Published in final edited form as:
IEEE Trans Med Imaging. 2020 June ; 39(6): 1856–1867. doi:10.1109/TMI.2019.2959609.

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

Zongwei Zhou [Member, IEEE],
Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA

Md Mahfuzur Rahman Siddiquee [Member, IEEE]
School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281 USA

Nima Tajbakhsh [Member, IEEE], **Jianming Liang [Senior Member, IEEE]**
Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA

Abstract

The state-of-the-art models for medical image segmentation are variants of U-Net and fully convolutional networks (FCN). Despite their success, these models have two limitations: (1) their optimal depth is apriori unknown, requiring extensive architecture search or inefficient ensemble of models of varying depths; and (2) their skip connections impose an unnecessarily restrictive fusion scheme, forcing aggregation only at the same-scale feature maps of the encoder and decoder sub-networks. To overcome these two limitations, we propose UNet++, a new neural architecture for semantic and instance segmentation, by (1) alleviating the unknown network depth with an efficient ensemble of U-Nets of varying depths, which partially share an encoder and co-learn simultaneously using deep supervision; (2) redesigning skip connections to aggregate features of varying semantic scales at the decoder sub-networks, leading to a highly flexible feature fusion scheme; and (3) devising a pruning scheme to accelerate the inference speed of UNet++. We have evaluated UNet++ using six different medical image segmentation datasets, covering multiple imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and electron microscopy (EM), and demonstrating that (1) UNet++ consistently outperforms the baseline models for the task of semantic segmentation across different datasets and backbone architectures; (2) UNet++ enhances segmentation quality of varying-size objects—an improvement over the fixed-depth UNet; (3) Mask RCNN++ (Mask R-CNN with UNet++ design) outperforms the original Mask R-CNN for the task of instance segmentation; and (4) pruned UNet++ models achieve significant speedup while showing only modest performance degradation. Our implementation and pre-trained models are available at https://github.com/MrGiovanni/UNetPlusPlus.

zongweiz@asu.edu.

Keywords

Neuronal Structure Segmentation; Liver Segmentation; Cell Segmentation; Nuclei Segmentation; Brain Tumor Segmentation; Lung Nodule Segmentation; Medical Image Segmentation; Semantic Segmentation; Instance Segmentation; Deep Supervision; Model Pruning

1. Introduction

The encoder-decoder networks are widely used in modern semantic and instance segmentation models [1], [2], [3], [4], [5], [6]. Their success is largely attributed to their skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network, and have proven to be effective in recovering fine-grained details of the target objects [7], [8], [9] even on complex background [10], [11]. Skip connections have also played a key role in the success of instance-level segmentation models such as [12], [13] where the idea is to segment and distinguish each instance of desired objects.

However, these encoder-decoder architectures for image segmentation come with two limitations. First, 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. A simple approach would be to train models of varying depths separately and then ensemble the resulting models during the inference time [14], [15], [16]. However, this simple approach is inefficient from a deployment perspective, because these networks do not share a common encoder. Furthermore, being trained independently, these networks do not enjoy the benefits of multi-task learning [17], [18]. Second, 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. While striking as a natural design, 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.

In this paper, we present UNet++, a new general purpose image segmentation architecture that aims at overcoming the above limitations. As presented in Fig. 1(g), UNet++ consists of U-Nets of varying depths whose decoders are densely connected at the same resolution via the redesigned skip pathways. The architectural changes introduced in UNet++ enable the following advantages. First, UNet++ is not prone to the choice of network depth because it embeds U-Nets of varying depths in its architecture. All these U-Nets partially share an encoder, while their decoders are intertwined. By training UNet++ with deep supervision, all the constituent U-Nets are trained simultaneously while benefiting from a shared image representation. This design not only improves the overall segmentation performance, but also enables model pruning during the inference time. Second, UNet++ is not handicapped by unnecessarily restrictive skip connections where only the same-scale feature maps from the encoder and decoder can be fused. The redesigned skip connections introduced in UNet++ present feature maps of varying scales at a decoder node, allowing the aggregation layer to decide how various feature maps carried along the skip connections should be fused with the decoder feature maps. The redesigned skip connections are realized in UNet++ by densely connecting the decoders of the constituents U-Nets at the same resolution. We have extensively evaluated UNet++ across six segmentation datasets and multiple backbones of different depths. Our results demonstrate that UNet++ powered by redesigned skip connections and deep supervision enables a significantly higher level of performance for both semantic and instance segmentation. This significant improvement of UNet++ over the classical UNet architecture is ascribed to the advantages offered by the redesigned skip connections and the extended decoders, which together enable gradual aggregation of the image features across the network, both horizontally and vertically.

In summary, we make the following five contributions:

1. We introduce a built-in ensemble of U-Nets of varying depths in UNet++, enabling improved segmentation performance for varying size objects—an improvement over the fixed-depth U-Net (see Section II-B).
2. We redesign skip connections in UNet++, enabling flexible feature fusion in decoders—an improvement over the restrictive skip connections in U-Net that require fusion of only same-scale feature maps (see Section II-B).
3. We devise a scheme to prune a trained UNet++, accelerating its inference speed while maintaining its performance (see Section IV-C).
4. We discover that simultaneously training multi-depth U-Nets embedded within the UNet++ architecture stimulates collaborative learning among the constituent U-Nets, leading to much better performance than individually training isolated U-Nets of the same architecture (see Section IV-D and Section V-C).
5. We demonstrate the extensibility of UNet++ to multiple backbone encoders and further its applicability to various medical imaging modalities including CT, MRI, and electron microscopy (see Section IV-A and Section IV-B).

II. Proposed Network Architecture: UNet++

Fig. 1 shows how UNet++ evolves from the original U-Net. In the following, we first trace this evolution, motivating the need for UNet++, and then explain its technical and implementation details.

A. Motivation behind the new architecture

We have done a comprehensive ablation study to investigate the performance of U-Nets of varying depths (Fig. 1(a–d)). For this purpose, we have used three relatively small datasets, namely CELL, EM, and BRAIN TUMOR (detailed in Section III-A). Table I summarizes the results. For the cell and brain tumor segmentation, a shallower network (U-Net L^3) outperforms the deep U-Net. For the EM dataset, on the other hand, the deeper U-Nets consistently outperform the shallower counterparts, but the performance gain is only marginal. Our experimental results suggest two key findings: 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. While these findings may encourage an automated neural architecture search, such an approach is hindered by the limited computational resources

[19], [20], [21], [22], [23]. Alternatively, we propose an ensemble architecture, which combines U-Nets of varying depths into one unified structure. We refer to this architecture as U-Net^e (Fig. 1(e)). We train U-Net^e by defining a separate loss function for each U-Net in the ensemble, *i.e.*, $X^{0,j}$, $j \in \{1, 2, 3, 4\}$. Our deep supervision scheme differs from the commonly used deep supervision in deep image classification and image segmentation networks; in [24], [25], [26], [27] the auxiliary loss functions are added to the nodes along the decoder network, *i.e.* $X^{4-j,j}$, $j \in \{0, 1, 2, 3, 4\}$, whereas we apply them on $X^{0,j}$, $j \in \{1, 2, 3, 4\}$. At the inference time, the output from each U-Net in the ensemble is averaged.

The ensemble architecture (U-Net^e) outlined above benefits from knowledge sharing, because all U-Nets within the ensemble partially share the same encoder even though they have their own decoders. However, this architecture still suffers from two drawbacks. First, the decoders are disconnected—deeper U-Nets do not offer a supervision signal to the decoders of the shallower U-Nets in the ensemble. Second, the common design of skip connections used in the U-Net^e is unnecessarily restrictive, requiring the network to combine the decoder feature maps with only the same-scale feature maps from the encoder. While striking as a natural design, there is no guarantee that the same-scale feature maps are the best match for the feature fusion.

To overcome the above limitations, we remove long skip connections from the U-Net^e and connect every two adjacent nodes in the ensemble, resulting in a new architecture, which we refer to as UNet+ (Fig. 1(f)). Owing to the new connectivity scheme, UNet+ connects the disjoint decoders, enabling gradient back-propagation from the deeper decoders to the shallower counterparts. UNet+ further relaxes the unnecessarily restrictive behaviour of skip connections by presenting each node in the decoders with the aggregation of all feature maps computed in the shallower stream. While using aggregated feature maps at a decoder node is far less restrictive than having only the same-scale feature map from the encoder, there is still room for improvement. We further propose to use dense connectivity in UNet+, resulting in our final architecture proposal, which we refer to as UNet++ (Fig. 1(g)). 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. As such, the aggregation layer in the decoder node may learn to use only the same-scale encoder feature maps or use all collected feature maps available at the gate. Unlike U-Net^e, deep supervision is not required for UNet+ and UNet++, however, as we will describe later, deep supervision enables model pruning during the inference time, leading to a significant speedup with only modest drop in performance.

B. Technical details

1) Network connectivity: Let $x^{i,j}$ denote the output of node $X^{i,j}$ where i indexes the down-sampling layer along the encoder and j indexes the convolution layer of the dense block along the skip connection. The stack of feature maps represented by $x^{i,j}$ is computed as

$$x^{i,j} = \begin{cases} \mathcal{H}\left(\mathcal{D}\left(x^{i-1,j}\right)\right), & j = 0 \\ \mathcal{H}\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}\left(x^{i+1,j-1}\right)\right]\right), & j > 0 \end{cases} \qquad (1)$$

where function $\mathcal{H}(\cdot)$ is a convolution operation followed by an activation function, $\mathcal{D}(\cdot)$ and $\mathcal{U}(\cdot)$ denote a down-sampling layer and an up-sampling layer respectively, and [ ] denotes the concatenation layer. Basically, as shown in Fig. 1(g), nodes at level $j = 0$ receive only one input from the previous layer of the encoder; nodes at level $j = 1$ receive two inputs, both from the encoder sub-network but at two consecutive levels; and nodes at level $j > 1$ receive $j + 1$ inputs, of which j inputs are the outputs of the previous j nodes in the same skip connection and the $j + 1^{th}$ input is the up-sampled output from the lower skip connection. The reason that all prior feature maps accumulate and arrive at the current node is because we make use of a dense convolution block along each skip connection.

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:

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

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:$\mathcal{L} = \sum_{i=1}^{d}\eta_i \cdot \mathcal{L}\left(Y, P^i\right)$, 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.

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.

III. Experiments

A. Datasets

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

1) Electron Microscopic (EM): The dataset is provided by the EM segmentation challenge [30] as a part of ISBI 2012. The dataset consists of 30 images (512×512 pixels) from serial section transmission electron microscopy of the Drosophila firt instar larva ventral nerve cord (VNC). Referring to the example in Fig. 3, each image comes with a corresponding fully annotated ground truth segmentation map for cells (white) and membranes (black). The labeled images are split into training (24 images), validation (3 images), and test (3 images) datasets. Both training and inference are done based on 96×96 patches, which are chosen to overlap by half of the patch size via sliding windows. Specifically, during the inference, we aggregate predictions across patches by voting in the overlapping areas.

2) Cell: The dataset is acquired with a Cell-CT imaging system [31]. Two trained experts manually segment the collected images, so each image in the dataset comes with two binary cell masks. For our experiments, we select a subset of 354 images that have the highest level of agreement between the two expert annotators. The selected images are then split into training (212 images), validation (70 images), and test (72 images) subsets.

3) Nuclei: The dataset is provided by the Data Science Bowl 2018 segmentation challenge and consists of 670 segmented nuclei images from different modalities (brightfield vs. fluorescence). This is the only dataset used in this work with instance-level annotation where each nucleolus is marked in a different color. Images are randomly assigned into a training set (50%), a validation set (20%), and a test set (30%). We then use a sliding window mechanism to extract 96×96 patches from the images, with 32-pixel stride for training and validating model, and with 1-pixel stride for testing.

4) Brain Tumor: The dataset is provided by BraTS 2013 [32], [34]. To ease the comparison with other approaches, the models are trained using 20 High-grade (HG) and 10 Low-grade (LG) with Flair, T1, T1c, and T2 scans of MR images from all patients, resulting in a total of 66,348 slices. We further pre-process the dataset by re-scaling the slices to 256×256. Finally, the 30 patients available in the dataset are randomly assigned into five folds, each having images from six patients. We then randomly assign these five folds into a training set (3-fold), a validation set (1-fold), and a test set (1-fold). The ground truth segmentation have four different labels: necrosis, edema, non-enhancing tumor, and enhancing tumor. Following the BraTS-2013, the “complete” evaluation is done by considering all four labels as positive class and others as negative class.

5) Liver: The dataset is provided by MICCAI 2017 LiTS Challenge and consists of 331 CT scans, which we split into training (100 patients), validation (15 patients), and test (15 patients) subsets. The ground truth segmentation provides two different labels: liver and lesion. For our experiments, we only consider liver as positive class and others as negative class.

6) Lung Nodule: The dataset is provided by the Lung Image Database Consortium image collection (LIDC-IDRI) [33] and consists of 1018 cases collected by seven academic centers and eight medical imaging companies. Six cases with ground truth issues were identified and removed. The remaining cases were split into training (510), validation (100), and test (408) sets. Each case is a 3D CT scan and the nodules have been marked as volumetric binary masks. We have re-sampled the volumes to 1-1-1 spacing and then extracted a 64×64×64 crop around each nodule. These 3D crops are used for model training and evaluation.

B. Baselines and implementation

For comparison, we use the original U-Net [35] and a customized wide U-Net architecture for 2D segmentation tasks, and V-Net [28] and a customized wide V-Net architecture for 3D segmentation tasks. We choose U-Net (or V-Net for 3D) because it is a common performance baseline for image segmentation. We have also designed a wide U-Net (or wide V-Net for 3D) with similar number of parameters to our suggested architecture. This is to ensure that the performance gain yielded by our architecture is *not* simply due to increased number of parameters. Table III details the U-Net and wide U-Net architectures. We have further compared the performance of UNet++ against UNet+, which is our intermediate architecture proposal. The numbers of kernels in the intermediate nodes have been given in Table III.

Our experiments are implemented in Keras with Tensorflow backend. We use *early-stop* mechanism on the validation set to avoid over-fitting and evaluate the results using Dice-coefficient and Intersection over Union (IoU). Alternative measurement metrics, such as pixel-wise sensitivity, specificity, F1, and F2 scores, along with the statistical analysis can be found in Appendix Section A. Adam is used as the optimizer with a learning rate of 3e-4. Both UNet+ and UNet++ are constructed from the original U-Net architecture. All the experiments are performed using three NVIDIA TITAN X (Pascal) GPUs with 12 GB memory each.

IV. Results

A. Semantic segmentation results

Table IV compares U-Net, wide U-Net, UNet+, and UNet++ in terms of the number parameters and segmentation results measured by IoU (mean±s.d) for the six segmentation tasks under study. As seen, 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 of neuronal structure (↑0.62±0.10, ↑0.55±0.01), cell (↑2.30±0.30, ↑2.12±0.09), nuclei (↑1.87±0.06, ↑1.71±0.06), brain tumor (↑2.00±0.87, ↑1.86±0.81), liver (↑2.62±0.09, ↑2.26±0.02), and lung nodule (↑5.06±1.42, ↑3.12±0.88) segmentation. Using deep supervision and average voting further improves UNet++, increasing the IoU by up to 0.8

points. Specifically, neuronal structure and lung nodule segmentation benefit the most from deep supervision because they appear at varying scales in EM and CT slices. Deep supervision, however, is only marginally effective for other datasets at best. Fig. 3 depicts a qualitative comparison between the results of U-Net, wide U-Net, and UNet++.

We have further investigated the extensibility of UNet++ for semantic segmentation by applying redesigned skip connections to an array of modern CNN architectures: vgg-19 [36], resnet-152 [8], and densenet-201 [9]. Specifically, we have turned each architecture above into a U-Net model by adding a decoder sub-network, and then replaced the plain skip connections of U-Net with the redesigned connections of UNet++. For comparison, we have also trained U-Net and UNet+ with the aforementioned backbone architectures. For a comprehensive comparison, we have used `EM, CELL, NUCLEI, BRAIN TUMOR` and `LIVER` segmentation datasets. As seen in Fig. 4, UNet++ consistently outperforms U-Net and UNet+ across all backbone architectures and applications under study. Through 20 trials, we further present statistical analysis based on the independent two-sample *t*-test on each pair among U-Net, UNet+, and UNet++. Our results suggest that UNet++ is an effective, backbone-agnostic extension to U-Net. To facilitate reproducibility and model reuse, we have released the implementation[1] of U-Net, UNet+, and UNet++ for various traditional and modern backbone architectures.

B. Instance segmentation results

Instance segmentation consists in segmenting and distinguishing all object instances; hence, more challenging than semantic segmentation. We use Mask R-CNN [12] as the baseline model for instance segmentation. Mask R-CNN utilizes feature pyramid network (FPN) as backbone to generate object proposal at multiple scales, and then outputs the segmentation masks for the collected proposals via a dedicated segmentation branch. We modify Mask R-CNN by replacing the plain skip connections of FPN with the redesigned skip connections of UNet++. We refer to this model as Mask RCNN++. We use resnet101 as the backbone for Mask R-CNN in our experiments.

Table V compares the performance of Mask R-CNN and Mask RCNN++ for nuclei segmentation. We have chosen the NUCLEI dataset because multiple nucleolus instances can be present in an image, in which case each instance is annotated in a different color, and thus marked as a distinct object. Therefore, this dataset is amenable to both semantic segmentation where all nuclei instances are treated as foreground class, and also instance segmentation where each individual nucleus is to be segmented separately. As seen in Table V, Mask RCNN++ outperforms its original counterpart, achieving 1.82 points increase in IoU (93.28% to 95.10%), 3.45 points increase in Dice (87.91% to 91.36%), and 0.013 points increase in the leaderboard score (0.401 to 0.414). To put this performance in perspective, we have also trained a U-Net and UNet++ model for semantic segmentation with a resnet101 backbone. As seen in Table V, Mask R-CNN models achieve higher segmentation performance than semantic segmentation models. Furthermore, as expected, UNet++ outperforms U-Net for semantic segmentation.

[1] The project page: https://github.com/MrGiovanni/UNetPlusPlus

C. Model pruning

Once UNet++ is trained, the decoder path for depth d at inference time is completely independent from the decoder path for depth $d + 1$. As a result, we can completely remove the decoder for depth d+1, obtaining a shallower version of the trained UNet++ at depth d, owing to the introduced deep supervision. This pruning can significantly reduce the inference time, but segmentation performance may degrade. As such, the level of pruning should be determined by evaluating the model's performance on the validation set. We have studied the inference speed-IoU trade-off for UNet++ in Fig. 5. We use UNet++ L^d to denote UNet++ pruned at depth d (see Fig. 2 for further details). As seen, UNet++ L^3 achieves on average 32.2% reduction in inference time and 75.6% reduction in memory footprint while degrading IoU by only 0.6 points. More aggressive pruning further reduces the inference time but at the cost of significant IoU degradation. More importantly, this observation has the potential to exert important impact on computer-aided diagnosis (CAD) on mobile devices, as the existing deep convolutional neural network models are computationally expensive and memory intensive.

D. Embedded vs. isolated training of pruned models

In theory, 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. Referring to Fig. 2, embedded training of a sub-network consists of training all graph nodes (both yellow and grey components) with deep supervision, but we then use only the yellow sub-network during the inference time. In contrast, isolated training consists of removing the grey nodes from the graph, basing the training and test solely on the yellow sub-network.

We have compared the isolated and embedded training schemes for various levels of UNet++ pruning across two datasets in Fig. 6. We have discovered that the embedded training of UNet++ L^d results in a higher performing model than training the same architecture in isolation. The observed superiority is more pronounced under aggressive pruning when the full UNet++ is pruned to UNet++ L^1. In particular, the embedded training of UNet++ L^1 for liver segmentation achieves 5-point increase in IoU over the isolated training scheme. This finding suggests that supervision signal coming from the deep downstream enables training higher performing shallower models. This finding is also related to knowledge distillation where the knowledge learned by a deep teacher network is learned by a shallower student network.

V. Discussions

A. Performance analysis on stratified lesion sizes

Fig. 7 compares U-Net and UNet++ for segmenting different sizes of brain tumors. To avoid clutter in the figure, we group the tumors by size into seven buckets. As seen, UNet++ consistently outperforms U-Net across all the buckets. We also adopt t-test on each bucket based on 20 different trials to measure the significance of the improvement, concluding that 5 out of the 7 comparisons are statistically significant ($p < 0.05$). 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.

B. Feature maps visualization

In Section II-A, we explained that the redesigned skip connections enable the fusion of semantically rich decoder feature maps with feature maps of varying semantic scales from the intermediate layers of the architecture. In this section, we illustrate this privilege of our re-designed skip connections by visualizing the intermediate feature maps.

Fig. 8 shows representative feature maps from early, intermediate, and late layers along the top most skip connection (*i.e.*, $X^{0,i}$) for a brain tumor image. The representative feature map for a layer is obtained by averaging all its feature maps. Also note that architectures in the left side of Fig. 8 are trained using only loss function appended to the deepest decoder layer ($X^{0,4}$) whereas the architectures in the right side of Fig. 8 are trained with deep supervision. Note that these feature maps are not the final outputs. We have appended an additional 1×1 convolutional layer on top of each decoder branch to form the final segmentation. We observe that the outputs of U-Net's intermediate layers are semantically dissimilar whereas for UNet+ and UNet++ the outputs are formed gradually. The output of node $X^{0,0}$ in U-Net undergoes slight transformation (few convolution operations only) whereas the output of $X^{1,3}$, the input of $X^{0,4}$, goes through nearly every transformation (four down-sampling and three up-sampling stages) learned by the network. Hence, there is a large gap between the representation capability of $X^{0,0}$ and $X^{1,3}$. So, simply concatenating the outputs of $X^{0,4}$ and $X^{1,3}$ is not an optimal solution. In contrast, redesigned skip connections in UNet+ and UNet++ help refine the segmentation result gradually. We further present the learning curves of all six medical applications in Appendix Section B, revealing that the addition of dense connections in UNet++ encourages a better optimization and reaches lower validation loss.

C. Collaborative learning in UNet++

Collaborative learning is known as training multiple classifier heads of the same network simultaneously on the same training data. It is found to improve the generalization power of deep neural networks [37]. UNet++ naturally embodies collaborative learning through aggregating multi-depth networks and supervising segmentation heads from each of the constituent networks. Besides, the segmentation heads, for example $X^{0,2}$ in Fig. 2, receive gradients from both strong (loss from ground truth) and soft (losses propagated from adjacent deeper nodes) supervision. As a result, the shallower networks improve their segmentation (Fig. 6) and provide more informative representation to deeper counterparts. Basically, deeper and shallower networks regularize each other via collaborative learning in UNet++. Training multi-depth embedded networks together results in improved segmentation than training them individually as isolated network which is evident in Section IV-D. The embedded design of UNet++ makes it amenable to auxiliary training, multi-task learning, and knowledge distillation [17], [38], [37].

VI. Related Works

In the following, we review the works related to redesigned skip connections, feature aggregation, and deep supervision, which are the main components of our new architecture.

A. Skip connections

Skip connections were first introduced in the seminal work of Long *et al.* [39] where they proposed a fully convolutional networks (FCN) for semantic segmentation. Shortly after, building on skip connections, Ronneberger *et al.* [35] proposed U-Net architecture for semantic segmentation in medical images. The FCN and U-Net architectures however differ in how the up-sampled decoder feature maps were fused with the same-scale feature maps from the encoder network. While FCN [39] uses the summation operation for feature fusion, U-Net [35] concatenates the features followed by the application of convolutions and non-linearities. The skip connections have shown to help recover the full spatial resolution, making fully convolutional methods suitable for semantic segmentation [40], [41], [42], [43]. Skip connections have further been used in modern neural architectures such as residual networks [8], [44] and dense networks [9], facilitating the gradient flow and improving the overall performance of classification networks.

B. Feature aggregation

The exploration of aggregating hierarchical feature has recently been the subject of research. Fourure *et al.* [45] propose GridNet, which is an encoder-decoder architecture wherein the feature maps are wired in a grid fashion, generalizing several classical segmentation architectures. Despite GridNet contains multiple streams with different resolutions, it lacks up-sampling layers between skip connections; and thus, it does not represent UNet++. Full-resolution residual networks (FRRN) [46] employs a two-stream system, where full-resolution information is carried in one stream and context information in the other pooling stream. In [47], two improved versions of FRRN are proposed, *i.e.*, incremental MRRN with 28.6M parameters and dense MRRN with 25.5M parameters. These 2D architectures however have similar number of parameters to our 3D VNet++ and three times more parameters than 2D UNet++; and thus, simply upgrading these architectures to a 3D manner may not be amenable to the common 3D volumetric medical imaging applications. We would like to note that our redesigned dense skip connections are completely different from those used in MRRN, which consists of a common residual stream. Also, it's not flexible to apply the design of MRRN to other backbone encoders and meta framework such as Mask R-CNN [12]. DLA[2] [48], topologically equivalent to our intermediate architecture UNet+ (Fig. 1(f)), sequentially connects the same resolution of feature maps, without long skip connections as used in U-Net. Our experimental results demonstrate that by densely connecting the layers, UNet++ achieves higher segmentation performance than UNet+/DLA (see Table IV).

[2] Deep Layer Aggregation—a simultaneous but independent work published in CVPR-2018 [48].

C. Deep supervision

He *et al.* [8] suggested that the depth *d* of network can act as a regularizer. Lee *et al.* [27] demonstrated that deeply supervised layers can improve the learning ability of the hidden layer, enforcing the intermediate layers to learn discriminative features, enabling fast convergence and regularization of the network [26]. DenseNet [9] performs a similar deep supervision in an implicit fashion. Deep supervision can be used in U-Net like architecture as well. Dou *et al.* [49] introduce a deep supervision by combining predictions from varying resolutions of feature maps, suggesting that it can combat potential optimization difficulties and thus reach faster convergence rate and more powerful discrimination capability. Zhu *et al.* [50] used eight additional deeply supervised layers in their proposed architecture. Our nested networks are however more amenable to training under deep supervision: 1) multiple decoders automatically generate full resolution segmentation maps; 2) the networks are embedded various different depths of U-Net so that it grasps multiple-resolution features; 3) densely connected feature maps help smooth the gradient flow and give relatively consistent predicting mask; 4) the high dimension features have effects on every outputs through back-propagation, allowing us to prune the network in the inference phase.

D. Our previous work

We first presented UNet++ in our DLMIA 2018 paper [51]. UNet++ has since been quickly adopted by the research community, either as a strong baseline for comparison [52], [53], [54], [55], or as a source of inspiration for developing newer semantic segmentation architectures [56], [57], [58], [59], [60], [61]; it has also been utilized for multiple applications, such as segmenting objects in biomedical images [62], [63], natural images [64], and satellite images [65], [66]. Recently, Shenoy [67] has independently and systematically investigated UNet++ for the task of “contact prediction model PconsC4”, demonstrating significant improvement over widely-used U-Net.

Nevertheless, to further strengthen UNet++ on our own, the current work presents several extensions to our previous work: (1) we present a comprehensive study on network depth, motivating the need for the proposed architecture (Section II-A); (2) we compare the embedded training schemes with the isolated ones at various levels of pruned UNet++, and discover that training embedded U-Nets of multi-depths leads to improved performance than individually training them in isolation (Section IV-D); (3) we strengthen our experiments by including a new magnetic resonance imaging (MRI) dataset for brain tumor segmentation (Section IV); (4) we demonstrate the effectiveness of UNet++ in Mask R-CNN, resulting in a more powerful model namely Mask RCNN++ (Section IV-B); (5) we investigate the extensibility of UNet++ to multiple advanced encoder backbones for semantic segmentation (Section IV-A); (6) we study the effectiveness of UNet++ in segmenting lesions of varying sizes (Section V-A); and (7) we visualize the feature propagation along the resigned skip connection to explain the performance (Section V-B).

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 redesigned 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 backbones for semantic segmentation and meta framework for instance segmentation.

Acknowledgments

This research has been supported partially by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and partially by NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. We thank Mohammad Reza Hosseinzadeh Taher and Fatemeh Haghighi for their verification of liver segmentation performance and the ablation study of embedded and isolated UNet++. We also thank Michael G. Meyer for allowing us to test our ideas on the Cell-CT dataset. The content of this paper is covered by US patents pending.

APPENDIX A: Additional Measurements

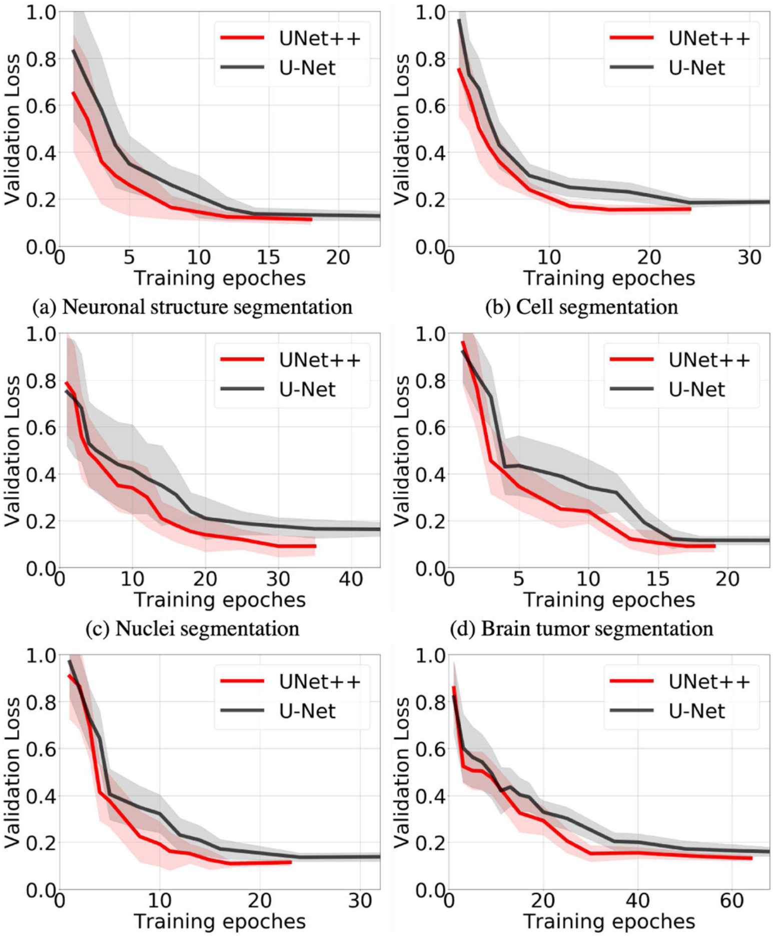
TABLE VI:

Pixel-wise sensitivity, specificity, F1, and F2 scores for all six applications under study. Note that the *p*-values are calculated between our UNet++ with deep supervision vs. the original U-Net. As seen, powered by redesigned skip connections and deep supervision, UNet++ achieves a significantly higher level of segmentation performance over U-Net across all the biomedical applications under study.

EM	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**
U-Net	91.21±2.18	83.55±1.62	87.21±1.88	89.56±2.06
UNet++	92.87±2.08	84.94±1.55	88.73±1.79	91.17±1.96
p-value	0.018	0.008	0.013	0.016
Cell	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**
U-Net	94.04±2.36	96.10±0.75	81.25±2.62	88.47±2.49
UNet++	95.88±2.59	96.76±0.65	84.34±2.52	90.90±2.57
p-value	0.025	0.005	5.00*e*-4	0.004
Nuclei	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**
U-Net	93.57±4.30	93.94±0.87	83.64±2.97	89.33±3.71
UNet++	97.28±4.85	96.30±0.94	90.14±3.82	94.29±4.41
p-value	0.015	5.35*e*-10	6.75*e*-7	4.47*e*-4
Brain Tumor	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**
U-Net	94.00±1.15	97.52±0.78	88.42±2.61	91.68±1.77
UNet++	95.81±1.25	98.01±0.67	90.83±2.46	93.75±1.77
p-value	2.90*e*-5	0.042	0.005	7.03*e*-3
Liver	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**
U-Net	91.22±2.02	98.48±0.43	86.19±2.84	89.14±2.37
UNet++	93.15±1.88	98.74±0.36	88.54±2.57	91.25±2.18
p-value	0.003	0.046	0.010	0.006
Lung Nodule	**Sensitivity**	**Specificity**	**F1 score**	**F2 score**

EM	Sensitivity	Specificity	F1 score	F2 score
U-Net	94.95±1.31	97.27±0.47	83.98±1.94	90.24±1.60
UNet++	95.83±0.86	97.81±0.40	86.78±1.66	91.99±1.22
p-value	0.018	3.25*e*-3	1.92*e*-5	4.27*e*-3

APPENDIX B: Learning Curves



(a) Neuronal structure segmentation

(b) Cell segmentation

(c) Nuclei segmentation

(d) Brain tumor segmentation

(e) Liver segmentation

(f) Lung nodule segmentation

Fig. 9:

UNet++ enables a better optimization than U-Net evidenced by the learning curves for the tasks of neuronal structure, cell, nuclei, brain tumor, liver, and lung nodule segmentation. We have plotted the validation losses averaged by 20 trials for each application. As seen, UNet++ with deep supervision accelerates the convergence speed and yields the lower validation loss due to the new design of the intermediate layers and dense skip connections.

REFERENCES

[1]. Zhou SK, Greenspan H, and Shen D, Deep learning for medical image analysis. Academic Press, 2017.

[2]. Shen D, Wu G, and Suk H-I, “Deep learning in medical image analysis,” Annual review of biomedical engineering, vol. 19, pp. 221–248, 2017.

[3]. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, and Sánchez CI, “A survey on deep learning in medical image analysis,” Medical image analysis, vol. 42, pp. 60–88, 2017. [PubMed: 28778026]

[4]. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, and Tang A, “Deep learning: a primer for radiologists,” Radiographics, vol. 37, no. 7, pp. 2113–2131, 2017. [PubMed: 29131760]

[5]. Falk T, Mai D, Bensch R, O. Çiçek, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K et al., “U-net: deep learning for cell counting, detection, and morphometry,” Nature methods, p. 1, 2018.

[6]. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang J, Wu Z, and Ding X, “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,” arXiv preprint arXiv:1908.10454, 2019.

[7]. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, and Pal C, “The importance of skip connections in biomedical image segmentation,” in Deep Learning and Data Labeling for Medical Applications. Springer, 2016, pp. 179–187.

[8]. He K, Zhang X, Ren S, and Sun J, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[9]. Huang G, Liu Z, Weinberger KQ, and van der Maaten L, “Densely connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017, p. 3.

[10]. Hariharan B, Arbeláez P, Girshick R, and Malik J, “Hypercolumns for object segmentation and fine-grained localization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.

[11]. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, and Belongie S, “Feature pyramid networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017, p. 4.

[12]. He K, Gkioxari G, Dollár P, and Girshick R, “Mask r-cnn,” in Proceedings of the IEEE International Conference on Computer Vision IEEE, 2017, pp. 2980–2988.

[13]. Hu R, Dollár P, He K, Darrell T, and Girshick R, “Learning to segment every thing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4233–4241.

[14]. Dietterich TG, “Ensemble methods in machine learning,” in International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.

[15]. Hoo-Chang S, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, and Summers RM, “Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,” IEEE transactions on medical imaging, vol. 35, no. 5, p. 1285, 2016. [PubMed: 26886976]

[16]. Ciompi F, de Hoop B, van Riel SJ, Chung K, Scholten ET, Oudkerk M, de Jong PA, Prokop M, and van Ginneken B, “Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box,” Medical image analysis, vol. 26, no. 1, pp. 195–202, 2015. [PubMed: 26458112]

[17]. Bengio Y et al., “Learning deep architectures for ai,” Foundations and trends R in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[18]. Zhang Y and Yang Q, “A survey on multi-task learning,” arXiv preprint arXiv:1707.08114, 2017.

[19]. Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li L-J, Fei-Fei L, Yuille A, Huang J, and Murphy K, “Progressive neural architecture search,” in Proceedings of the European Conference on Computer Vision, 2018, pp. 19–34.

[20]. Zoph B, Vasudevan V, Shlens J, and Le QV, “Learning transferable architectures for scalable image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710.

[21]. Liu C, Chen L-C, Schroff F, Adam H, Hua W, Yuille AL, and Fei-Fei L, “Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 82–92.

[22]. Zhang Y, Qiu Z, Liu J, Yao T, Liu D, and Mei T, “Customizable architecture search for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11 641–11 650.

[23]. Li X, Zhou Y, Pan Z, and Feng J, “Partial order pruning: for best speed/accuracy trade-off in neural architecture search,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9145–9153.

[24]. Xie S and Tu Z, “Holistically-nested edge detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.

[25]. Chen H, Qi XJ, Cheng JZ, and Heng PA, “Deep contextual networks for neuronal structure segmentation,” in Thirtieth AAAI conference on artificial intelligence, 2016.

[26]. Dou Q, Yu L, Chen H, Jin Y, Yang X, Qin J, and Heng P-A, “3d deeply supervised network for automated segmentation of volumetric medical images,” Medical image analysis, vol. 41, pp. 40–54, 2017. [PubMed: 28526212]

[27]. Lee C-Y, Xie S, Gallagher P, Zhang Z, and Tu Z, “Deeply-supervised nets,” in Artificial Intelligence and Statistics, 2015, pp. 562–570.

[28]. Milletari F, Navab N, and Ahmadi S-A, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in 2016 Fourth International Conference on 3D Vision (3DV) IEEE, 2016, pp. 565–571.

[29]. Sudre CH, Li W, Vercauteren T, Ourselin S, and Cardoso MJ, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2017, pp. 240–248.

[30]. Cardona A, Saalfeld S, Preibisch S, Schmid B, Cheng A, Pulokas J, Tomancak P, and Hartenstein V, “An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy,” PLoS biology, vol. 8, no. 10, p. e1000502, 2010. [PubMed: 20957184]

[31]. Meyer MG, Hayenga JW, Neumann T, Katdare R, Presley C, Steinhauer DE, Bell TM, Lancaster CA, and Nelson AC, “The cell-ct 3-dimensional cell imaging technology platform enables the detection of lung cancer using the noninvasive luced sputum test,” Cancer cytopathology, vol. 123, no. 9, pp. 512–523, 2015. [PubMed: 26148817]

[32]. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R et al., “The multimodal brain tumor image segmentation benchmark (brats),” IEEE transactions on medical imaging, vol. 34, no. 10, p. 1993, 2015. [PubMed: 25494501]

[33]. Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA et al., “The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans,” Medical physics, vol. 38, no. 2, pp. 915–931, 2011. [PubMed: 21452728]

[34]. Kistler M, Bonaretti S, Pfahrer M, Niklaus R, and Büchler P, “The virtual skeleton database: an open access repository for biomedical research and collaboration,” Journal of medical Internet research, vol. 15, no. 11, p. e245, 2013. [PubMed: 24220210]

[35]. Ronneberger O, Fischer P, and Brox T, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234–241.

[36]. Simonyan K and Zisserman A, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[37]. Song G and Chai W, “Collaborative learning for deep neural networks,” in Neural Information Processing Systems (NeurIPS), 2018.

[38]. Hinton G, Vinyals O, and Dean J, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.

[39]. Long J, Shelhamer E, and Darrell T, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[40]. Chaurasia A and Culurciello E, “Linknet: Exploiting encoder representations for efficient semantic segmentation,” in 2017 IEEE Visual Communications and Image Processing (VCIP) IEEE, 2017, pp. 1–4.

[41]. Lin G, Milan A, Shen C, and Reid ID, “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation.” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017, p. 5.

[42]. Zhao H, Qi X, Shen X, Shi J, and Jia J, “Icnet for real-time semantic segmentation on high-resolution images,” in Proceedings of the European Conference on Computer Vision, 2018, pp. 405–420.

[43]. Tajbakhsh N, Lai B, Ananth S, and Ding X, “Errornet: Learning error representations from limited data to improve vascular segmentation,” arXiv preprint arXiv:1910.04814, 2019.

[44]. He K, Zhang X, Ren S, and Sun J, “Identity mappings in deep residual networks,” in Proceedings of the European Conference on Computer Vision Springer, 2016, pp. 630–645.

[45]. Fourure D, Emonet R, Fromont E, Muselet D, Trémeau A, and Wolf C, “Residual conv-deconv grid network for semantic segmentation,” in Proceedings of the British Machine Vision Conference, 2017, 2017.

[46]. Pohlen T, Hermans A, Mathias M, and Leibe B, “Full-resolution residual networks for semantic segmentation in street scenes,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4151–4160.

[47]. Jiang J, Hu Y-C, Liu C-J, Halpenny D, Hellmann MD, Deasy JO, Mageras G, and Veeraraghavan H, “Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images,” IEEE transactions on medical imaging, vol. 38, no. 1, pp. 134–144, 2019. [PubMed: 30040632]

[48]. Yu F, Wang D, Shelhamer E, and Darrell T, “Deep layer aggregation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2018, pp. 2403–2412.

[49]. Dou Q, Chen H, Jin Y, Yu L, Qin J, and Heng P-A, “3d deeply supervised network for automatic liver segmentation from ct volumes,” in International Conference on Medical Image Computing and Computer-Assisted Intervention Springer, 2016, pp. 149–157.

[50]. Zhu Q, Du B, Turkbey B, Choyke PL, and Yan P, “Deeply-supervised cnn for prostate segmentation,” in International Joint Conference on Neural Networks (IJCNN) IEEE, 2017, pp. 178–184.

[51]. Zhou Z, Siddiquee MMR, Tajbakhsh N, and Liang J, “Unet++: A nested u-net architecture for medical image segmentation,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2018, pp. 3–11.

[52]. Sun K, Zhao Y, Jiang B, Cheng T, Xiao B, Liu D, Mu Y, Wang X, Liu W, and Wang J, “High-resolution representations for labeling pixels and regions,” CoRR, vol. abs/1904.04514, 2019.

[53]. Fang Y, Chen C, Yuan Y, and Tong K.-y., “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 302–310.

[54]. Fang J, Zhang Y, Xie K, Yuan S, and Chen Q, “An improved mpb-cnn segmentation method for edema area and neurosensory retinal detachment in sd-oct images,” in International Workshop on Ophthalmic Medical Image Analysis. Springer, 2019, pp. 130–138.

[55]. Meng C, Sun K, Guan S, Wang Q, Zong R, and Liu L, “Multiscale dense convolutional neural network for dsa cerebrovascular segmentation,” Neurocomputing, vol. 373, pp. 123–134, 2020.

[56]. Zhang J, Jin Y, Xu J, Xu X, and Zhang Y, “Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation,” arXiv preprint arXiv:1812.00352, 2018.

[57]. Chen F, Ding Y, Wu Z, Wu D, and Wen J, “An improved framework called du++ applied to brain tumor segmentation,” in 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) IEEE, 2018, pp. 85–88.

[58]. Zhou C, Chen S, Ding C, and Tao D, “Learning contextual and attentive information for brain tumor segmentation,” in International MICCAI Brainlesion Workshop. Springer, 2018, pp. 497–507.

[59]. Wu S, Wang Z, Liu C, Zhu C, Wu S, and Xiao K, “Automatical segmentation of pelvic organs after hysterectomy by using dilated convolution u-net++,” in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C) IEEE, 2019, pp. 362–367.

[60]. Song T, Meng F, Rodríguez-Patón A, Li P, Zheng P, and Wang X, “U-next: A novel convolution neural network with an aggregation u-net architecture for gallstone segmentation in ct images,” IEEE Access, vol. 7, pp. 166 823–166 832, 2019.

[61]. Yang C and Gao F, “Eda-net: Dense aggregation of deep and shallow information achieves quantitative photoacoustic blood oxygenation imaging deep in human breast,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 246–254.

[62]. Zyuzin V and Chumarnaya T, “Comparison of unet architectures for segmentation of the left ventricle endocardial border on two-dimensional ultrasound images,” in 2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) IEEE, 2019, pp. 110–113.

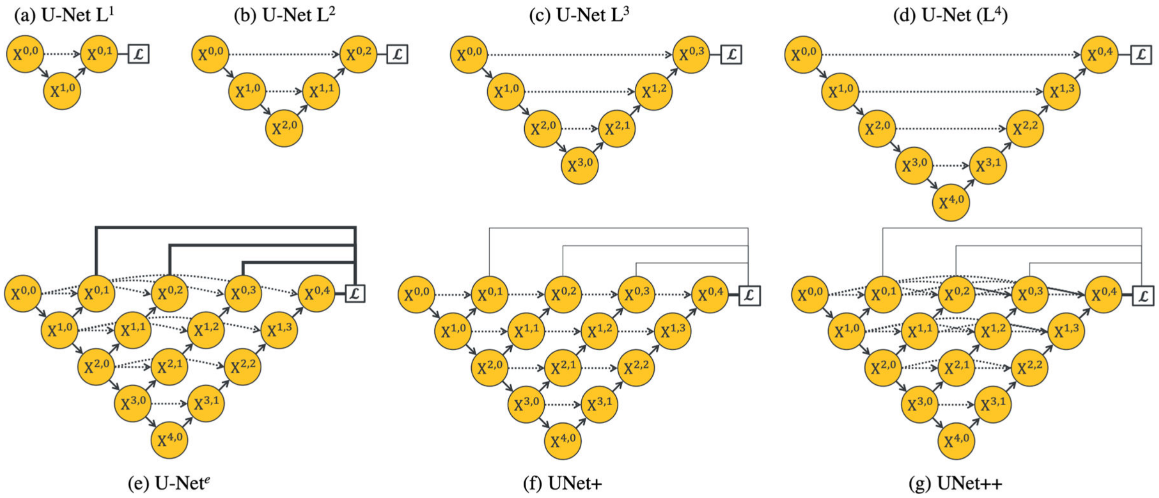
[63]. Cui H, Liu X, and Huang N, “Pulmonary vessel segmentation based on orthogonal fused u-net++ of chest ct images,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 293–300.

[64]. Sun K, Xiao B, Liu D, and Wang J, “Deep high-resolution representation learning for human pose estimation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019.

[65]. Peng D, Zhang Y, and Guan H, “End-to-end change detection for high resolution satellite images using improved unet++,” Remote Sensing, vol. 11, no. 11, p. 1382, 2019.

[66]. Zhang Y, Gong W, Sun J, and Li W, “Web-net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries,” Remote Sensing, vol. 11, no. 16, p. 1897, 2019.

[67]. Shenoy AA, “Feature optimization of contact map predictions based on inter-residue distances and u-net++ architecture.”


(a) U-Net L^1

(b) U-Net L^2

(c) U-Net L^3

(d) U-Net (L^4)

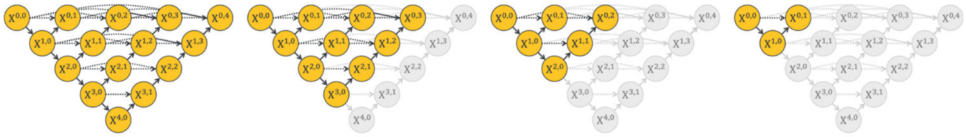
(e) U-Net^e

(f) UNet+

(g) 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 by connecting the decoders of U-Net^e, enabling the deeper decoders to send supervision signals to the shallower decoders. (g) UNet++ is constructed by adding dense skip connections to UNet+, 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++.



(a) UNet++ (L^4)

(b) UNet++ L^3

(c) UNet++ L^2

(d) UNet++ L^1

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.

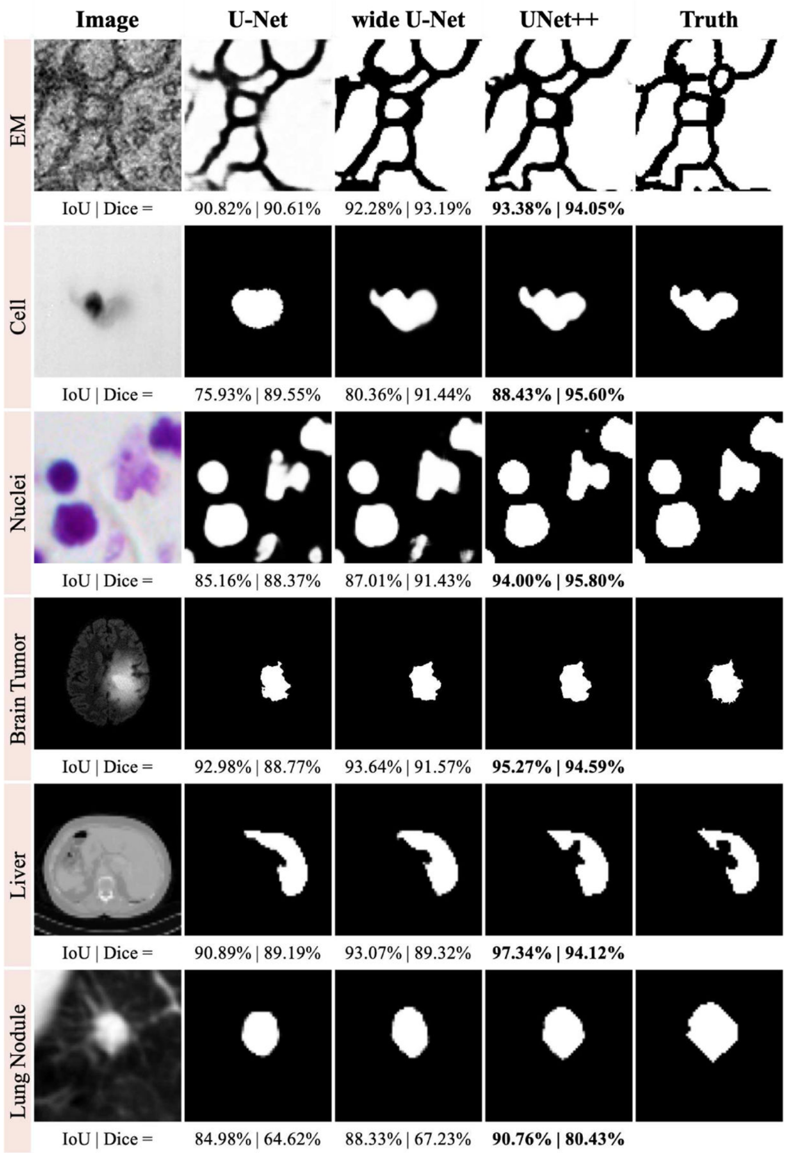
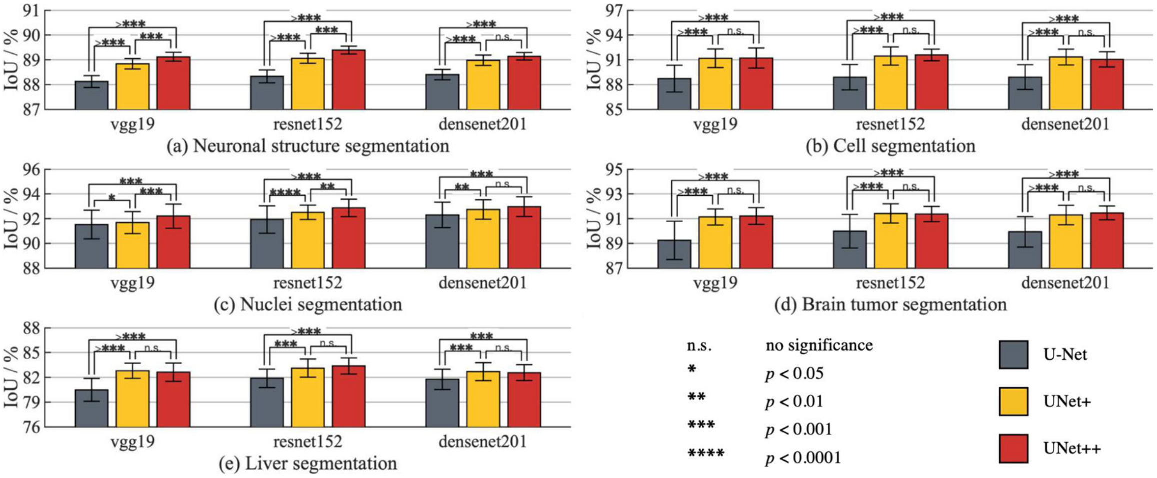


Fig. 3:

Qualitative comparison among U-Net, wide U-Net, and UNet++; showing segmentation results for our six distinct biomedical image segmentation applications. They include various 2D and 3D modalities. The corresponding quantitative scores are provided at the bottom of each prediction (IoU | Dice).



(a) Neuronal structure segmentation

(b) Cell segmentation

(c) Nuclei segmentation

(d) Brain tumor segmentation

(e) Liver segmentation

Fig. 4:

Comparison between U-Net, UNet+, and UNet++ when applied to the state-of-the-art backbones for the tasks of neuronal structure, cell, nuclei, brain tumor, and liver segmentation. UNet++, trained with deep supervision, consistently outperforms U-Net across all backbone architectures and applications under study. By densely connecting the intermediate layers, UNet++ also yields higher segmentation performance than UNet+ in most experimental configurations. The error bars represent the 95% confidence interval and the number of * on the bridge indicates the level of significance measured by *p*-value (“n.s.” stands for “not statistically significant”).

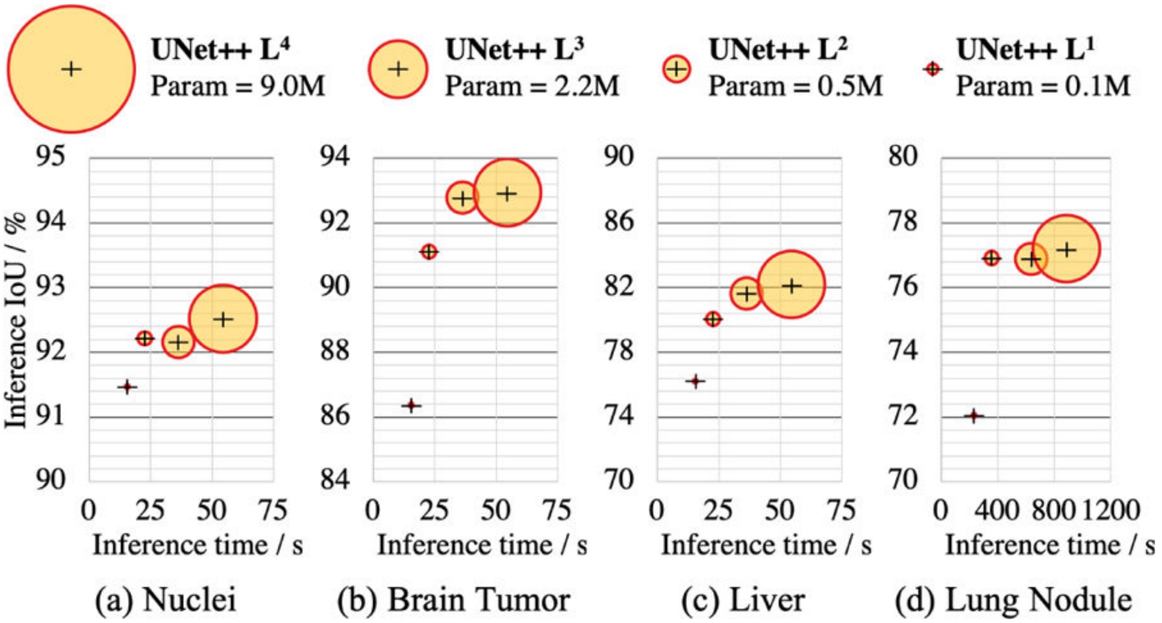


Fig. 5:

Complexity (size ∝ 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.

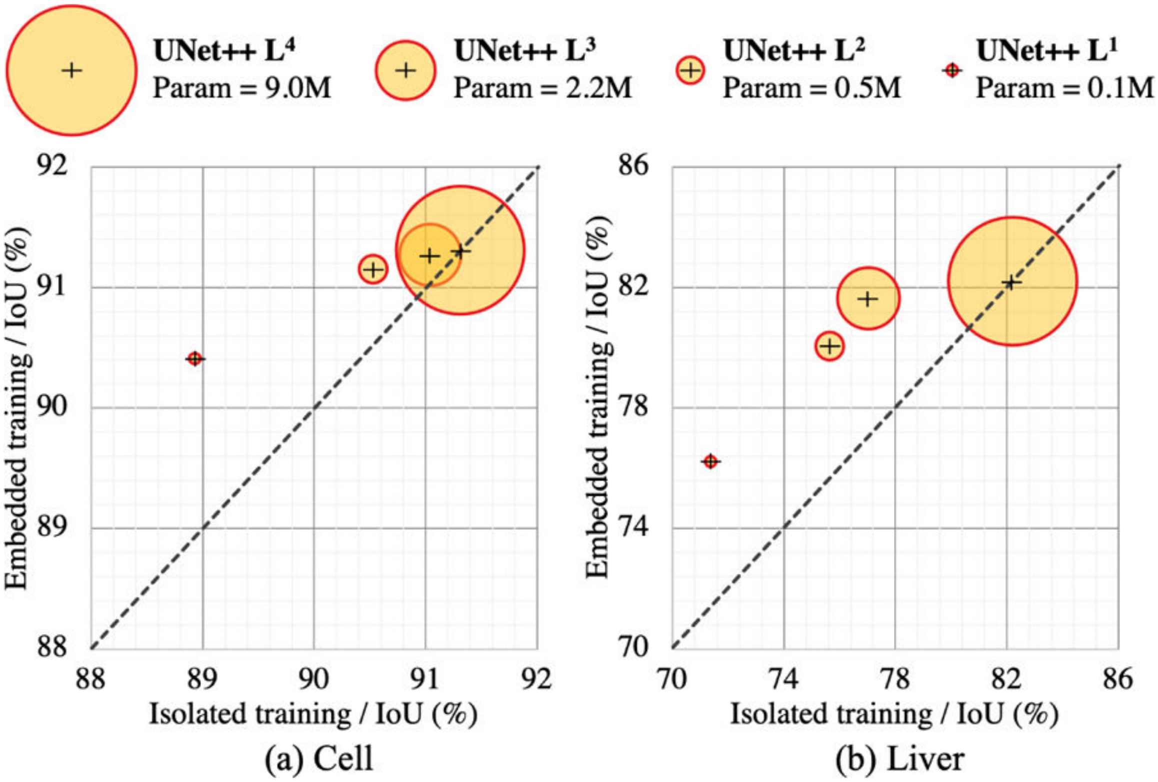


Fig. 6:

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.

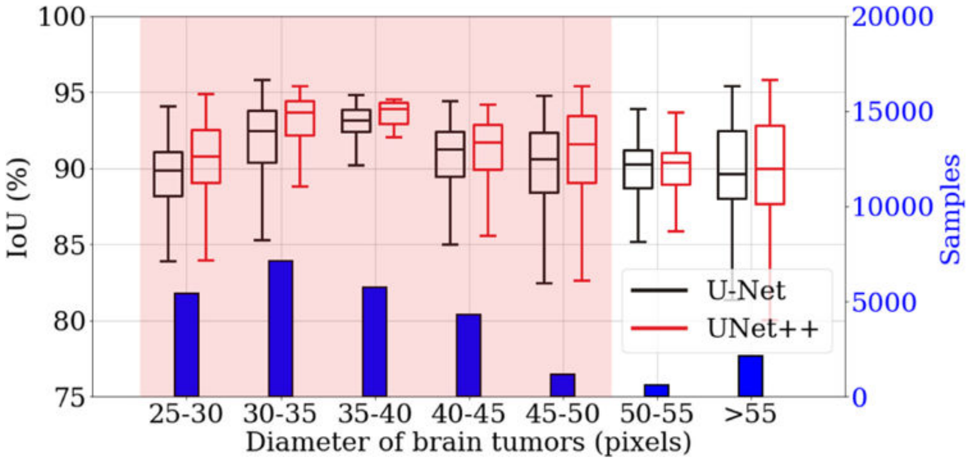


Fig. 7:

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).

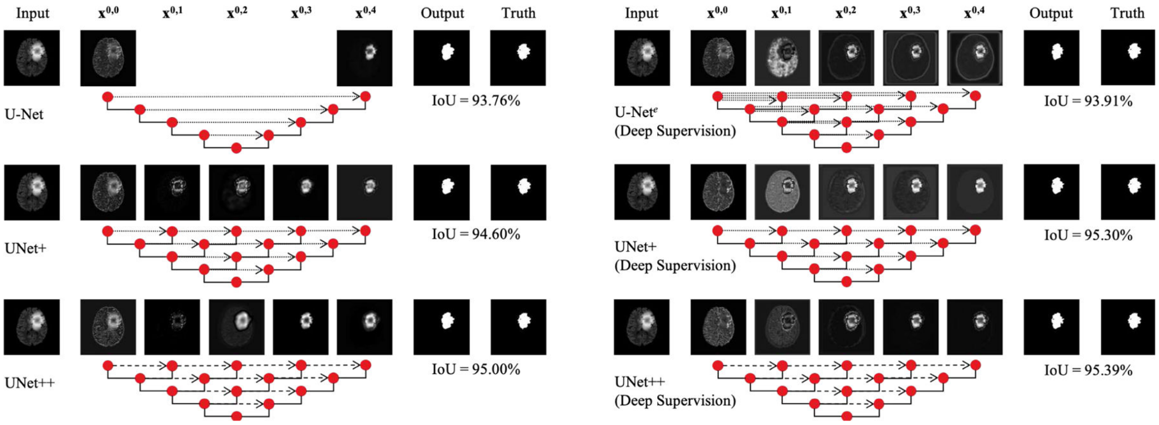


Fig. 8:

Visualization and comparison of feature maps from early, intermediate, and late layers along the top most skip connection for brain tumor images. Here, the dot arrows denote plain skip connection in U-Net and UNet+, while the dash arrows denote dense connections introduced in UNet++.

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	𝒳	0.1M	86.83±0.43	88.58±1.68	86.90±2.25
U-Net L^2	𝒳	0.5M	87.59±0.34	89.39±1.64	88.71±1.45
U-Net L^3	𝒳	1.9M	88.16±0.29	90.14±1.57	89.62±1.41
U-Net (L^4)	𝒳	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+	𝒳	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++	𝒳	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**

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	BraTS2013 [32]
Liver	331	96×96	CT	MICCAI 2017 LiTS
Lung Nodule	1,012	64×64×64	CT	LIDC-IDRI [33]

TABLE III:

Details of the architectures used in our study. Wider version of U-Net and V-Net are designed to have comparable number of parameters to UNet++ and VNet++.

Architecture	**Params**	$\mathbf{X^{0,0}}$ $\mathbf{X^{0,4}}$	$\mathbf{X^{1,0}}$ $\mathbf{X^{1,3}}$	$\mathbf{X^{2,0}}$ $\mathbf{X^{2,2}}$	$\mathbf{X^{3,0}}$ $\mathbf{X^{3,1}}$	$\mathbf{X^{4,0}}$ $\mathbf{X^{4,0}}$
U-Net	7.8M	32	64	128	256	512
wide U-Net	9.1M	35	70	140	280	560
V-Net	22.6M	32	64	128	256	512
wide V-Net	27.0M	35	70	140	280	560
Architecture	**Params**	$\mathbf{X^{0,0-4}}$	$\mathbf{X^{1,0-3}}$	$\mathbf{X^{2,0-2}}$	$\mathbf{X^{3,0-1}}$	$\mathbf{X^{4,0}}$
UNet+	8.7M	32	64	128	256	512
UNet++	9.0M	32	64	128	256	512
VNet+	25.3M	32	64	128	256	512
VNet++	26.2M	32	64	128	256	512

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).

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.