U-Net: Convolutional Networks for Biomedical Image Segmentation

Ronneberger, et al, 2015

Fateme Sadat Haghpanah
Vasudev Sharma
ML in Healthcare, Fall 2021

Computer Science
UNIVERSITY OF TORONTO

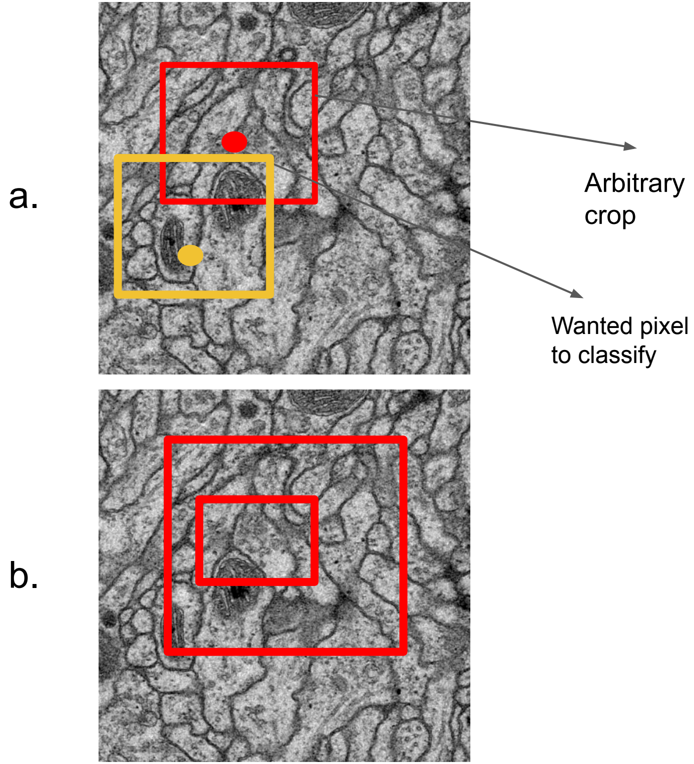
Outline

- Previous works lead to U-Net
  - Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images, Ciresan et al, 2012
  - Fully Convolutional Networks for Semantic Segmentation, Long et al, 2014
- U-Net Architecture
- Training Strategies
- Results
- U-Net Variations
- Summary and Limitations

Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images, Ciresan et al, 2012

Sliding Window Setup

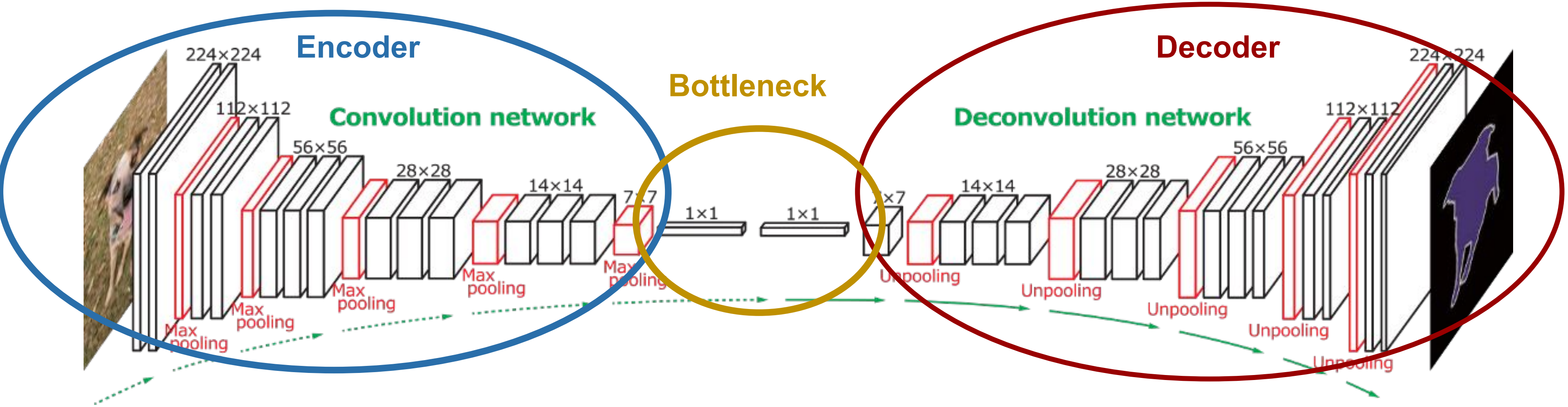
- Advantage:
  - Improve localization
  - Increase number of data training
- Drawbacks:
  - slow to run
  - Redundancy due to overlap (a)
  - Tradeoff on localization and use of context (b)



Sample slice of a dataset, Figure 5, Ciresan et al, 2012

Fully Convolutional Networks for Semantic Segmentation, Long et al, 2014

- Architecture:
  - Capable of being trained on arbitrary size of input (no fully connected layer in network)
  - Consists of Use upsampling / transposed convolution
  - Skip connection



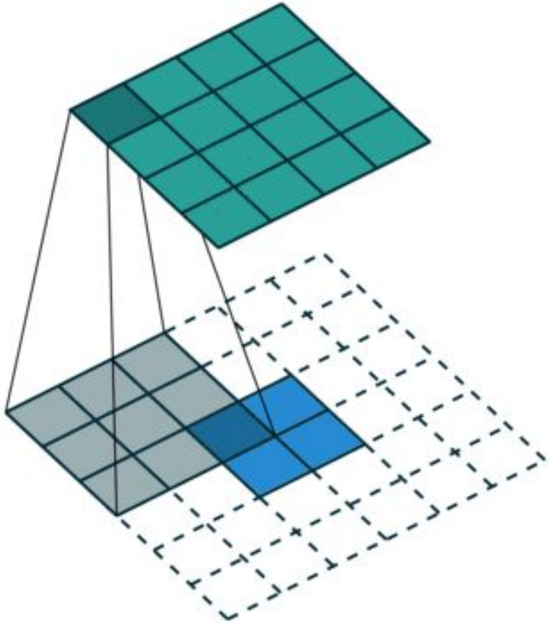
https://medium.com/@wilburdes/semantic-segmentation-using-fully-convolutional-neural-networks-86e45336f99b

Transposed Convolution (Deconvolution/ Unpooling)

Convolution (3*3 kernel)

https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1

Transposed Conv (3*3 kernel)



https://datascience.stackexchange.com/questions/6107/what-are-deconvolutional-layers

Skip Connections

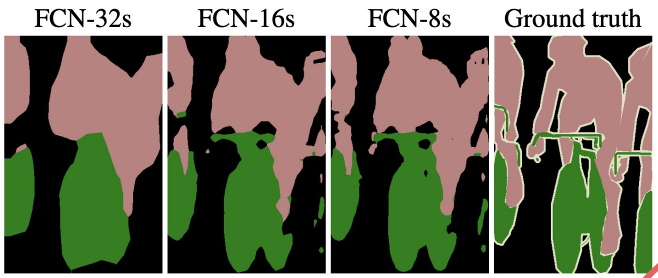


Figure 4, Long et al, 2014

To go up from the bottleneck layer and construct the segmentation labels

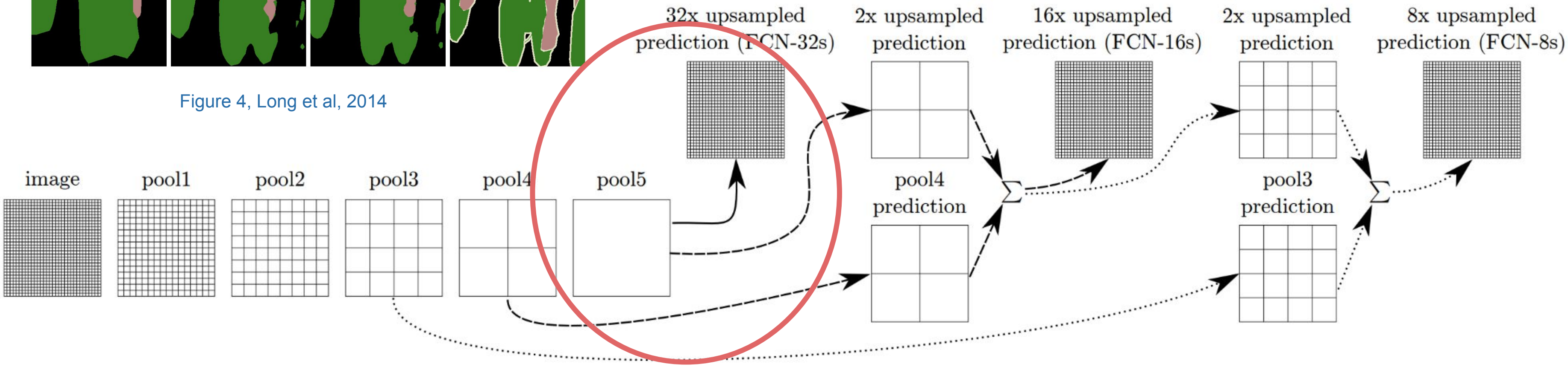


Figure 3, Long et al, 2014

Left half: Encoder
Right half: Decoder

Number of feature channels per layer

Input size of each layer

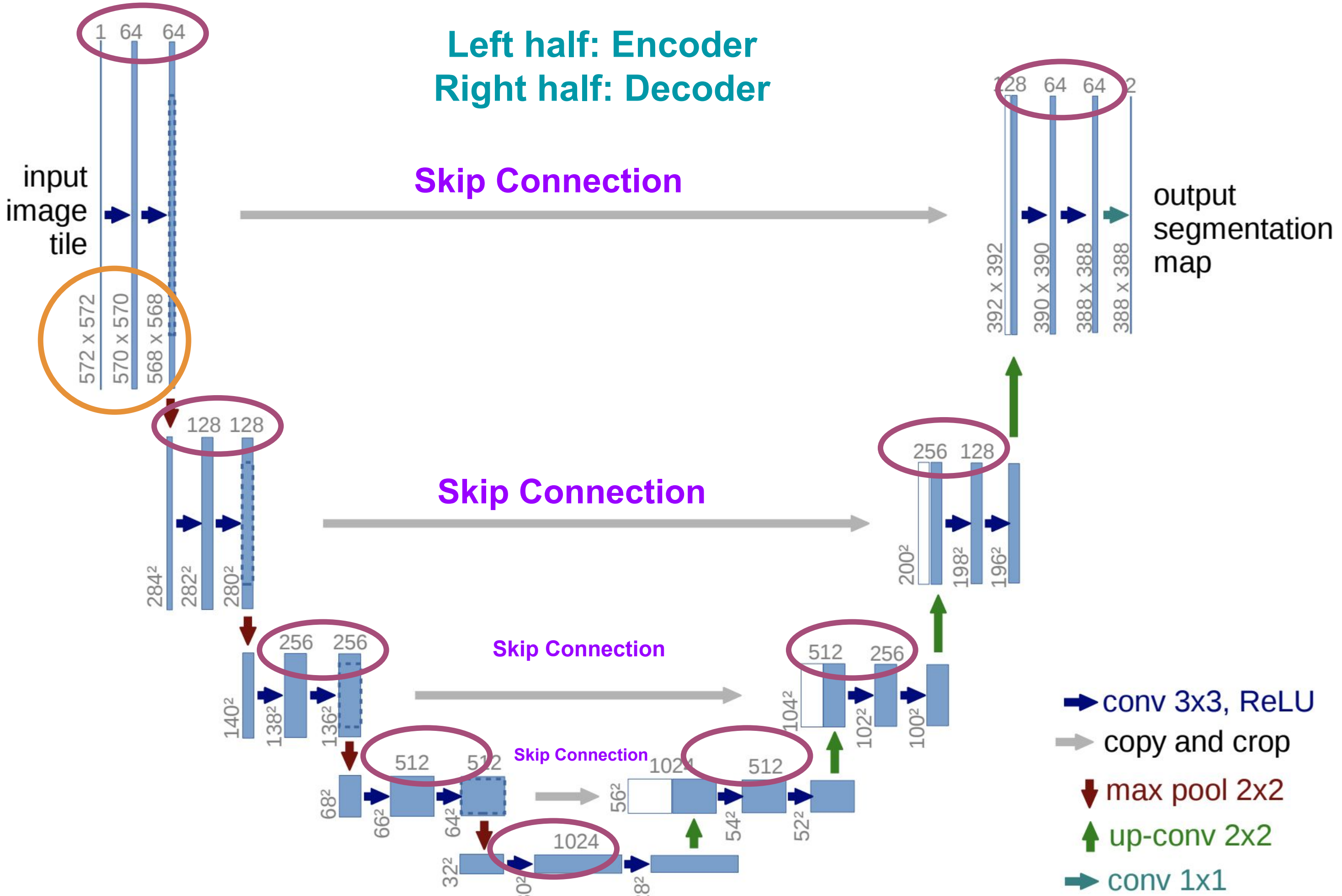


Figure 2, Ronneberger, et al, 2015

Figure 1, Ronneberger et al, 2015

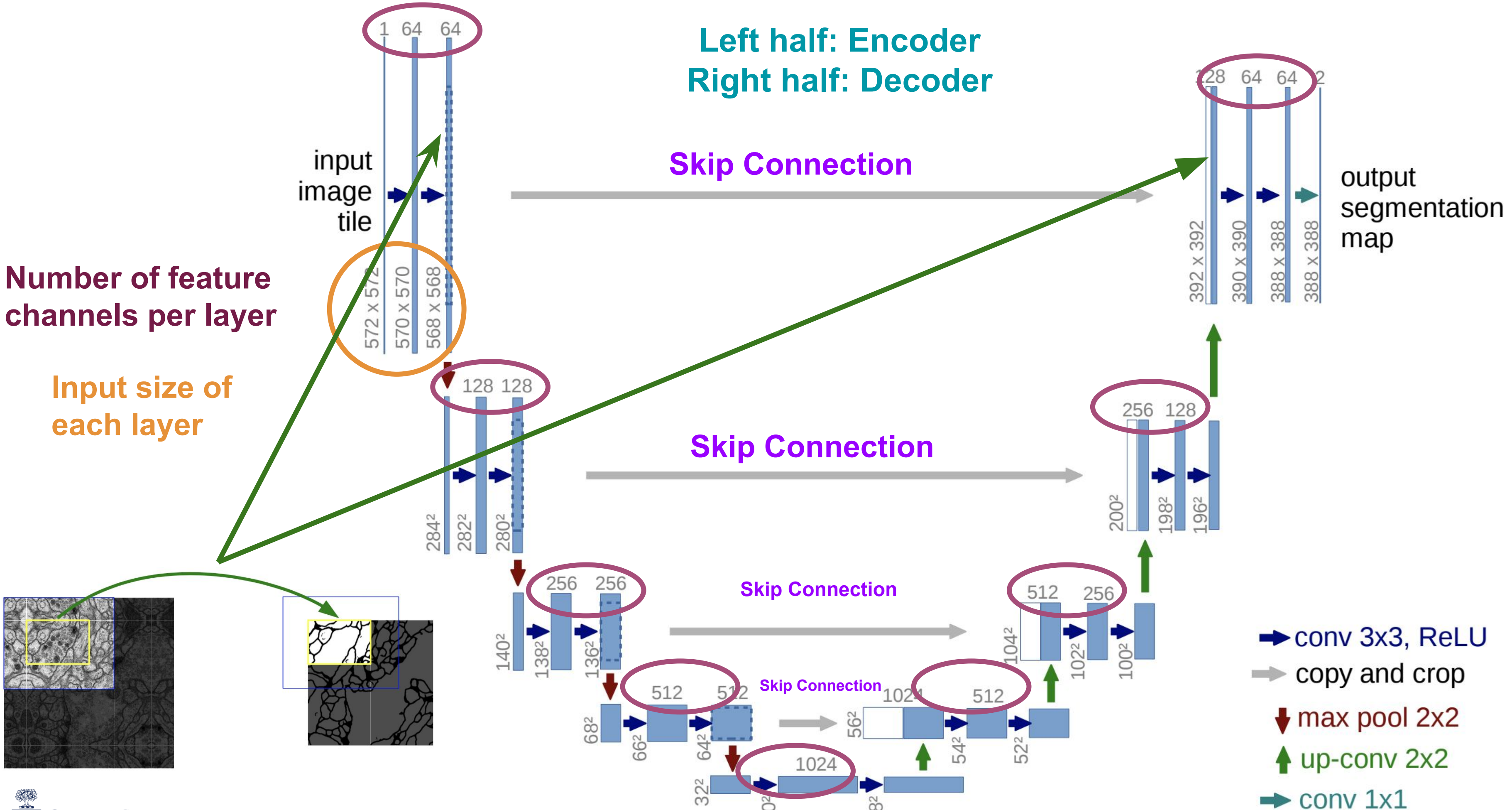
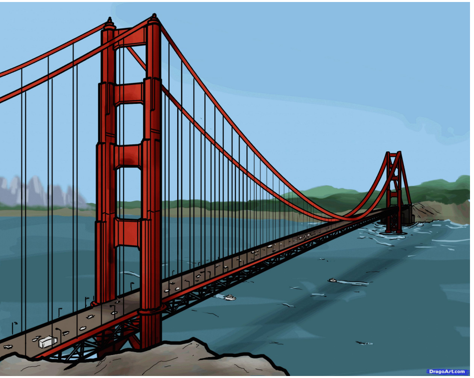
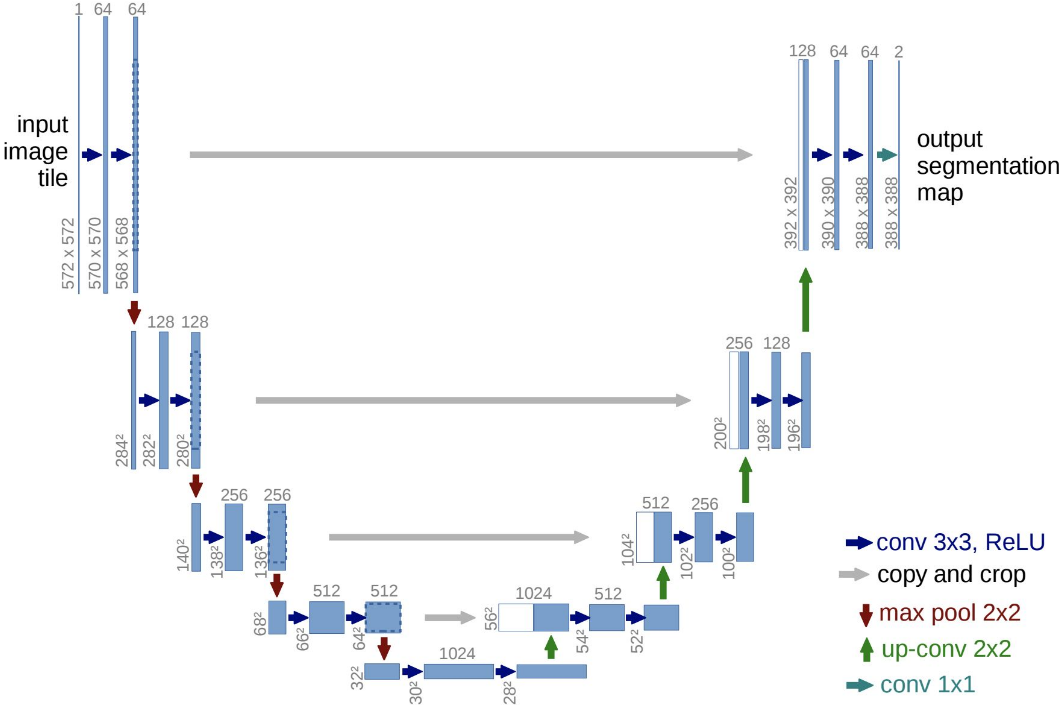


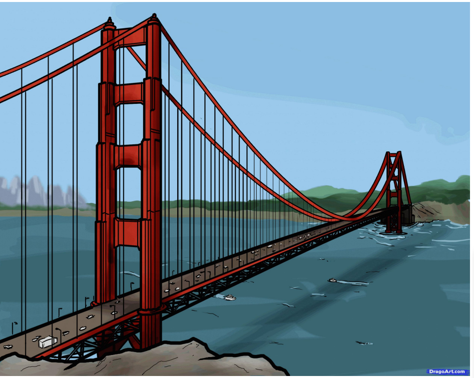
Figure 2, Ronneberger, et al, 2015

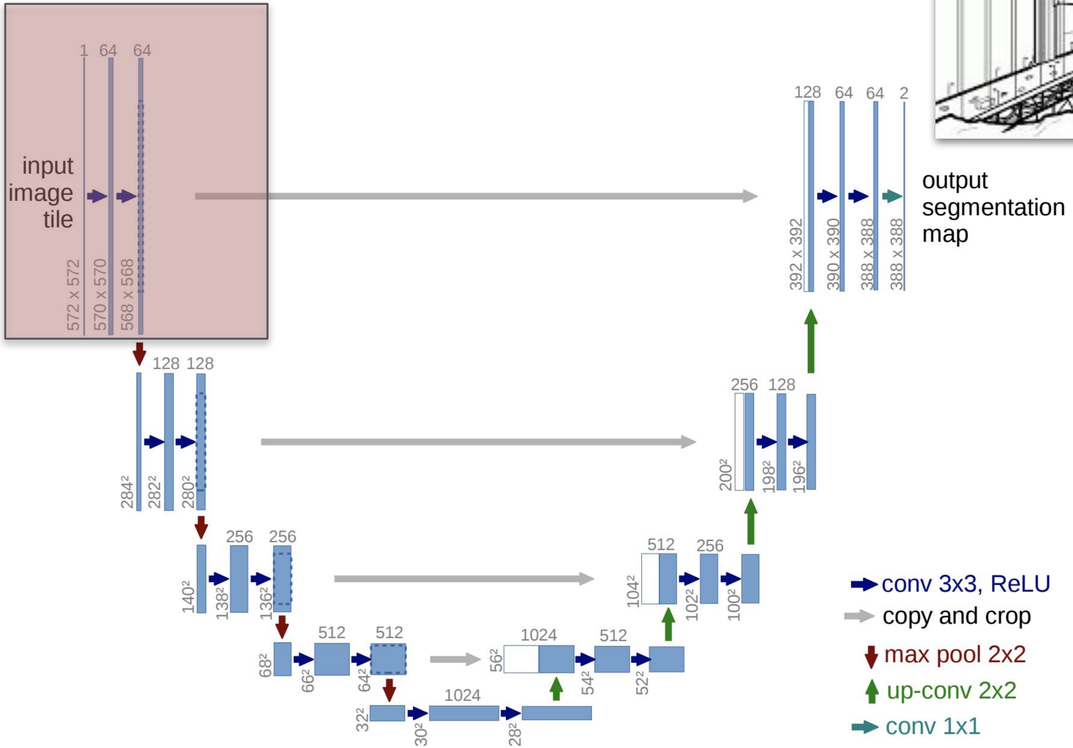
Figure 1, Ronneberger et al, 2015



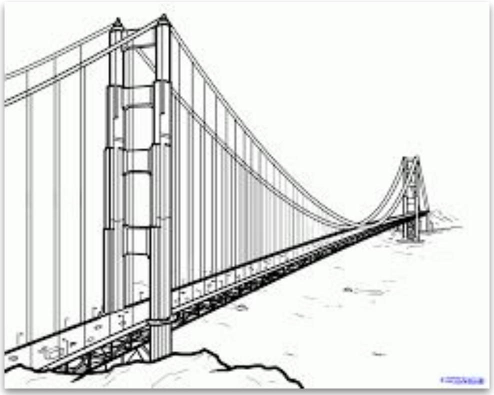


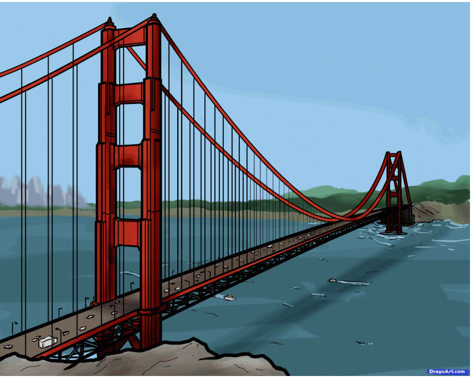
Computer Science
UNIVERSITY OF TORONTO

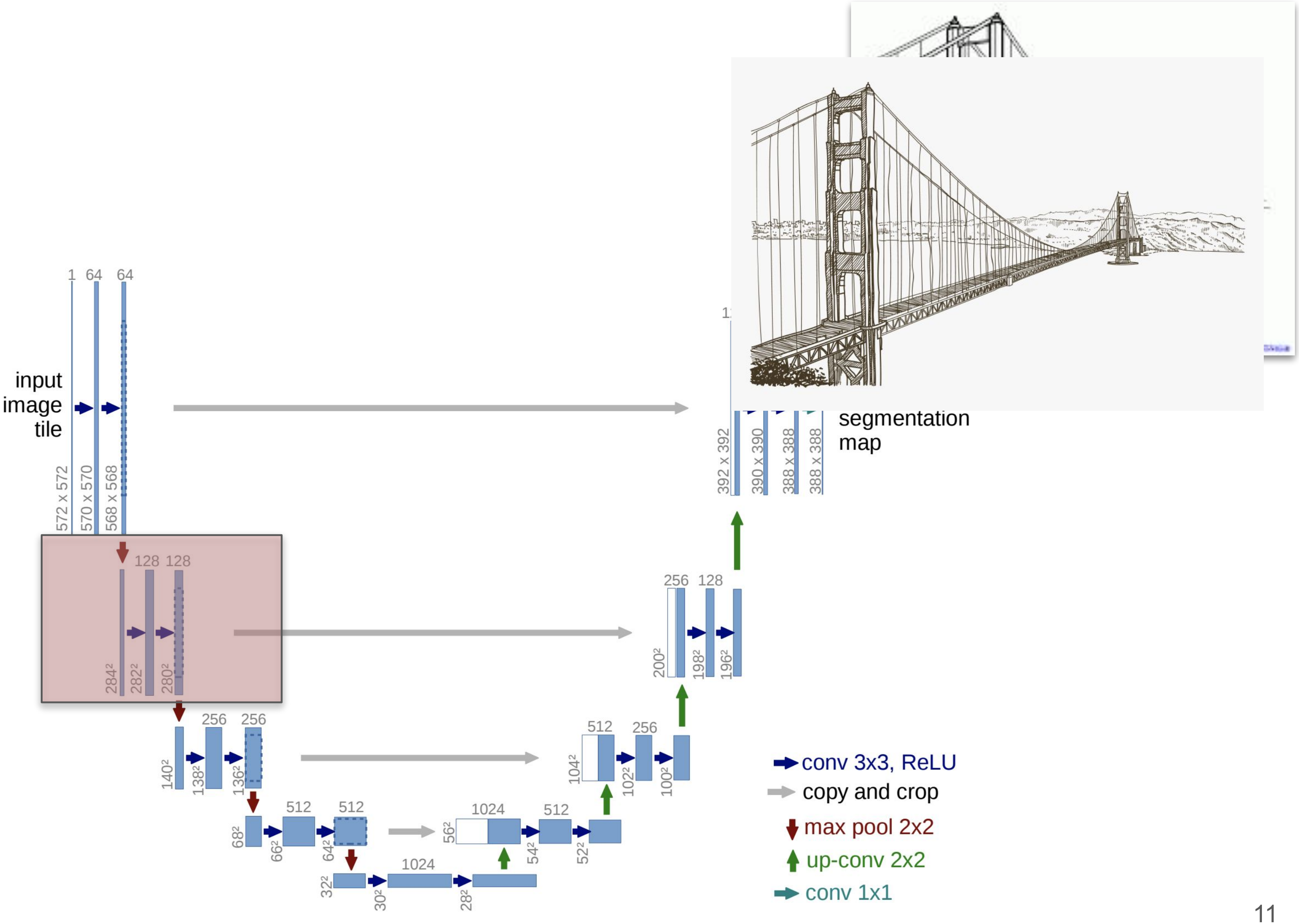




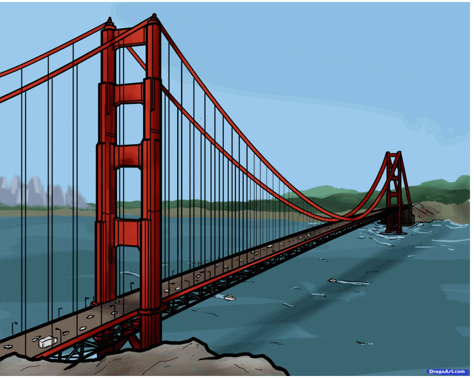
input image tile
1 64 64
572 x 572
570 x 570
568 x 568
128 128
284^2 282^2 280^2
256 256
140^2 138^2 136^2
512 512
68^2 66^2 64^2
1024
32^2 30^2 28^2
1024 512
56^2 54^2 52^2
512 256
104^2 102^2 100^2
256 128
200^2 198^2 196^2
128 64 64 2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

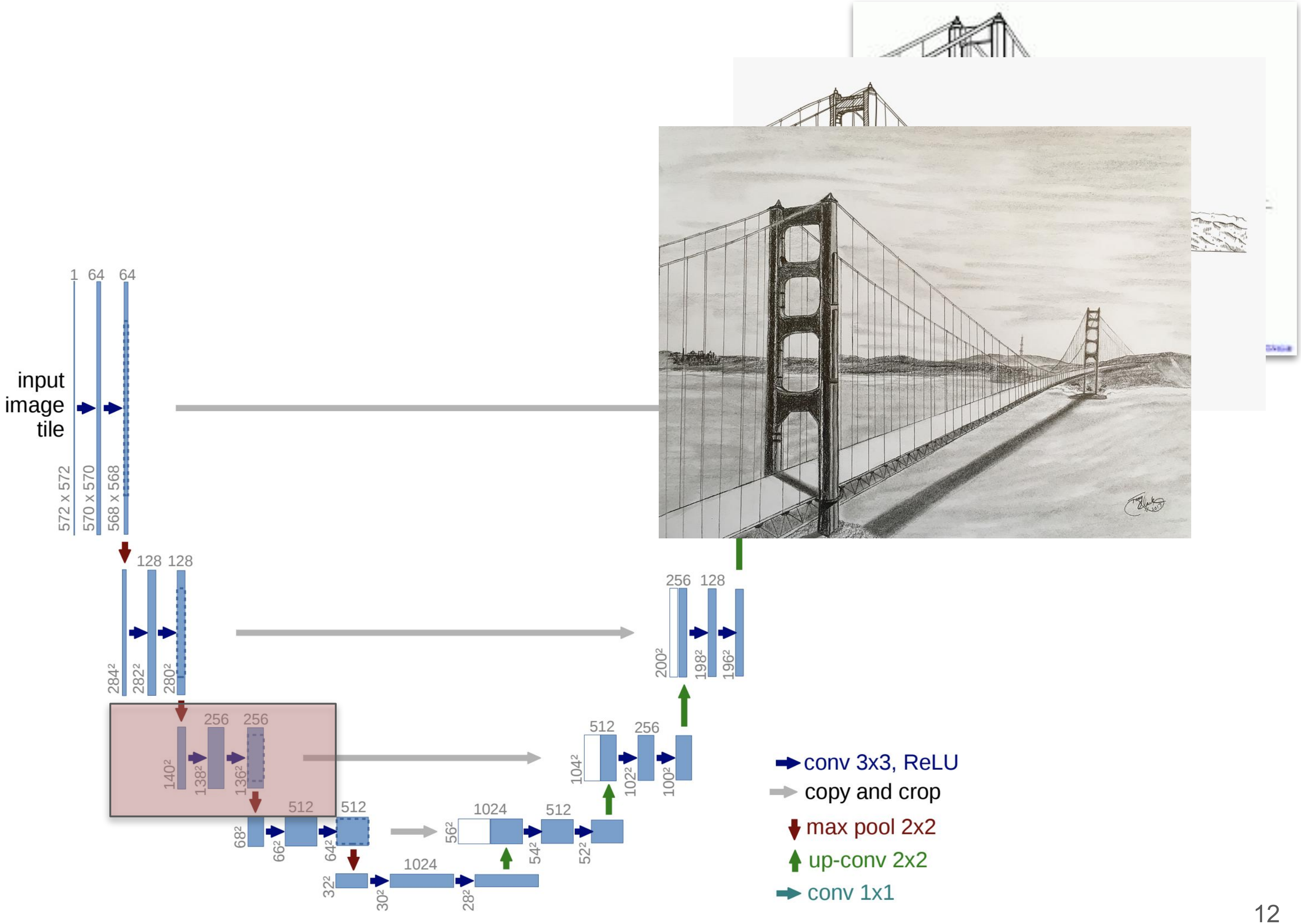


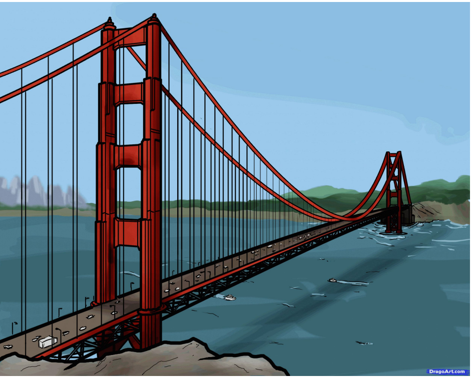


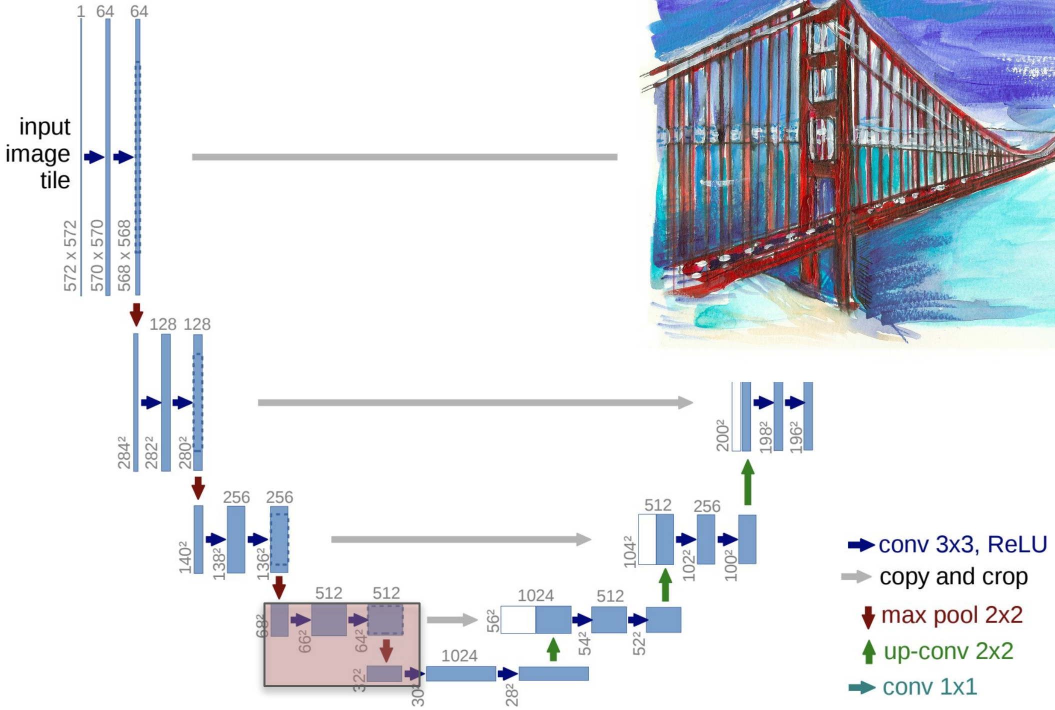


1 64 64
input
image
tile
572 x 572
570 x 570
568 x 568
128 128
284^2
282^2
280^2
256 256
140^2
138^2
136^2
512 512
68^2
66^2
64^2
1024
32^2
30^2
28^2
1024 512
56^2
54^2
52^2
512 256
104^2
102^2
100^2
256 128
200^2
198^2
196^2
segmentation
map
392 x 392
390 x 390
388 x 388
388 x 388
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

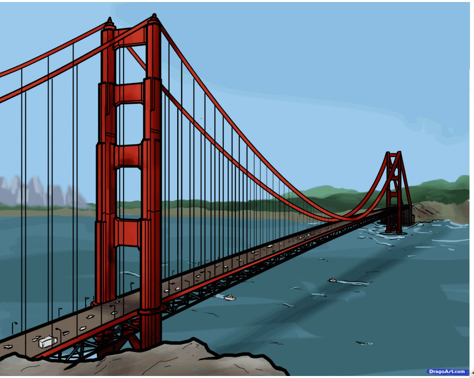


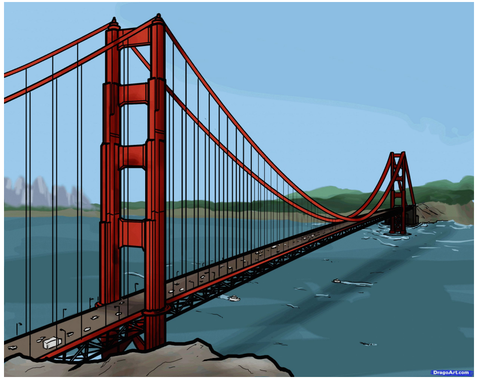

1 64 64
input
image
tile
572 x 572
570 x 570
568 x 568
128 128
284²
282²
280²
256 256
140²
138²
136²
512 512
68²
66²
64²
32²
1024
30²
28²
1024 512
56²
54²
52²
512 256
104²
102²
100²
256 128
200²
198²
196²
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

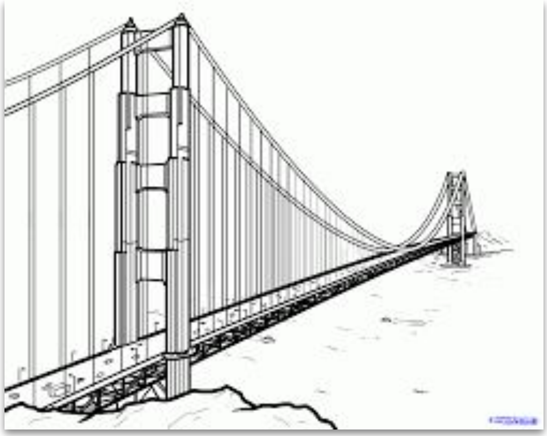


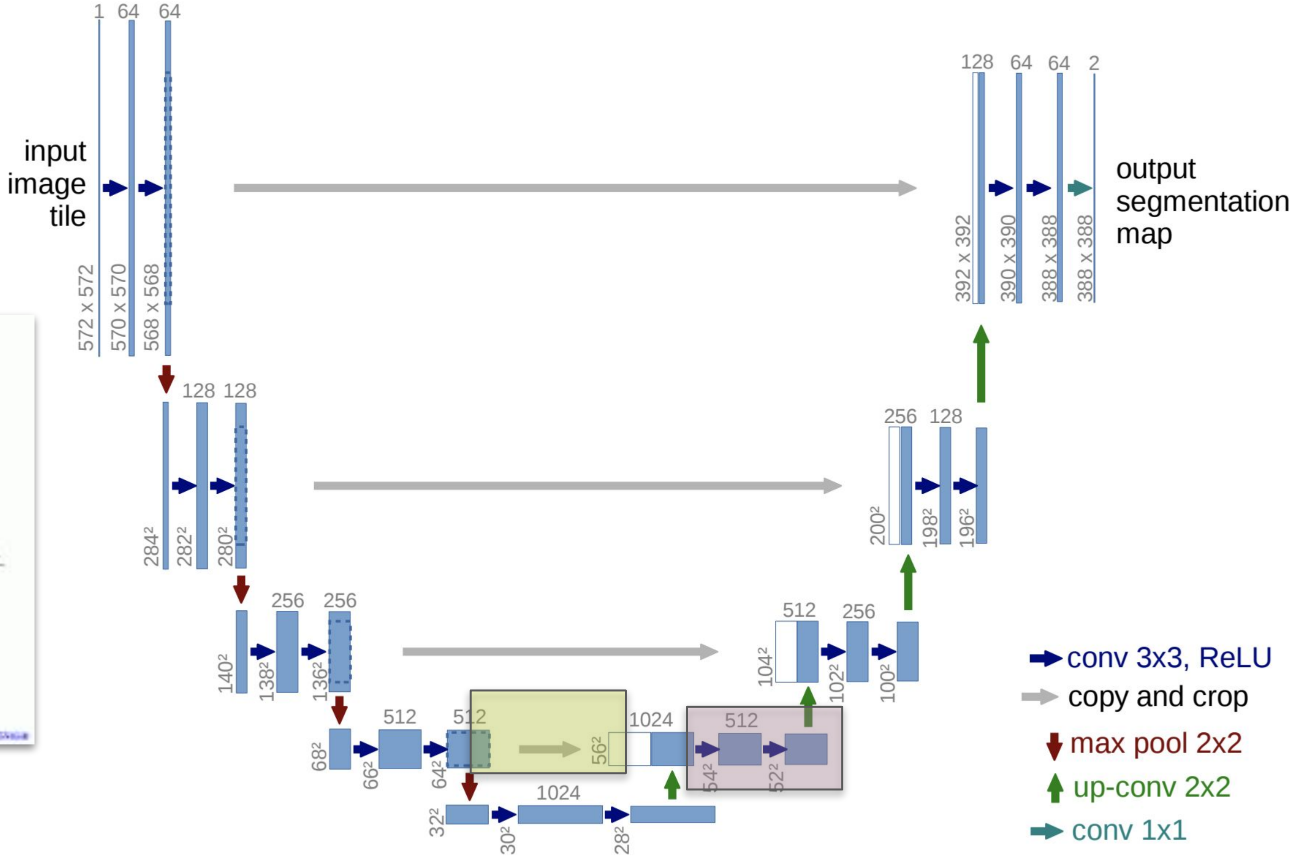


1
64
64
input
image
tile
572 x 572
570 x 570
568 x 568
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
30²
28²
1024
56²
512
54²
52²
512
256
104²
102²
100²
200²
198²
196²
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

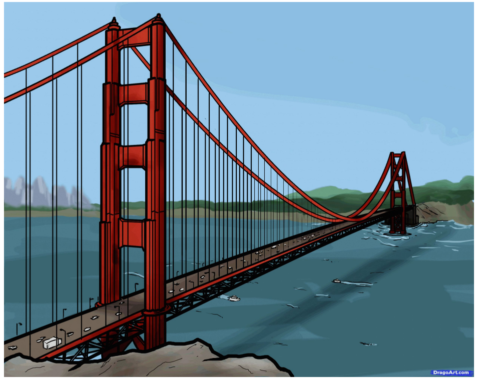

DragoArt.com

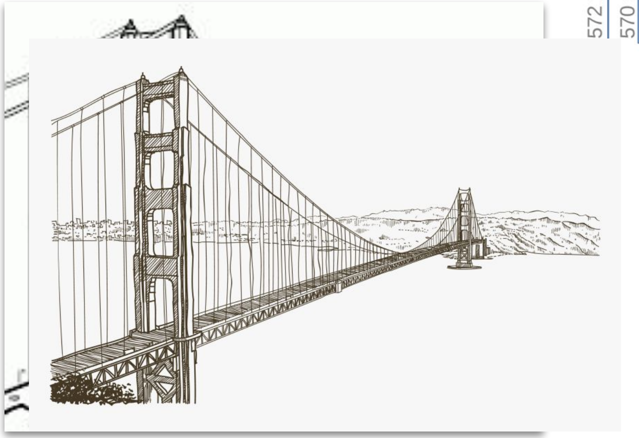


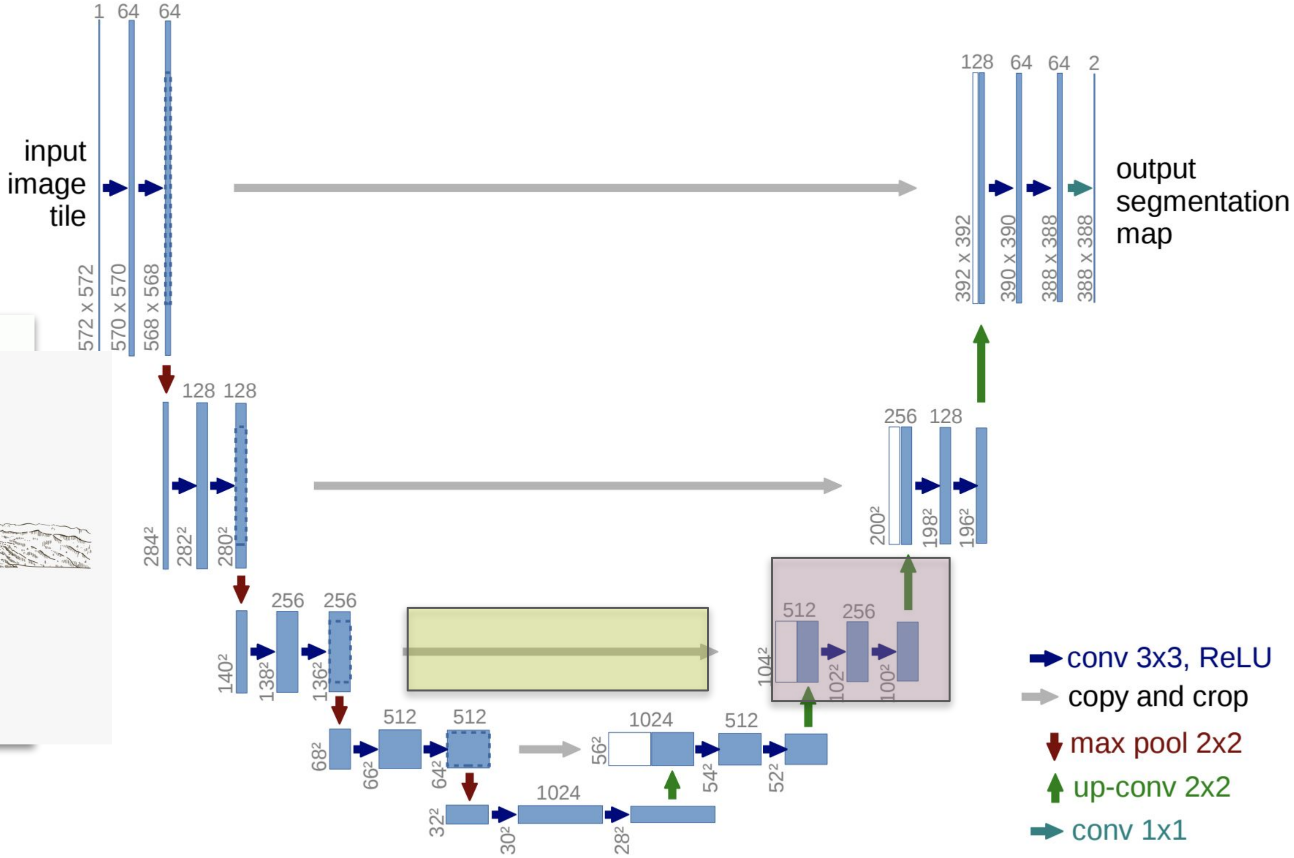




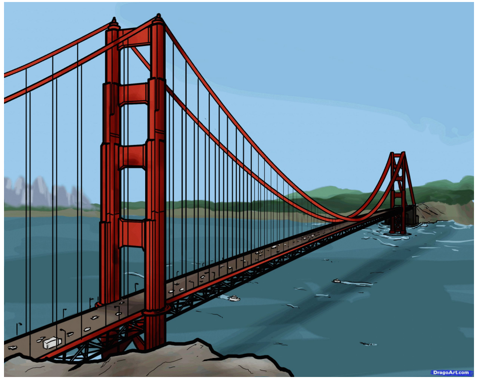
input image tile
output segmentation map
1 64 64
572 x 572
570 x 570
568 x 568
128 128
284^2 282^2 280^2
256 256
140^2 138^2 136^2
512 512
68^2 66^2 64^2
1024
32^2 30^2 28^2
1024 512
56^2 54^2 52^2
512 256
104^2 102^2 100^2
256 128
200^2 198^2 196^2
128 64 64 2
392 x 392
390 x 390
388 x 388
388 x 388
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

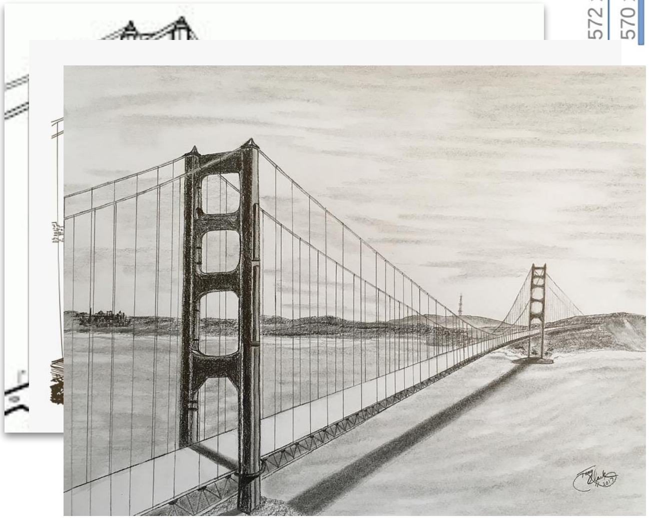


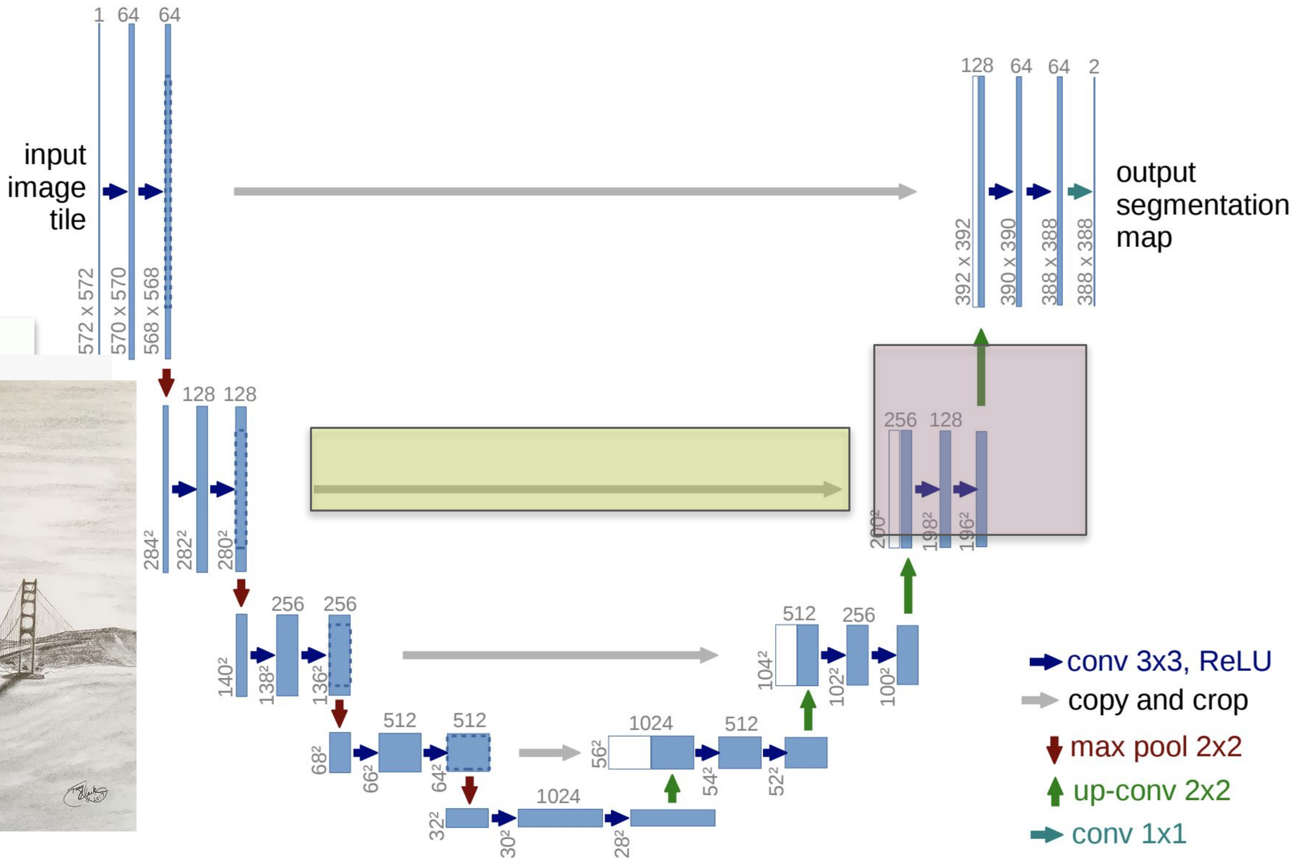




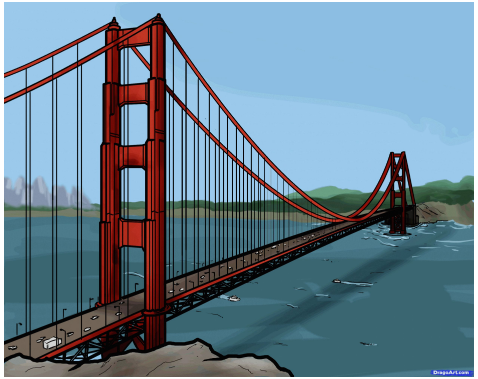
input image tile
output segmentation map
1 64 64
572 x 572
570 x 570
568 x 568
128 128
284²
282²
280²
256 256
140²
138²
136²
512 512
68²
66²
64²
1024
32²
30²
28²
1024 512
56²
54²
52²
512 256
104²
102²
100²
256 128
200²
198²
196²
128 64 64 2
392 x 392
390 x 390
388 x 388
388 x 388
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

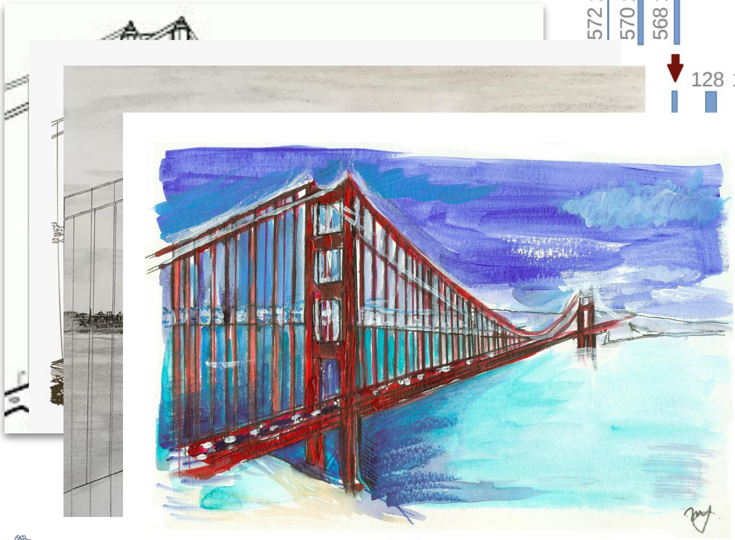


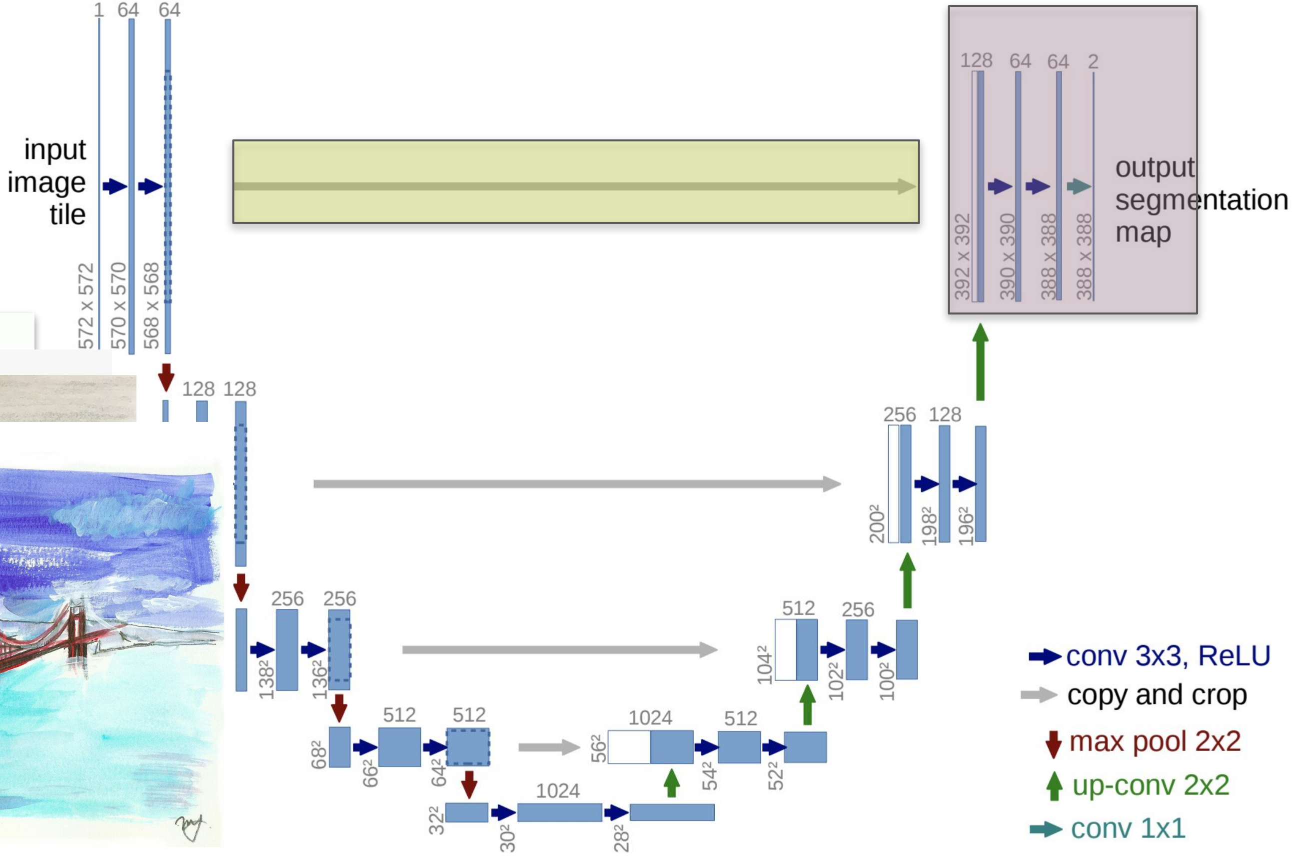


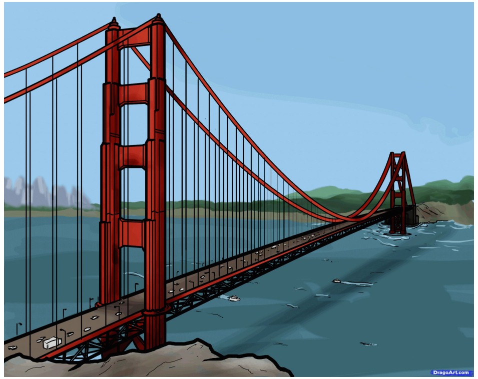


input image tile
1 64 64
572 x 572
570 x 570
568 x 568
128 128
284²
282²
280²
256 256
140²
138²
136²
512 512
68²
66²
64²
1024
32²
30²
28²
1024 512
56²
54²
52²
512 256
104²
102²
100²
256 128
200²
198²
196²
128 64 64 2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

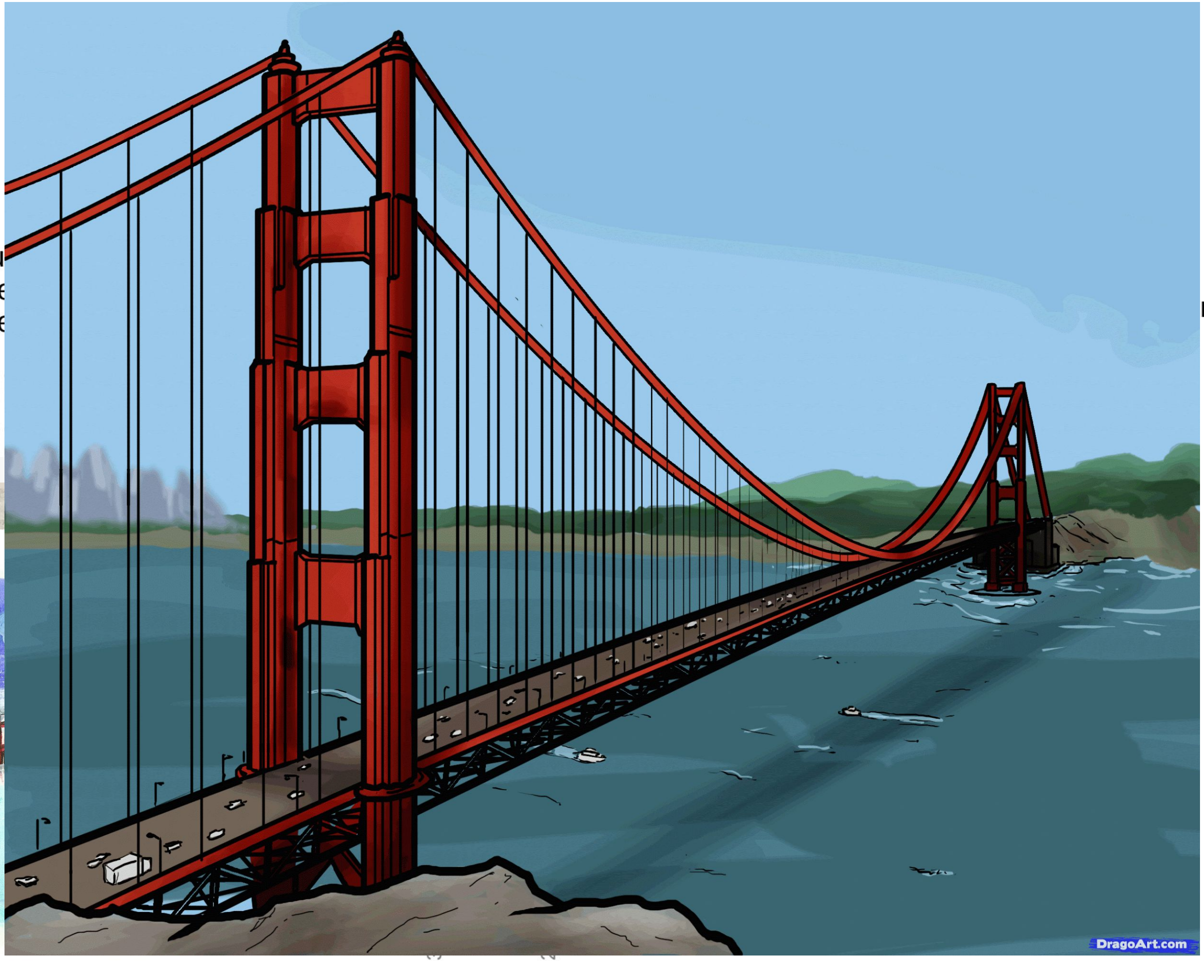





input image tile
1 64 64
572 x 572
570 x 570
568 x 568
128 128
256 256
512 512
1024
512
256
128
128 64 64 2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1



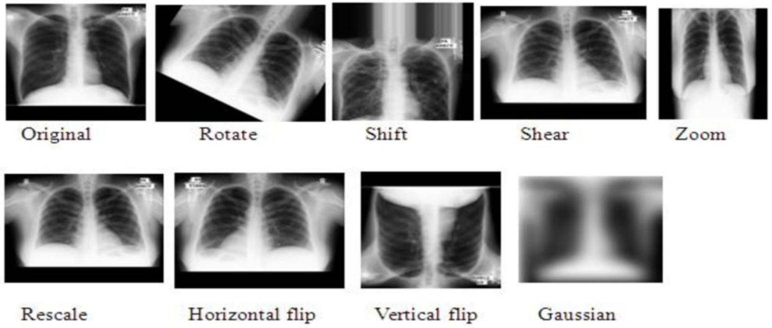
inpu
image
tile



n

Training strategy: Data Augmentation

Teach model invariance and robustness properties

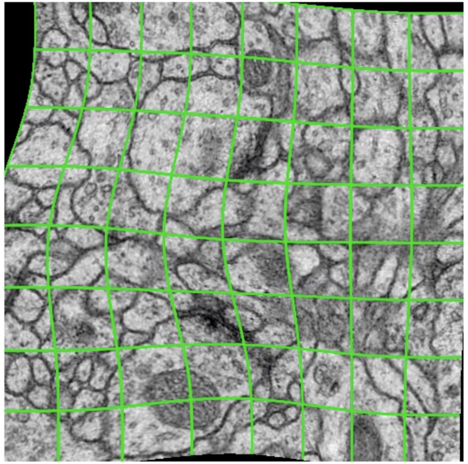


Data augmentations applied on a Chest X-ray image

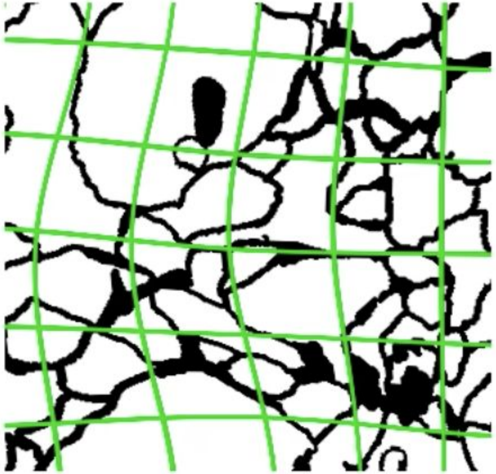
Microscopy Images: very less images in Unet paper

- **Shift** and **rotation invariance**
- Robustness to **deformation** and **gray value variations**

Data Augmentation: Random Elastic Deformation



Elastic transformation on raw image



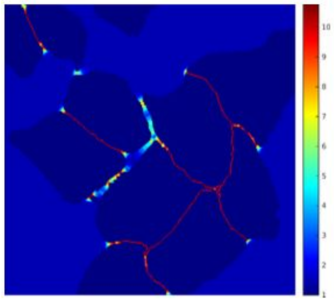
Elastic transformation on corresponding mask.

Credit: https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

Other training strategies

i) Touching cells: pixel-wise weighted loss

Segmentation mask: White(cells) and Black (background)



Loss weight for each pixel

ii) Favour larger input tiles over larger batch size

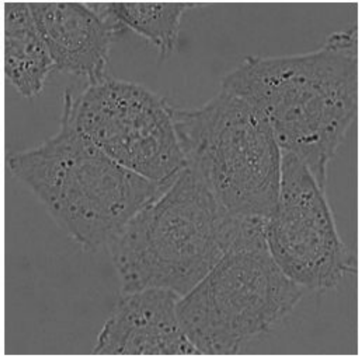
iii) Good weight initialization

Experimental Results: Segmentation of Neuronal Structures in EM stacks

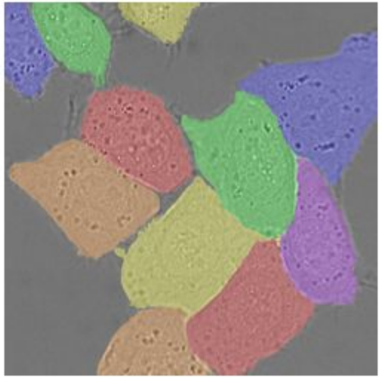
Dataset: EM Segmentation Challenge(ISBI 2012)

30 images (512x512 pixels)

Transmission electron microscopy (**TEM**) of Drosophila first instar larva ventral nerve cord (**VNC**)



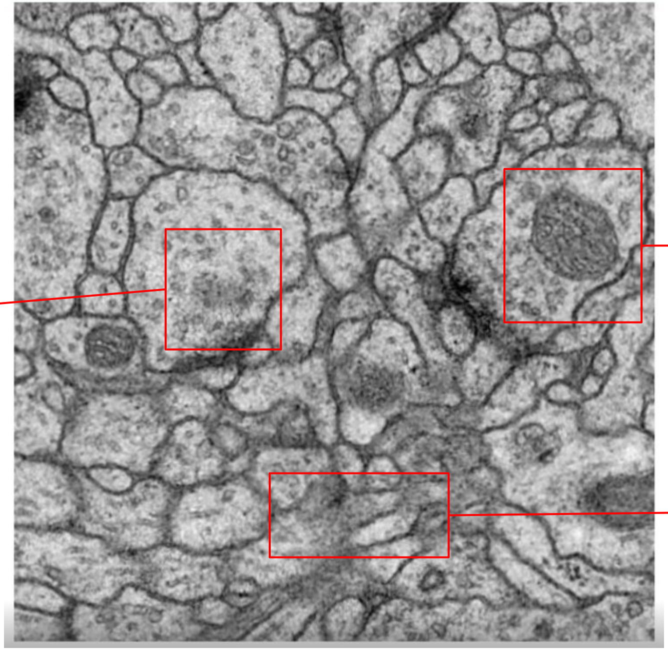
Raw TEM sample image



Overlay of ground truth on raw image

Challenges in the dataset

Structures with very low contrast



Other structures

Fuzzy membranes

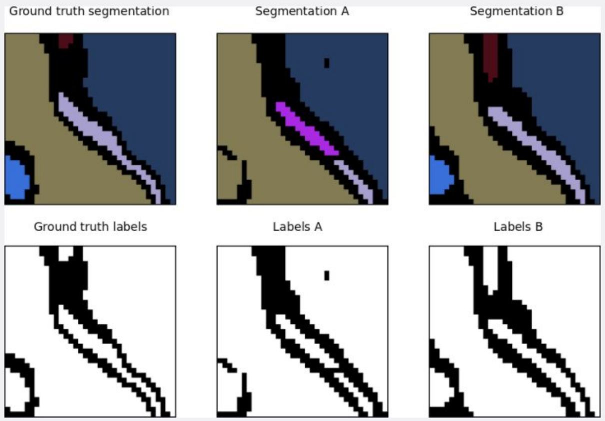
Raw image

Evaluation: EM stacks

Penalizes **topological disagreements,** and used to compare the performance of boundary labellings

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Ranking in EM segmentation challenge, sorted by warping error



Application of the topology-preserving warping error. Example A and B have almost the same amount of pixel error with respect to the ground truth, however, example B has no topological error.

Credit: https://imagej.net/plugins/tws/topology-preserving-warping-error

IDSIA [1]: Sliding Window Convolution Network

Evaluation: EM stacks

Penalizes **connectivity errors**

Compares segmentations in which regions are non-contiguous clusters of pixels

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Ranking in EM segmentation challenge, sorted by warping error

Given 2 segmentations: S1 and S2 of image **I** with n pixels:

$$RI = \frac{a+b}{\binom{n}{2}}$$

$$RE = 1 - RI$$

a = number of pixel pairs in **I** that are in the **same** object in S1 as in **same** object of S2 (same label)

b = number of pixel pairs in **I** that are in the **different** object in S1 as in **different** object of S2 (different labels)

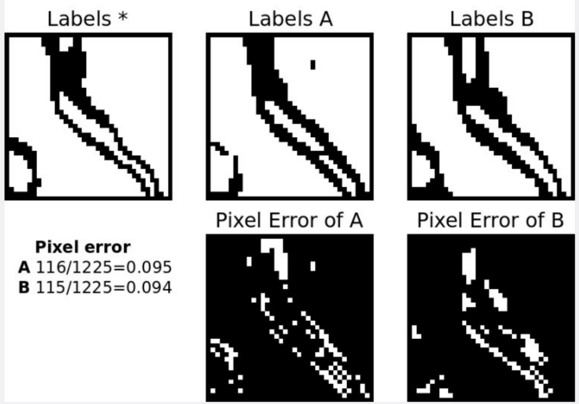
Evaluation: EM stacks

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Ranking in EM segmentation challenge, sorted by warping error

Focuses of **pixel level disagreement**

Measures pixel differences between the segmented and original image



Pixel error between two different segmentations labels (A and B) with respect to the original labels (*, ground truth).

Results: ISBI cell Tracking challenge (2014 and 2015)

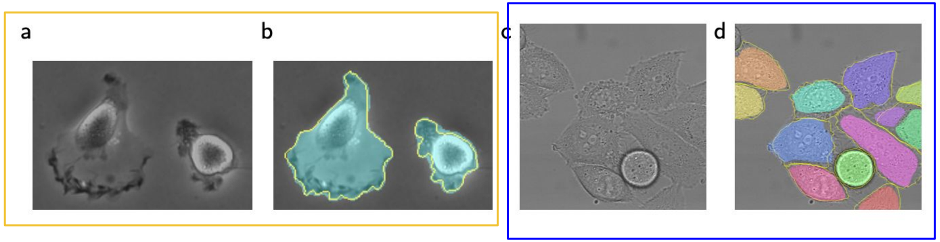


Fig. 4. Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

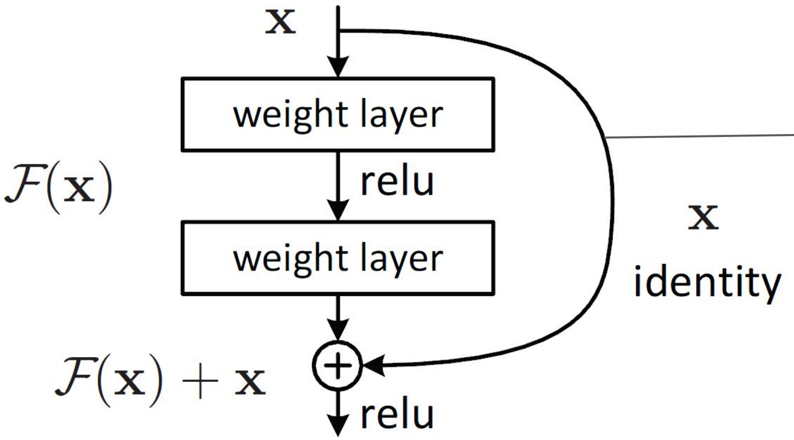
Segmentation results - IOU (Intersection over union) on ISBI

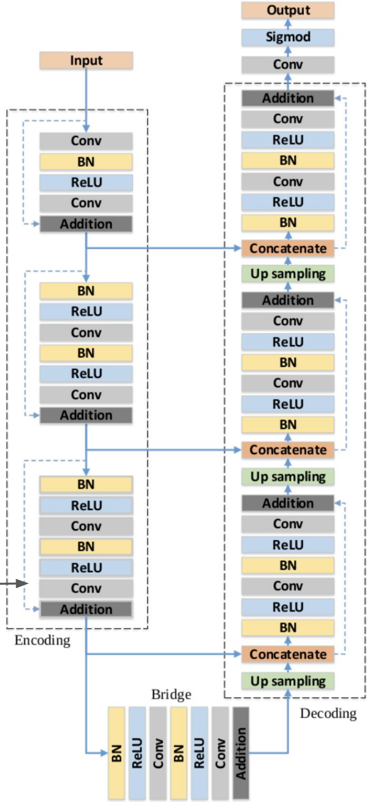
- DIC-HeLa - **20 partially** annotated training images (DIC - Differential Inference Contrast) microscope
- PhC-U373 - **35 partially** annotated training images, phase contrast microscopy

Limitations: U-Net’s variants

a) Residual U-Net

- Residual networks are proposed to overcome the problem of Deep CNN's (vanishing gradients)
- Residual U-Net borrows residual blocks from ResNet[1] paper
- Train deeper networks, leading to faster convergence





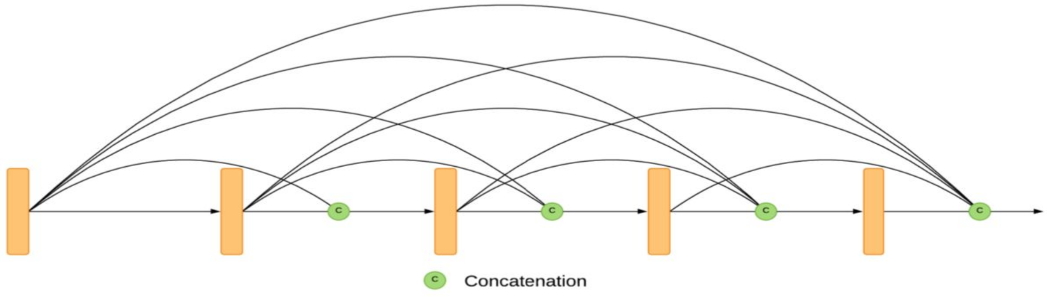
[1]: https://arxiv.org/abs/1512.03385

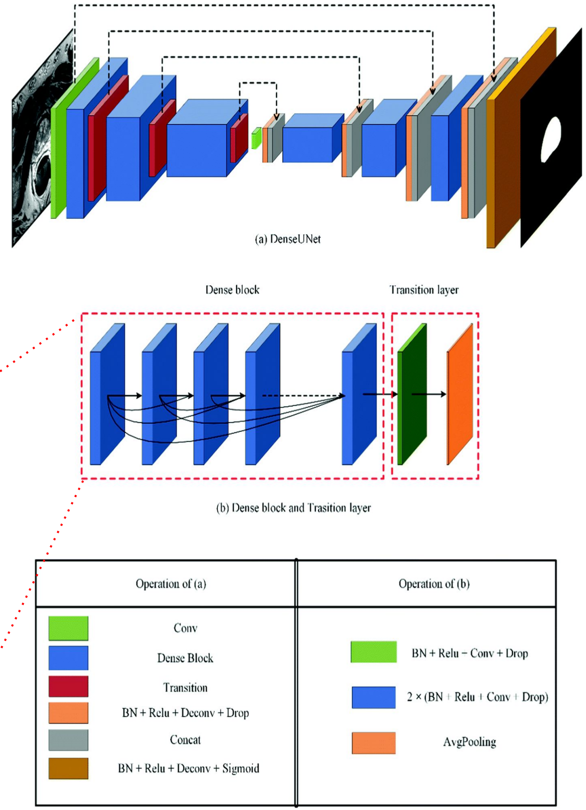
b) Dense-Unet

Dense blocks instead of Conv blocks

Dense-UNet = UNet backbone + 2 modifications

a) Every layer receives features from previous layers
b) Identity maps combined with channel wise concatenation





DenseNet-Unet Architecture

DenseNet: https://arxiv.org/abs/1608.06993

Summary

- Unet makes accurate biomedical semantic segmentations feasible with few training examples
- Encoder captures context, while decoder helps in maintaining localization
  - (localization and use of context at same time)
- Fast inference (1s per image)
- Training strategies
  - data augmentations
  - pixel-wise weighted loss (seems to be key concepts to train network with few images)

Limitations (addressed by other architectures):

- Residual Unet: Train larger models (skip connections)

- Dense Unet: Every layer has contextual information, better segmentation accuracy

Limitations (Our point of view)

a) Determining the **depth** of the network **apriori** is difficult (ablation study was missing)

b) Data Augmentation: how to select the transformation that are suited for a given task?

c) Missing ablation studies for the pre-processing / post processing in EM stacks evaluation

d) Why not dice loss for training the network?

Thank You :)

Questions?!