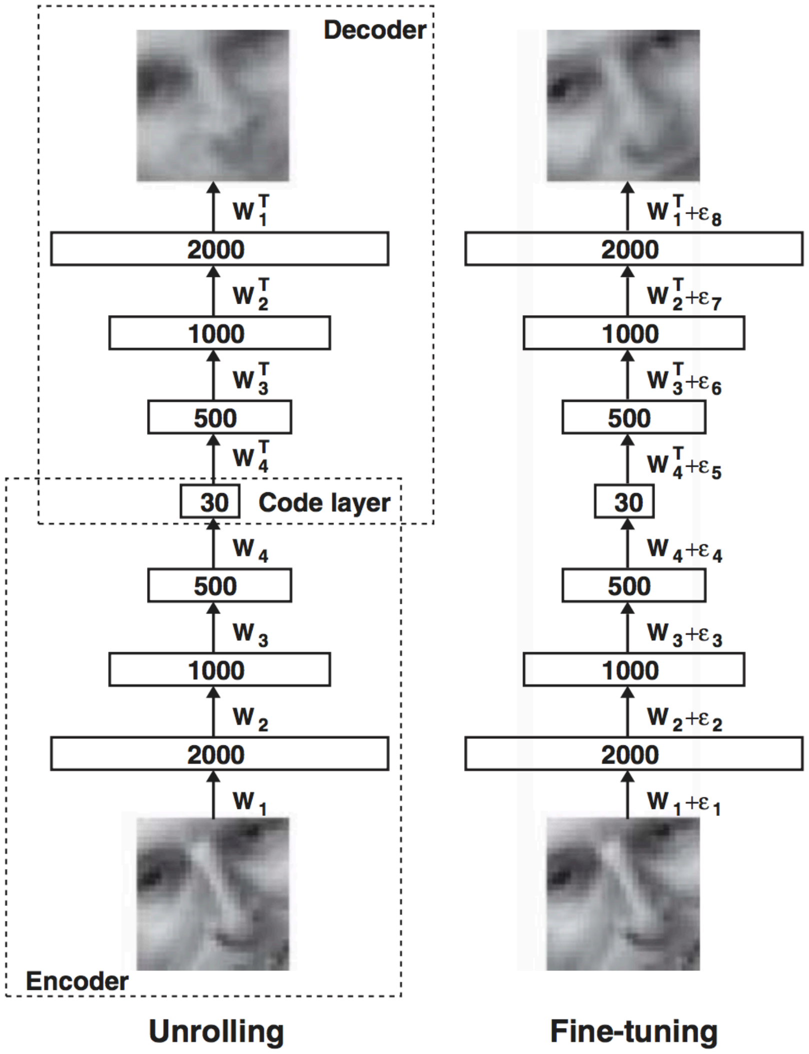
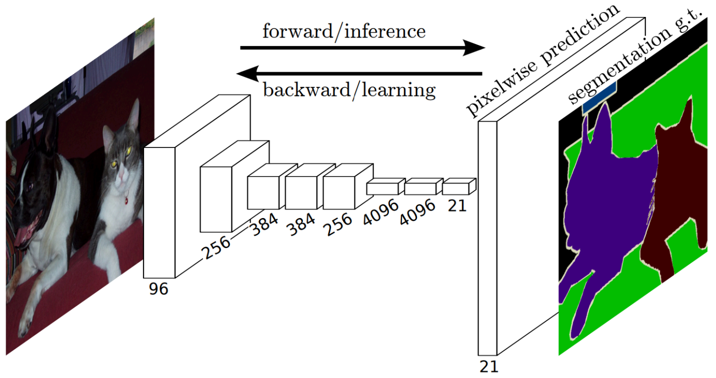
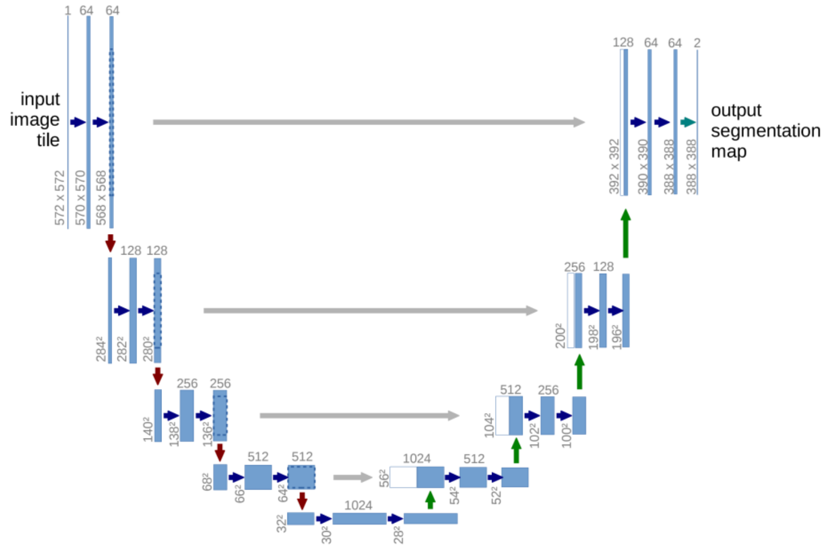
UNet++: A Nested U-Net Architecture for Medical Image Segmentation

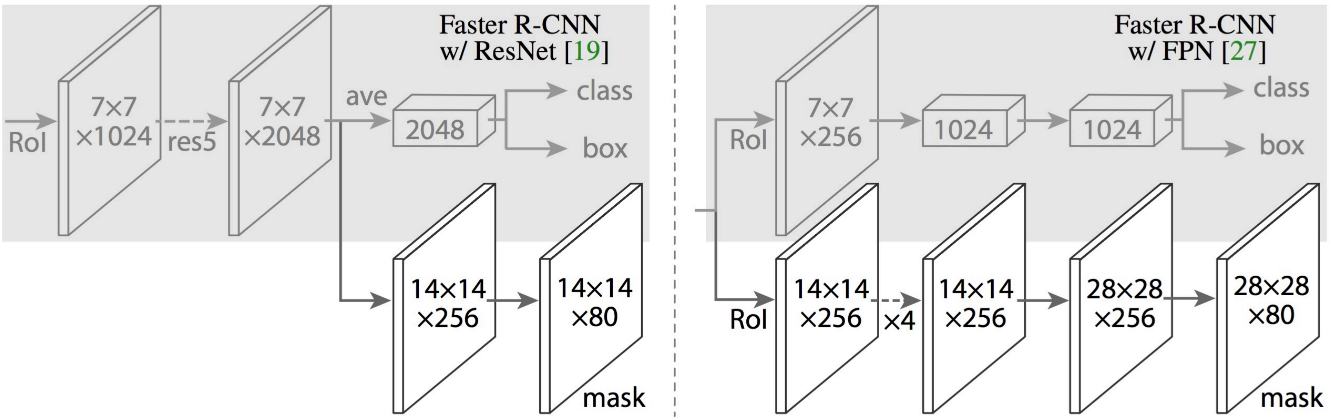
Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang

Arizona State University



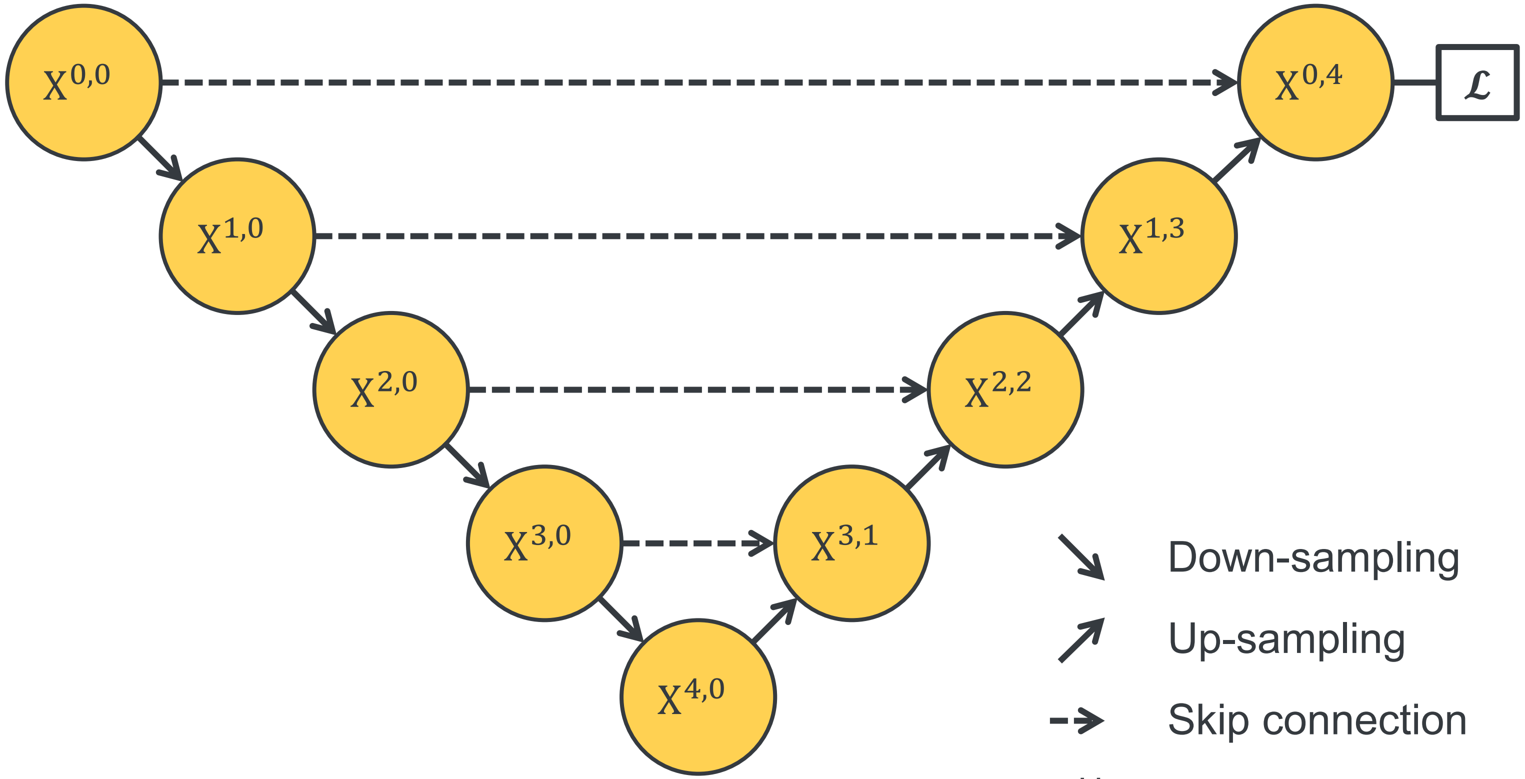






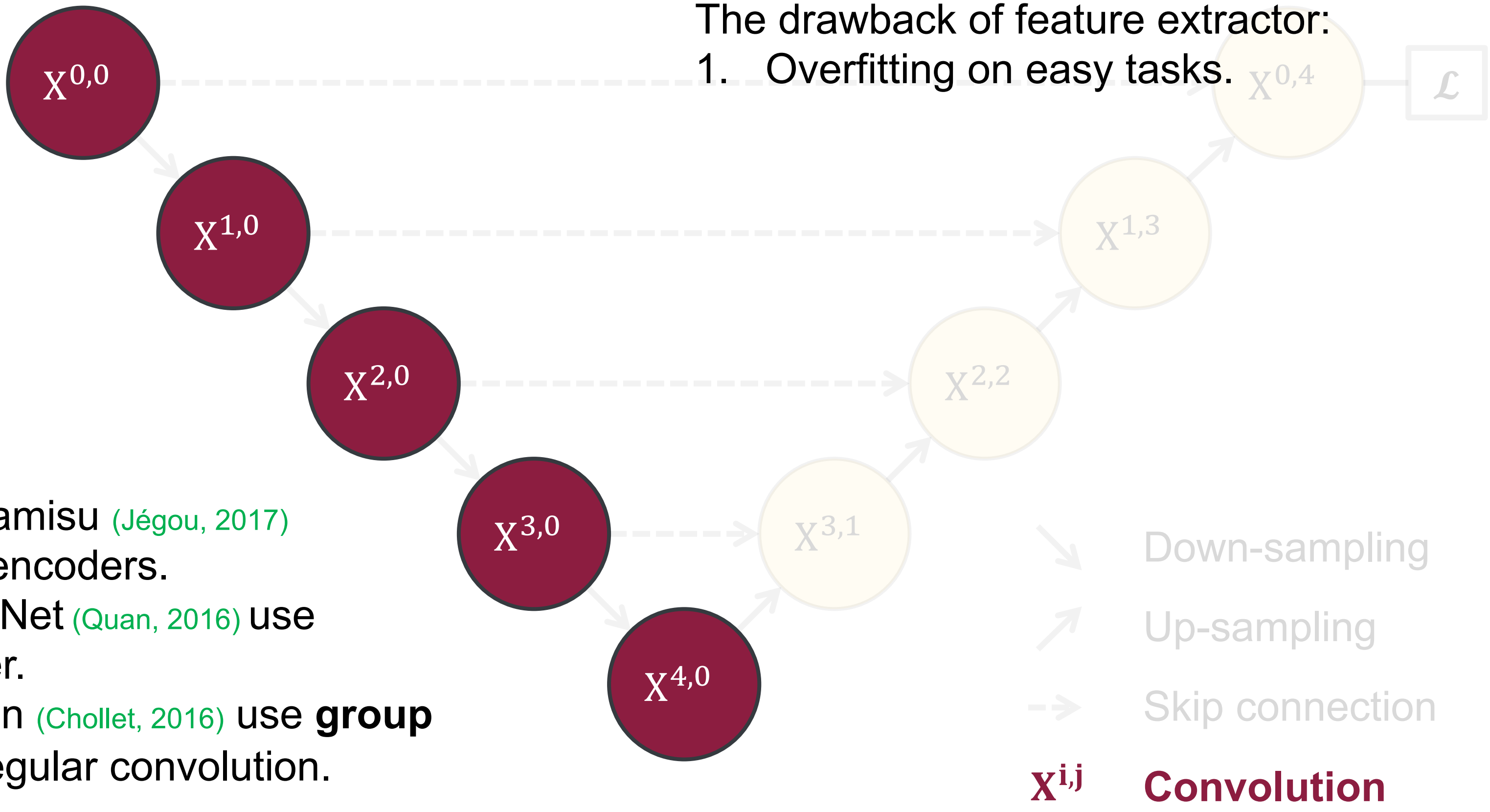
(Ronneberger, 2015)

(Hinton, 2006) — (Long, 2015) — (He, 2018)


$X^{0,0}$
$X^{1,0}$
$X^{2,0}$
$X^{3,0}$
$X^{4,0}$
$X^{3,1}$
$X^{2,2}$
$X^{1,3}$
$X^{0,4}$
$\mathcal{L}$
Down-sampling
Up-sampling
Skip connection
$X^{i,j}$ Convolution

The feature extractor is important:
1. Good feature representation.
2. Fast convergence speed.

The drawback of feature extractor:
1. Overfitting on easy tasks.



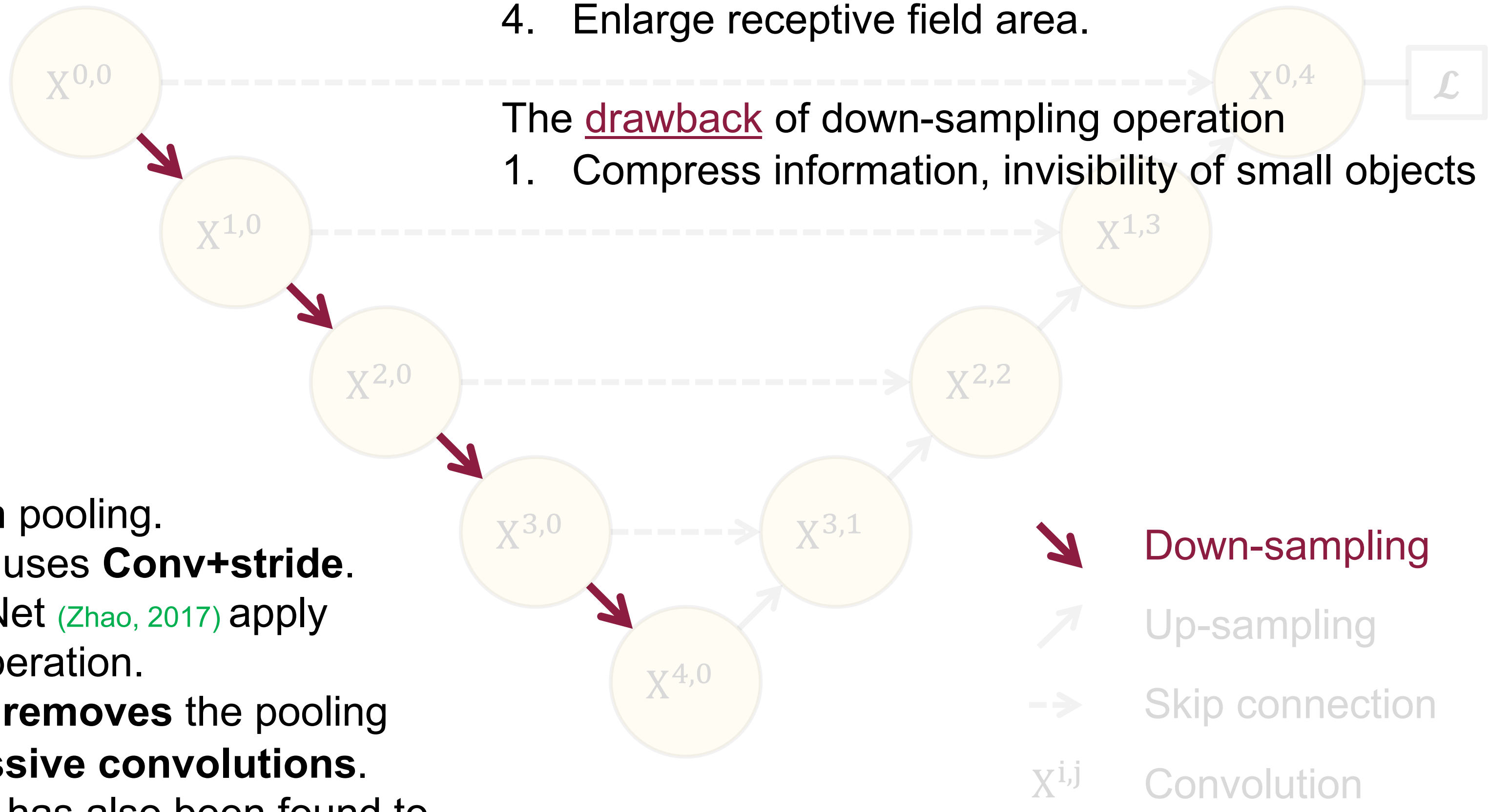
- H-DenseUNet (Li, 2017), Tiramisu (Jégou, 2017) use **dense units** in their encoders.
- PSPNet (Zhao, 2017), FusionNet (Quan, 2016) use **residual units** as encoder.
- ResNext (Xie, 2016), Xception (Chollet, 2016) use **group convolution** instead of regular convolution.

The down-sampling is important:

1. Robust against small input variance.
2. Reduce overfitting.
3. Reduce computation cost.
4. Enlarge receptive field area.

The drawback of down-sampling operation

1. Compress information, invisibility of small objects



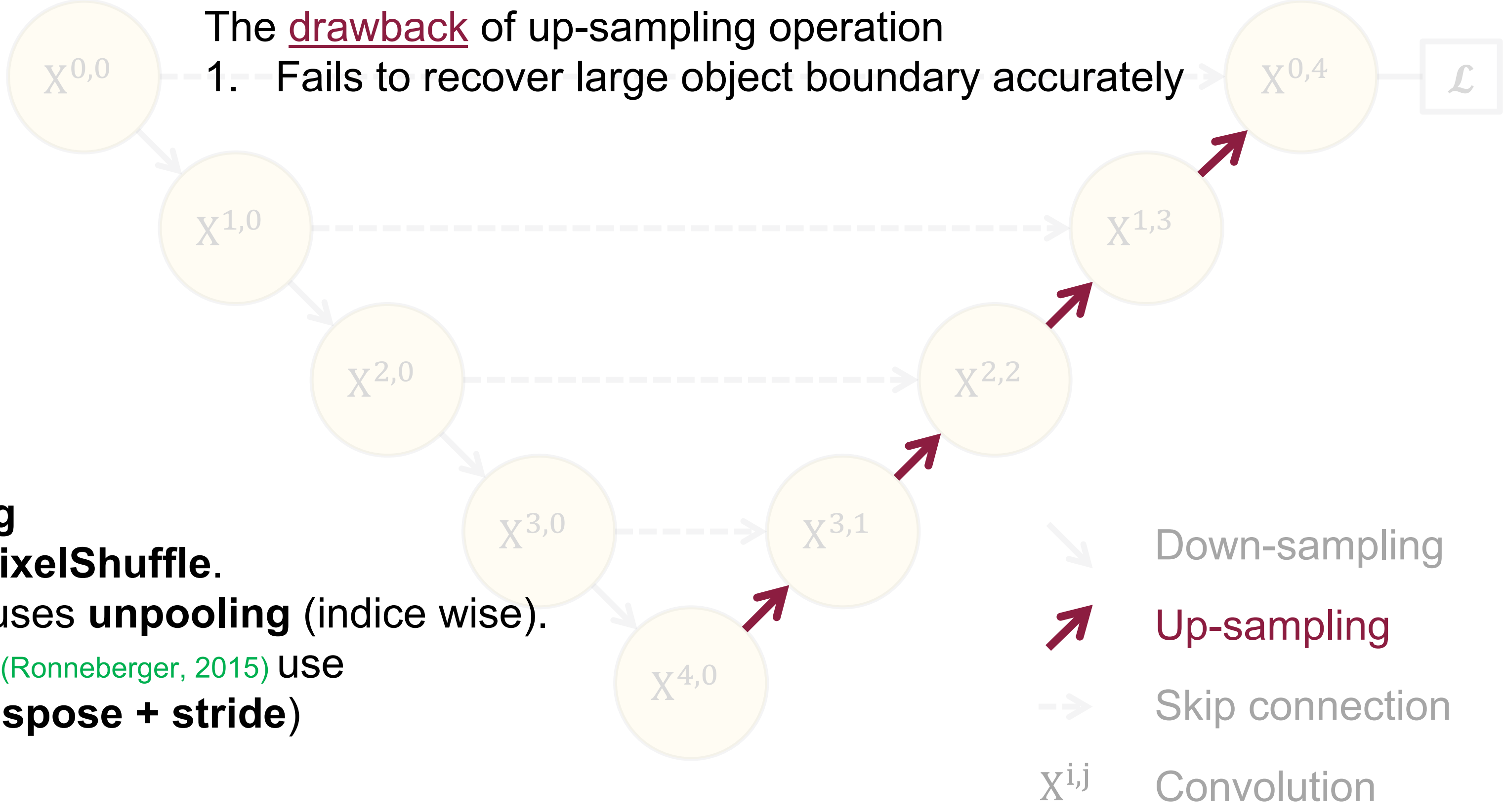
- **Max** vs. **Ave** vs. **L2-norm** pooling.
- ALL-CNN (Springenberg, 2015) uses **Conv+stride**.
- DeepLab (Chen, 2017), PSPNet (Zhao, 2017) apply **dilated convolutional** operation.
- HyperDenseNet (Dolz, 2018) **removes** the pooling layers, only leave **successive convolutions**.
- Discarding pooling layers has also been found to be important in training **VAEs** or **GANs**.

The up-sampling is important:

1. Recovers lost resolution in down-sampling
2. Guides encoder to select important information

The drawback of up-sampling operation

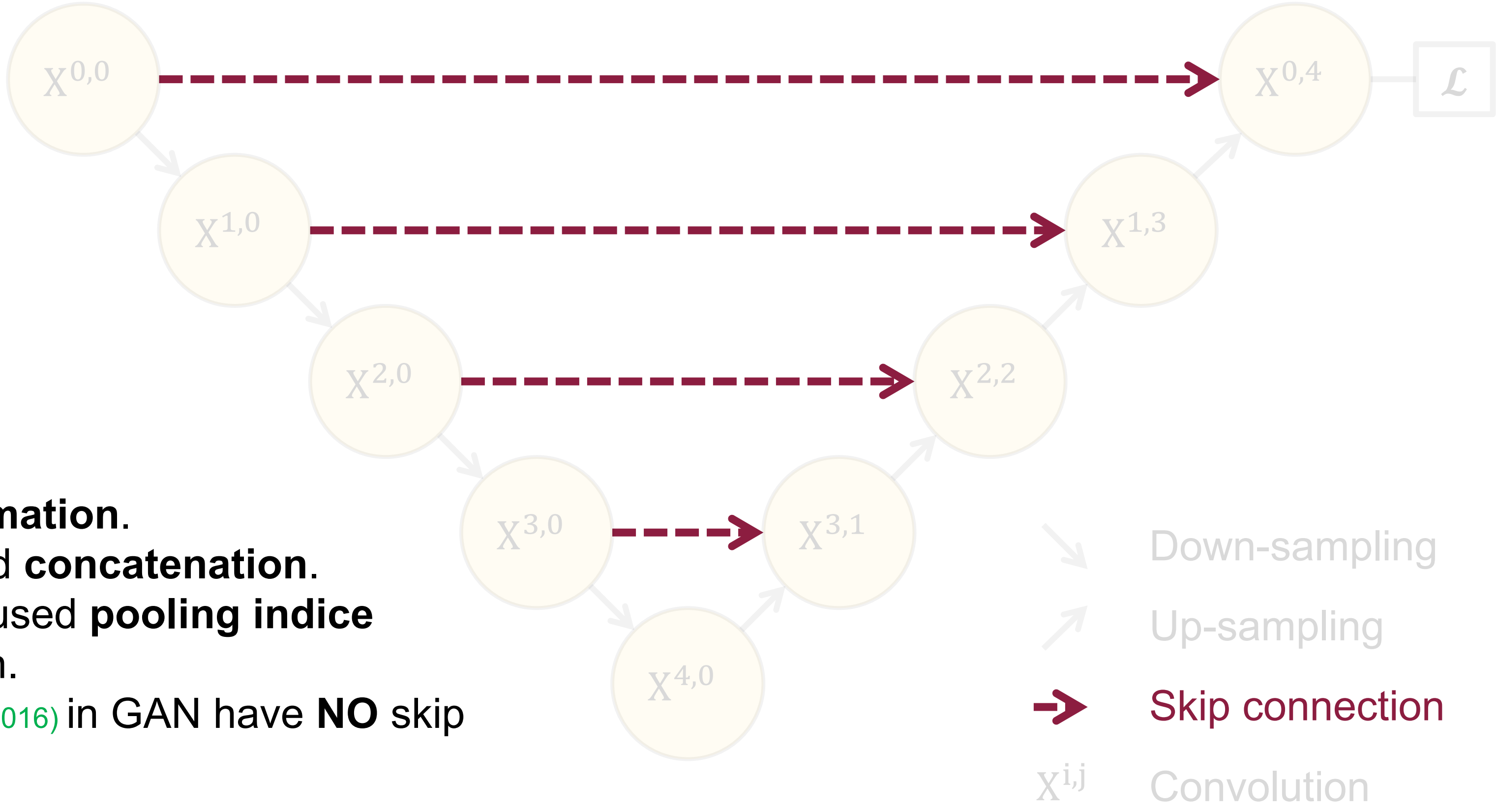
1. Fails to recover large object boundary accurately



- Up-sampling by **repeating**
- PixelDCL (Gao, 2017) uses **PixelShuffle**.
- SegNet (Badrinarayanan, 2016) uses **unpooling** (indice wise).
- FCN (Long, 2015) and U-Net (Ronneberger, 2015) use deconvolution (**ConvTranspose + stride**)

The skip connection is important:

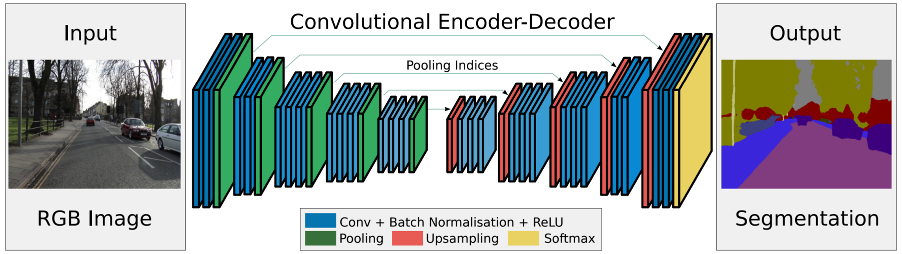
1. Fights the vanishing gradient problem.
2. Learns pyramid level features.
3. Recover info loss in down-sampling.



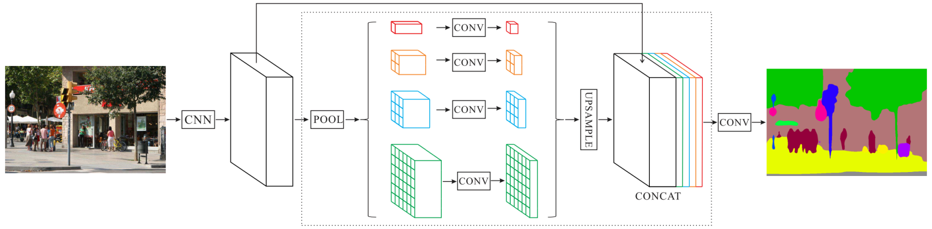
- FCN (Long, 2015) used **summation**.
- U-Net (Ronneberger, 2015) used **concatenation**.
- SegNet (Badrinarayanan, 2016) used **pooling indice** instead of skip connection.
- Some generators (Johnson 2016) in GAN have **NO** skip connection.

In summary, this **encoder-decoder** like architecture is very popular.

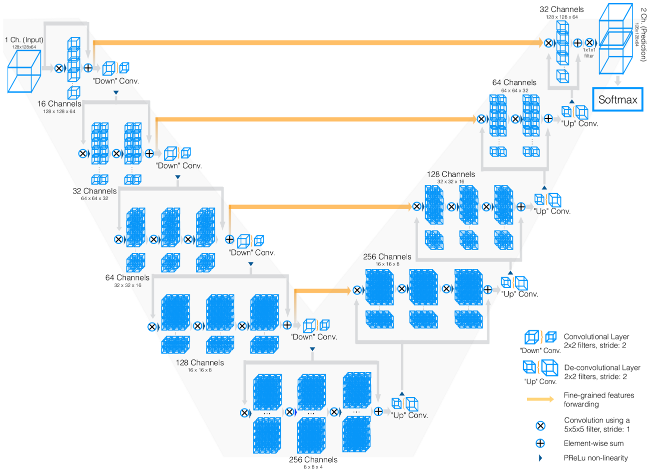
1. Performs consistently.
2. Continuous improvements.



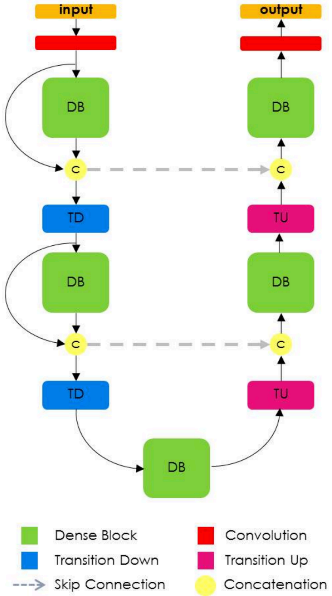
(Badrinarayanan, 2016)



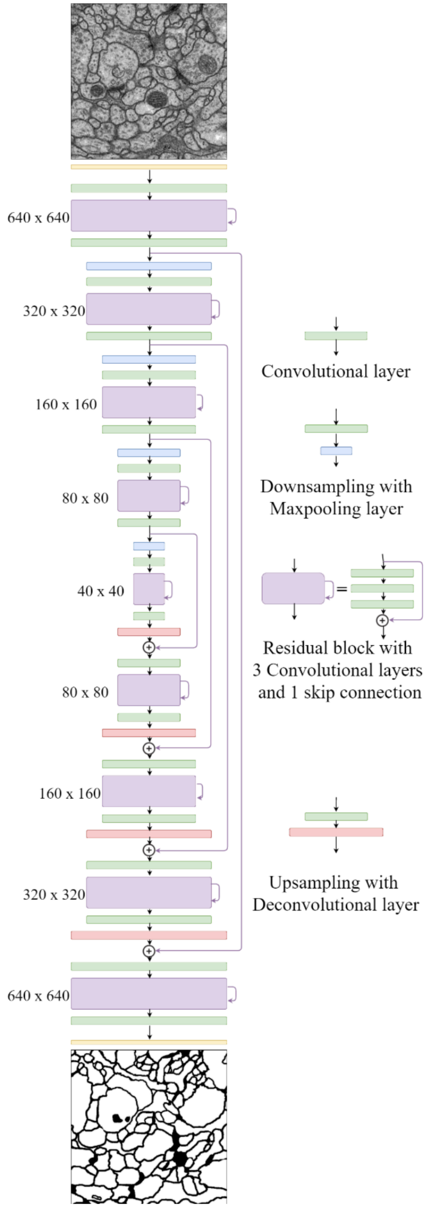
(Zhao, 2017)



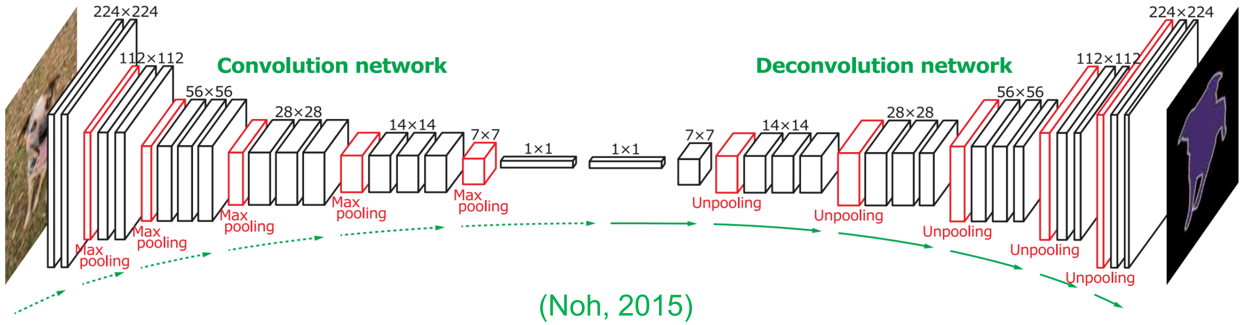
(Milletari, 2016)



(Jegou, 2017)



(Quan, 2016)

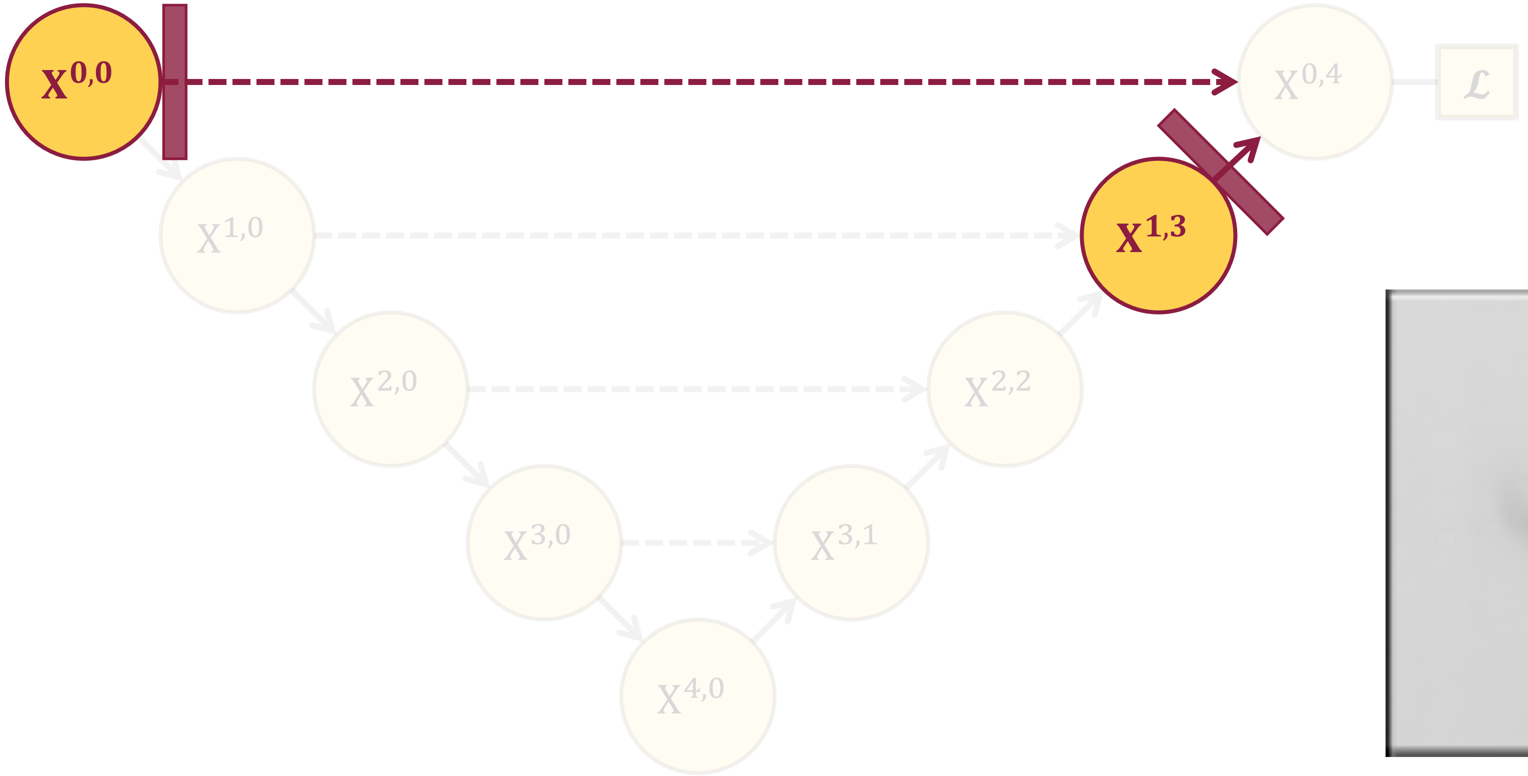


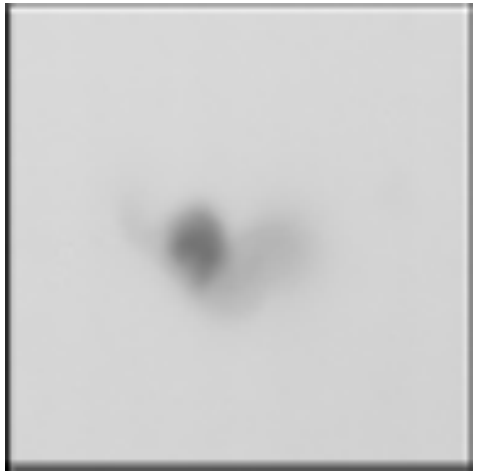
(Noh, 2015)

In summary, this **encoder-decoder** like architecture is very popular.

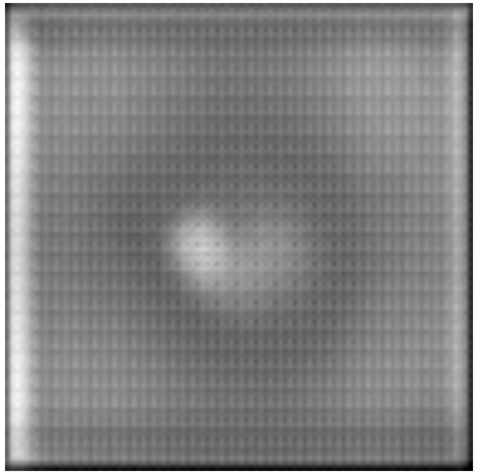
1. Performs consistently.
2. Continuous improvements.

However, we still find **semantic gap** along the skip connections.

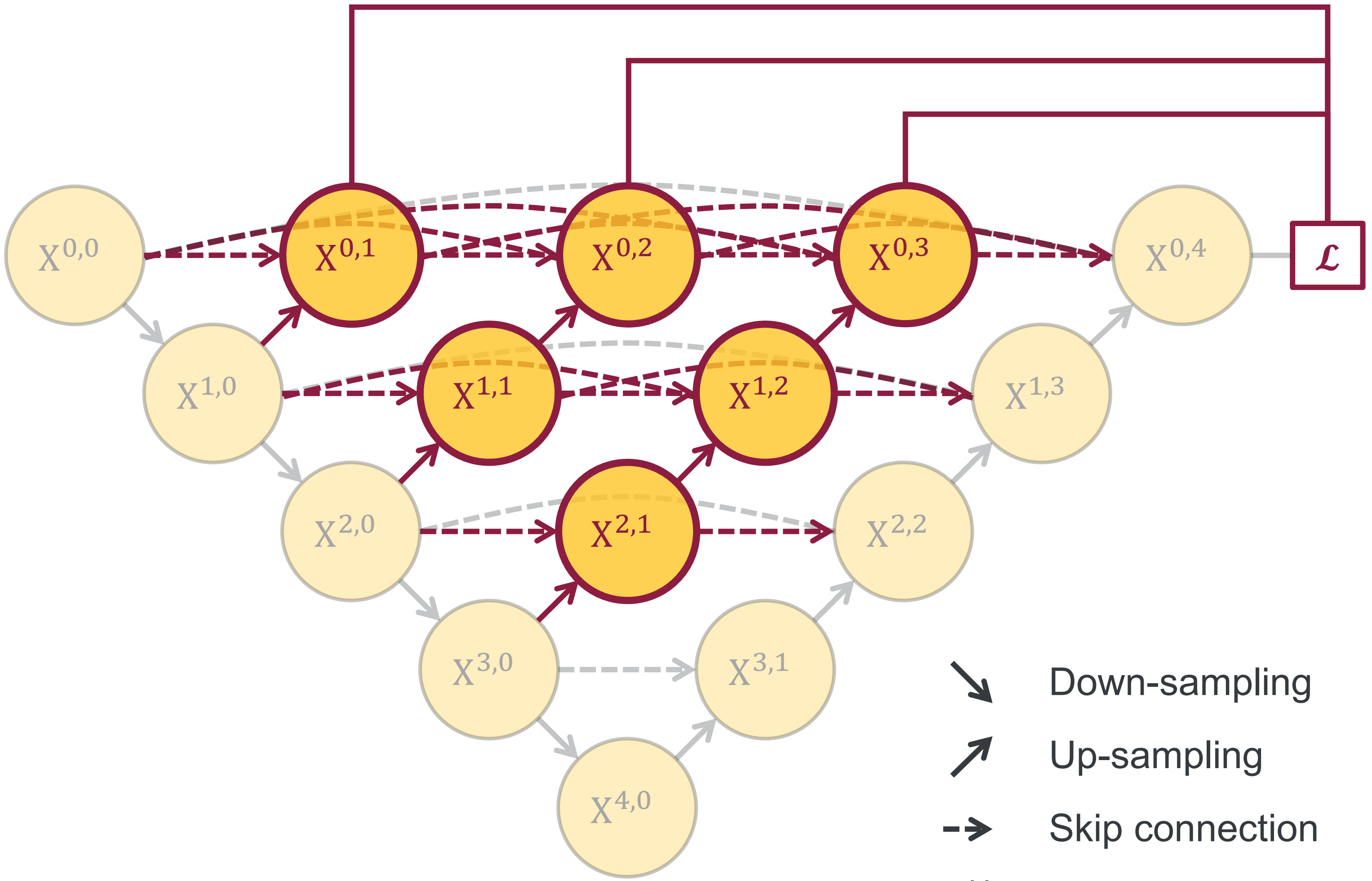




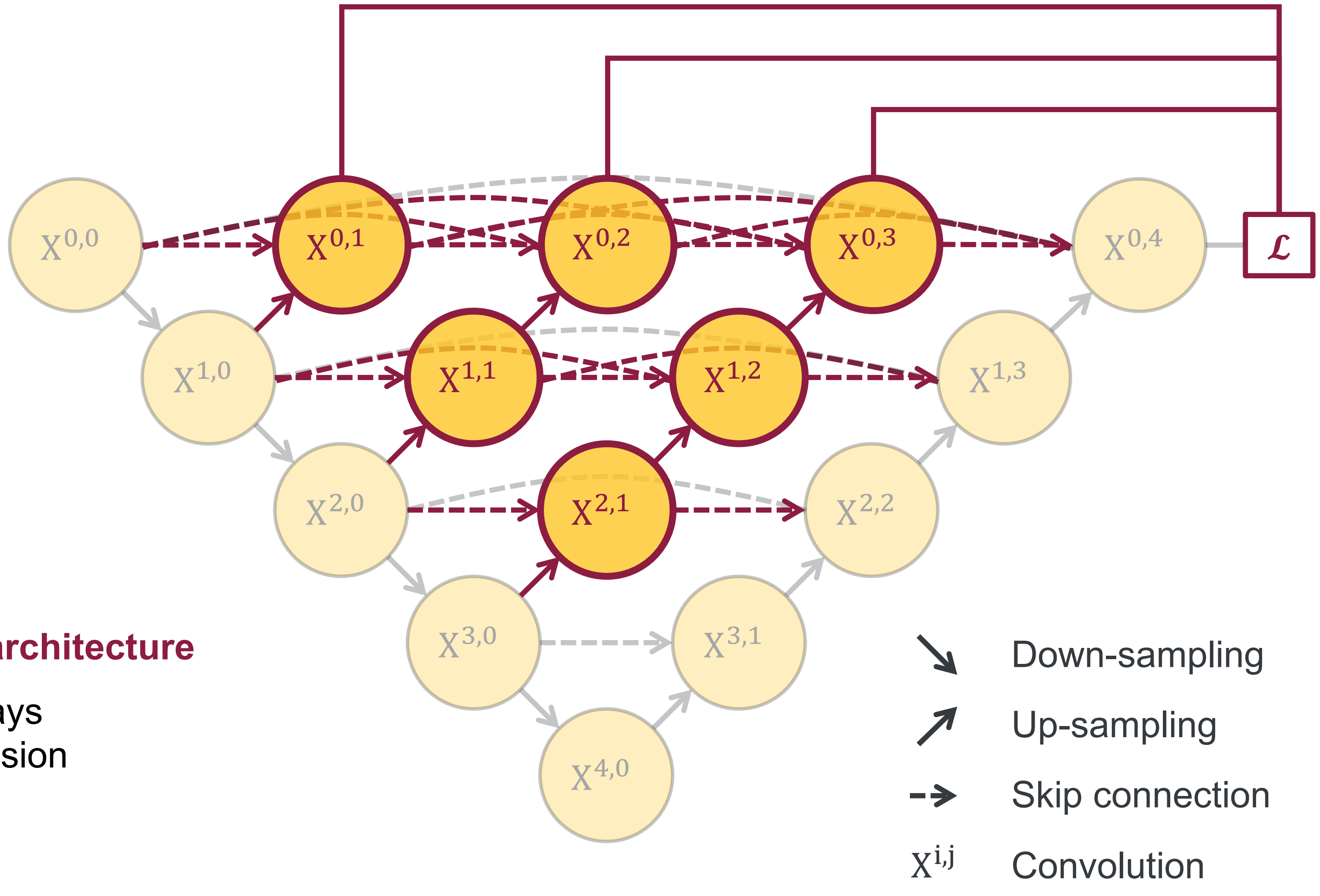
The output of $X^{0,0}$



The output of $X^{0,0}$

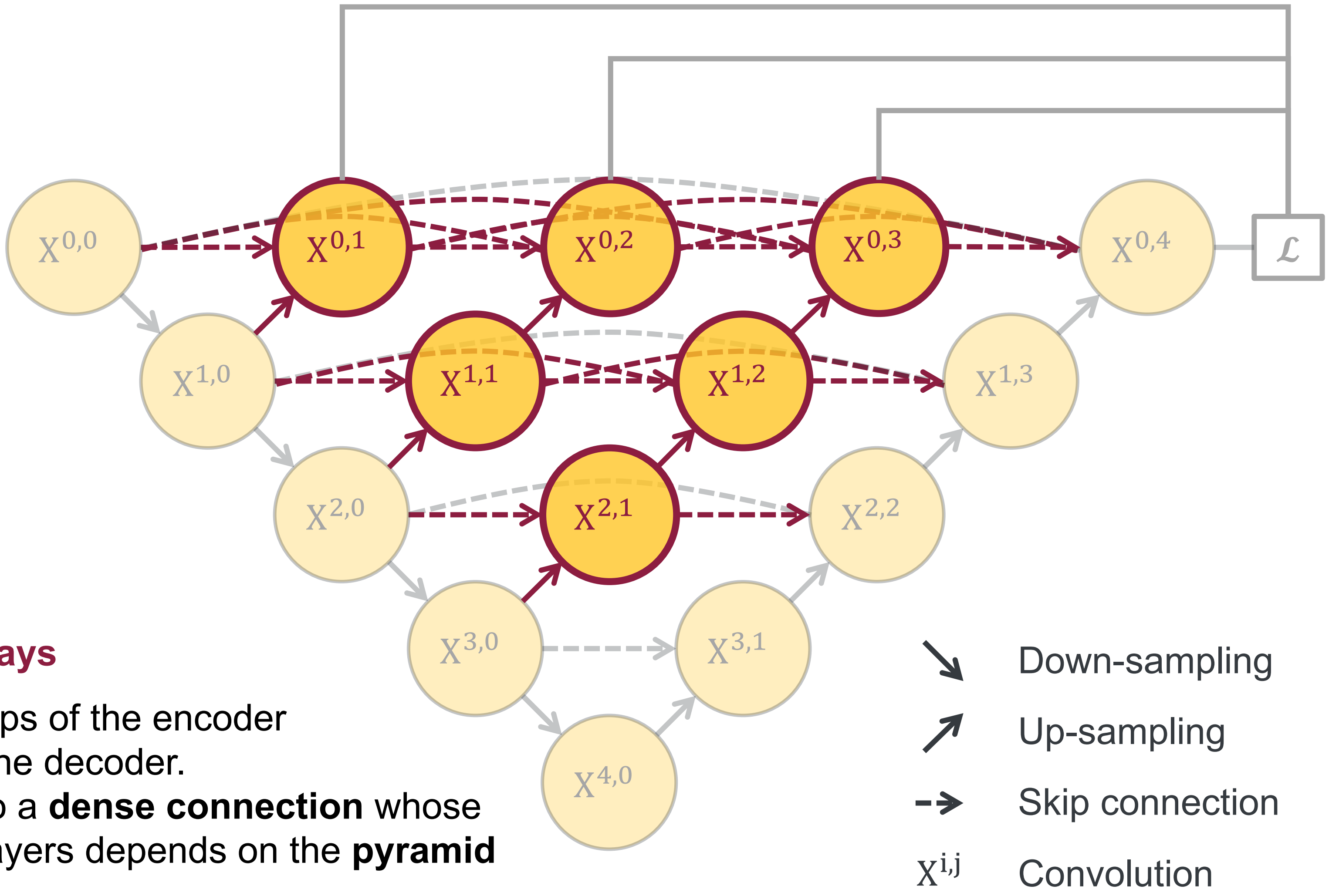


$X^{0,0}$
$X^{0,1}$
$X^{0,2}$
$X^{0,3}$
$X^{0,4}$
$\mathcal{L}$
$X^{1,0}$
$X^{1,1}$
$X^{1,2}$
$X^{1,3}$
$X^{2,0}$
$X^{2,1}$
$X^{2,2}$
$X^{3,0}$
$X^{3,1}$
$X^{4,0}$
Down-sampling
Up-sampling
Skip connection
$X^{i,j}$ Convolution



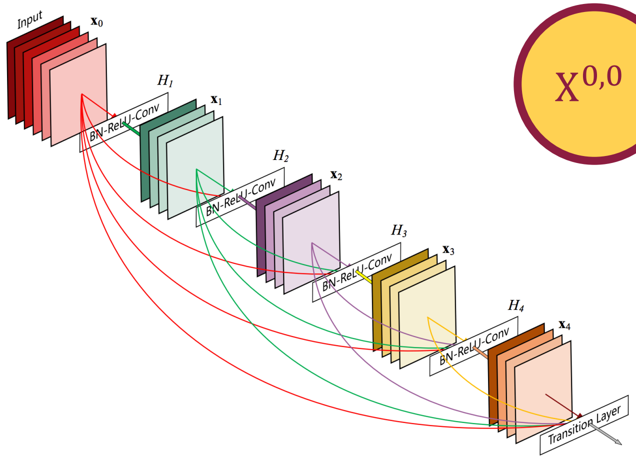
The novelties of UNet++ architecture

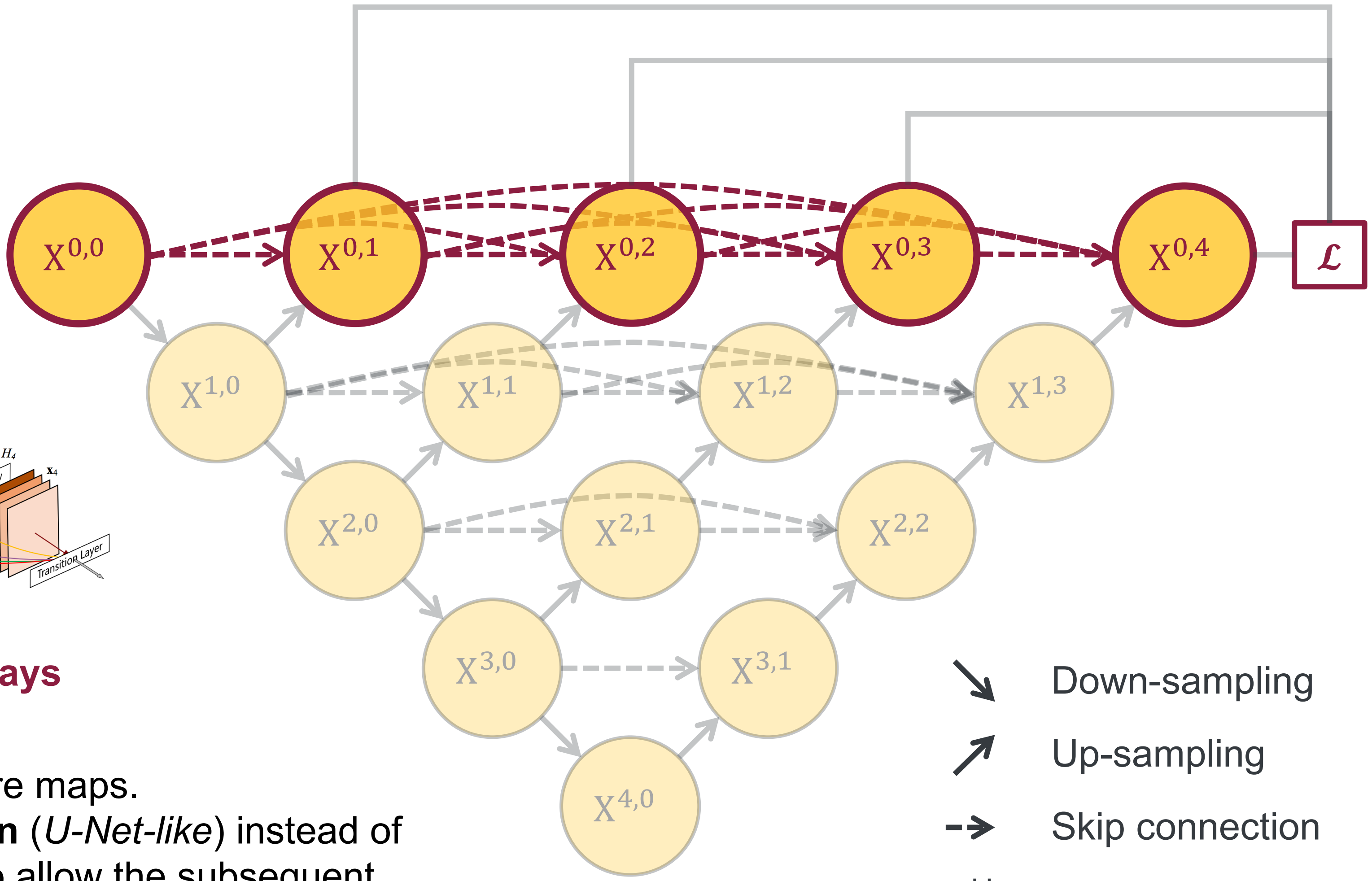
- Redesigned skip pathways
- The use of deep supervision



Re-design the skip pathways

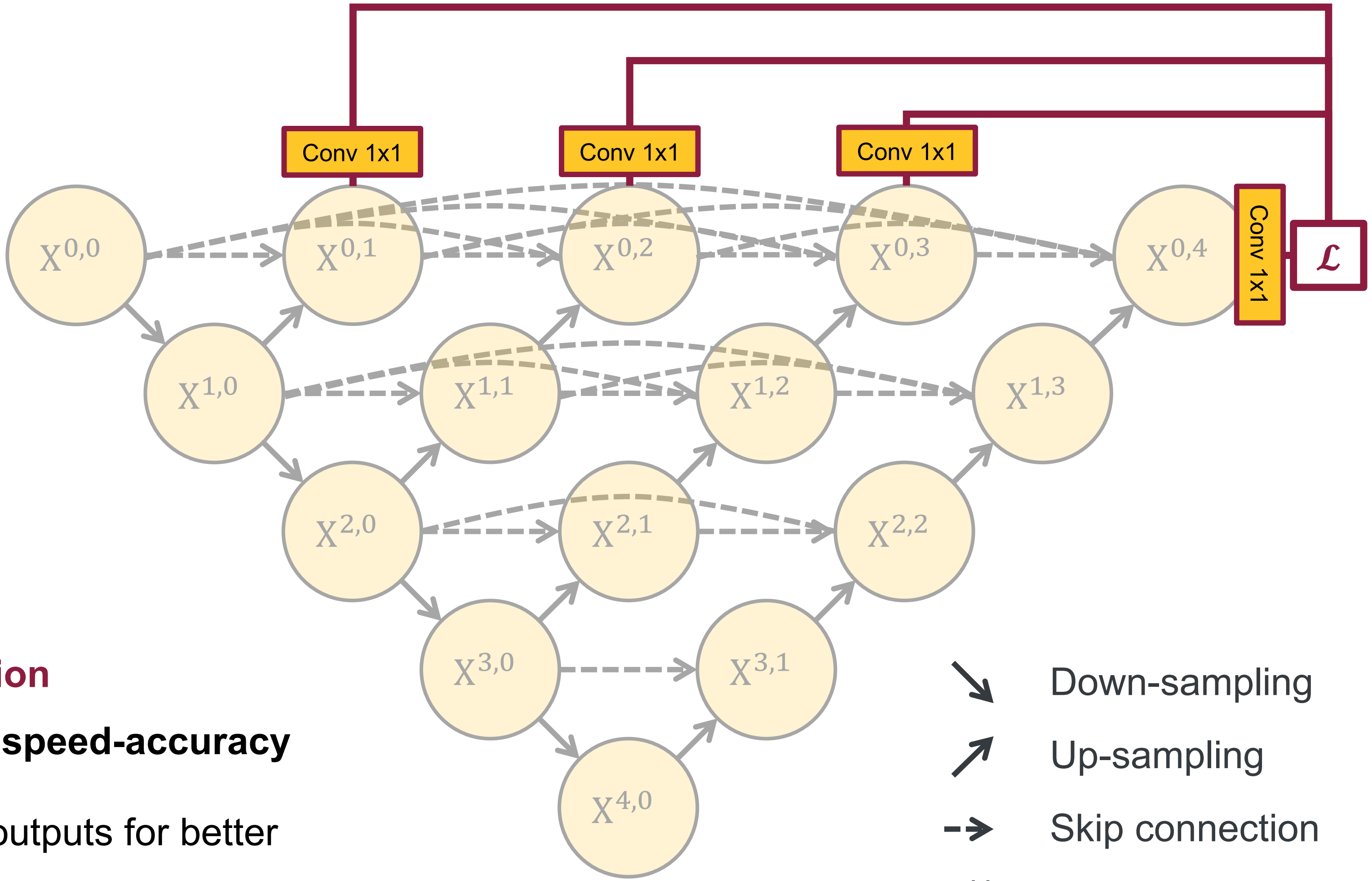
- In U-Net, the feature maps of the encoder are directly received in the decoder.
- In UNet++, they undergo a **dense connection** whose number of convolution layers depends on the **pyramid level**.





Re-design the skip pathways

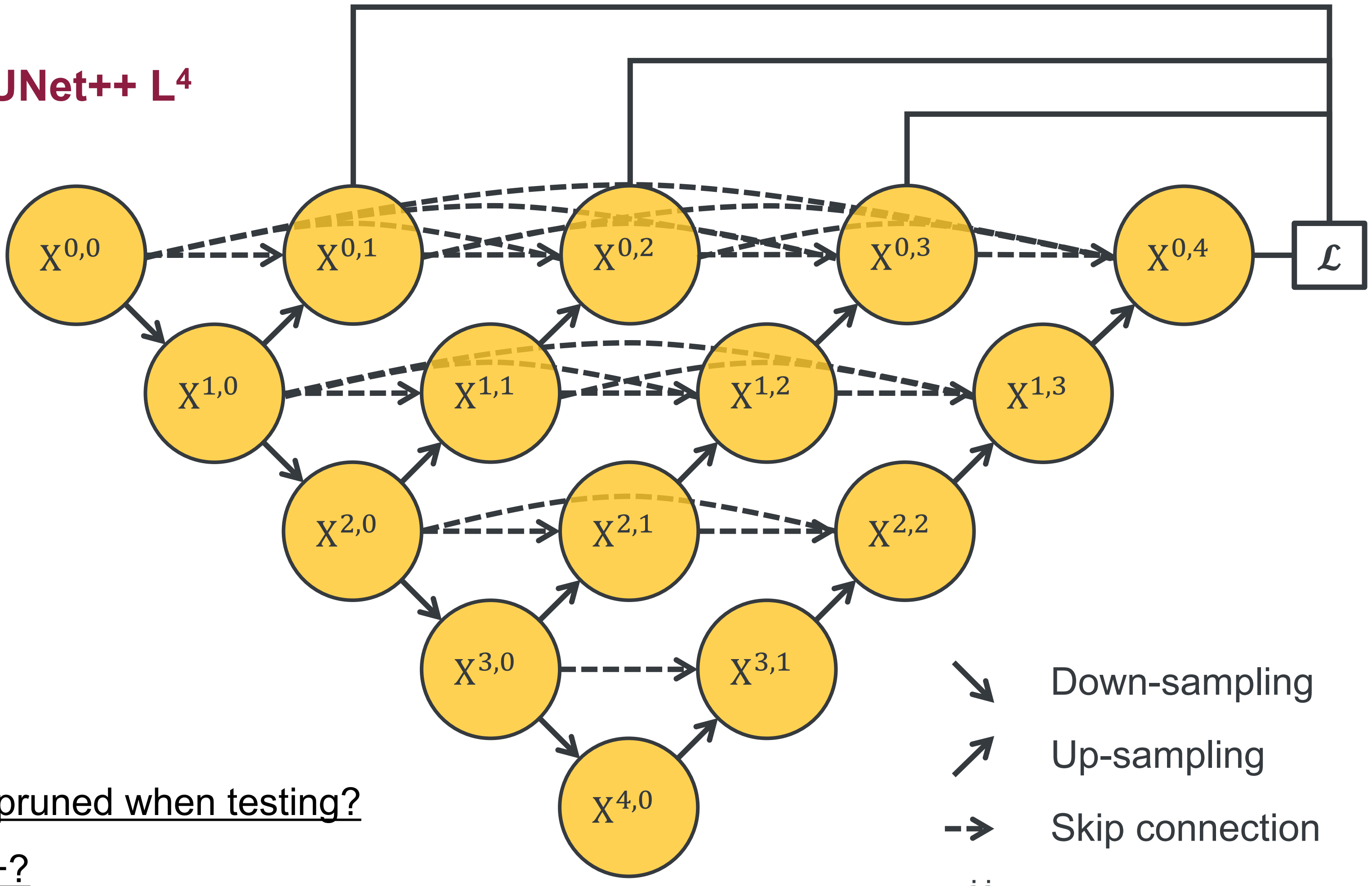
- Increase feature **fusion**.
- **Densely connect** feature maps.
- We apply **concatenation** (*U-Net-like*) instead of summation (*FCN-like*) to allow the subsequent layers to **re-use** intermediate representations.



The use of deep supervision

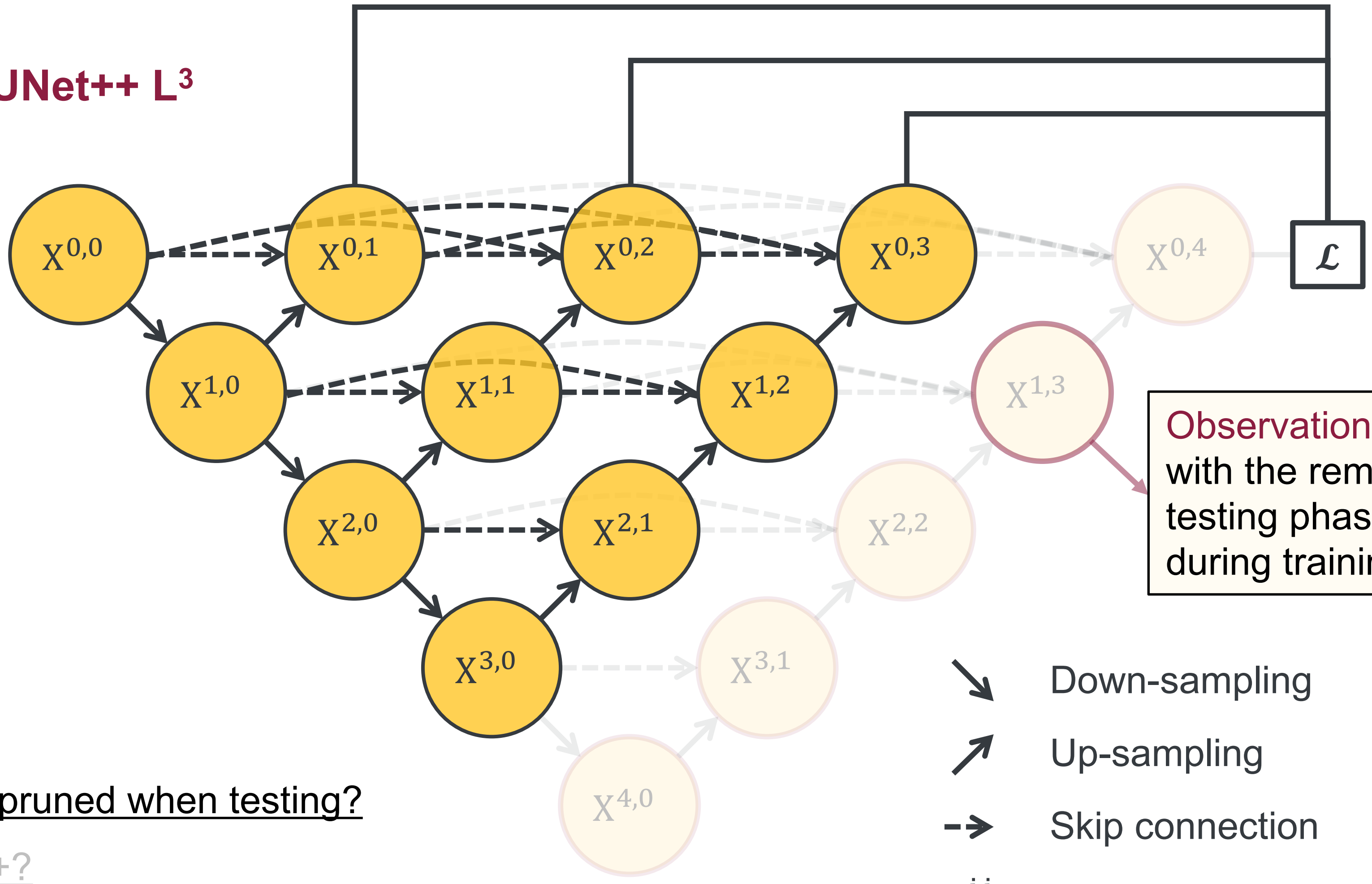
- Allow model pruning via **speed-accuracy** trade-off.
- Ensemble **multi-depth** outputs for better accuracy.

UNet++ L^4



- Q: Why UNet++ can be pruned when testing?
- Q: How to prune UNet++?
- Q: What's the benefit?

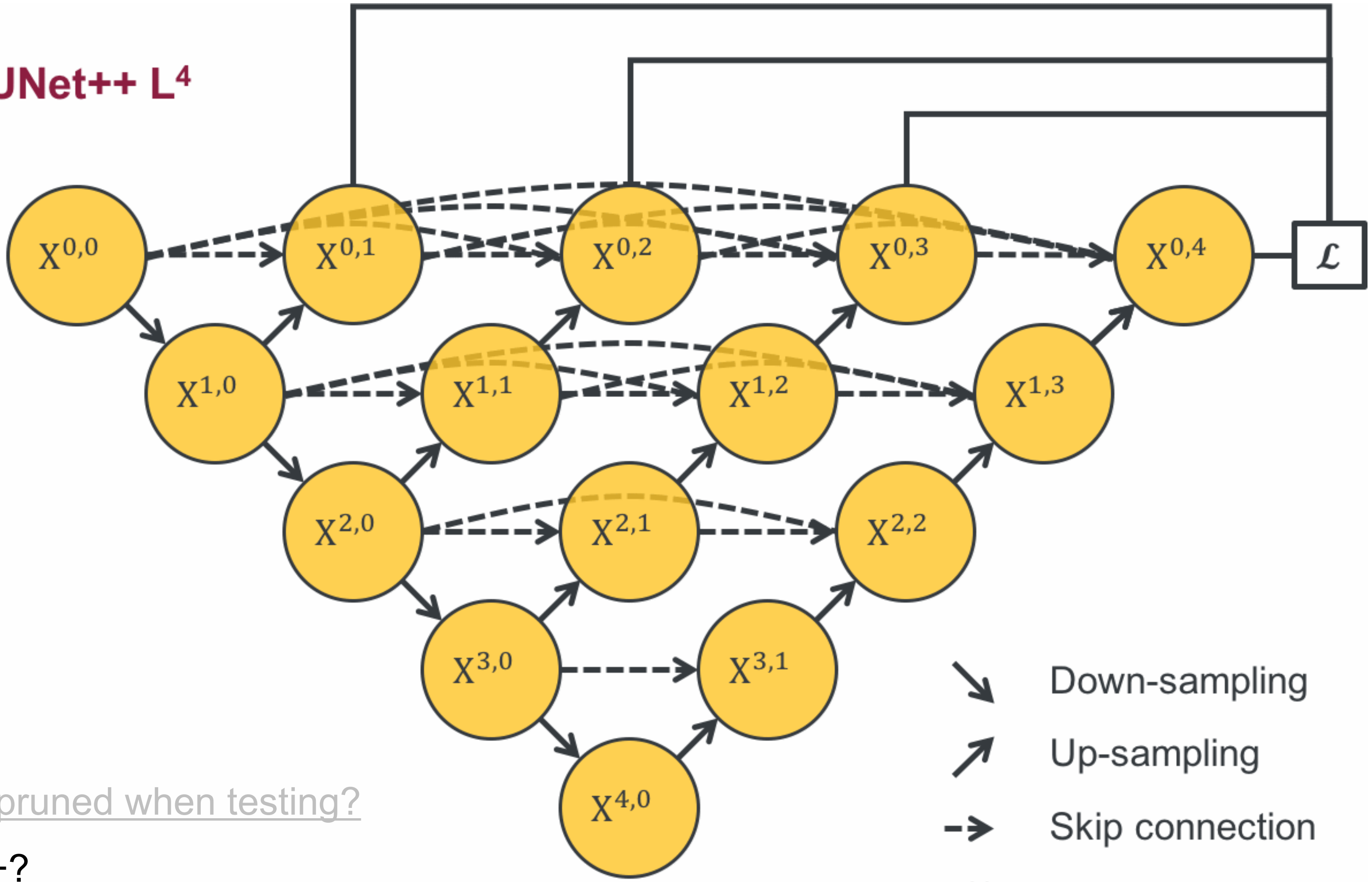
UNet++ L^3



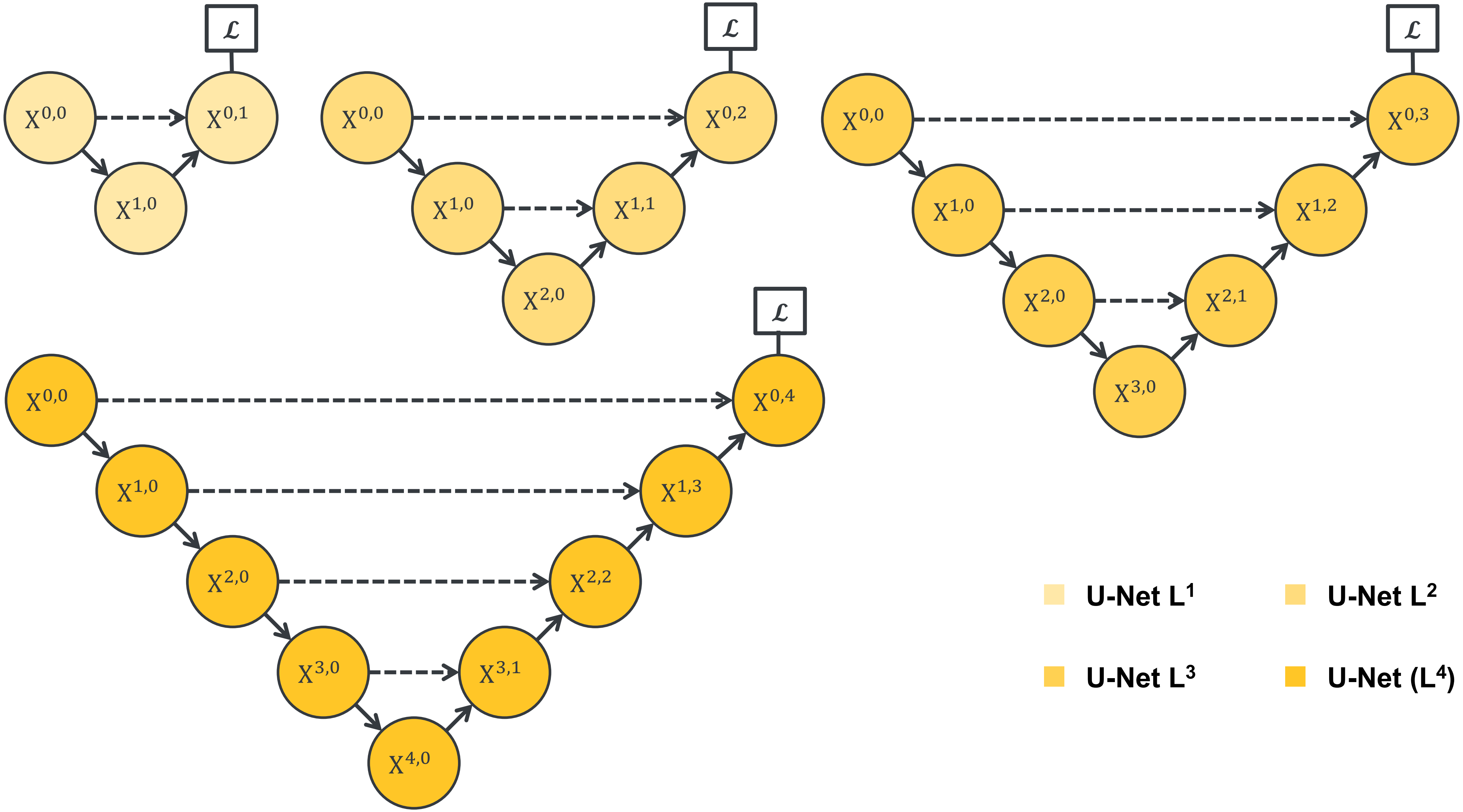
Observation: It is **independent** with the remaining parts during testing phase, but **contributes** during training phase.

- Why UNet++ can be pruned when testing?
- How to prune UNet++?
- What's the benefit?

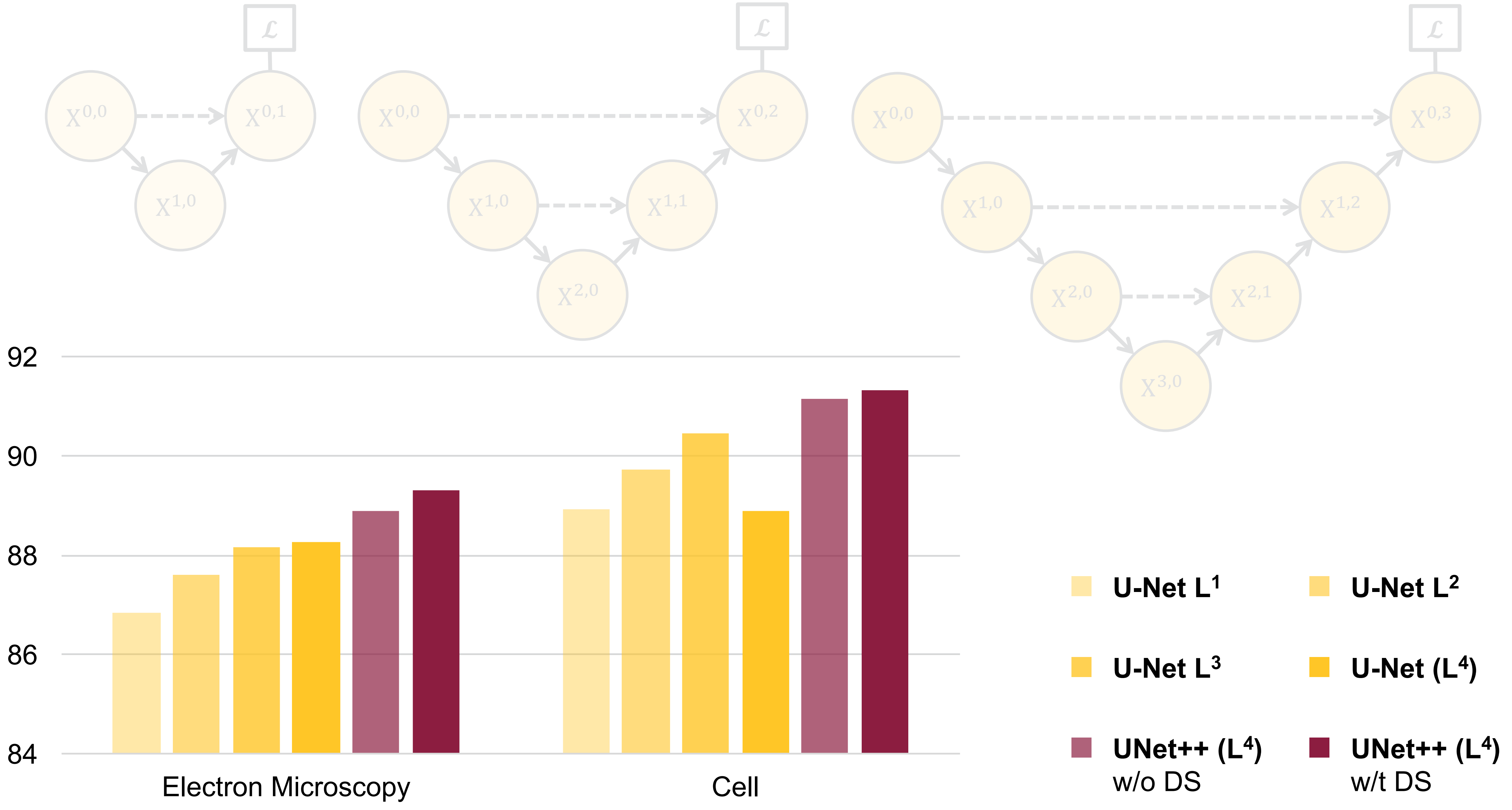
UNet++ L4



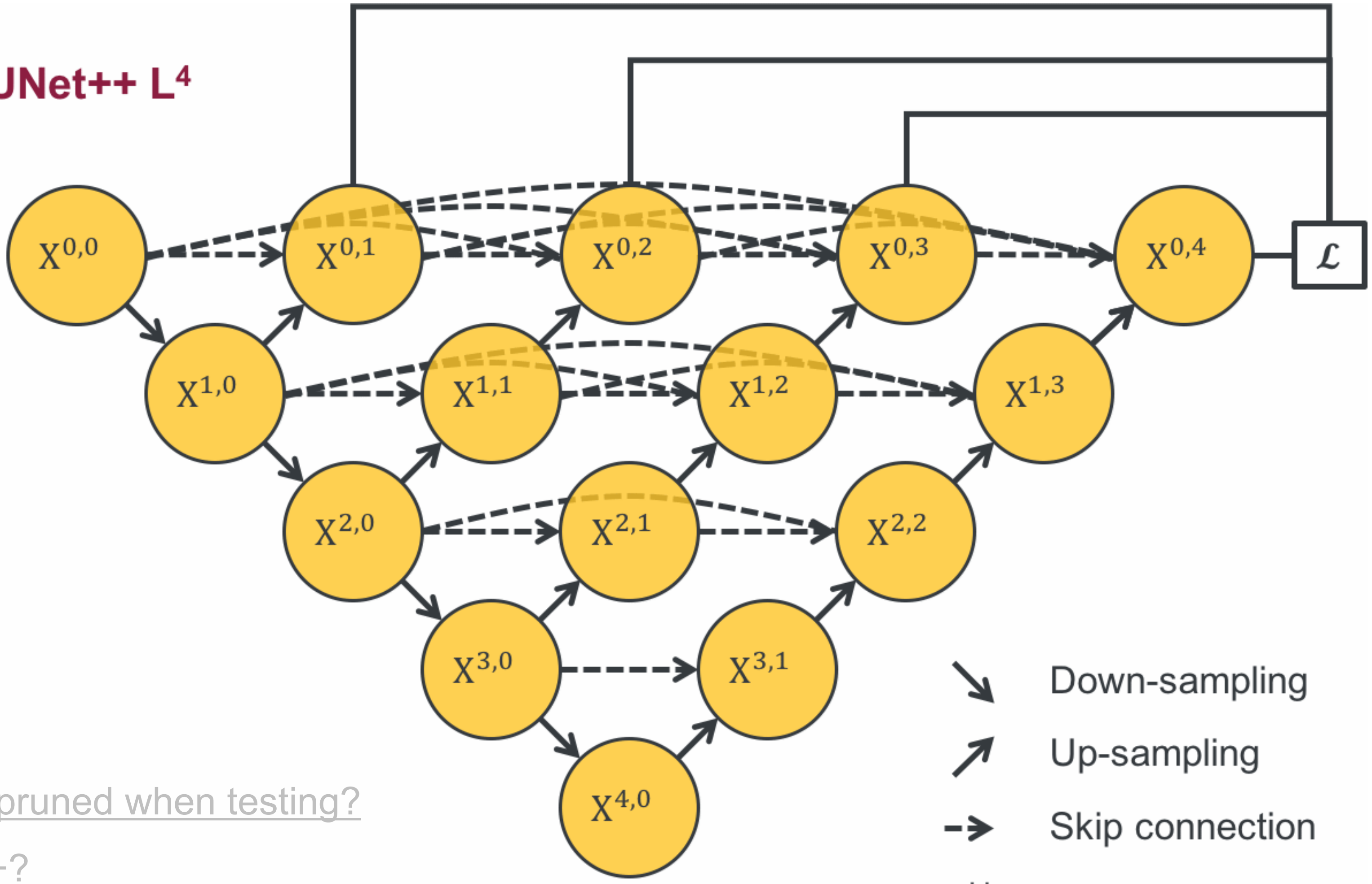
- Q: Why UNet++ can be pruned when testing?
- Q: How to prune UNet++?
- Q: What's the benefit?



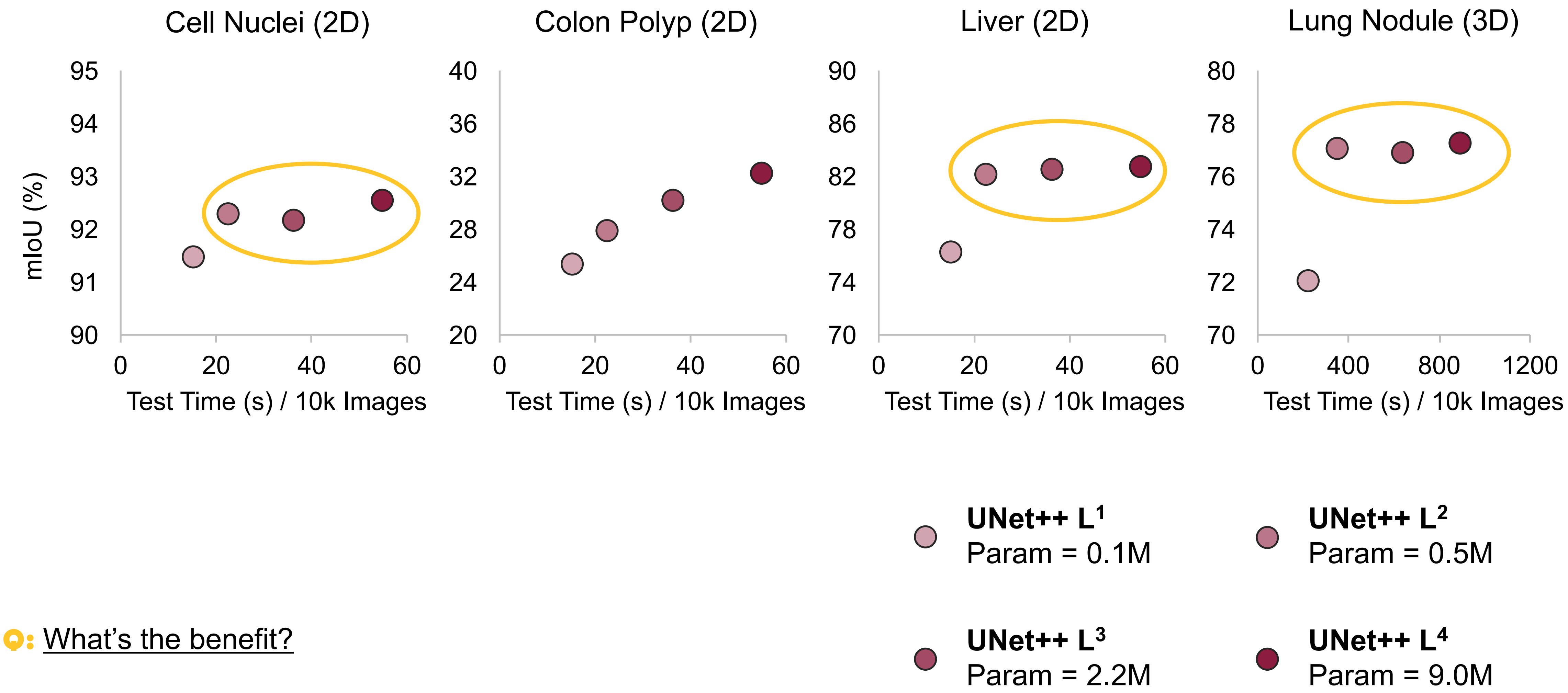
ℒ
$X^{0,0}$
$X^{0,1}$
$X^{1,0}$
ℒ
$X^{0,0}$
$X^{0,2}$
$X^{1,0}$
$X^{1,1}$
$X^{2,0}$
ℒ
$X^{0,0}$
$X^{0,3}$
$X^{1,0}$
$X^{1,2}$
$X^{2,0}$
$X^{2,1}$
$X^{3,0}$
ℒ
$X^{0,0}$
$X^{0,4}$
$X^{1,0}$
$X^{1,3}$
$X^{2,0}$
$X^{2,2}$
$X^{3,0}$
$X^{3,1}$
$X^{4,0}$
U-Net L^1
U-Net L^2
U-Net L^3
U-Net (L^4)


92
90
88
86
84
Electron Microscopy
Cell
U-Net L1
U-Net L2
U-Net L3
U-Net (L4)
UNet++ (L4) w/o DS
UNet++ (L4) w/t DS

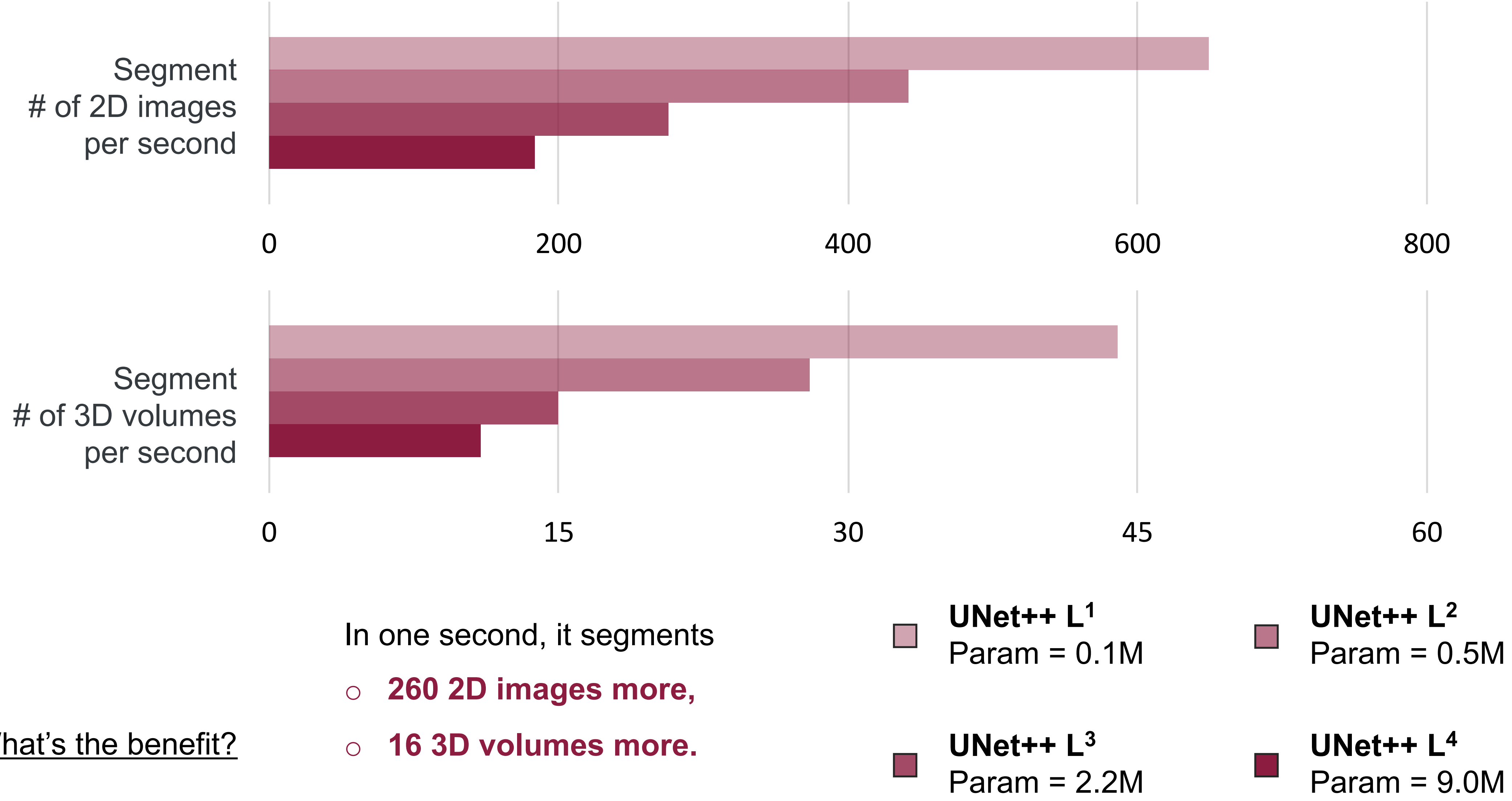
UNet++ L^4



- Q: Why UNet++ can be pruned when testing?
- Q: How to prune UNet++?
- Q: What's the benefit?



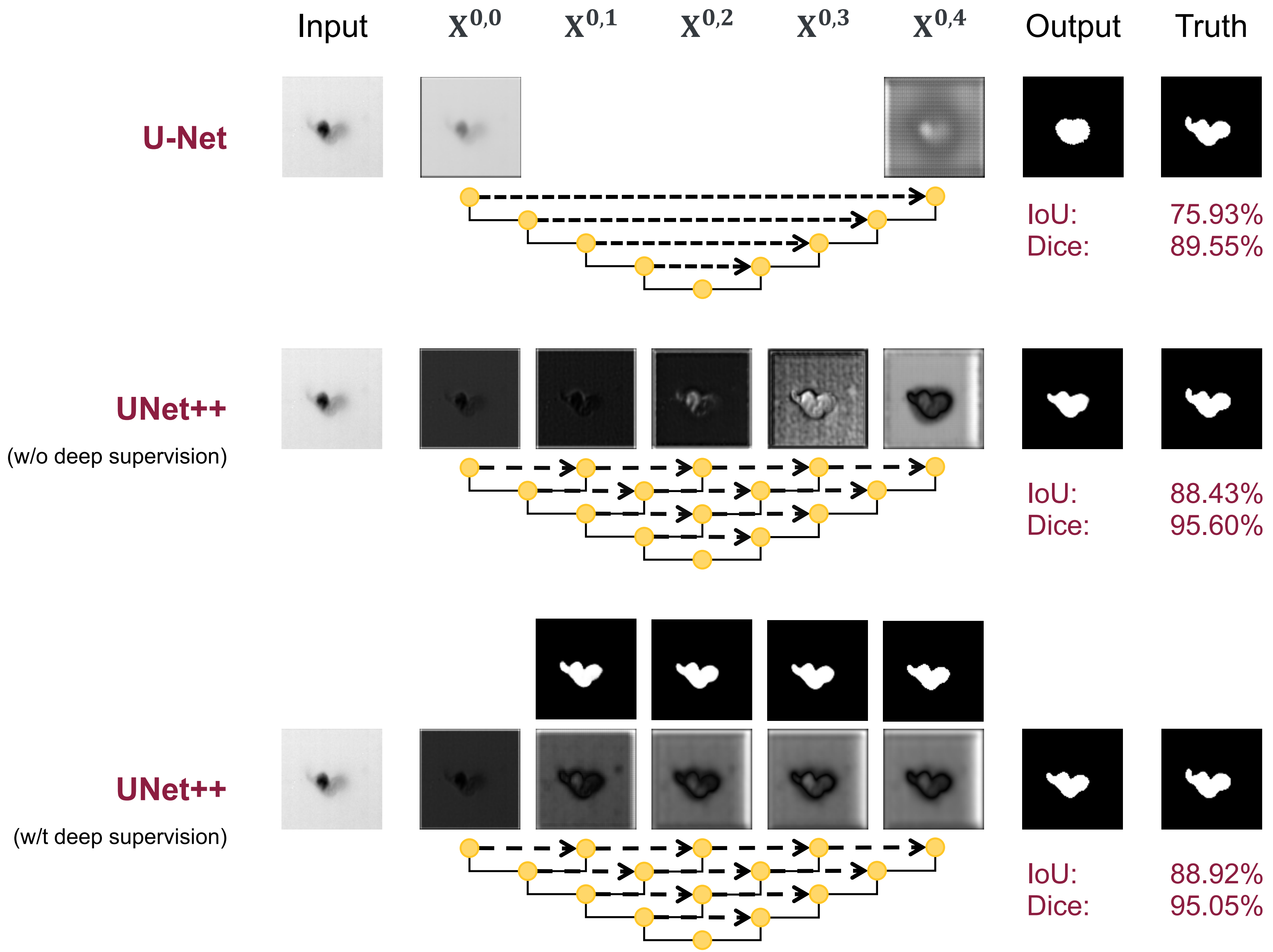
Q: What's the benefit?

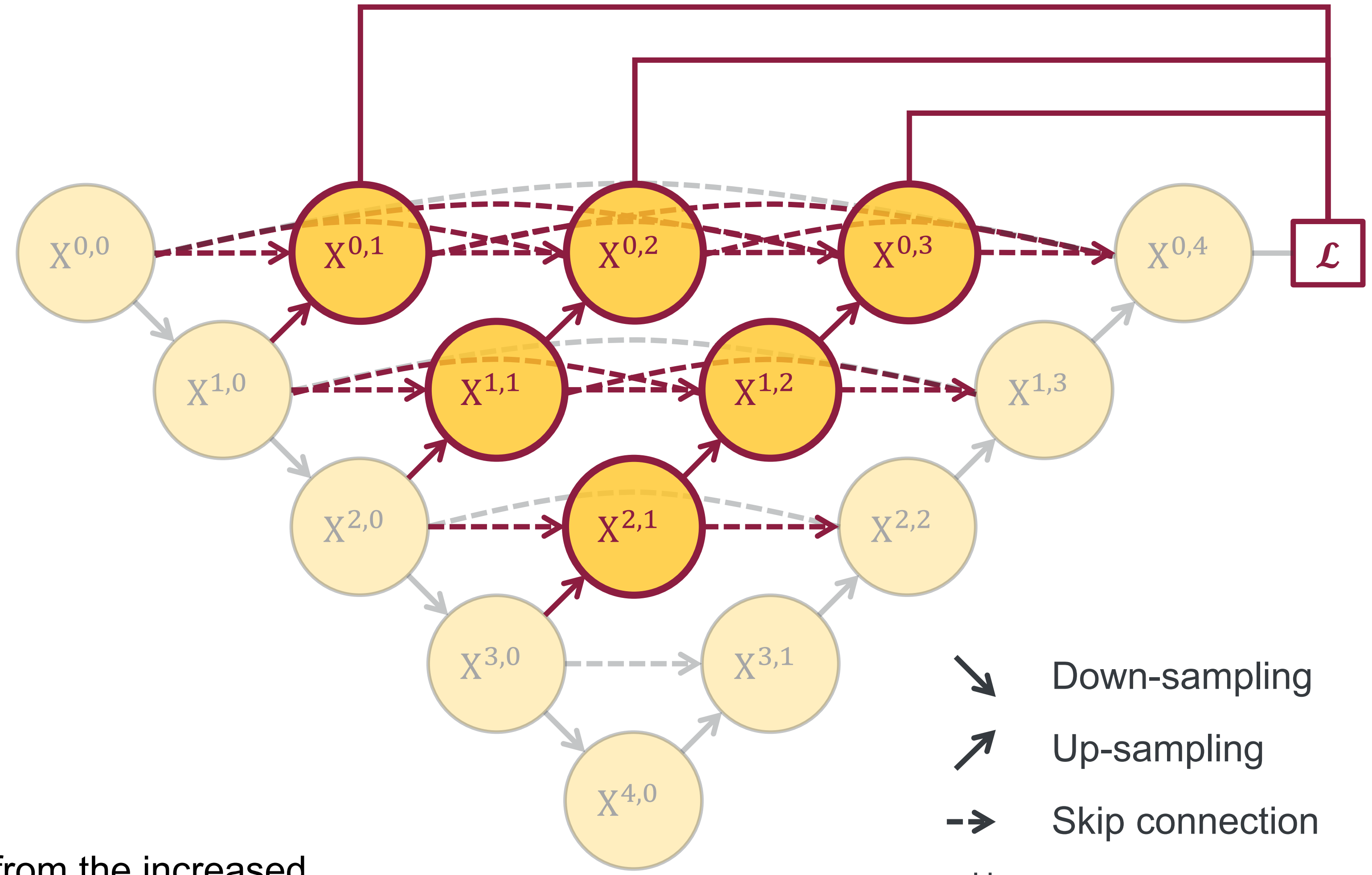


In one second, it segments

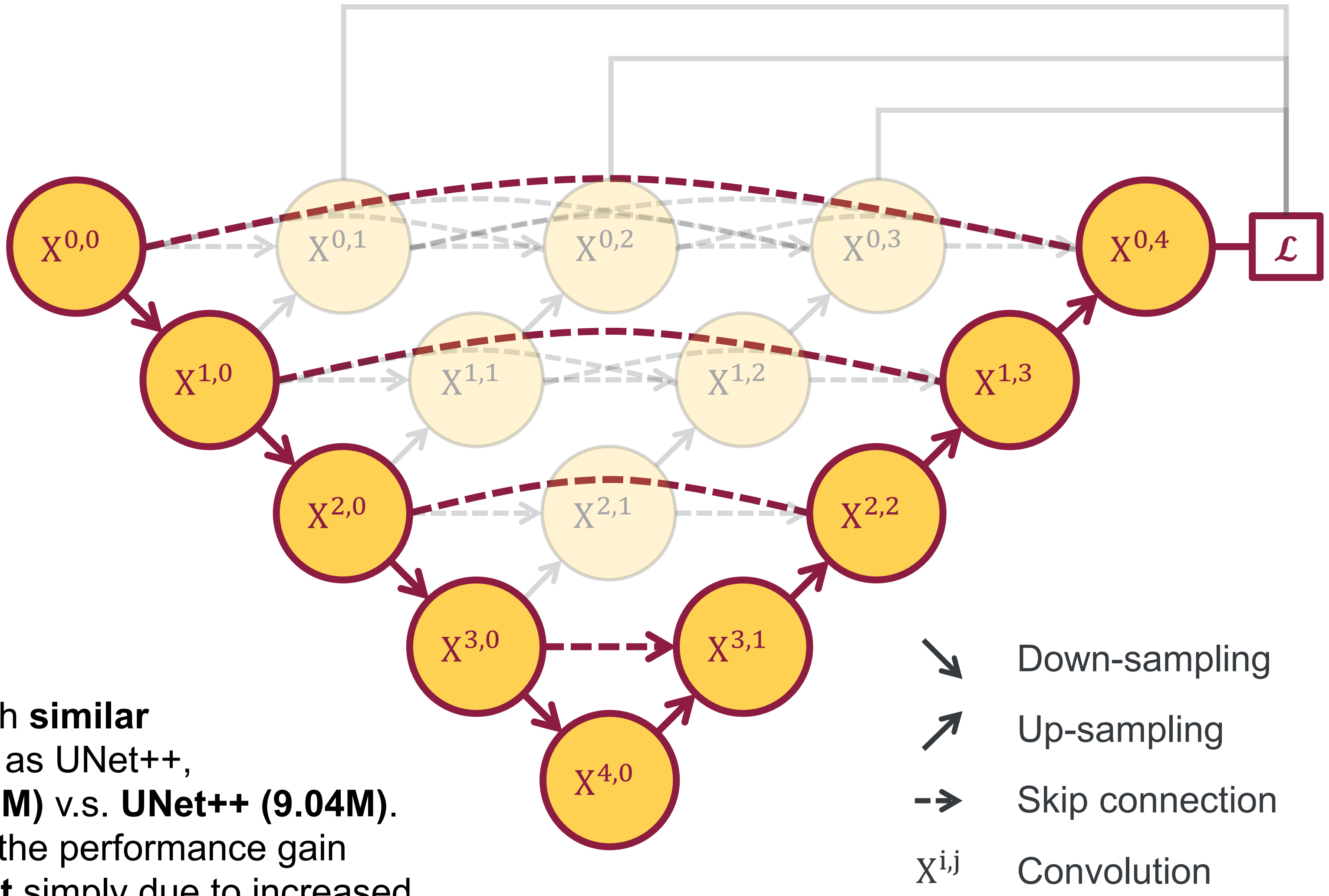
- **260 2D images more,**
- **16 3D volumes more.**

Q: <u>What's the benefit?</u>


Input
$X^{0,0}$
$X^{0,1}$
$X^{0,2}$
$X^{0,3}$
$X^{0,4}$
Output
Truth
U-Net
IoU: 75.93%
Dice: 89.55%
UNet++
(w/o deep supervision)
IoU: 88.43%
Dice: 95.60%
UNet++
(w/t deep supervision)
IoU: 88.92%
Dice: 95.05%



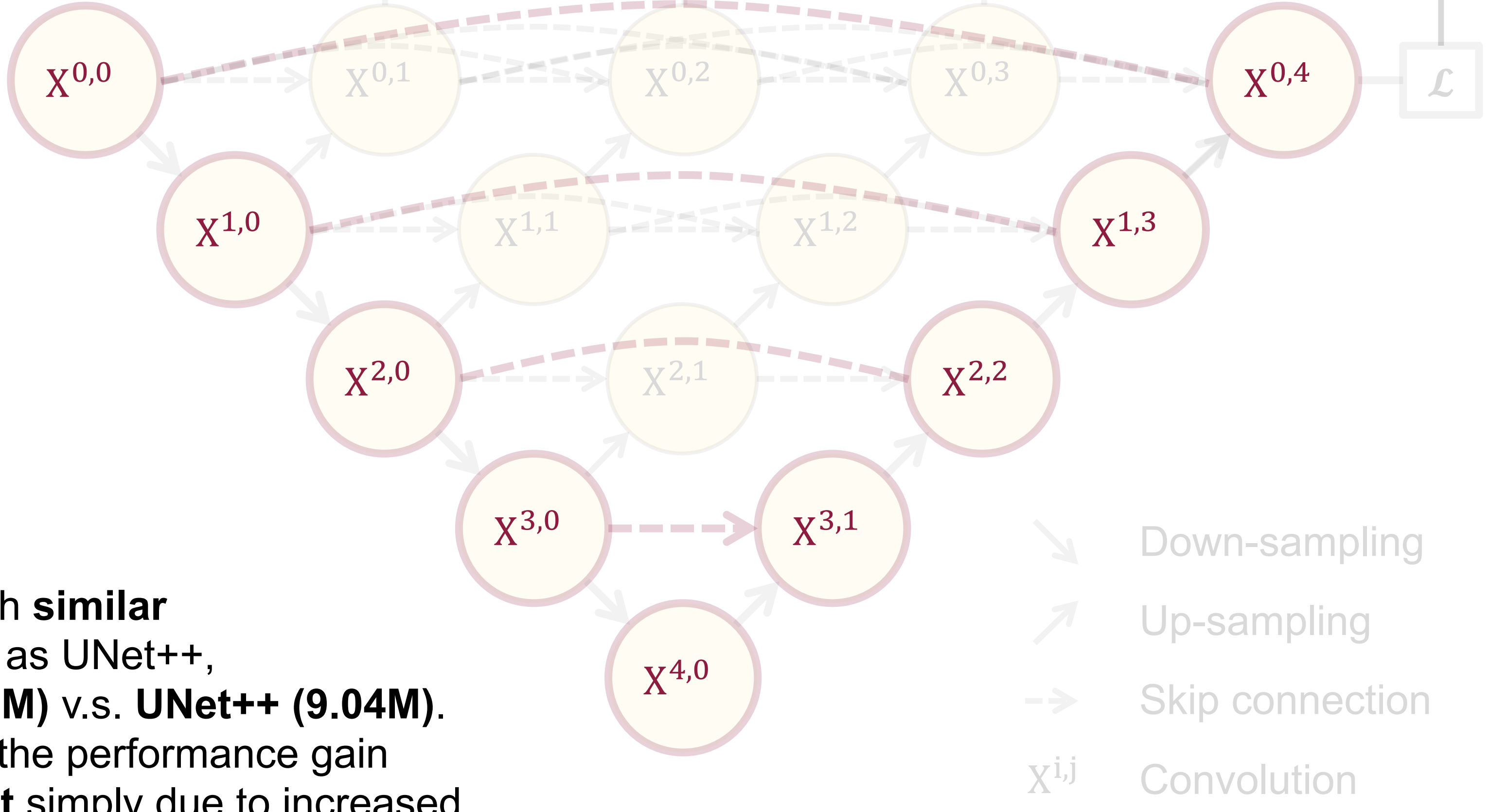
Q: Does the benefit come from the increased number of parameters? That is, U-Net (7.76M) vs. UNet++ (9.04M)?



"Wide" U-Net

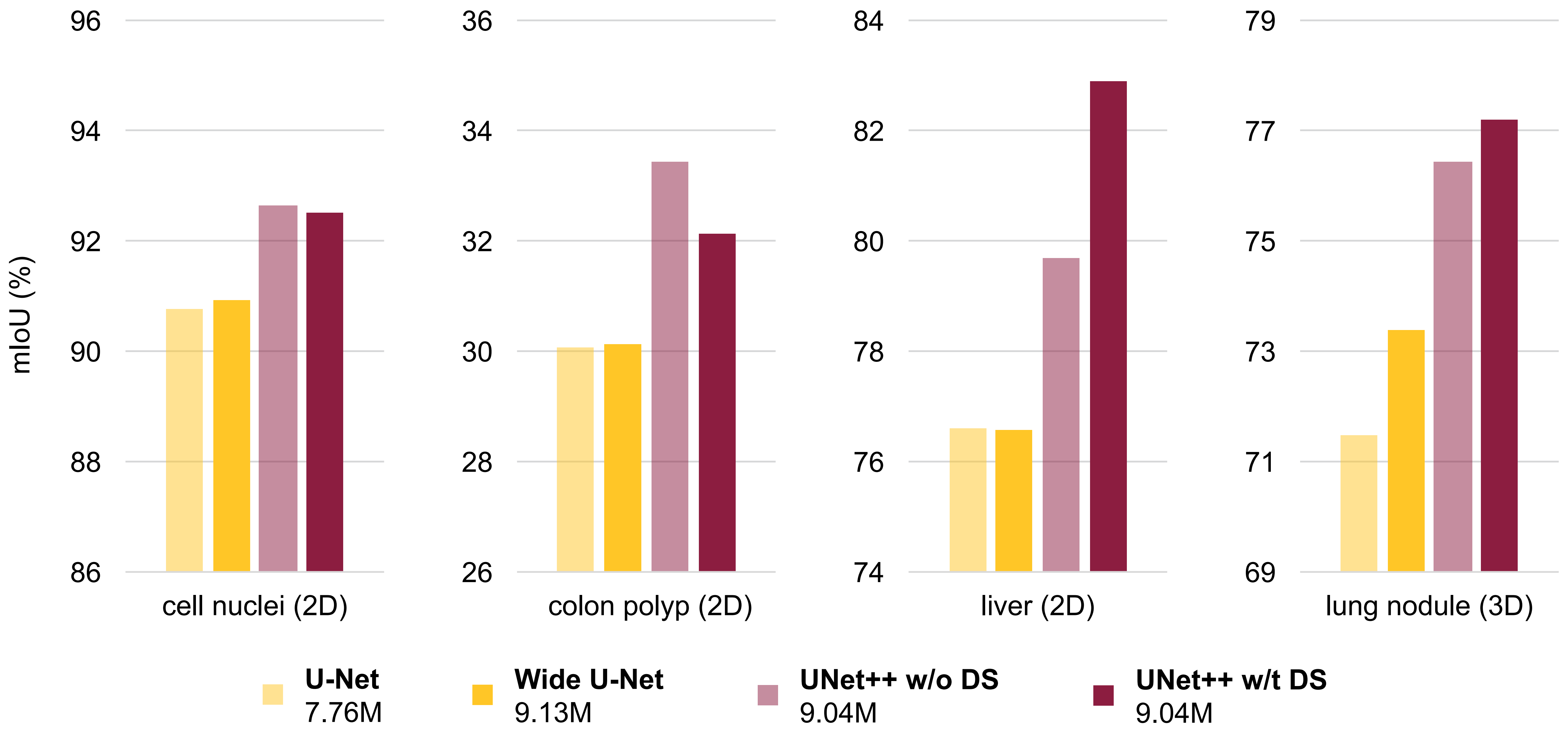
- Design a wide U-Net with **similar number of parameters** as UNet++, where **wide U-Net (9.13M)** v.s. **UNet++ (9.04M)**.
- This was to ensure that the performance gain yielded by UNet++ is **not** simply due to increased number of parameters.

encoder / decoder	$X^{0,0}$ / $X^{0,4}$	$X^{1,0}$ / $X^{1,3}$	$X^{2,0}$ / $X^{2,2}$	$X^{3,0}$ / $X^{3,1}$	$X^{4,0}$ / $X^{4,0}$
U-Net **[7.76M]**	32	64	128	256	512
"Wide" U-Net **[9.13M]**	35	70	140	280	560
UNet++ **[9.04M]**	32	64	128	256	512

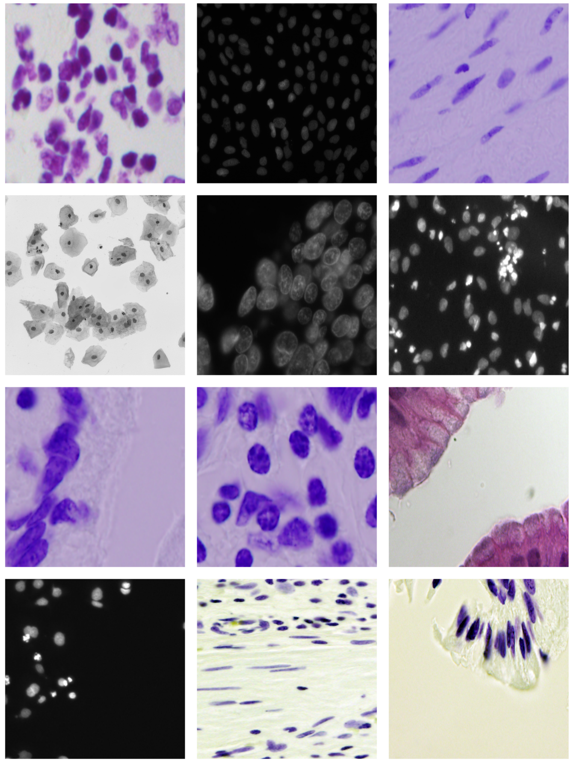


"Wide" U-Net

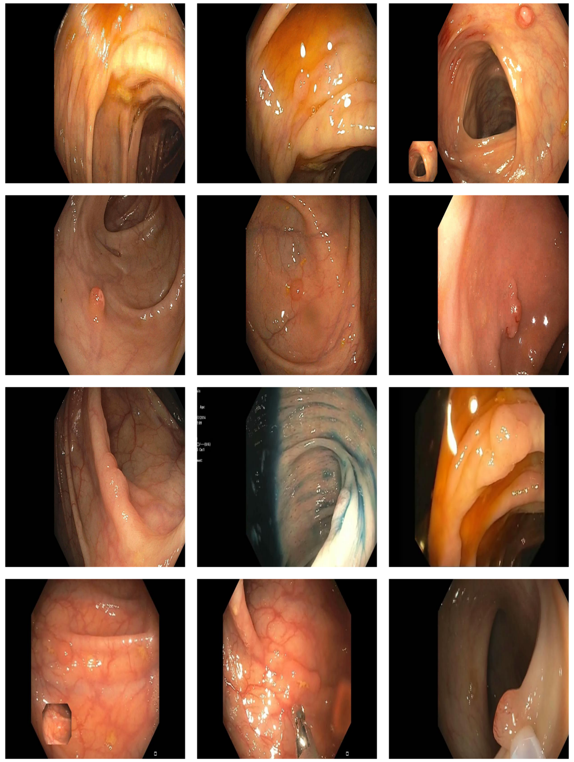
- Design a wide U-Net with **similar number of parameters** as UNet++, where **wide U-Net (9.13M)** v.s. **UNet++ (9.04M)**.
- This was to ensure that the performance gain yielded by UNet++ is **not** simply due to increased number of parameters.


mIoU (%)
96
94
92
90
88
86
cell nuclei (2D)
36
34
32
30
28
26
colon polyp (2D)
84
82
80
78
76
74
liver (2D)
79
77
75
73
71
69
lung nodule (3D)
U-Net 7.76M
Wide U-Net 9.13M
UNet++ w/o DS 9.04M
UNet++ w/t DS 9.04M

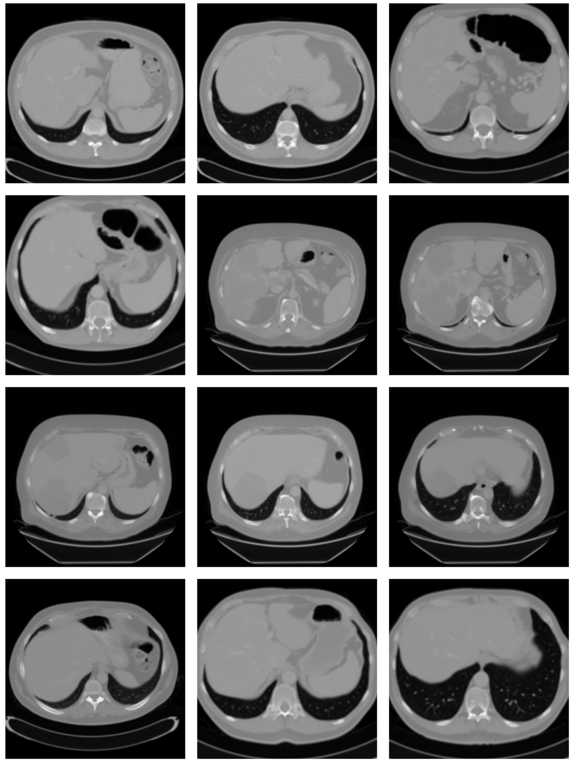
Dataset	Images	Input Size	Modality	Provider
cell nuclei	670	96×96	microscopy	Data Science Bowl 2018
colon polyp	7,379	224×224	RGB video	ASU-Mayo
liver	331	512×512	CT	MICCAI 2018 LiTS Challenge
lung nodule	1,012	64×64×64	CT	LIDC-IDRI



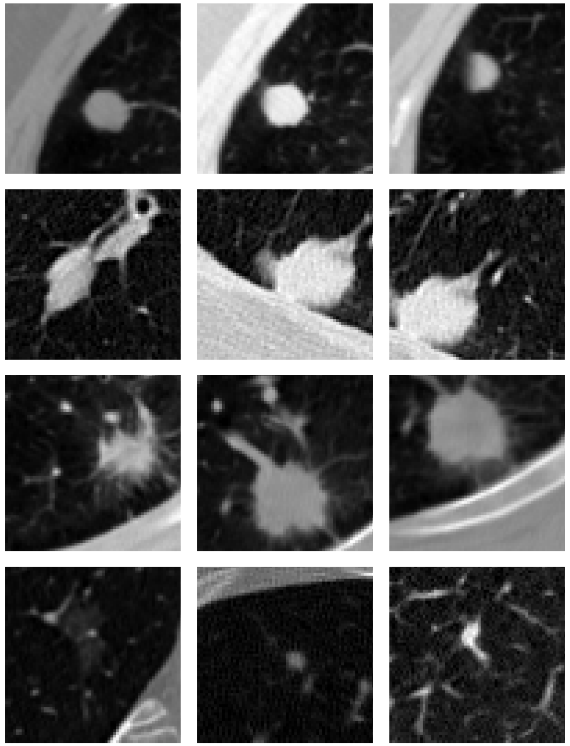
Cell Nuclei



Colon Polyp



Liver



Lung Nodule

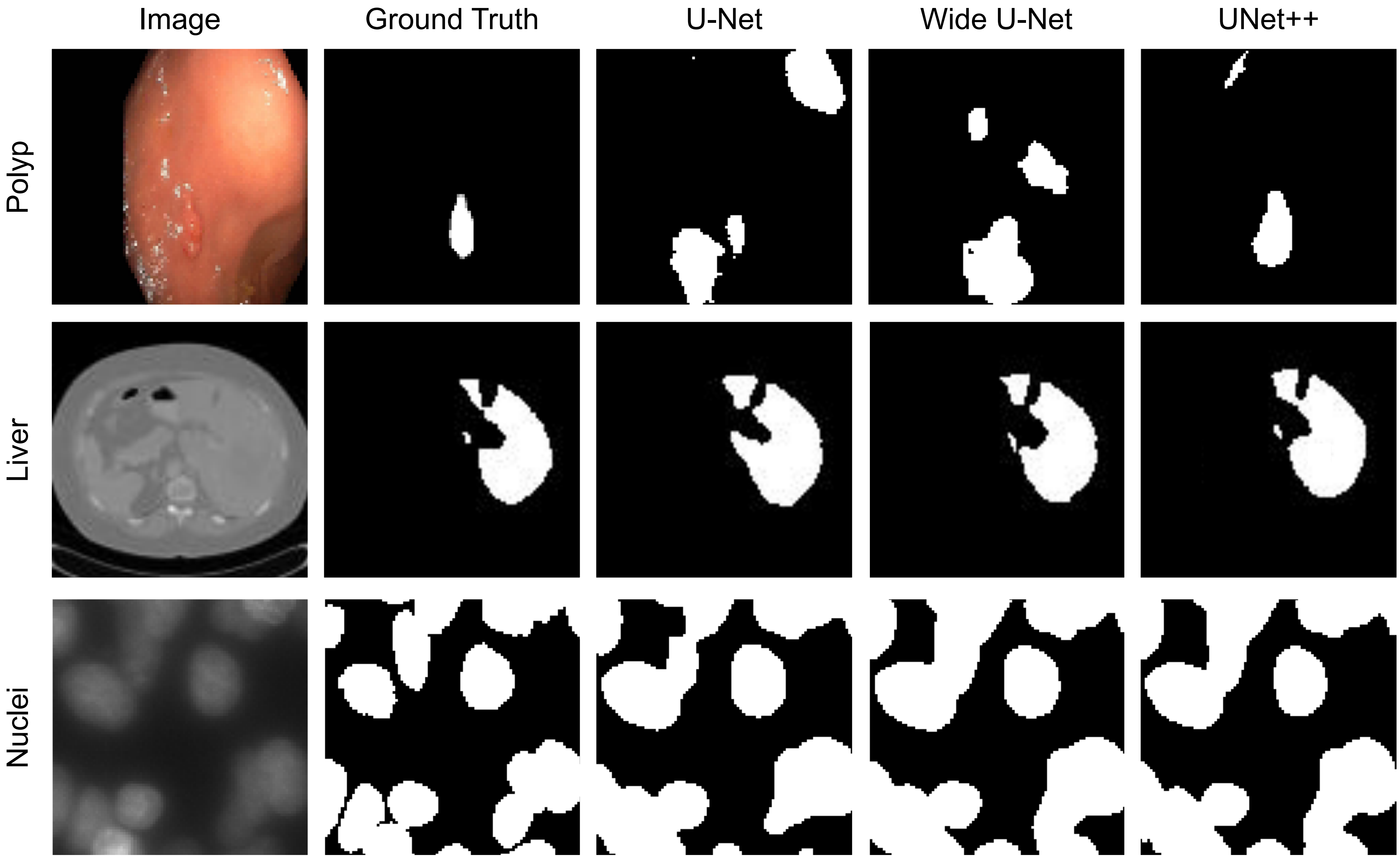


Image
Ground Truth
U-Net
Wide U-Net
UNet++
Polyp
Liver
Nuclei

Take Home Messages

- The re-designed skip pathways aim at reducing the **semantic gap** between the feature maps of the encoder and decoder sub-network.
- Deep supervision exposes the model to flexible **accuracy-speed** trade-off.

Zongwei Zhou

He received the BSc degree with honors in Computer Science from Dalian University of Technology in 2016. I'm currently a PhD student in the Department of Biomedical Informatics, Arizona State University reported to Dr. Jianming Liang. His research interests lie predominately in the area of Computer Vision, Deep Learning, and Medical Image Analysis.

Mahfuzur Rahman Siddiquee

He received the BSc degree with honors in Computer Science from North South University in 2015. He is currently pursuing my PhD in Computer Science at Arizona State University. His PhD advisor is Dr. Jianming Liang. His research interest is in Deep Learning and Medical Image Analysis.

Nima Tajbakhsh

He is a research scientist with keen interest in computer vision, machine learning, and medical imaging. He has developed methods based on deep learning and hand-crafted features for detecting structures/lesions in CT scans, mammograms, digital histopathology images, ultrasound videos, and colonoscopy videos, resulting in 12 issued U.S. patents and several pending applications.

Jianming Liang

He is an Associate Professor at Arizona State University. Drawing upon computer vision, machine learning, visualization, and mathematics, his research focuses on developing computational methodologies for addressing a profound challenge facing biomedicine: image data explosion, through a multidisciplinary team-based approach. In addition to his 70+ peer-reviewed publications, he holds 50 US patents and patents pending. He received an ASU President's Award for Innovation.

- ❑ **Paper** https://arxiv.org/abs/1807.10165
- ❑ **Code** https://github.com/MrGiovanni/Nested-UNet
- ❑ **Contact** zongweiz@asu.edu; mrahmans@asu.edu

Thanks