
Image Segmentation with U-Net

ECE 57000 — Fall 2026

David I. Inouye

| Segmentation needs a prediction at every image location

Arc 5: CNN fundamentals → evidence → ResNet classification → **U-Net segmentation**.

Task → Upsampling → Architecture

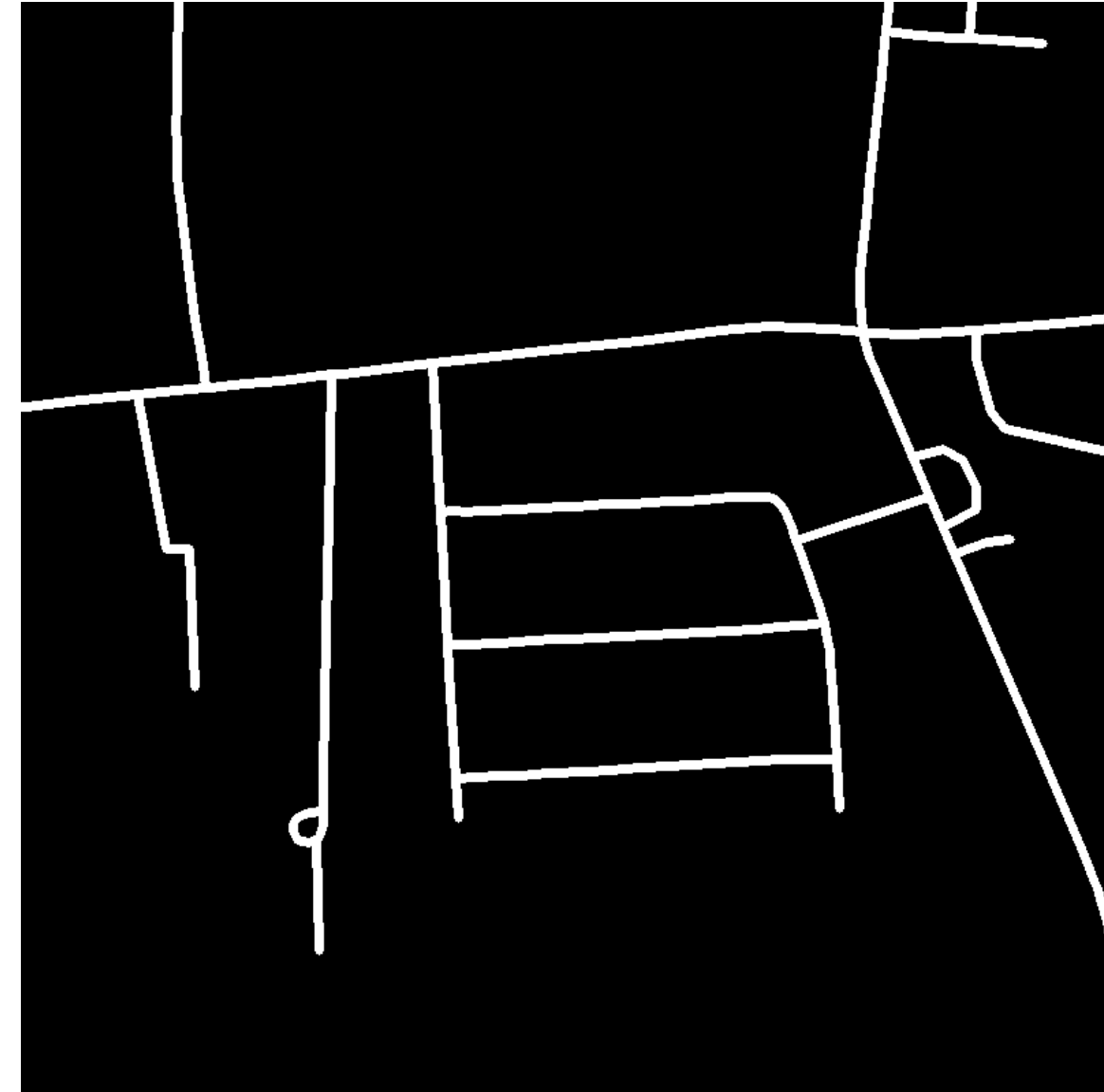
A classifier summarizes an image. A segmentation model must retain where each class occurs.

| A road map needs a **label** at every pixel

Input: aerial RGB image



Target: road / background



Provided annotation, not a model prediction • Massachusetts Roads, Mnih (2013)

Mnih, Massachusetts Roads (2013), validation tile 10228690_15; aligned crop.

| The input can stay the same while **the output becomes spatial**

	Image classification	Semantic segmentation
RGB input	$[B, 3, H, W]$	$[B, 3, H, W]$
Model logits	$[B, K]$	$[B, K, H, W]$
Integer targets	$[B]$	$[B, H, W]$

For roads: **two logits per pixel**, road or background.

Reuse cross-entropy, averaged over labeled pixels. **Keep the spatial grid instead of globally pooling it away.**

| Transposed convolution uses **the transpose of the linear operator**

Convolution: 6 inputs $\rightarrow$ 2 outputs

Transpose: 2 inputs $\rightarrow$ 6 outputs

a	b	0	c	d	0
0	a	b	0	c	d

A_K

rows become columns



a	0
b	a
0	b
c	0
d	c
0	d

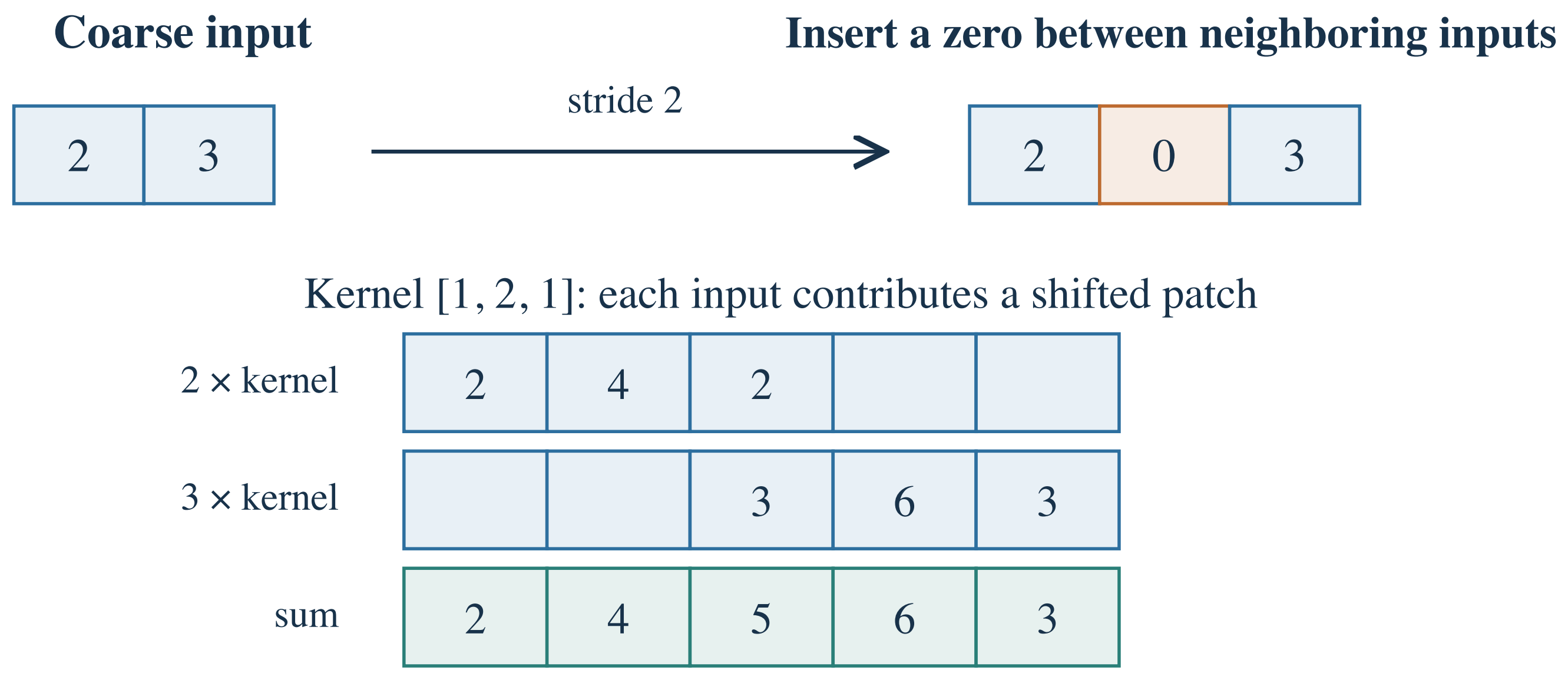
A_K^T

Collect weighted sums $\rightarrow$ scatter weighted patches; overlaps add.

Transpose is not inverse: increasing size does not undo information loss.

$$y = A_K x \quad \longrightarrow \quad u = A_K^T z$$

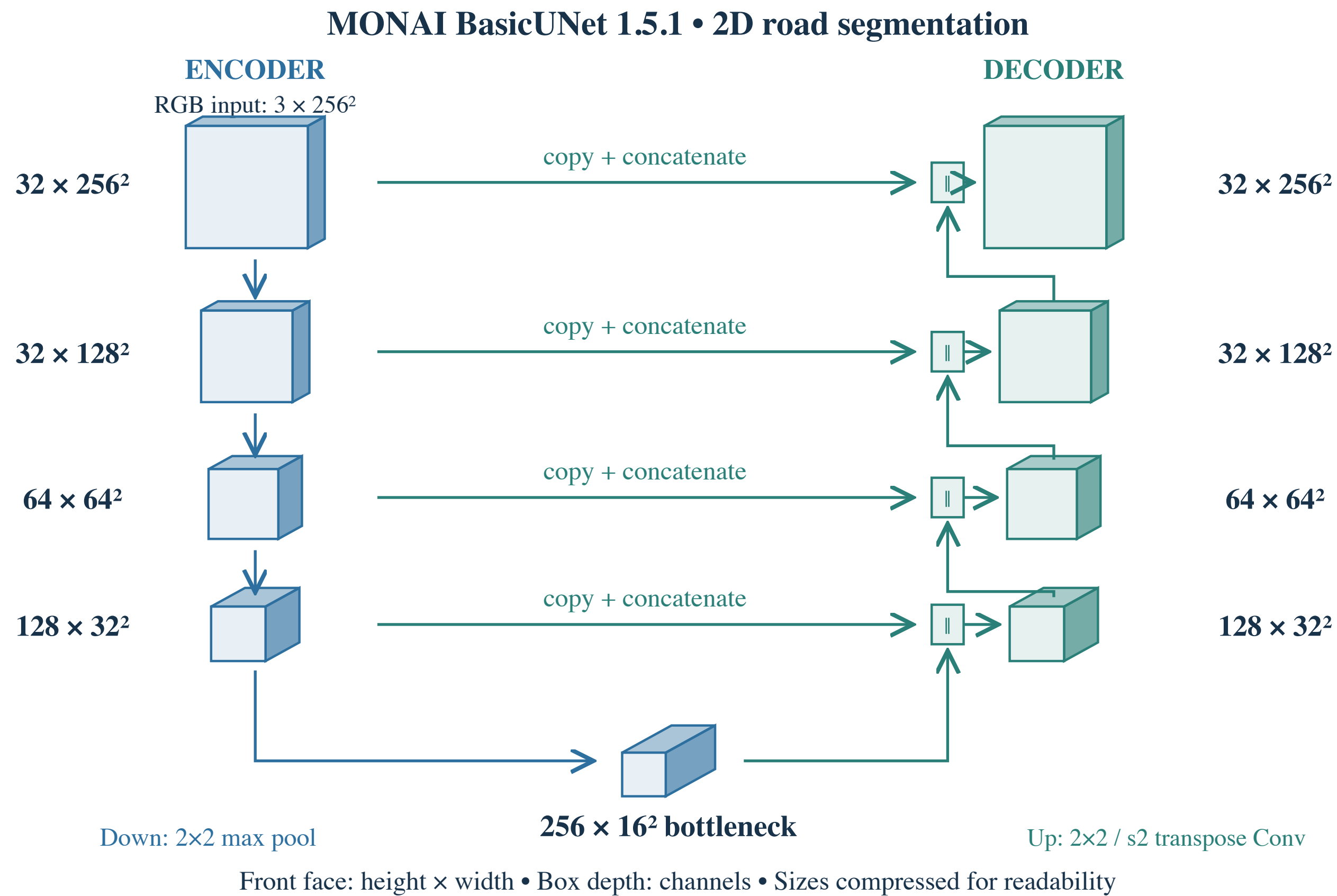
Fractional-stride intuition explains learned spatial expansion



The overlap adds: $2 + 3 = 5$. The fine-grid values are learned combinations.

In 2D, `ConvTranspose2d(16, 8, 2, stride=2)` maps $[B, 16, 16, 16]$ to $[B, 8, 32, 32]$.

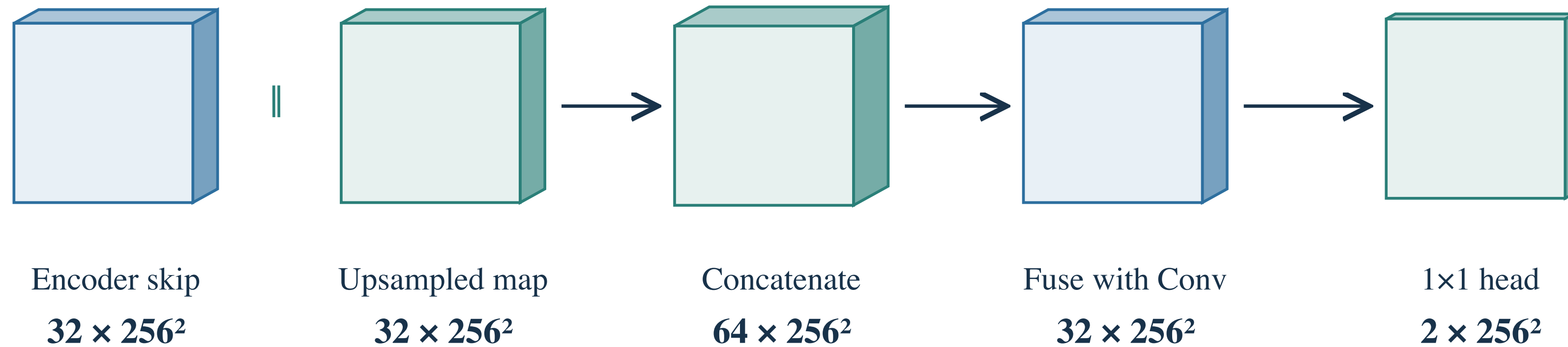
| U-Net combines coarse context with fine-resolution features



Original illustration of MONAI BasicUNet 1.5.1, 2D configuration; shapes are channels $\times$ spatial size.

| The U-Net head keeps **one score vector at every pixel**

Last decoder stage: preserve locations, combine channels



Concatenation doubles channels; the height and width stay fixed.

Front face: height $\times$ width • Box depth: channels • Sizes compressed for readability

| The task determines **what the architecture must preserve**

- **Classification:** aggregate spatial features into an image-level decision.
- **Segmentation:** combine broad context with location-specific detail.
- **Both:** architectural choices need evidence on the intended data and task.

We can now read complete architectures from familiar operations. **Knowing the pieces does not show that a model solves the application.**