UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

Sesión 2.0

Segmentación

U-net, R-CNN, DeepLab

TRANSFORMATEC

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

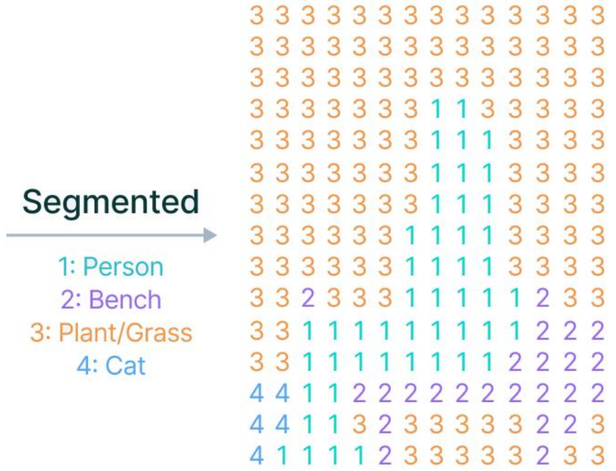
1.

Segmentación
y detección

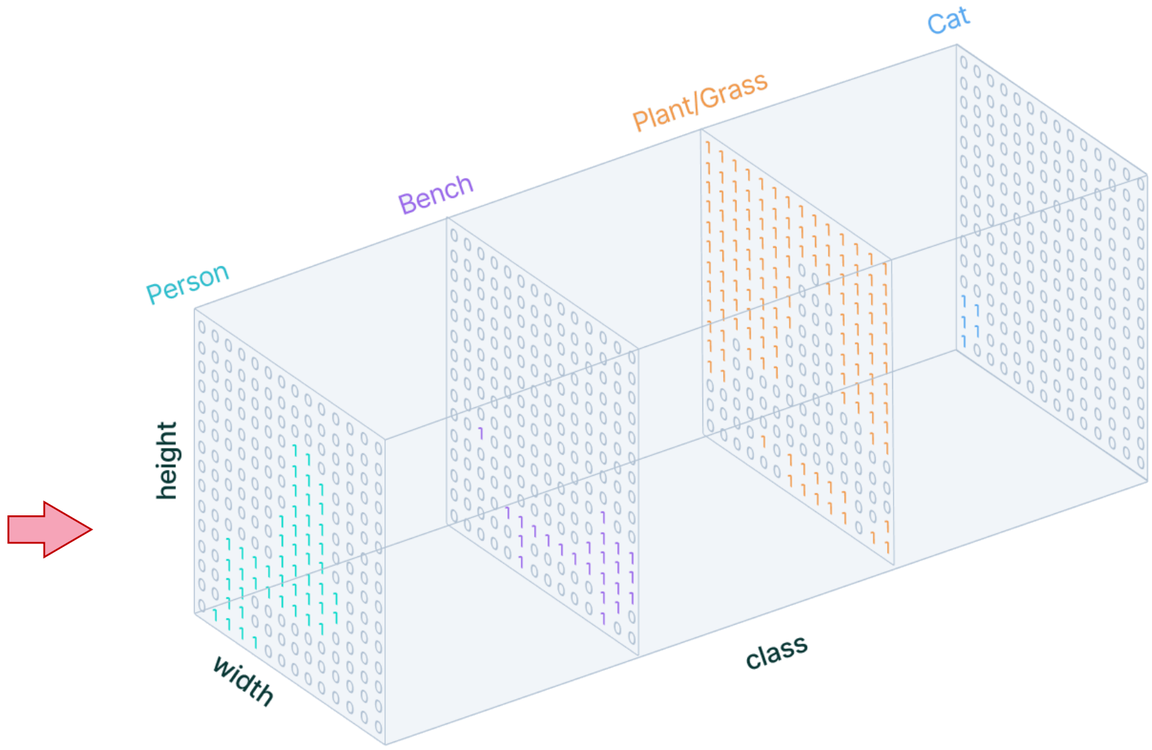
TRANSFORMATEC

Segmentación *de imágenes*

Entrada



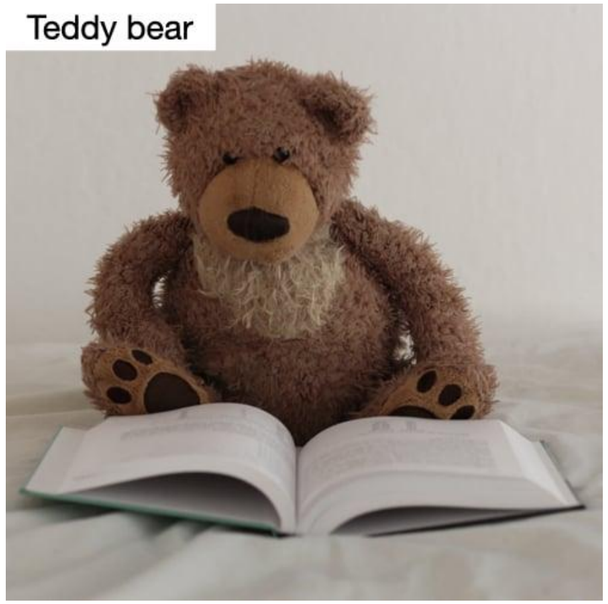
Etiquetas semánticas



One-hot

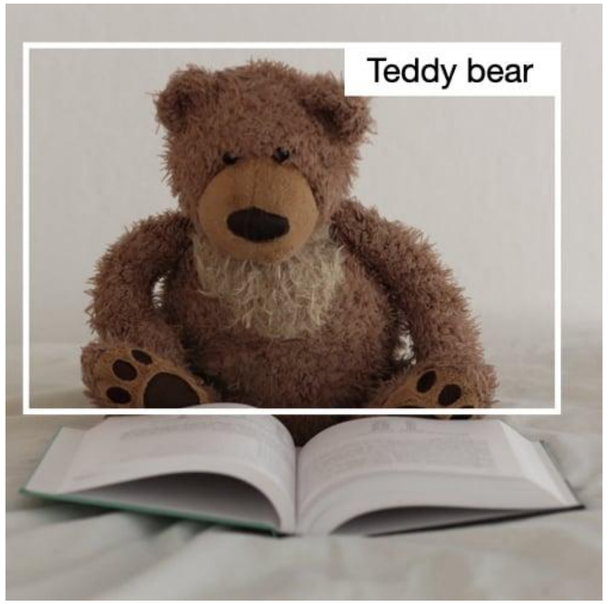
Segmentación *y detección*

Clasificación de imágenes



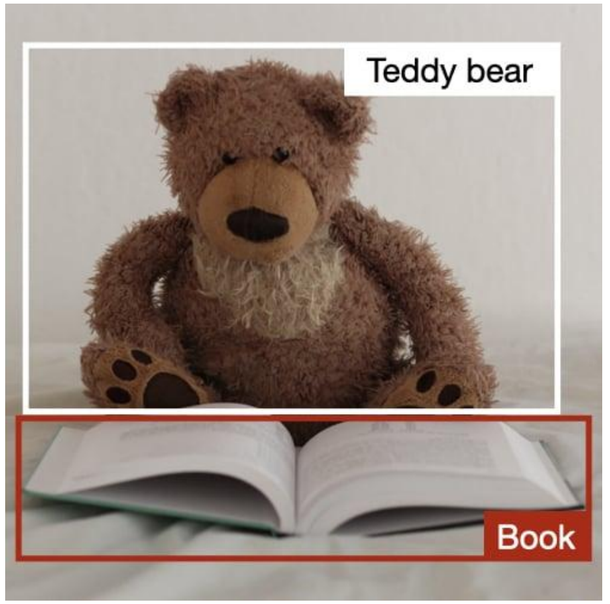
- Clasifica una imagen.
- Predice la probabilidad de que un objeto esté en una imagen.

Clasificación con localización



- Detecta un objeto en una imagen.
- Predice la probabilidad de que un objeto esté en una imagen, además de su ubicación.

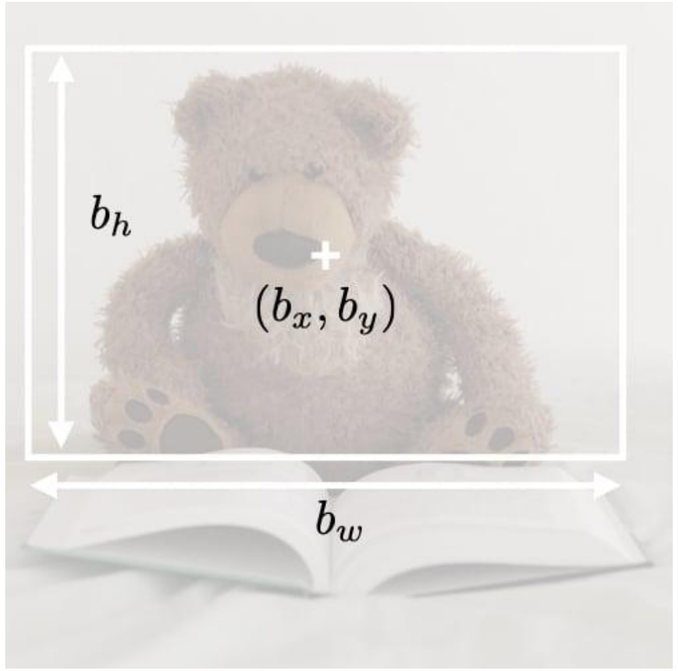
Detección



- Detecta varios objetos en una imagen.
- Predice la probabilidad de que algunos objetos estén en una imagen, además estima sus ubicaciones.

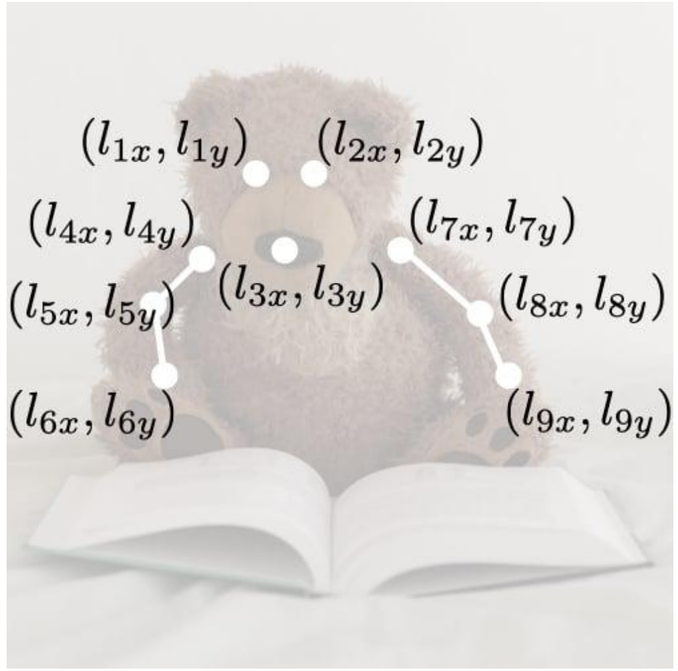
Segmentación *y detección*

Bounding box



- Detecta la parte de la imagen donde se encuentra el objeto.

Landmark

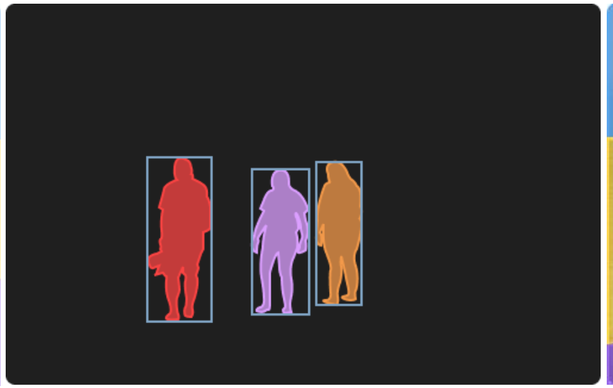


- Detecta una forma o características de un objeto (por ejemplo, ojos).
- Más granular.

Segmentación

Imagen de entrada

Semantic segmentation

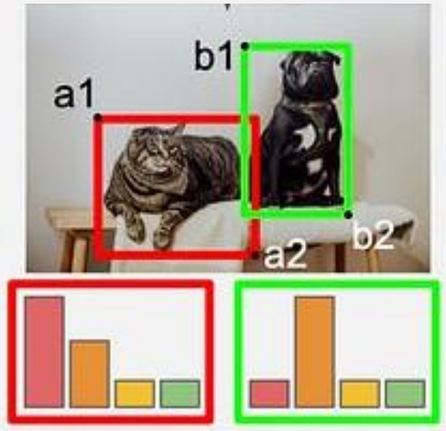


Instance segmentation

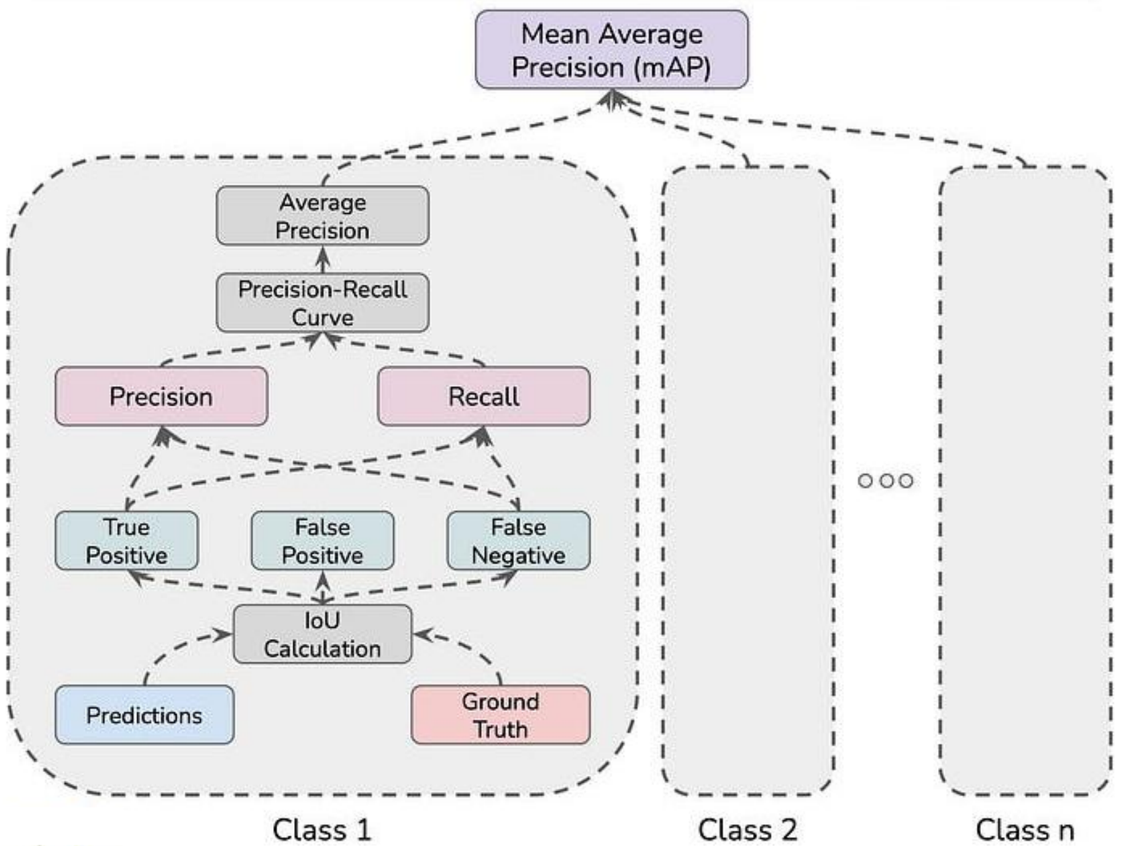
Panoptic segmentation

Mean average *precision (mAP)*

$$\text{mAP} = \frac{1}{n}\sum_{k=1}^{n} \text{AP}_k$$

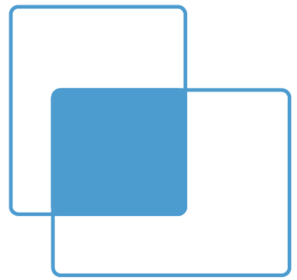


Mean average *precision (mAP)*

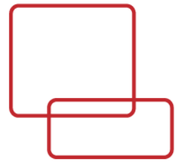


Intersection *over Union (IoU)*

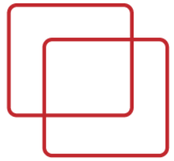
$$IOU = \frac{\text{Area of intersection}}{\text{Area of union}}$$



Examples



IOU = 0.1

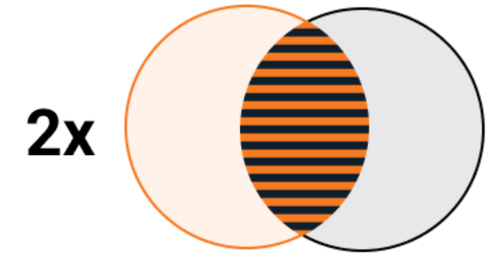


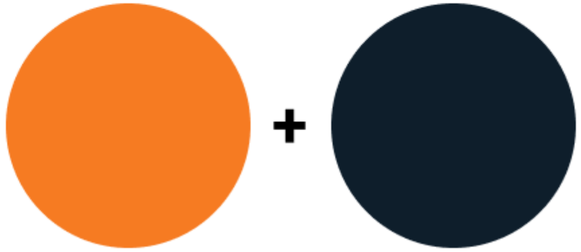
IOU = 0.3

IOU = 0.6

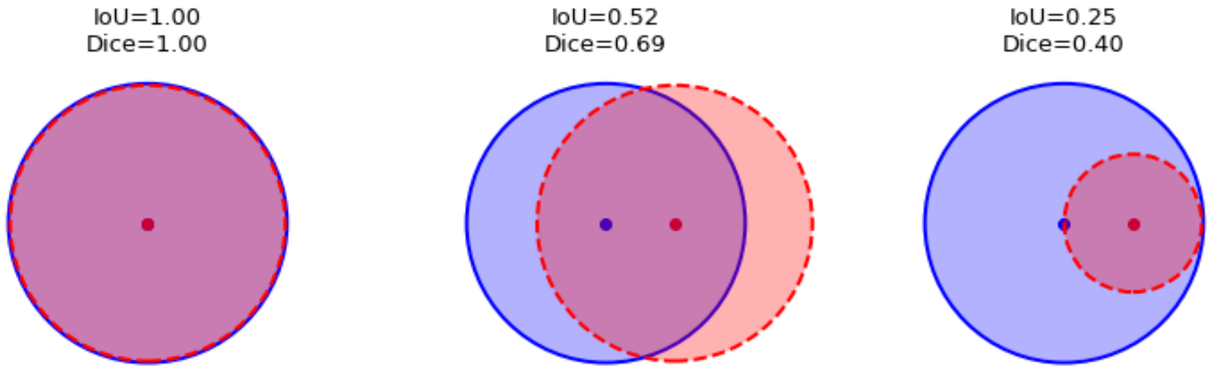
Métricas

$$\text{DICE} = \frac{2|A \cap B|}{|A| + |B|} = \frac{\text{2x}}{\quad + \quad}$$





IoU vs *DICE*



UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

2.

Métodos *Tradicionales*

TRANSFORMATEC

Threshold *Segmentation*

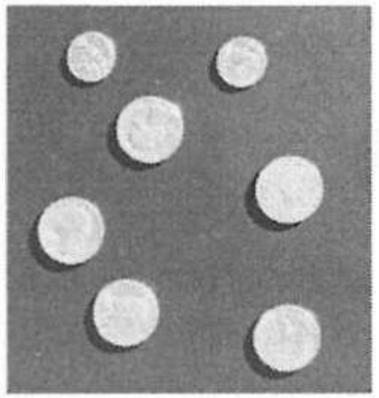
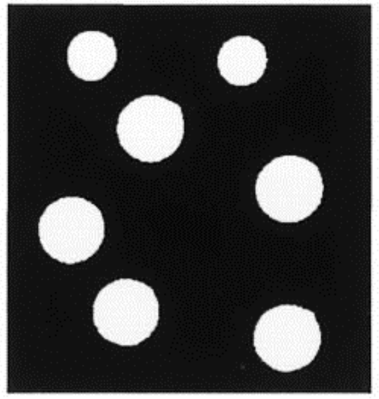
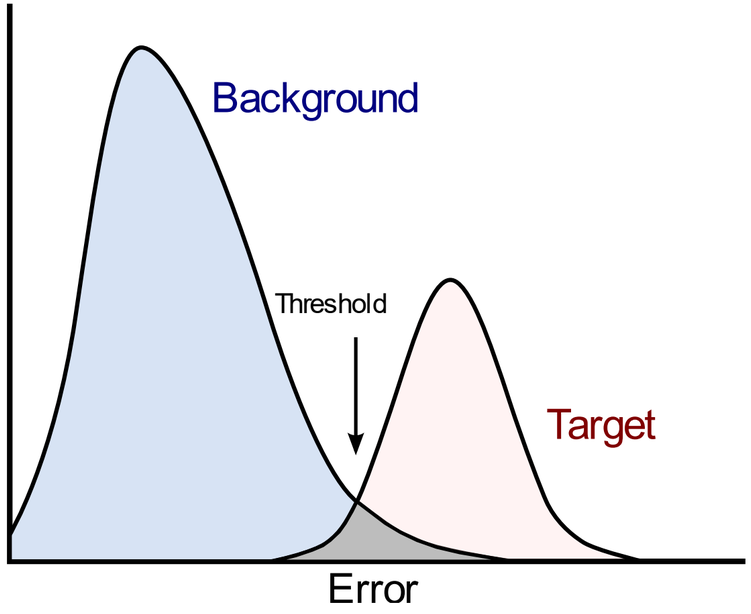


Imagen de entrada

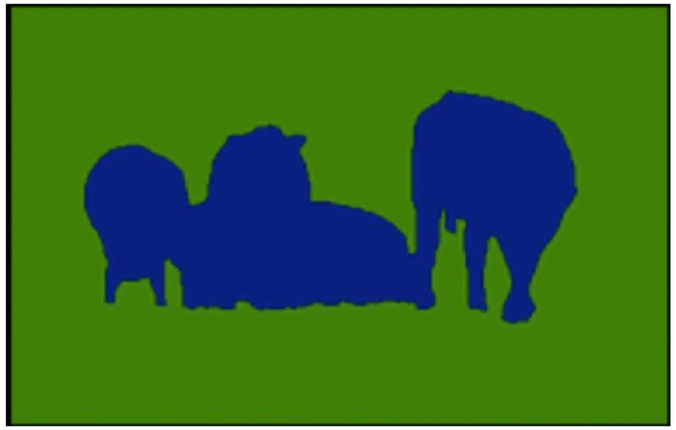


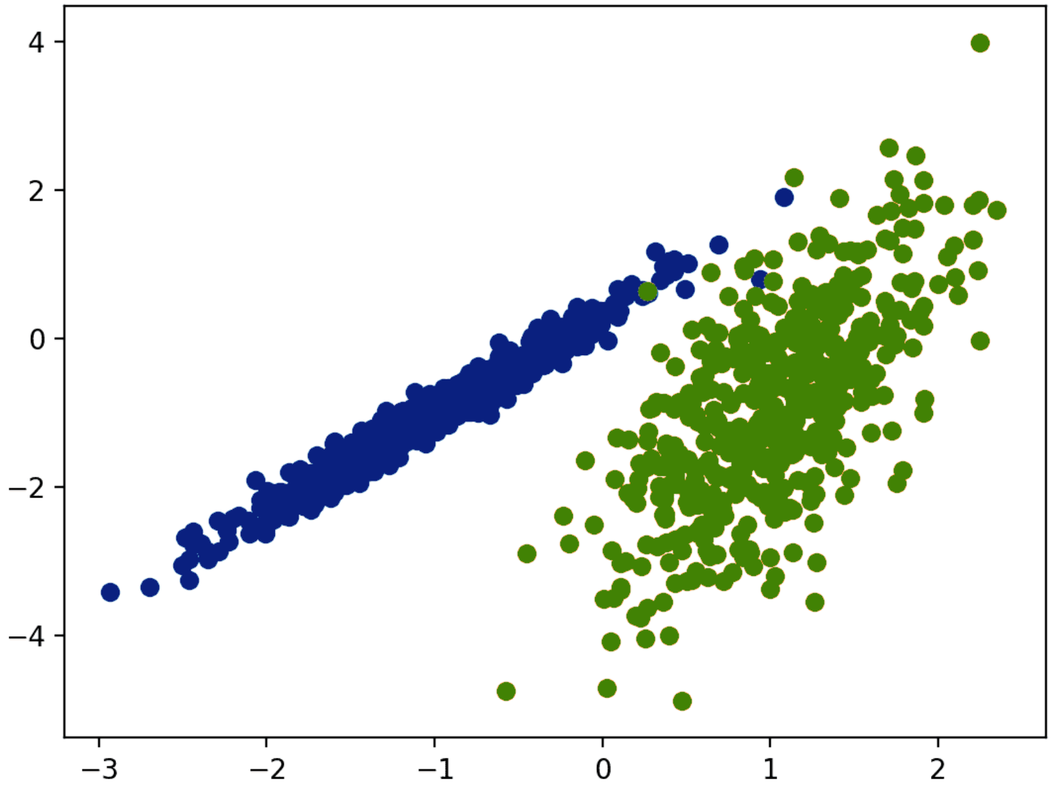
Resultado



Edge-Based *Segmentation*

Cluster-Based *Segmentation*





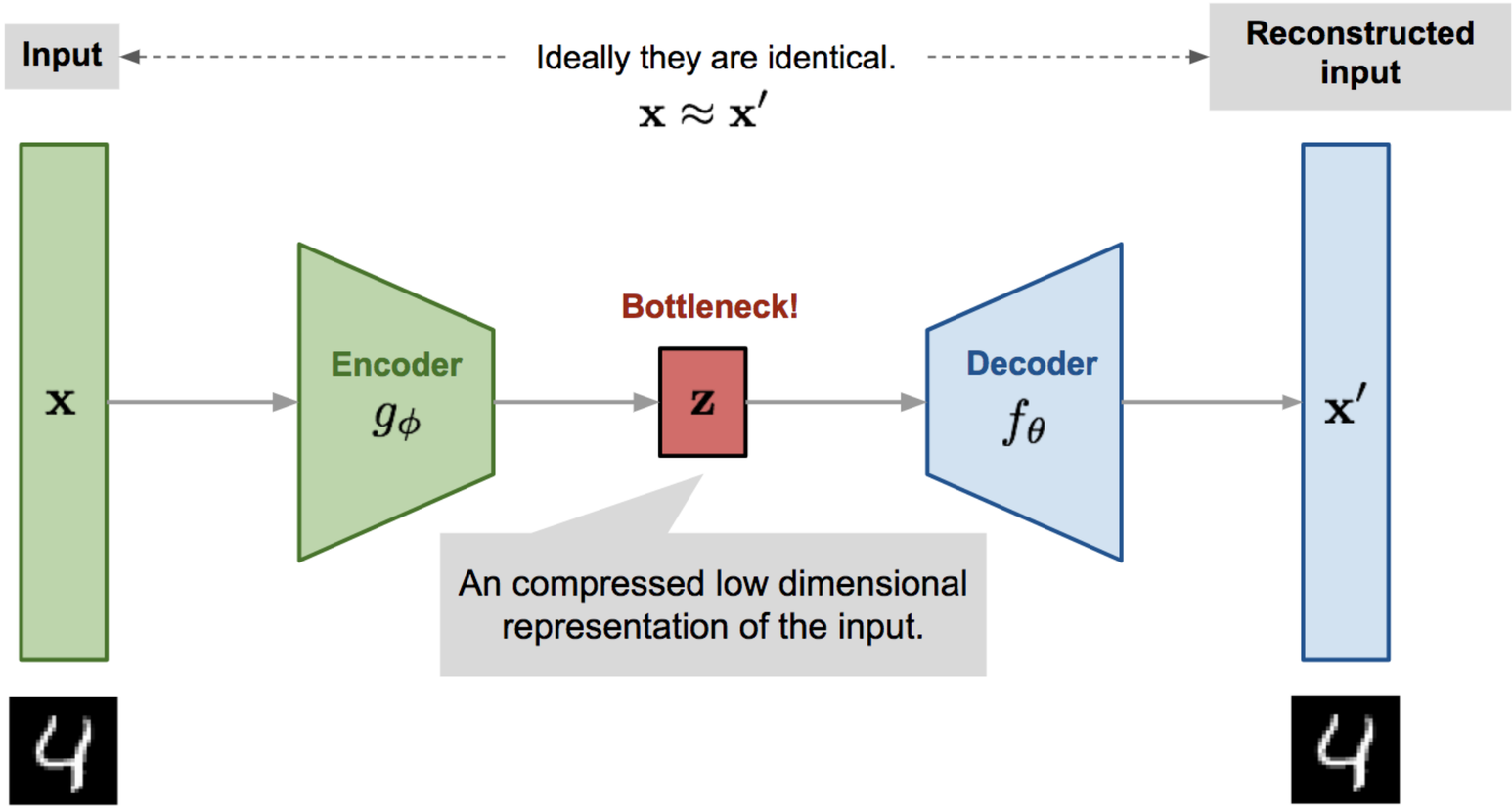
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

3.

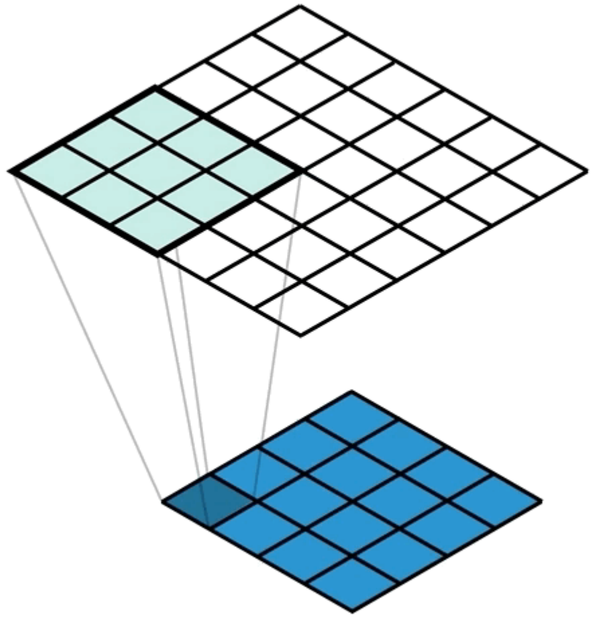
Encoder-*Decoder*

TRANSFORMATEC

Autoencoder

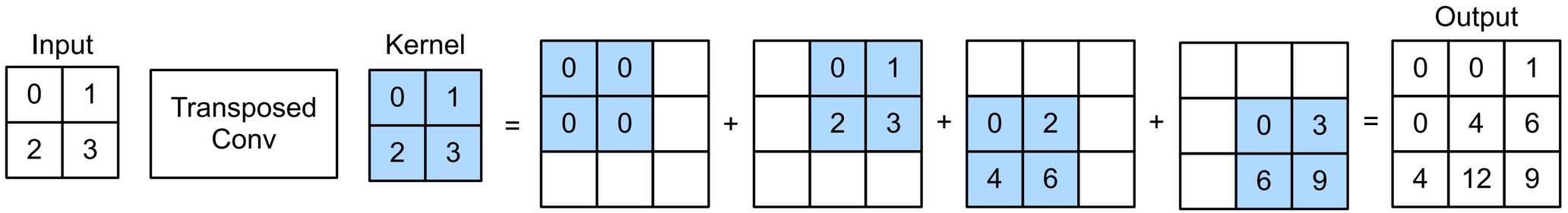


Transposed *Convolution*

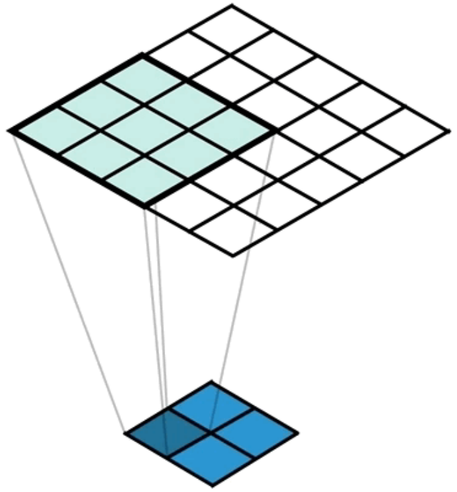


$$y(m,n) = \sum_{i,j} x(i,j) \cdot w(m-i, n-j)$$

Transposed *Convolution*



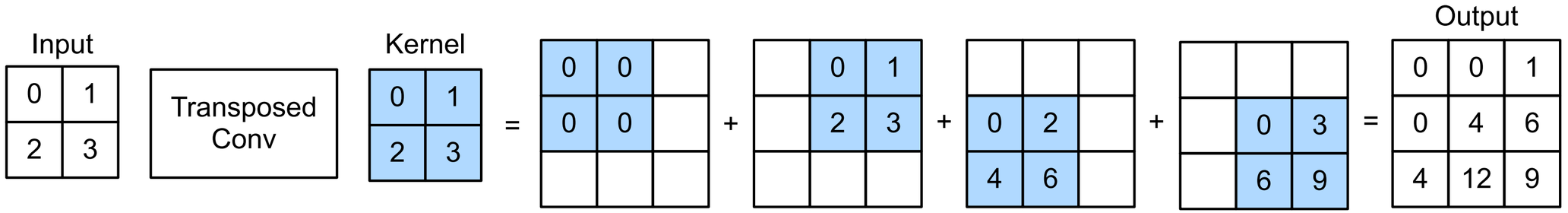
Transposed *Convolution*

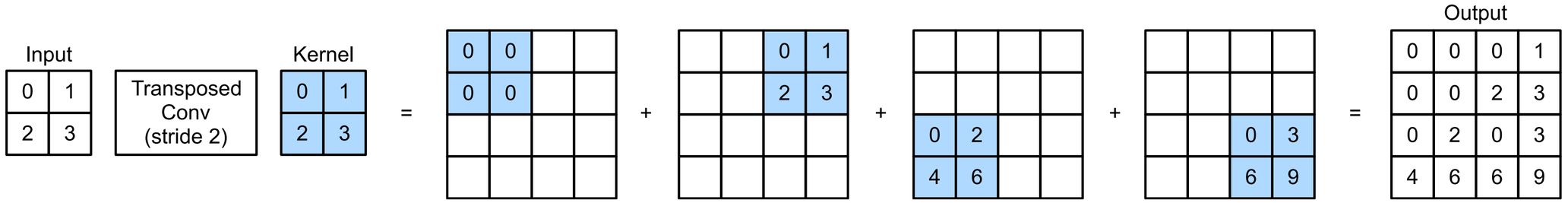


$$y(m,n) = \sum_{i,j} x(i,j) \cdot w(m - i \cdot s + p, n - j \cdot s + p)$$

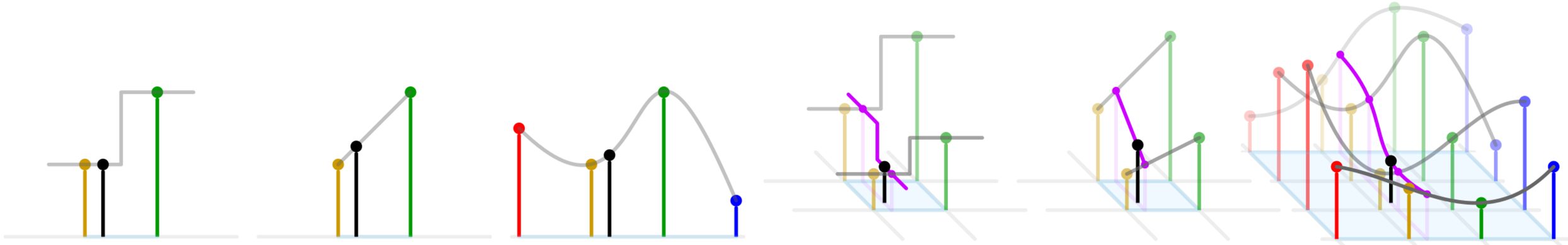
- s: stride
- p: padding

Transposed *Convolution*

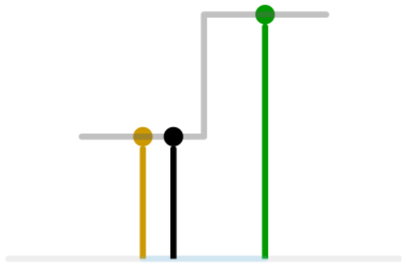




Inter*polation*

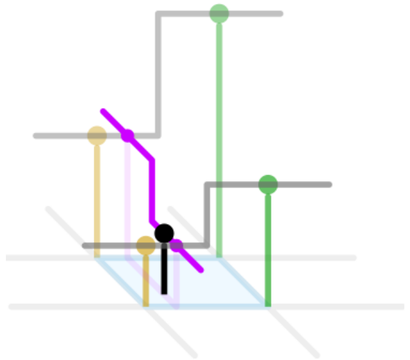


Interpolation



1D nearest-neighbour

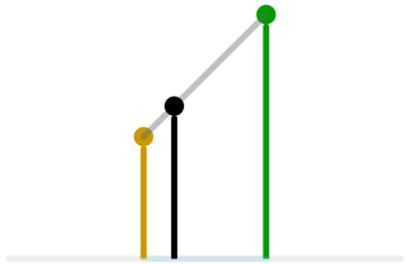
$$y_{\text{out}}(i) = y_{\text{in}}\left(\text{round}\left[\frac{i}{s}\right]\right)$$



2D nearest-neighbour

$$y_{\text{out}}(i,j) = y_{\text{in}}\left(\text{round}\left[\frac{i}{s}\right], \text{round}\left[\frac{j}{s}\right]\right)$$

Interpolation



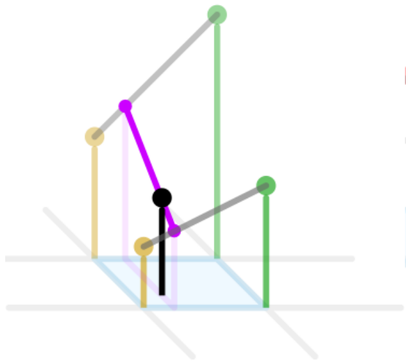
Linear

$$y_{\text{out}}(i) = (1-a)\cdot y_{\text{in}}(x_1) + a\cdot y_{\text{in}}(x_2)$$

Donde:

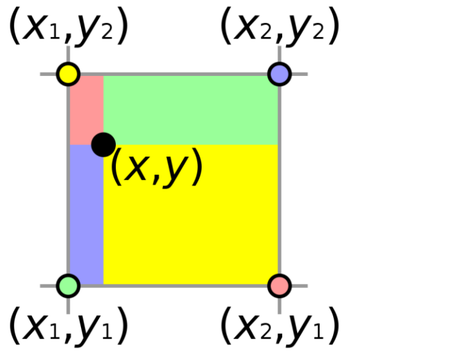
$x_1 = \lfloor i \cdot s \rfloor$ $\quad$ $x_2 = x_1 + 1$

$a = i \cdot s - x_1$



Bilinear

$$\begin{aligned} y_{\text{out}}(i,j) = {} & (1-a)(1-b)\cdot y_{\text{in}}(x_1,y_1) \\ & +a(1-b)\cdot y_{\text{in}}(x_2,y_1) \\ & +(1-a)b\cdot y_{\text{in}}(x_1,y_2) \\ & +ab\cdot y_{\text{in}}(x_2,y_2) \end{aligned}$$



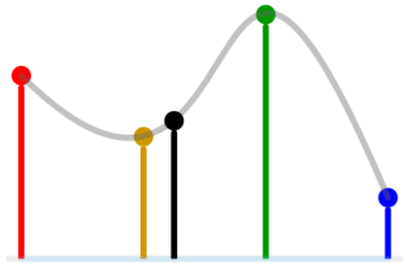
Donde:

$x_1 = \lfloor i \cdot s_h \rfloor$ $\quad$ $x_2 = x_1 + 1$

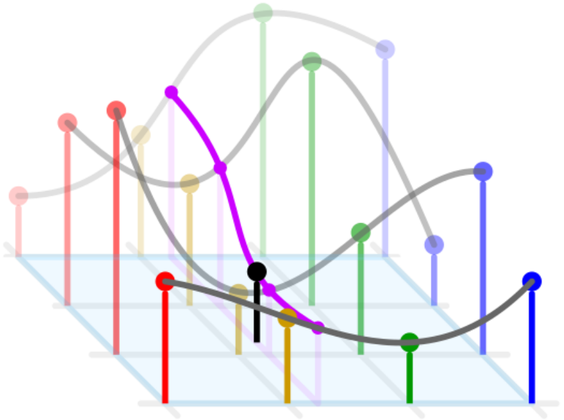
$y_1 = \lfloor f \cdot s_w \rfloor$ $\quad$ $y_2 = y_1 + 1$

$a = i \cdot s_h - x_1$ $\quad$ $b = j \cdot s_w - y_1$

Interpolation



Cubic

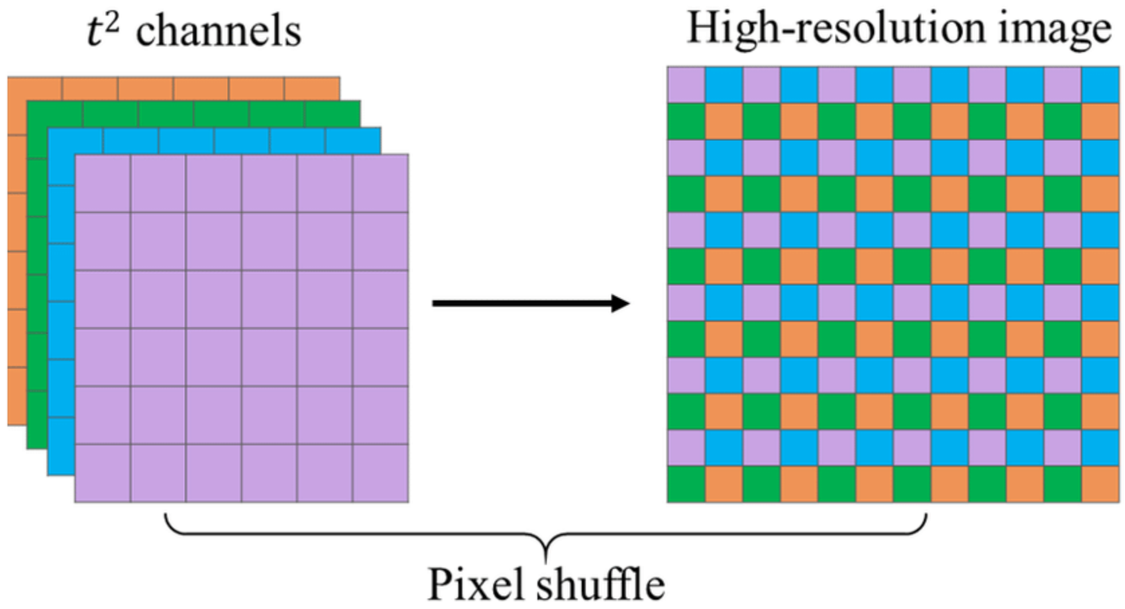


Bicubic

$$y_{\text{out}}(i) = \sum_{k=-1}^{2} c_k \cdot y_{\text{in}}(x_k)$$

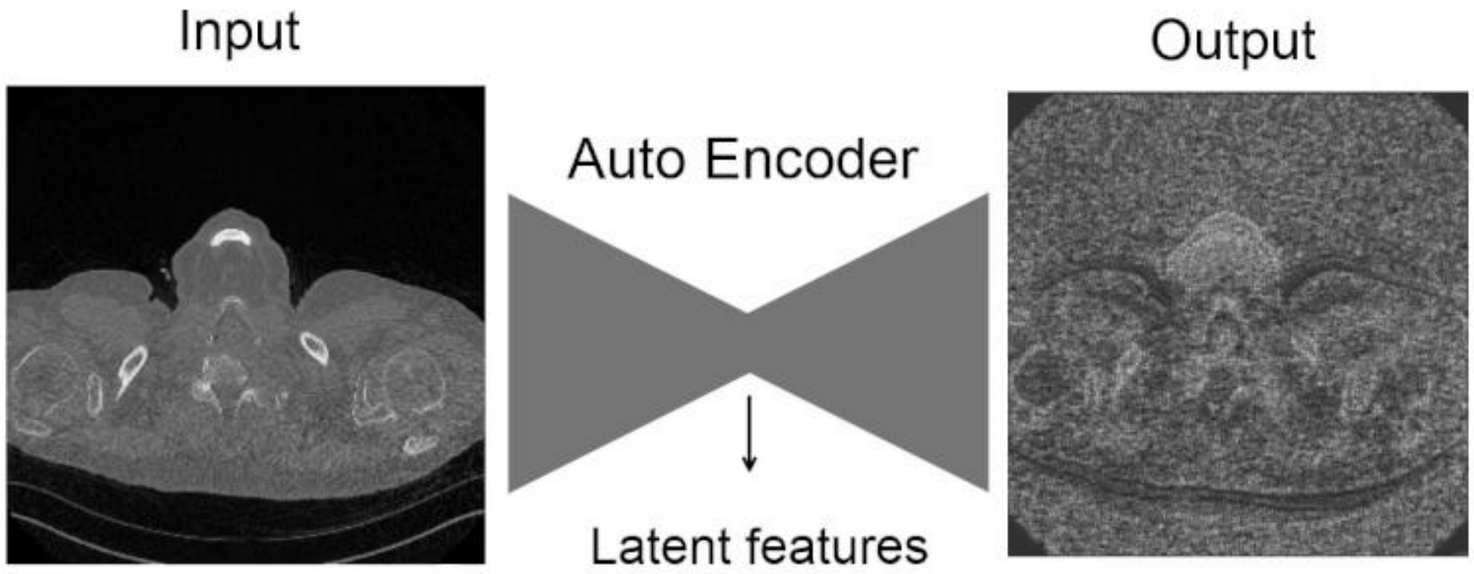
$$y_{\text{out}}(i,j) = \sum_{m=-1}^{2} \sum_{n=-1}^{2} c_{mn} \cdot y_{\text{in}}(x_m, y_n)$$

Pixel *Shuffle*

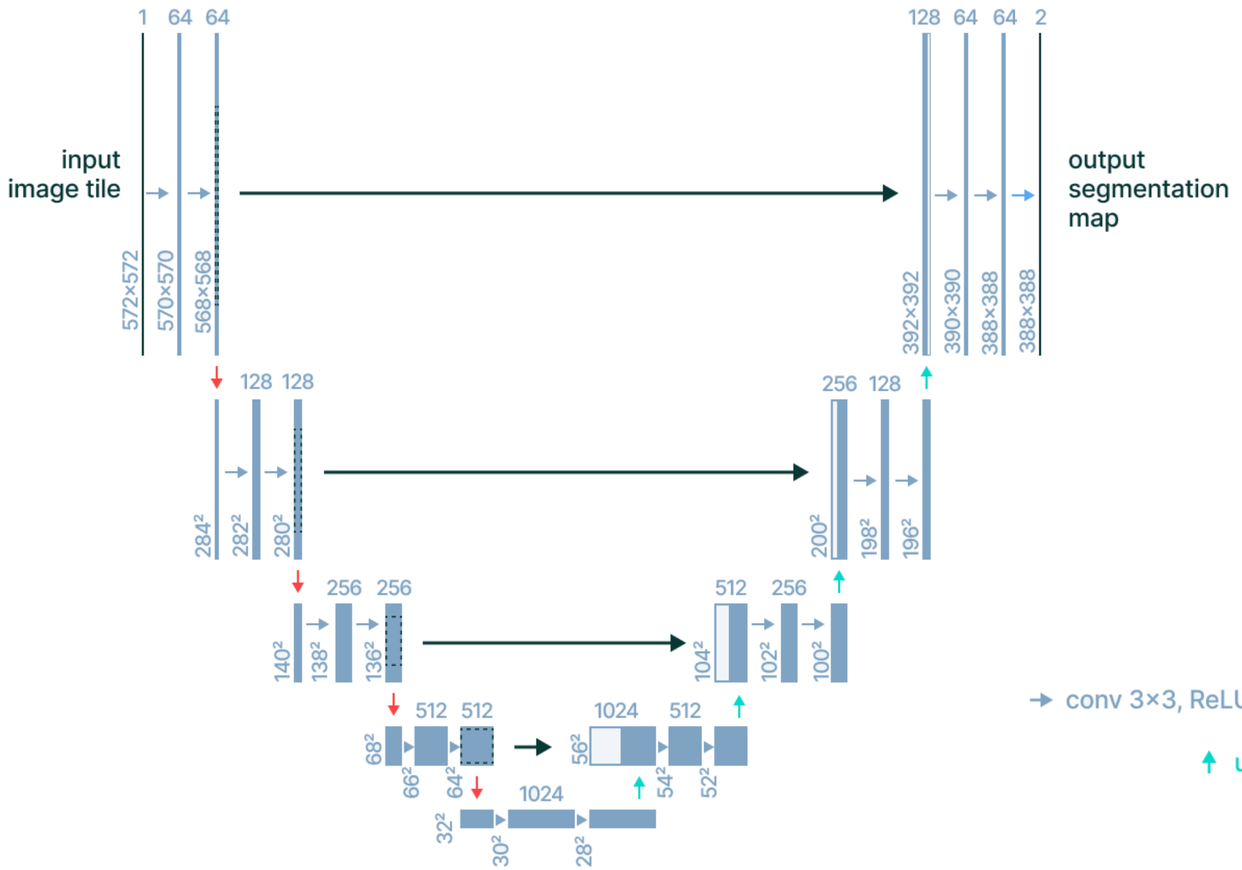


Wenzhe Shi et al. (2016). "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1874-1883).

Auto*encoder*

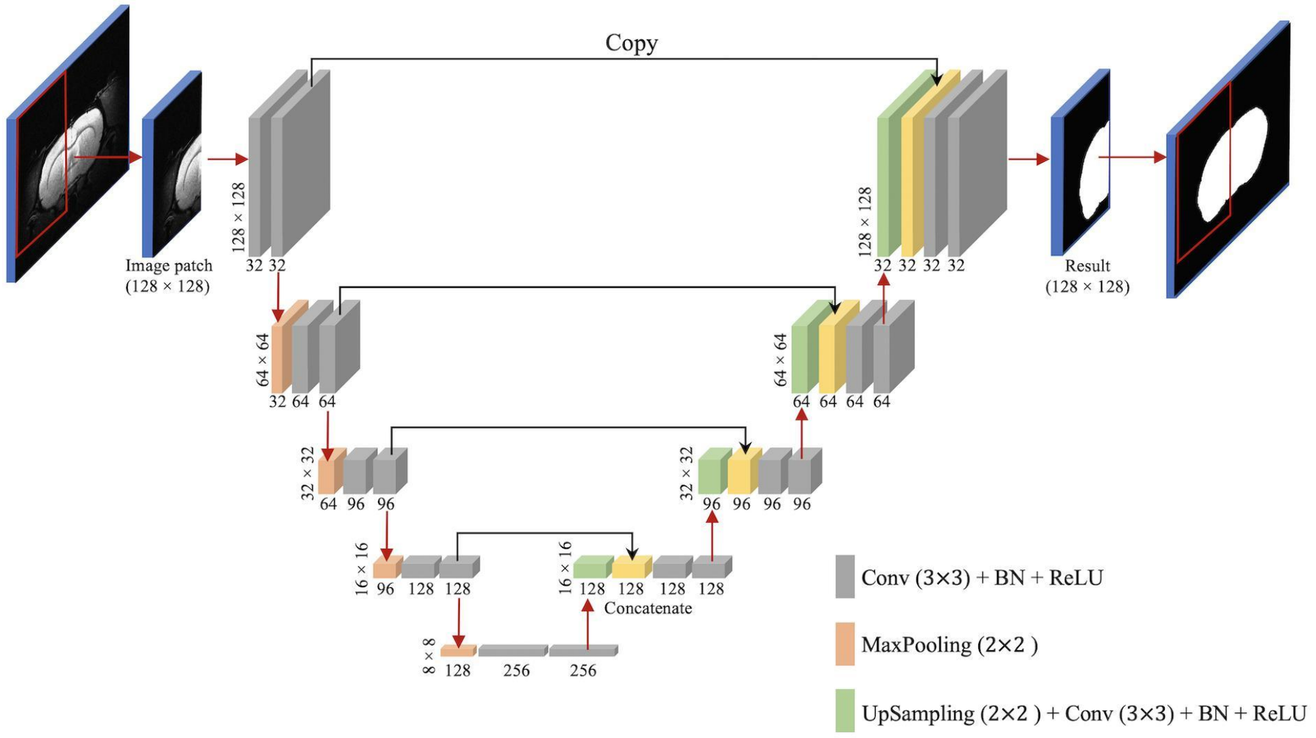


U-*Net*



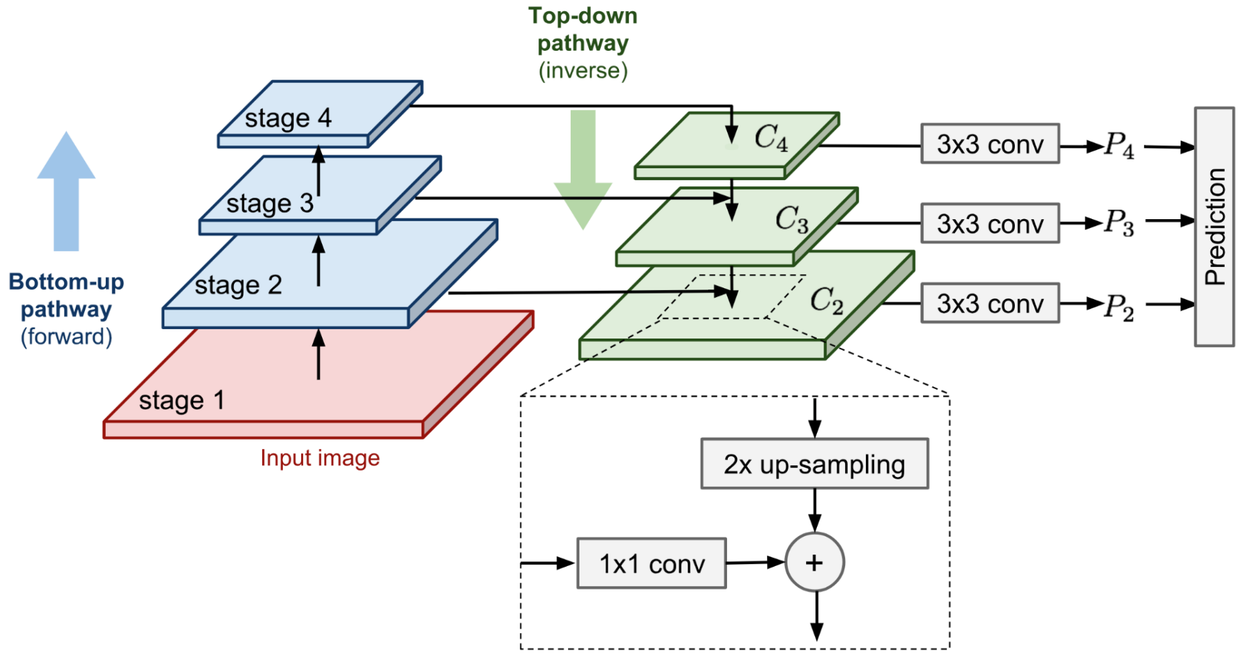
Olaf Ronneberger et al. (2015) "U-Net: Convolutional Networks for Biomedical Image Segmentation". International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

U-*Net*



Olaf Ronneberger et al. (2015) "U-Net: Convolutional Networks for Biomedical Image Segmentation". International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Feature Pyramid *Network (FPN)*

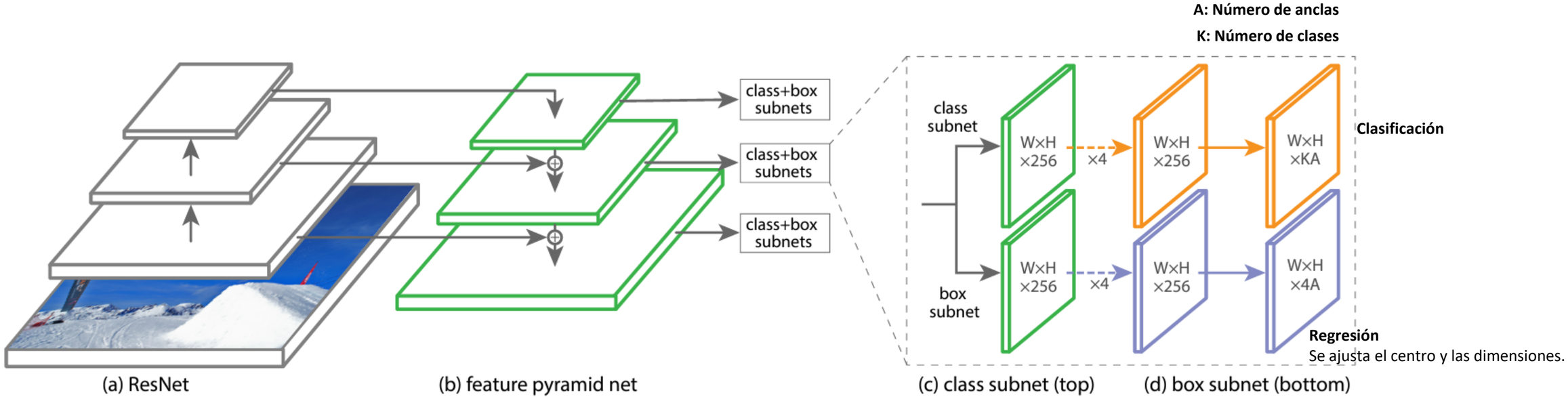


Clasificación

Regresión
Se ajusta el centro y las dimensiones.

Tsung-Yi Lin et al. (2016) "Feature Pyramid Networks for Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

RetinaNet



Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\mathrm{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases}$$

Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\mathrm{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases} \qquad p_t = \begin{cases} p, & y = 1 \\ 1-p, & \text{otro caso} \end{cases}$$

Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\text{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases} \qquad p_t = \begin{cases} p, & y = 1 \\ 1-p, & \text{otro caso} \end{cases} \qquad p_t \to 1$$

La clasificación es correcta

Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\mathrm{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases} \qquad p_t = \begin{cases} p, & y = 1 \\ 1-p, & \text{otro caso} \end{cases} \qquad p_t \to 1$$

La clasificación es correcta

Resumiendo: $\mathrm{CE} = -\log p_t$

Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection". IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\mathrm{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases} \qquad p_t = \begin{cases} p, & y = 1 \\ 1-p, & \text{otro caso} \end{cases} \qquad p_t \to 1$$

La clasificación es correcta

Resumiendo: $\mathrm{CE} = -\log p_t$

Asignamos pesos basados en p_t : $\mathrm{FL} = -(1-p_t)^{\gamma} \log p_t$

Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

RetinaNet

Cross-entropy

$$\mathrm{CE} = \begin{cases} -\log p\,, & y = 1 \\ -\log(1-p)\,, & \text{otro caso} \end{cases} \qquad p_t = \begin{cases} p, & y = 1 \\ 1-p, & \text{otro caso} \end{cases} \qquad p_t \to 1$$

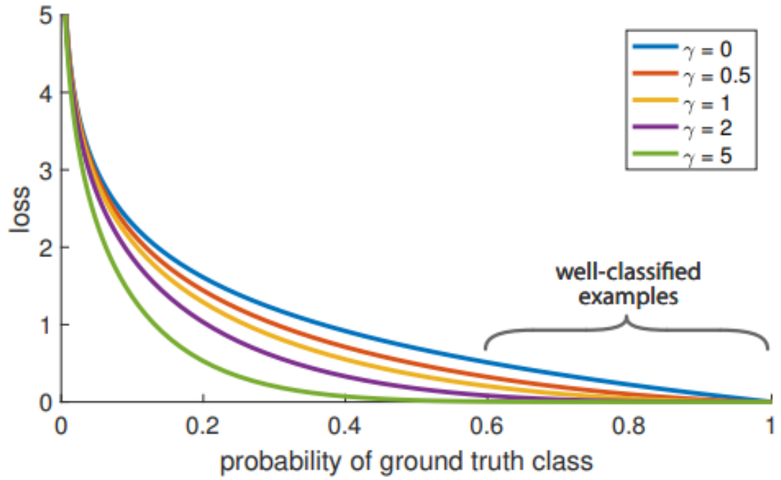
La clasificación es correcta

Resumiendo: $\mathrm{CE} = -\log p_t$

Asignamos pesos basados en p_t : $\mathrm{FL} = -(1-p_t)^{\gamma} \log p_t$

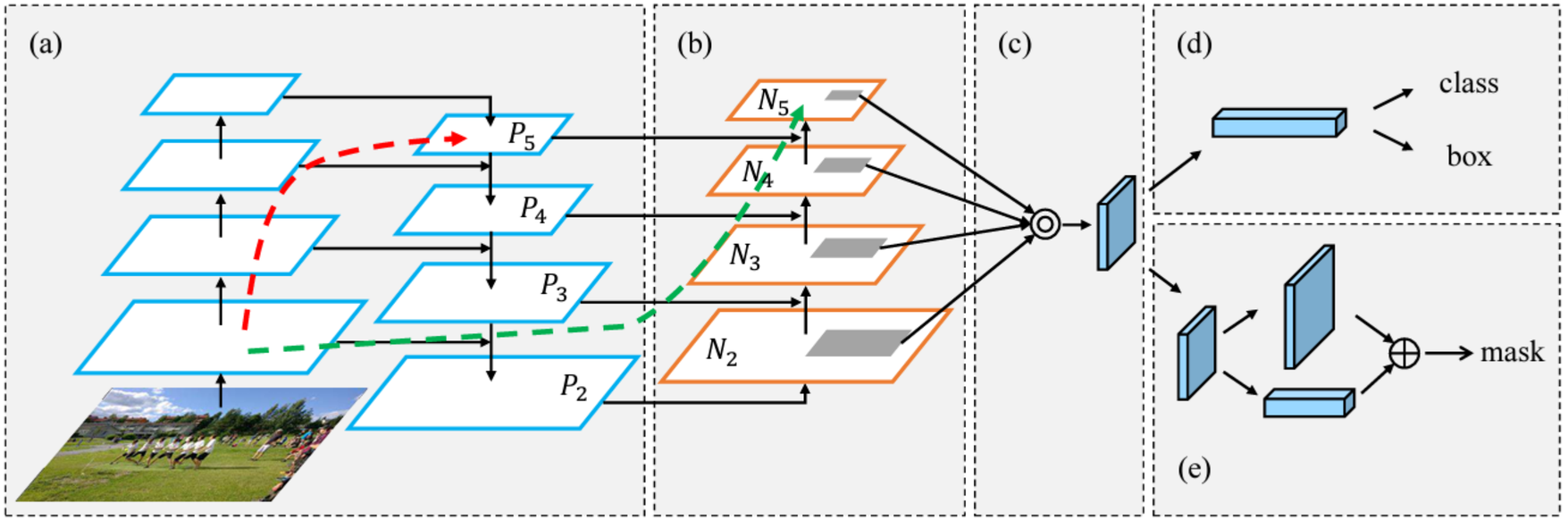
$(1-p_t) \to 0$ **Caso conocido**

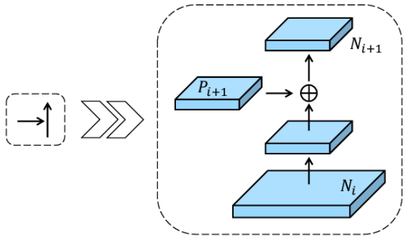
$(1-p_t) \to 1$ **Caso desconocido**

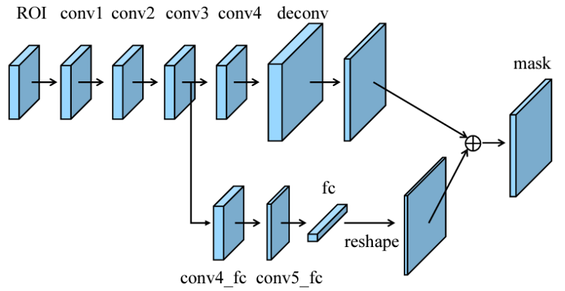


Tsung-Yi Lin et al. (2018) "Focal Loss for Dense Object Detection".
IEEE International Conference on Computer Vision ICCV.

PANet







FPN

Bottom-up path augmentation

Adaptive feature pooling

Fully-connected fusion

Shu Liu et al. (2018) "Path Aggregation Network for Instance Segmentation".
Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8759-8768).

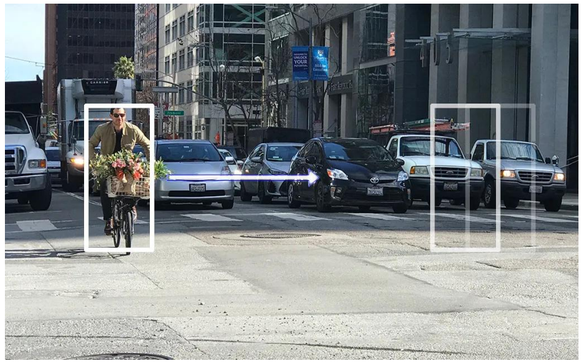
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

4.

Mask *R-CNN*

TRANSFORMATEC

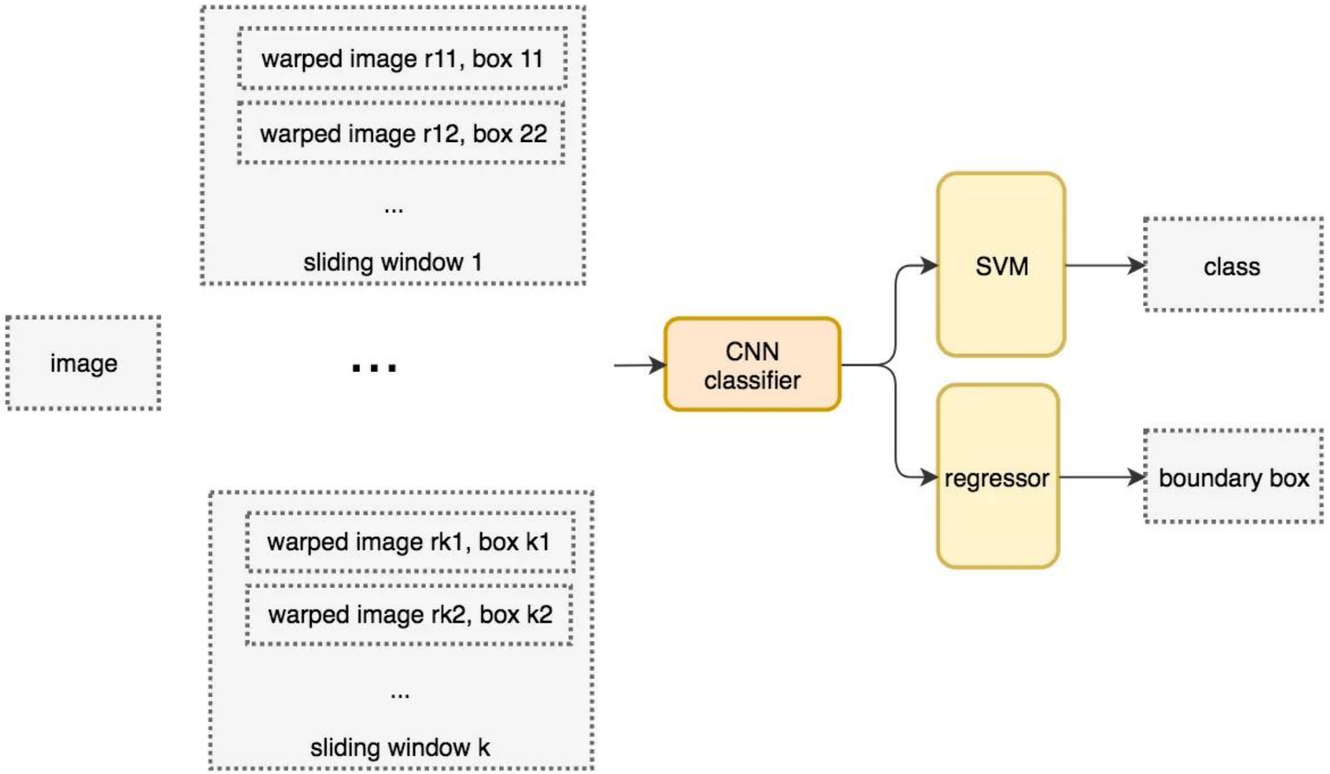
Sliding-window *detectors*



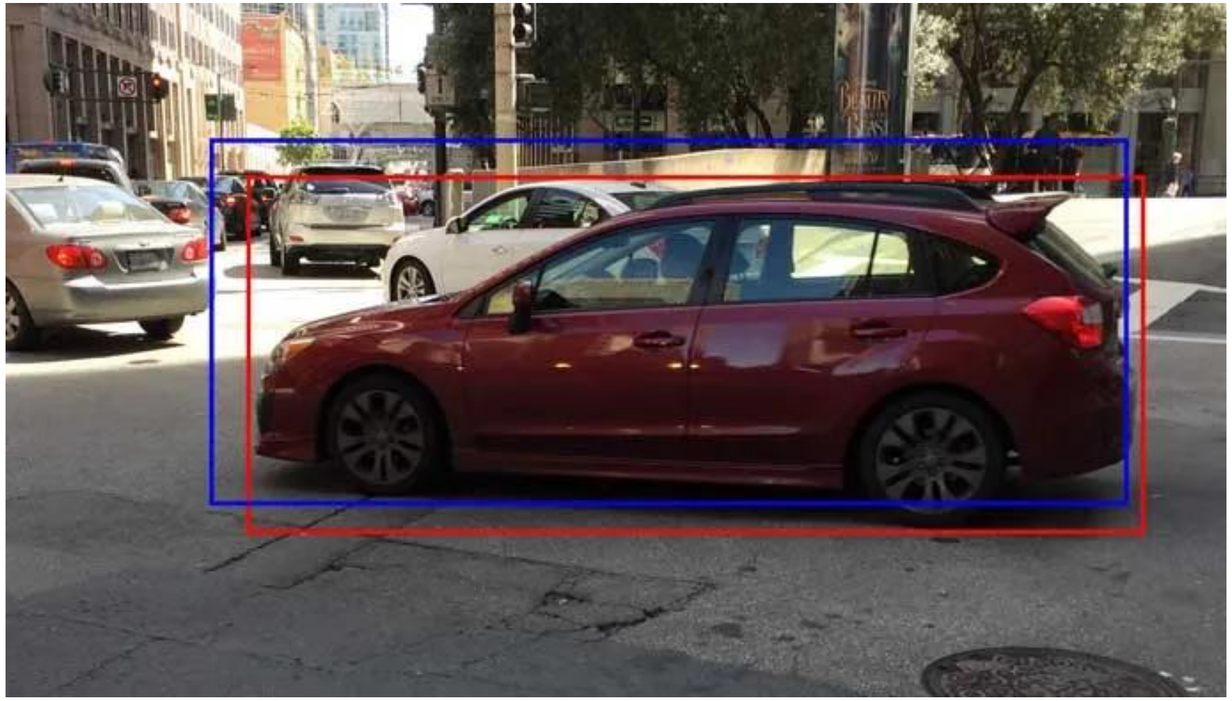
Sliding window

Warp

Sliding-window *detectors*



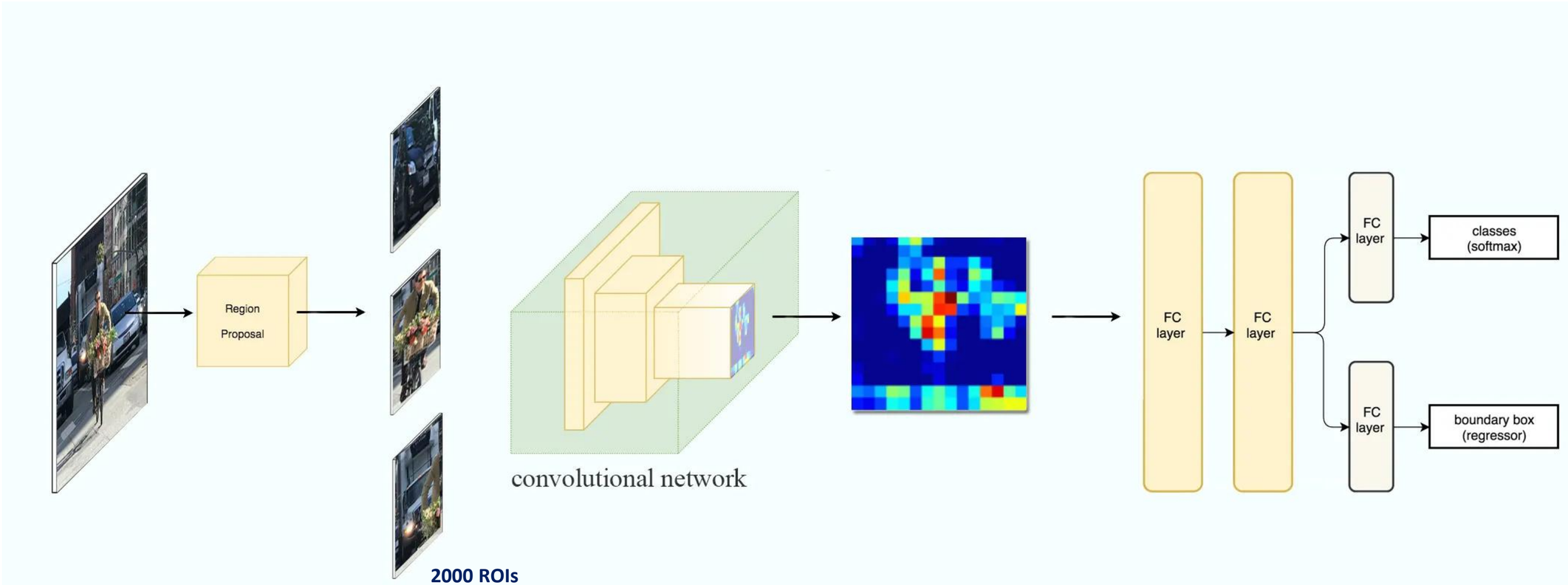
Sliding-window *detectors*



Selective *Search*

Koen E. A. van de Sande, et al. (2011) "Segmentation as Selective Search for Object Recognition". 2011 International Conference on Computer Vision (pp. 1879-1886). IEEE.

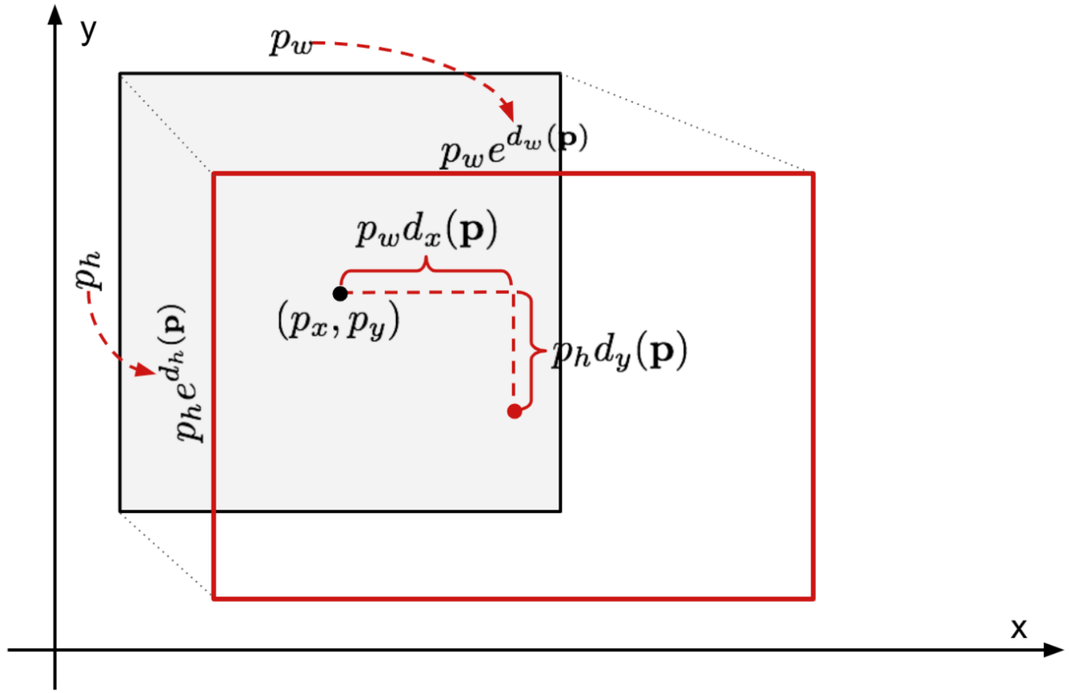
R-CNN



Ross Girshick, et al. (2014) "Rich feature hierarchies for accurate object detection and semantic segmentation". Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.

R-CNN

Bounding Box Regression

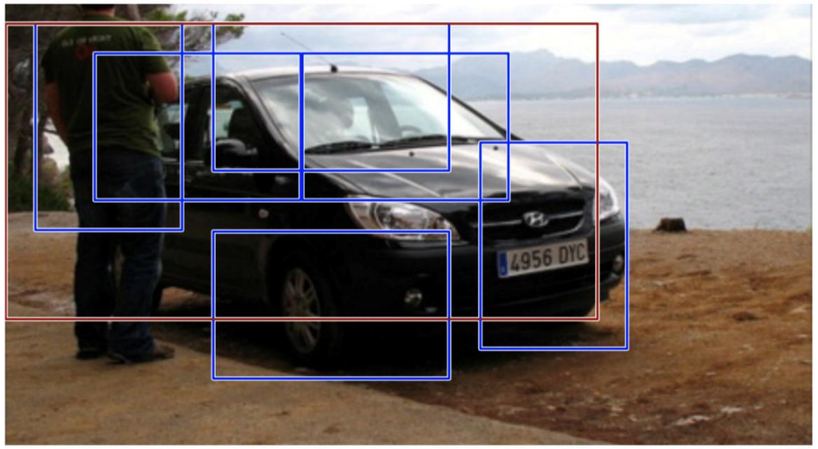


Ross Girshick, et al. (2014) "Rich feature hierarchies for accurate object detection and semantic segmentation". Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.

R-CNN

Non-Maximum Suppression

1. Los cuadros delimitadores se ordenan de acuerdo a sus puntuaciones, de mayor a menor.
2. Se selecciona el ROI con la puntuación más alta y se designa como una detección real.
3. Se eliminan todos los ROIs restantes que tienen IoU altos con los ROIs previamente seleccionados.



Before non-max suppression

After non-max suppression

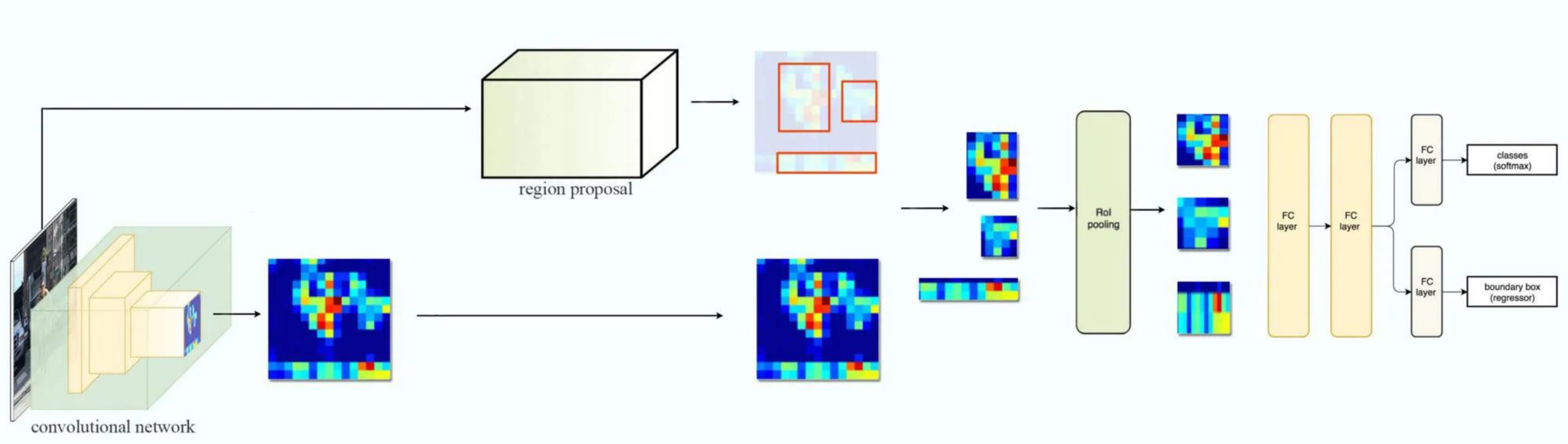
Ross Girshick, et al. (2014) "Rich feature hierarchies for accurate object detection and semantic segmentation". Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.

R-CNN

R-CNN es lento en entrenamiento e inferencia

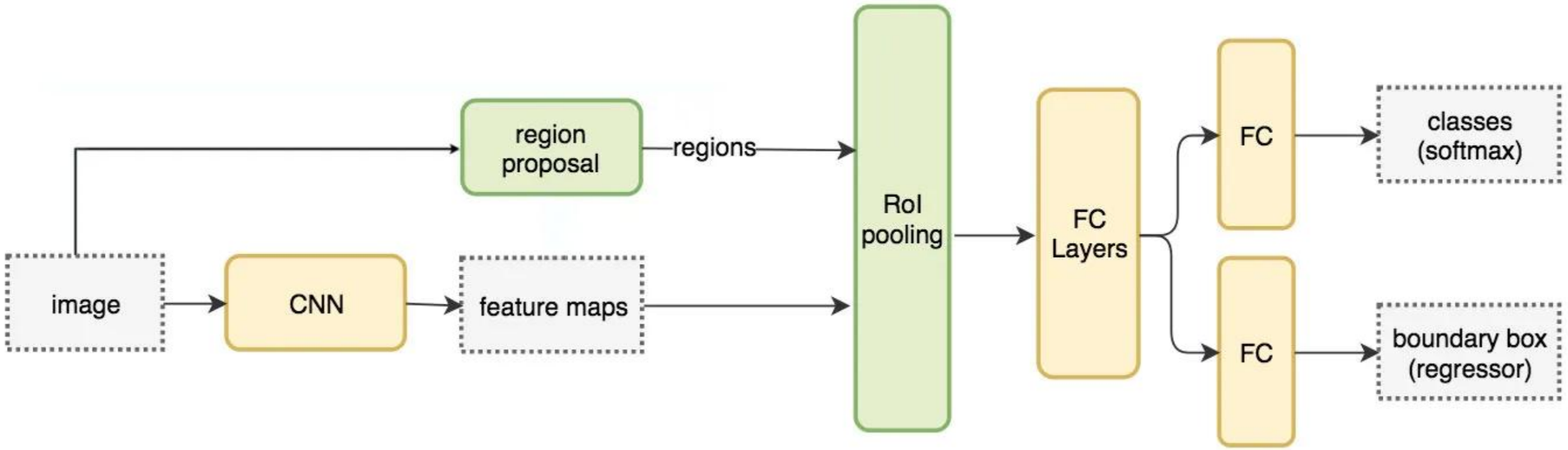
Ross Girshick, et al. (2014) "Rich feature hierarchies for accurate object detection and semantic segmentation". Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.

Fast R-*CNN*



Ross Girshick (2015) "Fast R-CNN". Proceedings of the IEEE international conference on computer vision. 2015. p. 1440-1448.

Fast R-*CNN*



Ross Girshick (2015) "Fast R-CNN". Proceedings of the IEEE international conference on computer vision. 2015. p. 1440-1448.

Fast R-*CNN*

ROI pooling

0.1	0.3	0.2	0.3	0.2	0.6	0.8	0.9
0.4	0.5	0.1	0.4	0.7	0.1	0.4	0.3
0.2	0.1	0.3	0.8	0.6	0.2	0.1	0.1
0.4	0.6	0.2	0.1	0.3	0.6	0.1	0.2
0.1	0.8	0.3	0.3	0.5	0.3	0.3	0.3
0.2	0.9	0.4	0.5	0.1	0.1	0.1	0.2
0.3	0.1	0.8	0.6	0.3	0.3	0.6	0.5
0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.2

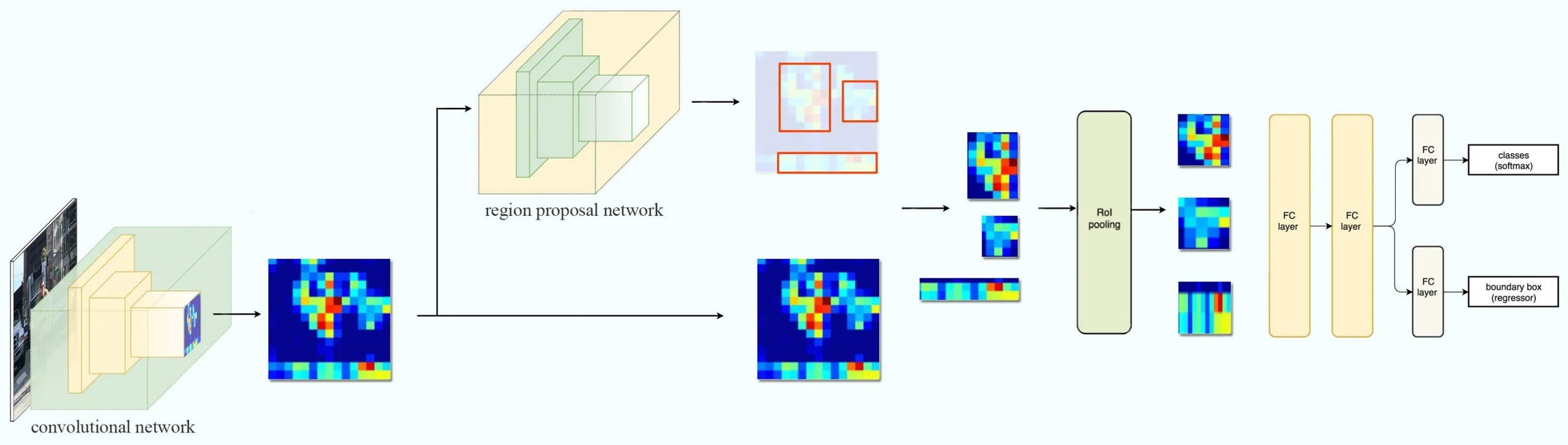
0.1	0.3	0.2	0.3	0.2	0.6	0.8	0.9
0.4	0.5	0.1	0.4	0.7	0.1	0.4	0.3
0.2	0.1	0.3	0.8	0.6	0.2	0.1	0.1
0.4	0.6	0.2	0.1	0.3	0.6	0.1	0.2
0.1	0.8	0.3	0.3	0.5	0.3	0.3	0.3
0.2	0.9	0.4	0.5	0.1	0.1	0.1	0.2
0.3	0.1	0.8	0.6	0.3	0.3	0.6	0.5
0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.2

Buscamos el máximo valor

0.8	0.6
0.9	0.6

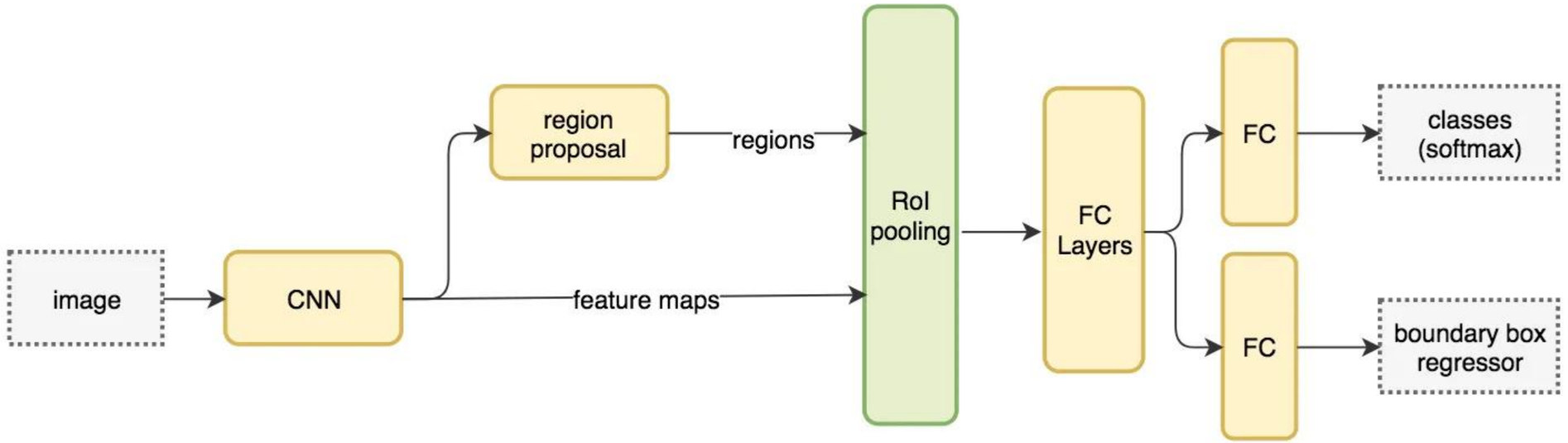
Ross Girshick (2015) "Fast R-CNN".
Proceedings of the IEEE international conference on computer vision. 2015. p. 1440-1448.

Faster *R-CNN*



Shaoqing Ren et al. (2015) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". Advances in neural information processing systems, 2015, vol. 28.

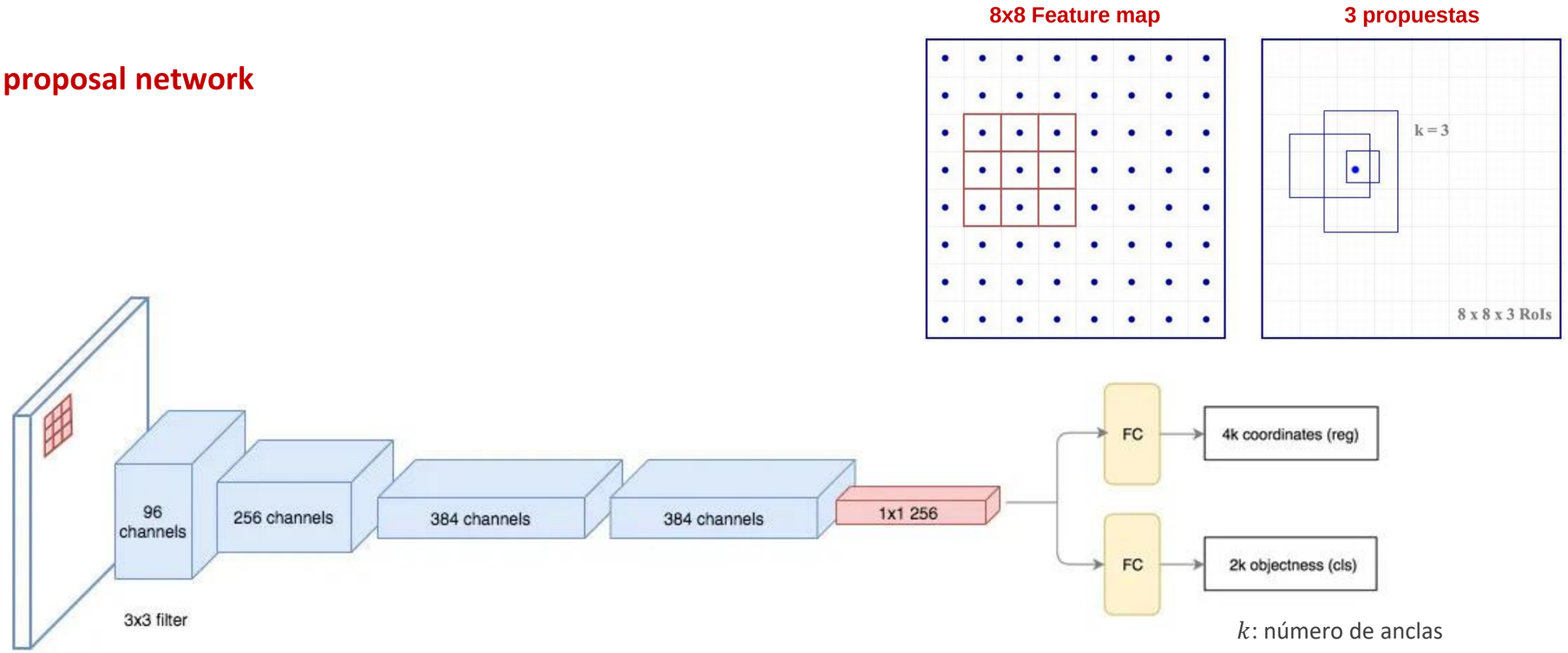
Faster *R-CNN*



Shaoqing Ren et al. (2015) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". Advances in neural information processing systems, 2015, vol. 28.

Faster *R-CNN*

Region proposal network

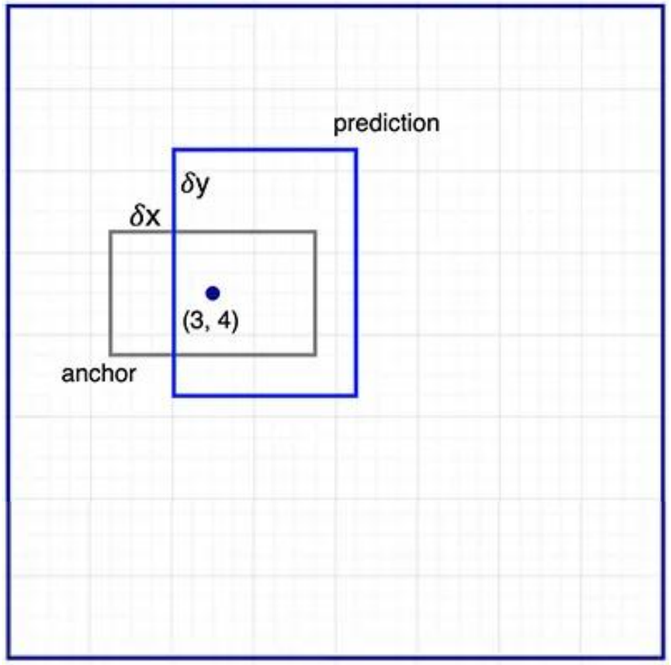


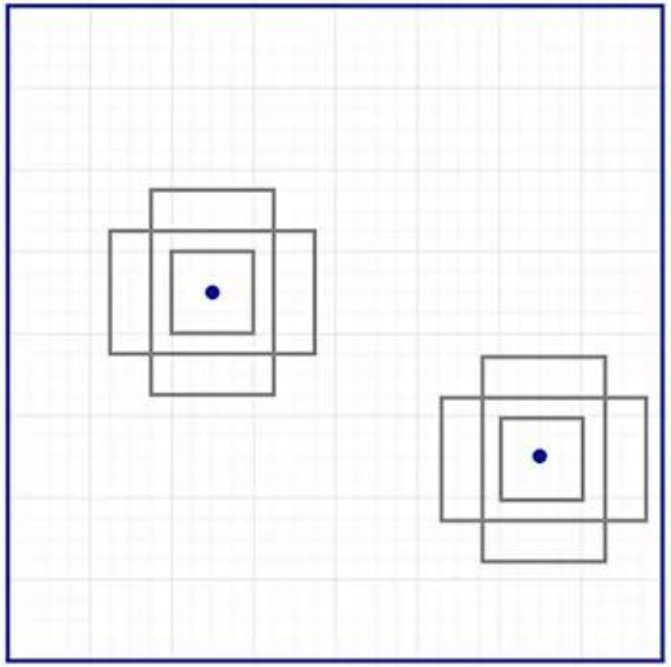
k: número de anclas

Shaoqing Ren et al. (2015) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". Advances in neural information processing systems, 2015, vol. 28.

Faster *R-CNN*

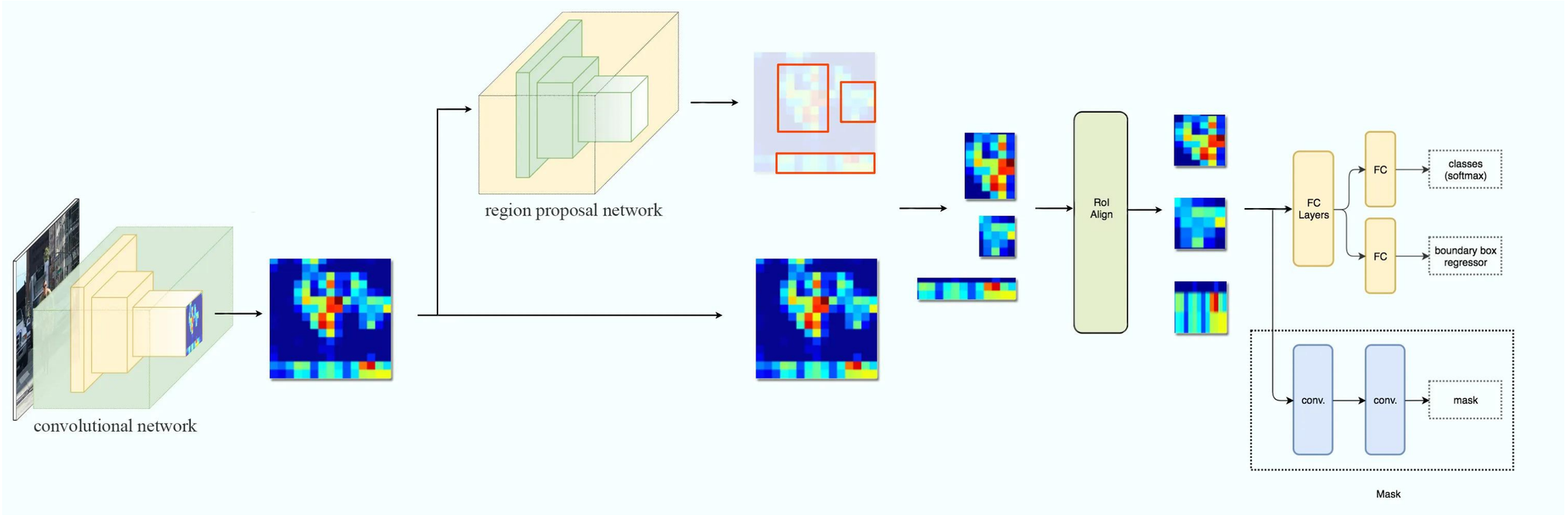
Region proposal network





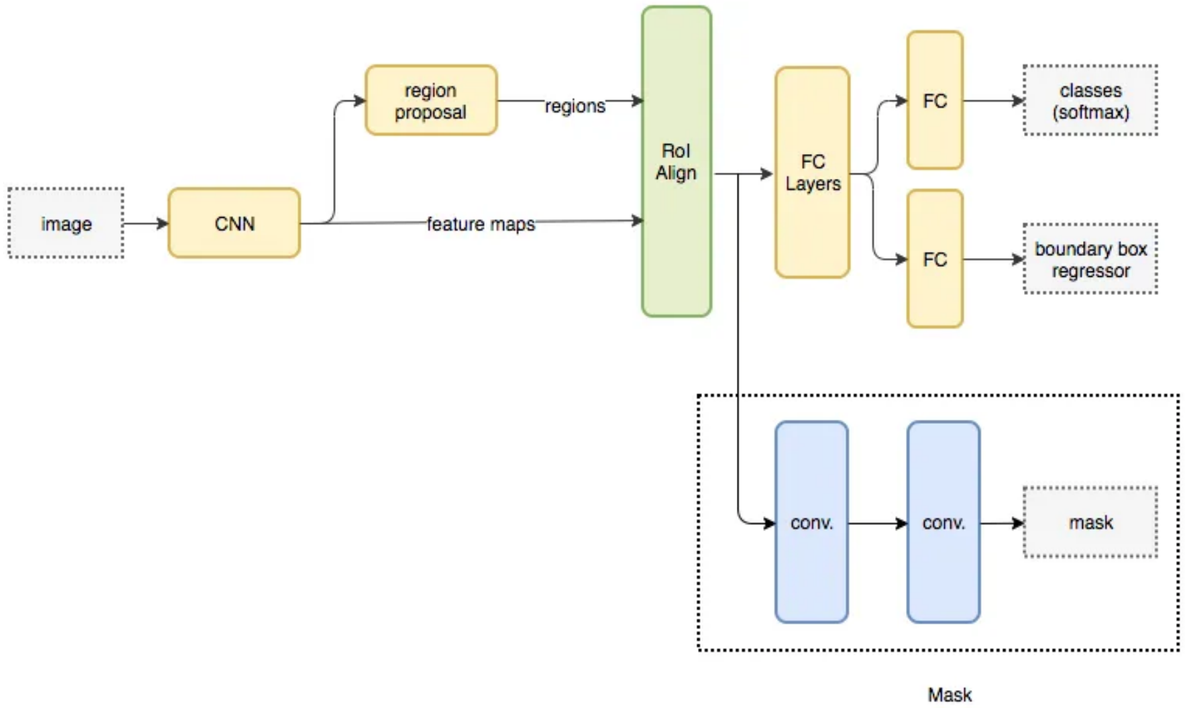
Shaoqing Ren et al. (2015) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". Advances in neural information processing systems, 2015, vol. 28.

Mask *R-CNN*



Kaiming He et al. (2017) "Mask R-CNN". Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Mask *R-CNN*



Kaiming He et al. (2017) "Mask R-CNN".
Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Mask *R-CNN*

ROI pooling

0.1	0.3	0.2	0.3	0.2	0.6	0.8	0.9
0.4	0.5	0.1	0.4	0.7	0.1	0.4	0.3
0.2	0.1	0.3	0.8	0.6	0.2	0.1	0.1
0.4	0.6	0.2	0.1	0.3	0.6	0.1	0.2
0.1	0.8	0.3	0.3	0.5	0.3	0.3	0.3
0.2	0.9	0.4	0.5	0.1	0.1	0.1	0.2
0.3	0.1	0.8	0.6	0.3	0.3	0.6	0.5
0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.2

0.8	0.6
0.9	0.6

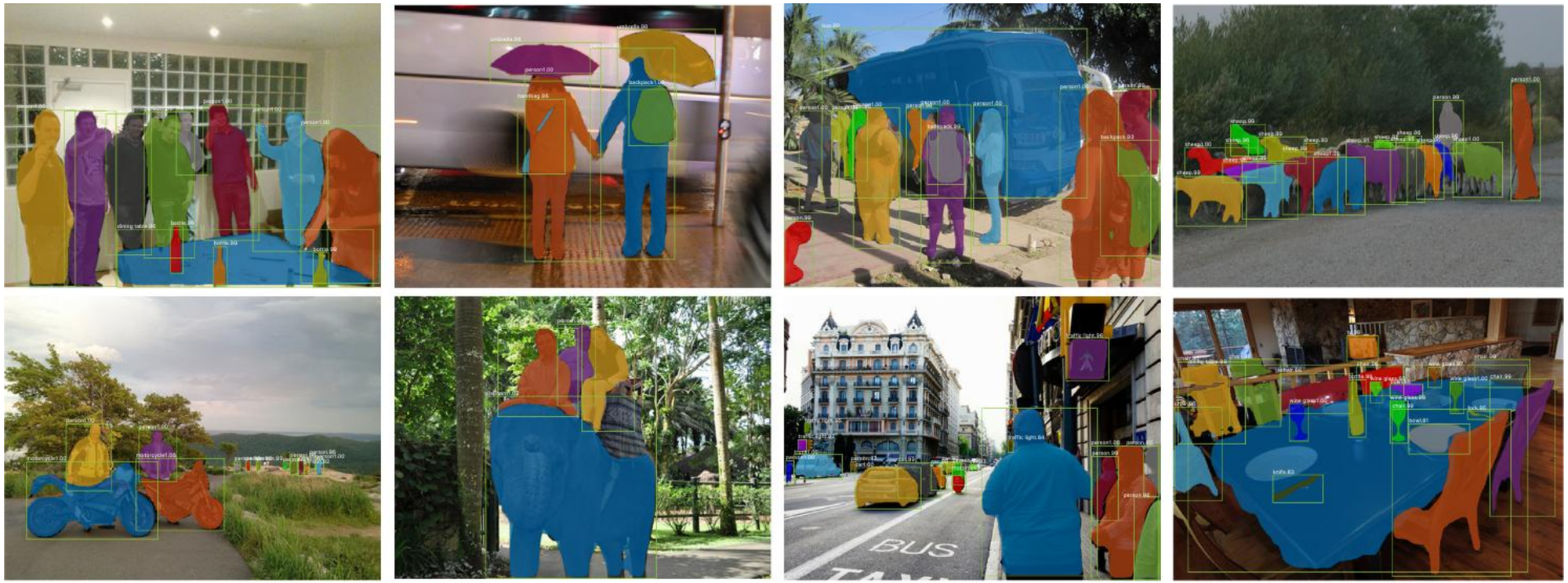
ROI Align

0.1	0.3	0.2	0.3	0.2	0.6	0.8	0.9
0.4	0.5	0.1	0.4	0.7	0.1	0.4	0.3
0.2	0.1	0.3	0.8	0.6	0.2	0.1	0.1
0.4	0.6	0.2	0.1	0.3	0.6	0.1	0.2
0.1	0.8	0.3	0.3	0.5	0.3	0.3	0.3
0.2	0.9	0.4	0.5	0.1	0.1	0.1	0.2
0.3	0.1	0.8	0.6	0.3	0.3	0.6	0.5
0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.2

0.88	0.6
0.9	0.6

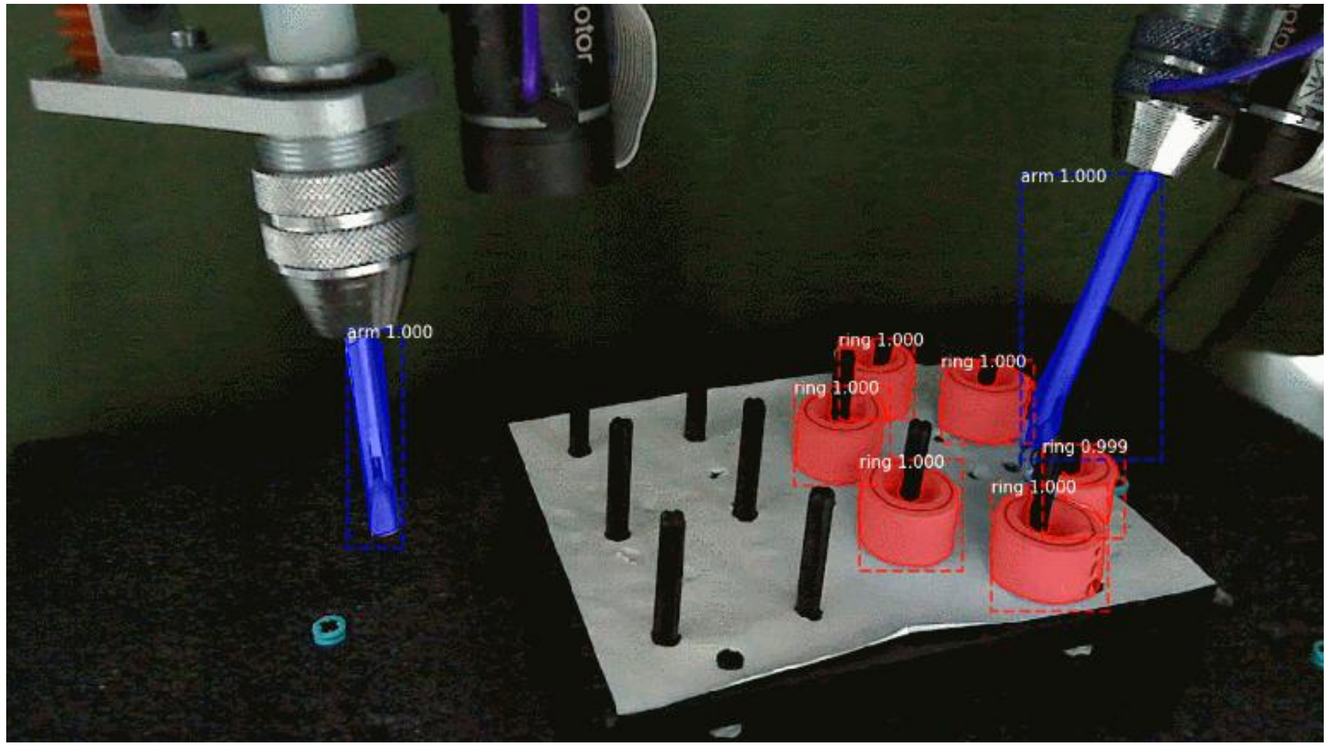
Kaiming He et al. (2017) "Mask R-CNN".
Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Mask *R-CNN*



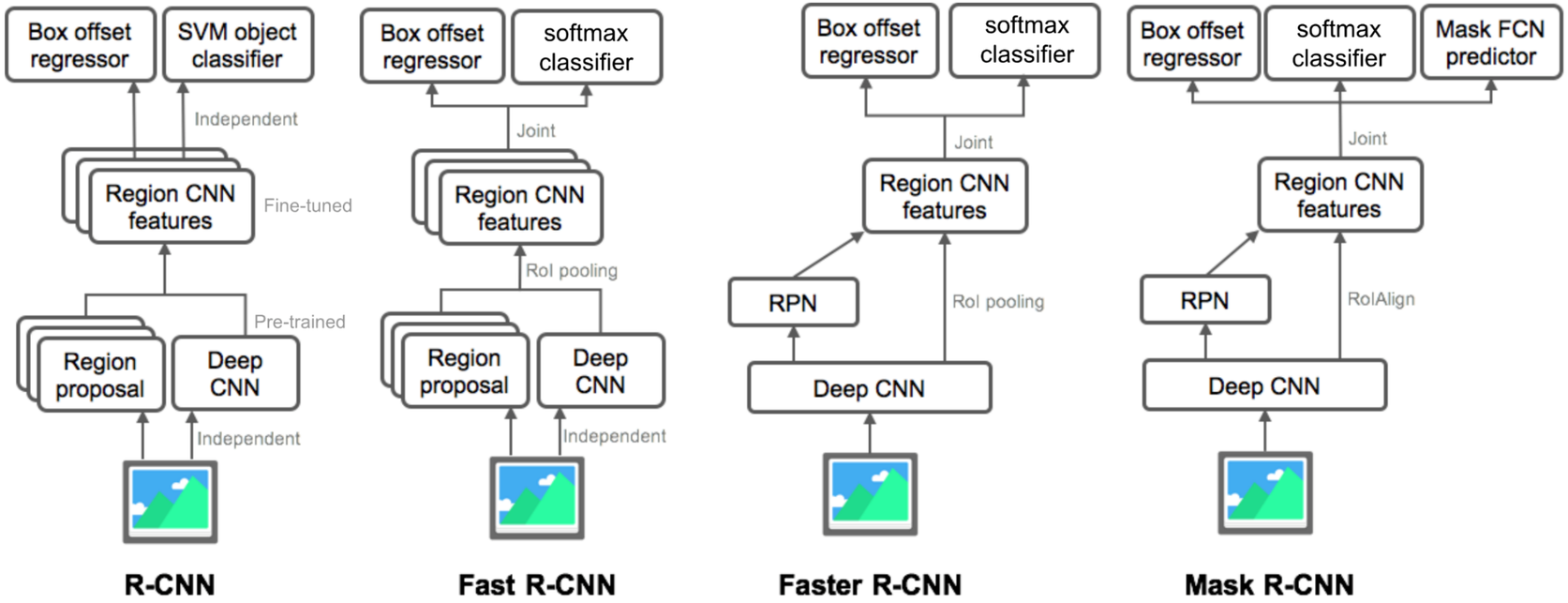
Kaiming He et al. (2017) "Mask R-CNN".
Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Mask *R-CNN*



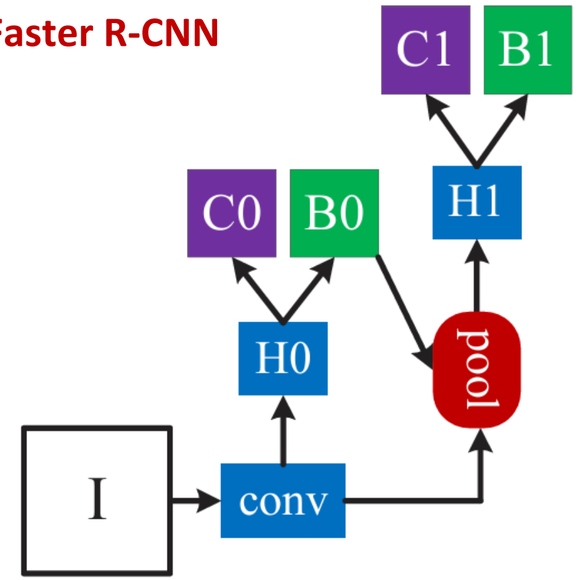
Kaiming He et al. (2017) "Mask R-CNN".
Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

R-CNN *family*


Box offset regressor
SVM object classifier
Independent
Region CNN features
Fine-tuned
Pre-trained
Region proposal
Deep CNN
Independent
R-CNN
Box offset regressor
softmax classifier
Joint
Region CNN features
RoI pooling
Region proposal
Deep CNN
Independent
Fast R-CNN
Box offset regressor
softmax classifier
Joint
Region CNN features
RPN
RoI pooling
Deep CNN
Faster R-CNN
Box offset regressor
softmax classifier
Mask FCN predictor
Joint
Region CNN features
RPN
RoIAlign
Deep CNN
Mask R-CNN

Cascade *R-CNN*

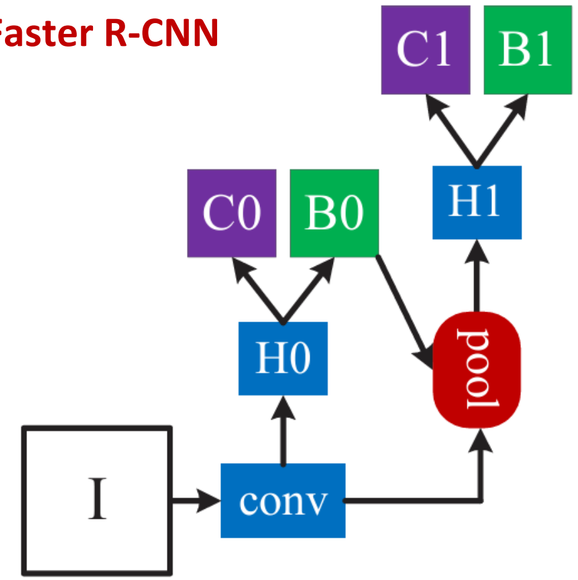
Faster R-CNN

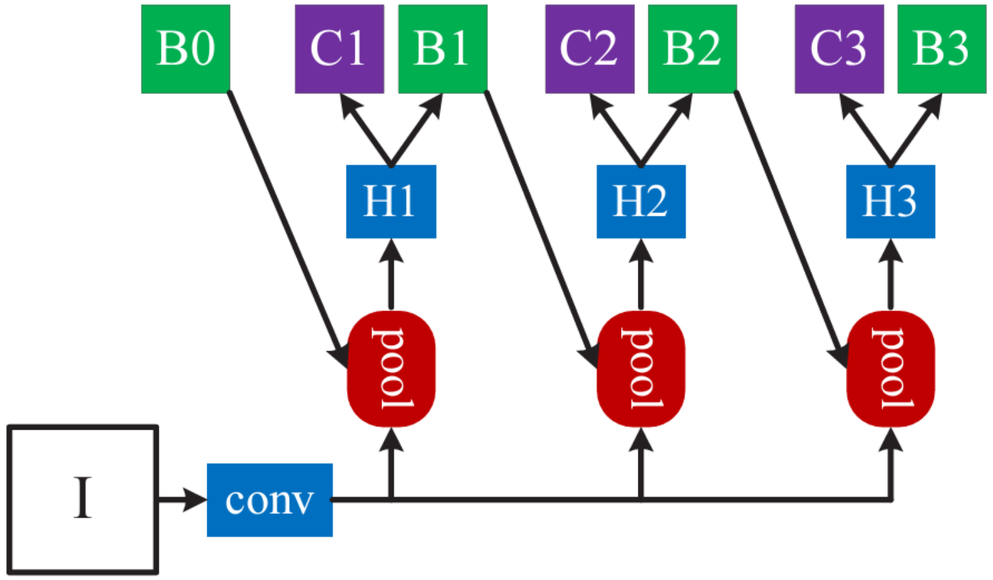


Zhaowei Cai and Nuno Vasconcelos (2018) "Cascade R-CNN: Delving into High Quality Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

Cascade *R-CNN*

Faster R-CNN





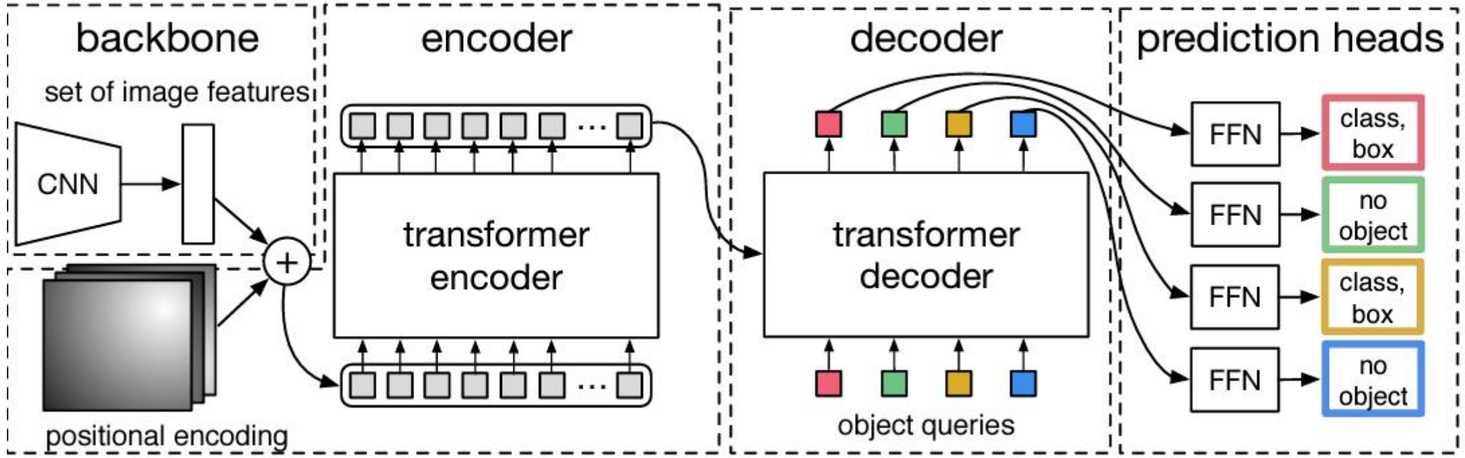
Zhaowei Cai and Nuno Vasconcelos (2018) "Cascade R-CNN: Delving into High Quality Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

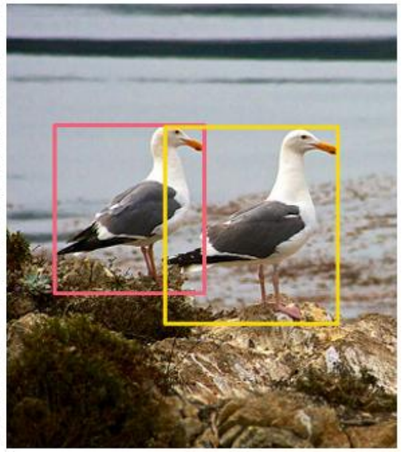
DETR

(DEtection TRansformer)

Sin anclas

Sin Non-Maximum Suppression (NMS)





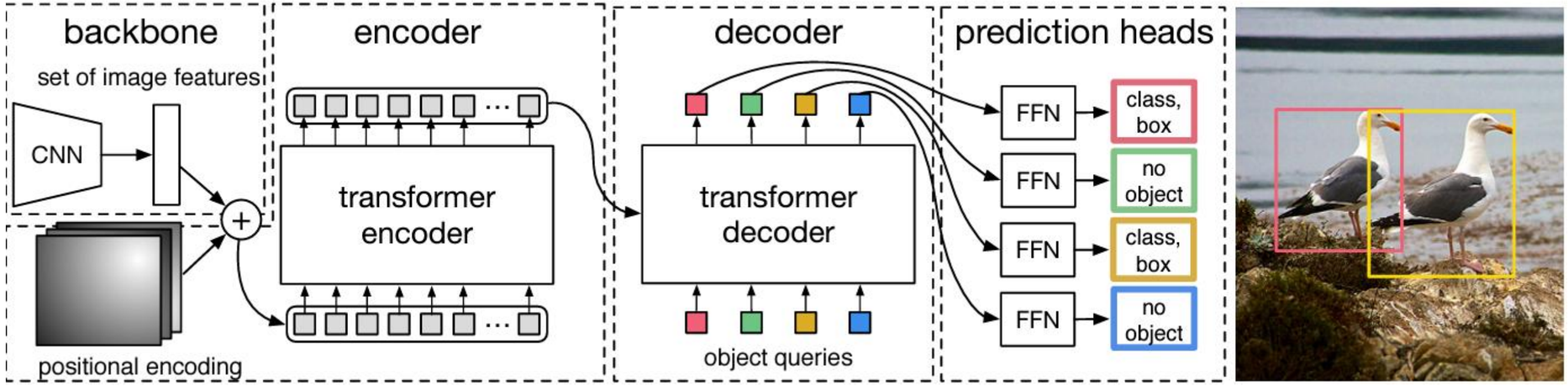
Object queries

Son representaciones vectoriales que están diseñados para preguntar qué objetos están presentes en una imagen y dónde están ubicados.

Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

(DEtection TRansformer)

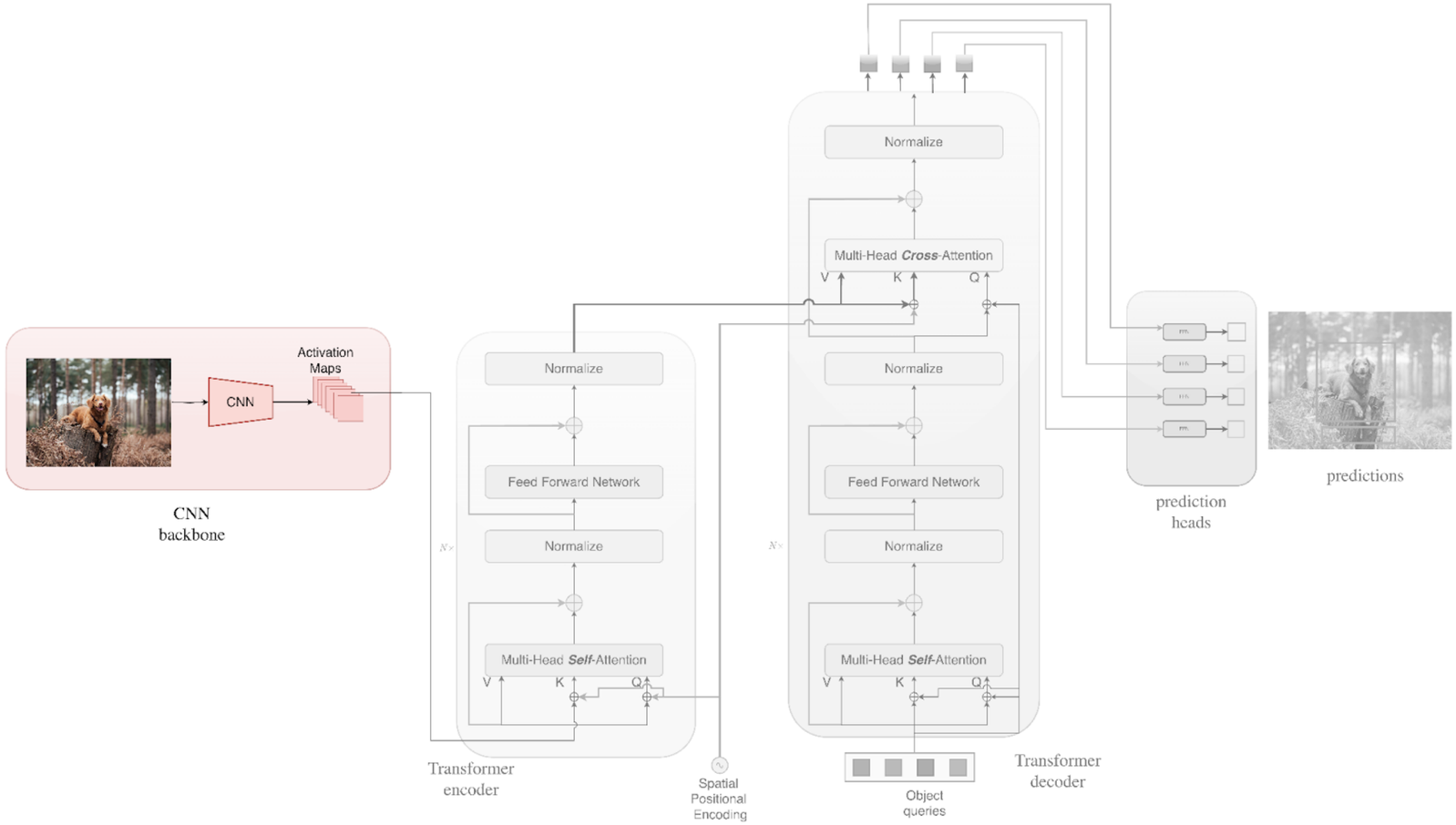


$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

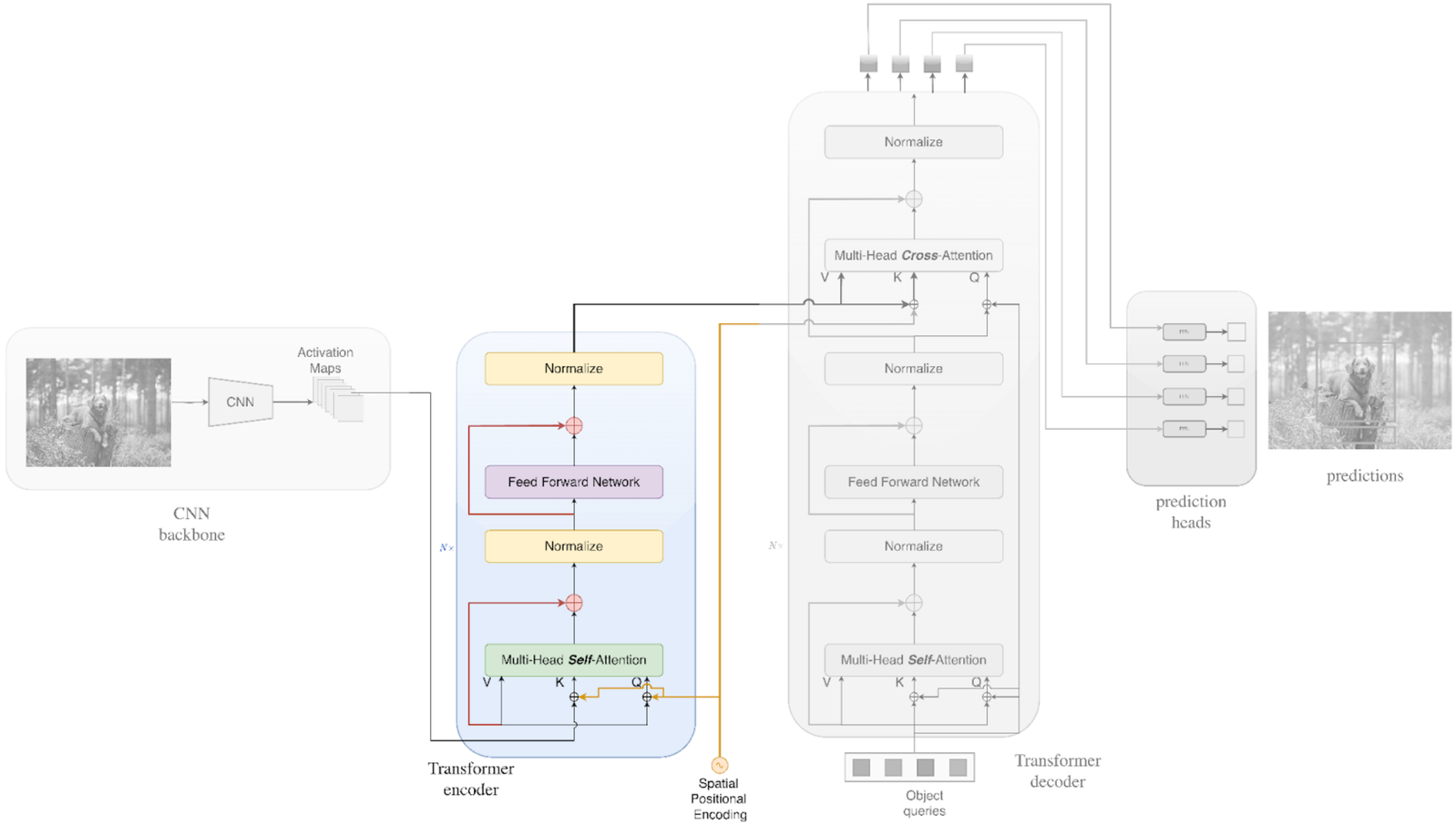
(DEtection TRansformer)



Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

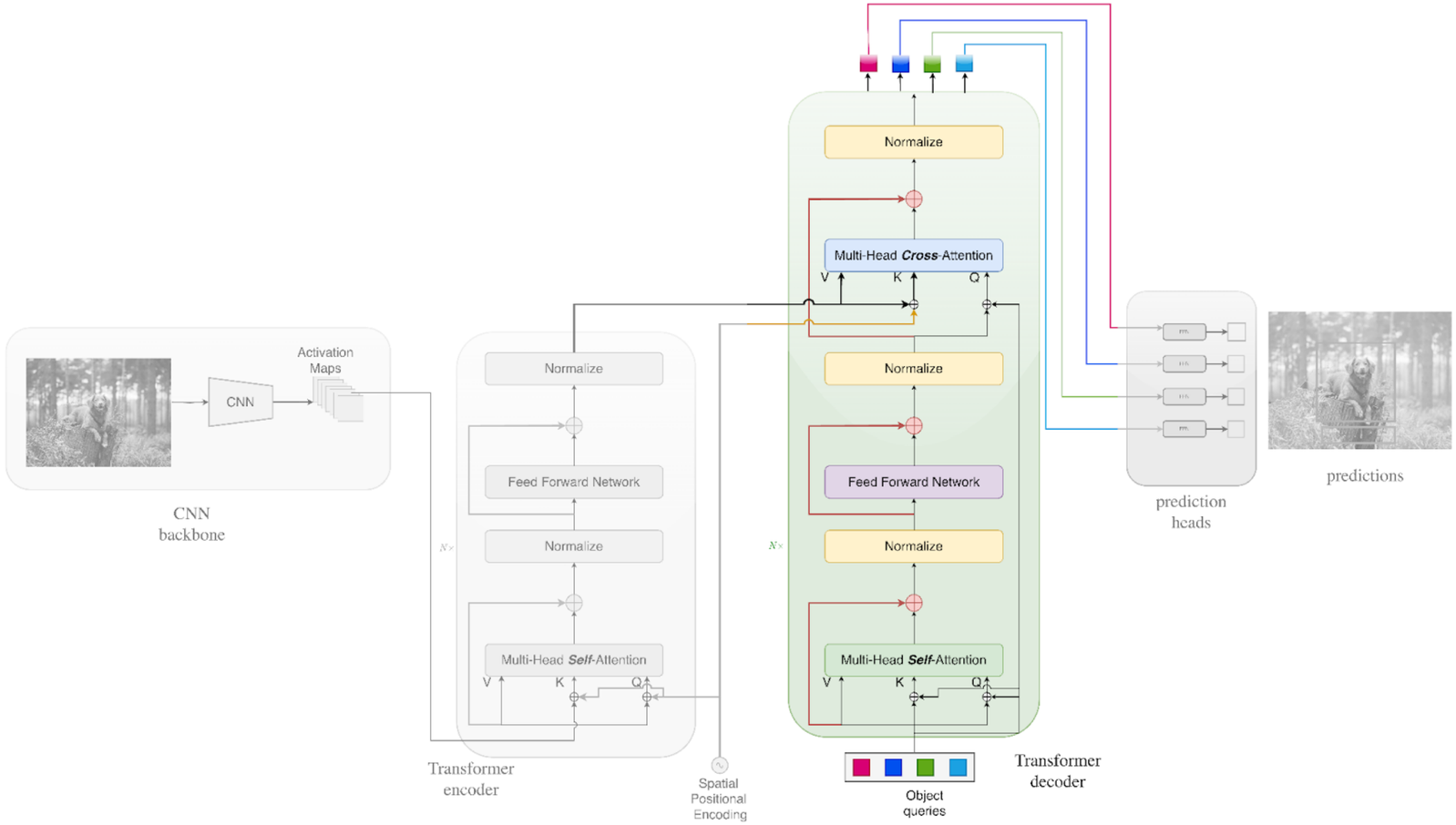
(DEtection TRansformer)



Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

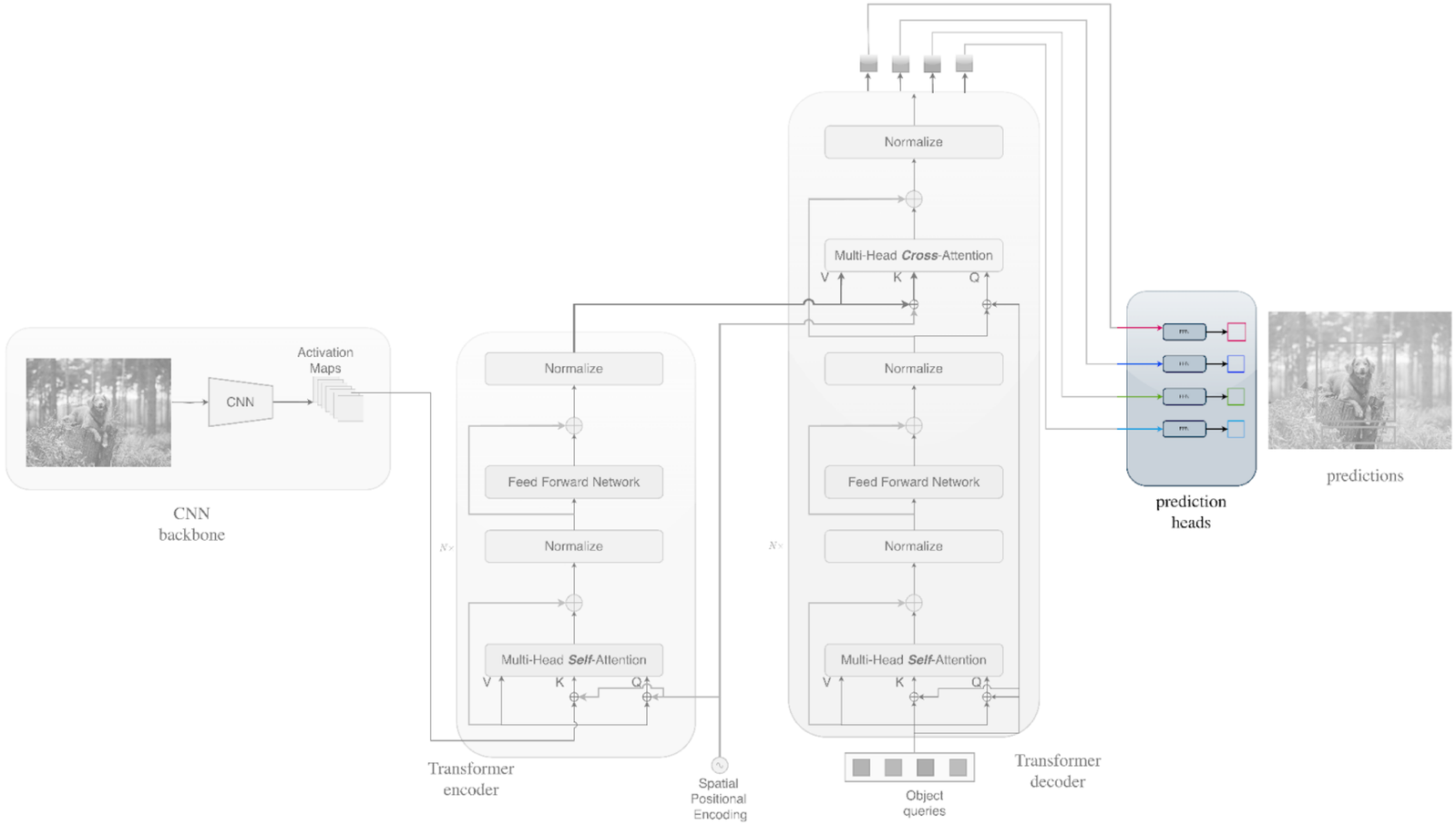
(DEtection TRansformer)



Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

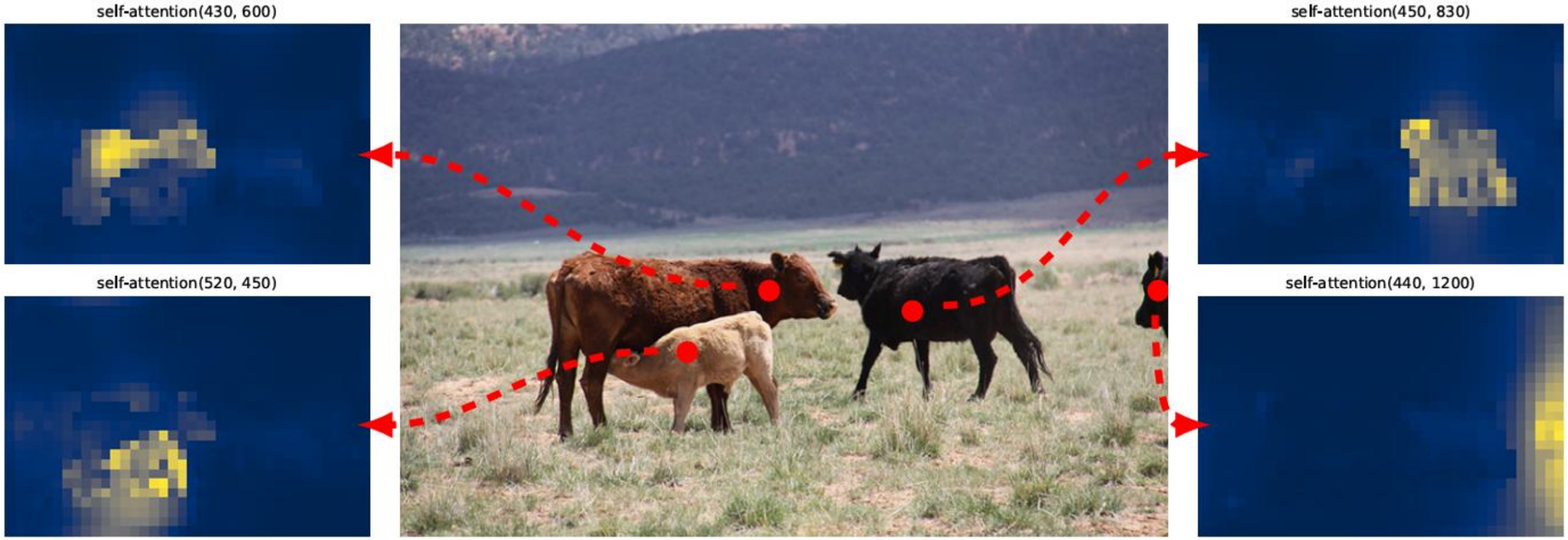
(DEtection TRansformer)



Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

DETR

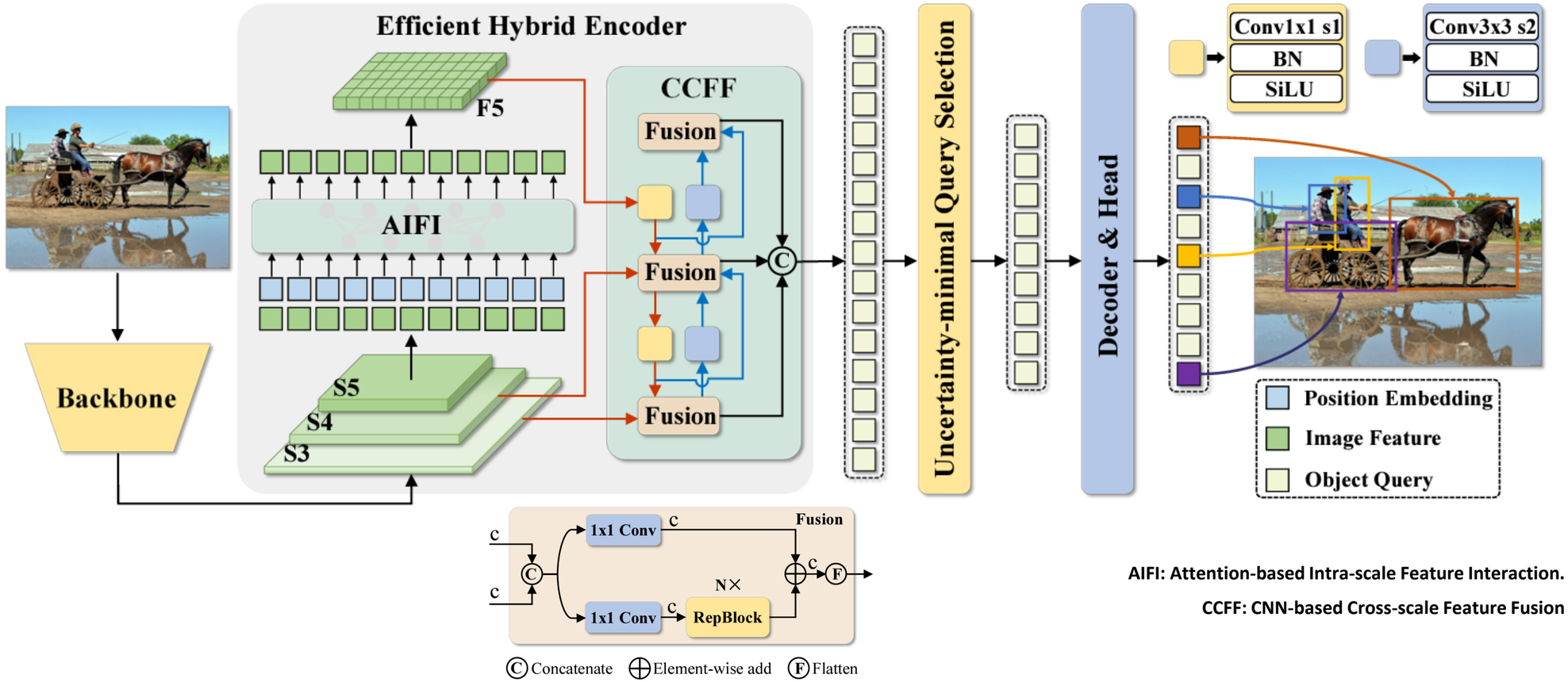
(DEtection TRansformer)



Nicolas Carion et al. (2020) "End-to-End Object Detection with Transformers". European conference on computer vision. Cham: Springer International Publishing, 2020.

RT-*DETR*

(Real-Time DEtection TRansformer)

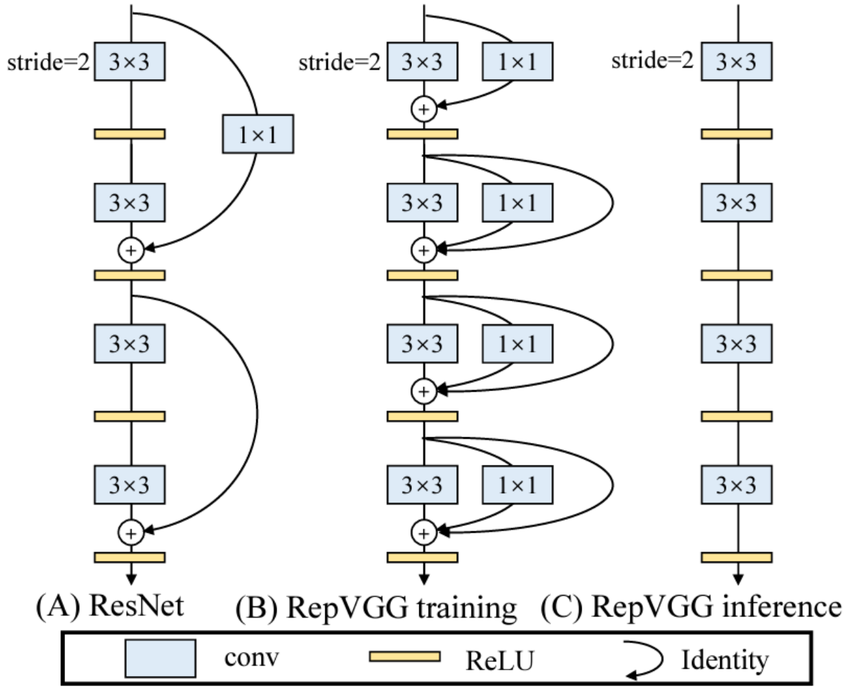


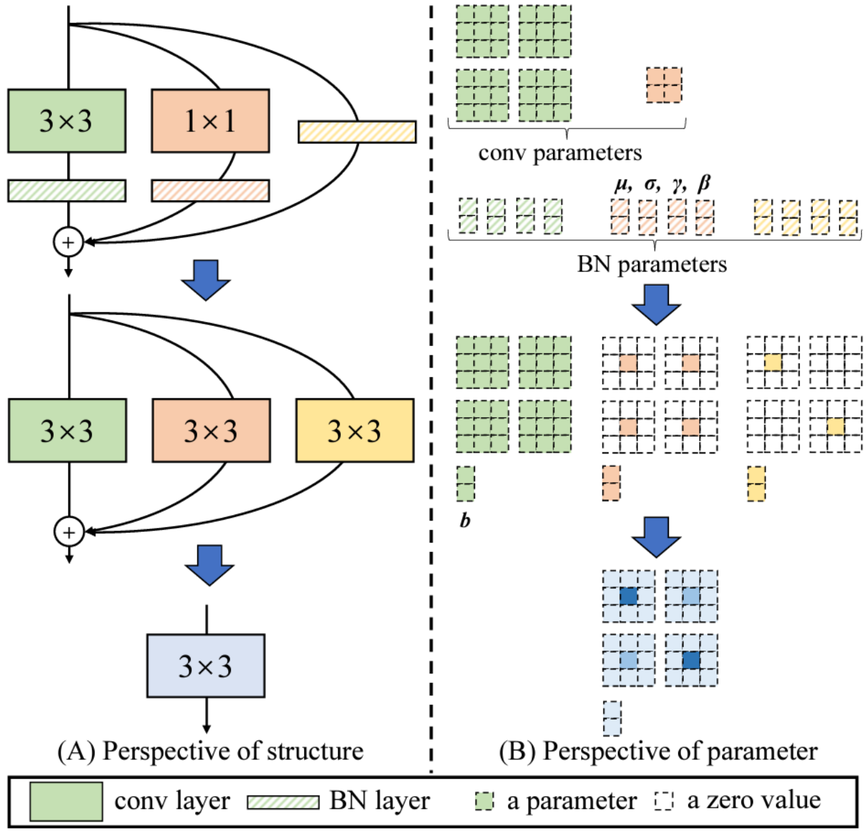
AIFI: Attention-based Intra-scale Feature Interaction.

CCFF: CNN-based Cross-scale Feature Fusion

Zhaowei Cai and Nuno Vasconcelos (2018) "Cascade R-CNN: Delving into High Quality Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

RepVGG

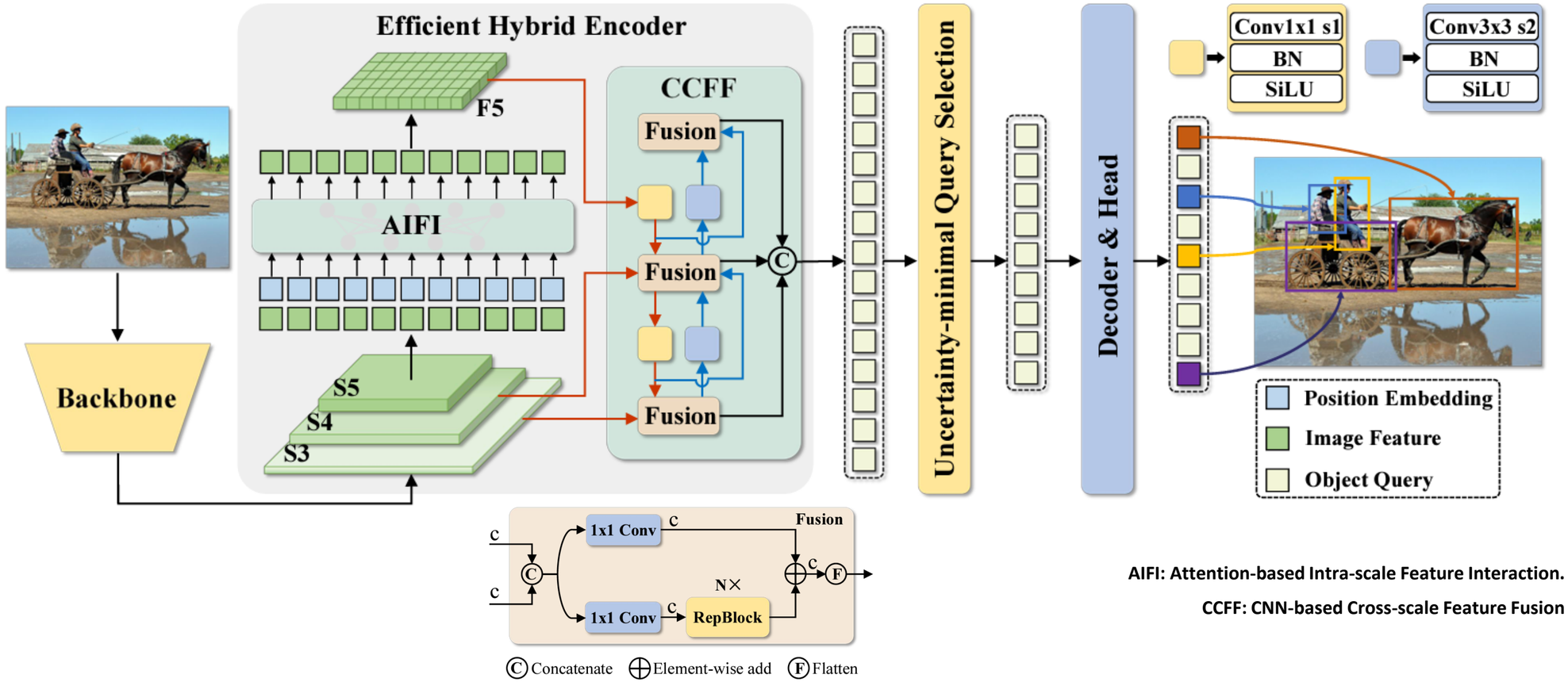




Xiaohan Ding (2021) "RepVGG: Making VGG-style ConvNets Great Again". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13733-13742).

RT-*DETR*

(Real-Time DEtection TRansformer)



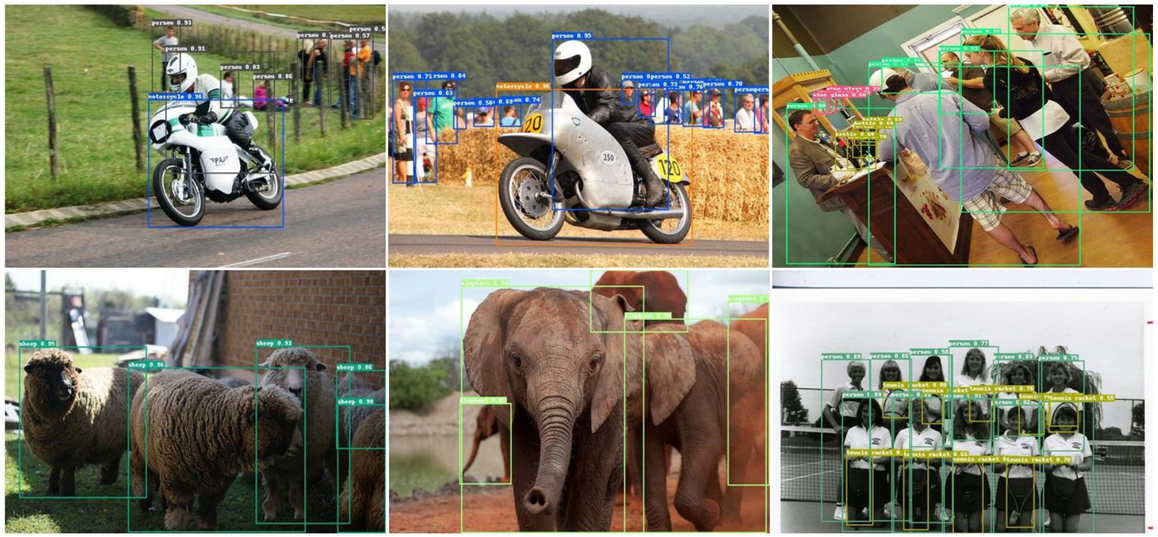
AIFI: Attention-based Intra-scale Feature Interaction.

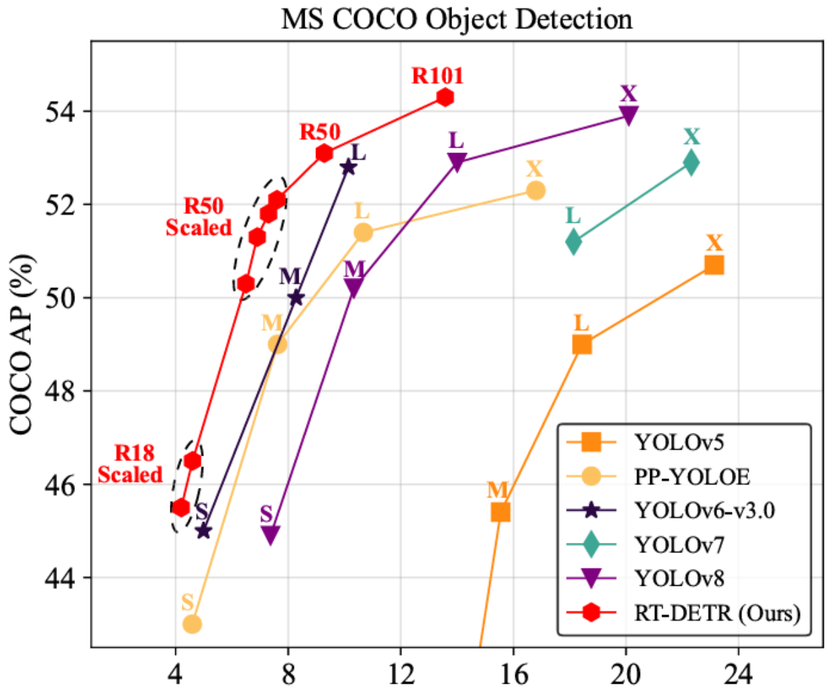
CCFF: CNN-based Cross-scale Feature Fusion

Zhaowei Cai and Nuno Vasconcelos (2018) "Cascade R-CNN: Delving into High Quality Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

RT-*DETR*

(Real-Time DEtection TRansformer)





Zhaowei Cai and Nuno Vasconcelos (2018) "Cascade R-CNN: Delving into High Quality Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

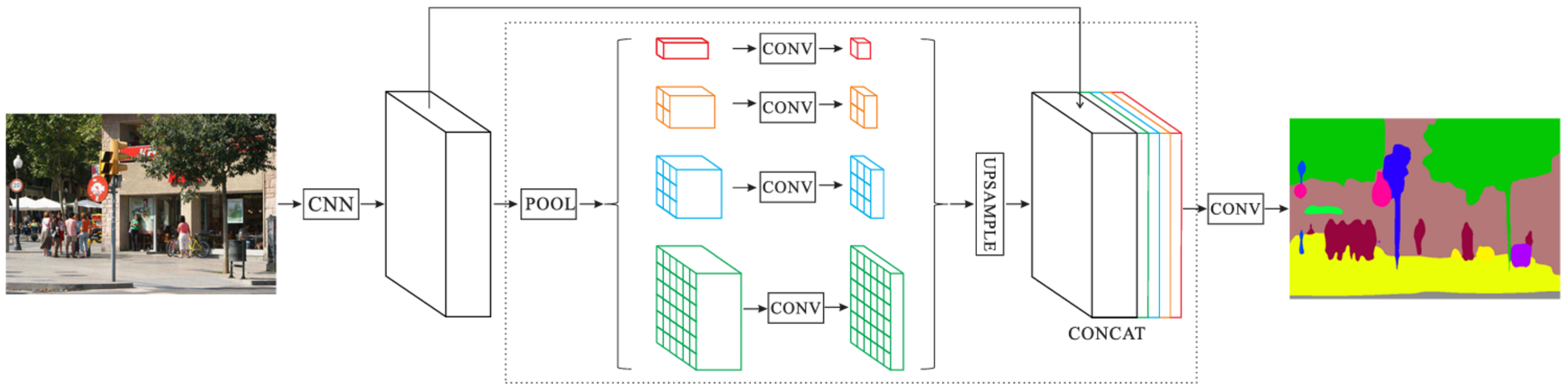
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

5.

DeepLab

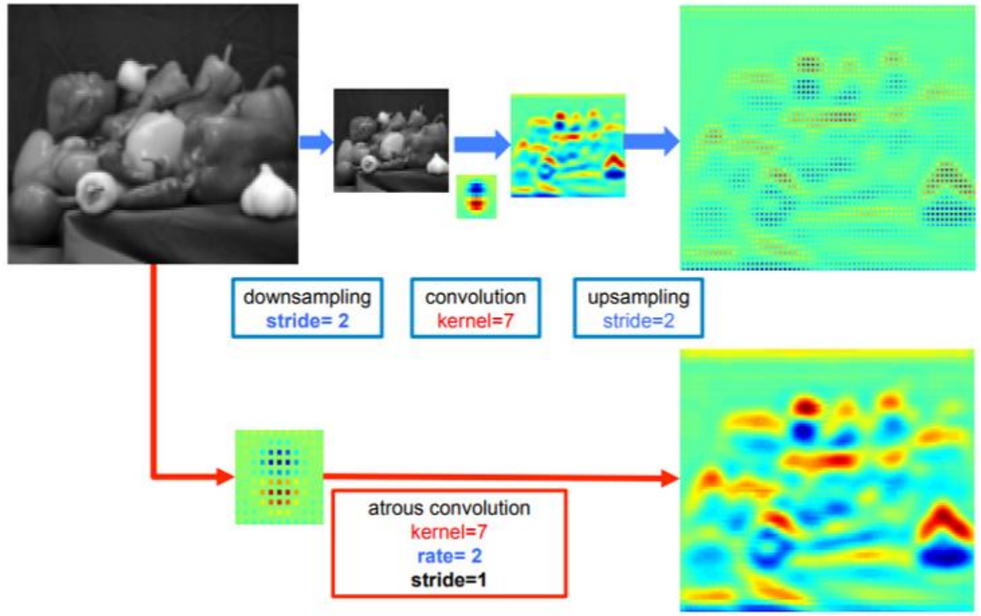
TRANSFORMATEC

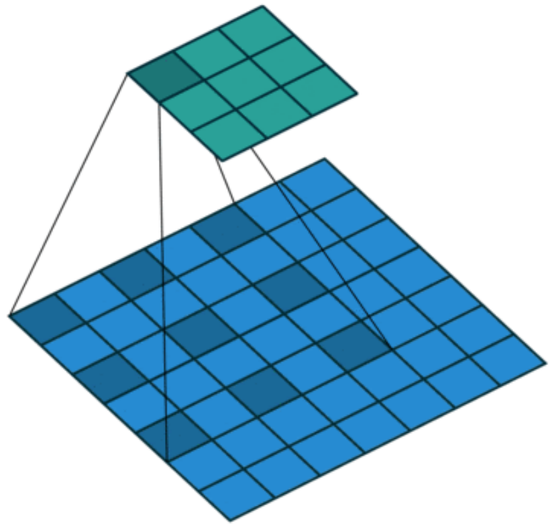
Pyramid Scene Parsing *Network (PSPNet)*



Hengshuang Zhao et al. (2017) "Pyramid scene parsing network".
Proceedings of the IEEE conference on computer vision and pattern recognition.

DeepLab



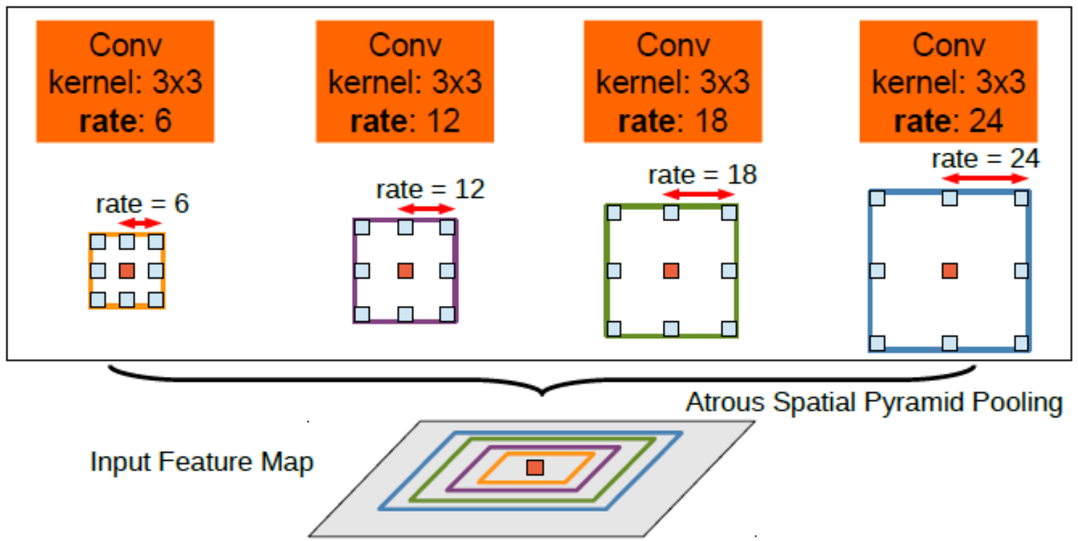


Atrous convolution

Liang-Chieh Chen et al. (2017) "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs".
IEEE transactions on pattern analysis and machine intelligence, 2017, vol. 40, no 4, p. 834-848.

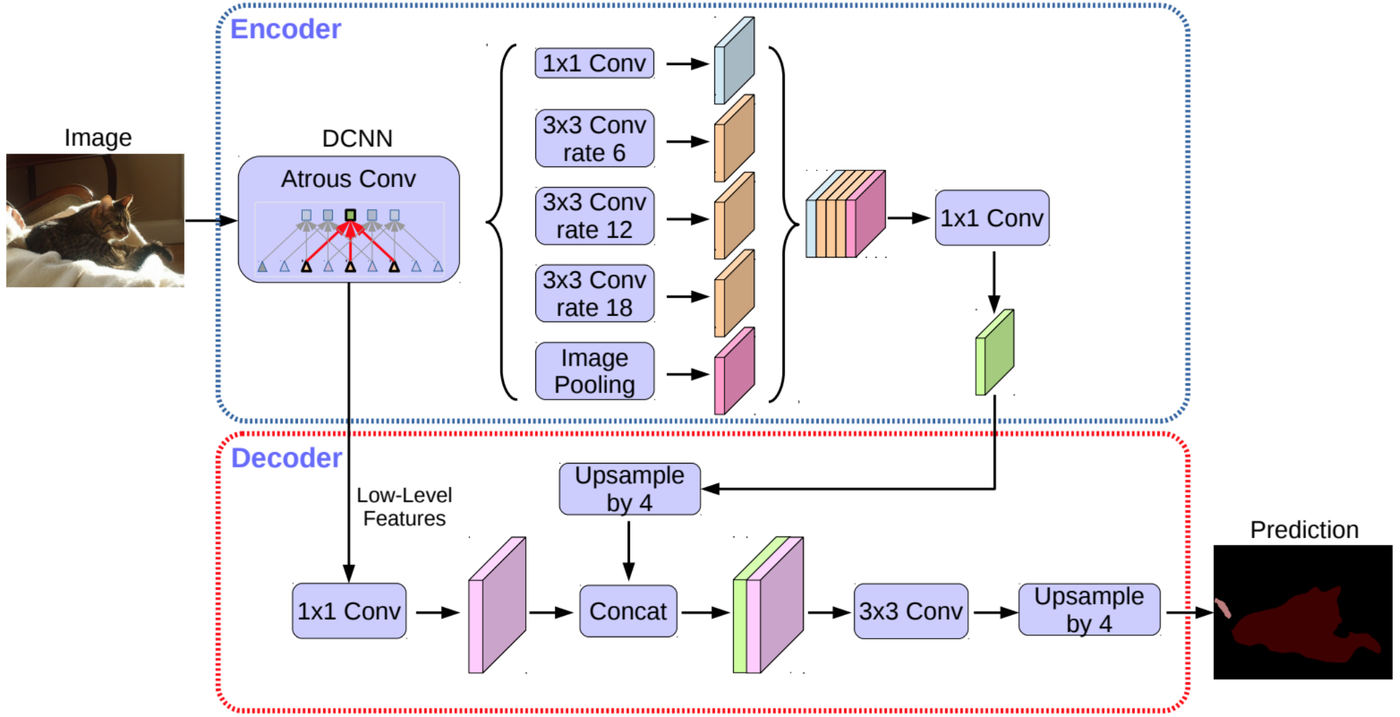
DeepLab

Atrous Spatial Pyramid Pooling (ASPP)



Liang-Chieh Chen et al. (2017) "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs".
IEEE transactions on pattern analysis and machine intelligence, 2017, vol. 40, no 4, p. 834-848.

DeepLabV3+



Chen, Liang-Chieh, et al. (2018) "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". Proceedings of the European conference on computer vision (ECCV). 2018. p. 801-818.

Deep*LabV3+*

Chen, Liang-Chieh, et al. (2018) "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". Proceedings of the European conference on computer vision (ECCV). 2018. p. 801-818.

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

GRACIAS

Victor Flores Benites

TRANSFORMATEC