Encoder The U-Net-encoder has a structure that is typical for CNNs and is made up of five layers l with $l \in [1, 2, 3, 4, 5]$. The input image serves as input to the first layer. Layers l implement two 3×3 unpadded convolutions ($\text{Conv}_{3\times3}(\cdot)$), each followed by a Rectified Linear Unit ($\text{ReLU}(\cdot)$) activation function. The output E_l of layer l is downsampled with a 2×2 max pooling operation ($\text{MaxPool}_{2\times2}(\cdot)$) and serves as input to the next layer. The encoder is formulated as follows:

$$E_1 = \text{ReLU}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(Input)))) \tag{2.1}$$

$$E'_l = \text{MaxPool}_{2\times2}(E_l) \tag{2.2}$$

$$E_l = \text{ReLU}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(E'_{l-1})))) \tag{2.3}$$

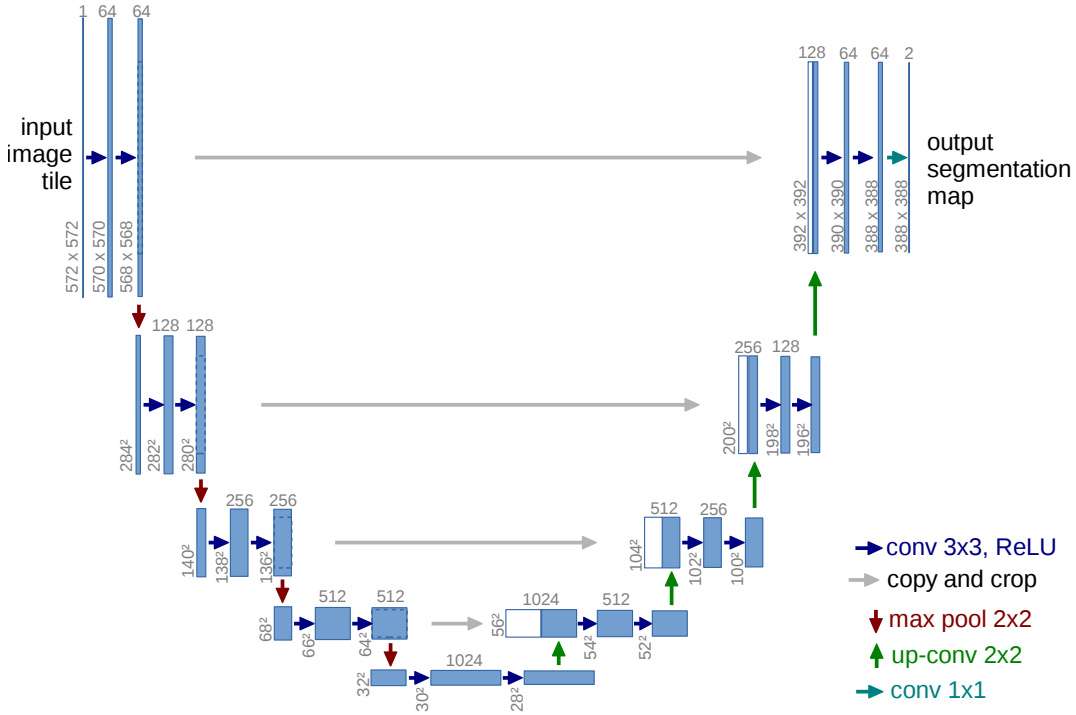

Figure 2.1.: UNet architecture taken from [3]

U-Net [3] was introduced in 2015 and is a extension of FCN [1]. The U-Net architecture implements a nearly symmetrical encoder (also described as the contracting part of the network) and decoder (also described as the expanding part of the network). The decoder in this network combines features extracted at 5 different encoder stages, all with varying resolution and depth. These features get concatenated and up-sampled with the use of convolutions step by step with a

$$D_5 = \text{ReLU}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(E_5)))) \tag{2.4}$$