Deep Learning Frameworks

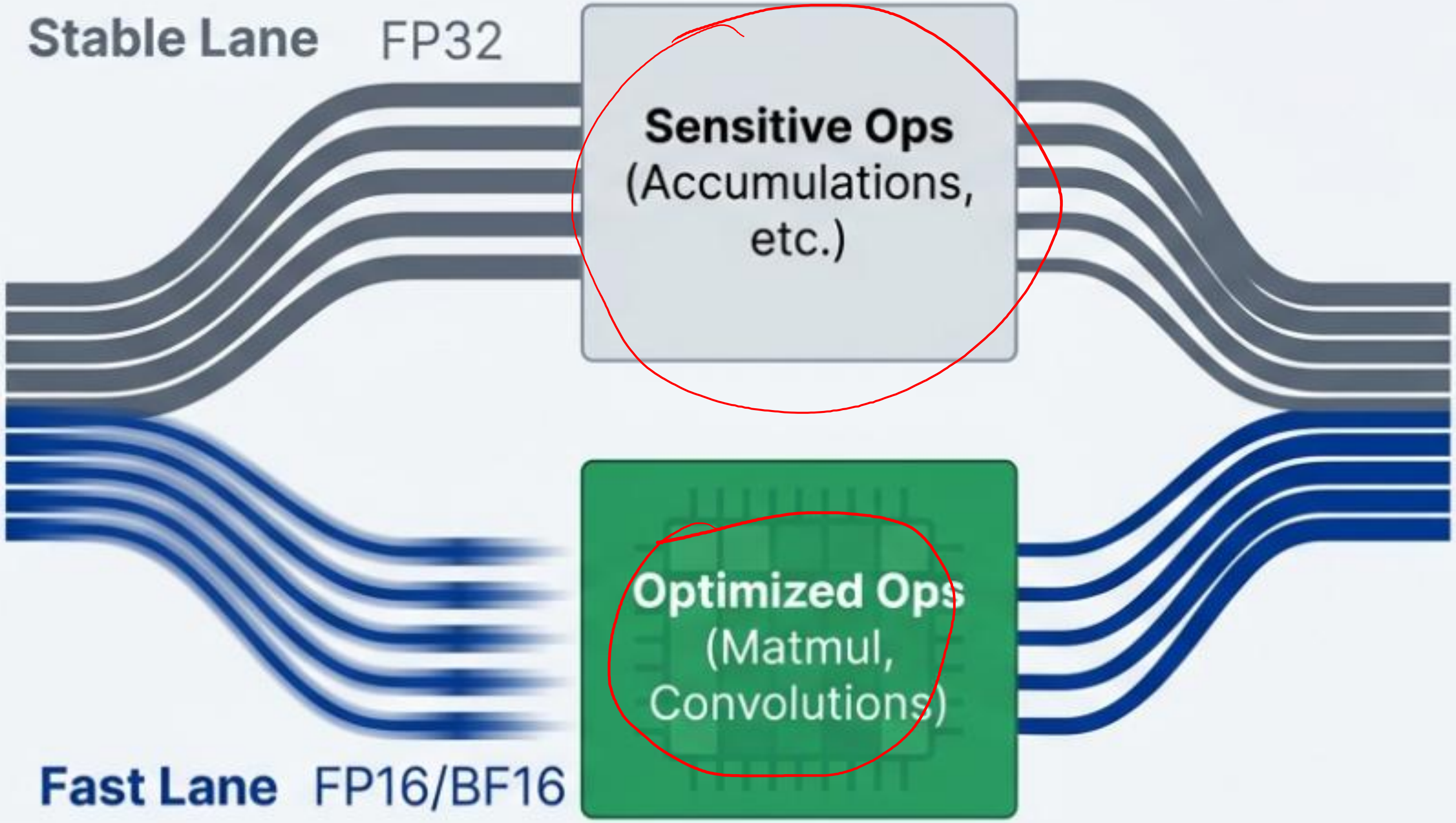
Segmentation – UNet, Deeplabv3

Mixed Precision Training- Autograd, GradScaler

https://tinyurl.com/dlframeworks

https://github.com/sakharamg/DeepLearningFrameworks

Mixed Precision Training



- **Faster Training:** Modern GPUs (like those with Tensor Cores) are specifically optimized for FP16/BF16 matrix multiplies and convolutions.
- **Lower Memory Usage:** The biggest win is often in **activation memory** (intermediate tensors from the forward pass), allowing for larger models or bigger batch sizes.

Code

```
# 1. Initialize the GradScaler
scaler = torch.amp.GradScaler("cuda")

# --- Inside your training loop ---
optimizer.zero_grad(set_to_none=True)

# 2. Forward pass with autocast
with torch.autocast(device_type="cuda", dtype=torch.float16):
    out = model(x)
    loss = criterion(out, y)

# 3. Scale the loss and call backward()
scaler.scale(loss).backward()

# 4. Unscale gradients and call optimizer.step()
scaler.step(optimizer)

# 5. Update the scale factor for the next iteration
scaler.update()
```

bfloat16

Initialize the **scaler** once, outside the loop.

Enables **mixed precision** for the forward pass.

Scales loss to prevent gradient underflow during backprop.

Unscales gradients and steps the optimizer. Skips if grads are invalid.

Prepares scaler for the next iteration by adjusting the scale factor.

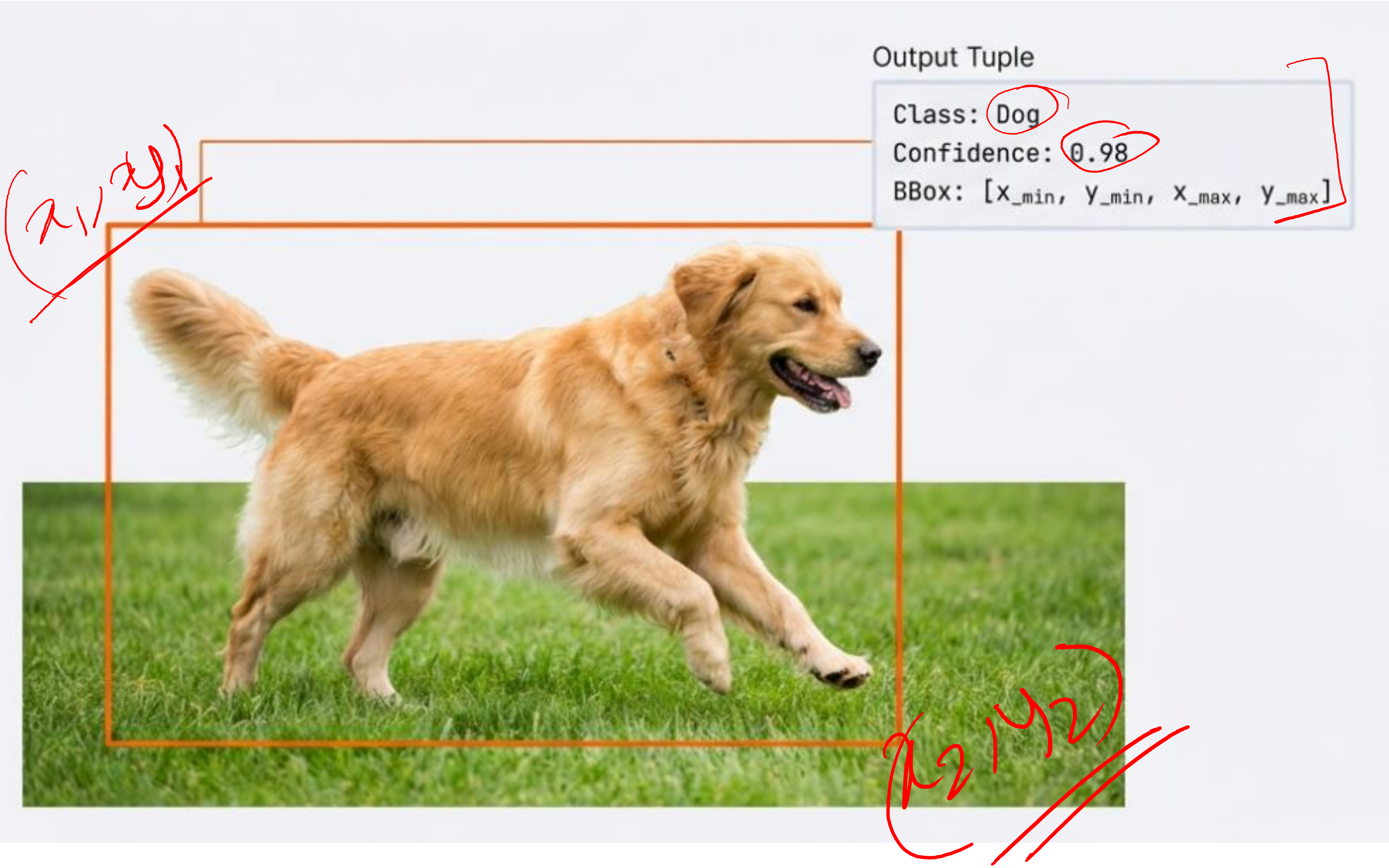
Floating Point Formats

Type	Total bits	Sign	Exponent	Mantissa (fraction)
float32 (FP32 / binary32)	32	1	8	23
float16 (FP16 / binary16)	16	1	5	10
bfloat16 (BF16)	16	1	8	7

A6000 A100, H100

https://www.geeksforgeeks.org/computer-organization-architecture/ieee-standard-754-floating-point-numbers/

Recap: Object Detection



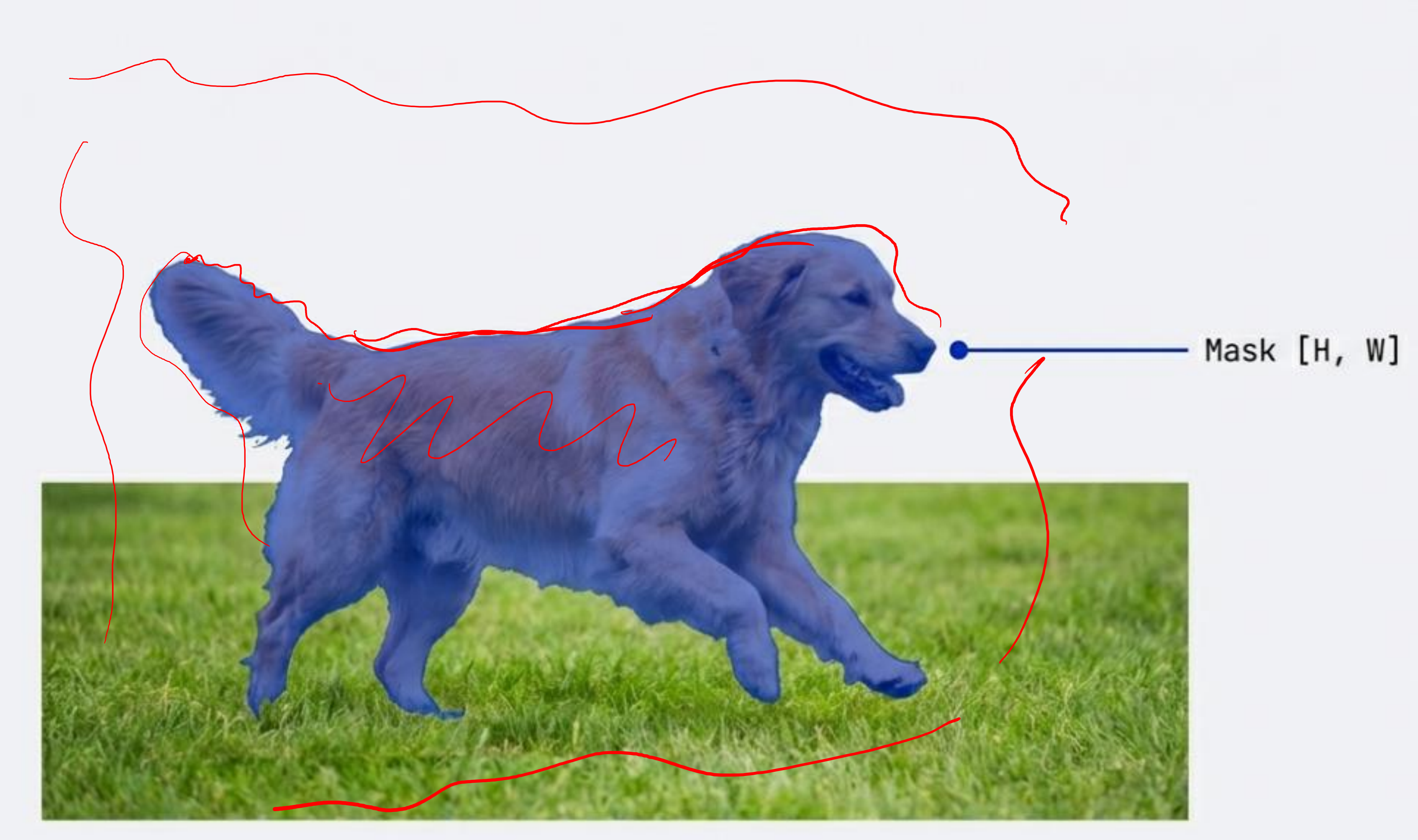
Limitations of Object Detection

Use Cases (What it's good at)

Limitations (What it misses)

- \- Cannot handle irregular shapes or thin objects (like hair or wire).
- \- Cannot separate foreground from background inside the box.
- \- Lacks pixel-level scene understanding (e.g., distinguishing road vs. sidewalk).

Segmentation



Core Concept: Assigning a **semantic label** to every single pixel in the image, rather than generating a list of boxes.

Segmentation vs Object Detection

Segmentation

0, 1 → DOG

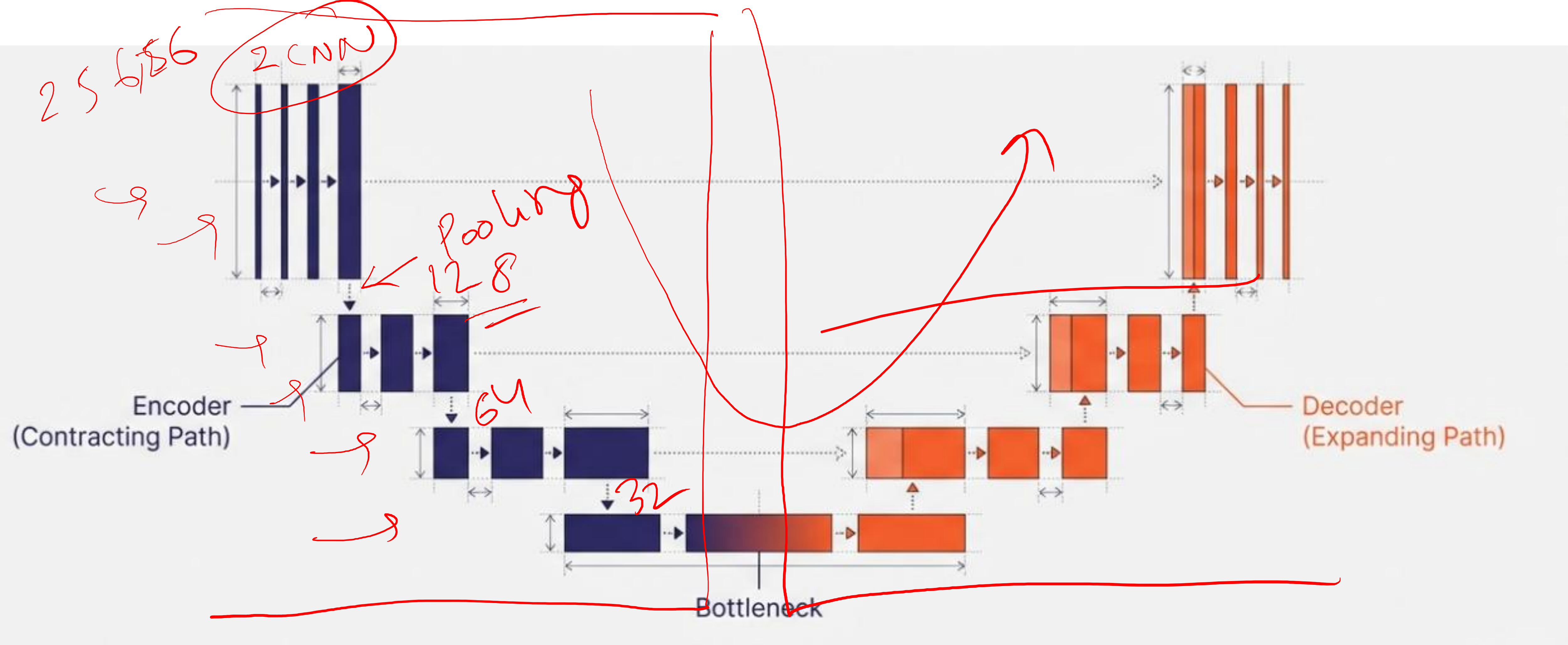
R, G, B

- Input: [3,H,W]
- Output: [C,H,W]   C → # classes
- Loss: Cross Entropy

0, 1, 2 ← CAT

DOG

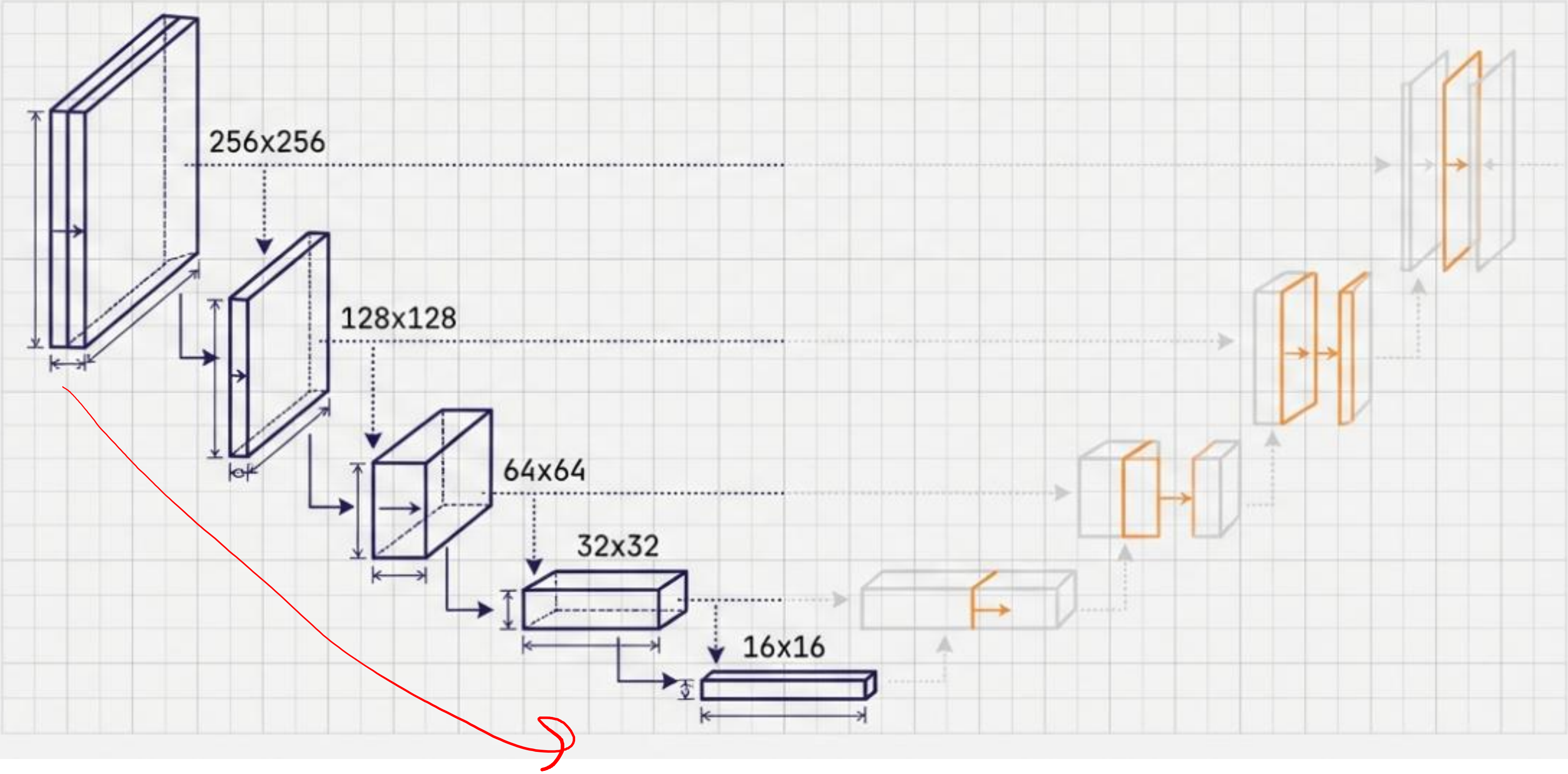
UNet



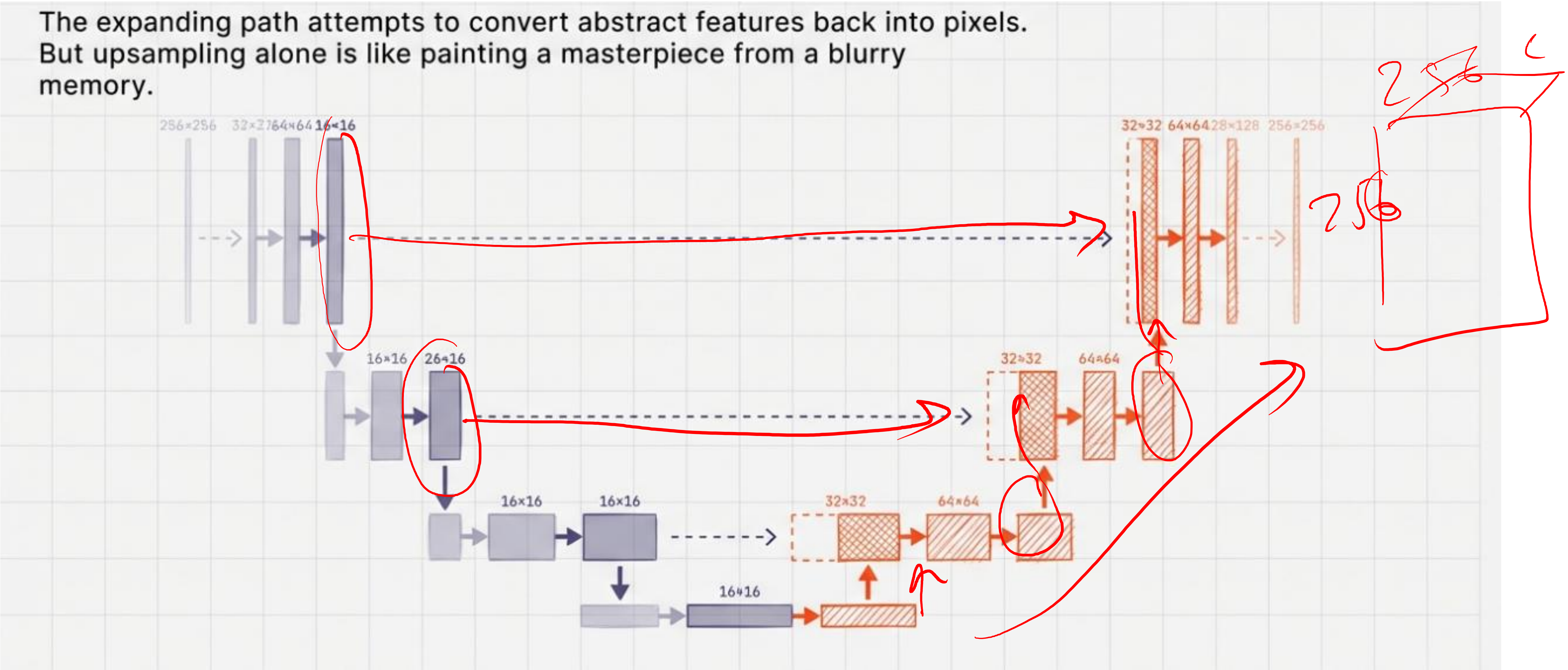
2 CNN
256
Pooling
128
64
32
Encoder
(Contracting Path)
Decoder
(Expanding Path)
Bottleneck

UNet - Encoder

```
The contracting path shrinks the image to increase the receptive field.
The model stops looking at pixels and starts seeing patterns.
```



UNet - Decoder



Up Sampling: Interpolation

1	1
1	1

Pooled feature map

0.96	2.0	0
0.96	2.0	0
0.78	1.5	0

3x3

Upsample (Nearest Neighbor)

0.96	0.96	2.0	2.0		
0.96	0.96	2.0	2.0		

6x6

Up Sampling: Interpolation

Pooled feature map

0.96	2.0	0
0.96	2.0	0
0.78	1.5	0

Upsample (Nearest Neighbor)

0.96	0.96				
0.96	0.96				

Up Sampling: Interpolation

Pooled feature map

0.96	2.0	0
0.96	2.0	0
0.78	1.5	0

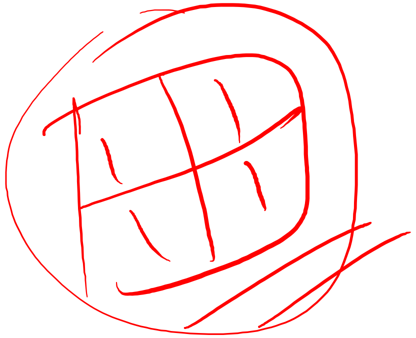
Upsample (Nearest Neighbor)

0.96	0.96	2.0	2.0	0	0
0.96	0.96	2.0	2.0	0	0
0.96	0.96	2.0	2.0	0	0
0.96	0.96	2.0	2.0	0	0
0.78	0.78	1.5	1.5	0	0
0.78	0.78	1.5	1.5	0	0

Up Sampling: Transposed Convolution

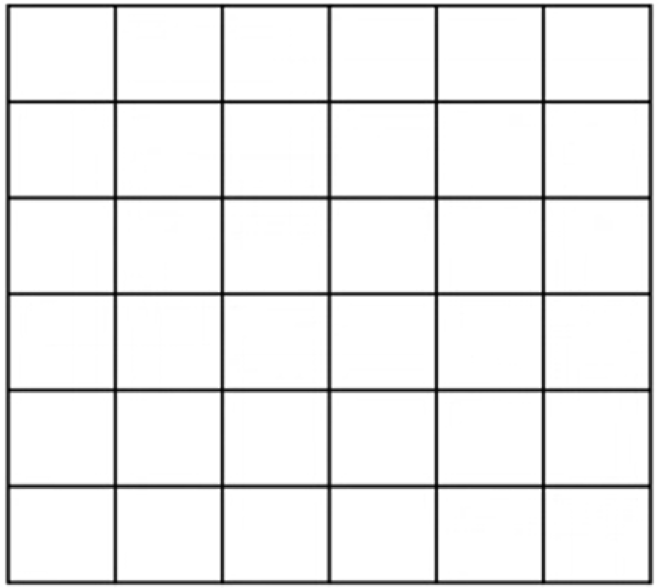
Pooled feature map

4.23	4.03	0.65
4.23	4.03	0.65
2.40	2.67	0.65



0.84	-1.0
0.94	-0.65

Upsample (Transposed convolution)



Up Sampling: Transposed Convolution

Upsample (Transposed convolution)

4.98	-2.8				
5.40	-1.3				

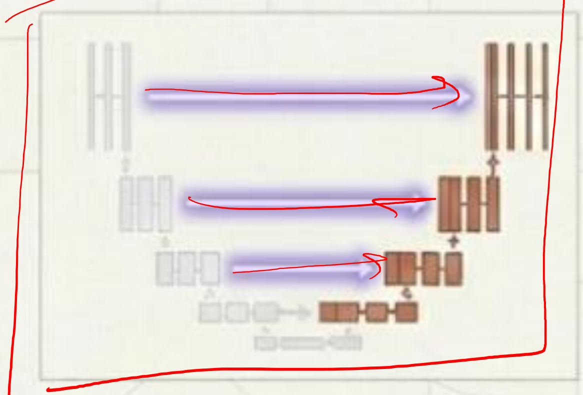
Pooled feature map

4.23	4.03	0.65
4.23	4.03	0.65
2.40	2.67	0.65

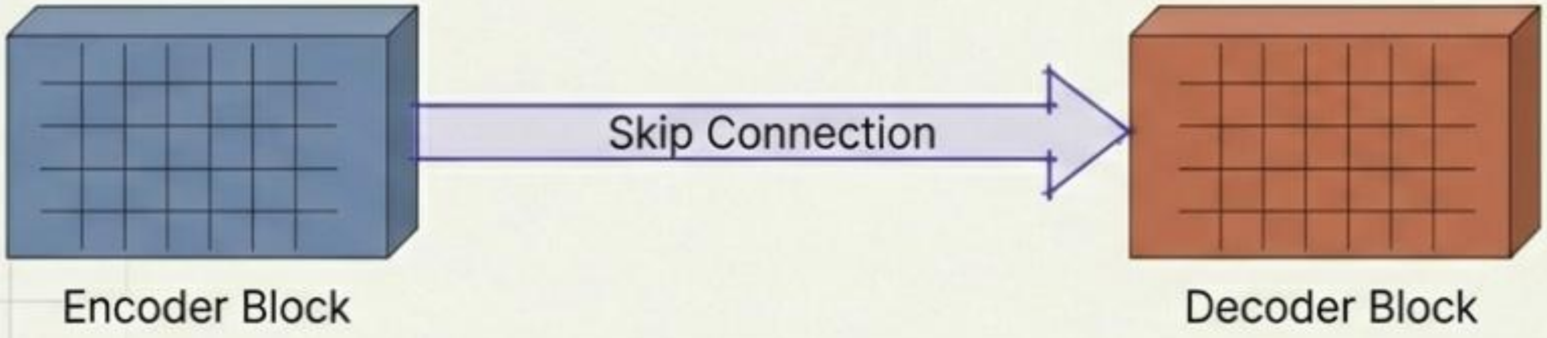
0.84	-1.0
0.94	-0.65

Bias = 1.43

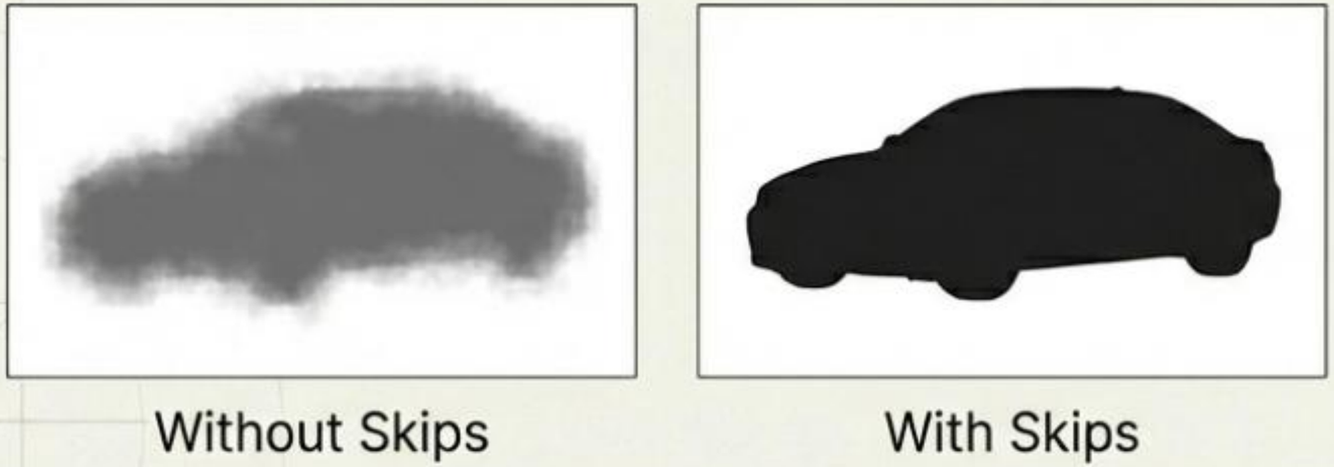
Skip Connections bridge the gap between “What” and “Where”



Horizontal Bridge



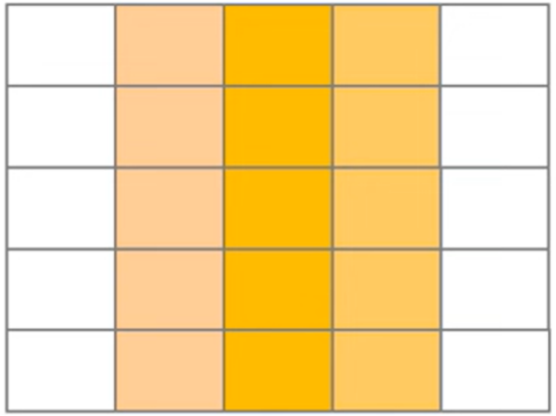
DECODER OUTPUT



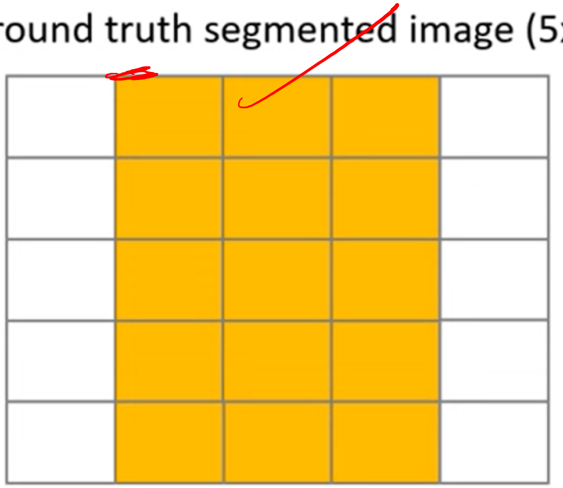
- **Skip Connections** copy high-resolution feature maps from the Encoder and **CONCATENATE** them to the Decoder.
- **Effect:** The decoder receives a spatial ‘tracer sketch’ (edges, boundaries) from the encoder to guide its upsampling.

Intersection Over Union

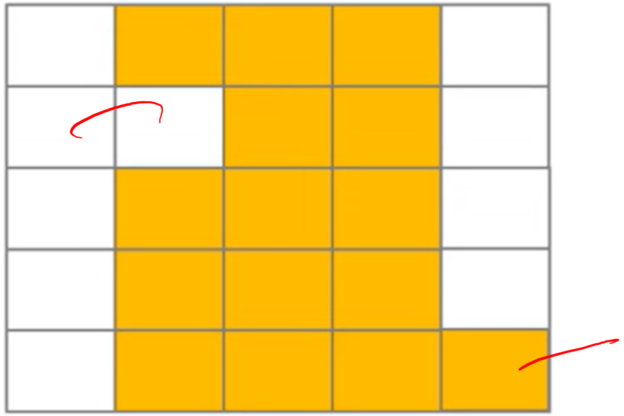
Image (5x5)



Ground truth segmented image (5x5)



Predicted segmentation image (5x5)



$$\boxed{\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{Intersection}}{\text{Union}}} = \frac{14}{14 + 1 + 1} = 0.875$$

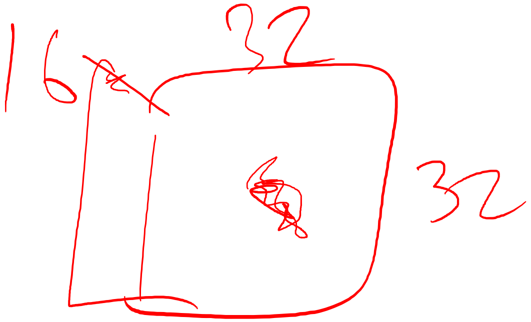
16

Lab: Segmentation using UNet

https://tinyurl.com/dlframeworks

https://github.com/sakharamg/DeepLearningFrameworks

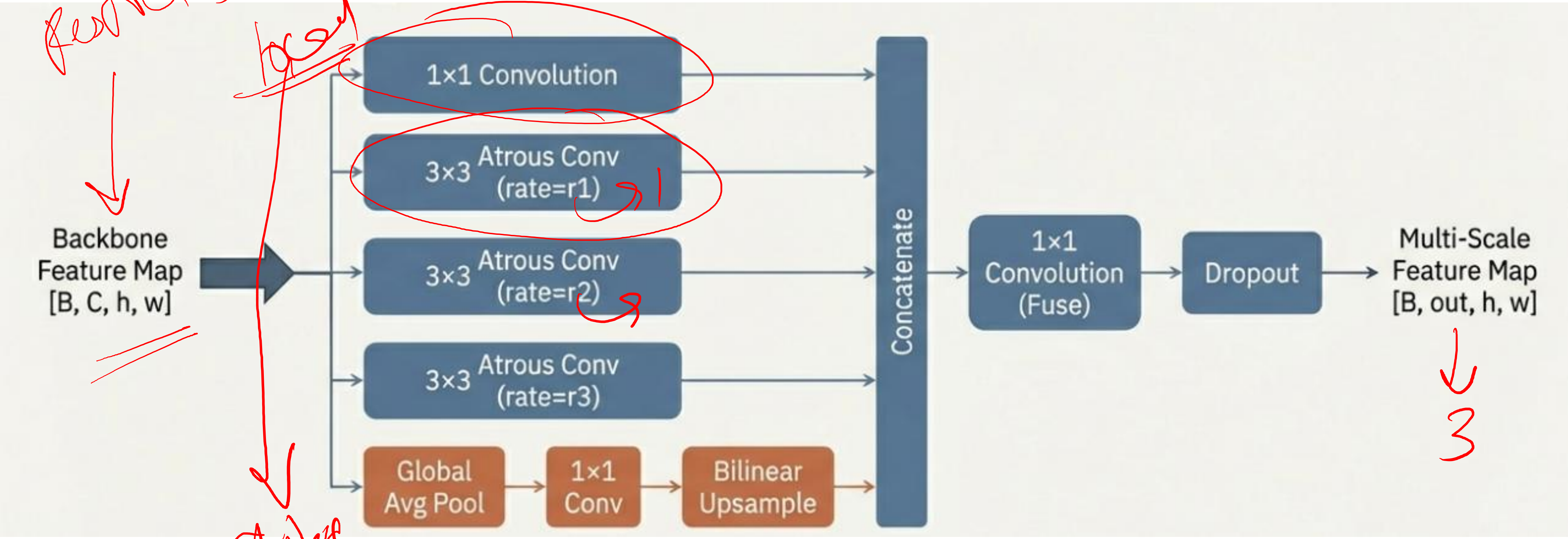
Deep Lab v3



16 x 1 x 1

Resnet 50

local

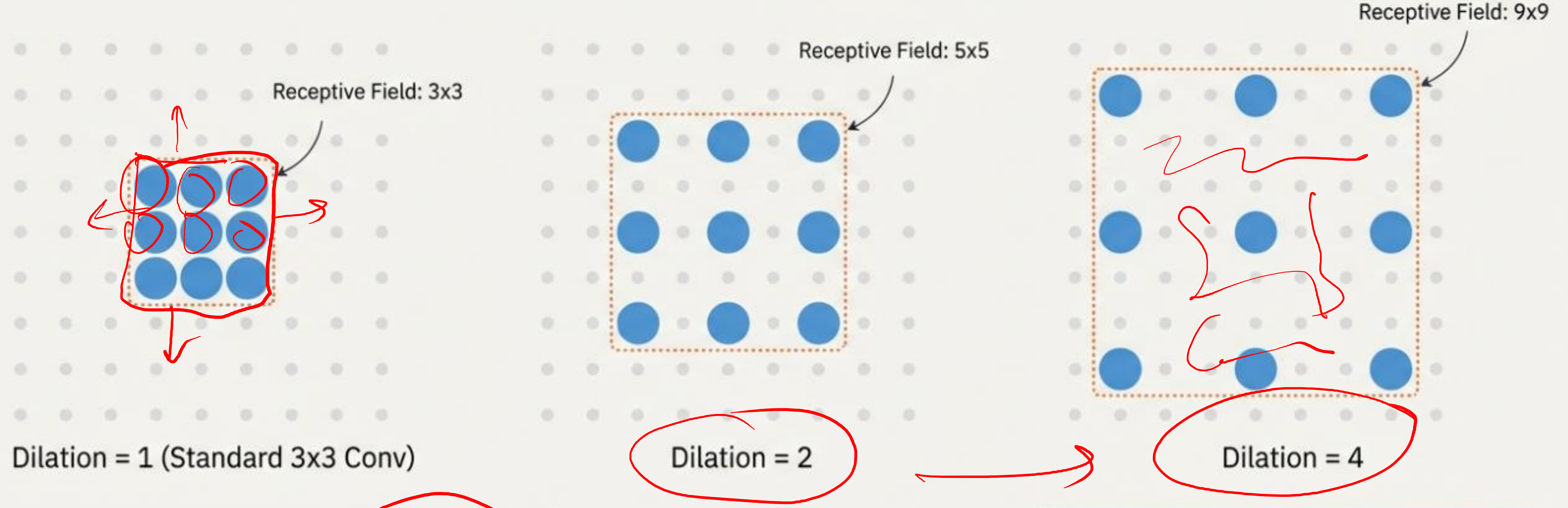


global

3x3

Atrous or Dilation Convolution

An atrous convolution introduces a "dilation rate" (d) that defines the spacing between kernel points.
It effectively "inflates" the kernel's view without adding parameters.



Effective Kernel Size: $k\_eff = k + (k-1)(d-1)$

Bigger receptive field with zero extra parameters and zero downsampling.

Lab: Segmentation using DeepLabv3

https://tinyurl.com/dlframeworks
https://github.com/sakharamg/DeepLearningFrameworks

Thank You