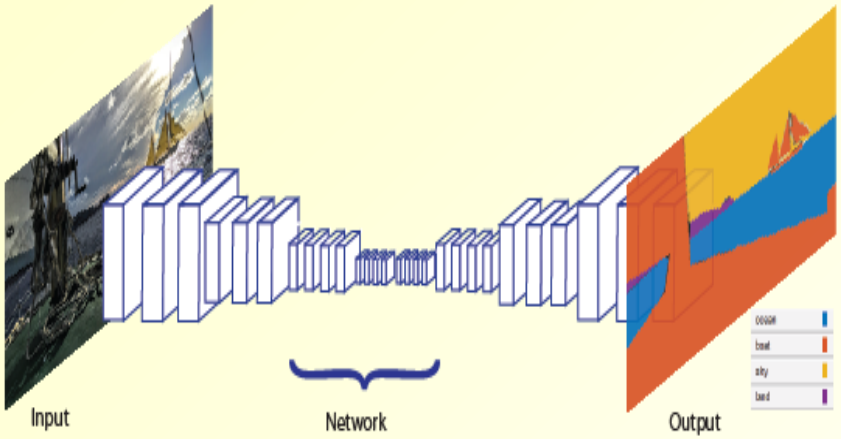
Segmentación con Deep Learning

Conceptos

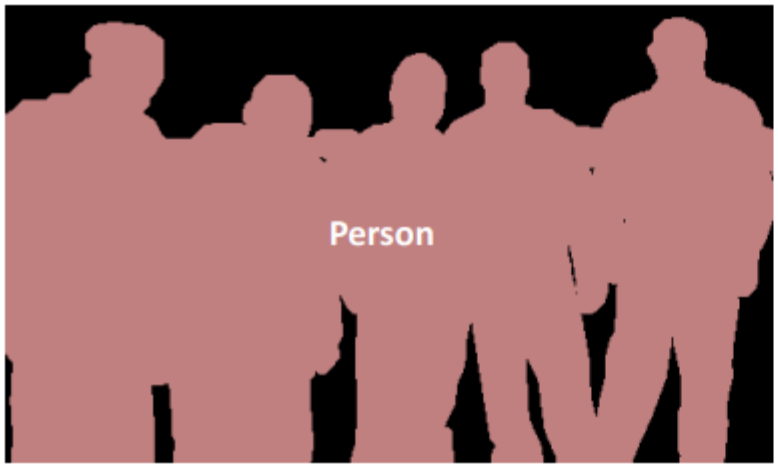
- Redes neuronales Autocodificadoras
- U-net
- DeepLabv3+
- Métricas. IoU
- Vision Transformer
- MaskFormer

Segmentación Semántica es el proceso de asociar a cada pixel de la imagen una etiqueta de clase (p.e flores,personas calle, cielo, océano, coche, etc).

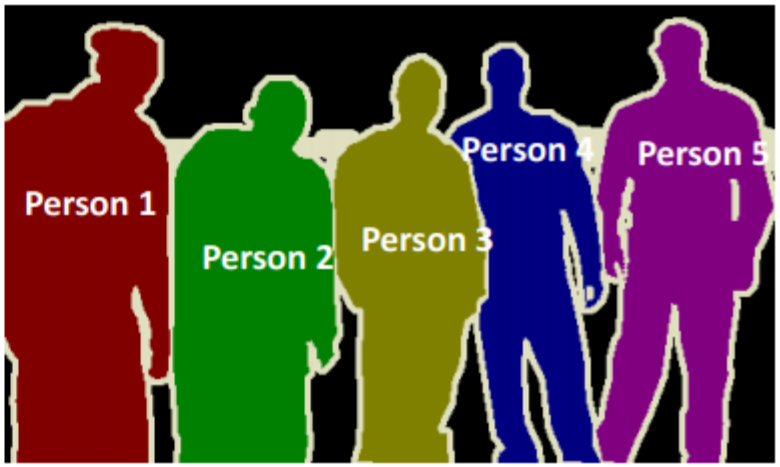


Aplicaciones

Conducción autónoma
Inspección industrial
Clasificación de terrenos con imágenes satélites
Análisis de imágenes médicas (X-rayos para el COVID-19).



Segmentación Semántica

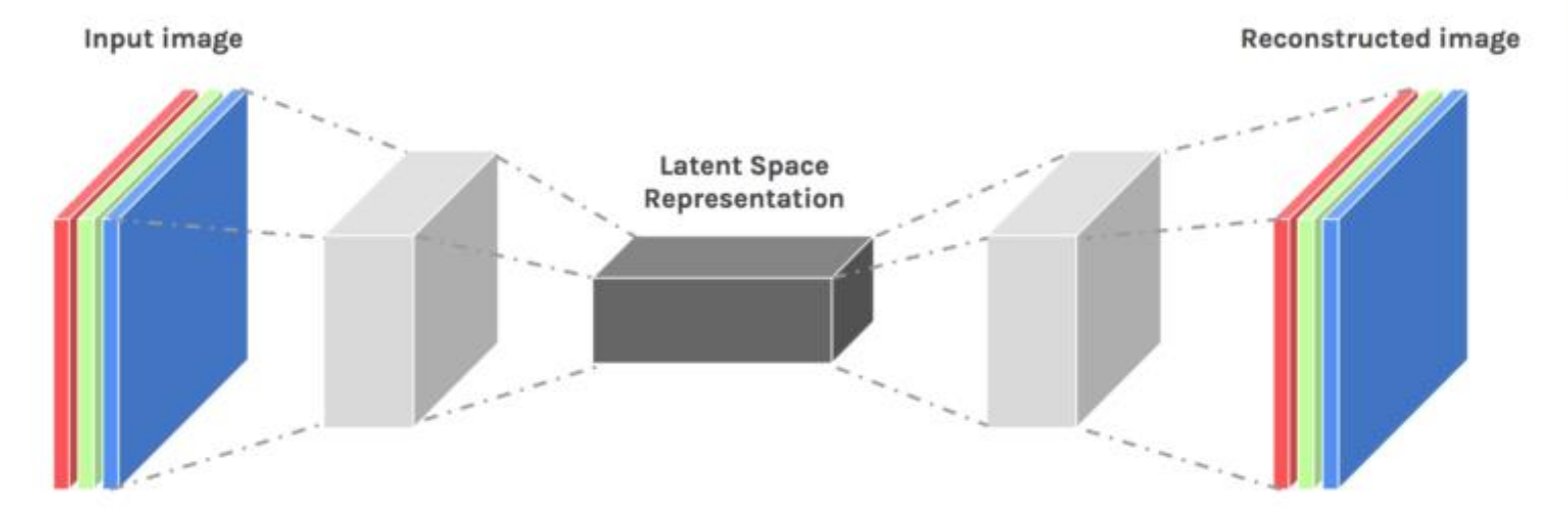


Segmentación por Instancias

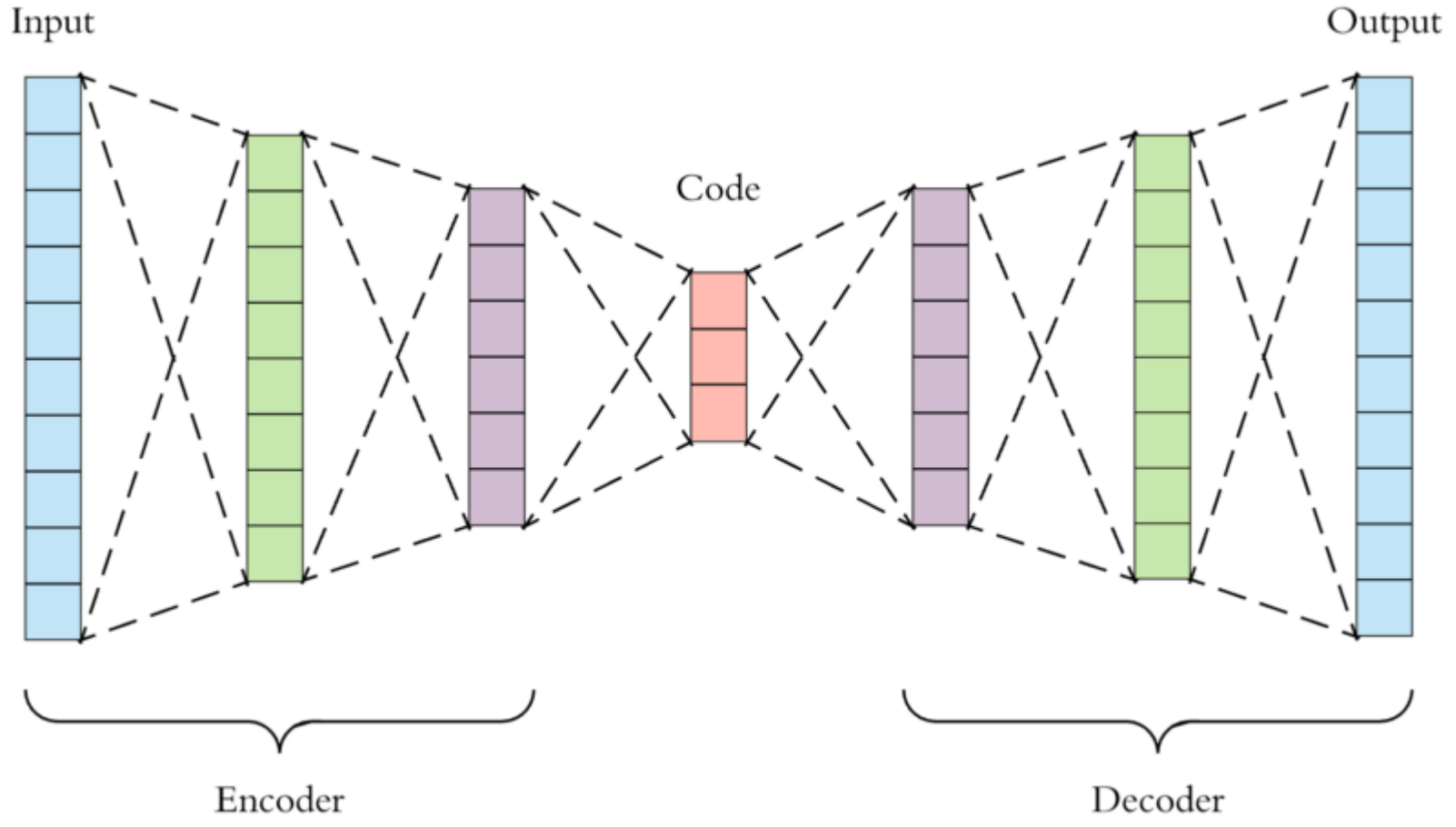
Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

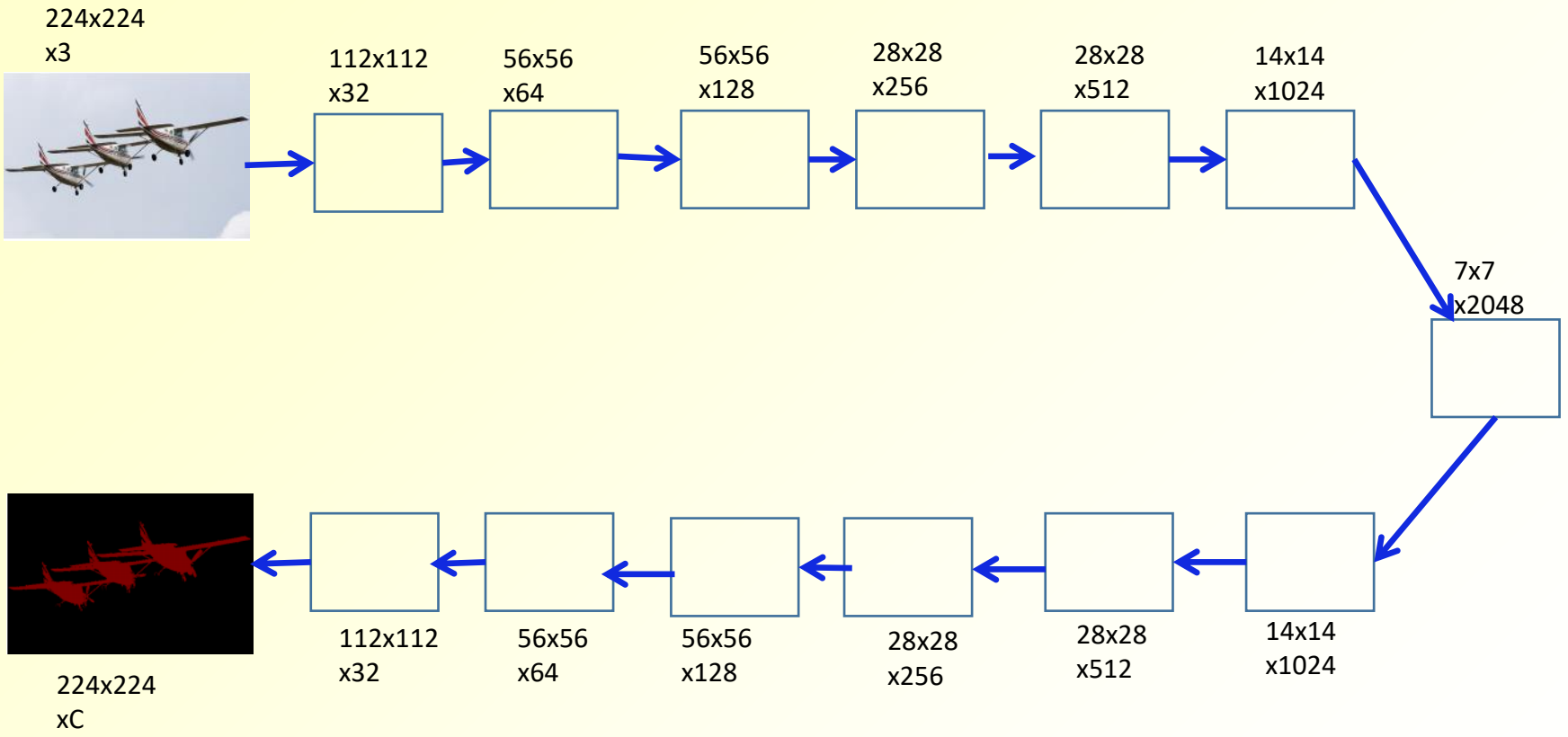
Red neuronal Autocodificadora: es un algoritmo de aprendizaje no supervisado que aplica propagación hacia atrás, donde los valores objetivos son iguales a las entradas. Los autocodificadores se usan pare reducir el tamaño de las entradas para obtener una representación con menos datos.



Arquitectura Autocodificador (Red Codificadora-Decodificadora)



Arquitectura Autocodificador (Red Codificadora-Decodificadora)



Capas de la red autocodificadora

Convolución

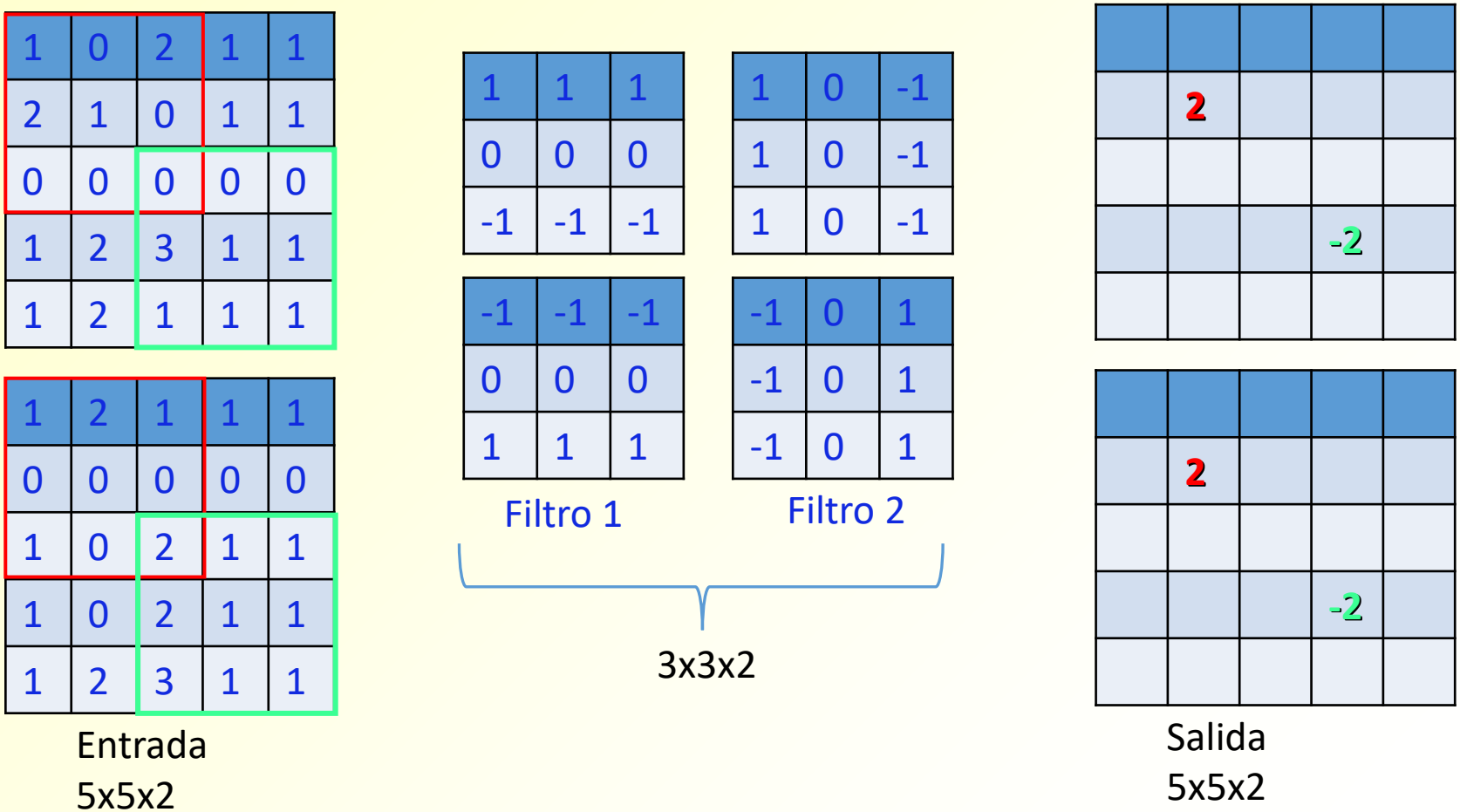
Sub-muestreo

Sobre-muestreo

Deconvolución

Capas de la red autocodificadora

Convolución: Es una operación de convolución que actua sobre un volumen



Capas de la red autocodificadora

Sub-muestreo:Nos permite reducir la dimensión de la entrada. Las posibles capas para realizar submuestreo son: Max-Pooling, Average-Pooling, Strided Convolution

1	0	2	1
2	1	0	1
0	0	0	0
1	2	3	1

Max-Pooling

2	2
2	3

Average-Pooling

1	1
0.75	1

Strided Convolution (s=2)

1	0
0	1

Filtro

2	3
2	1

Capas de la red autocodificadora

Sobre-muestreo:Nos permite aumentar la dimensión de la entrada. Las posibles capas para realizar sobre-muestreo son: Un-Pooling y deconvolución.

	0	1	2	3
0	1	0	2	1
1	2	1	0	1
2	0	0	0	0
3	1	2	3	1

Max-Pooling

2	2
2	3

Máscara

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

Un-Pooling

0	0	2	0
2	0	0	0
0	0	0	0
0	2	3	0

Capas de la red autocodificadora

Deconvolución

	0	1	2	3
0	1	0	2	1
1	2	1	0	1
2	0	0	0	0
3	1	2	3	1

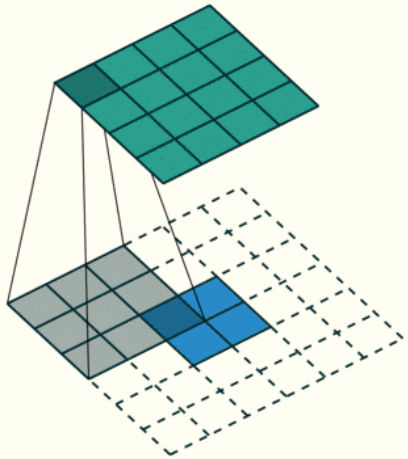
Max-Pooling

2	2
2	3

0	0	0	0	0	0
0	0	0	0	0	0
0	0	2	2	0	0
0	0	2	3	0	0
0	0	0	0	0	0
0	0	0	0	0	0

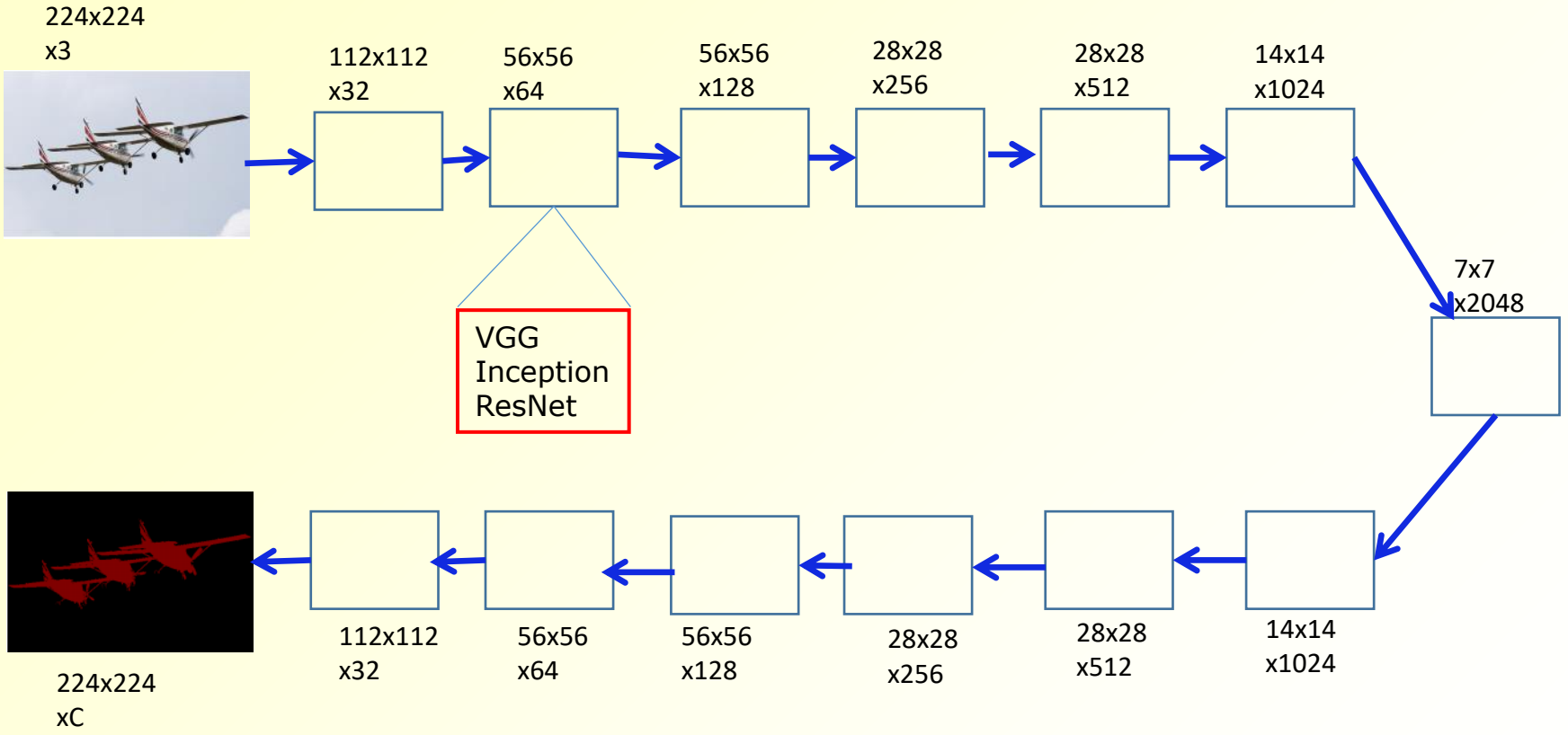
1	**1**	**1**
0	0	0
1	1	1

Filtro

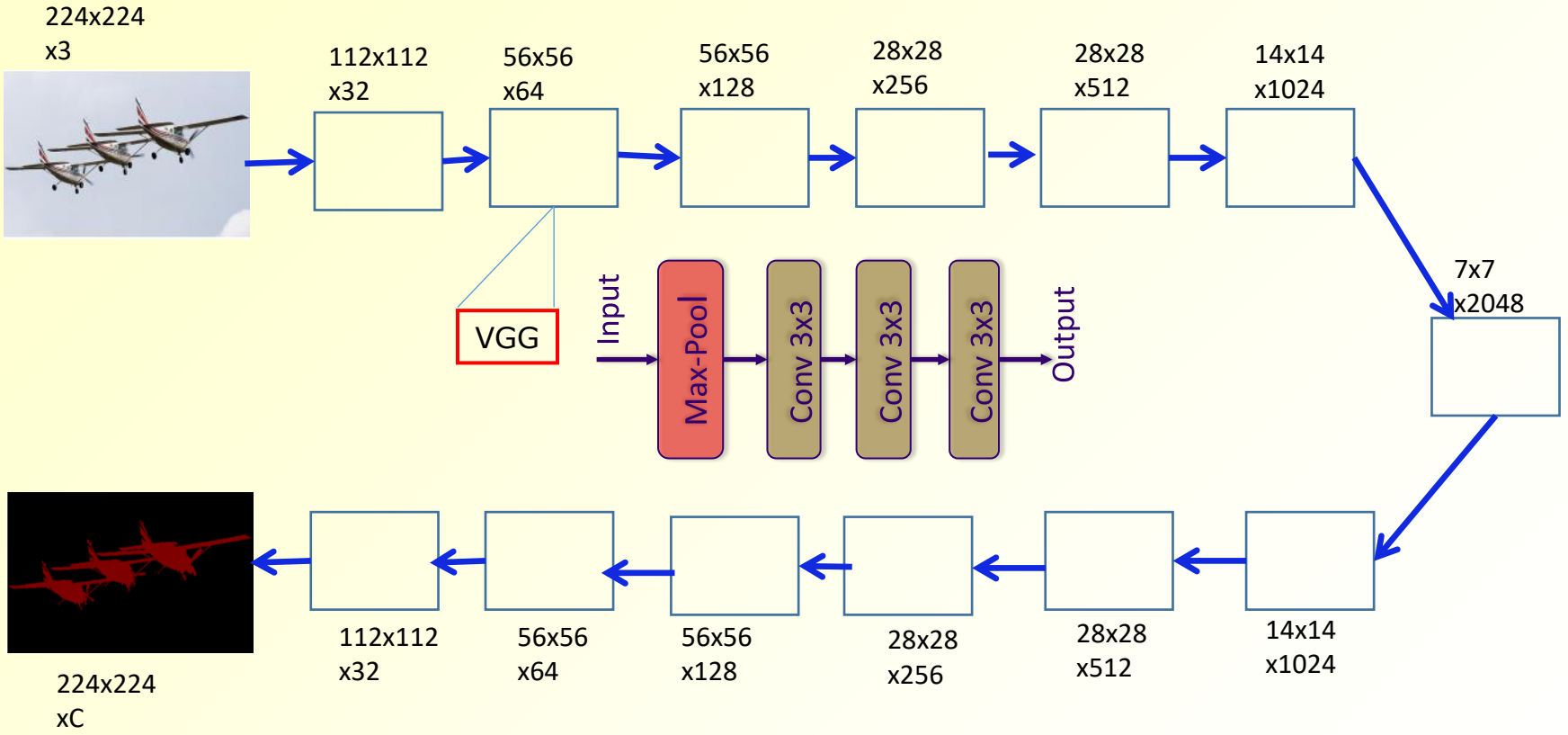


	0	1	2	3
0	2	4	4	2
1	2	5	5	3
2	2	4	4	4
3	2	5	5	3

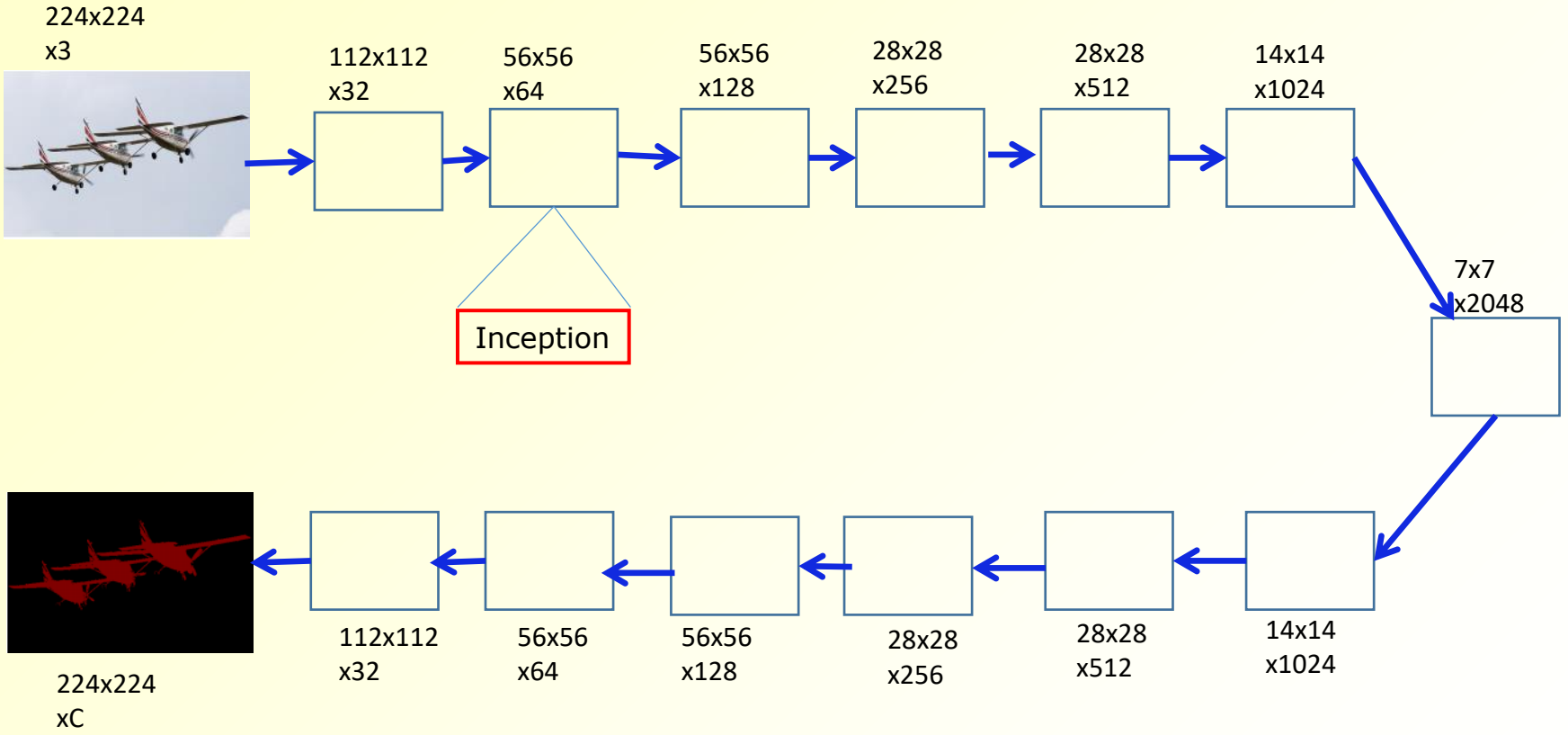
Arquitectura Autocodificador (Red Codificadora-Decodificadora)



Arquitectura Autocodificador (Red Codificadora-Decodificadora)

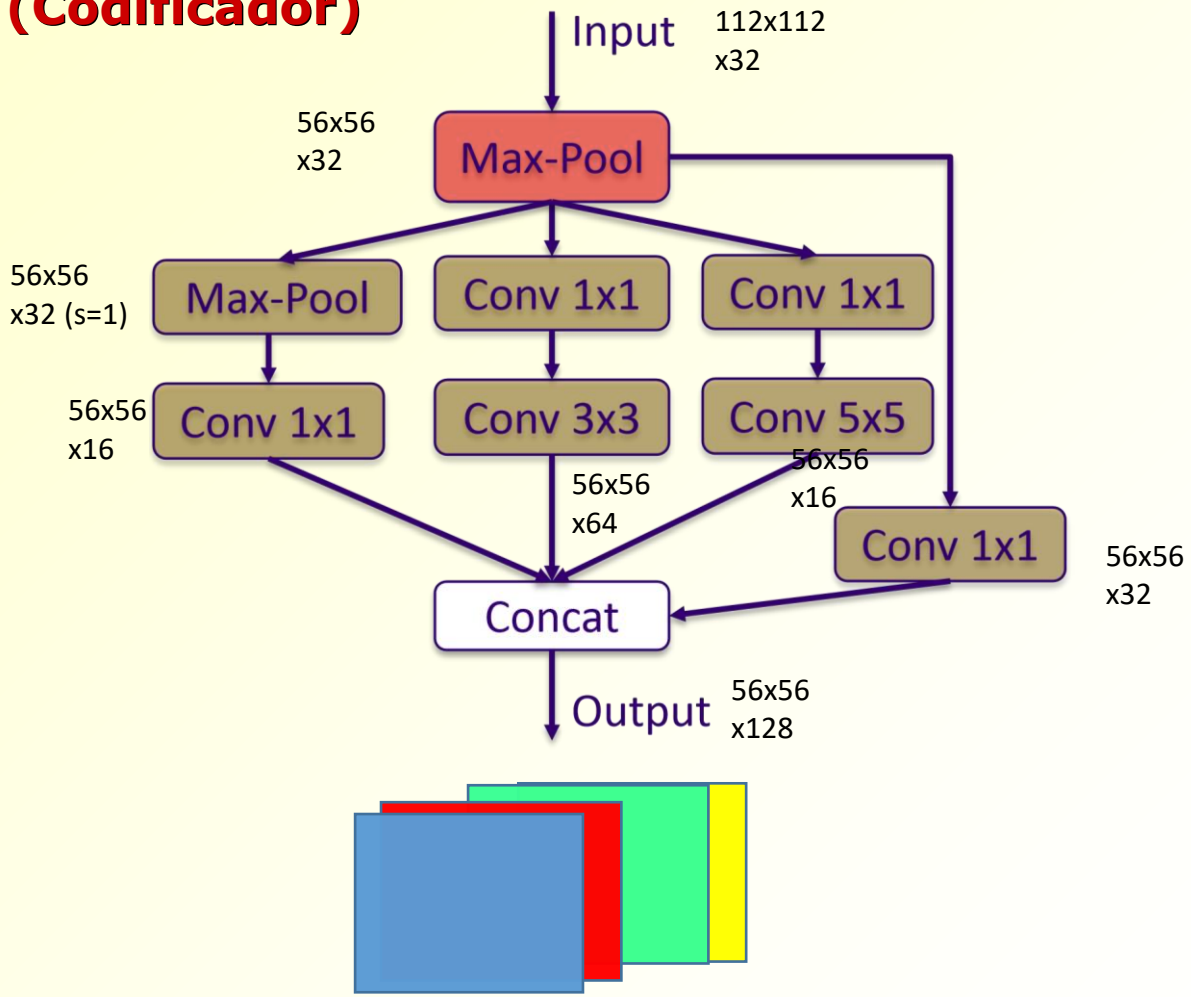


Arquitectura Autocodificador (Red Codificadora-Decodificadora)

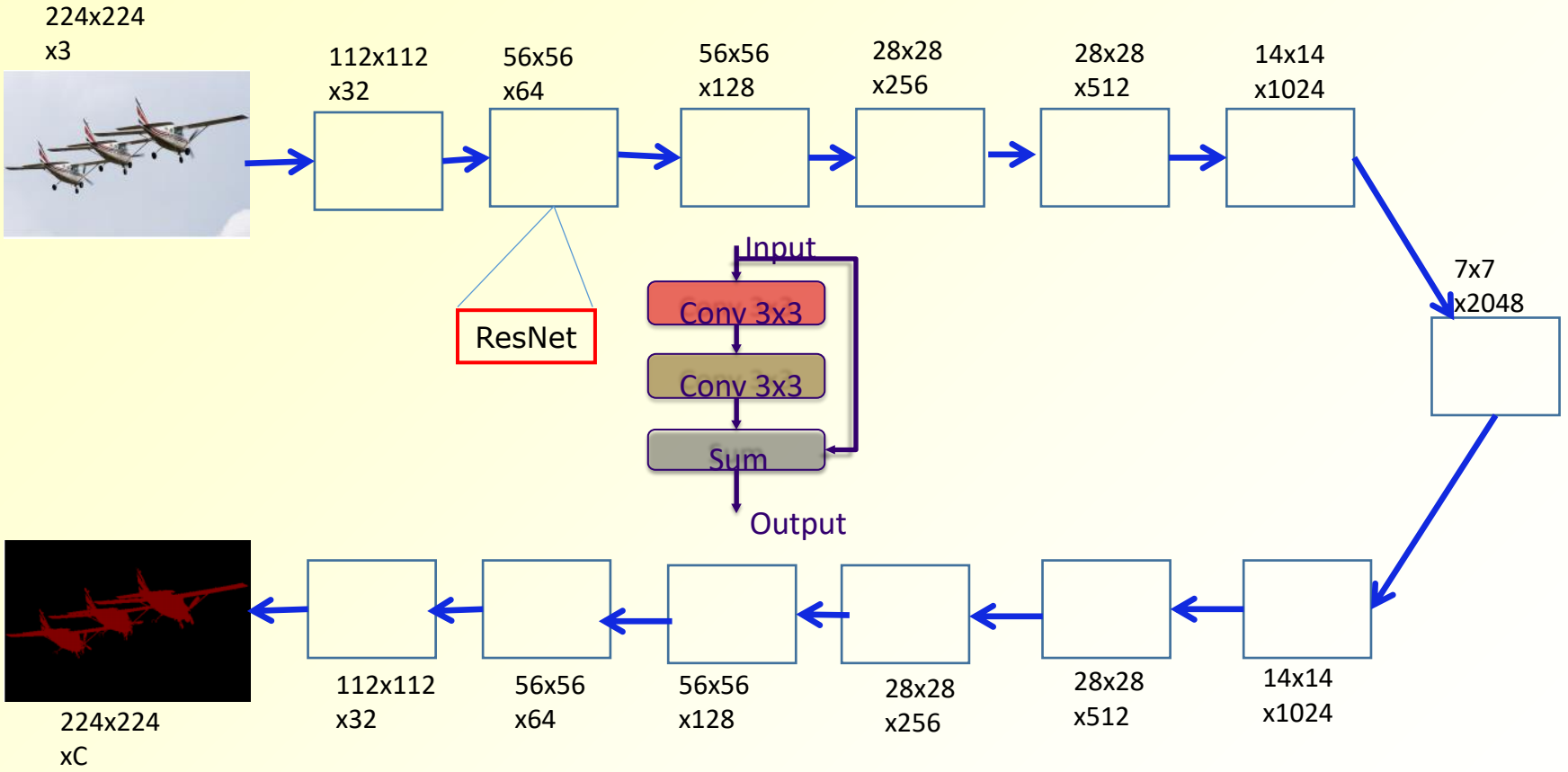


Arquitectura Autocodificador (Red Codificadora-Decodificadora)

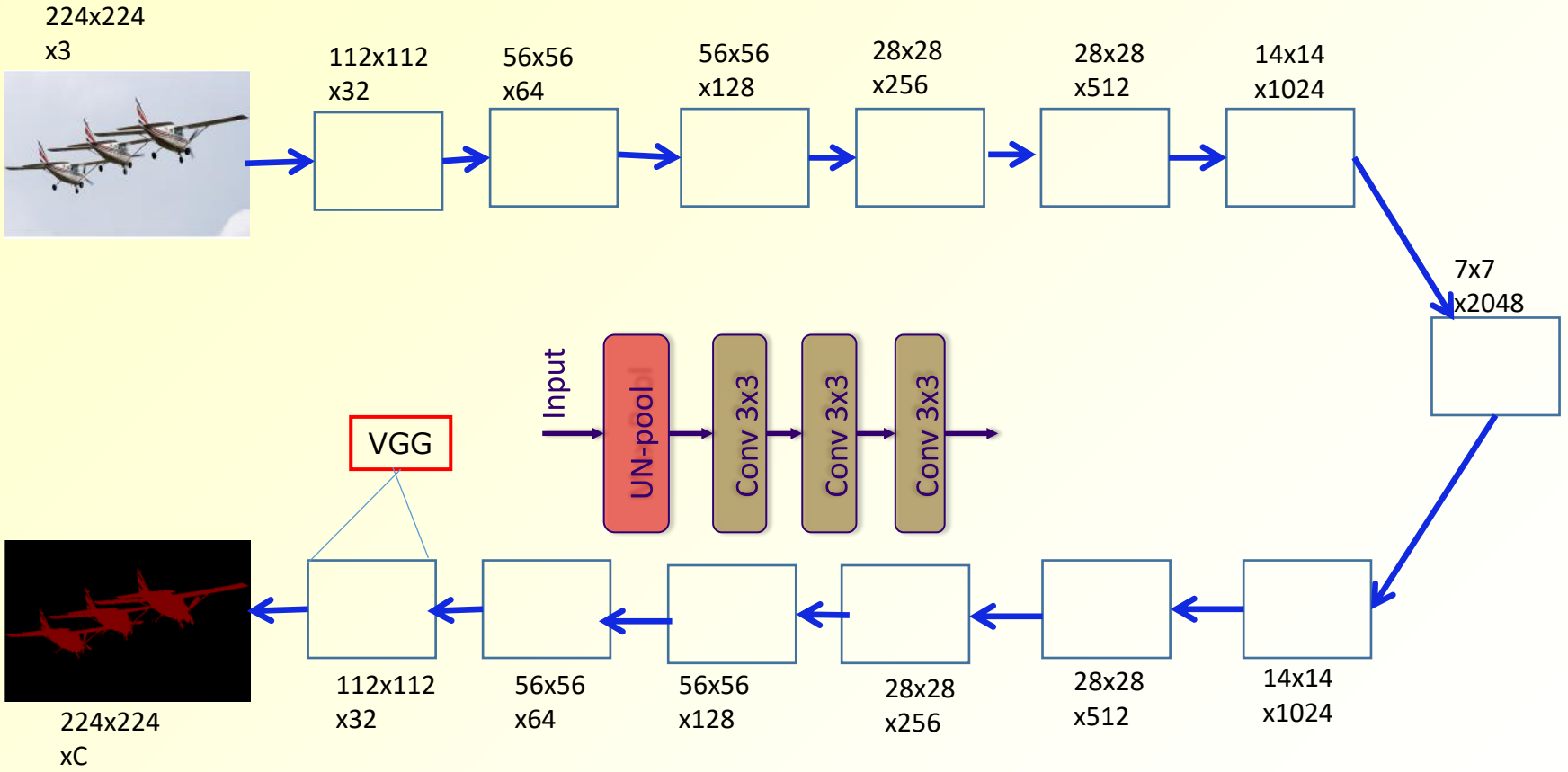
Inception (Codificador)



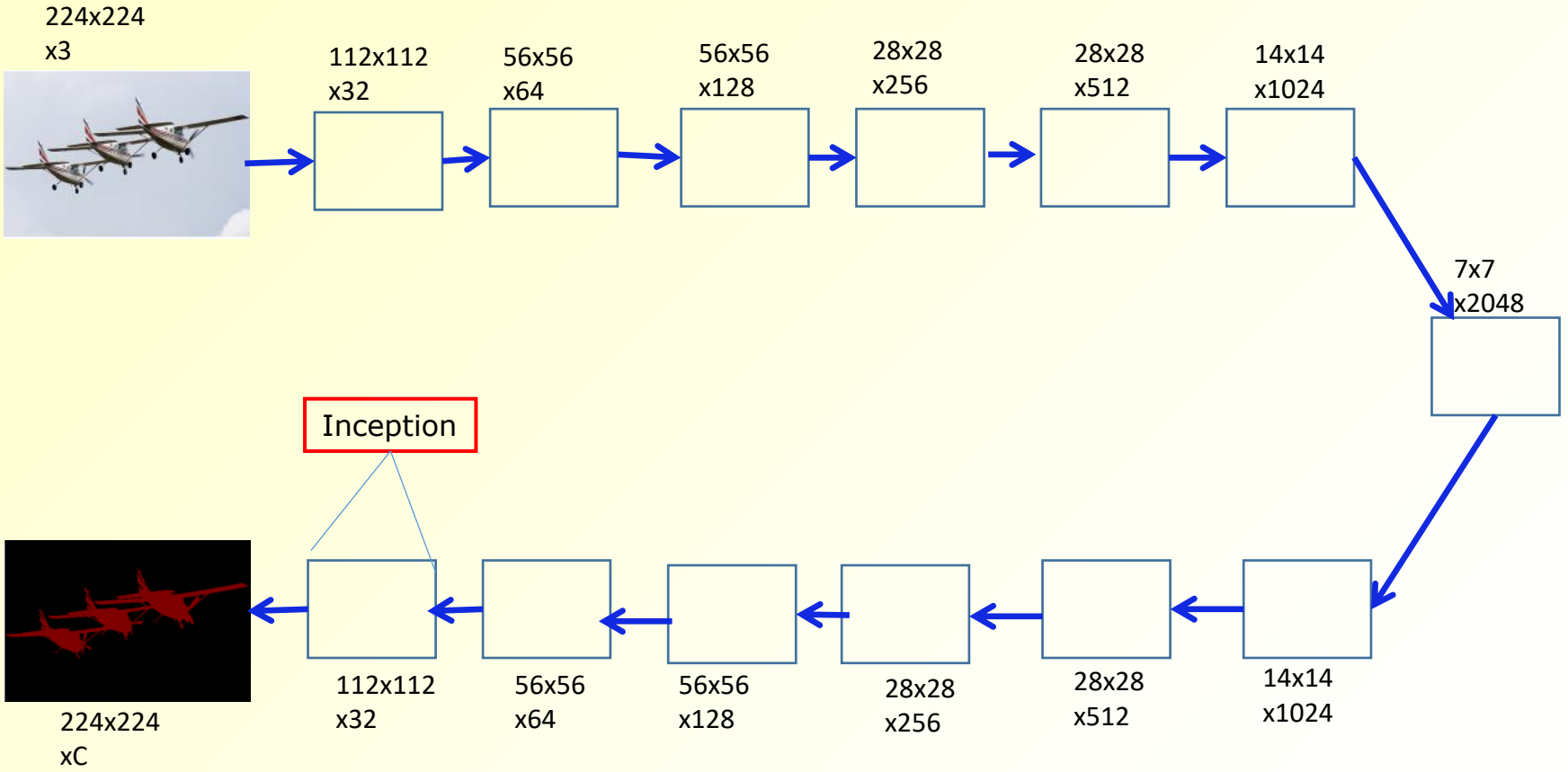
Arquitectura Autocodificador (Red Codificadora-Decodificadora)



Arquitectura Autocodificador (Red Codificadora-Decodificadora)

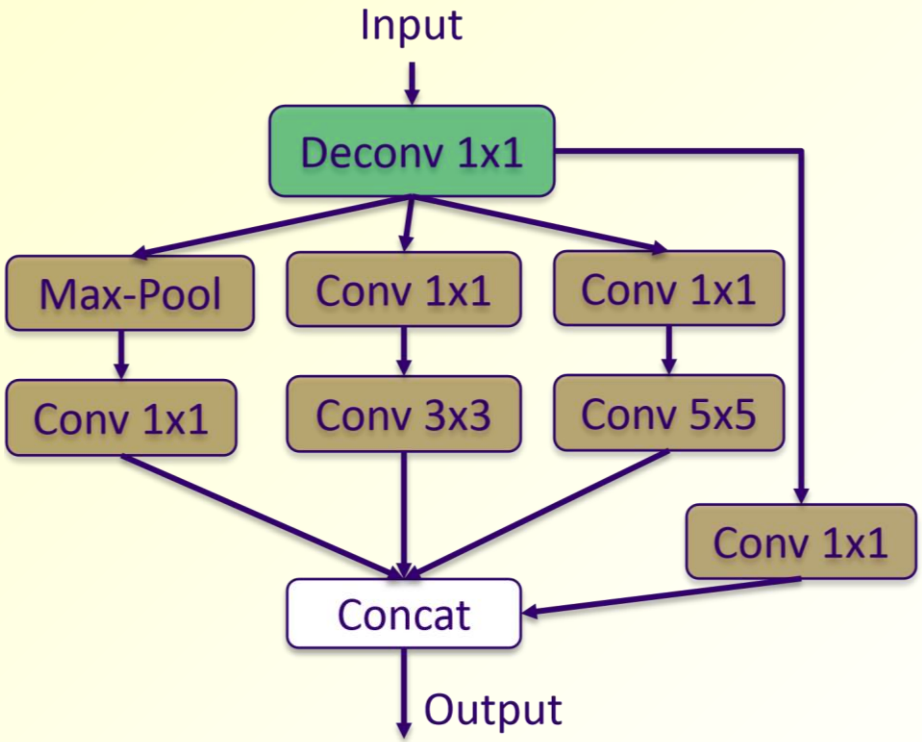


Arquitectura Autocodificador (Red Codificadora-Decodificadora)

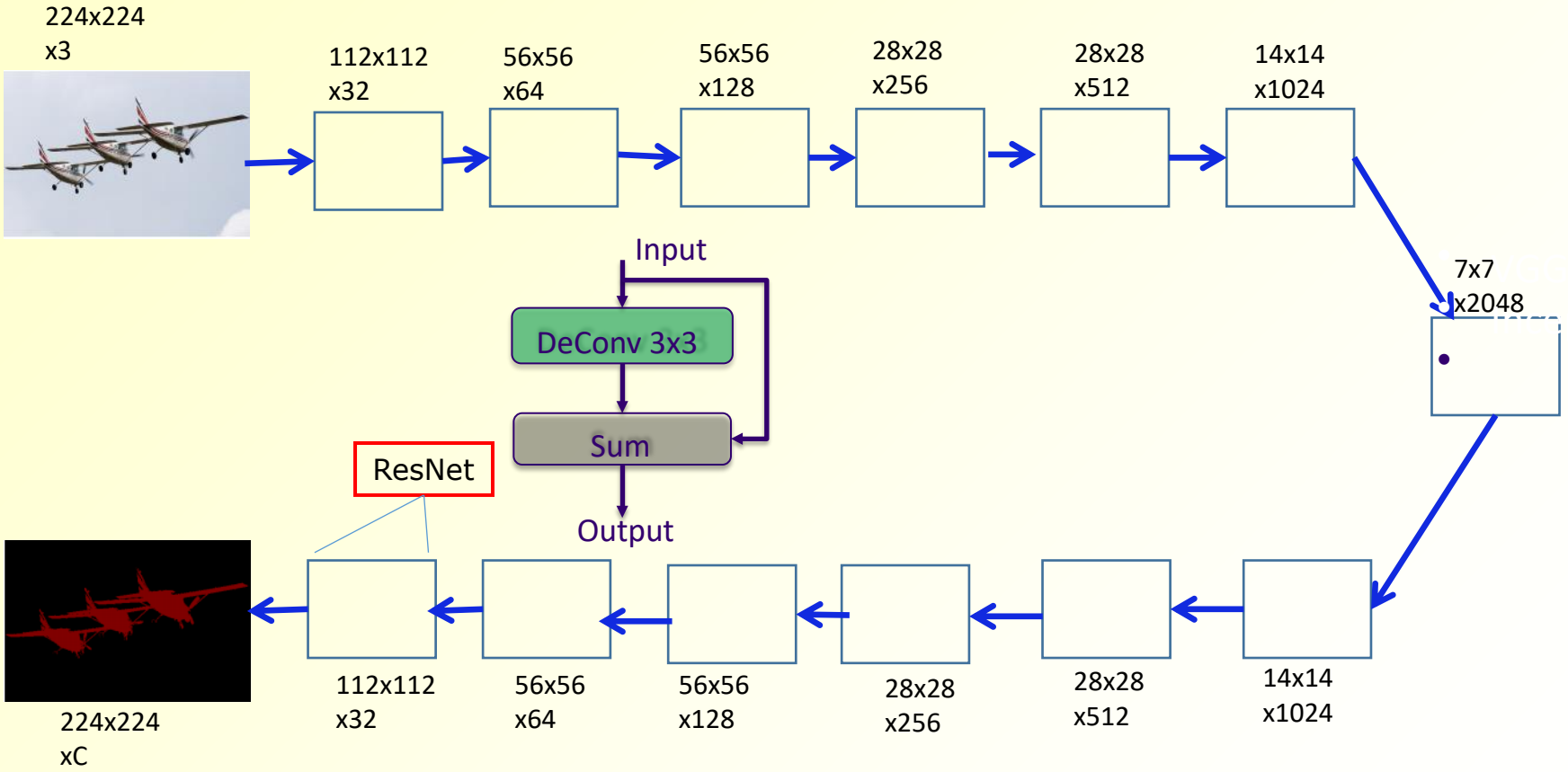


Arquitectura Autocodificador (Red Codificadora-Decodificadora)

Inception (Decodificador)



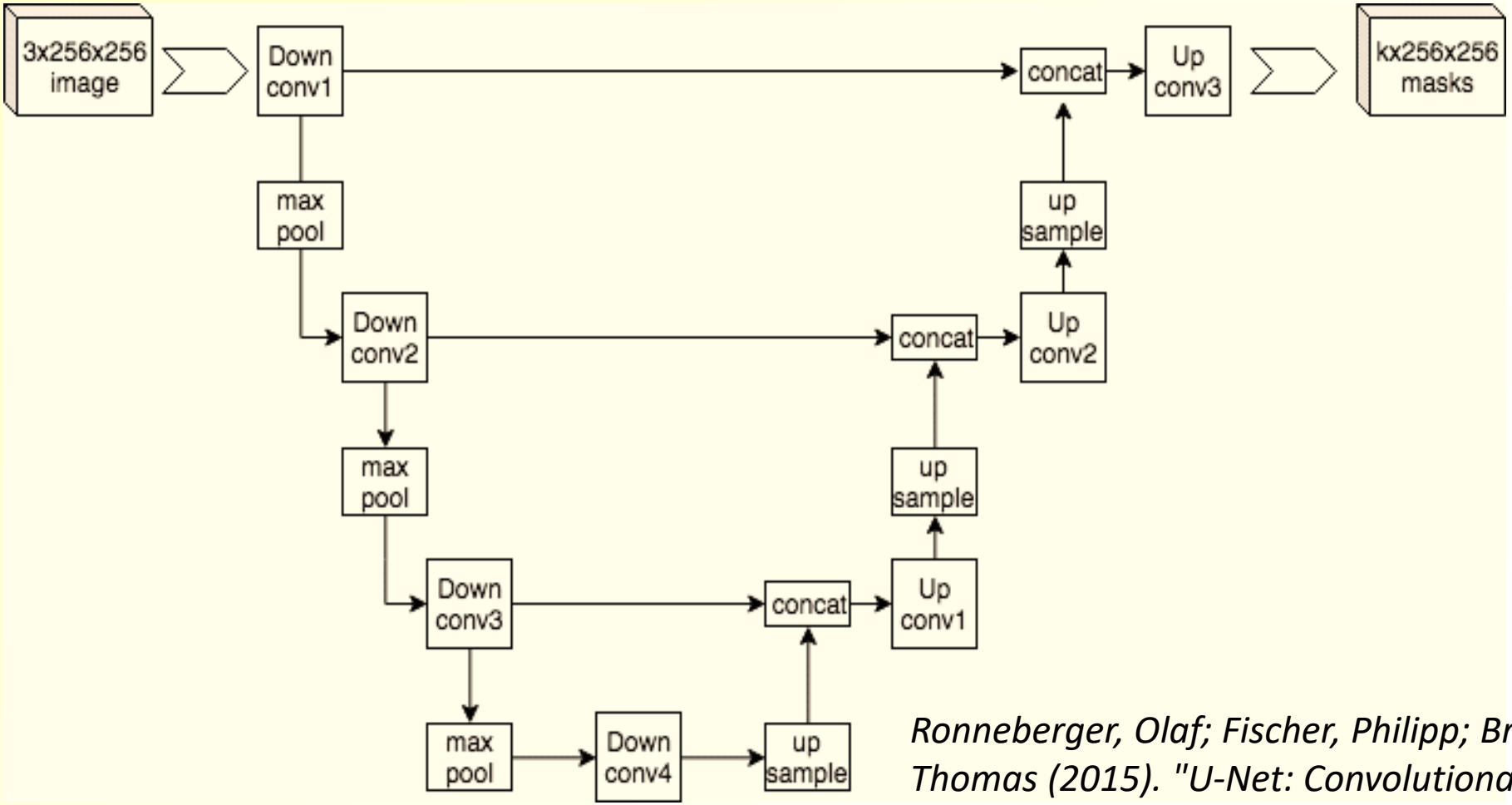
Arquitectura Autocodificador (Red Codificadora-Decodificadora)



Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

U-Net



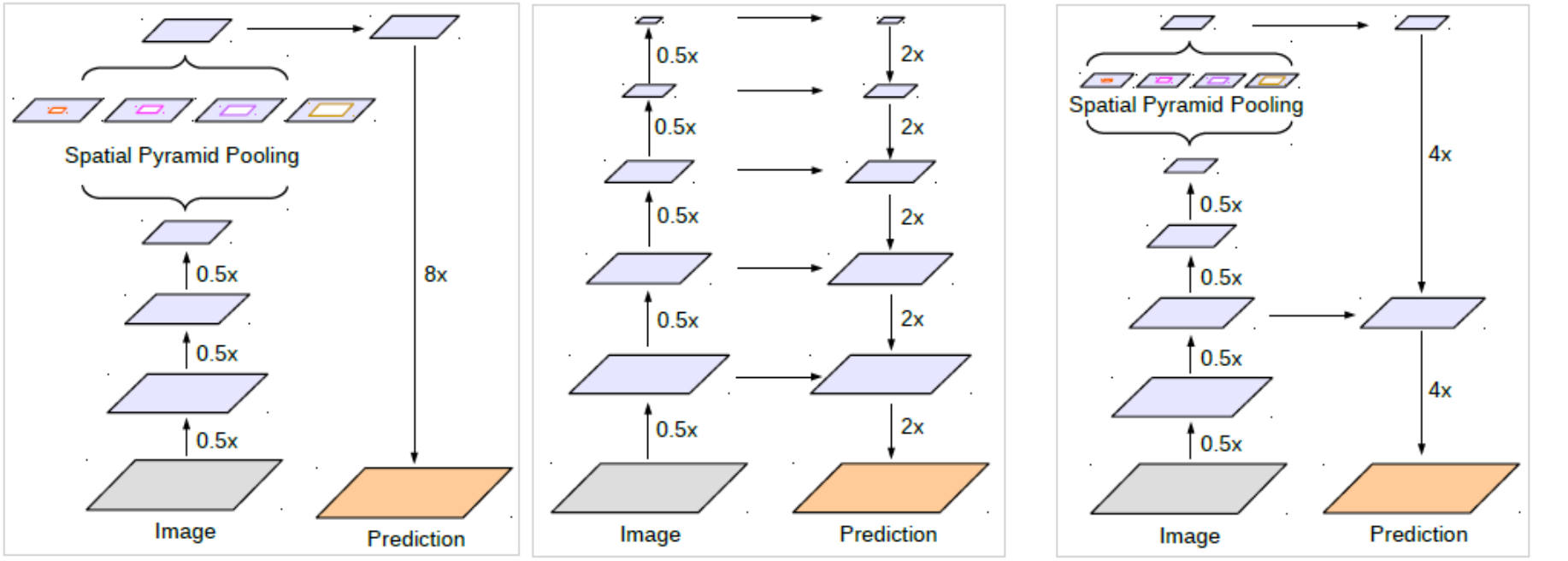
Ronneberger, Olaf; Fischer, Philipp; Brox, Thomas (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation".

Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

DeepLabv3: es un arquitectura codificadora-decodificadora con el objetivo de dar una segmentación semántica de la imagen de entrada.

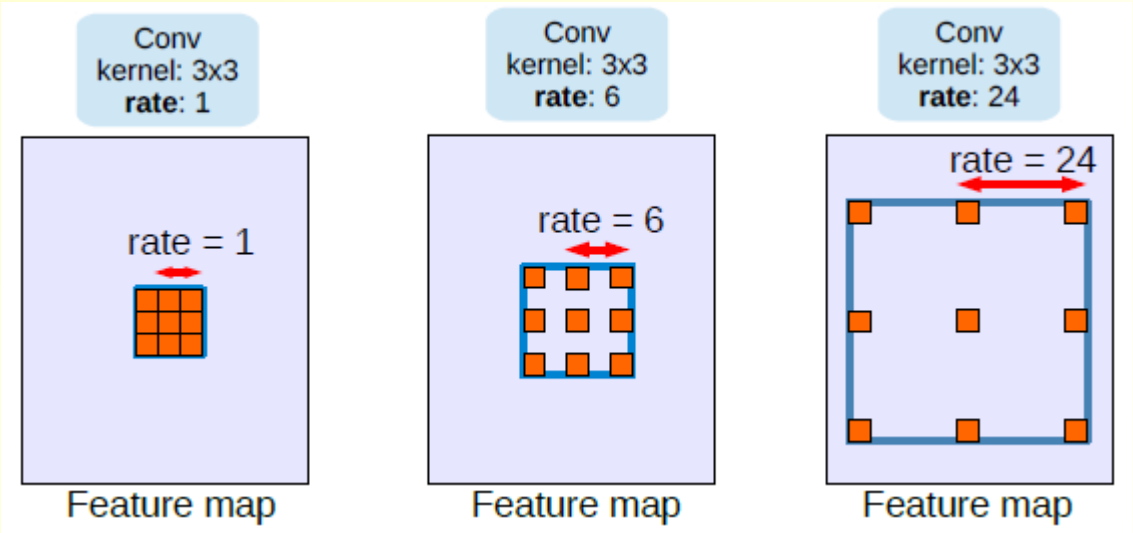
Encoder-Decoder with Atrous SeparableConvolution for Semantic Image SegmentationLiang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, andHartwig Adam. Google

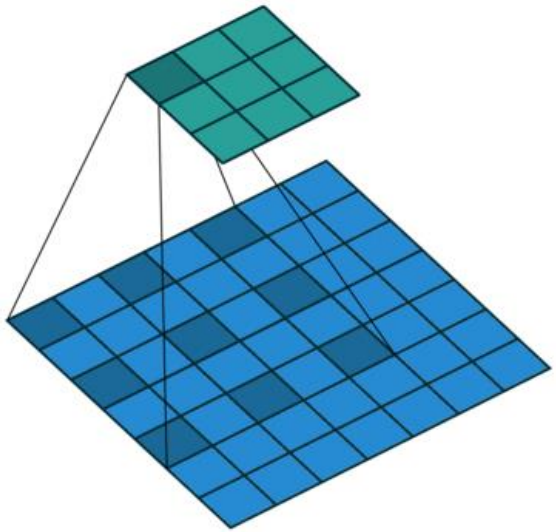


(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Convolución Atrous (convoluciones dilatadas)

$$y[i] = \sum_{k} x[i + r \cdot k] w[k]$$



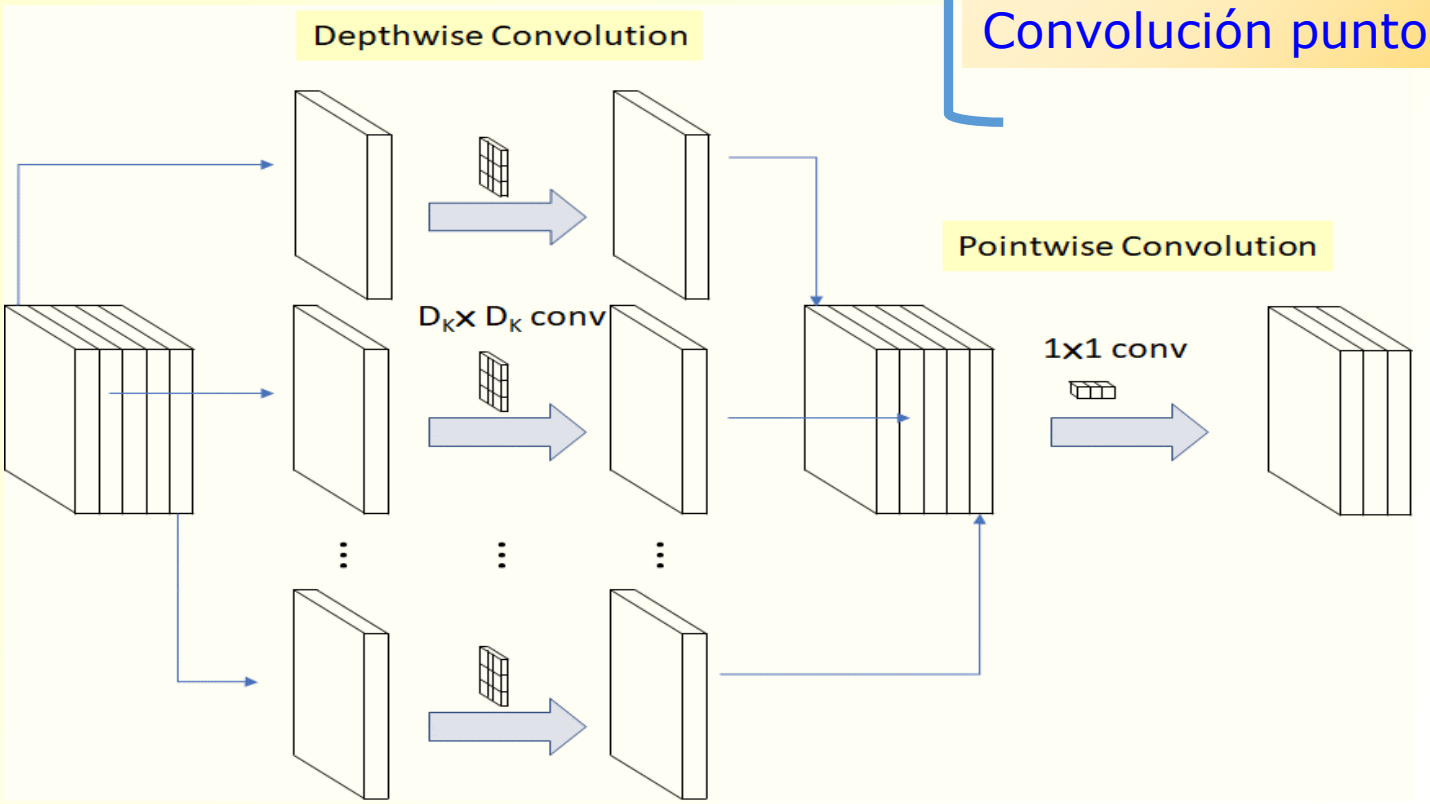


Convolución 2D con un filtro 3x3, radio de dilatación r=2 y descartando los bordes (no padding).

Convolución Atrous Separable

Convolución es separable

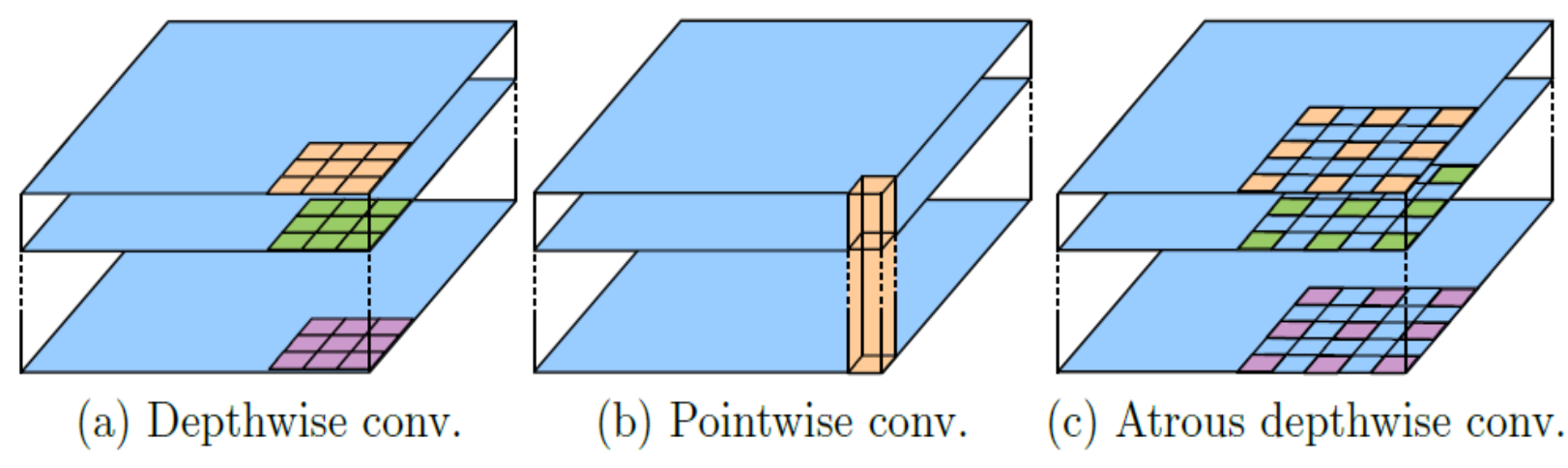
- Convolución en profundidad
- Convolución punto a punto



Convolución Atrous Separable :mejora la eficiencia computacional

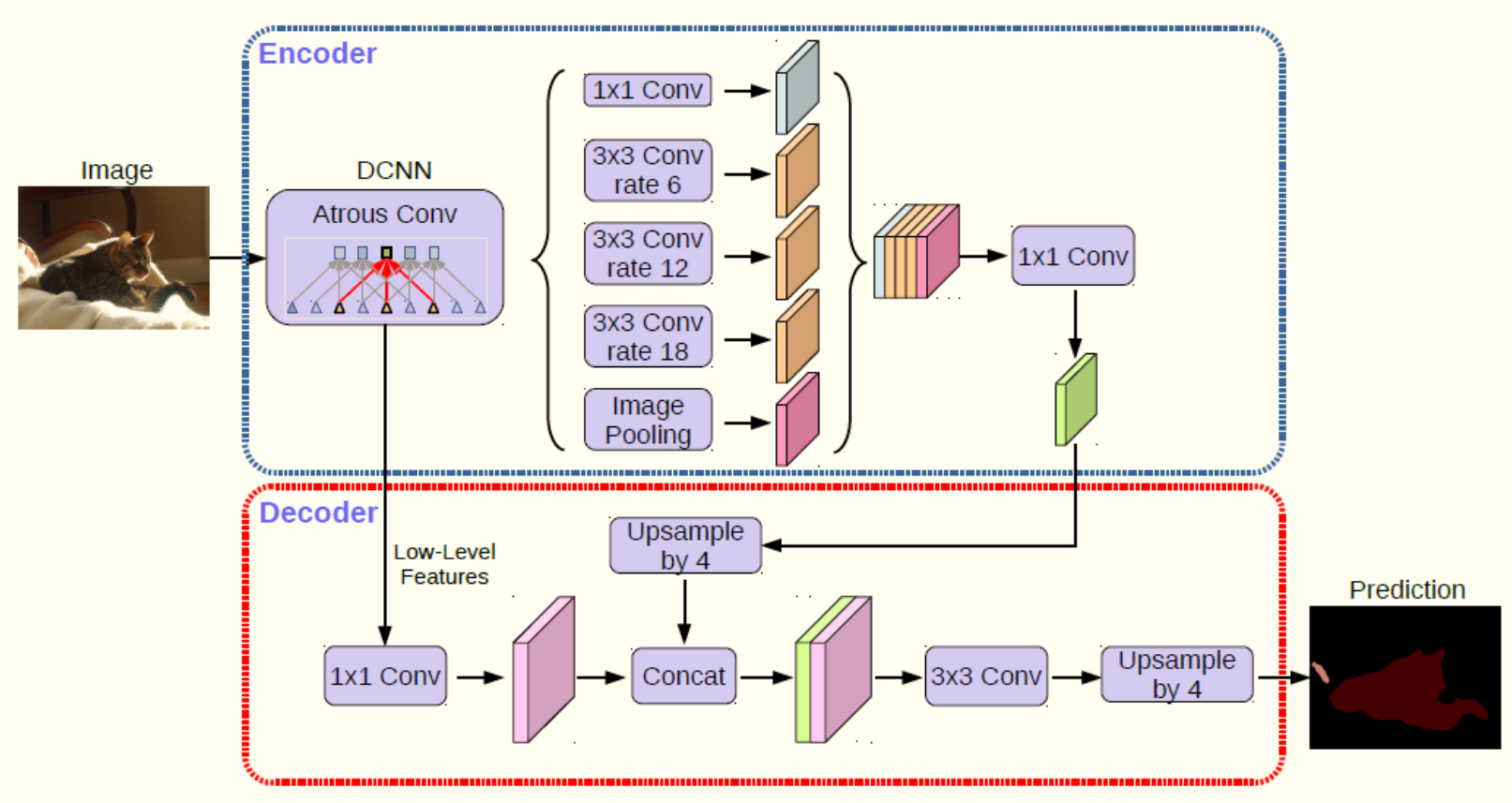
Convolución es separable

- Convolución en profundidad
- Convolución punto a punto



(a) Depthwise conv. (b) Pointwise conv. (c) Atrous depthwise conv.

Codificador-Decodificador



Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

Metricas. IoU

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

Visión Transformers

"An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale," published at ICLR 2021
“Attention Is All You Need” (Vaswani et al., 2017).

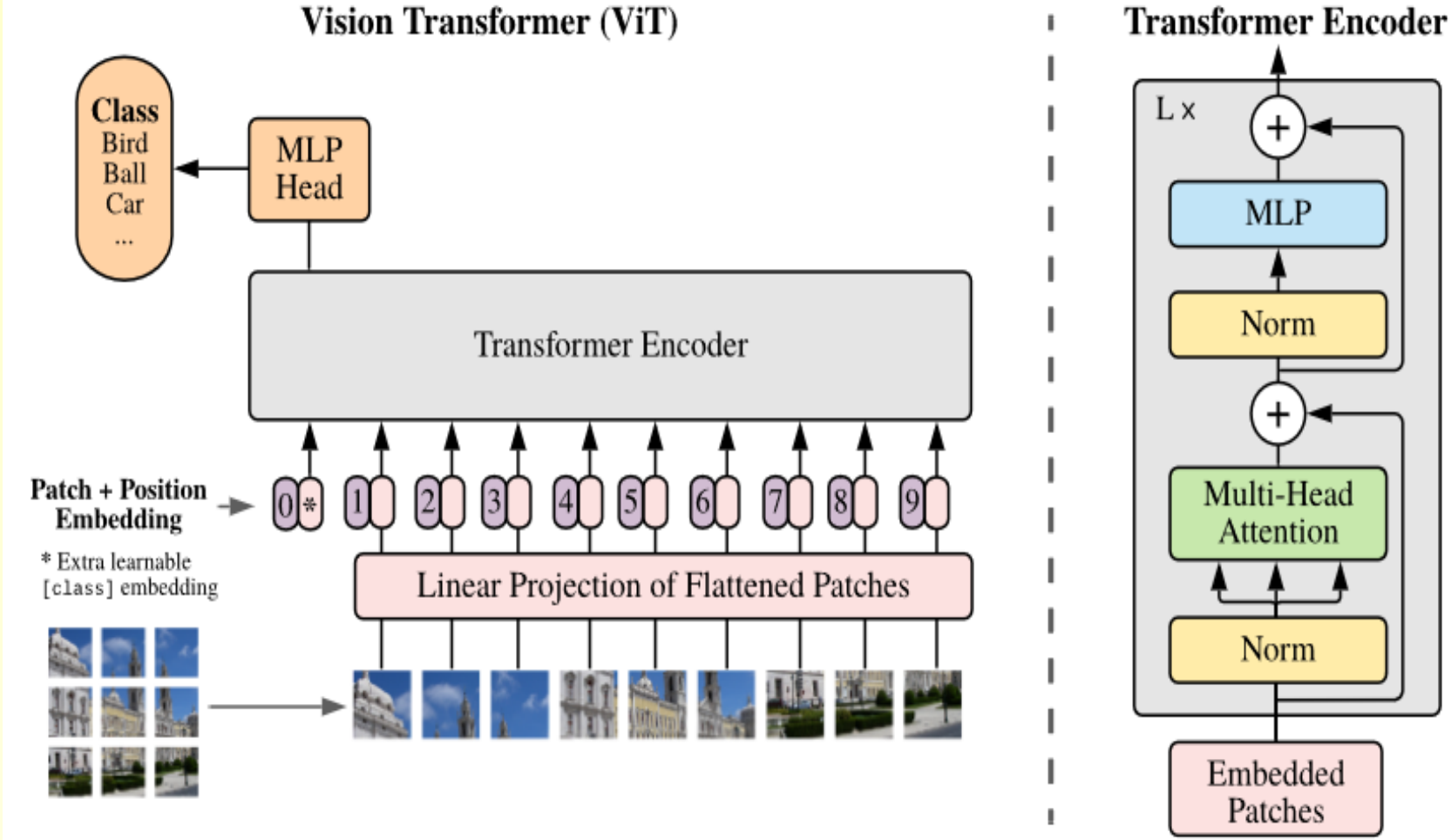
Transformers.- Es un modelo de aprendizaje profundo que adopta mecanismos de auto-atención para aprender relaciones entre elementos de una secuencia, ponderando de forma diferente el significado de cada parte de la entrada. Usan capas denominadas *capas de atención*

Pasos que se realizan en el transformers.

1. Dividir una imagen en regiones (patches)
2. Aplanar las regions y proyectar en un espacio de más baja dimensión
3. Añadir informacion de posicion
4. Introducir la secuencia como entrada al encoder del transformer
5. Entrenar con las etiquetas de la imagen
6. Fine-tune

Visión Transformers

"An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale," published at ICLR 2021



Caracteristicas Transformers.-

- Auto-Atención: captura las dependencias a “largo-término” entre elementos de la secuencia
  - Dado un conjunto de trozos en la imagen estima la relevancia de una región con respecto a otra.
  - Capa de auto-atención codifica cada trozo en términos de todos los trozos.

Sea X la secuencia con los trozos de la imagen ya aplanados. X se proyecta usando tres matrices:

$$Q = XW^Q, K = XW^K, V = XW^V$$

$$Z = \text{softmax}\left(\frac{\mathbb{Q}K^T}{\sqrt{d_q}}\right)V$$

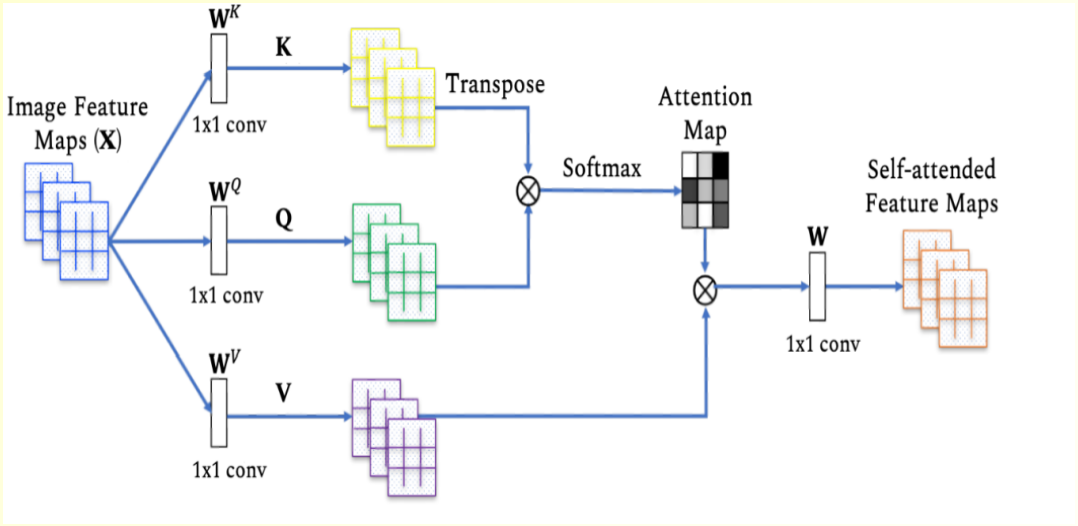
Capa de Auto-Atención

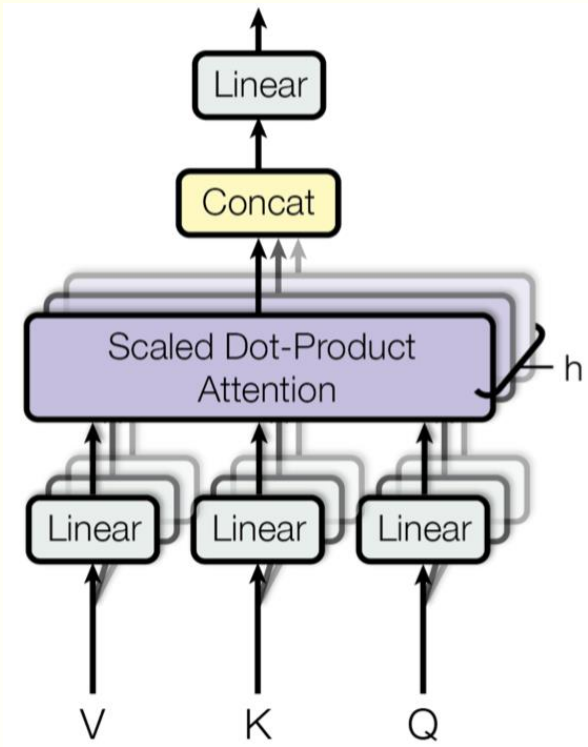
Transformers.-

Sea X la secuencia con los trozos de la imagen ya aplanados. X se proyecta usando tres matrices: $Q = XW^Q$, $K = XW^K$, $V = XW^V$

Capa de Auto-Atención

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{dq}}\right)V$$





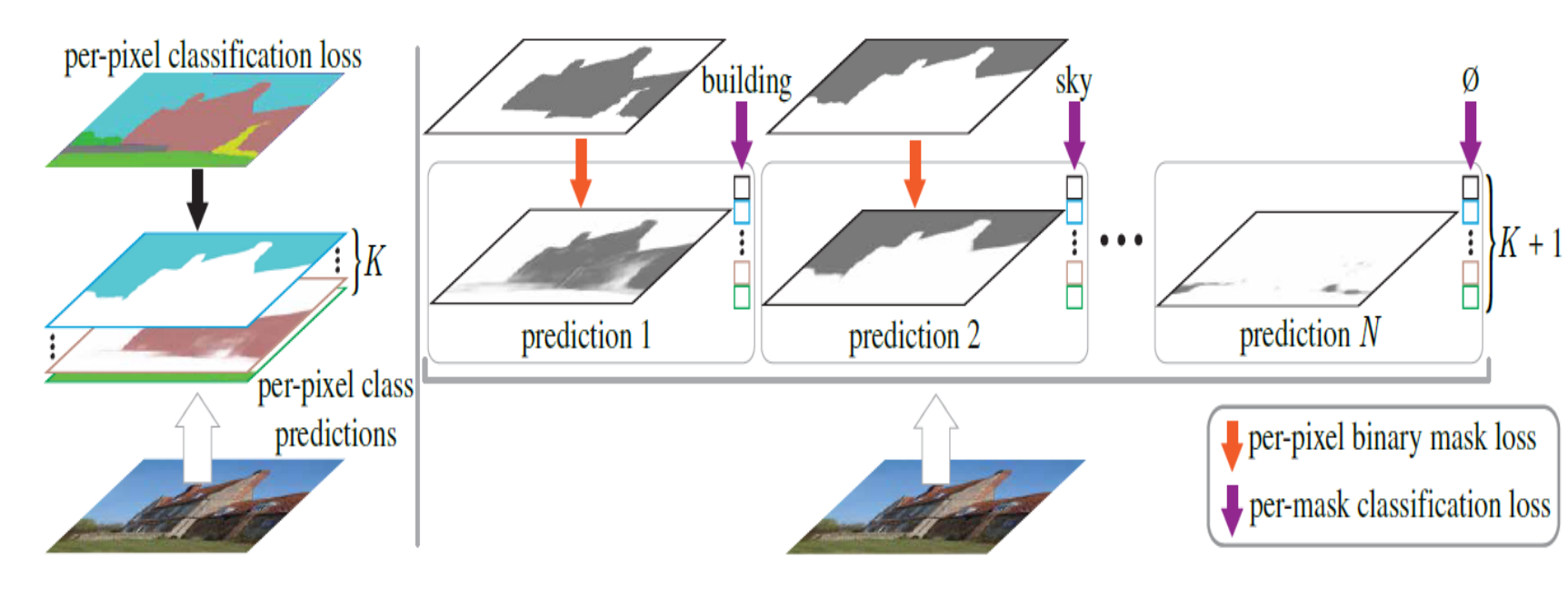
Multi-head Attention

Segmentación con Deep Learning

- *Conceptos*
- *Redes neuronales Autocodificadoras*
- *U-net*
- *DeepLabv3+*
- *Métricas. IoU*
- *Vision Transformer*
- *MaskFormer*

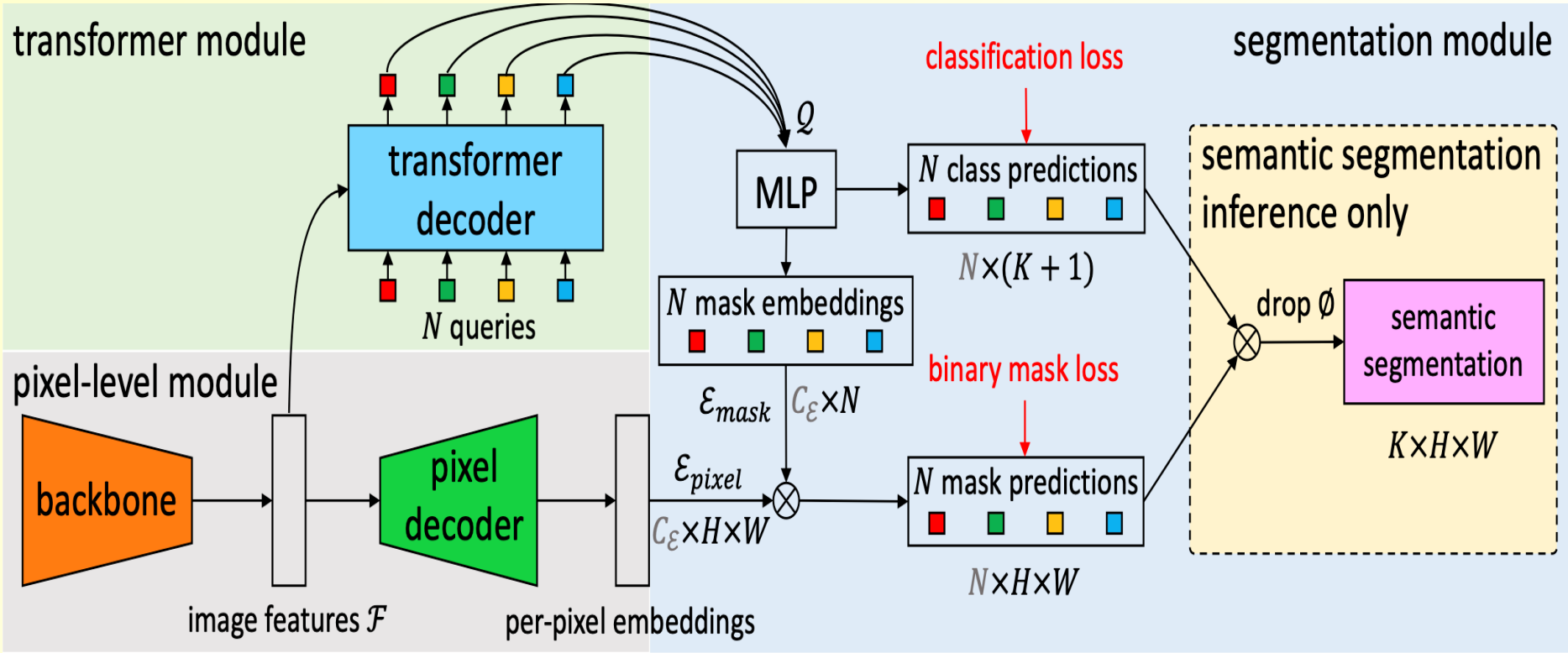
Cheng, B., Schwing, A. G., & Kirillov, A. (2021). Per-Pixel Classification is Not All You Need for Semantic Segmentation https://doi.org/https://arxiv.org/abs/2107.06278v2

MaskFormer es un modelo de clasificación de máscaras que predice un conjunto de máscaras binarias, cada una de ellas asociada a un etiqueta (persona, gato,etc).



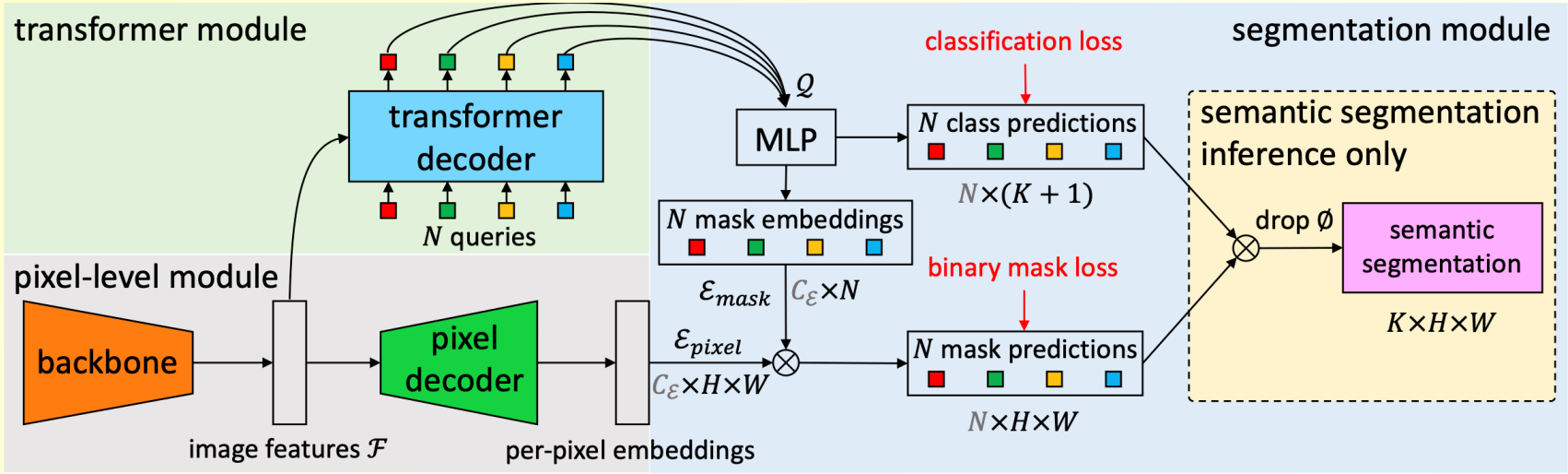
Clasificación por pixel vs Clasificación de máscaras

Cheng, B., Schwing, A. G., & Kirillov, A. (2021). Per-Pixel Classification is Not All You Need for Semantic Segmentation https://doi.org/https://arxiv.org/abs/2107.06278v2



MaskFormer

Cheng, B., Schwing, A. G., & Kirillov, A. (2021). Per-Pixel Classification is Not All You Need for Semantic Segmentation https://doi.org/https://arxiv.org/abs/2107.06278v2

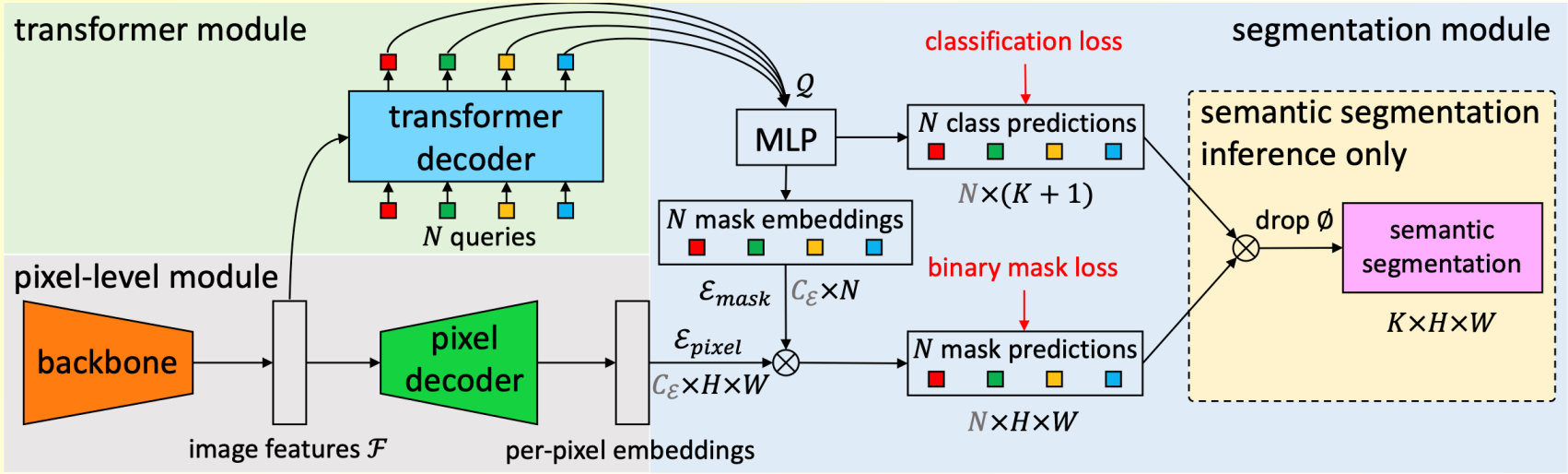


Backbone: Uso ResNet con 50 y 101 capas. Aunque admite otras arquitecturas.

Pixel decoder: Usa un decodificador semántico (Unet). Sobre-muestrea x2 los rasgos de más baja resolución y lo suma con los rasgos obtenidos en el Backbone a la correspondiente resolución

MaskFormer

Cheng, B., Schwing, A. G., & Kirillov, A. (2021). Per-Pixel Classification is Not All You Need for Semantic Segmentation https://doi.org/https://arxiv.org/abs/2107.06278v2



Transformer decoder.-Entrada son N queries. Cada queries esta asociada con una posición encapsulada. Se inicia a 0.

Modulo de Segmentación.- EL bloque MLP tiene 2 capas ocultas de 256 canales.

Ecuaciones para la Clasificación con Máscaras .- K clases

1. Divide la imagen en N regiones $\{m_i | m_i \in \{0,1\}^{H \times W}\}^N_{i=1}$
2. A cada región le asigna un vector de probabilidades indicando la probabilidad de que esa región represente un objeto de un tipo. Siendo K el total de clases

$$\mathcal{L}_{\text{mask-cls}}(z, z^{\text{gt}}) = \sum_{j=1}^{N} \left[-\log p_{\sigma(j)}(c_j^{\text{gt}}) + \mathbb{1}_{c_j^{\text{gt}} \neq \varnothing} \mathcal{L}_{\text{mask}}(m_{\sigma(j)}, m_j^{\text{gt}}) \right]$$

z: $regiones\ predichas$ z^{gt}: $mascaras\ reales$

$p_{\sigma(j)}$: probabilidad de la mascara $\sigma(j)$ de ser de la clase j

$\mathcal{L}_{\text{mask}}(m_{\sigma(j)}, m_j^{\text{gt}})$: error como IoU o RMSE entre dos máscaras