Semantic and Instance Segmentation

Pixel evidence becomes a class field—and, when needed, separate object masks

Prof. Nipun Batra

ES 667 · Deep Learning · IIT Gandhinagar

Detection drew rectangles; segmentation must assign pixels

One real photograph changes the supervision from a box to pixels

OBSERVED PIXELS + GT ANNOTATION

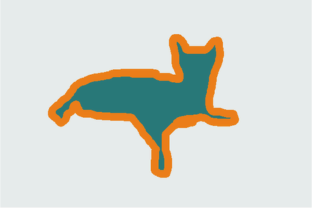
One real photograph, one official pixel contract

OBSERVED

Oxford-IIIT Pet RGB photograph

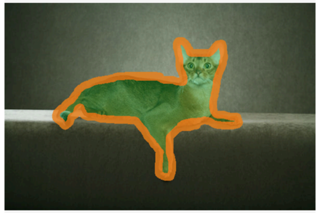
GT VALUES

1 foreground · 2 background · 3 boundary



GT OVERLAY

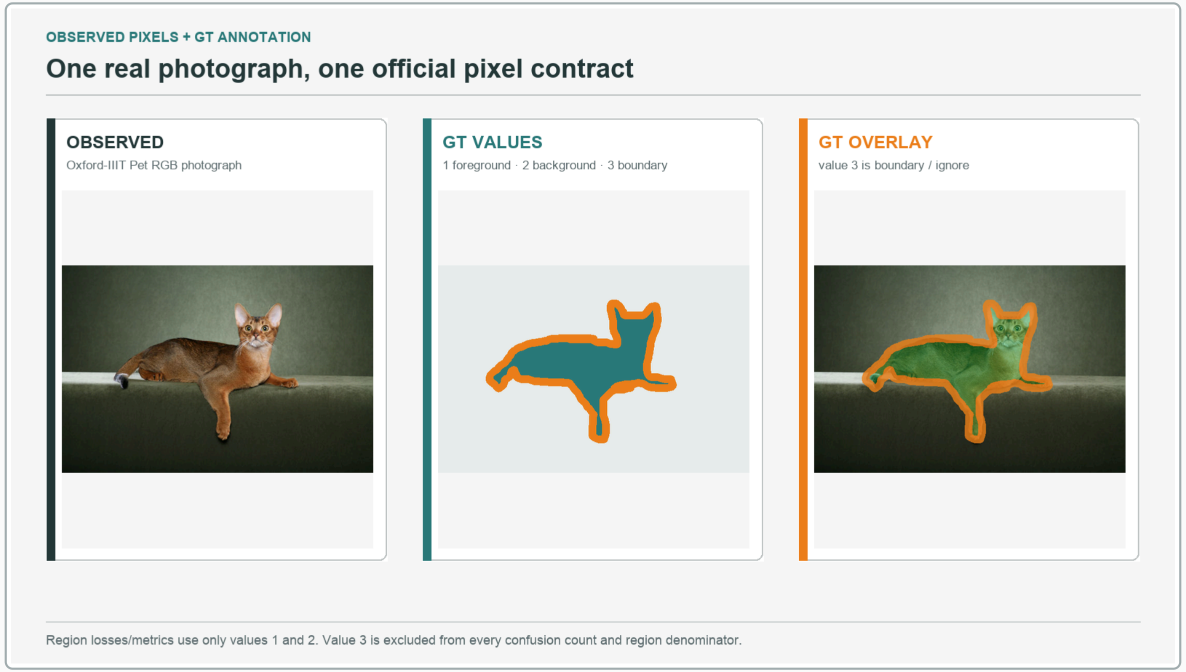
value 3 is boundary / ignore



Region losses/metrics use only values 1 and 2. Value 3 is excluded from every confusion count and region denominator.

One real photograph changes the supervision from a box to pixels

V VISUAL



OBSERVED Oxford pixels + official GT trimap. Value 1 is foreground, 2 background, and 3 boundary/ignore; no model output appears here.

Commit: which output can answer both requirements?

Q QUESTION

HELD CASE · GENERATED CONCEPT SCENE Two touching crowns form one region. Report total canopy area **and** count two trees.

A	B	C	D
one global class	one class + one box	one $H \times W$ foreground map	a scored set of separate object masks

Commit: which output can answer both requirements?

Q QUESTION

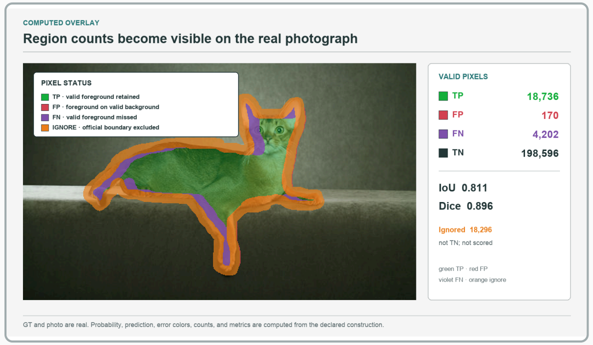
HELD CASE · GENERATED CONCEPT SCENE Two touching crowns form one region. Report total canopy area **and** count two trees.

A	B	C	D
one global class	one class + one box	one $H \times W$ foreground map	a scored set of separate object masks

Commit silently; the same scene returns after semantic and instance outputs.

Outline

V VISUAL



OBSERVED pixels + GT · CONSTRUCTED probabilities · COMPUTED errors

01 · OUTPUT CONTRACT

Separate semantic labels, instances, and panoptic records.

02 · PIXEL LEARNING

Turn logits into probabilities; train with declared losses.

03 · SPATIAL DETAIL

Decode resolution with skips, upsampling, and alignment.

04 · METRICS

Threshold once; compute IoU, Dice, and boundary failures.

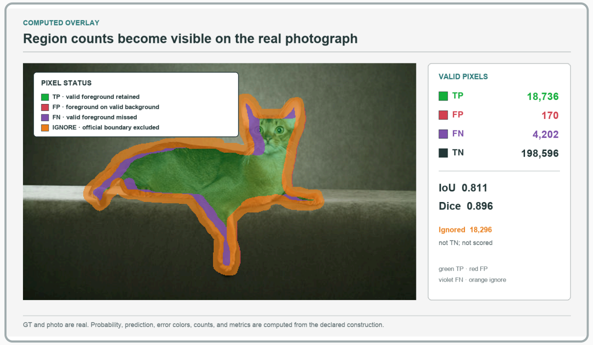
05 · INSTANCES

Predict a scored set of separate object masks.

06 · EVALUATION

Write void, empty-mask, matching, and averaging policies.

Outline



OBSERVED pixels + GT · CONSTRUCTED probabilities · COMPUTED errors

01 · OUTPUT CONTRACT

Separate semantic labels, instances, and panoptic records.

02 · PIXEL LEARNING

Turn logits into probabilities; train with declared losses.

03 · SPATIAL DETAIL

Decode resolution with skips, upsampling, and alignment.

04 · METRICS

Threshold once; compute IoU, Dice, and boundary failures.

05 · INSTANCES

Predict a scored set of separate object masks.

06 · EVALUATION

Write void, empty-mask, matching, and averaging policies.

truth / TRAIN · logit, probability / INFER · hard prediction · overlap / TP · FP or FN · EVAL

Three clocks prevent a probability from becoming a metric

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Three clocks prevent a probability from becoming a metric

V VISUAL

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP
TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

Three clocks prevent a probability from becoming a metric

V VISUAL

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP
TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP
TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

Three clocks prevent a probability from becoming a metric

V VISUAL

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP
TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP
TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

TRAIN compares with labels; INFER makes an output; EVAL scores that output under a declared protocol.

The exact arithmetic case is constructed—not an orchard annotation

V VISUAL

generated concept

not a measured orchard sample

constructed truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

$|G| = 6$ pixels

constructed identities

0	1	2	0
0	1	2	0
0	1	2	0
0	0	0	0

IDs withheld from G.

The exact arithmetic case is constructed—not an orchard annotation

V VISUAL

generated concept

not a measured orchard sample

constructed truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

$|G| = 6$ pixels

constructed identities

0	1	2	0
0	1	2	0
0	1	2	0
0	0	0	0

IDs withheld from G.

The 4×4 ledger isolates the arithmetic: one union supports area; two identities support counting.

Semantic, instance, and panoptic outputs answer different questions

V VISUAL

Task	Output contract	Question answered
semantic	$H \times W$ class labels	Which class owns this pixel?
instance	variable scored set of object masks	Which pixels belong to object j?
panoptic	$H \times W$ class–identity pairs	What and which object, everywhere?

Semantic, instance, and panoptic outputs answer different questions

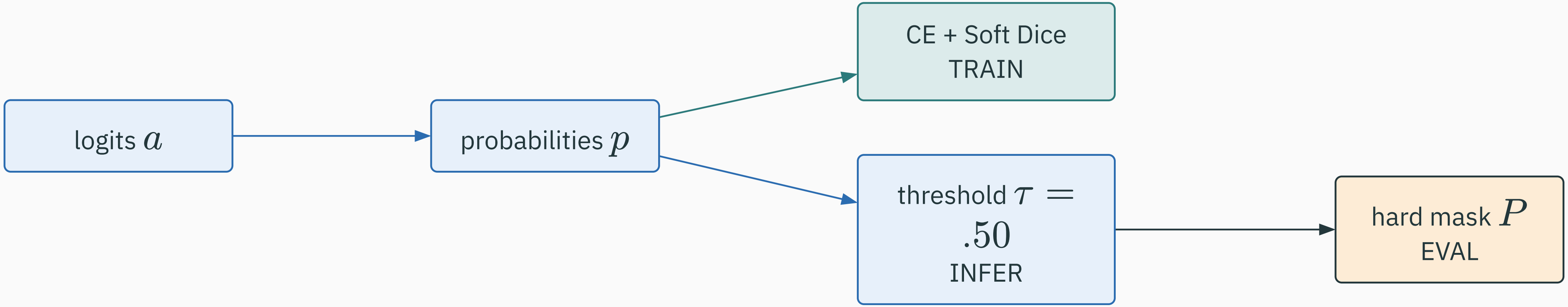
V VISUAL

Task	Output contract	Question answered
semantic	$H \times W$ class labels	Which class owns this pixel?
instance	variable scored set of object masks	Which pixels belong to object j?
panoptic	$H \times W$ class–identity pairs	What and which object, everywhere?

> “Stuff” such as sky usually needs a class; countable “things” such as trees also need an identity.

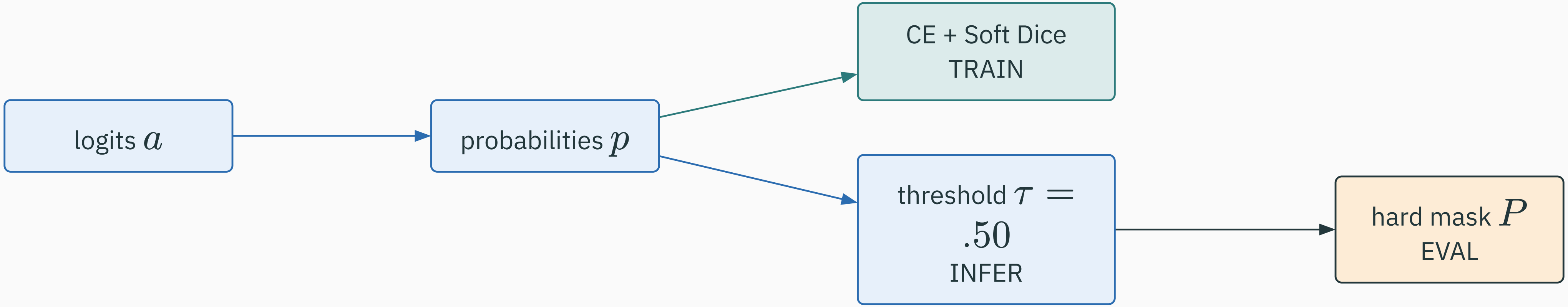
One prediction will cross all three clocks

V VISUAL



One prediction will cross all three clocks

V VISUAL



HELD CASE · PROMISE The exact G, p, threshold, and P will remain fixed through loss, inference, and evaluation.

TRAIN: classify every spatial location

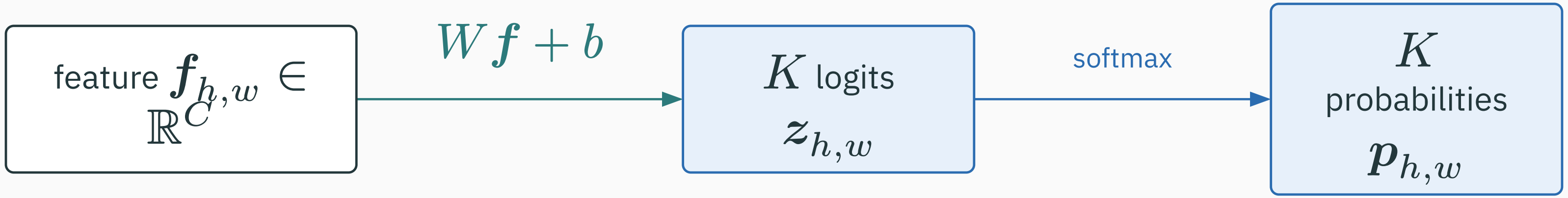
A semantic head emits one score vector per pixel

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



A semantic head emits one score vector per pixel

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

feature $\boldsymbol{f}_{h,w} \in \mathbb{R}^C$ $\xrightarrow{W\boldsymbol{f}+b}$ K logits $\boldsymbol{z}_{h,w}$ $\xrightarrow{\text{softmax}}$ K probabilities $\boldsymbol{p}_{h,w}$

Segmentation shares one classifier across locations; evidence changes, weights do not. Argmax waits for INFER.

Softmax acts across classes at one location—not across pixels

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · ONE PIXEL Classes are **background, road, car** and $\boldsymbol{z} = \begin{pmatrix} .2 \\ 1.4 \\ -.3 \end{pmatrix}$.

Softmax acts across classes at one location—not across pixels

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · ONE PIXEL Classes are **background, road, car** and $\boldsymbol{z} = \begin{pmatrix} .2 \\ 1.4 \\ -.3 \end{pmatrix}$.

EXPONENTIATE

$$\exp(\boldsymbol{z}) \approx \begin{pmatrix} 1.22 \\ 4.05 \\ .74 \end{pmatrix} \quad \textstyle\sum_k \exp(z_k) \approx 6.01$$

NORMALIZE

$$\boldsymbol{p} \approx \begin{pmatrix} .203 \\ .674 \\ .123 \end{pmatrix} \quad p_{\text{road}} = .674$$

Softmax acts across classes at one location—not across pixels

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · ONE PIXEL Classes are **background, road, car** and $\boldsymbol{z} = \begin{pmatrix} .2 \\ 1.4 \\ -.3 \end{pmatrix}$.

EXPONENTIATE

$$\exp(\boldsymbol{z}) \approx \begin{pmatrix} 1.22 \\ 4.05 \\ .74 \end{pmatrix} \quad \textstyle\sum_k \exp(z_k) \approx 6.01$$

NORMALIZE

$$\boldsymbol{p} \approx \begin{pmatrix} .203 \\ .674 \\ .123 \end{pmatrix} \quad p_{\text{road}} = .674$$

$$\ell = -\log p_{\text{road}} = -\log .674 \approx .395.$$

Softmax acts across classes at one location—not across pixels

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · ONE PIXEL Classes are **background, road, car** and $\boldsymbol{z} = \begin{pmatrix} .2 \\ 1.4 \\ -.3 \end{pmatrix}$.

EXPONENTIATE

$$\exp(\boldsymbol{z}) \approx \begin{pmatrix} 1.22 \\ 4.05 \\ .74 \end{pmatrix} \quad \textstyle\sum_k \exp(z_k) \approx 6.01$$

NORMALIZE

$$\boldsymbol{p} \approx \begin{pmatrix} .203 \\ .674 \\ .123 \end{pmatrix} \quad p_{\text{road}} = .674$$

$$\ell = -\log p_{\text{road}} = -\log .674 \approx .395.$$

$\sum_k p_{h,w,k} = 1$ **independently at every** (h,w).

Binary foreground can use one logit and one sigmoid

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$p_i = \sigma(a_i) = \frac{1}{1+e^{-a_i}}, \quad a_i = \log \frac{p_i}{1-p_i}.$$

Binary foreground can use one logit and one sigmoid

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$p_i = \sigma(a_i) = \frac{1}{1+e^{-a_i}}, \quad a_i = \log \frac{p_i}{1-p_i}.$$

p	.90	.60	.30	.10	.05
$a = \mathrm{logit}(p)$	2.197	.405	$-.847$	-2.197	-2.944

Binary foreground can use one logit and one sigmoid

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$p_i = \sigma(a_i) = \frac{1}{1+e^{-a_i}}, \quad a_i = \log \frac{p_i}{1-p_i}.$$

p	$.90$	$.60$	$.30$	$.10$	$.05$
$a = \text{logit}(p)$	2.197	$.405$	$-.847$	-2.197	-2.944

> The held binary mask uses one foreground probability. A K-class semantic task instead keeps K logits and softmax.

The held foreground probabilities are now fixed

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · CONSTRUCTED PROBABILITY FIELD p_i is a teaching construction—not model output; blue cells satisfy $p_i \geq .50$.

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

$\rightarrow$

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

The held foreground probabilities are now fixed

V VISUAL

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

HELD CASE · CONSTRUCTED PROBABILITY FIELD p_i is a teaching construction—not model output; blue cells satisfy $p_i \geq .50$.

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

→

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

$$\sum_i p_i = 7.2, \quad \sum_i g_i = 6, \quad \sum_i p_i g_i = 4.8.$$

Binary cross-entropy scores each pixel locally

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\ell_i = -[g_i \log p_i + (1 - g_i)\log(1 - p_i)].$$

Binary cross-entropy scores each pixel locally

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

$$\ell_i = -[g_i \log p_i + (1 - g_i) \log(1 - p_i)].$$

FOREGROUND · $g_i = 1$

penalty $= -\log p_i$

confident $.90 \to .105$

missed $.30 \to 1.204$

BACKGROUND · $g_i = 0$

penalty $= -\log(1 - p_i)$

safe $.10 \to .105$

false alarm $.60 \to .916$

Binary cross-entropy scores each pixel locally

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

$$\ell_i = -[g_i \log p_i + (1-g_i)\log(1-p_i)].$$

FOREGROUND · $g_i = 1$

penalty $= -\log p_i$

confident $.90 \to .105$

missed $.30 \to 1.204$

BACKGROUND · $g_i = 0$

penalty $= -\log(1-p_i)$

safe $.10 \to .105$

false alarm $.60 \to .916$

CE asks whether each probability is calibrated to that pixel's label.

Group the held pixels to reproduce dense BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

HELD CASE · SAME G, p Five foreground pixels have $.90$; one has $.30$. Background has $.60 \times 3$, $.10 \times 5$, $.05 \times 2$.

Group the held pixels to reproduce dense BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · SAME G, p Five foreground pixels have $.90$; one has $.30$. Background has $.60 \times 3$, $.10 \times 5$, $.05 \times 2$.

$$16\mathcal{L}_{\text{BCE}} = -10\log .90 - \log .30 - 3\log .40 - 2\log .95$$

Group the held pixels to reproduce dense BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · SAME G, p Five foreground pixels have $.90$; one has $.30$. Background has $.60 \times 3$, $.10 \times 5$, $.05 \times 2$.

$$16\mathcal{L}_{\mathrm{BCE}} = -10\log .90 - \log .30 - 3\log .40 - 2\log .95$$

$$\approx 1.0536 + 1.2040 + 2.7489 + .1026 \approx 5.1090.$$

Group the held pixels to reproduce dense BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · SAME G, p Five foreground pixels have $.90$; one has $.30$. Background has $.60 \times 3$, $.10 \times 5$, $.05 \times 2$.

$$16\mathcal{L}_{\text{BCE}} = -10\log .90 - \log .30 - 3\log .40 - 2\log .95$$

$$\approx 1.0536 + 1.2040 + 2.7489 + .1026 \approx 5.1090.$$

$$\mathcal{L}_{\text{BCE}} \approx \frac{5.109}{16} \approx .319.$$

The tensor keeps spatial axes separate from the class axis

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Axis	Size	Meaning
h	H	vertical location
w	W	horizontal location
k	K	class score or probability

The tensor keeps spatial axes separate from the class axis

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Axis	Size	Meaning
h	H	vertical location
w	W	horizontal location
k	K	class score or probability

$$Z[h, w, :] \in \mathbb{R}^K \to \boldsymbol{p}[h, w, :] \in [0, 1]^K.$$

The tensor keeps spatial axes separate from the class axis

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Axis	Size	Meaning
h	H	vertical location
w	W	horizontal location
k	K	class score or probability

$$Z[h, w, :] \in \mathbb{R}^K \to \boldsymbol{p}[h, w, :] \in [0, 1]^K.$$

Normalize over k; reduce the loss over valid (h, w) locations.

TRAIN checkpoint: the labels supervise probabilities, not hard masks

Q QUESTION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

Which object enters the derivative during training?

A

the thresholded mask P

B

the differentiable probability p

C

only pixel accuracy

TRAIN checkpoint: the labels supervise probabilities, not hard masks

Q QUESTION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Which object enters the derivative during training?

A

the thresholded mask P

B

the differentiable probability p

C

only pixel accuracy

B. CE and Soft Dice operate on p; thresholding belongs to INFER.

TRAIN: preserve detail while acquiring context

A classification head destroys the layout segmentation needs

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

feature map $H' \times W' \times C$ → global pool → one vector $\mathbb{R}^K$

A classification head destroys the layout segmentation needs

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

feature map $H' \times W' \times C$ → global pool → one vector $\mathbb{R}^K$

After locations collapse, different pixels cannot receive different labels.

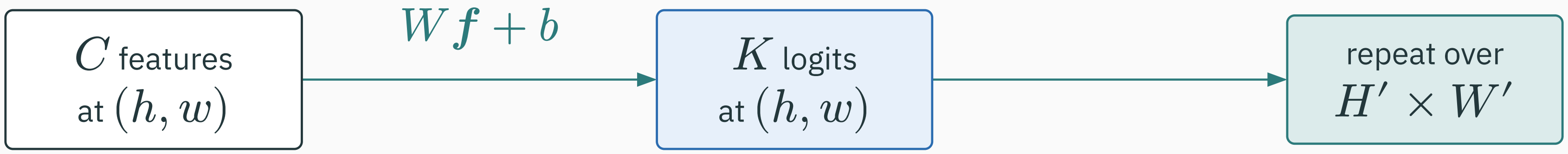
A 1×1 convolution is the shared pixel classifier

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



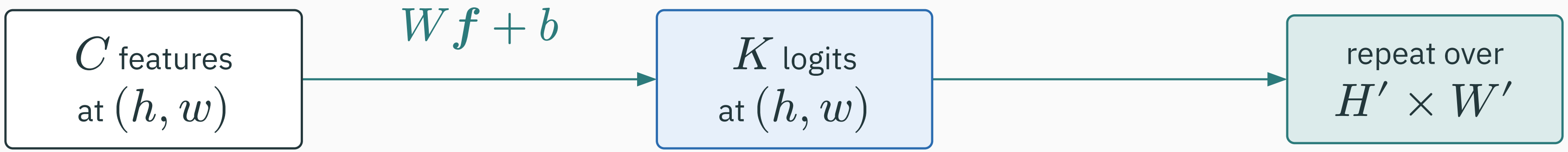
A 1×1 convolution is the shared pixel classifier

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



> A 1×1 kernel mixes channels at one location. Earlier spatial convolutions already gathered neighborhood evidence.

Encoder–decoder networks exchange coordinates for context

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

INPUT $256 \times 256 \times 3$ → **ENCODE** $32 \times 32 \times 256$ broad context → **DECODE** $256 \times 256 \times 32$ localize → **LOGITS** $256 \times 256 \times K$

Encoder–decoder networks exchange coordinates for context

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

INPUT $256 \times 256 \times 3$ → **ENCODE** $32 \times 32 \times 256$ broad context → **DECODE** $256 \times 256 \times 32$ localize → **LOGITS** $256 \times 256 \times K$

Downsample to understand what surrounds a pixel; upsample to decide which pixel.

A real boundary audit exposes what region overlap can hide

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

COMPUTED AUDIT

Region overlap can be strong while boundaries still shift

REGION · valid pixels only

BCE	**0.1487**
Soft Dice	**0.5739**
IoU	**0.8108**
Hard Dice	**0.8955**
Accuracy	**0.9803**

BOUNDARY · contour distances

Precision @ 5px	**0.8098**
Recall @ 5px	**0.3027**
Boundary F1 @ 5px	**0.4407**
ASSD (px)	**6.940**
HD95 (px)	**25.495**

Two contracts, one lesson

IoU/Dice reduce valid-region confusion. Boundary F1/ASSD/HD95 compare contours; the ignored GT band is the reference contour.

Boundary metrics are a diagnostic extension; they are not part of the deck's exact 4 x 4 numerical spine.

Does a larger grid recreate discarded detail?

Q QUESTION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

A 2×2 feature map is repeated into a 4×4 grid. Did upsampling recover which source pixels created each value?

1	2
3	4

$\rightarrow$

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Does a larger grid recreate discarded detail?

Q QUESTION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

A 2×2 feature map is repeated into a 4×4 grid. Did upsampling recover which source pixels created each value?

$$\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 & 4 \\ \hline \end{array} \quad \rightarrow \quad \begin{array}{|c|c|c|c|}\hline 1 & 1 & 2 & 2 \\ \hline 1 & 1 & 2 & 2 \\ \hline 3 & 3 & 4 & 4 \\ \hline 3 & 3 & 4 & 4 \\ \hline \end{array}$$

Commit before the next slide: resolution and information are not synonyms.

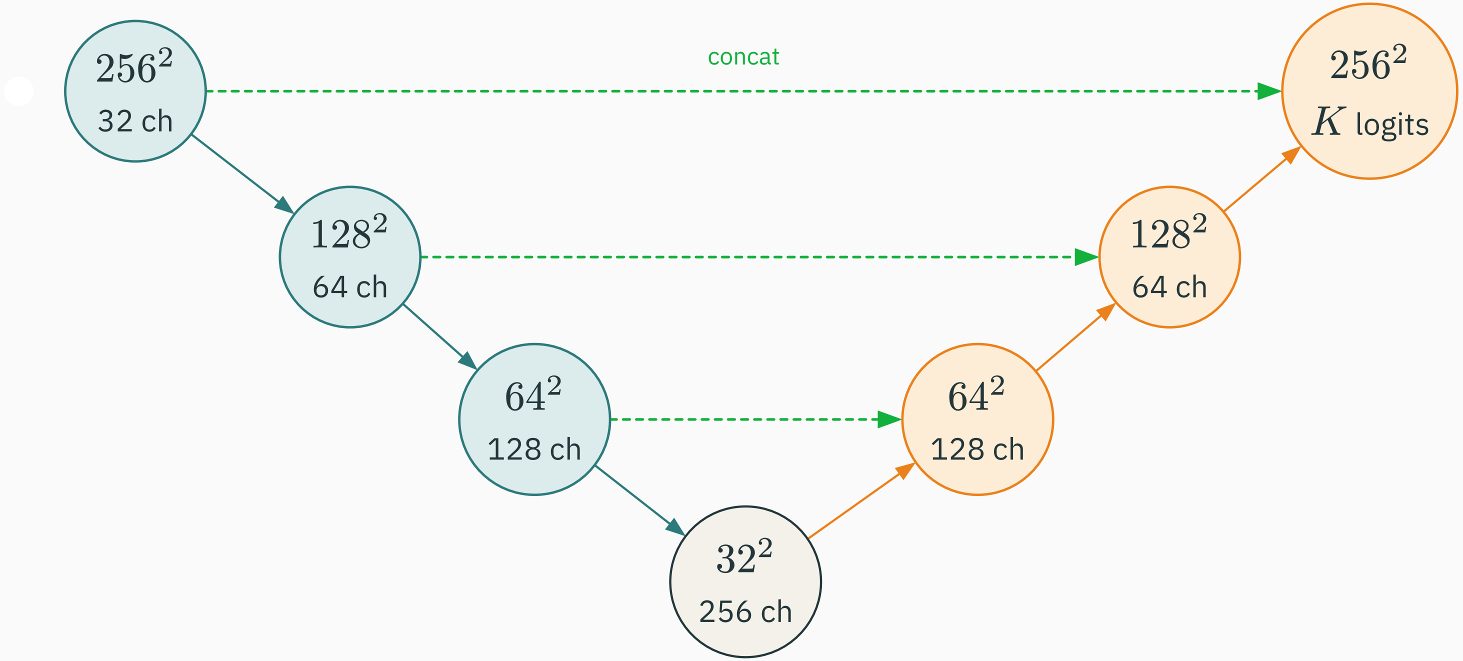
U-Net routes fine detail around the bottleneck

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



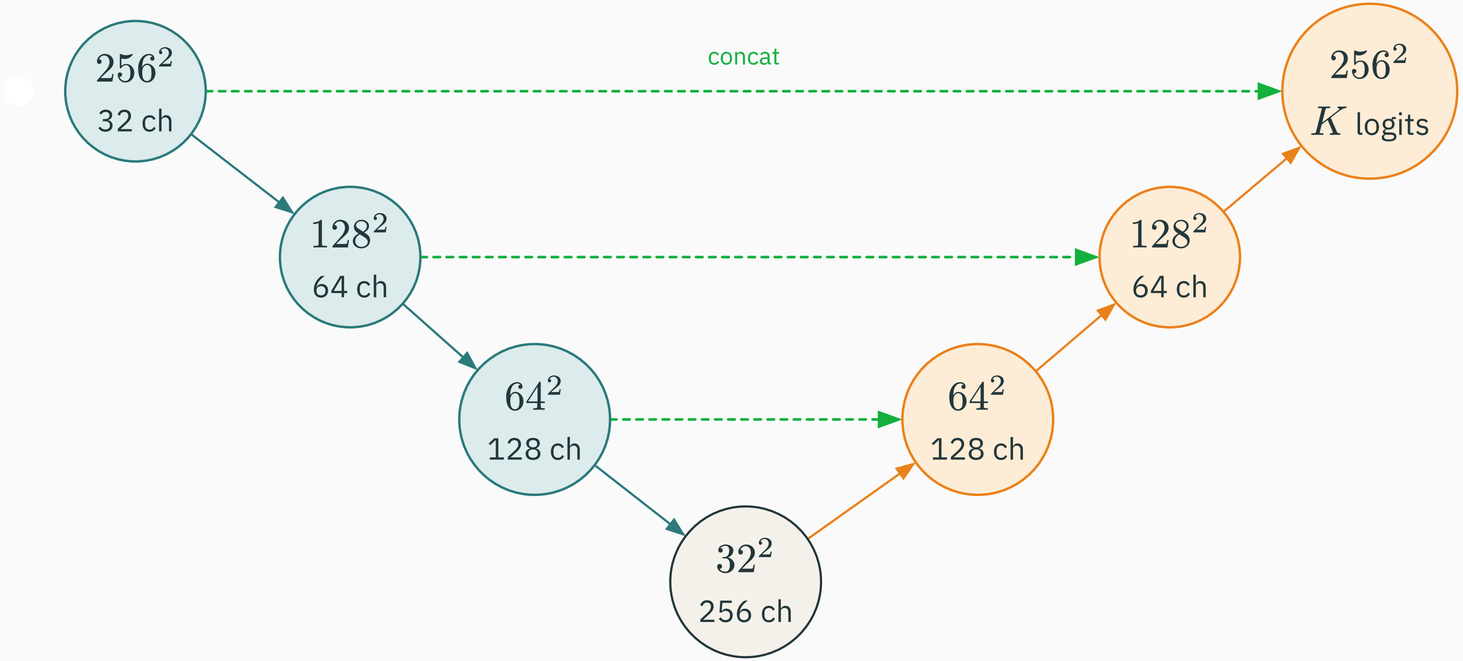
U-Net routes fine detail around the bottleneck

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



Teal learns context; orange restores resolution; green skips carry aligned fine features.

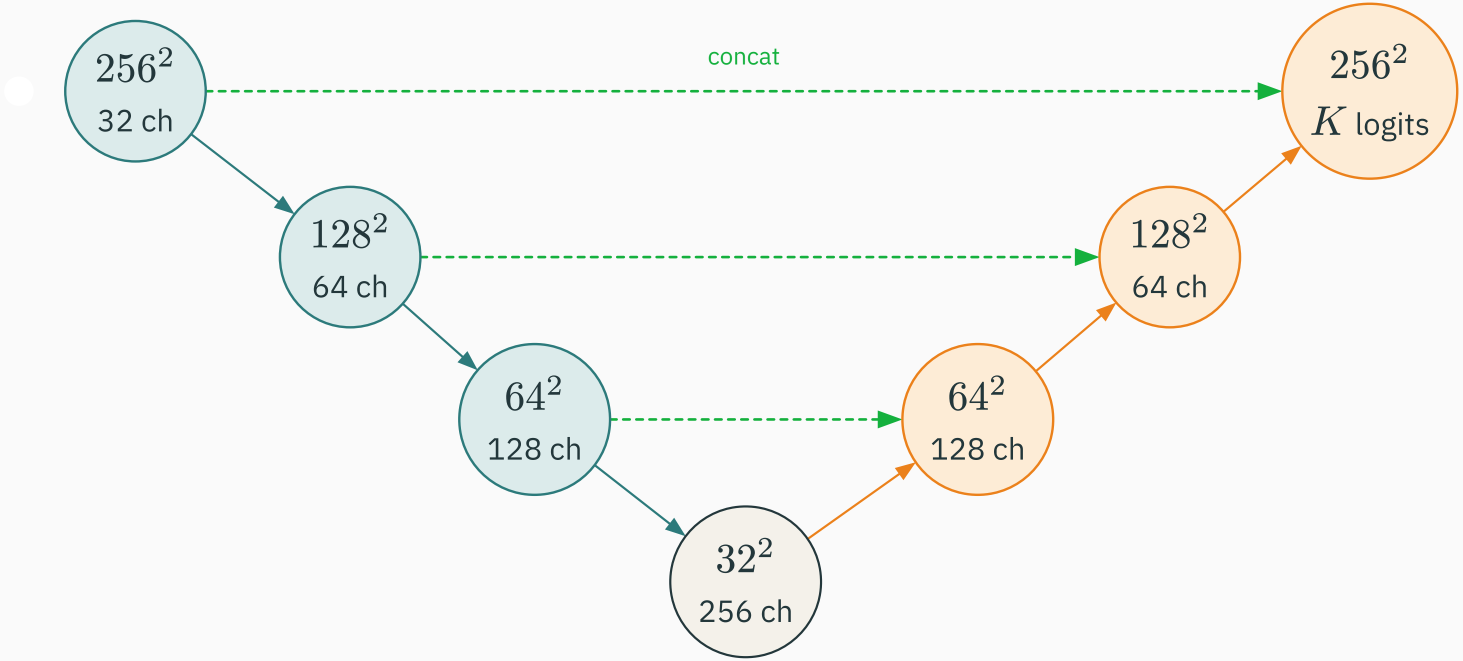
U-Net routes fine detail around the bottleneck

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



Upsampling adds locations; skips restore evidence that the bottleneck discarded.

This ledger is a same-padding teaching U-Net—not the 2015 pixel sizes

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

TOY LEDGER USED HERE

same-padding convolutions preserve each stage size; matching encoder and decoder tensors concatenate directly

ORIGINAL U-NET

valid convolutions shrink tensors; the encoder feature is copied **and cropped** before concatenation

This ledger is a same-padding teaching U-Net—not the 2015 pixel sizes

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

TOY LEDGER USED HERE

same-padding convolutions preserve each stage size; matching encoder and decoder tensors concatenate directly

ORIGINAL U-NET

valid convolutions shrink tensors; the encoder feature is copied **and cropped** before concatenation

$$U, E \in \mathbb{R}^{64\times64\times128} \quad \to \quad \mathrm{concat}(U, E) \in \mathbb{R}^{64\times64\times256}.$$

This ledger is a same-padding teaching U-Net—not the 2015 pixel sizes

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

TOY LEDGER USED HERE

same-padding convolutions preserve each stage size; matching encoder and decoder tensors concatenate directly

ORIGINAL U-NET

valid convolutions shrink tensors; the encoder feature is copied **and cropped** before concatenation

$$U, E \in \mathbb{R}^{64\times 64\times 128} \quad \to \quad \mathrm{concat}(U, E) \in \mathbb{R}^{64\times 64\times 256}.$$

The skip operation is concatenation in original U-Net; "skip connection" is not universally synonymous with addition.

Boundary failure has a smallest useful diagnostic

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Symptom	Suspect	Smallest test
correct class; blurred edge	coarse stride or missing skip	overlay boundary; ablate one skip
checkerboard texture	transposed-conv overlap	compare resize+conv at same output size
shifted instance mask	RoI quantization	compare RoIPool with RoIAlign

Boundary failure has a smallest useful diagnostic

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

Symptom	Suspect	Smallest test
correct class; blurred edge	coarse stride or missing skip	overlay boundary; ablate one skip
checkerboard texture	transposed-conv overlap	compare resize+conv at same output size
shifted instance mask	RoI quantization	compare RoIPool with RoIAlign

Inspect spatial alignment before changing the loss.

Architecture checkpoint: context and coordinates play different roles

Q QUESTION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

If the deep path is removed but every high-resolution skip remains, what is most at risk?

A	B	C
semantic meaning	number of pixels	softmax normalization

Architecture checkpoint: context and coordinates play different roles

Q QUESTION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

If the deep path is removed but every high-resolution skip remains, what is most at risk?

A	B	C
semantic meaning	number of pixels	softmax normalization

A. A crisp local edge does not by itself say whether the region is road, cat, or shadow.

INFER makes a mask; EVAL measures its agreement

TRAIN ends with probabilities; INFER begins without labels

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$G, p \to \mathcal{L}_{\text{BCE}}, \mathcal{L}_{\text{Dice}}$$

TRAIN ends with probabilities; INFER begins without labels

V VISUAL

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$G, p \to \mathcal{L}_{\mathrm{BCE}}, \mathcal{L}_{\mathrm{Dice}}$$

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$p, \tau \to P$; no truth and no loss

TRAIN ends with probabilities; INFER begins without labels

V VISUAL

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$G, p \to \mathcal{L}_{\text{BCE}}, \mathcal{L}_{\text{Dice}}$$

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$p, \tau \to P$; no truth and no loss

The same p crosses a clock boundary; its meaning does not change.

Commit: threshold the held probabilities at $.50$

Q QUESTION

TRAIN	probability ↔ truth; CE / Dice
INFER	probability → label / object masks
EVAL	hard output ↔ truth; IoU / AP

HELD CASE · DECISION RULE $P_i = 1[p_i \geq .50]$. Equality is foreground by convention.

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

$\rightarrow$

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

Commit: threshold the held probabilities at $.50$

Q QUESTION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · DECISION RULE $P_i = 1[p_i \geq .50]$. Equality is foreground by convention.

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

$\rightarrow$

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

Before revealing P, predict the four counts: TP, FP, FN, TN.

Thresholding yields this exact hard mask

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

→

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

→

hard P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

Thresholding yields this exact hard mask

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

→

foreground p

.10	.90	.90	.10
.60	.90	.90	.10
.10	.90	.30	.60
.10	.60	.05	.05

→

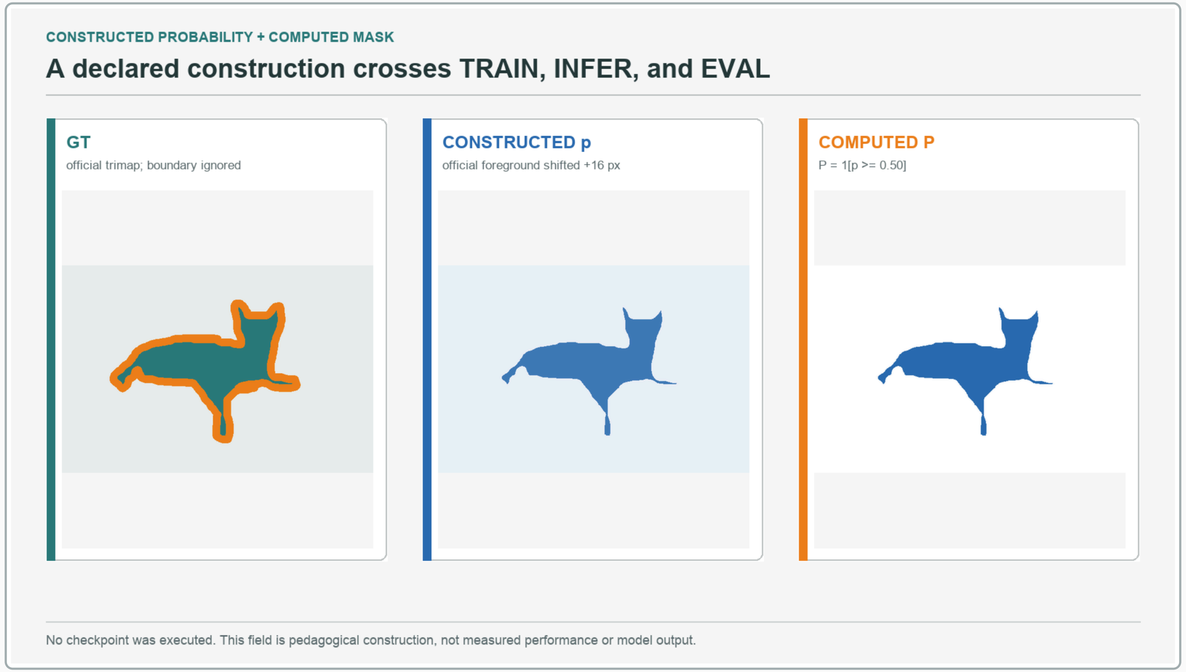
hard P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

Thresholding is nondifferentiable; it belongs to inference or metric computation, not the training path.

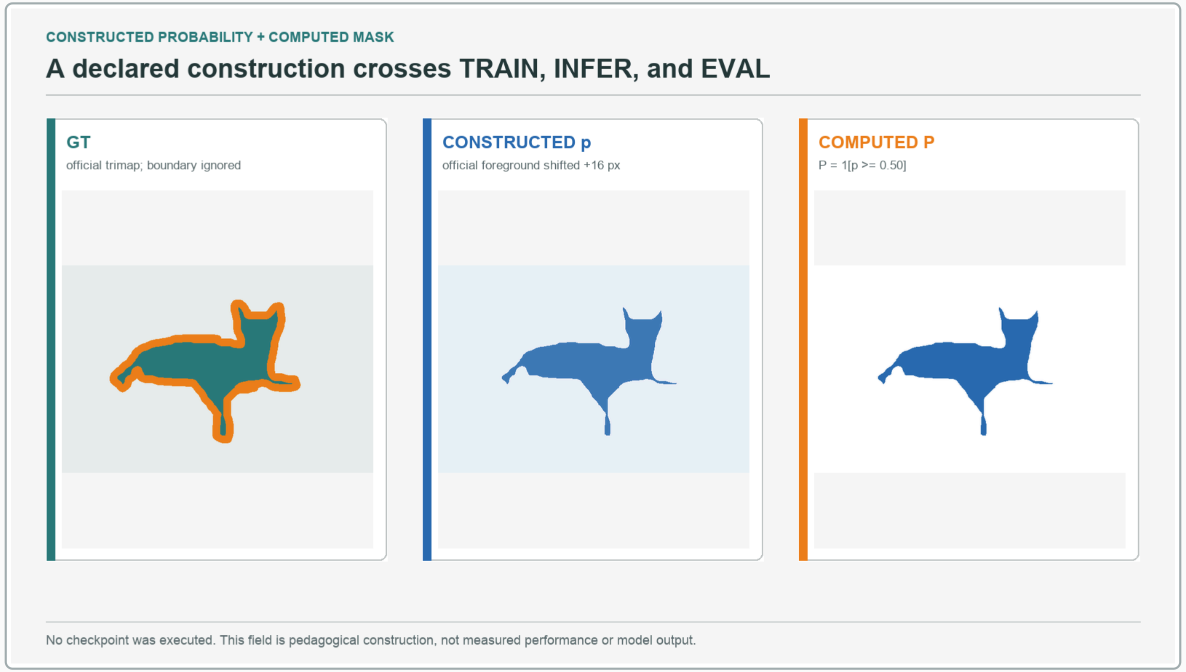
The same clocks operate on actual pixels

V VISUAL



The same clocks operate on actual pixels

V VISUAL



> GT is official; p is the declared +16 px construction; $P = 1[p \geq .50]$ is computed. No checkpoint was executed.

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

$$\mathrm{TP} = 5$$

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

$$\mathrm{TP} = 5 \quad \mathrm{FP} = 3$$

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

$$\mathrm{TP} = 5 \quad \mathrm{FP} = 3 \quad \mathrm{FN} = 1$$

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

$$\mathrm{TP}=5 \quad \mathrm{FP}=3 \quad \mathrm{FN}=1 \quad \mathrm{TN}=7$$

Count errors before naming a metric

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

truth G

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

prediction P

0	1	1	0
1	1	1	0
0	1	0	1
0	1	0	0

$$\mathrm{TP} = 5 \quad \mathrm{FP} = 3 \quad \mathrm{FN} = 1 \quad \mathrm{TN} = 7$$

Every hard metric below is a different reduction of the same four counts.

Pixel accuracy can reward an empty foreground prediction

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · HELD MASK $\text{accuracy} = \frac{\text{TP+TN}}{16} = \frac{5+7}{16} = .750.$

Pixel accuracy can reward an empty foreground prediction

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · HELD MASK $\text{accuracy} = \frac{\text{TP}+\text{TN}}{16} = \frac{5+7}{16} = .750.$

Now imagine a 100×100 image with only 100 tumor pixels.

Pixel accuracy can reward an empty foreground prediction

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

HELD CASE · HELD MASK $\text{accuracy} = \frac{\text{TP}+\text{TN}}{16} = \frac{5+7}{16} = .750.$

Now imagine a 100×100 image with only 100 tumor pixels.

$$\text{predict background everywhere} \rightarrow \text{accuracy} = \frac{9900}{10000} = 99\%.$$

Pixel accuracy can reward an empty foreground prediction

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

HELD CASE · HELD MASK $\text{accuracy} = \frac{\text{TP+TN}}{16} = \frac{5+7}{16} = .750.$

Now imagine a 100×100 image with only 100 tumor pixels.

$$\text{predict background everywhere} \to \text{accuracy} = \frac{9900}{10000} = 99\%.$$

$$\text{tumor recall} = \frac{0}{100} = 0.$$

Pixel accuracy can reward an empty foreground prediction

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

HELD CASE · HELD MASK $\text{accuracy} = \frac{\text{TP}+\text{TN}}{16} = \frac{5+7}{16} = .750.$

Now imagine a 100×100 image with only 100 tumor pixels.

$$\text{predict background everywhere} \to \text{accuracy} = \frac{9900}{10000} = 99\%.$$

$$\text{tumor recall} = \frac{0}{100} = 0.$$

Accuracy weights the easy background by its frequency; it may not reflect the foreground task.

IoU removes true background from the denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

IoU removes true background from the denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P \cap G| = 5$ and $|P \cup G| = 5 + 3 + 1 = 9.$

IoU removes true background from the denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P \cap G| = 5$ and $|P \cup G| = 5 + 3 + 1 = 9.$

$$\mathrm{IoU} = \tfrac{5}{9} \approx .556.$$

IoU removes true background from the denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P \cap G| = 5$ and $|P \cup G| = 5 + 3 + 1 = 9.$

$$\mathrm{IoU} = \tfrac{5}{9} \approx .556.$$

Seven true-background pixels cannot inflate foreground IoU.

Real pixels reveal where every region count lives

V VISUAL

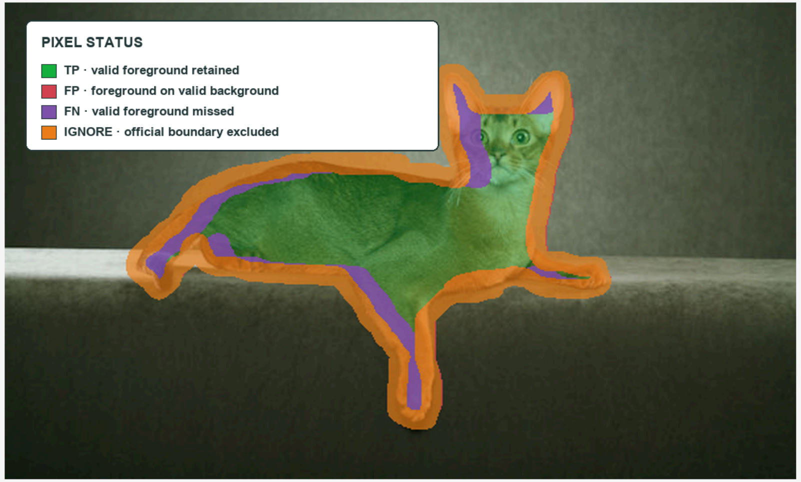
TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

COMPUTED OVERLAY

Region counts become visible on the real photograph



VALID PIXELS

TP	18,736
FP	170
FN	4,202
TN	198,596

IoU 0.811

Dice 0.896

Ignored 18,296

not TN; not scored

green TP · red FP
violet FN · orange ignore

GT and photo are real. Probability, prediction, error colors, counts, and metrics are computed from the declared construction.

Dice gives the intersection twice the weight

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

Dice gives the intersection twice the weight

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P| = 8$, $|G| = 6$, and $|P \cap G| = 5$.

Dice gives the intersection twice the weight

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P| = 8$, $|G| = 6$, and $|P \cap G| = 5$.

$$\mathrm{Dice} = \tfrac{10}{14} \approx .714.$$

Dice gives the intersection twice the weight

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

HELD CASE · SAME COUNTS $|P| = 8$, $|G| = 6$, and $|P \cap G| = 5$.

$$\mathrm{Dice} = \frac{10}{14} \approx .714.$$

For one binary mask, Dice is the foreground F1 score.

Dice is a monotone transform of IoU for one mask pair

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2\mathrm{IoU}}{1+\mathrm{IoU}}, \quad \mathrm{IoU} = \frac{\mathrm{Dice}}{2-\mathrm{Dice}}.$$

Dice is a monotone transform of IoU for one mask pair

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

$$\text{Dice} = \frac{2\text{IoU}}{1+\text{IoU}}, \quad \text{IoU} = \frac{\text{Dice}}{2-\text{Dice}}.$$

$$\frac{2\left(\frac{5}{9}\right)}{1+\frac{5}{9}} = \frac{10}{14} \approx .714.$$

Dice is a monotone transform of IoU for one mask pair

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{Dice} = \frac{2\mathrm{IoU}}{1+\mathrm{IoU}}, \quad \mathrm{IoU} = \frac{\mathrm{Dice}}{2-\mathrm{Dice}}.$$

$$\frac{2\left(\frac{5}{9}\right)}{1+\frac{5}{9}} = \frac{10}{14} \approx .714.$$

Qualifier: this preserves ranking for a single class and mask pair. Macro-averaging nonlinear scores over classes or images can change a model ranking.

Soft Dice scores the held probabilities before thresholding

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{SoftDice} = \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i + \sum_i g_i + \varepsilon}.$$

Soft Dice scores the held probabilities before thresholding

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\mathrm{SoftDice} = \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i + \sum_i g_i + \varepsilon}.$$

HELD CASE · SAME G, p $\sum p = 7.2, \sum g = 6, \sum pg = 4.8$; omit tiny ε in the displayed arithmetic.

Soft Dice scores the held probabilities before thresholding

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{SoftDice} = \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i + \sum_i g_i + \varepsilon}.$$

HELD CASE · SAME G, p $\sum p = 7.2, \sum g = 6, \sum pg = 4.8$; omit tiny ε in the displayed arithmetic.

$$\mathrm{SoftDice} = \frac{9.6}{13.2} \approx .727.$$

Soft Dice scores the held probabilities before thresholding

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$\mathrm{SoftDice} = \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i + \sum_i g_i + \varepsilon}.$$

HELD CASE · SAME G, p $\sum p = 7.2, \sum g = 6, \sum pg = 4.8$; omit tiny ε in the displayed arithmetic.

$$\mathrm{SoftDice} = \frac{9.6}{13.2} \approx .727.$$

$$\mathcal{L}_{\mathrm{Dice}} = 1 - .727 = .273.$$

CE and Soft Dice expose different failure surfaces

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

LOCAL CALIBRATION · CE

scores every valid pixel; strong gradients identify which probabilities are wrong

REGION OVERLAP · DICE

couples foreground pixels through one region denominator; resists background domination

CE and Soft Dice expose different failure surfaces

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

LOCAL CALIBRATION · CE

scores every valid pixel; strong gradients identify which probabilities are wrong

REGION OVERLAP · DICE

couples foreground pixels through one region denominator; resists background domination

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{Dice}}.$$

CE and Soft Dice expose different failure surfaces

V VISUAL

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

LOCAL CALIBRATION · CE

scores every valid pixel; strong gradients identify which probabilities are wrong

REGION OVERLAP · DICE

couples foreground pixels through one region denominator; resists background domination

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{Dice}}.$$

Combining them does not make either definition optional: declare λ, reduction, and empty-mask policy.

Semantic mIoU averages declared per-class IoUs

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

For each class k,

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}, \quad \mathrm{mIoU} = \frac{1}{|K_{\mathrm{report}}|} \sum_{k \in K_{\mathrm{report}}} \mathrm{IoU}_k.$$

Semantic mIoU averages declared per-class IoUs

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

For each class k,

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}, \quad \mathrm{mIoU} = \frac{1}{|K_{\mathrm{report}}|} \sum_{k \in K_{\mathrm{report}}} \mathrm{IoU}_k.$$

CLASSES	VOID	REDUCTION
include or exclude background?	remove ignored pixels from all counts	image mean or dataset-level confusion?

Semantic mIoU averages declared per-class IoUs

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

For each class k,

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}, \quad \mathrm{mIoU} = \frac{1}{|K_{\mathrm{report}}|} \sum_{k \in K_{\mathrm{report}}} \mathrm{IoU}_k.$$

CLASSES	**VOID**	**REDUCTION**
include or exclude background?	remove ignored pixels from all counts	image mean or dataset-level confusion?

“mIoU” is incomplete until its class set and aggregation rule are named.

Diagnose a surprising score on the correct clock

V VISUAL

Symptom	Suspect	Smallest test
99% accuracy; IoU $= 0$	rare foreground	foreground recall; class counts
loss falls; mIoU flat	threshold or class collapse	histogram p by class; confusion
IoU $.811$; boundary F1 $.441$	spatial displacement	overlay FP/FN; inspect contour distances

Diagnose a surprising score on the correct clock

V VISUAL

Symptom	Suspect	Smallest test
99% accuracy; IoU $= 0$	rare foreground	foreground recall; class counts
loss falls; mIoU flat	threshold or class collapse	histogram p by class; confusion
IoU $.811$; boundary F1 $.441$	spatial displacement	overlay FP/FN; inspect contour distances

Region and boundary metrics interrogate different geometry; choose the diagnostic on EVAL, then inspect actual pixels.

One semantic region is not yet two objects

A generated touching-object scene isolates the identity question

V VISUAL

generated concept scene

constructed union

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

one connected component

constructed identities

0	1	2	0
0	1	2	0
0	1	2	0
0	0	0	0

id 1 vs id 2

A generated touching-object scene isolates the identity question

V VISUAL

generated concept scene

constructed union

0	1	1	0
0	1	1	0
0	1	1	0
0	0	0	0

one connected component

constructed identities

0	1	2	0
0	1	2	0
0	1	2	0
0	0	0	0

id 1 vs id 2

Oxford supplies real semantic GT; this explicitly generated scene demonstrates why a semantic union cannot count touching instances.

Instance segmentation returns a scored set—not one class grid

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$x \to \left\{ \left(\hat{y}_j, \hat{b}_j, \hat{m}_j, \hat{s}_j \right) \right\}_{j=1}^{N}.$$

Instance segmentation returns a scored set—not one class grid

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

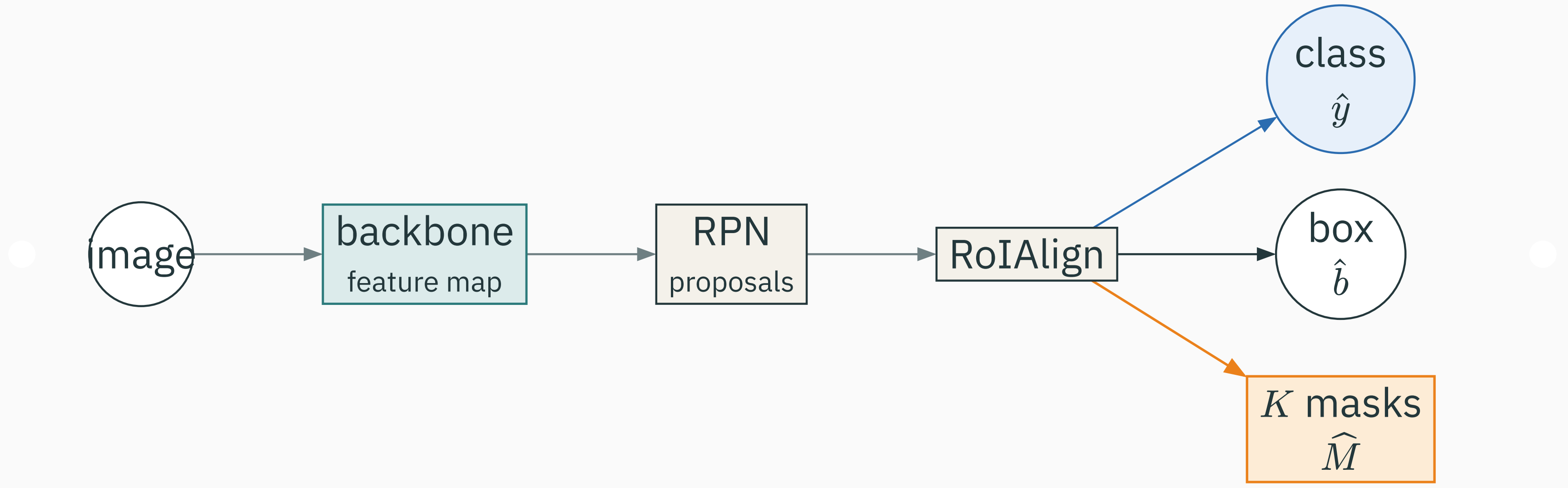
EVAL hard output ↔ truth; IoU / AP

$$x \to \left\{ \left(\hat{y}_j, \hat{b}_j, \hat{m}_j, \hat{s}_j \right) \right\}_{j=1}^{N}.$$

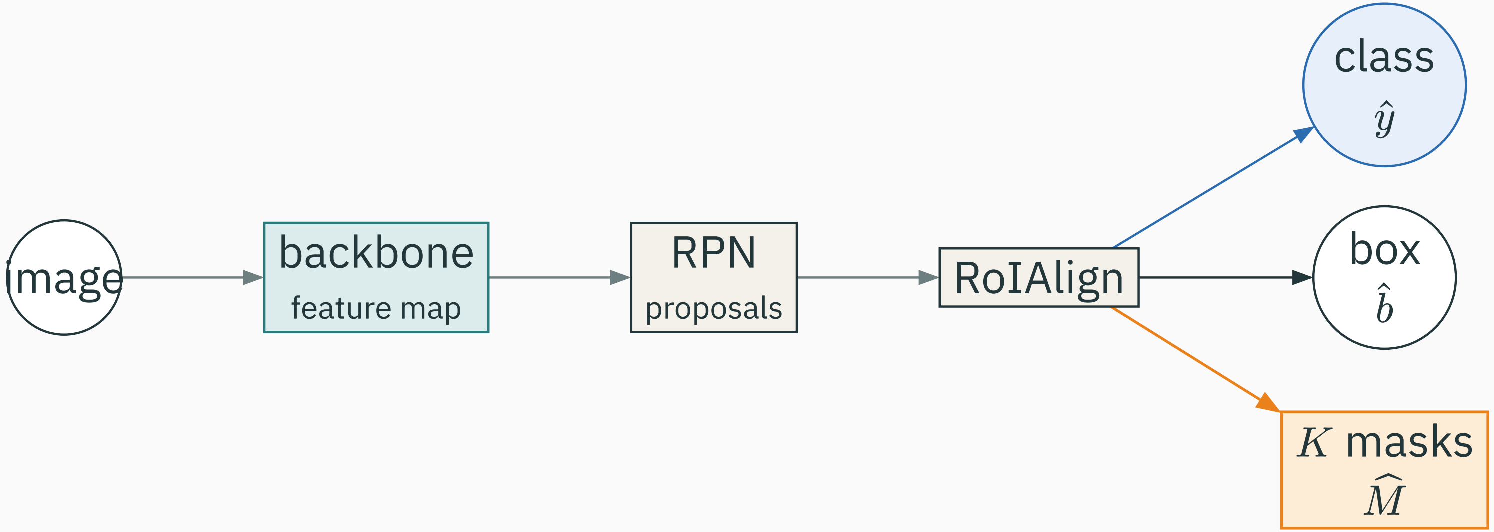
Symbol	Shape	Meaning
$\hat{y}_j$	scalar	object class
$\hat{b}_j$	four coordinates	box that places the mask
$\hat{m}_j$	$m \times m$ then resized	shape in the RoI frame
$\hat{s}_j$	scalar	ranking confidence

Mask R-CNN adds a mask branch parallel to class and box

V VISUAL

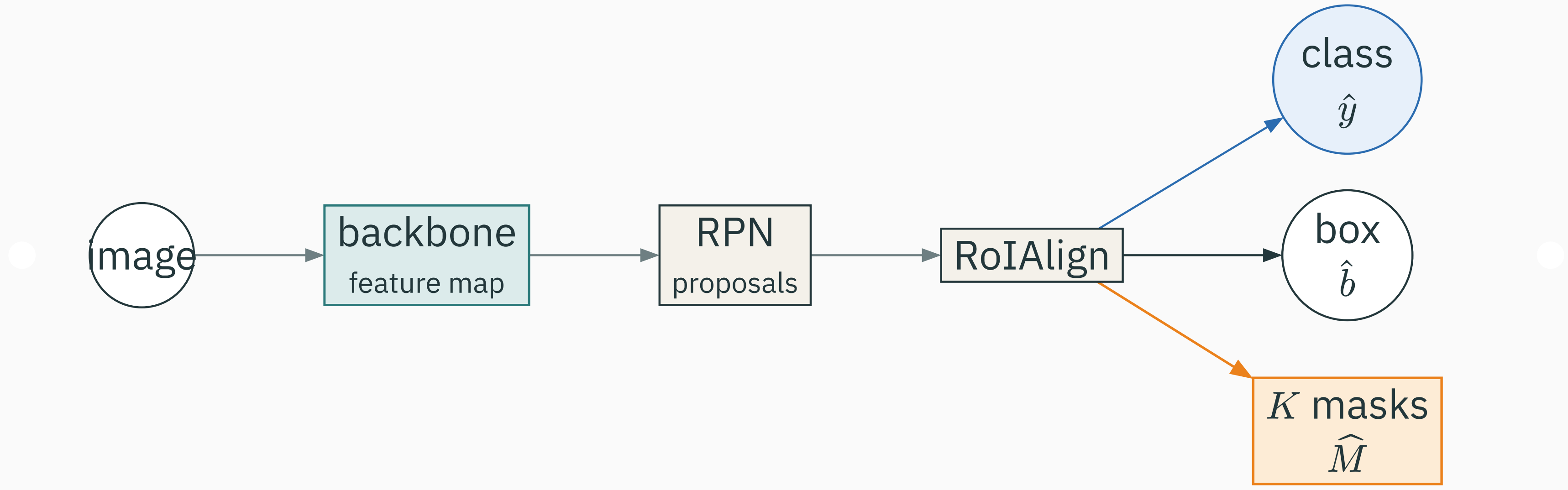


Mask R-CNN adds a mask branch parallel to class and box



RPN = Region Proposal Network; RoI = Region of Interest.

Mask R-CNN adds a mask branch parallel to class and box



RPN = Region Proposal Network; RoI = Region of Interest.

Detection proposes which object; the mask branch predicts which RoI pixels.

Separate the RPN loss from the second-stage RoI loss

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

STAGE 1 · RPN

$\mathcal{L}_{\text{RPN-obj}}$ on sampled anchors
$\mathcal{L}_{\text{RPN-box}}$ on positive anchors

STAGE 2 · SAMPLED ROIS

$\mathcal{L}_{\text{cls}}$ on positive + negative RoIs
$\mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$ on positive RoIs

Separate the RPN loss from the second-stage RoI loss

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

STAGE 1 · RPN

$\mathcal{L}_{\text{RPN-obj}}$ on sampled anchors
$\mathcal{L}_{\text{RPN-box}}$ on positive anchors

STAGE 2 · SAMPLED ROIS

$\mathcal{L}_{\text{cls}}$ on positive + negative RoIs
$\mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$ on positive RoIs

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RoI}}.$$

Separate the RPN loss from the second-stage RoI loss

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

STAGE 1 · RPN

$\mathcal{L}_{\text{RPN-obj}}$ on sampled anchors
$\mathcal{L}_{\text{RPN-box}}$ on positive anchors

STAGE 2 · SAMPLED ROIS

$\mathcal{L}_{\text{cls}}$ on positive + negative RoIs
$\mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$ on positive RoIs

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RoI}}.$$

The familiar three-term Mask R-CNN equation is the second-stage RoI loss, not the whole RPN-plus-RoI total.

Original Mask R-CNN predicts K independent binary masks

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\widehat{M} \in \mathbb{R}^{K\times m\times m}, \quad \widehat{M}_{k,u,v} = \sigma\big(a_{k,u,v}\big).$$

Original Mask R-CNN predicts K independent binary masks

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\widehat{M} \in \mathbb{R}^{K\times m\times m}, \quad \widehat{M}_{k,u,v} = \sigma\left(a_{k,u,v}\right).$$

For a training RoI with ground-truth class k, only channel k contributes to $\mathcal{L}_{\text{mask}}$.

Original Mask R-CNN predicts K independent binary masks

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

$$\widehat{M} \in \mathbb{R}^{K \times m \times m}, \quad \widehat{M}_{k,u,v} = \sigma(a_{k,u,v}).$$

For a training RoI with ground-truth class k, only channel k contributes to $\mathcal{L}_{\text{mask}}$.

CLASS HEAD

softmax chooses the object category

MASK CHANNEL k

sigmoid makes foreground/ background decisions inside that RoI

Original Mask R-CNN predicts K independent binary masks

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$\widehat{M} \in \mathbb{R}^{K\times m\times m}, \quad \widehat{M}_{k,u,v} = \sigma\big(a_{k,u,v}\big).$$

For a training RoI with ground-truth class k, only channel k contributes to $\mathcal{L}_{\text{mask}}$.

CLASS HEAD

softmax chooses the object category

MASK CHANNEL k

sigmoid makes foreground/ background decisions inside that RoI

> A class-agnostic one-mask head is a valid later variant; it is not the original K-mask formulation.

Calculate one selected-channel mask BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \widehat{M} = \begin{pmatrix} .8 & .6 \\ .3 & .1 \end{pmatrix}.$$

Calculate one selected-channel mask BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \widehat{M} = \begin{pmatrix} .8 & .6 \\ .3 & .1 \end{pmatrix}.$$

$$\mathcal{L}_{\text{mask}} = -\frac{1}{4}[\log .8 + \log .6 + \log(1 - .3) + \log(1 - .1)]$$

Calculate one selected-channel mask BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \widehat{M} = \begin{pmatrix} .8 & .6 \\ .3 & .1 \end{pmatrix}.$$

$$\mathcal{L}_{\text{mask}} = -\frac{1}{4}[\log .8 + \log .6 + \log(1 - .3) + \log(1 - .1)]$$

$$= \frac{1}{4}[.223 + .511 + .357 + .105] \approx .299.$$

Calculate one selected-channel mask BCE

D DERIVATION

TRAIN probability ↔ truth; CE / Dice | **INFER** probability → label / object masks | **EVAL** hard output ↔ truth; IoU / AP

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \widehat{M} = \begin{pmatrix} .8 & .6 \\ .3 & .1 \end{pmatrix}.$$

$$\mathcal{L}_{\text{mask}} = -\frac{1}{4}[\log .8 + \log .6 + \log(1 - .3) + \log(1 - .1)]$$

$$= \frac{1}{4}[.223 + .511 + .357 + .105] \approx .299.$$

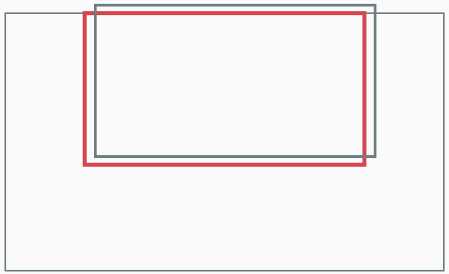
Pixels in this RoI and the selected ground-truth class channel define the mask loss.

RoIAlign reduces quantization error for boxes and especially masks

V VISUAL

ROIPOOL

round RoI and bin coordinates; convenient, but spatially quantized



ROIALIGN

keep fractional coordinates; sample neighboring features bilinearly

RoIAlign reduces quantization error for boxes and especially masks

V VISUAL

ROIPOOL

round RoI and bin coordinates; convenient, but spatially quantized

ROIALIGN

keep fractional coordinates; sample neighboring features bilinearly

> The Mask R-CNN paper reports box-AP gains too; the alignment benefit is larger for pixel-accurate masks.

Instance evaluation reuses L10 AP with mask IoU

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

score-sort predicted masks → same-class one-to-one match → **mask** IoU threshold → precision–recall mask AP

Instance evaluation reuses L10 AP with mask IoU

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

score-sort predicted masks → same-class one-to-one match → **mask** IoU threshold → precision–recall mask AP

Box AP asks whether rectangles overlap; mask AP asks whether object pixels overlap.

Policies, alignment, and panoptic output

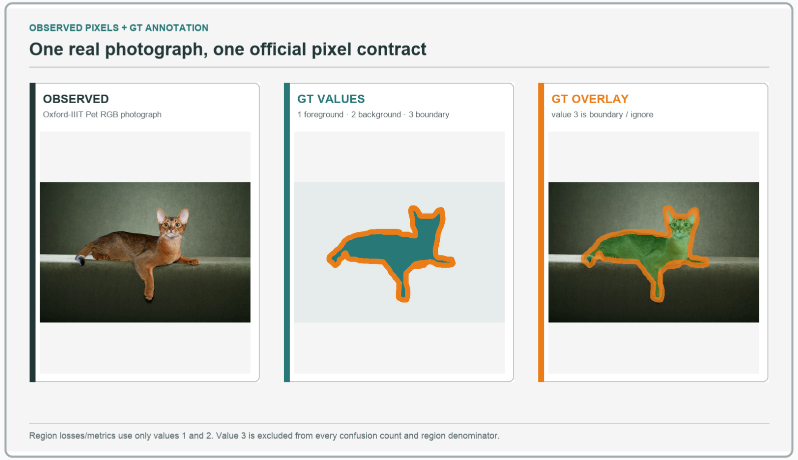
Void pixels leave both the loss sum and denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



Let V be the valid labeled pixels:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|V|}\sum_{i\in V}\log p_{i,y_i}.$$

Oxford values 1/2 are valid: $|V| = 221,704$. Value 3 contributes $18,296$ ignored boundary pixels.

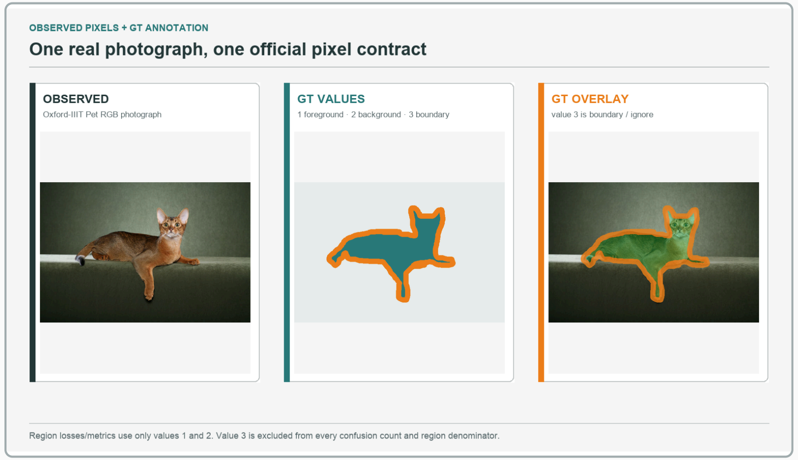
Void pixels leave both the loss sum and denominator

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP



Let V be the valid labeled pixels:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|V|}\sum_{i\in V}\log p_{i,y_i}.$$

Oxford values 1/2 are valid: $|V| = 221,704$. Value 3 contributes $18,296$ ignored boundary pixels.

Remove ignored pixels from CE, confusion counts, IoU, and Dice—not only from the numerator.

Empty masks and class averaging need a written policy

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

EMPTY–EMPTY

raw Dice is $\frac{0}{0}$; choose smoothing, exclusion, or a declared perfect score

MULTI-CLASS

compute per class; declare background inclusion, empty-class handling, and macro or frequency weighting

Empty masks and class averaging need a written policy

D DERIVATION

TRAIN probability ↔ truth; CE / Dice

INFER probability → label / object masks

EVAL hard output ↔ truth; IoU / AP

EMPTY–EMPTY

raw Dice is $\frac{0}{0}$; choose smoothing, exclusion, or a declared perfect score

MULTI-CLASS

compute per class; declare background inclusion, empty-class handling, and macro or frequency weighting

Also declare whether scores are averaged per image or accumulated from a dataset-level confusion matrix.

Empty masks and class averaging need a written policy

D DERIVATION

TRAIN probability ↔ truth; CE / Dice	**INFER** probability → label / object masks	**EVAL** hard output ↔ truth; IoU / AP

EMPTY–EMPTY

raw Dice is $\frac{0}{0}$; choose smoothing, exclusion, or a declared perfect score

MULTI-CLASS

compute per class; declare background inclusion, empty-class handling, and macro or frequency weighting

Also declare whether scores are averaged per image or accumulated from a dataset-level confusion matrix.

The reduction rule is part of the metric—not an implementation footnote.

Upsampling restores resolution by different mechanisms

V VISUAL

Method	What creates locations?	What learns?
nearest / bilinear	fixed interpolation	following convolution refines
transposed convolution	learned overlap pattern	kernel and mixing jointly
skip + resize	decoder plus aligned encoder detail	fusion after concatenation

Upsampling restores resolution by different mechanisms

V VISUAL

Method	What creates locations?	What learns?
nearest / bilinear	fixed interpolation	following convolution refines
transposed convolution	learned overlap pattern	kernel and mixing jointly
skip + resize	decoder plus aligned encoder detail	fusion after concatenation

Every route needs evidence to refine the larger grid; interpolation alone cannot reverse pooling.

Bilinear interpolation is a weighted four-neighbor sum

D DERIVATION

Suppose a RoIAlign sample lies halfway in both directions between

$$\begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix}.$$

Bilinear interpolation is a weighted four-neighbor sum

D DERIVATION

Suppose a RoIAlign sample lies halfway in both directions between

$$\begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix}.$$

All four weights equal $.25$:

$$f = .25(1) + .25(3) + .25(5) + .25(7) = 4.$$

Bilinear interpolation is a weighted four-neighbor sum

D DERIVATION

Suppose a RoIAlign sample lies halfway in both directions between

$$\begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix}.$$

All four weights equal $.25$:

$$f = .25(1) + .25(3) + .25(5) + .25(7) = 4.$$

Moving the fractional sample changes the weights smoothly; no coordinate is rounded.

Bilinear interpolation is a weighted four-neighbor sum

D DERIVATION

Suppose a RoIAlign sample lies halfway in both directions between

$$\begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix}.$$

All four weights equal $.25$:

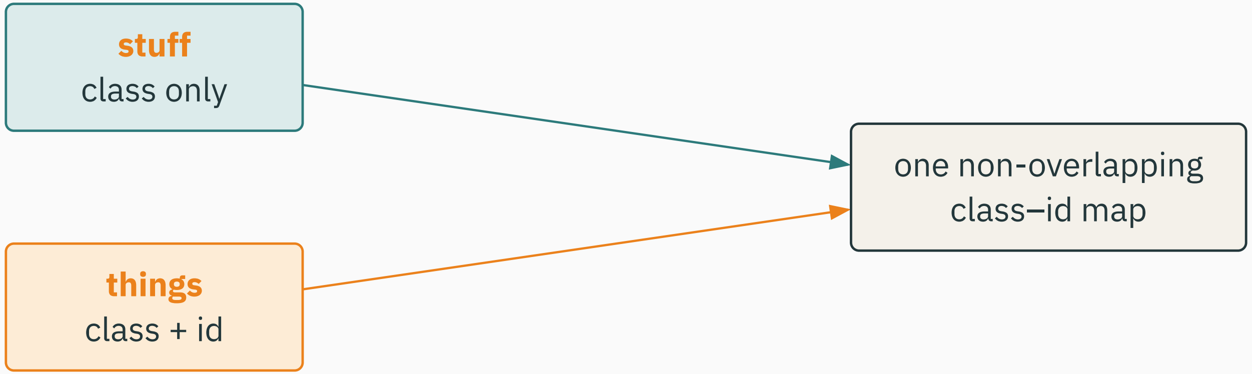
$$f = .25(1) + .25(3) + .25(5) + .25(7) = 4.$$

Moving the fractional sample changes the weights smoothly; no coordinate is rounded.

Precise sampling preserves alignment; it does not invent image detail.

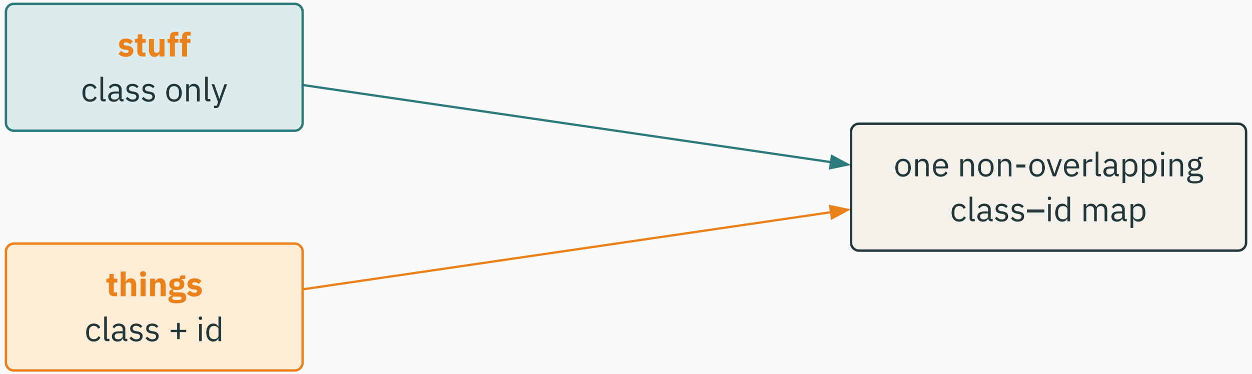
Panoptic output assigns one class–identity record per pixel

V VISUAL



Panoptic output assigns one class–identity record per pixel

V VISUAL



Overlapping instance hypotheses must be fused; datasets may reserve a void label for uncertain pixels.

Choose the output and metric from the question

D DERIVATION

Need	Output	Evaluation
road area	$H \times W$ semantic classes	per-class IoU / mIoU
count touching trees	scored set of instance masks	mask AP under mask IoU
whole labeled scene	$H \times W$ class–id map	PQ or declared panoptic protocol

Choose the output and metric from the question

D DERIVATION

Need	Output	Evaluation
road area	$H \times W$ semantic classes	per-class IoU / mIoU
count touching trees	scored set of instance masks	mask AP under mask IoU
whole labeled scene	$H \times W$ class–id map	PQ or declared panoptic protocol

Architecture names follow the output contract; they do not define the task.

Primary papers establish the architectural claims

V VISUAL

FCN + U-NET

Long et al. · pixels-to-pixels; Ronneberger et al. · context plus localization

V-NET + MASK R-CNN

Milletari et al. · Dice objective; He et al. · mask branch + RoIAlign

PANOPTIC + COMPANION

Kirillov et al. · stuff + things; exact L10 Colab · toy and real ledgers

REAL EVIDENCE

Oxford-IIIT Pet · Abyssinian_1 · official trimap · CC BY-SA 4.0

REPRODUCIBILITY

`shared/vision-evidence/oxford-iiit-pet/l10/` · builder + manifest + arrays + metric ledger

HONEST CLAIM

model output = none · measured performance = none · +16 px probability field = constructed

OBSERVED = photo · GT = trimap · CONSTRUCTED = orchard / toy p / shifted p · COMPUTED = masks, errors, metrics

Variants change mechanisms, not the contract

V VISUAL

Variant	What changes	What stays explicit
class-agnostic mask head	$1 \times m \times m$ rather than $K \times m \times m$	object class comes from class head
dilated / pyramid decoder	context without identical downsampling	output grid and pixel loss
query-based mask sets	learned queries replace RPN RoIs	one-to-one identity and mask metric
panoptic fusion	resolve overlaps and stuff regions	one class–identity assignment per pixel

Variants change mechanisms, not the contract

V VISUAL

Variant	What changes	What stays explicit
class-agnostic mask head	$1 \times m \times m$ rather than $K \times m \times m$	object class comes from class head
dilated / pyramid decoder	context without identical downsampling	output grid and pixel loss
query-based mask sets	learned queries replace RPN RoIs	one-to-one identity and mask metric
panoptic fusion	resolve overlaps and stuff regions	one class–identity assignment per pixel

> Do not turn variants into a catalogue: ask which output, alignment, loss, or matching assumption changed.

Revisit the orchard commitment before revealing the answer

Q QUESTION

HELD CASE · SAME GENERATED CONCEPT Report total canopy area **and** count two touching trees.

A	B	C	D
global class	class + box	semantic map	separate object masks

Revisit the orchard commitment before revealing the answer

Q QUESTION

HELD CASE · SAME GENERATED CONCEPT Report total canopy area **and** count two touching trees.

A	B	C	D
global class	class + box	semantic map	separate object masks

D. Instance masks preserve the two identities; their union gives total area. If every background/stuff pixel also needs a class, use a panoptic class–identity map.

Final checklist and public L12 handoff

V VISUAL

SEGMENTATION CHECKLIST

1. What is the output structure?
2. Which axis holds classes?
3. What enters TRAIN loss?
4. Which INFER threshold/fusion applies?
5. Which EVAL classes, masks, and reductions define the score?

PUBLIC L12 · NEXT TOKEN

Pixel prediction shared one classifier over spatial positions (h, w).
Next-token prediction shares one classifier over sequence positions t.
Pixel classes become vocabulary tokens; **causality** becomes the new constraint.

Final checklist and public L12 handoff

V VISUAL

SEGMENTATION CHECKLIST

1. What is the output structure?
2. Which axis holds classes?
3. What enters TRAIN loss?
4. Which INFER threshold/fusion applies?
5. Which EVAL classes, masks, and reductions define the score?

PUBLIC L12 · NEXT TOKEN

Pixel prediction shared one classifier over spatial positions (h, w).
Next-token prediction shares one classifier over sequence positions t.
Pixel classes become vocabulary tokens; **causality** becomes the new constraint.

A dense model is trustworthy only when output support, three clocks, alignment, and metric protocol agree.