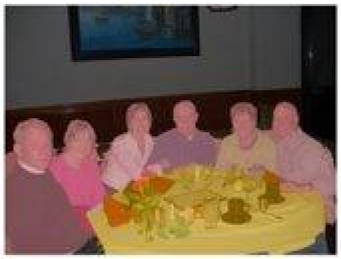
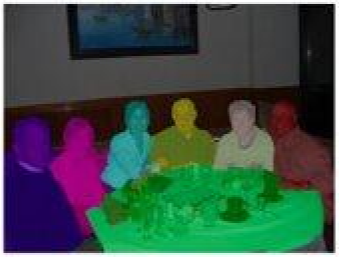
What is Object Segmentation ?

- Image segmentation is the process of classifying each pixel in the image as belonging to a specific category.
  - Semantic segmentation : We treats multiple objects within a single category as one entity
  - Instance segmentation : We identify individual objects within these categories



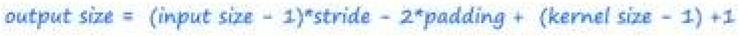
Semantic Segmentation



Instance Segmentation

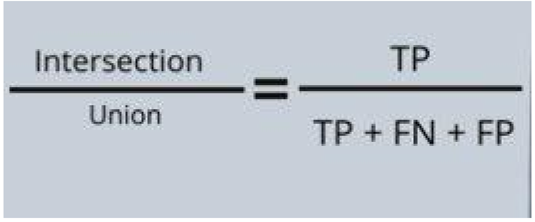
Transposed Convolution

- Transposed Convolutions is a method to up-sample the output. It can be considered as an opposite process to any simple CNN.
  - Output Shape:

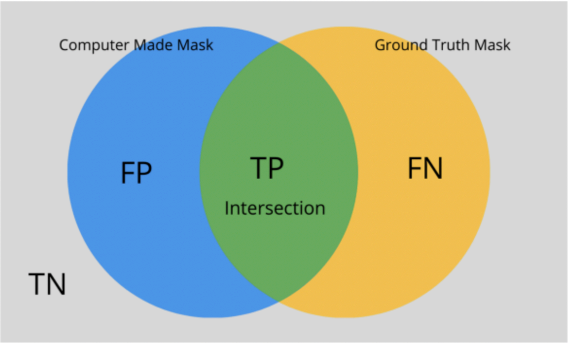


$$output\ size = (input\ size - 1)*stride - 2*padding + (kernel\ size - 1) + 1$$

Intersection Over Union (IoU)/ Dice Coefficient Metrics



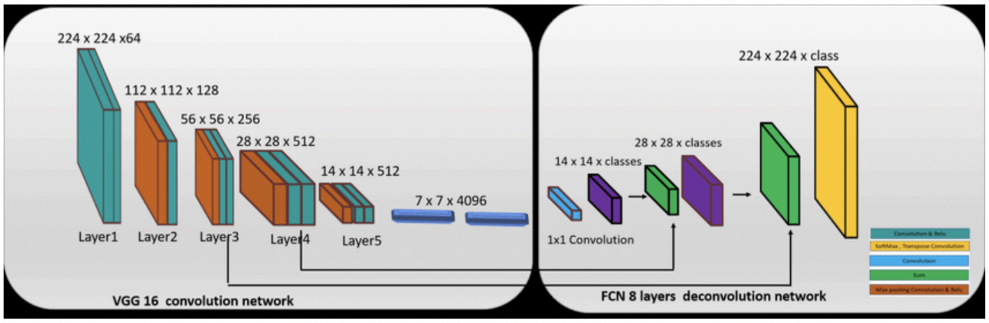
$$\frac{Intersection}{Union} = \frac{TP}{TP + FN + FP}$$



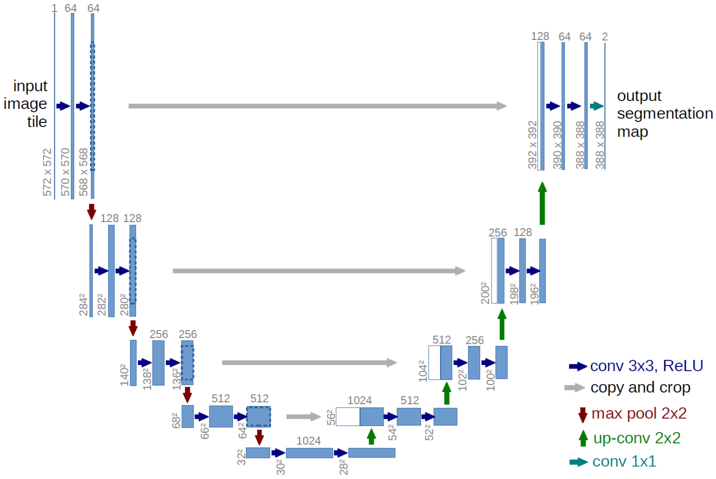
$$\text{Dice Coefficient} = \frac{2\text{xIntersection}}{\text{Union + Intersection}} = \frac{2TP}{2TP + FN + FP}$$

Different types of Encoder-Decoder Network

- **FCN:** The FCN model uses convolutional layers as feature extractor
  - When used in the Encoder part of the FCN model by downsampling the image
  - Now, the last layers contains all the key features we pass it to a decoder part of the model instead of Fully Connected Layers
  - The decoder consists of Deconvolutional layers that upsamples these key features to the original size of the image
  - to further retain any information loss caused due to upsampling, Fusing output/Skip connection are used
    - Deep features can be obtained when going deeper, spatial location information is also lost when going deeper.
    - That means output from shallower layers have more location information. If we combine both, we can enhance the result.

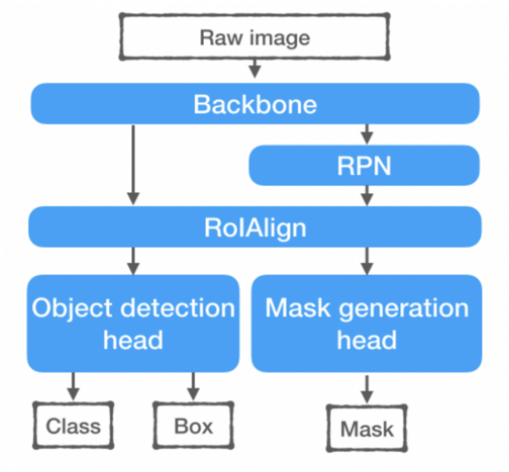


- **U-Net:** U-net Model has a “U” shape architecture with a symmetric Encoder and Decoder.
    - It uses Skip Connections between layers of Encoder and Decoder are used to make the information loss as minimal as possible.
    - The final output layer produces a per-pixel prediction of the target mask or segmentation.
    - UpSampling2D is a simple scaling up of images by using nearest neighbor or bilinear upsampling.



Mask R-CNN

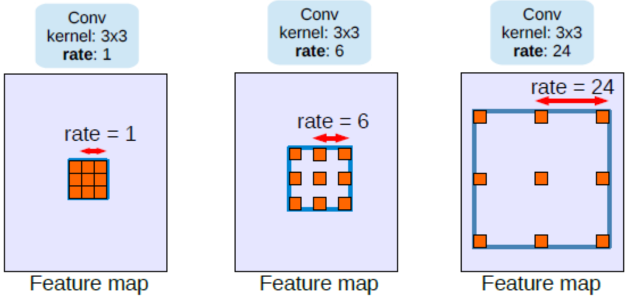
- Mask R-CNN is an extension of the Faster R-CNN model for object detection and instance segmentation.
- The RPN generates object proposals by predicting object scores and bounding box coordinates.
- The RoIAlign layer resamples object proposal features to a fixed size to ensure they can be processed by the fully connected layers.
- The FC layers predict the class labels and object masks using the features from ROIAlign layer



Atrous Convolution

- For each location i on the output y and a filter w, atrous convolution is applied over the input feature map x where the atrous rate r corresponds to the stride with which we sample the input signal.

- It is also called dilated convolution.
- Useful as it maintains the Field-of-View (FOV) at each layer of the network



Atrous convolution with different rates

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

Atrous convolution formula