

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, ...



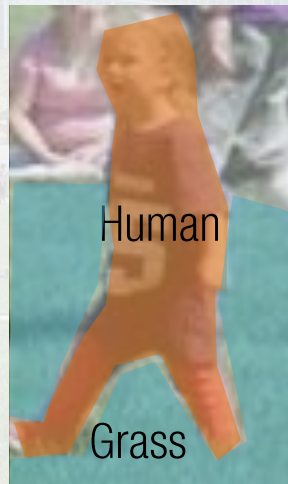
CHALLENGES OF VISUAL RECOGNITION

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



Human

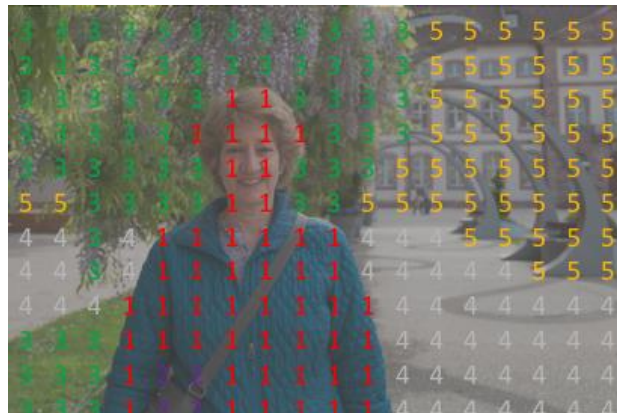
$$f(I) = l_{human}$$



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

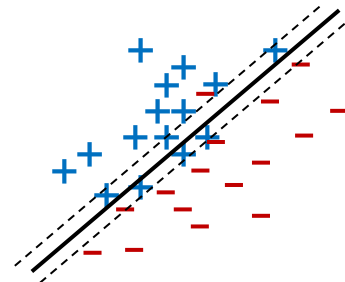


x_i

Visual feature

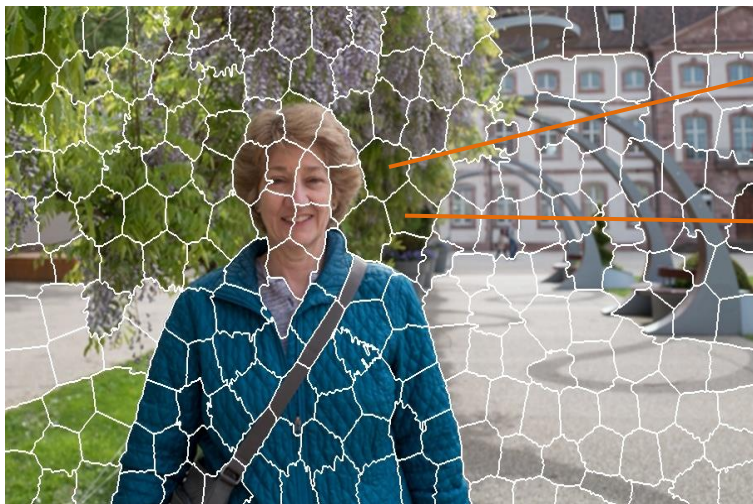
- Color histogram
- BoW
- SIFT
- HOG

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

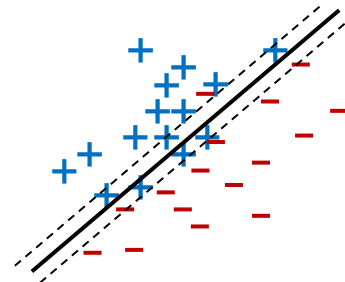


x_j

Visual feature

- Color histogram
- BoW
- SIFT
- HOG

Classification

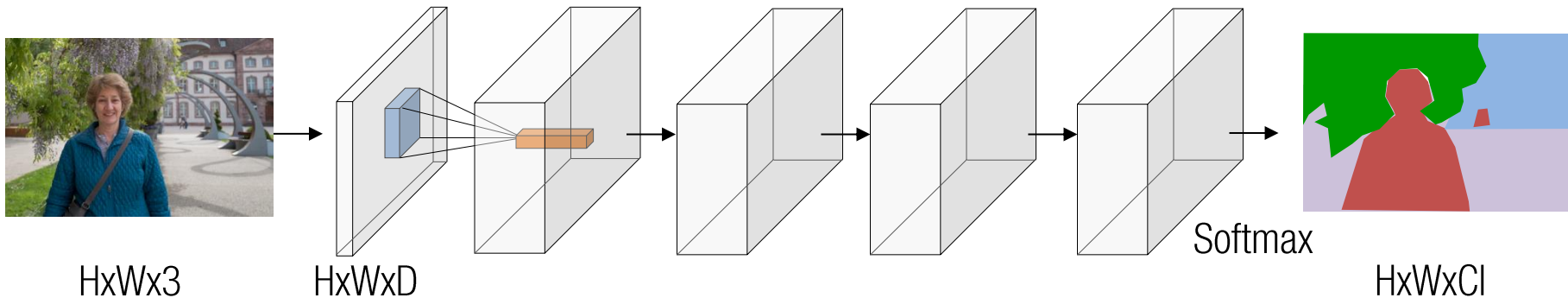


$$L = \phi(x_i) + \underbrace{\varphi(x_i, x_j)}_{\text{Context}}$$

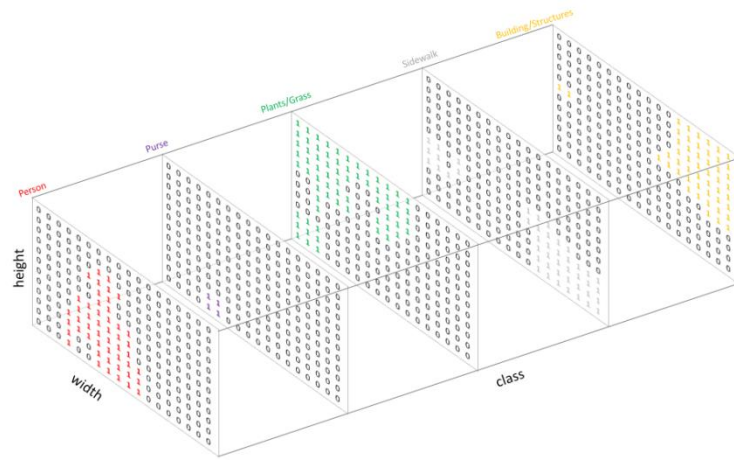
Context

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

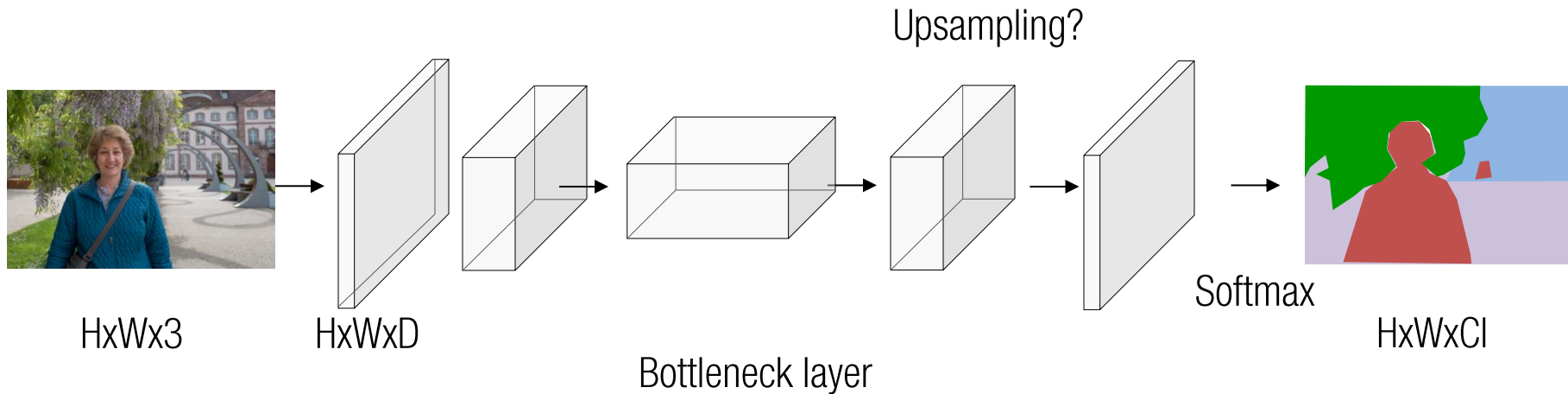
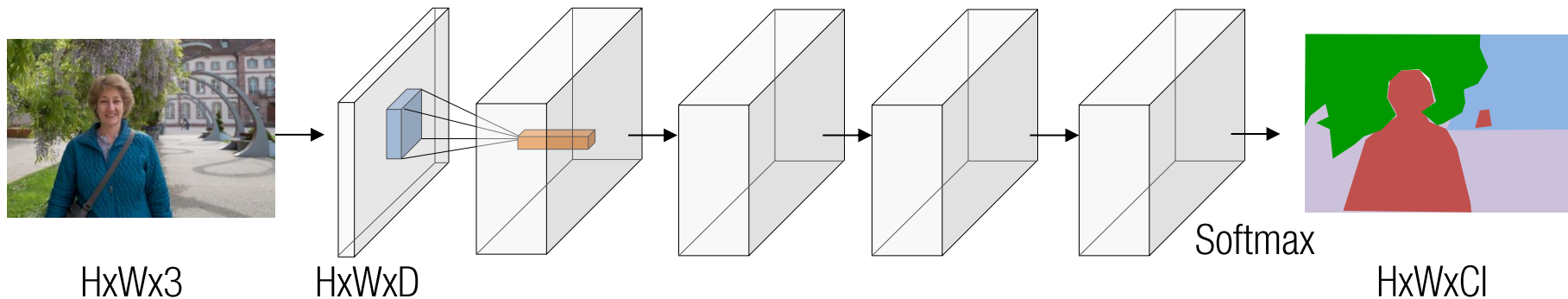
HOLISTIC PREDICTION VIA DEEP LEARNING



Computationally inefficient



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)

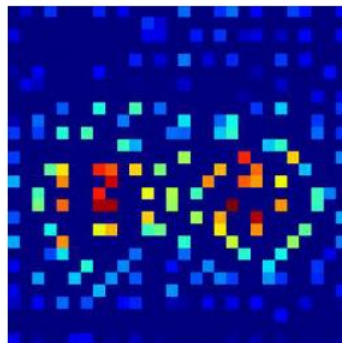
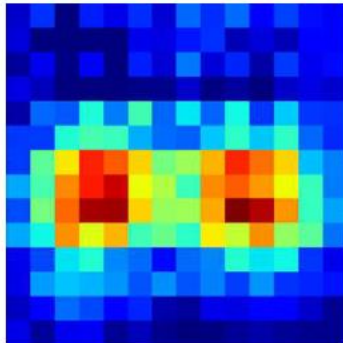
Max-**un**pooling (window size 2x2, stride 2)

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?

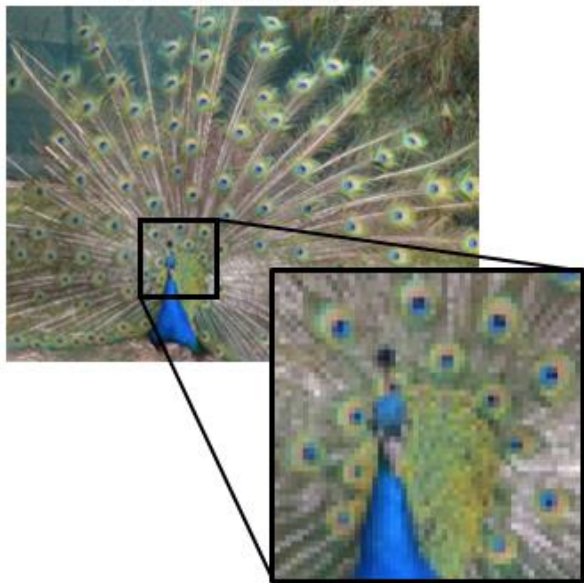


LEARNING UPSAMPLING

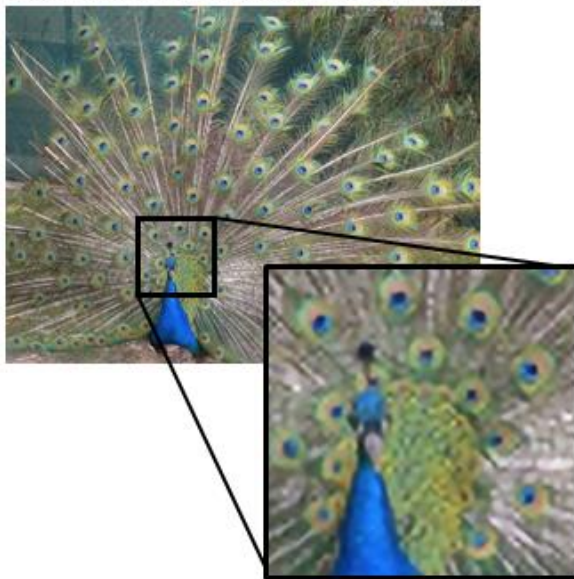
Low-Resolution Image



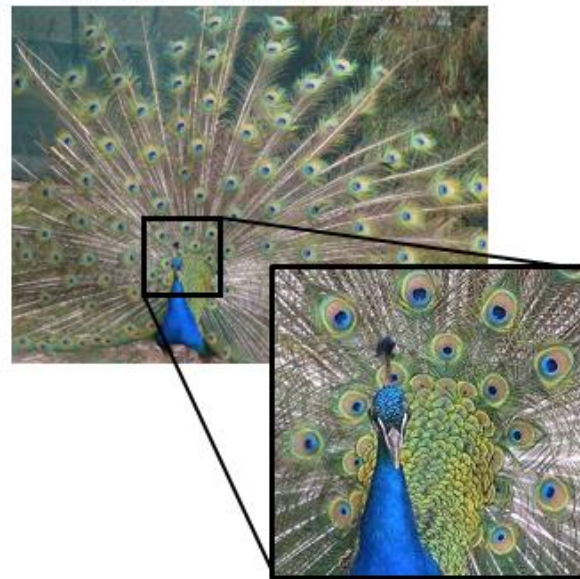
High-Resolution Image



High-Resolution Super-Resolved Image



High-Resolution Reference Image



SPATIAL UNPOOLING (UP-SAMPLING)

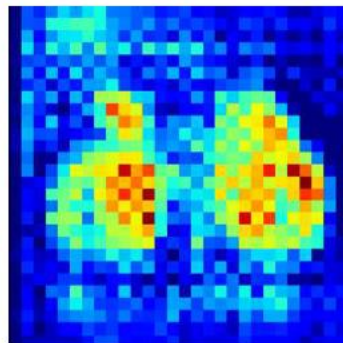
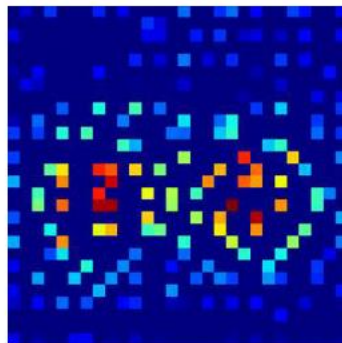
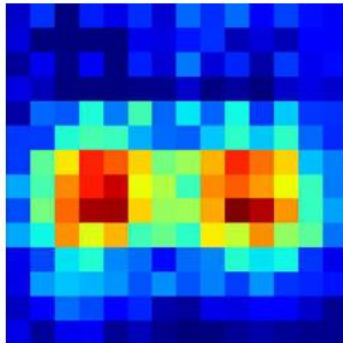
Max-**un**pooling (window size 2x2, stride 2)

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}

$$y_{11} = w_{11}x_{11} + w_{12}x_{12} + \cdots + w_{33}x_{33}$$

$$y_{12} = w_{11}x_{12} + w_{12}x_{13} + \cdots + w_{33}x_{34}$$

$$y_{21} = w_{11}x_{21} + w_{12}x_{22} + \cdots + w_{33}x_{43}$$

$$y_{22} = w_{11}x_{22} + w_{12}x_{23} + \cdots + w_{33}x_{44}$$

REVISITED: CONVOLUTION

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

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

REVISITED: CONVOLUTION

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

$$\text{Y} = \begin{array}{|c|c|c|c|c|} \hline \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\ \hline \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\ \hline \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\ \hline \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\ \hline \end{array} \begin{array}{c} \text{W} \end{array} \begin{array}{|c|} \hline \text{X} \\ \hline \end{array}$$

Non-zero entries of each row encode local connectivity.

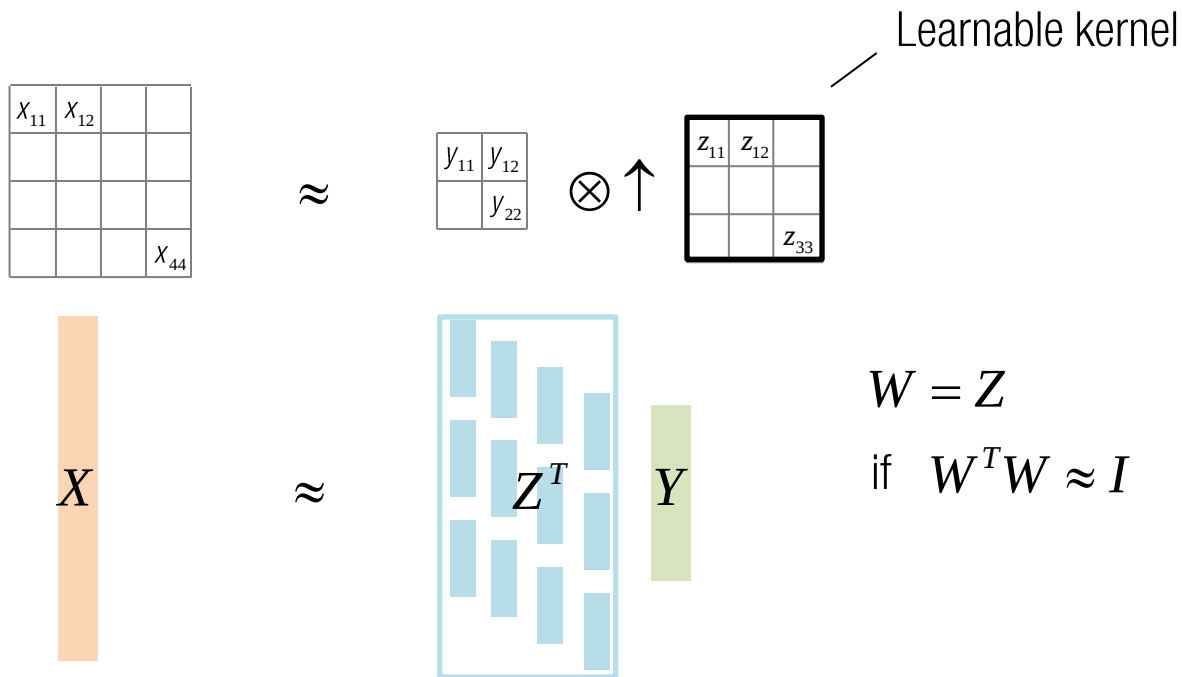
INVERSE OF CONVOLUTION

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

$$\begin{array}{c} \text{X} \end{array} = \begin{array}{c} W^+ \end{array} \begin{array}{c} Y \end{array}$$

Inverse of W does not preserve local connectivity.

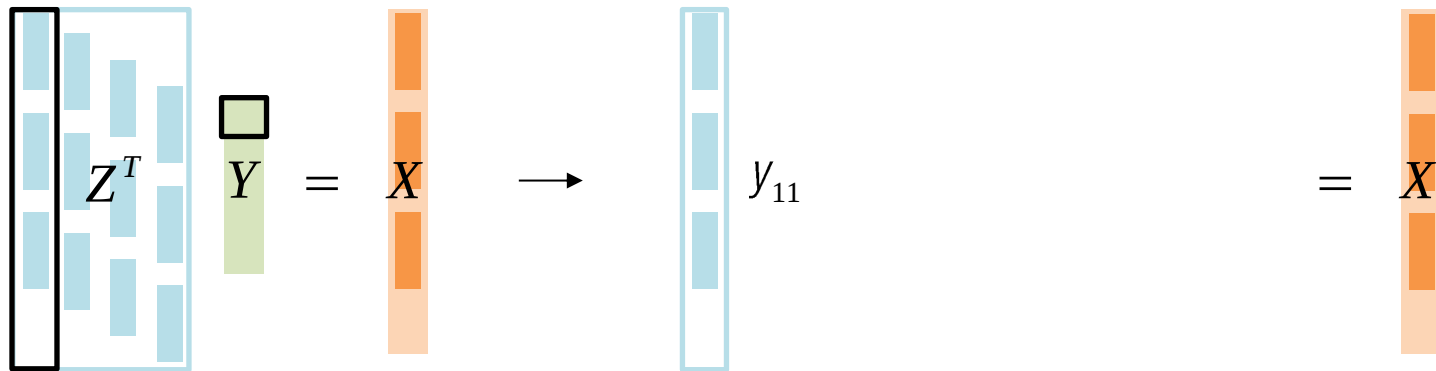
UPCONVOLUTION ~ INVERSE OF CONVOLUTION



Non-zero entries of column encode local connectivity.

UPCONVOLUTION ~ INVERSE OF CONVOLUTION

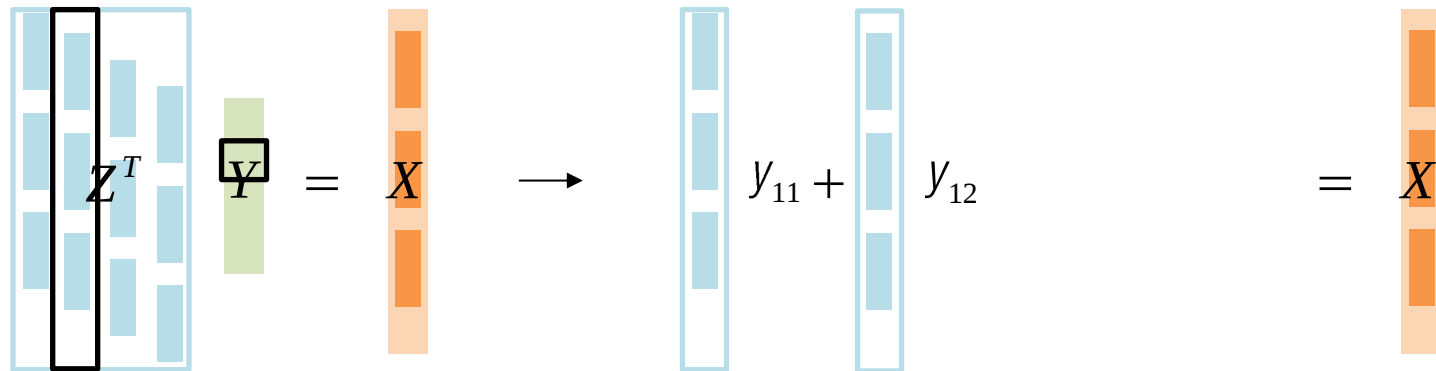
$$\begin{array}{|c|c|c|} \hline x_{11} & x_{12} & \\ \hline & & \\ \hline & & \\ \hline & & x_{44} \\ \hline \end{array} \approx \begin{array}{|c|c|} \hline y_{11} & y_{12} \\ \hline & y_{22} \\ \hline \end{array} \otimes \uparrow \begin{array}{|c|c|c|} \hline z_{11} & z_{12} & \\ \hline & & \\ \hline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



Non-zero entries of column encode local connectivity.

TRANSPPOSED CONVOLUTION ~ INVERSE OF CONVOLUTION

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



Non-zero entries of column encode local connectivity.

TRANPOSED CONVOLUTION ~ INVERSE OF CONVOLUTION

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

$$\begin{array}{|c|c|c|c|c|c|c|c|} \hline \text{[Blue blocks]} & Y & = & X & \rightarrow & \text{[Blue block]} & y_{11} + & \text{[Blue block]} & y_{12} + & \text{[Blue block]} & y_{21} + & \text{[Blue block]} & y_{22} = & X \\ \hline \end{array}$$

Non-zero entries of column encode local connectivity.

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

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

Upconvolution

$$\begin{array}{|c|c|c|c|} \hline \boxed{x_{11}} & \boxed{x_{12}} & & \\ \hline & & & \\ \hline & & & \\ \hline & & & x_{44} \\ \hline \end{array} \otimes \begin{array}{|c|c|} \hline w_{11} & w_{12} \\ \hline & \\ \hline & w_{13} \\ \hline \end{array} = \begin{array}{|c|c|} \hline \boxed{y_{11}} & y_{12} \\ \hline & y_{22} \\ \hline \end{array}$$

Cf) convolution

$$\begin{array}{|c|c|c|c|c|} \hline \text{blue bar} & \text{blue bar} & \text{blue bar} & \text{blue bar} & \text{blue bar} \\ \hline \end{array} Z^T \quad Y = X$$

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

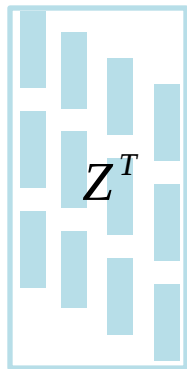
x_{11} is dependent only on y_{11} .

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

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

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

Upconvolution



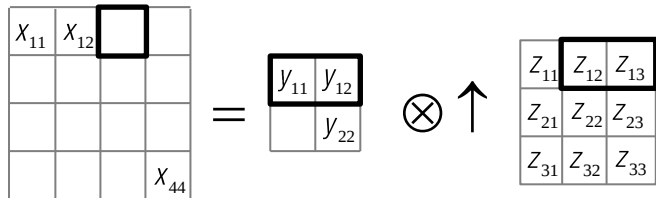
=



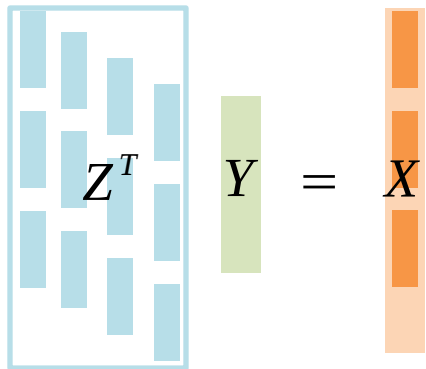
$$x_{11} = z_{11}y_{11}$$

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

TRANSPOSED CONVOLUTION AS UPCONVOLUTION



Upconvolution



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

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

$$x_{13} = z_{13}y_{11} + z_{12}y_{12}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

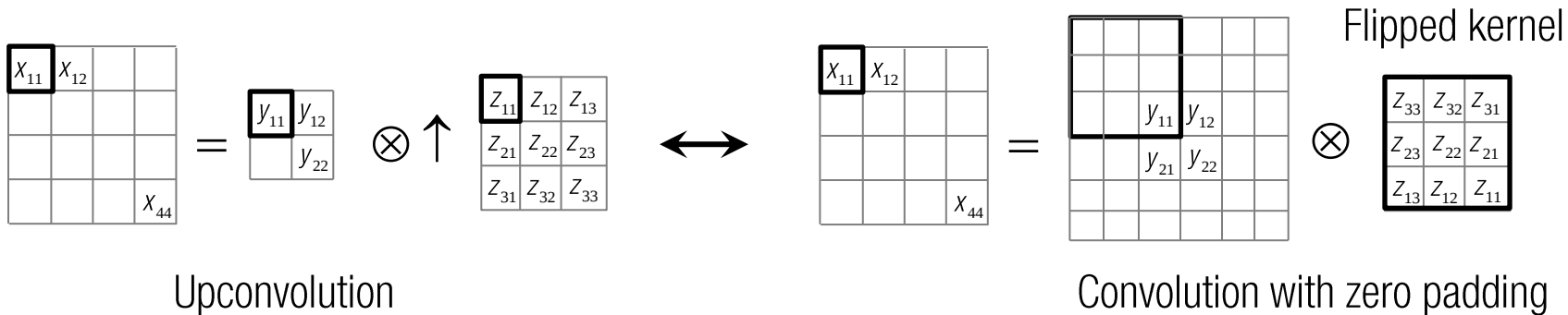
$$\begin{array}{|c|c|c|c|} \hline x_{11} & x_{12} & & \boxed{} \\ \hline & & & \\ \hline & & & \\ \hline & & & \\ \hline & & & x_{44} \\ \hline \end{array} = \begin{array}{|c|c|} \hline y_{11} & \boxed{y_{12}} \\ \hline & y_{22} \\ \hline \end{array} \otimes \uparrow \begin{array}{|c|c|c|} \hline z_{11}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

Upconvolution

$$\begin{array}{|c|c|c|c|c|c|} \hline \text{blue bar} & \text{blue bar} & \text{blue bar} & \text{blue bar} & \text{blue bar} & \text{blue bar} \\ \hline \end{array} Z^T \quad Y = X$$

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

TRANSPOSED CONVOLUTION AS UPCONVOLUTION

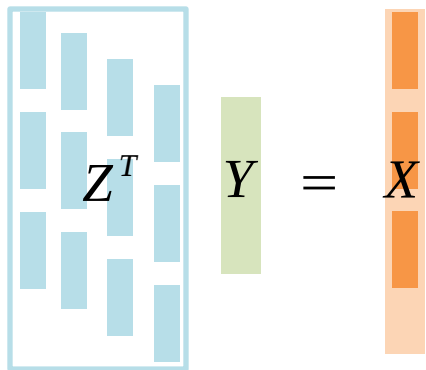
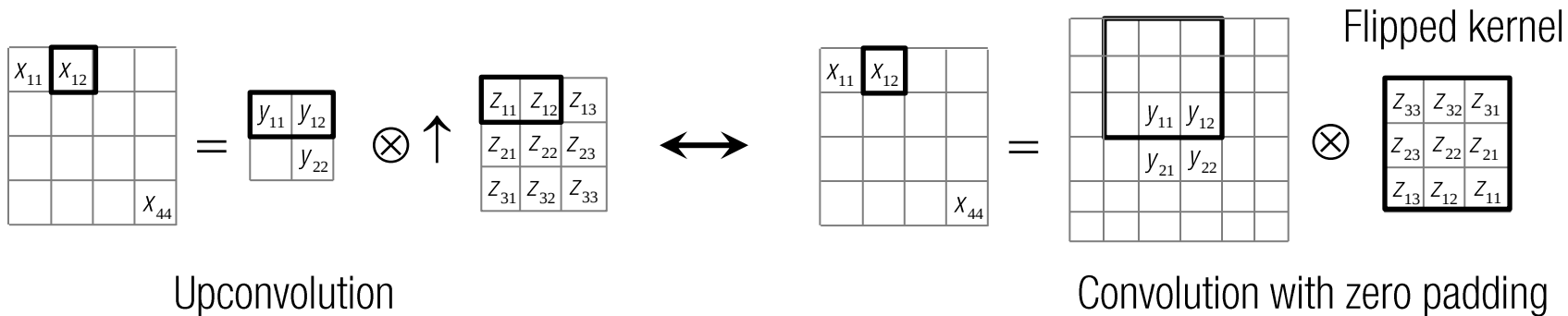


A vector equation illustrating the relationship between the input Y , the kernel Z^T , and the output X . The input Y is represented by a green vertical bar, the kernel Z^T by a blue vertical bar with diagonal elements, and the output X by an orange vertical bar. The equation is $Y = X$, with the kernel Z^T positioned between them. To the right, the specific element relationship is given as $x_{11} = z_{11}y_{11}$.

$$Y = X$$

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

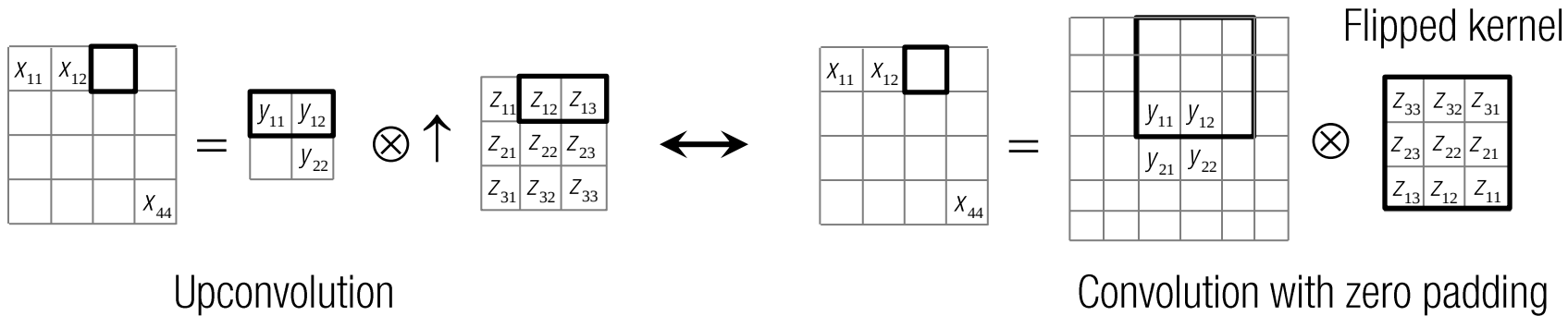
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



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

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

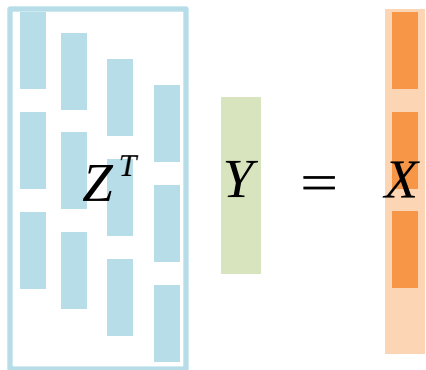
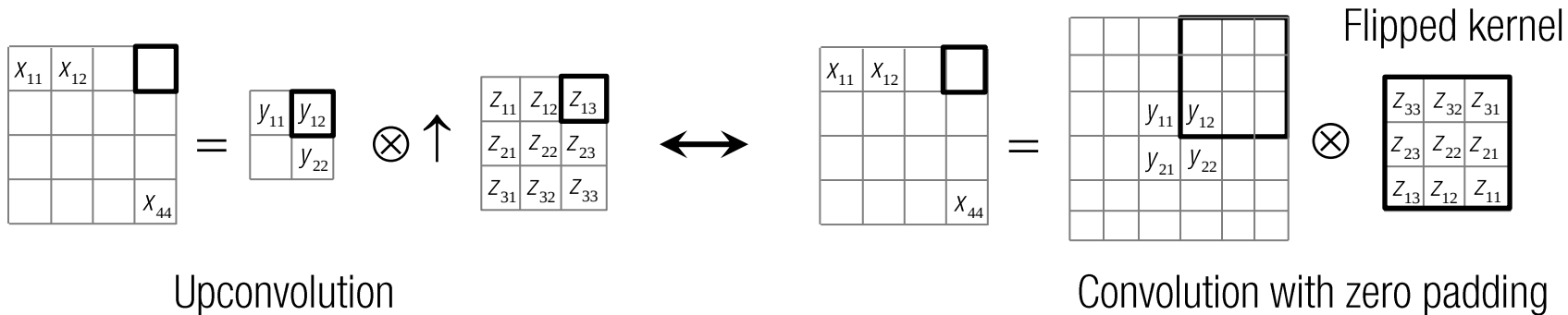
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



A diagram showing a matrix equation $Y = Z^T X$. The matrix Z^T is represented by a light blue box containing several vertical blue bars of varying heights. The matrix Y is represented by a single vertical green bar. The matrix X is represented by a single vertical orange bar. The equation is shown as $Y = X$ with Z^T placed to the left of Y .

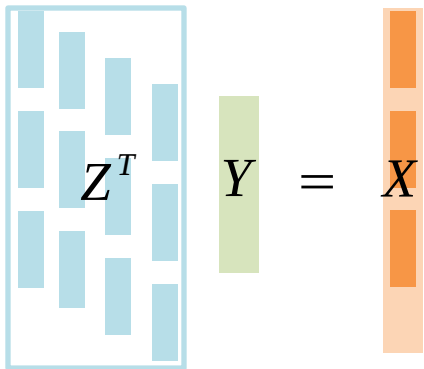
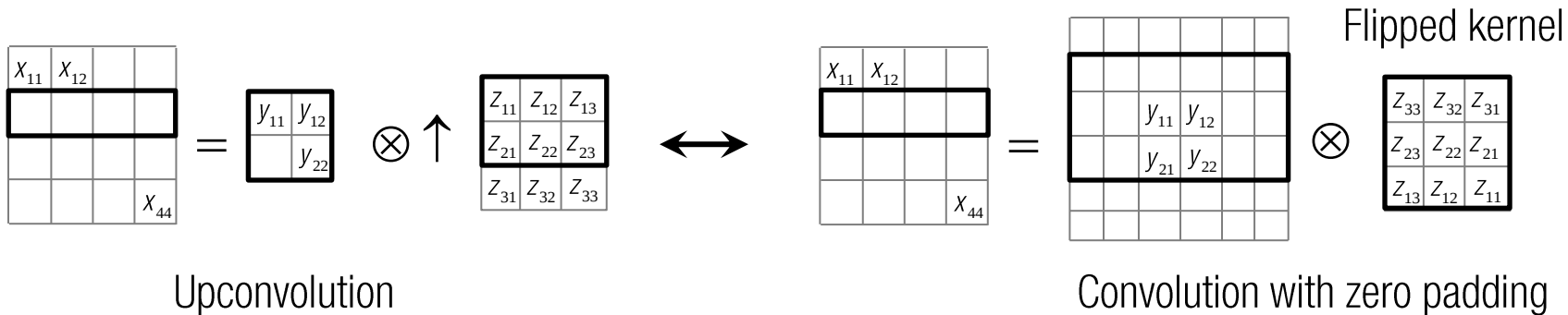
$$\begin{aligned} x_{11} &= z_{11}y_{11} \\ x_{12} &= z_{12}y_{11} + z_{11}y_{12} \\ x_{13} &= z_{13}y_{11} + z_{12}y_{12} \end{aligned}$$

TRANSPOSED CONVOLUTION AS UPCONVOLUTION



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

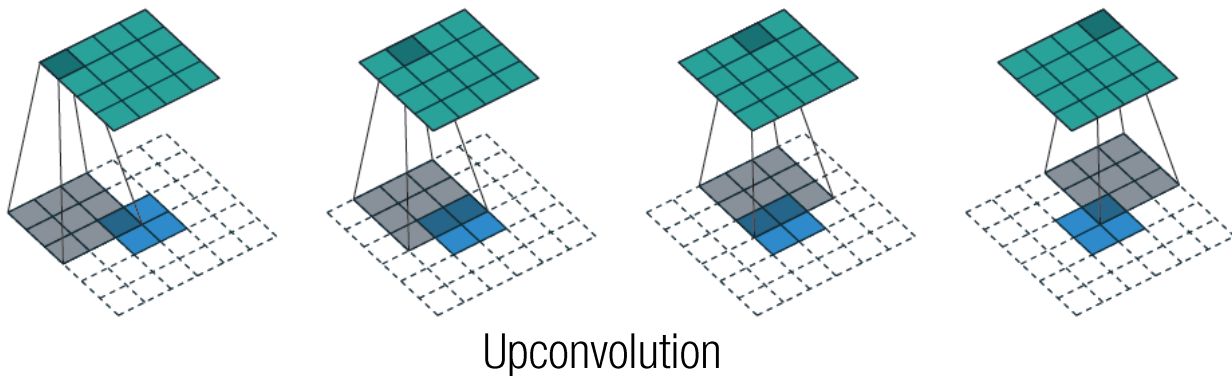
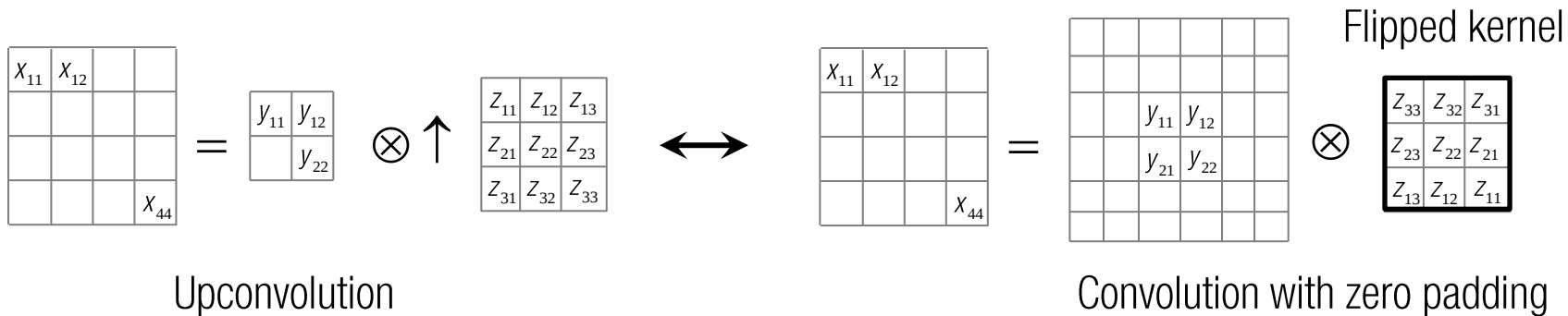
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



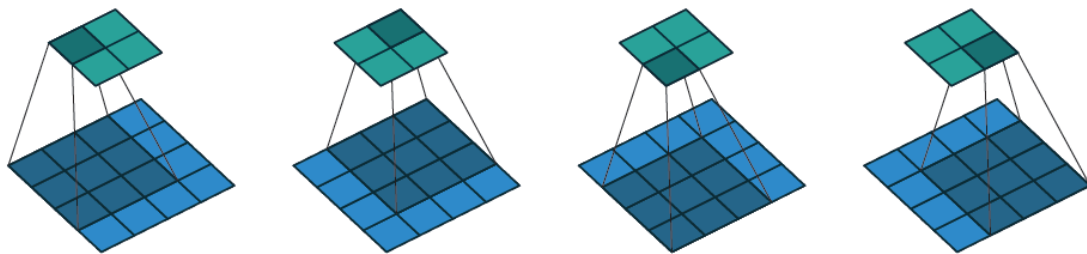
$$\begin{aligned} x_{11} &= z_{11}y_{11} \\ x_{12} &= z_{12}y_{11} + z_{11}y_{12} \\ x_{13} &= z_{13}y_{11} + z_{12}y_{12} \\ x_{14} &= z_{13}y_{12} \end{aligned}$$

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

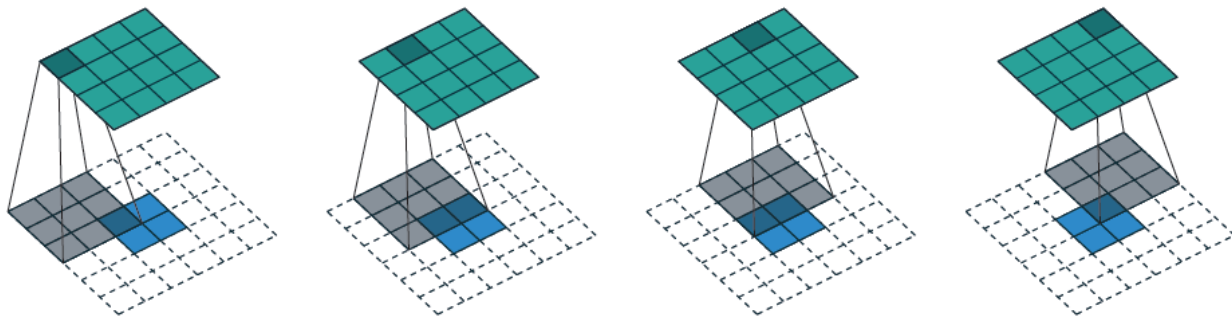
TRANSPOSED CONVOLUTION AS UPCONVOLUTION



DUALITY: CONVOLUTION VS. UPCONVOLUTION

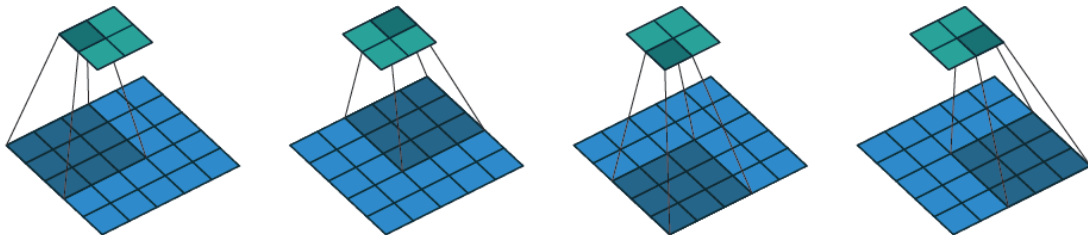


Convolution (3x3 kernel, no zero padding, 1 stride)

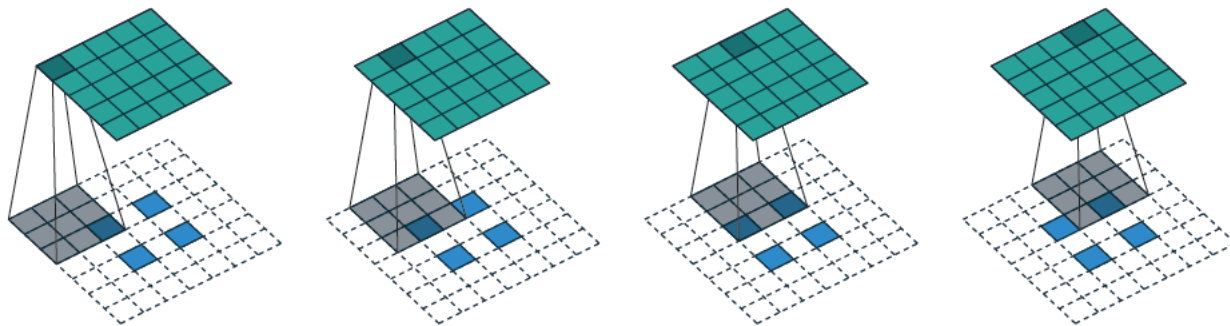


Upconvolution (3x3 kernel, 2x2 zero padding, 1 stride)

DUALITY: CONVOLUTION VS. UPCONVOLUTION

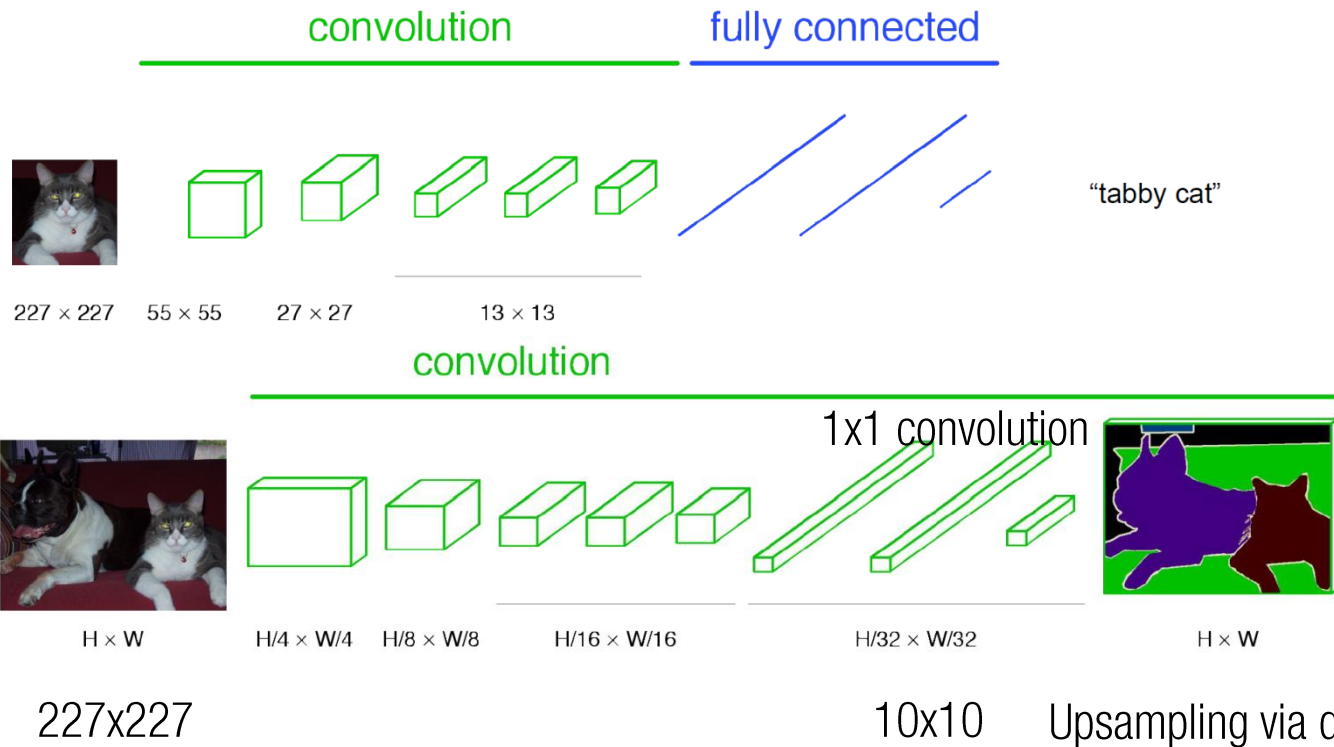


Convolution (3x3 kernel, no zero padding, 2 stride)

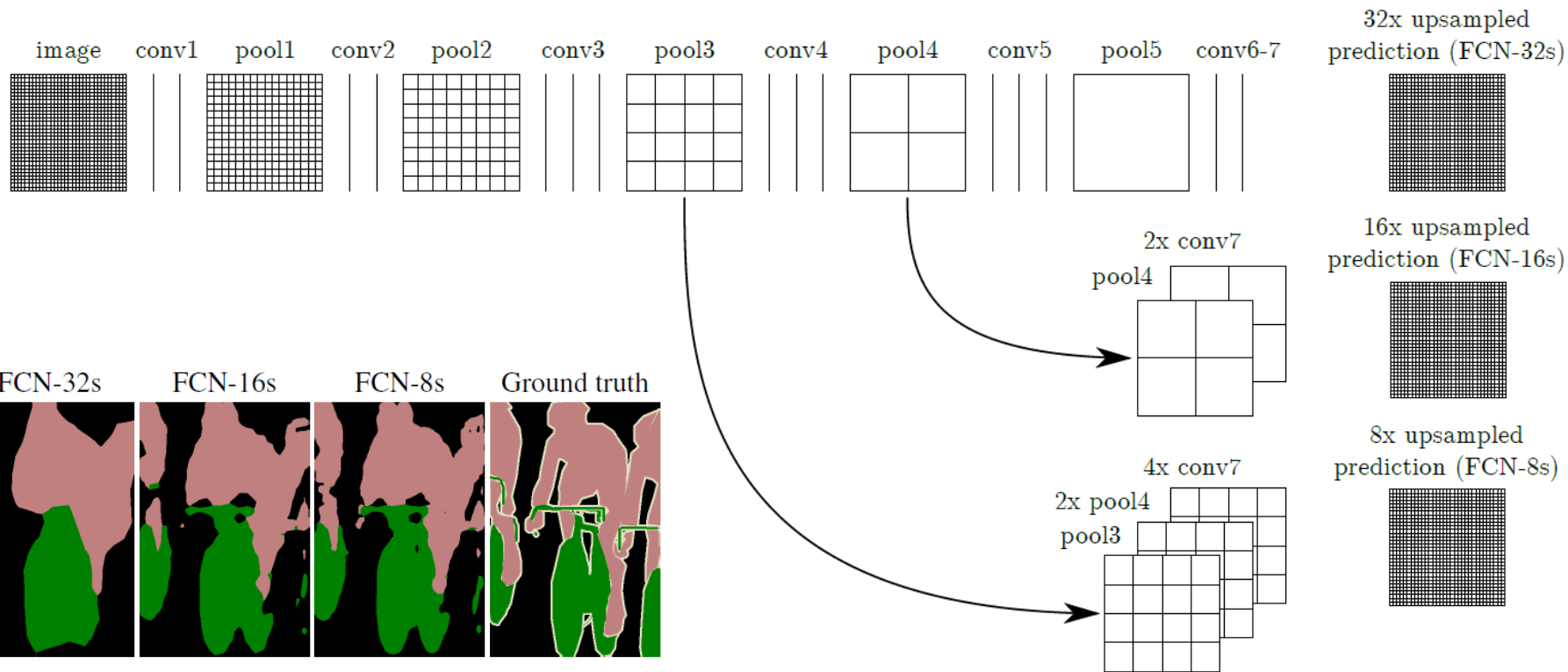


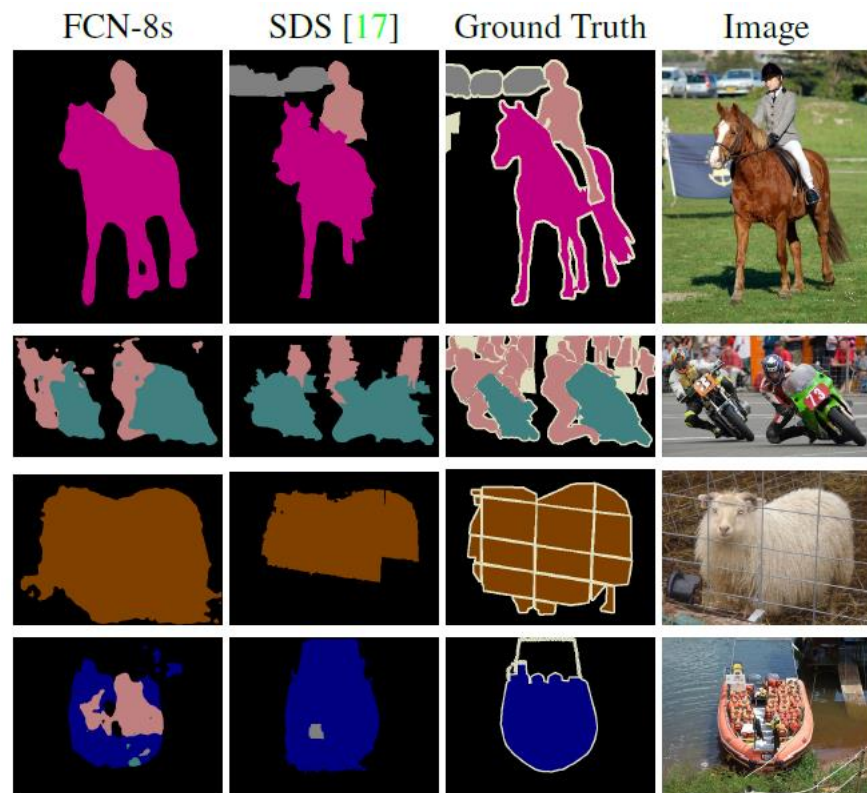
Upconvolution (3x3 kernel, 2x2 zero padding, 1 zero between input, 2 stride)

FULLY CONVOLUTIONAL NETWORK

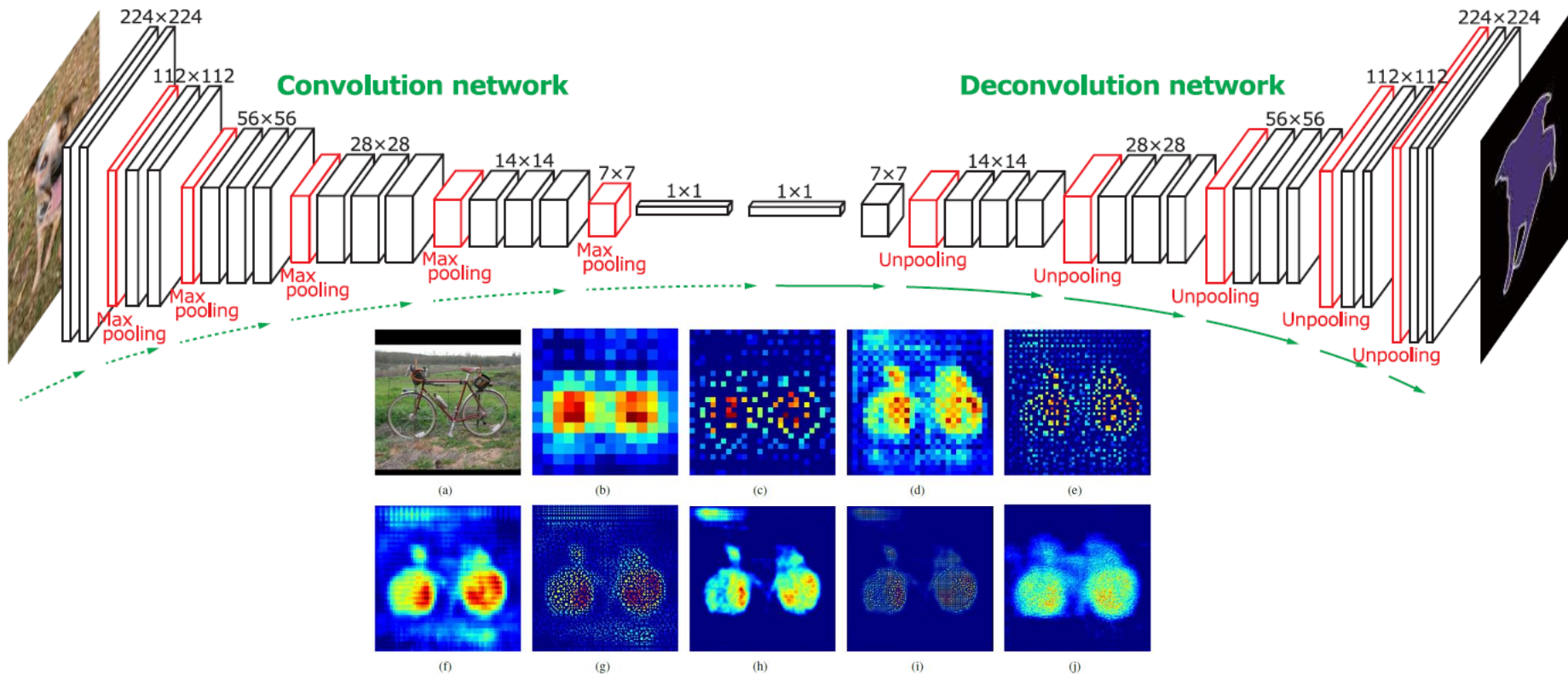


FULLY CONVOLUTIONAL NETWORK





DEEPER UPCONVOLUTIONAL LAYERS



Input image

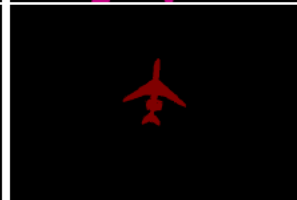
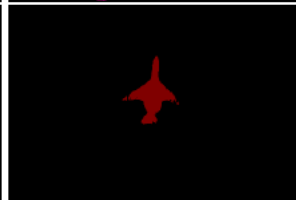
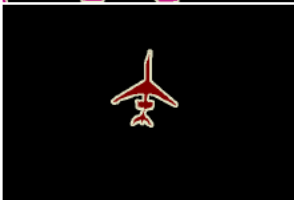
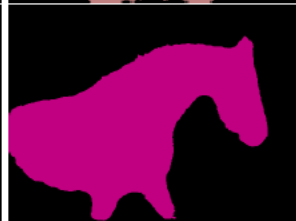
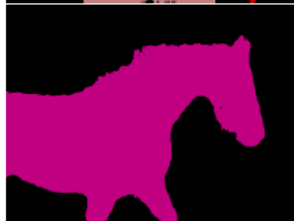
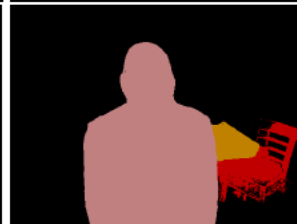
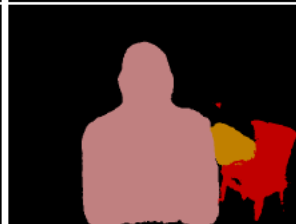
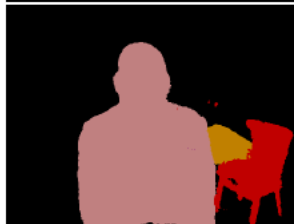
Ground-truth

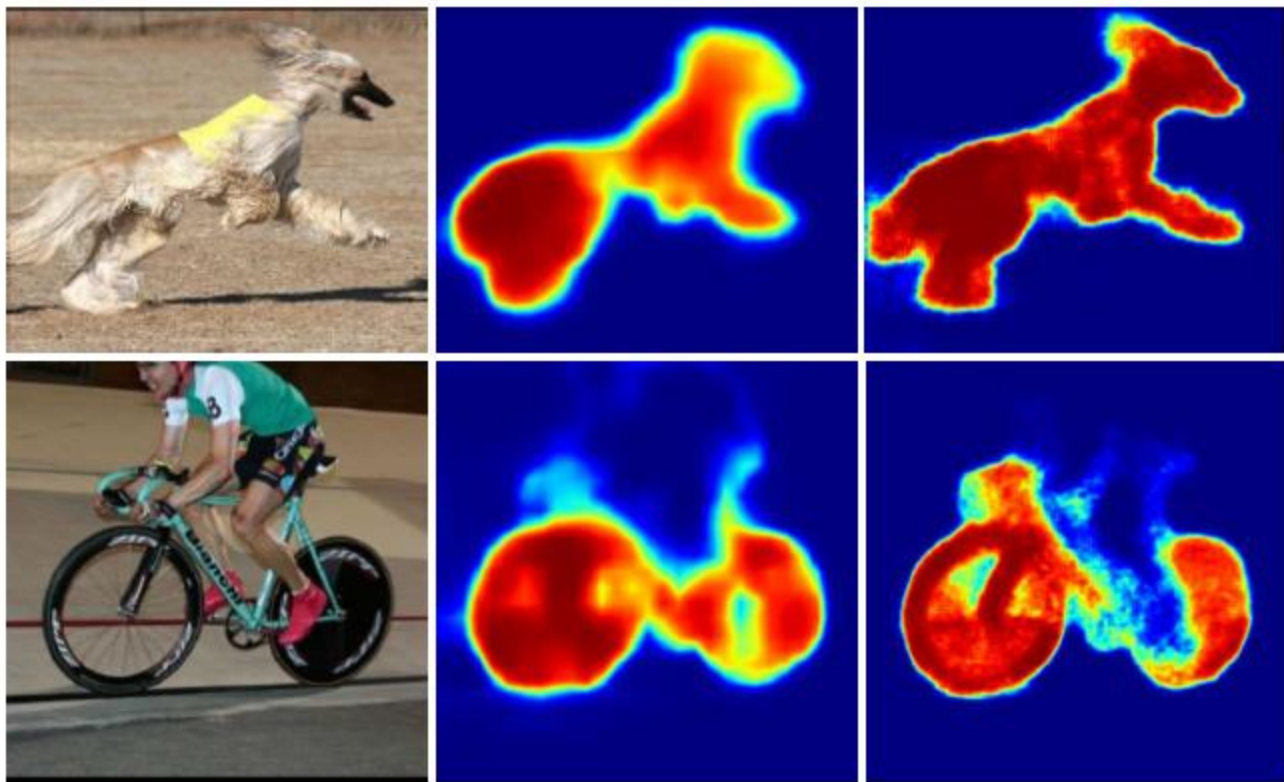
FCN

DeconvNet

EDeconvNet

EDeconvNet+CRF



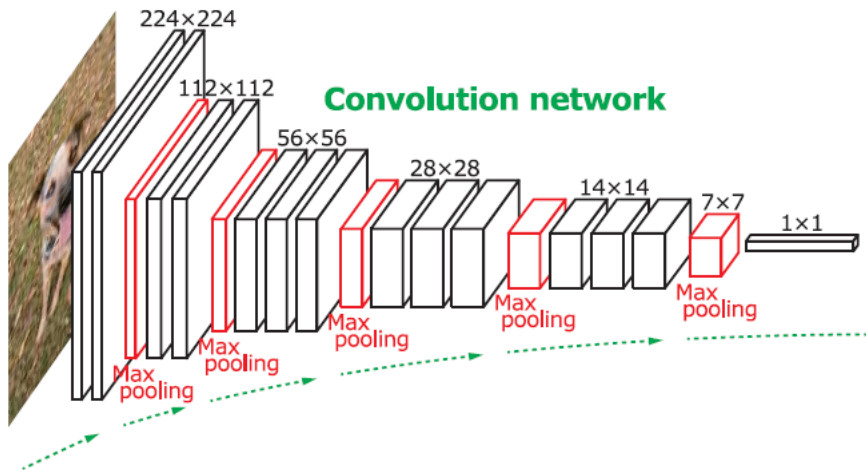


(a) Input image

(b) FCN-8s

(c) Ours

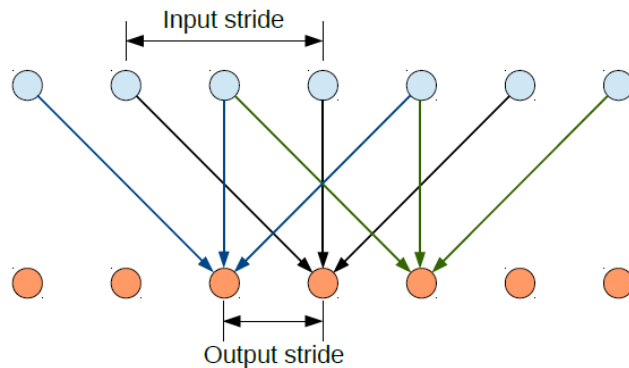
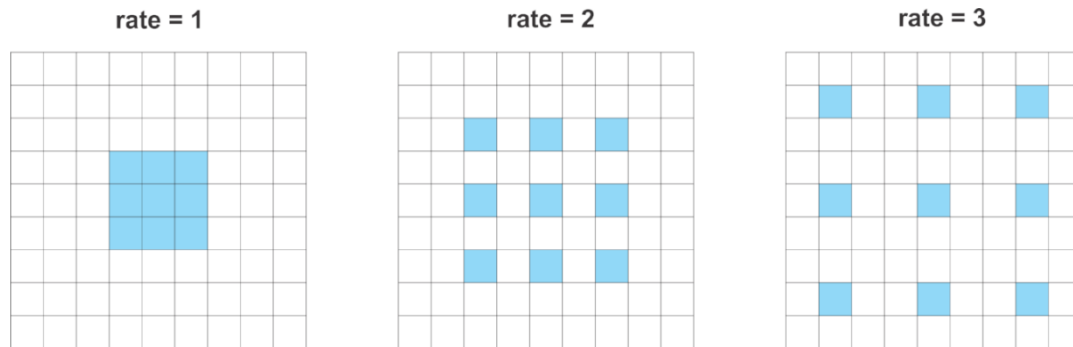
RECEPTIVE FIELD VS. RESOLUTION



Desired properties:

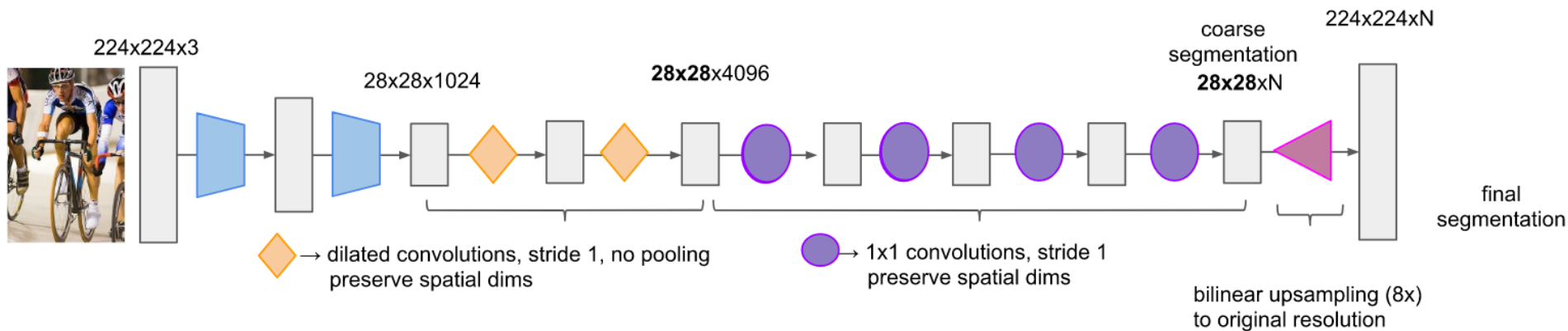
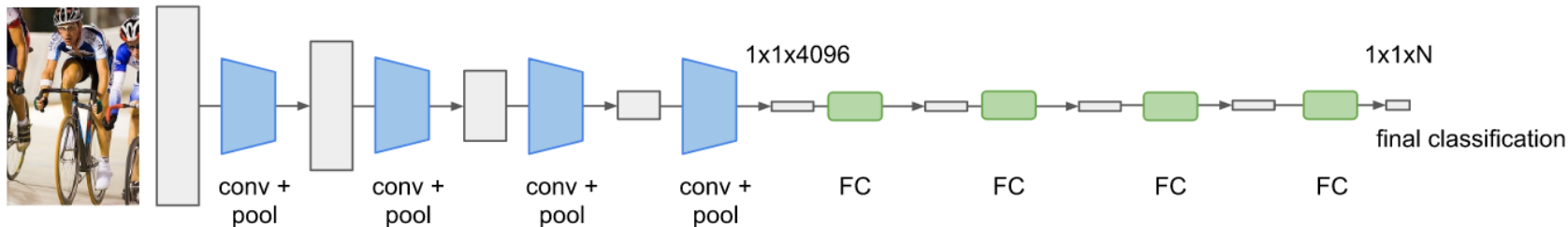
- Larger receptive field (bigger spatial context) $\longrightarrow$ reducing resolution
- Higher output resolution $\longrightarrow$ reducing receptive field

DEEPLAB: DILATED CONVOLUTION

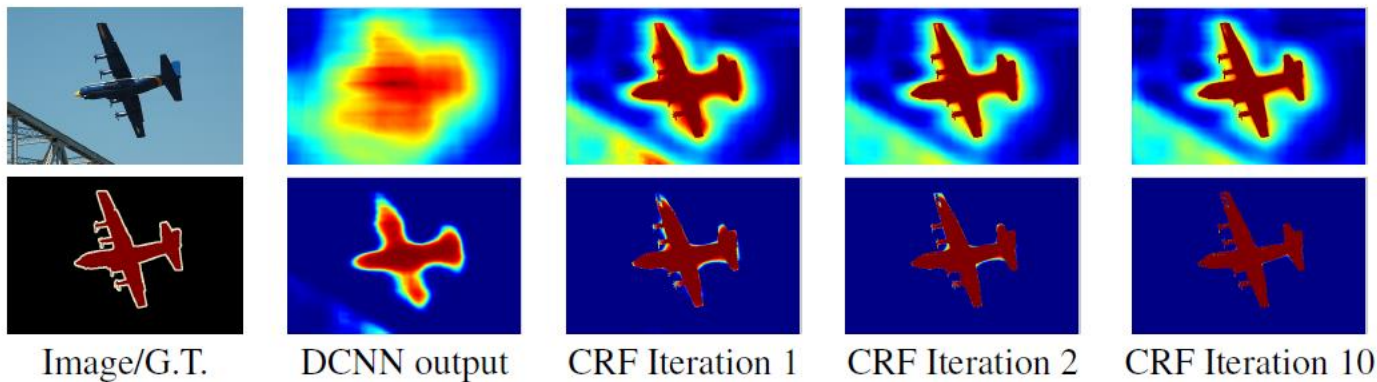


Enlarging receptive field without increasing reducing resolution

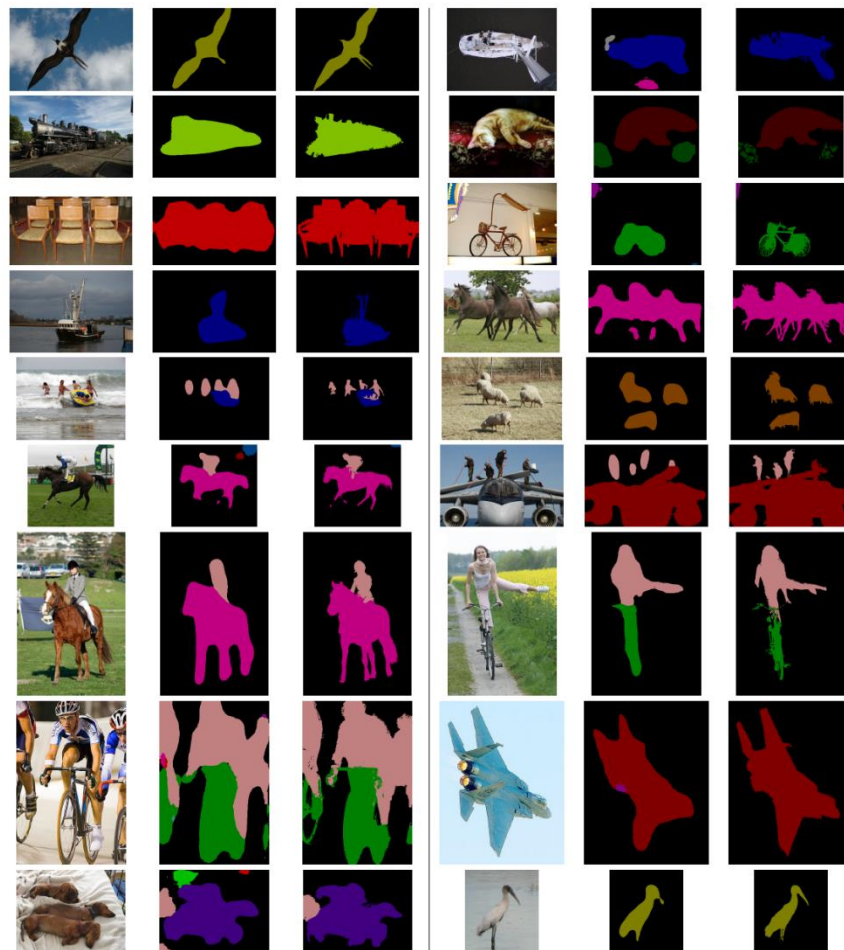
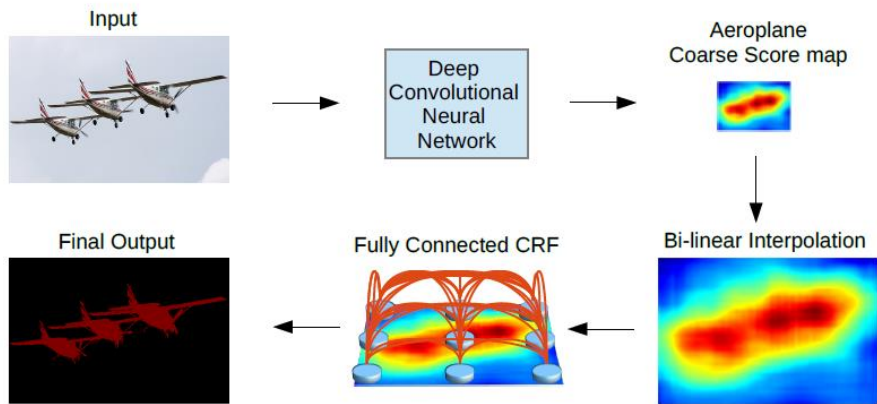
DEEPLAB: DILATED CONVOLUTION



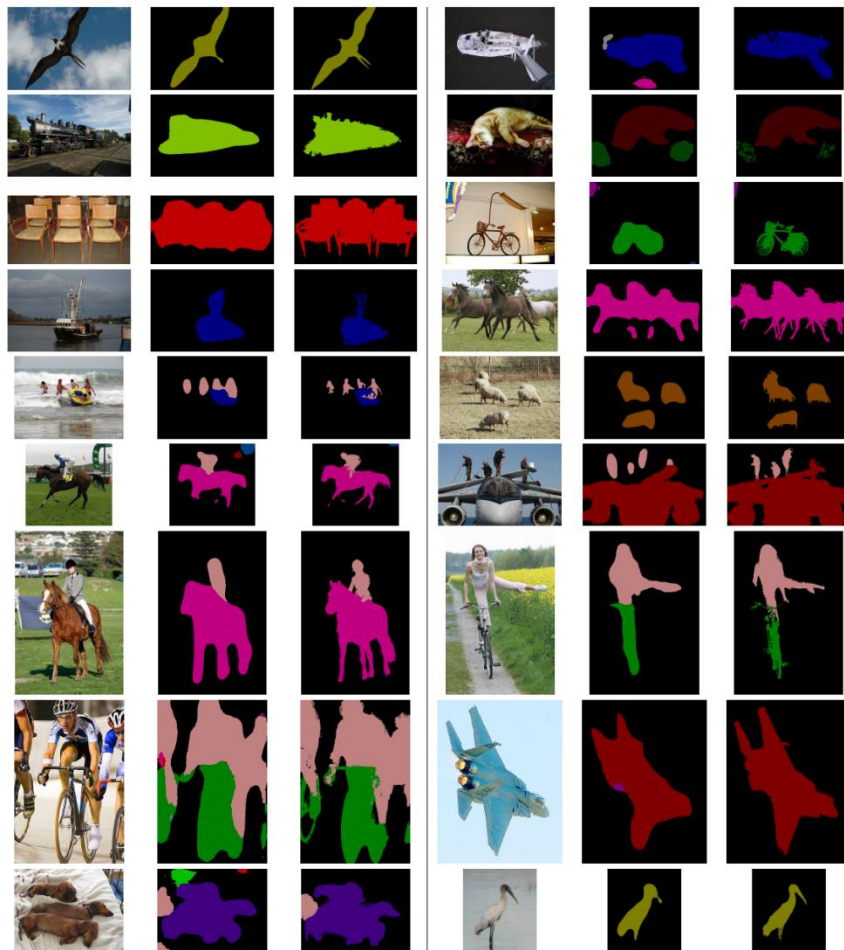
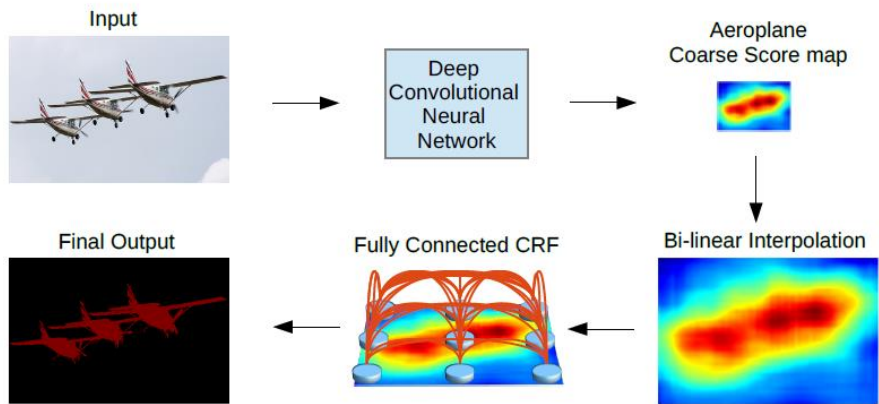
DEEPLAB: CRF POST-PROCESSING



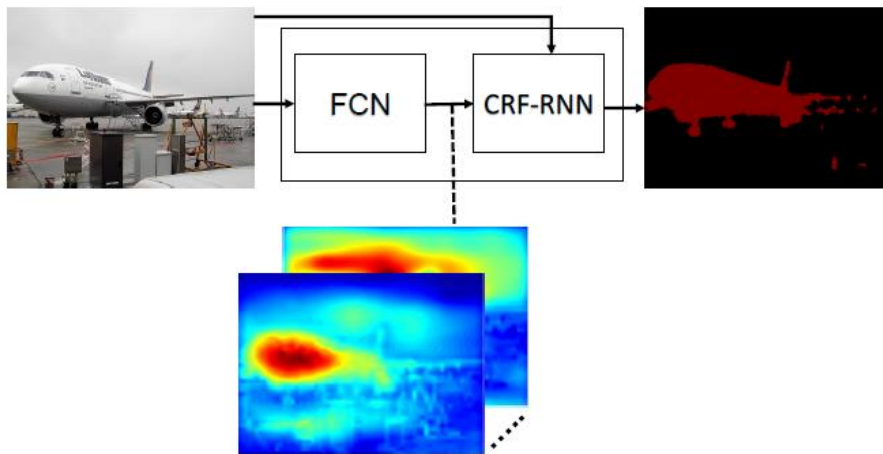
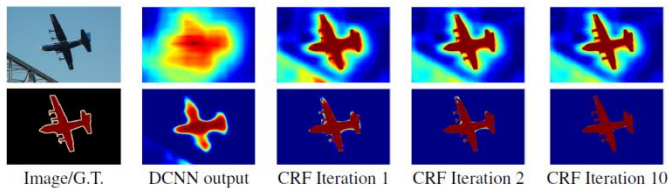
DEEPLAB



DEEPLAB



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



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

