

Segmentation d'image et Estimation de posture

A. Carlier

2025

Plan du cours

1 Segmentation d'image

2 Estimation de pose

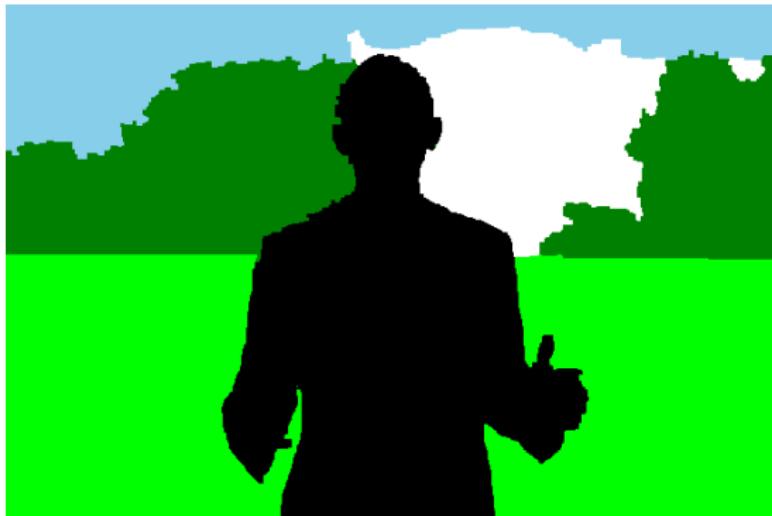
La base de données MS COCO



plus de 300K images, et plus de 2,5M d'instances d'objets segmentés (plus de 22 heures-humain pour 1000 segmentations)

[Lin et al.] Microsoft COCO : Common Objects in Context.

Les défis de la segmentation



Un grand nombre de prédictions : une pour chaque pixel de l'image.

Les défis de la segmentation



Pixels du ciel

Image originale

Pixels de la mer

La prédiction pour un pixel donné nécessite une analyse d'information locale conjointe à une compréhension plus globale de la scène.

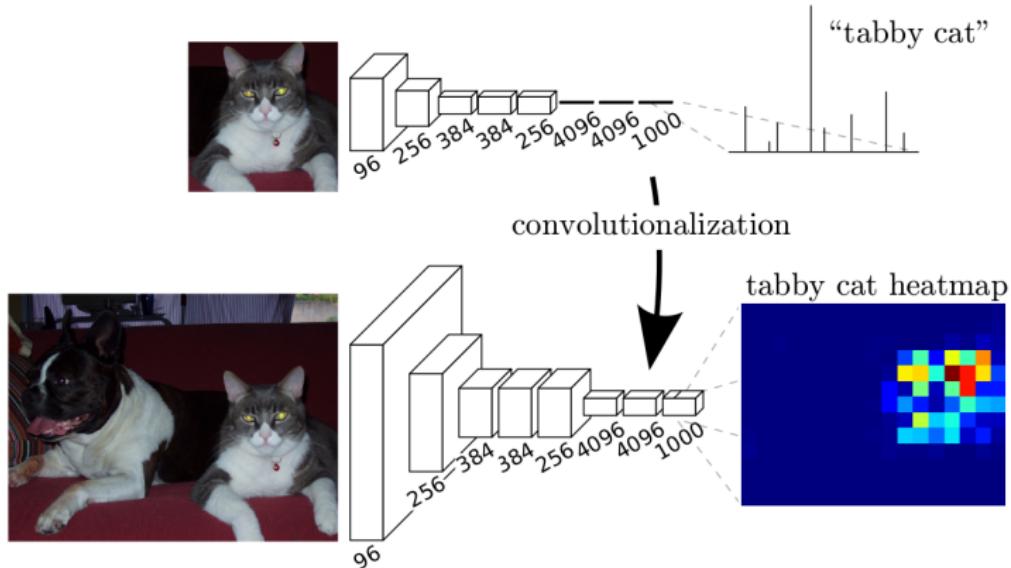
Segmentation par fenêtre glissante



Inconvénients :

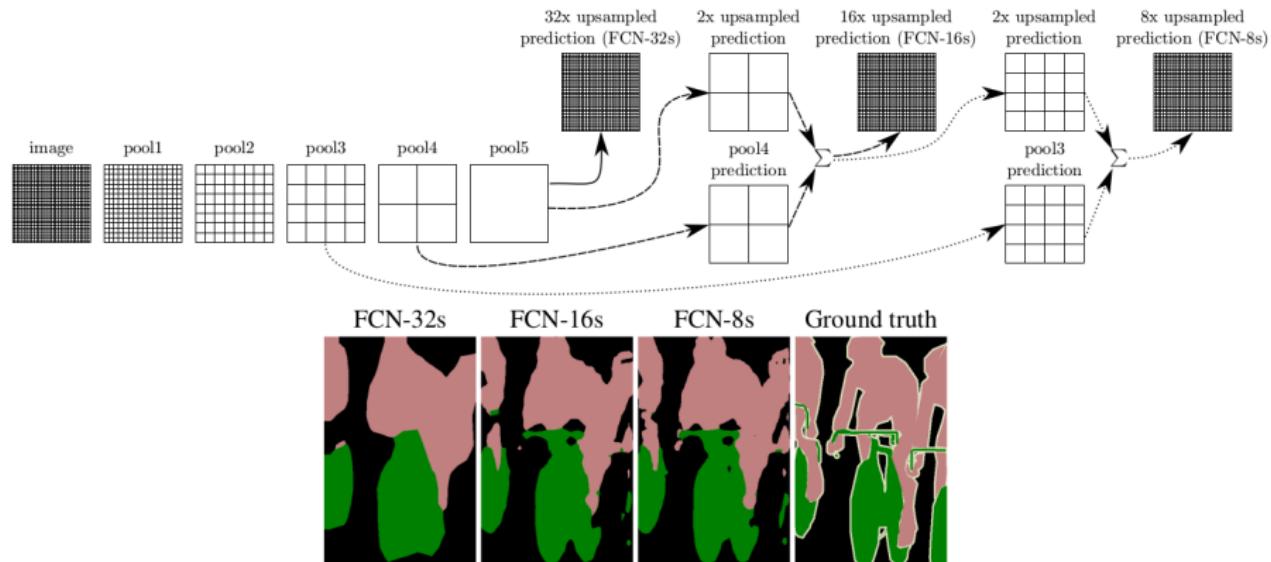
- Très coûteux.
- Pas de prise en compte de l'image globale.

Fully Convolutional Networks (2014)



- Premier algorithme à planter une segmentation *end-to-end*.
- Repose sur une reformulation des réseaux entièrement convolutifs.
- Base VGG-16.

Fully Convolutional Networks (2014)



- Résultat affiné en prenant en compte des prédictions intermédiaires.

[Long et al.] Fully Convolutional Networks for Semantic Segmentation.

L'opération de déconvolution (ou convolution transposée)

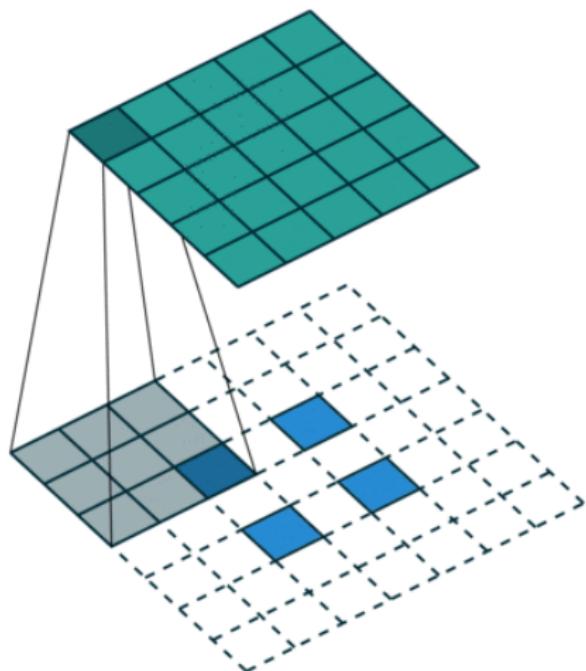
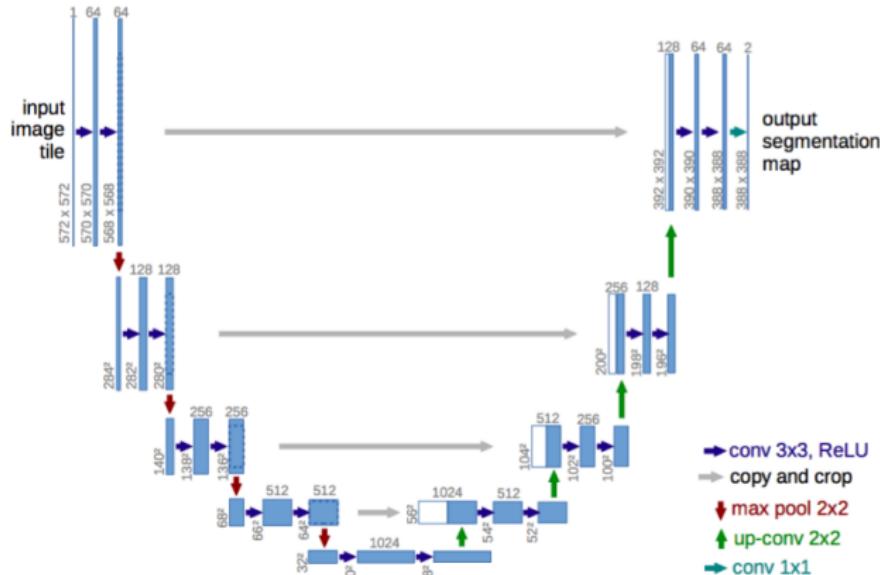


Image de https://github.com/vdumoulin/conv_arithmetic

Un mot sur les fonctions de coût

- Chaque pixel porte une fonction softmax afin de déterminer quelle classe est la plus probable.
- On utilise l'entropie croisée pour comparer la classe prédite d'un pixel et la classe réelle.
- La fonction de coût finale est la moyenne des entropies croisées sur l'ensemble des pixels de l'image.

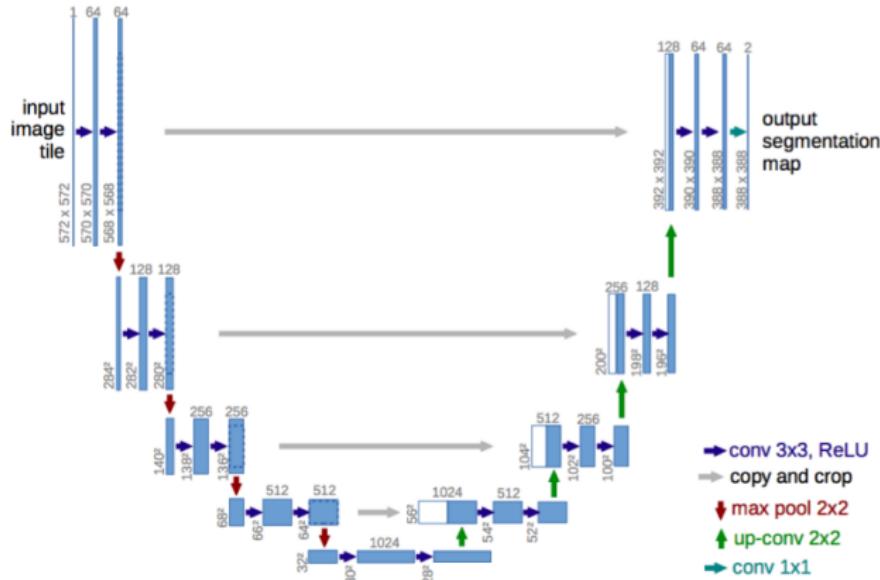
UNet (2015)



- Une évolution de FCN.
- Probablement un des réseaux les plus utilisés pour la segmentation !

[Ronneberger et al.] U-Net : Convolutional Networks for Biomedical Image Segmentation.

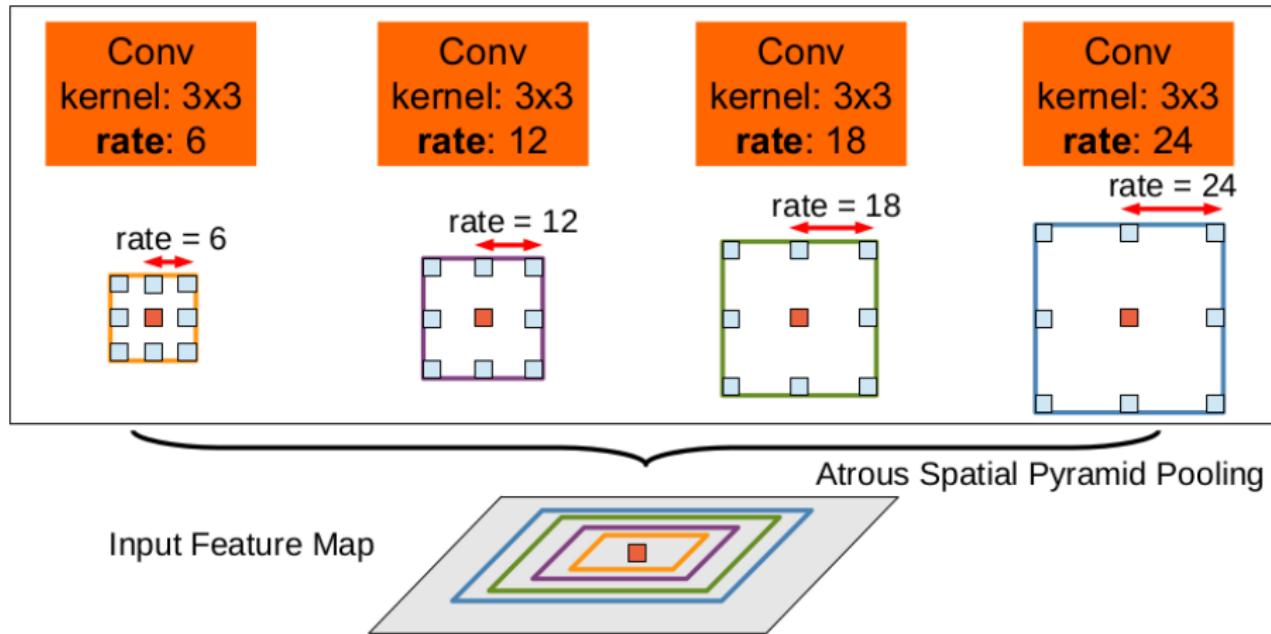
UNet (2015)



Architecture adaptable avec n'importe quelle base convulsive !

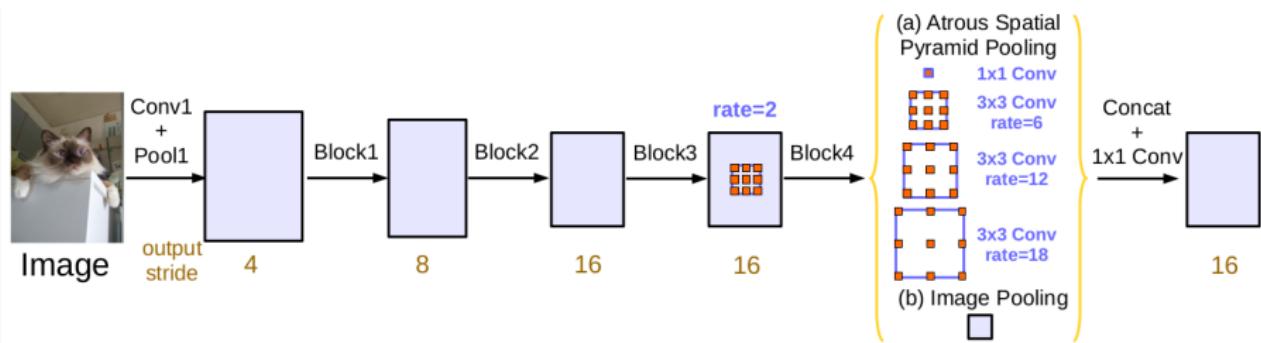
[Ronneberger et al.] U-Net : Convolutional Networks for Biomedical Image Segmentation.

Pyramide spatiale (DeepLab v3)



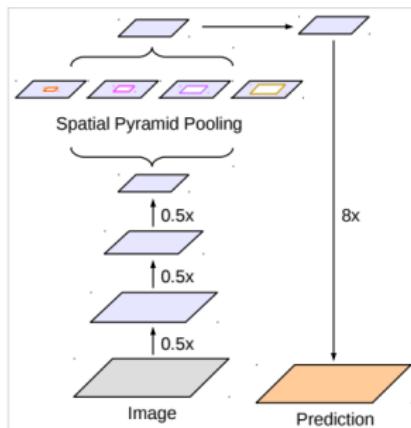
[Chen et al.] Rethinking Atrous Convolution for Semantic Image Segmentation

DeepLab v3

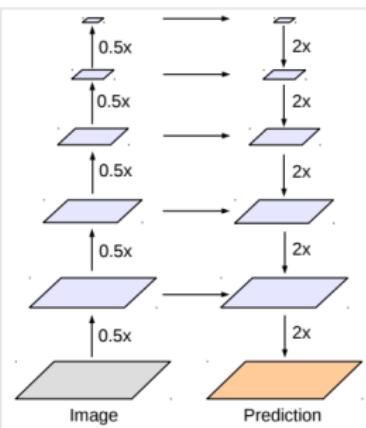


[Chen et al.] Rethinking Atrous Convolution for Semantic Image Segmentation

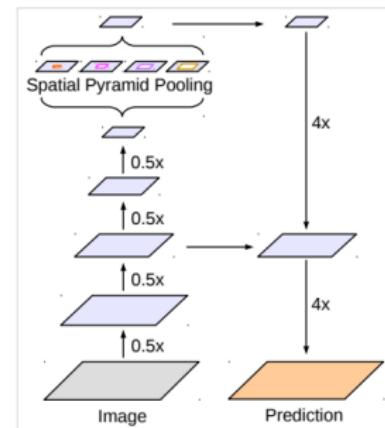
DeepLab v3+



(a) Spatial Pyramid Pooling



(b) Encoder-Decoder



(c) Encoder-Decoder with Atrous Conv

[Chen et al.] Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

Bilan sur la segmentation

Deux enjeux principaux :

- **Densité de la couche de sortie** : augmentation de la dimension par deconvolution ou convolution "à trous".
- **Traitements multi-échelle** : forme en U (*auto-encoder*), pyramide spatiale.

Challenges :



Plan du cours

1 Segmentation d'image

2 Estimation de pose

Keypoints Challenge MSCOCO 2019



200000 images, 250000 personnes, 1,7M de joints.

Estimation de pose

Problème de localisation des "joints", ou articulations, d'une personne sur une image.

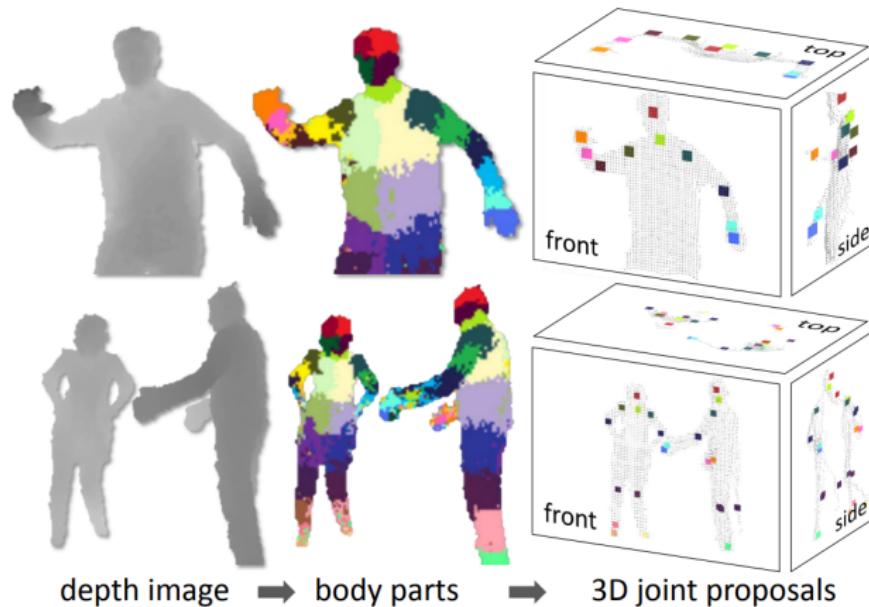


Nombreuses applications : reconnaissance d'actions, animation, jeux (Just Dance), *sports analytics*, etc.

Difficulté : prédictions locales mais nécessité d'une compréhension globale de la scène (comme pour la segmentation)

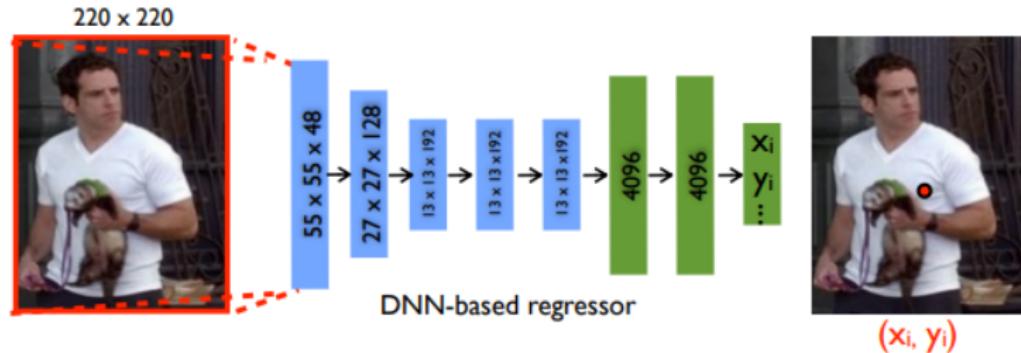
Estimation de pose

Approche "classique" implantée dans la Kinect (2011) :



Segmentation par parties du corps (forêts aléatoires) puis estimation de mode (*mean-shift*) pour l'extraction des joints.

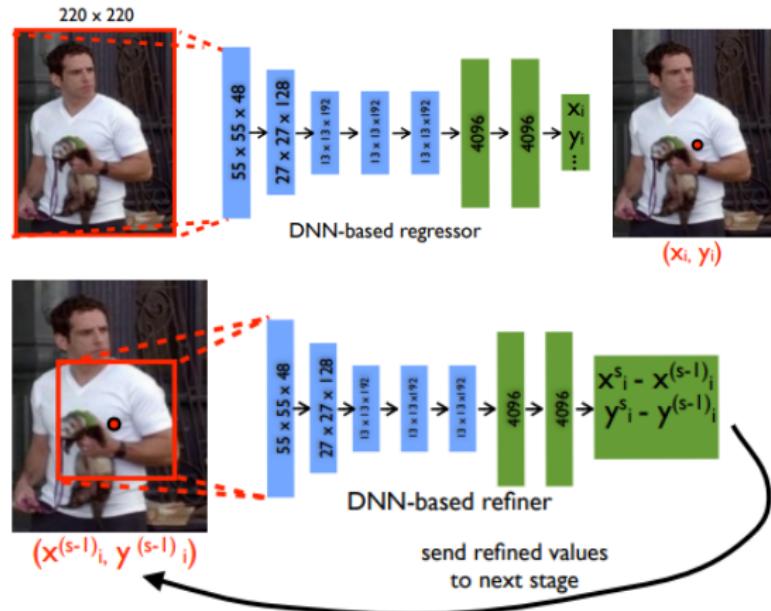
DeepPose (2014)



Utilisation d'AlexNet comme base convolutive, prédition directe des coordonnées des joints. Augmentation **massive** des données ($\times 40$) !

[Toshev et al.] DeepPose : Human Pose Estimation via Deep Neural Networks

DeepPose (2014)

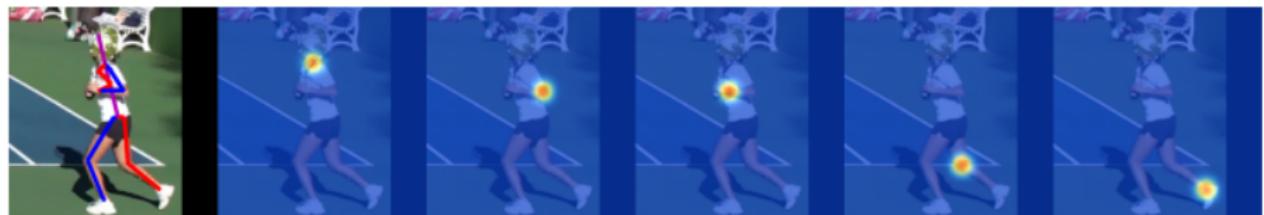


Affinage de la prédiction en repassant une sous-région de l'image dans le réseau de neurones.

[Toshev et al.] DeepPose : Human Pose Estimation via Deep Neural Networks

Prédiction de cartes de chaleur

Une idée introduite en 2015 consiste à prédire des cartes de chaleur pour chaque joint plutôt qu'une position.



Le problème se rapproche alors d'un problème de segmentation.

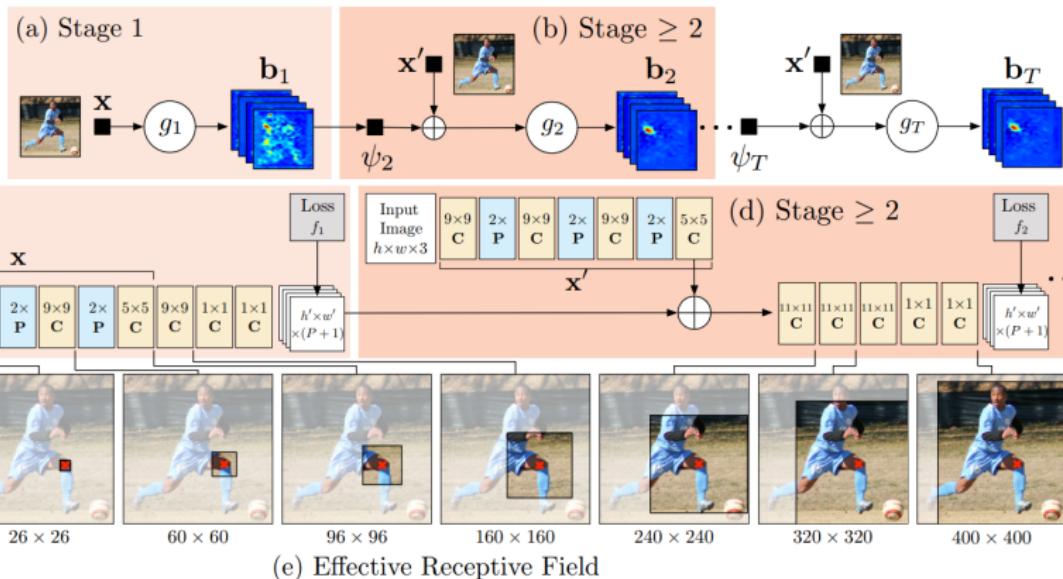
[Tompson et al.] Efficient Object Localization Using Convolutional Networks

OpenPose (2016-2018)

Convolutional
Pose Machines
(T -stage)

P Pooling

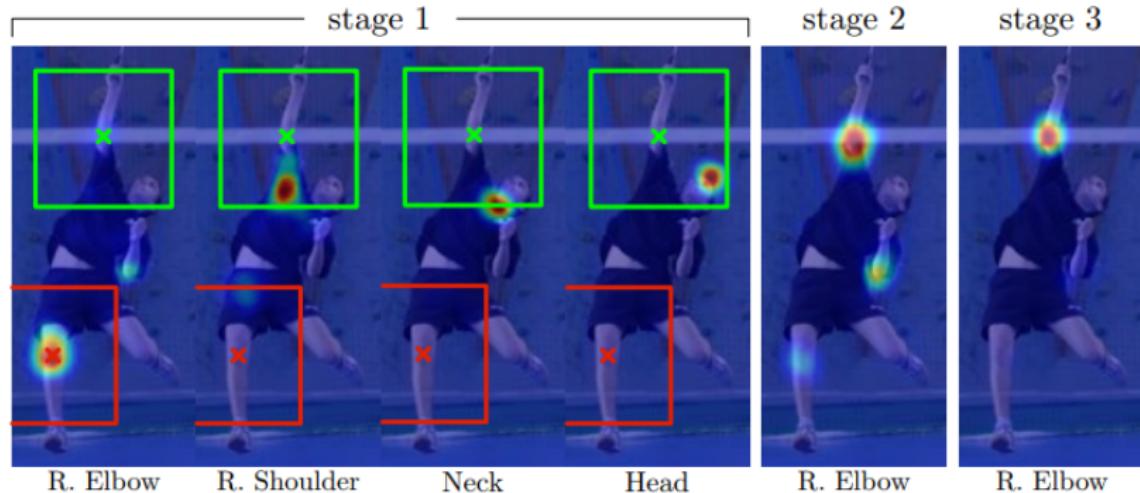
C Convolution



Affinage des prédictions en prenant en compte des premières versions des prédictions.

[Wei et al.] Convolutional Pose Machines

OpenPose (2016-2018)



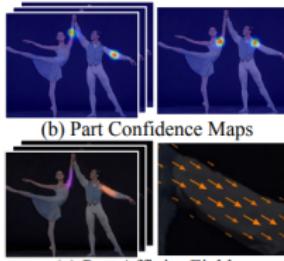
Les joints estimés avec certitude servent d'*a priori* aux joints pour lesquels il existe une ambiguïté.

[Wei et al.] Convolutional Pose Machines

OpenPose (2016-2018)



(a) Input Image



(b) Part Confidence Maps



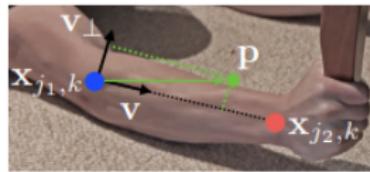
(c) Part Affinity Fields



(d) Bipartite Matching



(e) Parsing Results



Estimation conjointe des probabilités de présence des joints, et d'un champ de vecteur permettant dans une phase de post-traitement de connecter les joints.

[Wei et al.] OpenPose : Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields