UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

Sesión 3.1

Object recognition

YOLO. Pose & Depth estimation

TRANSFORMATEC

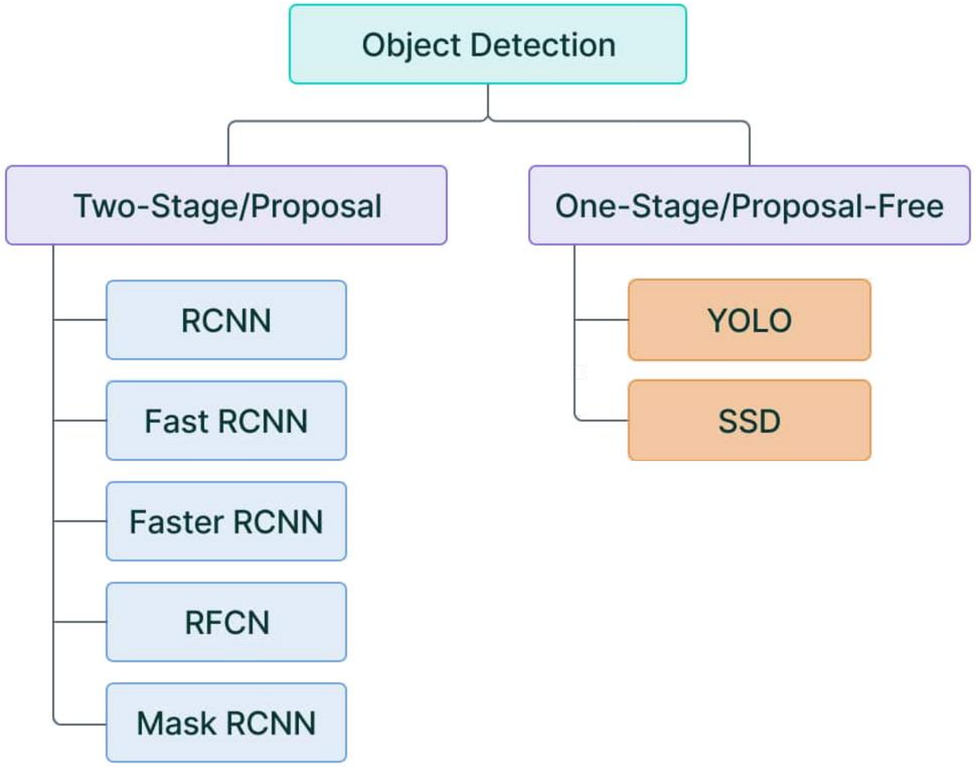
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

1.

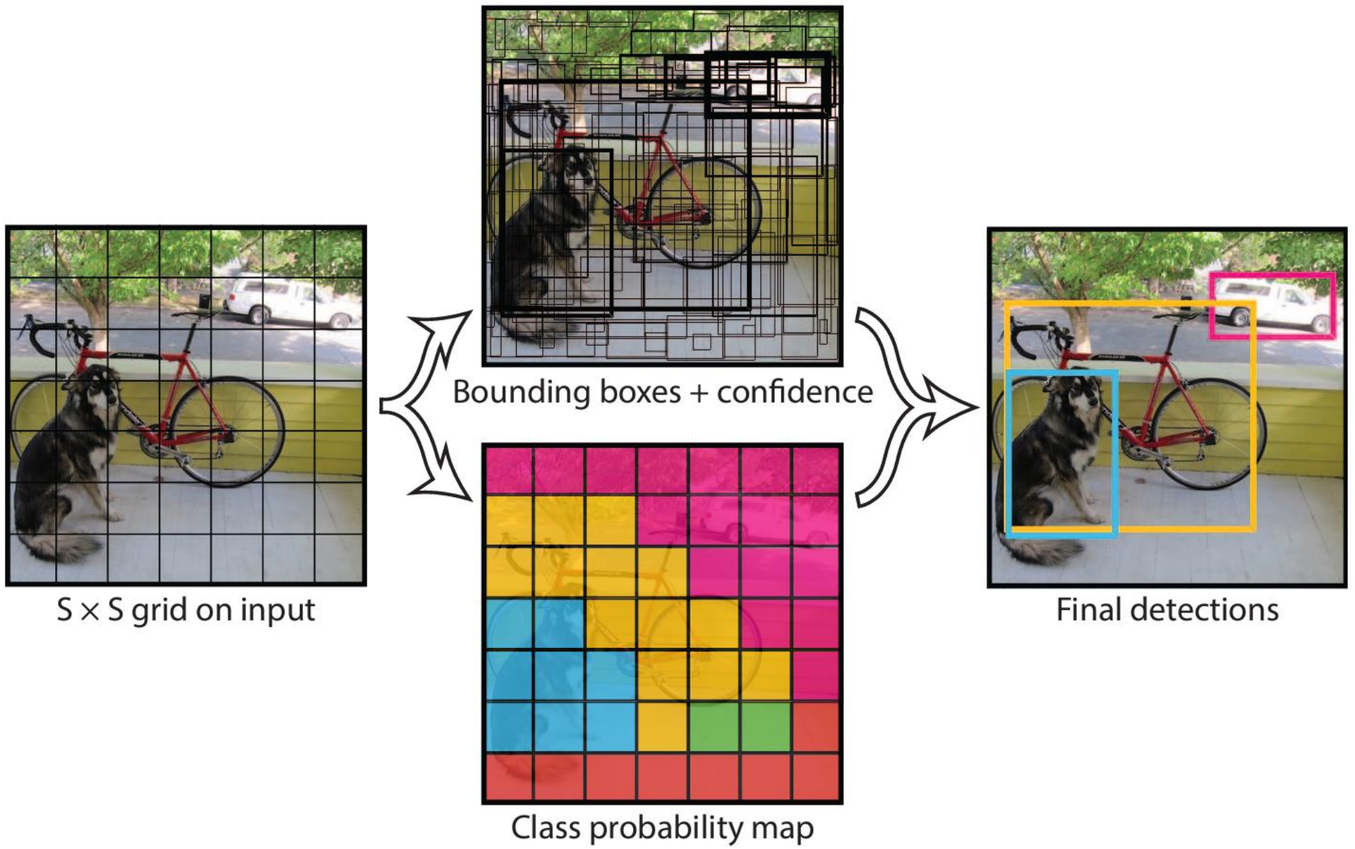
YOLO

TRANSFORMATEC

YOLO: *You Only Look Once*

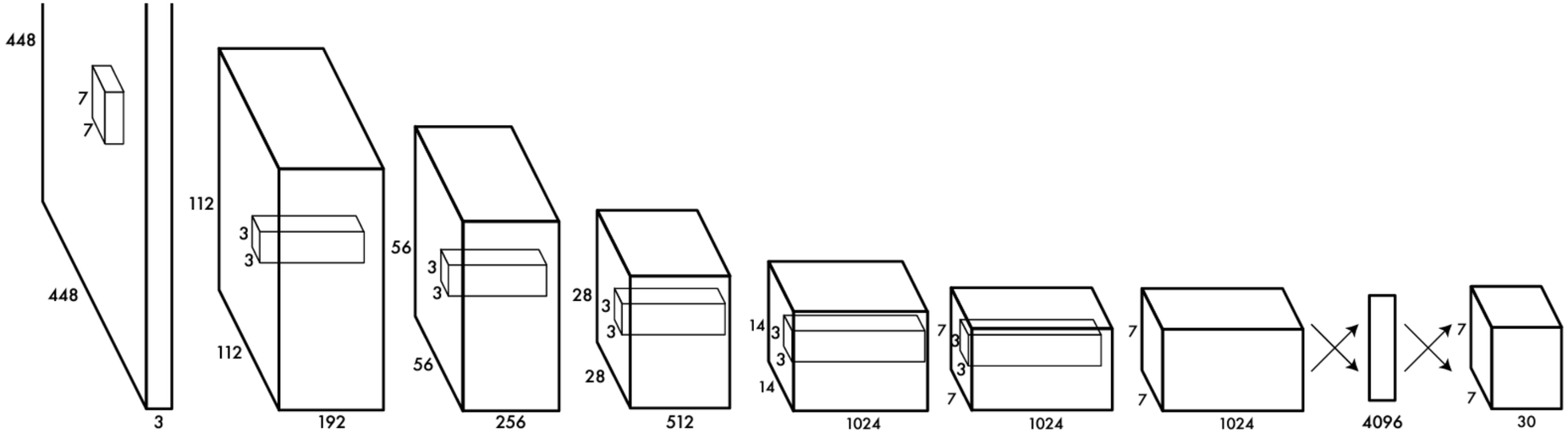


YOLO: *You Only Look Once*



Joseph Redmon et al. (2016) "You only look once: unified, real-time object detection". Proceedings of the IEEE conference on computer vision and pattern recognition.

YOLO: *You Only Look Once*



Joseph Redmon et al. (2016) "You only look once: unified, real-time object detection". Proceedings of the IEEE conference on computer vision and pattern recognition.

Localization loss

GT bbox x-coordinate in the ith cell

GT bbox y-coordinate in the ith cell

Predicted bbox x-coordinate in the ith cell

Predicted bbox y-coordinate in the ith cell

Set to 5 to increase the loss of bounding box predictions

Sum-squared error

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

For each grid cell

For each grid box

'1' if object appears in the ith cell and the jth box detect it, '0' otherwise

GT bbox height in the ith cell

$$+\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

Square root to reduce the range of the values

GT bbox width in the ith cell

Predicted bbox width in the ith cell

Predicted bbox height in the ith cell

Confidence loss

$$+\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[\left(C_i - \hat{C}_i \right)^2 \right]$$

GT confidence score

Predicted confidence score

Confidence error when an object is detected in the ith cell

Set to 0.5 to decrease the loss for empty boxes

$$+\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} \left[\left(C_i - \hat{C}_i \right)^2 \right]$$

'1' if there is no object in the ith cell, '0' otherwise

Confidence error when an object not detected in the ith cell

Clasification loss

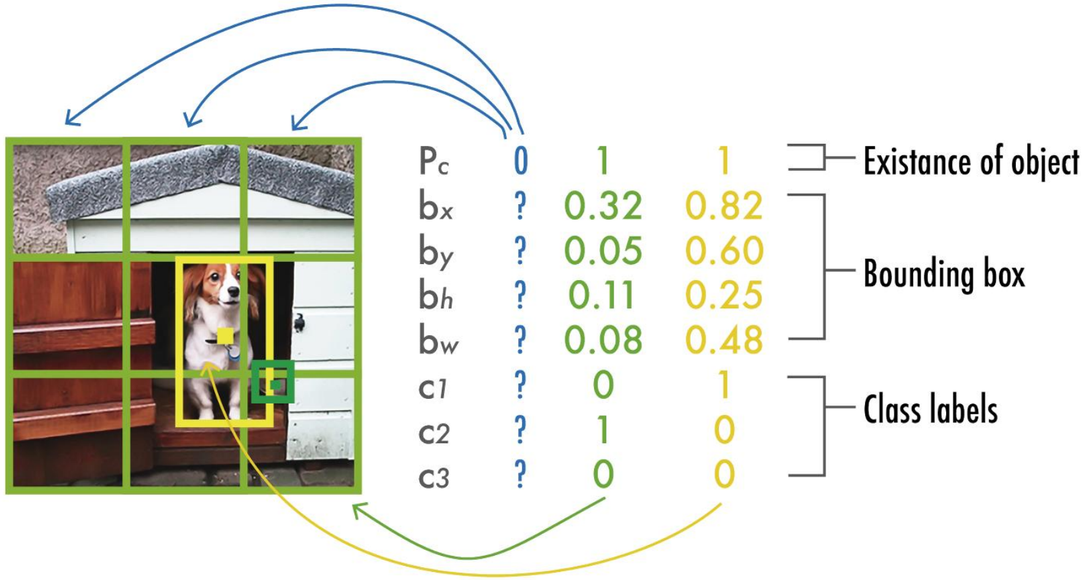
Predicted conditional probability of an object of class c appearing in the ith cell

$$+\sum_{i=0}^{S^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} \left[(p_i(c) - \hat{p}_i(c))^2 \right]$$

For each grid cell

For each class

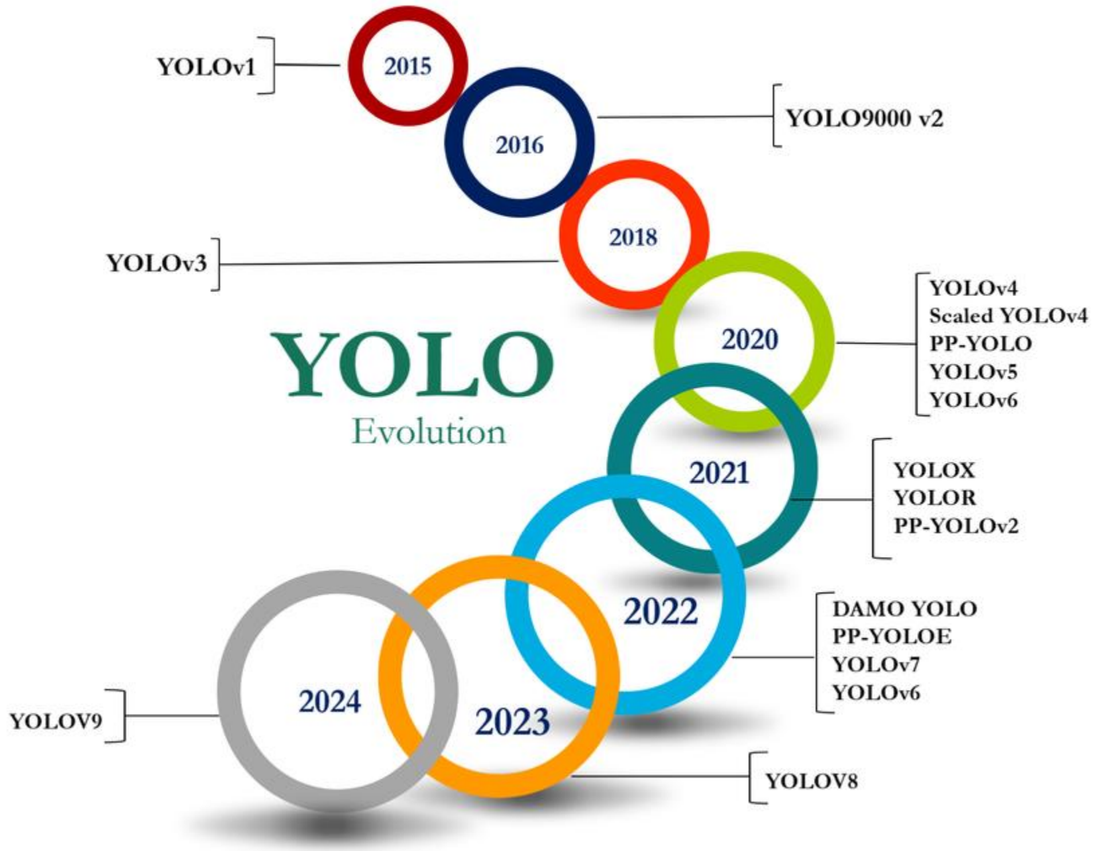
GT conditional probability of class c appearing in the ith cell



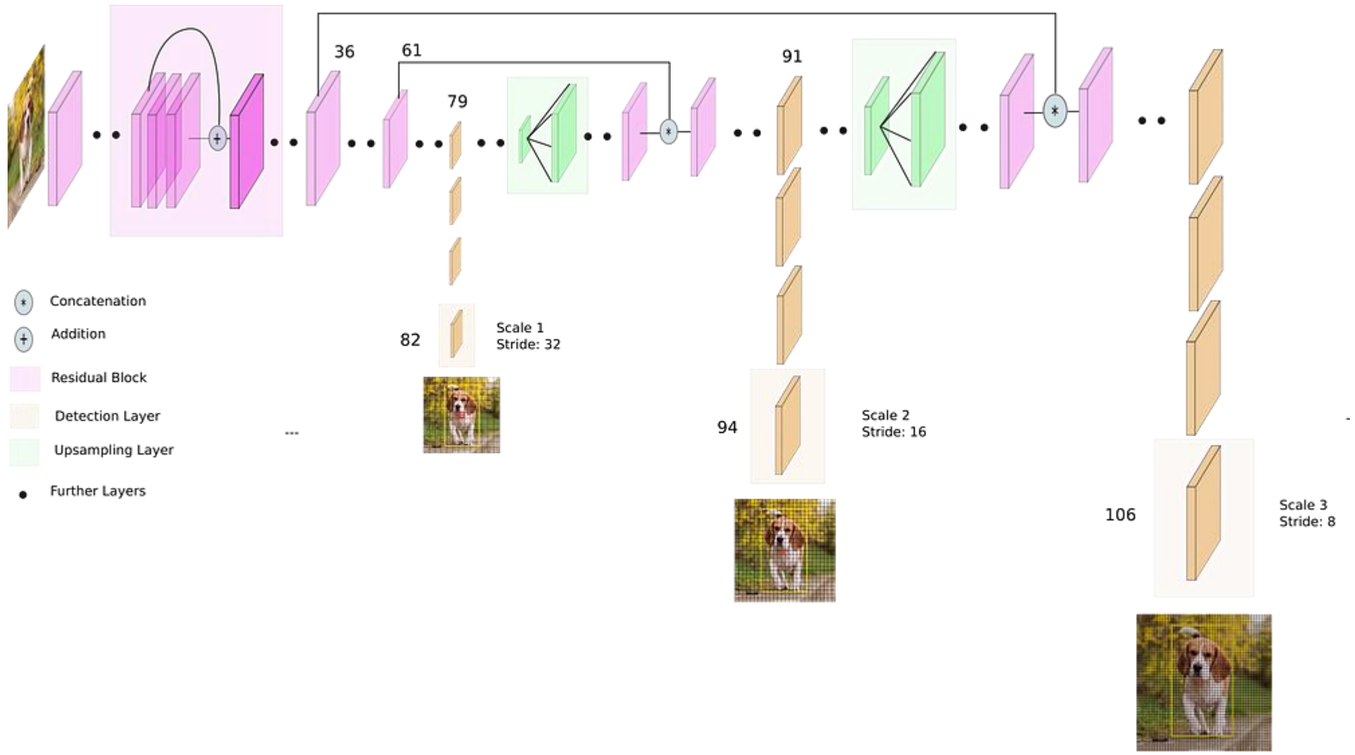
TRANSFORMATEC

Joseph Redmon et al. (2016) "You only look once: unified, real-time object detection". Proceedings of the IEEE conference on computer vision and pattern recognition.

YOLO: *You Only Look Once*

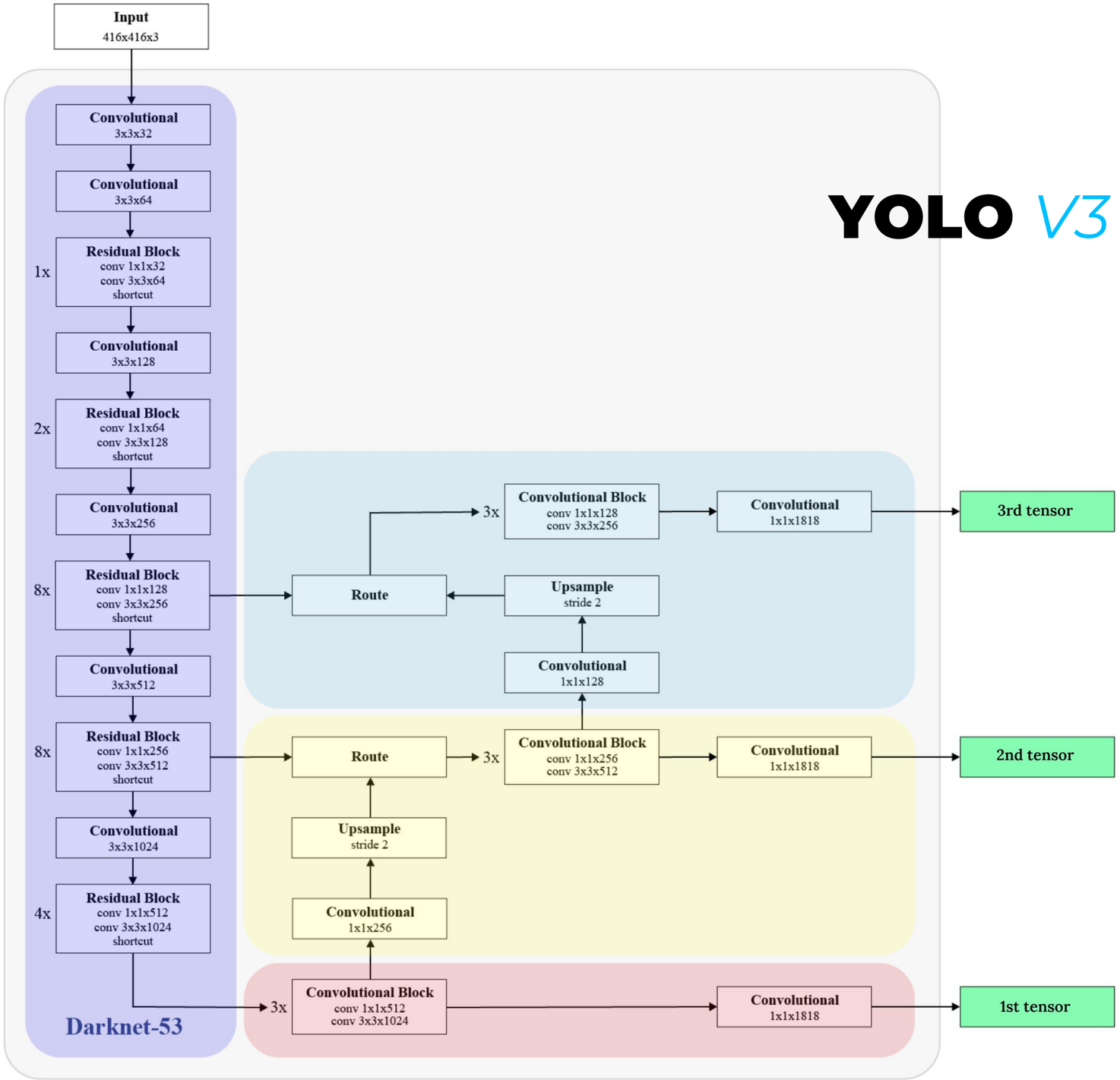


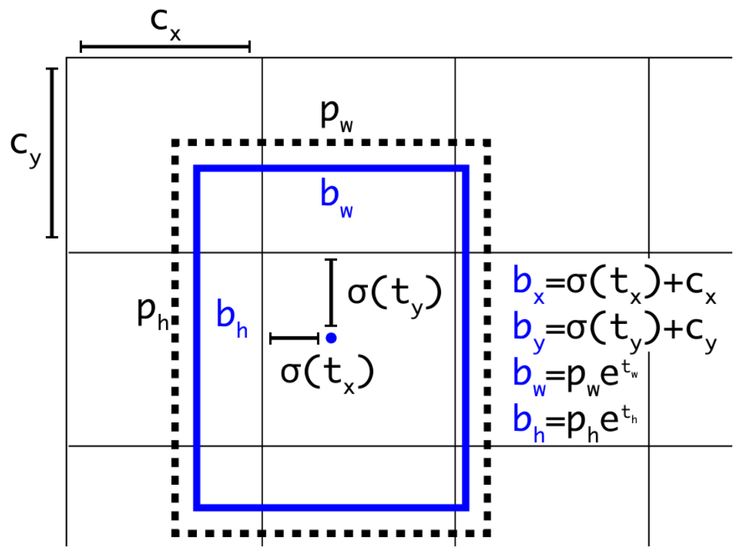
YOLO *V3*



Joseph Redmon and Ali Farhadi (2018) "YOLOv3: An Incremental Improvement". arXiv preprint arXiv:1804.02767

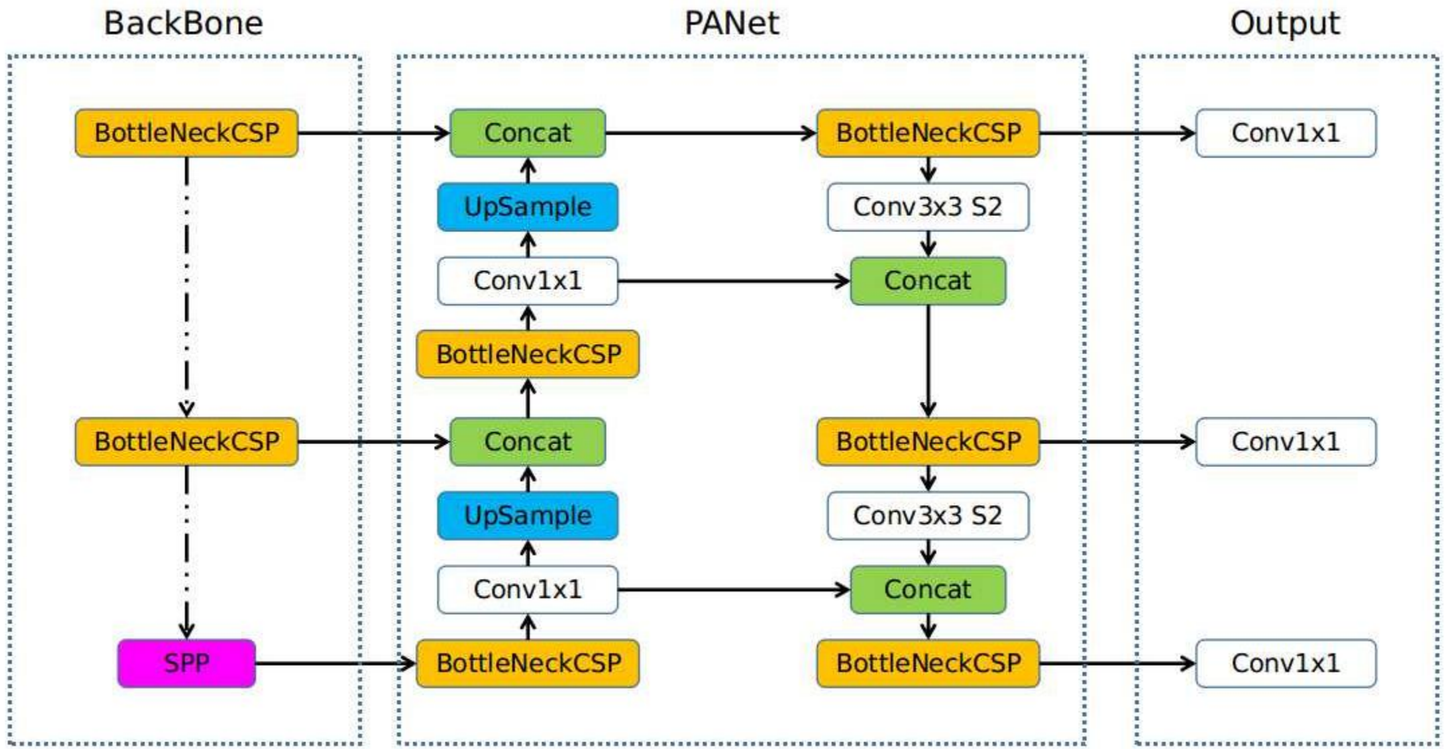
YOLO V3



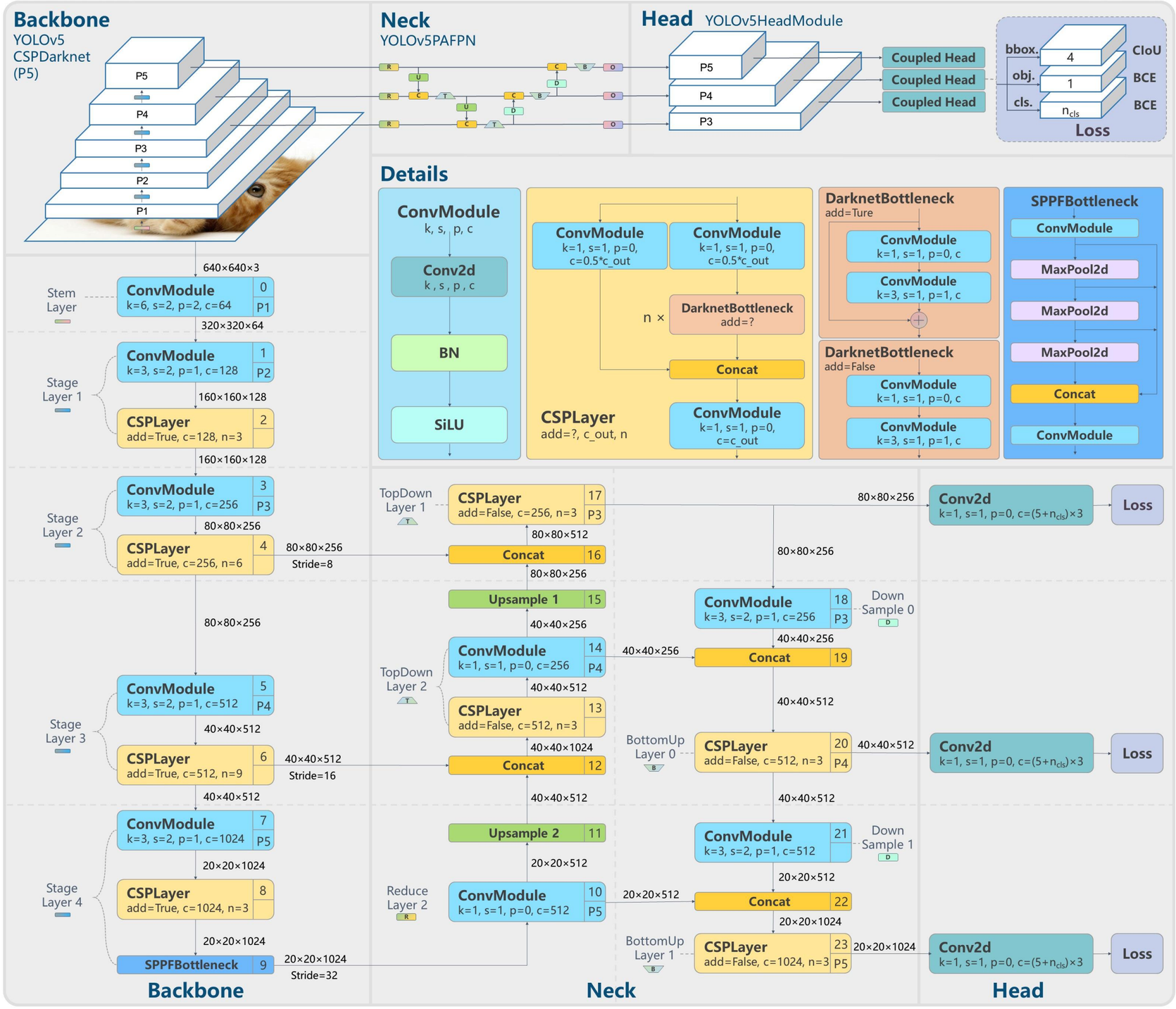


Joseph Redmon and Ali Farhadi (2018) "YOLOv3: An Incremental Improvement". arXiv preprint arXiv:1804.02767

YOLO *V5*



G. Jocher (2020) "YOLOv5 by Ultralytics".
https://github.com/ultralytics/yolov5

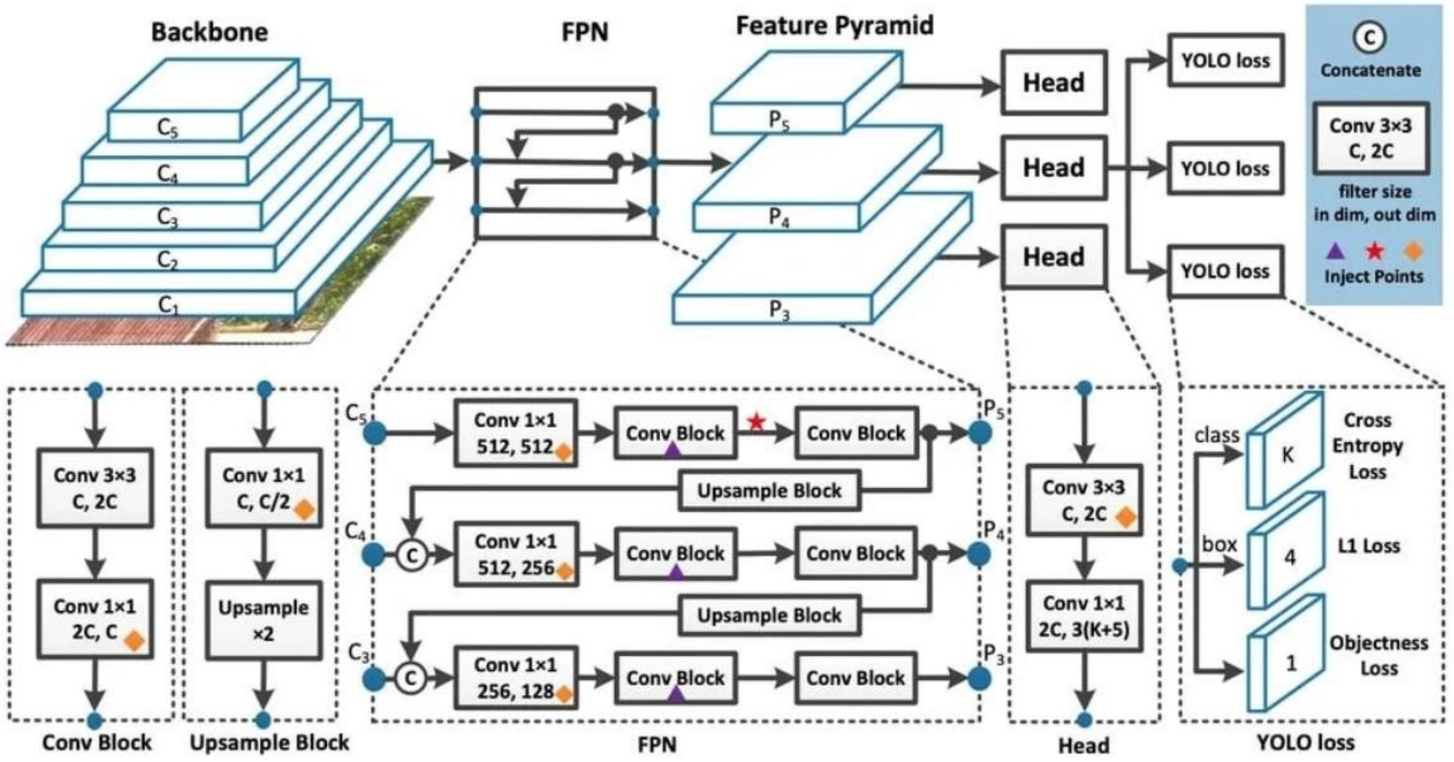


> Reinventa el mundo <

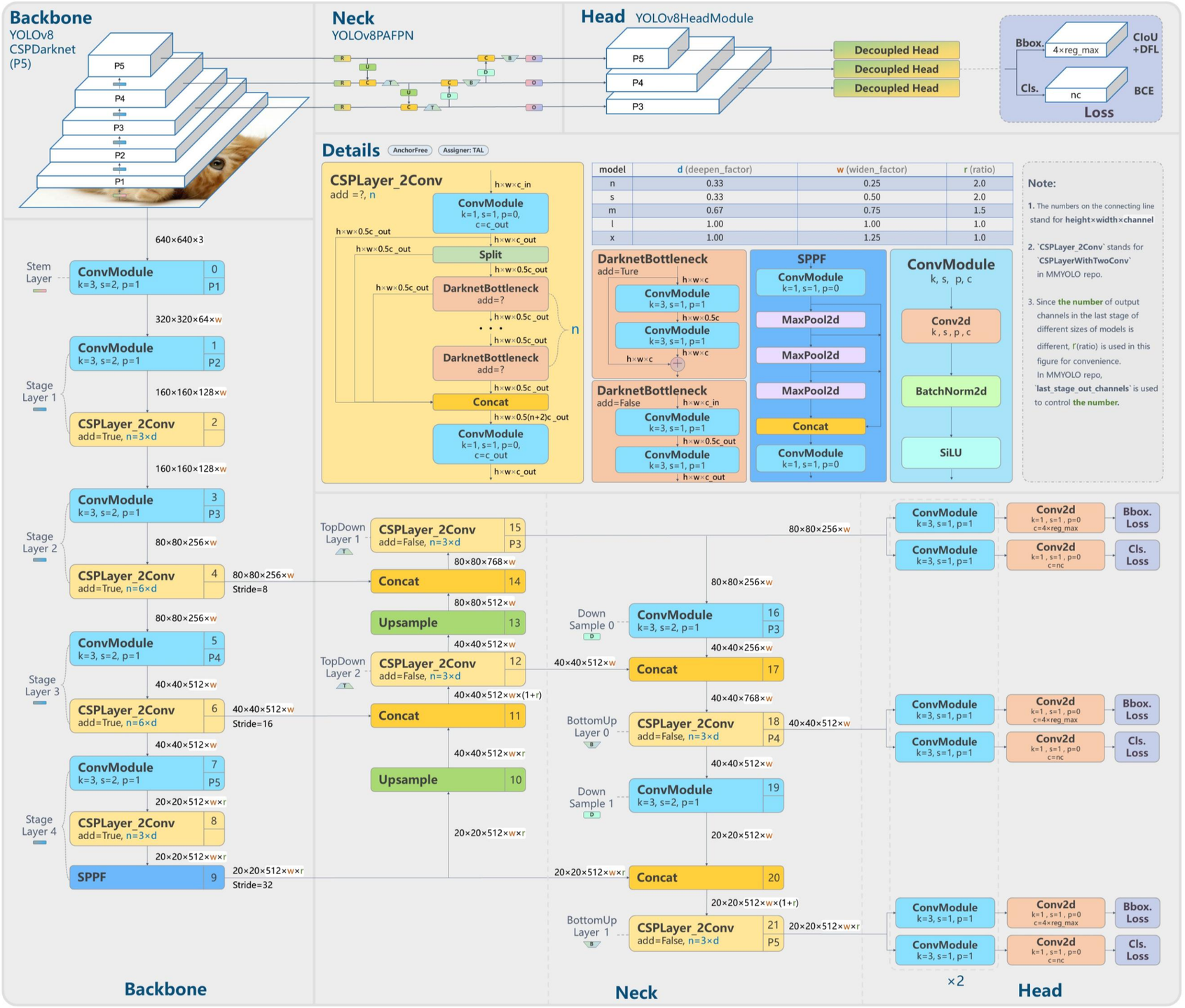
TRANSFORMATEC

G. Jocher (2020) "YOLOv5 by Ultralytics".
https://github.com/ultralytics/yolov5

YOLO *V8*



G. Jocher (2023) "YOLO by Ultralytics".
https://github.com/ultralytics/ultralytics



> Reinventa el mundo <

TRANSFORMATEC

G. Jocher (2023) "YOLO by Ultralytics".
https://github.com/ultralytics/ultralytics

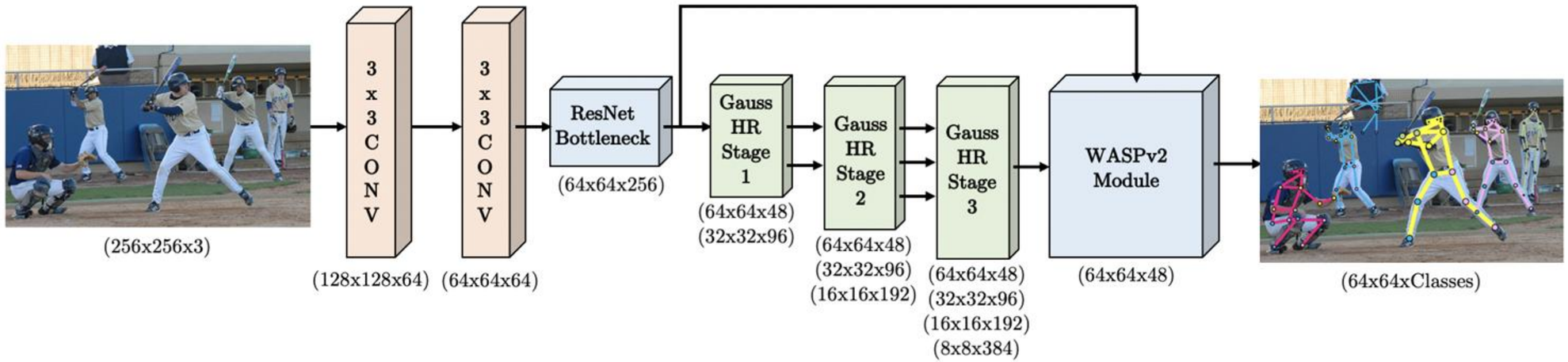
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

2.

Pose *Estimation*

TRANSFORMATEC

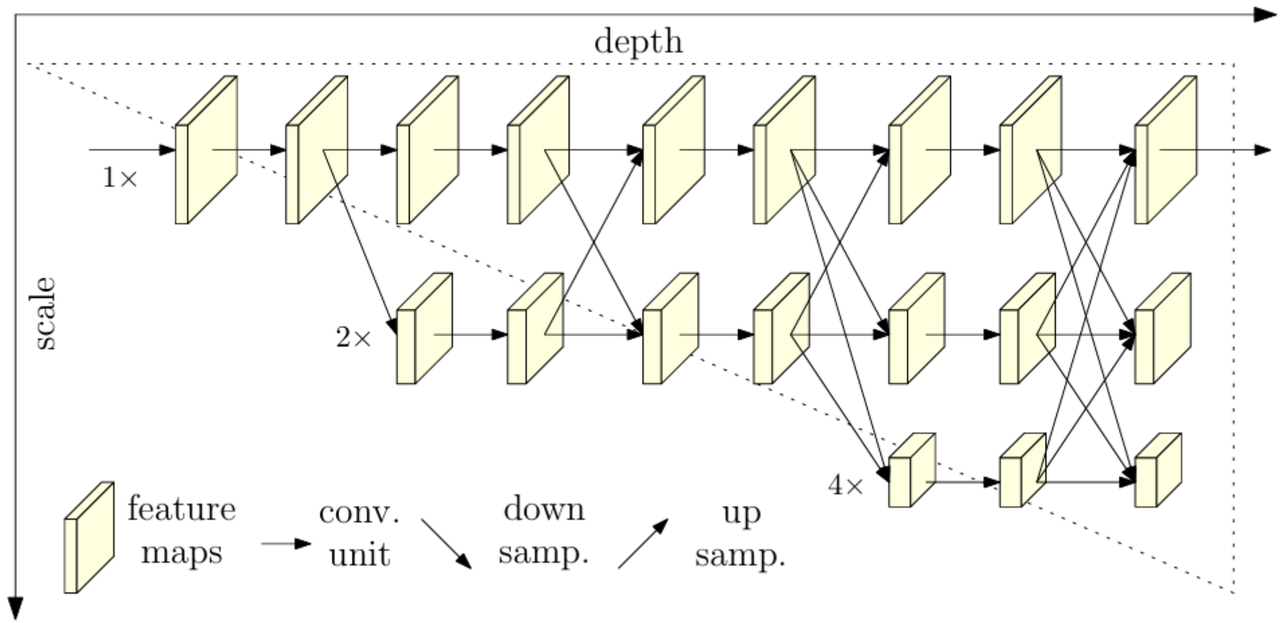
OmniPose



Bruno Artacho et al. (2021) "OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation". arXiv preprint arXiv:2103.10180.

OmniPose

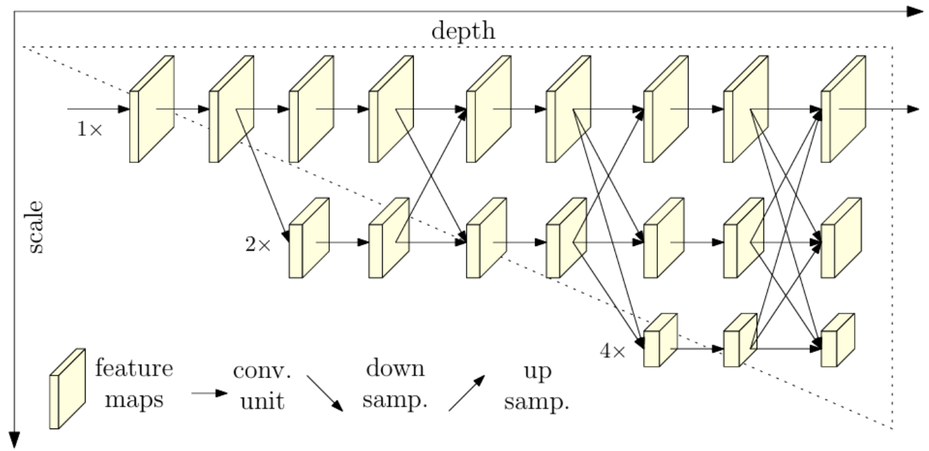
HR stage



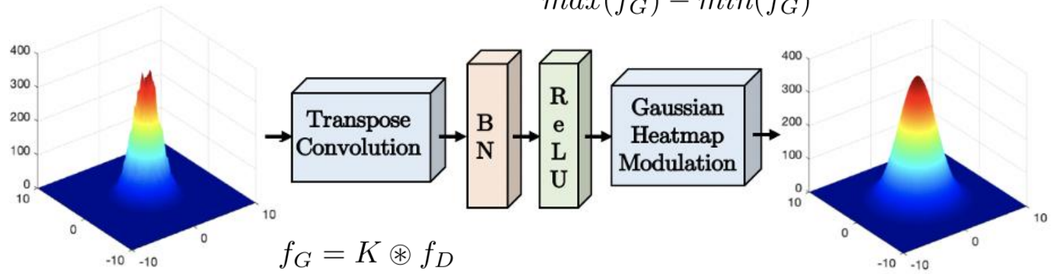
Ke Sun et al. (2019) "Deep High-Resolution Representation Learning for Human Pose Estimation". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5693-5703).

OmniPose

Gauss HR stage



$$f_{G_s} = \frac{f_G - min(f_G)}{max(f_G) - min(f_G)} * max(f_D)$$

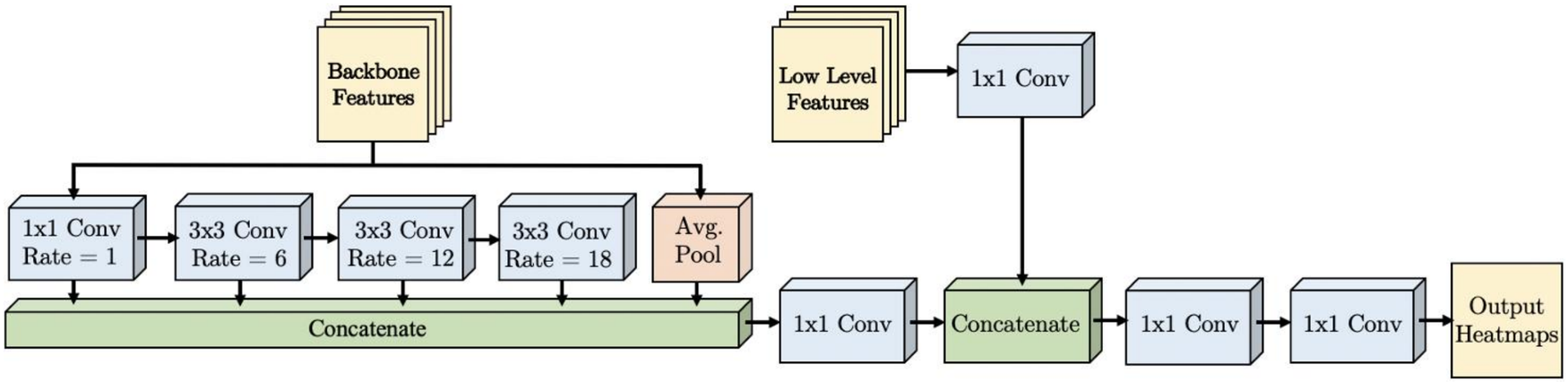


$$f_G = K \circledast f_D$$

Ke Sun et al. (2019) "Deep High-Resolution Representation Learning for Human Pose Estimation". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5693-5703).

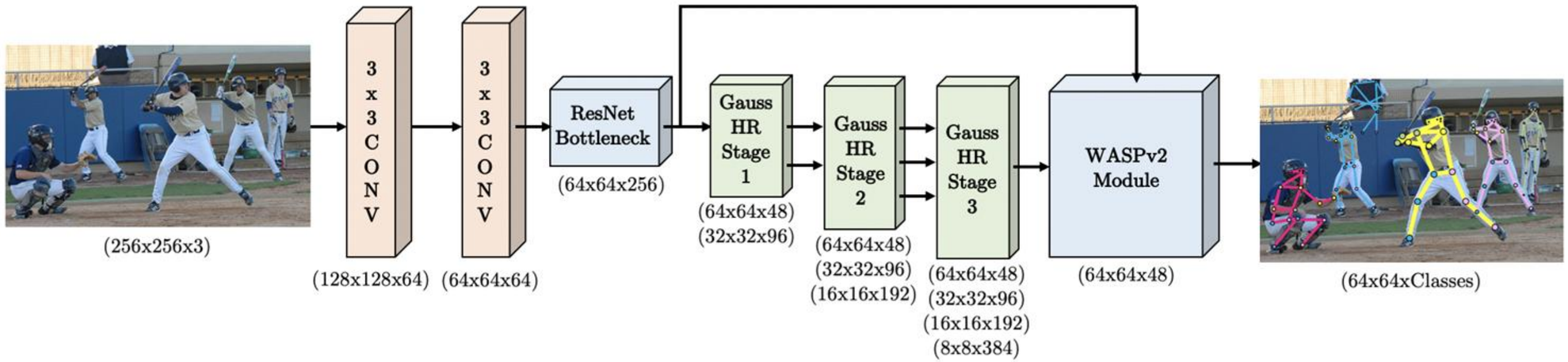
OmniPose

WASPv2 module



Bruno Artacho et al. (2021) "OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation". arXiv preprint arXiv:2103.10180.

Omni*Pose*



Bruno Artacho et al. (2021) "OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation". arXiv preprint arXiv:2103.10180.

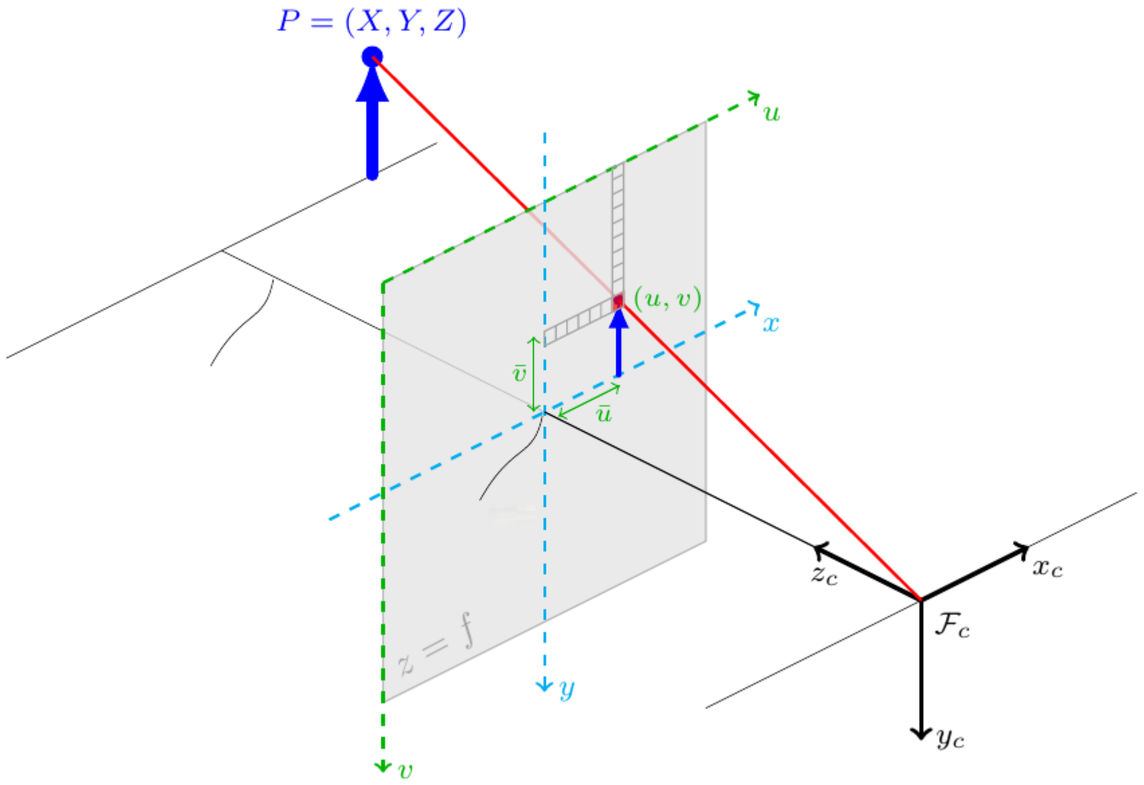
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

3.

Depth *Estimation*

TRANSFORMATEC

Modelado de la Proyección en Cámaras



Pinhole camera model

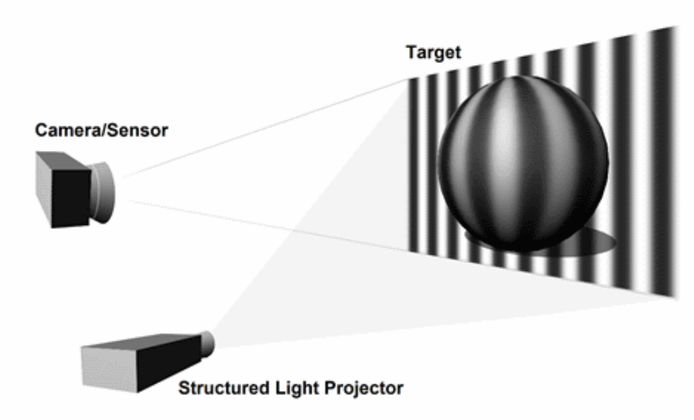
$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

donde:

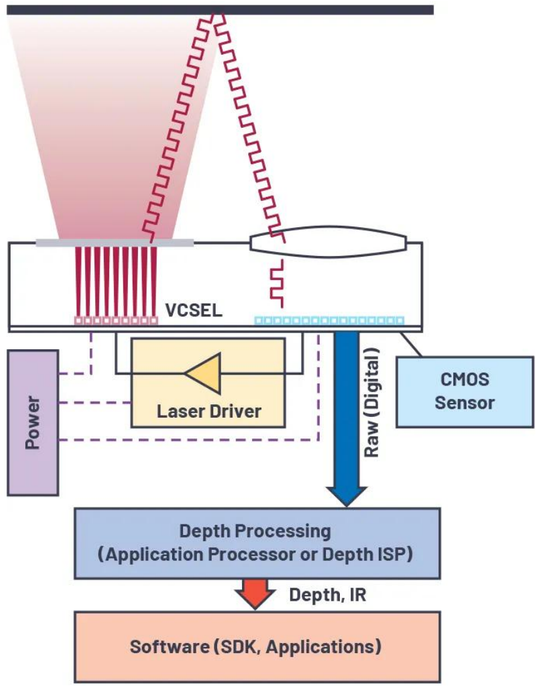
- (X, Y, Z) son las coordenadas 3D del punto en el espacio.
- (u, v) son las coordenadas proyectadas en la imagen.
- K es la matriz intrínseca de la cámara, que contiene la distancia focal f y el centro óptico.
- R y t representan la rotación y traslación de la cámara con respecto al mundo.
- s es un factor de escala.

Depth *images*

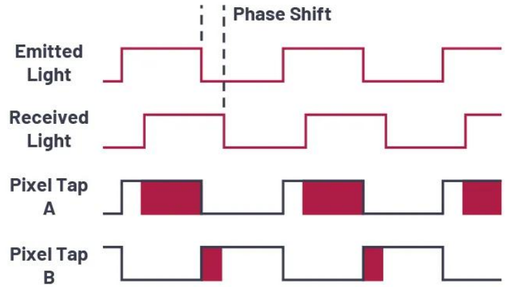
Structured Light



Time-of-flight (ToF)

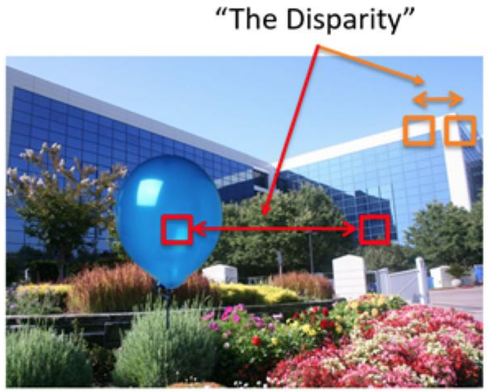


Sensor infrarrojo



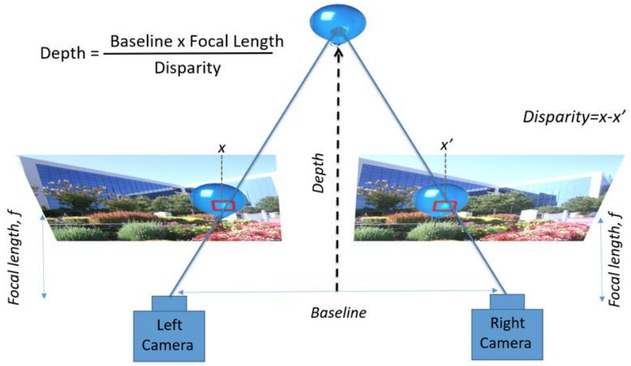
Passive *Stereo*

Left camera

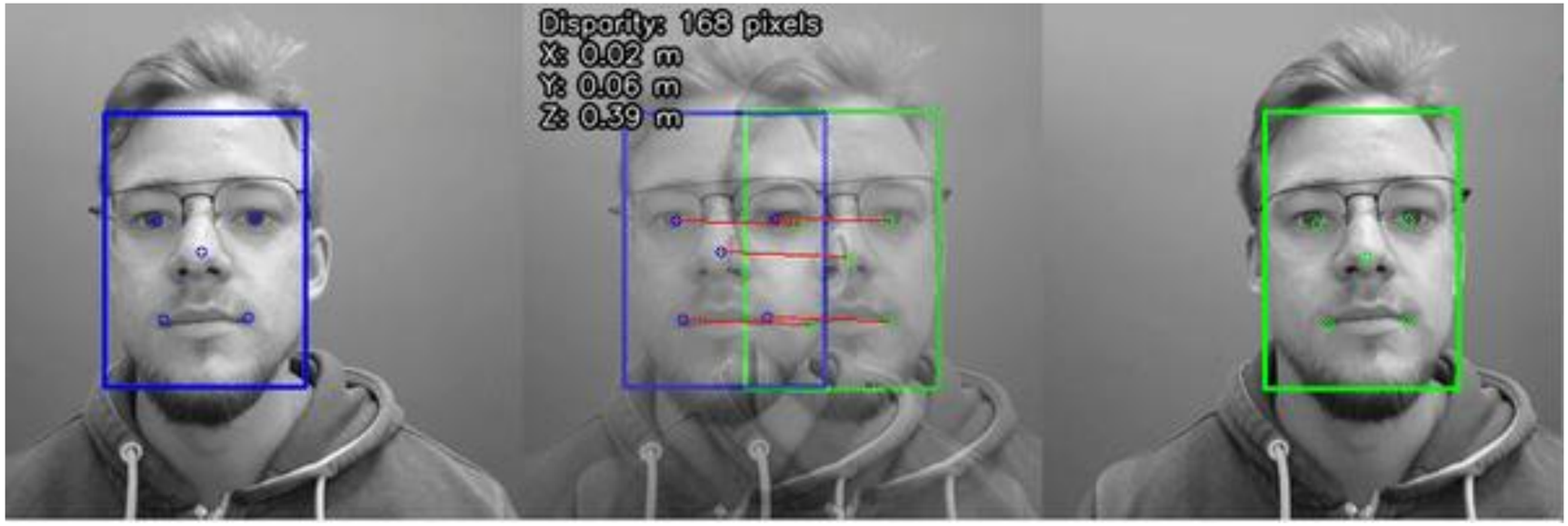


Right camera

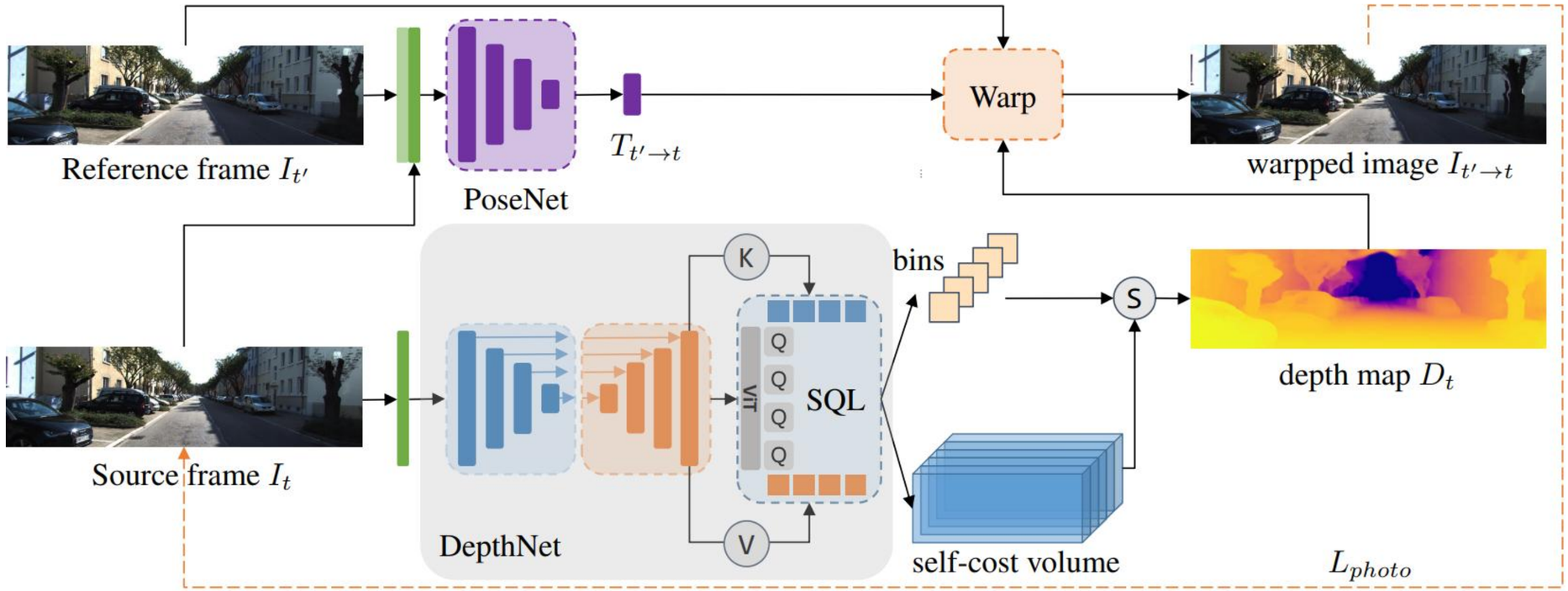
Disparity Map referenced to Right camera



Passive *Stereo*



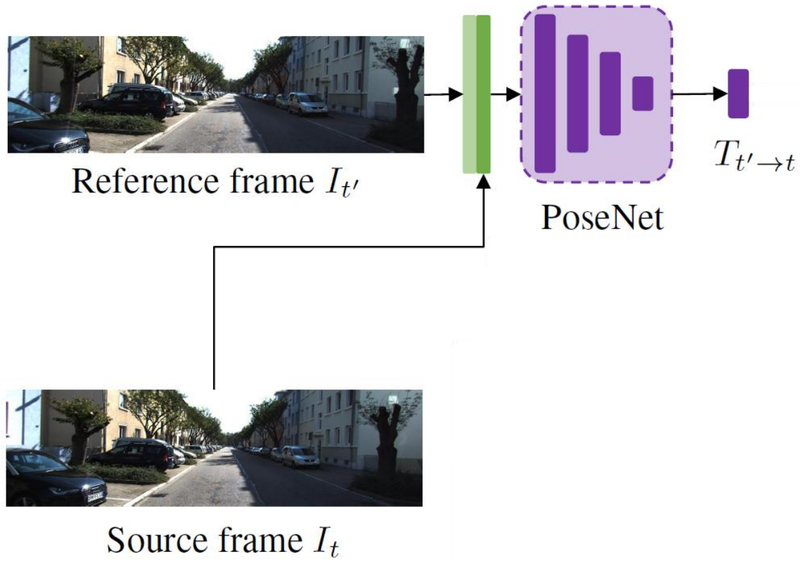
SQLdepth



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

PoseNet



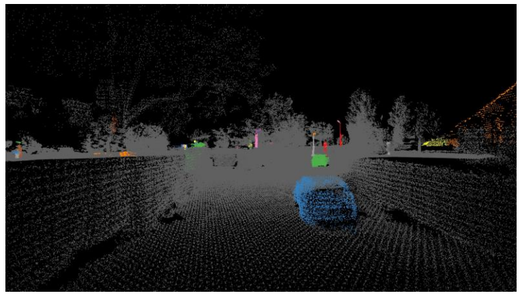
PoseNet sigue la arquitectura típica de una red de regresión convolucional y su objetivo es estimar la matriz de transformación

$$T_{t'\to t} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$$

Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

Geometric warp

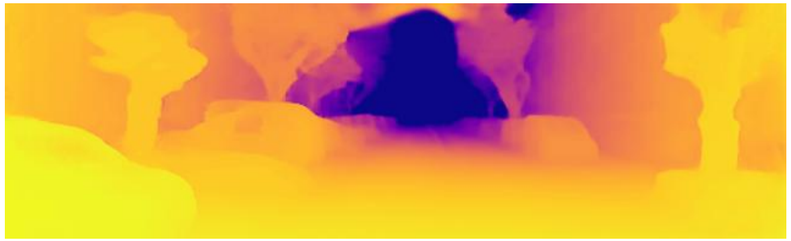


Source frame I_t

$$X_{t'} = T_{t'\to t} \cdot X_t$$

$$X_t = K^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \cdot D_t(u, v)$$

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = K \cdot X_{t'}$$



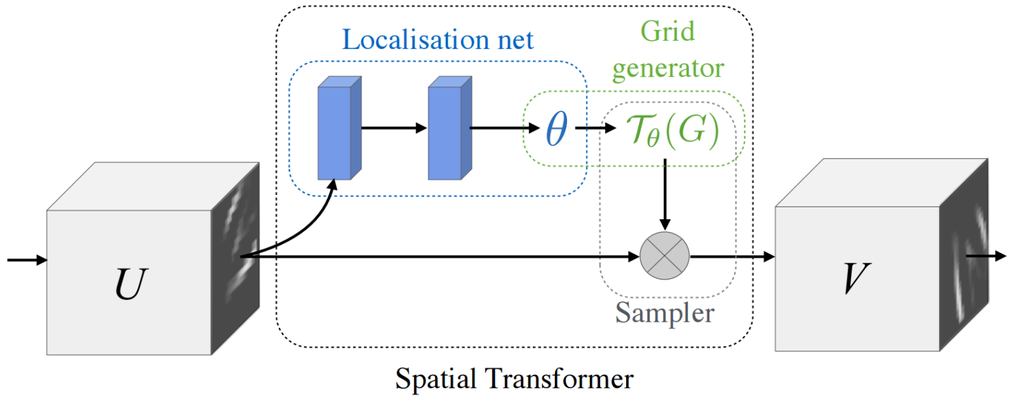
Depth image D_t

Source frame $I_{t'}$

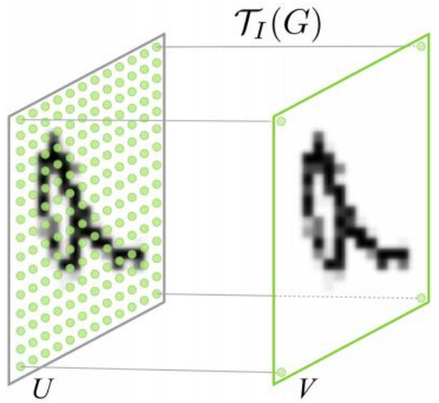
Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

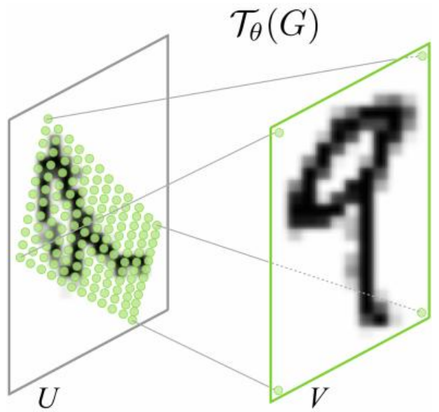
SQLdepth

Geometric warp



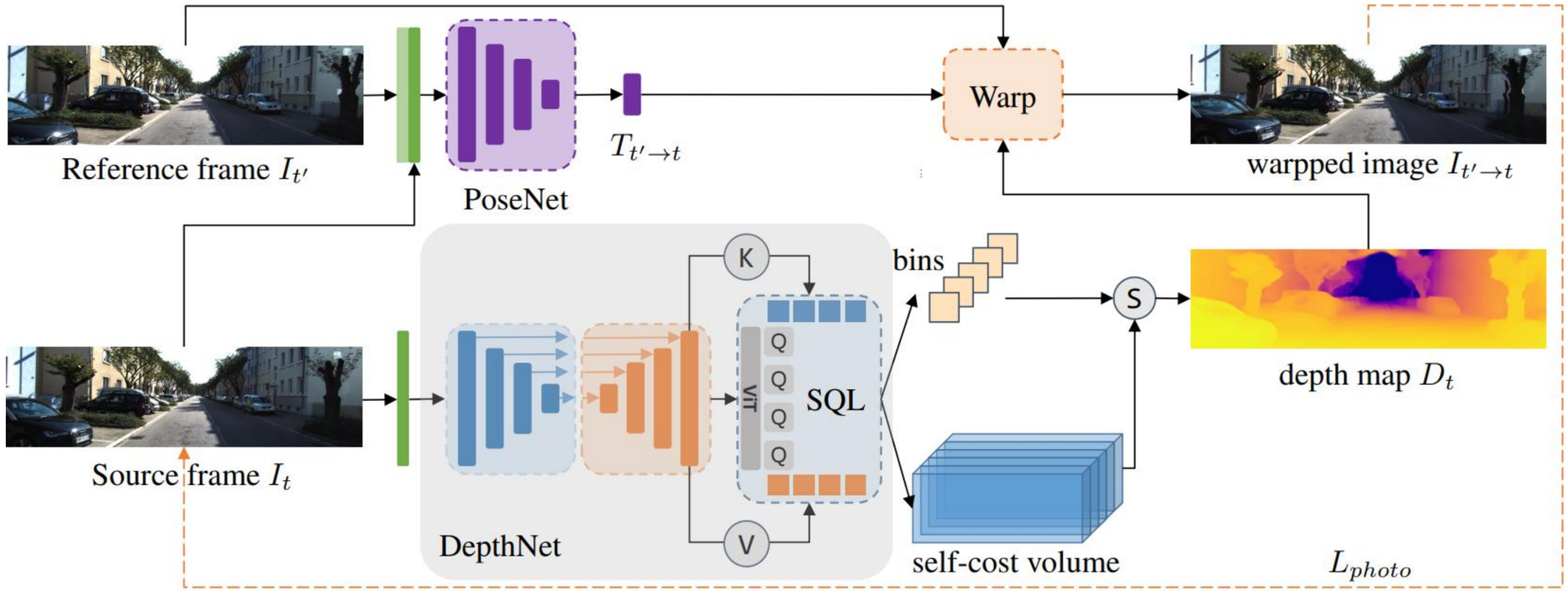
Spatial Transformer





Max Jaderberg et al. (2015) "Spatial Transformer Networks". Advances in neural information processing systems, 28.

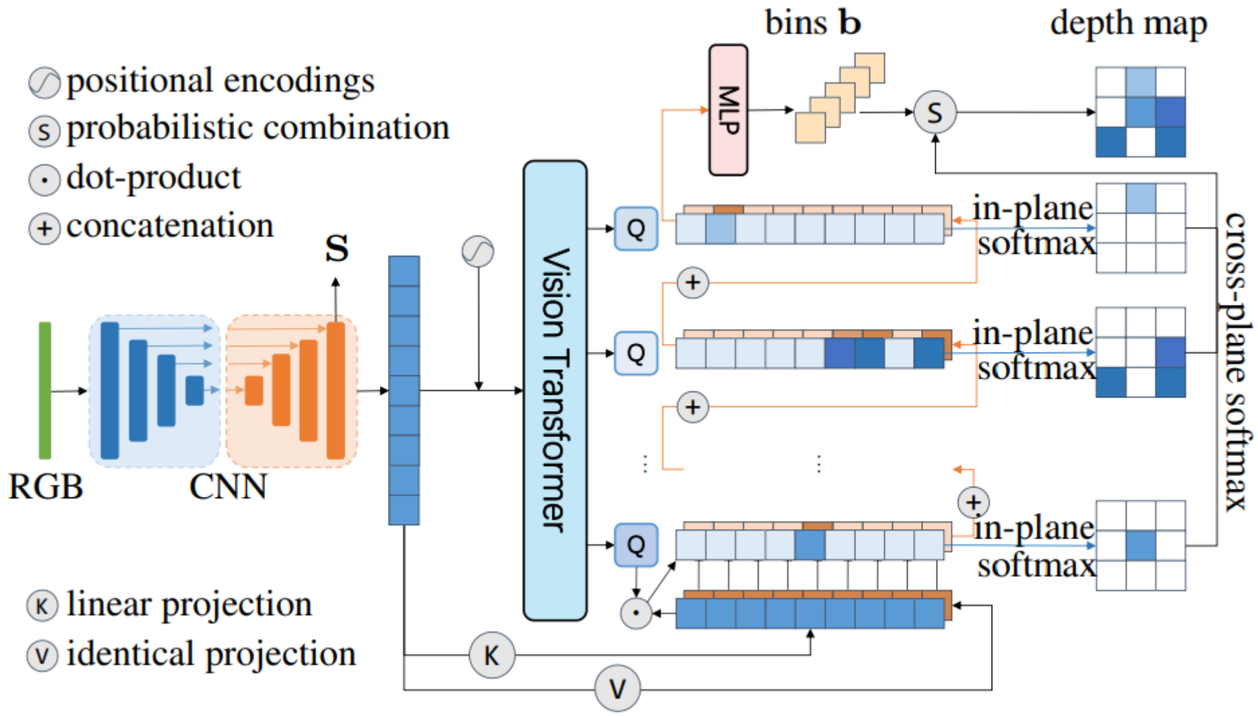
SQLdepth



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

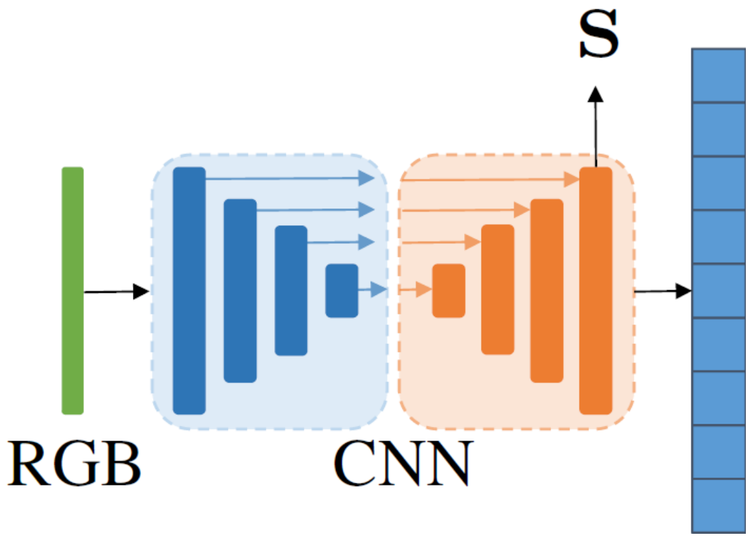
DepthNet with the Self Query Layer



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

U-net



Input:

Imagen RGB $I \in \mathbb{R}^{3\times H\times W}$.

Patche:

Feature map $F \in \mathbb{R}^{C\times\frac{h}{p}\times\frac{w}{p}}$, donde p es el tamaño del patch.

ViT input $F \in \mathbb{R}^{N\times C}$, donde $N = \frac{h}{p}\times\frac{w}{p}$.

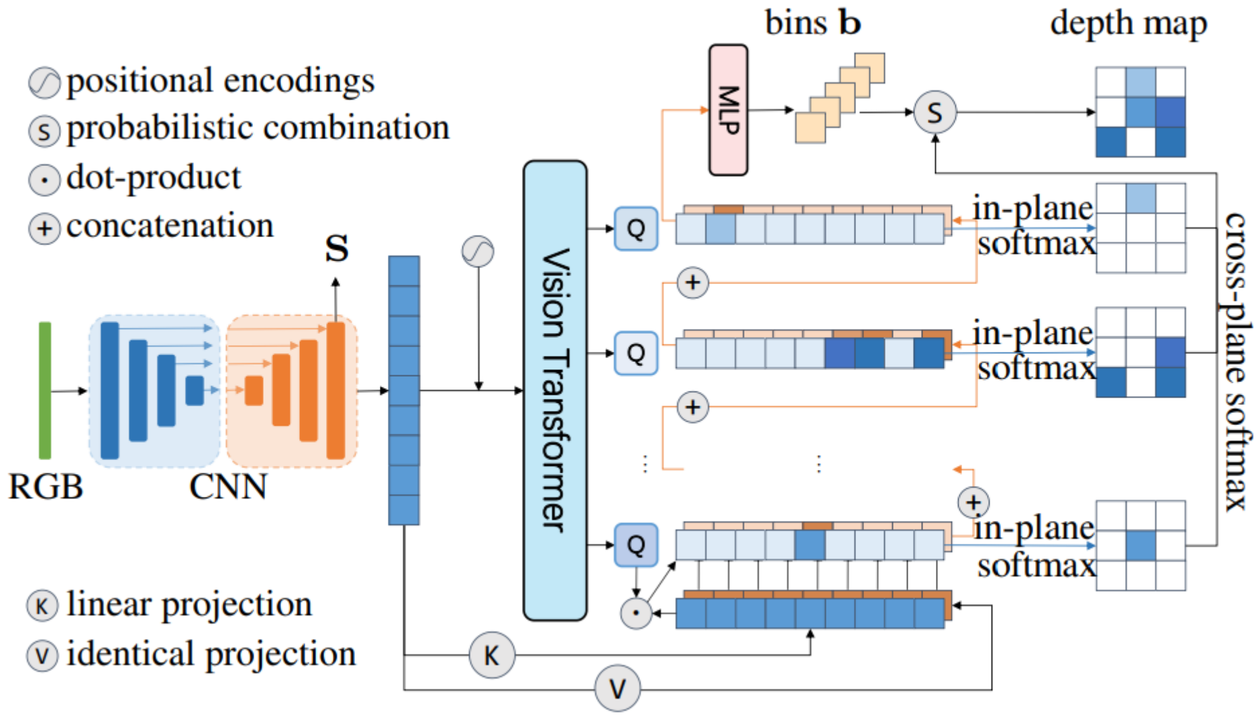
Queries:

Feature map $Q \in \mathbb{R}^{C\times Q}$, el cual abstrae estructuras espaciales importantes.

Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

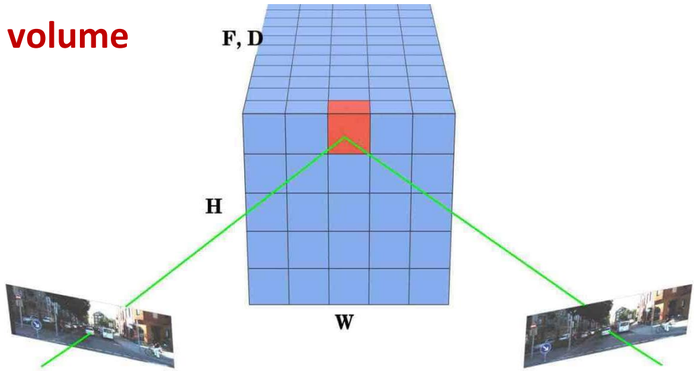
DepthNet with the Self Query Layer



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

Self-Query *Layer*

Cost volume



Para cada píxel en I_L, buscamos su posible correspondencia en I_R desplazándolo en diferentes niveles d.

$$V_{cost}(x, y, d) = \text{Similitud}\big(F_L(x, y), F_R(x - d, y)\big)$$

donde:

- F_L, F_R son los feature maps de las imágenes.
- d es el desplazamiento (disparidad en visión estéreo).
- $V_{cost} \in R^{H \times W \times D}$, cada posición almacena una similitud entre píxeles de ambas imágenes.

Self-cost volume

En lugar de comparar píxeles entre dos imágenes, se compara cada píxel con representaciones globales de proto-objetos en la imagen.

$$V(i, j, k) = Q_k^T \cdot S_{i,j}$$

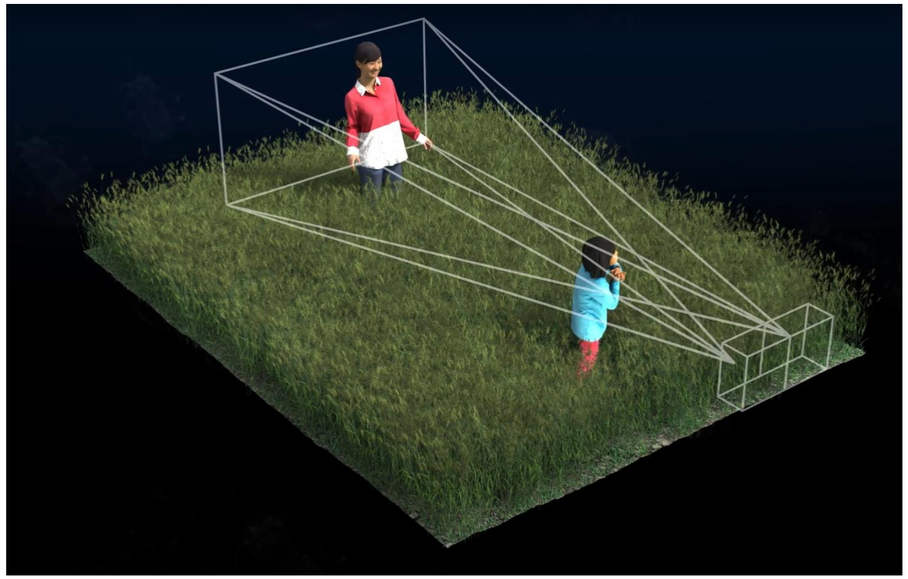
donde:

- $S_{i,j}$ es la representación de características en el píxel (i, j).
- Q_k es el vector de características del proto-objeto k (obtenido de ViT).
- $V \in R^{h \times w \times Q}$ indica la relación de similitud entre el píxel y cada proto-objeto.

Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

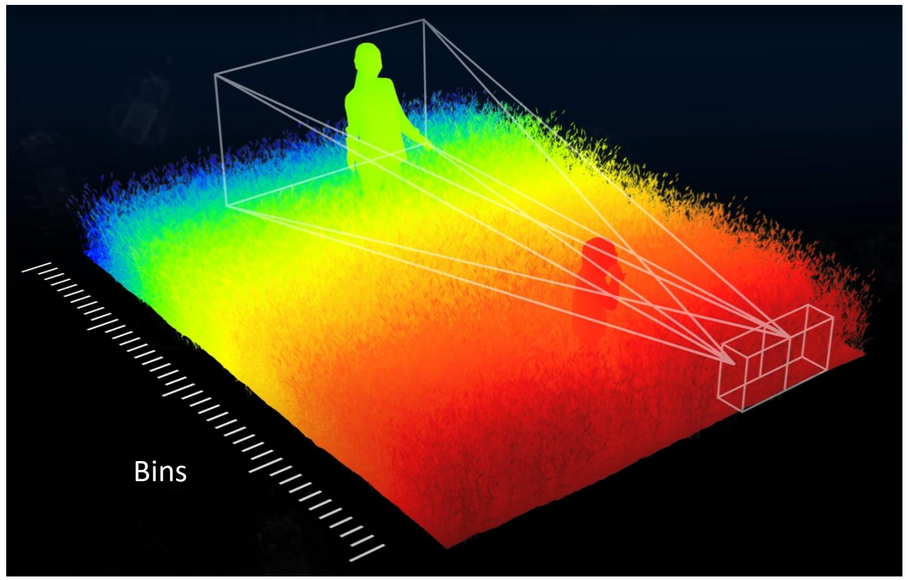
Depth bins



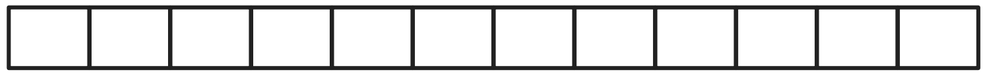
Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

Depth bins



K depth bins b



$$\mathrm{b} = \mathrm{MLP}\left(\bigoplus_{k=1}^{Q} \sum_{i,j} \mathrm{Softmax}(V_k)_{i,j} \cdot S_{i,j}\right)$$

Center depth of the i-th bin: $c(b_k) = d_{min} + (d_{max} - d_{min})\left(\frac{b_k}{2} + \sum_{t=1}^{k-1} b_t\right)$

Probabilidad de que el píxel (i, j) pertenezca al bin k

$$p_{i,j,k} = \mathrm{softmax}(\hat{V}_{i,j,k})$$

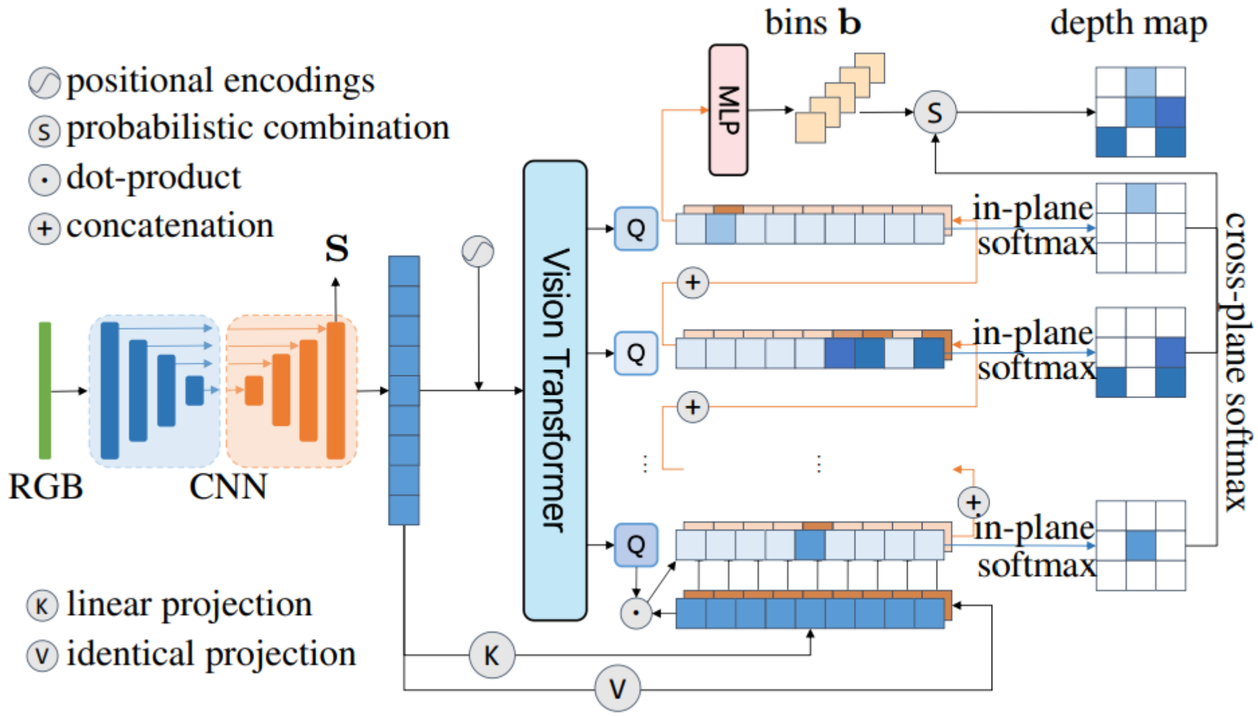
Profundidad estimada para el píxel (i, j)

$$\tilde{d}_{i,j} = \sum_{k=1}^{K} c(b_k) \cdot p_{i,j,k}$$

Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

DepthNet with the Self Query Layer



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

SQLdepth

Loss function $L = \mu L_{photo} + \lambda L_s$

Photometric loss: $L_{photo} = \min_{t'} pe(I_t, I_{t'\to t}), \quad I_{t'} \in \{t-1, t+1\}$

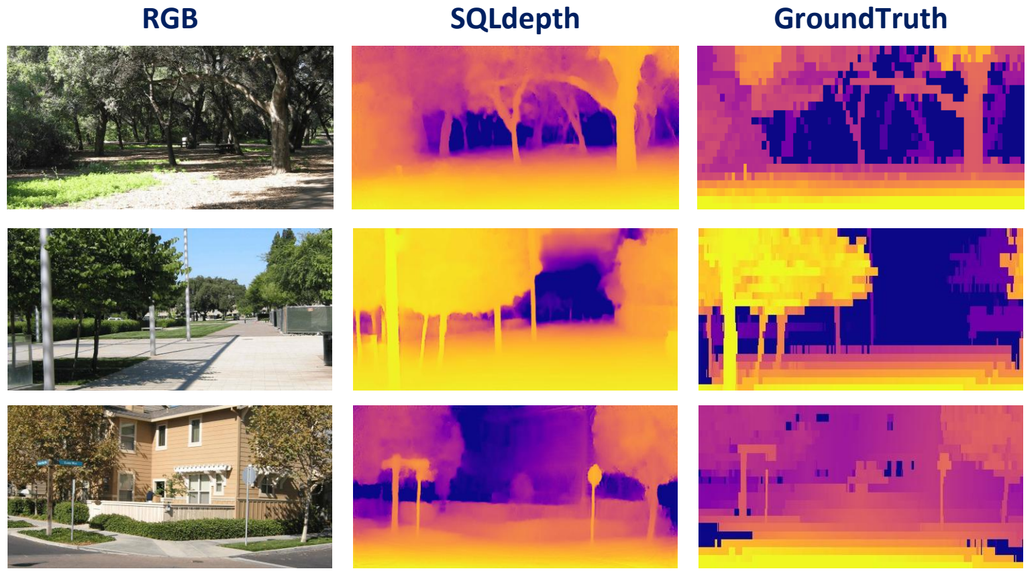
Auto-masking: $\mu = \left[\min_{t'} pe(I_t, I_{t'\to t}) < \min_{t'} pe(I_t, I_{t'})\right]$

Photometric error:

$$pe(I_a, I_b) = \frac{\alpha}{2}\left(1 - SSIM(I_a, I_b)\right) + (1-\alpha)|I_a - I_b|_1$$

Edge-aware smooth loss:

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}$$



Youhong Wang et al. (2024) "SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation". Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 5713-5721).

Structural similarity *index measure (SSIM)*

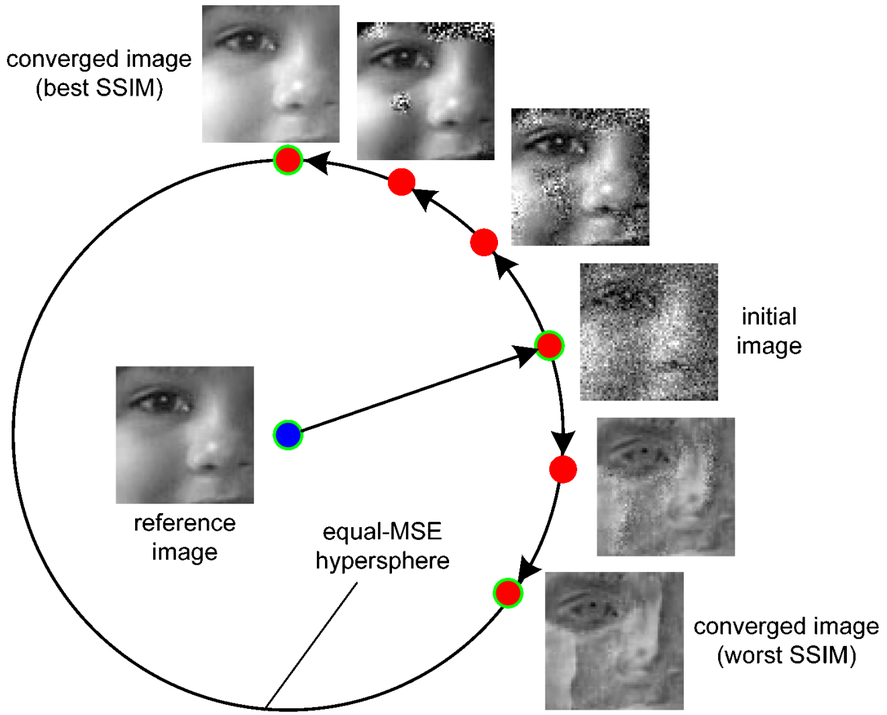
$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Donde:

- μ_x y μ_y son las medias locales de las imágenes x e y.
- σ_x^2 y σ_y^2 son las varianzas locales de x e y, respectivamente.
- σ_{xy} es la covarianza entre x e y.
- C_1 y C_2 son constantes de estabilización para evitar divisiones por cero, generalmente definidas como:

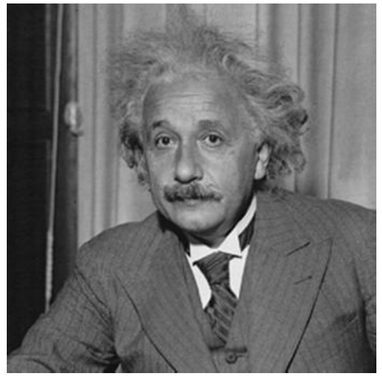
$$C_1 = (K_1 L)^2, \quad C_2 = (K_2 L)^2$$

donde L es el rango dinámico de los valores de píxeles (por ejemplo, 255 para imágenes de 8 bits), y K_1 y K_2 son pequeñas constantes típicamente 0.01 y 0.03, respectivamente.

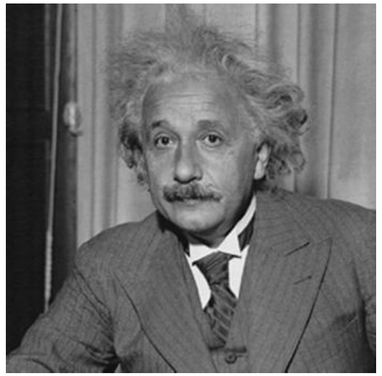


Zhou Wang et al. (2004) "Image quality assessment: from error visibility to structural similarity".
IEEE transactions on image processing, 13(4), 600-612.

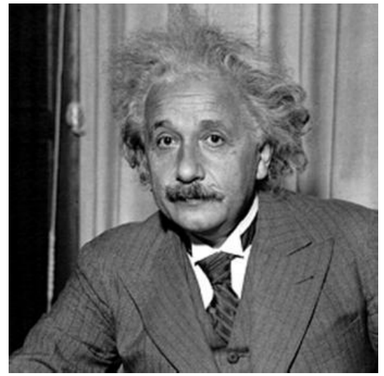
Structural similarity *index measure (SSIM)*



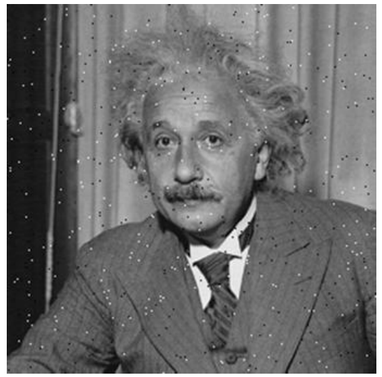
Original, MSE = 0, SSIM = 1



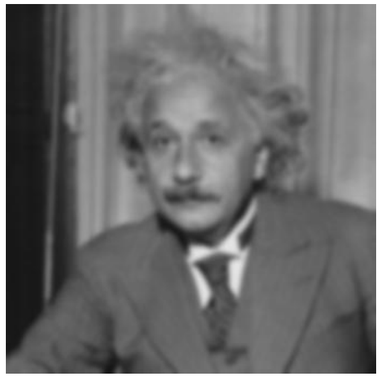
MSE = 144, SSIM = 0.988



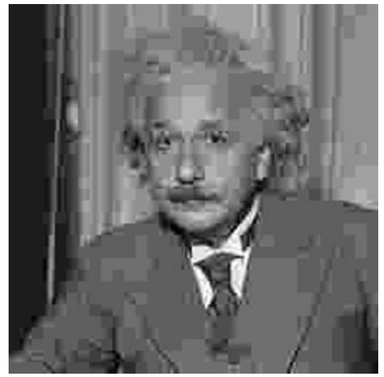
MSE = 144, SSIM = 0.913



MSE = 144, SSIM = 0.840



MSE = 144, SSIM = 0.694



MSE = 142, SSIM = 0.662

Zhou Wang et al. (2004) "Image quality assessment: from error visibility to structural similarity".
IEEE transactions on image processing, 13(4), 600-612.

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

GRACIAS

Victor Flores Benites

TRANSFORMATEC