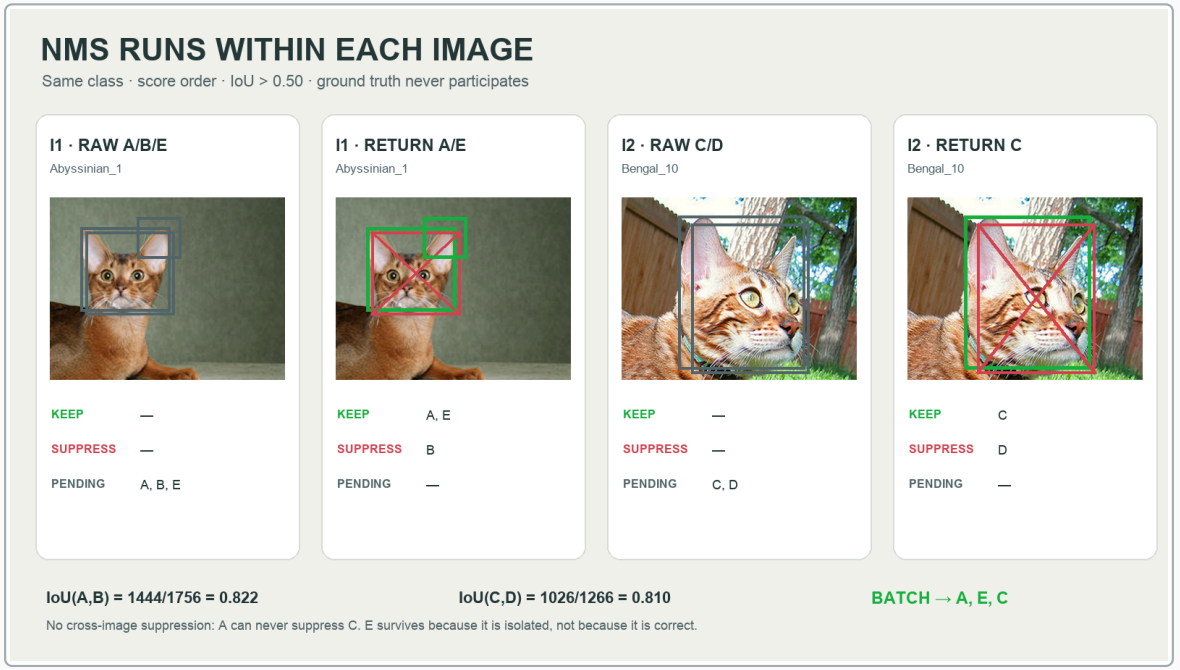
$$\mathrm{IoU}(C, D) = \frac{1026}{1266} \approx .810 > .50$$

Keep C; suppress D. No candidates remain.

NMS returns A, E, C—not “the correct boxes”

A ANSWER

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

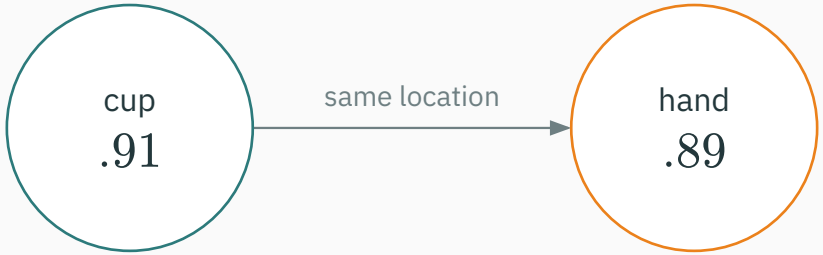


NMS is grouped by $(\text{image\_id}, \text{class})$: A can never suppress C across images. The global returned score order is A, E, C; correctness waits for EVAL.

Class-aware NMS prevents a cup from deleting a hand

Q QUESTION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



Usually compare boxes within the same class. Cross-class overlap may describe two real, interacting objects.

A low NMS threshold can still merge two nearby objects of the **same** class. Increasing τ_{NMS} often keeps more boxes for a fixed score order, but keep-count is not a global monotonic law across changed scores/classes/pipelines.

Follow along: print every NMS decision, then break it

I INTERACTIVE

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

INTERACTIVE

↗ colab.research.google.com/github/nipunbatra/dl-teaching/blob/master/notebooks/L09/02_nms_from_scratch.ipynb

Open the exact Colab. Predict A, E, C; print the keep/suppress trace; verify class-aware NMS; then construct two distinct same-class objects that a low threshold wrongly merges.

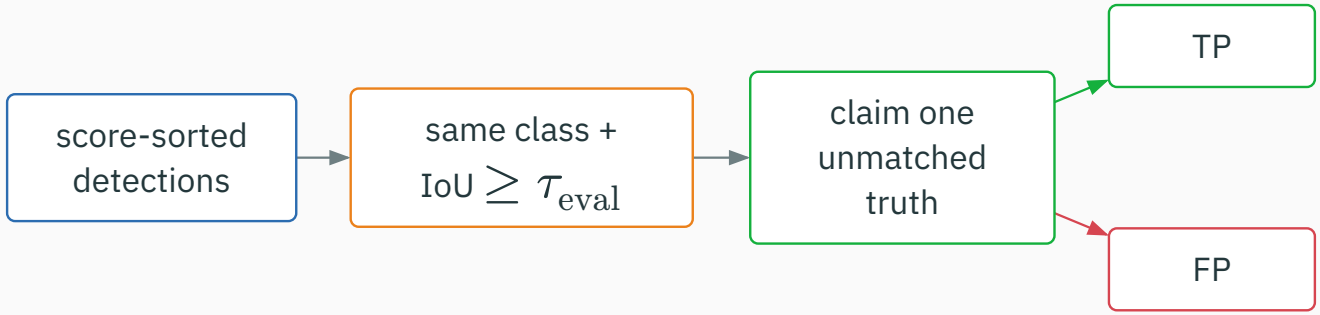
KEEP	SUPPRESS	LIMIT
A .95, E .89, C .88	B by A; D by C	isolated false alarm E survives

EVAL: match the returned set to truth, one claim at a time

Evaluation matching is not training assignment

V VISUAL

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

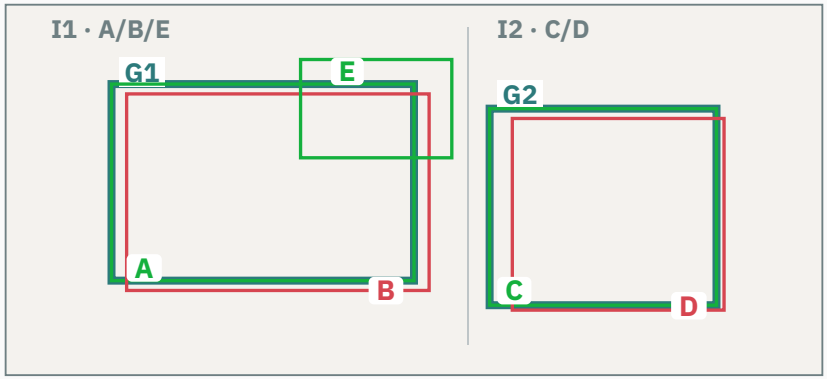


At $\tau_{\text{eval}} = .50$, each truth may be claimed once. A duplicate would be an FP even if its IoU were high.

Commit: which of A, E, C are true positives?

Q QUESTION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



detection	score	best same-class IoU	commit
A	.95	1.000 with G_1	?
E	.89	.127 with G_1	?
C	.88	1.000 with G_2	?

Match only within image_id, then concatenate survivors into global score order. A truth claimed by an earlier detection is unavailable to later detections in that image.

Evaluation exposes the false alarm NMS could not see

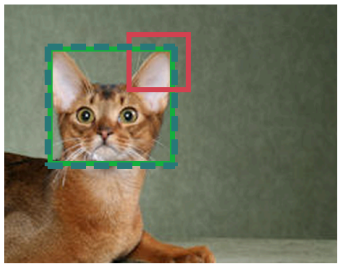
A ANSWER

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

MATCH THE RETURNED BATCH TO BOTH REAL TRUTHS

Global score order A, E, C · one claim per XML box

I1 · G1



A = TP · E = FP

I2 · G2

C = TP

RANKED MATCHING

RANK	BOX	IOU	RESULT
1	A 0.95	1.000	TP
2	E 0.89	0.127	FP
3	C 0.88	1.000	TP

AVERAGE PRECISION

AP = (1 + 2/3) / 2
= 5/6 ≈ 0.833

FINAL COUNTS

2 TP · 1 FP · 0 FN

Two-image teaching batch; not a benchmark.

E overlaps G1 by only 9/71 ≈ 0.127; NMS could not know it was wrong.

Evidence taxonomy: photos/XML = observed · candidates/scores/crops = constructed · IoU/NMS/matching/AP = computed · model output = none.

GT is revealed only now: A claims I1/G1, E is an I1 false alarm at $\frac{9}{71}$, and C claims I2/G2. Two TPs, one FP, zero FNs.

Precision and recall evolve with the score ranking

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

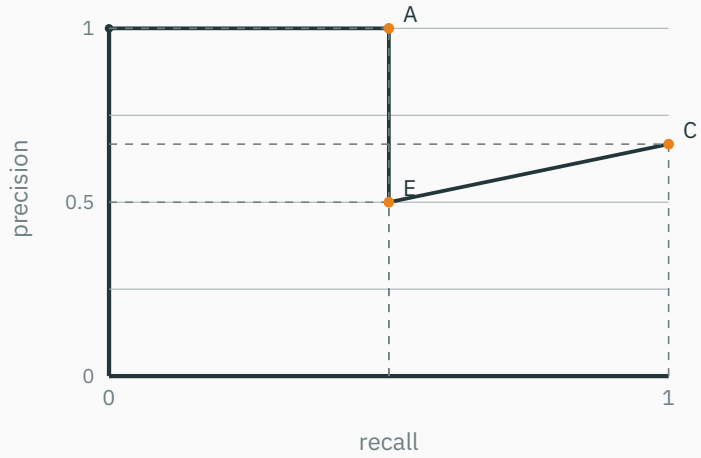
after	cum. TP	cum. FP	precision	recall
A	1	0	$\frac{1}{1} = 1.000$	$\frac{1}{2} = .500$
E	1	1	$\frac{1}{2} = .500$	$\frac{1}{2} = .500$
C	2	1	$\frac{2}{3} = .667$	$\frac{2}{2} = 1.000$

Lowering the score threshold admitted E before C: precision fell, then recall rose when C arrived.

The toy AP convention averages precision at TP ranks

D DERIVATION

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP



At the two TP ranks:

$$p_A = 1$$
$$p_C = \frac{2}{3}$$

$$\mathrm{AP}_{\mathrm{toy}} = \frac{1+\frac{2}{3}}{2} = \frac{5}{6} \approx .833$$

After NMS: A, E, C; B and D were suppressed before evaluation. Axes run from 0 to 1.

This TP-rank average is a transparent teaching convention. It is not a claim about every benchmark's interpolation rule.

COCO AP fixes the full evaluation protocol

D DERIVATION

TRAIN candidate ↔ truth; losses

INFER prediction ↔ prediction; NMS

EVAL prediction ↔ truth; AP

LOCALIZATION	INTERPOLATION	AGGREGATION
10 IoU thresholds $.50, .55, \ldots, .95$	101 recall thresholds $0, .01, \ldots, 1$	average over classes; declare area + max detections

Say AP50 for one IoU cutoff. “COCO AP” normally means the average across IoU thresholds; the official API also declares area ranges and maxDets.

Official COCO evaluation implementation

Diagnose metric surprises on the correct clock

V VISUAL

TRAIN candidate ↔ truth; losses | **INFER** prediction ↔ prediction; NMS | **EVAL** prediction ↔ truth; AP

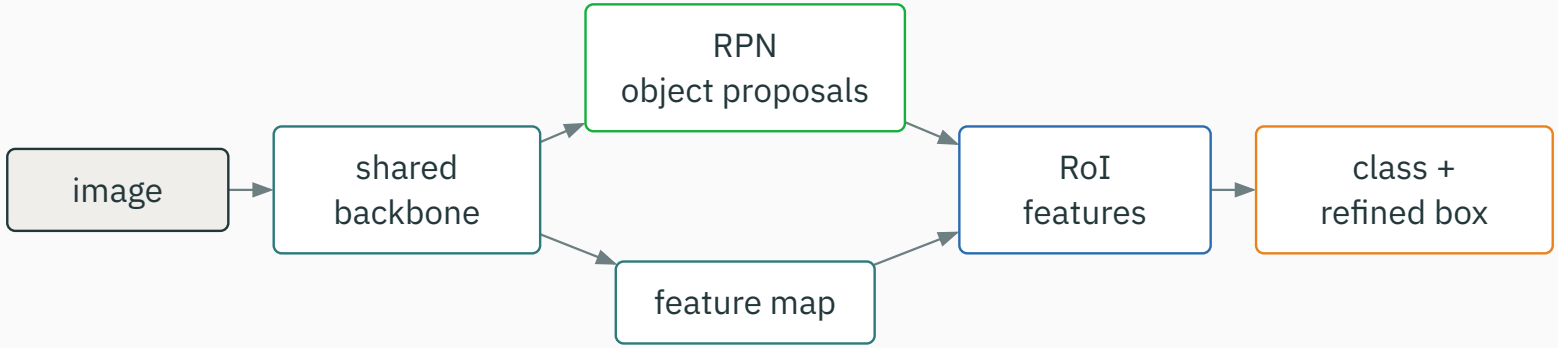
symptom	suspect	smallest discriminating test
high scores; low AP	ranking is confidently wrong	inspect first FP for class, IoU, duplicate status
AP50 high; COCO AP low	boxes are loose	compare AP50 vs AP75; histogram matched IoU
recall caps below 1	thresholding / missing candidates	lower score threshold before NMS; count unmatched truths

Loss, NMS, and AP expose different failure surfaces. Name the clock before changing a threshold.

Architectures rearrange the same evidence

Two-stage means propose, then refine

V VISUAL



RPN

Region Proposal Network: predicts objectness + proposal boxes densely.

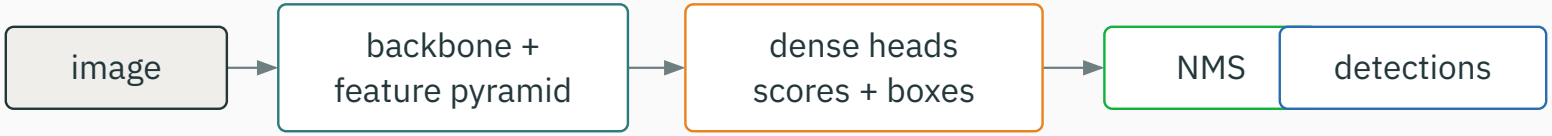
ROI

Region of Interest: samples a fixed-size feature for each proposal.

Stage 1 asks “where might an object be?” Stage 2 spends more computation to classify and refine a smaller proposal set.

One-stage predicts every candidate directly

V VISUAL

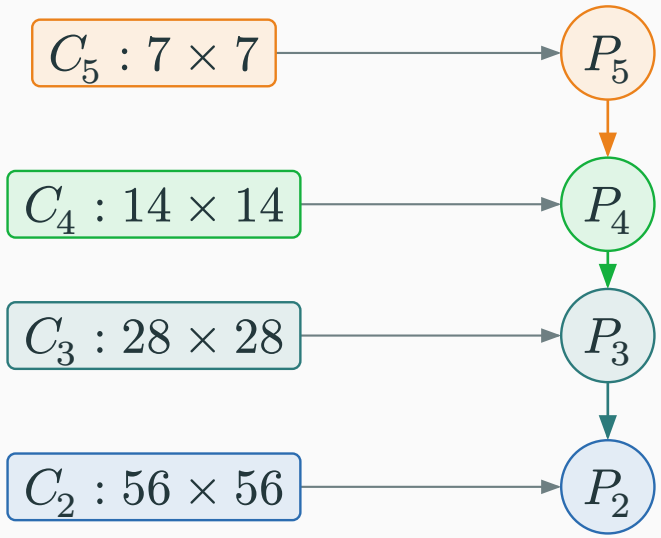


There is no learned proposal stage or per-RoI second head: classification and localization happen densely in one pass.

> One-stage versus two-stage describes where proposal/refinement computation occurs—not a timeless guarantee about speed or accuracy.

FPN gives small and large objects strong features

V VISUAL



FPN = **Feature Pyramid Network**: lateral 1×1 projections + top-down upsampling + same-shape addition.

High-resolution levels localize small objects; top-down semantics make those levels useful for recognition.

Anchor-free heads change the box parameterization

V VISUAL

ANCHOR-BASED	ANCHOR-FREE
start from (x_a, y_a, w_a, h_a); regress normalized translations and log scales	predict a center/keypoint or distances to left, top, right, bottom sides

Both still need spatial features, class evidence, box geometry, a training matcher, and an evaluation protocol.

Duplicate handling is a design choice, not a law

V VISUAL

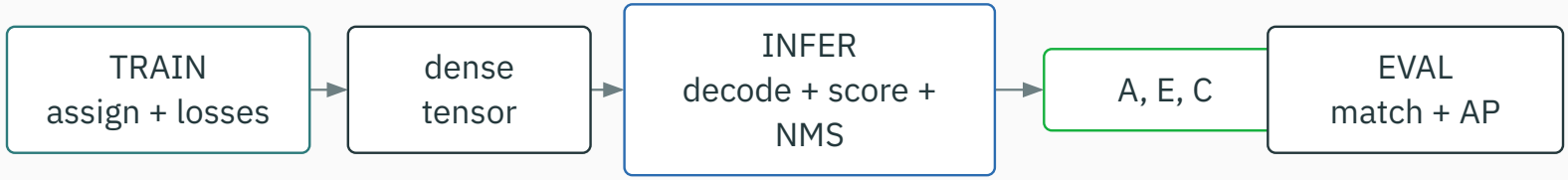
SOFT-NMS	SET PREDICTION
decay a neighbour’s score instead of deleting it; nearby objects may survive	train a fixed query set with one-to-one assignment; duplicate avoidance moves into learning

These alternatives change inference and training, but the evaluation clock still matches scored predictions to truths under a declared protocol.

One pipeline, three clocks, one returned set

The complete pipeline changes what IoU is comparing

V VISUAL



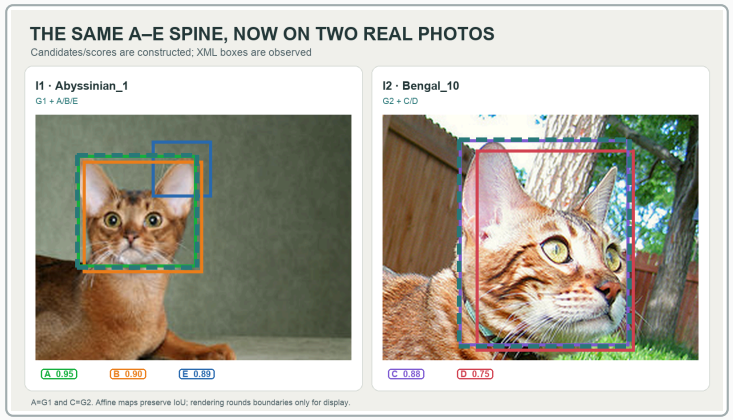
clock	IoU compares	decision
TRAIN	candidate ↔ truth	which targets and losses apply
INFER	prediction ↔ prediction	which neighbours are redundant
EVAL	prediction ↔ truth	TP / FP / FN and AP

The formula is the same; the objects, threshold, and consequence are not.

Revisit the opening commitment before revealing it

Q QUESTION

HELD CASE · SAME FIVE CANDIDATES A .95, B .90, E .89, C .88, D .75; $\tau_{\text{NMS}} = .50$, $\tau_{\text{eval}} = .50$.

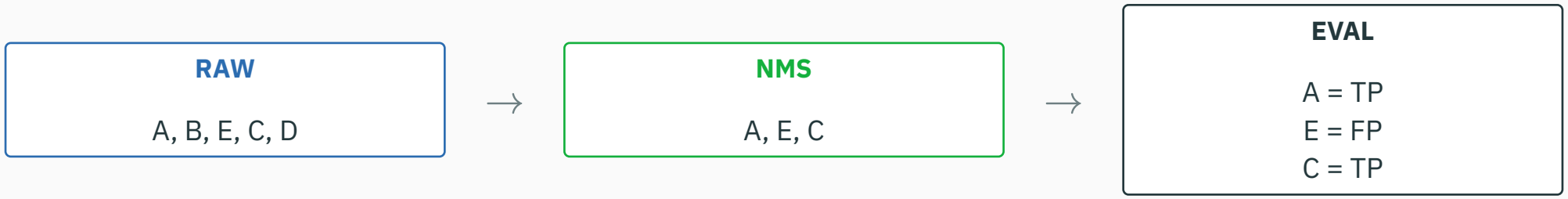


A · THRESHOLD	B · GEOMETRY	C · LOSS
all five pass .70	decode → score → NMS → evaluation match	requires truths unavailable at inference

Defend your original choice using the exact boxes and clocks—not a detector name.

Answer: B is complete—and still cannot remove E

A ANSWER



$$\mathrm{AP}_{\mathrm{toy}} = \frac{5}{6} \approx .833$$

Thresholds control admission, NMS controls redundancy, and matching measures correctness. No one operation can substitute for the others.

Final detection checklist

Q QUESTION

Given a new detector, ask in order:

1. What box coordinate and area convention does every component use?
2. What is the dense output shape, and what does each scalar mean?
3. Which TRAIN matcher creates positive, ignored, and background targets?
4. Which loss reduction and weights determine gradient mass?
5. Which INFER score and NMS policies create the returned set?
6. Which EVAL IoU, interpolation, class, area, and max-detection protocol defines AP?
7. Which symptom → suspect → test check will distinguish geometry from learning failure?

A detector is trustworthy only when the tensor, geometry, three clocks, and metric protocol are all explicit.

L11 handoff: boxes become a prediction at every pixel

V VISUAL

DETECTION

variable set
class + box + score

→

SEGMENTATION

dense spatial output
class evidence at each pixel

Both reuse the L8B backbone. L11 will replace four-number rectangles with masks, then recover boundaries that pooling made coarse.

Today asked “which objects, and where?” Next asks “which class owns every pixel—and which instance?”

Primary provenance and exact companions

V VISUAL

DETECTOR FAMILIES

Girshick et al. · R-CNN
Girshick · Fast R-CNN
Ren et al. · Faster R-CNN
Redmon et al. · YOLO

LOSSES, PYRAMIDS + CODE

Lin et al. · Focal Loss / RetinaNet
Lin et al. · Feature Pyramid Networks
Rezatofighi et al. · GIoU
Exact boxes / IoU Colab · Exact NMS Colab

REAL EVIDENCE

Oxford-IIIT Pet · Parkhi et al. (CVPR 2012)
I1 Abyssinian_1 · I2 Bengal_10
official XML tight-head ROIs · CC BY-SA 4.0

REPRODUCIBILITY

`shared/vision-evidence/oxford-iiit-pet/l9/`
builder + manifest + CSV decision ledgers
A–E/scores = constructed · model output = none

OBSERVED = photos/XML · CONSTRUCTED = crops/candidates/scores · COMPUTED = IoU/NMS/matching/AP