UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

Sesión 3.1

Visual Attention

ViT, Swin transformer, CrossViT

TRANSFORMATEC

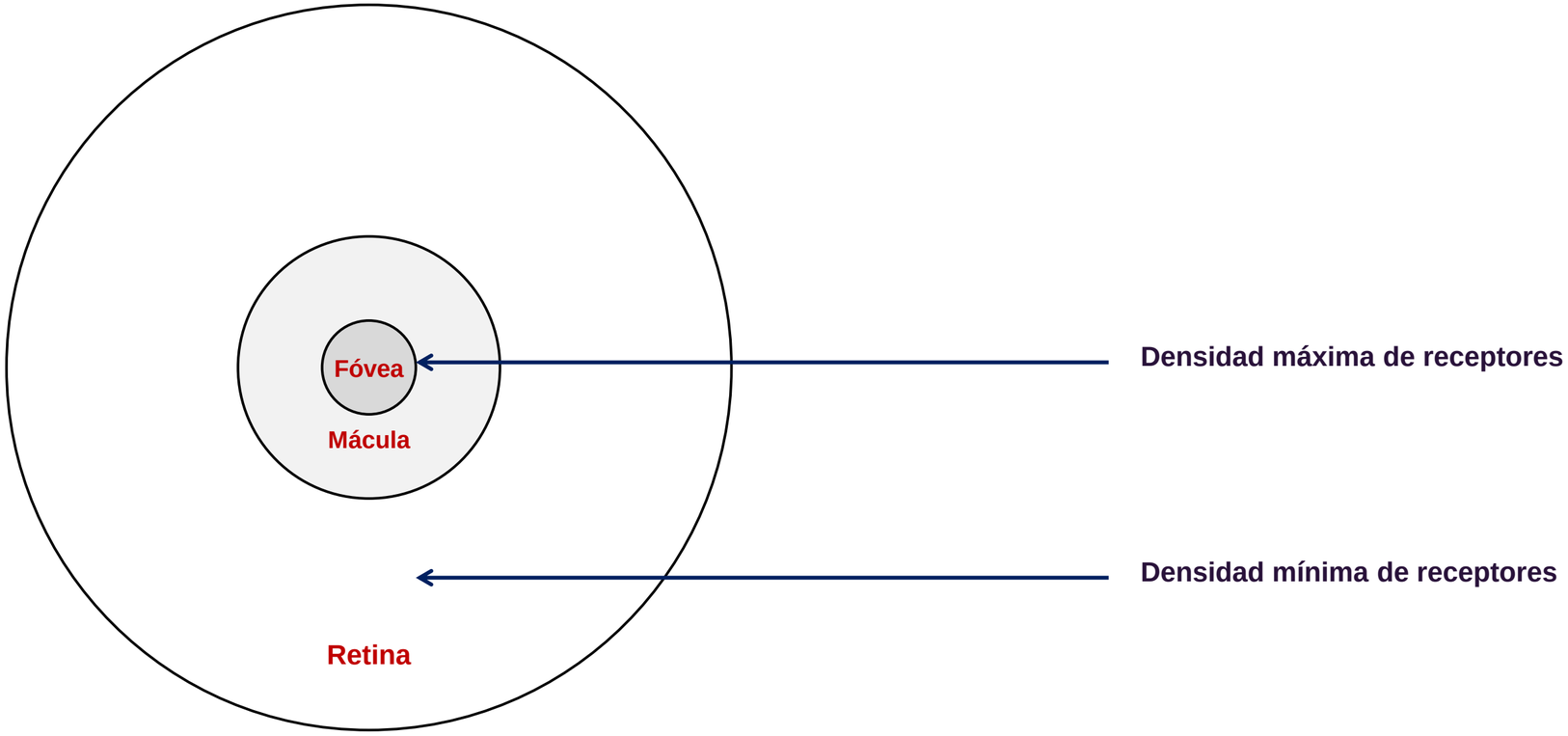
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

1.

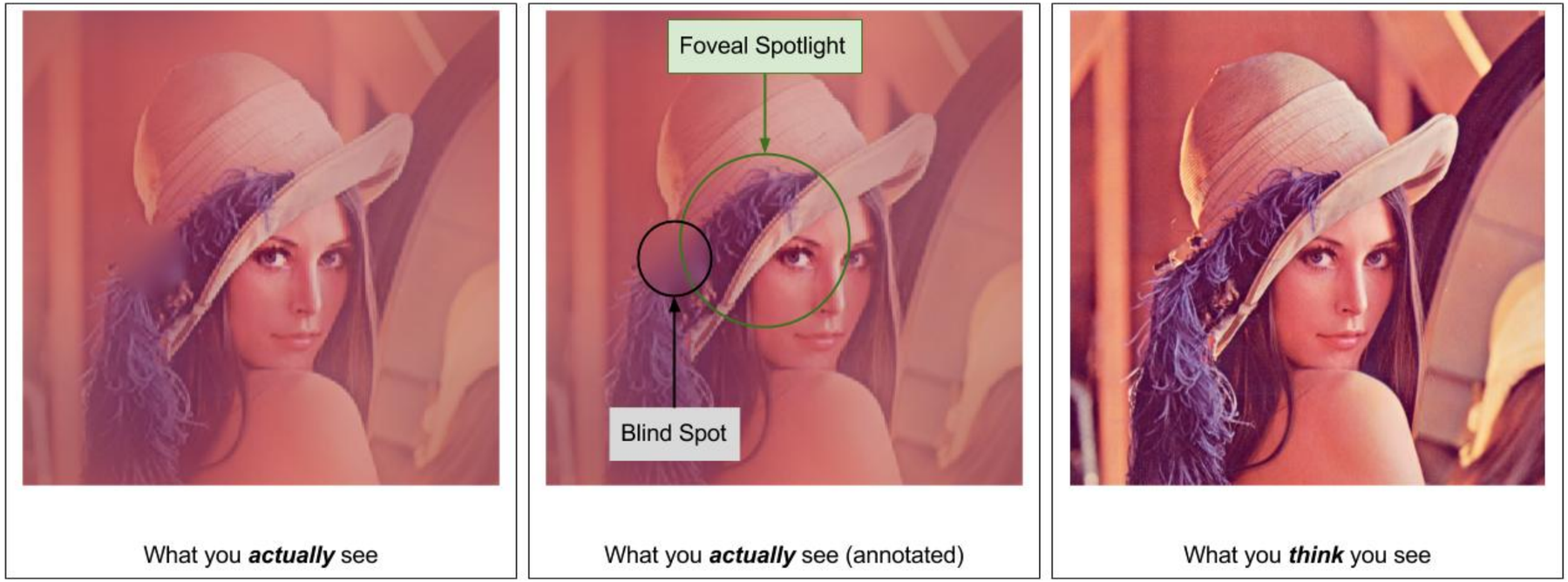
Atención *humana*

TRANSFORMATEC

Atención *visual humana*



Atención *visual humana*

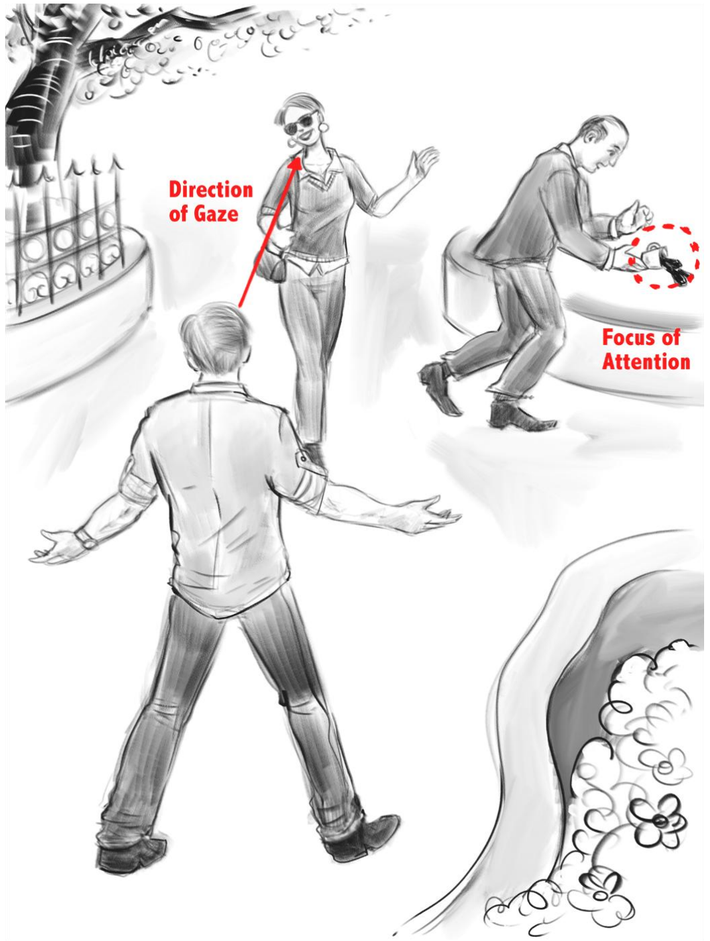


Modalidades *de atención visual*

Overt Attention

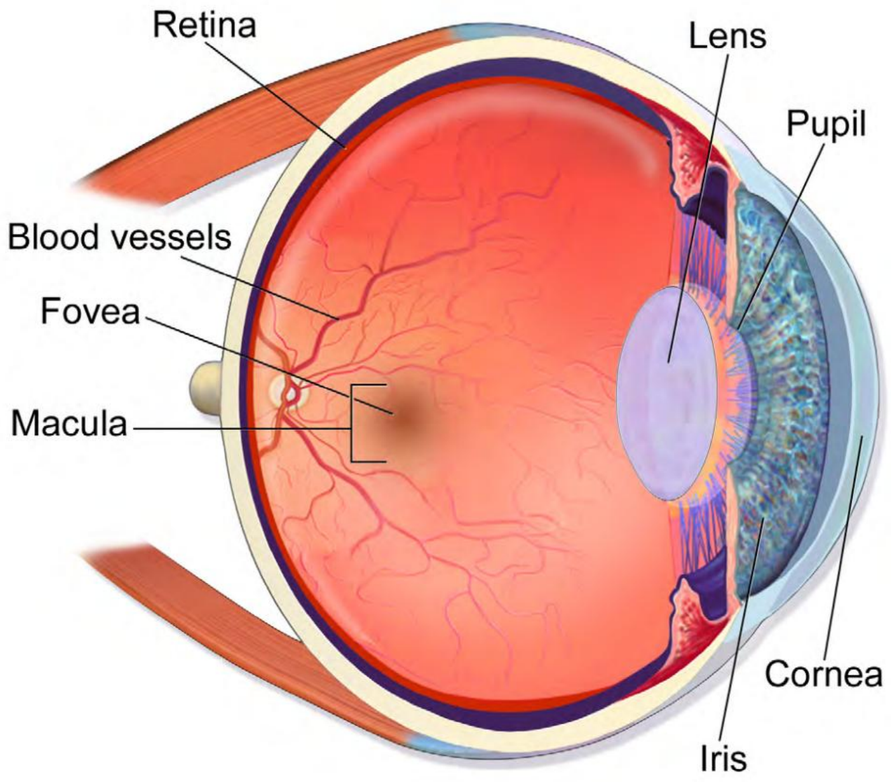
vs

Covert Attention

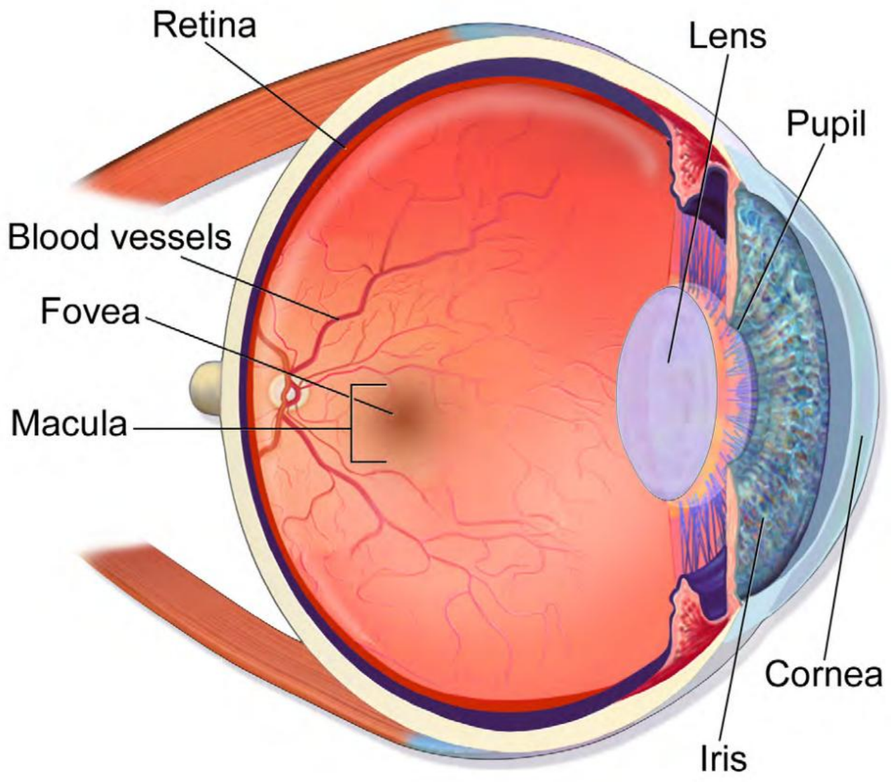


Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

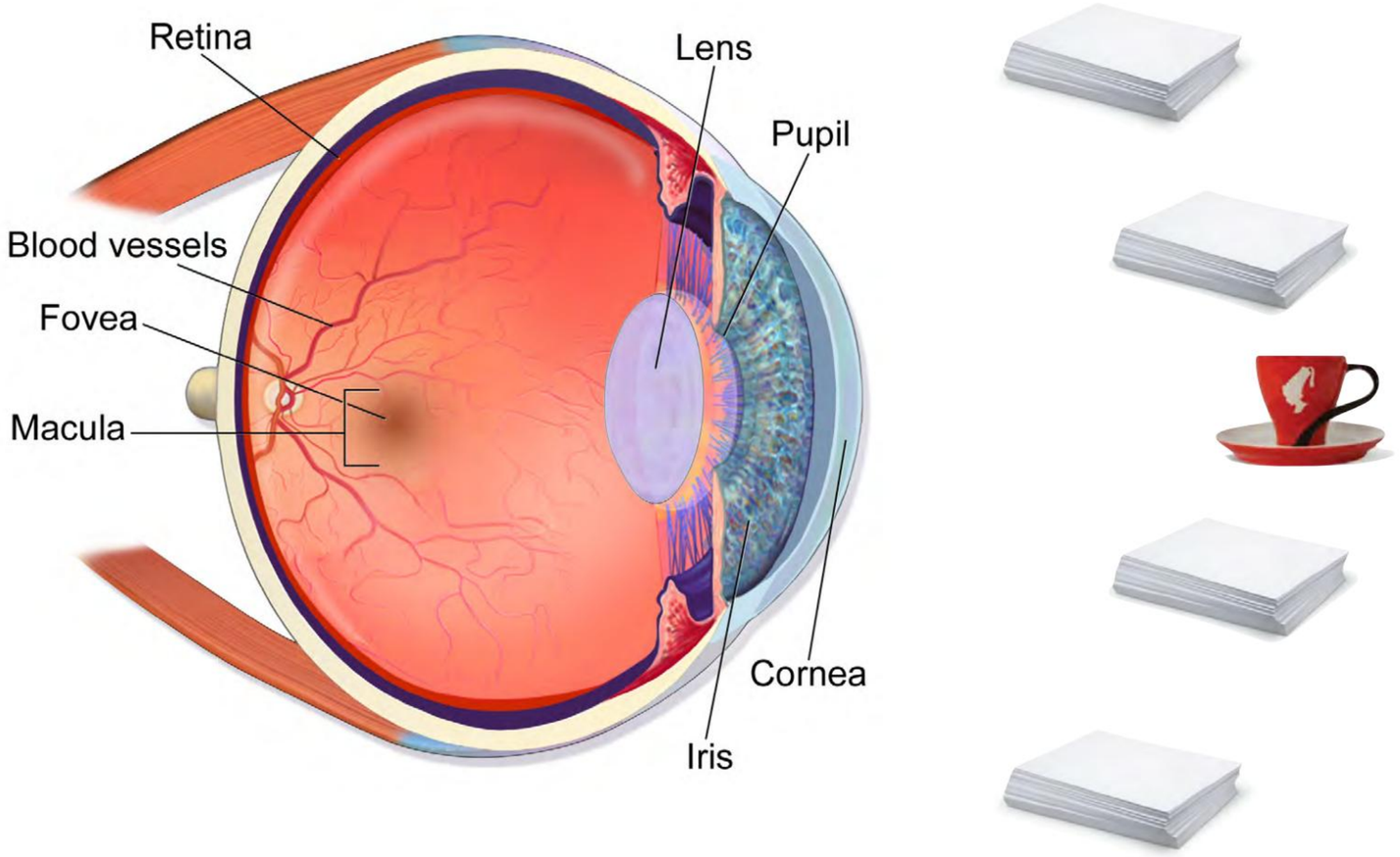
Overt *Attention*



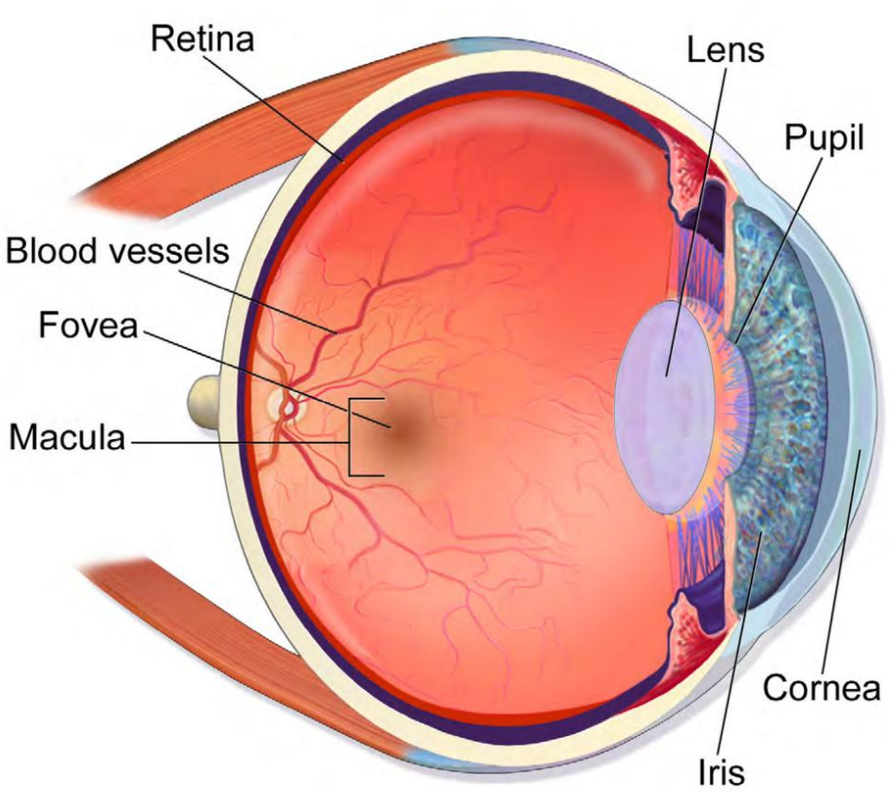
Overt *Attention*

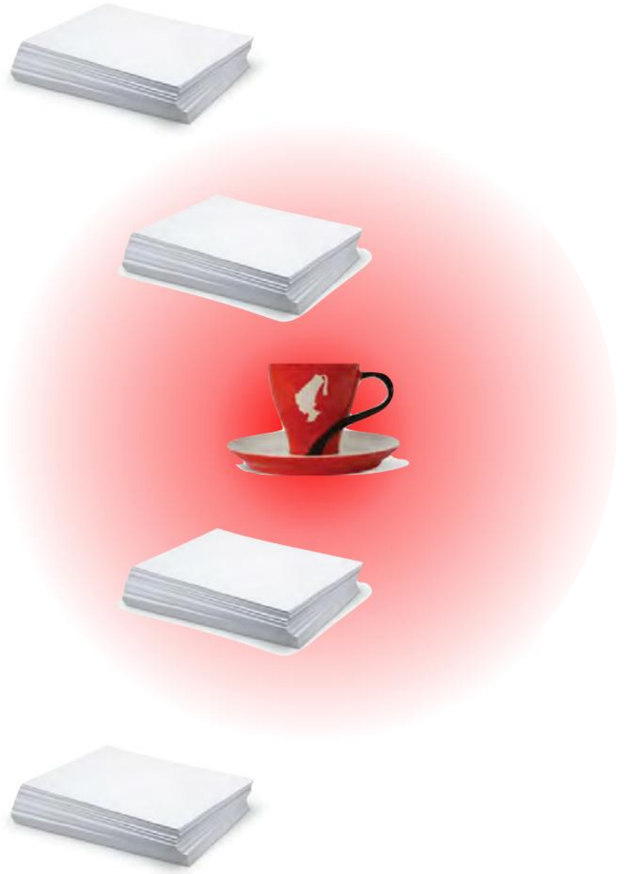


Overt *Attention*

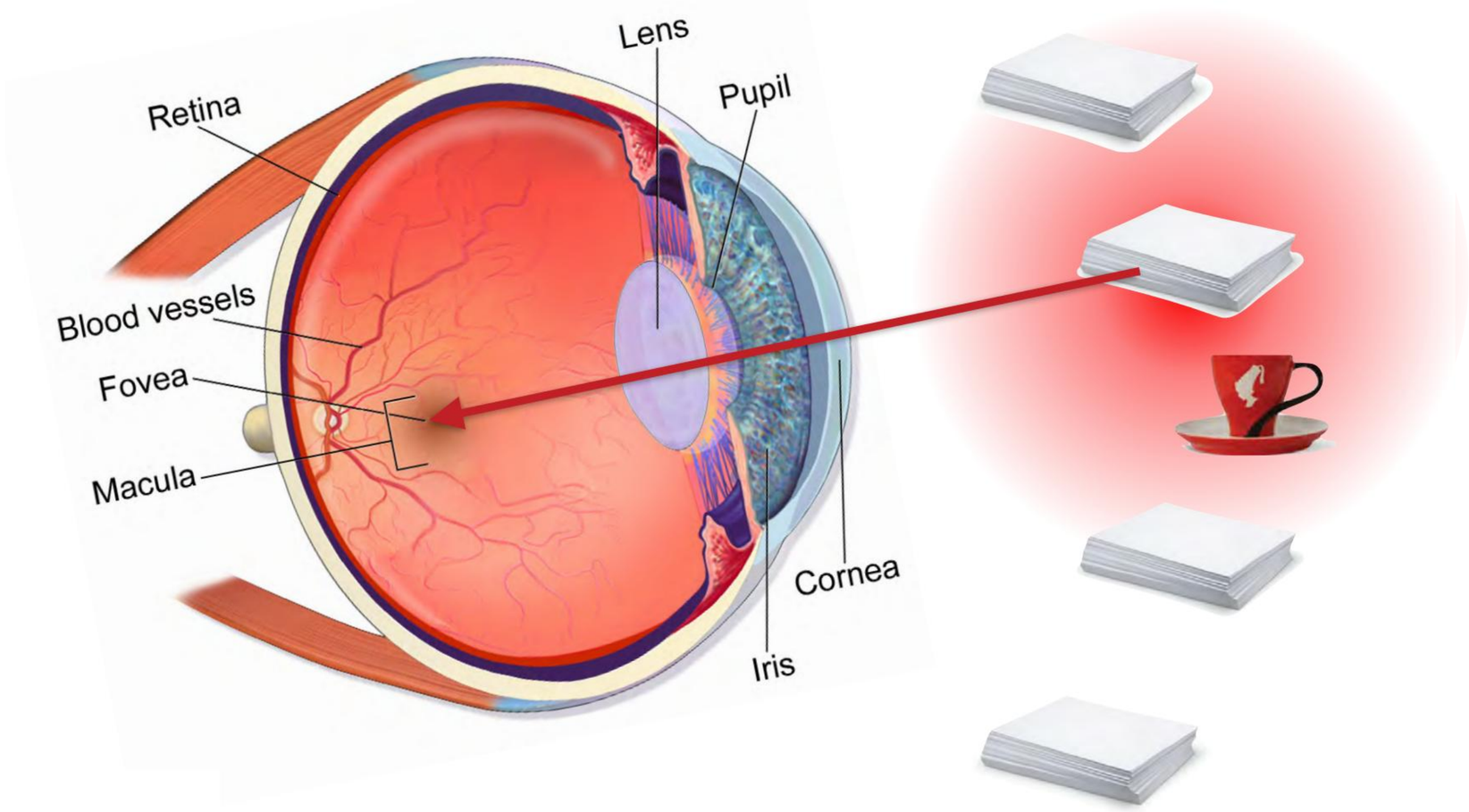


Overt *Attention*

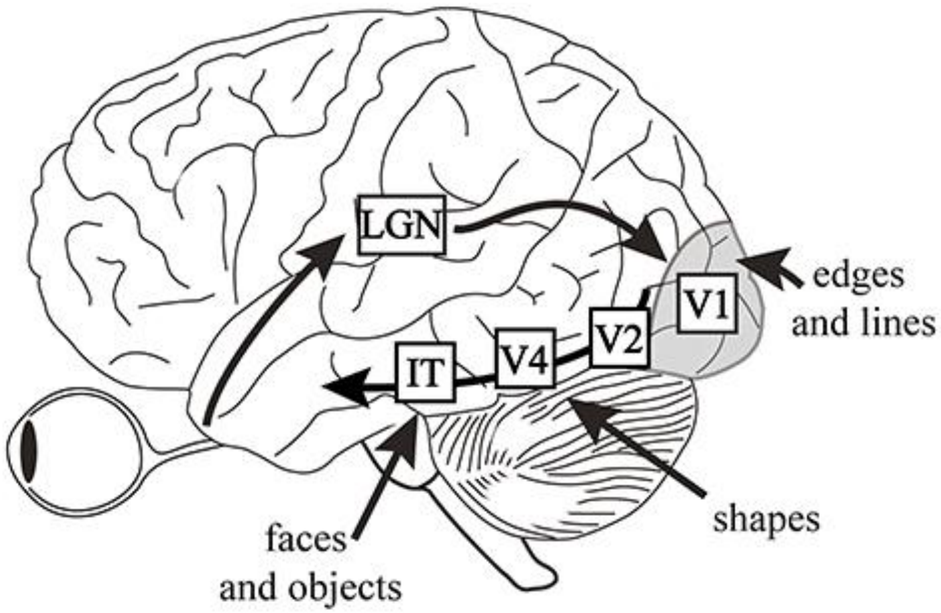


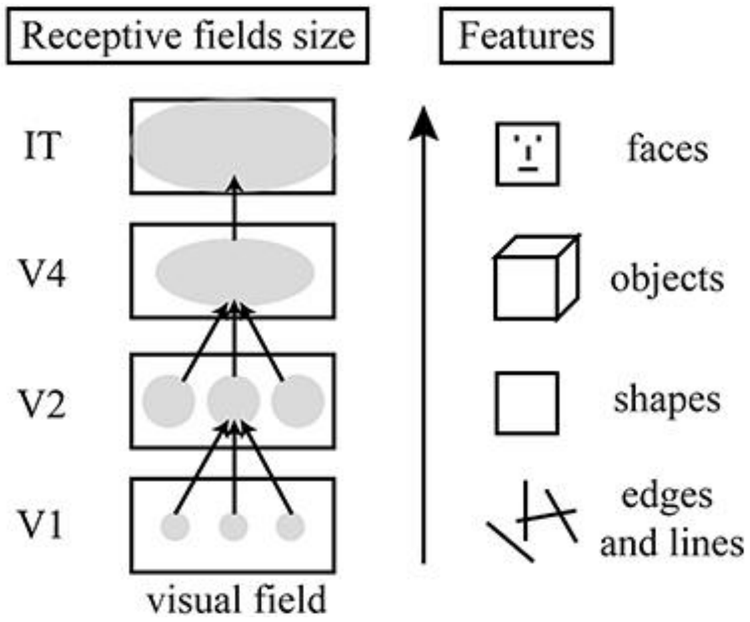


Overt *Attention*



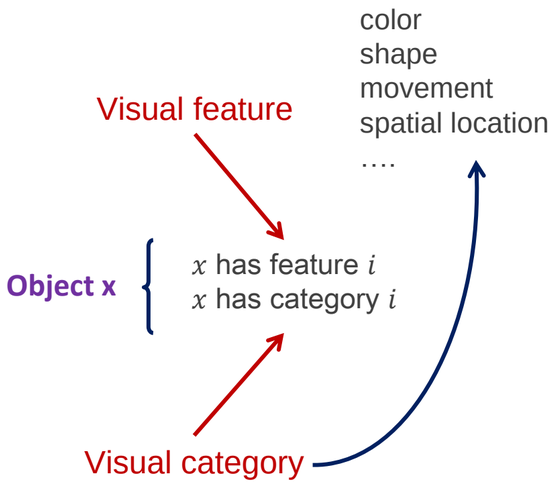
Visual *Cortex*

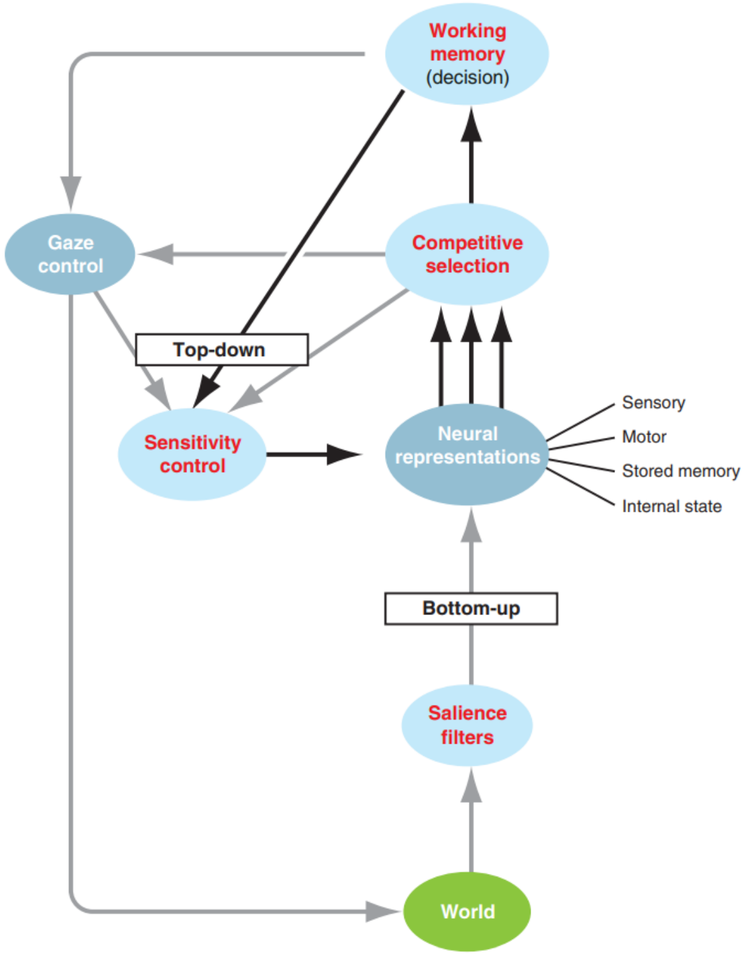




Covert *Attention*

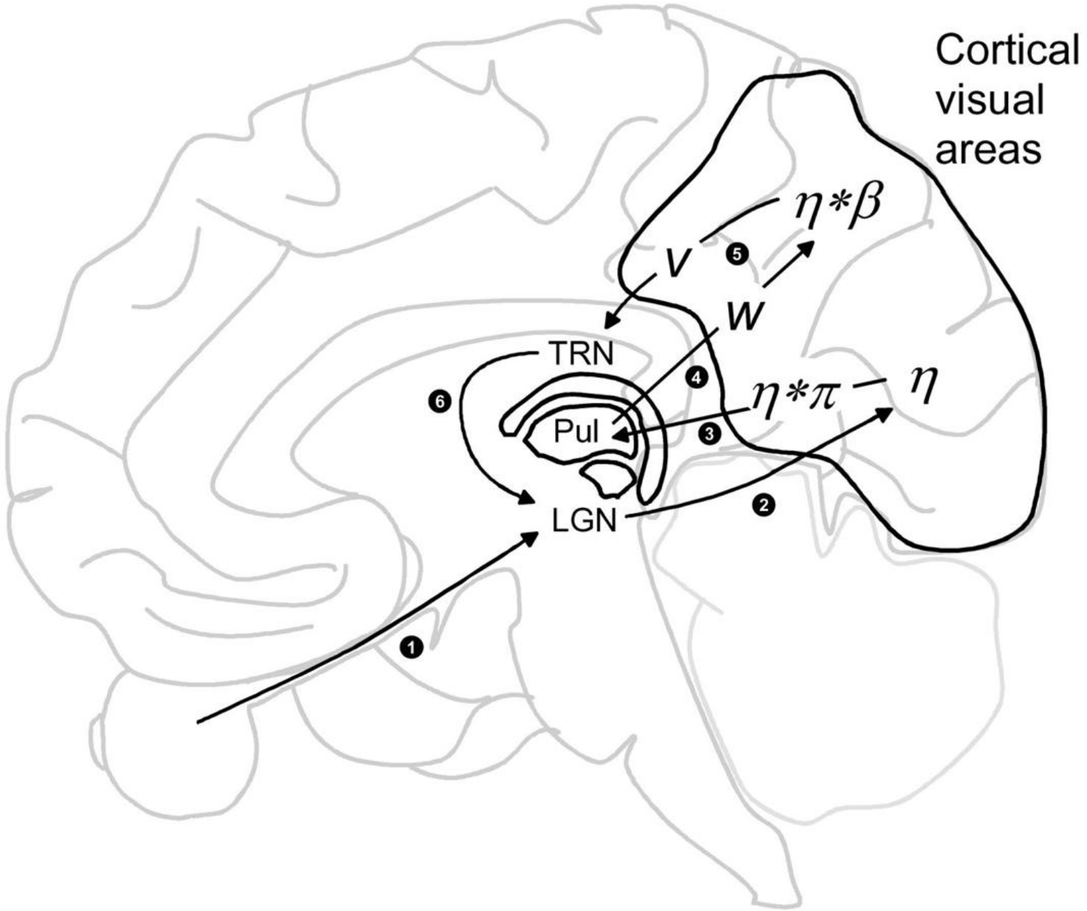
Theory of Visual Attention





Bundesen et al. (2011) "A neural theory of visual attention and short-term memory (NTVA)". Neuropsychologia, 49(6), 1446-1457.

Covert *Attention*



Bundesen et al. (2011) "A neural theory of visual attention and short-term memory (NTVA)". Neuropsychologia, 49(6), 1446-1457.

Atención *Humana*

“La atención es un proceso cognitivo básico en el cerebro, y consiste en la selección dinámica de entidades de entrada mediante la ponderación adaptativa acorde a la importancia de cada entrada.”

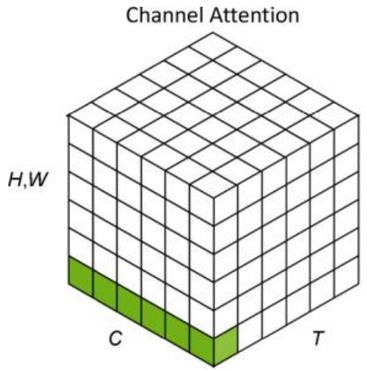
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

2.

Beyond *Transformers*

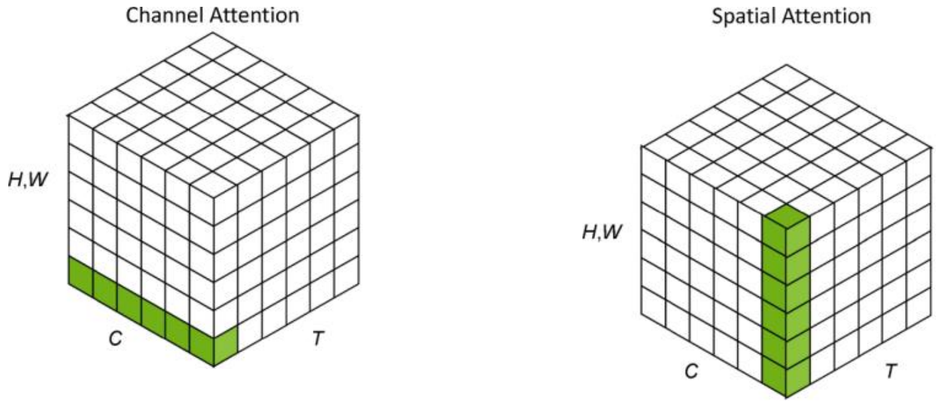
TRANSFORMATEC

Modalidades *de atención*



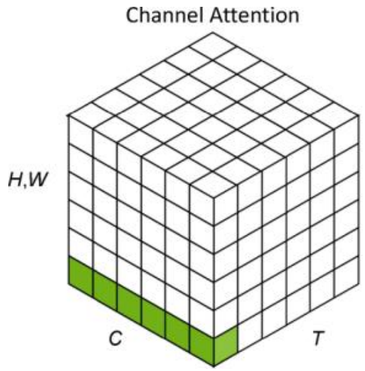
Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

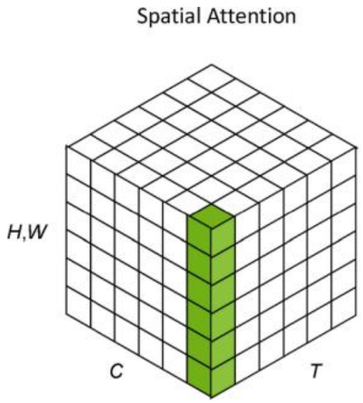
Modalidades *de atención*

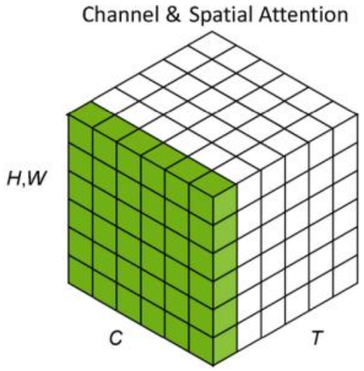


Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

Modalidades *de atención*

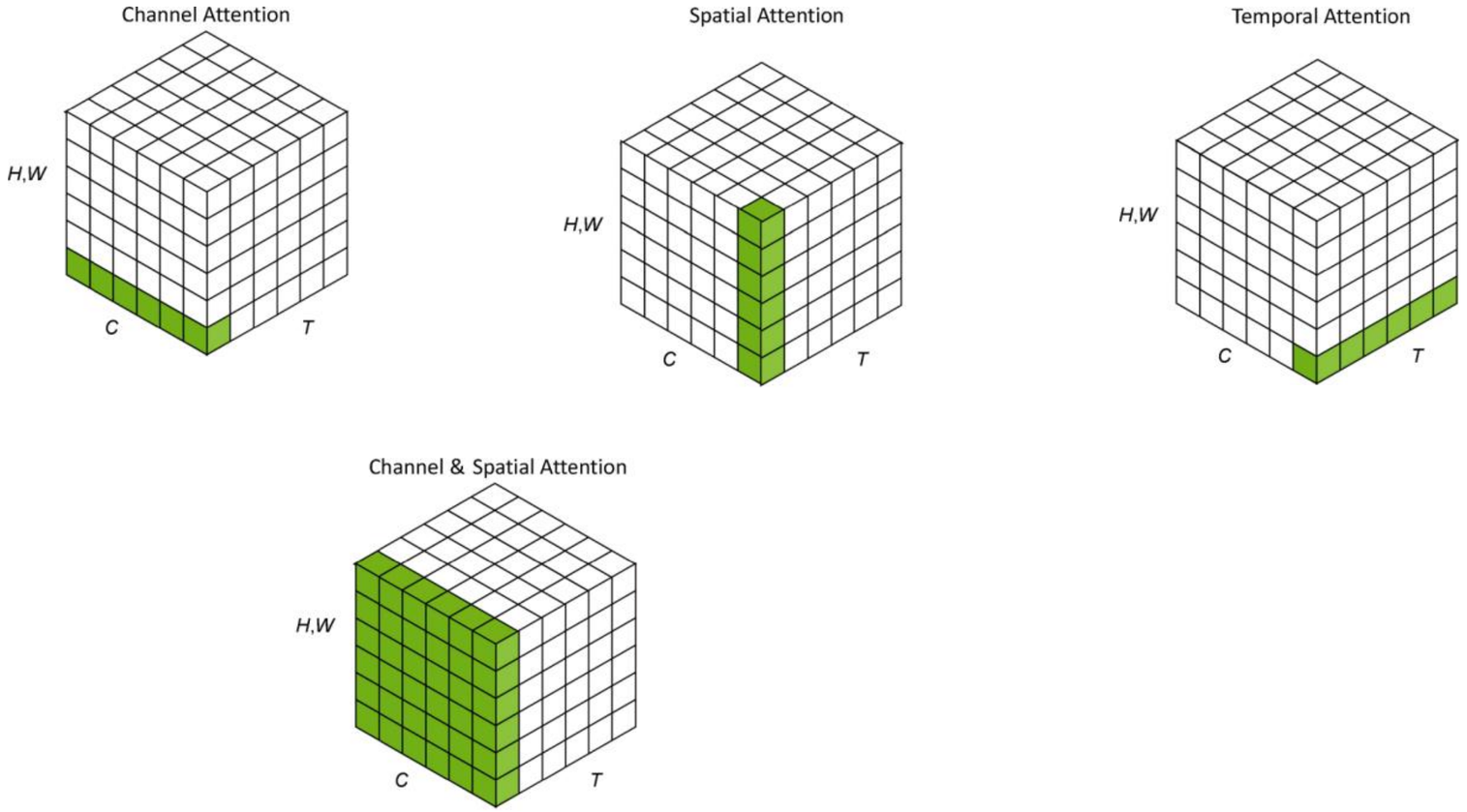






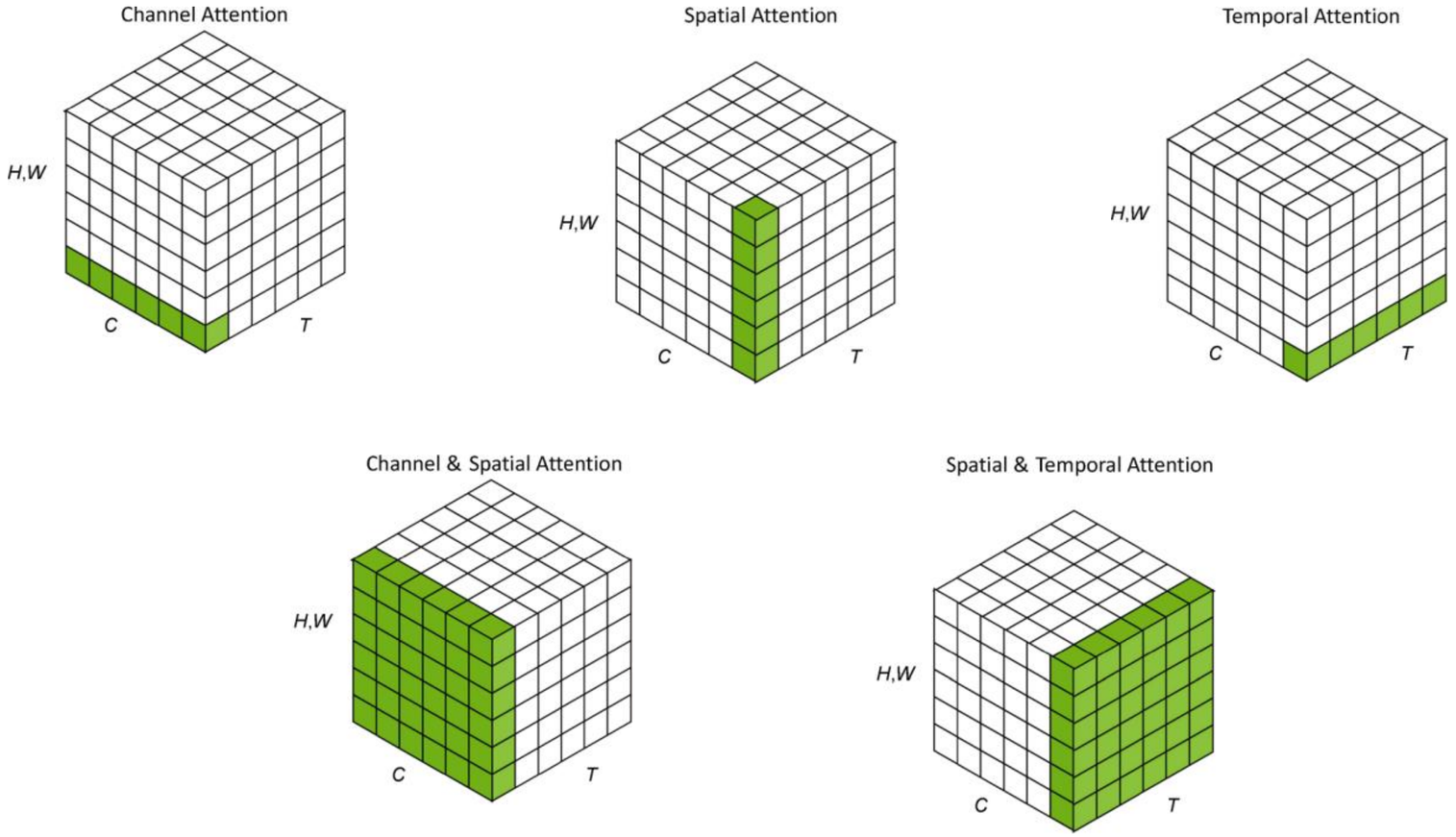
Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

Modalidades *de atención*



Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

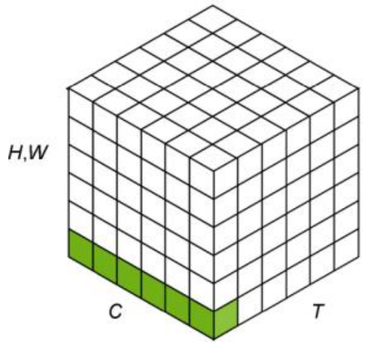
Modalidades *de atención*



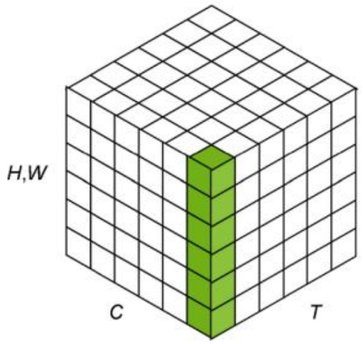
Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

Modalidades *de atención*

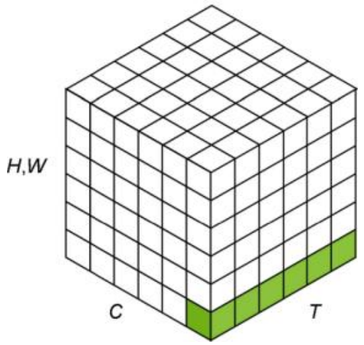
Channel Attention



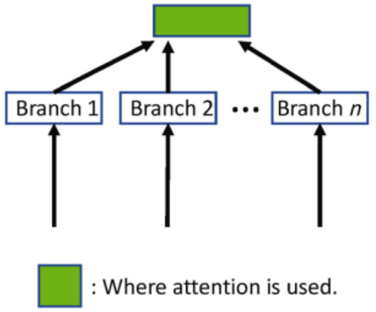
Spatial Attention



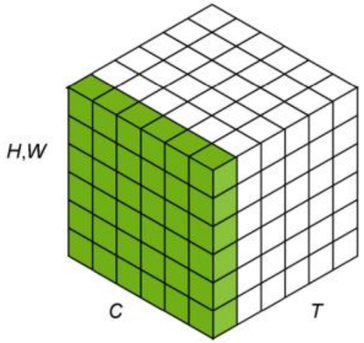
Temporal Attention



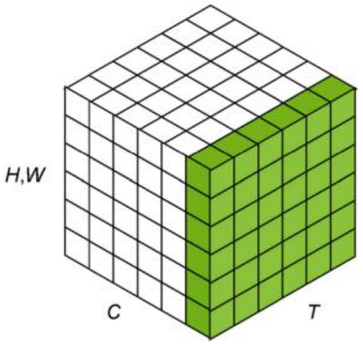
Branch Attention



Channel & Spatial Attention



Spatial & Temporal Attention



Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

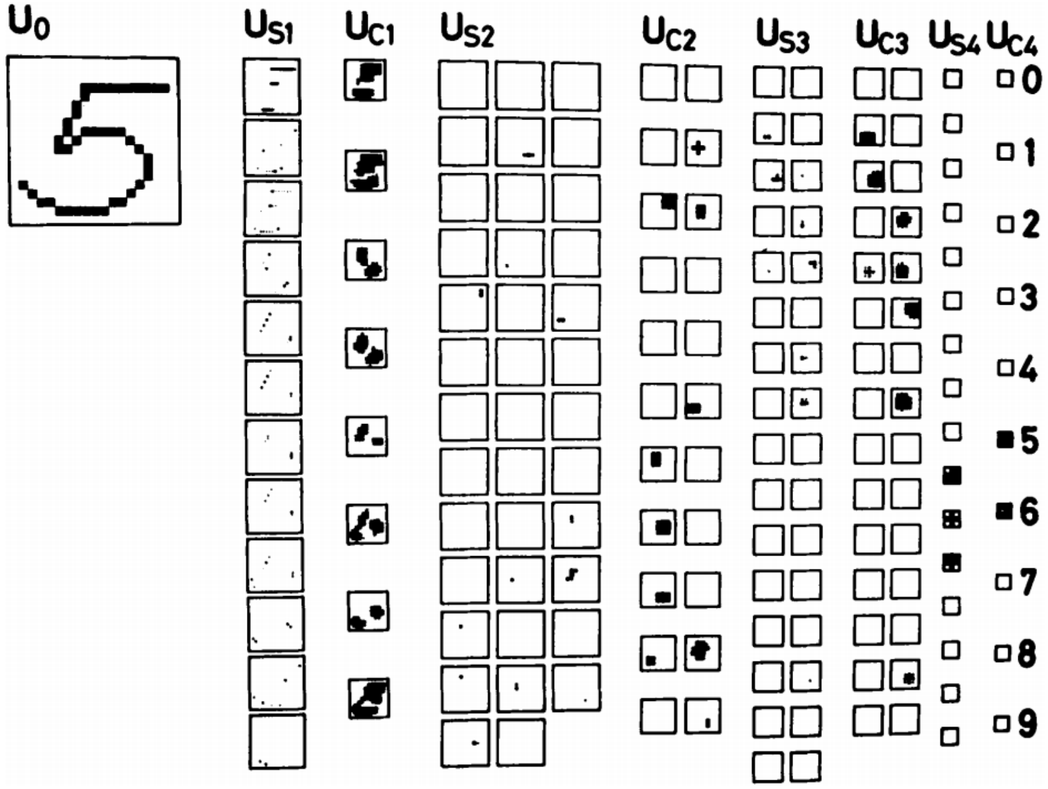
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

3.

Hard *attention*

TRANSFORMATEC

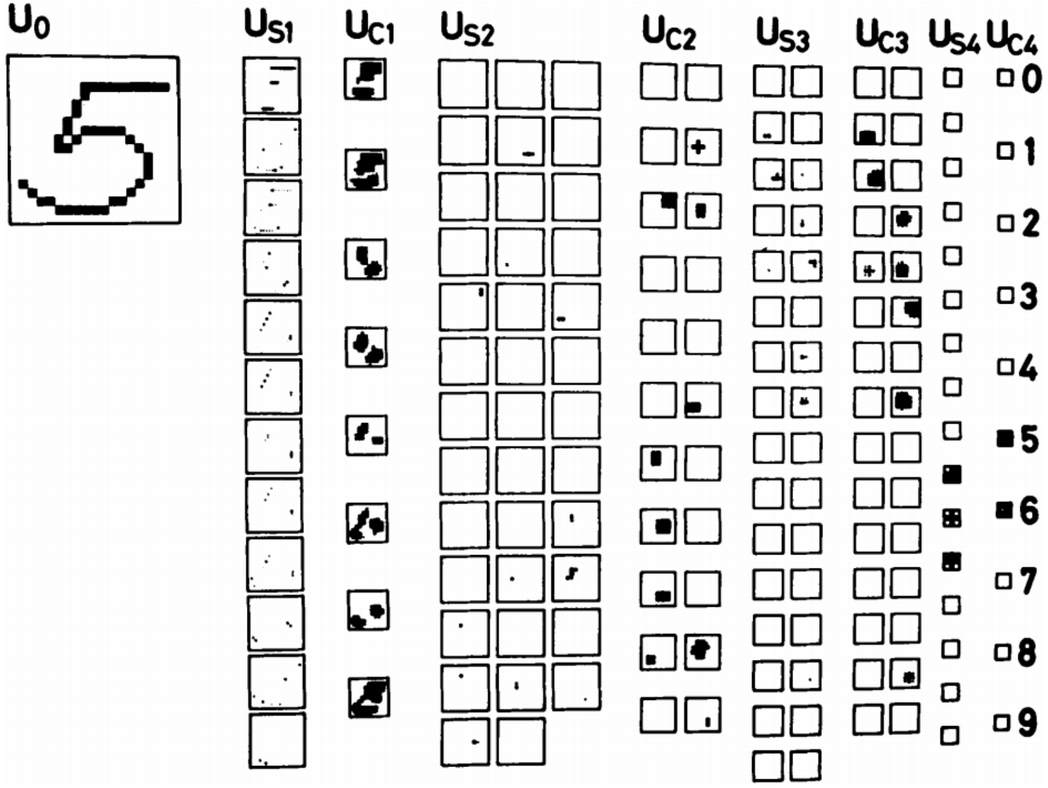
Neocognitron



Kunihiko Fukushima

Fukushima, Kunihiko and Sei Miyake (1980) "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". Biological Cybernetics .

Neocognitron



S-cell:

$$u_{Sl}(k_l, \boldsymbol{n}) = r_l \cdot \varphi\left(\frac{1 + \sum_{k_l=1}^{K_l-1} \sum_{v \in S_l} a_l(k_{l-1}, \boldsymbol{v}, k_l) \cdot u_{Cl-1}(k_{l-1}, \boldsymbol{n}+\boldsymbol{v})}{1 + \frac{2r_l}{1+r_l} \cdot b_l(k_l) \cdot v_{Cl-1}(\boldsymbol{n})}\right)$$

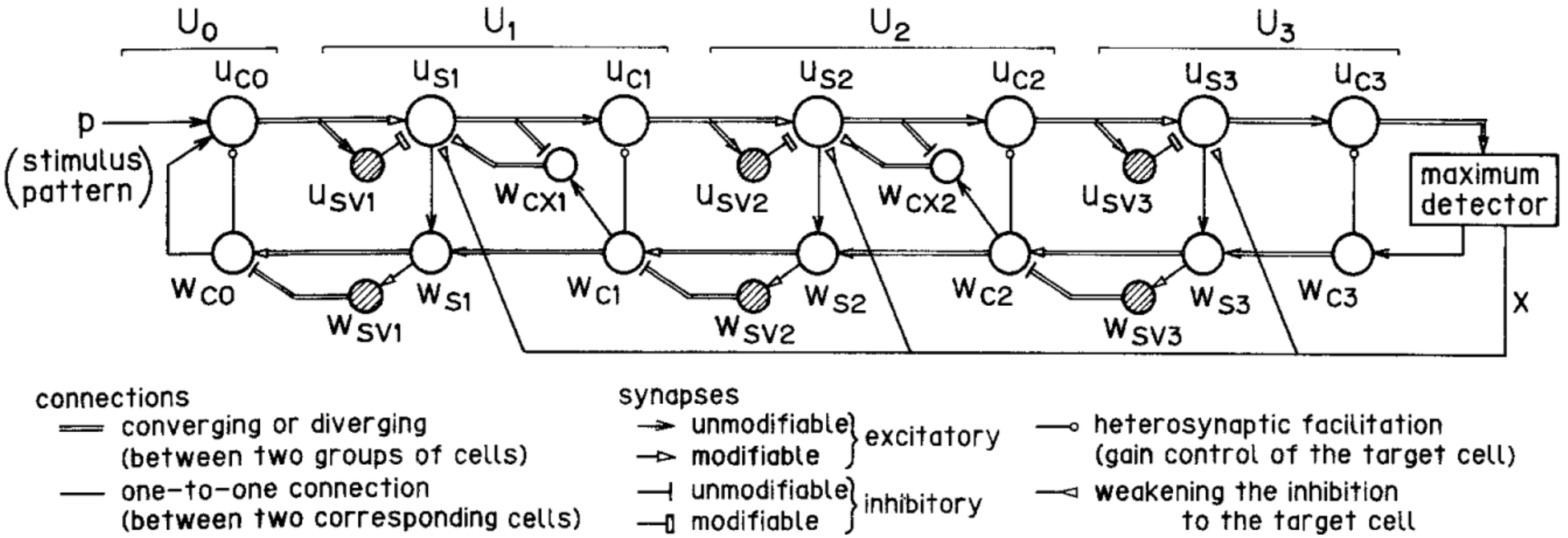
C-cell:

$$u_{Cl}(k_l, \boldsymbol{n}) = \psi\left(\frac{1 + \sum_{v \in D_l} d_l(\boldsymbol{v}) \cdot u_{Sl}(k_l, \boldsymbol{n}+\boldsymbol{v})}{1 + v_{Sl}(\boldsymbol{n})}\right)$$

donde $\varphi(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$ $\psi(x) = \varphi\left(\frac{x}{\alpha + x}\right)$

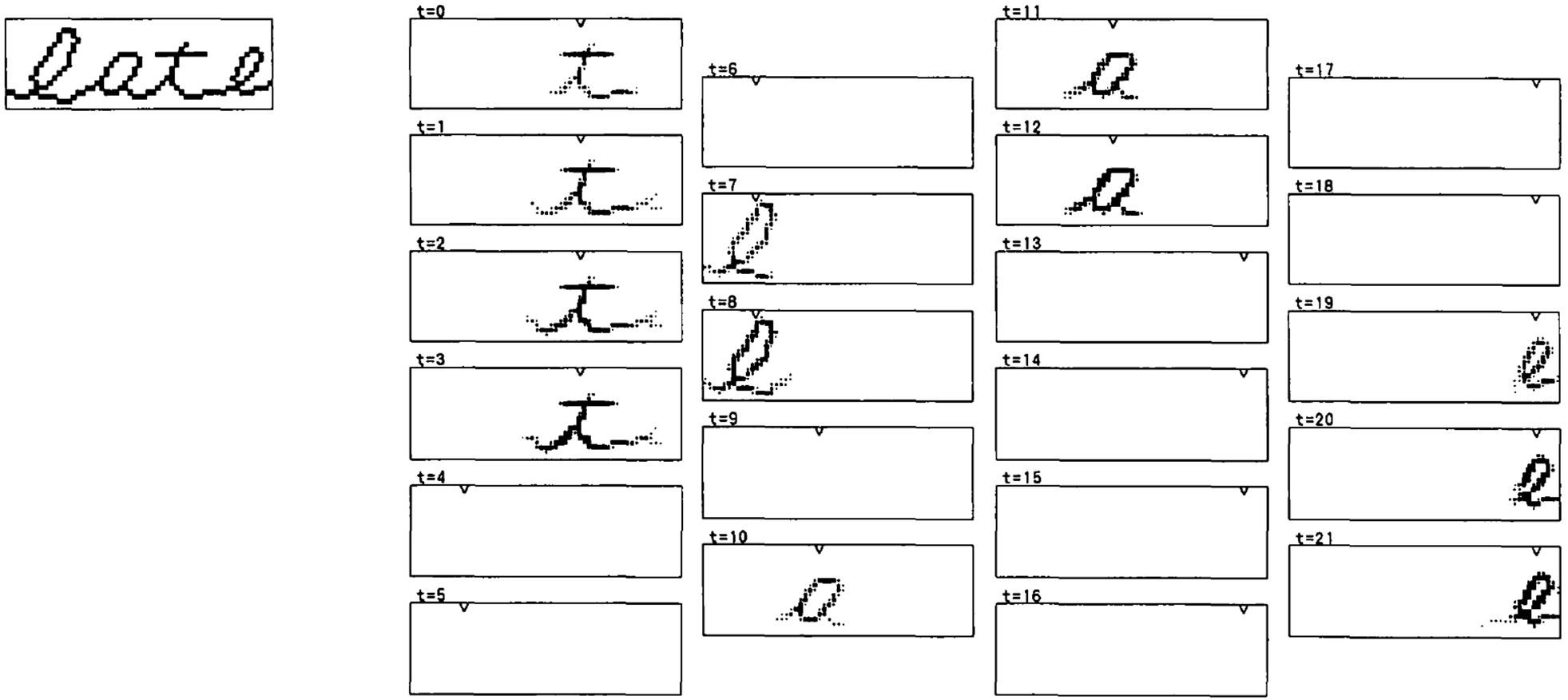
Fukushima, Kunihiko et al. (1983) "Neocognitron: A neural network model for a mechanism of visual pattern recognition". IEEE Transactions on Systems, Man, and Cybernetics .

Neocognitron



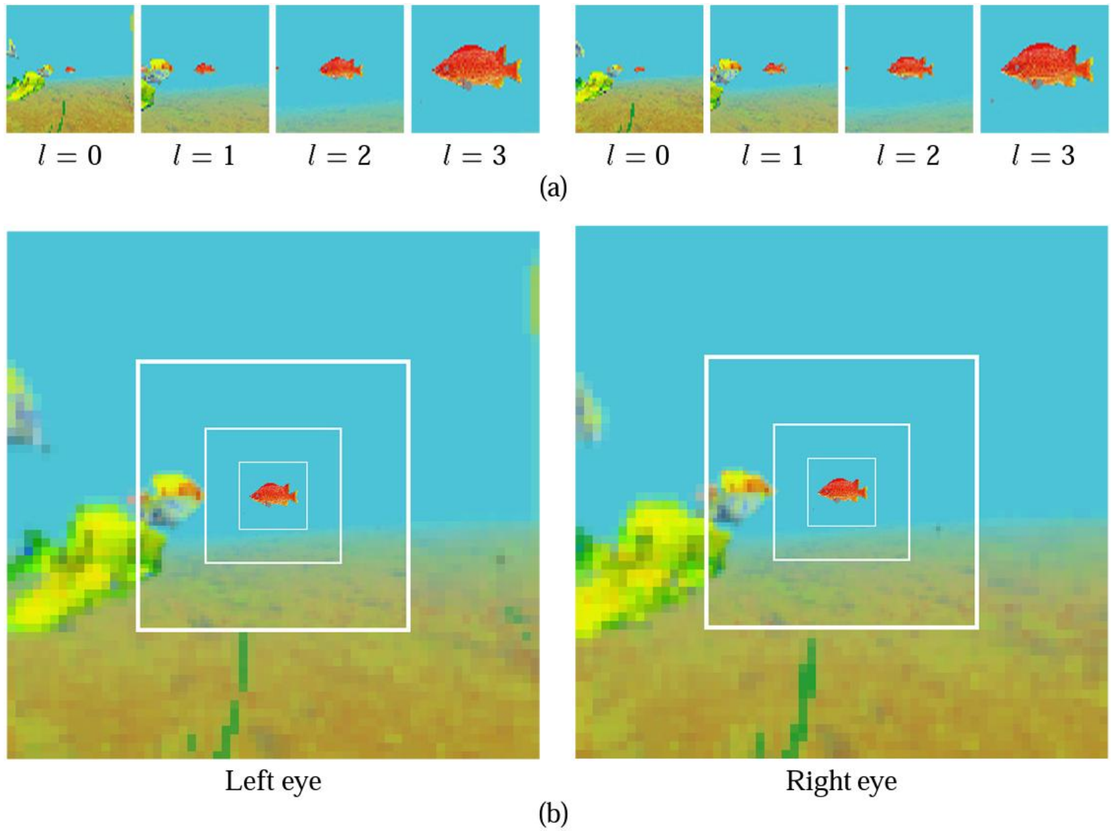
Fukushima, Kunihiko et al. (1986) "A Neural Network Model for Selective Attention in Visual Pattern Recognition". Biological Cybernetics, 55(1), 5-15.

Neocog*nitron*



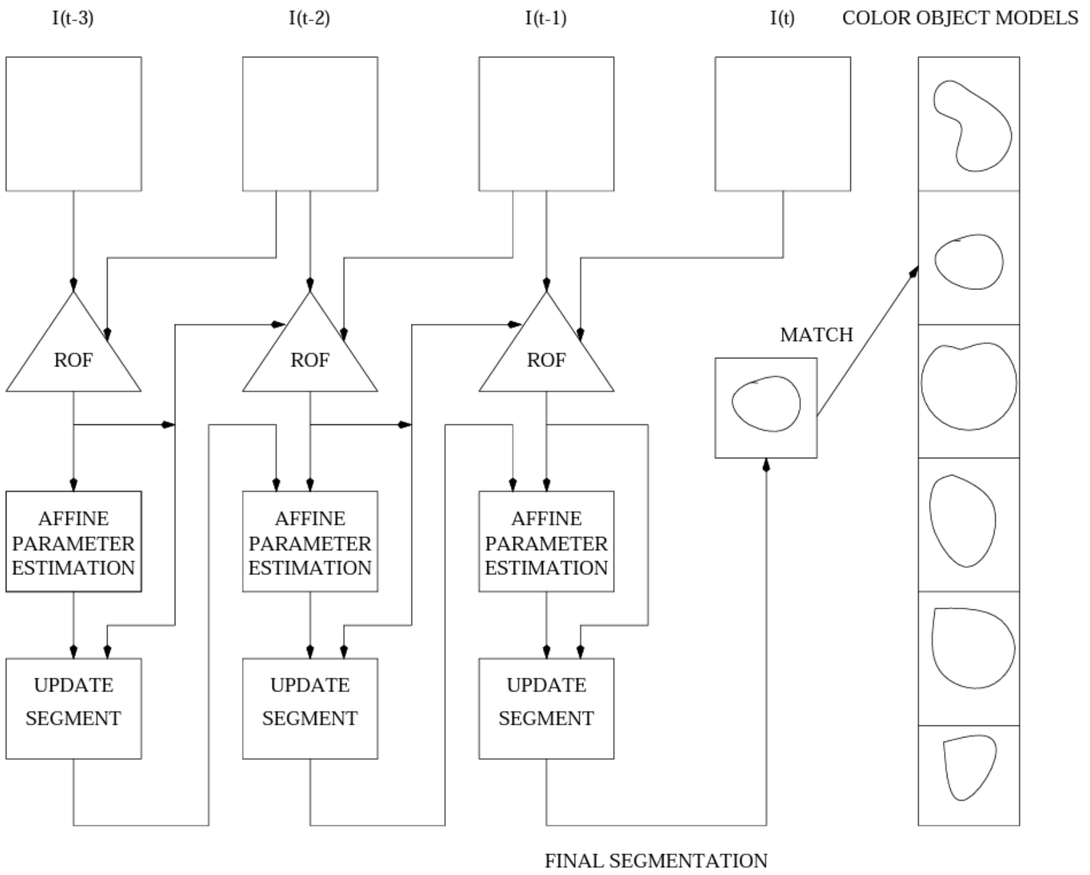
Fukushima, Kunihiko et al. (1986) "A Neural Network Model for Selective Attention in Visual Pattern Recognition". Biological Cybernetics, 55(1), 5-15.

Overt *Attention*



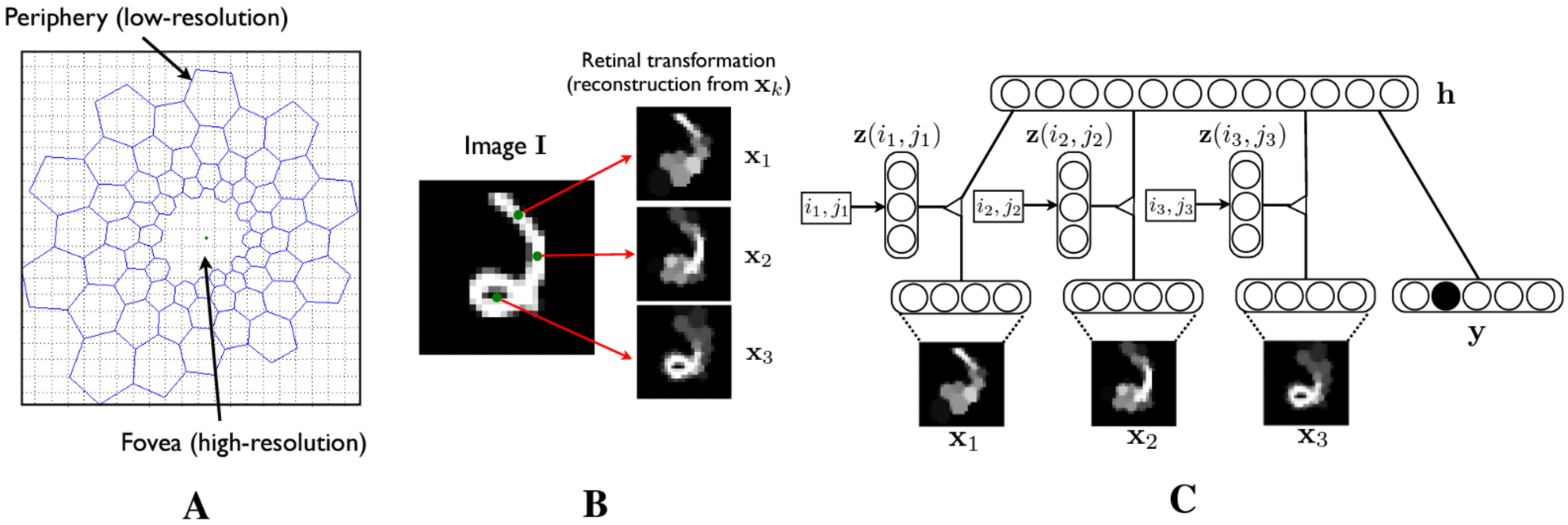
Rabie, Tamer F., and Demetri Terzopoulos (1996) "Motion and Color Analysis for Animal Perception". Proceeding of the National Conference on Artificial Intelligence. 1996. p. 1090-1097.

Overt *Attention*



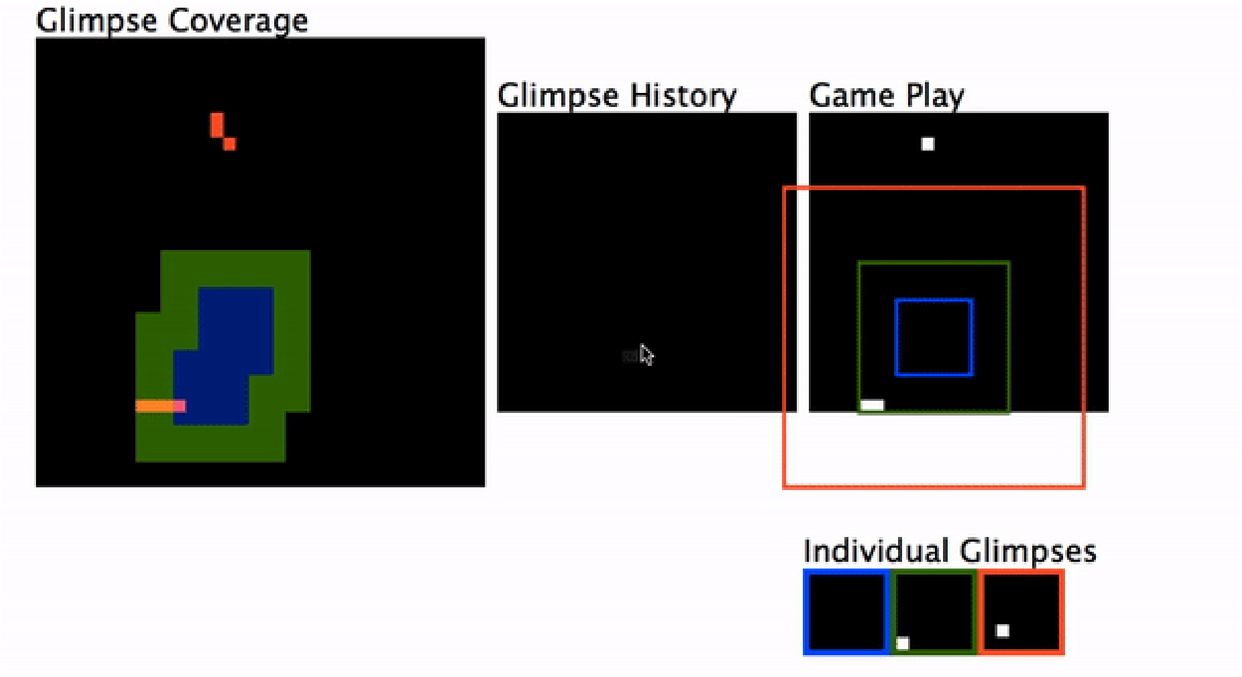
Rabie, Tamer F., and Demetri Terzopoulos (1996) "Motion and Color Analysis for Animal Perception". Proceeding of the National Conference on Artificial Intelligence. 1996. p. 1090-1097.

Foveal *Glimpse*



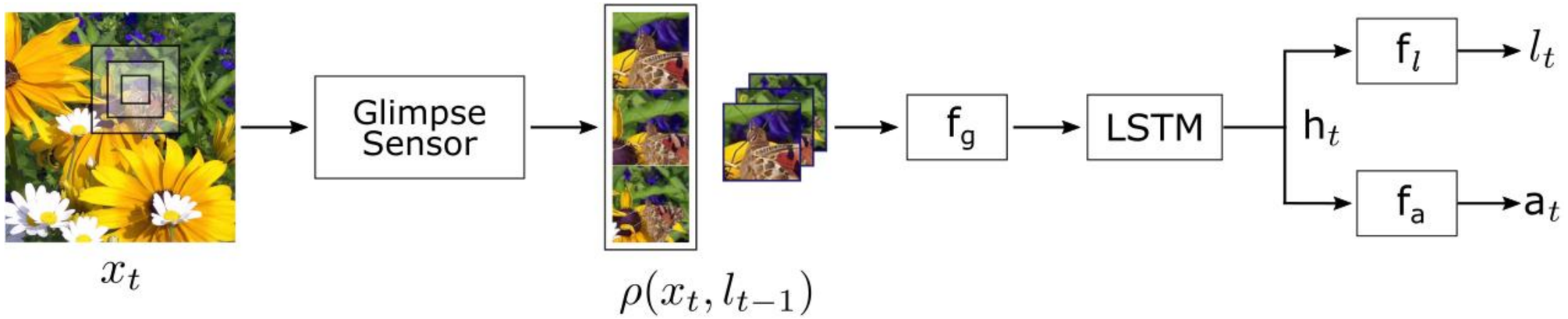
Hugo Larochelle and Geoffrey E. Hinton (2010) "Learning to combine foveal glimpses with a third-order Boltzmann machine". Advances in neural information processing systems, 2010, vol. 23

Recurrent Attention *Model (RAM)*



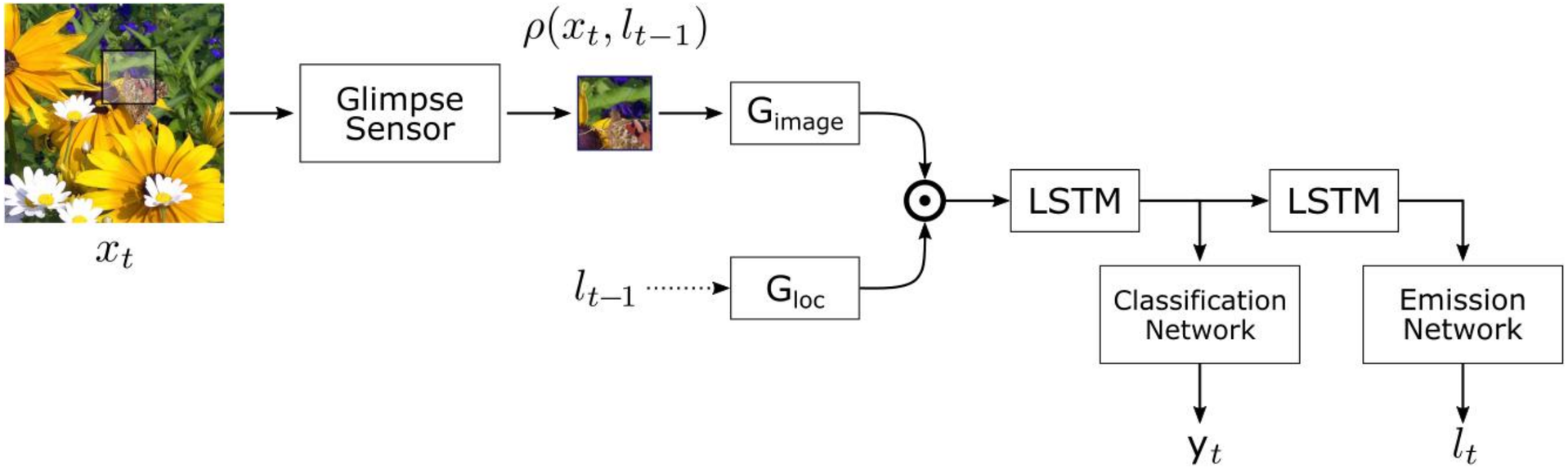
Mnih et al. (2014) "Recurrent Models of Visual Attention". Advances in neural information processing systems (pp. 2204-2212).

Recurrent Attention *Model (RAM)*



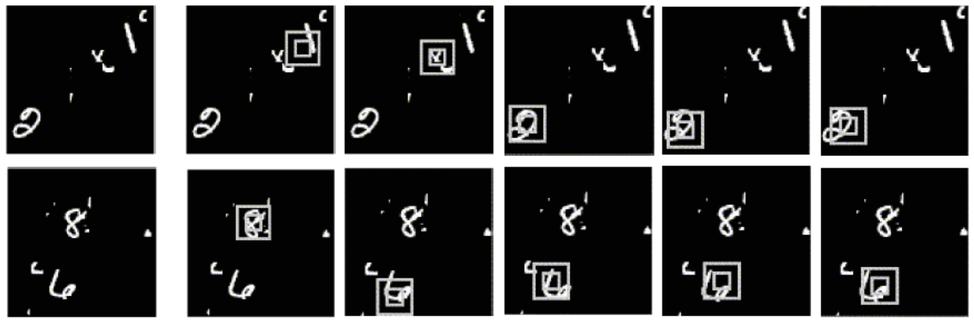
Mnih et al. (2014) "Recurrent Models of Visual Attention". Advances in neural information processing systems (pp. 2204-2212).

Deep Recurrent Visual *Attention Model (DRAM)*



Ba et al. (2015) "Multiple object recognition with visual attention". arXiv preprint arXiv:1412.7755.

Deep Recurrent Visual *Attention Model (DRAM)*



Model	Test Err.
RAM Mnih et al. (2014)	9%
DRAM w/o context	7%
DRAM	**5%**

Model	Test Err.
ConvNet 64-64-64-512	3.2%
DRAM	**2.5%**

Model	Test Err.
11 layer CNN Goodfellow et al. (2013)	3.96%
10 layer CNN	4.11%
Single DRAM	5.1%
Single DRAM MC avg.	4.4%
forward-backward DRAM MC avg.	**3.9%**

Ba et al. (2015) "Multiple object recognition with visual attention". arXiv preprint arXiv:1412.7755.

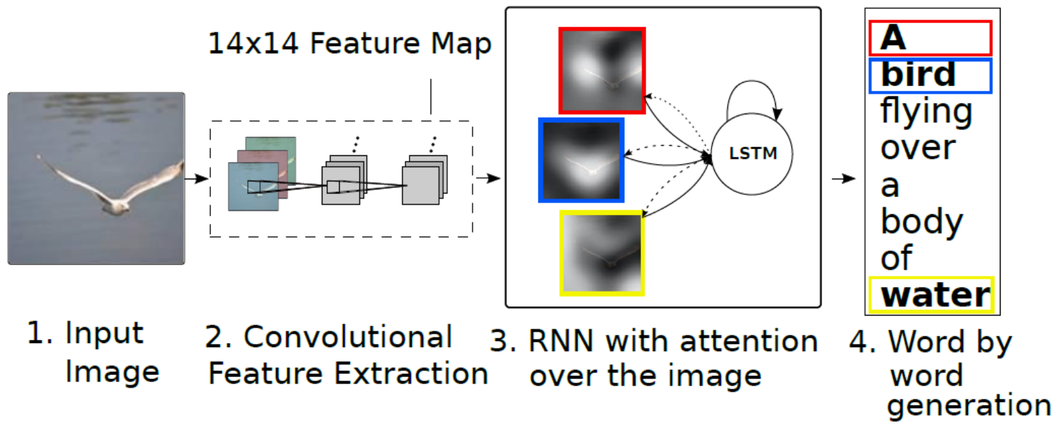
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

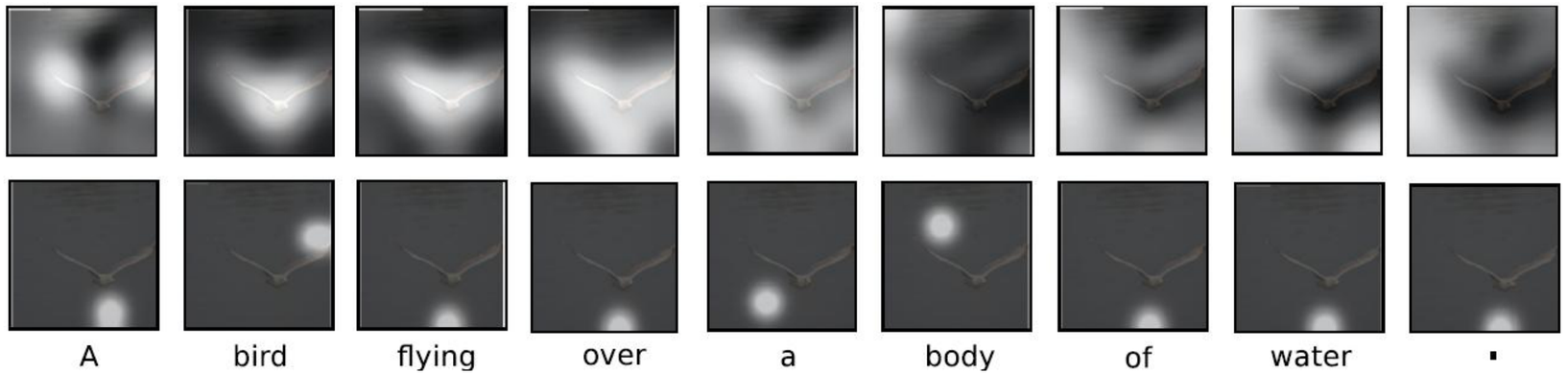
4.

Soft *attention*

TRANSFORMATEC

Show, Attend *and Tell*



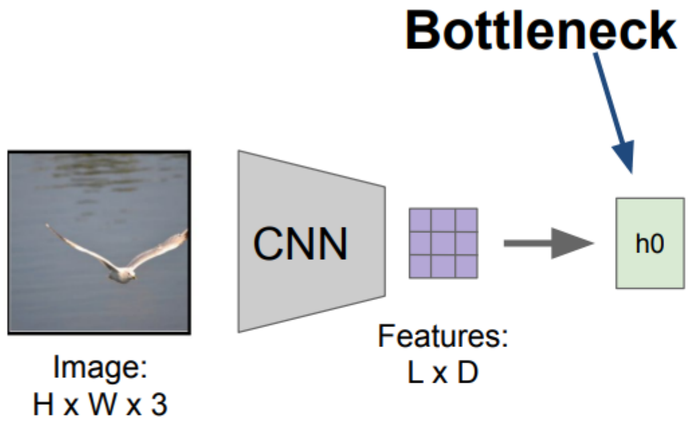


Soft Attention

Hard Attention

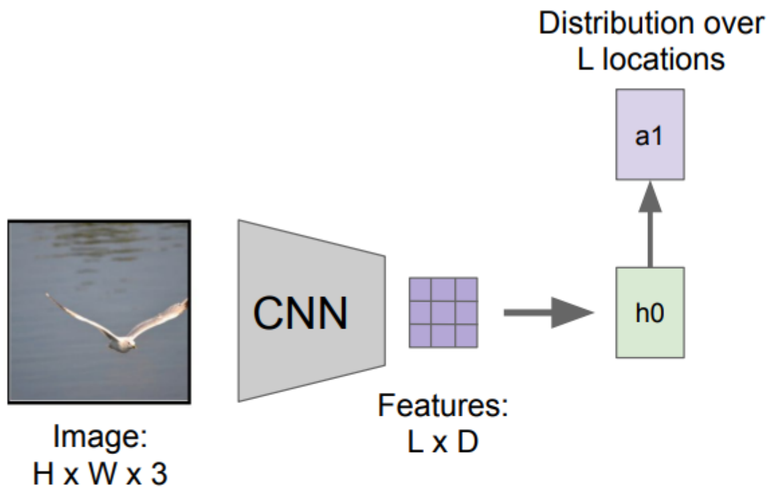
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention".
Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Show, Attend *and Tell*



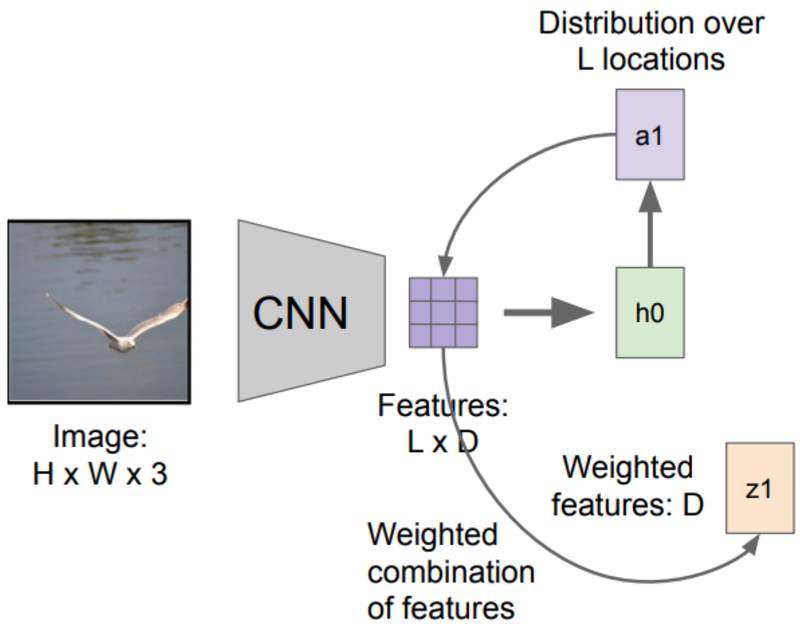
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Show, Attend *and Tell*



Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

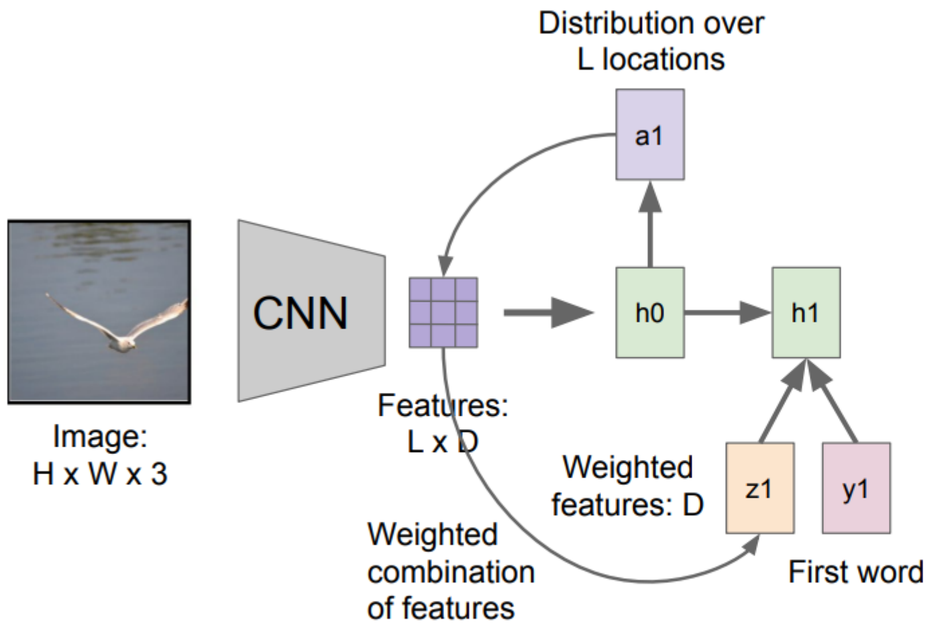
Show, Attend *and Tell*



$$z = \sum_{i=1}^{L} p_i v_i$$

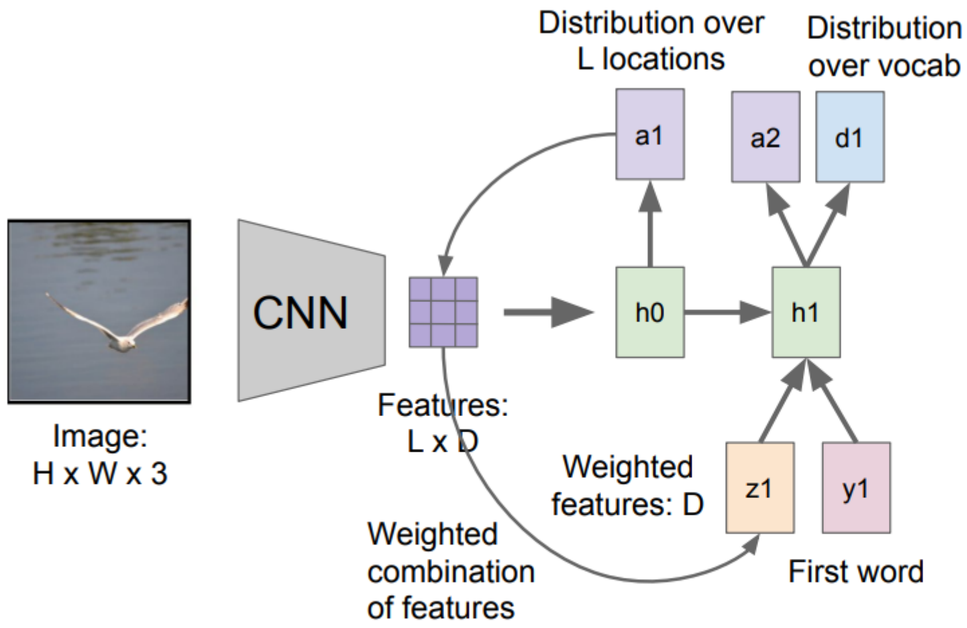
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention".
Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Show, Attend *and Tell*



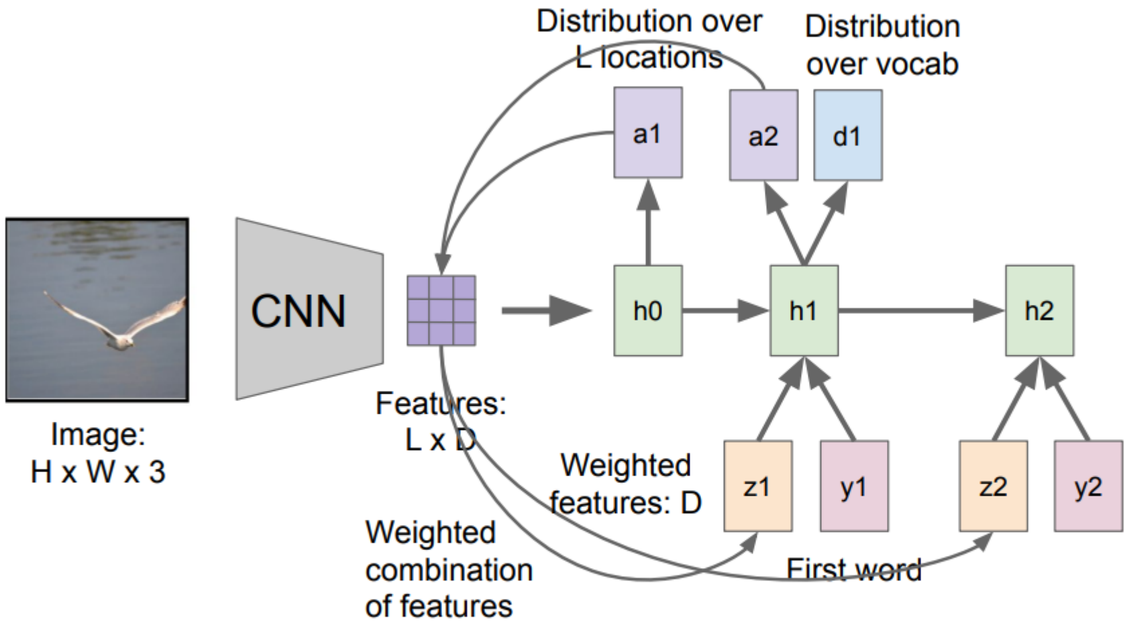
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Show, Attend *and Tell*



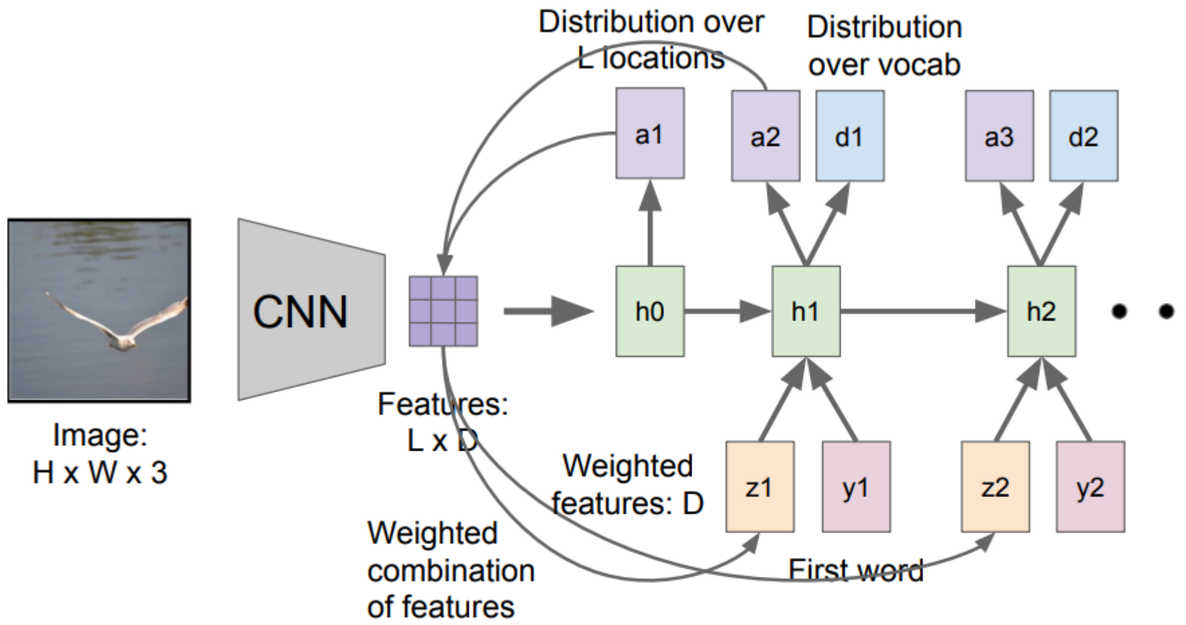
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Show, Attend *and Tell*



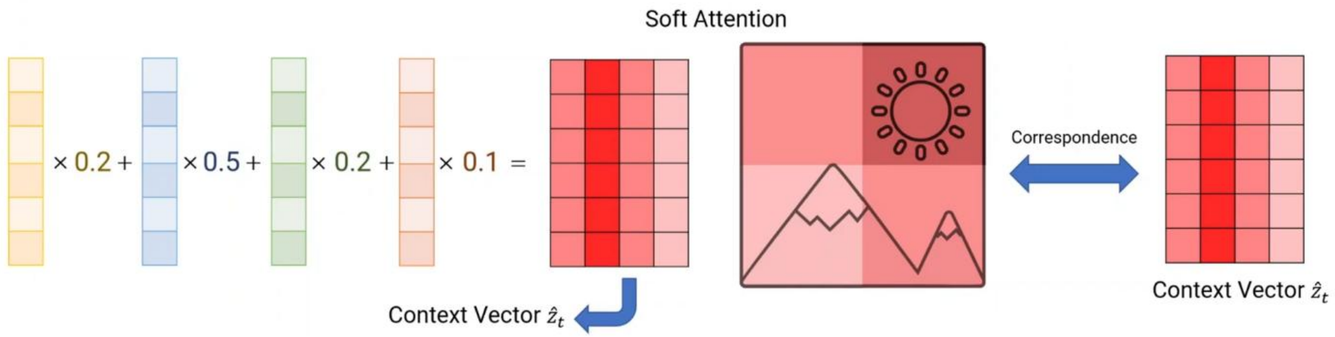
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

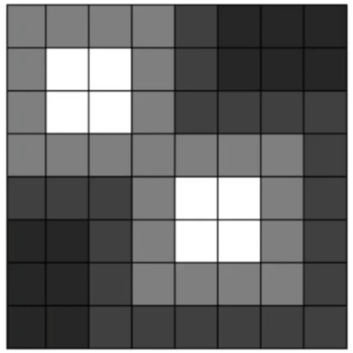
Show, Attend *and Tell*



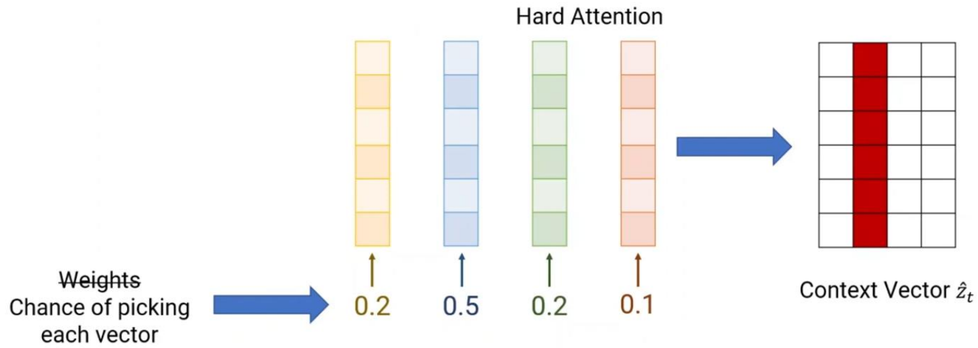
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

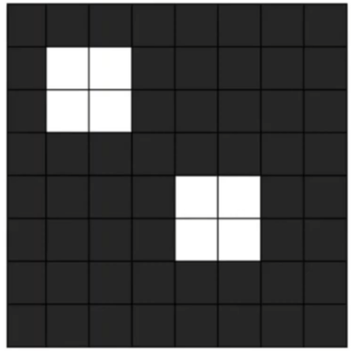
Show, Attend *and Tell*





Soft





Hard

Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention".
Proceedings of the 32nd International Conference on International Conference on Machine Learning.

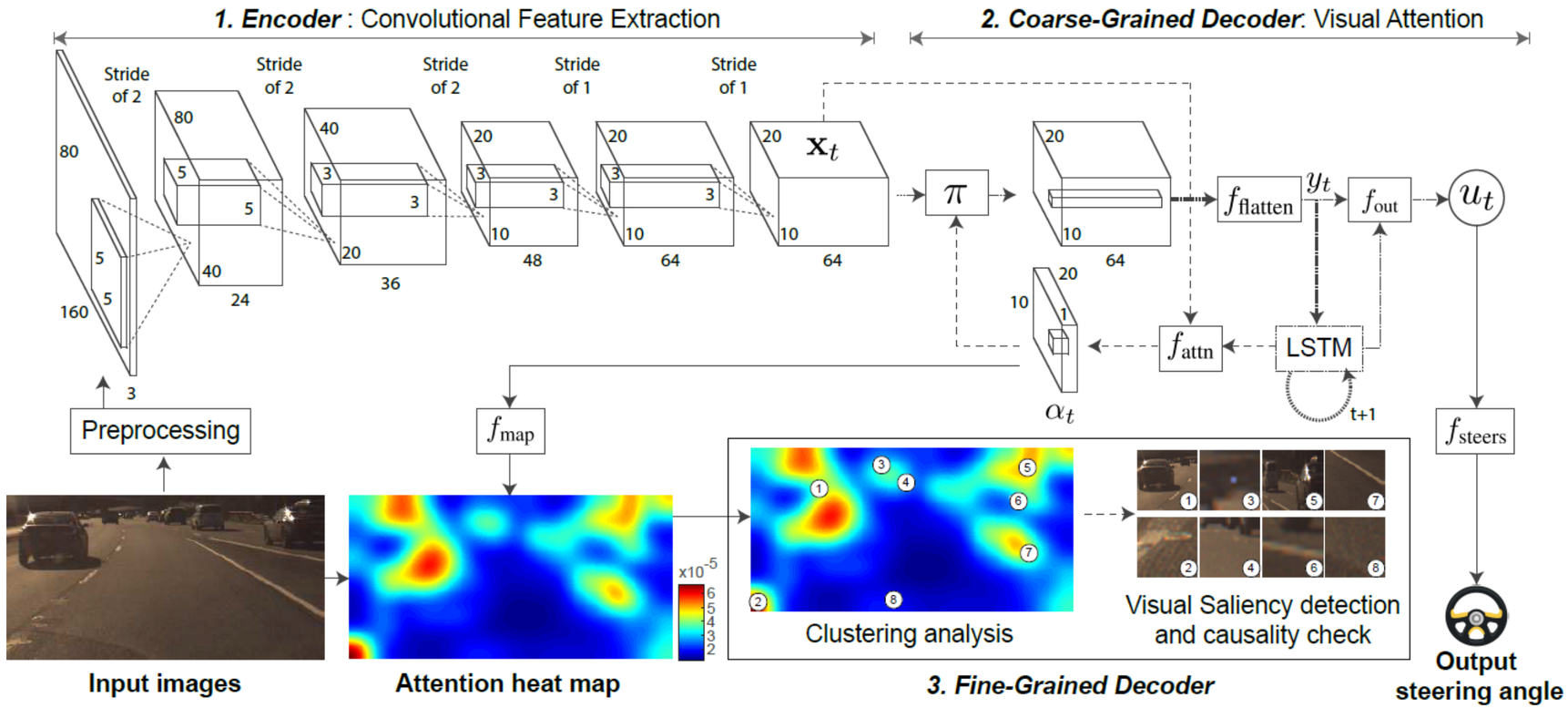
Show, Attend *and Tell*

(a) A man and a woman playing frisbee in a field.

(b) A woman is throwing a frisbee in a park.

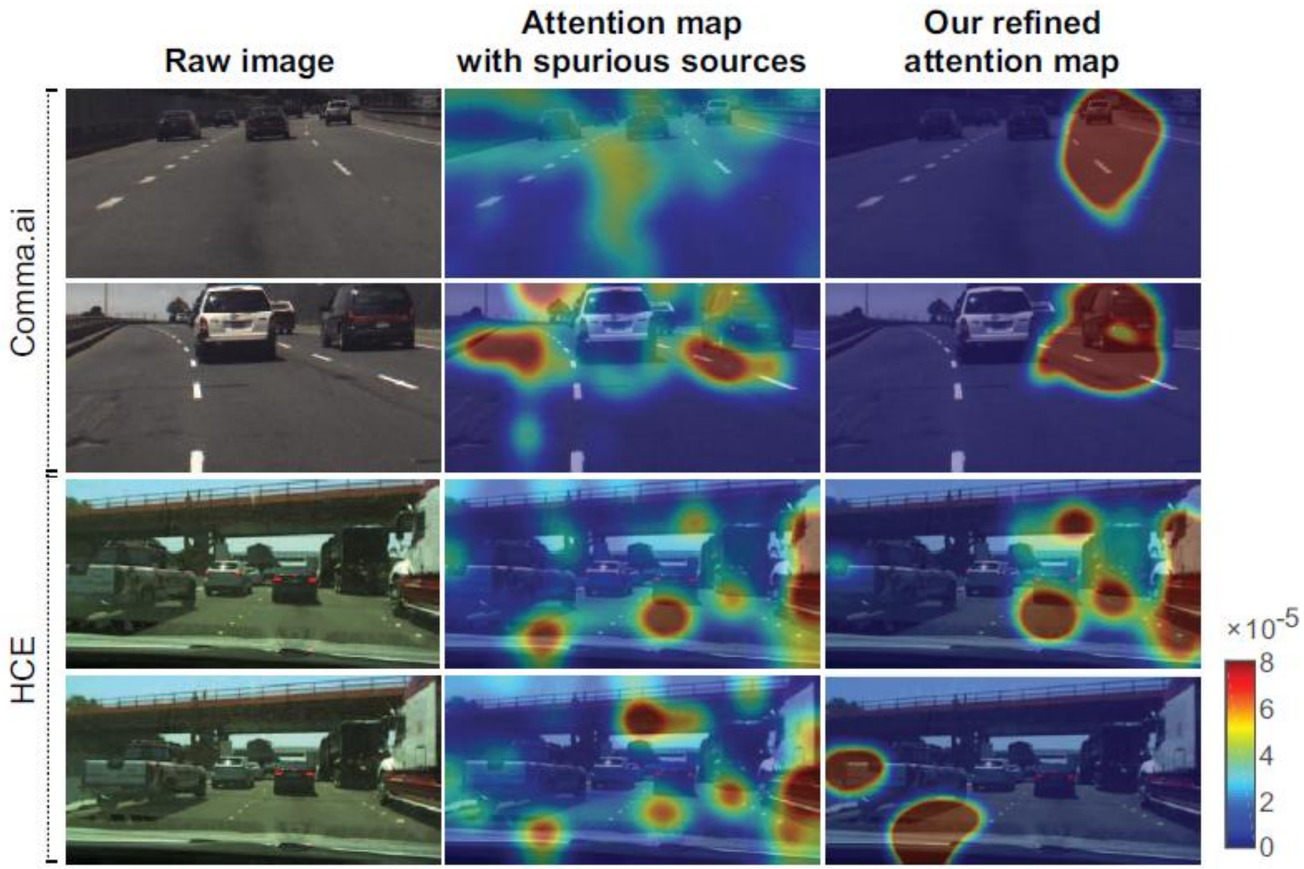
Xu et al. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". Proceedings of the 32nd International Conference on International Conference on Machine Learning.

Self-driving car



Kim et al. (2017) "Interpretable learning for self-driving cars by visualizing causal attention". Proceedings of the IEEE international conference on computer vision (pp. 2942-2950).

Self-driving car



Kim et al. (2017) "Interpretable learning for self-driving cars by visualizing causal attention". Proceedings of the IEEE international conference on computer vision (pp. 2942-2950).

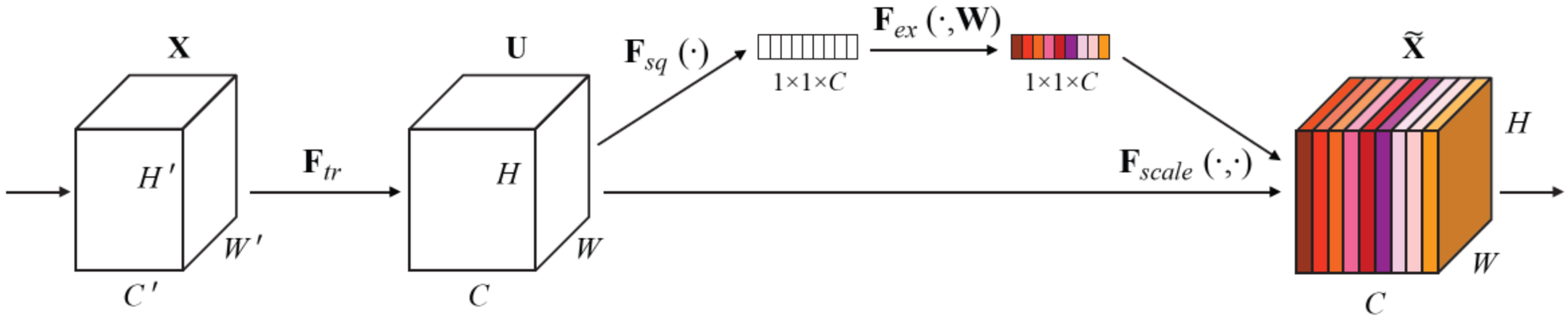
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

5.

Channel *attention*

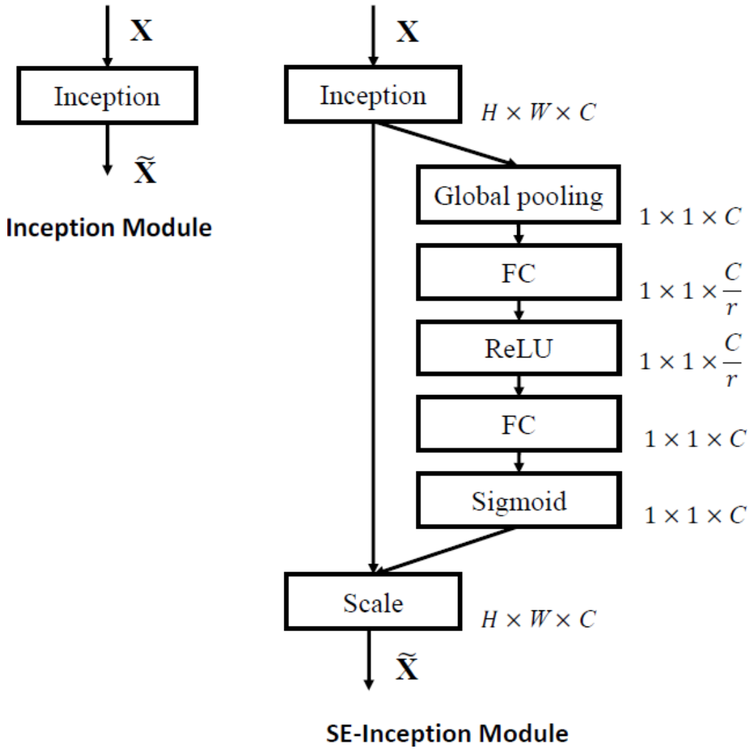
TRANSFORMATEC

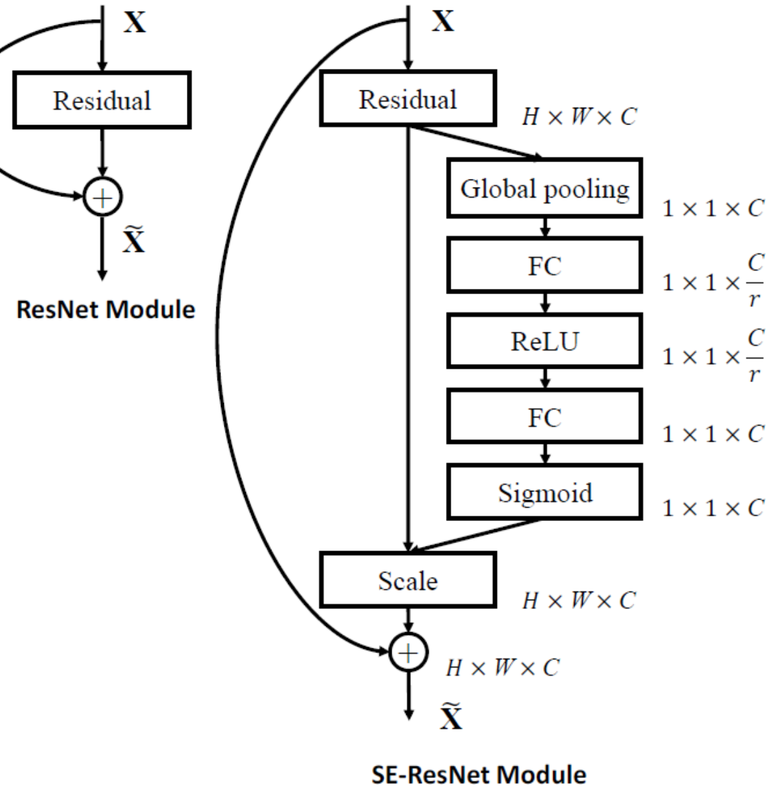
Squeeze-and-*Excitation*



Jie Hu et al. (2018) "Squeeze-and-Excitation Networks". Proceedings of the IEEE conference on computer vision and pattern recognition.

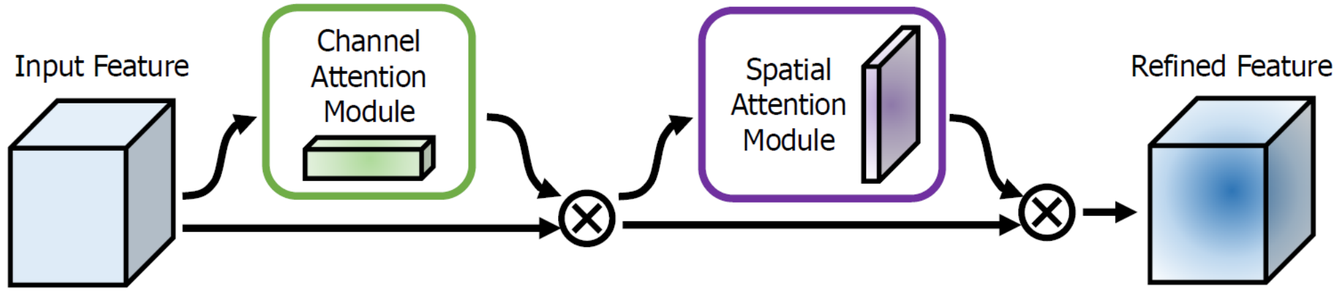
Squeeze-and-*Excitation*

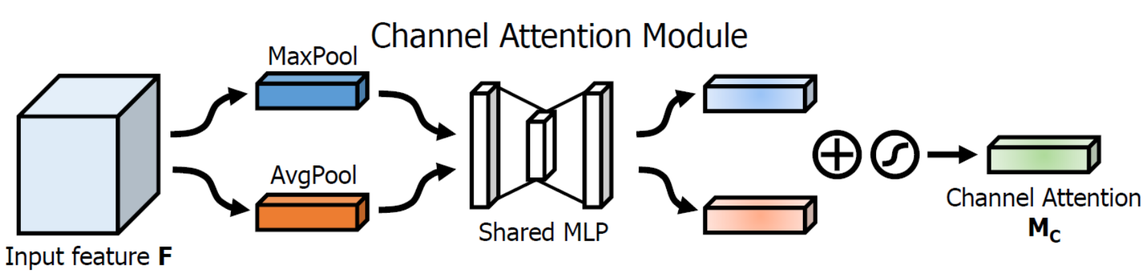


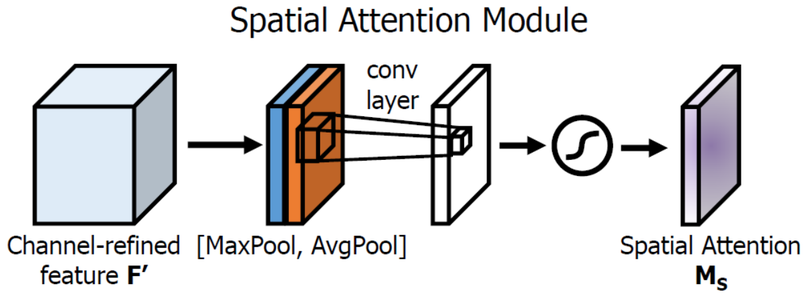


Jie Hu et al. (2018) "Squeeze-and-Excitation Networks".
Proceedings of the IEEE conference on computer vision and pattern recognition.

Convolutional Block *Attention Module*

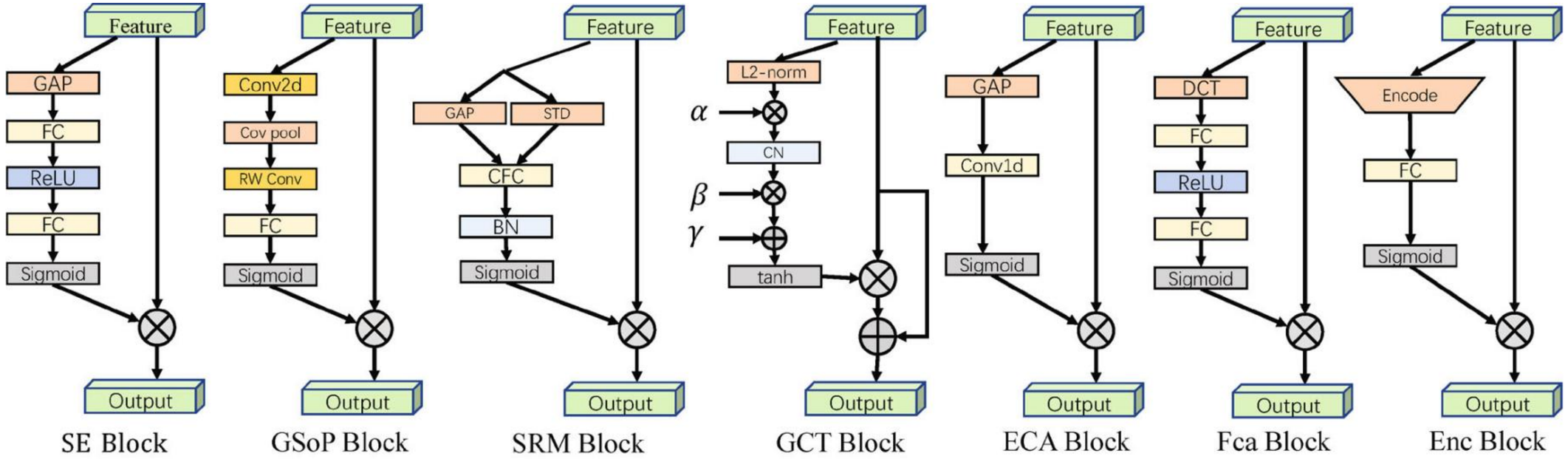






Sanghyun Woo et al. (2018) "CBAM: Convolutional Block Attention Module". Proceedings of the European conference on computer vision (ECCV). 2018. p. 3-19.

Channel *attention*



Meng-Hao Guo et al. (2022) "Attention Mechanisms in Computer Vision: A Survey". Computational Visual Media. Springer.

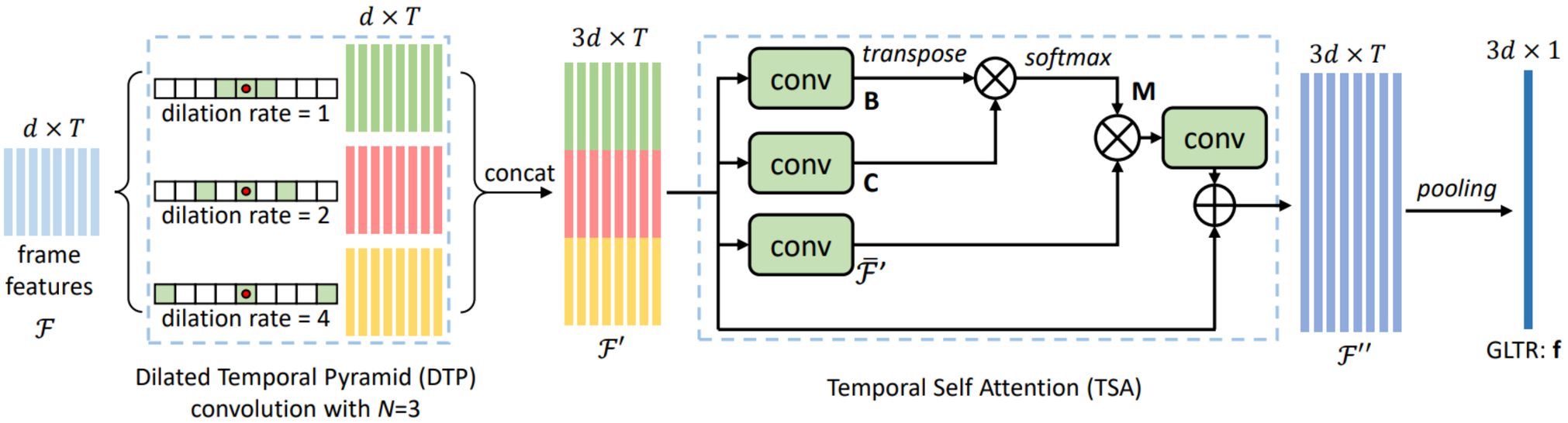
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

6.

Temporal *attention*

TRANSFORMATEC

Dilated Tem*poral Pyramid*



Jianing Li et al. (2019) "Global-Local Temporal Representations For Video Person Re-Identification". Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 3958-3967.

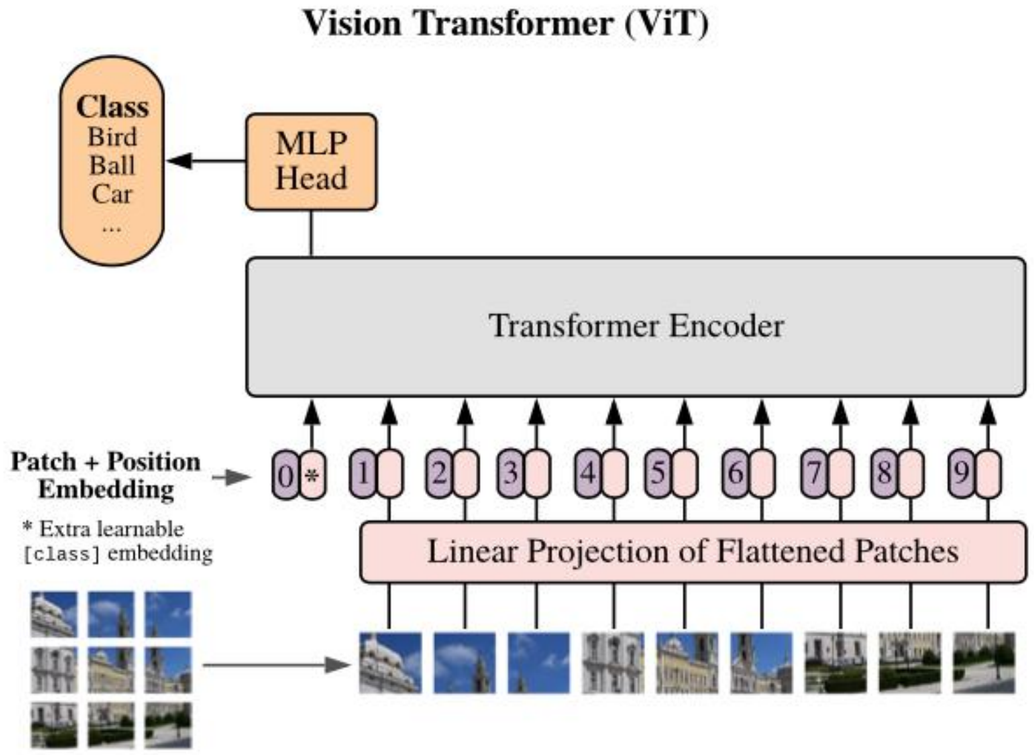
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

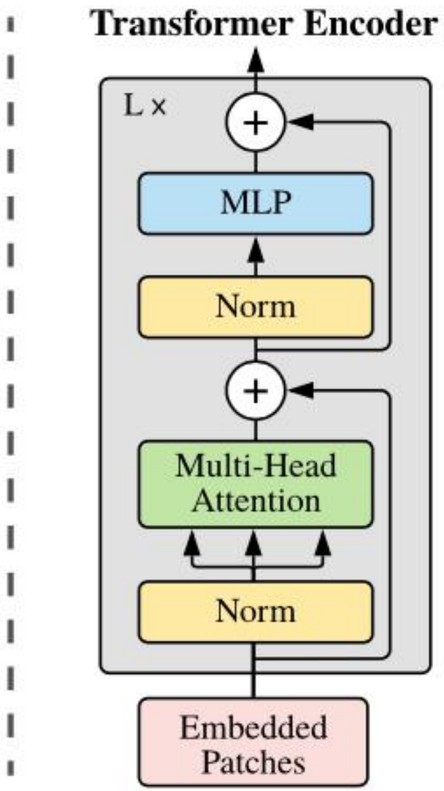
7.

Visual *Transformer*

TRANSFORMATEC

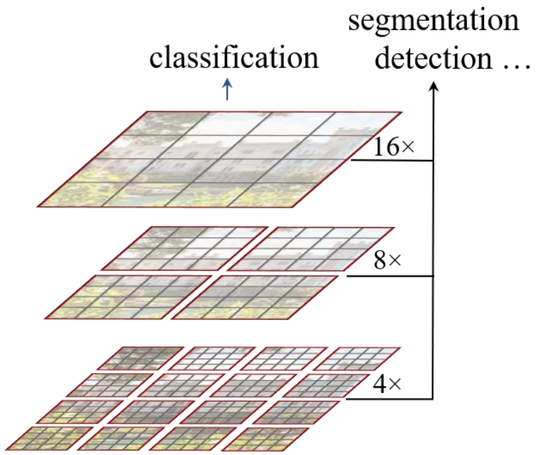
ViT



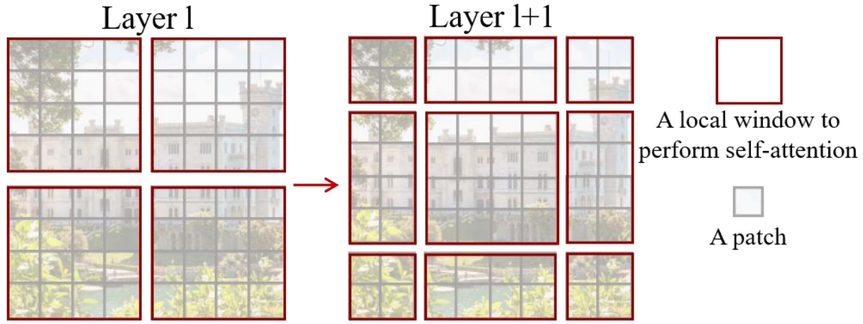


Alexey Dosovitskiy et al. (2021) "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". Proceedings of the IEEE/CVF international conference on computer vision. 2019. p. 3286-3295.

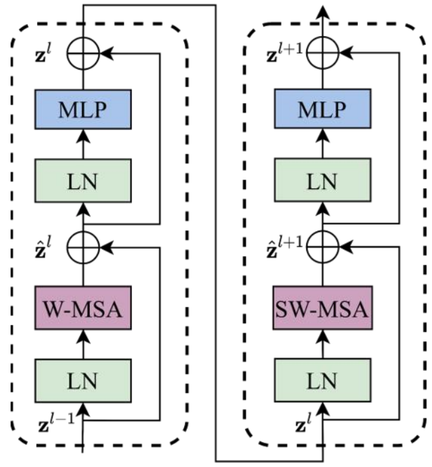
Swin *Transformer*



(a) Swin Transformer



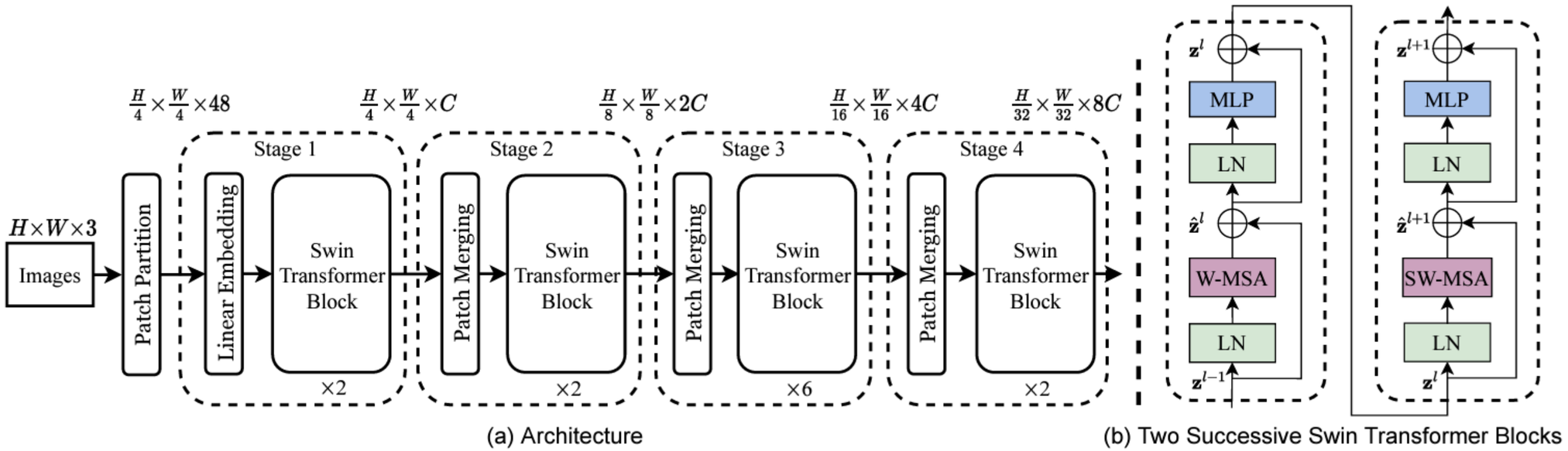
(b) Shifted Window



(c) Two Successive Swin Transformer Blocks

Ze Liu et al. (2021) "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

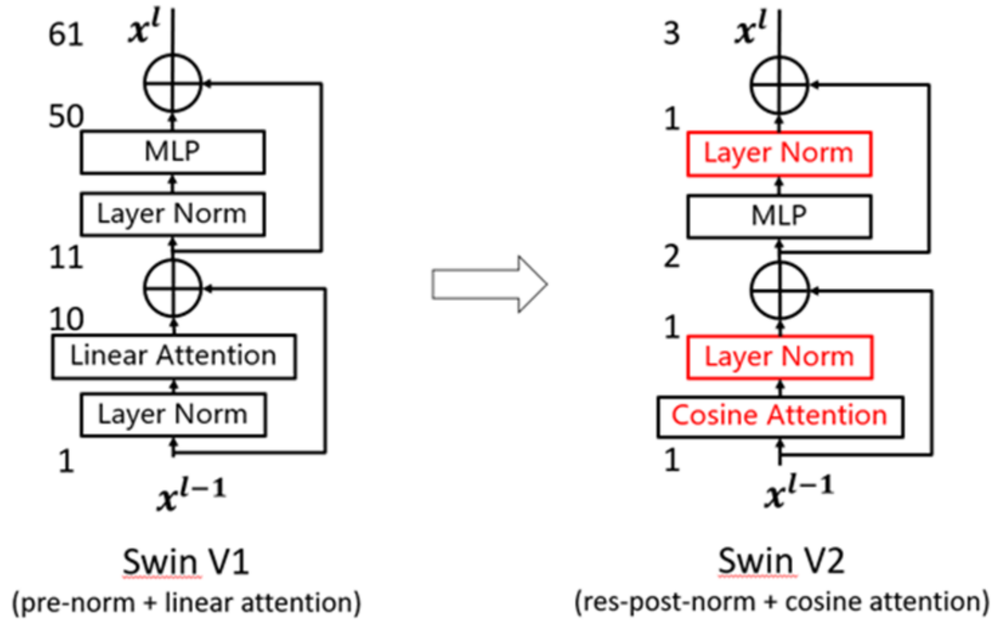
Swin *Transformer*



TRANSFORMATEC

Ze Liu et al. (2021) "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

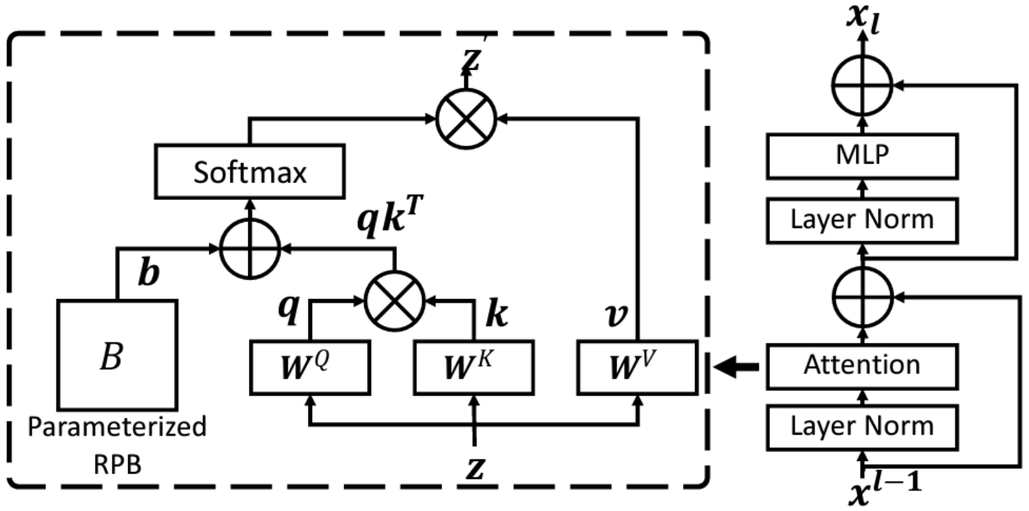
Swin *Transformer v2*



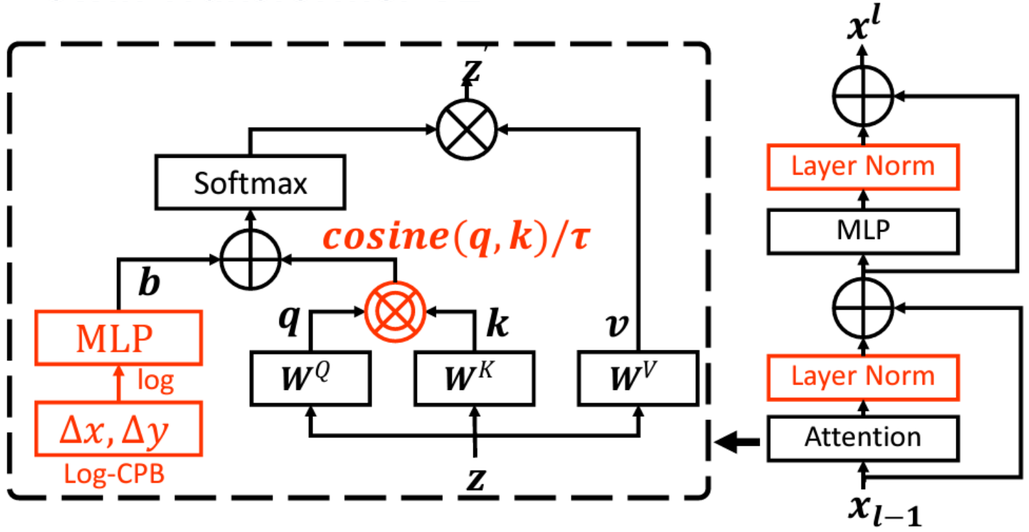
Ze Liu et al. (2022) "Swin Transformer V2: Scaling Up Capacity and Resolution".
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12009-12019).

Swin *Transformer v2*

Swin Transformer



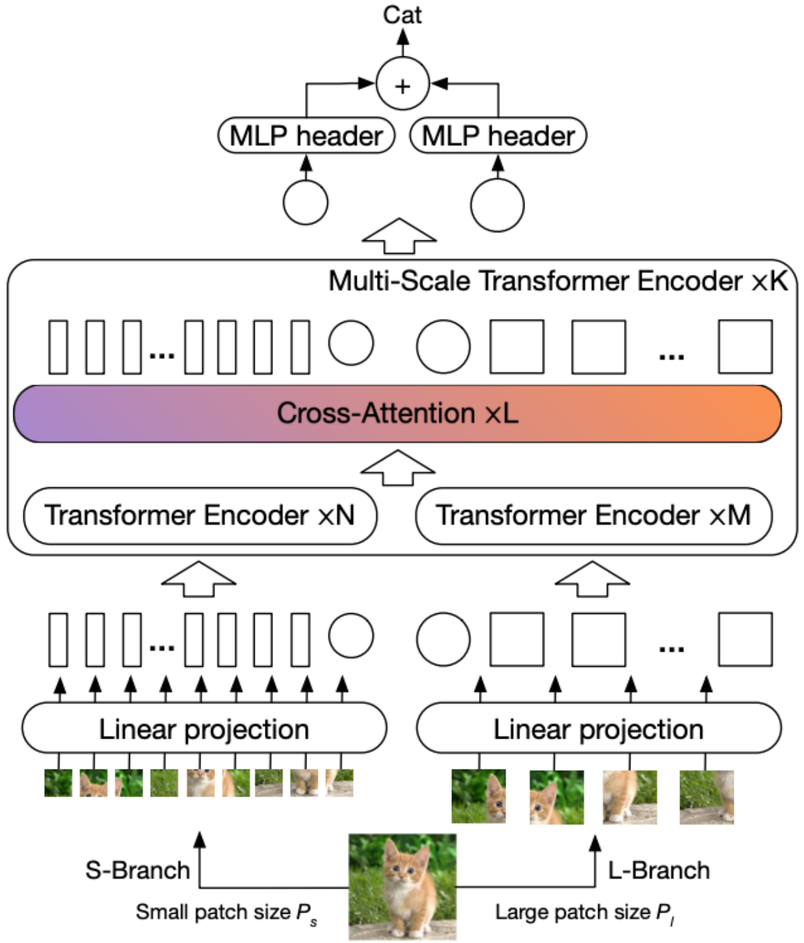
Swin Transformer V2

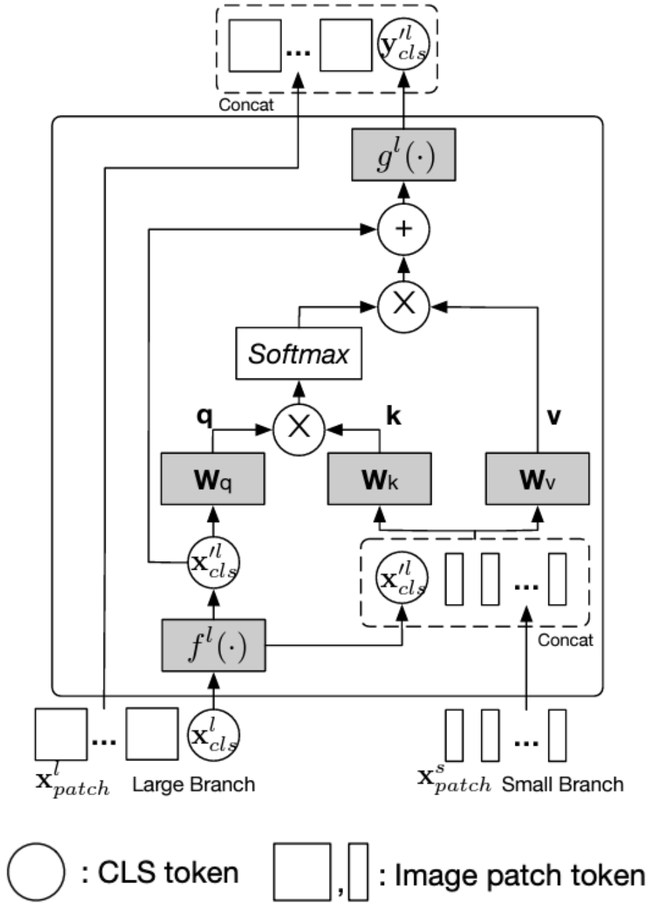


TRANSFORMATEC

Ze Liu et al. (2022) "Swin Transformer V2: Scaling Up Capacity and Resolution". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12009-12019).

Cross *ViT*





Chun-Fu Chen et al. (2021) "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification". Proceedings of the IEEE/CVF international conference on computer vision (pp. 357-366).

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

GRACIAS

Victor Flores Benites

TRANSFORMATEC