

STRUCTURED OUT PREDICTION (SEMANTIC SEGMENTATION)

HYUN SOO PARK



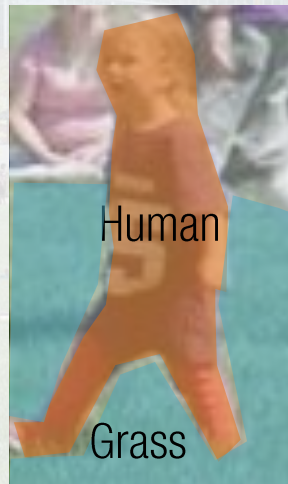
CHALLENGES OF VISUAL RECOGNITION

- Appearance
 - DOF: texture, illumination, material, shading, ...
- Shape
 - DOF: object category, geometric pose, viewpoint, ...



Human

$$f(I) = l_{human}$$



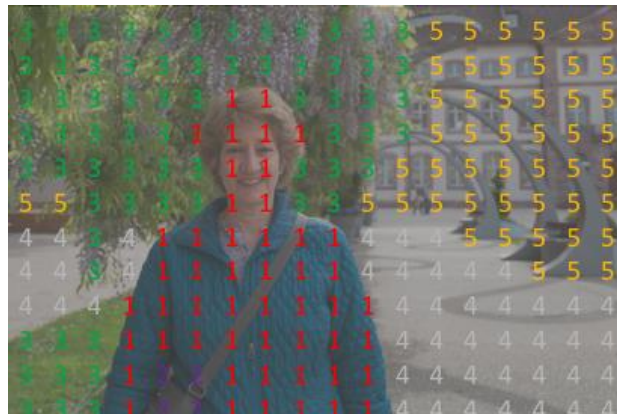
Human

Grass

Semantic segmentation



SEMANTIC SEGMENTATION: PIXEL CLASSIFICATION



- 0: Background/Unknown
- 1: Person
- 2: Purse
- 3: Plants/Grass
- 4: Sidewalk
- 5: Building/Structures

SEMANTIC SEGMENTATION FORMULATION



Unsupervised superpixel segmentation

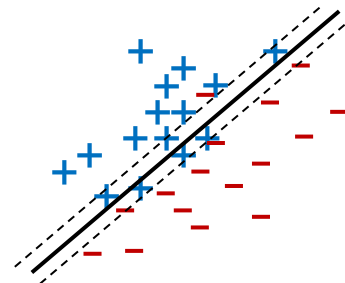


x_i

Visual feature

- Color histogram
- BoW
- SIFT
- HOG

Classification



$$L = \phi(x_i) \quad \text{e.g., } \phi(x_i) = w_{tree} \cdot x_i$$

SEMANTIC SEGMENTATION FORMULATION



Unsupervised superpixel segmentation



x_i

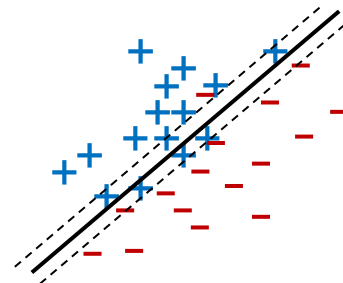


x_j

Visual feature

- Color histogram
- BoW
- SIFT
- HOG

Classification

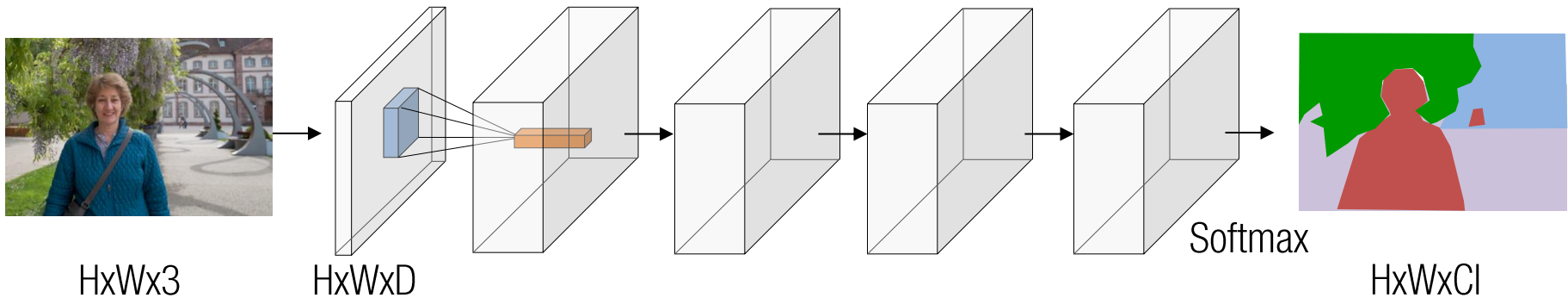


$$L = \phi(x_i) + \underbrace{\varphi(x_i, x_j)}_{\text{Context}}$$

Context

CRF: Conditional Random Field
aka. joint classifications (structured pred.)

HOLISTIC PREDICTION VIA DEEP LEARNING



$H \times W \times 3$

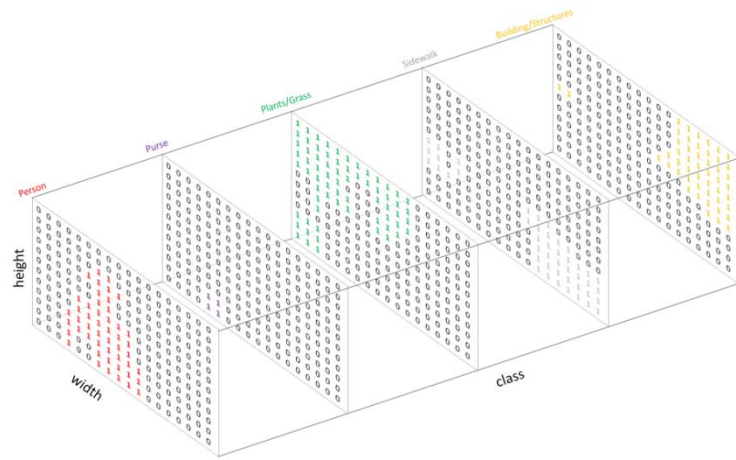
$H \times W \times D$

Fully convolutional layers

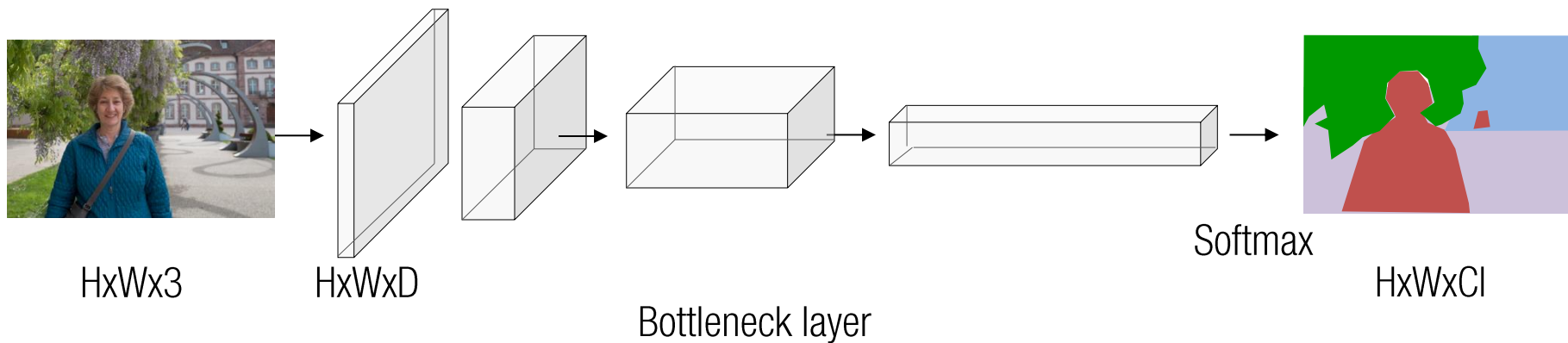
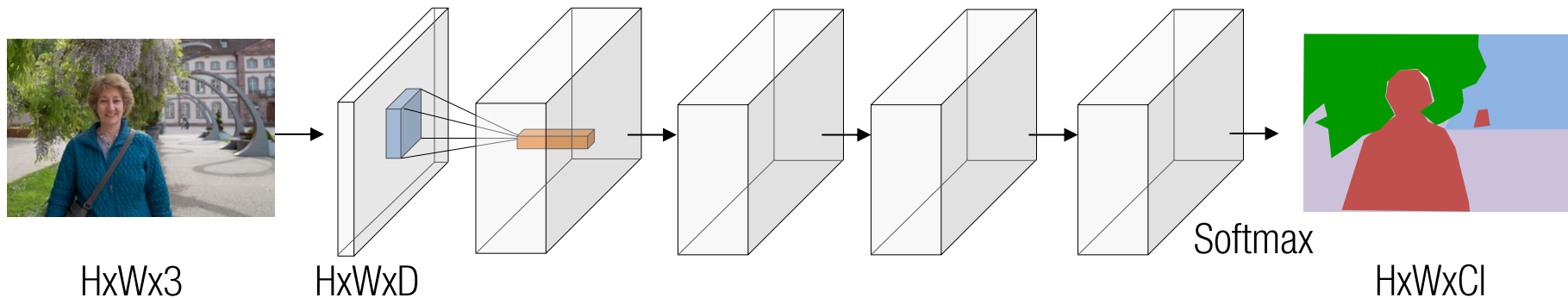
Softmax

$H \times W \times C$

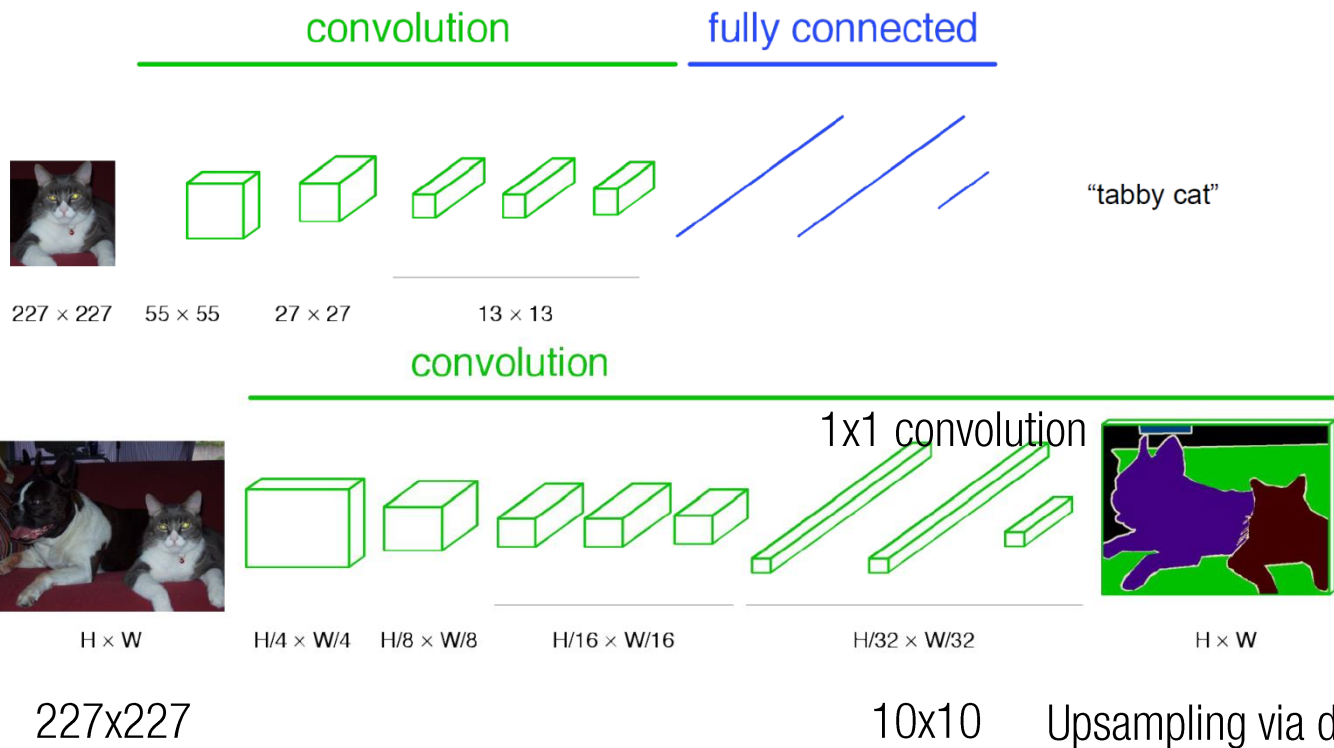
Computationally inefficient



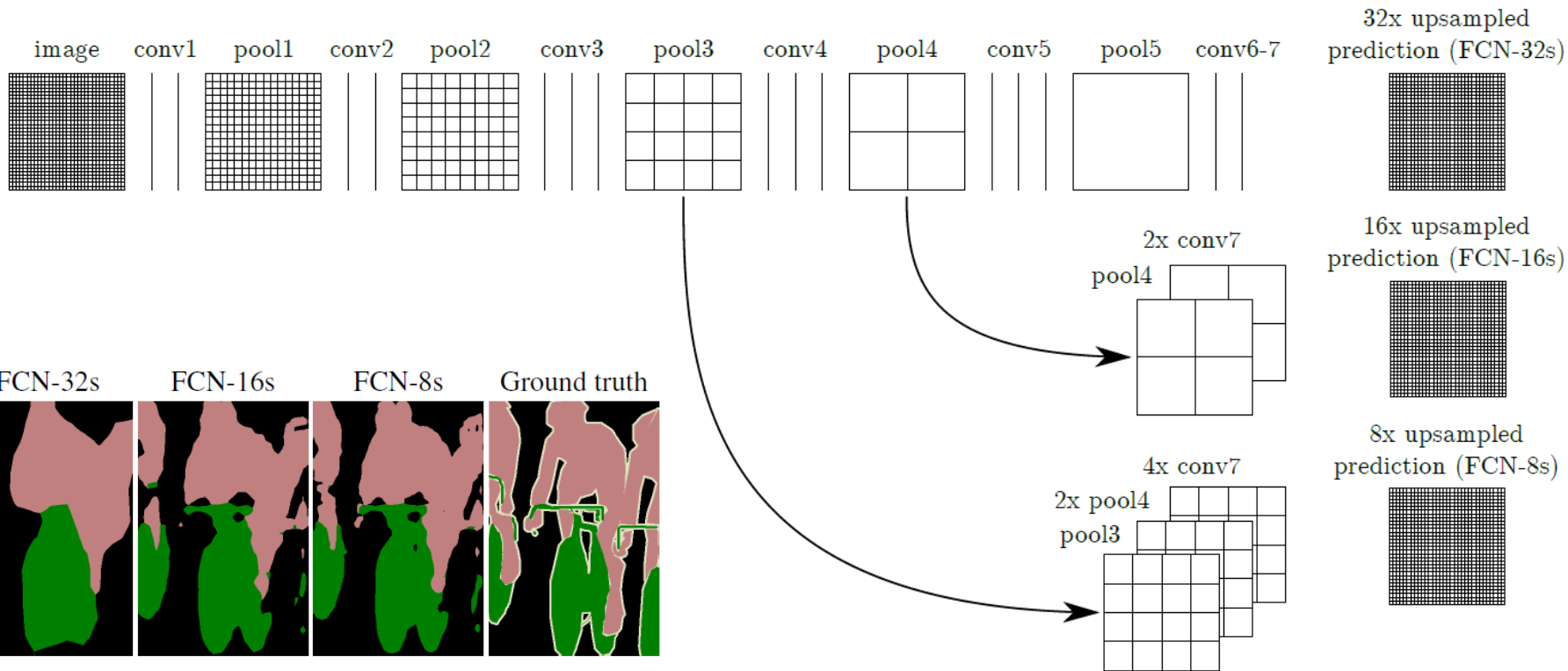
HOLISTIC PREDICTION VIA DEEP LEARNING

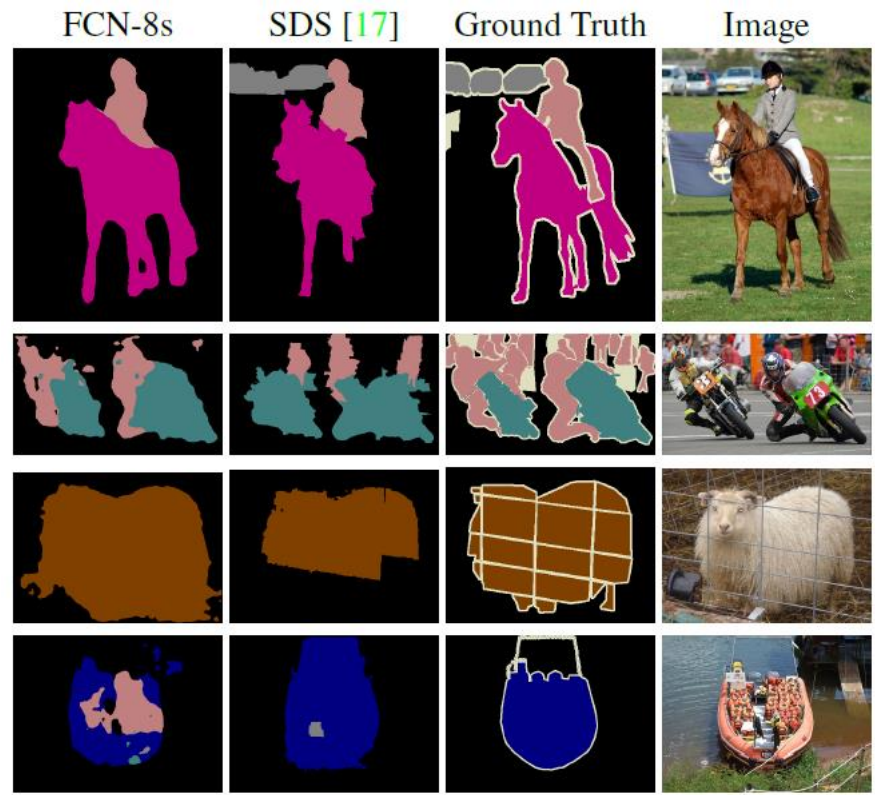


FULLY CONVOLUTIONAL NETWORK

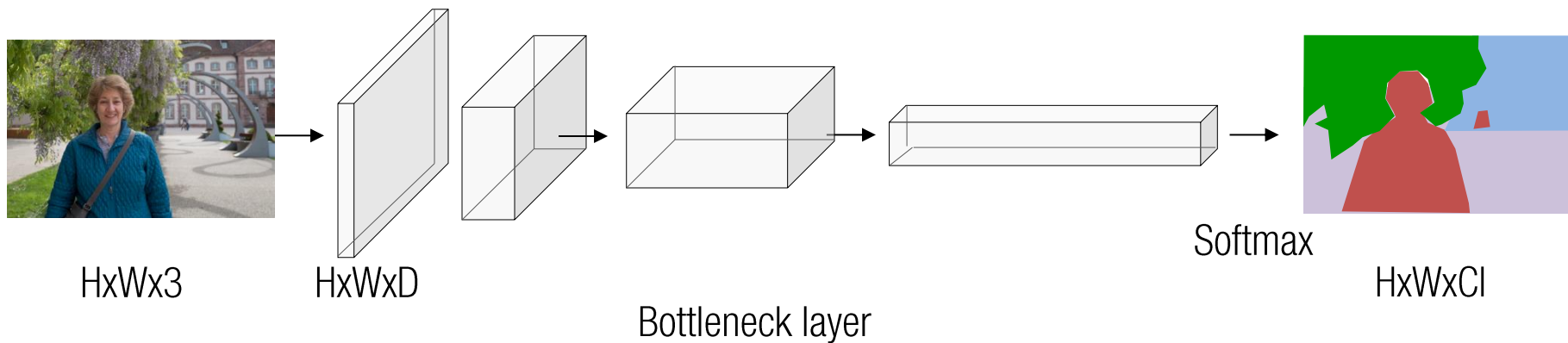
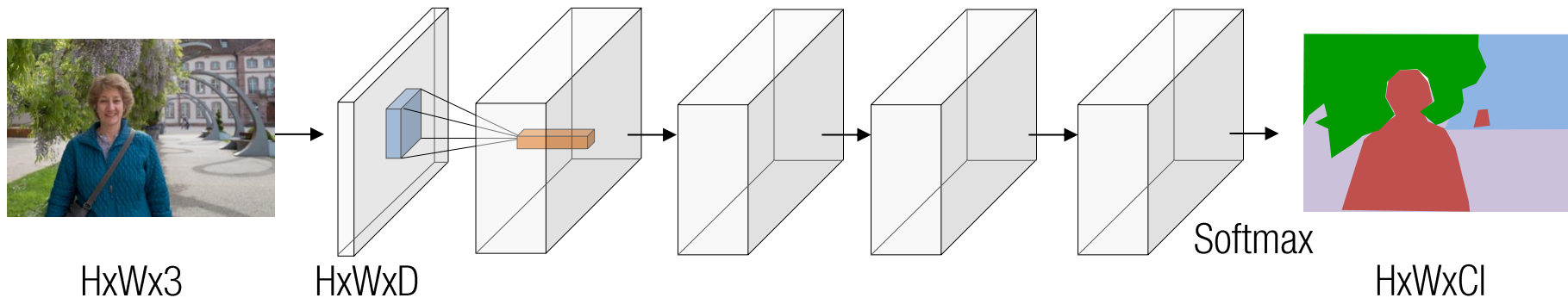


FULLY CONVOLUTIONAL NETWORK

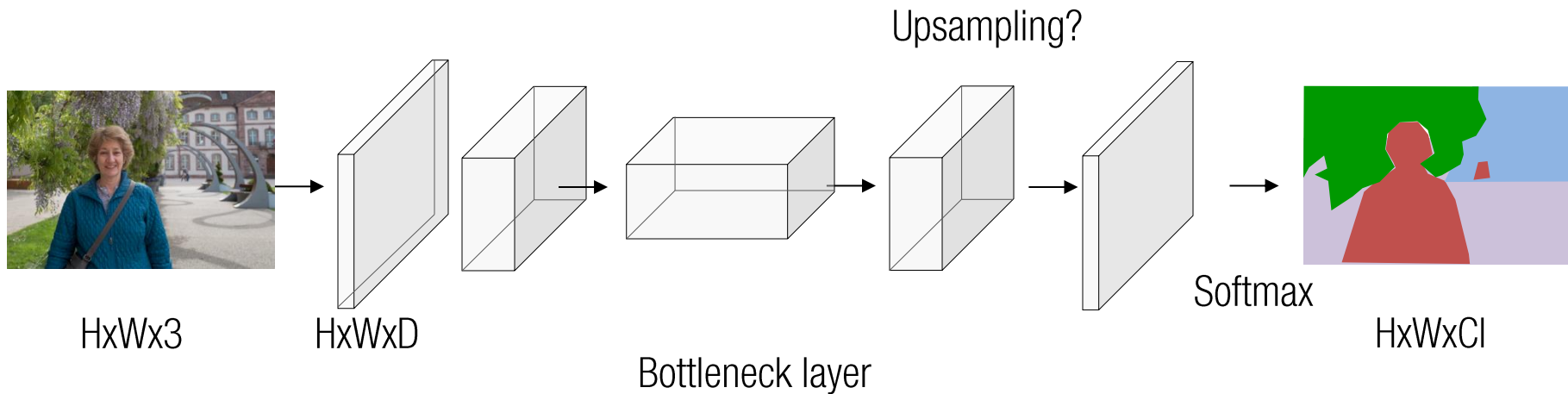
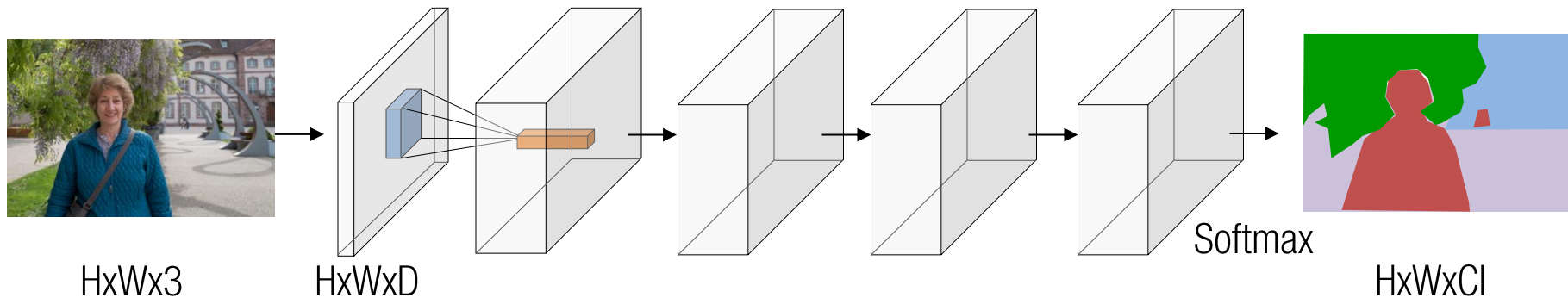




HOLISTIC PREDICTION VIA DEEP LEARNING



HOLISTIC PREDICTION VIA DEEP LEARNING



REVISITED: SPATIAL POOLING (DOWN-SAMPLING)

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

4×4

5	4
7	7

2×2

Max-pooling (window size 2x2, stride 2)

SPATIAL UNPOOLING (UP-SAMPLING)

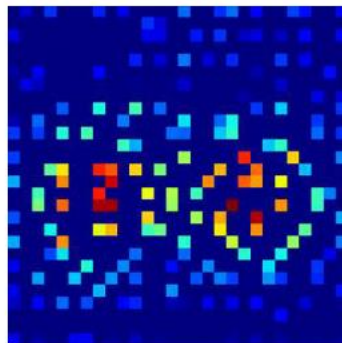
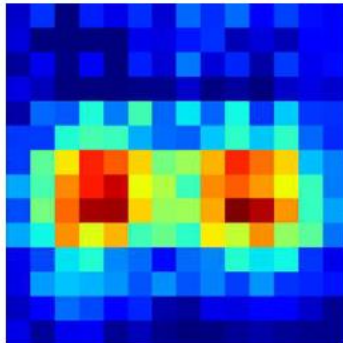
Max-**un**pooling (window size 2x2, stride 2)

5	4
7	7

0	0	0	0
0	5	0	4
0	0	0	0
7	0	0	7

Learnable parameter?

Can we learn upsampling?

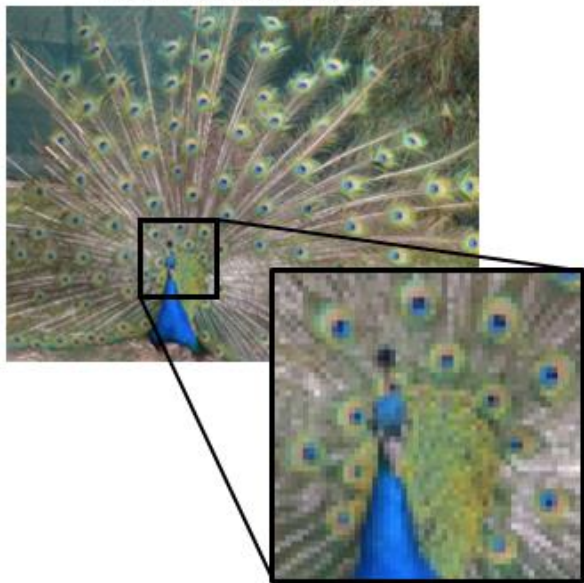


LEARNING UPSAMPLING

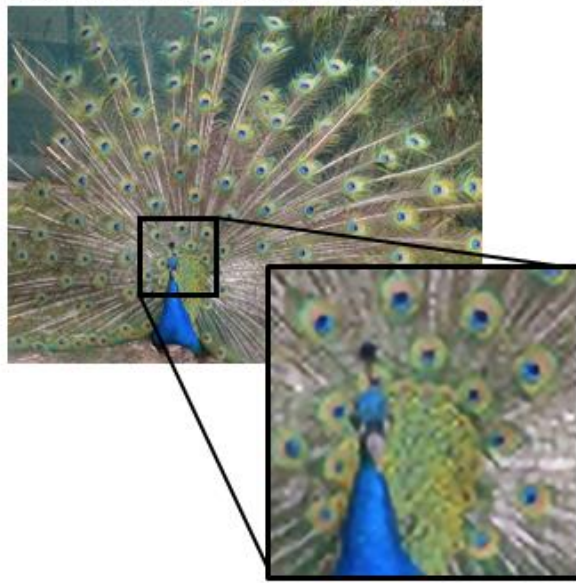
Low-Resolution Image



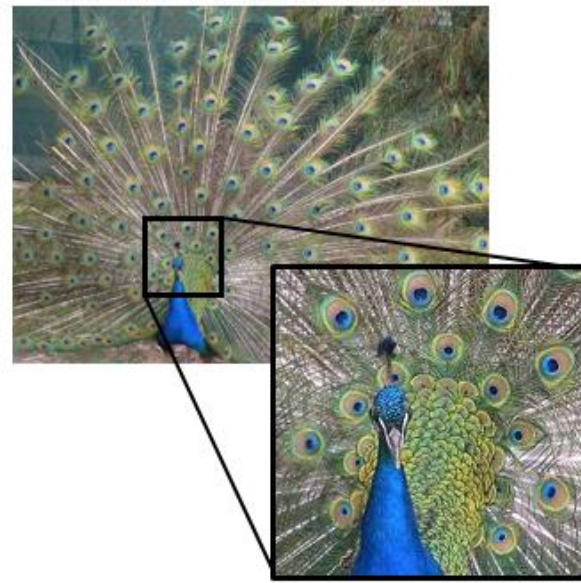
High-Resolution Image



High-Resolution Super-Resolved Image



High-Resolution Reference Image



SPATIAL UNPOOLING (UP-SAMPLING)

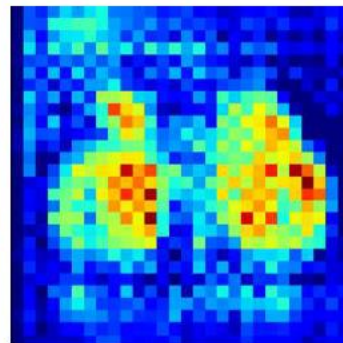
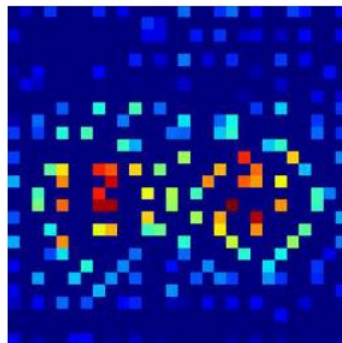
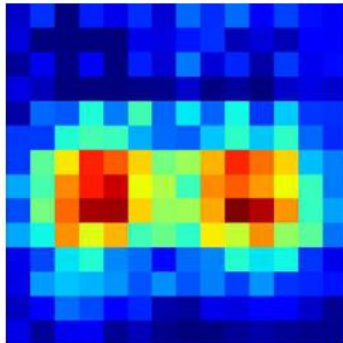
Max-**un**pooling (window size 2x2, stride 2)

5	4
7	7

0	0	0	0
0	5	0	4
0	0	0	0
7	0	0	7

Learnable parameter?

Can we learn upsampling?



REVISITED: CONVOLUTION

X_{11}	X_{12}		
			X_{44}

 \otimes

W_{11}	W_{12}		
			W_{13}

 $=$

y_{11}	y_{12}
	y_{22}

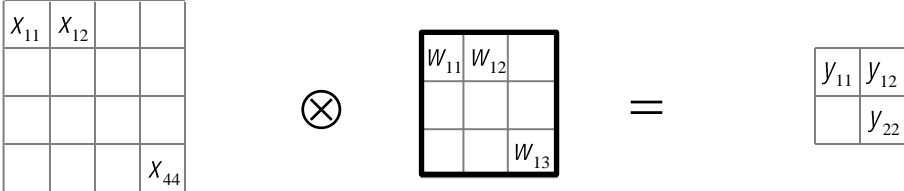
$$y_{11} = W_{11}X_{11} + W_{12}X_{12} + \cdots + W_{33}X_{33}$$

$$y_{12} = W_{11}X_{12} + W_{12}X_{13} + \cdots + W_{33}X_{34}$$

$$y_{21} = W_{11}X_{21} + W_{12}X_{21} + \cdots + W_{33}X_{43}$$

$$y_{22} = W_{11}X_{22} + W_{12}X_{23} + \cdots + W_{33}X_{44}$$

REVISITED: CONVOLUTION



$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{11} & w_{12} & w_{13} & 0 & w_{21} & w_{22} & w_{23} & 0 & w_{31} & w_{32} & w_{33} \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{44} \end{bmatrix}$$

REVISITED: CONVOLUTION

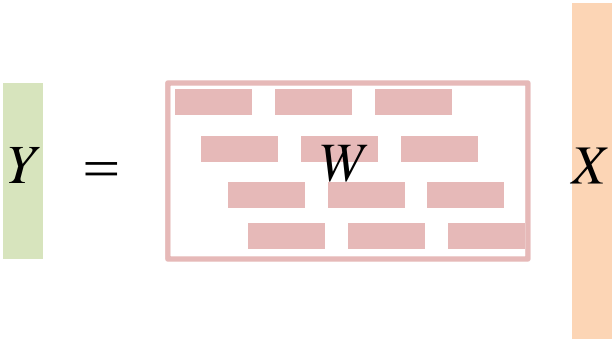
X_{11}	X_{12}		
			X_{44}

 \otimes

W_{11}	W_{12}		
		W_{13}	

 $=$

y_{11}	y_{12}		



Non-zero entries of each row encode local connectivity.

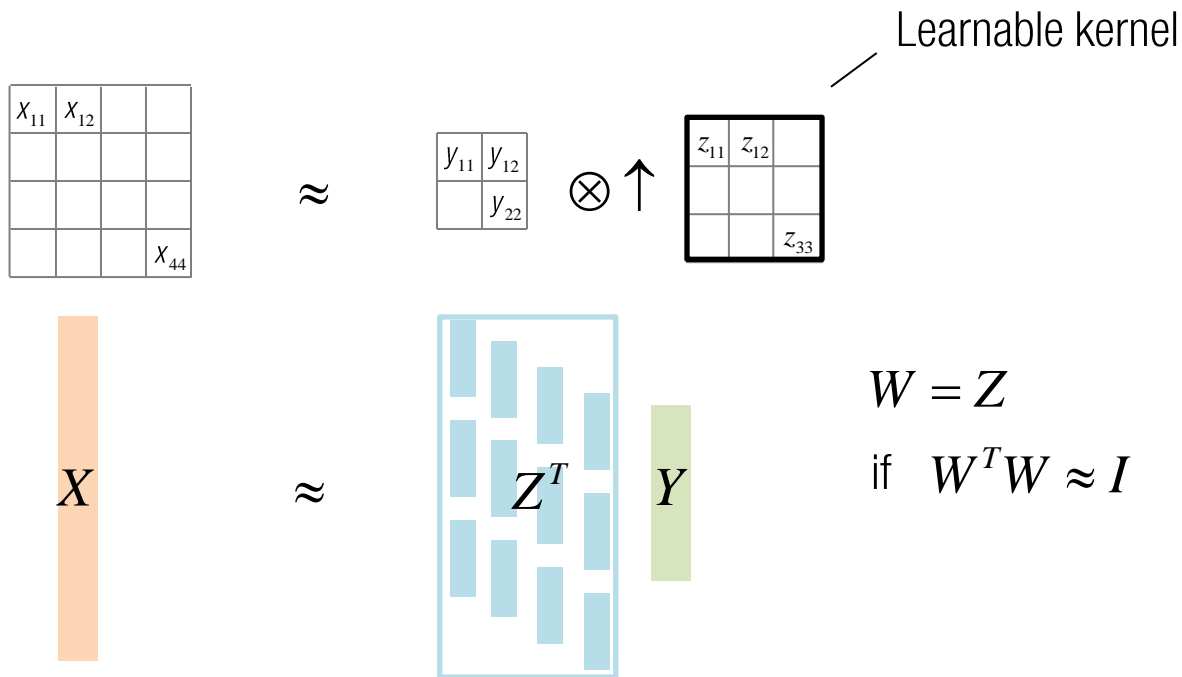
INVERSE OF CONVOLUTION

$$\begin{bmatrix} x_{11} & x_{12} & & \\ & & & \\ & & & \\ & & & x_{44} \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ & y_{22} \end{bmatrix} \left[\otimes \begin{bmatrix} w_{11} & w_{12} \\ & \\ & w_{13} \end{bmatrix} \right]^{-1}$$

$$\mathbf{X} = \mathbf{W}^+ \mathbf{Y}$$

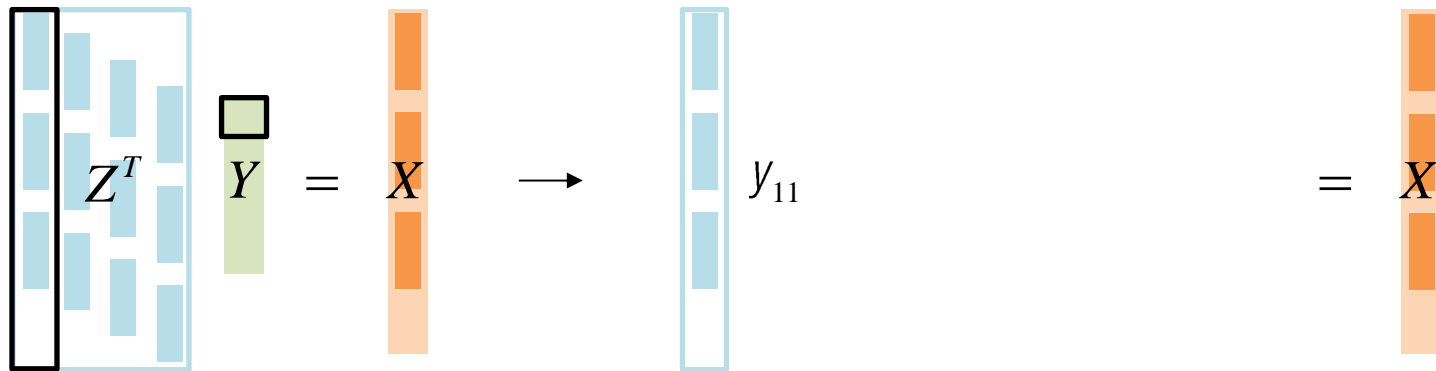
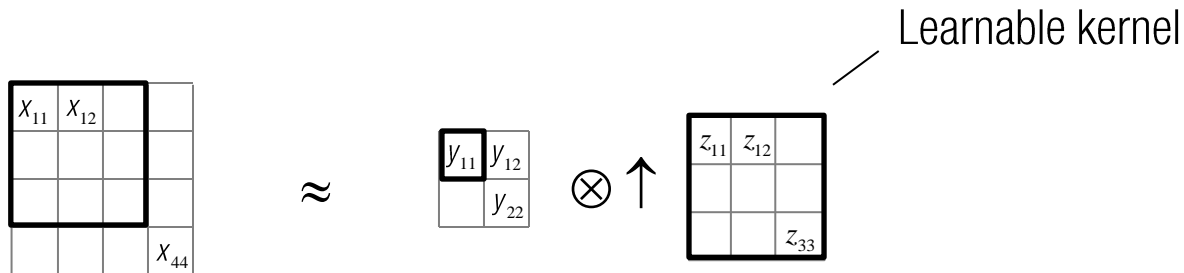
Inverse of W does not preserve local connectivity.

UPCONVOLUTION ~ INVERSE OF CONVOLUTION



Non-zero entries of column encode local connectivity.

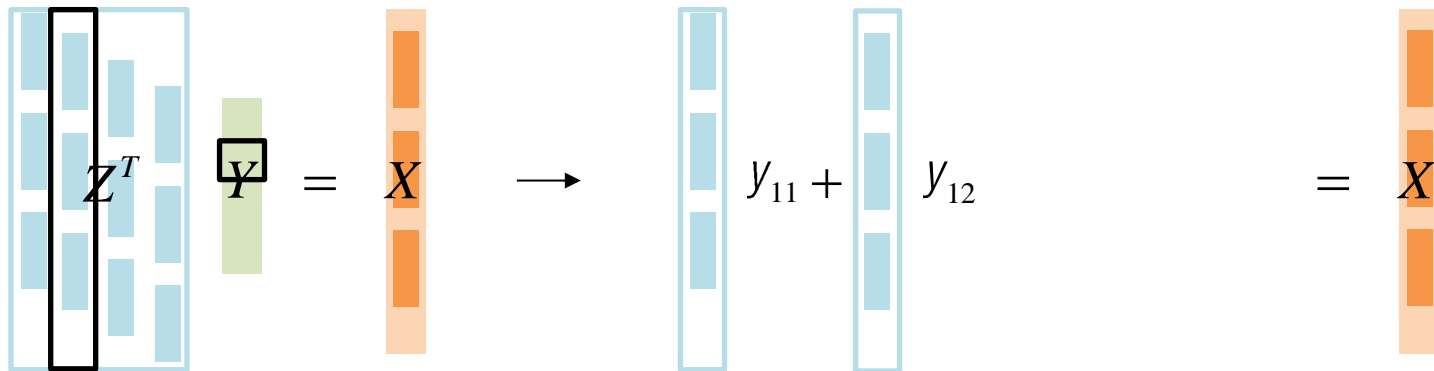
UPCONVOLUTION ~ INVERSE OF CONVOLUTION



Non-zero entries of column encode local connectivity.

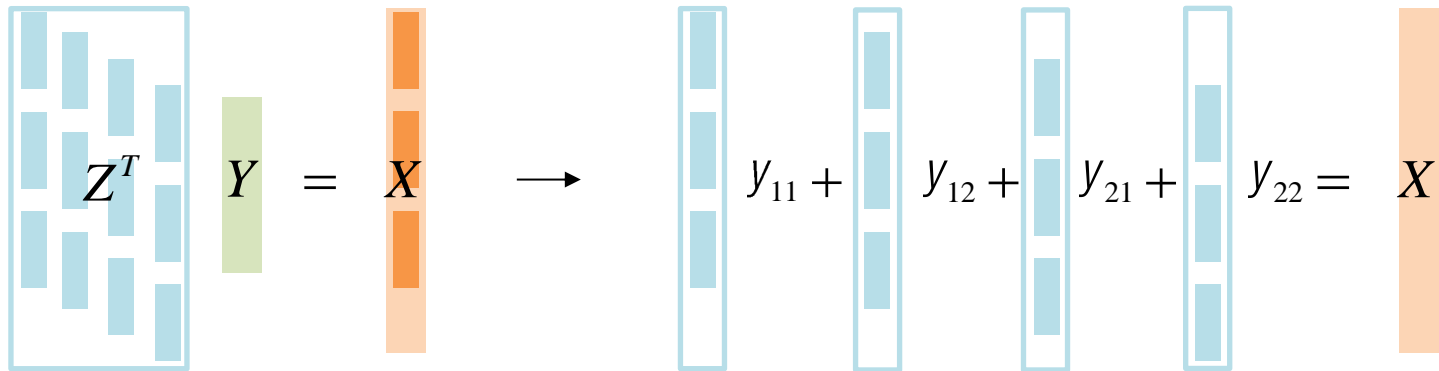
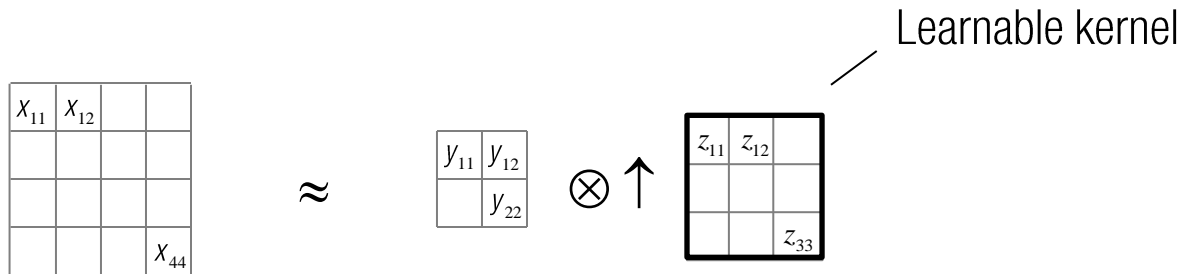
TRANSPOSED CONVOLUTION ~ INVERSE OF CONVOLUTION

$$\begin{array}{|c|c|c|} \hline X_{11} & X_{12} & \\ \hline & & \\ \hline & & \\ \hline & & X_{44} \\ \hline \end{array} \approx \begin{array}{|c|c|} \hline y_{11} & y_{12} \\ \hline & y_{22} \\ \hline \end{array} \otimes \uparrow \begin{array}{|c|c|} \hline z_{11} & z_{12} \\ \hline & \\ \hline & z_{33} \\ \hline \end{array} \quad \text{Learnable kernel}$$



Non-zero entries of column encode local connectivity.

TRANSPOSED CONVOLUTION ~ INVERSE OF CONVOLUTION



Non-zero entries of column encode local connectivity.

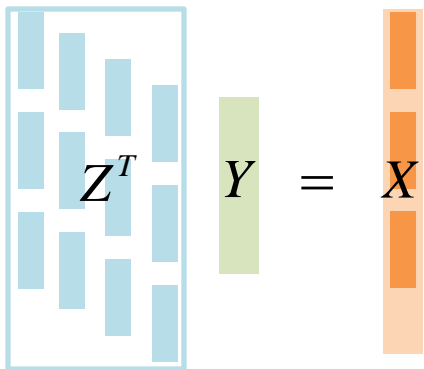
TRANSPOSED CONVOLUTION AS UPCONVOLUTION

$$\begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & & \\ \hline & & & \\ \hline & & & \\ \hline & & & X_{44} \\ \hline \end{array} = \begin{array}{|c|c|} \hline y_{11} & y_{12} \\ \hline & y_{22} \\ \hline \end{array} \otimes \uparrow \begin{array}{|c|c|c|} \hline Z_{11} & Z_{12} & Z_{13} \\ \hline Z_{21} & Z_{22} & Z_{23} \\ \hline Z_{31} & Z_{32} & Z_{33} \\ \hline \end{array}$$

Upconvolution

$$\begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & & \\ \hline & & & \\ \hline & & & \\ \hline & & & X_{44} \\ \hline \end{array} \otimes \begin{array}{|c|c|} \hline W_{11} & W_{12} \\ \hline & \\ \hline & W_{13} \\ \hline \end{array} = \begin{array}{|c|c|} \hline y_{11} & y_{12} \\ \hline & y_{22} \\ \hline \end{array}$$

Cf) convolution



$$x_{11} = z_{11}y_{11}$$

x_{11} is dependent only on y_{11} .

x_{11} contributes only to y_{11} .

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

The diagram illustrates the operation of a transposed convolution as an upconvolution. It shows three 3x3 grids. The first grid on the left represents the input X , with elements x_{11} and x_{12} highlighted in a black box. This is followed by an equals sign, a 2x2 grid representing the kernel Y with elements y_{11} , y_{12} , y_{21} , and y_{22} , also highlighted in a black box. To the right of the kernel is a circled cross symbol \otimes and an upward-pointing arrow \uparrow . The final grid on the right represents the output Z , with elements z_{11} and z_{12} highlighted in a black box. The output grid contains elements z_{11}, z_{12}, z_{13} in the first row, z_{21}, z_{22}, z_{23} in the second row, and z_{31}, z_{32}, z_{33} in the third row.

Upconvolution

The diagram shows a matrix equation $Y = X$ using colored bars. On the left, a light blue box contains several vertical bars of varying heights, representing the matrix Z^T . To its right is a green vertical bar representing the vector Y . An equals sign follows, and to its right is an orange vertical bar representing the vector X .

$$x_{11} = z_{11}y_{11}$$

$$x_{12} = z_{12}y_{11} + z_{11}y_{12}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

The diagram illustrates the operation of a transposed convolution as an upconvolution. On the left, a 3x3 grid represents the input X with elements X_{11} , X_{12} , and X_{44} . A 2x2 kernel Y with elements y_{11} , y_{12} , and y_{22} is applied to the input. This is followed by a convolution operation \otimes and an upscaling operation \uparrow to produce a 6x6 grid representing the output Z with elements Z_{11} , Z_{12} , Z_{13} , Z_{21} , Z_{22} , Z_{23} , Z_{31} , Z_{32} , and Z_{33} .

Upconvolution

The diagram shows the matrix equation $X = Z^T Y$. On the left, a blue-bordered box contains a 6x3 grid of light blue vertical bars, representing the matrix Z^T . To its right is a green vertical bar representing the vector Y . An equals sign follows, and to the right is an orange vertical bar representing the vector X .

$$X_{11} = Z_{11}y_{11}$$

$$X_{12} = Z_{12}y_{11} + Z_{11}y_{12}$$

$$X_{13} = Z_{13}y_{11} + Z_{12}y_{12}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

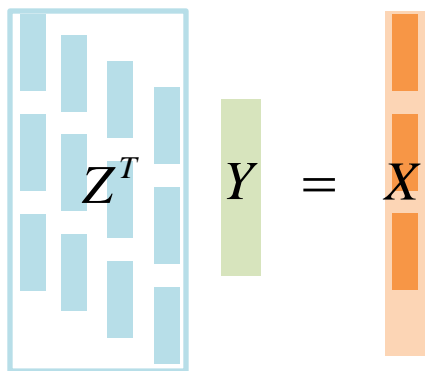
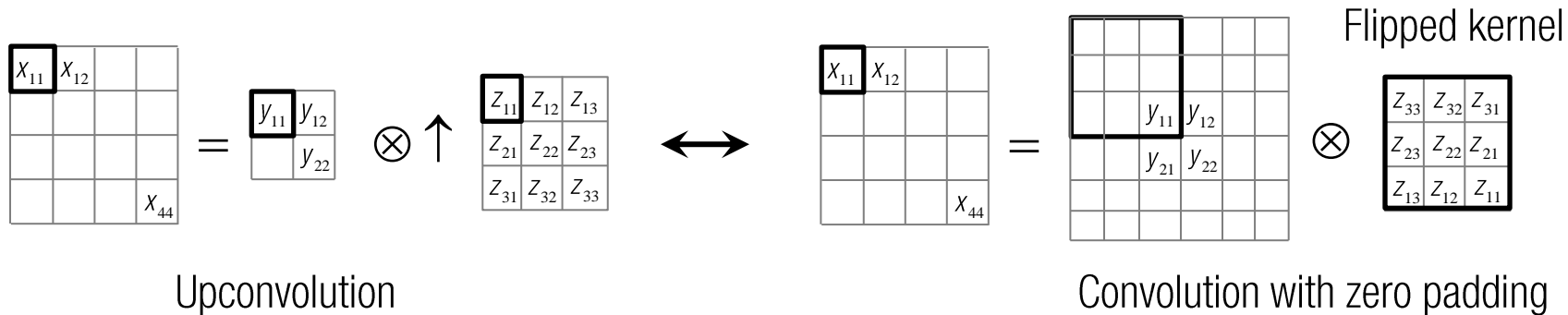
$$\begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & & \square \\ \hline & & & \\ \hline & & & \\ \hline & & & X_{44} \\ \hline \end{array} = \begin{array}{|c|c|} \hline y_{11} & y_{12} \\ \hline & y_{22} \\ \hline \end{array} \otimes \uparrow \begin{array}{|c|c|c|} \hline Z_{11} & Z_{12} & Z_{13} \\ \hline Z_{21} & Z_{22} & Z_{23} \\ \hline Z_{31} & Z_{32} & Z_{33} \\ \hline \end{array}$$

Upconvolution

$$Z^T Y = X$$

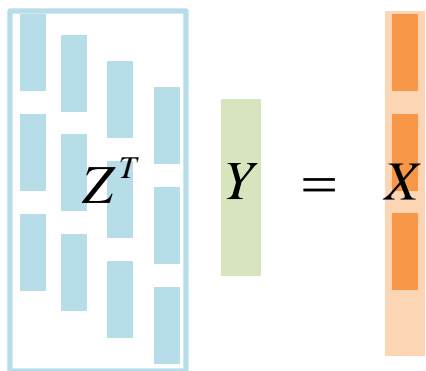
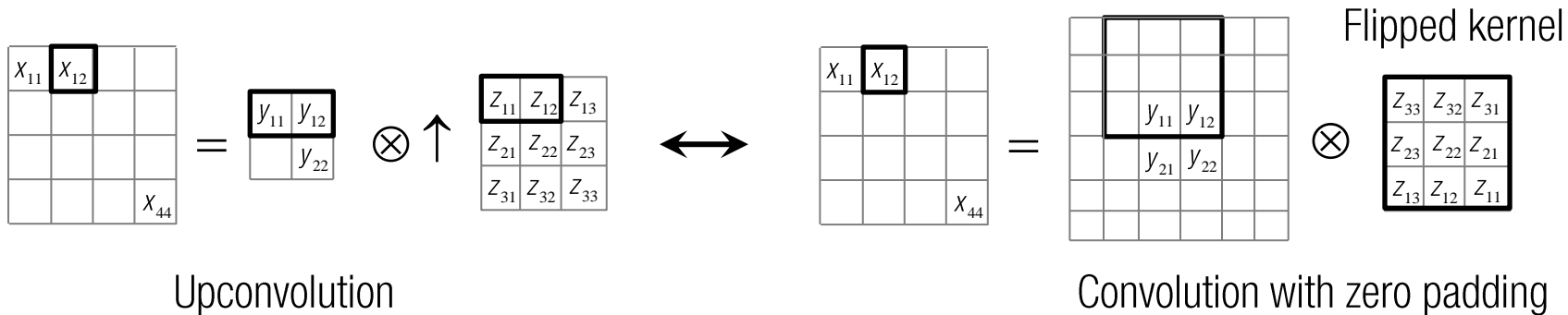
$$\begin{aligned}
 X_{11} &= Z_{11}y_{11} \\
 X_{12} &= Z_{12}y_{11} + Z_{11}y_{12} \\
 X_{13} &= Z_{13}y_{11} + Z_{12}y_{12} \\
 X_{14} &= Z_{13}y_{12}
 \end{aligned}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION



$$x_{11} = z_{11}y_{11}$$

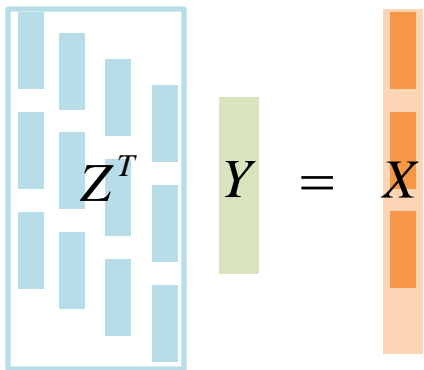
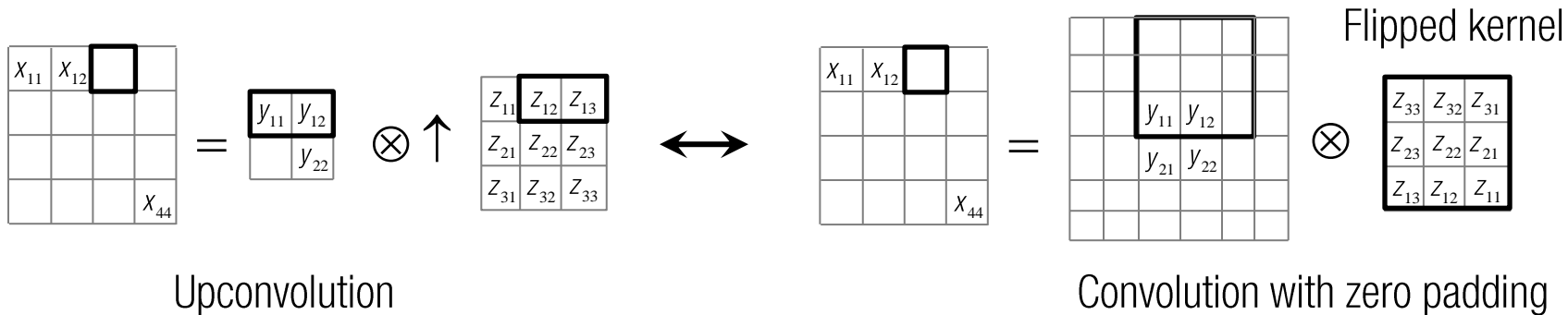
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



$$x_{11} = z_{11}y_{11}$$

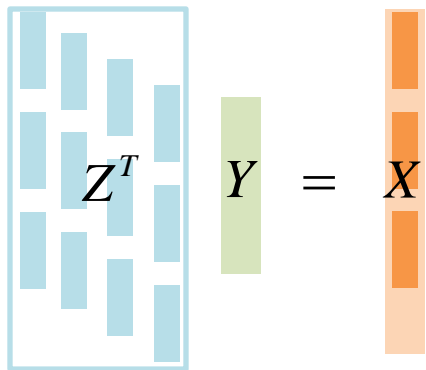
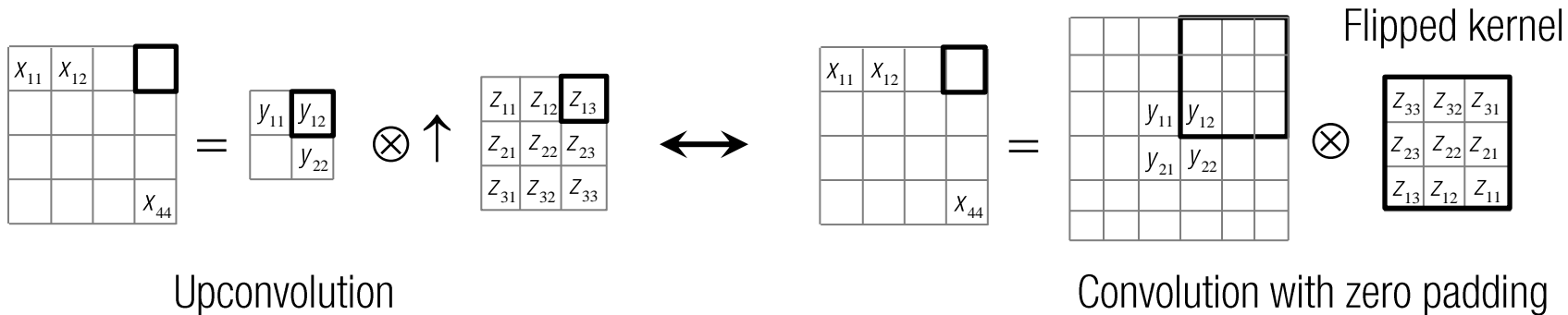
$$x_{12} = z_{12}y_{11} + z_{11}y_{12}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION



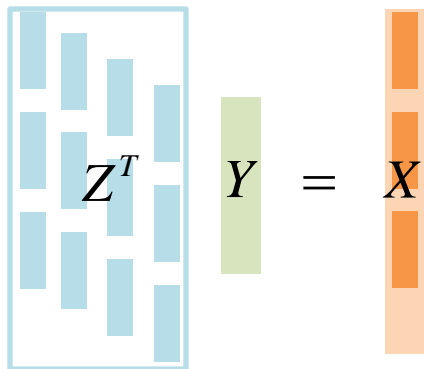
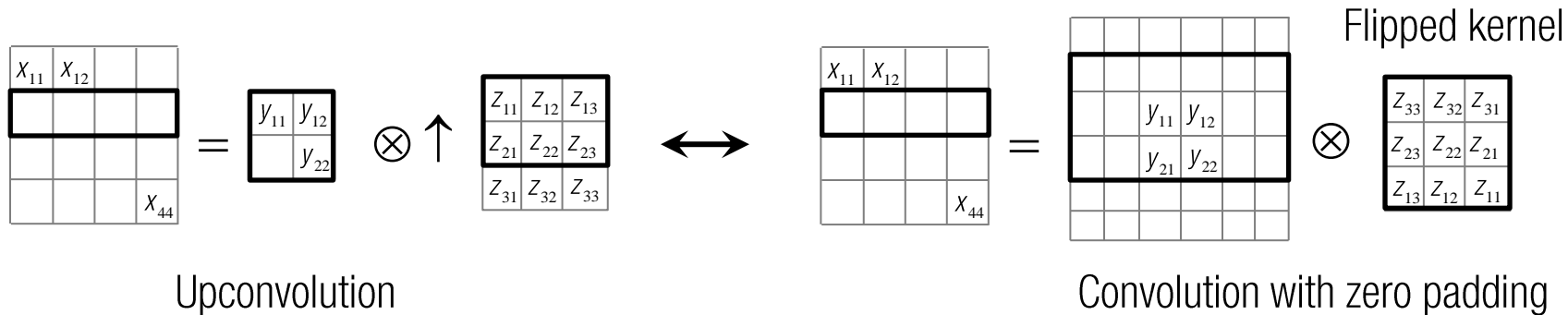
$$\begin{aligned}
 X_{11} &= Z_{11}y_{11} \\
 X_{12} &= Z_{12}y_{11} + Z_{11}y_{12} \\
 X_{13} &= Z_{13}y_{11} + Z_{12}y_{12}
 \end{aligned}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION



$$\begin{aligned}
 X_{11} &= Z_{11}y_{11} \\
 X_{12} &= Z_{12}y_{11} + Z_{11}y_{12} \\
 X_{13} &= Z_{13}y_{11} + Z_{12}y_{12} \\
 X_{14} &= Z_{13}y_{12}
 \end{aligned}$$

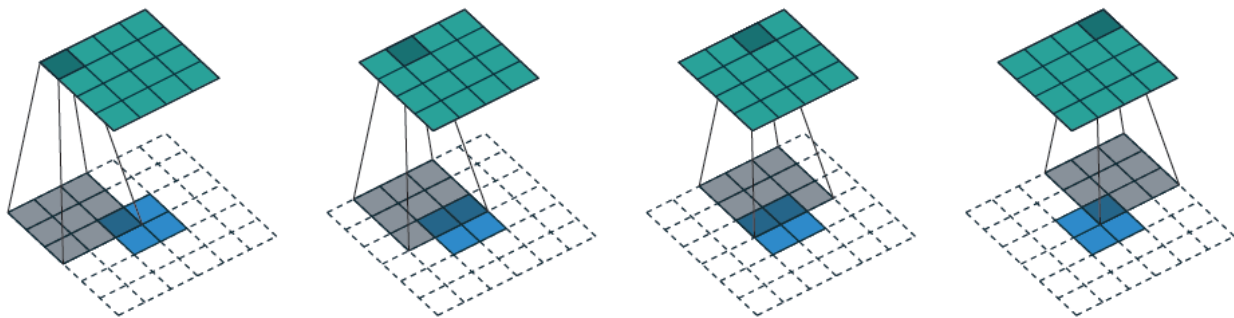
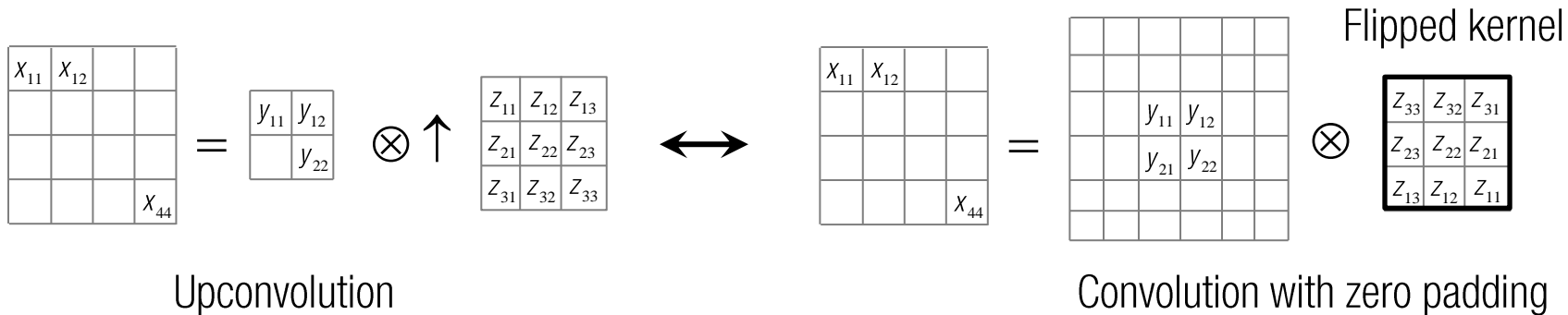
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



$$\begin{aligned} x_{11} &= z_{11}y_{11} \\ x_{12} &= z_{12}y_{11} + z_{11}y_{12} \\ x_{13} &= z_{13}y_{11} + z_{12}y_{12} \\ x_{14} &= z_{13}y_{12} \end{aligned}$$

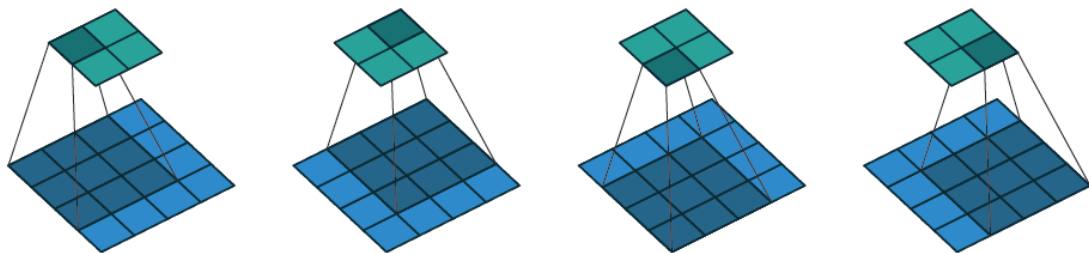
$$\begin{aligned} x_{21} &= z_{21}y_{11} + z_{11}y_{21} \\ x_{22} &= z_{22}y_{11} + z_{21}y_{12} + z_{12}y_{21} + z_{11}y_{22} \\ x_{23} &= z_{23}y_{11} + z_{22}y_{12} + z_{13}y_{21} + z_{12}y_{22} \\ x_{24} &= z_{23}y_{12} + z_{13}y_{22} \end{aligned}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

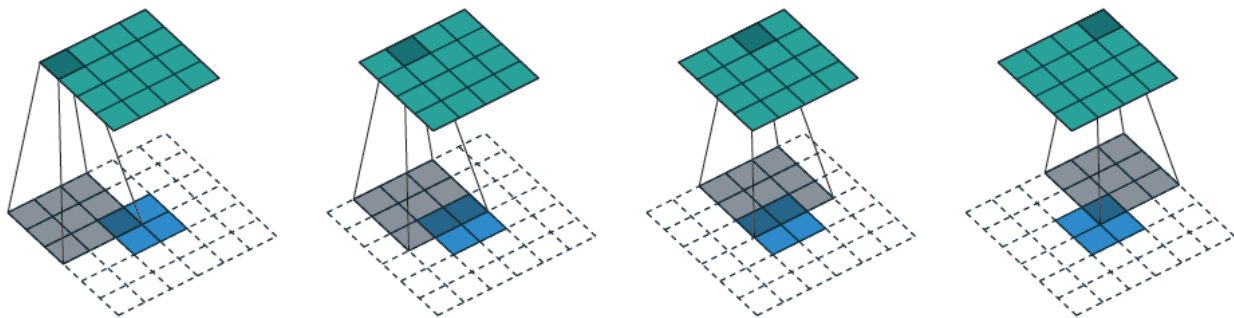


Upconvolution

DUALITY: CONVOLUTION VS. UPCONVOLUTION

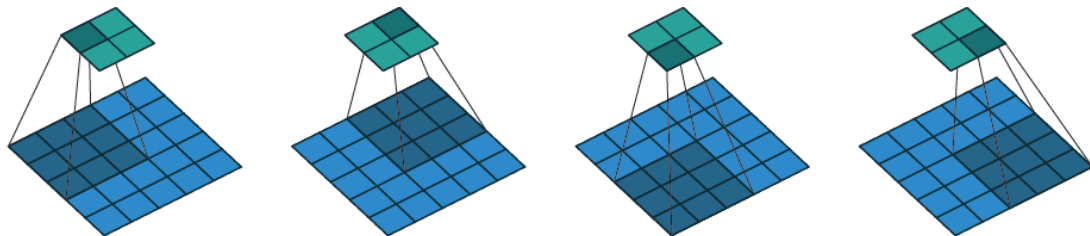


Convolution (3x3 kernel, no zero padding, 1 stride)

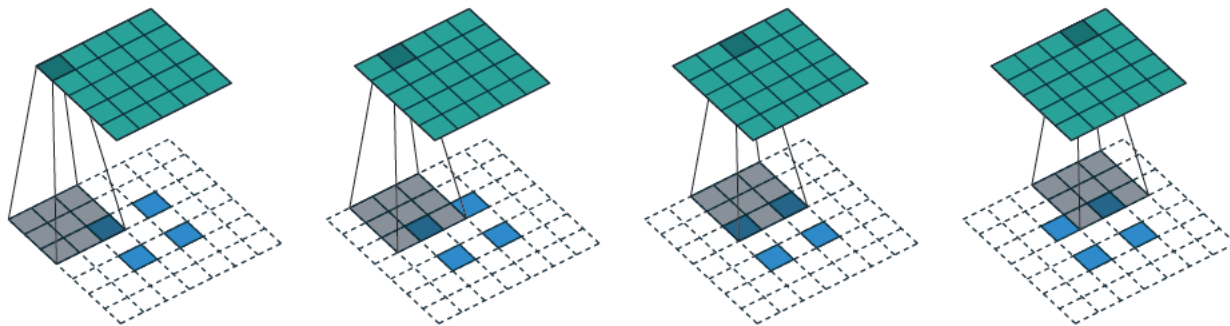


Upconvolution (3x3 kernel, 2x2 zero padding, 1 stride)

DUALITY: CONVOLUTION VS. UPCONVOLUTION

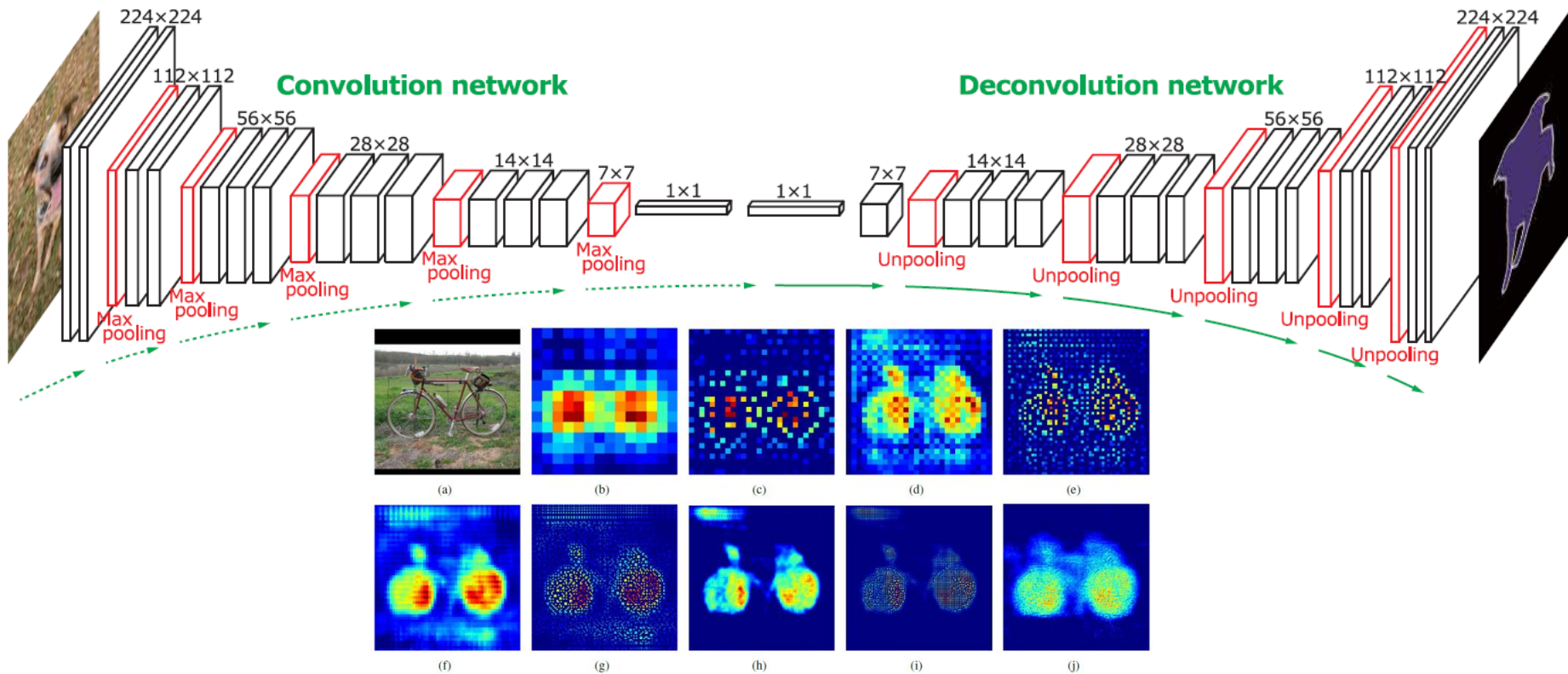


Convolution (3x3 kernel, no zero padding, 2 stride)



Upconvolution (3x3 kernel, 2x2 zero padding, 1 zero between input, 2 stride)

DEEPER UPCONVOLUTIONAL LAYERS



Input image

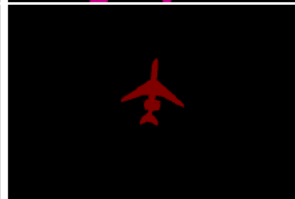
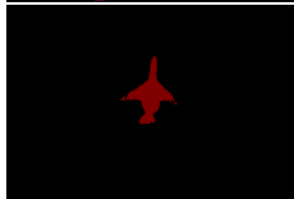
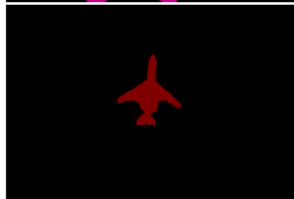
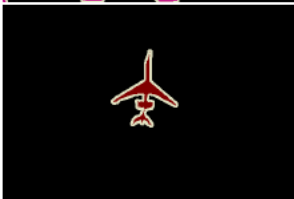
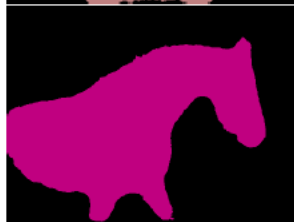
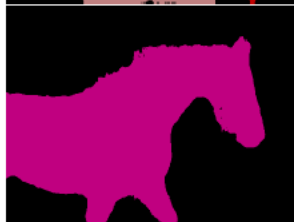
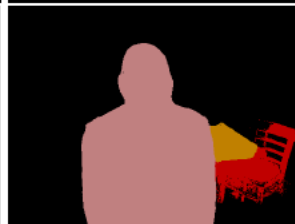
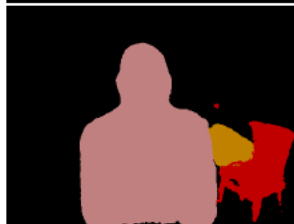
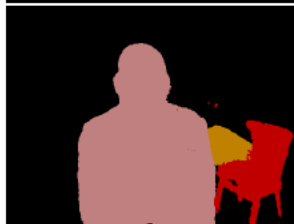
Ground-truth

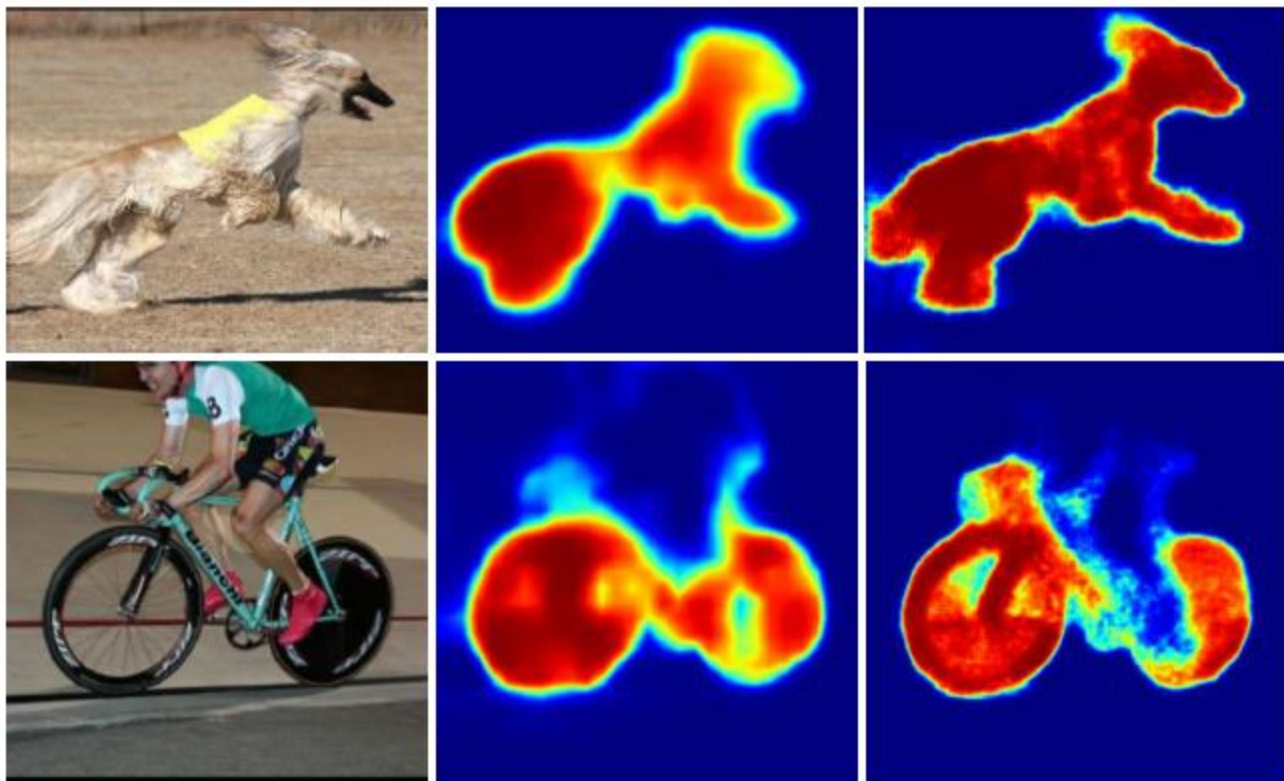
FCN

DeconvNet

EDeconvNet

EDeconvNet+CRF



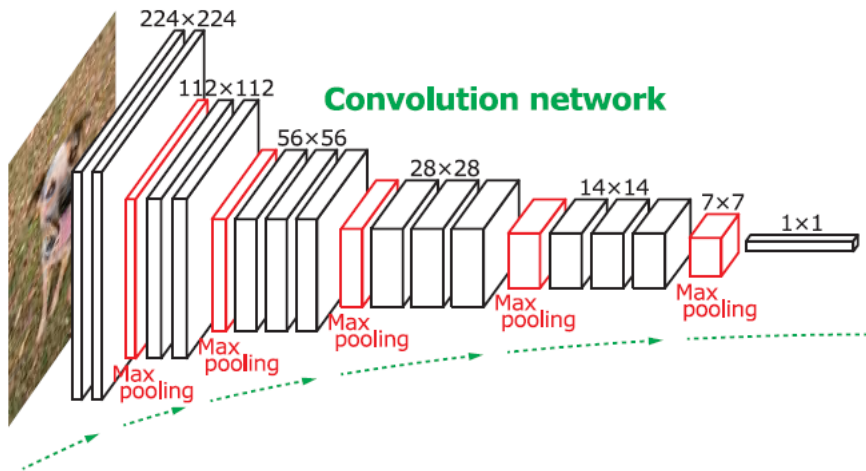


(a) Input image

(b) FCN-8s

(c) Ours

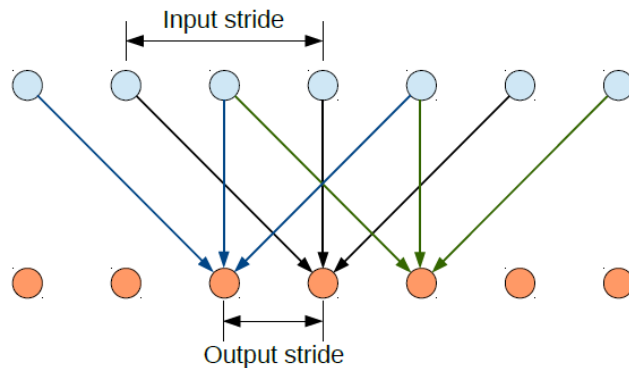
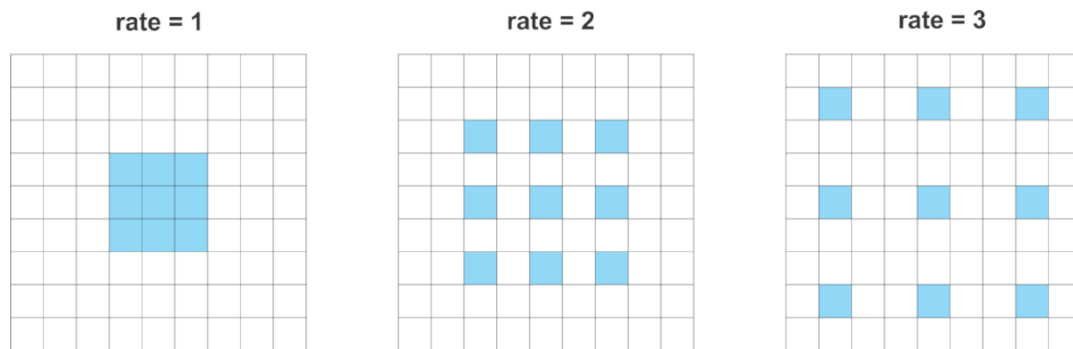
RECEPTIVE FIELD VS. RESOLUTION



Desired properties:

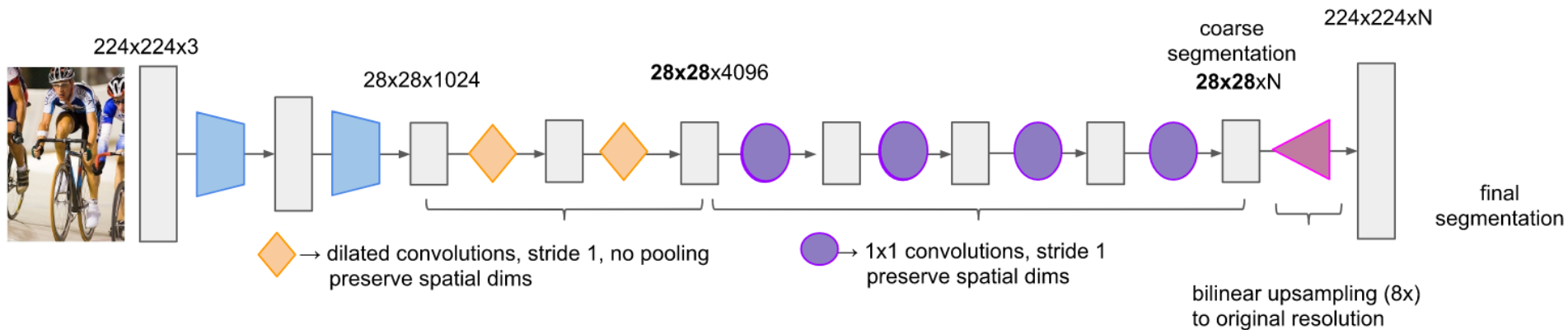
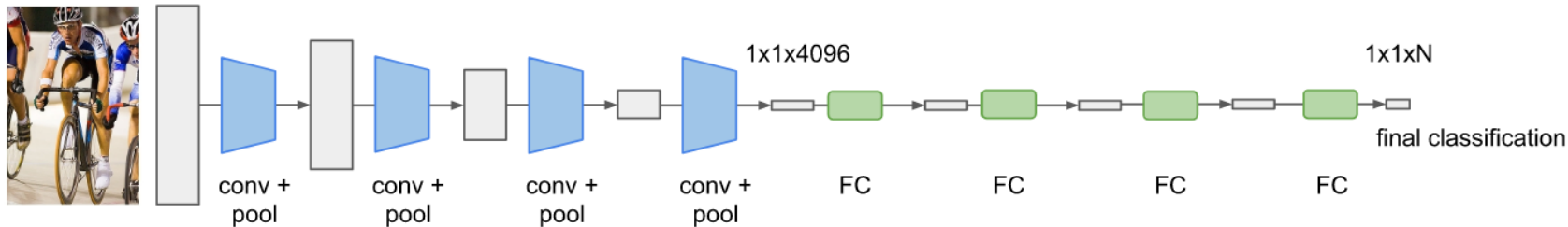
- Larger receptive field (bigger spatial context) \longrightarrow reducing resolution
- Higher output resolution \longrightarrow reducing receptive field

DEEPLAB: DILATED CONVOLUTION

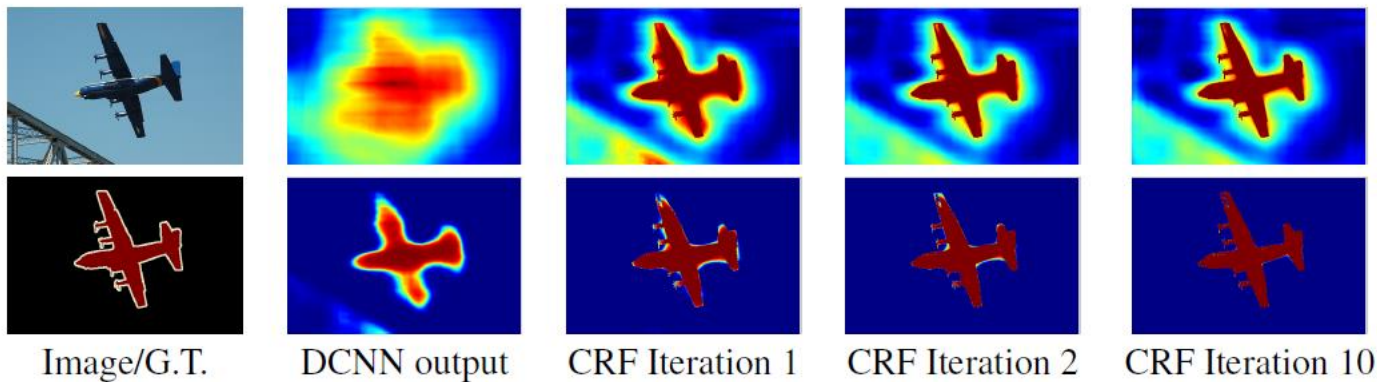


Enlarging receptive field without increasing reducing resolution

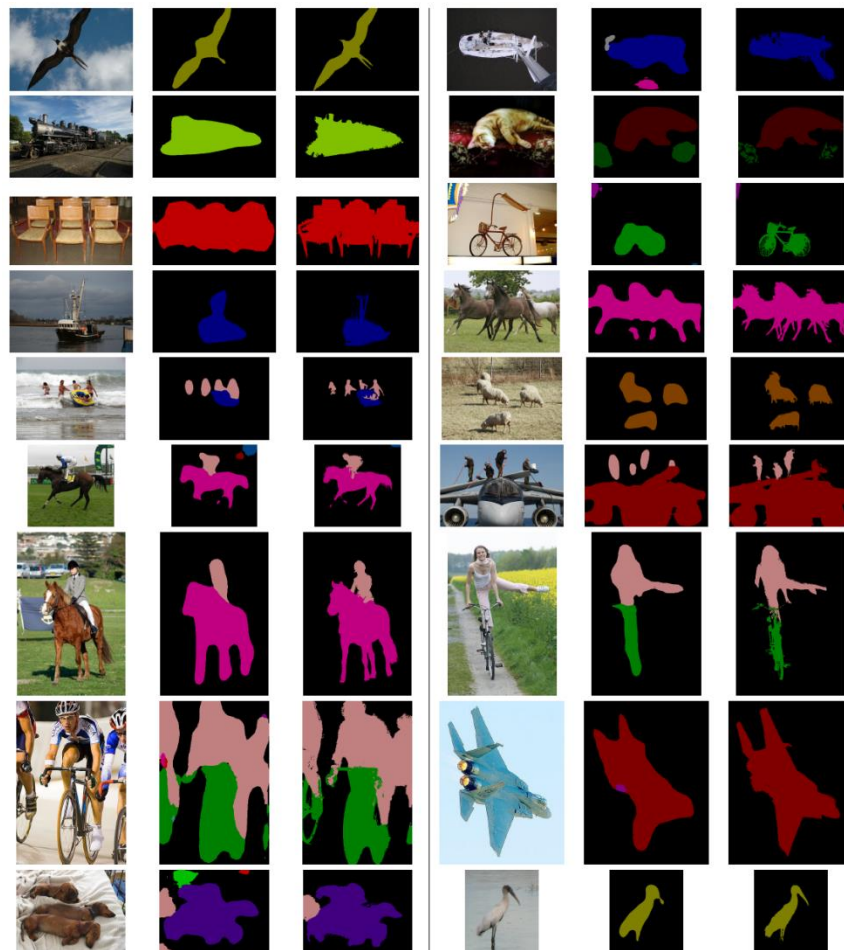
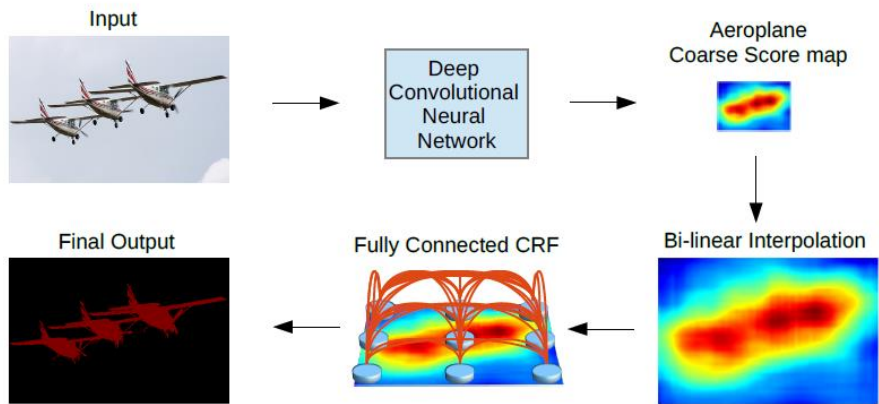
DEEPLAB: DILATED CONVOLUTION



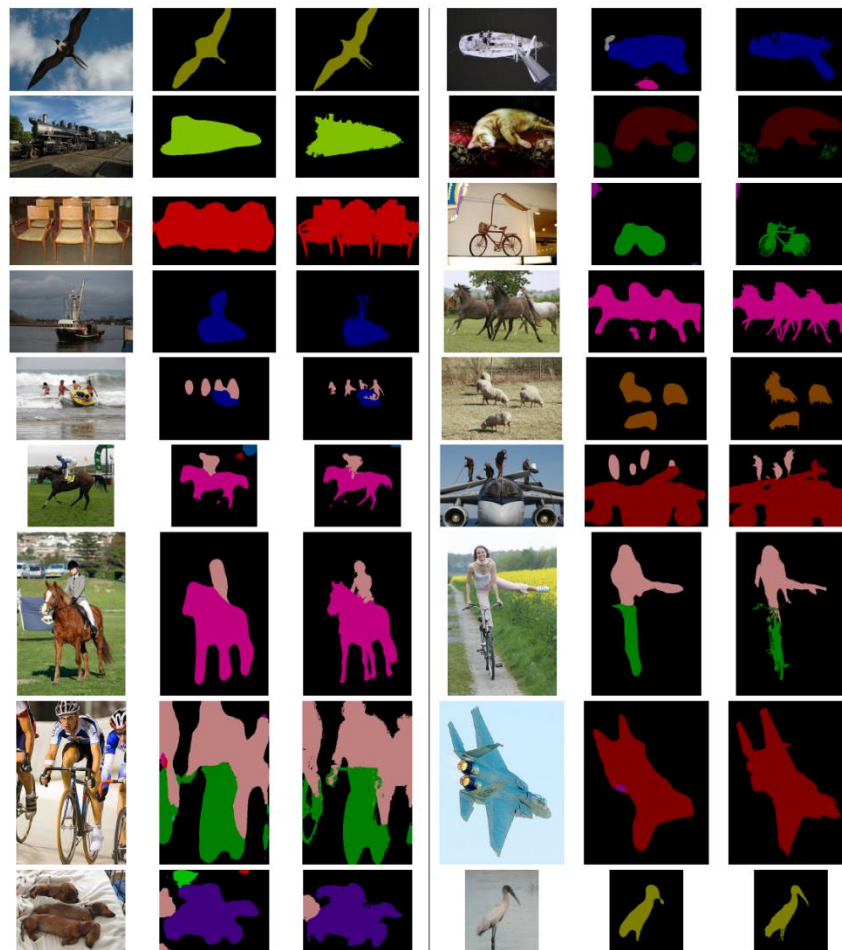
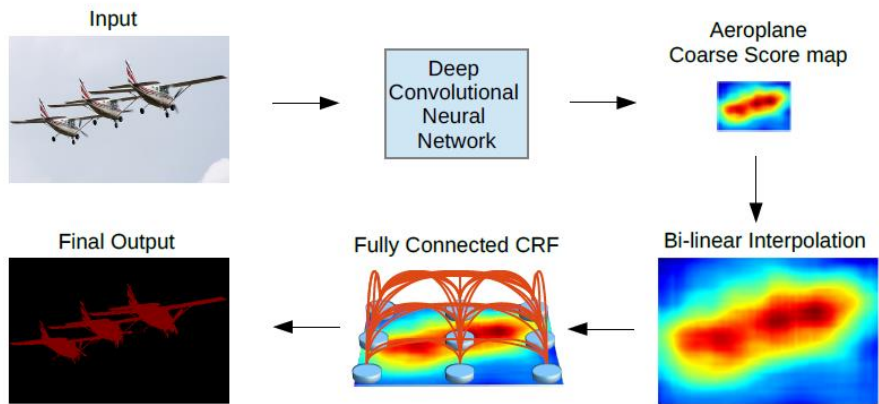
DEEPLAB: CRF POST-PROCESSING



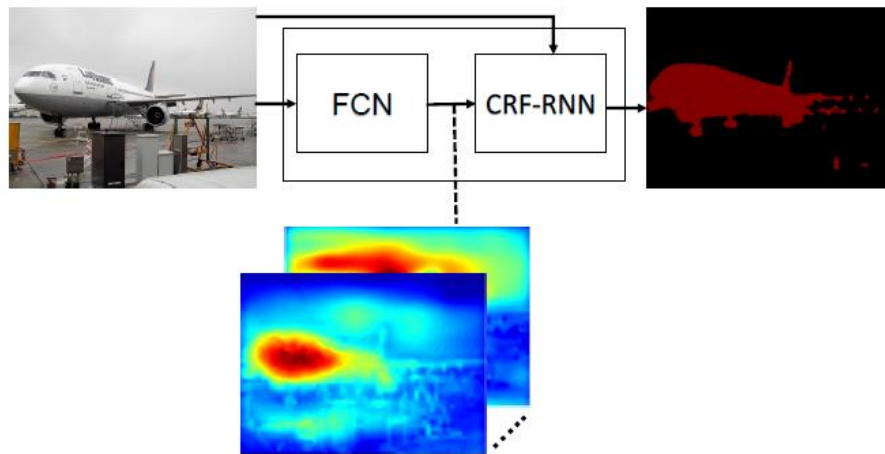
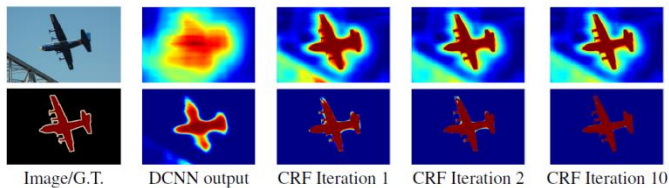
DEEPLAB



DEEPLAB



CRF AS RECURRENT NEURAL NET (END-TO-END)



CRF AS RECURRENT NEURAL NET (END-TO-END)

