165B
Machine Learning
Learning CNN

Lei Li (leili@cs)
UCSB

Acknowledgement: Slides borrowed from Bhiksha Raj's 11485 and Mu Li & Alex Smola's 157 courses on Deep Learning, with modification

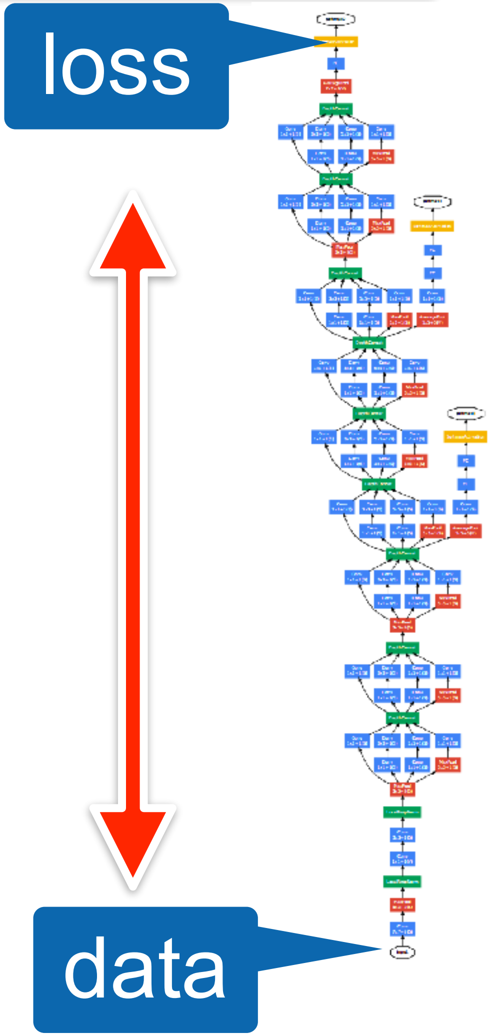
Recap

- AlexNet
  - 11 layers, bigger convolusion
  - ReLu, Dropout, preprocessing
- VGG
  - Bigger and deeper AlexNet (repeated VGG blocks)
  - VGG-16 and VGG-19
- ResNet
  - 50 or 153 layers
  - Residual connection

Batch Normalization

Batch Normalization

- Loss occurs at last layer
  - Last layers learn quickly
- Data is inserted at first layer
  - Input layers change - **everything** changes
  - Last layers need to relearn many times
  - Slow convergence
- This is like covariate shift
  - The distribution of each layer shift across over training process



Batch Normalization

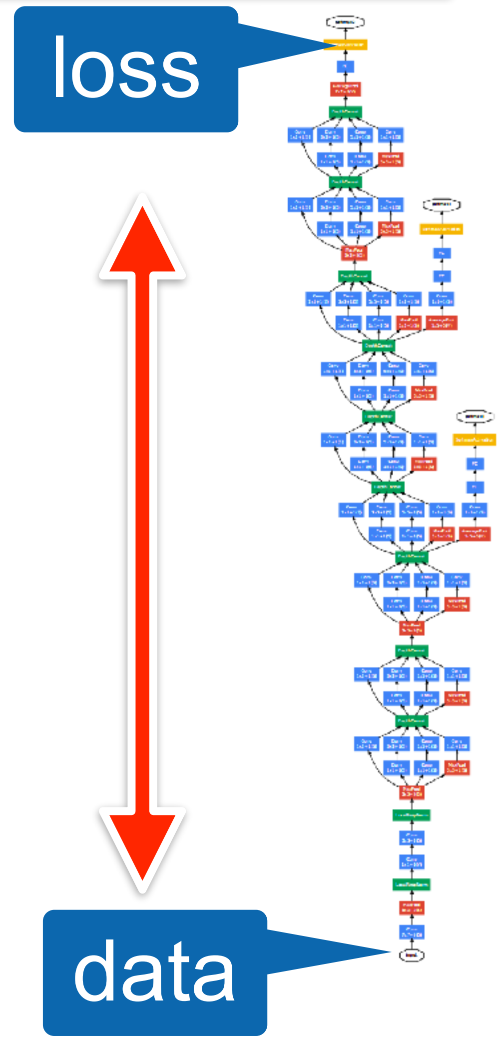
- For each layer, compute mean and variance

$$\mu_B = \frac{1}{|B|}\sum_{i\in B} x_i \text{ and } \sigma_B^2 = \frac{1}{|B|}\sum_{i\in B}(x_i - \mu_B)^2 + \epsilon$$

and adjust it separately

$$x_{i+1} = \gamma\frac{x_i - \mu_B}{\sigma_B} + \beta$$

- γ and β are learnable parameters



Sergey Ioffe, Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015

This was the original motivation …

What Batch Norms really do

- Doesn't really reduce covariate shift (Lipton et al., 2018)
- Regularization by noise injection

Random offset

$$x_{i+1} = \gamma \frac{x_i - \hat{\mu}_B}{\hat{\sigma}_B} + \beta$$

Random scale

  - Random shift per minibatch
  - Random scale per minibatch
- No need to mix with dropout (both are capacity control)
- Ideal minibatch size of 64 to 256

Code

```
torch.nn.BatchNorm1d(num_features)
```

```
torch.nn.BatchNorm2d(num_features)
>>> m = nn.BatchNorm2d(100)
>>> input = torch.randn(20, 100, 32, 32)
>>> output = m(input)
```

Quiz

- https://edstem.org/us/courses/16390/lessons/29420/slides/168304

Learning CNN

Recap: Learning the Model

- Finding the parameter θ to minimize the empirical risk over training data $D = \{(x_n, y_n)\}_{n=1}^{N}$
  $$\hat{\theta} \leftarrow \arg\min_{\theta} L(\theta) = \frac{1}{N} \sum_{n} \ell(y_n, f(x_n; \theta))$$
- Update rule: $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Gradient Descent

```
learning rate eta.
1.set initial parameter θ ← θ_0
2.for epoch = 1 to maxEpoch or until
  converge:
3.   for each data (x, y) in D:
4.      compute forward y_hat = f(x; θ)
5.      compute gradient g = ∂err(y_hat, y)/∂θ using
  back-propagation
6.      total_g += g
7.   update θ = θ − eta * total_g / num_sample
```

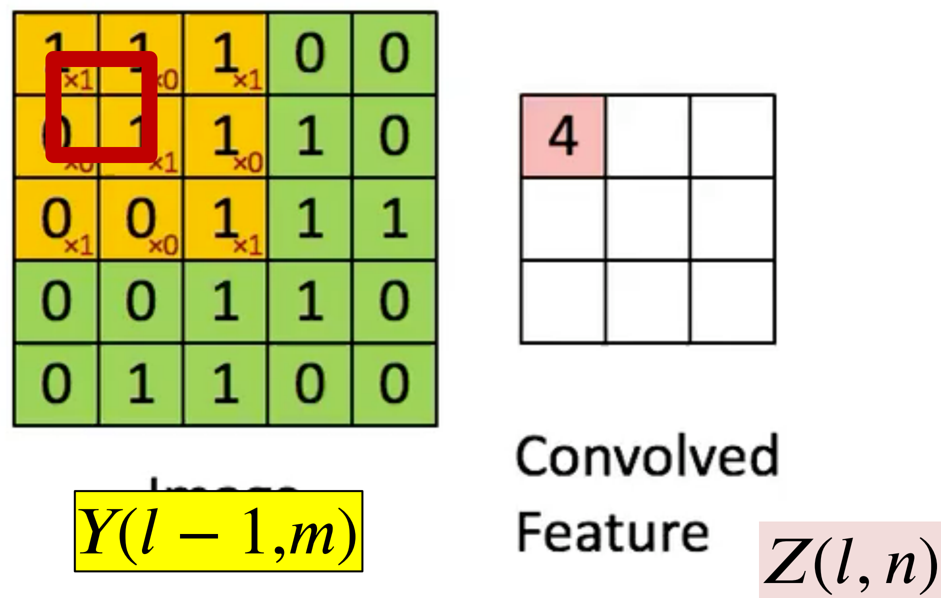
Step 5: $g = \frac{\partial \mathrm{err}(y_{hat}, y)}{\partial \theta}$

Backpropagation for Convolutional layers

- Forward: compute the network output for each layer

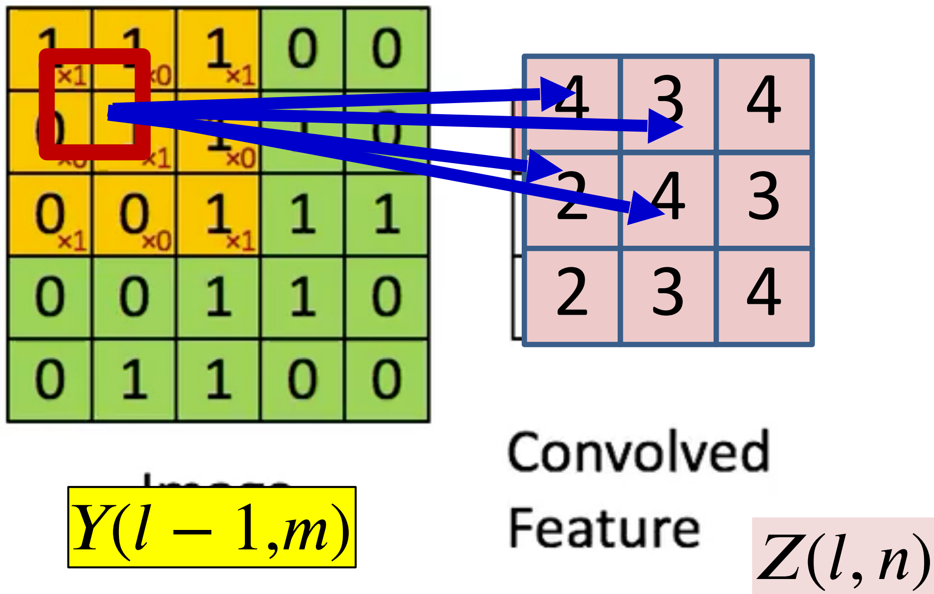
- Backword:
  - How to compute the derivatives w.r.t. the activation (easy, since element-wise)
  - How to compute the derivative w.r.t. input $Y(l-1)$ and and conv kernel $w(l)$
  - FFN layers are already covered as previous

Back-Propagation for Convolutional layer



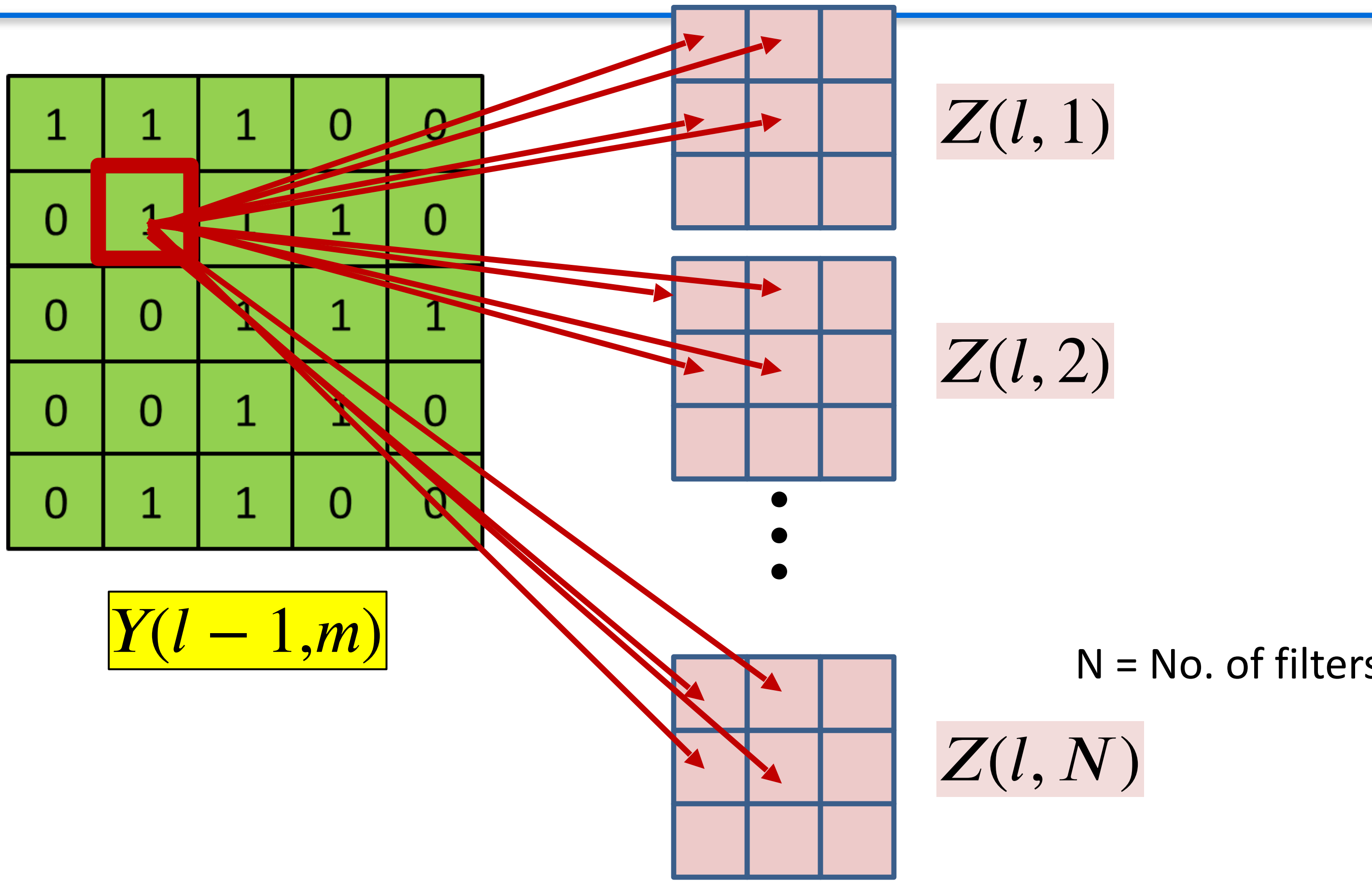
- Each $Y(l-1,m,x,y)$ affects several $z(l,n,x',y')$ terms

BP: Convolutional layer



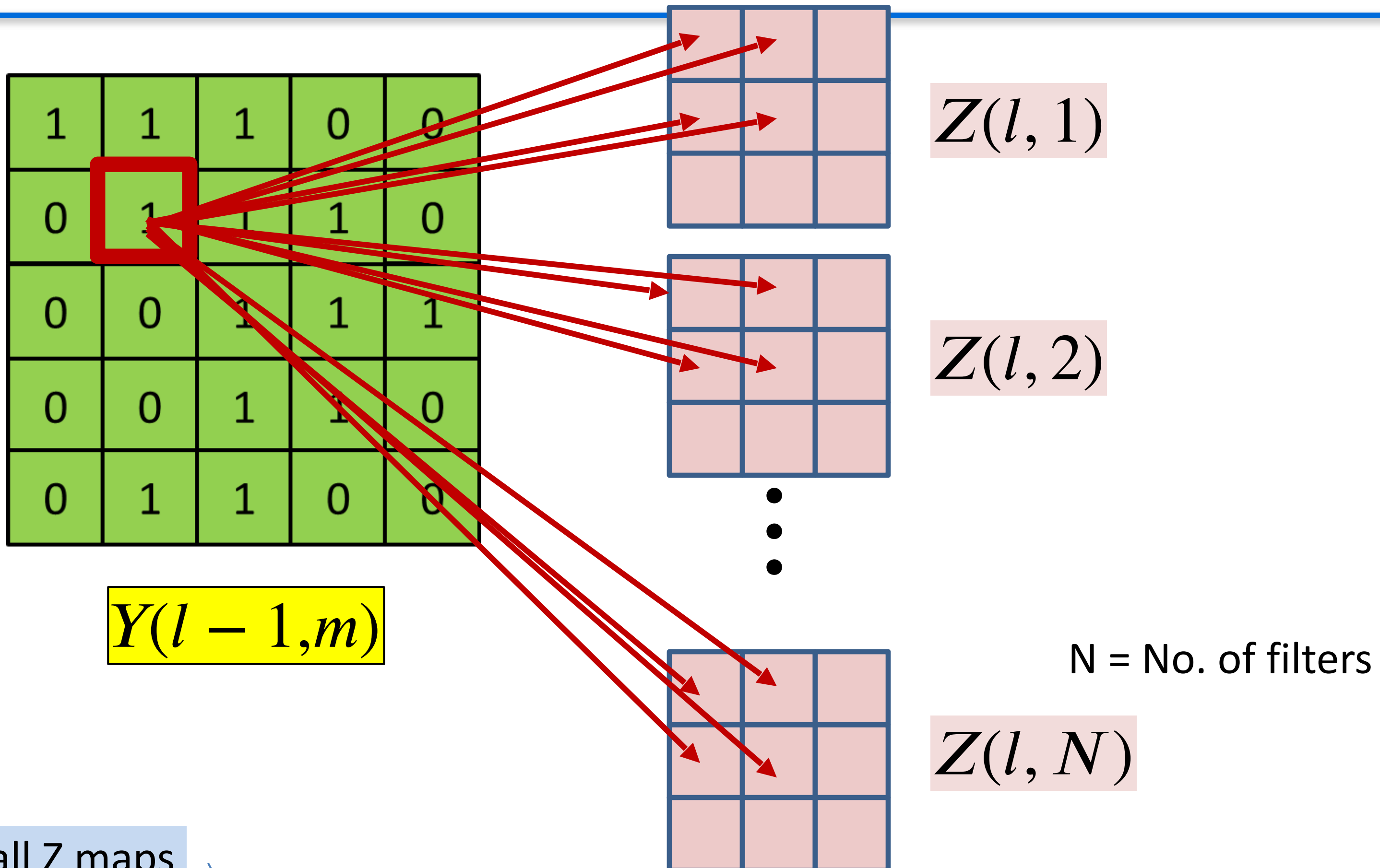
- Each $Y(l-1,m,x,y)$ affects several $z(l,n,x',y')$ terms

BP: Convolutional layer



- Each $Y(l-1,m,x,y)$ affects several $z(l,n,x',y')$ terms
  - Affects terms in *all* l th layer Z maps

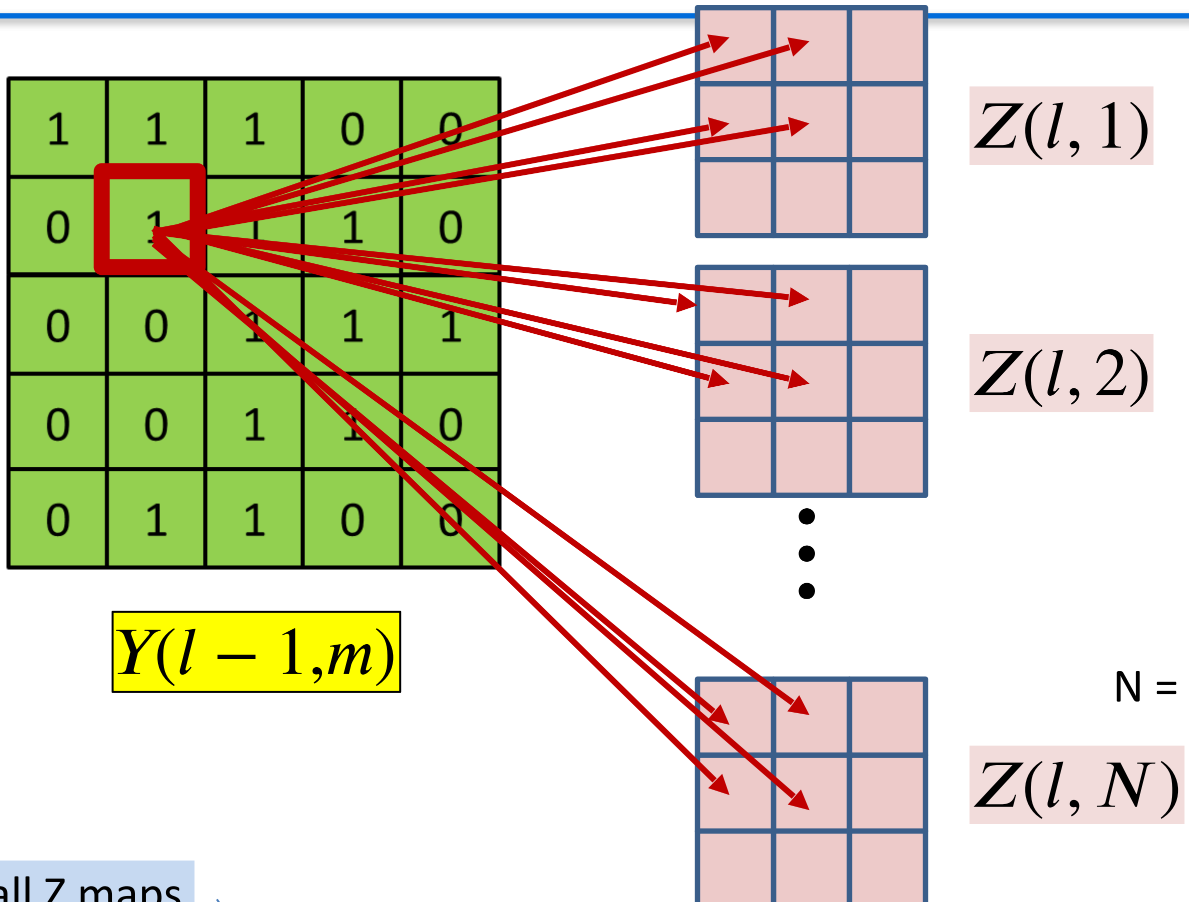
BP: Convolutional layer



Summing over all Z maps

$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_{n} \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} \frac{\partial z(l,n,x',y')}{\partial Y(l-1,m,x,y)}$$

BP: Convolutional layer



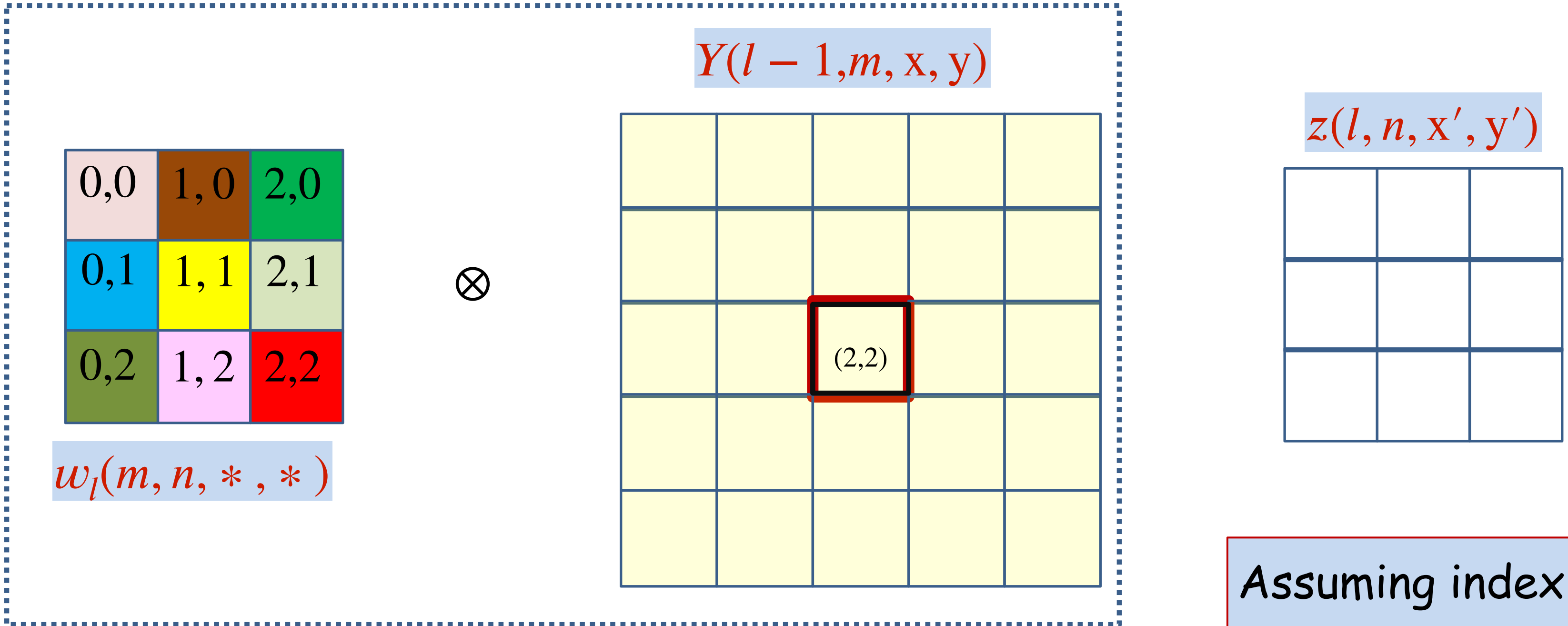
N = No. of filters

Summing over all Z maps

What is this?

$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_{n} \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} \frac{\partial z(l,n,x',y')}{\partial Y(l-1,m,x,y)}$$

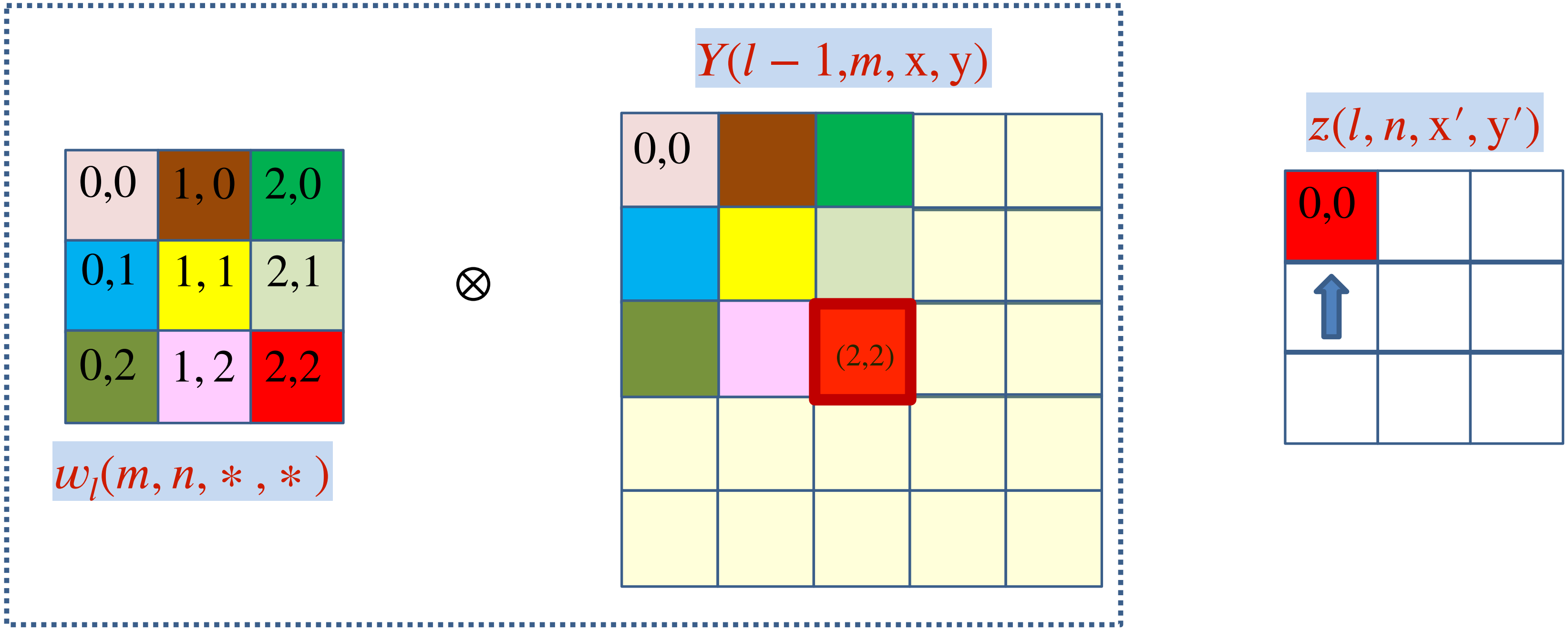
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



Assuming indexing begins at 0

- Compute how *each* x, y in Y influences various locations of z

How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$

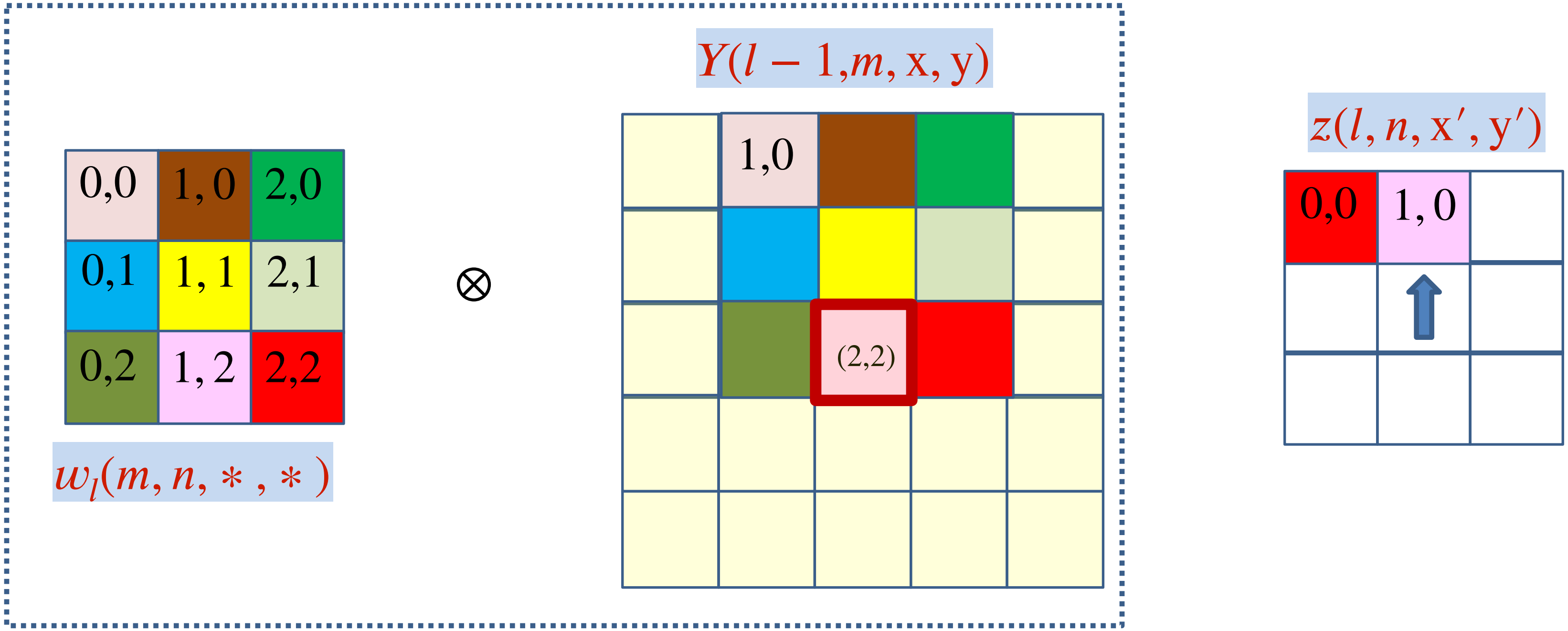


$$z(l, n, 0{,}0) += Y(l-1, m, 2, 2)\, w_l(m, n, 2, 2)$$

- **Note:** The coordinates of $z(l, n)$ and $w_l(m, n)$ sum to the coordinates of $Y(l-1, m)$

$$z(l, n, x', y') += Y(l-1, m, 2{,}2)\, w_l(m, n, 2-x', 2-y')$$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

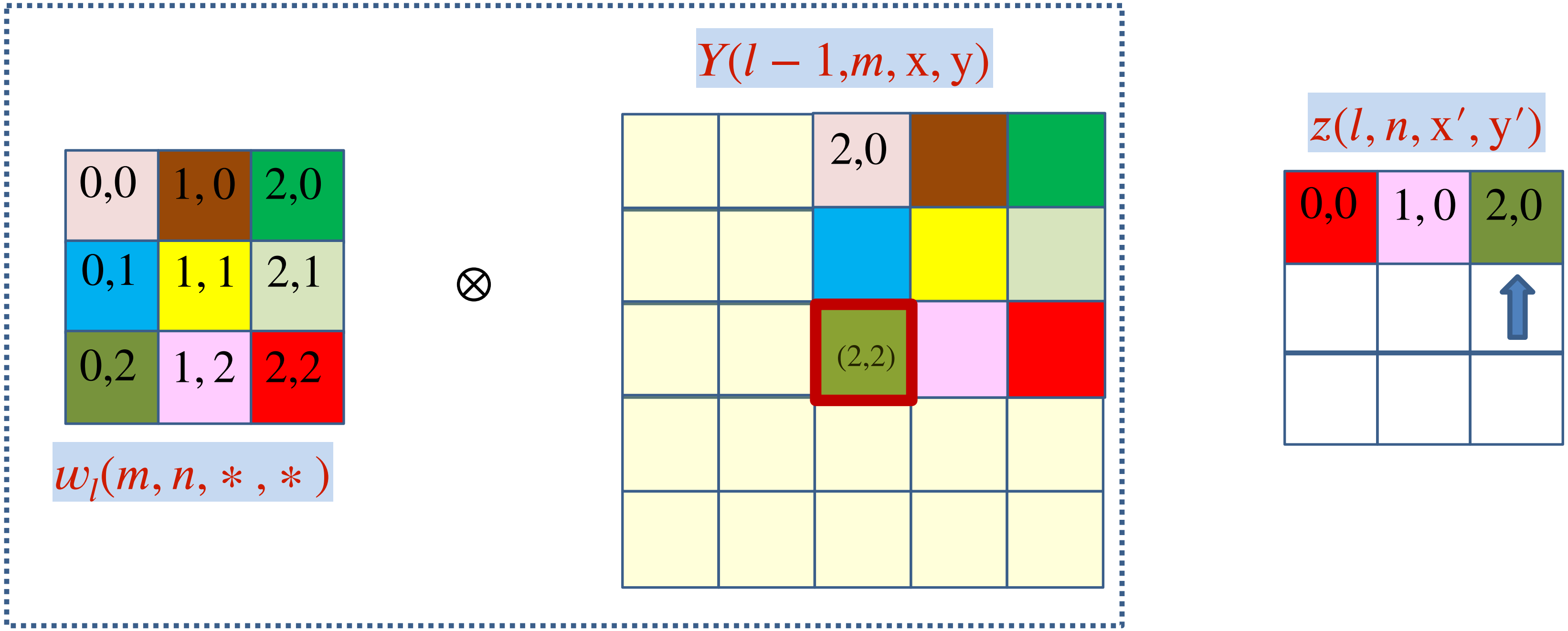


$$z(l,n,1,0) += Y(l-1,m,2,2)w_l(m,n,1,2)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$z\left(l,n,x',y'\right) += Y(l-1,m,2,2)w_l(m,n,2-x',2-y')$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

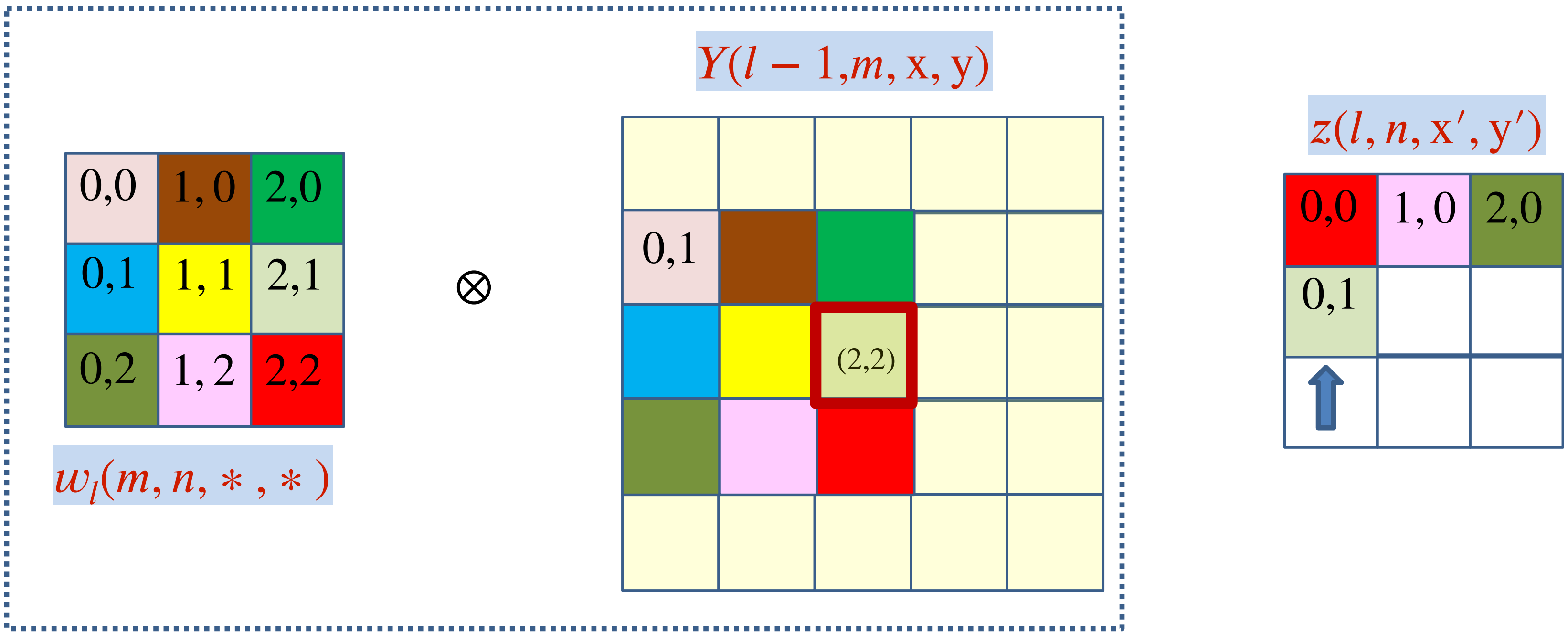


$$z(l,n,2{,}0) += Y(l-1,m,2,2)w_l(m,n,0,2)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$$z\left(l,n,x',y'\right) += Y(l-1,m,2{,}2)w_l(m,n,2-x',2-y')$$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

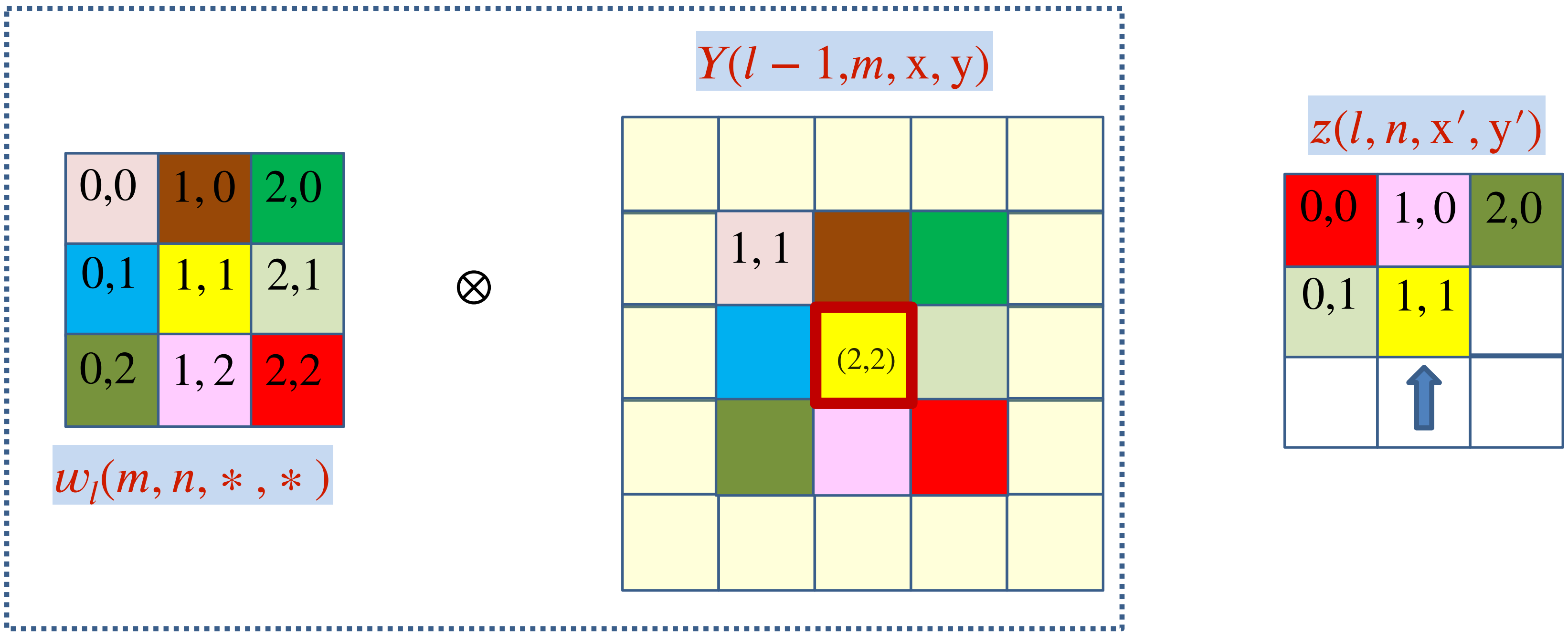


$$z(l,n,0,1) += Y(l-1,m,2,2)w_l(m,n,2,1)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$$z(l,n,x',y') += Y(l-1,m,2,2)w_l(m,n,2-x',2-y')$$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

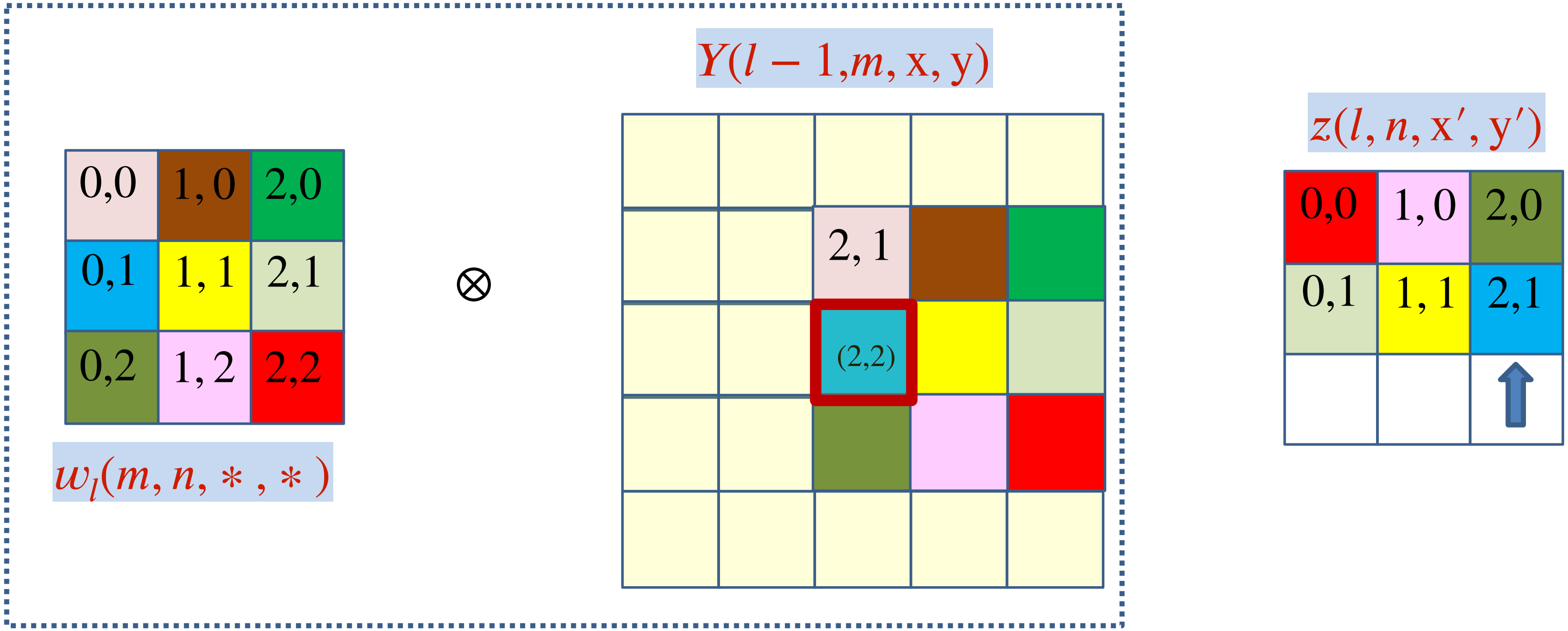


$$z(l,n,1,1) += Y(l-1,m,2,2)w_l(m,n,1,1)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$$z(l,n,x',y') += Y(l-1,m,2,2)w_l(m,n,2-x',2-y')$$

How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$

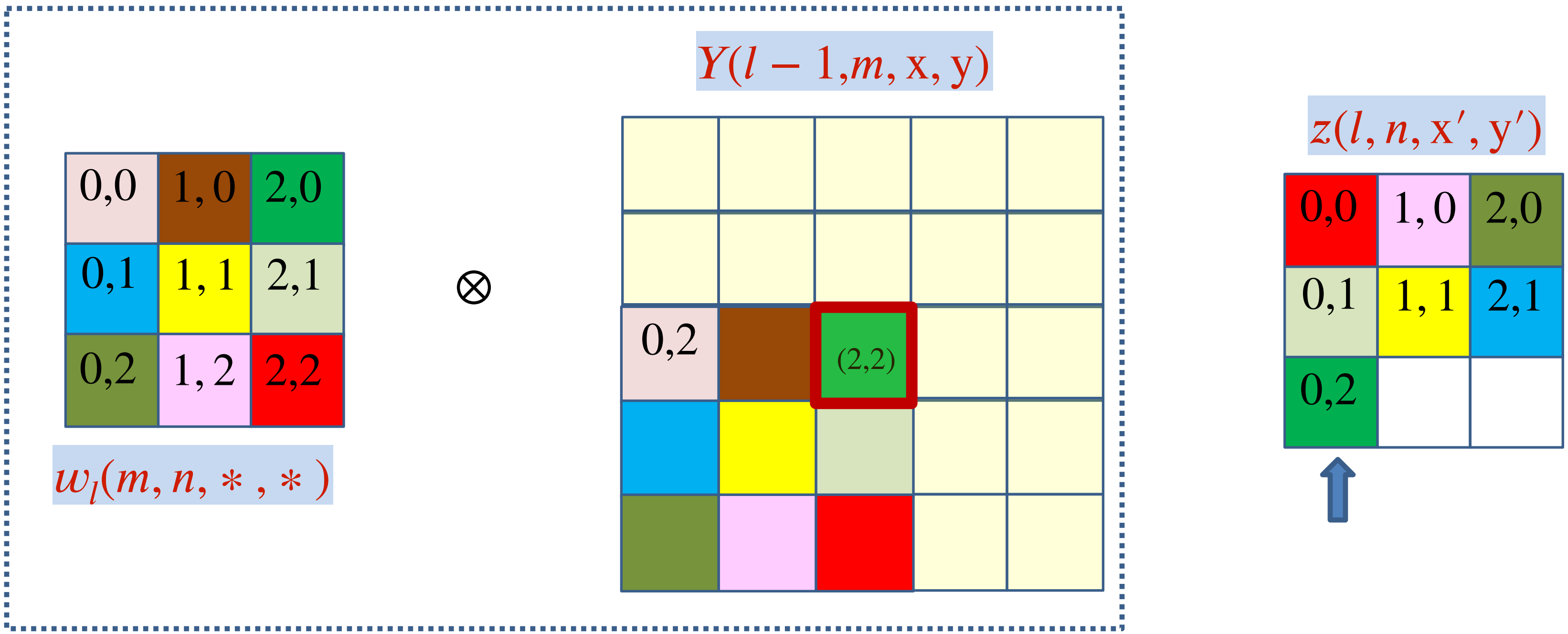


$$z(l, n, 2, 1) += Y(l-1, m, 2, 2)w_l(m, n, 0, 1)$$

- **Note:** The coordinates of $z(l, n)$ and $w_l(m, n)$ sum to the coordinates of $Y(l-1, m)$

$z(l, n, x', y') += Y(l-1, m, 2, 2)w_l(m, n, 2-x', 2-y')$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

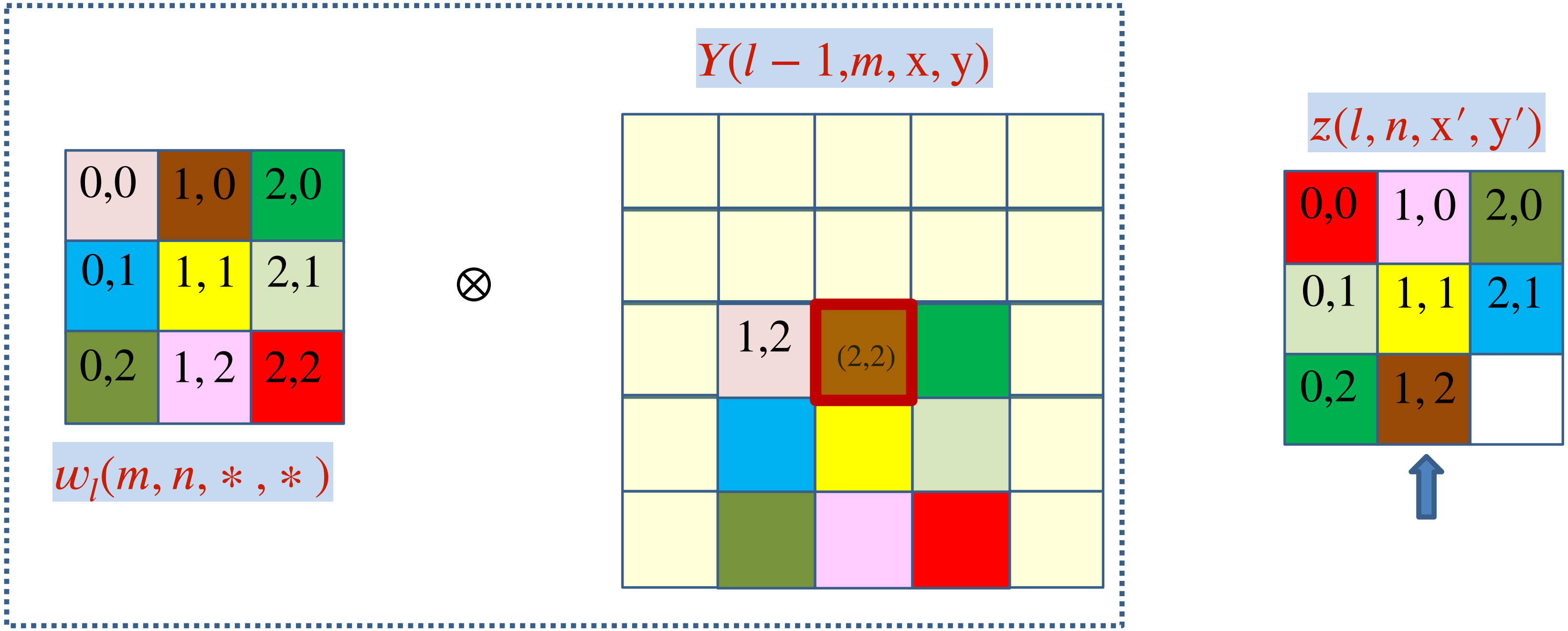


$$z(l,n,0,2) += Y(l-1,m,2,2)w_l(m,n,2,0)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$z\left(l,n,x',y'\right) += Y(l-1,m,2,2)w_l(m,n,2-x',2-y')$

How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$

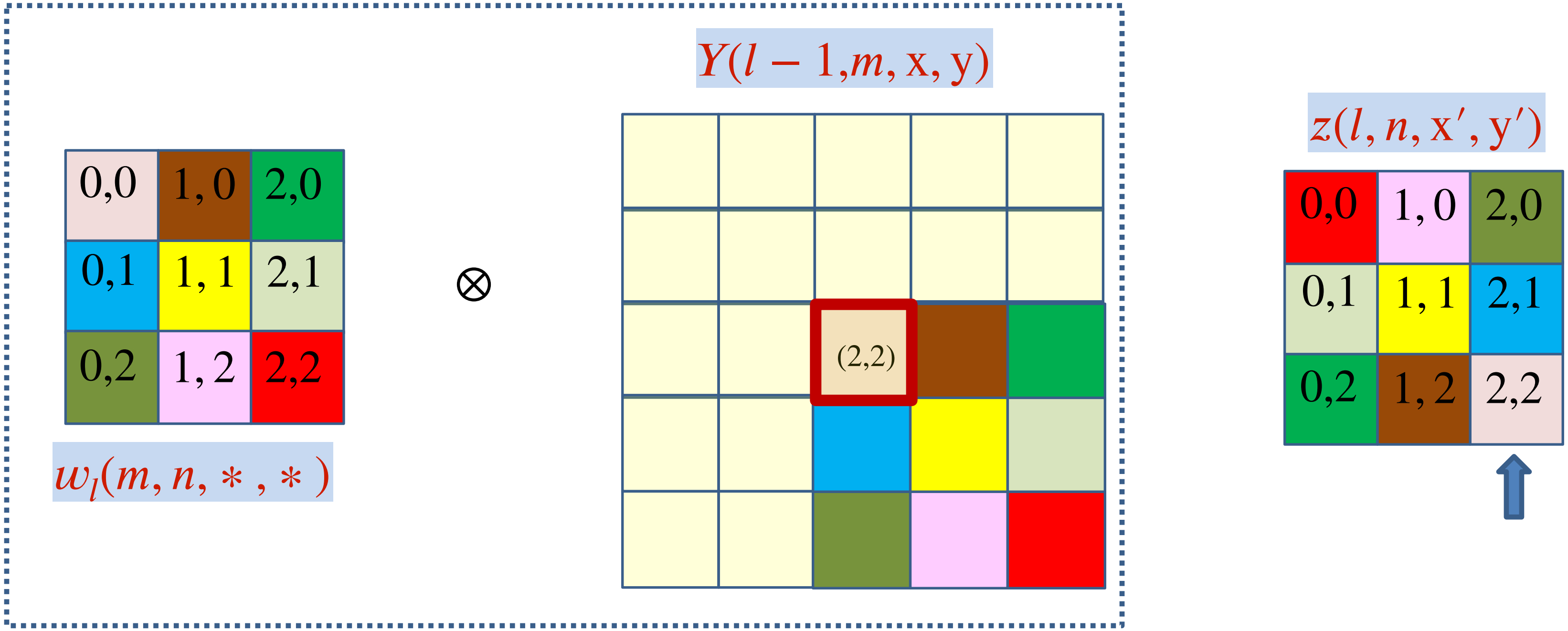


$$z(l, n, 1,2) += Y(l-1, m, 2, 2)w_l(m, n, 2, 1)$$

- **Note:** The coordinates of $z(l, n)$ and $w_l(m, n)$ sum to the coordinates of $Y(l-1, m)$

$z(l, n, x', y') += Y(l-1, m, 2,2)w_l(m, n, 2-x', 2-y')$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

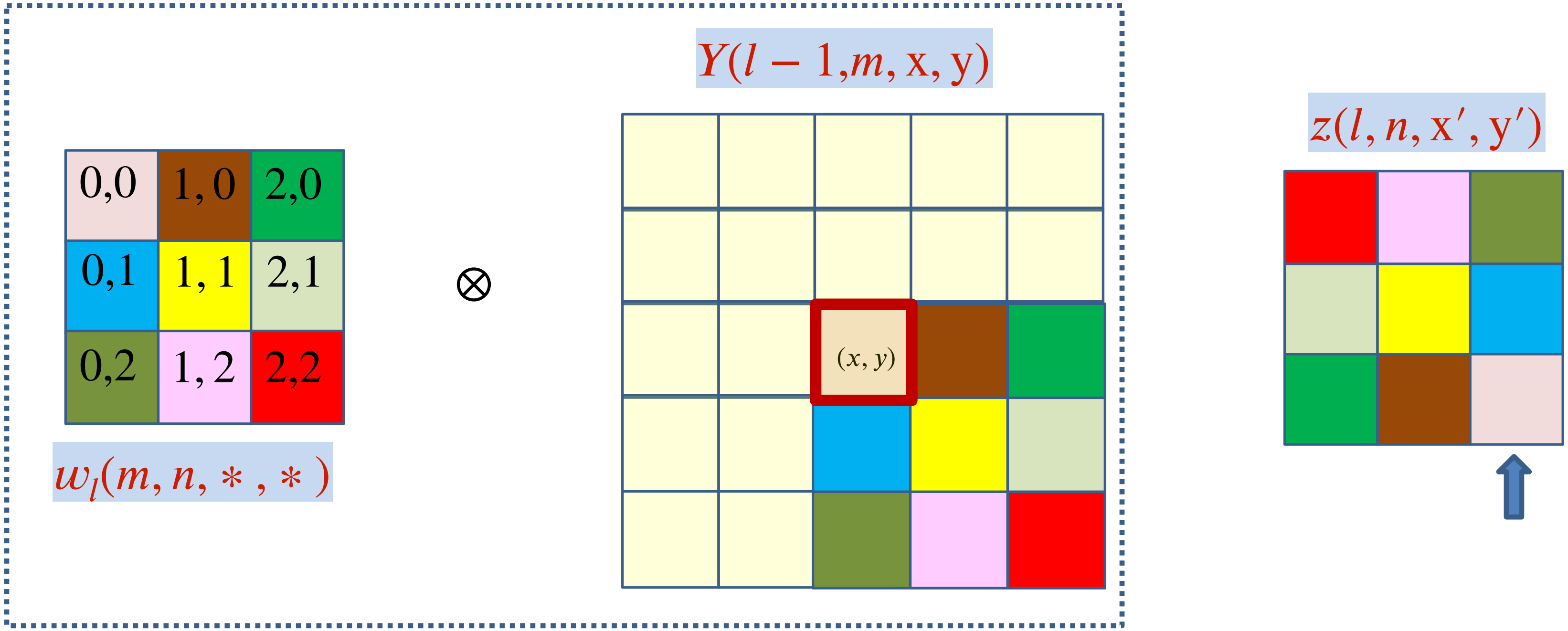


$$z(l,n,2,2) += Y(l-1,m,2,2)w_l(m,n,0,0)$$

- **Note:** The coordinates of $z(l,n)$ and $w_l(m,n)$ sum to the coordinates of $Y(l-1,m)$

$z(l,n,x',y') += Y(l-1,m,2,2)w_l(m,n,2-x',2-y')$

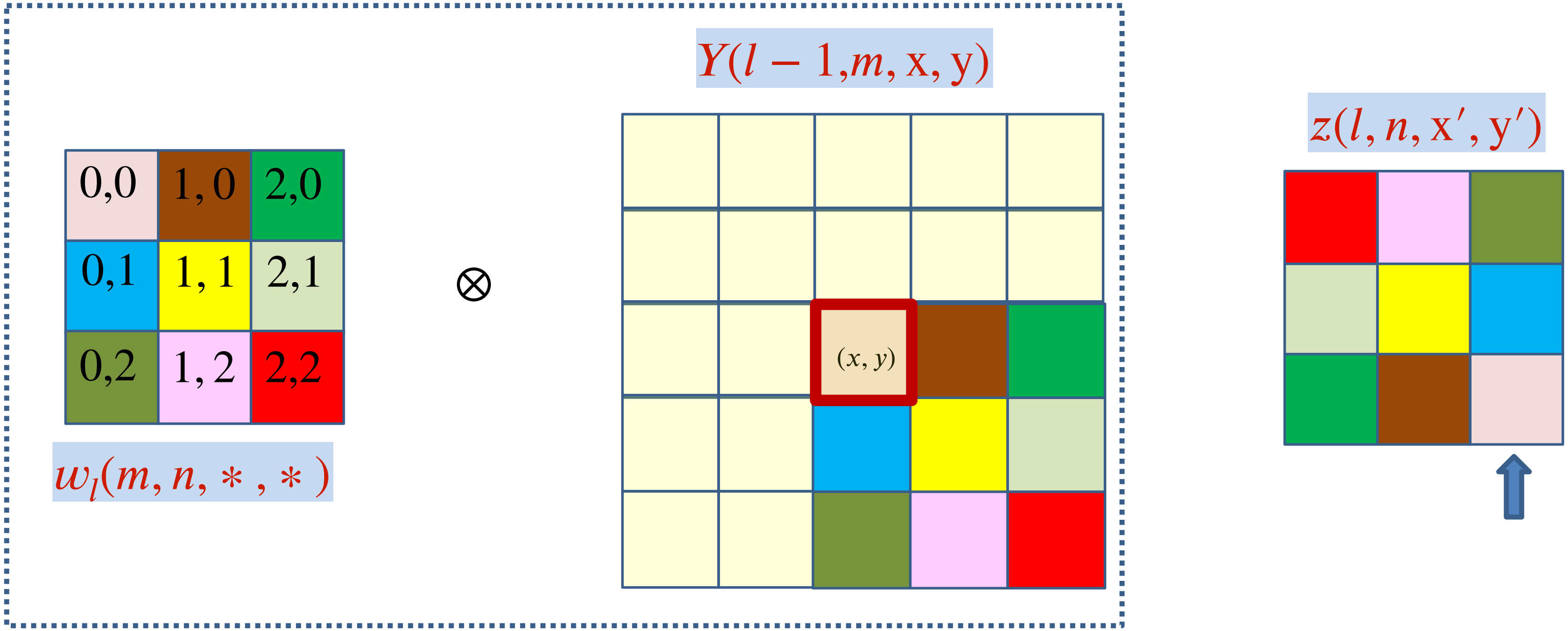
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, x', y') += Y(l-1, m, x, y)\, w_l(m, n, x-x', y-y')$$

- **Note:** The coordinates of $z(l, n)$ and $w_l(m, n)$ sum to the coordinates of $Y(l-1, m)$

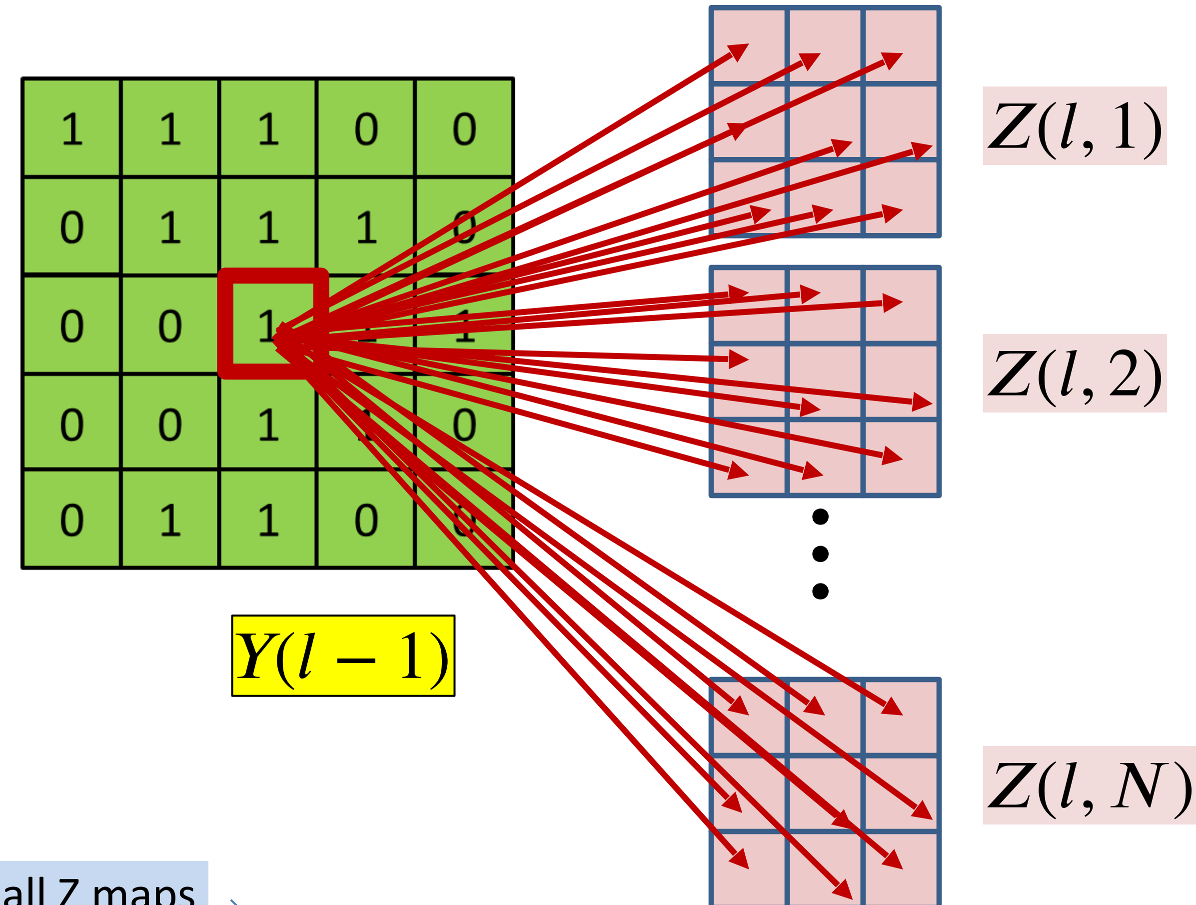
How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$



$$z\left(l,n,x',y'\right) += Y\left(l-1,m,x,y\right) w_l(m,n,x-x',y-y')$$

$$\frac{\partial z(l,n,x',y')}{\partial Y(l-1,m,x,y)} = w_l(m,n,x-x',y-y')$$

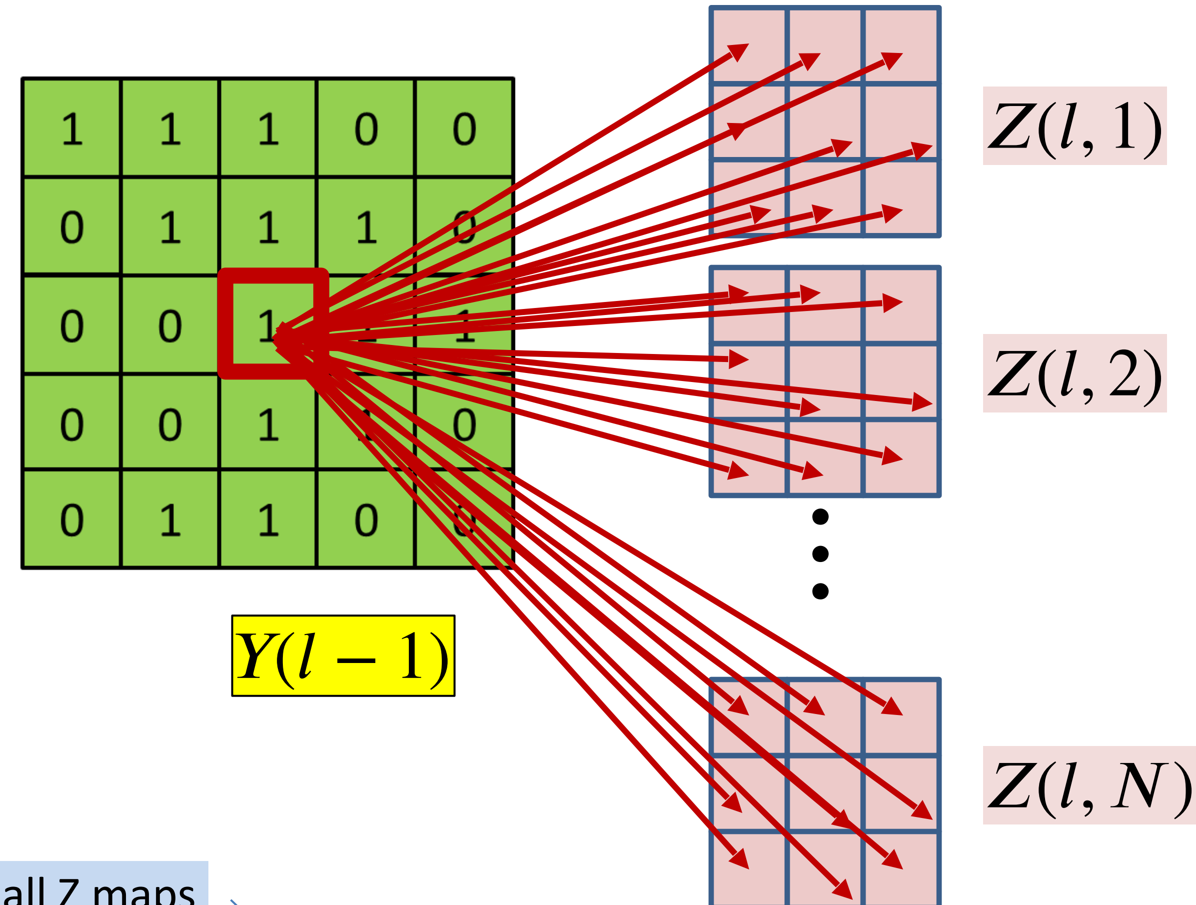
BP: Convolutional layer



Summing over all Z maps

$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_n \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} \frac{\partial z(l,n,x',y')}{\partial Y(l-1,m,x,y)}$$

BP: Convolutional layer



Summing over all Z maps

$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_{n} \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

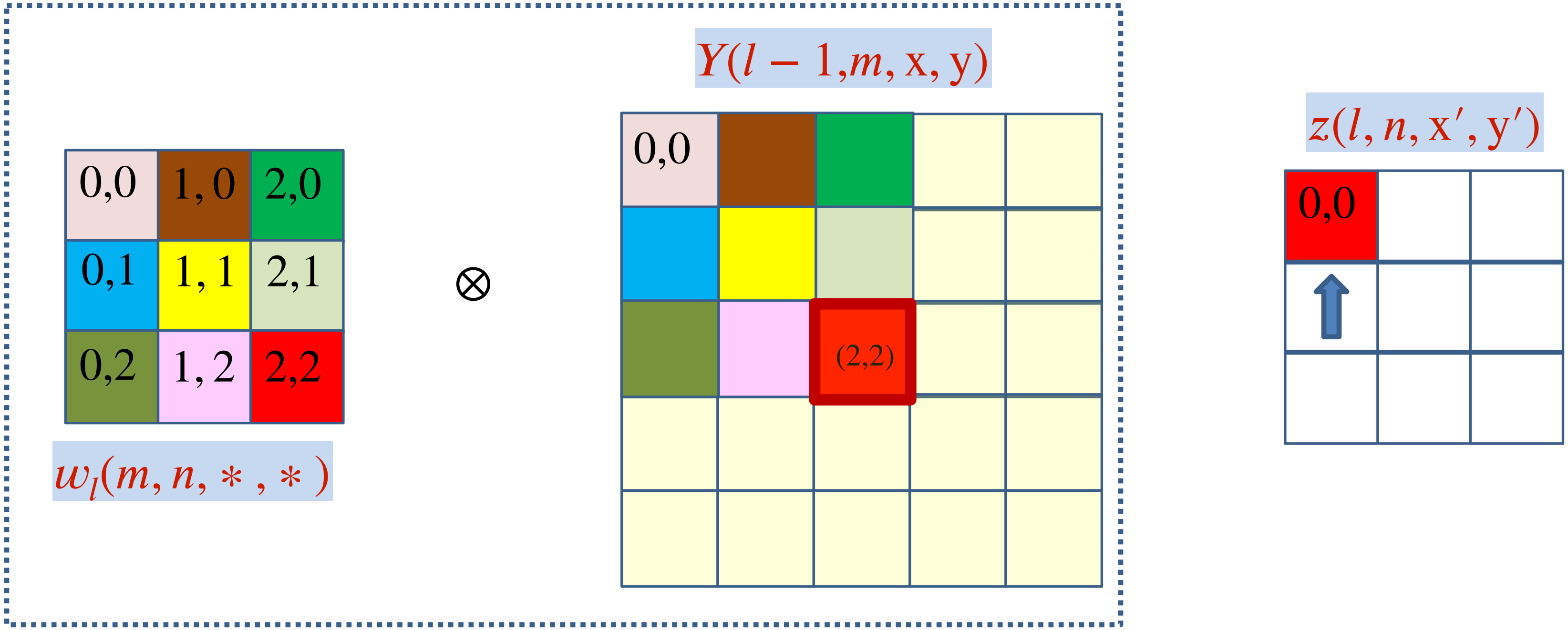
Computing derivative for $\boldsymbol{Y(l-1, m, *, *)}$

- The derivatives for every element of every map in $Y(l-1)$ by direct implementation of the formula:

$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$

- But this is actually a convolution!
  - Let's see how

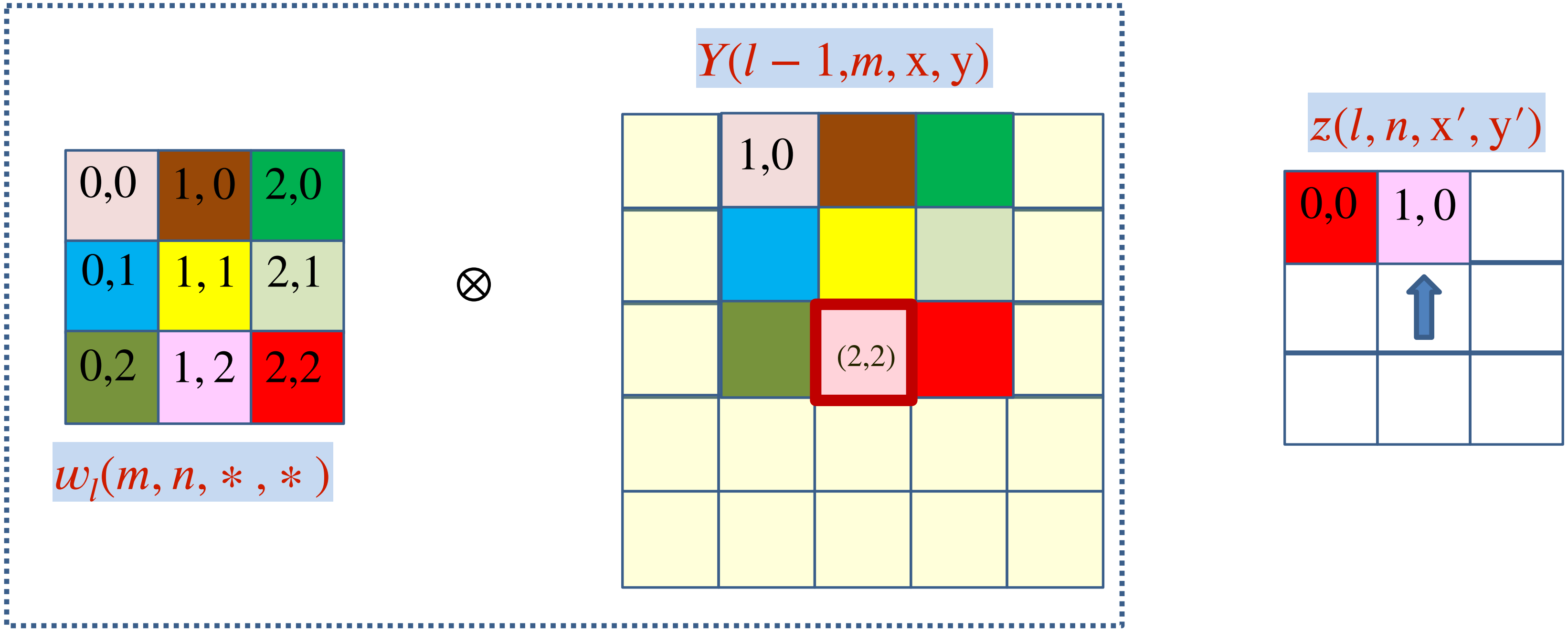
How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$



$$z(l,n,0,0) += Y(l-1,m,2,2)\,w_l(m,n,2,2)$$

$$\frac{\partial \ell}{\partial Y(l-1,m,2,2)} += \frac{\partial \ell}{\partial z(l,n,0,0)} w_l(m,n,2,2)$$

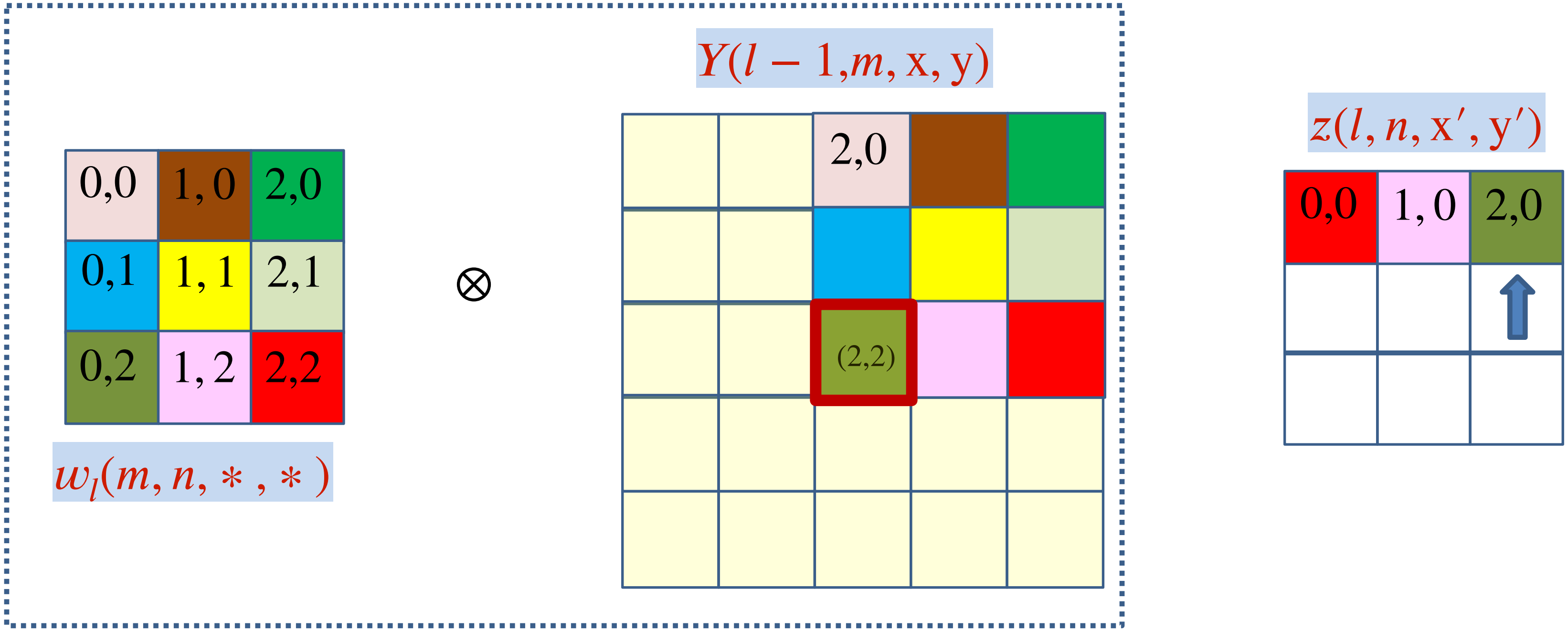
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, 1,0) += Y(l-1, m, 2, 2)\, w_l(m, n, 1, 2)$$

$$\frac{\partial \ell}{\partial Y(l-1, m, 2, 2)} += \frac{\partial \ell}{\partial z(l, n, 1, 0)} w_l(m, n, 1, 2)$$

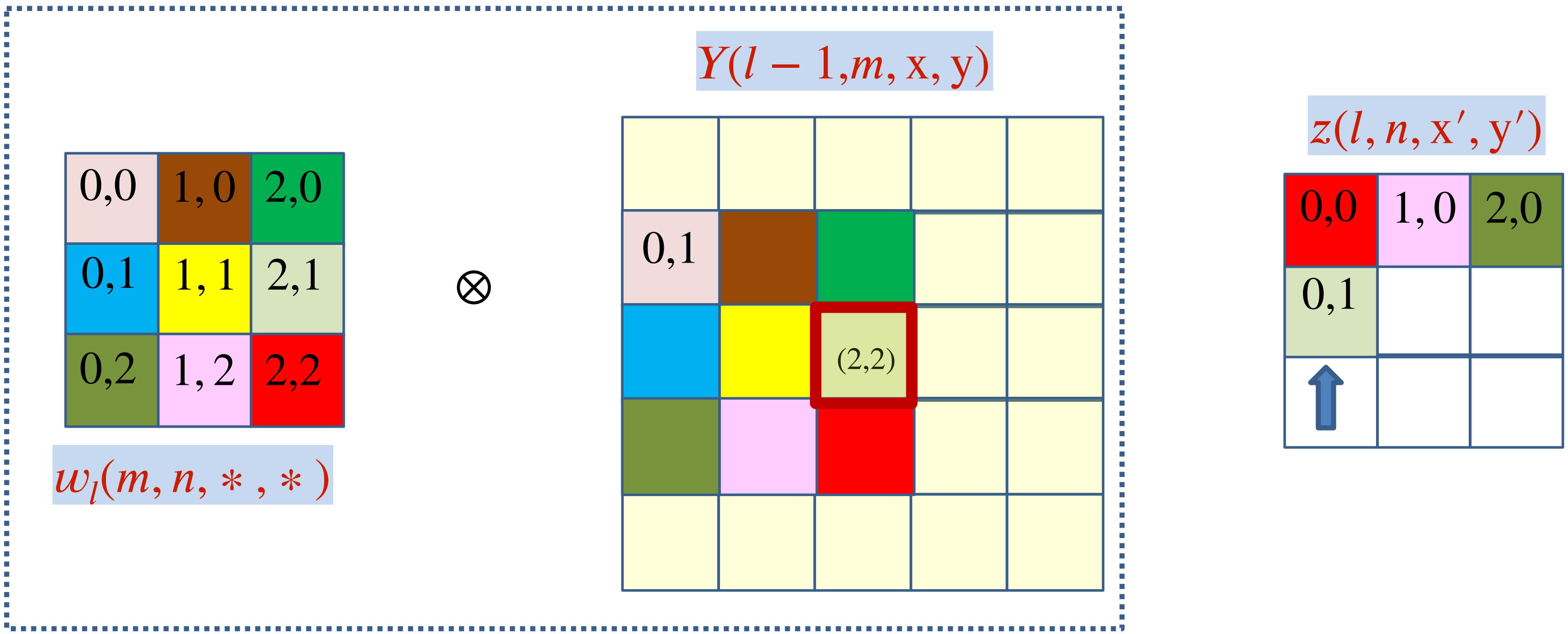
How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$



$$z(l,n,2,0) += Y(l-1,m,2,2)w_l(m,n,0,2)$$

$$\frac{\partial \ell}{\partial Y(l-1,m,2,2)} += \frac{\partial \ell}{\partial z(l,n,2,0)} w_l(m,n,0,2)$$

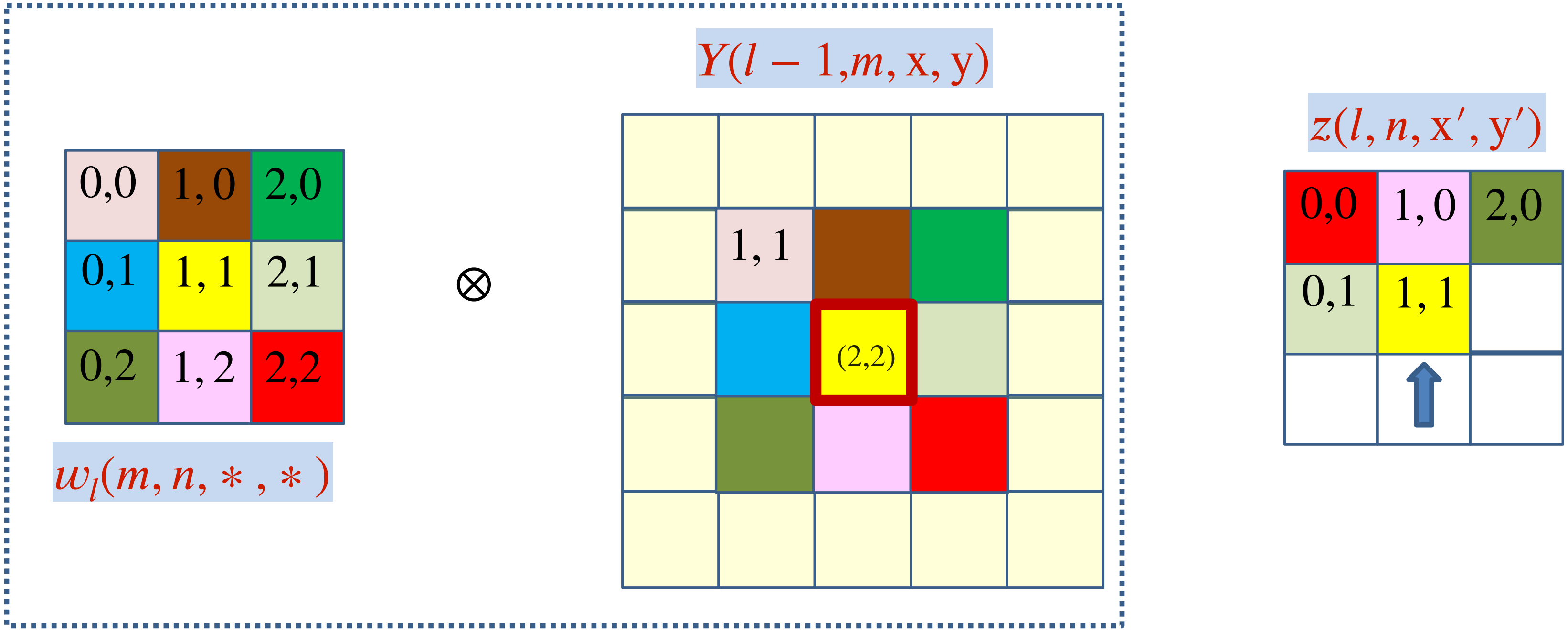
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, 0,1) += Y(l-1, m, 2, 2)w_l(m, n, 2, 1)$$

$$\frac{\partial \ell}{\partial Y(l-1, m, 2, 2)} += \frac{\partial \ell}{\partial z(l, n, 0, 1)} w_l(m, n, 2, 1)$$

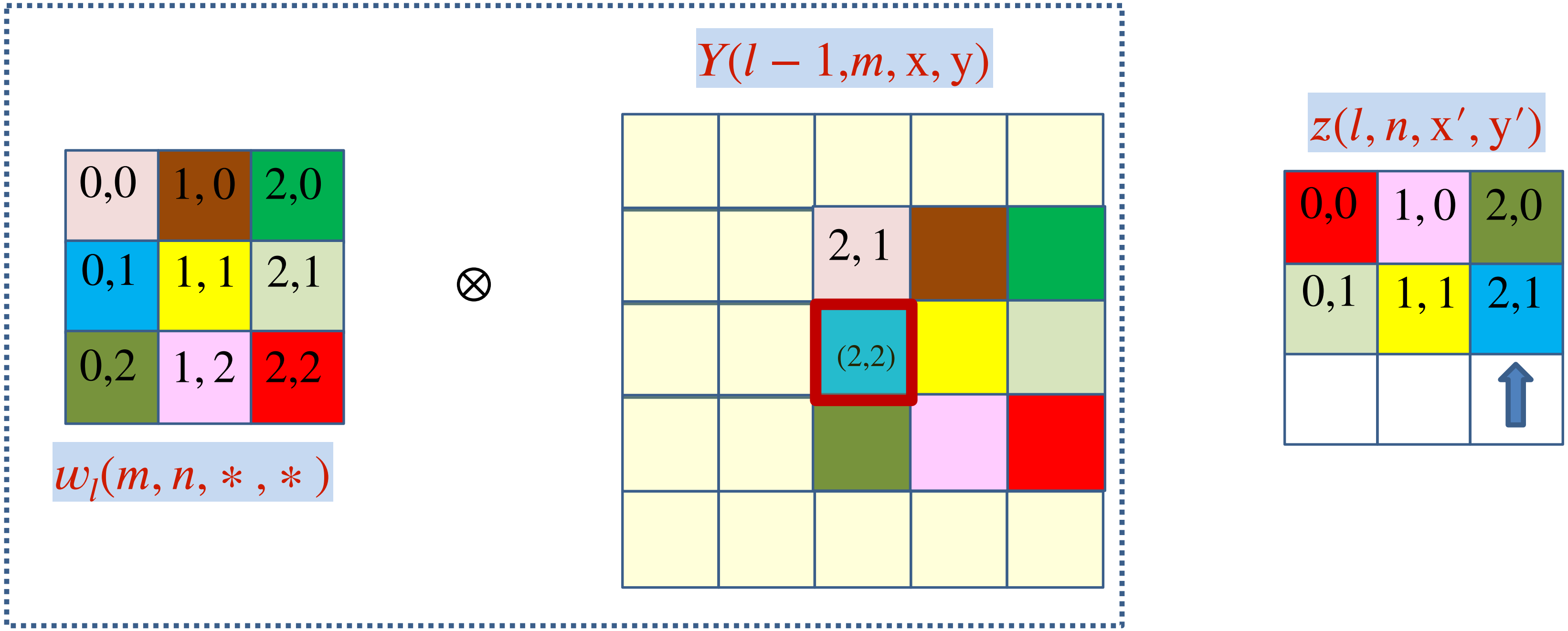
How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$



$$z(l,n,1,1) += Y(l-1,m,2,2)\,w_l(m,n,1,1)$$

$$\frac{\partial \ell}{\partial Y(l-1,m,2,2)} += \frac{\partial \ell}{\partial z(l,n,1,1)} w_l(m,n,1,1)$$

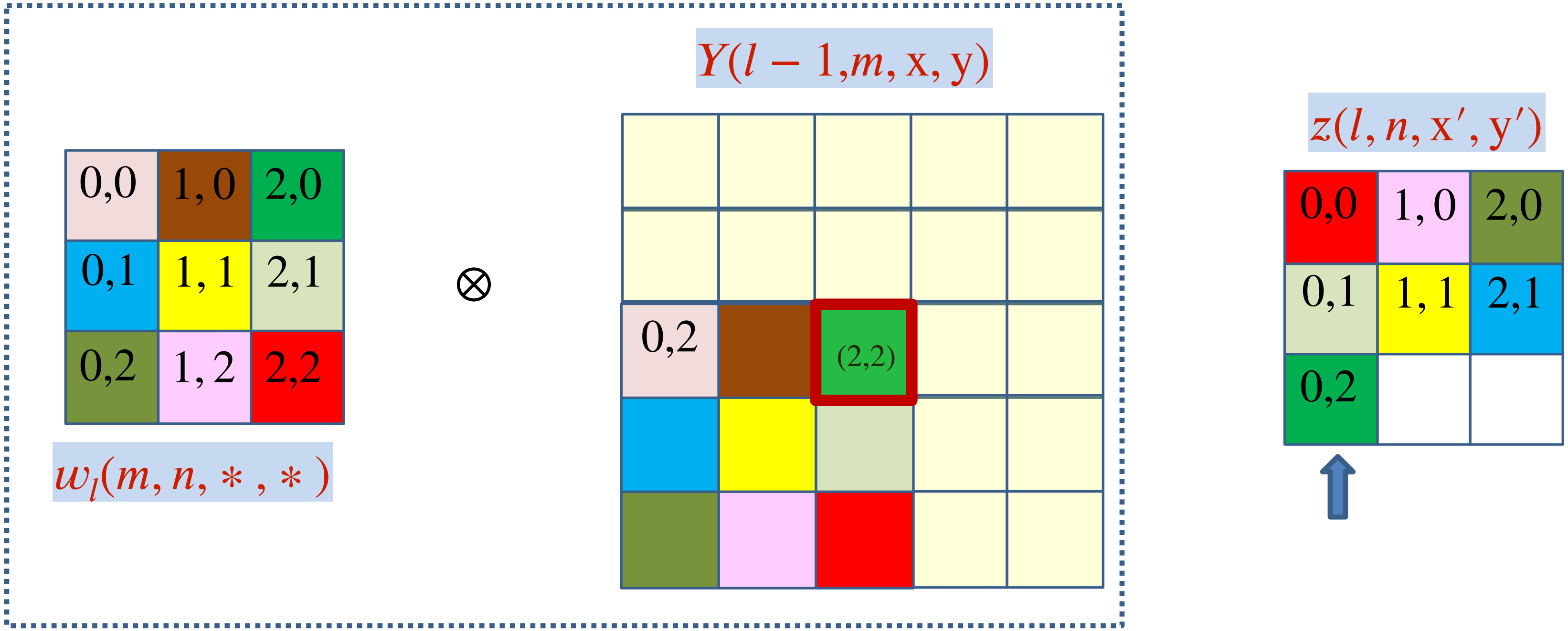
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, 2, 1) += Y(l-1, m, 2, 2)\, w_l(m, n, 0, 1)$$

$$\frac{\partial \ell}{\partial Y(l-1, m, 2, 2)} += \frac{\partial \ell}{\partial z(l, n, 2, 1)} w_l(m, n, 0, 1)$$

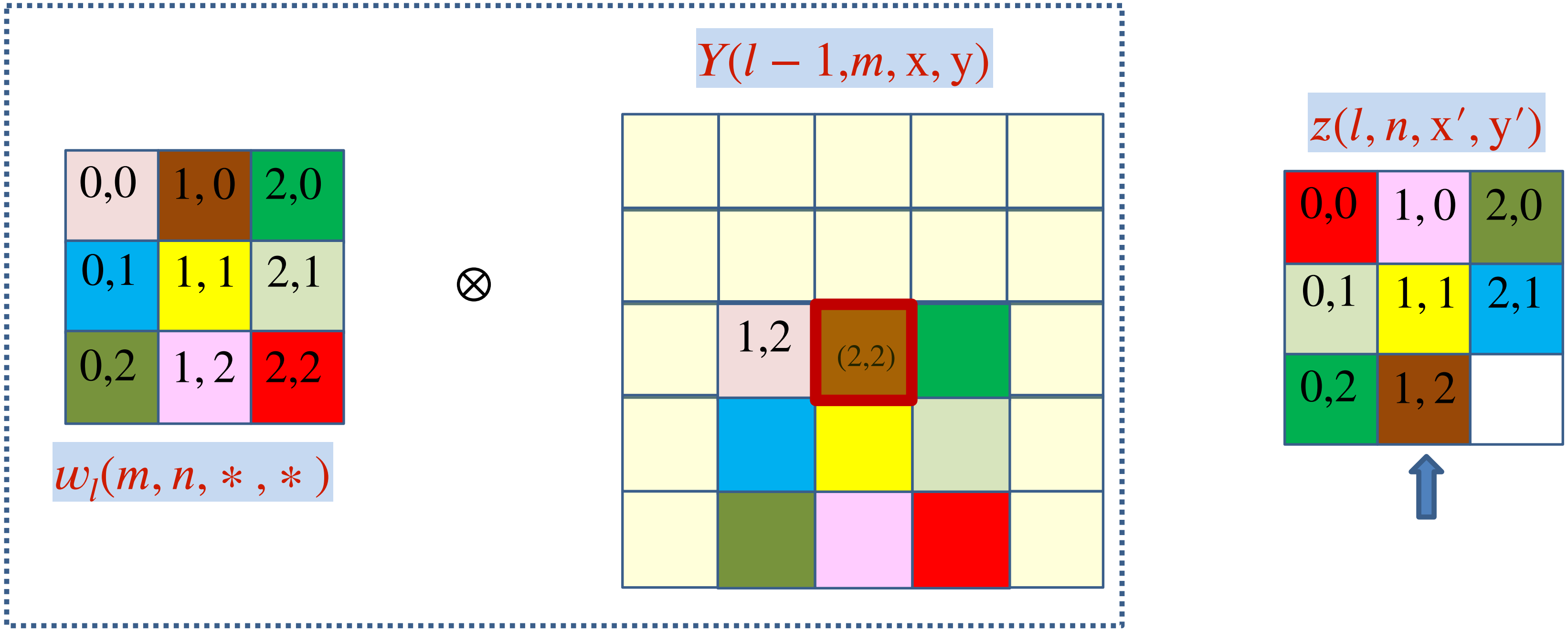
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, 0,2) += Y(l-1, m, 2, 2)\, w_l(m, n, 2, 0)$$

$$\frac{\partial \ell}{\partial Y(l-1, m, 2, 2)} += \frac{\partial \ell}{\partial z(l, n, 0, 2)} w_l(m, n, 2, 0)$$

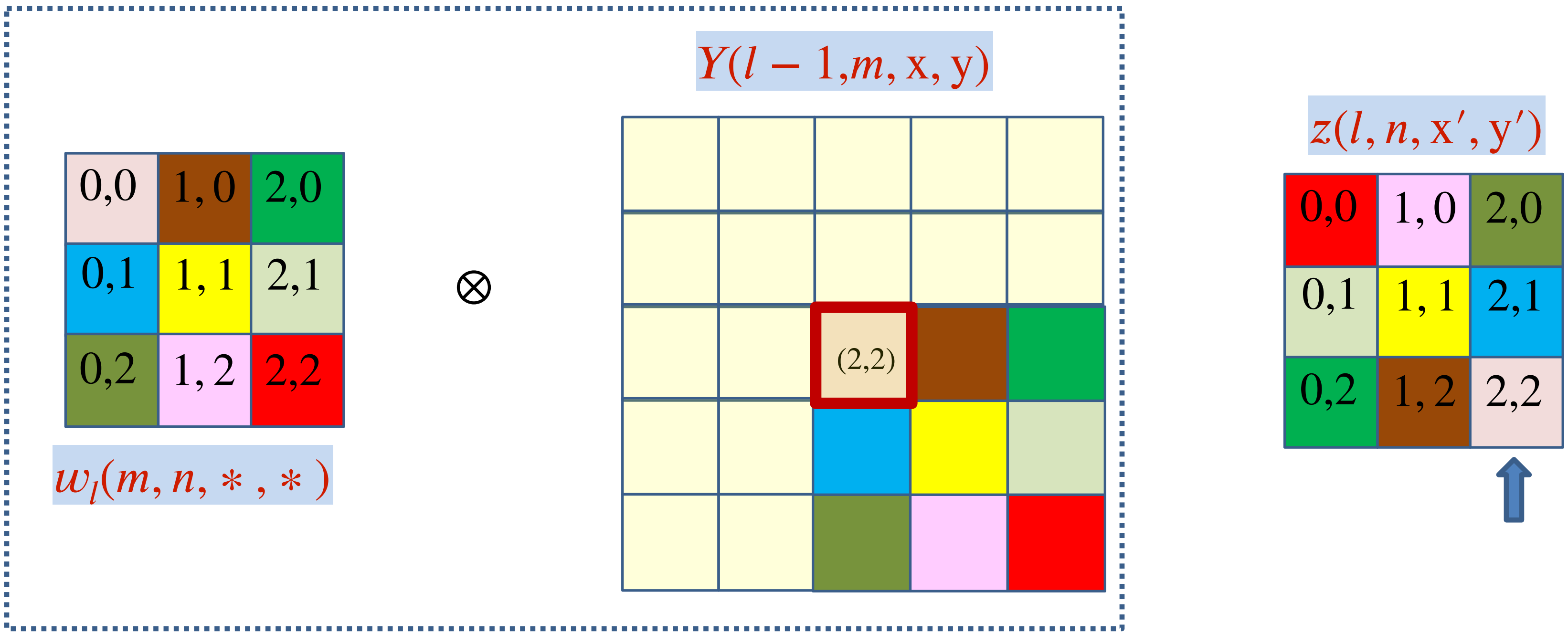
How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$



$$z(l,n,1{,}2) += Y(l-1,m,2,2)w_l(m,n,2,1)$$

$$\frac{\partial \ell}{\partial Y(l-1,m,2,2)} += \frac{\partial \ell}{\partial z(l,n,1,2)} w_l(m,n,1,0)$$

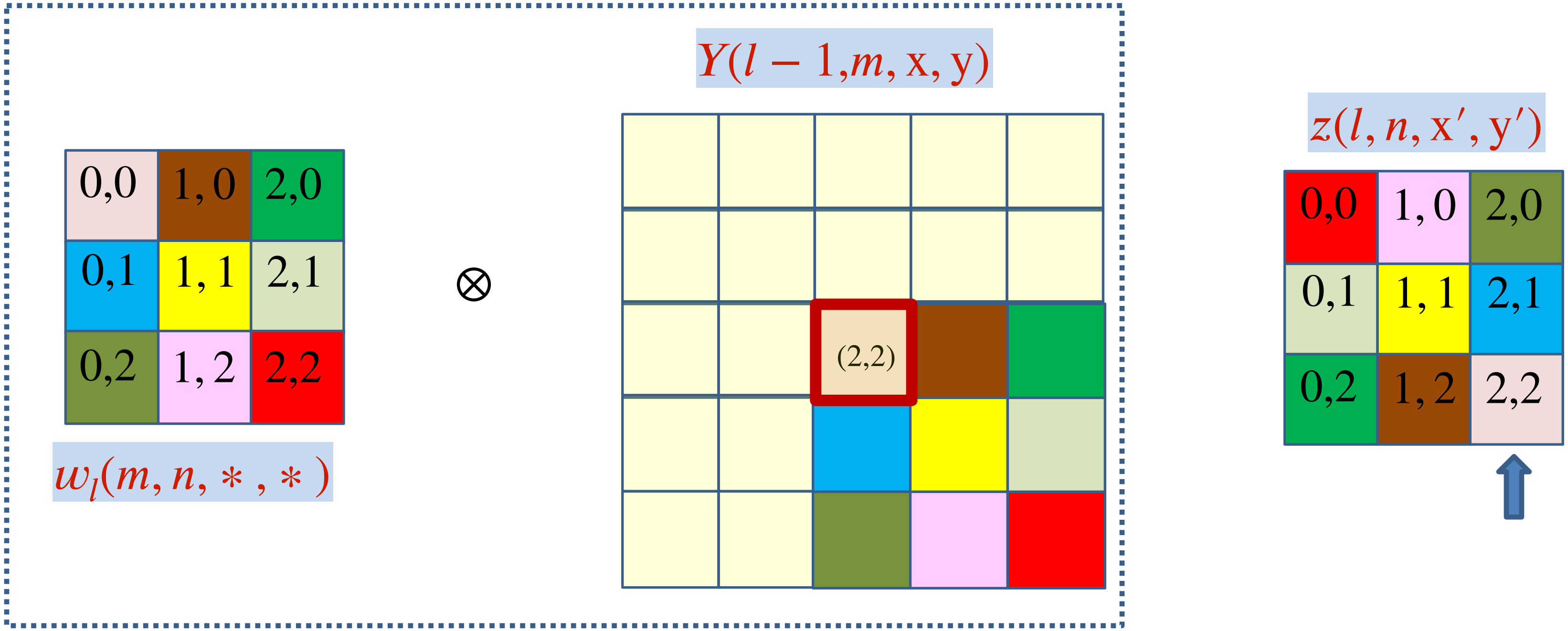
How a single $Y(l-1, m, x, y)$ influences $z(l, n, x', y')$



$$z(l, n, 2,2) += Y(l-1, m, 2, 2)w_l(m, n, 0, 0)$$

$$\frac{\partial \ell}{\partial Y(l-1, m, 2, 2)} += \frac{\partial \ell}{\partial z(l, n, 2, 2)} w_l(m, n, 0, 0)$$

How a single $Y(l-1,m,x,y)$ influences $z(l,n,x',y')$

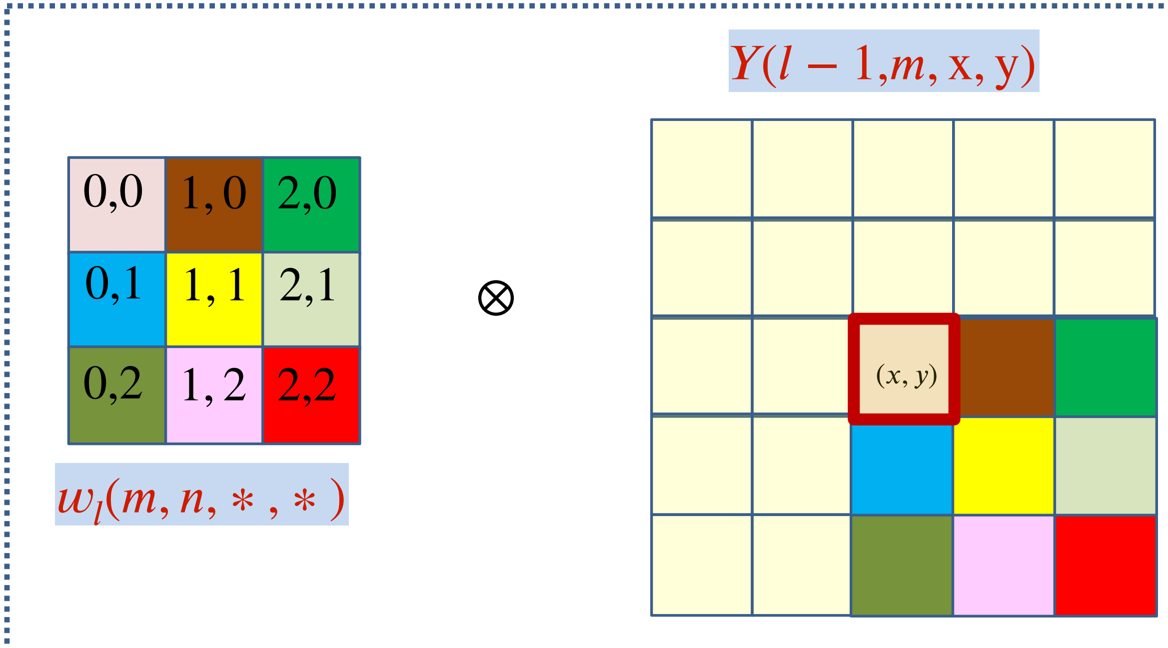


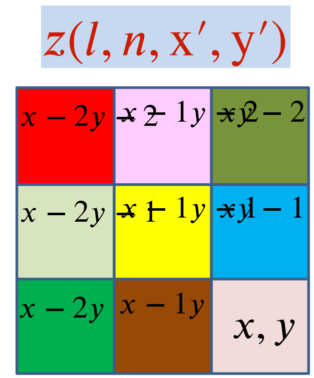
$$z\left(l,n,x',y'\right) += Y(l-1,m,2,2)w_l(m,n,2-\mathrm{x}',2-\mathrm{y}')$$

$$\frac{\partial \ell}{\partial Y(l-1,m,2,2)} += \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,2-x',2-y')$$

- The derivative at $Y(l-1,m,2,2)$ is the sum of component-wise product of the filter elements and the elements of the derivative at $z(l,m,.,.,.)$

Derivative at $Y(l-1,m,x,y)$ from a single $Z(l,n)$ map





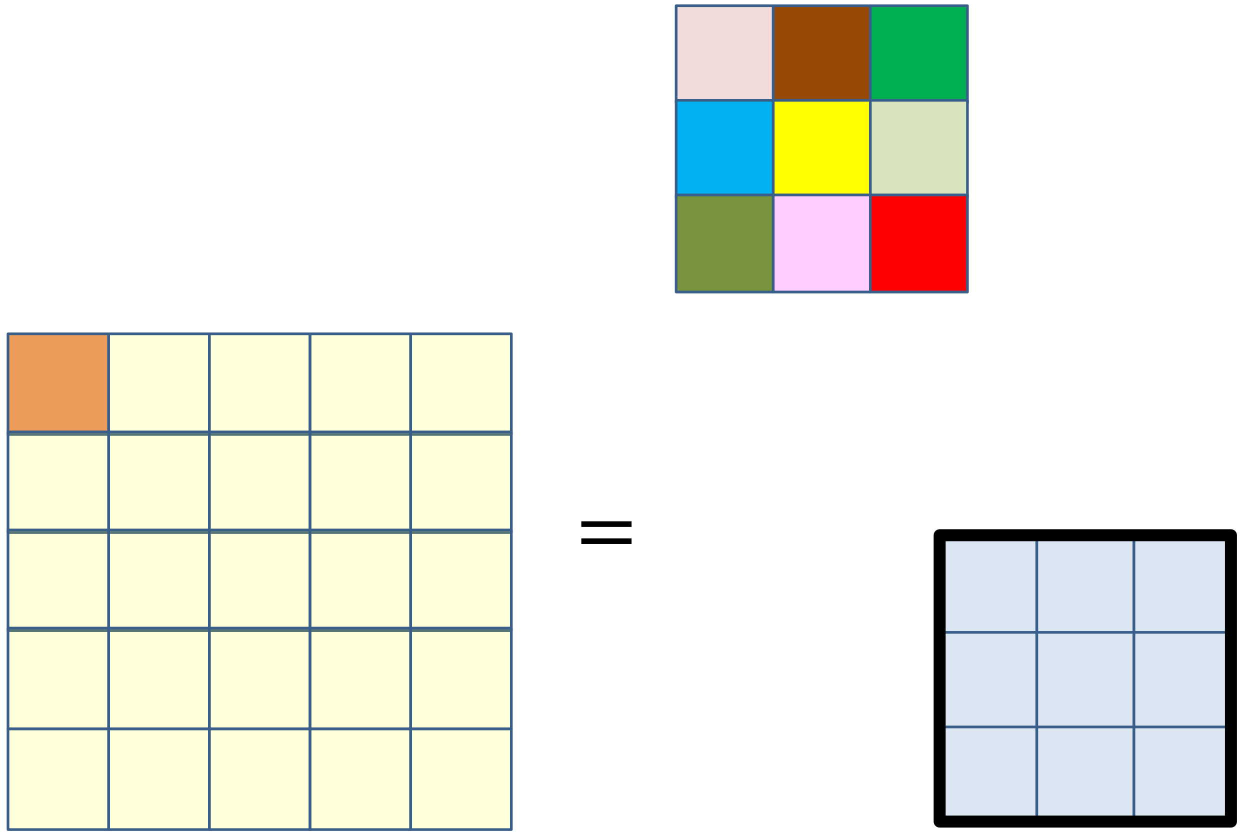
$$z\left(l,n,x',y'\right) += Y\left(l-1,m,x,y\right) w_l(m,n,x-x',y-y')$$

$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} += \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

Contribution of the entire nth affine map $z(l,n,*,*)$

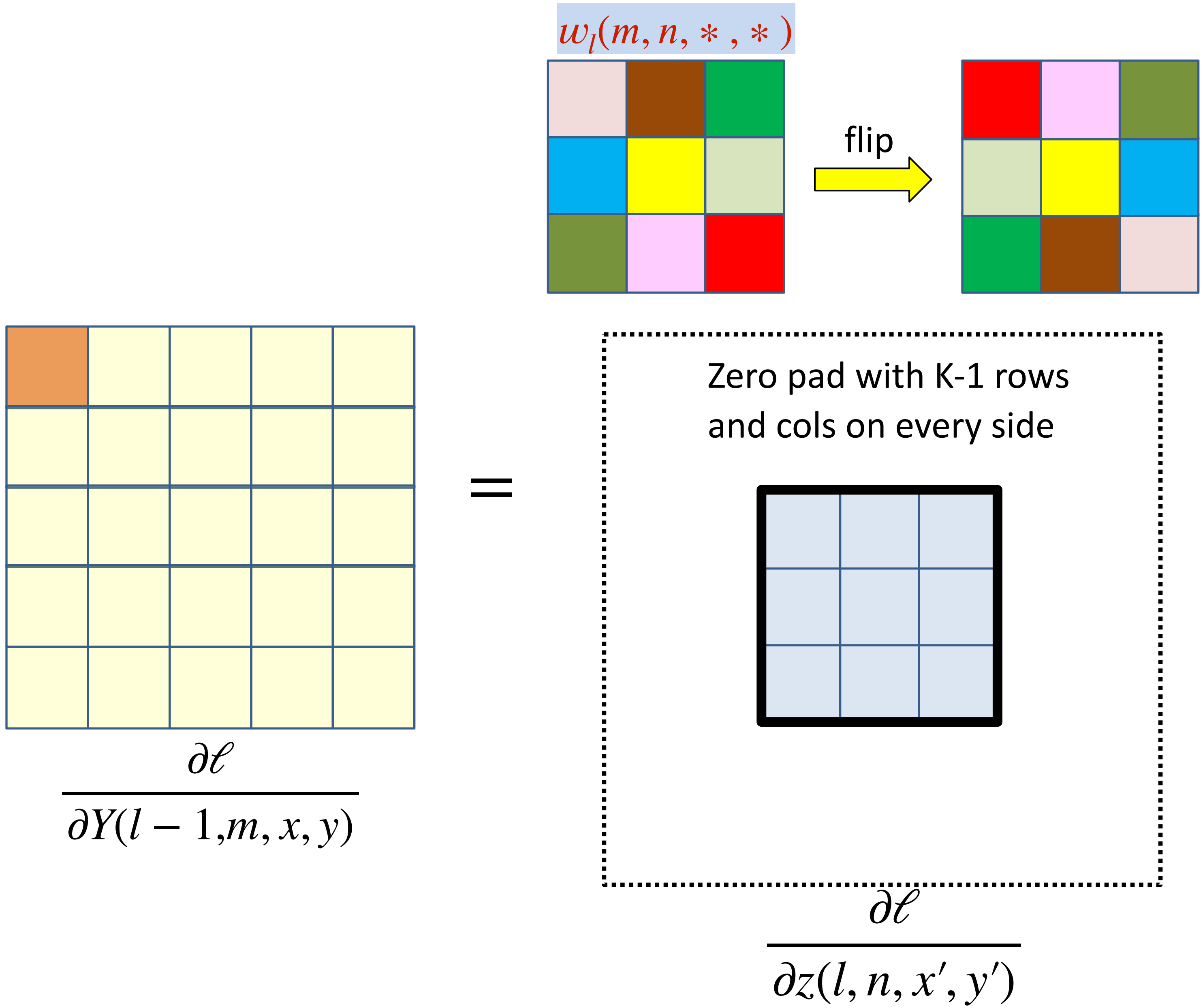
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

$w_l(m,n,*,*)$

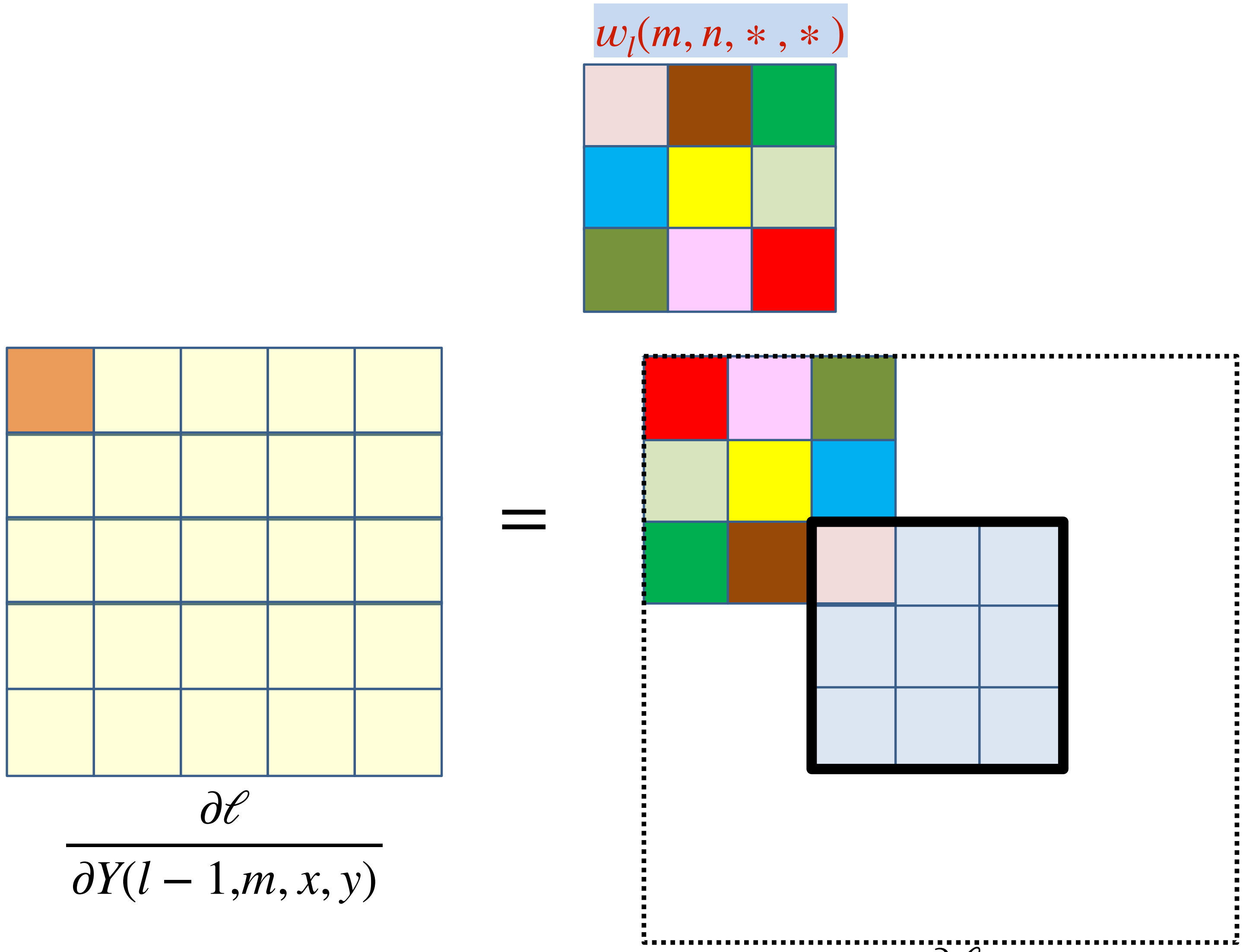


$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} \qquad = \qquad \frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

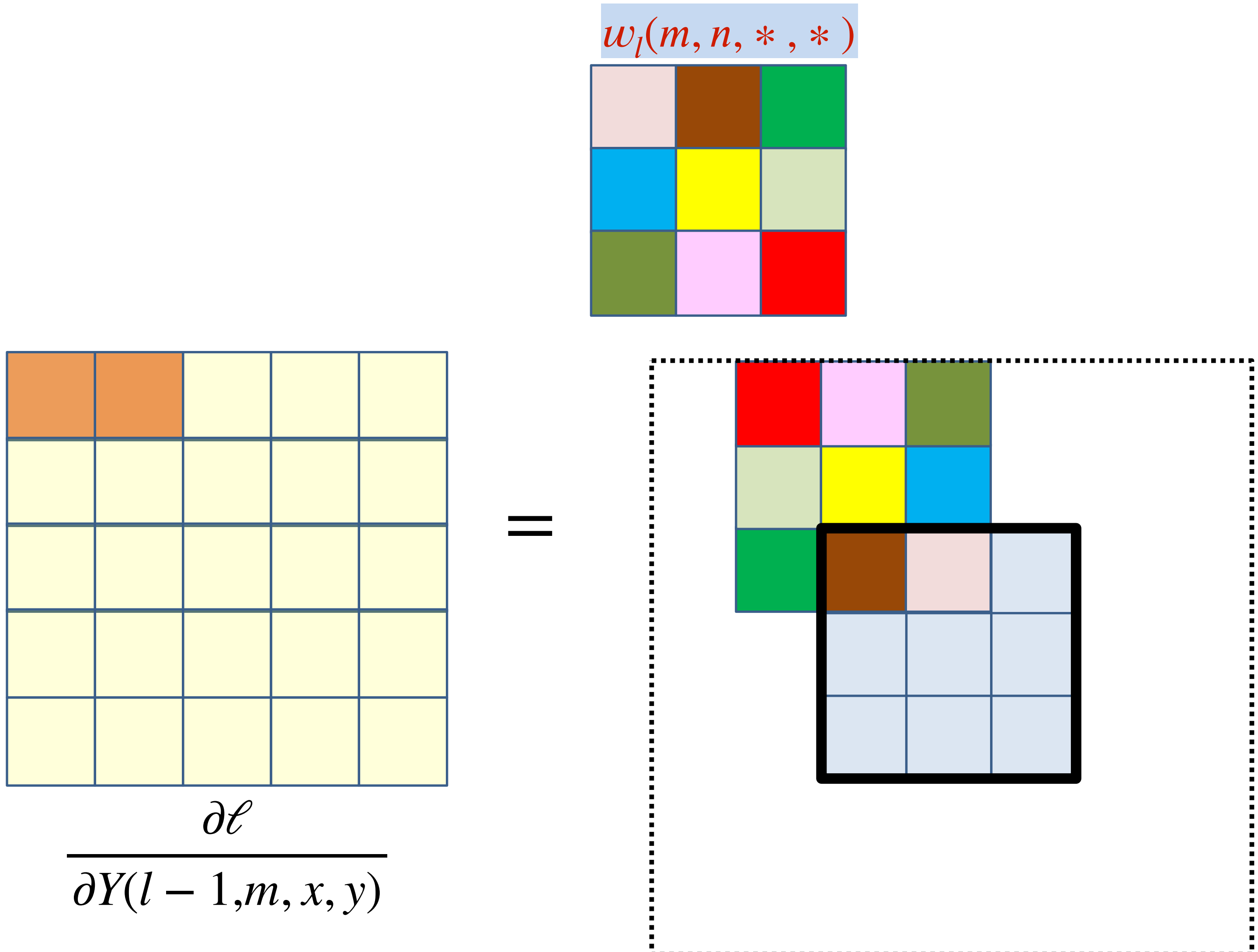


Derivative at $Y(l-1, m)$ from a single $Z(l, n)$ map



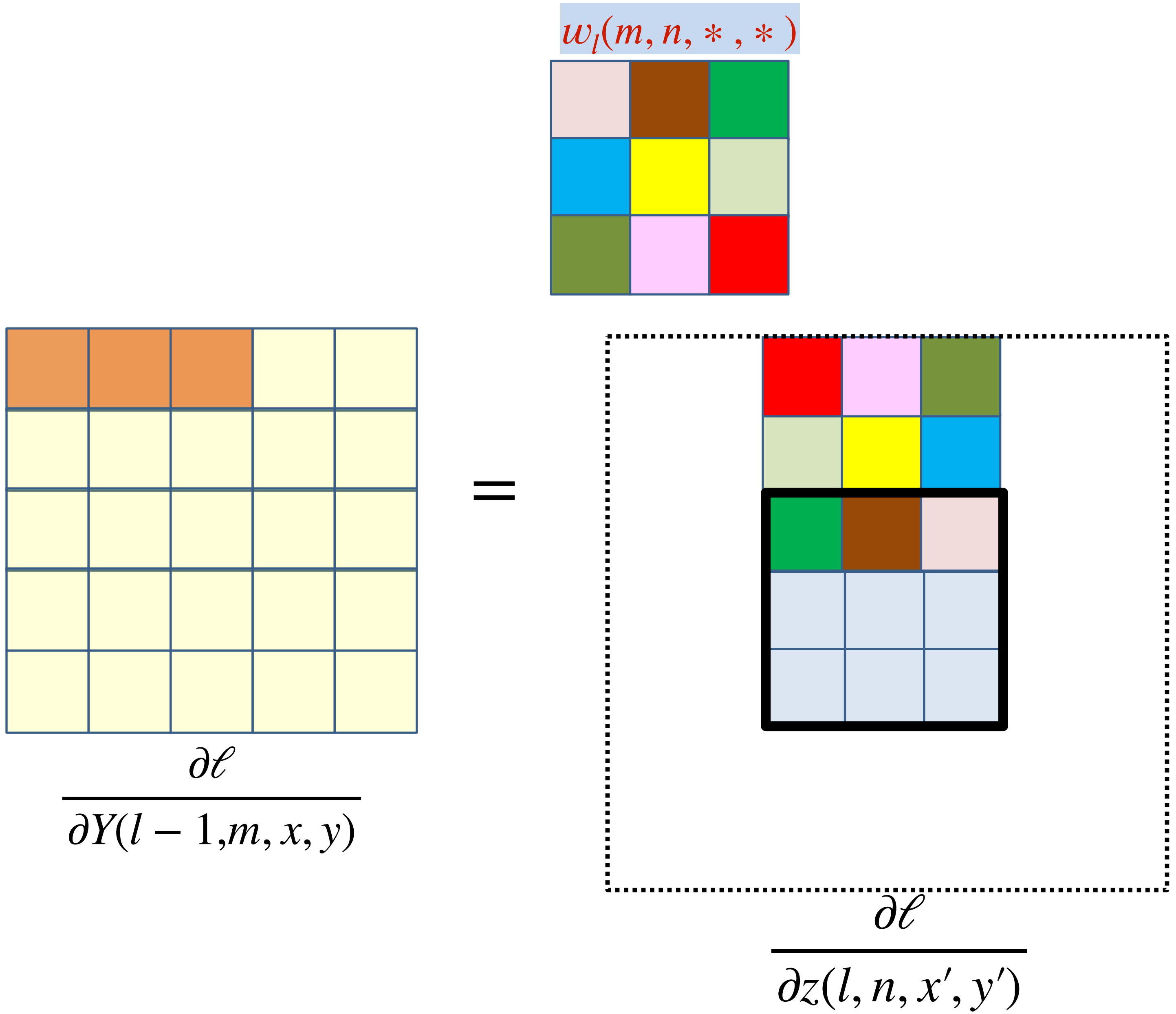
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \frac{\partial \ell}{\partial z(l, n, x', y')}$$

Derivative at $Y(l-1, m)$ from a single $Z(l, n)$ map



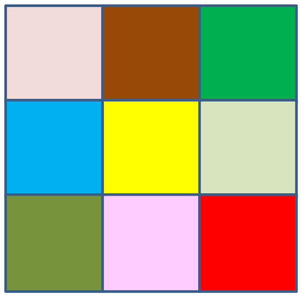
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \frac{\partial \ell}{\partial z(l, n, x', y')}$$

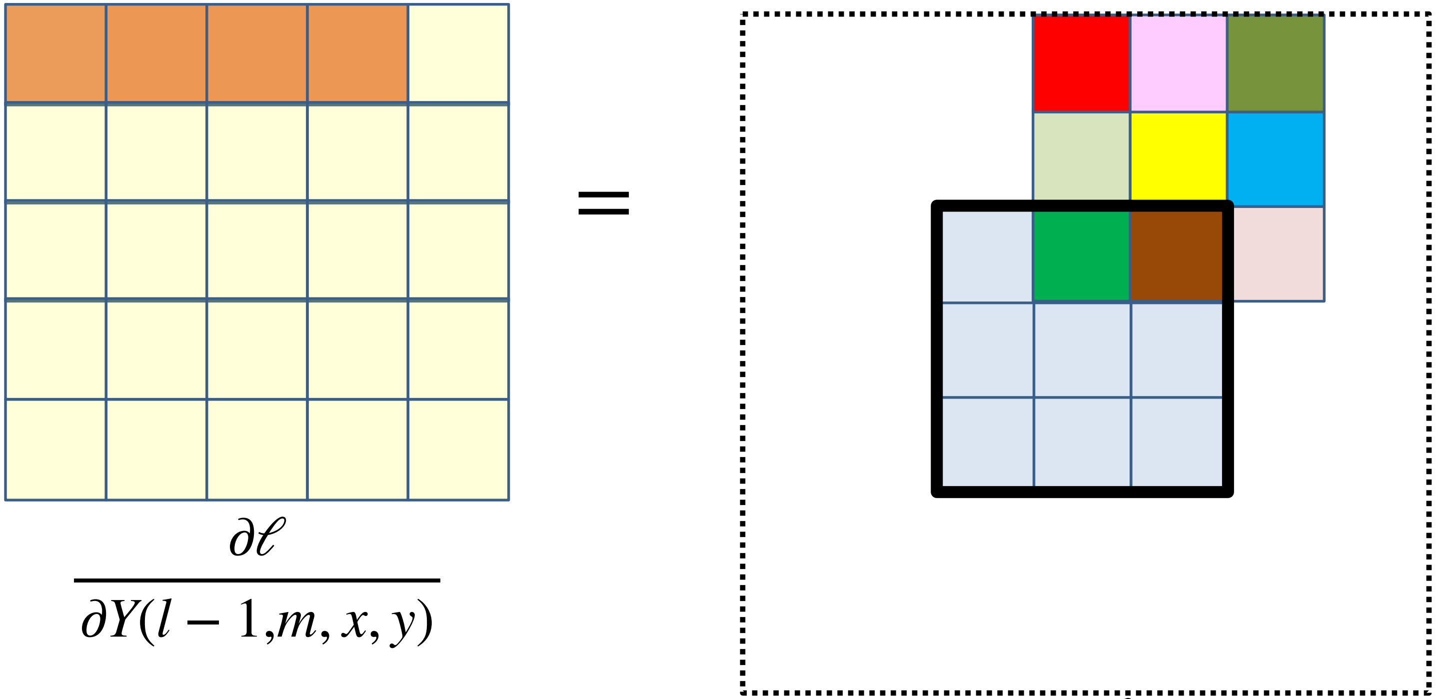
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

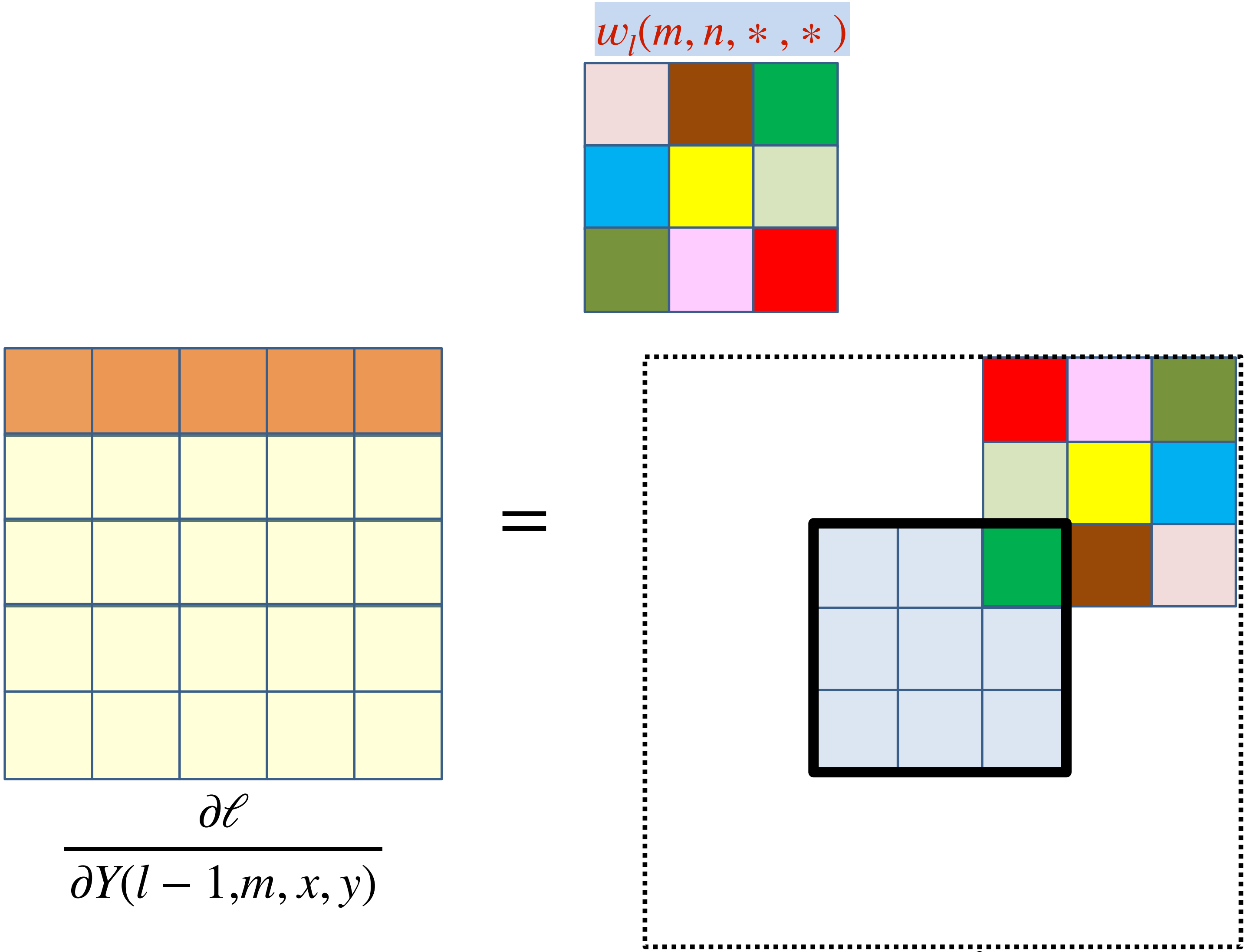
$w_l(m,n,*,*)$





$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

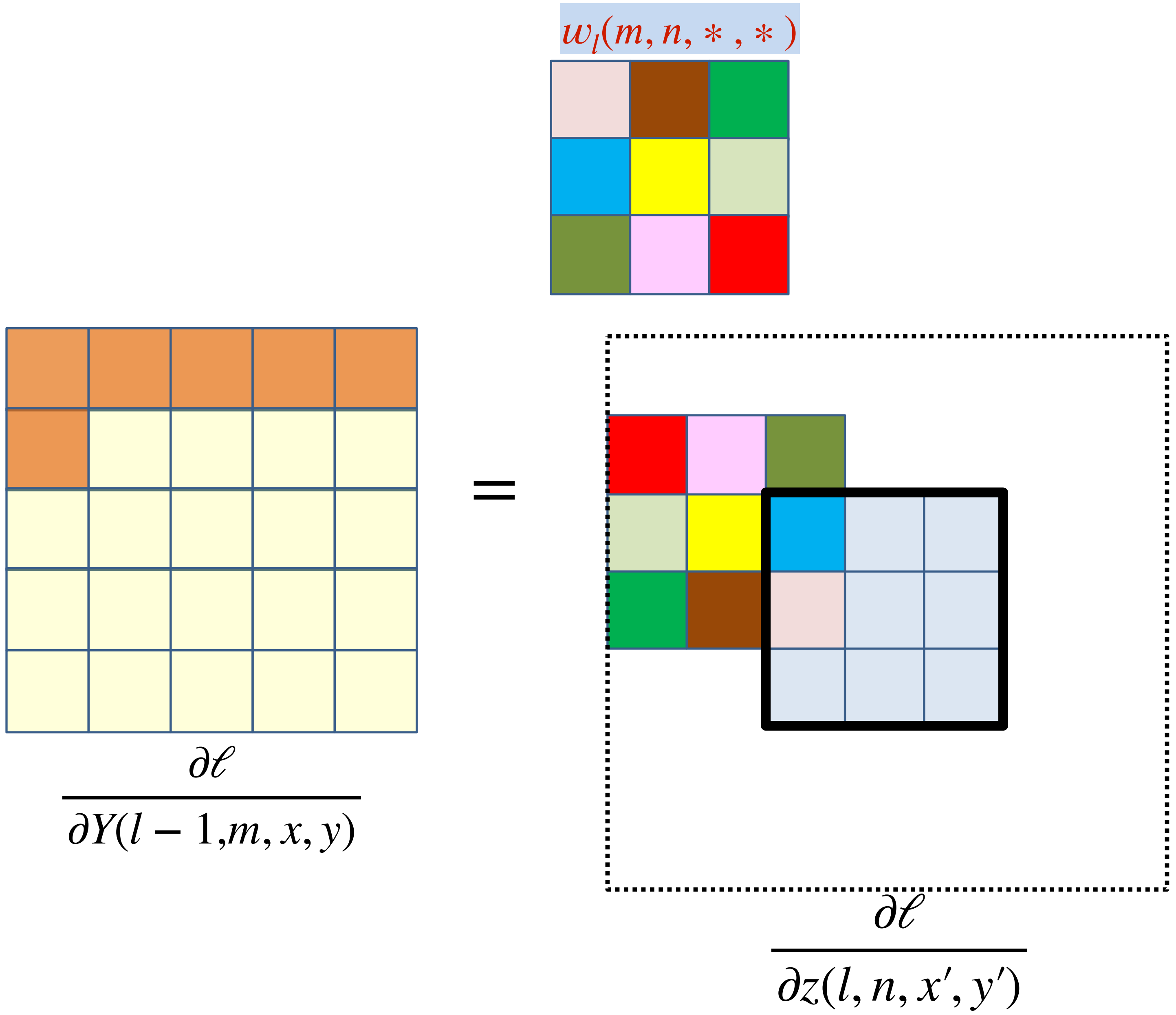
Derivative at $Y(l-1, m)$ from a single $Z(l, n)$ map



$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

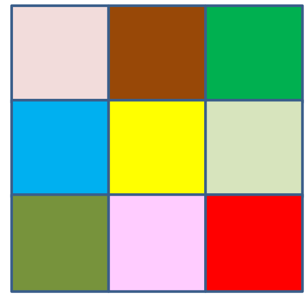
$$\frac{\partial \ell}{\partial z(l, n, x', y')}$$

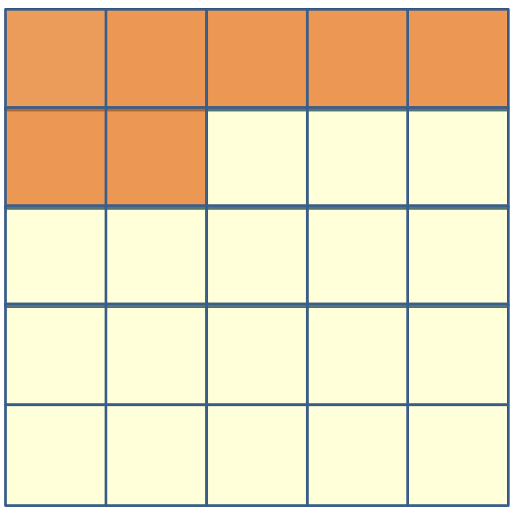
Derivative at $Y(l-1, m)$ from a single $Z(l, n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

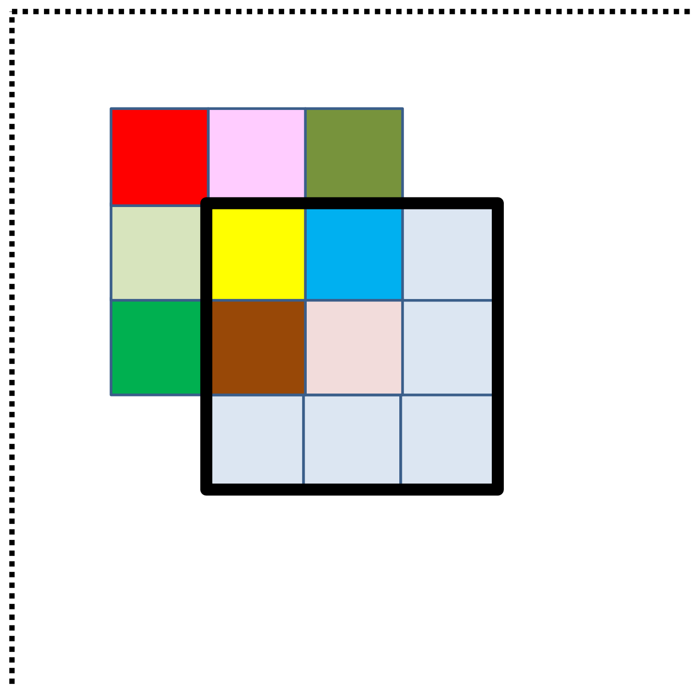
$w_l(m,n,*,*)$





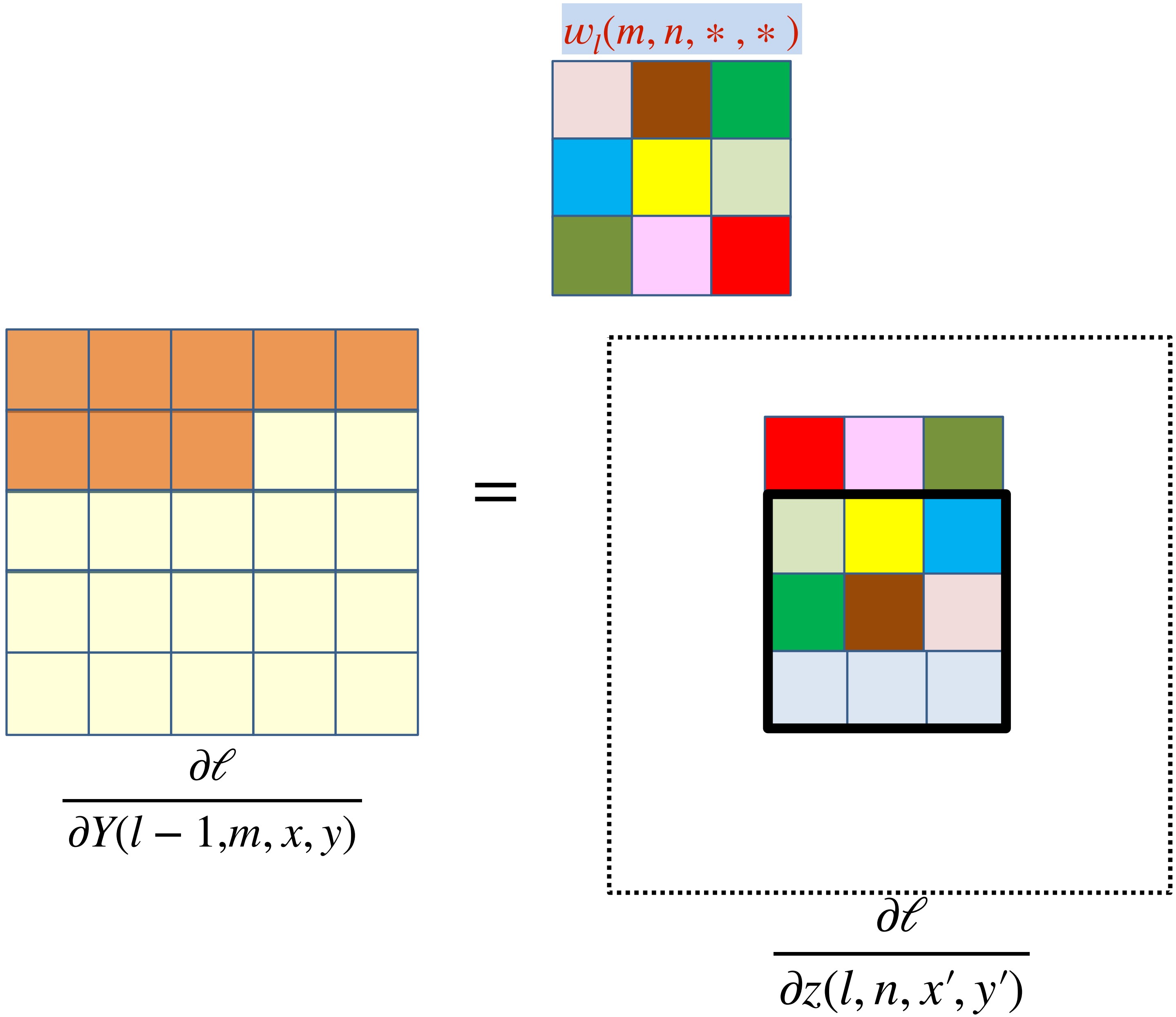
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)}$$

$$=$$

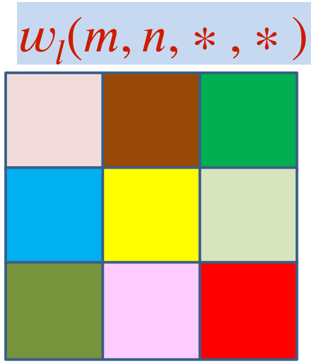


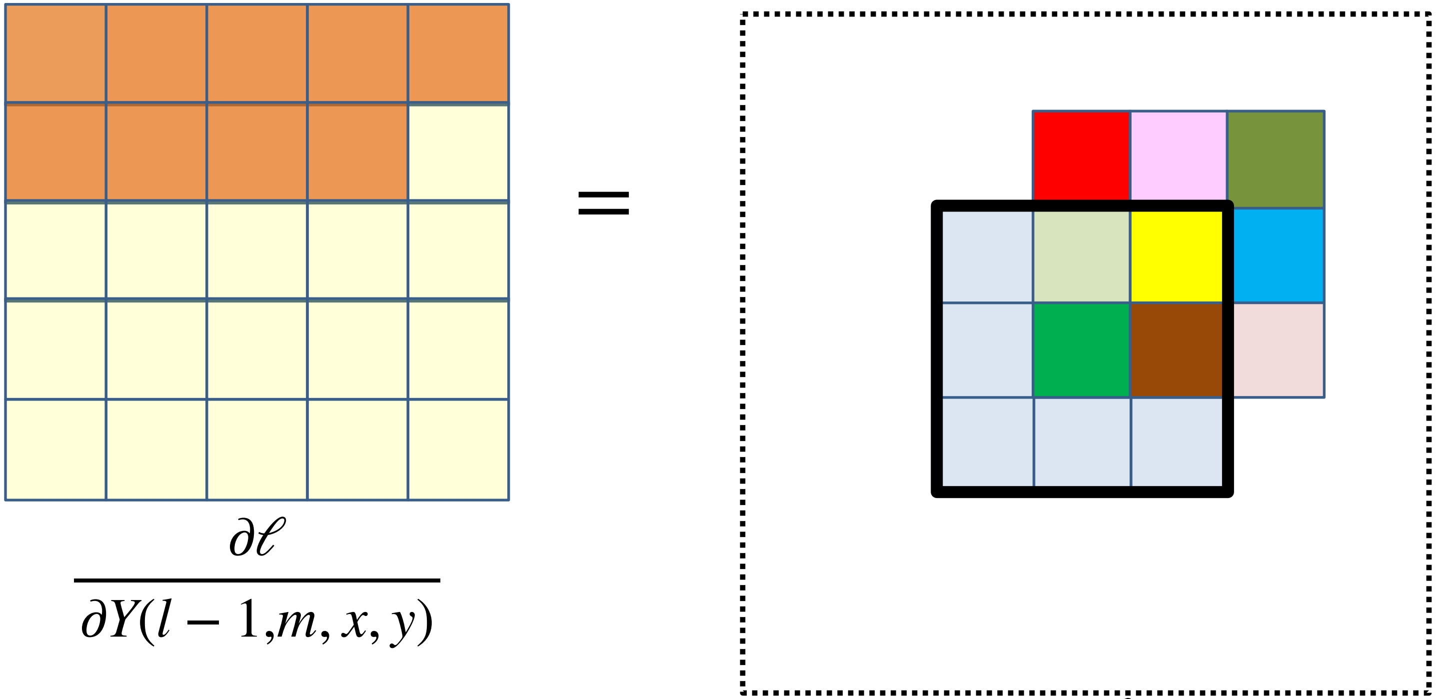
$$\frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



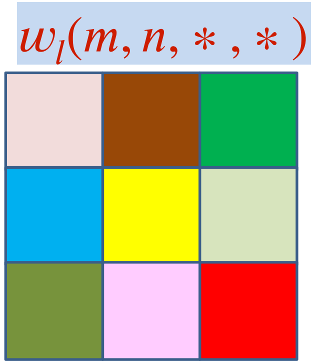
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

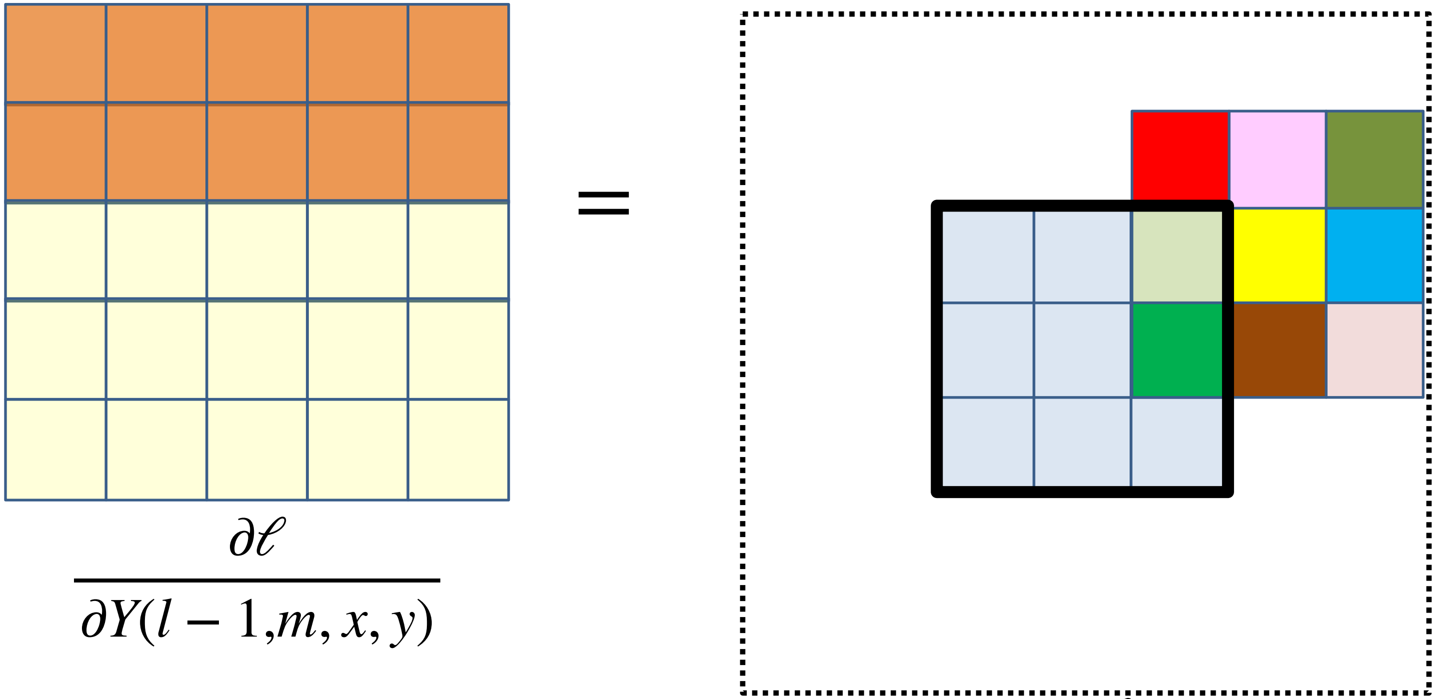




$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

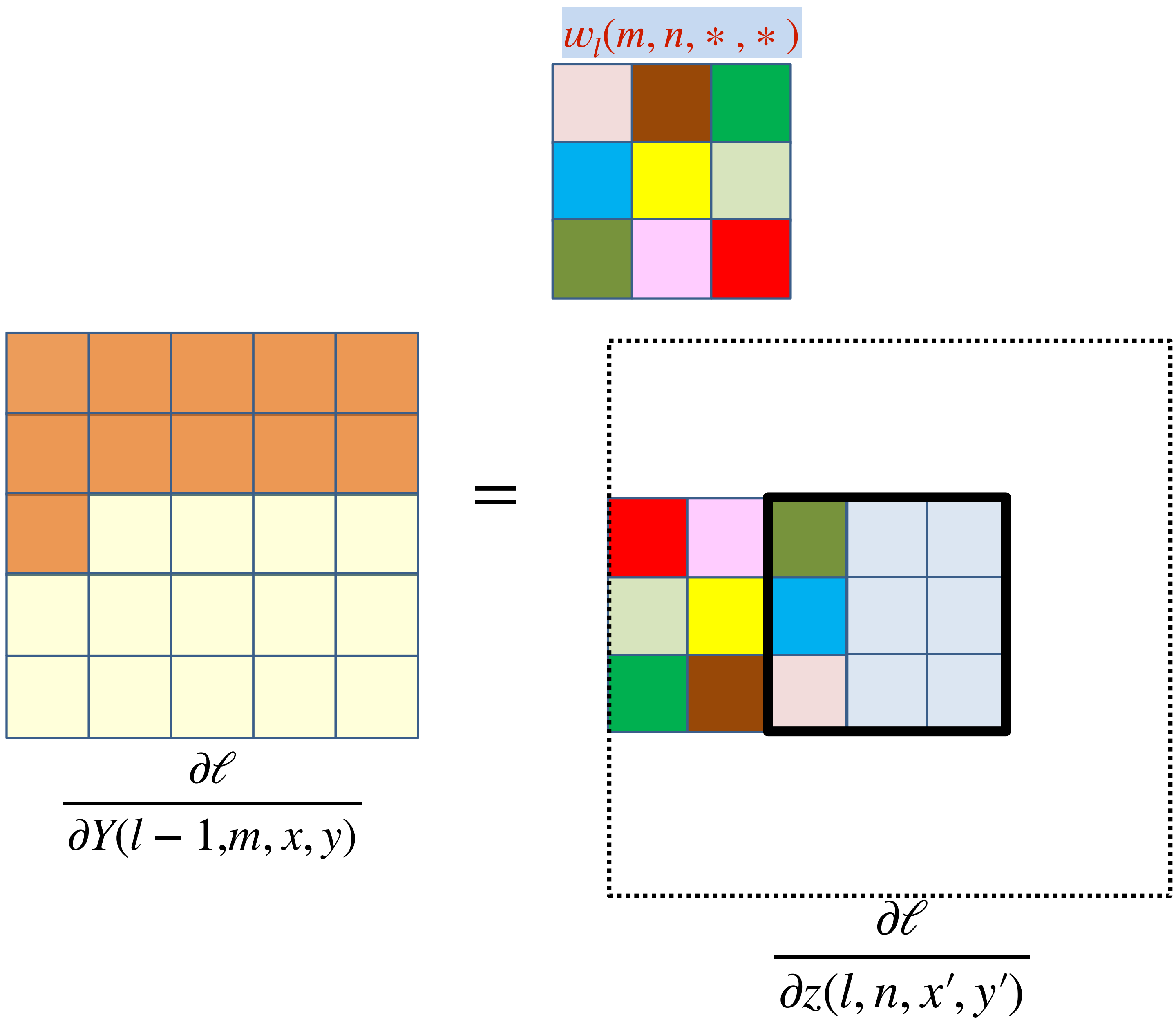
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



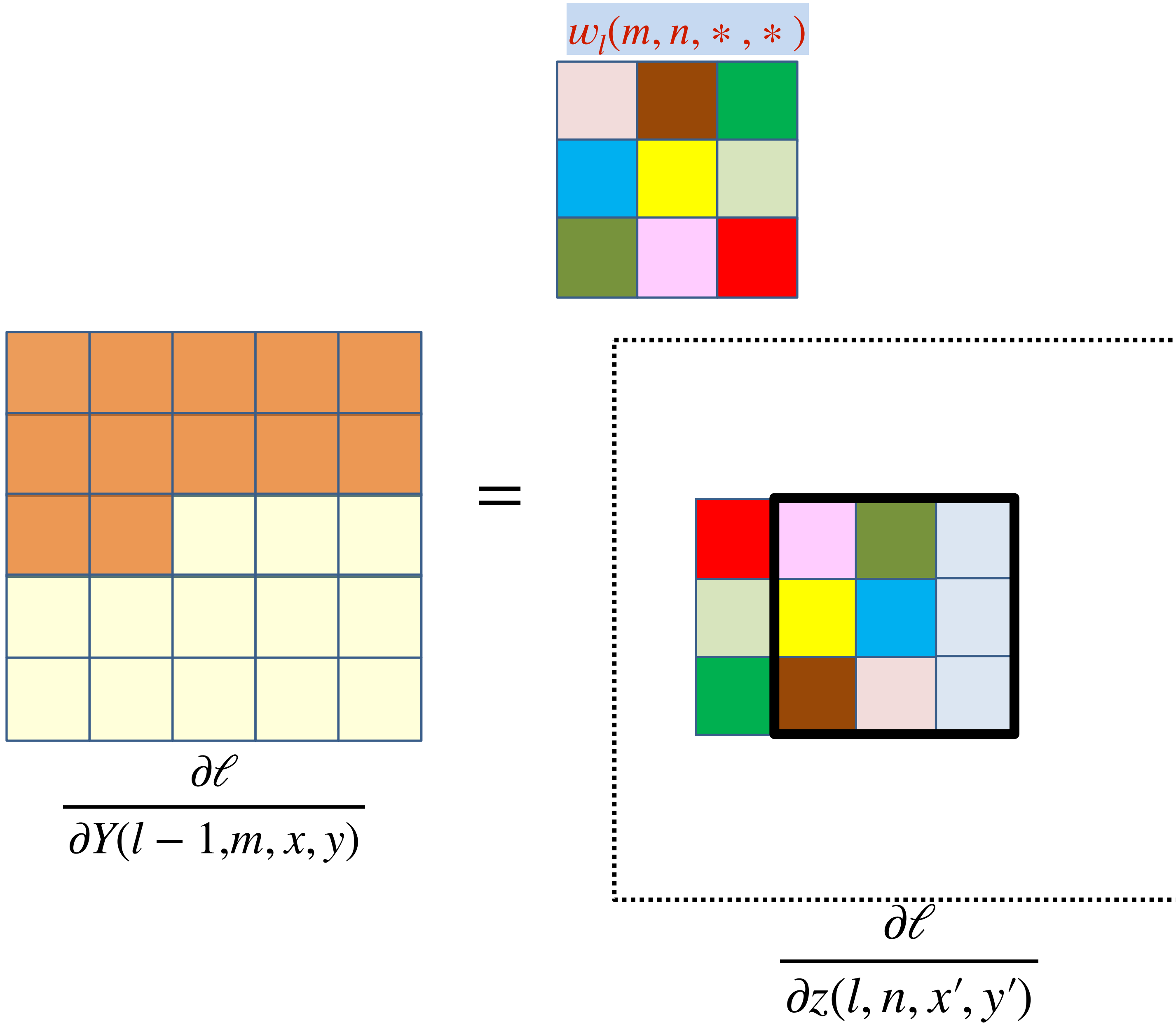


$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1, m)$ from a single $Z(l, n)$ map

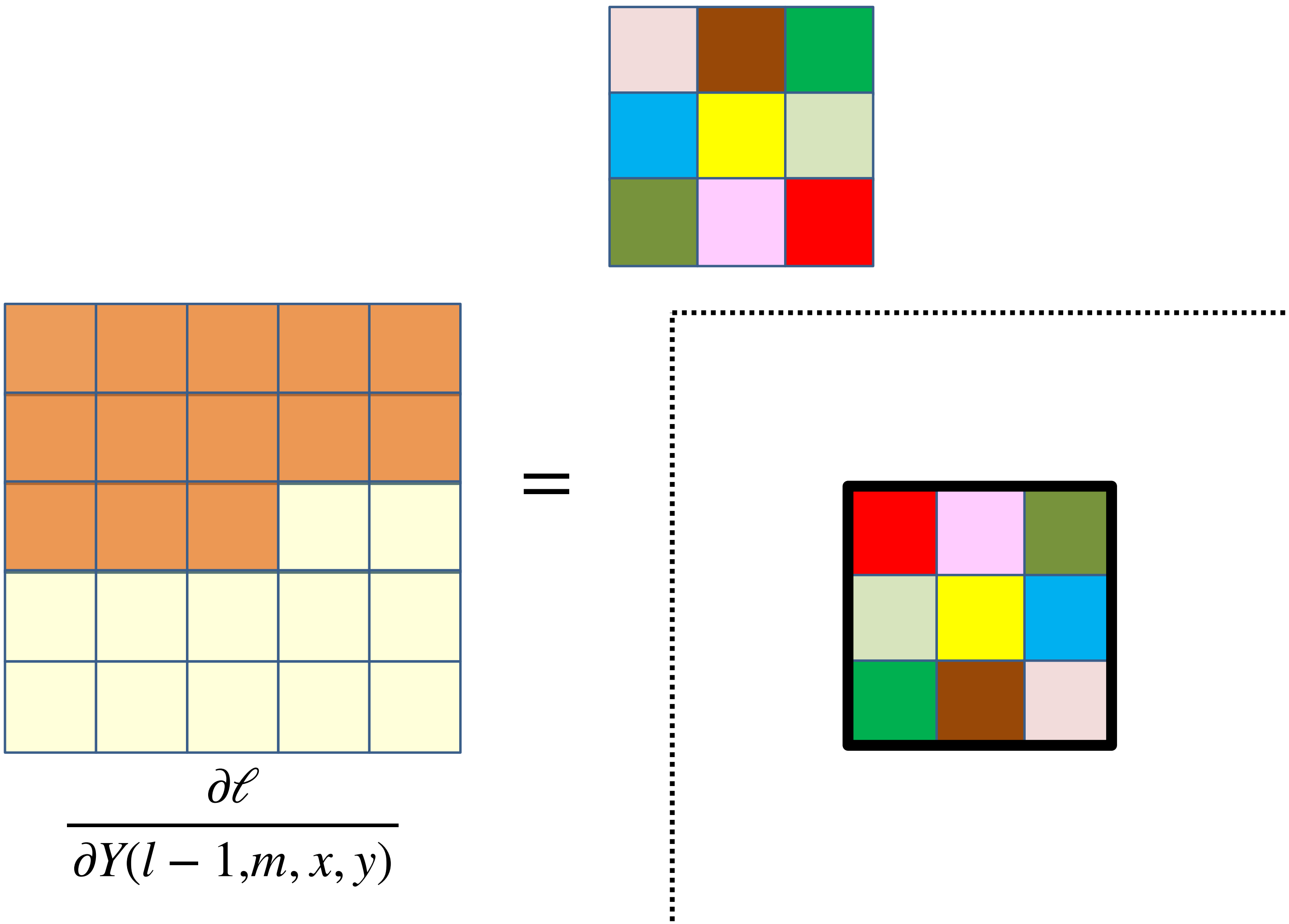


Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



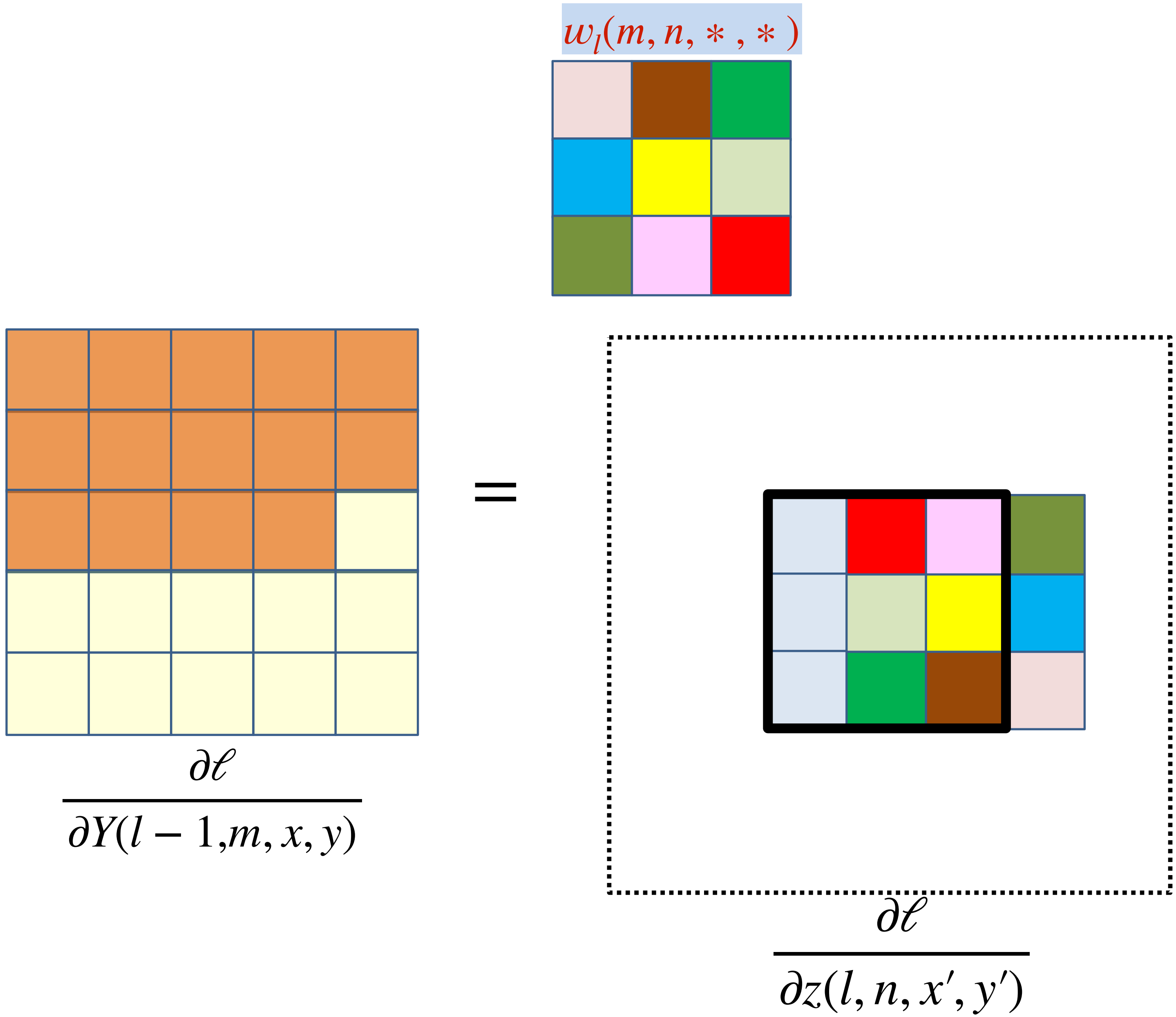
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

$w_l(m,n,*,*)$

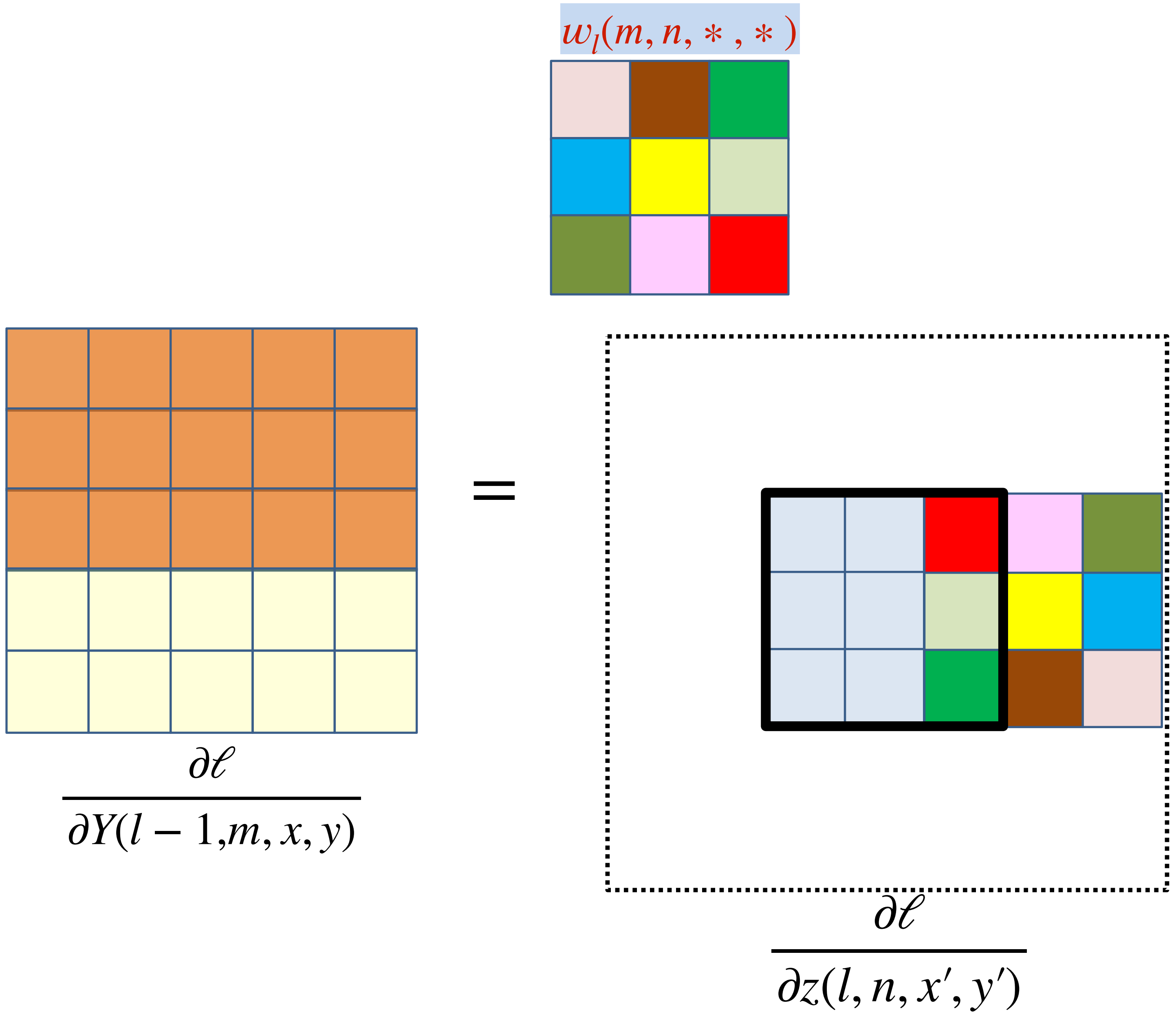


$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

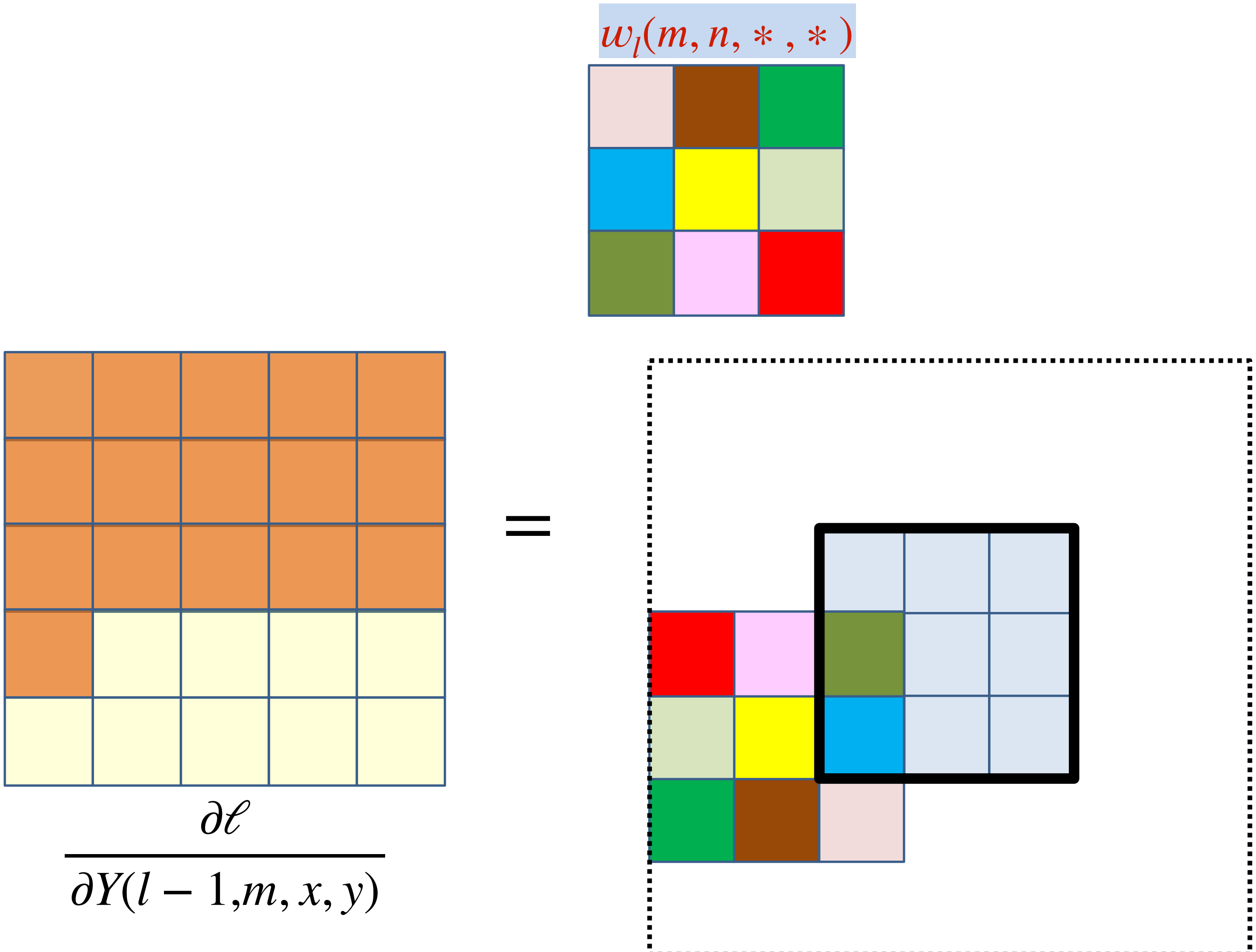
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



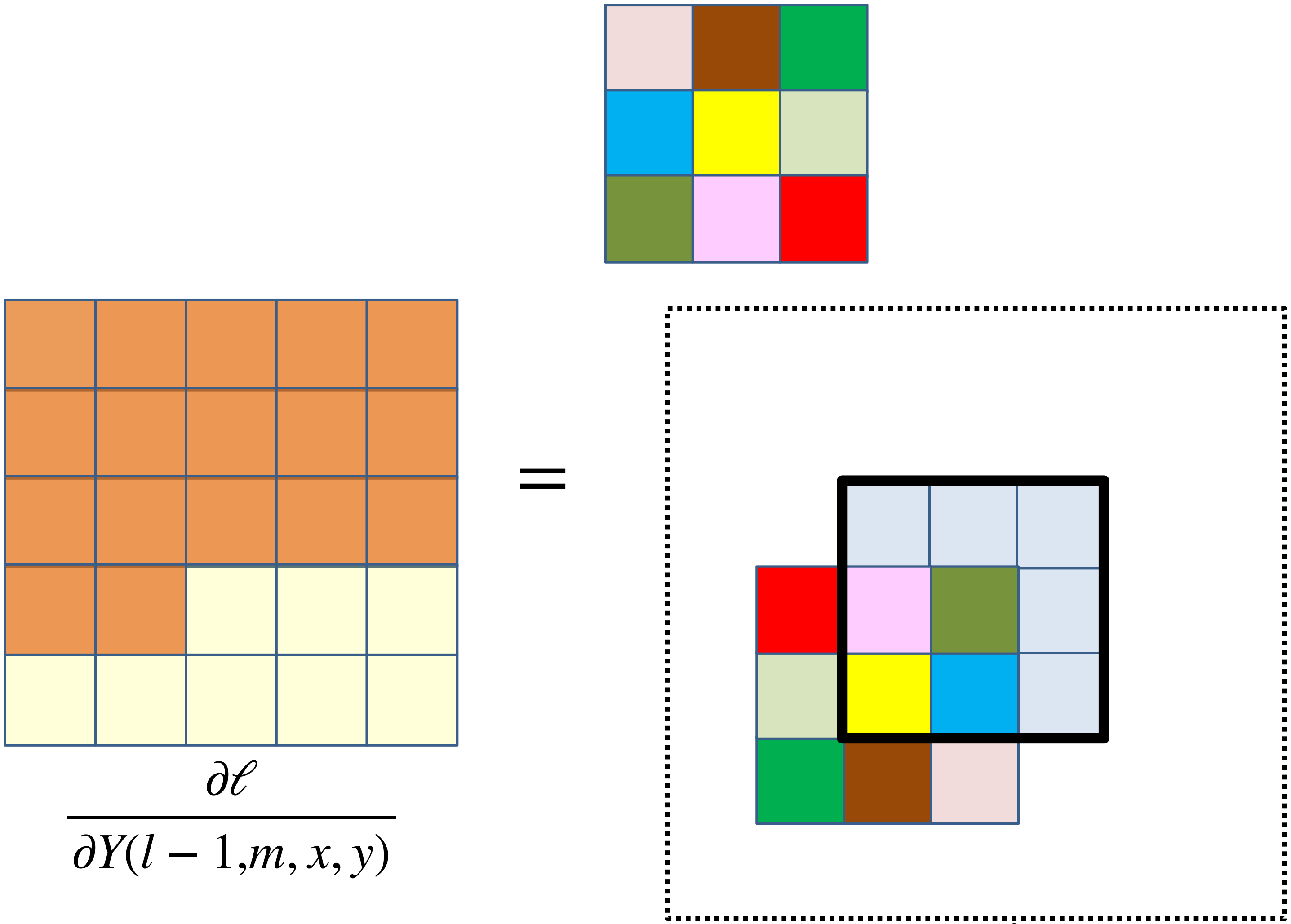
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \frac{\partial \ell}{\partial z(l, n, x', y')}$$

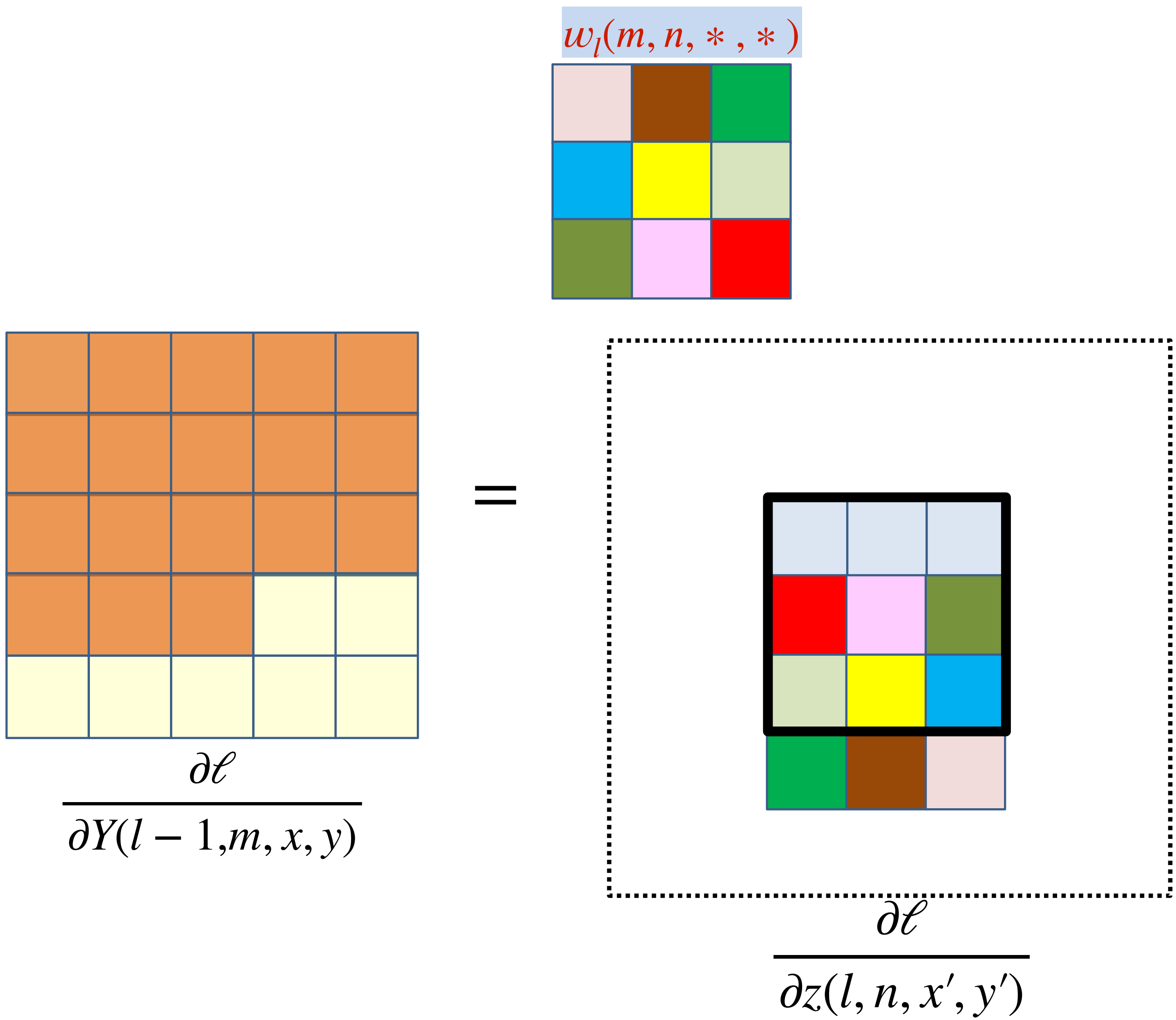
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

$w_l(m, n, *, *)$

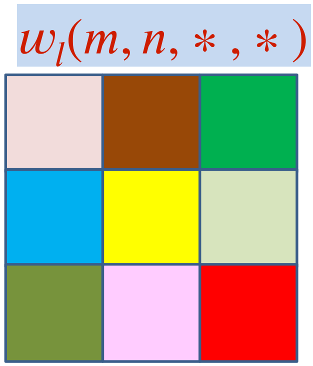


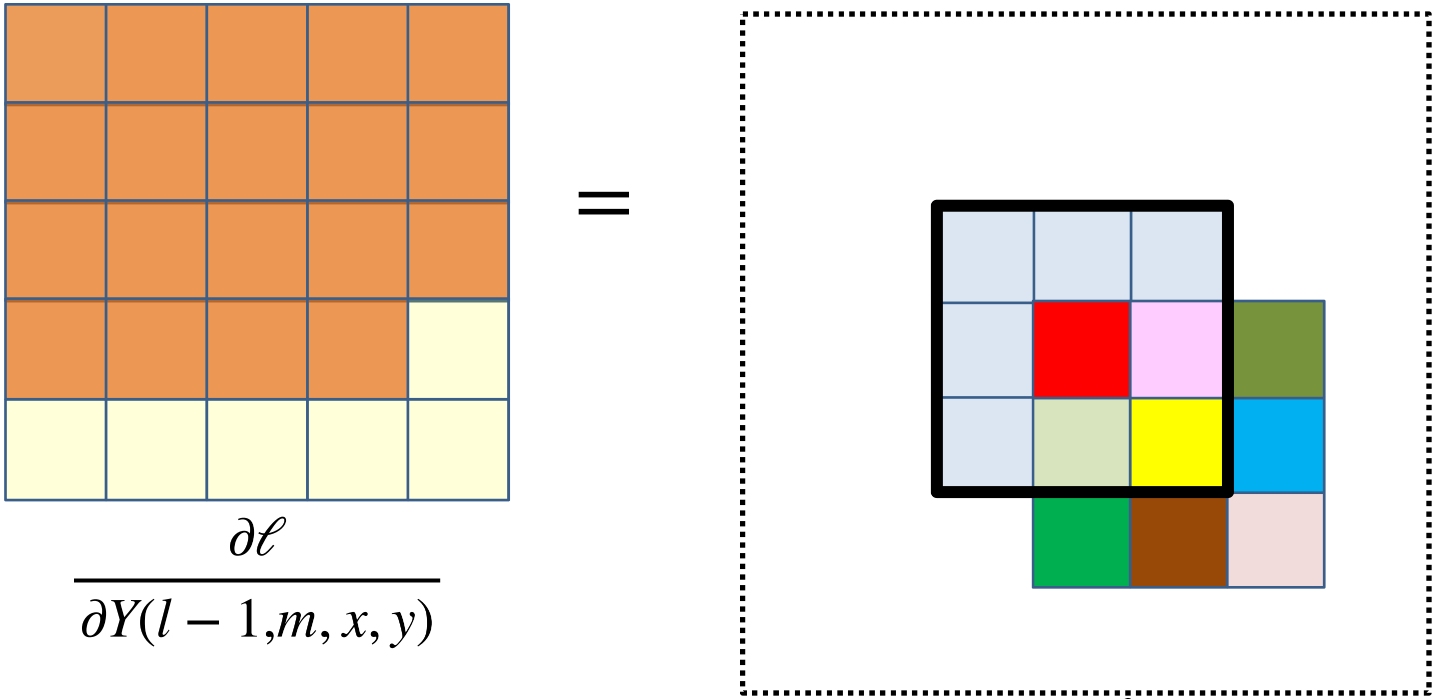
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

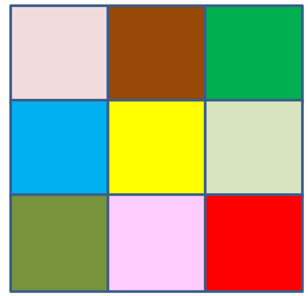


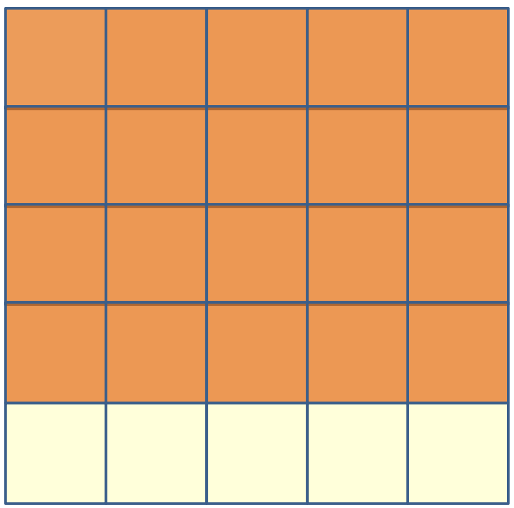


$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

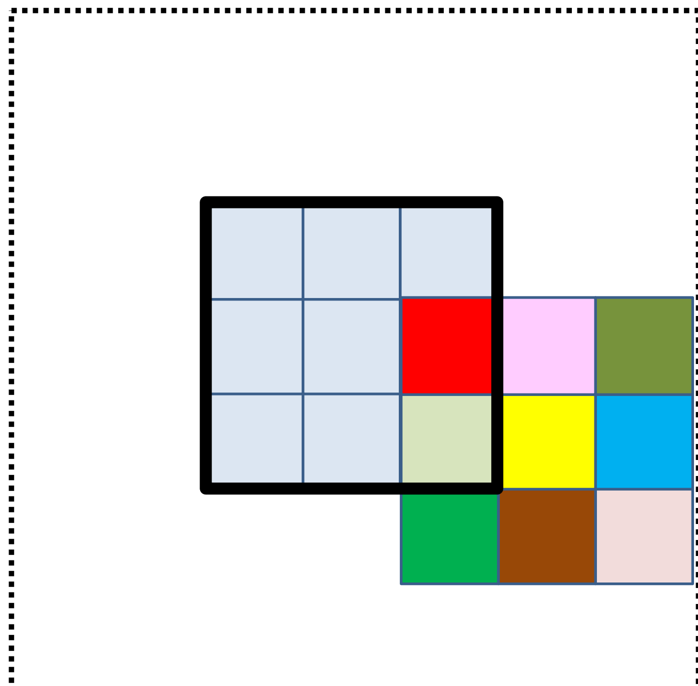
$w_l(m, n, *, *)$





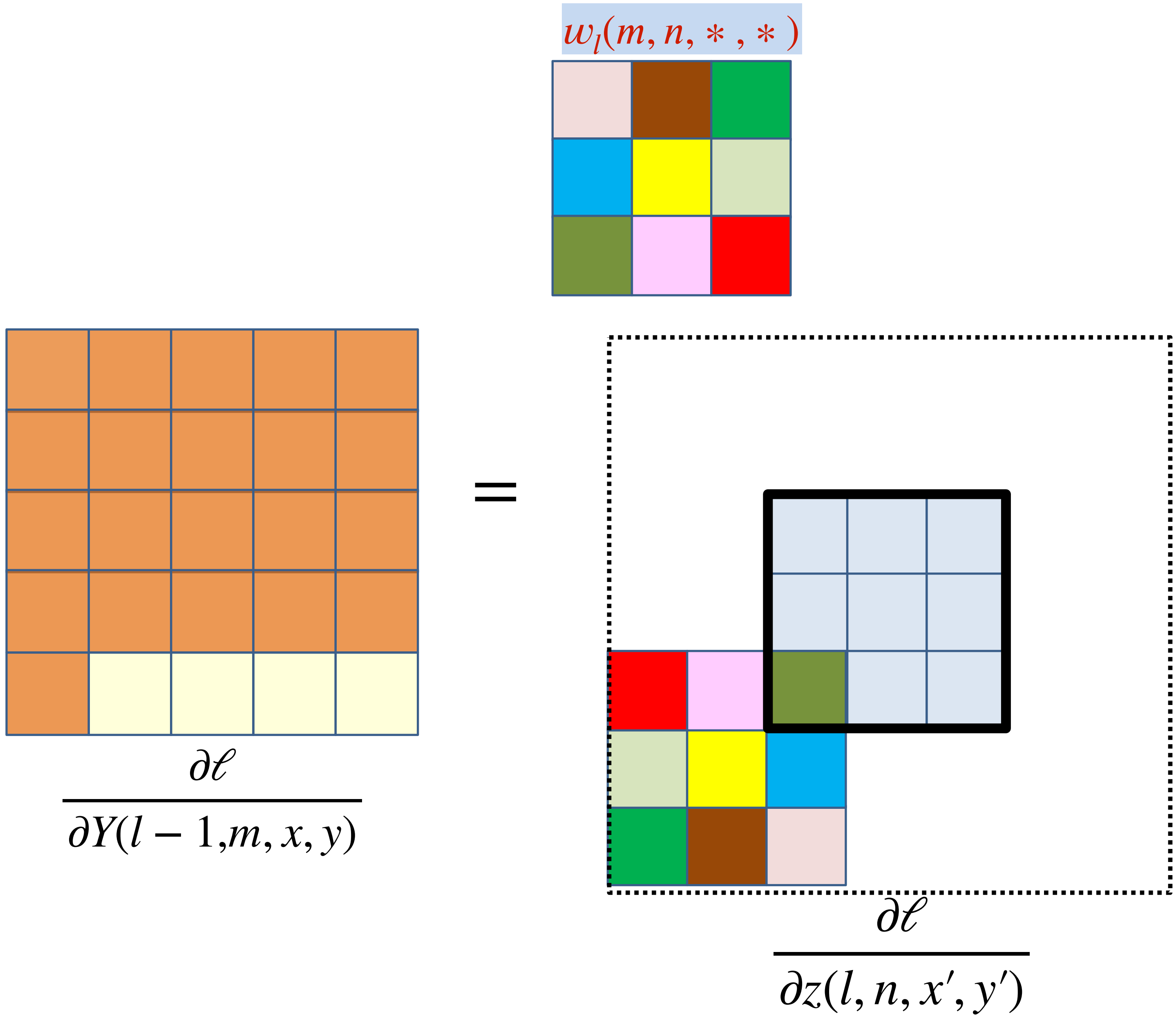
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$



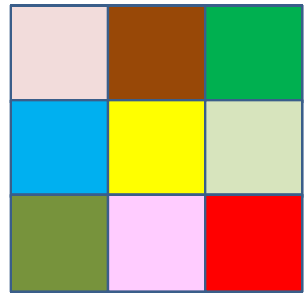
$$\frac{\partial \ell}{\partial z(l, n, x', y')}$$

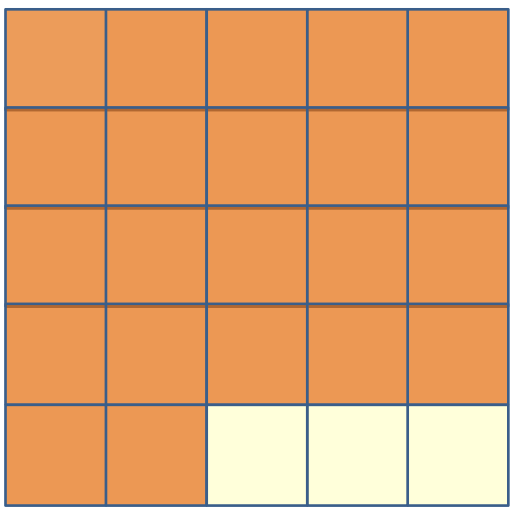
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

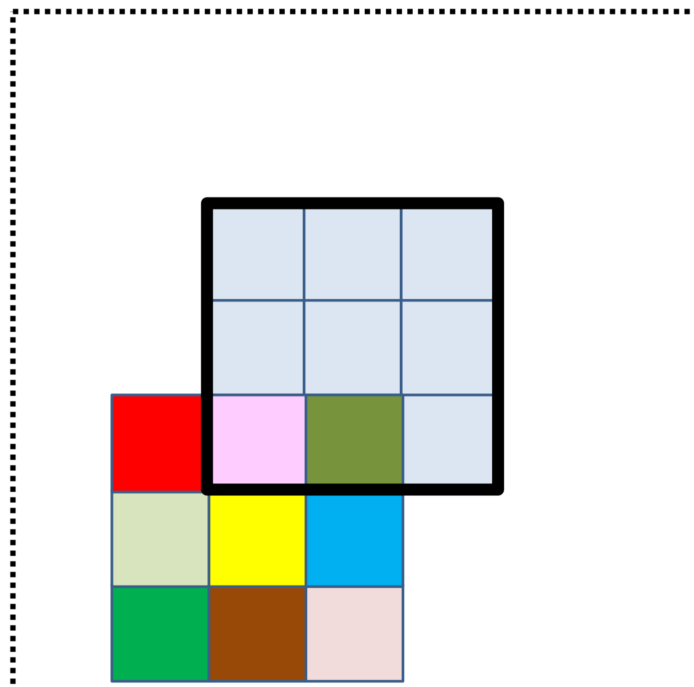
$w_l(m,n,*,*)$





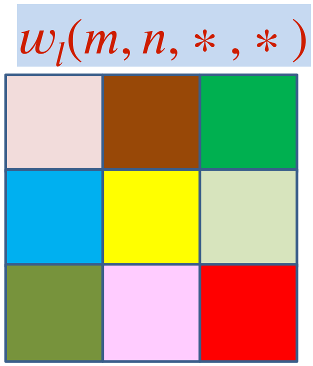
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)}$$

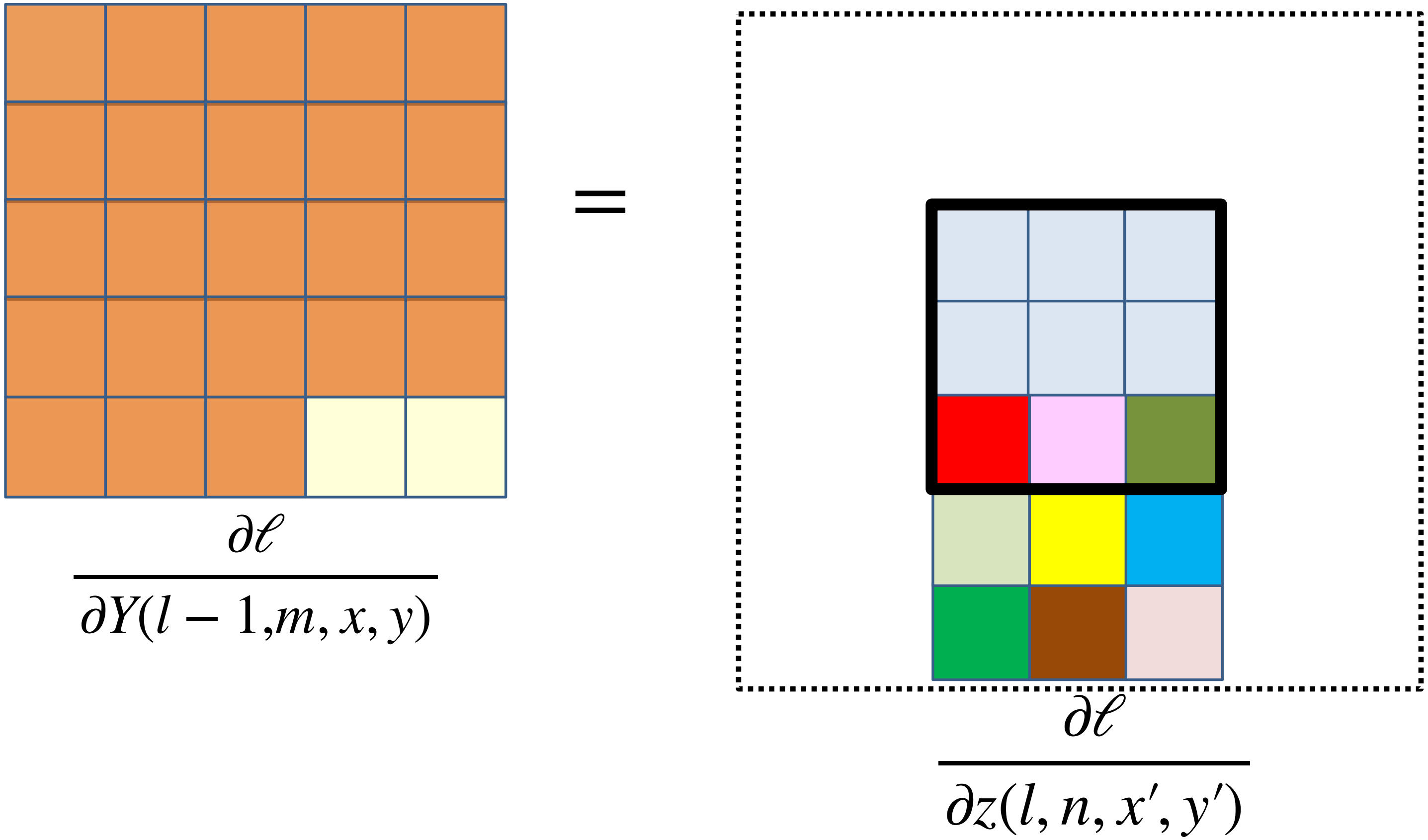
=



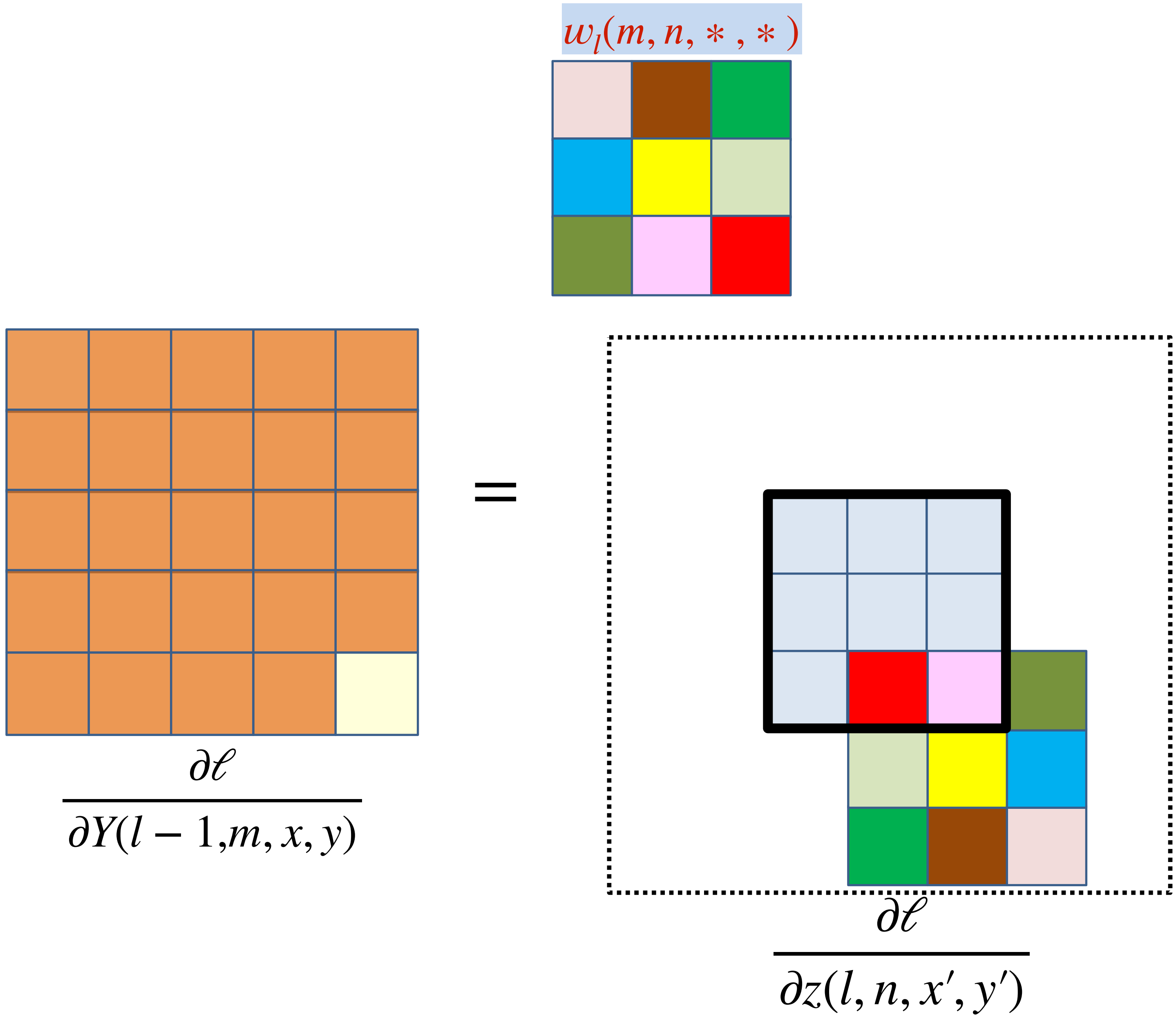
$$\frac{\partial \ell}{\partial z(l,n,x',y')}$$

Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map

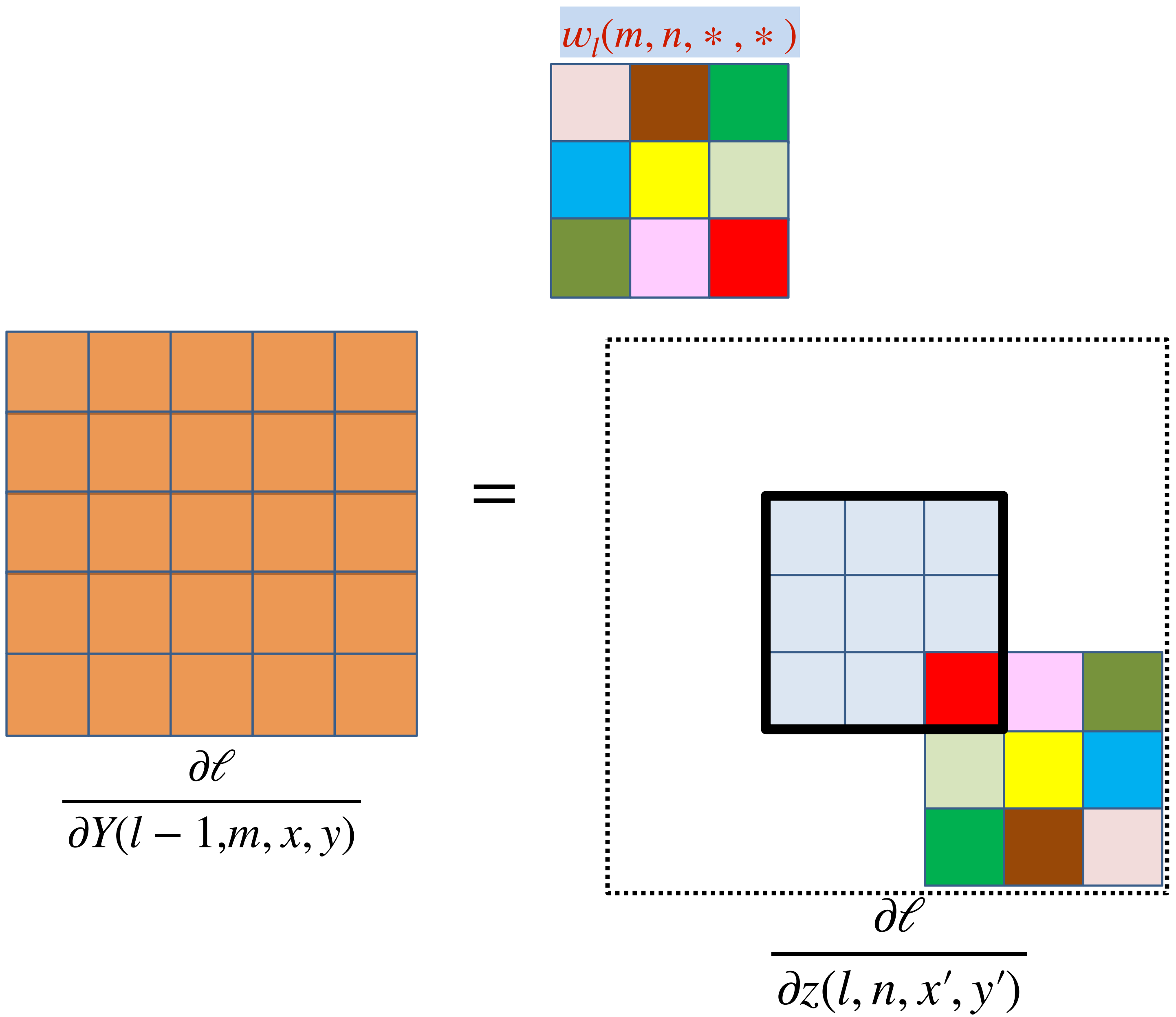




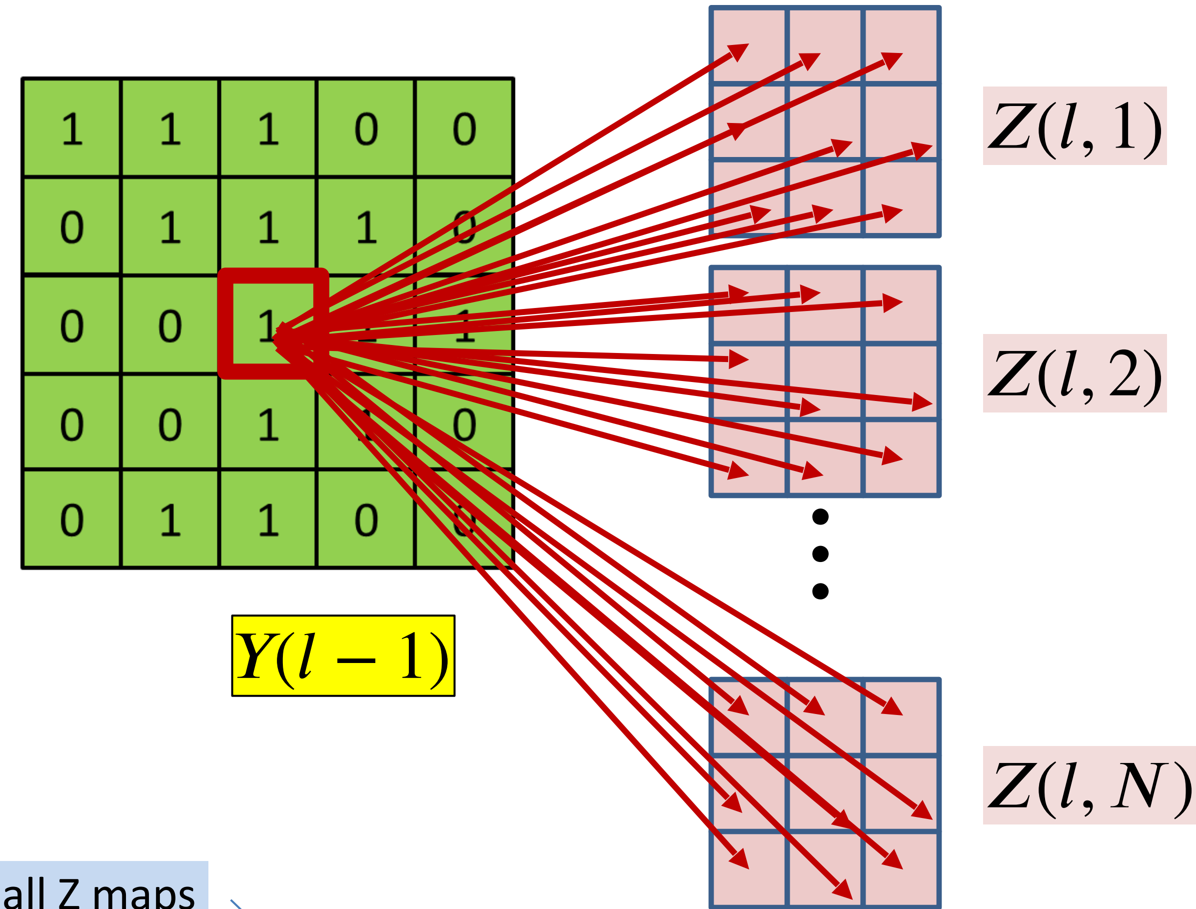
Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



Derivative at $Y(l-1,m)$ from a single $Z(l,n)$ map



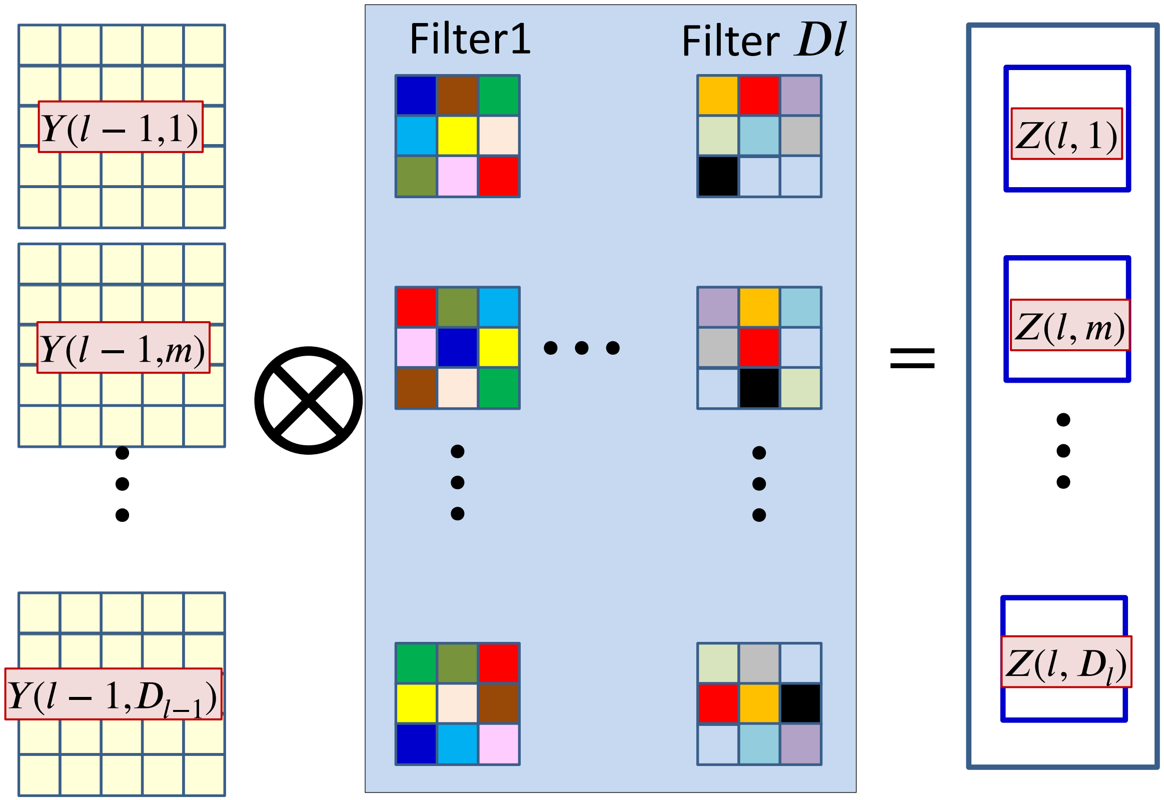
BP: Convolutional layer



Summing over all Z maps

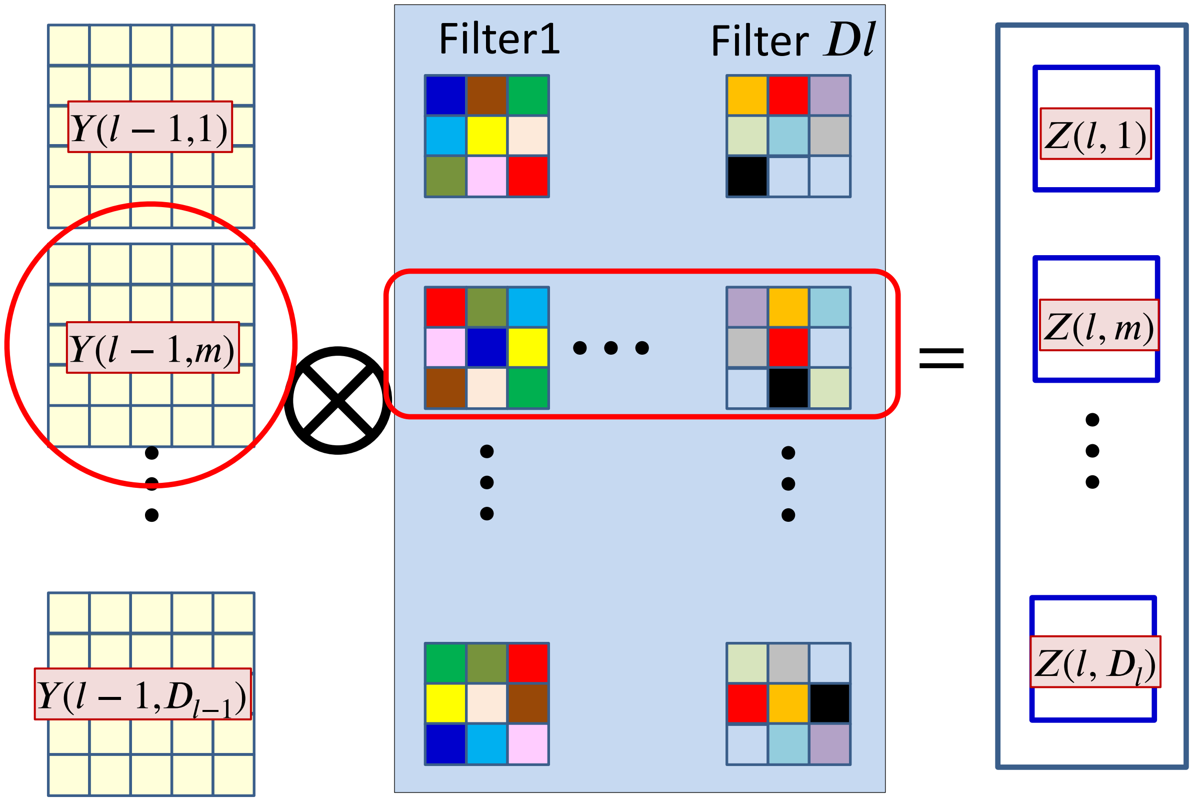
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_{n} \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

The actual convolutions

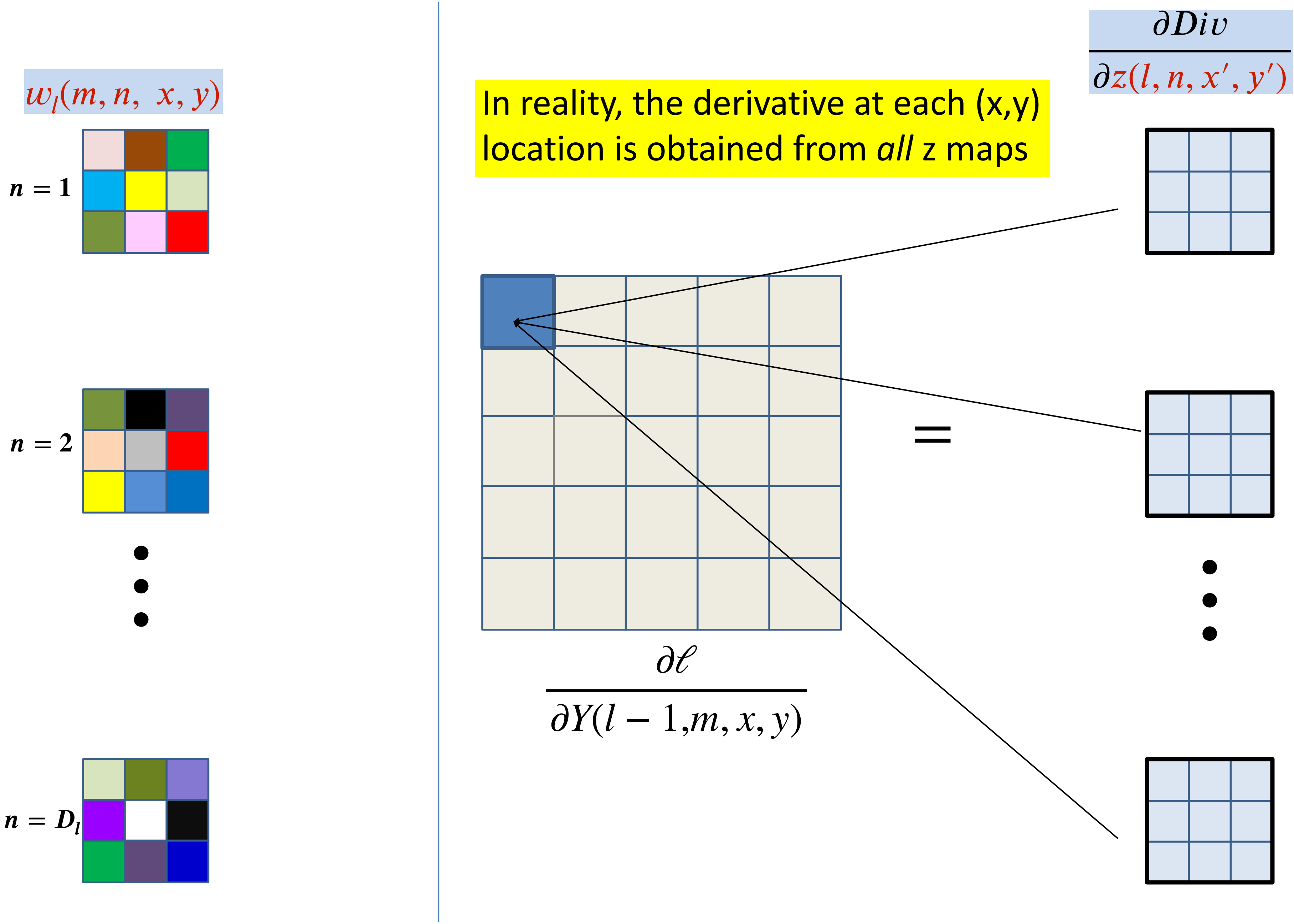


- The D_l affine maps are produced by convolving with D_l filters

The actual convolutions

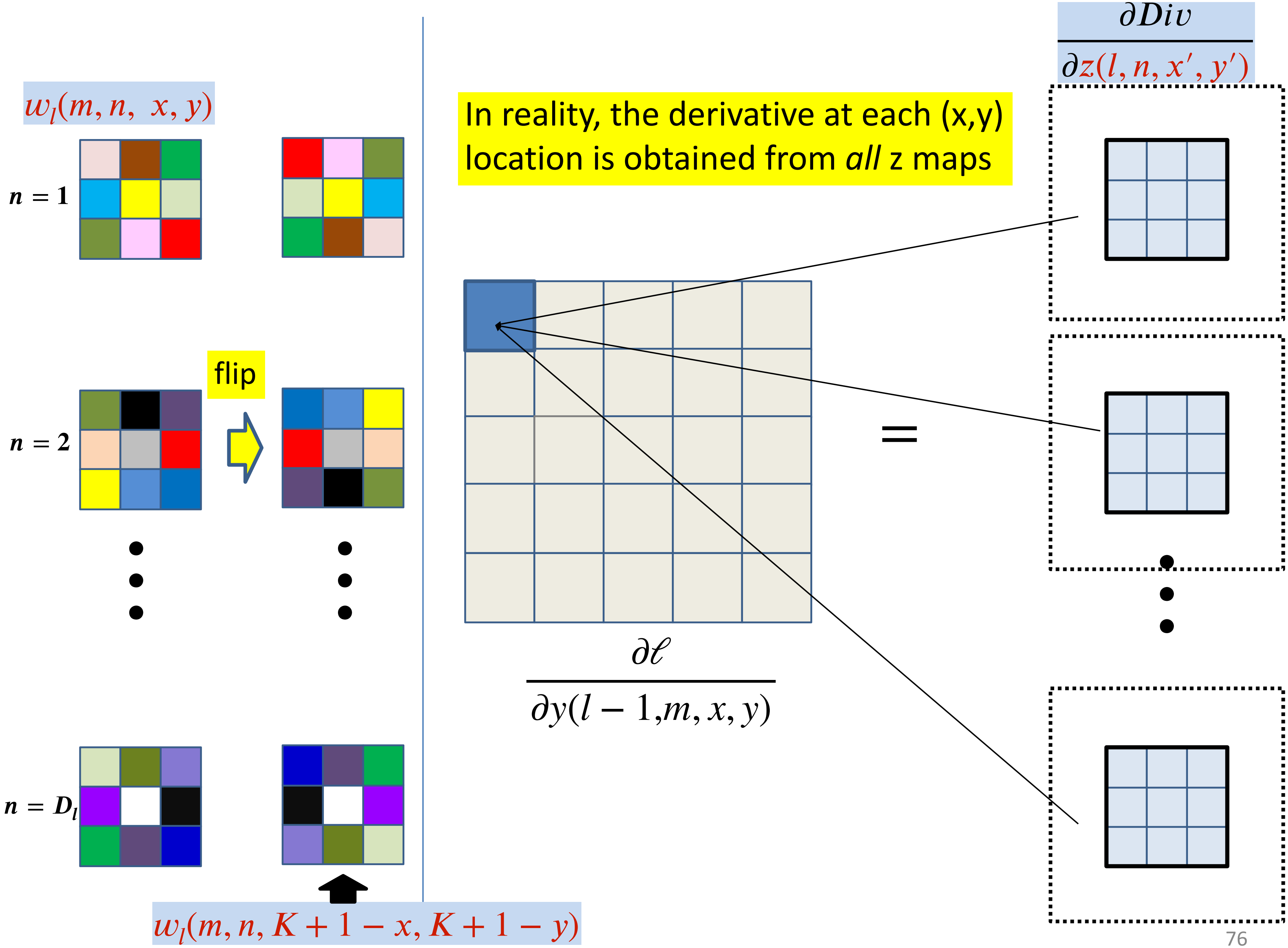


- The D_l affine maps are produced by convolving with D_l filters
- The m^{th} Y map always convolves the m^{th} plane of the filters
- The derivative for the m^{th} Y map will invoke the m^{th} plane of *all* the filters

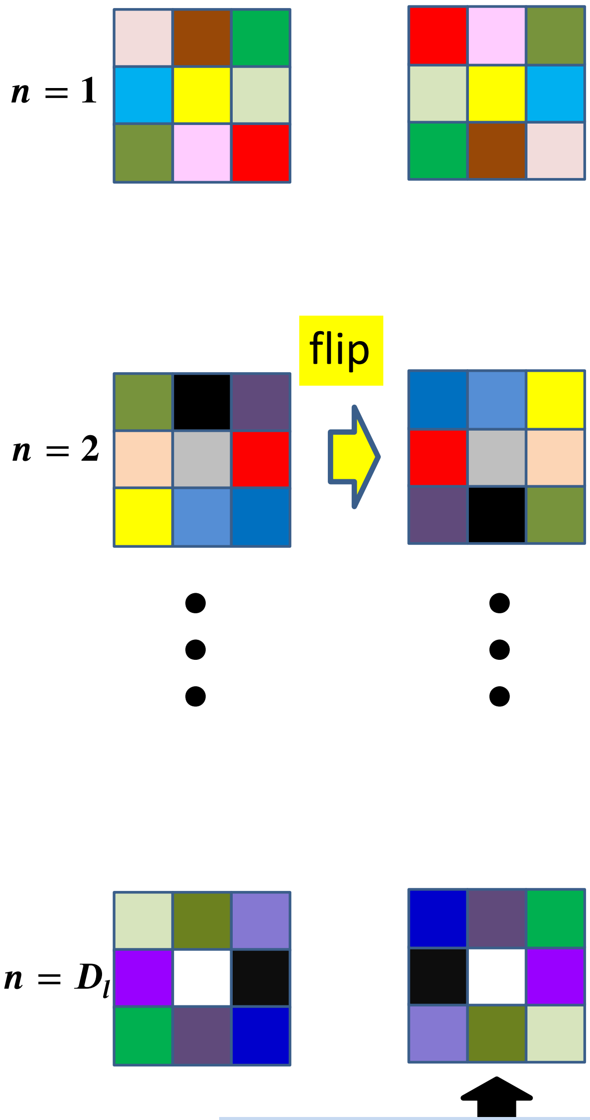


$w_l(m, n, x, y)$
$n = 1$
$n = 2$
$n = D_l$
In reality, the derivative at each (x,y) location is obtained from *all* z maps
$\frac{\partial Div}{\partial z(l, n, x', y')}$
$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$
=

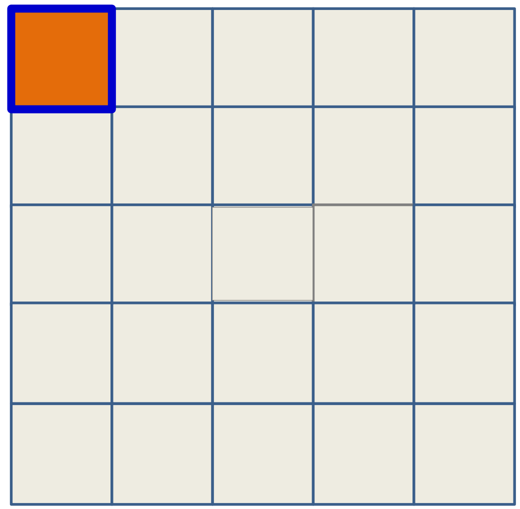
In reality, the derivative at each (x,y) location is obtained from *all* z maps



$w_l(m, n, x, y)$

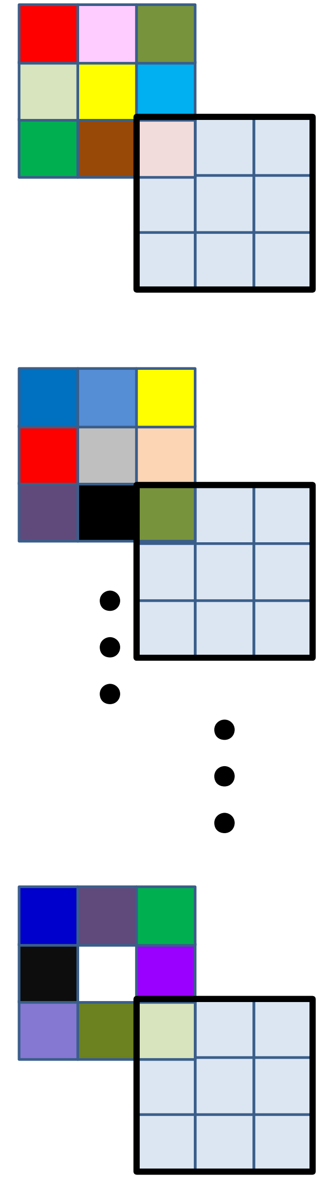


$w_l(m, n, K+1-x, K+1-y)$

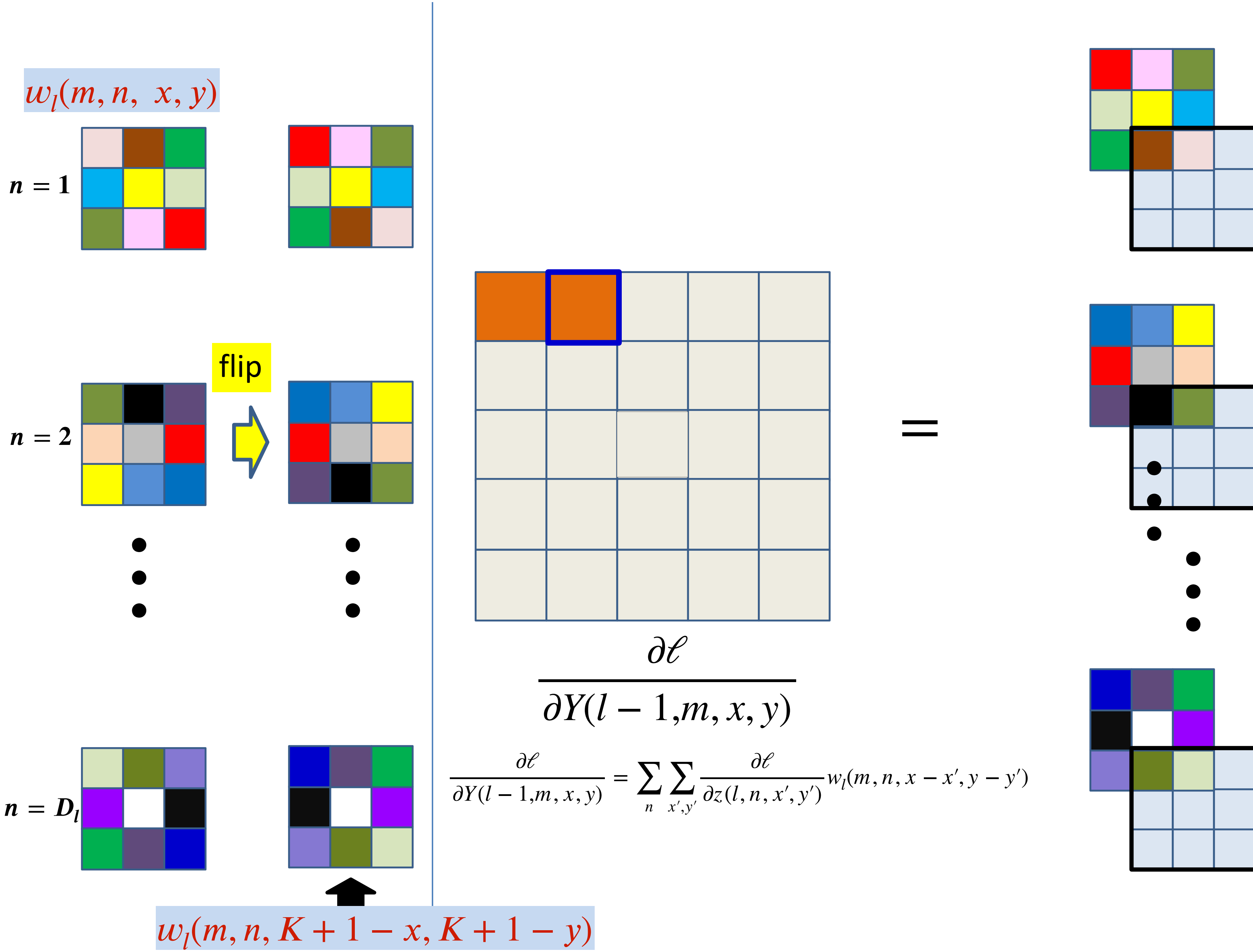


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$

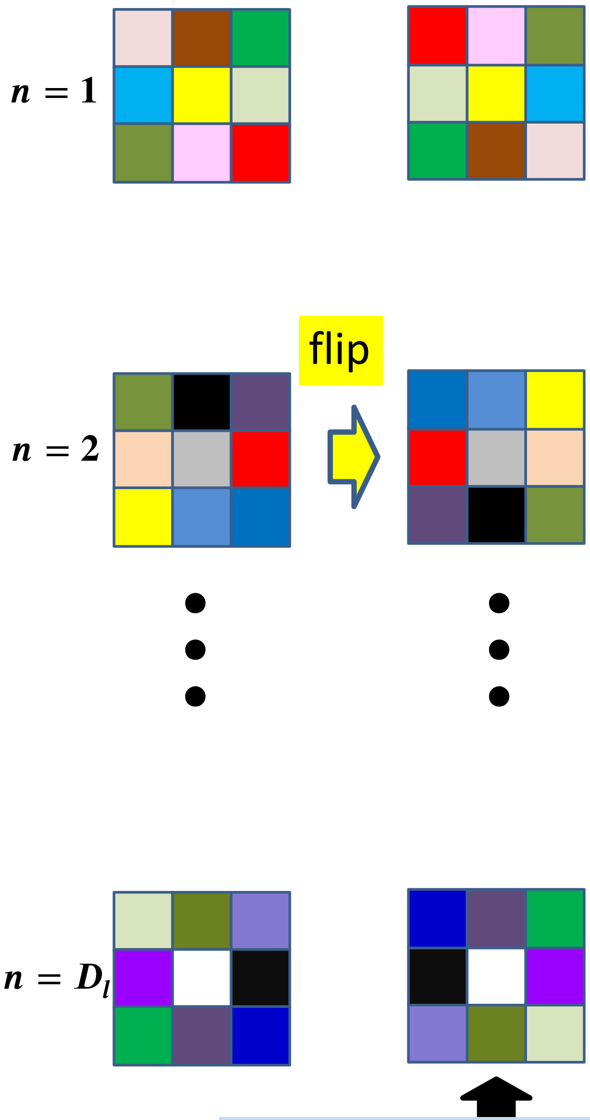


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$

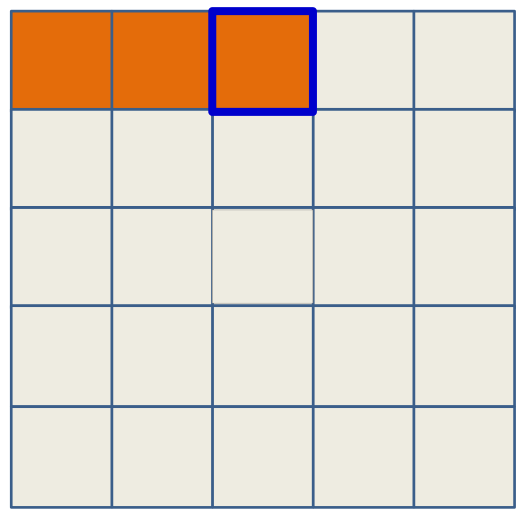


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$

$w_l(m, n, x, y)$

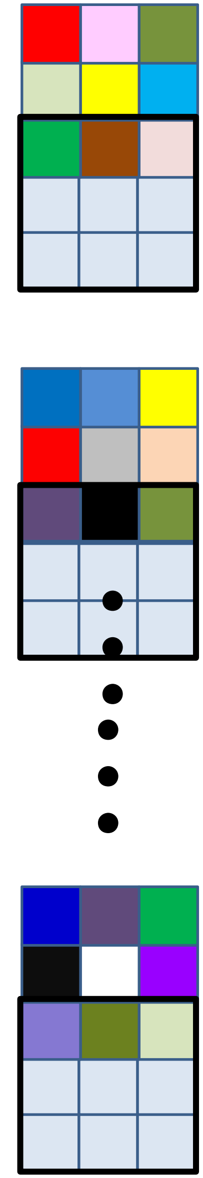


$w_l(m, n, K+1-x, K+1-y)$

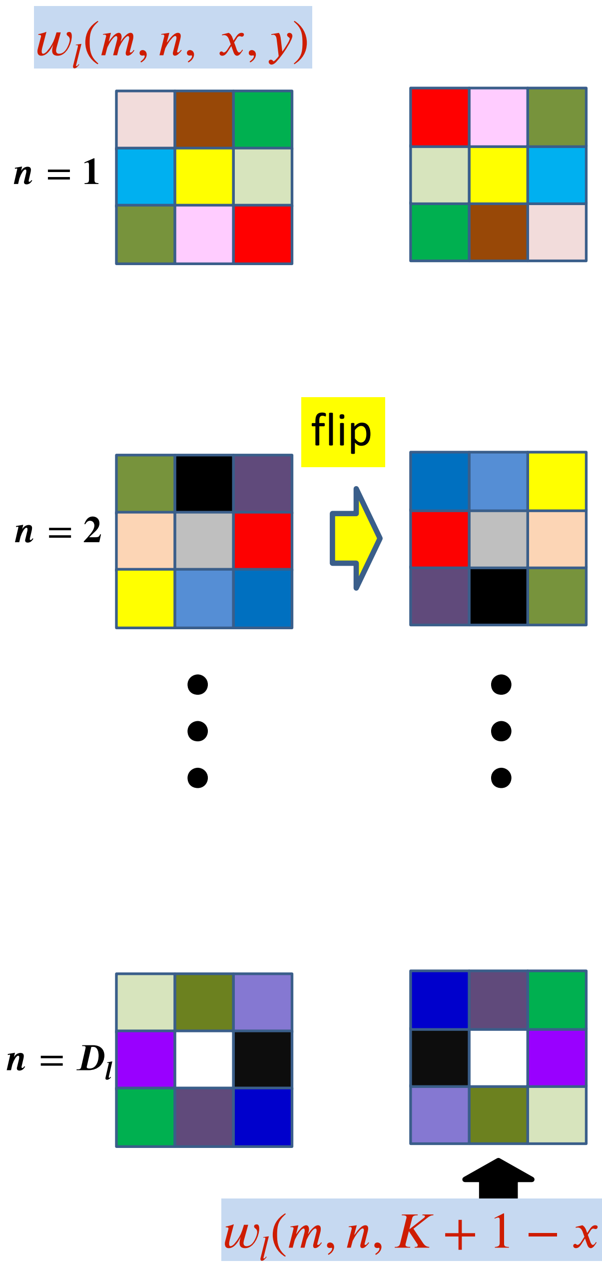


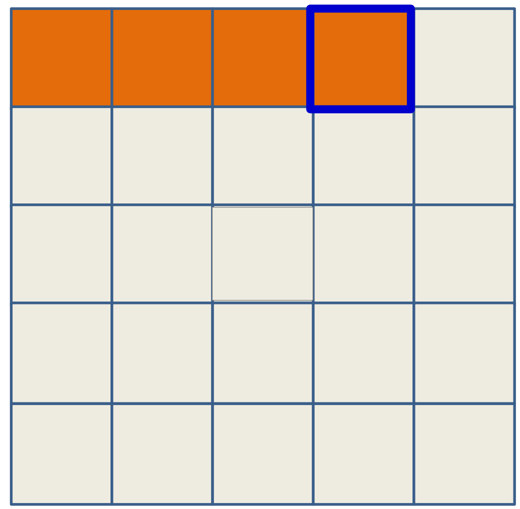
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

=



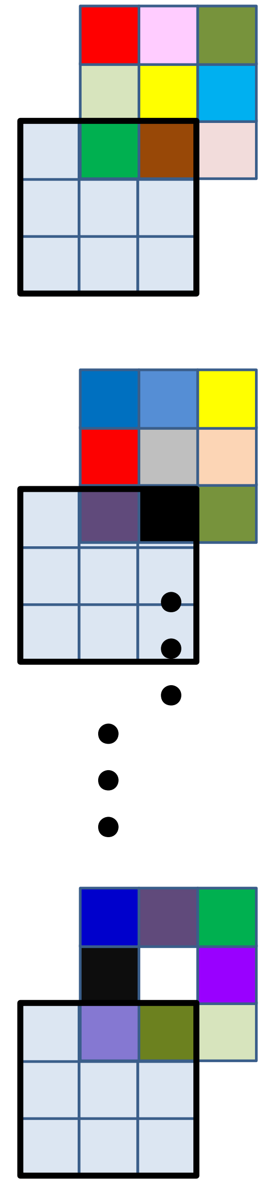
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$



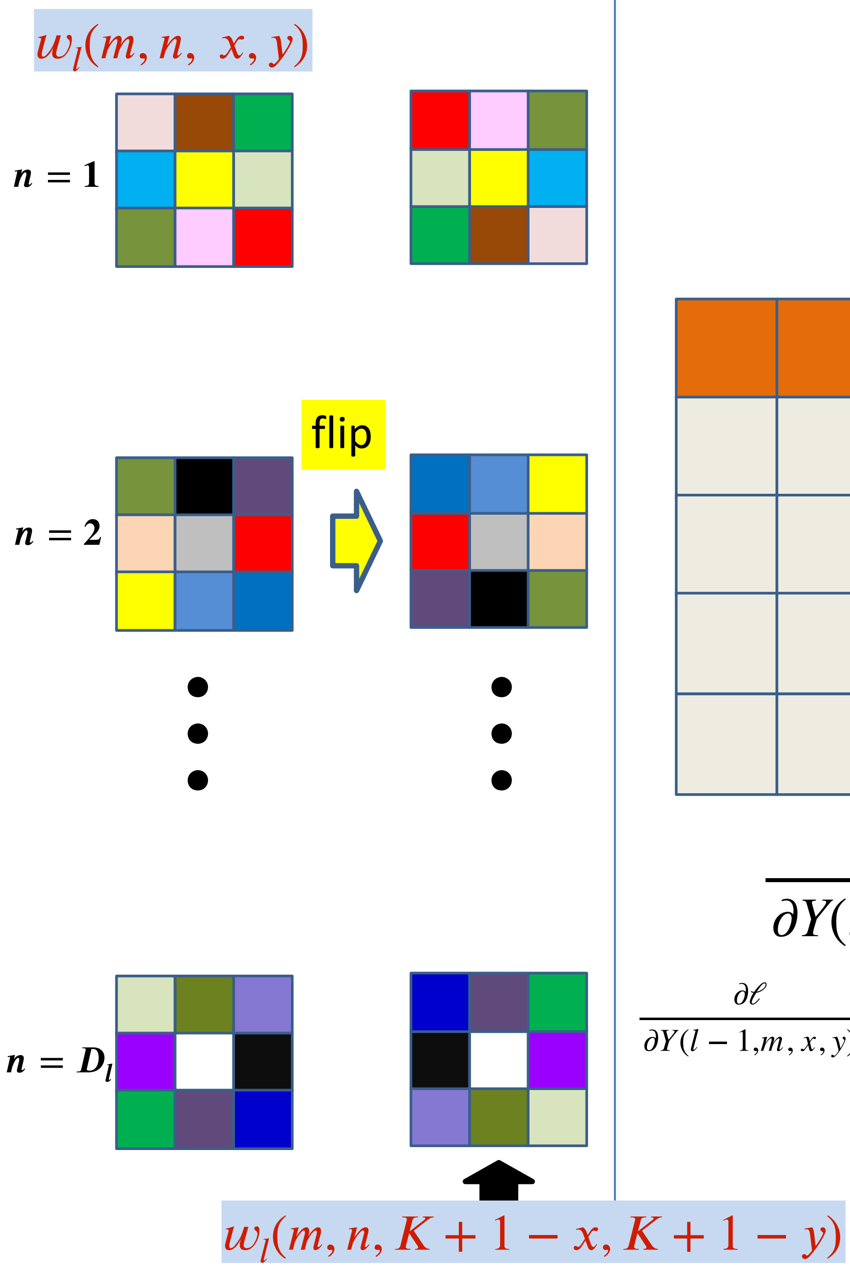


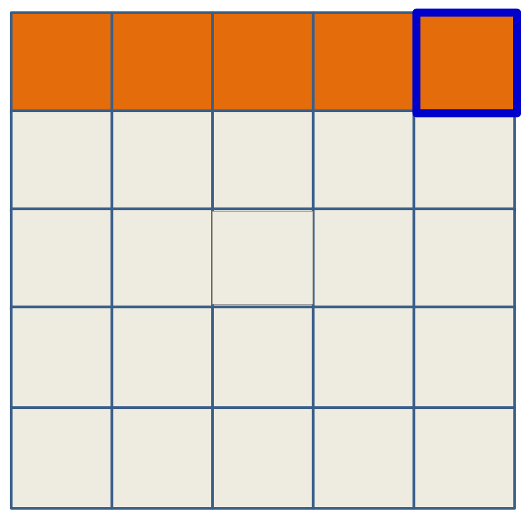
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$



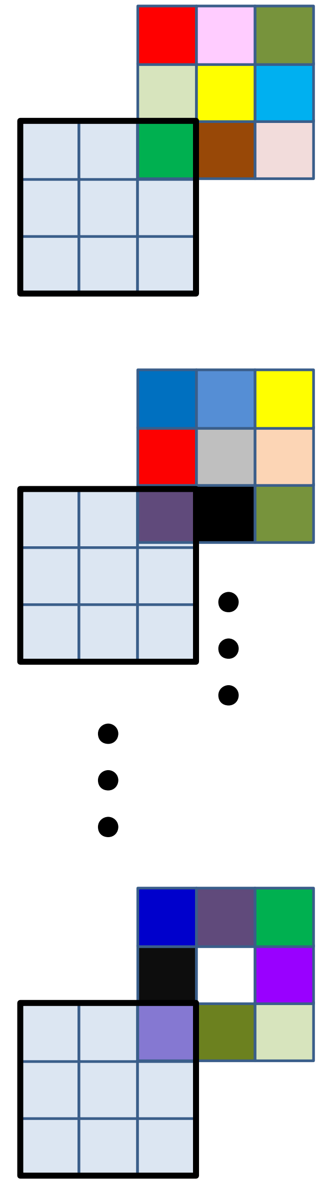
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$



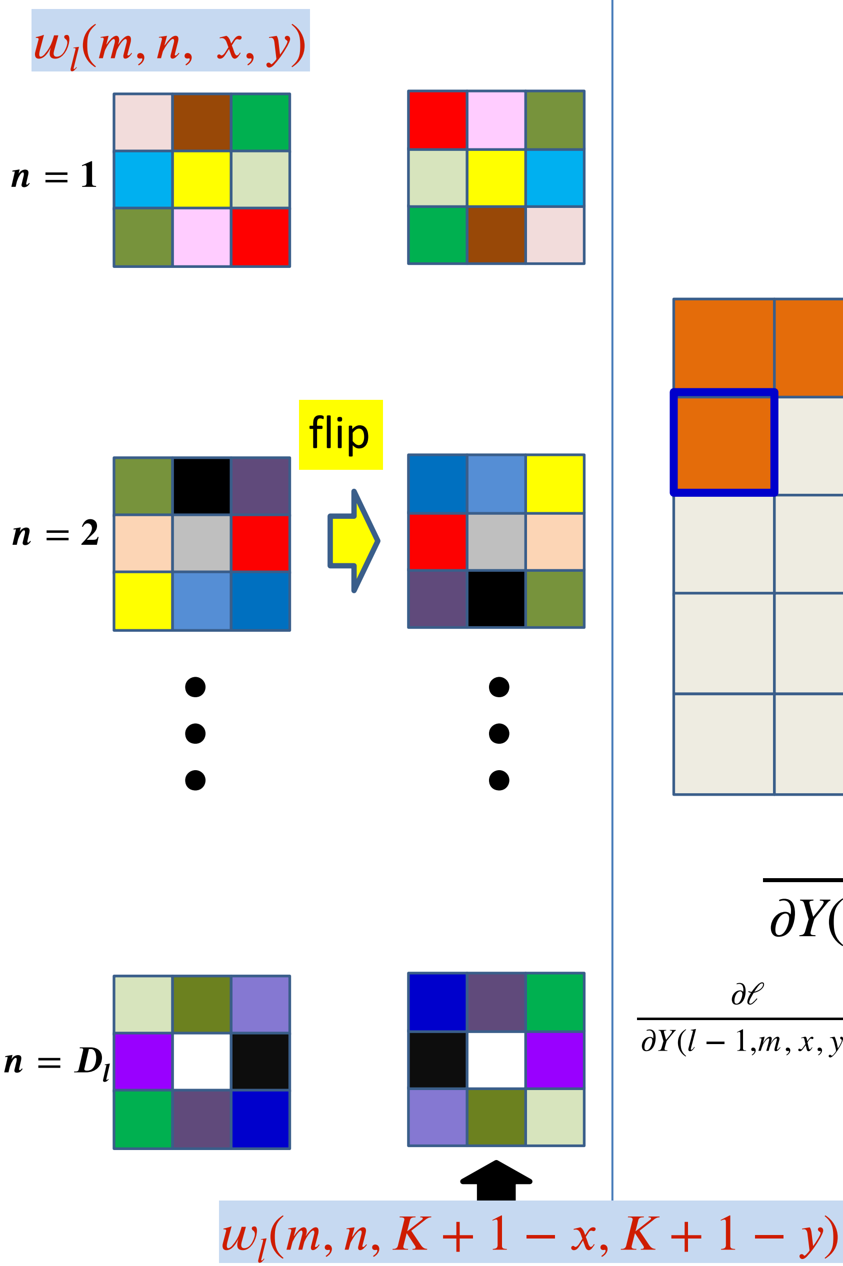


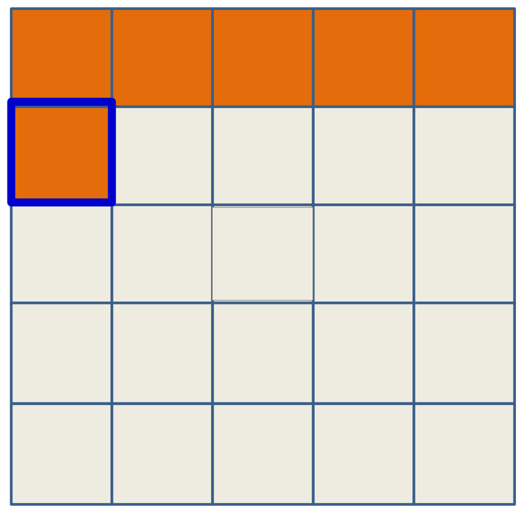
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$



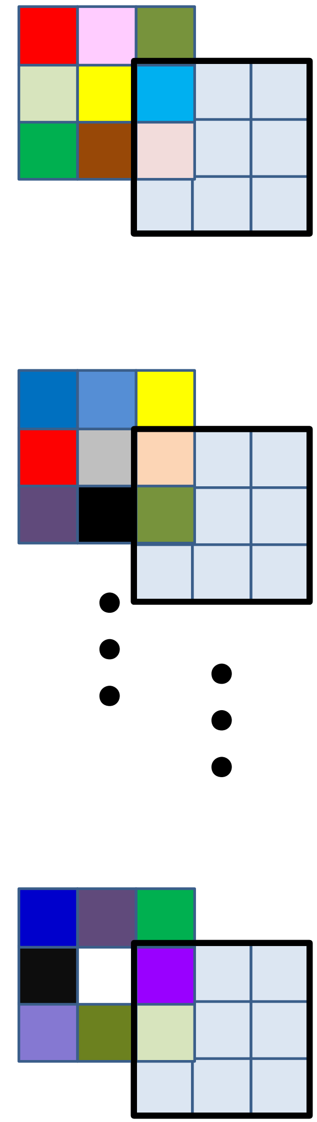
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$



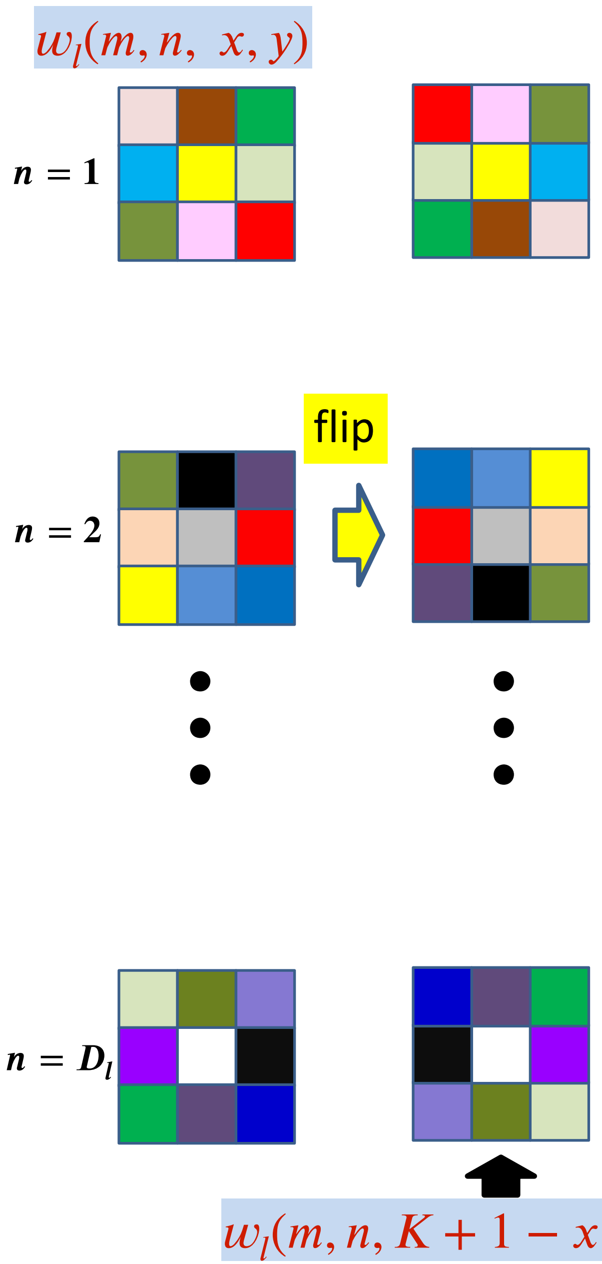


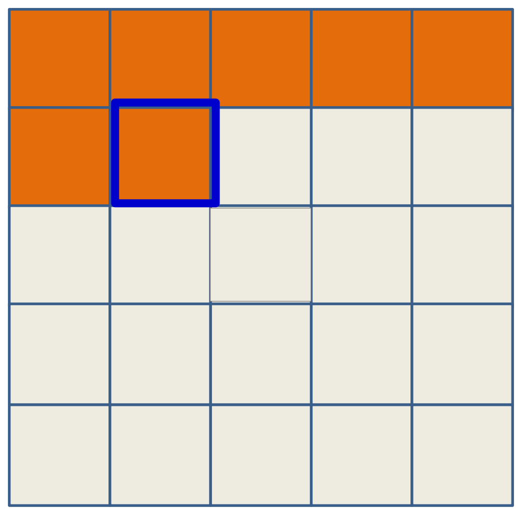
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

=



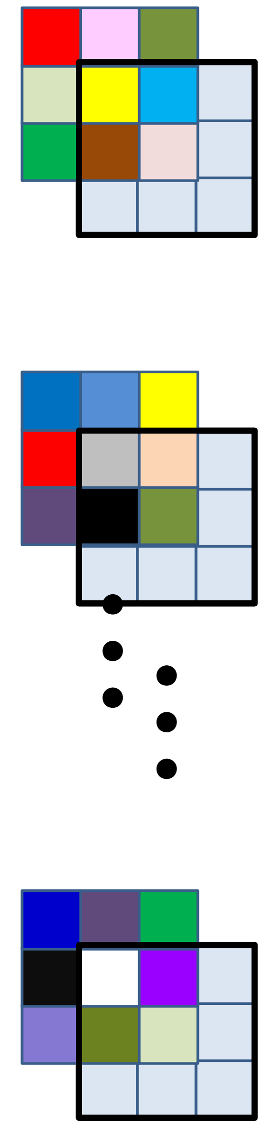
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$



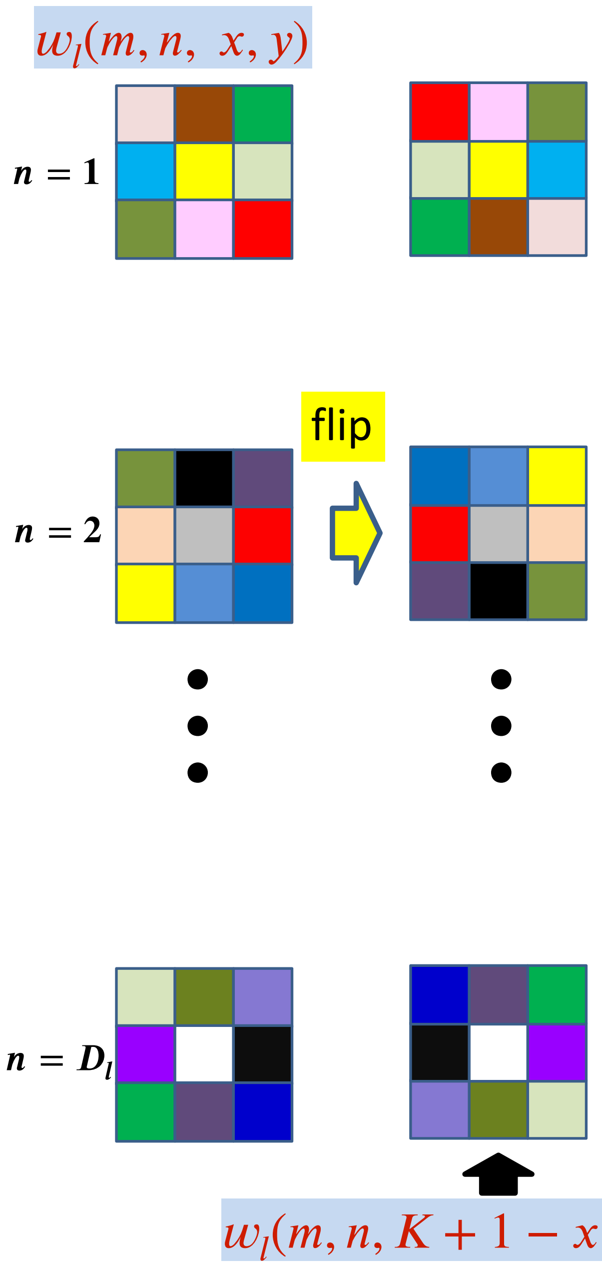


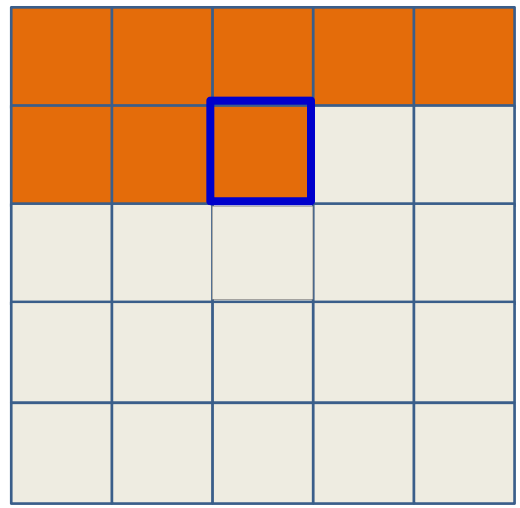
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

=



$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$

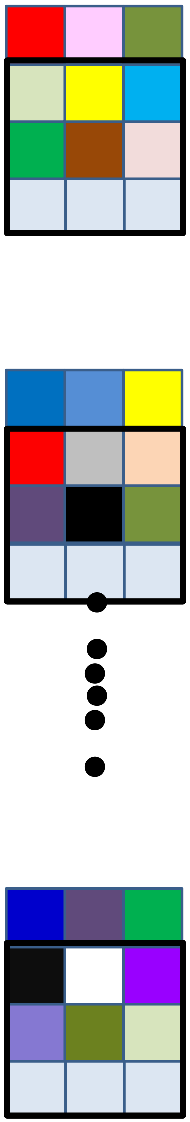


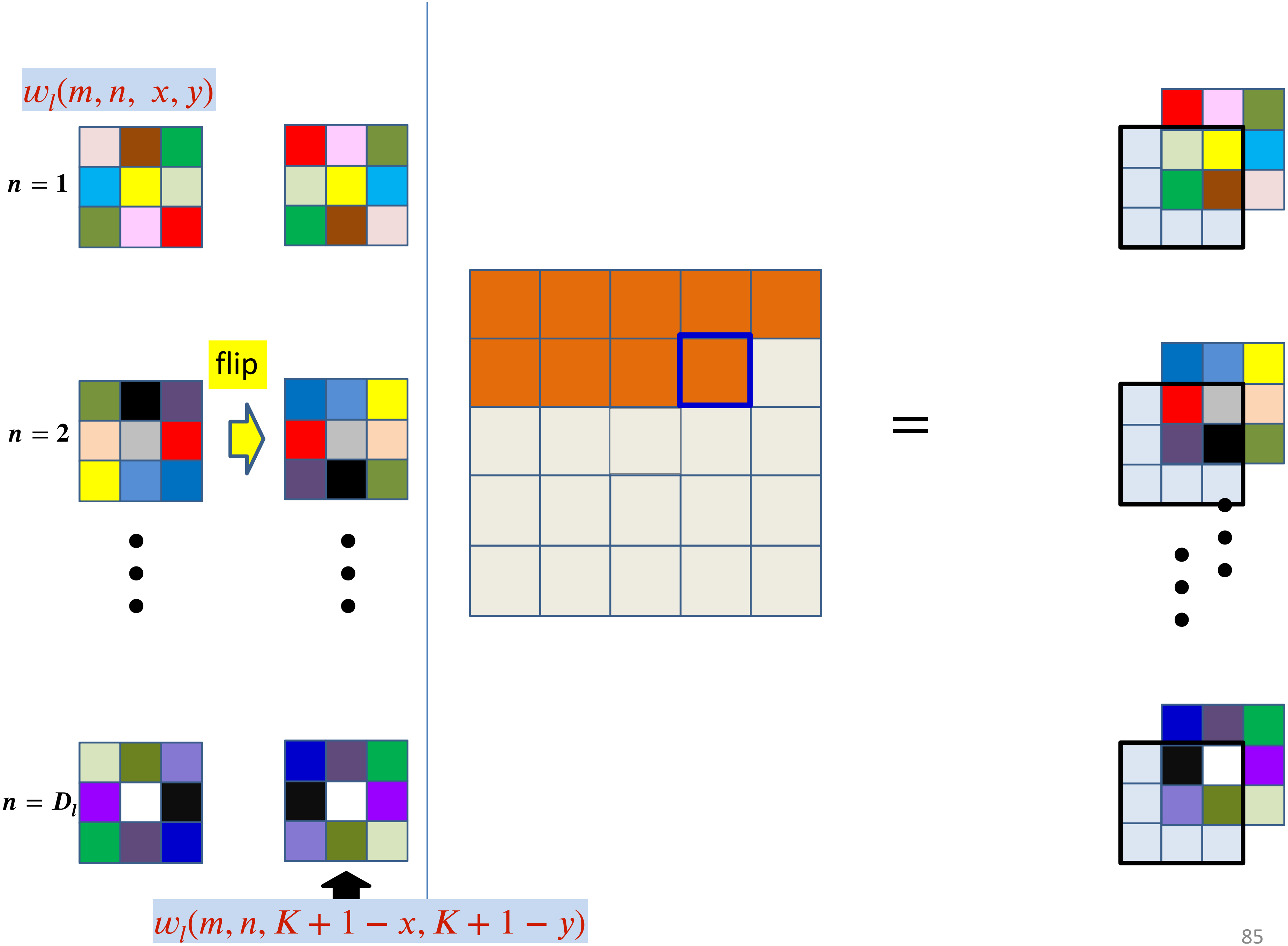


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$

$=$





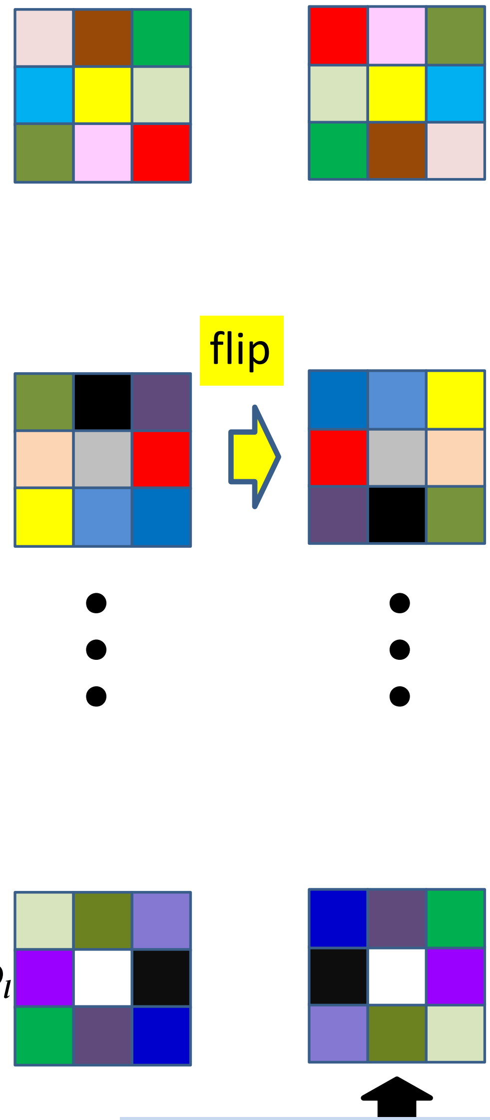
$w_l(m, n, x, y)$
$n = 1$
$n = 2$
flip
$n = D_l$
$w_l(m, n, K+1-x, K+1-y)$
=

$w_l(m, n,\ x, y)$

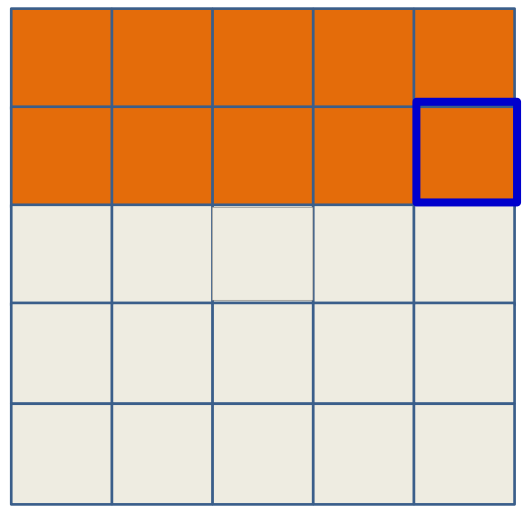
$\boldsymbol{n} = 1$

$\boldsymbol{n} = 2$

$\boldsymbol{n} = \boldsymbol{D_l}$

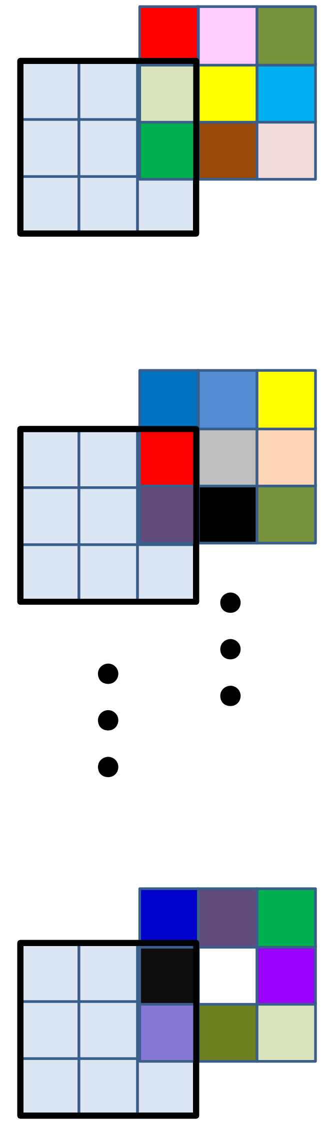


$w_l(m, n, K+1-x, K+1-y)$



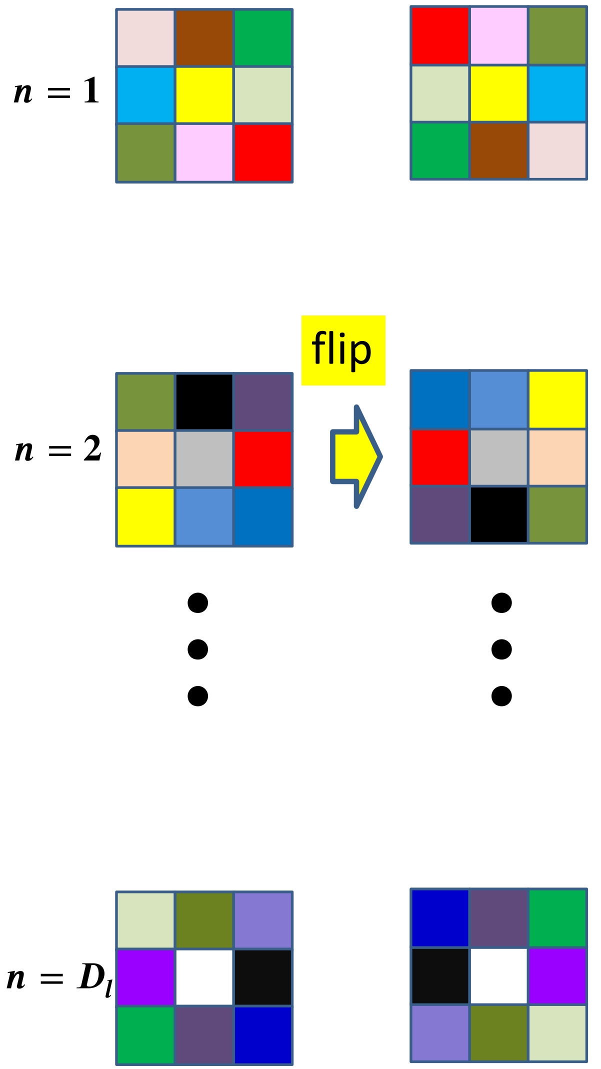
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$

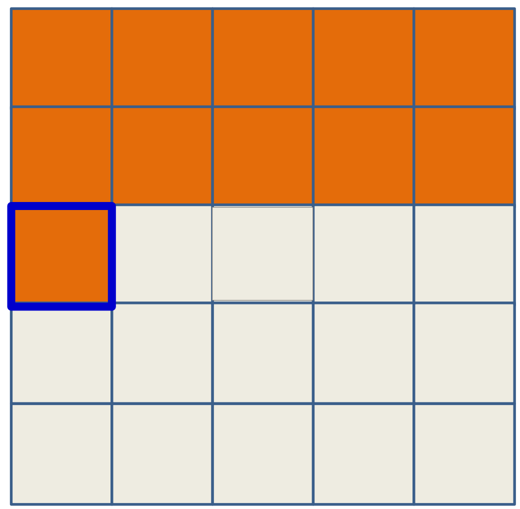


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$

$w_l(m, n, x, y)$

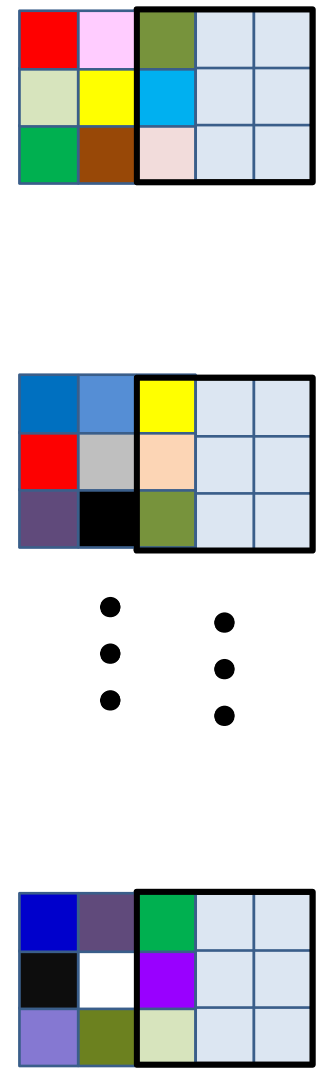


$w_l(m, n, K+1-x, K+1-y)$

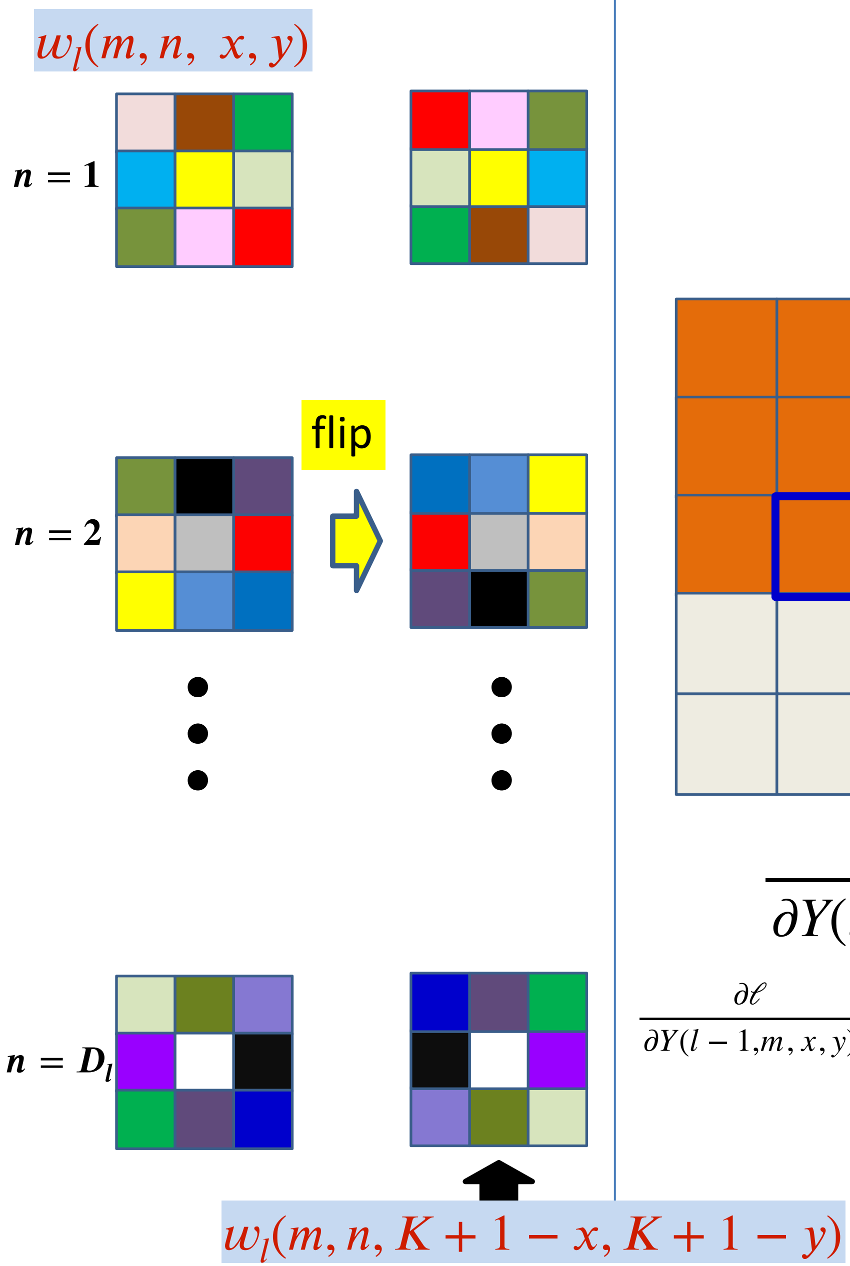


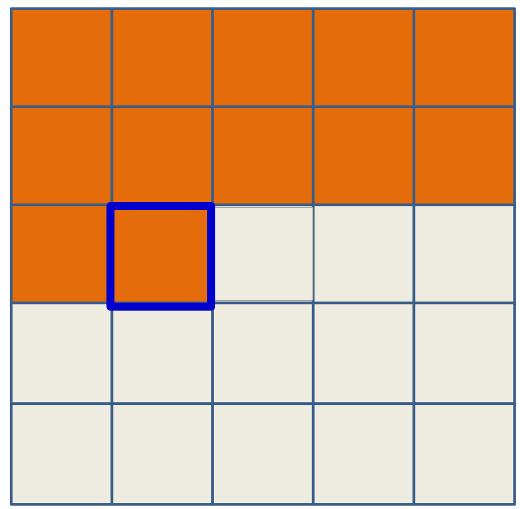
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

=

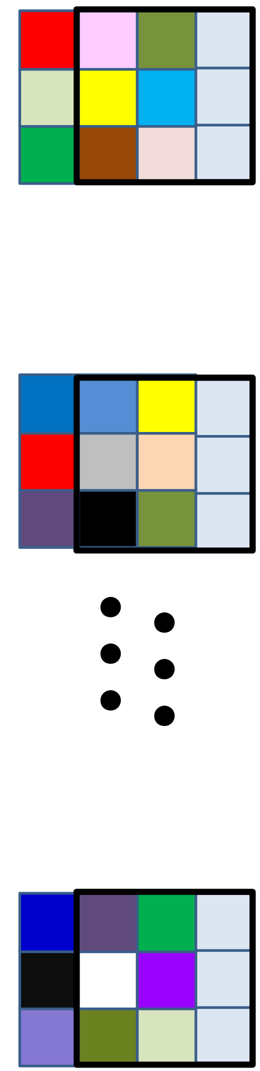


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_n \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$



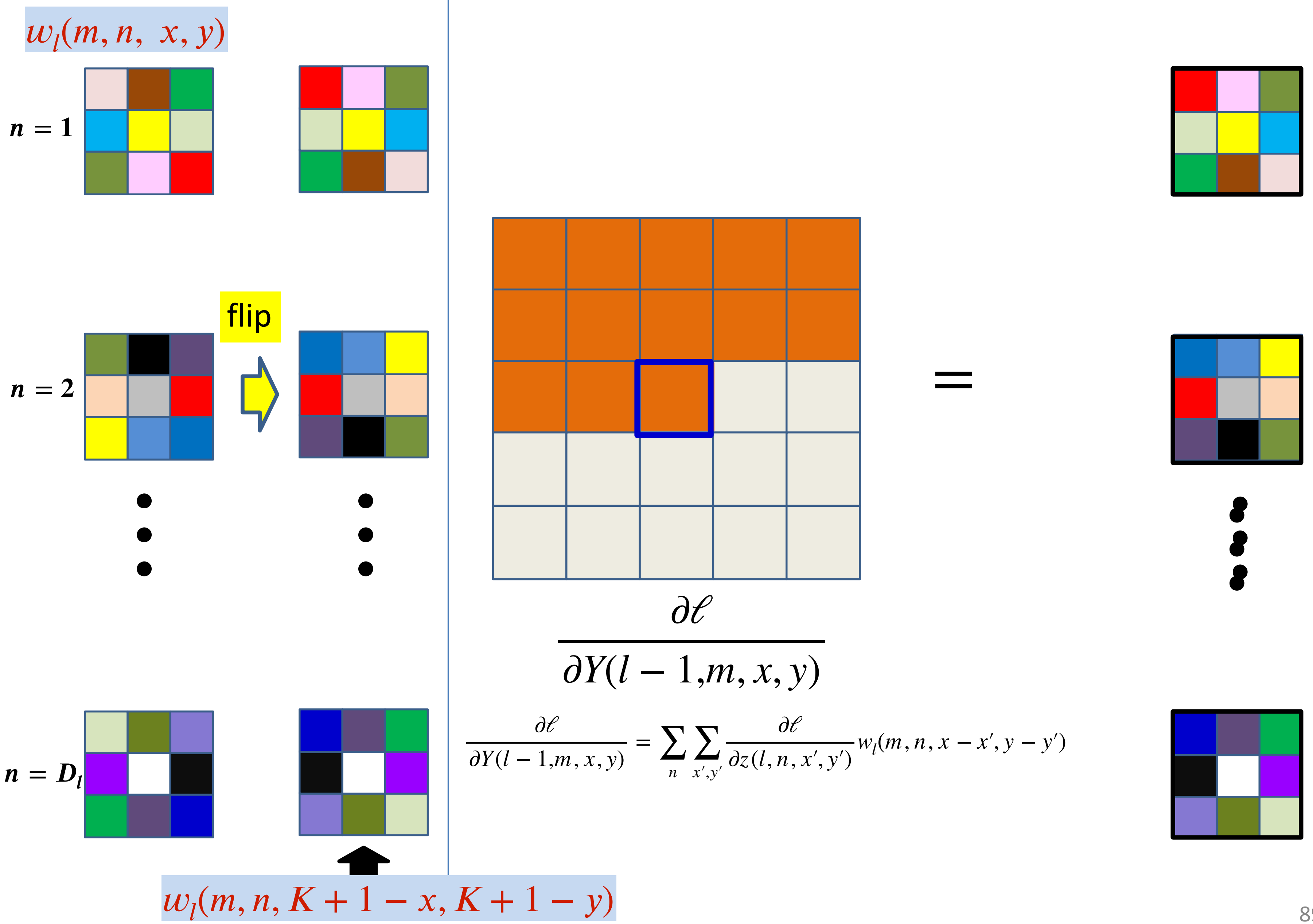


=

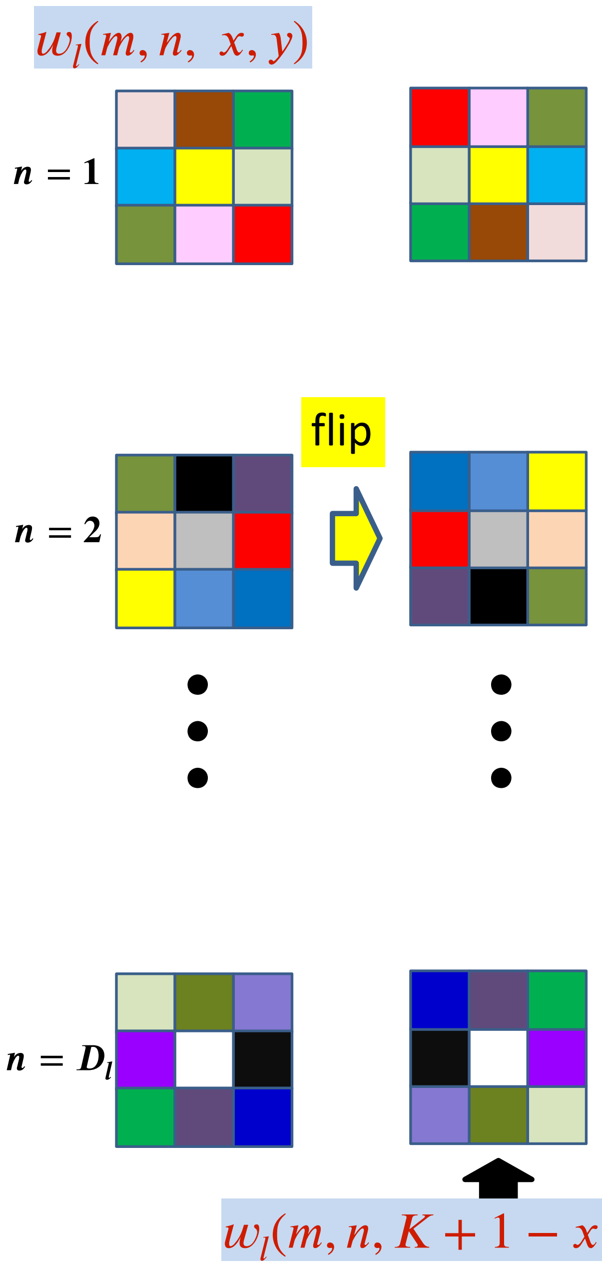


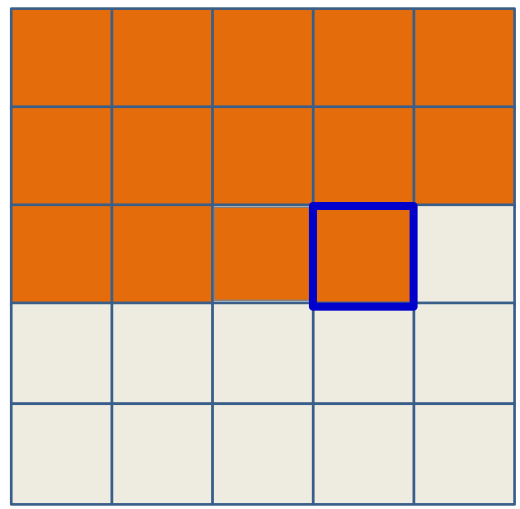
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$



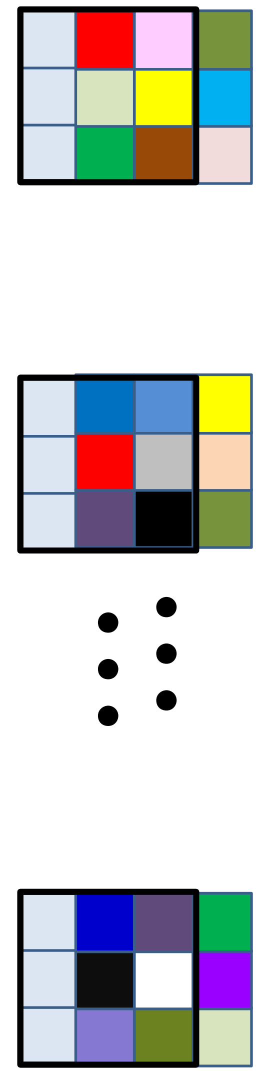
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$





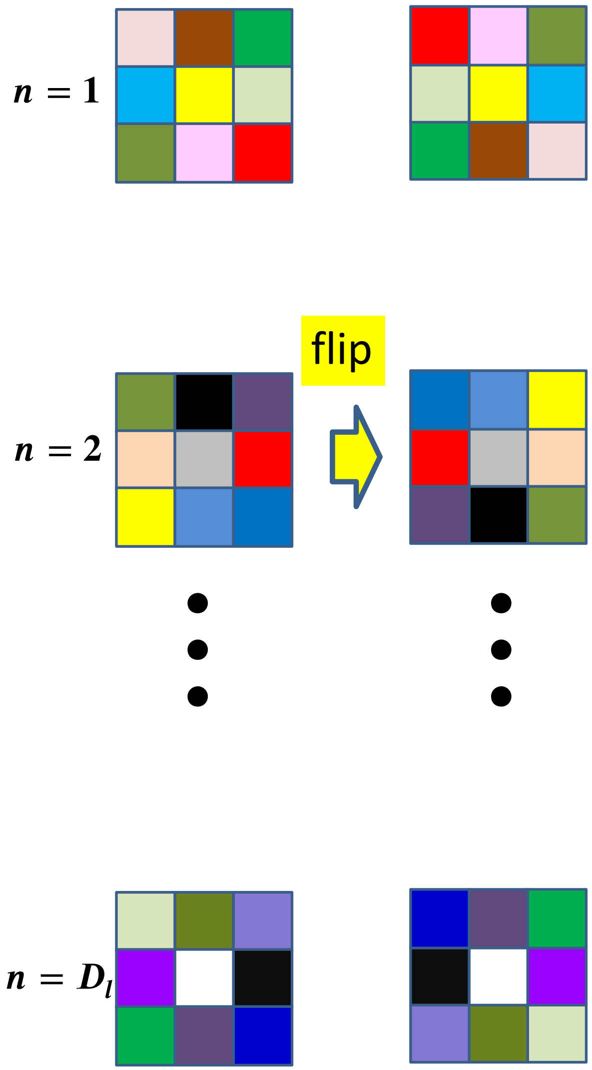
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$

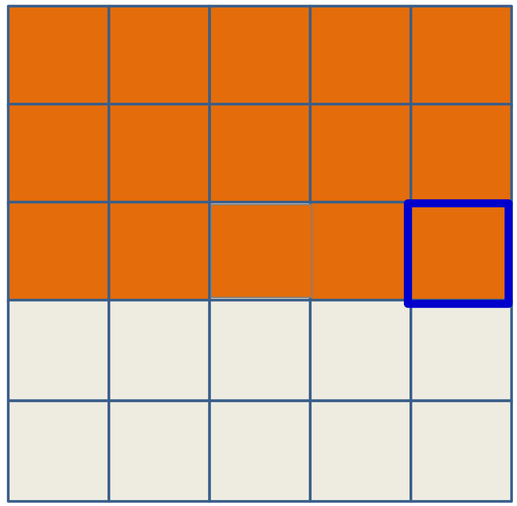


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_n \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$

$w_l(m, n, \ x, y)$

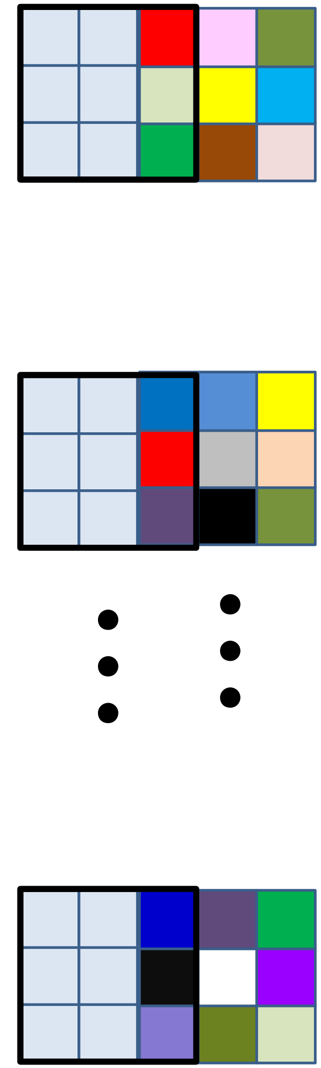


$w_l(m, n, K+1-x, K+1-y)$

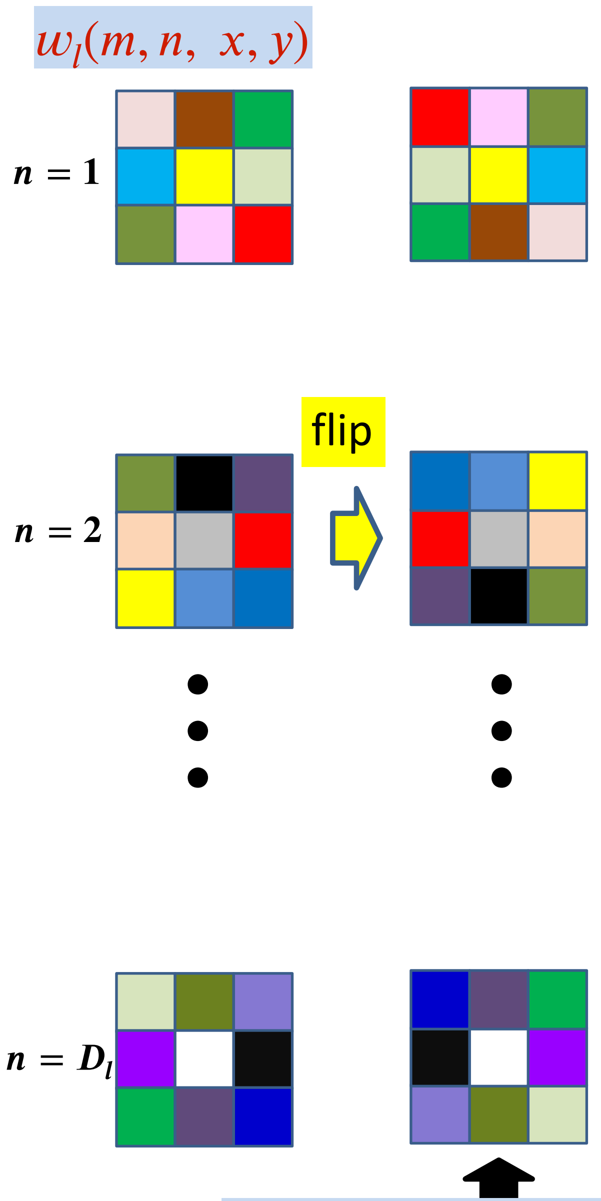


$=$

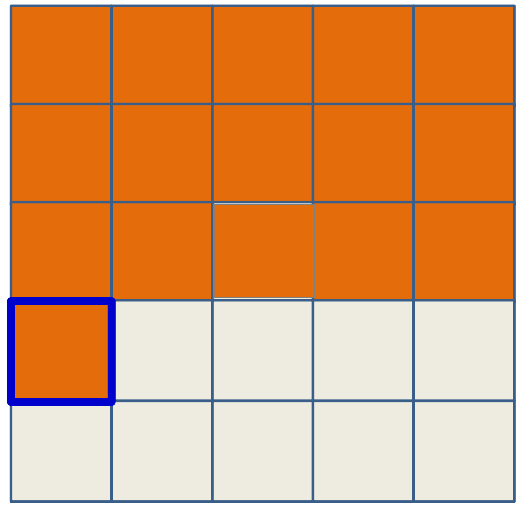
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$



$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_n \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$

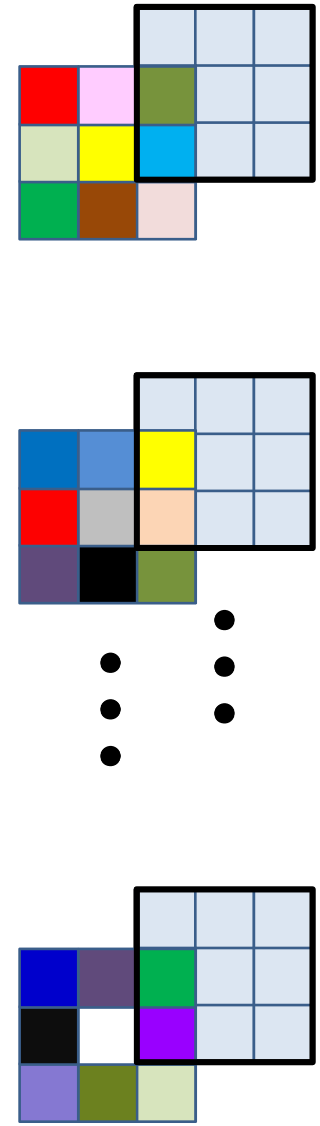


$w_l(m,n,K+1-x,K+1-y)$



$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)}$$

$=$



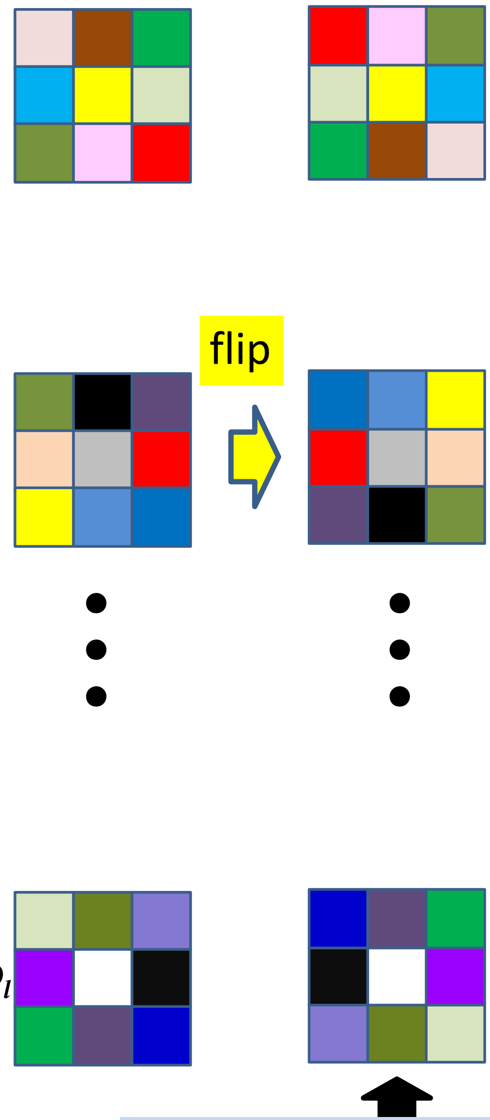
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_n \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

$w_l(m,n,\ x,y)$

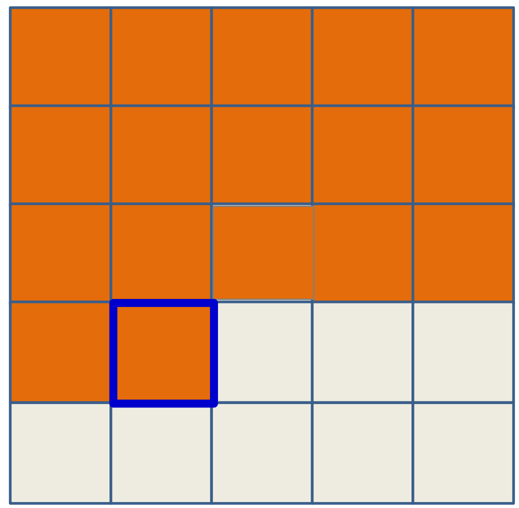
$\boldsymbol{n} = 1$

$\boldsymbol{n} = 2$

$\boldsymbol{n} = \boldsymbol{D_l}$

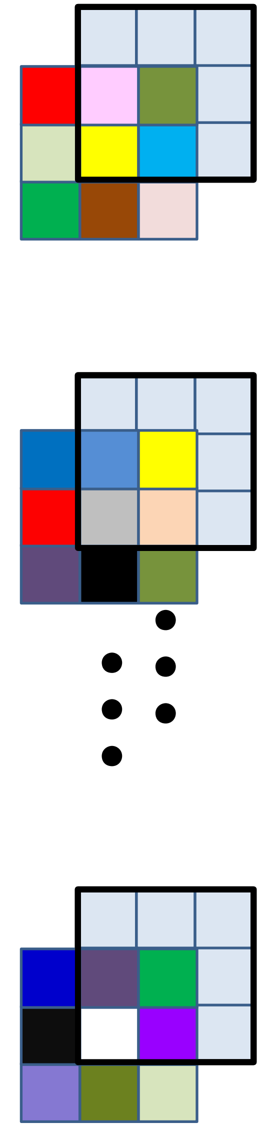


$w_l(m,n,K+1-x,K+1-y)$

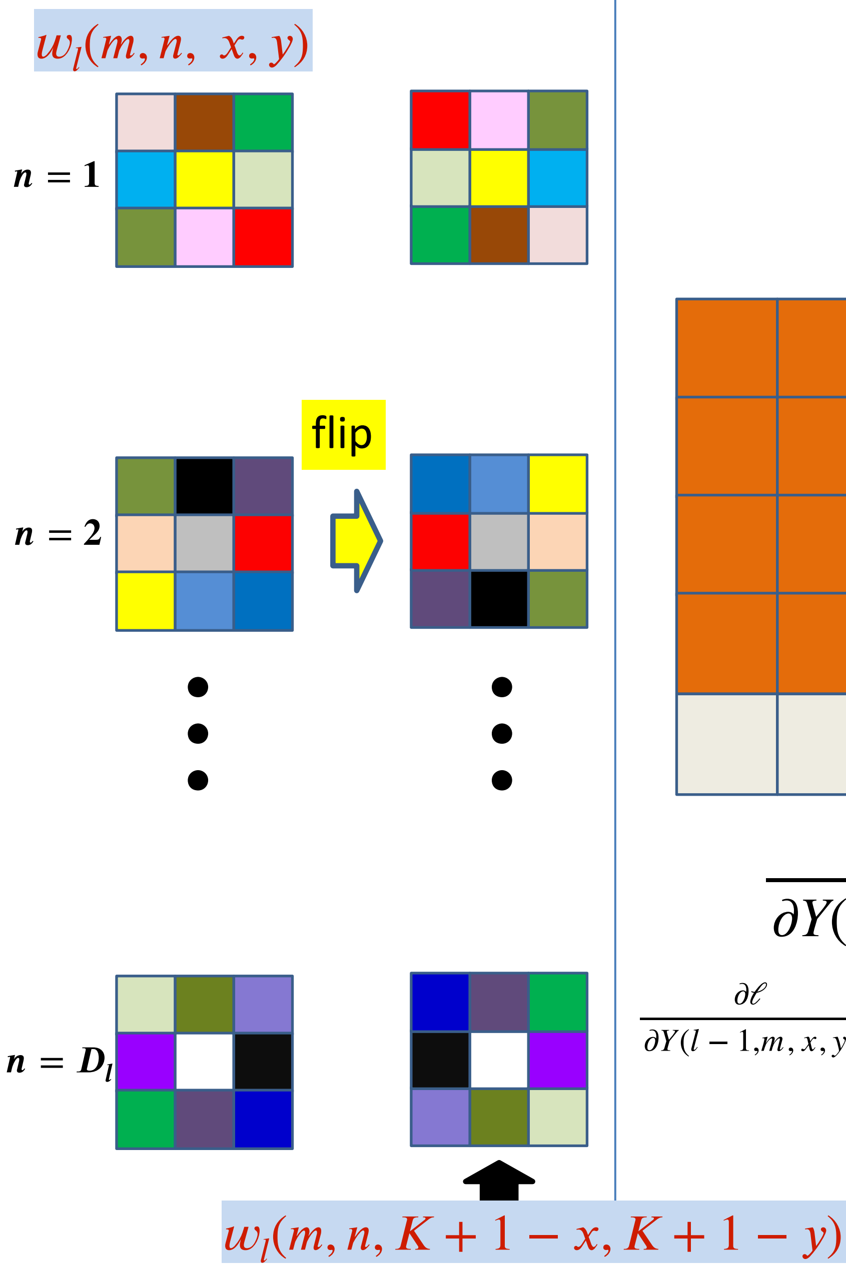


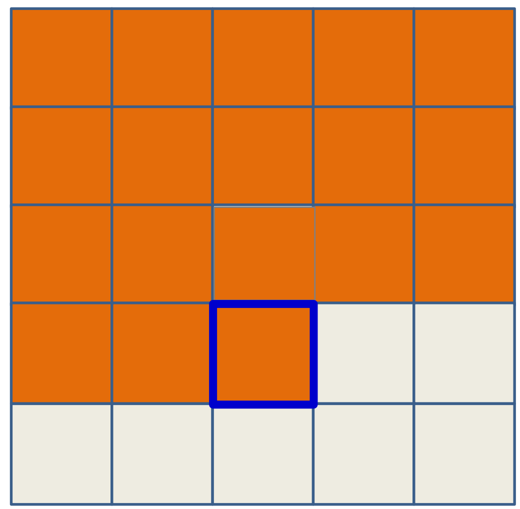
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)}$$

$=$



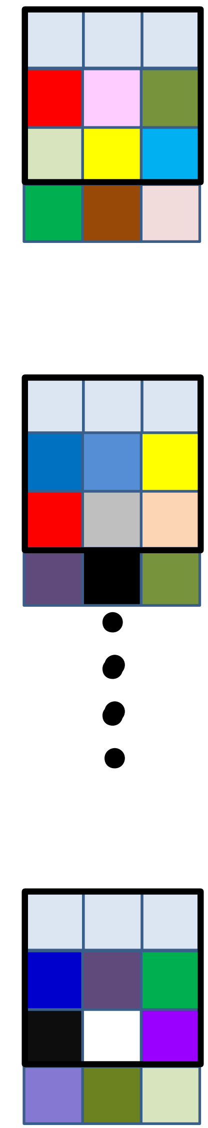
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_n \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$





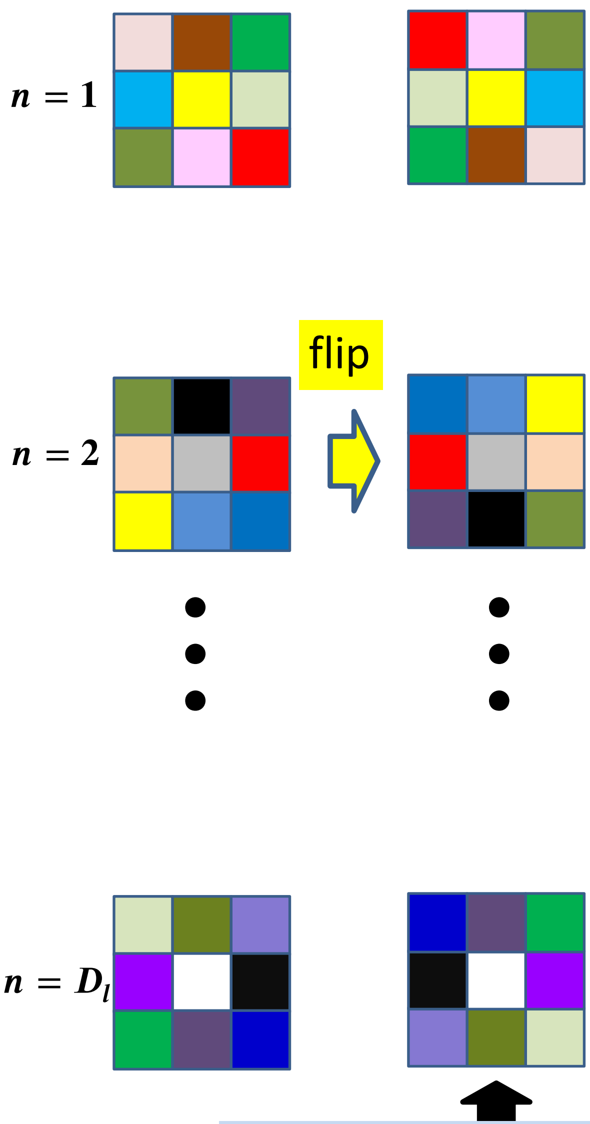
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$

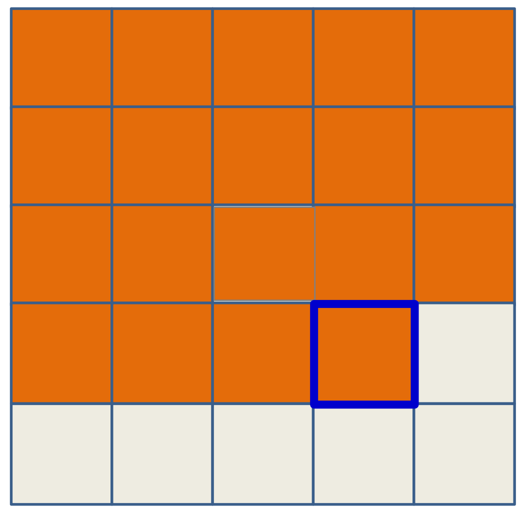


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$

$w_l(m, n, x, y)$

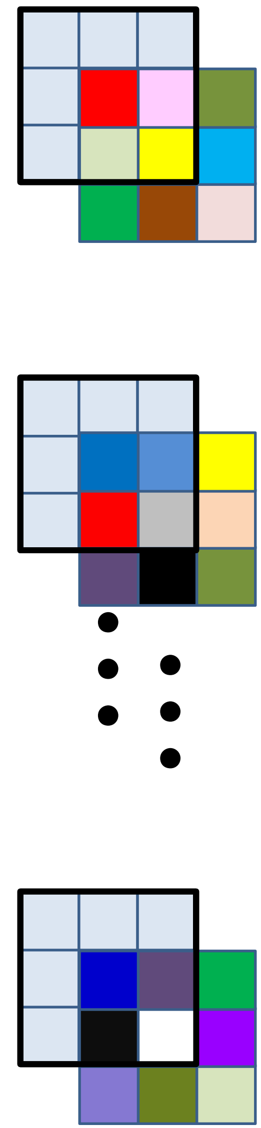


$w_l(m, n, K+1-x, K+1-y)$



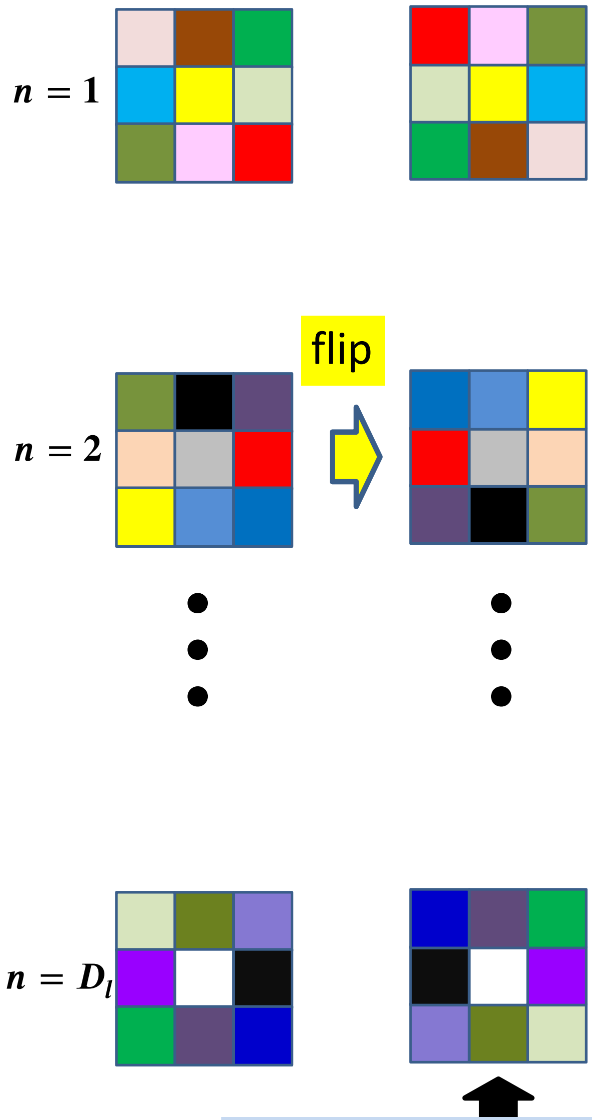
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

=

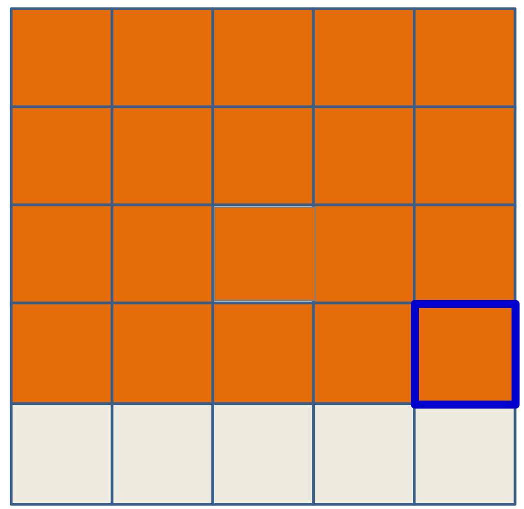


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$

$w_l(m,n,\ x,y)$

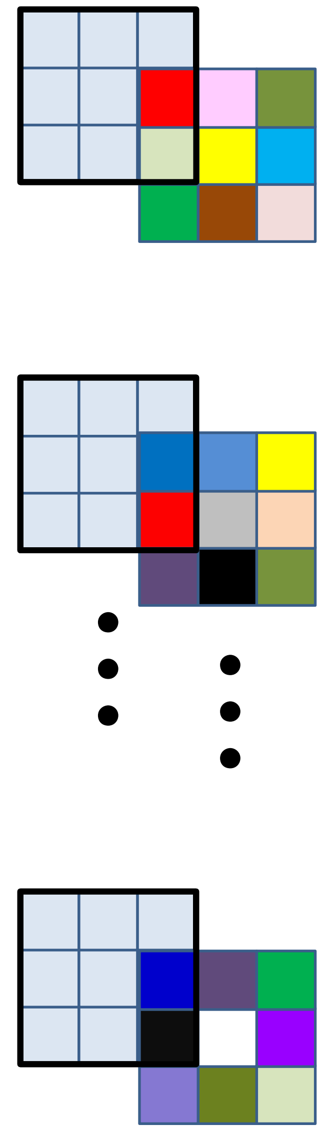


$w_l(m,n,K+1-x,K+1-y)$



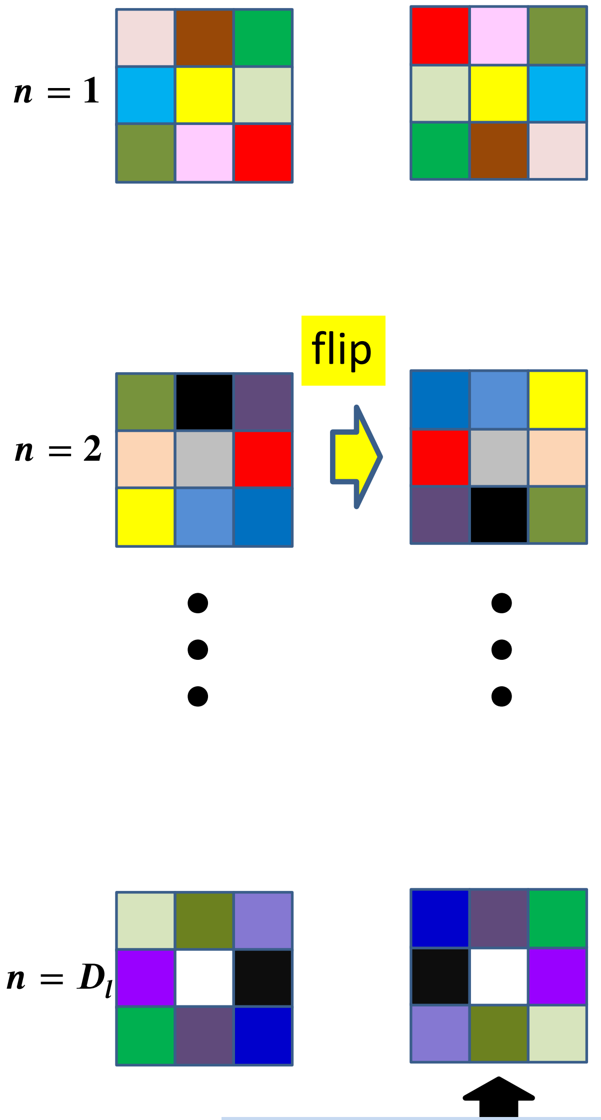
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)}$$

$=$

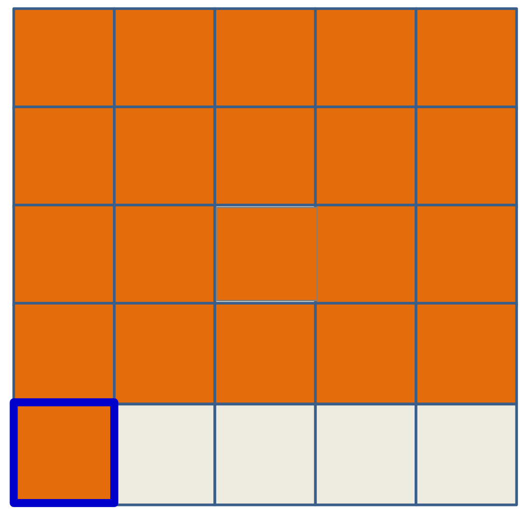


$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_{n} \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

$w_l(m, n, x, y)$

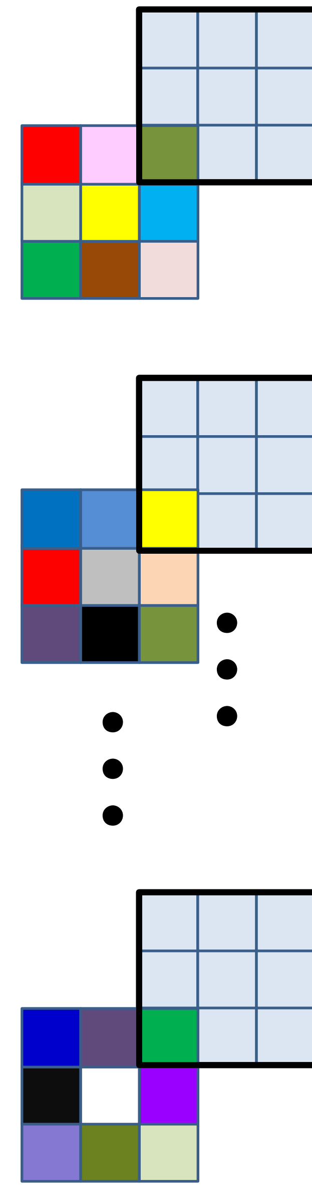


$w_l(m, n, K+1-x, K+1-y)$

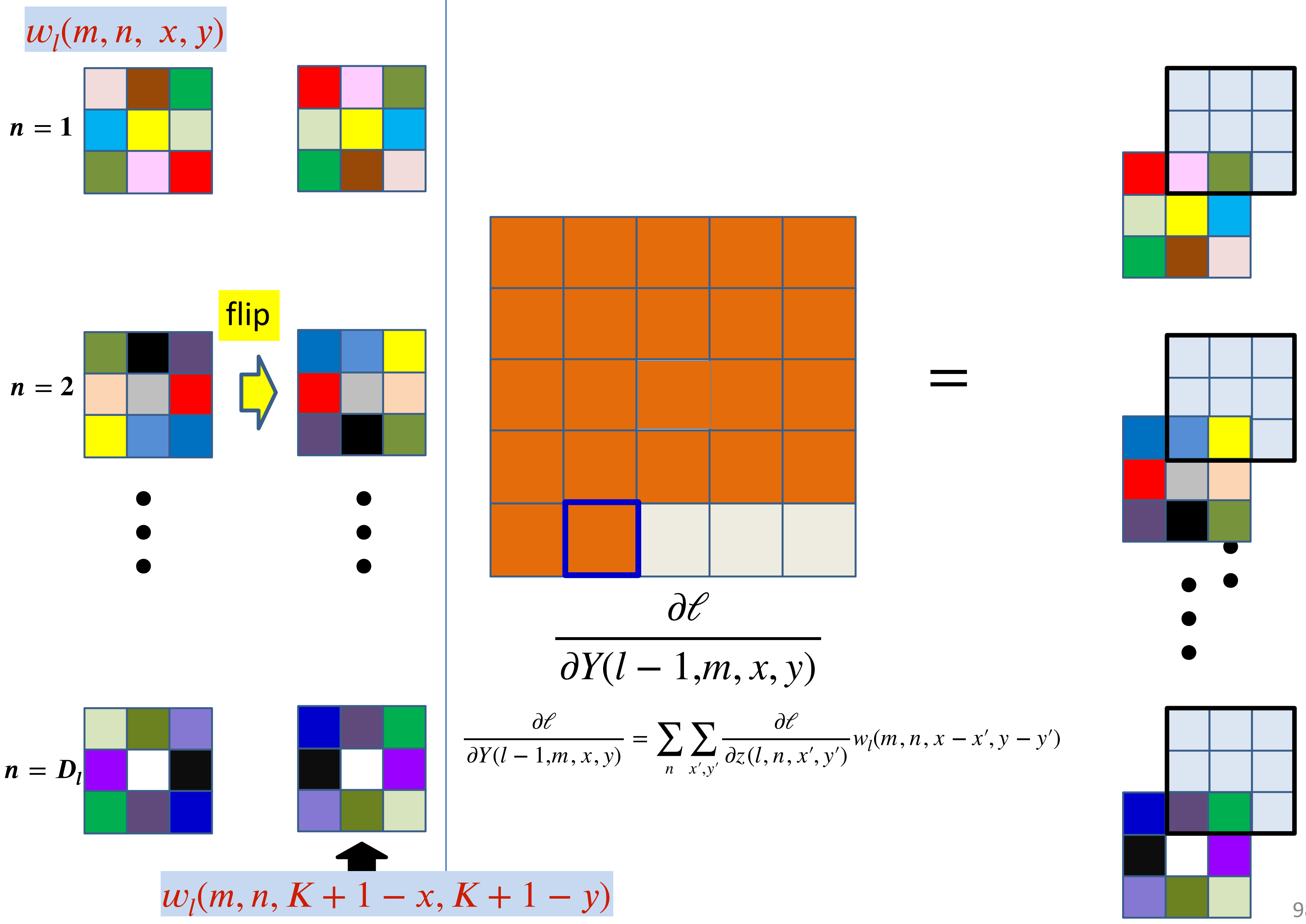


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

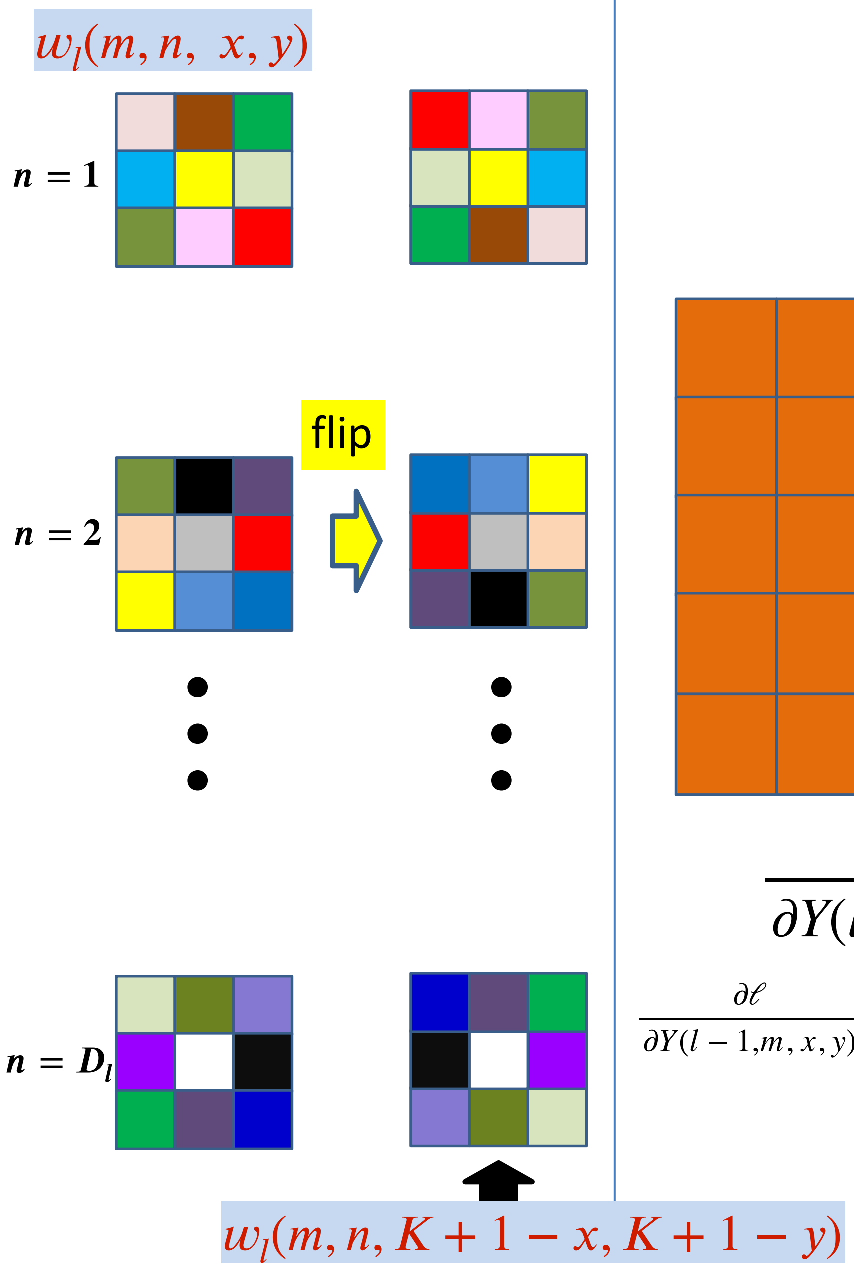
=

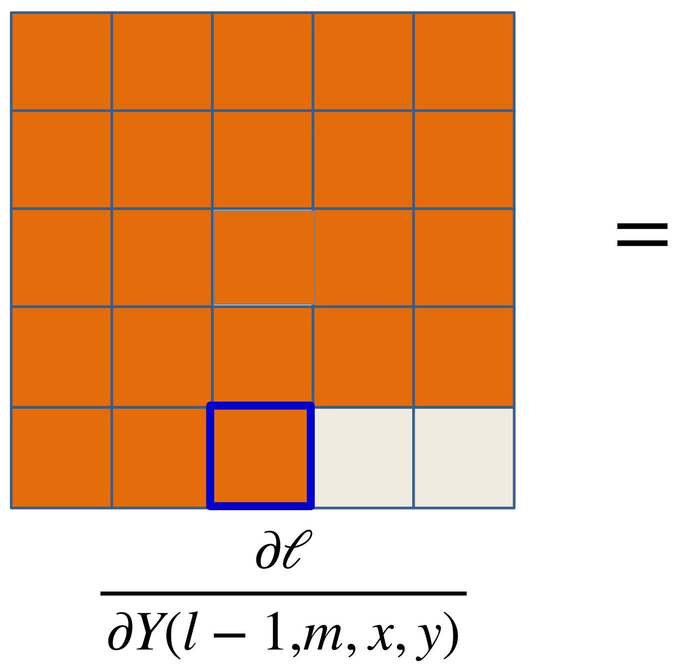


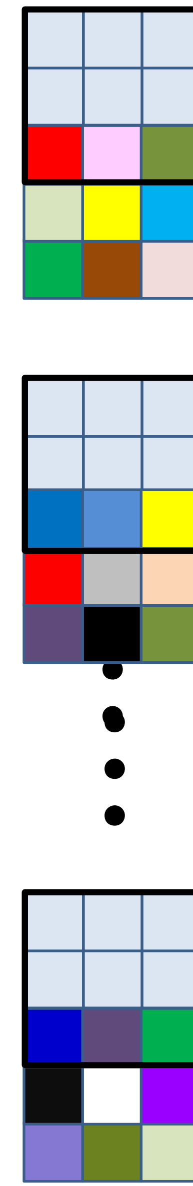
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_n \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$



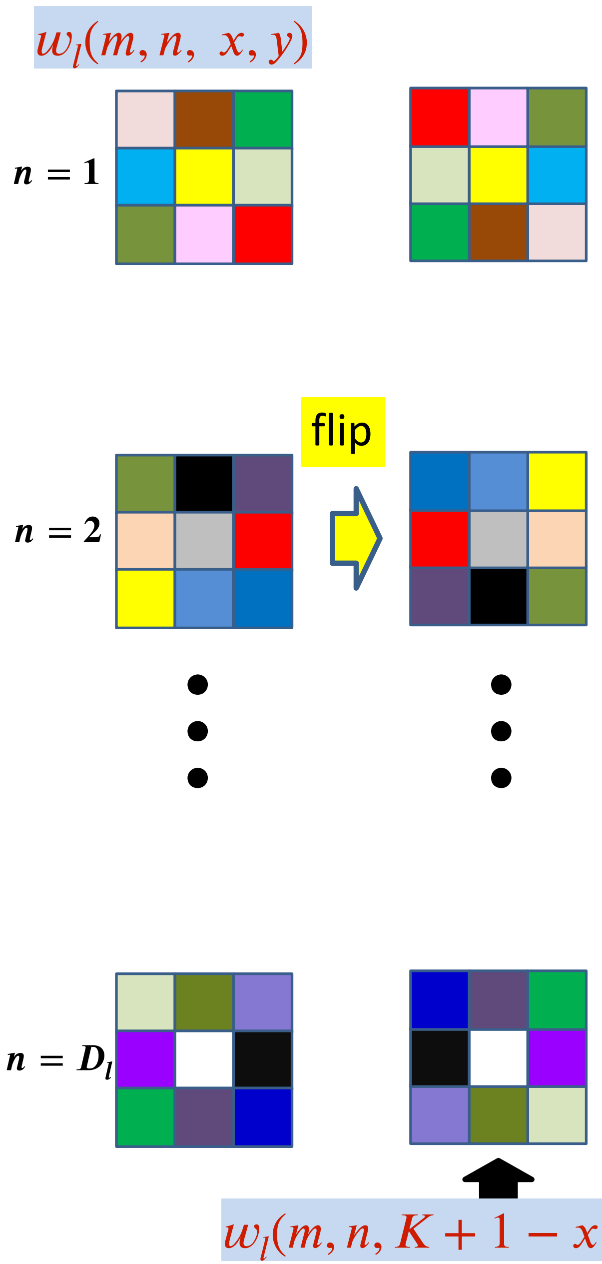
$$\frac{\partial \ell}{\partial Y(l-1,m,x,y)} = \sum_n \sum_{x',y'} \frac{\partial \ell}{\partial z(l,n,x',y')} w_l(m,n,x-x',y-y')$$

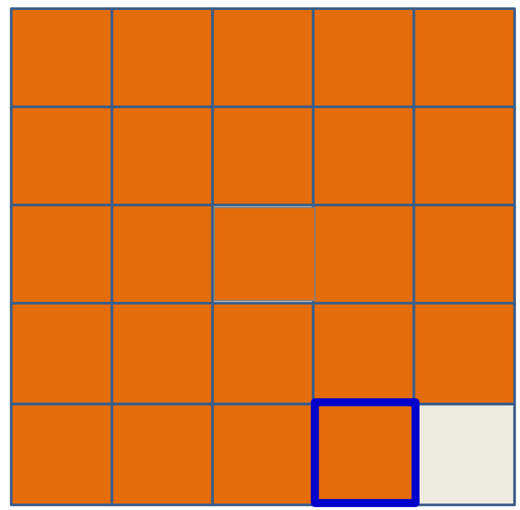






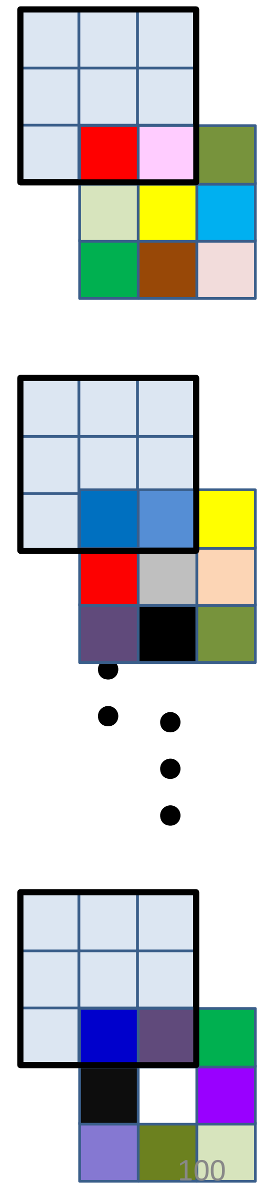
$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$$



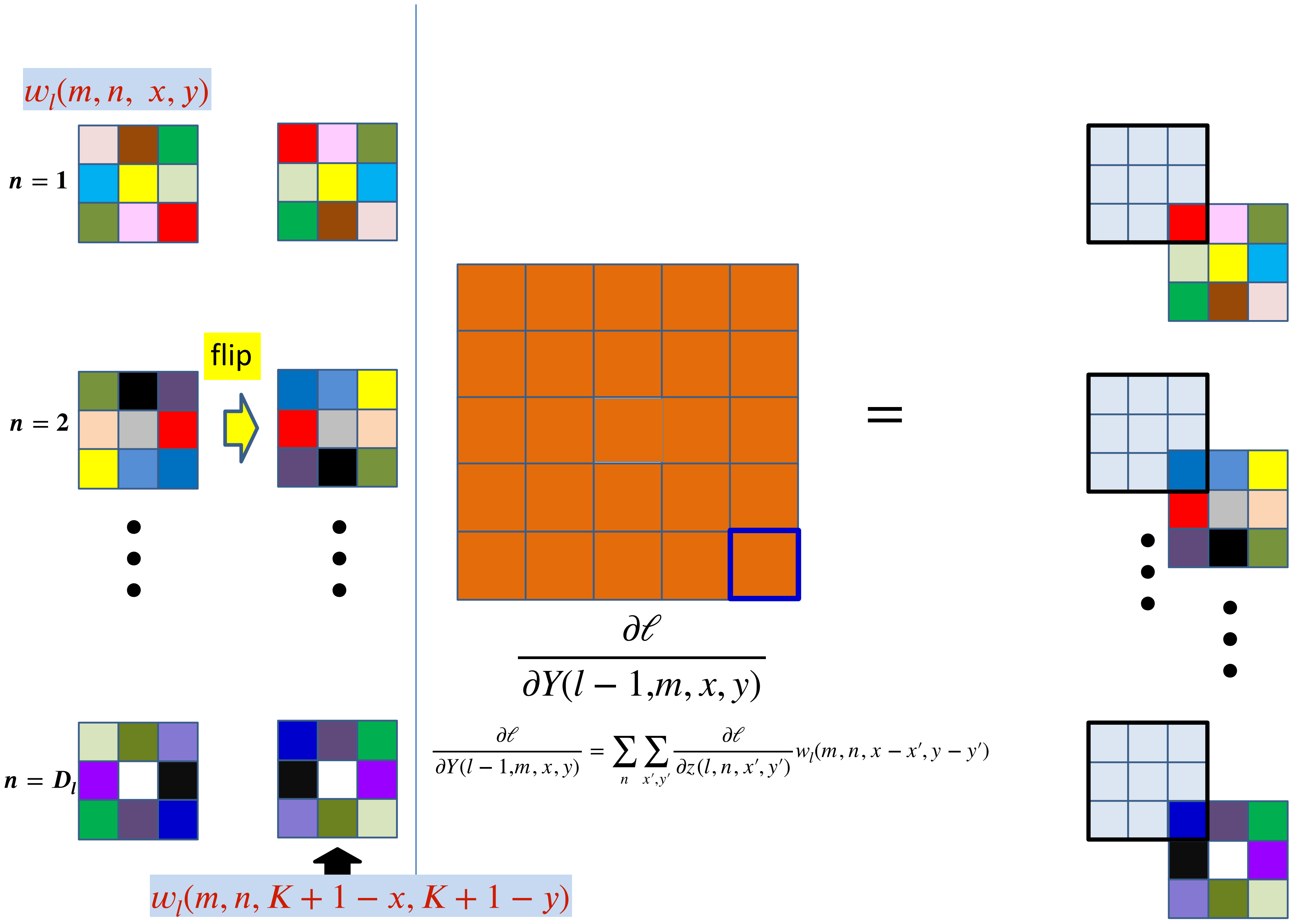


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$$

$=$

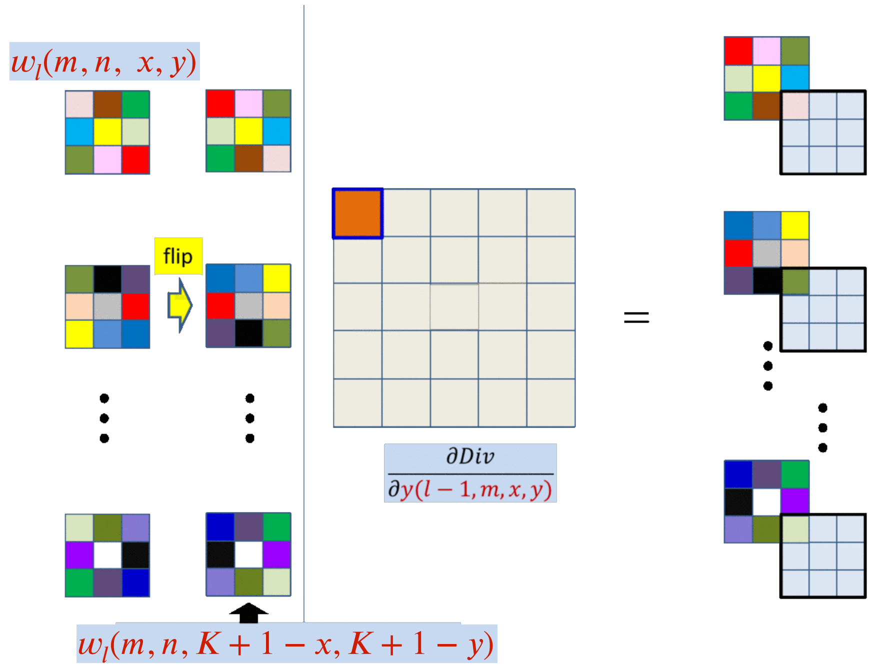


$$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x-x', y-y')$$



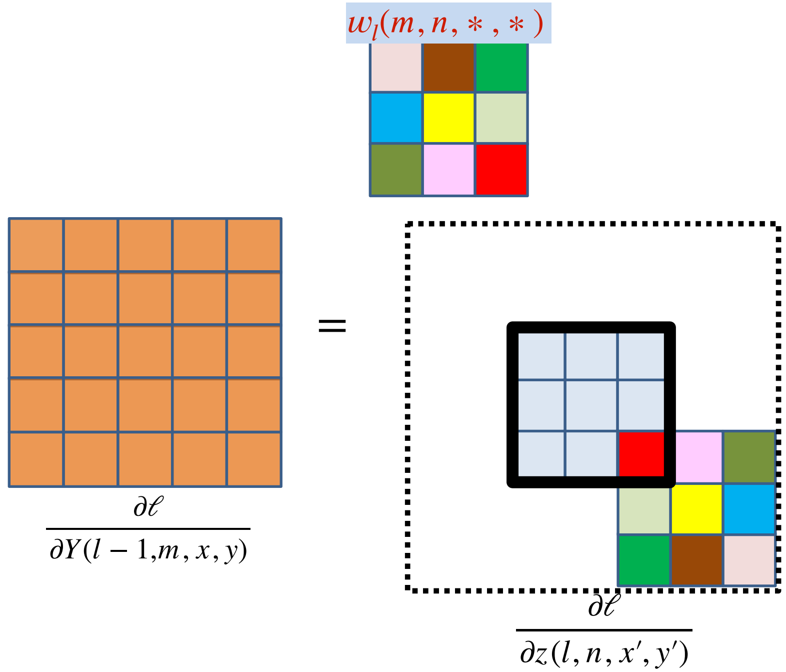
$w_l(m, n, x, y)$
$n = 1$
flip
$n = 2$
$n = D_l$
$w_l(m, n, K+1-x, K+1-y)$
$\frac{\partial \ell}{\partial Y(l-1, m, x, y)}$
=
$\frac{\partial \ell}{\partial Y(l-1, m, x, y)} = \sum_{n} \sum_{x', y'} \frac{\partial \ell}{\partial z(l, n, x', y')} w_l(m, n, x - x', y - y')$

Computing the derivative for $Y(l-1, m)$



- This is just a convolution of the zero-padded $\dfrac{\partial \ell}{\partial z(l,n,x',y')}$ maps by the transposed and flipped filter
  - After zero padding it first with $K-1$ zeros on every side

The size of the Y-derivative map

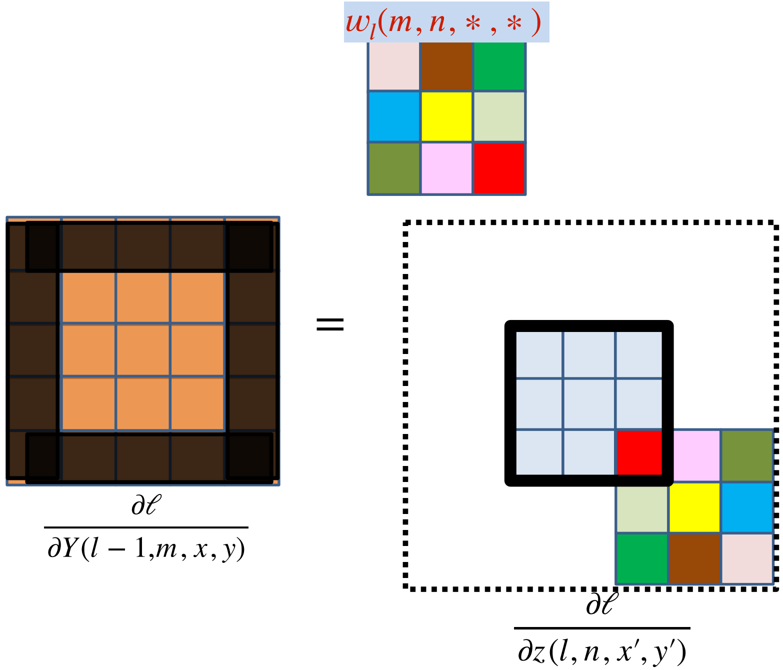


- We continue to compute elements for the derivative Y map as long as the (flipped) filter has at least one element in the (unpadded) derivative Zmap
  - I.e. so long as the Y derivative is non-zero

- The size of the Y derivative map will be $(H + K - 1) \times (W + K - 1)$
  - H and W are heidght and width of the Zmap

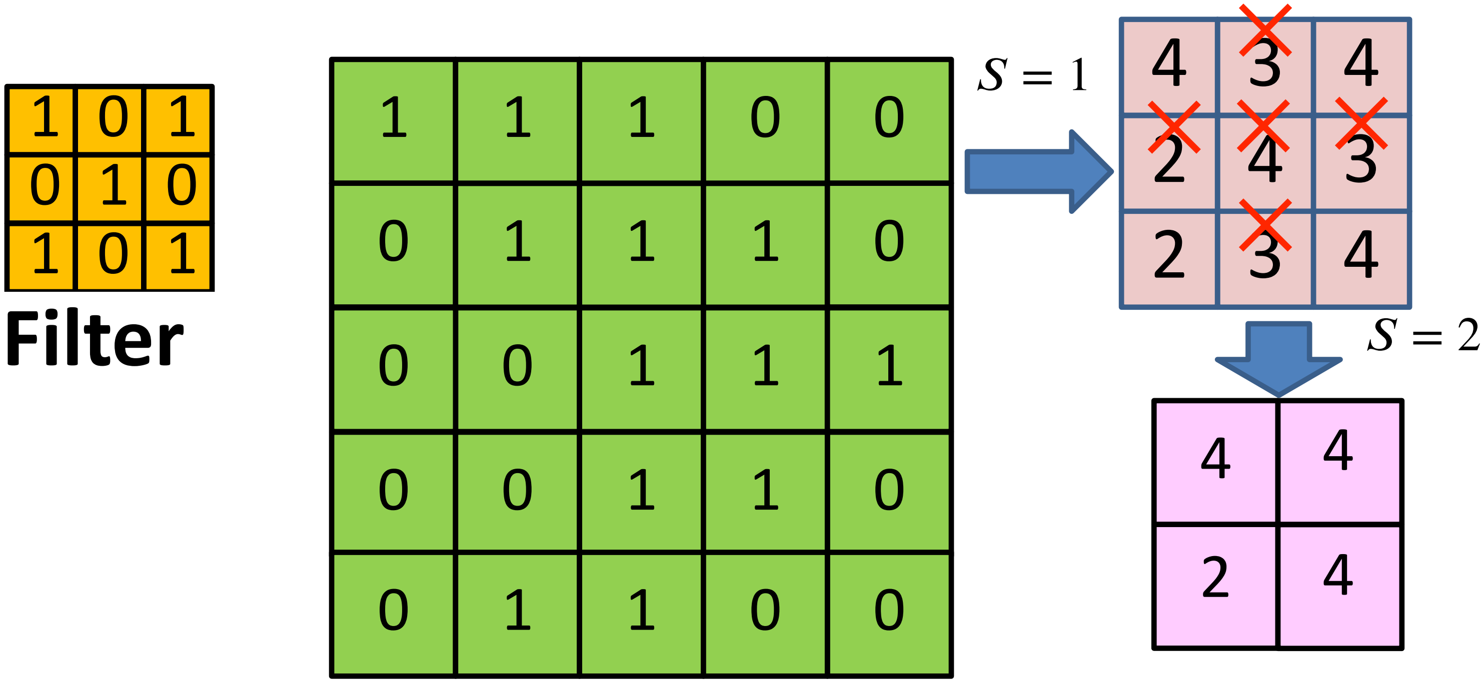
- This will be the size of the actual Y map that was originally convolved

The size of the Y-derivative map



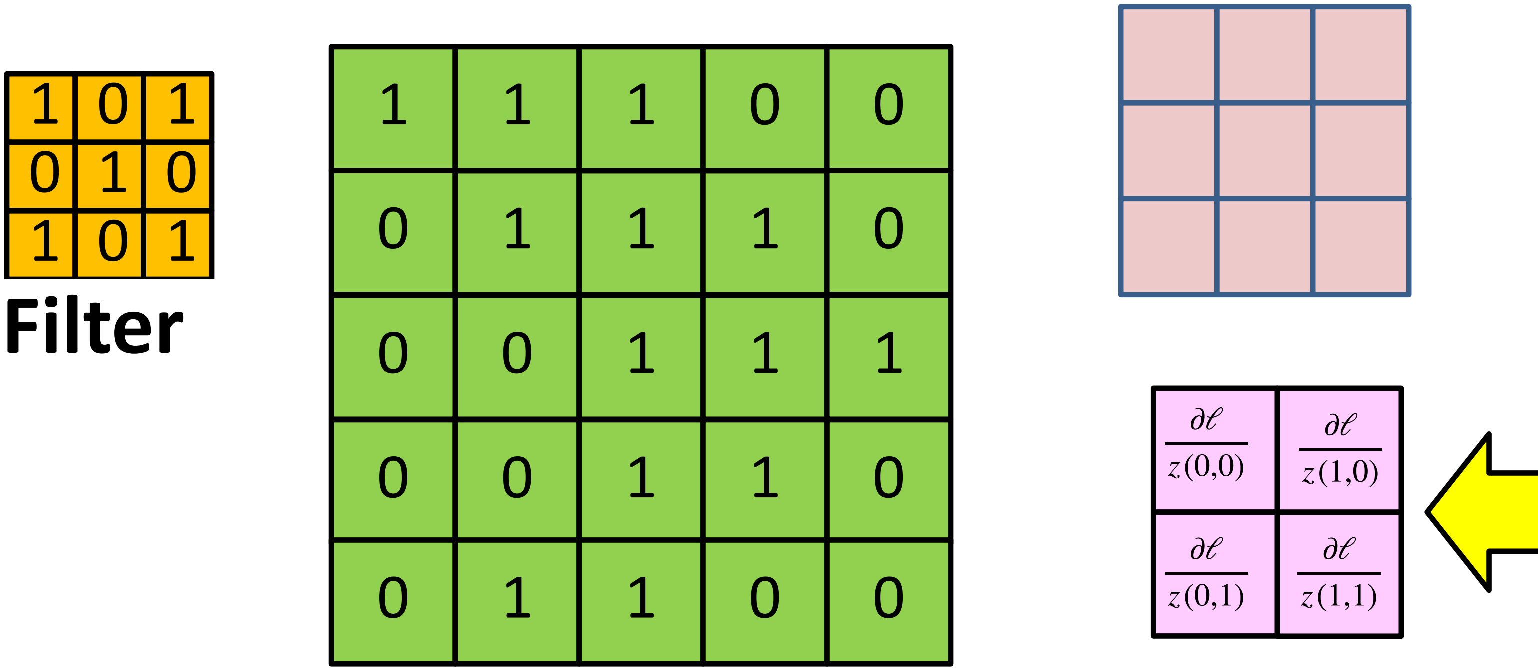
- If the Y map was zero-padded in the forward pass, the derivative map will be the size of the *zero-padded* map
  - The zero padding regions must be deleted before further backprop

Stride greater than 1



- **Observation:** Convolving with a stride S greater than 1 is the same as convolving with stride 1 and “dropping” $S-1$ out of every S rows, and $S-1$ of every S columns
  - **Downsampling by S**
  - E.g. for stride 2, it is the same as convolving with stride 1 and dropping every 2nd entry

Derivatives with Stride greater than 1



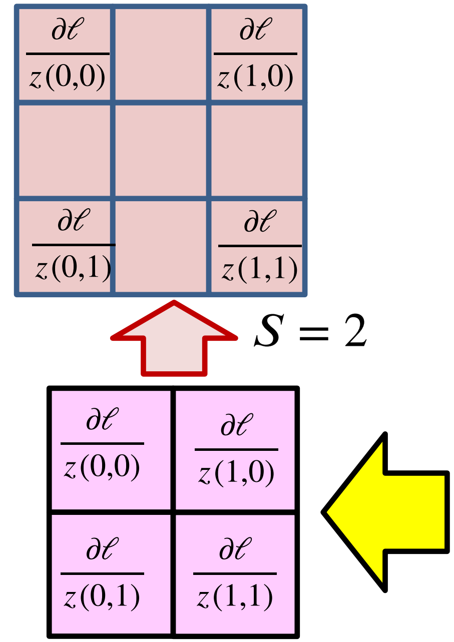
- **Derivatives:** Backprop gives us the derivatives of the divergence with respect to the elements of the *downsampled* (strided) Z map

Derivatives with Stride greater than 1

1	0	1
0	1	0
1	0	1

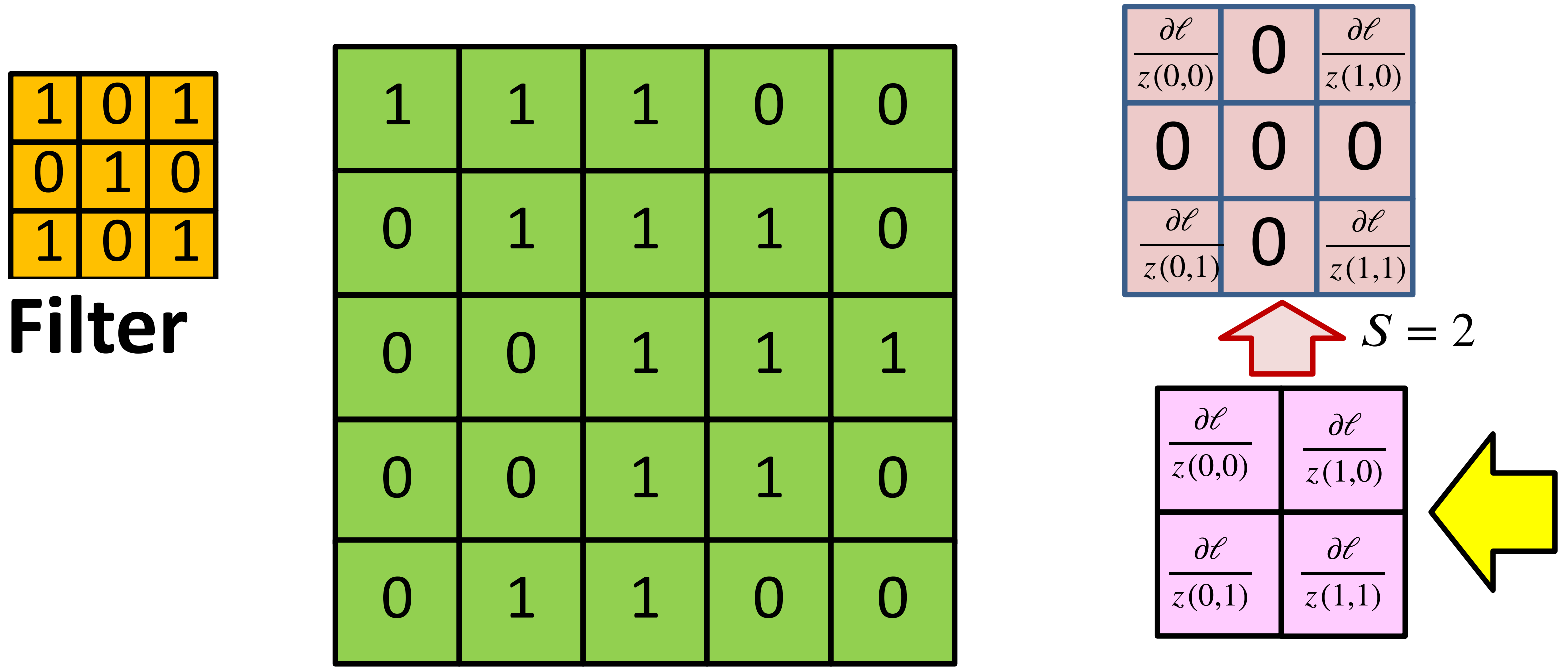
Filter

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



- **Derivatives:** Backprop gives us the derivatives of the divergence with respect to the elements of the *downsampled* (strided) Z map
- We can place these derivative values back into their original locations of the full-sized Z map

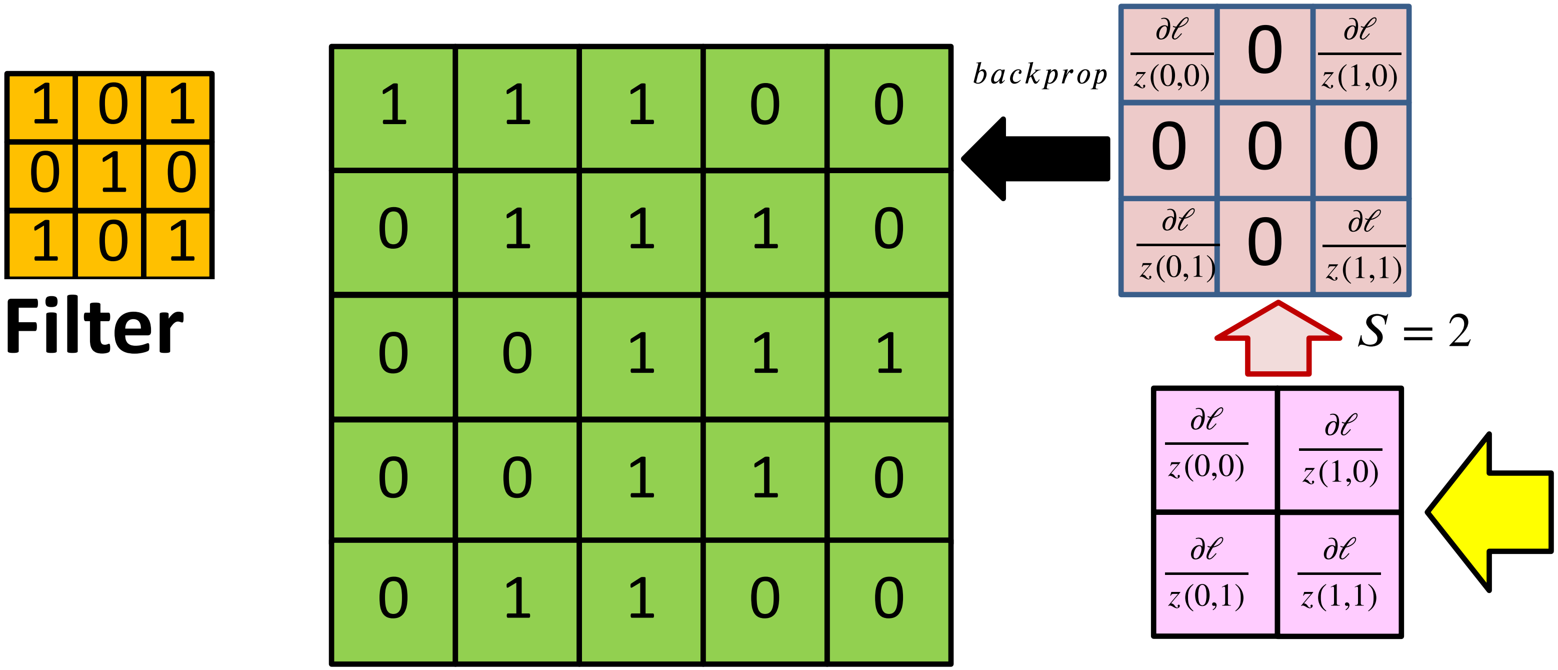
Derivatives with Stride greater than 1



- **Derivatives:** Backprop gives us the derivatives of the divergence with respect to the elements of the *downsampled* (strided) Z map
- We can place these values back into their original locations of the full-sized Z map

- The remaining entries of the Z map do not affect the divergence
  - Since they get dropped out
- The derivative of the divergence w.r.t. these values is 0

Computing derivatives with Stride > 1

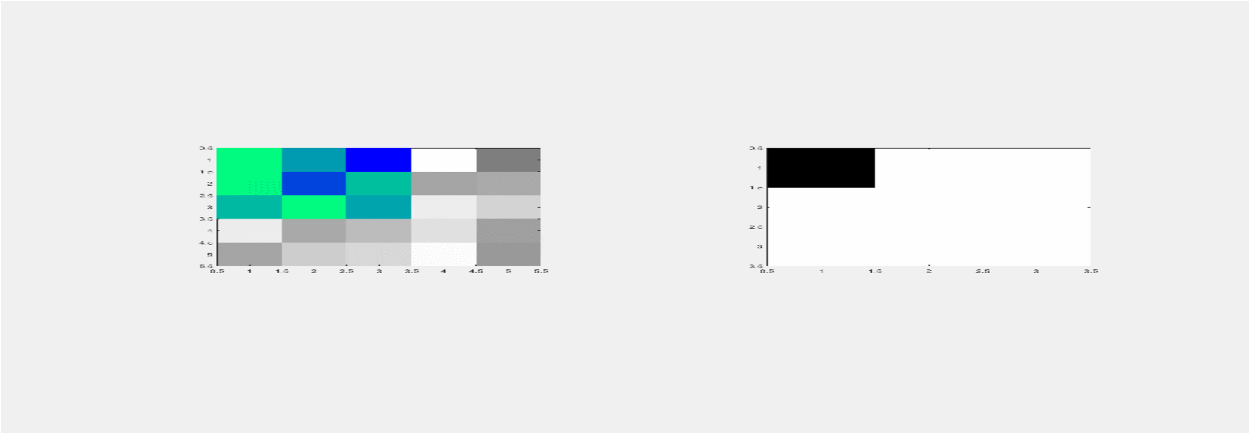


- **Upsampling derivative map:**
  - Upsample the downsampled derivatives
  - Insert zeros into the "empty" slots
  - This gives us the derivatives w.r.t. all the entries of a full-sized (stride 1) Z map
- We can compute the derivatives for Y, using the full map

Computing $\dfrac{\partial \ell}{\partial w_l(m,n,x,y)}$

The derivatives for the weights

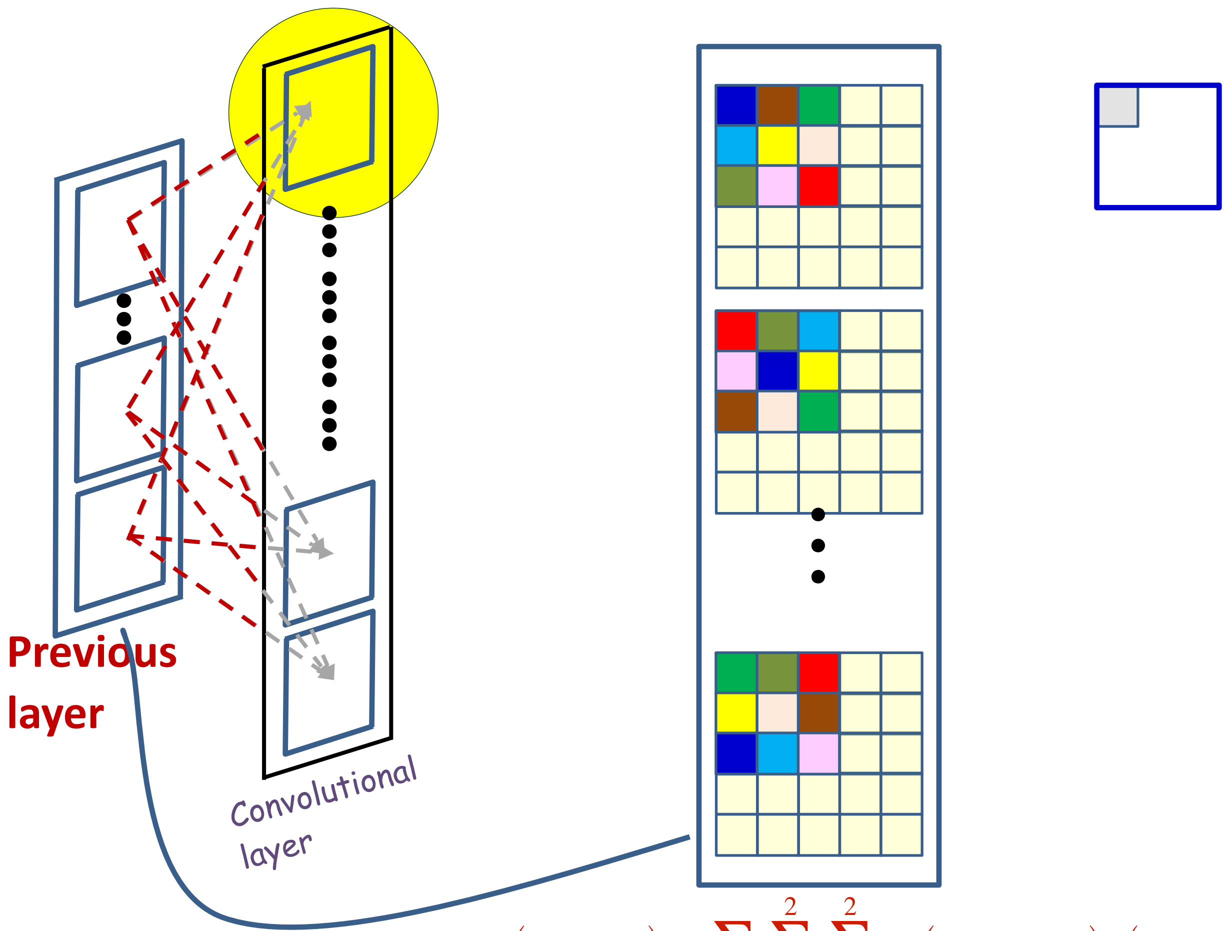
$Y(l-1,m) \otimes w_l(m,n)$ $Z(l,n)$



$$z(l,n,x,y) = \sum_m \sum_{x',y'} w_l(m,n,x',y')\, y(l-1,m,x+x',y+y') + b_l(n)$$

- Each ***weight*** $w_l(m,n,x',y')$ affects several $z(l,n,x,y)$
  – Consider the contribution of one filter components: $w_l(m,n,i,j)$ (e.g. $w_l(m,n,1,2)$)

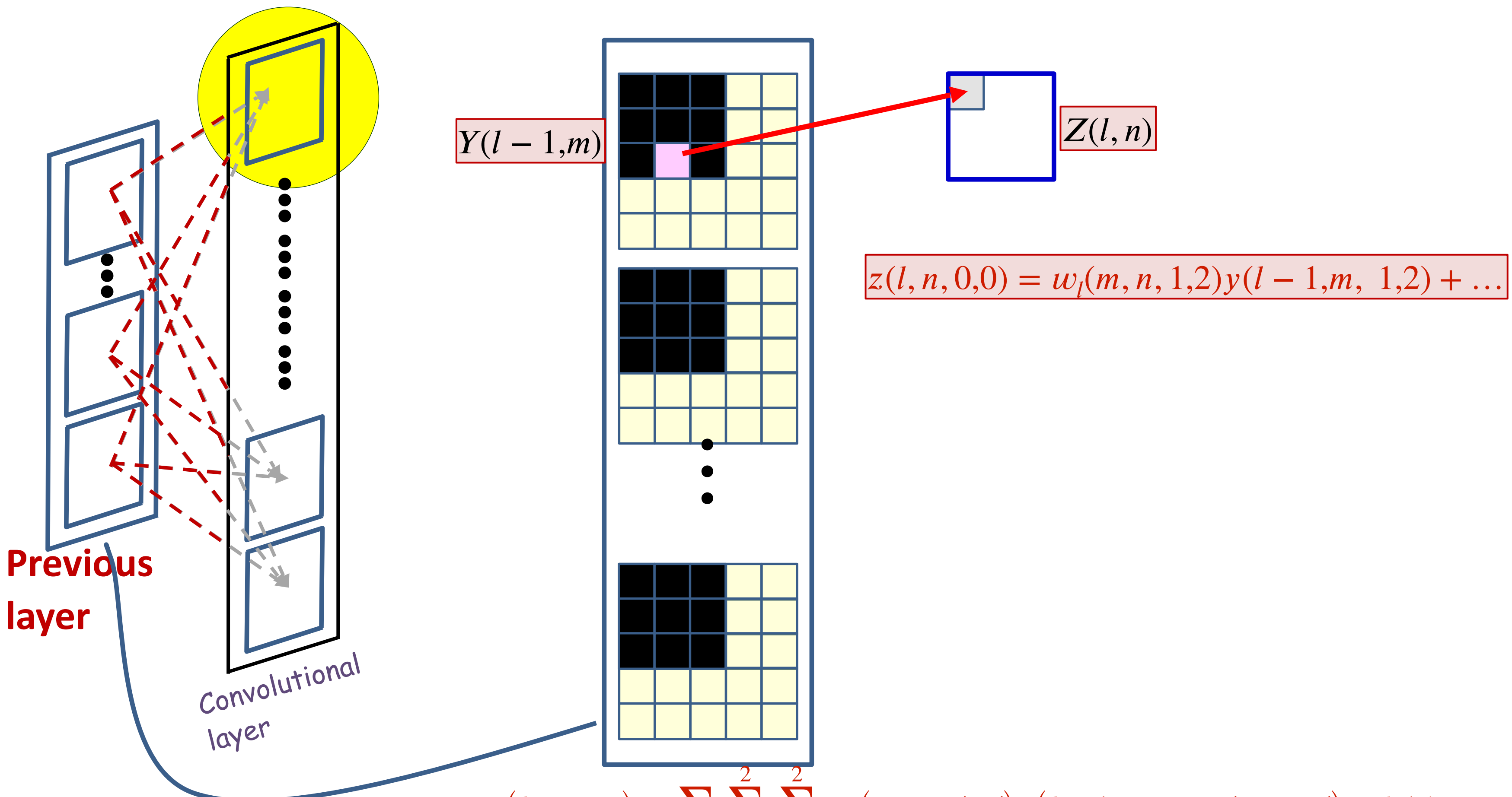
Convolution: the contribution of a single weight



$$z(l,n,x,y) = \sum_{m} \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each affine output is computed from multiple input maps simultaneously
- Each ***weight*** $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$

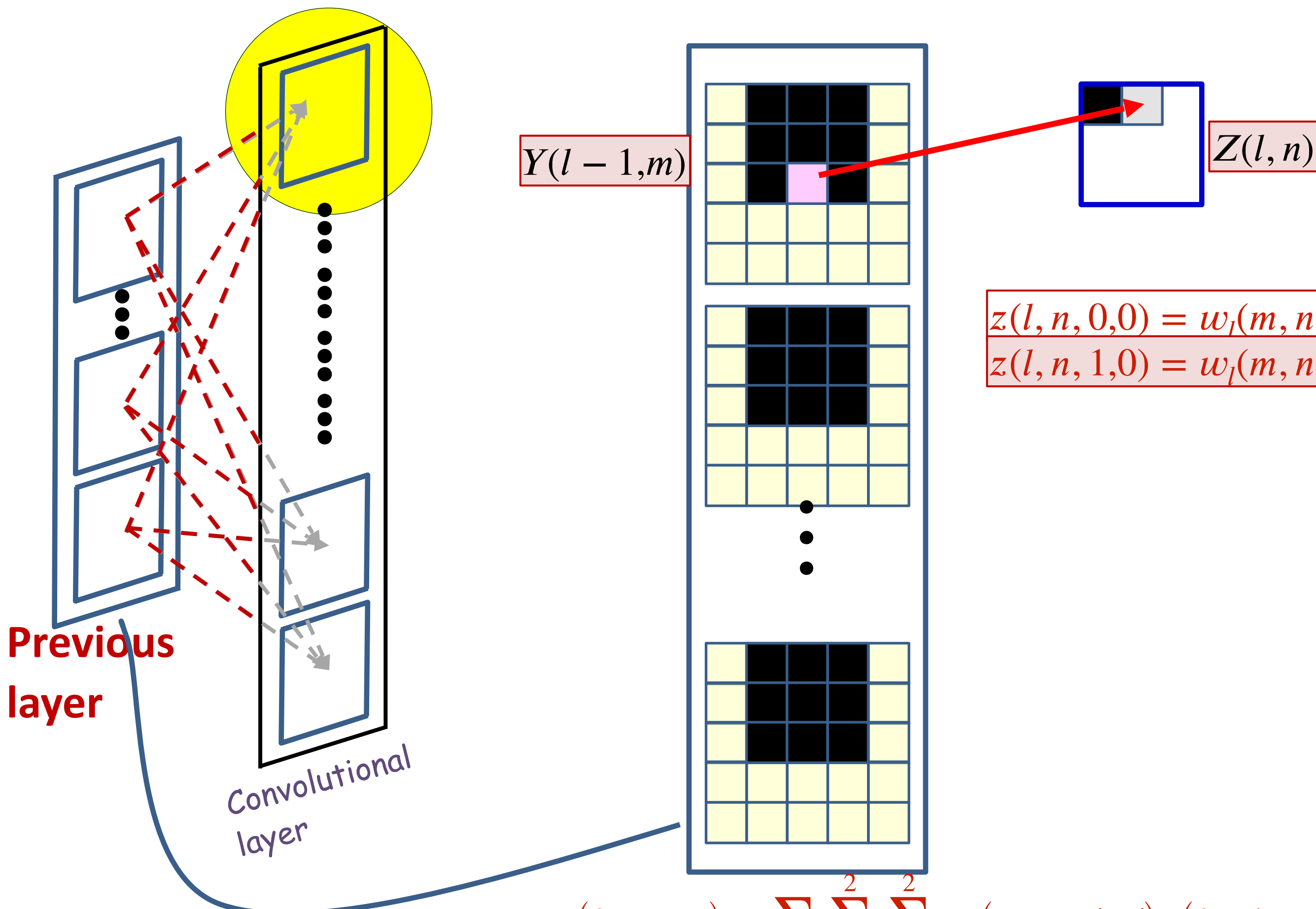
Convolution: the contribution of a single weight



$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight



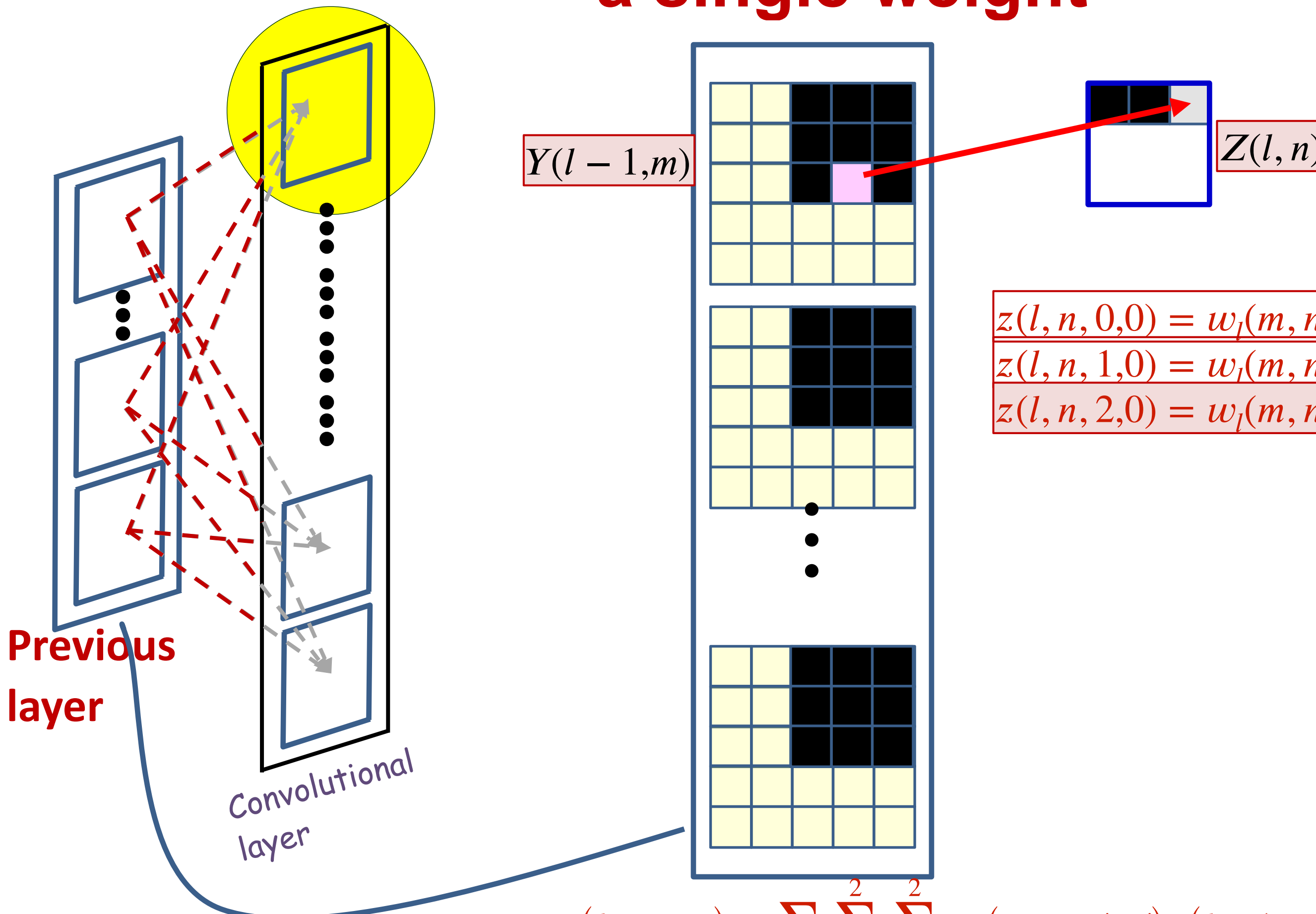
$z(l,n,0,0) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}2) + \ldots$

$z(l,n,1{,}0) = w_l(m,n,1{,}2)y(l-1,m,\ 2{,}2) + \ldots$

$$z(l,n,x,y) = \sum_{m}\sum_{x'=0}^{2}\sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1{,}2)$

Convolution: the contribution of a single weight



$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$

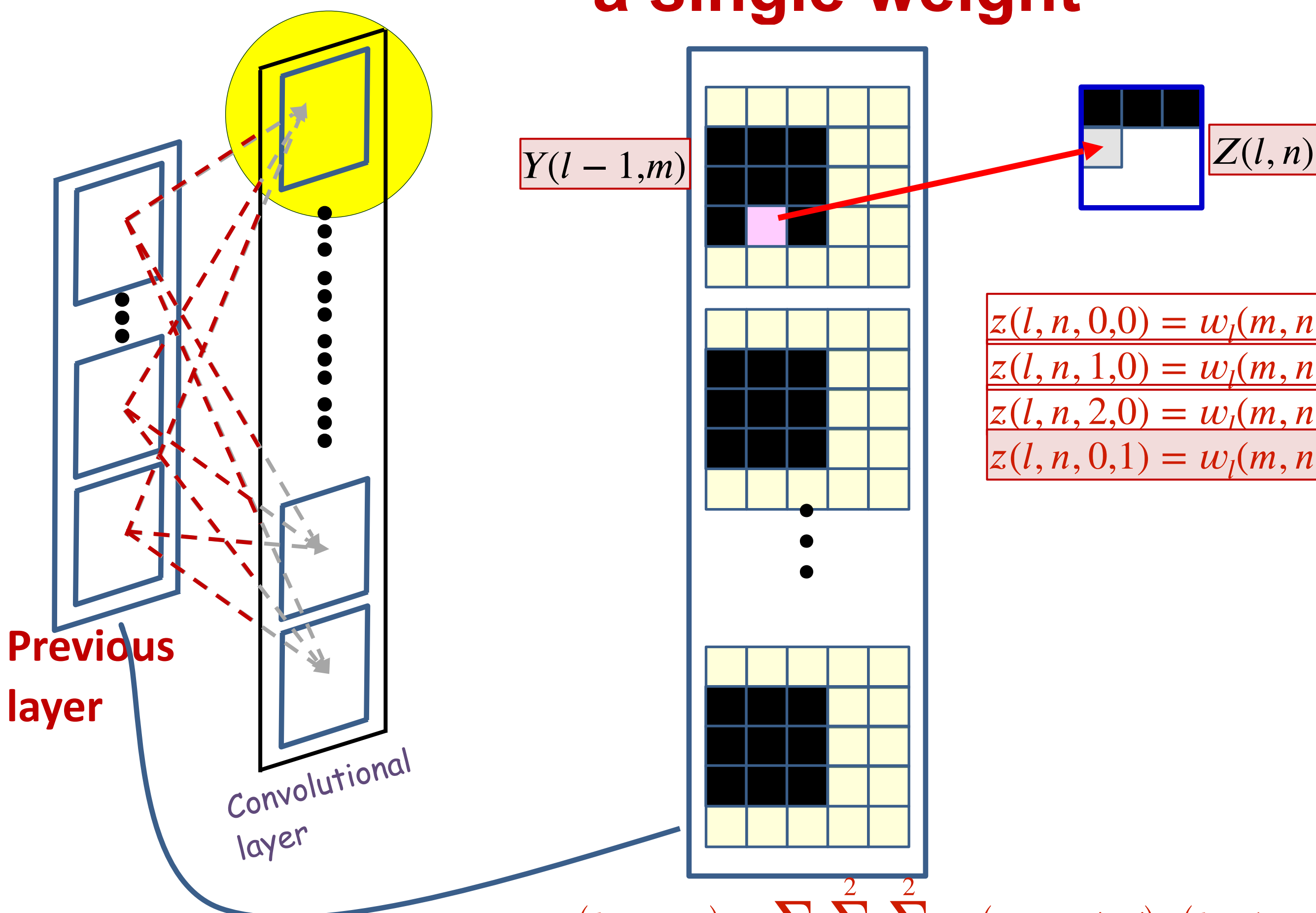
$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$

$$z\left(l,n,x,y\right) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l\left(m,n,x',y'\right) y\left(l-1,m,x+x',y+x'\right) + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight



$z(l,n,0{,}0) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}2) + \ldots$

$z(l,n,1{,}0) = w_l(m,n,1{,}2)y(l-1,m,\ 2{,}2) + \ldots$

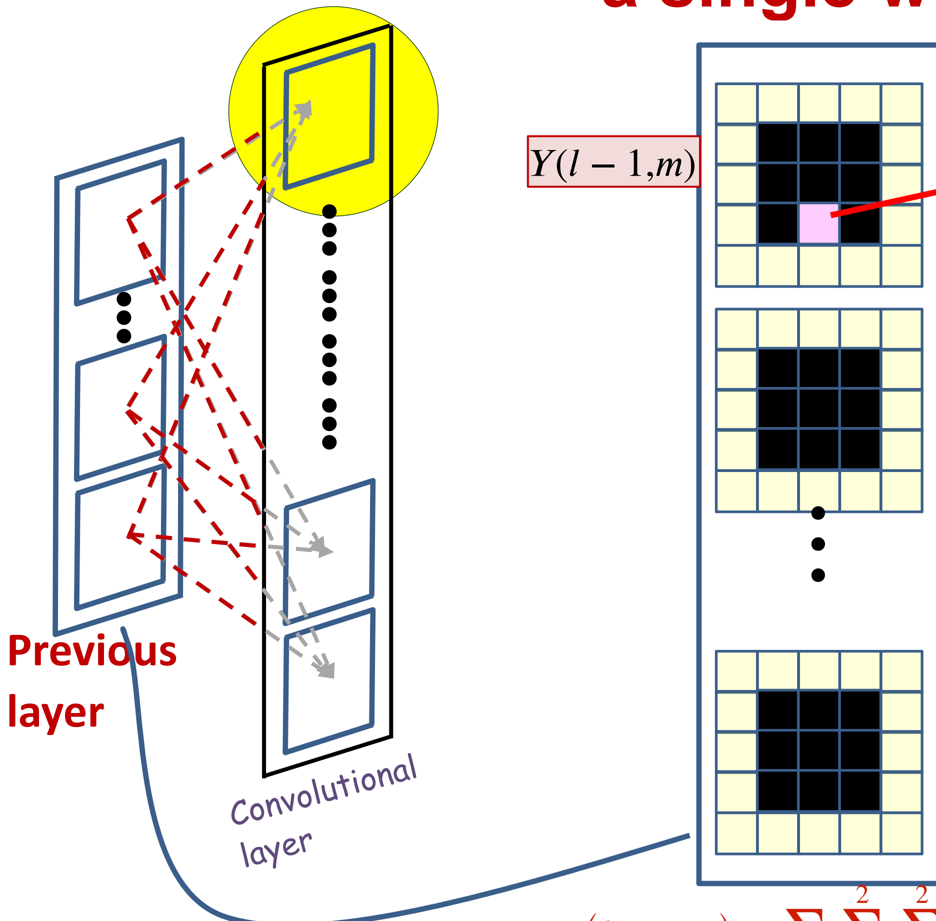
$z(l,n,2{,}0) = w_l(m,n,1{,}2)y(l-1,m,\ 3{,}2) + \ldots$

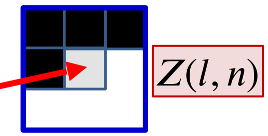
$z(l,n,0{,}1) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}3) + \ldots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\, y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1{,}2)$

Convolution: the contribution of a single weight





$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$

$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$

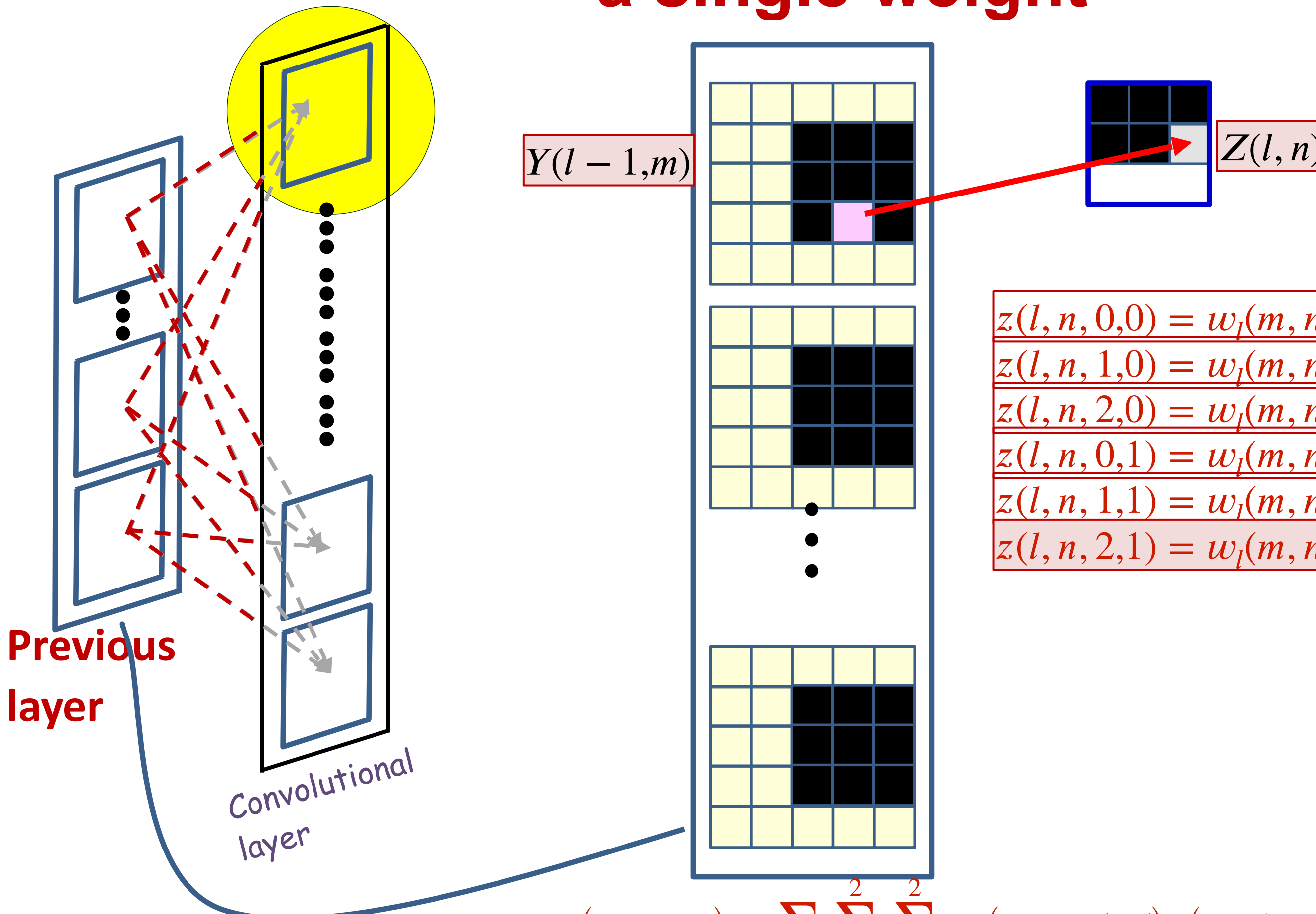
$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \ldots$

$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \ldots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight



$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$

$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$

$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \ldots$

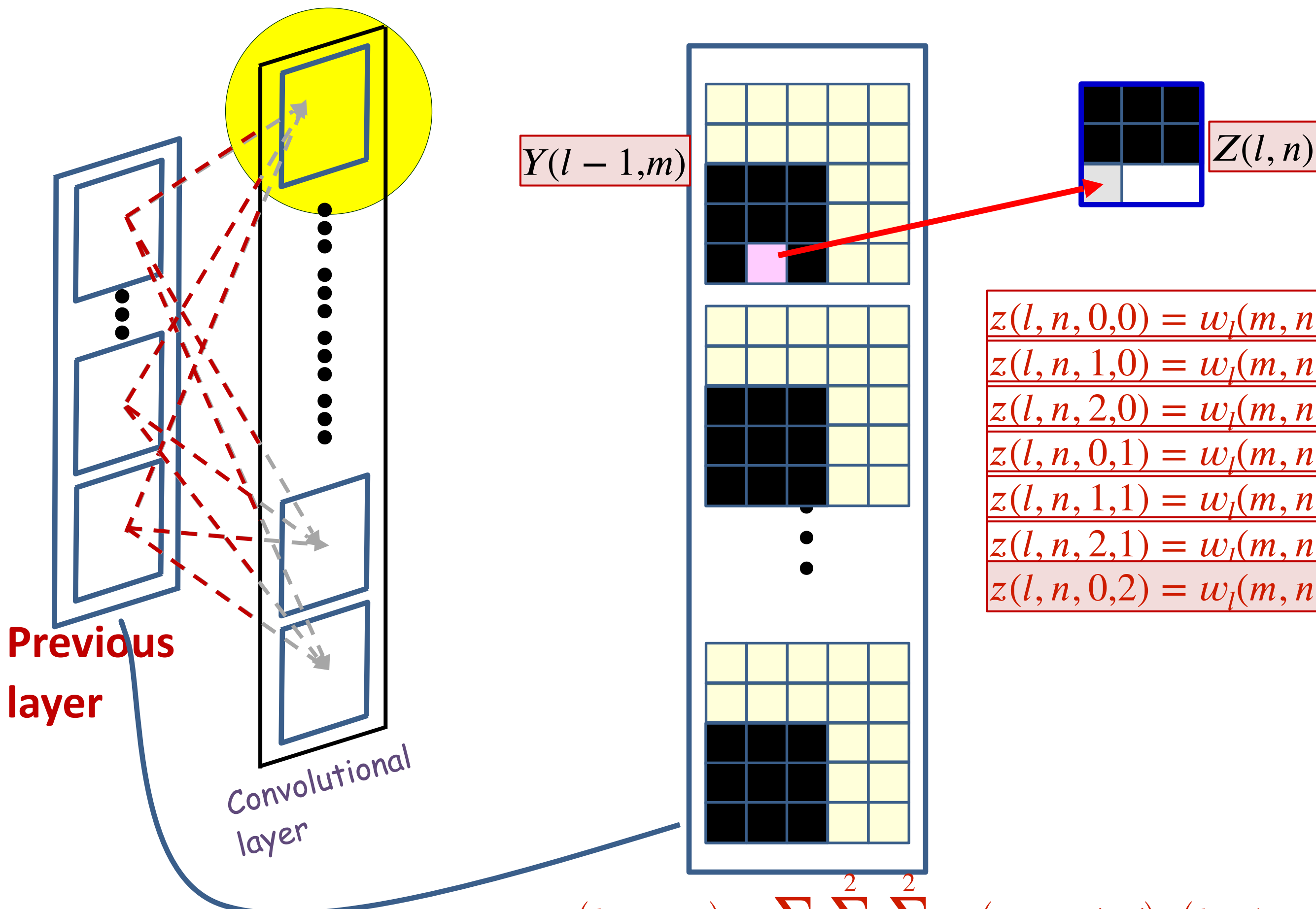
$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \ldots$

$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \ldots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\, y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight

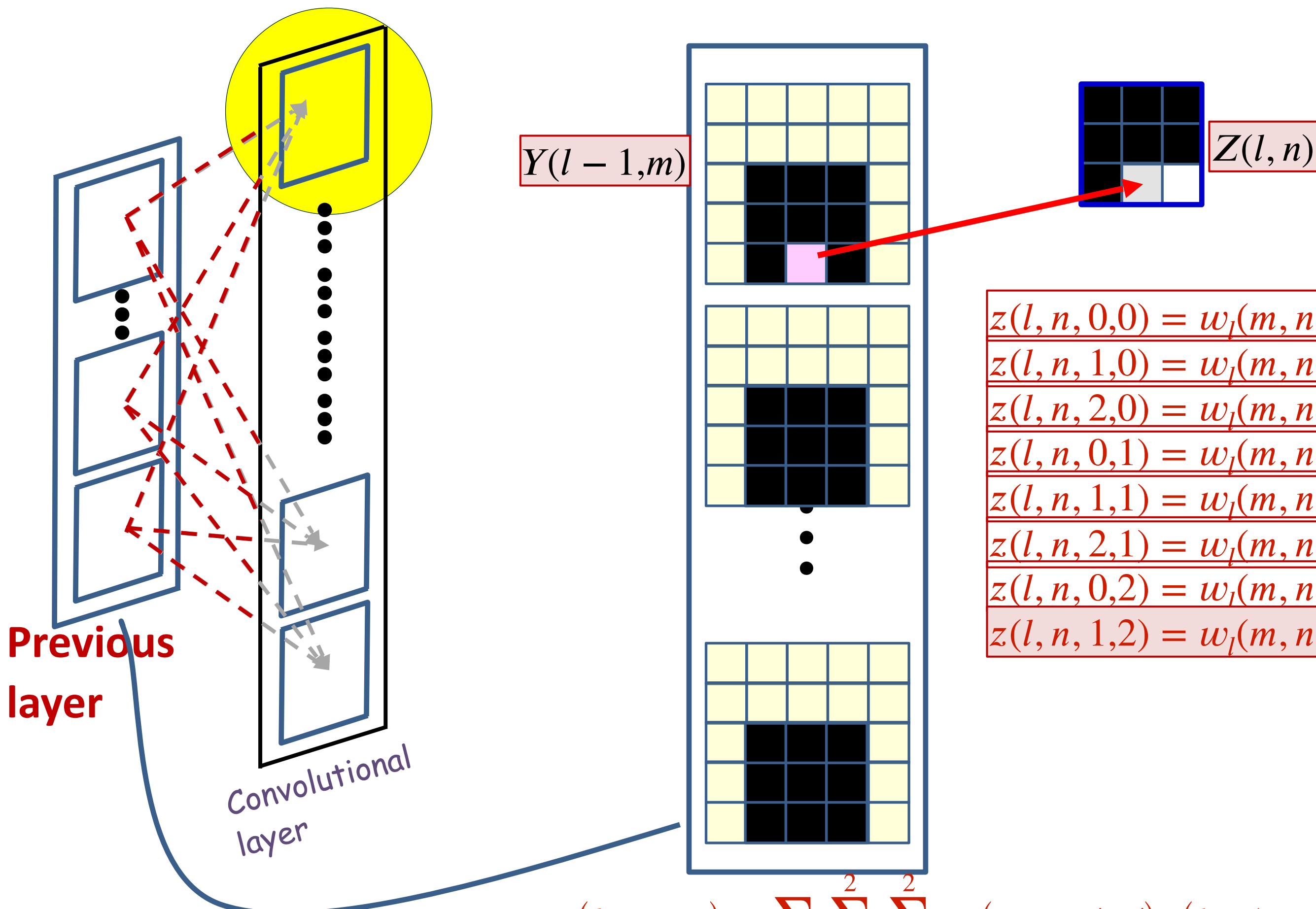


$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$
$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$
$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$
$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \ldots$
$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \ldots$
$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \ldots$
$z(l,n,0,2) = w_l(m,n,1,2)y(l-1,m,\ 1,4) + \ldots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight

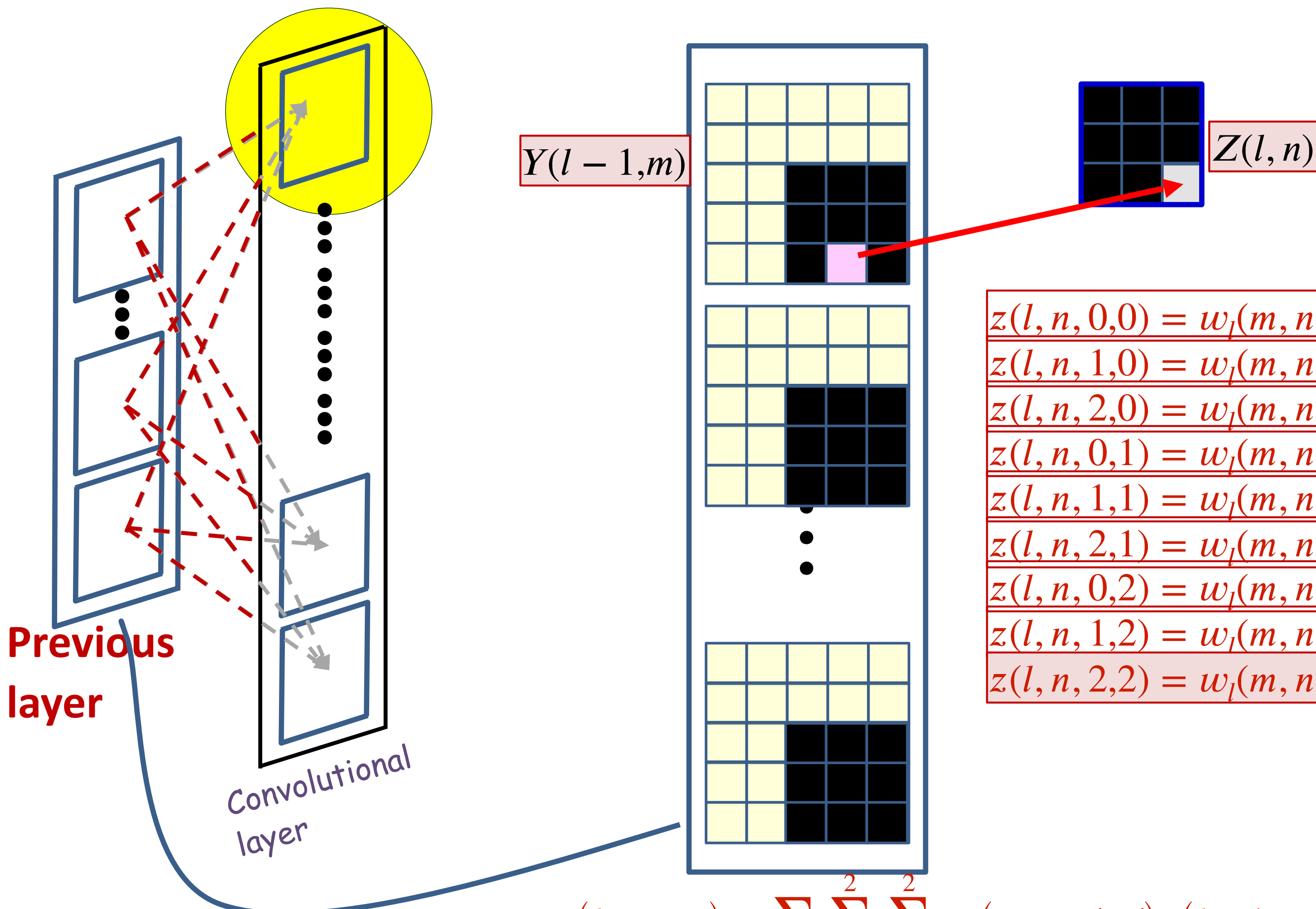


$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \dots$
$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \dots$
$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \dots$
$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \dots$
$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \dots$
$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \dots$
$z(l,n,0,2) = w_l(m,n,1,2)y(l-1,m,\ 1,4) + \dots$
$z(l,n,1,2) = w_l(m,n,1,2)y(l-1,m,\ 2,4) + \dots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight



$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$

$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$

$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \ldots$

$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \ldots$

$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \ldots$

$z(l,n,0,2) = w_l(m,n,1,2)y(l-1,m,\ 1,4) + \ldots$

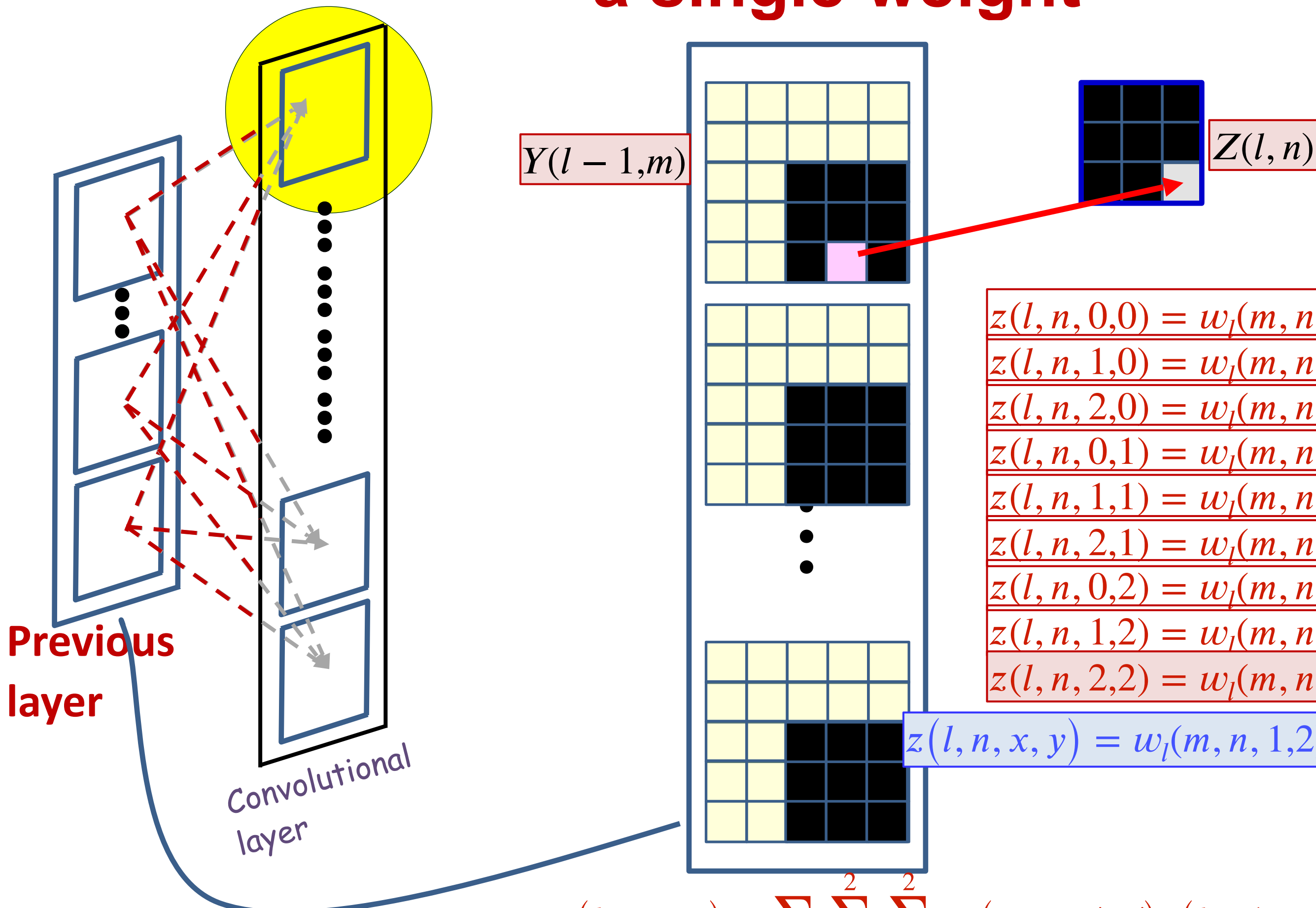
$z(l,n,1,2) = w_l(m,n,1,2)y(l-1,m,\ 2,4) + \ldots$

$z(l,n,2,2) = w_l(m,n,1,2)y(l-1,m,\ 3,4) + \ldots$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight



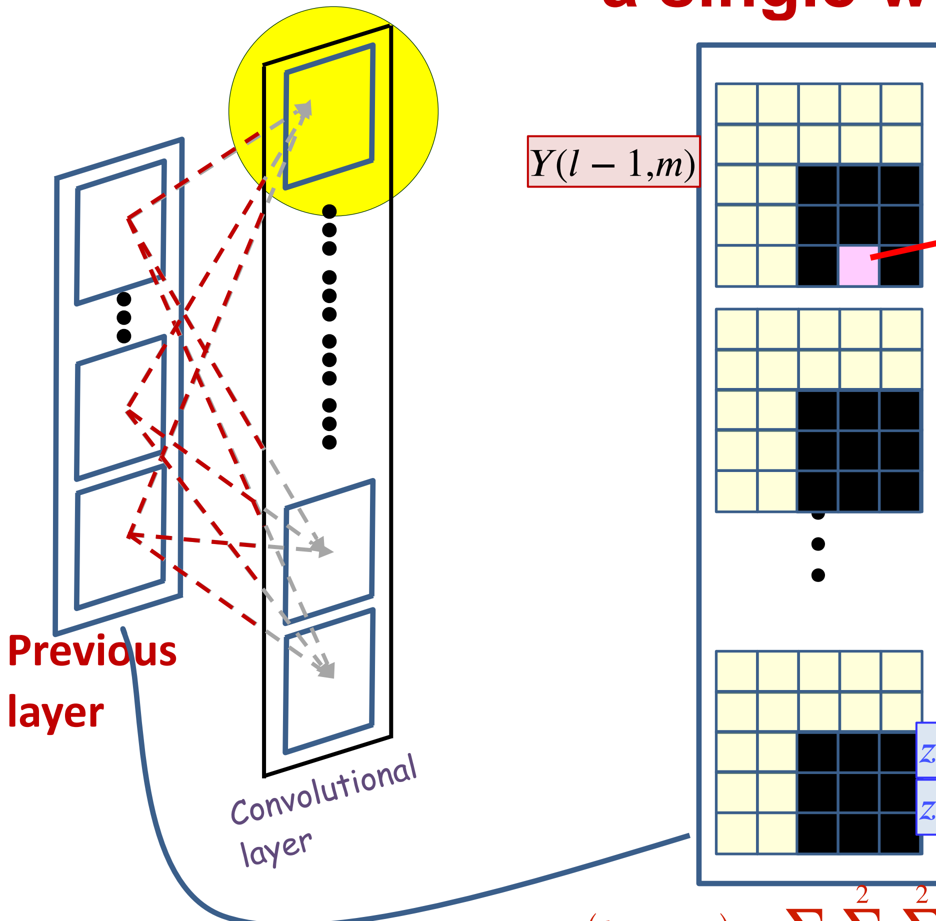
$$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \dots$$
$$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \dots$$
$$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \dots$$
$$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \dots$$
$$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \dots$$
$$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \dots$$
$$z(l,n,0,2) = w_l(m,n,1,2)y(l-1,m,\ 1,4) + \dots$$
$$z(l,n,1,2) = w_l(m,n,1,2)y(l-1,m,\ 2,4) + \dots$$
$$z(l,n,2,2) = w_l(m,n,1,2)y(l-1,m,\ 3,4) + \dots$$

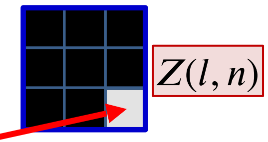
$$z(l,n,x,y) = w_l(m,n,1,2)y(l-1,m,\ x+1,y+2) + \dots$$

$$z(l,n,x,y) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l(m,n,x',y')\,y(l-1,m,x+x',y+x') + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1,2)$

Convolution: the contribution of a single weight





$z(l,n,0,0) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}2) + \ldots$

$z(l,n,1,0) = w_l(m,n,1{,}2)y(l-1,m,\ 2{,}2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1{,}2)y(l-1,m,\ 3{,}2) + \ldots$

$z(l,n,0,1) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}3) + \ldots$

$z(l,n,1,1) = w_l(m,n,1{,}2)y(l-1,m,\ 2{,}3) + \ldots$

$z(l,n,2,1) = w_l(m,n,1{,}2)y(l-1,m,\ 3{,}3) + \ldots$

$z(l,n,0,2) = w_l(m,n,1{,}2)y(l-1,m,\ 1{,}4) + \ldots$

$z(l,n,1,2) = w_l(m,n,1{,}2)y(l-1,m,\ 2{,}4) + \ldots$

$z(l,n,2,2) = w_l(m,n,1{,}2)y(l-1,m,\ 3{,}4) + \ldots$

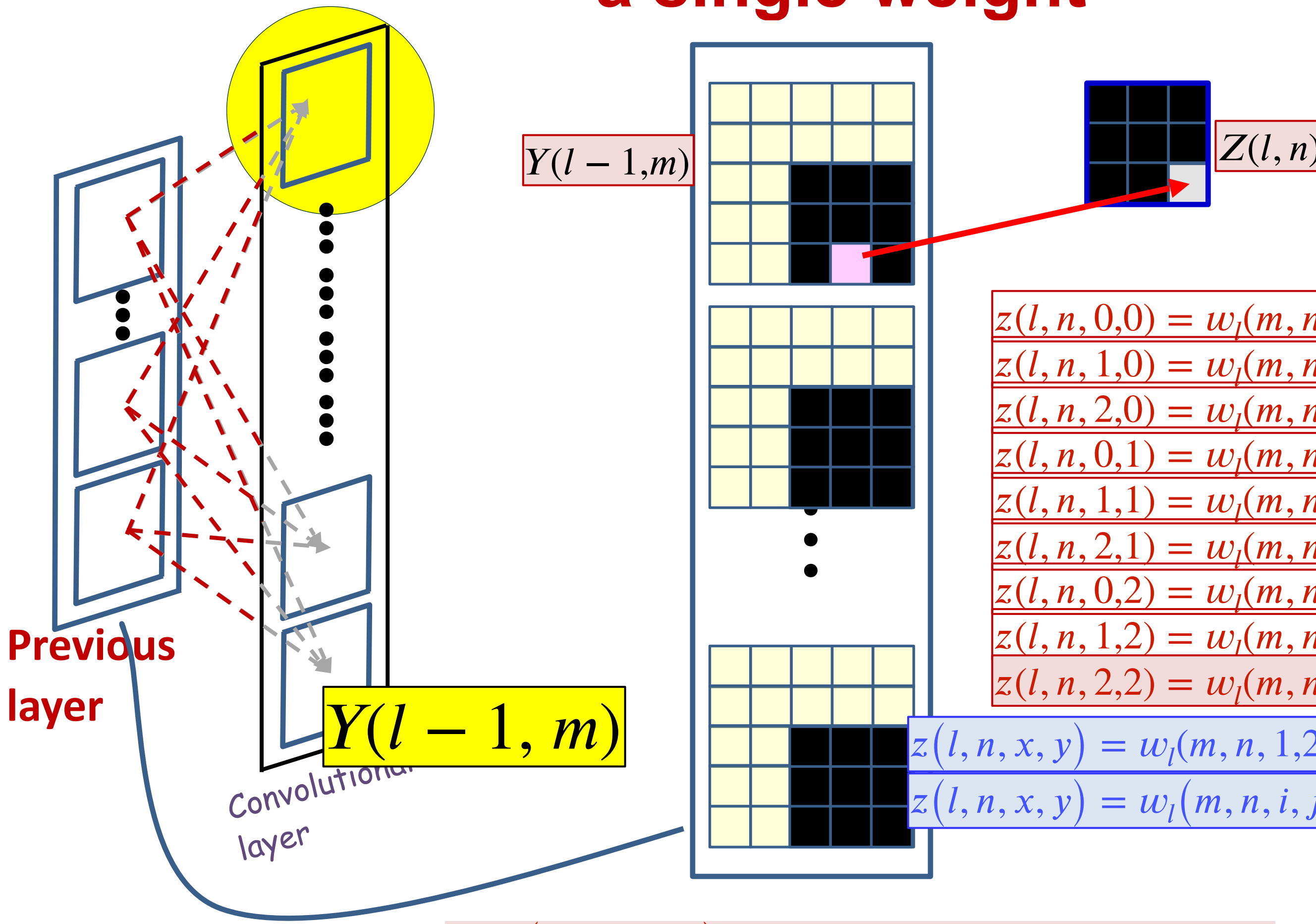
$z(l,n,x,y) = w_l(m,n,1{,}2)y\left(l-1,m,\ x+1,y+2\right) + \ldots$

$z(l,n,x,y) = w_l\left(m,n,i,j\right)y\left(l-1,m,\ x+i,y+j\right) + \ldots$

$$z\left(l,n,x,y\right) = \sum_m \sum_{x'=0}^{2} \sum_{y'=0}^{2} w_l\left(m,n,x',y'\right)y\left(l-1,m,x+x',y+x'\right) + b_l(n)$$

- Each weight $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - Consider the contribution of one filter components: e.g. $w_l(m,n,1{,}2)$

Convolution: the contribution of a single weight



$z(l,n,0,0) = w_l(m,n,1,2)y(l-1,m,\ 1,2) + \ldots$

$z(l,n,1,0) = w_l(m,n,1,2)y(l-1,m,\ 2,2) + \ldots$

$z(l,n,2,0) = w_l(m,n,1,2)y(l-1,m,\ 3,2) + \ldots$

$z(l,n,0,1) = w_l(m,n,1,2)y(l-1,m,\ 1,3) + \ldots$

$z(l,n,1,1) = w_l(m,n,1,2)y(l-1,m,\ 2,3) + \ldots$

$z(l,n,2,1) = w_l(m,n,1,2)y(l-1,m,\ 3,3) + \ldots$

$z(l,n,0,2) = w_l(m,n,1,2)y(l-1,m,\ 1,4) + \ldots$

$z(l,n,1,2) = w_l(m,n,1,2)y(l-1,m,\ 2,4) + \ldots$

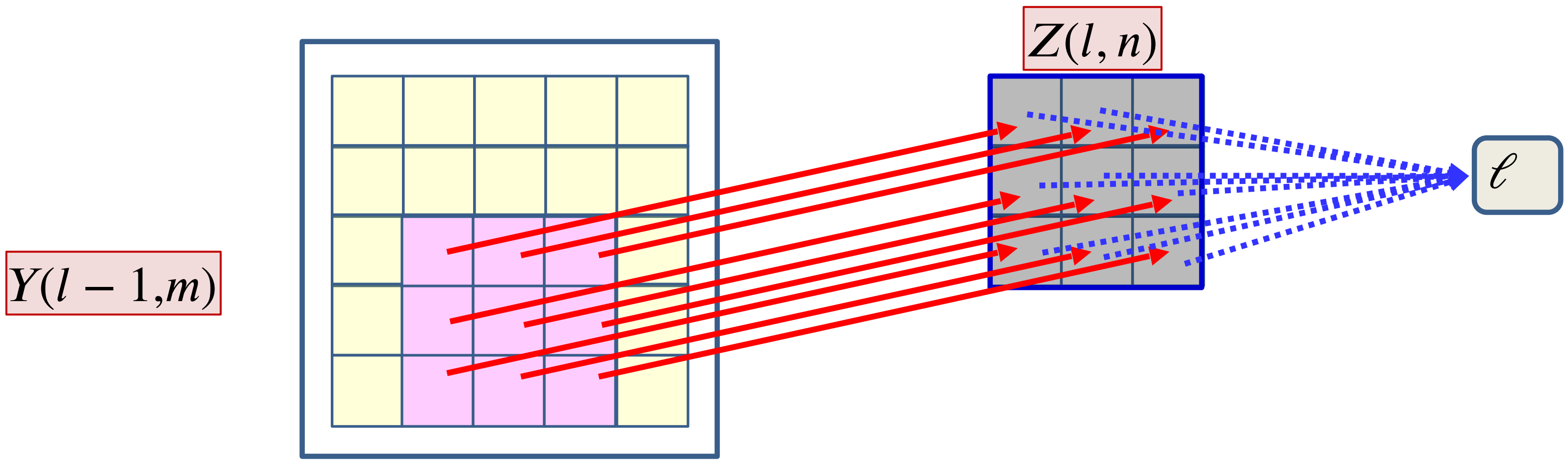
$z(l,n,2,2) = w_l(m,n,1,2)y(l-1,m,\ 3,4) + \ldots$

$z(l,n,x,y) = w_l(m,n,1,2)y\left(l-1,m,\ x+1,y+2\right) + \ldots$

$z(l,n,x,y) = w_l\left(m,n,i,j\right)y\left(l-1,m,\ x+i,y+j\right) + \ldots$

$$\frac{dz\left(l,n,x,y\right)}{dw_l\left(m,n,i,j\right)} = y\left(l-1,m,x+i,y+j\right)$$

The derivative for a single weight



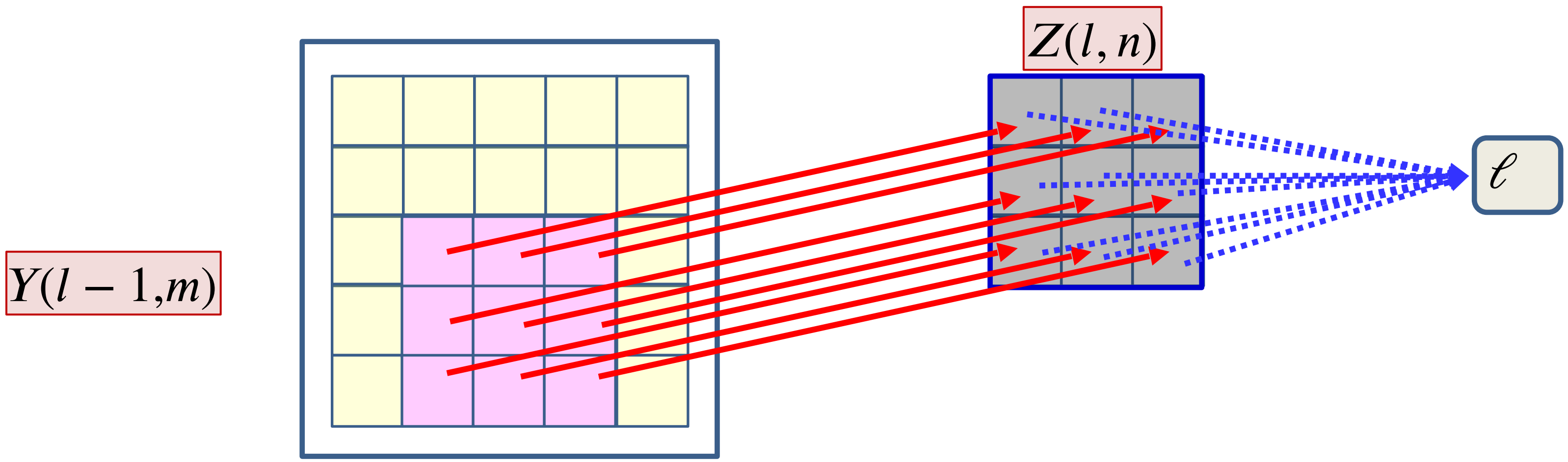
- Each filter component $w_l(m, n, i, j)$ affects several $z(l, n, x, y)$
  - The derivative of each $z(l, n, x, y)$ w.r.t. $w_l(m, n, i, j)$ is given by

$$\frac{dz(l, n, x, y)}{dw_l(m, n, i, j)} = y(l-1, m, x+i, y+j)$$

- The final divergence is influenced by *every* $z(l, n, x, y)$
- The derivative of the divergence w.r.t $w_l(m, n, i, j)$ must sum over all $z(l, n, x, y)$ terms it influences

$$\frac{\partial \ell}{\partial w_l(m, n, i, j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l, n, x, y)} \frac{\partial z(l, n, x, y)}{\partial w_l(m, n, i, j)}$$

The derivative for a single weight



- Each filter component $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - The derivative of each $z(l,n,x,y)$ w.r.t. $w_l(m,n,i,j)$ is given by

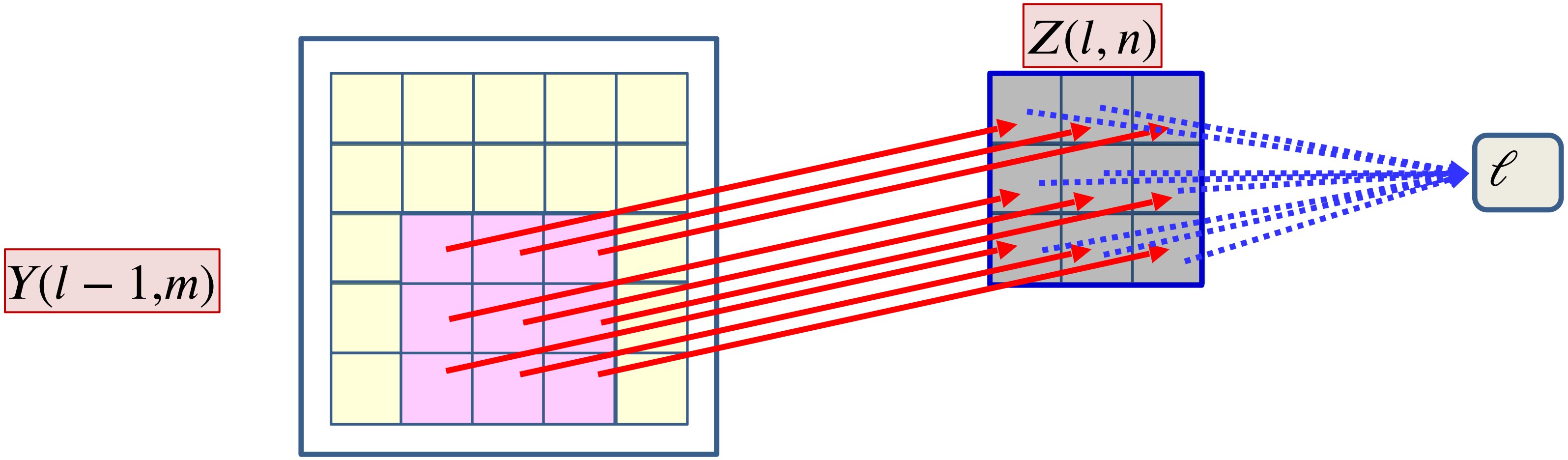
$$\frac{dz(l,n,x,y)}{dw_l(m,n,i,j)} = y(l-1,m,x+i,y+j)$$

- The final divergence is influenced by *every* $z(l,n,x,y)$
- The derivative of the divergence w.r.t $w_l(m,n,i,j)$ must sum over all $z(l,n,x,y)$ terms it influences

Already computed

$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \boxed{\frac{\partial \ell}{\partial z(l,n,x,y)}} \frac{\partial z(l,n,x,y)}{\partial w_l(m,n,i,j)}$$

The derivative for a single weight



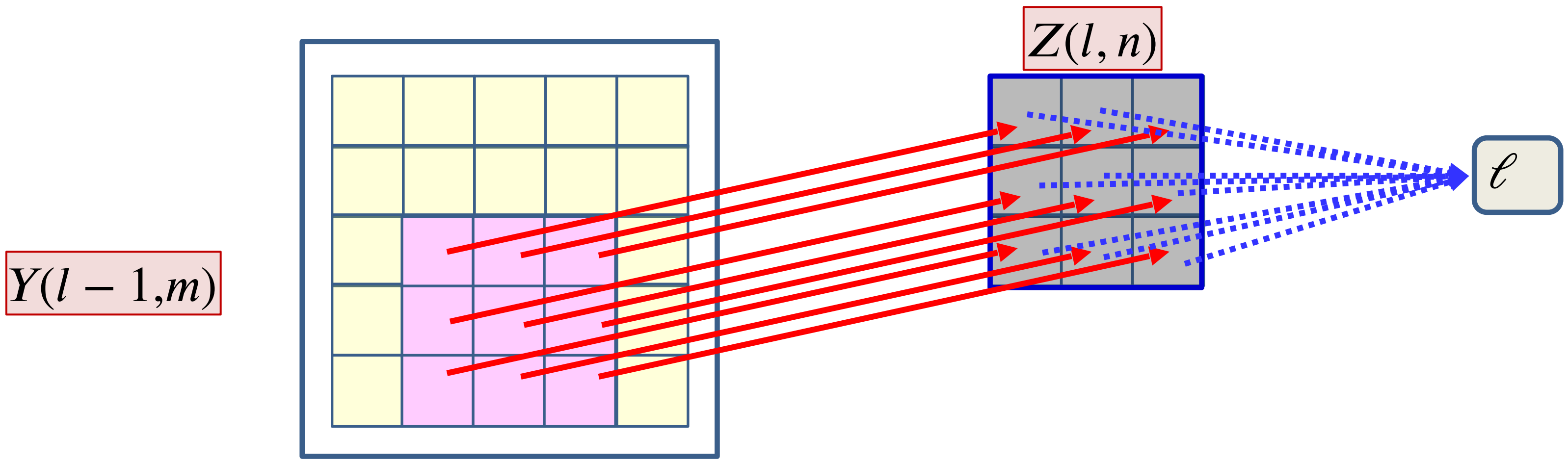
- Each filter component $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - The derivative of each $z(l,n,x,y)$ w.r.t. $w_l(m,n,i,j)$ is given by

$$\boxed{\frac{dz(l,n,x,y)}{dw_l(m,n,i,j)} = y(l-1,m,x+i,y+j)}$$

- The final divergence is influenced by *every* $z(l,n,x,y)$
- The derivativ Already computed e w.r.t $w_l(m,n,i,j)$ must sum over all $z(l,n,x,y)$ terms it influences

$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \boxed{\frac{\partial \ell}{\partial z(l,n,x,y)}} \boxed{\frac{\partial z(l,n,x,y)}{\partial w_l(m,n,i,j)}}$$

The derivative for a single weight



- Each filter component $w_l(m,n,i,j)$ affects several $z(l,n,x,y)$
  - The derivative of each $z(l,n,x,y)$ w.r.t. $w_l(m,n,i,j)$ is given by

$$\frac{dz(l,n,x,y)}{dw_l(m,n,i,j)} = Y(l-1,m,x+i,y+j)$$

- The final divergence is influenced by *every* $z(l,n,x,y)$
- The derivative of the divergence w.r.t $w_l(m,n,i,j)$ must sum over all $z(l,n,x,y)$ terms it influences

$$\boxed{\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)}$$

But this too is a convolution

$$\boxed{\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)}$$

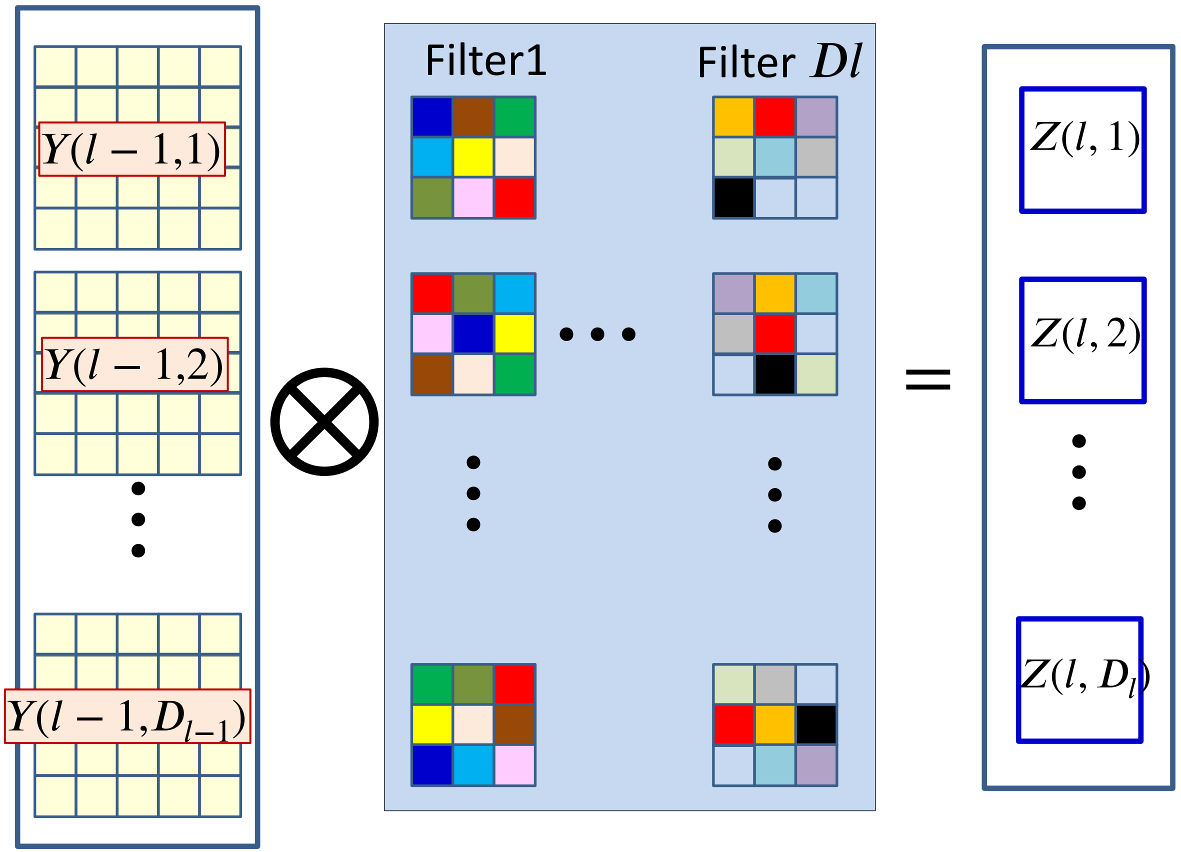
- The derivatives for all components of all filters can be computed directly from the above formula

- In fact it is just a convolution

$$\boxed{\frac{\partial \ell}{\partial w_l(m,n)} = \frac{\partial \ell}{\partial z(l,n)} \otimes Y(l-1,m)}$$

- How?

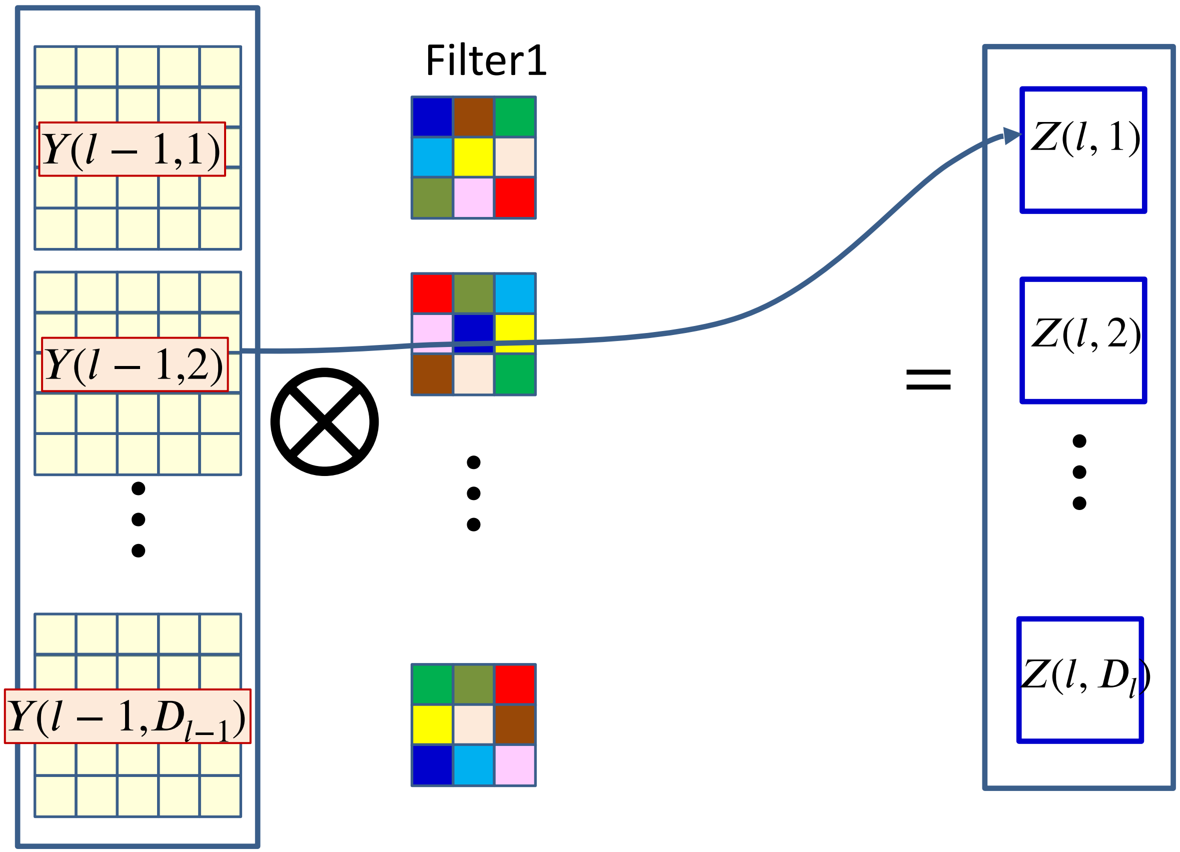
Recap: Convolution



$$z(l,n,x,y) = \sum_m \sum_{i=0}^{2} \sum_{j=0}^{2} w_l(m,n,i,j)\, y(l-1,m,x+i,y+j) + b_l(n)$$

- Forward computation: Each filter produces an affine map

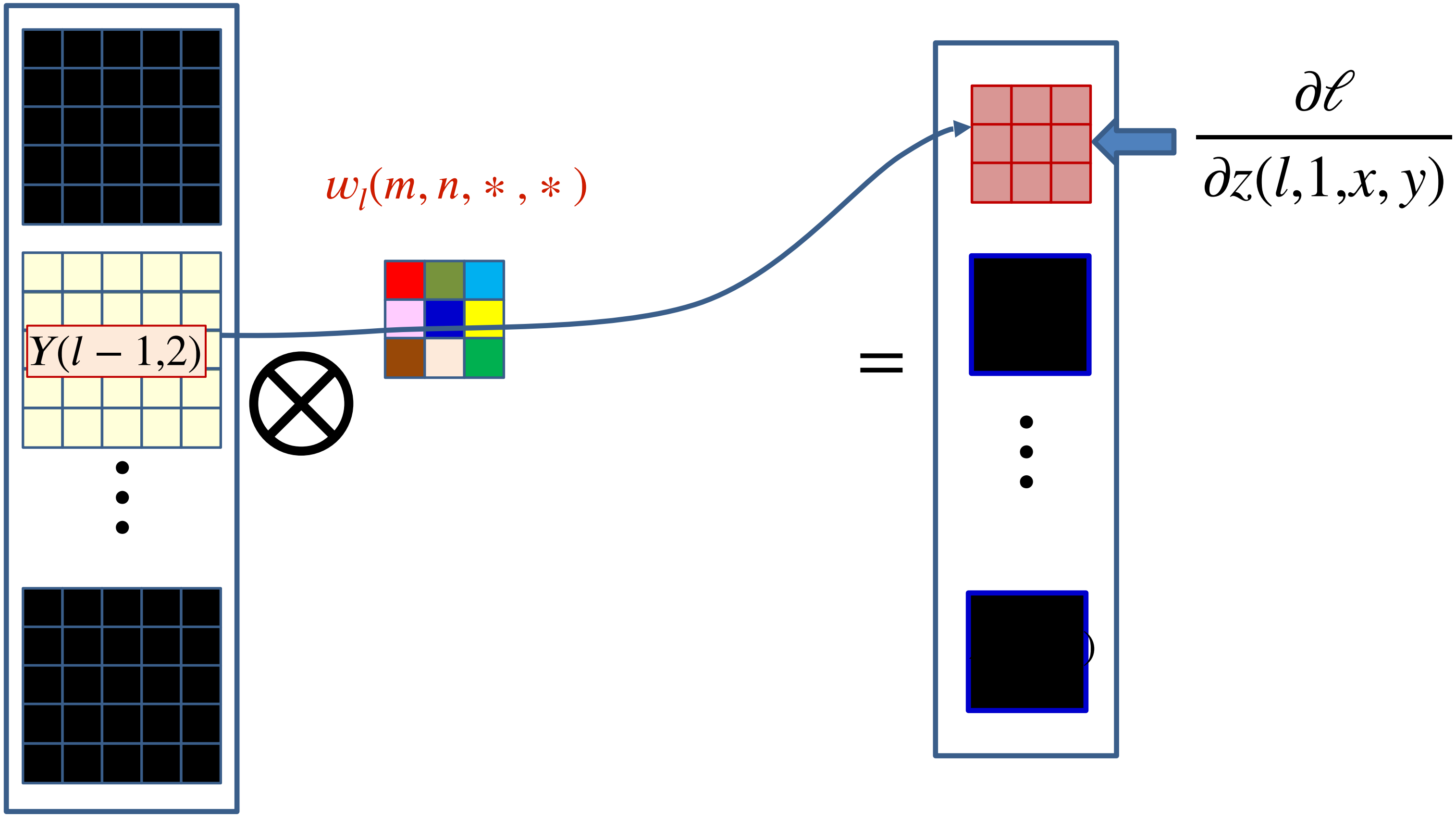
Recap: Convolution



$$z(l,n,x,y) = \sum_{m}\sum_{i=0}^{2}\sum_{j=0}^{2} w_l(m,n,i,j)\,y(l-1,m,x+i,y+j) + b_l(n)$$

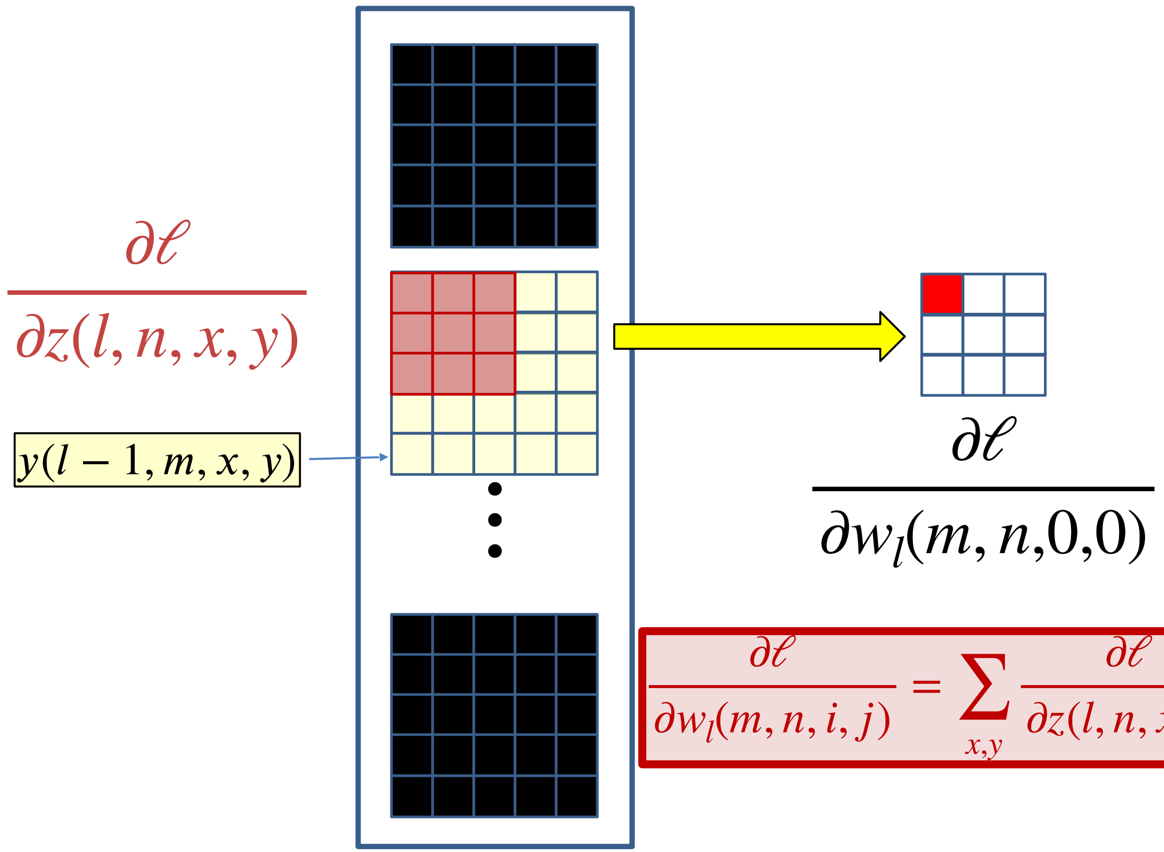
- $Y(l-1,m)$ influences $Z(l,n)$ through $w_l(m,n)$

The filter derivative



- The derivatives of the divergence w.r.t. every element of $Z(l, n)$ is known
  - Must use them to compute the derivative for $w_l(m, n, *, *)$

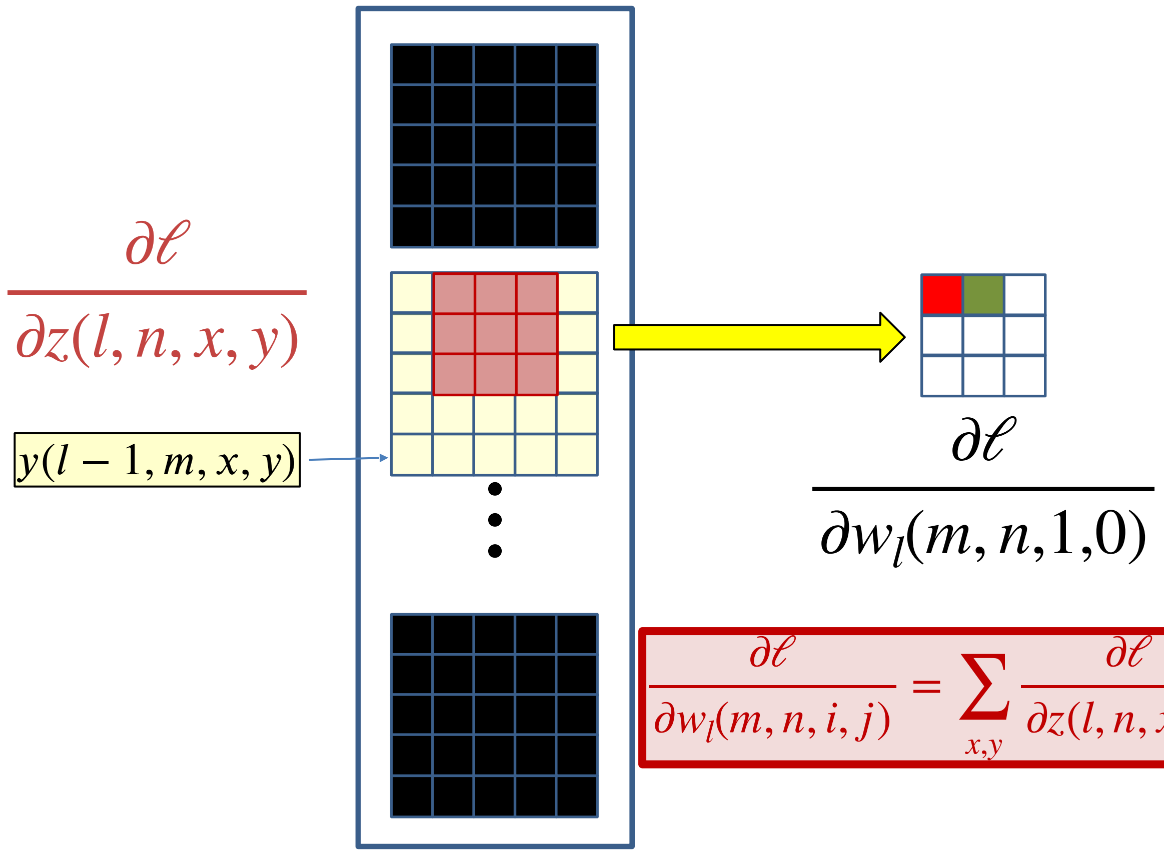
The filter derivative



$$\boxed{\frac{\partial \ell}{\partial w_l(m, n, i, j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l, n, x, y)} Y(l-1, m, x+i, y+j)}$$

- The derivatives of the divergence w.r.t. every element of $Z(l, n)$ is known
  - Must use them to compute the derivative for $w_l(m, n, *, *)$

The filter derivative



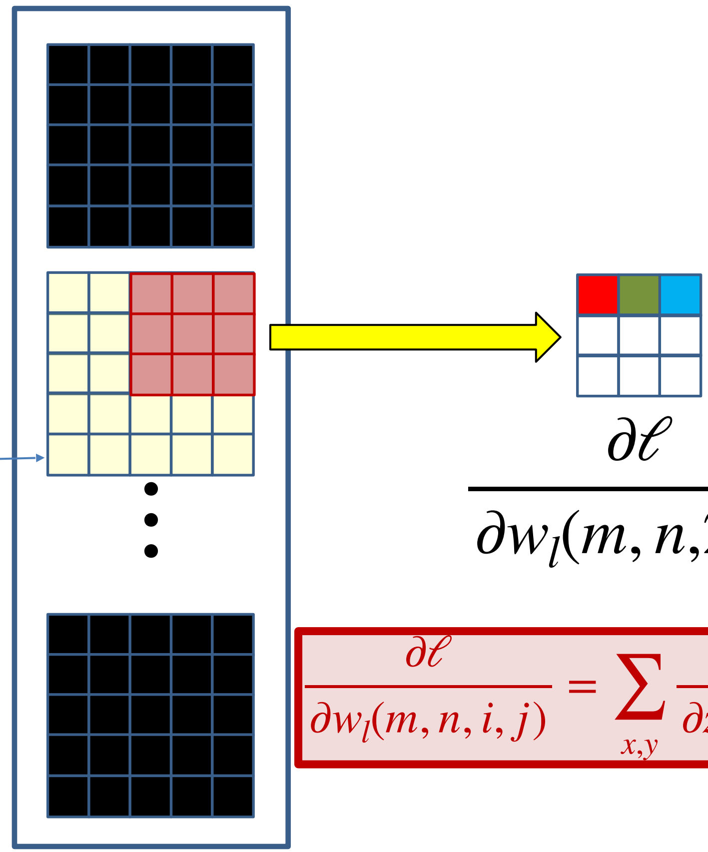
$$\frac{\partial \ell}{\partial w_l(m, n, i, j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l, n, x, y)} Y(l-1, m, x+i, y+j)$$

- The derivatives of the divergence w.r.t. every element of $Z(l, n)$ is known
  - Must use them to compute the derivative for $w_l(m, n, *, *)$

The filter derivative

$$\frac{\partial \ell}{\partial z(l,n,x,y)}$$

$y(l-1,m,x,y)$

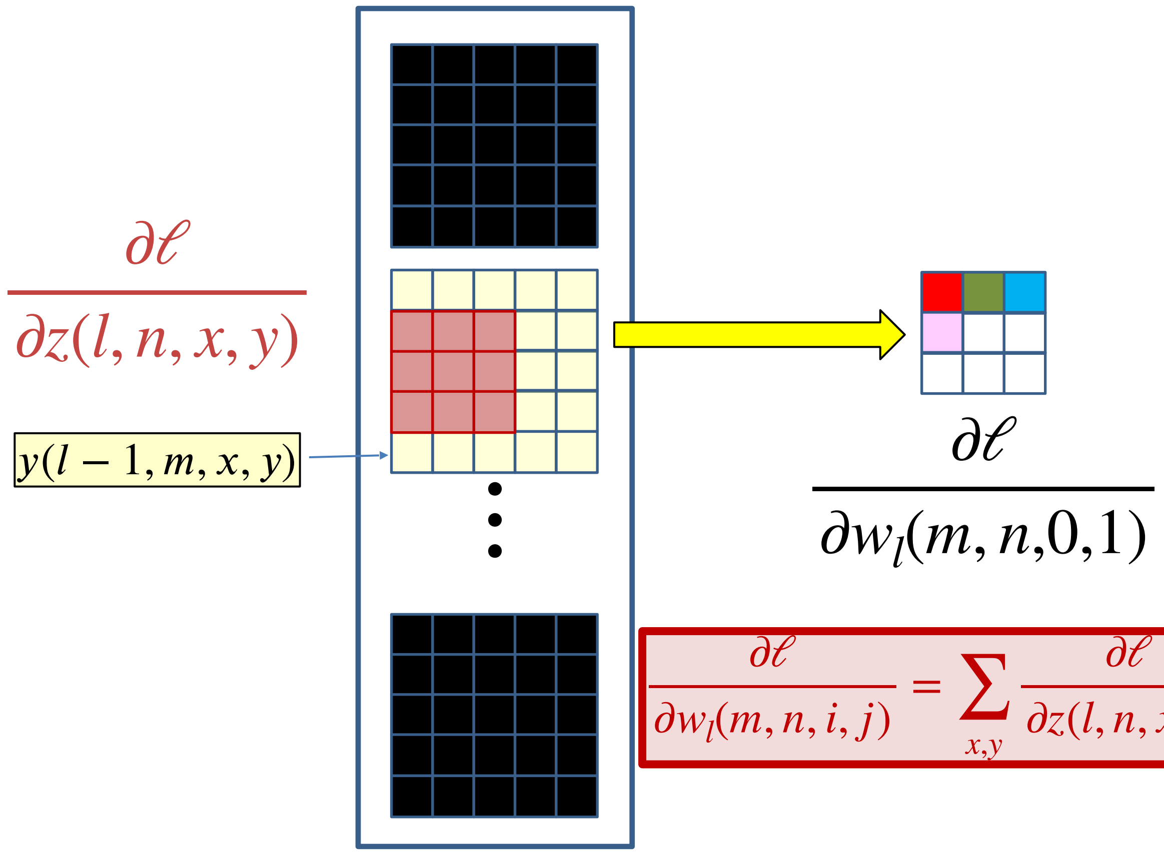


$$\frac{\partial \ell}{\partial w_l(m,n,2,0)}$$

$$\boxed{\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)}$$

- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

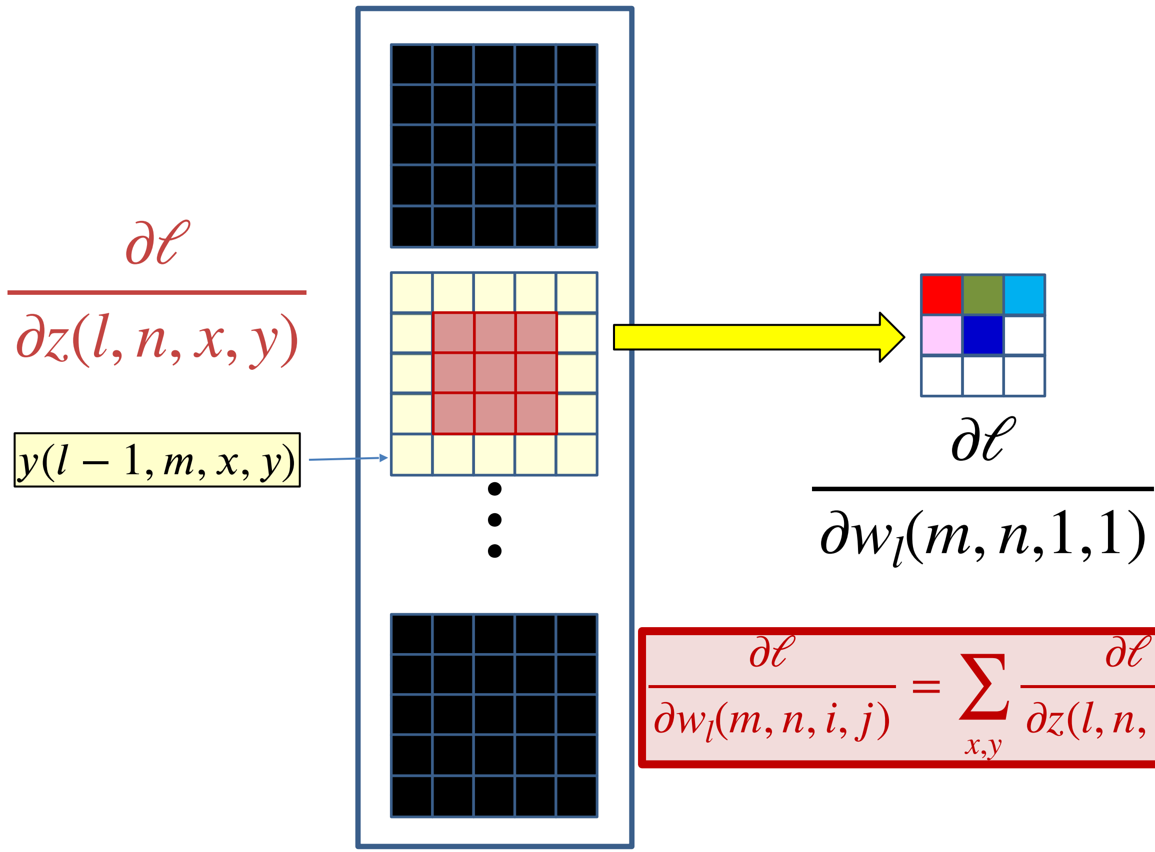
The filter derivative



$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)$$

- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

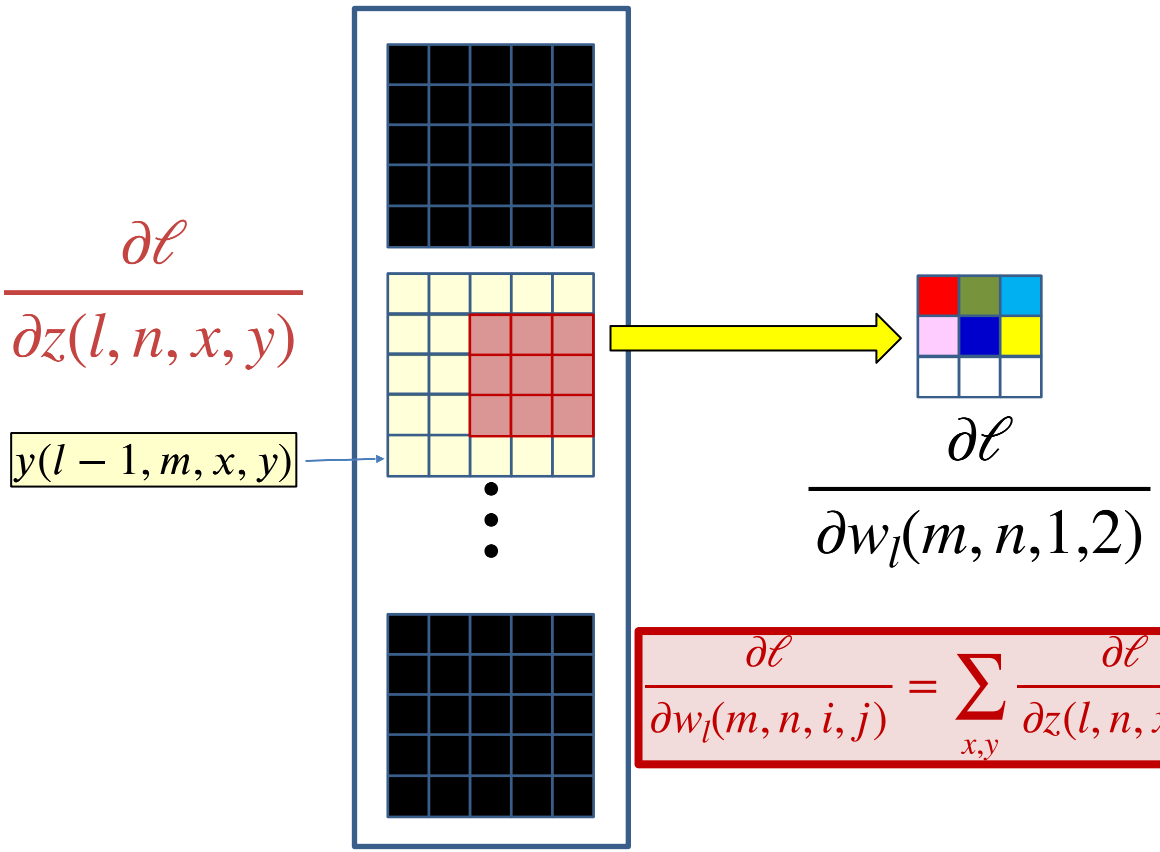
The filter derivative



$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)$$

- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

The filter derivative



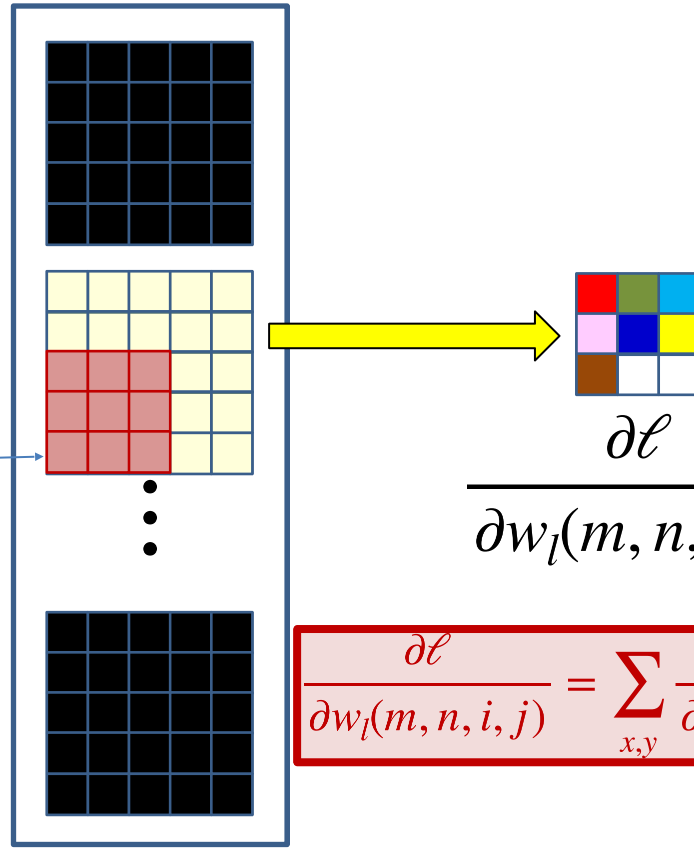
$$\frac{\partial \ell}{\partial w_l(m, n, i, j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l, n, x, y)} Y(l-1, m, x+i, y+j)$$

- The derivatives of the divergence w.r.t. every element of $Z(l, n)$ is known
  - Must use them to compute the derivative for $w_l(m, n, *, *)$

The filter derivative

$$\frac{\partial \ell}{\partial z(l,n,x,y)}$$

$y(l-1,m,x,y)$

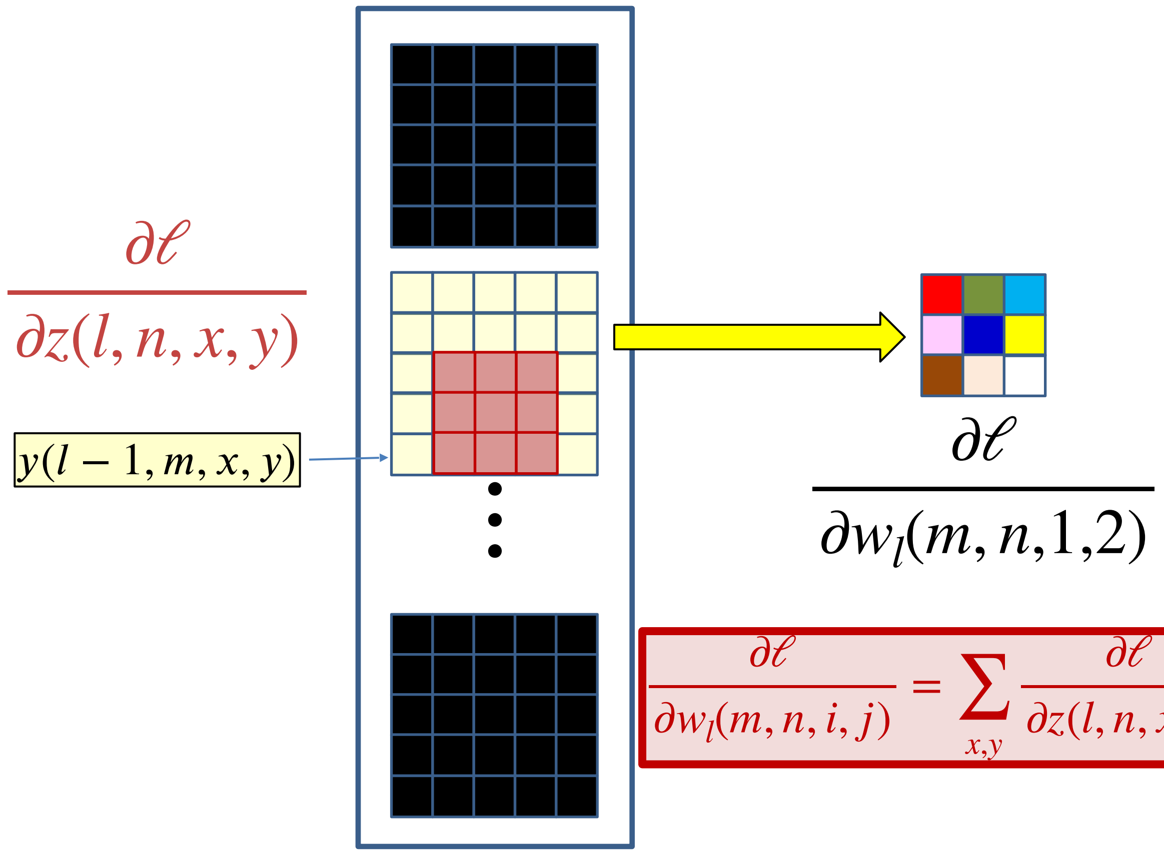


$$\frac{\partial \ell}{\partial w_l(m,n,0,2)}$$

$$\boxed{\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)}$$

- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

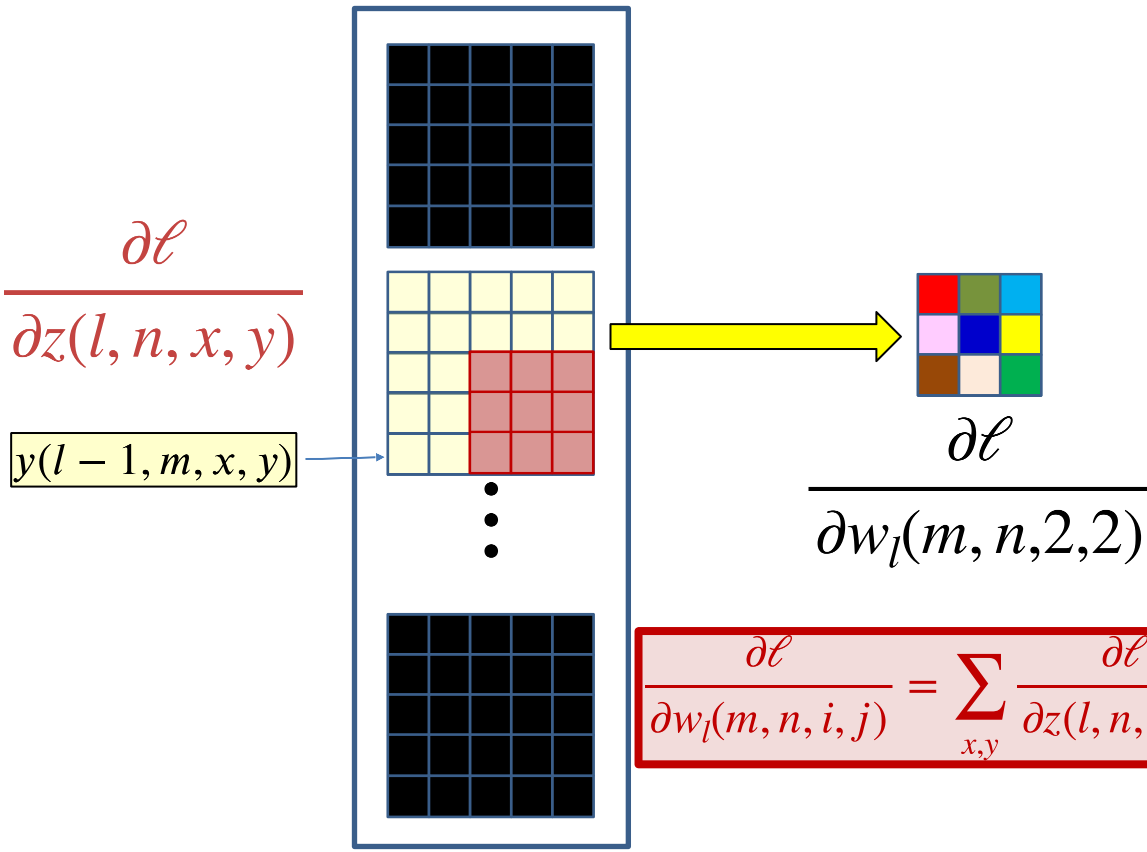
The filter derivative



$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)$$

- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

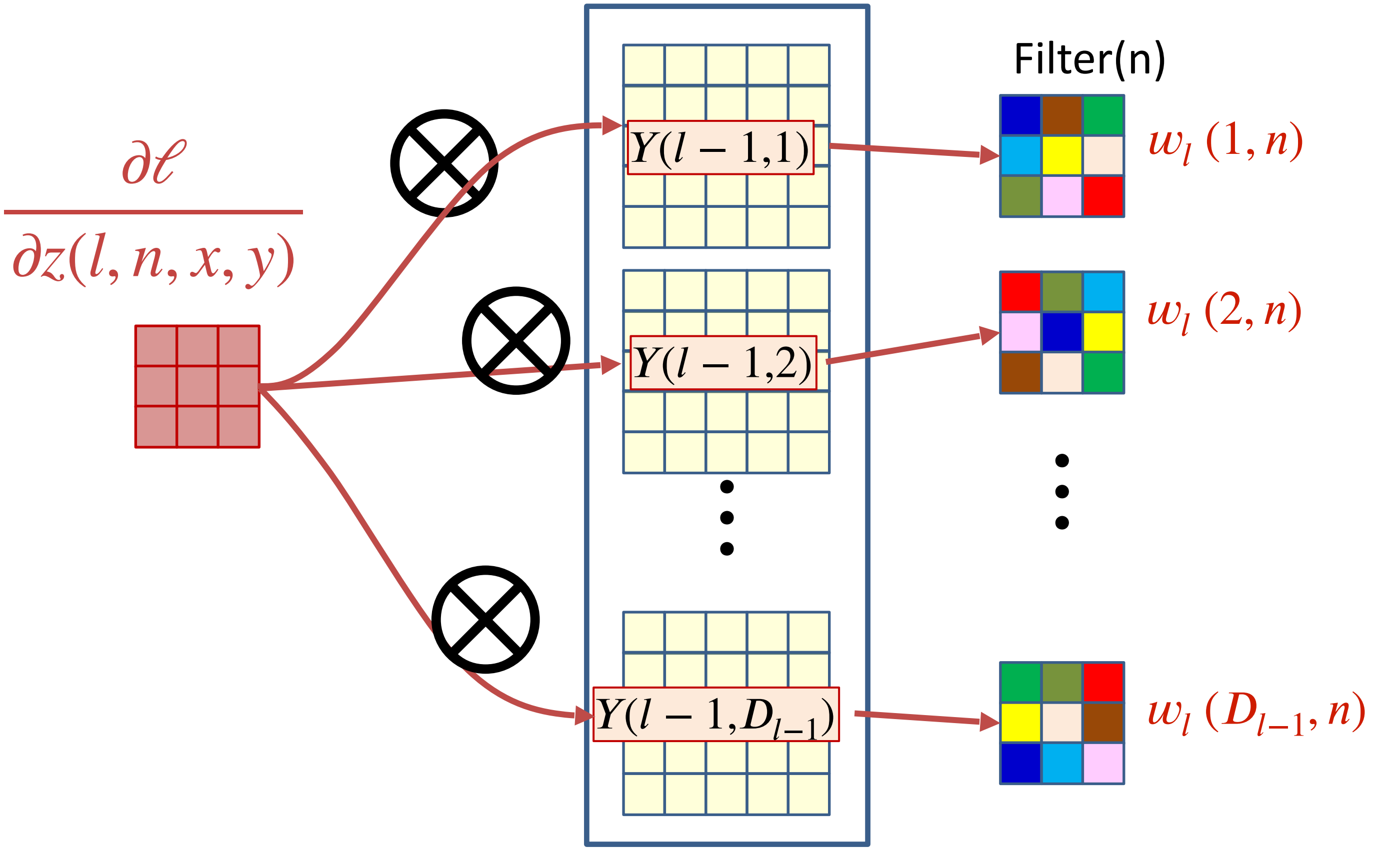
The filter derivative



$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j)$$

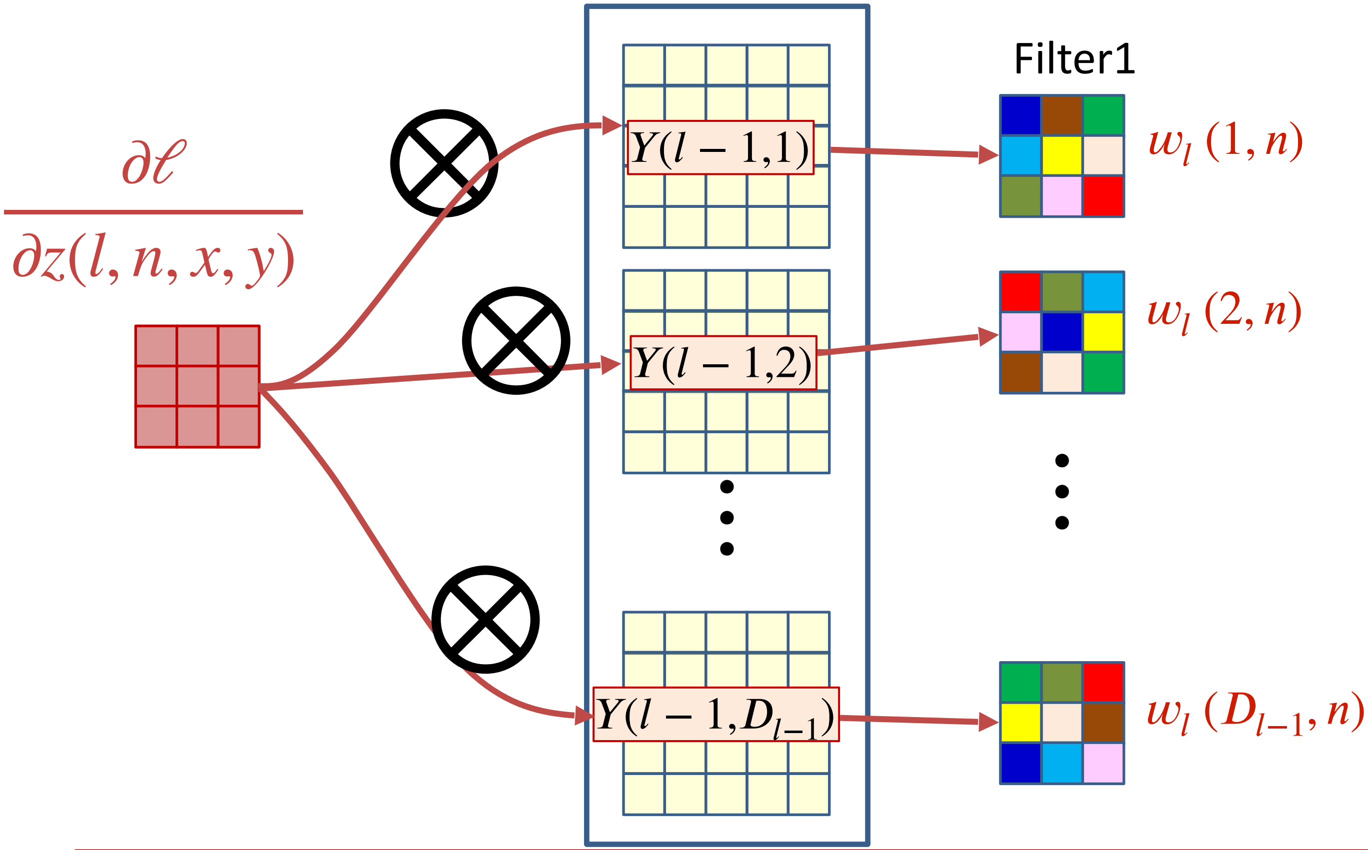
- The derivatives of the divergence w.r.t. every element of $Z(l,n)$ is known
  - Must use them to compute the derivative for $w_l(m,n,*,*)$

The filter derivative



- The derivative of the n^{th} affine map $Z(l,n)$ convolves with every output map $Y(l-1,m)$ of the $(l-1)^{\text{th}}$ layer, to get the derivative for $w_l(m,n)$, the m^{th} “plane” of the n^{th} filter

The filter derivative



$$\frac{\partial \ell}{\partial w_l(m,n,i,j)} = \sum_{x,y} \frac{\partial \ell}{\partial z(l,n,x,y)} Y(l-1,m,x+i,y+j) = \frac{\partial \ell}{\partial z(l,n)} \otimes Y(l-1,m)$$

$\frac{\partial \ell}{\partial z(l,n,x,y)}$ must be upsampled if the stride was greater than 1 in the forward pass

If $Y(l-1,m)$ was zero padded in the forward pass, it must be zero padded for backprop

Next Up

- Advanced optimization methods