Image Localization and Segmentation

Computer Vision

- Computer Vision (CV) is a field of artificial intelligence that trains computers to interpret and understand the visual world. By using digital images from cameras and videos, machines can accurately identify and classify objects — and then react to what they “see.”
- Important tasks of computer vision: object detection, image classification, image localization, and image segmentation.
- The techniques applied for CV are image processing techniques, machine learning, and, more recently, deep learning.

Applications of Computer Vision

- **Some important application of Computer Vision:**
  - Autonomous vehicles: Understanding the surroundings to navigate safely.
  - Medical imaging: Helping in diagnosis by accurately interpreting MRIs, CT scans, etc.
  - Surveillance: Enhancing security through automated monitoring and threat detection.
  - Agriculture: Monitoring crops and predicting yields through aerial imagery.
  - Retail: Improving shopping experiences through gesture recognition and personalized advertising.

Computer vision Tasks

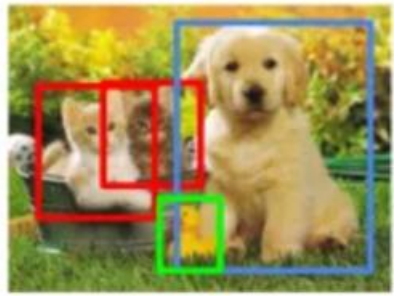
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

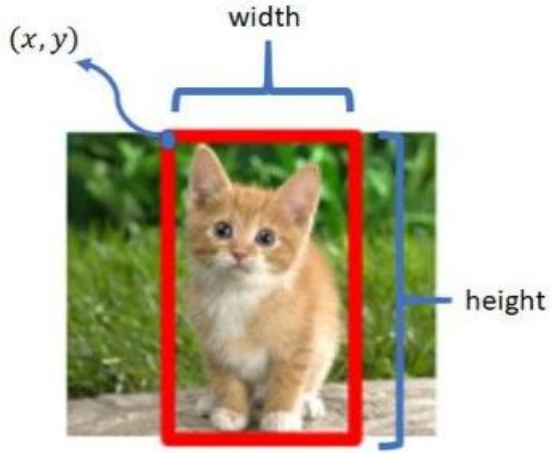
Instance Segmentation

CAT, DOG, DUCK

Semantic Segmentation

GRASS, CAT, TREE, SKY

Image Localization



- Image localization refers to the process of identifying the location of one or more objects within an image.
- Usually represented by bounding boxes around the objects.
- It is the processes of finding *where* an object is, rather than identifying *what* the object is.
- **Object Detection** combines object localization and object classification tasks by not only locating the objects in an image but also identifying them.

Applications of Image Localization

- **Automotive**: Locating and tracking vehicles and pedestrians for autonomous driving systems.
- **Retail**: Detecting products on shelves for inventory management.
- **Sports Analytics**: Identifying and tracking players and the ball in sports broadcasts.
- **Agriculture**: Locating areas of interest in aerial images for precision farming.
- **Medical Imaging**: Identifying specific anatomical regions or abnormalities within medical scans.

Image Localization techniques

- Commonly applied techniques for localization
  - Sliding Window Detection
  - Region Proposals
  - Advancements in deep learning have significantly improved the accuracy and efficiency of image localization tasks.
- Challenges of localization
  - Variability in object sizes, shapes, and appearances.
  - Occlusions and overlapping objects.
  - Complex backgrounds that may hide or camouflage objects.

Image Segmentation

- Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects) to simplify its representation into something that is more meaningful and easier to analyze.
- It aims to assign a label to every pixel in an image such that pixels with the same label share certain characteristics.
- Identifies the regions of interest in an image for various image analysis tasks.
- Facilitates a more precise object recognition and classification by isolating objects from the background or separating them from each other.

Types of Image Segmentation

- **Binary Segmentation:** Divides the image into two parts – the object and the background.
- **Semantic Segmentation:** Each pixel is classified into a predefined category, without differentiating between different objects of the same class.
- **Instance Segmentation:** Similar to semantic segmentation but dif me class.

Binary image | Instance segmentation | Semantic segmentation

Applications of Image Segmentation

- Medical imaging, for example, tumor detection or organ delineation.
- Scene understanding in autonomous driving systems.
- Object tracking in video surveillance.
- Crop disease detection in precision agriculture.

Segmentation Techniques

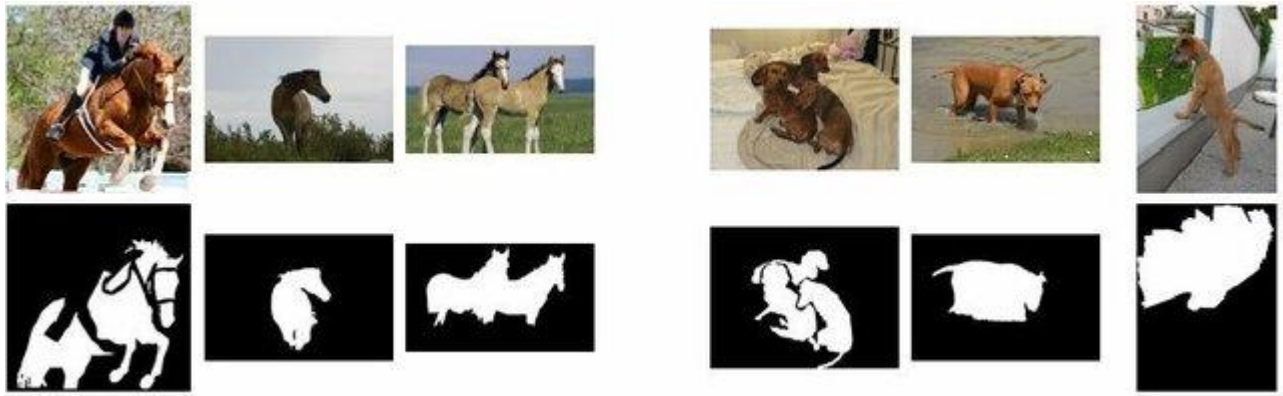
- Thresholding : Pixels are classified into foreground and background by comparing their intensity to a chosen threshold value.
- Region-Based Segmentation : Starts with a seed point and adds neighboring pixels to the region if they have similar properties

- Clustering methods (K-means, etc.) : Partitions the image into K clusters where each pixel is assigned to the cluster with the nearest mean intensity or color.

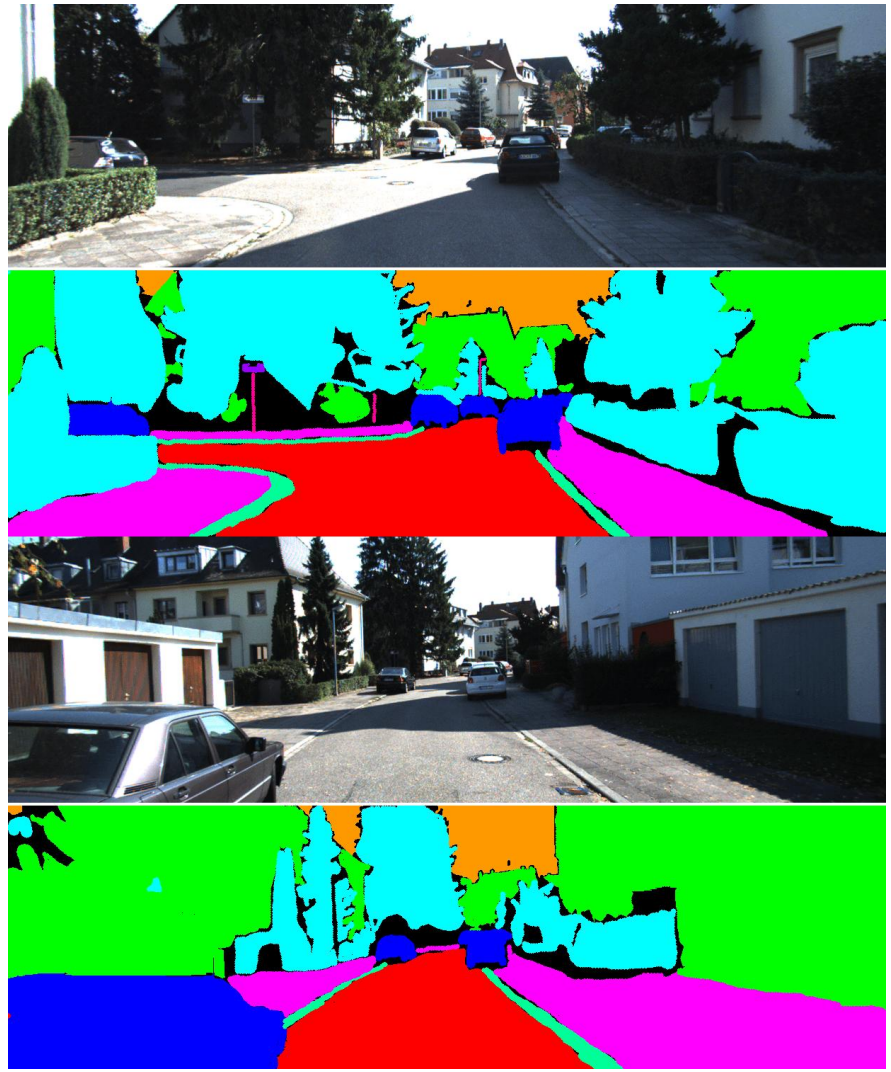
- Deep Learning approaches (U-Net, Vnet, Unet++).

Segmentation Masks

- A segmentation mask is a binary or multi-class image that indicates the category of each pixel in the original image. Essentially, it's a map where each pixel's value corresponds to the class of the object it belongs to.
- In binary masks, pixels are marked as either foreground (object) or background. In multi-class masks, different values represent different classes.
- Segmentation masks enable precise analysis of images by clearly delineating the boundaries of objects at the pixel level.



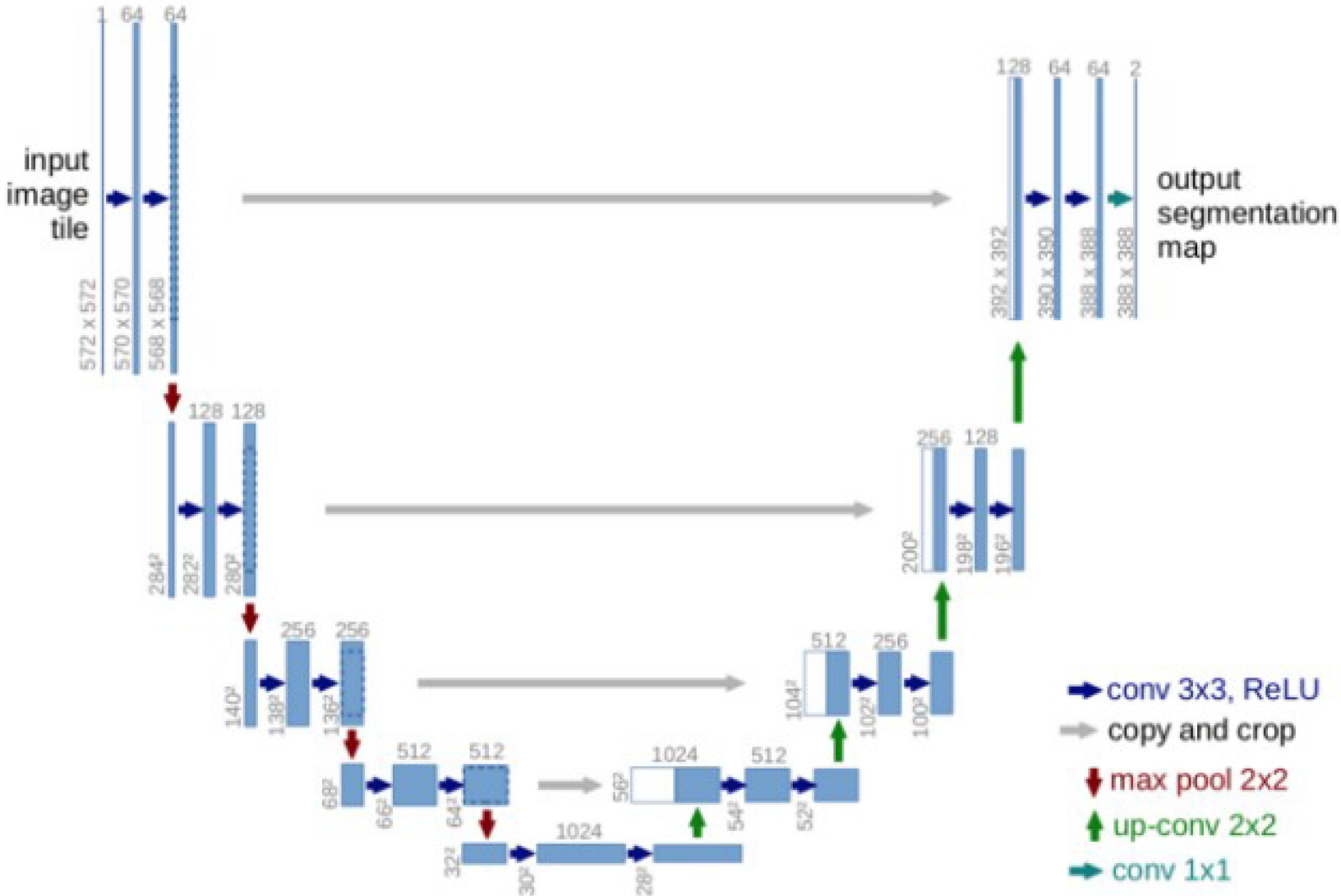
Binary Segmentation Mask



Multi class segmentation mask

UNet

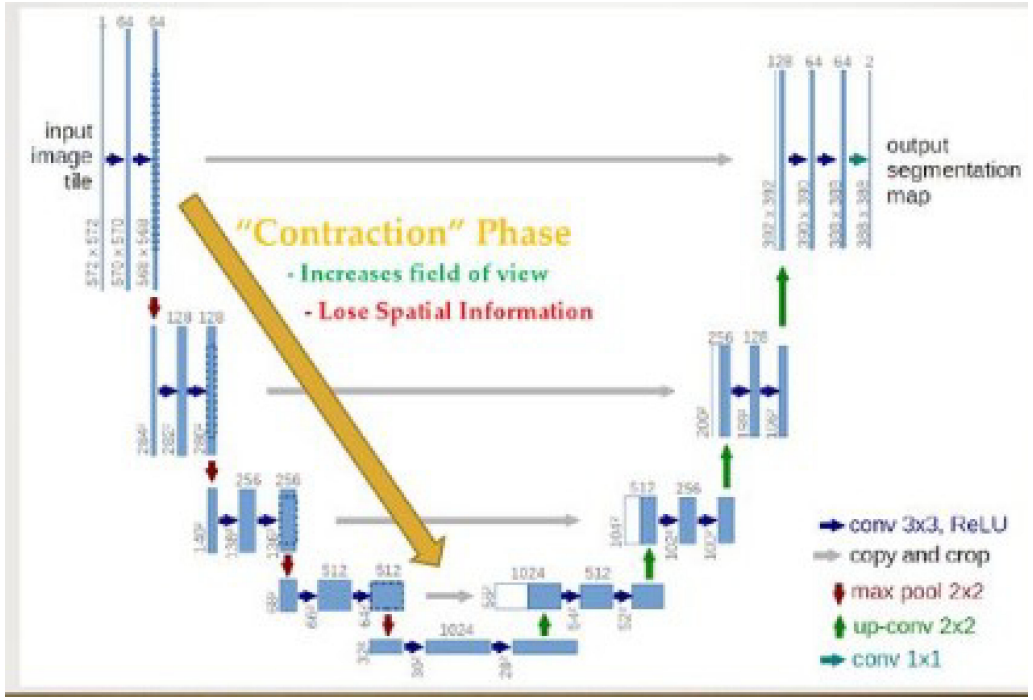
- The U-Net is convolutional network architecture for fast and precise segmentation of images
- First designed by Ronneberger and applied in 2015 to process biomedical images.
- It not only distinguish whether there is a disease, but also localize the area of abnormality.
- The U-Net owes its name to its symmetric shape, which is different from other FCN variants.
- U-Net architecture is separated in 3 parts:
  - The contracting/downsampling path (Encoder)
  - Bottleneck
  - The expanding/upsampling path (Decoder)



UNet architecture

Encoding path

- The contracting (Encoding) path is composed of 4 blocks.
- Consists of repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling.
- At each downsampling step, the number of feature channels is doubled to capture more complex features.

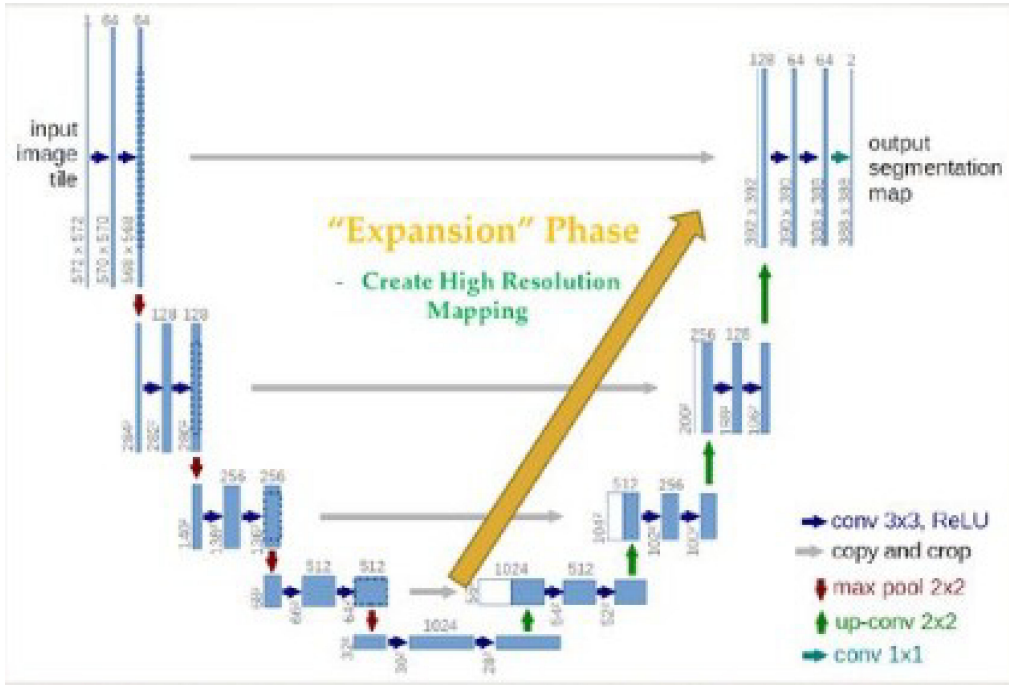


Bottleneck

- This is the part of U-Net where the transition between the downsampling and upsampling paths occurs.
- Typically involves two 3x3 convolutions, with ReLU activations. This section is crucial for processing the most abstracted form of the input image.

Decoding path

- The expanding (Decoding) path is also composed of 4 blocks.
- Consists of an upsampling of the feature map followed by a 2x2 convolution (“up-convolution”) that halves the number of feature channels.
- A concatenation with the correspondingly cropped feature map from the contracting path, followed by two 3x3 convolutions, each followed by a ReLU.
- This path increases the resolution of the output, allowing for precise



Skip Connections

- Feature maps from the contracting path are concatenated with the upsampled output at each level of the expansive path to ensure that fine-grained details are not lost during upsampling.
- These connections provide the local context to the global information, enabling precise localization.

- Output Layer:
    - A 1x1 convolution is used at the end of the network to map each feature vector to the desired number of classes.

Detailed explanation of each layer

I /P= 572 x 572 x 1

EC11 = 570 x 570 x 64

EC12= 568 x 568 x 64

EP1=284 x 284 x 128

EC21 = 282 x 282 x 128

EC22= 280 x 280 x 128

EP1=140 x 140 x 128

EC31 = 138 x 138 x 256

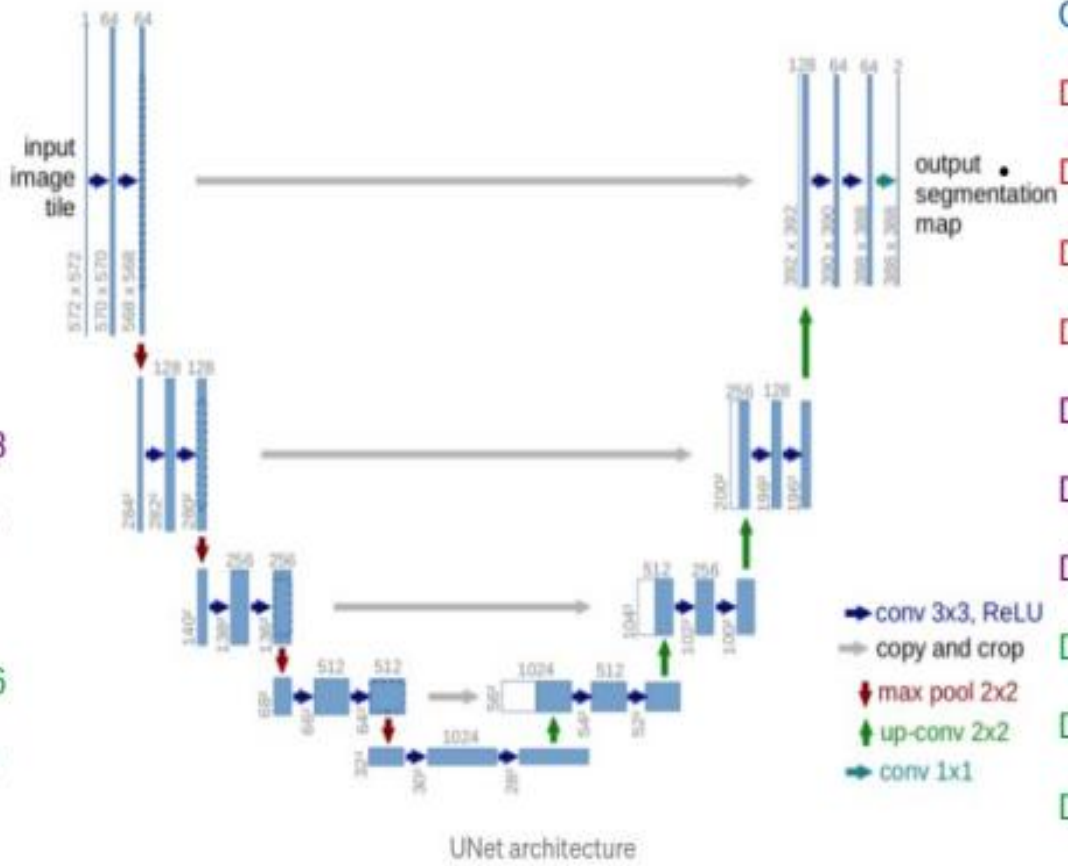
EC32= 136 x 136 x 256

EP3=68 x 68 x 256

EC41 = 66 x 66 x 512

EC42= 64 x 64 x 512

EP4= 32 x 32 x 512



UNet architecture

C51 = 30 x 30 x 1024

C52= 28 x 28 x 1024

O/P = 388 x 388 x 2

DC13= 388 x 388 x 2

DC12= 388 x 388 x 64

DC11= 390 x 390 x 64

DT1=392 x 392 x 64

DC21 = 196 x 196 x 128

DC21= 198 x 198 x 128

DT2=200x 200 x 128

DC32 = 100 x 100 x 256

DC31= 102 x 102 x 256

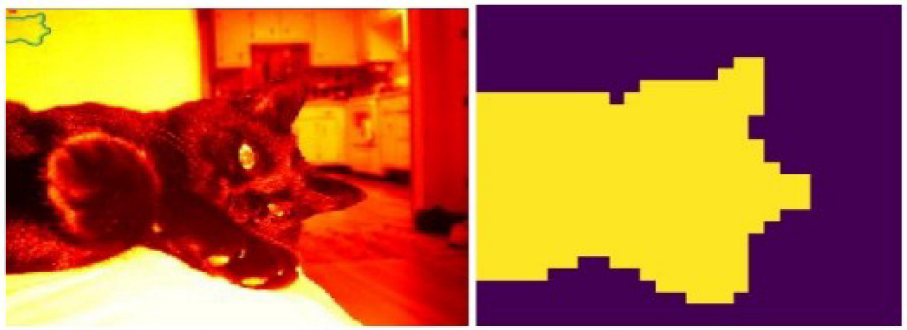
DT3=104 x 104 x 256

DC42 = 52x 52 x 512

DC41= 54 x 54 x 512

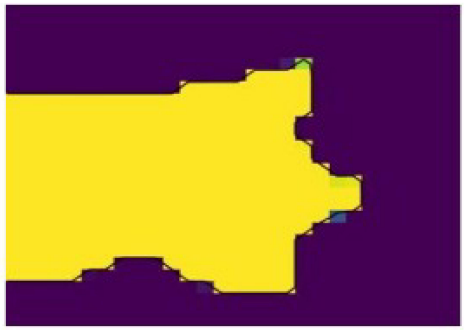
DT4=56 x 56 x 512

Segmentation Exampl

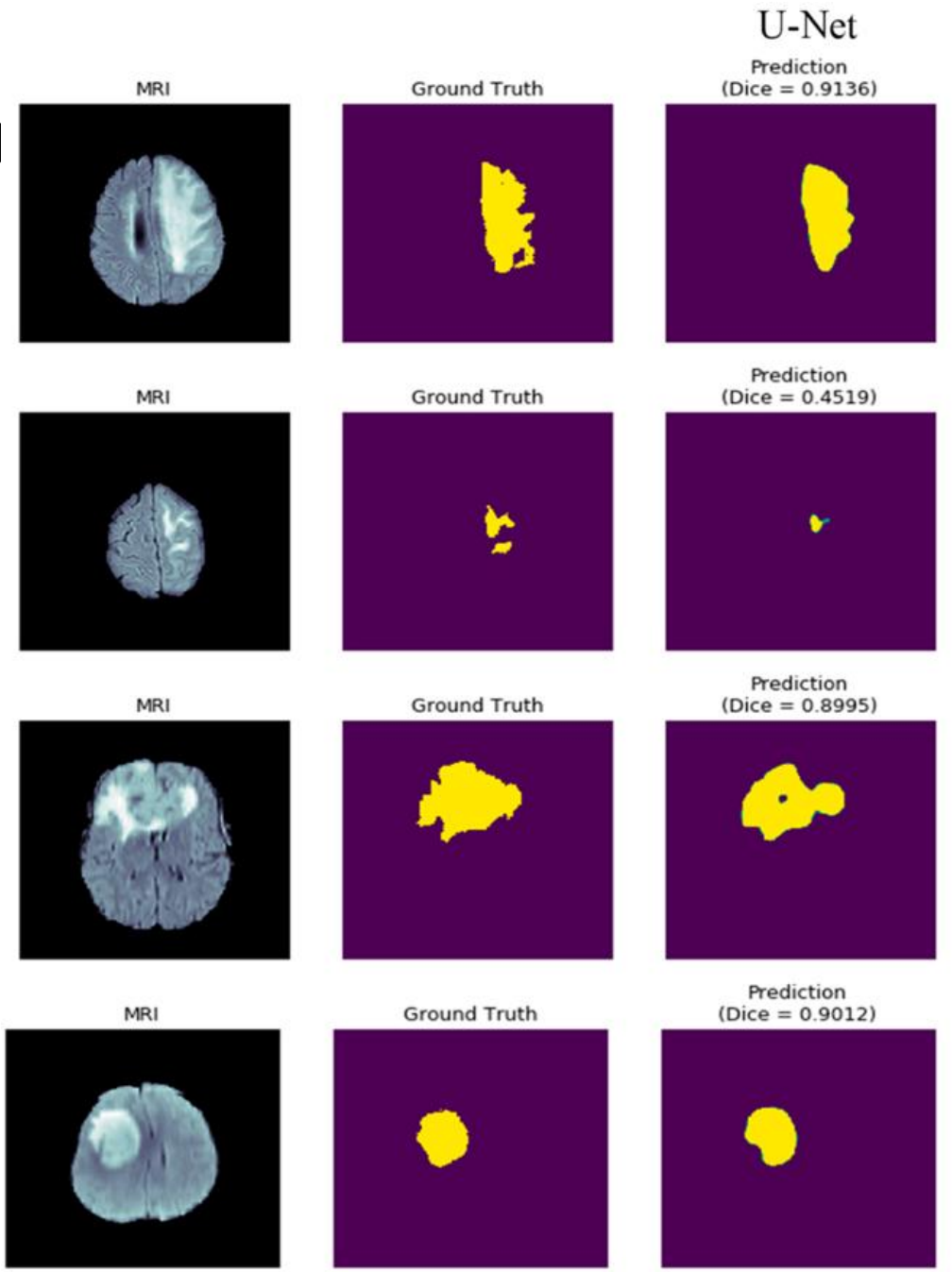


Cat

Ground Truth



Predicted Mask



Advantages of U-Net

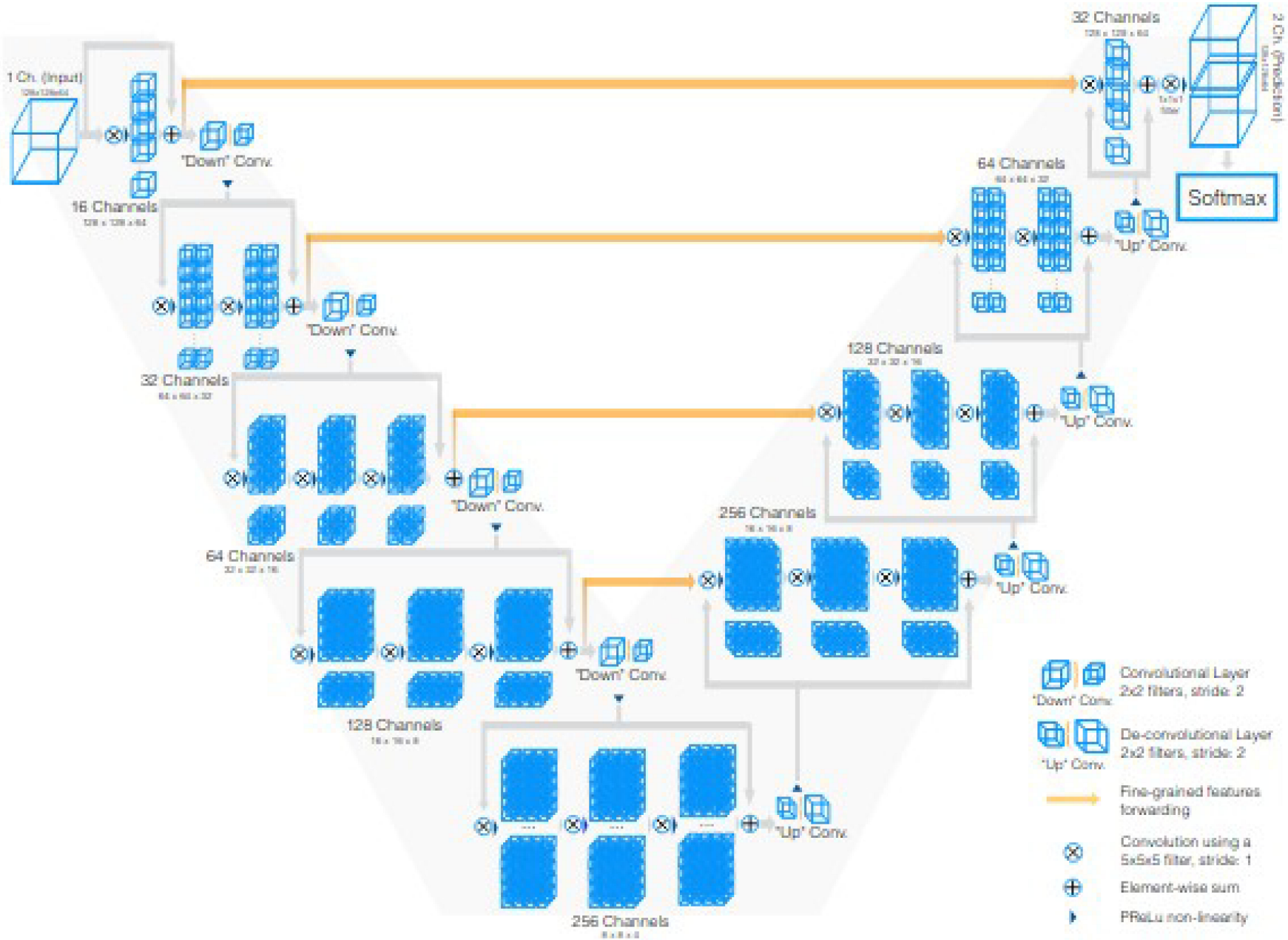
- Ability to learn from a small number of images.
- High accuracy and efficiency in segmenting images, making it especially useful in medical image analysis where fine details are crucial.
- Flexibility to be adapted for various segmentation tasks beyond biomedical imaging.

Vnet Architecture

- V-Net is a convolutional neural network (CNN) architecture tailored for 3D image segmentation, build on the principles of the U-Net architecture.
- It's primarily used for volumetric segmentation tasks, such as segmenting organs or tumors in 3D medical scans (CT, MRI).
- Introduced residual connections within each convolutional block, improving gradient flow and enabling deeper networks.

Architecture details

- **Contracting Path**: Similar to U-Net, it consists of successive blocks of 3D convolutions, followed by ReLU activation and 2x2x2 max pooling for downsampling. The number of feature channels is doubled after each block to capture more complex features.
- **Bottleneck**: The central part of the network where the highest level of feature abstraction occurs, followed by the first upsampling step.
- **Expansive Path**: Upsampling blocks that gradually increase the resolution of the volumetric feature maps. Includes skip connections from the contracting path to incorporate fine-grained details.
- **Final Segmentation Layer**: Uses a soft voxel-wise classification followed by a Dice coefficient loss function to accurately segment the target structures.


1 Ch. (Input)
16 Channels
32 Channels
64 Channels
128 Channels
256 Channels
"Down" Conv.
"Up" Conv.
Softmax
Convolutional Layer
2x2 filters, stride: 2
De-convolutional Layer
2x2 filters, stride: 2
Fine-grained features forwarding
Convolution using a 5x5x5 filter, stride: 1
Element-wise sum
PReLu non-linearity

Pros and Cons of Vnet

- Pros
  - **3D Contextual Information**: V-Net takes advantage of the full 3D context, leading to more accurate and coherent segmentation across volumetric data.
  - **High Precision**: V-Net demonstrates high precision in segmenting various structures, even in challenging conditions like low contrast or noise.
  - **Dice Loss Function**: The use of the Dice coefficient as a loss function directly optimizes the segmentation quality, making it effective for dealing with imbalanced classes.

- Cons
  - **Computational Intensity**: Processing 3D volumes requires significant computational resources in terms of memory and processing power
  - **Data Requirements**: Training deep 3D networks like V-Net typically requires large volumes of annotated volumetric data

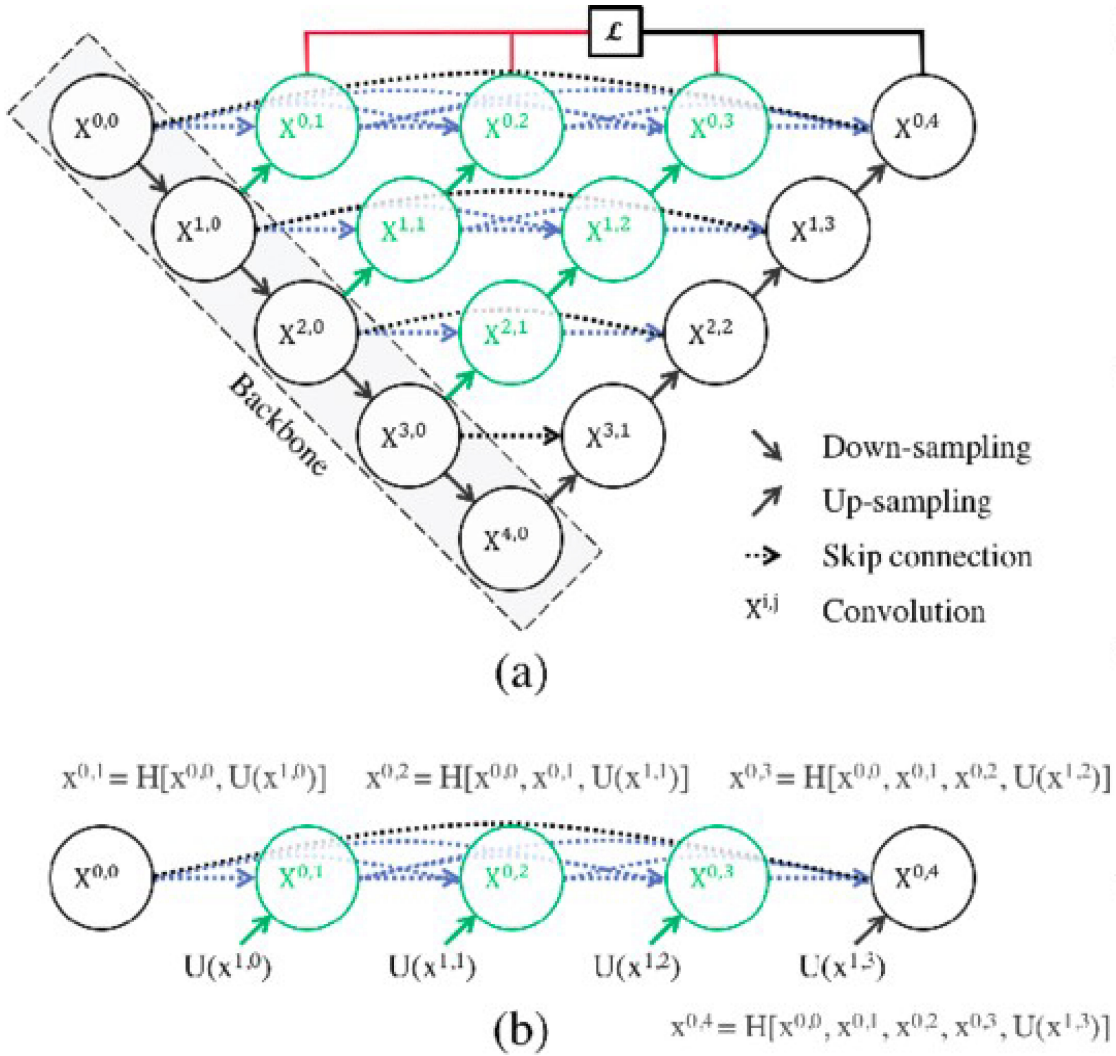
Unet++

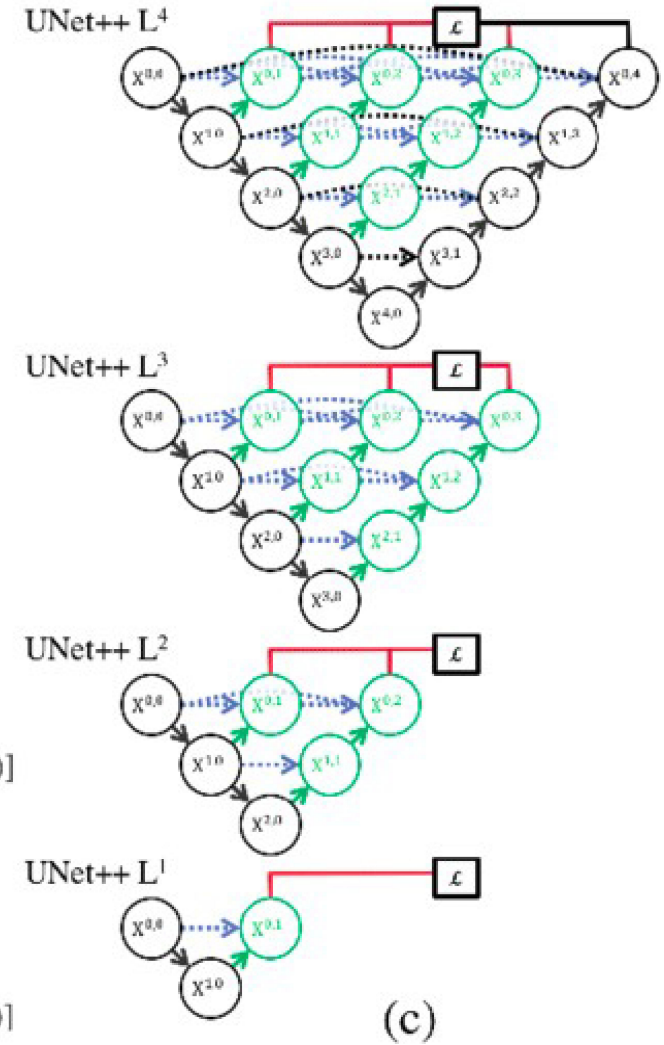
- U-Net++ is an optimized version of the U-Net architecture, developed to address specific challenges in medical image segmentation, though its applications extend beyond.
- Its main goal is to bridge the semantic gap between the encoder and decoder pathways, which is achieved through a series of nested, dense skip pathways.

Innovations in Unet++

- **Nested Skip Connections**: U-Net++ integrates multiple levels of skip connections, creating a dense network of paths between the encoder and decoder blocks. This structure enables the fusion of features at different resolutions more effectively.
- **Redesigned Skip Pathways**: Introduces intermediate convolutional layers within the skip pathways, allowing for the refinement of features before they are concatenated with upsampled features. This process facilitates more precise segmentation.
- **Deep Supervision**: Employs deep supervision by adding auxiliary segmentation branches at different depths of the network. This encourages the learning of more discriminative features at all levels, improving the segmentation performance, especially for small and intricate structures.

Unet++ Architecture

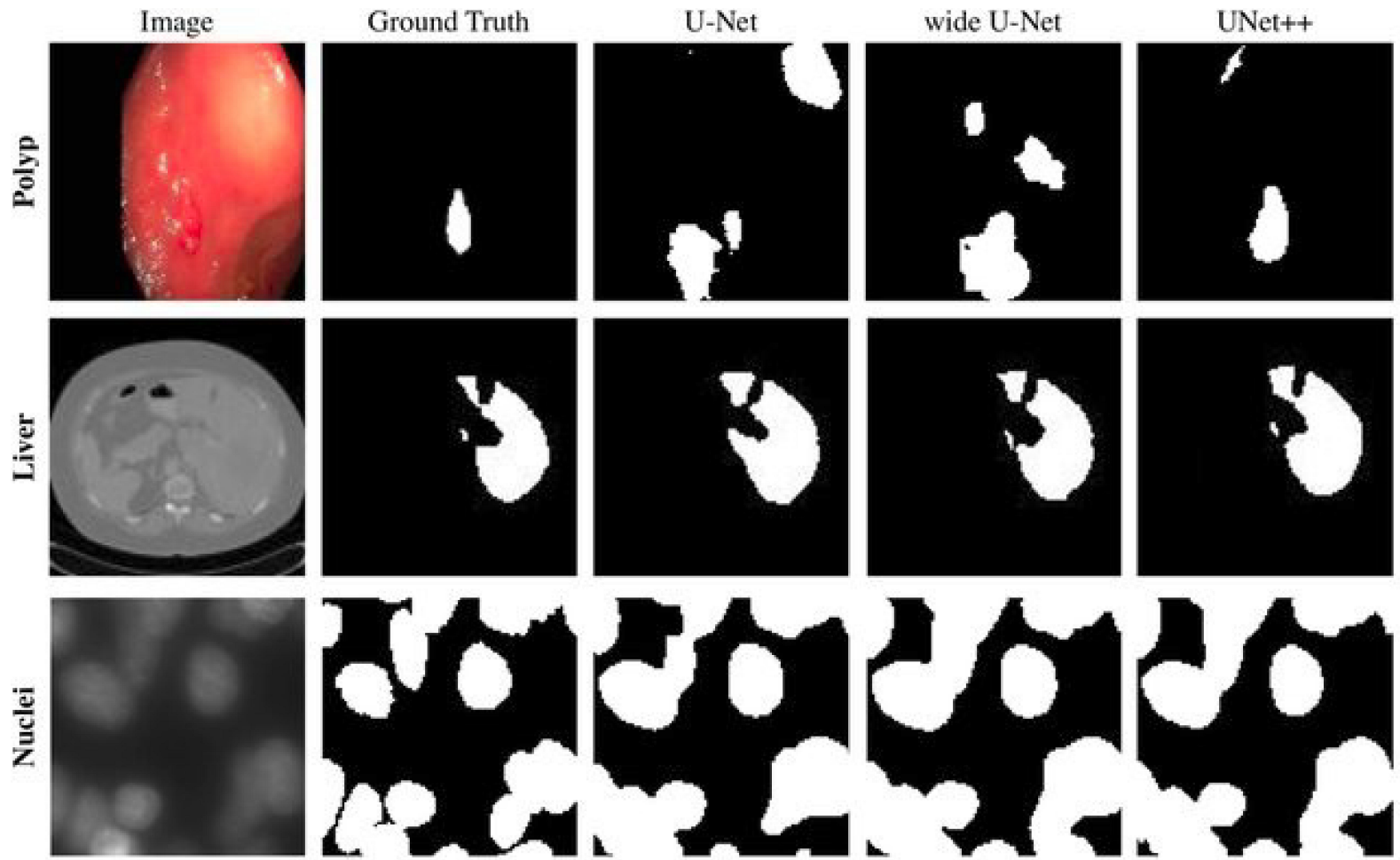




Pros and Cons of Unet++

- Pros
  - **Enhanced Feature Fusion**: The nested skip connections in U-Net++ improve feature fusion between the encoder and decoder paths, reducing the semantic gap and leading to more accurate segmentation, especially for complex and small structures.
  - **Deep Supervision**: by adding auxiliary outputs in intermediate layers. This encourages the network to learn useful features at multiple scales.
  - **Higher Accuracy**: Achieved higher segmentation accuracy compared to the original U-Net and other segmentation architectures
- Cons
  - **Increased Computational Complexity**: The nested skip pathways and additional convolutional layers in U-Net++ significantly increase the network's parameters and computational complexity.
  - **Training Time and Memory Requirement**: Complex architecture leads to longer training times and higher computational resource requirements.
  - **Model Size and Inference Time**: U-Net++ results in larger model sizes due to its increased number of parameters. This can affect inference time, making real-time applications more challenging.

Segmentation examples



References

- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234-241. Springer International Publishing, 2015.
- Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In *2016 fourth international conference on 3D vision (3DV)*, pp. 565-571. Ieee, 2016.
- Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A nested u-net architecture for medical image segmentation." In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3-11. Springer International Publishing, 2018.

Thank you