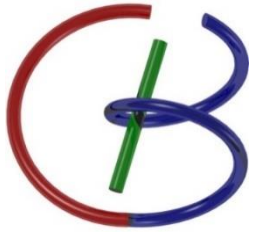
Rensselaer



UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation

Zongwei Zhou, *Member, IEEE,* Md Mahfuzur Rahman Siddiquee, *Member, IEEE,*
Nima Tajbakhsh, *Member, IEEE,* and Jianming Liang, *Senior Member, IEEE*

UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation
Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang
Arizona State University
IEEE Transactions on Medical Imaging (TMI)
paper | code

UNet++: A Nested U-Net Architecture for Medical Image Segmentation
Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang
Arizona State University
Deep Learning in Medical Image Analysis (DLMIA) 2018. **(Oral)**
paper | code | slides | poster | blog

Chuang Niu, May 6, 2020

Problems of encoder-decoder architectures

- The encoder-decoder networks are widely used in modern semantic and instance segmentation models. Their success is largely attributed to their skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder with shallow, low-level, fine-grained feature maps from the encoder.

- The encoder-decoder architectures for image segmentation have two limitations:

(1) The optimal depth of an encoder-decoder network can vary from one application to another, depending on the task difficulty and the amount of labeled data available for training.

(2) The design of skip connections used in an encoder-decoder network is unnecessarily restrictive, demanding the fusion of the same-scale encoder and decoder feature maps. The same-scale feature maps from the decoder and encoder networks are semantically dissimilar and no solid theory guarantees that they are the best match for feature fusion.

Architecture: from U-Net to UNet++

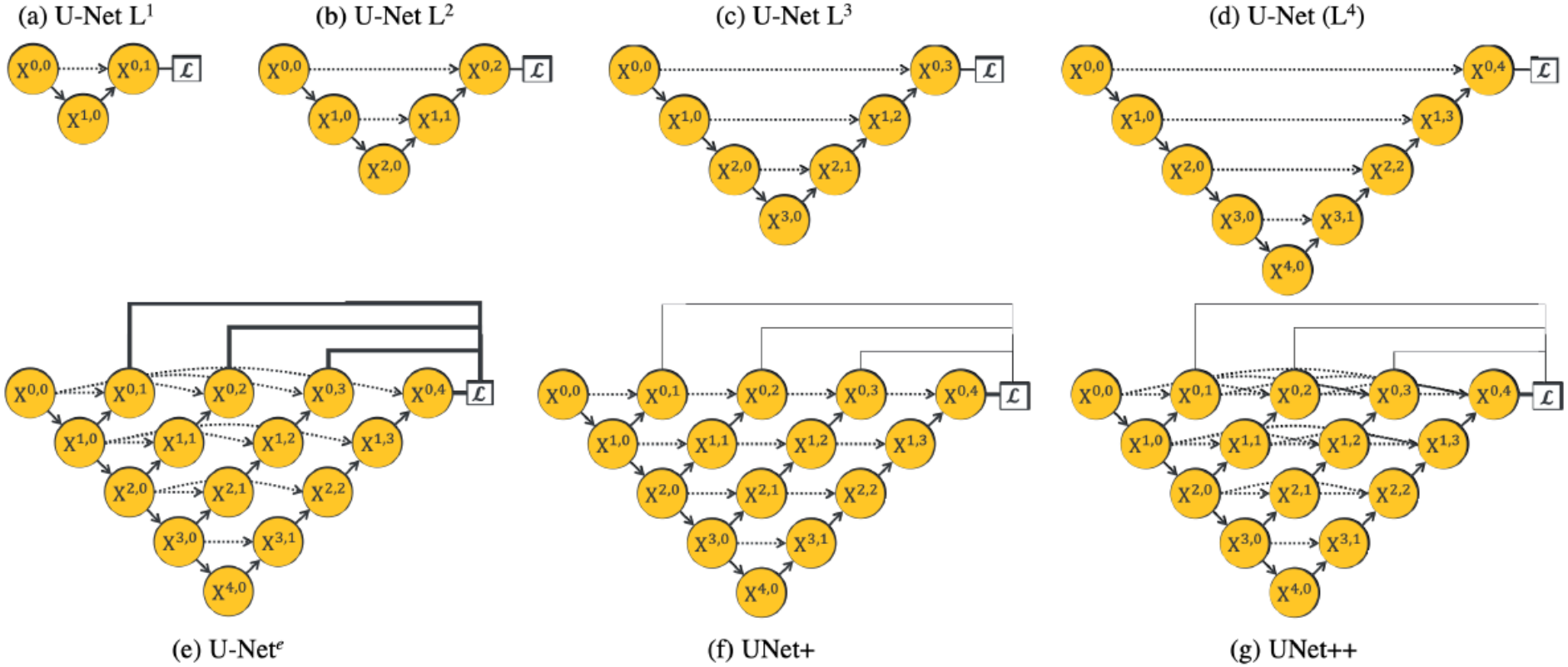


Fig. 1: Evolution from U-Net to UNet++. Each node in the graph represents a convolution block, downward arrows indicate down-sampling, upward arrows indicate up-sampling, and dot arrows indicate skip connections. (a–d) U-Nets of varying depths. (e) Ensemble architecture, U-Net^e, which combines U-Nets of varying depths into one unified architecture. All U-Nets (partially) share the same encoder, but have their own decoders. (f) UNet+ is constructed from U-Net^e by dropping the original skip connections and connecting every two adjacent nodes with a short skip connection, enabling the deeper decoders to send supervision signals to the shallower decoders. (g) UNet++ is constructed from U-Net^e by connecting the decoders, resulting in densely connected skip connections, enabling dense feature propagation along skip connections and thus more flexible feature fusion at the decoder nodes. As a result, each node in the UNet++ decoders, from a horizontal perspective, combines multiscale features from its all preceding nodes at the same resolution, and from a vertical perspective, integrates multiscale features across different resolutions from its preceding node, as formulated at Eq. 1. This multiscale feature aggregation of UNet++ gradually synthesizes the segmentation, leading to increased accuracy and faster convergence, as evidenced by our empirical results in Section IV. Note that, explicit deep supervision is required (bold links) to train U-Net^e but optional (pale links) for UNet+ and UNet++.

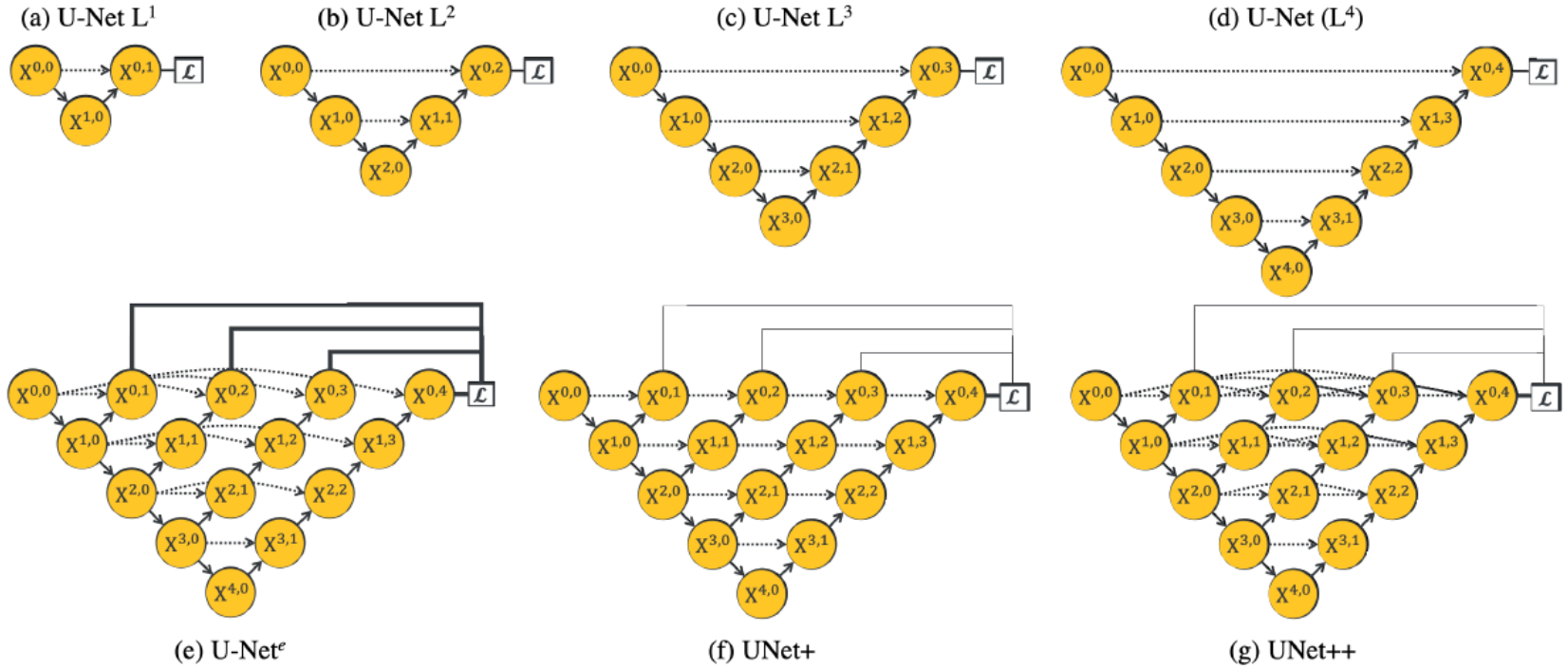
Motivation behind the UNet++

(1) Deeper U-Nets are not necessarily always better.
(2) The optimal depth of architecture depends on the difficulty and size of the dataset at hand.

TABLE I: Ablation study on U-Nets of varying depths alongside with the new variants of U-Nets proposed in this work. U-Net L^d refers to a U-Net with a depth of d (Fig. 1(a-d)). U-Net^e, UNet+, and UNet++ are the new variants of U-Net, which are depicted in Fig. 1(e-g). "DS" denotes deeply supervised training followed by average voting. Intersection over union (IoU) is used as the metric for comparison (mean±s.d. %).

Architecture	DS	Params	EM	Cell	Brain Tumor
U-Net L^1	✗	0.1M	86.83±0.43	88.58±1.68	86.90±2.25
U-Net L^2	✗	0.5M	87.59±0.34	89.39±1.64	88.71±1.45
U-Net L^3	✗	1.9M	88.16±0.29	90.14±1.57	89.62±1.41
U-Net (L^4)	✗	7.8M	88.30±0.24	88.73±1.64	89.21±1.55
U-Net^e	✓	8.7M	88.33±0.23	90.72±1.51	90.19±0.83
UNet+	✗	8.7M	88.39±0.15	90.71±1.25	90.70±0.91
UNet+	✓	8.7M	88.89±0.12	91.18±1.13	91.15±0.65
UNet++	✗	9.0M	88.92±0.14	91.03±1.34	90.86±0.81
UNet++	✓	9.0M	**89.33±0.10**	**91.21±0.98**	**91.21±0.68**

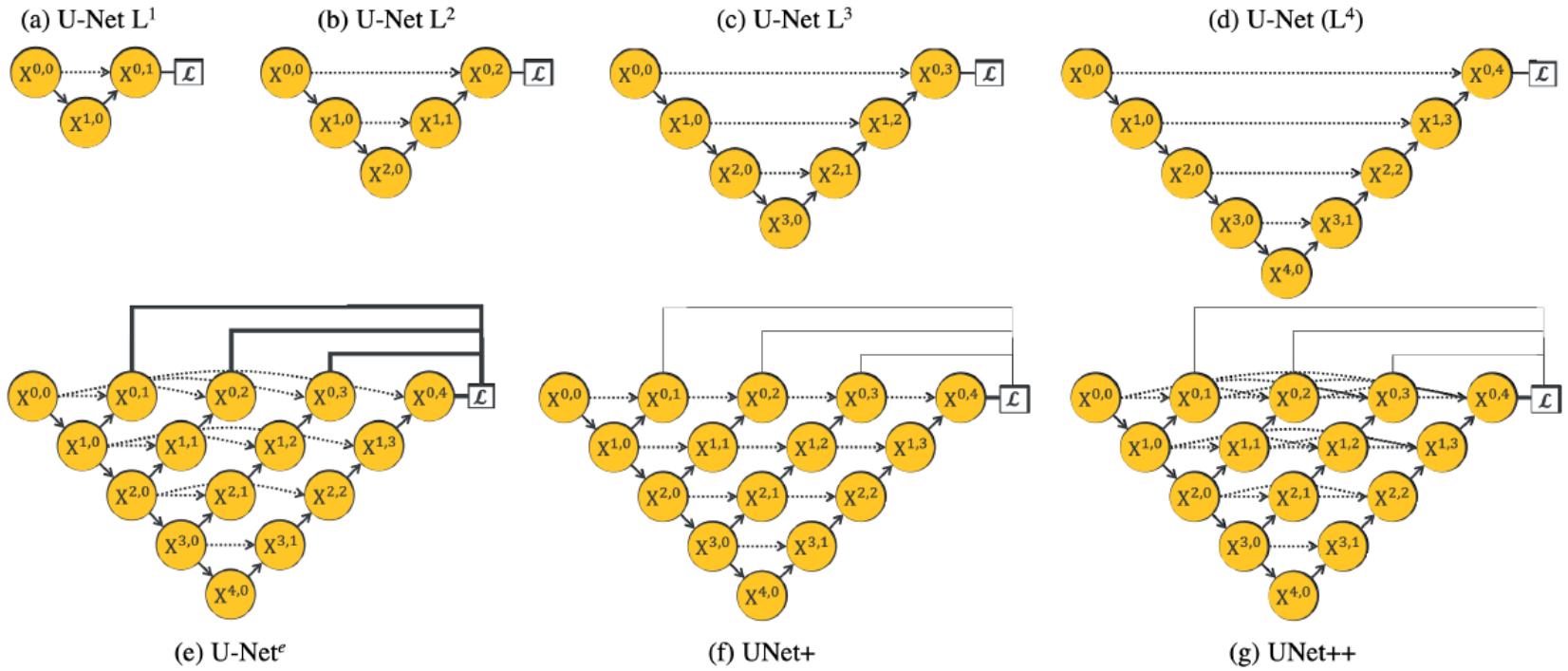
Motivation behind the UNet++



U-Net^e combines U-Nets of varying depths into one unified structure, which is trained with a separate loss function and the output from each U-Net is averaged in inference.

(1) The decoders are disconnected, deeper U-Nets do not offer a supervision signal to the decoders of the shallower U-Nets in the ensemble.
(2) The common design of skip connections is unnecessarily restrictive, requiring the network to combine the decoder feature maps with only the same-scale feature maps from the encoder.

Motivation behind the UNet++



UNet+ connects the disjoint decoders, enabling gradient back-propagation from the deeper decoders to the shallower counterparts.

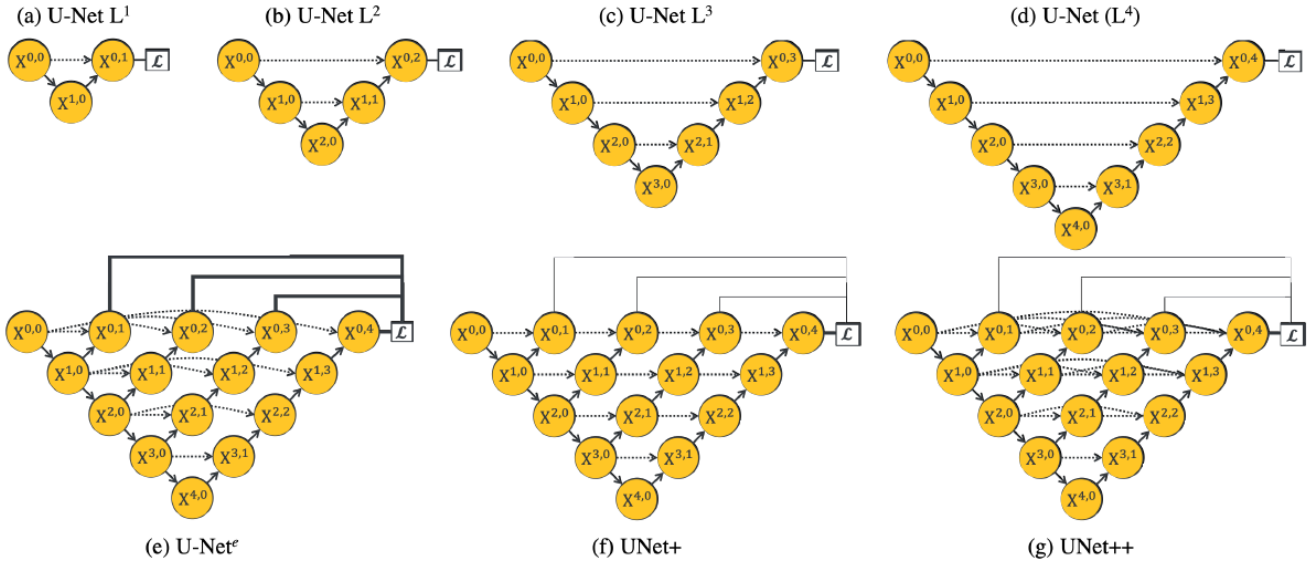
UNet++ uses dense connectivity in UNet+. With dense connectivity, each node in a decoder is presented with not only the final aggregated feature maps but also with the intermediate aggregated feature maps and the original same-scale feature maps from the encoder.

Deep Supervision

2) Deep supervision: We introduce deep supervision in UNet++. For this purpose, we append a 1×1 convolution with $\mathcal{C}$ kernels followed by a *Sigmoid* activation function to the outputs from nodes $X^{0,1}$, $X^{0,2}$, $X^{0,3}$, and $X^{0,4}$ where $\mathcal{C}$ is the number of classes observed in the given dataset. We then define a hybrid segmentation loss consisting of pixel-wise cross-entropy loss and soft dice-coefficient loss for each semantic scale. The hybrid loss may take advantages of what both loss functions have to offer: smooth gradient and handling of class imbalance [28], [29]. Mathematically, the hybrid loss is defined as:

$$\mathcal{L}(Y,P) = -\frac{1}{N}\sum_{c=1}^{\mathcal{C}}\sum_{n=1}^{N}\left(y_{n,c}\log p_{n,c} + \frac{2y_{n,c}p_{n,c}}{y_{n,c}^2 + p_{n,c}^2}\right) \quad (2)$$

where $y_{n,c} \in Y$ and $p_{n,c} \in P$ denote the target labels and predicted probabilities for class c and n^{th} pixel in the batch, N indicates the number of pixels within one batch. The overall loss function for UNet++ is then defined as the weighted summation of the hybrid loss from each individual decoders: $\mathcal{L} = \sum_{i=1}^{d}\eta_i \cdot \mathcal{L}(Y,P^i)$, where d indexes the decoder. In the experiments, we give same balanced weights η_i to each loss, *i.e.*, $\eta_i \equiv 1$, and do not process the ground truth for different outputs supervision like Gaussian blur.



Model pruning

3) Model pruning: Deep supervision enables model pruning. Owing to deep supervision, UNet++ can be deployed in two operation modes: 1) ensemble mode where the segmentation results from all segmentation branches are collected and then averaged, and 2) pruned mode where the segmentation output is selected from only one of the segmentation branches, the choice of which determines the extent of model pruning and speed gain. Fig. 2 shows how the choice of the segmentation branch results in pruned architectures of varying complexity. Specifically, taking the segmentation result from $X^{0,4}$ leads to no pruning whereas taking the segmentation result from $X^{0,1}$ leads to maximal pruning of the network.

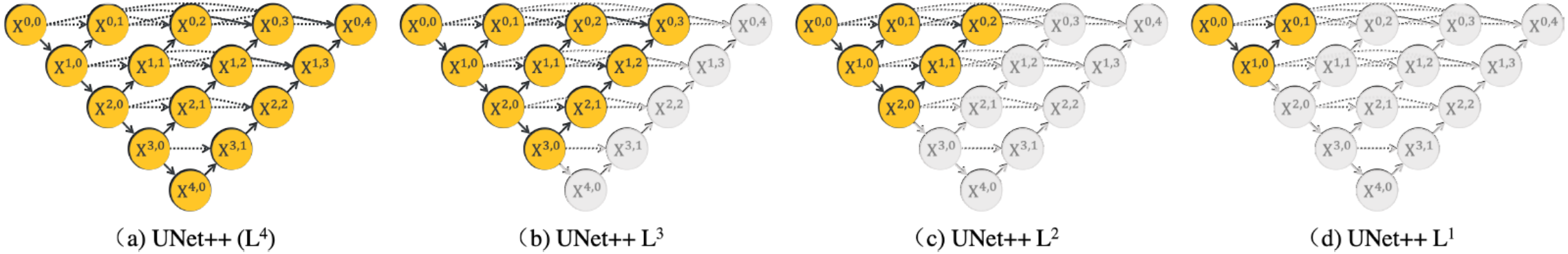


Fig. 2: Training UNet++ with deep supervision makes segmentation results available at multiple nodes $X^{0,j}$, enabling architecture pruning at inference time. Taking the segmentation result from $X^{0,4}$ leads to no pruning, UNet++ (L^4), whereas taking the segmentation result from $X^{0,1}$ results in a maximally pruned architecture, UNet++ L^1. Note that nodes removed during pruning are colored in gray.

Datasets

Six biomedical image segmentation datasets are used in this study, covering lesions/organs from most commonly used medical imaging modalities including microscopy, computed tomography (CT), and magnetic resonance imaging (MRI).

TABLE II: Summary of biomedical image segmentation datasets used in our experiments (see Section III-A for details).

Application	Images	Input Size	Modality	Provider
EM	30	96×96	microscopy	ISBI 2012 [30]
Cell	354	96×96	Cell-CT	VisionGate [31]
Nuclei	670	96×96	mixed	Data Science Bowl
Brain Tumor	66,348	256×256	MRI	BraTS 2013 [32]
Liver	331	96×96	CT	MICCAI 2017 LiTS
Lung Nodule	1,012	64×64×64	CT	LIDC-IDRI [33]

Semantic segmentation results

Wide U-Net consistently outperforms U-Net. This improvement is attributed to the larger number of parameters in wide U-Net. UNet++ without deep supervision achieves a significant IoU gain over both U-Net and wide U-Net for all the six tasks. Using deep supervision and average voting further improves UNet++, increasing the IoU by up to 0.8 points.

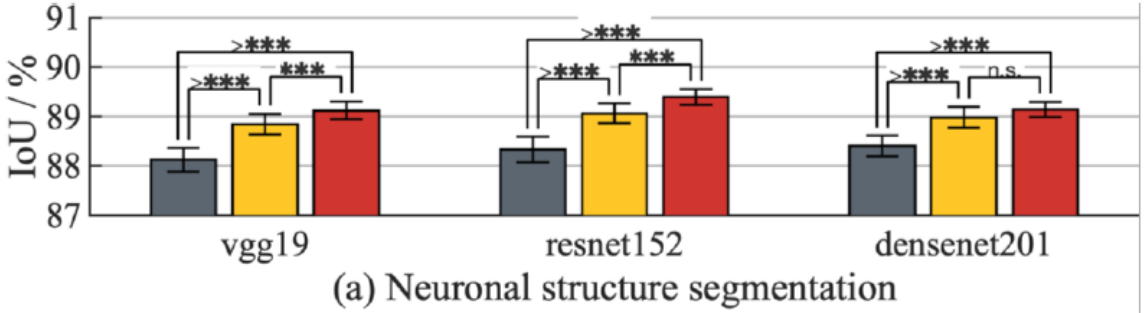
TABLE IV: Semantic segmentation results measured by IoU (mean±s.d. %) for U-Net, wide U-Net, UNet+ (our intermediate proposal), and UNet++ (our final proposal). Both UNet+ and UNet++ are evaluated with and without deep supervision (DS). We have performed independent two sample t-test between U-Net [5] vs. others for 20 independent trials and highlighted boxes in red when the differences are statistically significant ($p < 0.05$).

Architecture	DS	Params	2D Application					Architecture	DS	Params	3D Application
			EM	Cell	Nuclei	Brain Tumor†	Liver				Lung Nodule
U-Net [5]	✗	7.8M	88.30±0.24	88.73±1.64	90.57±1.26	89.21±1.55	79.90±1.38	V-Net [28]	✗	22.6M	71.17±4.53
wide U-Net	✗	9.1M	88.37±0.13	88.91±1.43	90.47±1.15	89.35±1.49	80.25±1.31	wide V-Net	✗	27.0M	73.12±3.99
UNet+	✗	8.7M	88.39±0.15	90.71±1.25	91.73±1.09	90.70±0.91	79.62±1.20	VNet+	✗	25.3M	75.93±2.93
UNet+	✓	8.7M	88.89±0.12	91.18±1.13	92.04±0.89	91.15±0.65	**82.83±0.92**	VNet+	✓	25.3M	76.72±2.48
UNet++	✗	9.0M	88.92±0.14	91.03±1.34	**92.44±1.20**	90.86±0.81	82.51±1.29	VNet++	✗	26.2M	76.24±3.11
UNet++	✓	9.0M	**89.33±0.10**	**91.21±0.98**	92.37±0.98	**91.21±0.68**	82.60±1.11	VNet++	✓	26.2M	**77.05±2.42**

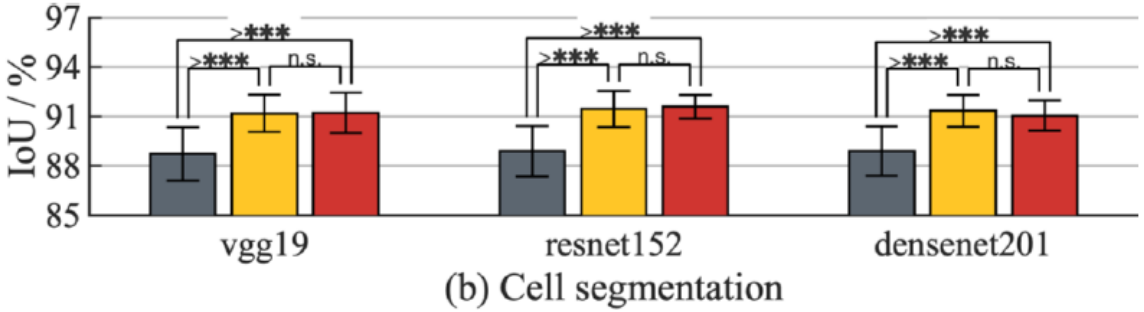
† The winner in BraTS 2013 holds a "complete" Dice of 92% vs. 90.83%±2.46% (our UNet++ with deep supervision).

Semantic segmentation results

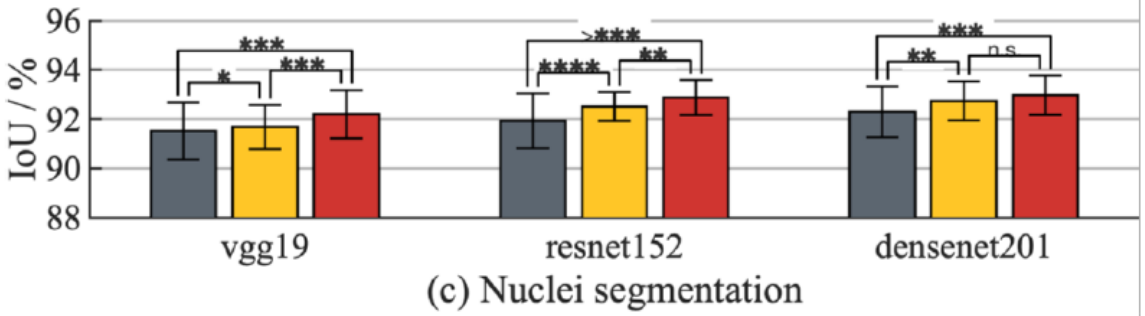
The paper investigated the extensibility of UNet++ for semantic segmentation by applying redesigned skip connections to an array of modern CNN architectures: vgg-19, resnet-152, and densenet-201. UNet++ consistently outperforms U-Net, and UNet+ across all backbone architectures and applications.



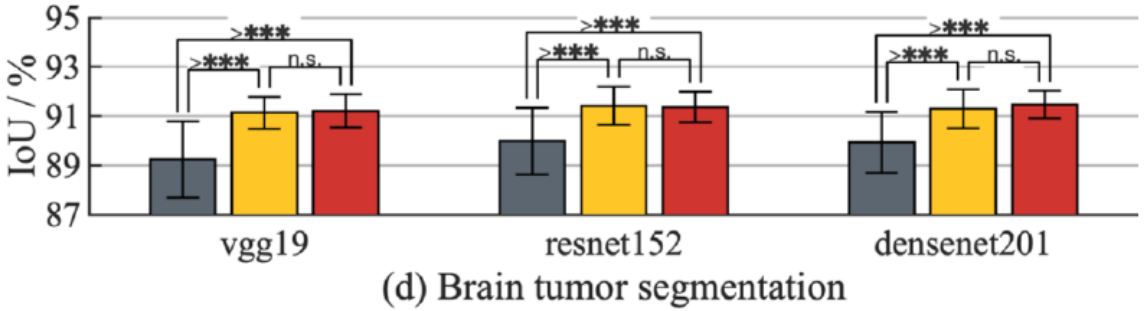
(a) Neuronal structure segmentation



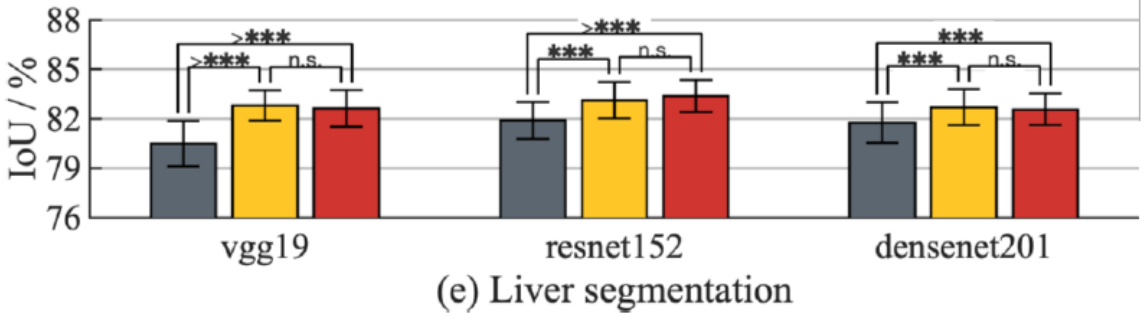
(b) Cell segmentation



(c) Nuclei segmentation



(d) Brain tumor segmentation



(e) Liver segmentation

n.s.	no significance
*	$p < 0.05$
**	$p < 0.01$
***	$p < 0.001$
****	$p < 0.0001$

U-Net

UNet+

UNet++

Semantic and Instance segmentation results

Both UNet++ and Mask RCNN++ outperform their original counterparts

TABLE V: Redesigned skip connections improve both semantic and instance segmentation for the task of nuclei segmentation. We use Mask R-CNN for instance segmentation and U-Net for semantic segmentation in this comparison.

Architecture	Backbone	IoU	Dice	Score
U-Net	resnet101	91.03	75.73	0.244
UNet++	resnet101	**92.55**	**89.74**	**0.327**
Mask R-CNN [12]	resnet101	93.28	87.91	0.401
Mask RCNN++†	resnet101	**95.10**	**91.36**	**0.414**

†Mask R-CNN with UNet++ design in its feature pyramid.

Model pruning results

Once UNet++ is trained, the decoder path for depth *d* at inference time is completely independent from the decoder path for depth *d* + *1*.

UNet++ L3 achieves on average 32.2% reduction in inference time and 75.6% reduction in memory footprint while degrading IoU by only 0.6 points. This observation has the potential to exert important impact on computer-aided diagnosis (CAD) on mobile devices

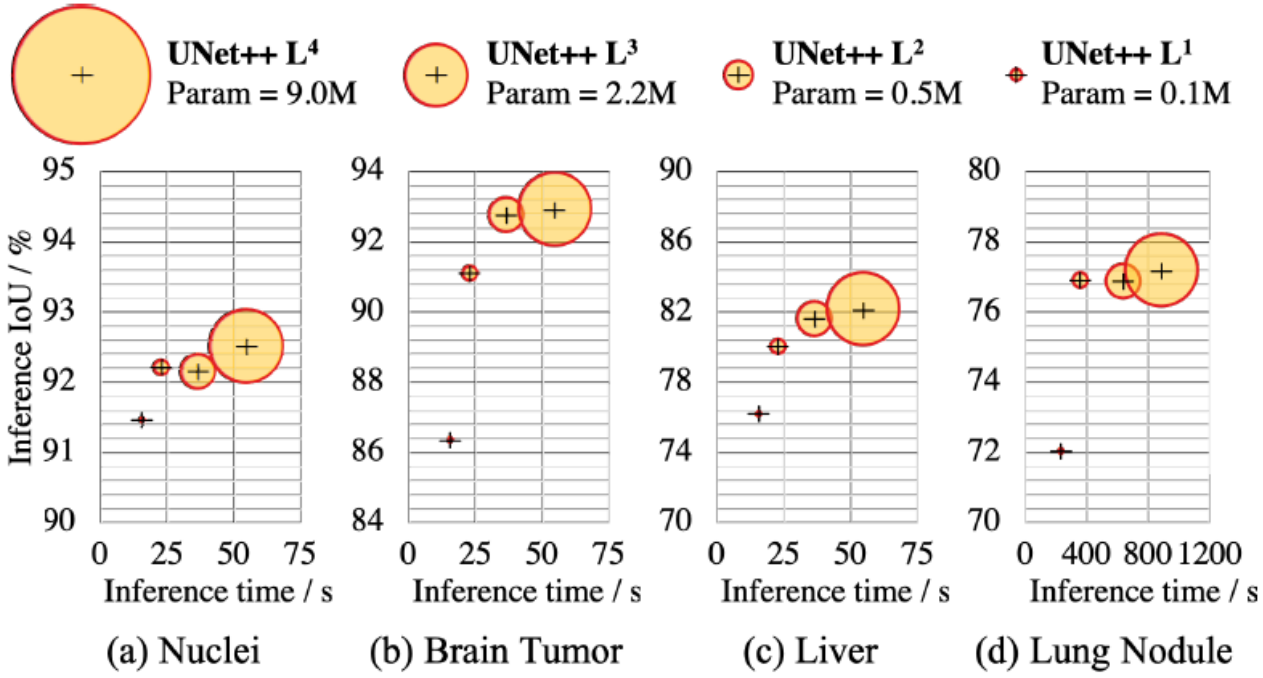


Fig. 5: Complexity (size $\propto$ parameters), inference time, and IoU of UNet++ under different levels of pruning. The inference time is calculated by the time taken to process 10K test images on a single NVIDIA TITAN X (Pascal) GPU with 12 GB memory.

Isolated training & Embedded training

UNet++ L*d* can be trained in two fashions: **1)** embedded training where the full UNet++ model is trained and then pruned at depth *d* to obtain UNet++ L*d*, **2)** isolated training where UNet++ L*d* is trained in isolation without any interactions with the deeper encoder and decoder nodes.

The embedded training of UNet++ L*d* results in a higher performing model than training the same architecture in isolation.

This finding suggests that supervision signal coming from the deep downstream enables training higher performing shallower models.

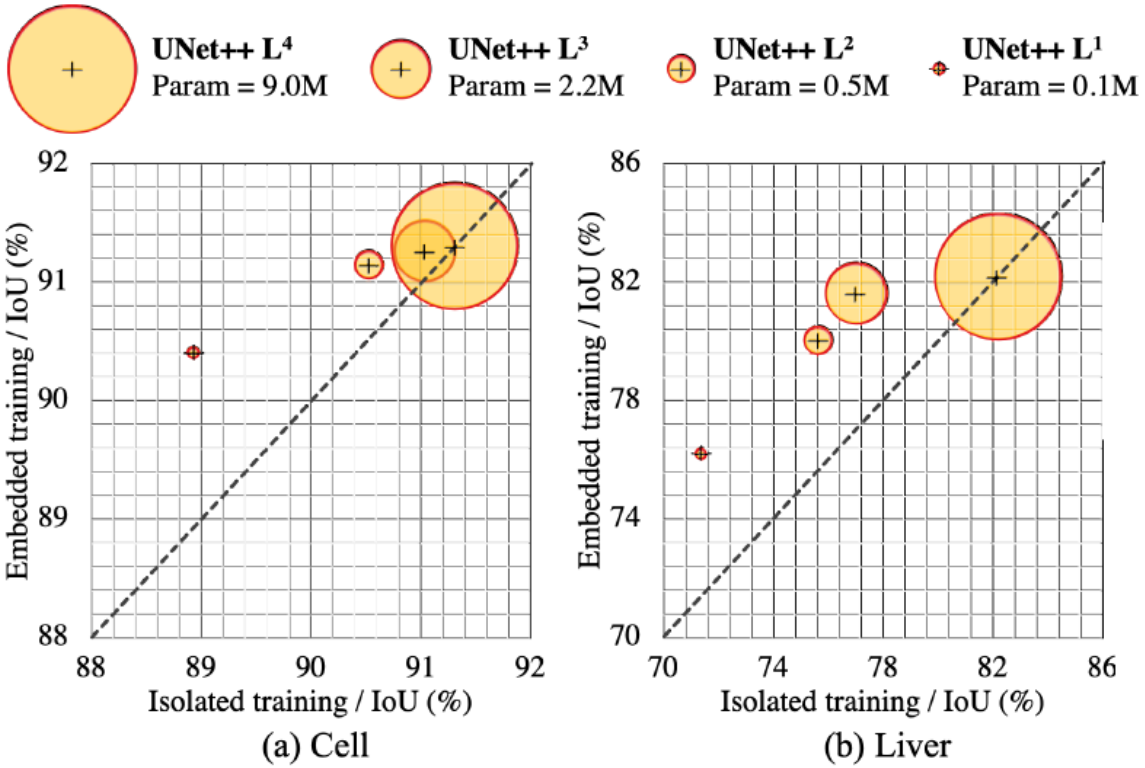


Fig. 6: We demonstrate that our architectural design improves the performance of each shallower network embedded in UNet++. The embedded shallower networks show improved segmentation when pruned from UNet++ in comparison to the same network trained isolated. Due to no pruning, UNet++ L^4 naturally achieves the same level of performance in isolated and embedded training modes.

Performance analysis on stratified lesion sizes

UNet++ consistently outperforms U-Net across all the sizes of brain tumors.

The capability of UNet++ in segmenting tumors of varying sizes is attributed to its built-in ensemble of U-Nets, which enables image segmentation based on multi-receptive field networks.

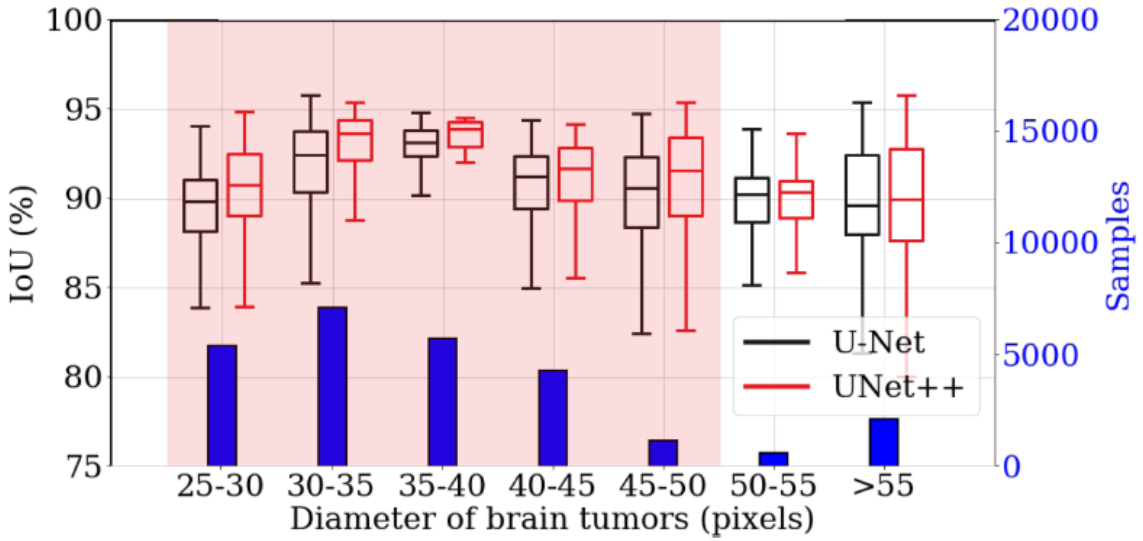


Fig. 7: UNet++ can better segment tumors of various sizes than does U-Net. We measure the size of tumors based on the ground truth masks and then divide them into seven groups. The histogram shows the distribution of different tumor sizes. The box-plot compares the segmentation performances of U-Net (black) and UNet++ (red) in each group. The t-test for two independent samples has been further performed on each group. As seen, UNet++ improves segmentation for all sizes of tumors and the improvement is significant ($p < 0.05$) for the majority of the tumor sizes (highlighted in red).

Conclusion

VII. Conclusion

We have presented a novel architecture, named UNet++, for more accurate image segmentation. The improved performance by our UNet++ is attributed to its nested structure and re-designed skip connections, which aim to address two key challenges of the U-Net: 1) unknown depth of the optimal architecture and 2) the unnecessarily restrictive design of skip connections. We have evaluated UNet++ using six distinct biomedical imaging applications and demonstrated consistent performance improvement over various state-of-the-art back-bones for semantic segmentation and meta framework for instance segmentation.

Thanks for your attention!