IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 28, 2020, accepted January 11, 2021, date of publication January 21, 2021, date of current version January 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053408

INet: Convolutional Networks for Biomedical Image Segmentation

WEIHAO WENG, (Graduate Student Member, IEEE), AND XIN ZHU, (Senior Member, IEEE)
Biomedical Information Engineering Laboratory, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan

Corresponding author: Xin Zhu (zhuxin@u-aizu.ac.jp)

This work was supported by JSPS KAKENHI Grant Numbers 18K11532 and 18K08010, and Competitive Research Fund of The University of Aizu Grant Number 2020-P-3.

ABSTRACT Encoder–decoder networks are state-of-the-art approaches to biomedical image segmentation, but have two problems: i.e., the widely used pooling operations may discard spatial information, and therefore low-level semantics are lost. Feature fusion methods can mitigate these problems but feature maps of different scales cannot be easily fused because down- and upsampling change the spatial resolution of feature map. To address these issues, we propose INet, which enlarges receptive fields by increasing the kernel sizes of convolutional layers in steps (e.g., from 3×3 to 7×7 and then 15×15) instead of downsampling. Inspired by an Inception module, INet extracts features by kernels of different sizes through concatenating the output feature maps of all preceding convolutional layers. We also find that the large kernel makes the network feasible for biomedical image segmentation. In addition, INet uses two overlapping max-poolings, i.e., max-poolings with stride 1, to extract the sharpest features. Fixed-size and fixed-channel feature maps enable INet to concatenate feature maps and add multiple shortcuts across layers. In this way, INet can recover low-level semantics by concatenating the feature maps of all preceding layers and expedite the training by adding multiple shortcuts. Because INet has additional residual shortcuts, we compare INet with a UNet system that also has residual shortcuts (ResUNet). To confirm INet as a backbone architecture for biomedical image segmentation, we implement dense connections on INet (called DenseINet) and compare it to a DenseUNet system with residual shortcuts (ResDenseUNet). INet and DenseINet require 16.9% and 37.6% fewer parameters than ResUNet and ResDenseUNet, respectively. In comparison with six encoder–decoder approaches using nine public datasets, INet and DenseINet demonstrate efficient improvements in biomedical image segmentation. INet outperforms DeepLabV3, which implementing atrous convolution instead of downsampling to increase receptive fields. INet also outperforms two recent methods (named HRNet and MS-NAS) that maintain high-resolution representations and repeatedly exchange the information across resolutions.

INDEX TERMS Biomedical image, convolutional networks, encoder–decoder networks, semantic segmentation.

I. INTRODUCTION

LeNet [1], AlexNet [2], VggNet [3], GoogleNet [4], ResNet [5], and DenseNet [6] represent a series of breakthroughs in image classification using convolutional neural networks (CNNs). A CNN is a neural network using convolution operations (Conv) in place of general matrix multiplication [7] in at least one layer (a Conv-layer). Semantic segmentation, also called pixel-level classification, is the task of predicting the corresponding category for each pixel in a digital image and outputting a pixelwise mask for each object in the image. Biomedical image segmentation is a critical step in biomedical image processing. It aims to provide a reliable basis for pathology research and to assist doctors in making clinical diagnoses more accurately. Recently, CNN-based systems have achieved great success in automated biomedical image segmentation. Application areas include brain magnetic resonance imaging (MRI) image segmentation [8], lung segmentation on chest x-ray images [9], cell segmentation in electron microscope recordings [10], and dermoscopic image segmentation [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Fahmi Khalifa.

A. ENCODER–DECODER NETWORKS

The most widely employed CNNs architectures for image segmentation are variants of so-called "encoder–decoder networks" proposed initially in [12] for unsupervised

feature learning. These encoder–decoder networks can obtain lower-level spatial resolution features together with a deep perception of the image (semantic recognition) via downsampling. They can also obtain higher-level spatial resolution features for highly accurate recovery of the image via upsampling.

The general semantic segmentation task is to partition an image into a set of coherent regions that are connected and nonoverlapping, and that enable homogeneous pixels to be clustered together [13]. A fully convolutional network (FCN) [14] is an end-to-end image segmentation method modified from VggNet [3] through replacing the fully connected layer in a classification network by a transposed Conv-layer. The single-layer deep decoder of the FCN is efficient when computing low-resolution representations and for coarse-segmentation map estimation. SegNet [15] improves upon FCN by using a 13-layer deep decoder to recover images, corresponding to a 13-layer deep encoder used for feature extraction.

B. BIOMEDICAL IMAGE SEGMENTATION

In the field of biomedical image analysis, the desired output often lies in distinguishing only interesting areas in an image, such as tumor regions [16] or organs [17]. Segmenting regions of interest enables doctors to analyze only the significant parts of otherwise incomprehensible multimodal biomedical images [18]. In practice, diagnosis only leverages the features extracted from segmented images [19]. Furthermore, biomedical image segmentation demands a higher accuracy than that for natural images. While an imprecise segmentation mask may be unimportant in a natural image, even marginal segmentation errors in a biomedical image may cause unreliable results in a clinical setting.

While improvements in the accuracy and quality of segmentation are of great importance in the domain of biomedical image processing, the acquisition of biomedical images is expensive and complicated, and accurate annotation is also difficult. Deep CNNs may suffer from overfitting problems when there are insufficient training data, motivating the introduction of UNet [20] to improve performance with very few annotated images. The encoder and decoder in UNet are symmetrical, as for SegNet, but comprise fewer Conv-layers (8 versus 13). In UNet, the output feature map of the Conv-layer in the encoder is copied and concatenated with the input feature map of the corresponding Conv-layer in the decoder. This concatenation aims to provide high-level spatial local information together with high-level semantic global information, which has been shown to be important for biomedical image segmentation [21].

C. RELATED WORK

We review encoder–decoder networks related to feature fusion from two aspects: spatial information recovering and multilevel semantics exploiting.

1) SPATIAL INFORMATION RECOVERING

Encoder–decoder networks are notorious for the fact that pooling causes much loss of valuable information and ignores the relationship between parts and wholes. Max-pooling is a widely used technique for downsampling in CNNs. Max-pooling separates feature maps into nonoverlapping regions, and outputs the maximum value from each region. This thereby causes the losing of spatial information that could be valuable. Several existing methods have tried to refine the coarse high-level semantics by exploiting high-level spatial resolution information. Stacked hourglass networks [22] implement repeated bottom-up and top-down processing in conjunction with multiscale fusion. Deeply fused nets [23] fuse intermediate representations of shallow layers as the input to deeper layers. HRNet [24] merges the representations produced by subnetworks with high-level resolution as the input to other parallel subnetworks. The global convolutional network [25] encodes rich spatial information from input images by using skip connections with large kernels. In general, to recover spatial information using encoder–decoder networks, modern methods concatenate the features of multiple layers before prediction computation [26]–[28].

2) MULTILEVEL SEMANTICS EXPLOITING

Higher-level semantics has a great impact on the output of an encoder–decoder network. In addition to high-level spatial information, feature fusion also recovers low-level semantics. ResNet [5] adds lower-level semantic input feature maps to higher-level semantic output feature maps to avoid degradation problems brought by increasing depth. DenseNet [6] concatenates multilevel semantics with the same level spatial information to produce smoother decision boundaries. H-DenseUNet [29] shows how low-level semantics can mitigate the difficulty of training encoder–decoder networks for biomedical image segmentation via improved information flow and parameter efficiency.

D. MOTIVATION

Encoder–decoder networks employ feature fusion to recover spatial information and exploit multilevel semantics, from which two problems arise. First, the feature maps of the deep Conv-layer contain lower-level spatial information used to recover the fused feature maps. Second, feature fusion methods only provide the semantics of feature maps at the same level of resolution. These problems are difficult to solve, because elementwise addition and channel concatenation lead to an unnecessarily restrictive fusion scheme, forcing aggregation only for same-scale feature maps. As encoder–decoder networks reduce the scale of feature maps by downsampling and increase them by upsampling, only the corresponding feature maps of the encoder and decoder are at the same scale.

UNet lacks global spatial information and multilevel semantics. As a result, UNet analyzes images pixel-wise and uses color contrasting to distinguish objects.

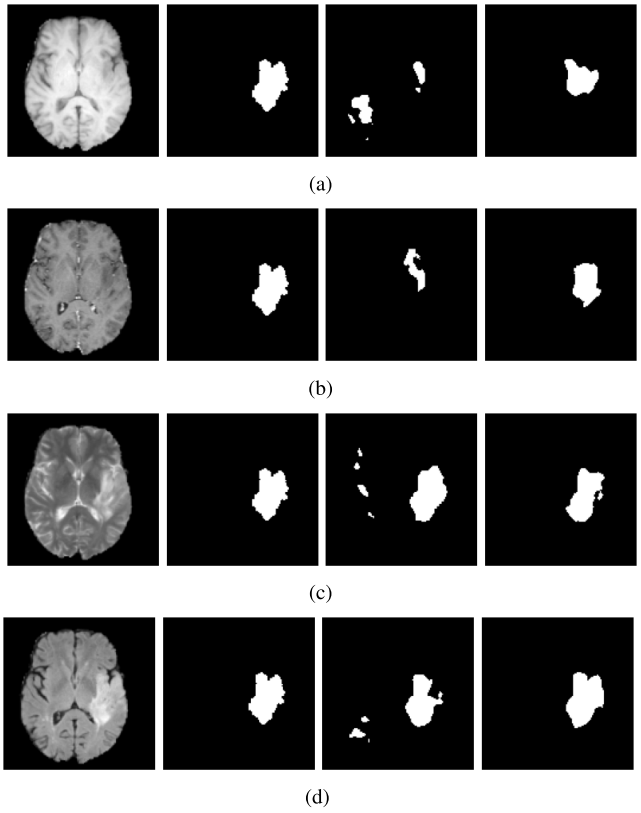

FIGURE 1. Qualitative results for (a) T1w, (b) T1C, (c) T2w, and (d) FLAIR image of the BraTS2013 test set. From left to right: input image, ground truth, ResUNet, and INet.

However, employing different colors to make objects stand out does not always enhance the boundaries of tumors. Fig. 1 shows some qualitative results using the BraTS2013 test set. The BraTS2013 dataset [30] provides four MRI modalities, including T1-weighted (T1w), T1w contrast-enhanced (T1C), T2-weighted (T2w), and fluid attenuation inversion recovery (FLAIR). Tumor is more noticeable in T2w (Fig. 1(c)) and FLAIR (Fig. 1(d)) images than those in T1w (Fig. 1(a)) and T1C (Fig. 1(b)) images. Networks for pixel-wise analysis are sensitive to color contrasting (e.g., in the T2w image). As a result, ResUNet wrongly segments the regions filled with cerebrospinal fluid (CSF), which is bright. Therefore, although ResUNet successfully identified tumor via strong signals in T2w and FLAIR images, it failed for T1w and T1C images. However, our proposed INet can distinguish tumor regions from healthy brain tissue when color contrasting is weak.

In this paper, we propose the INet architecture, which has the following advantages:

- INet maintains spatial information by fixing the sizes of feature maps. Instead of expanding receptive fields by downsampling, INet mimics an expanding procedure by gradually increasing the kernel size of a Conv-layer.
- INet fuses multilevel semantics by concatenating the feature maps of all preceding layers.
- INet enhances its optimization capability by enabling customized residual shortcuts and providing the proposed convolutional index.
- Derivative models of INet can be further developed for biomedical image segmentation.

II. METHOD

A. RECEPTIVE FIELD

The receptive field of a feature in the feature map is composed of an input space where the feature is extracted, and a group of feature maps that form the receptive field for a Conv-layer in CNNs. The Conv changes theoretical receptive fields (TRF) by the process as follows [31]:

$$\begin{aligned} j_i &= j_{i-1} \cdot s \\ r_i &= r_{i-1} + (k-1) \cdot j_{i-1} \end{aligned} \tag{1}$$

where r_i and r_{i-1} denote the sizes of the receptive field for the output features of the i^{th} and $(i-1)^{th}$ Convs, respectively; j_{i-1} and j_i represent the distance of two adjacent features in the output feature maps of the i^{th} and $(i-1)^{th}$ Convs, respectively; k is the kernel size, and s the stride size of the i^{th} Conv, respectively. The proposed method stacks Convs with stride 1×1 and increases the kernel sizes from 3×3 to 7×7 and then 15×15 in steps to follow the Convs used for downsampling, i.e., 3×3 Conv with a stride 2×2. Fig. 2(b) shows the receptive field changes for 3×3 Convs with stride 2×2. The first layer is the input layer, giving r_0 and j_0 as 1×1 and 1, respectively. By applying (1), r_1 is 3×3, r_2 7×7, and r_3 15×15. This is similar to Atrous convolution [32], but Atrous convolution increases the TRF while controlling the resolution of features by inserting zeros between the kernels. As a result, Atrous convolution has a so-called ''gridding issue''. That is, by padding zeros between two pixels in a kernel, the TRF for this kernel covers the area with checkerboard patterns (i.e., only sampling locations with nonzero values). This causes local spatial information loss.

The effective receptive field (ERF) [33] is the actual receptive field, occupying only a small fraction of the TRF and being affected mainly by the centural pixels of the receptive field. As the number of stacked Conv-layers grows, the size of the ERF relative to the TRF shrinks at a rate of $O(1/\sqrt{n})$. Stacked small filters such as 1×1 or 3×3 are more efficient than larger kernels, given the same computational complexity. However, only shrinking feature maps or increasing kernel sizes will expand the ERF in extreme cases. To address this issue, INet increases the kernel sizes of Conv-layers from 3×3 to 7×7 and then to 15×15.

B. NETWORK ARCHITECTURE

As shown in Fig. 2, INet comprises three convolutional subnetworks (Conv-subnetworks): 3×3, 7×7, and 15×15 Conv-subnetworks. Each Conv-subnetwork includes three Conv-layers, which are composed of 1–4 Convs. As shown in Fig. 2(c), the 2^{nd} Conv-layer is within the 3×3 Conv-subnetwork and contains two 3×3 Convs. The output feature

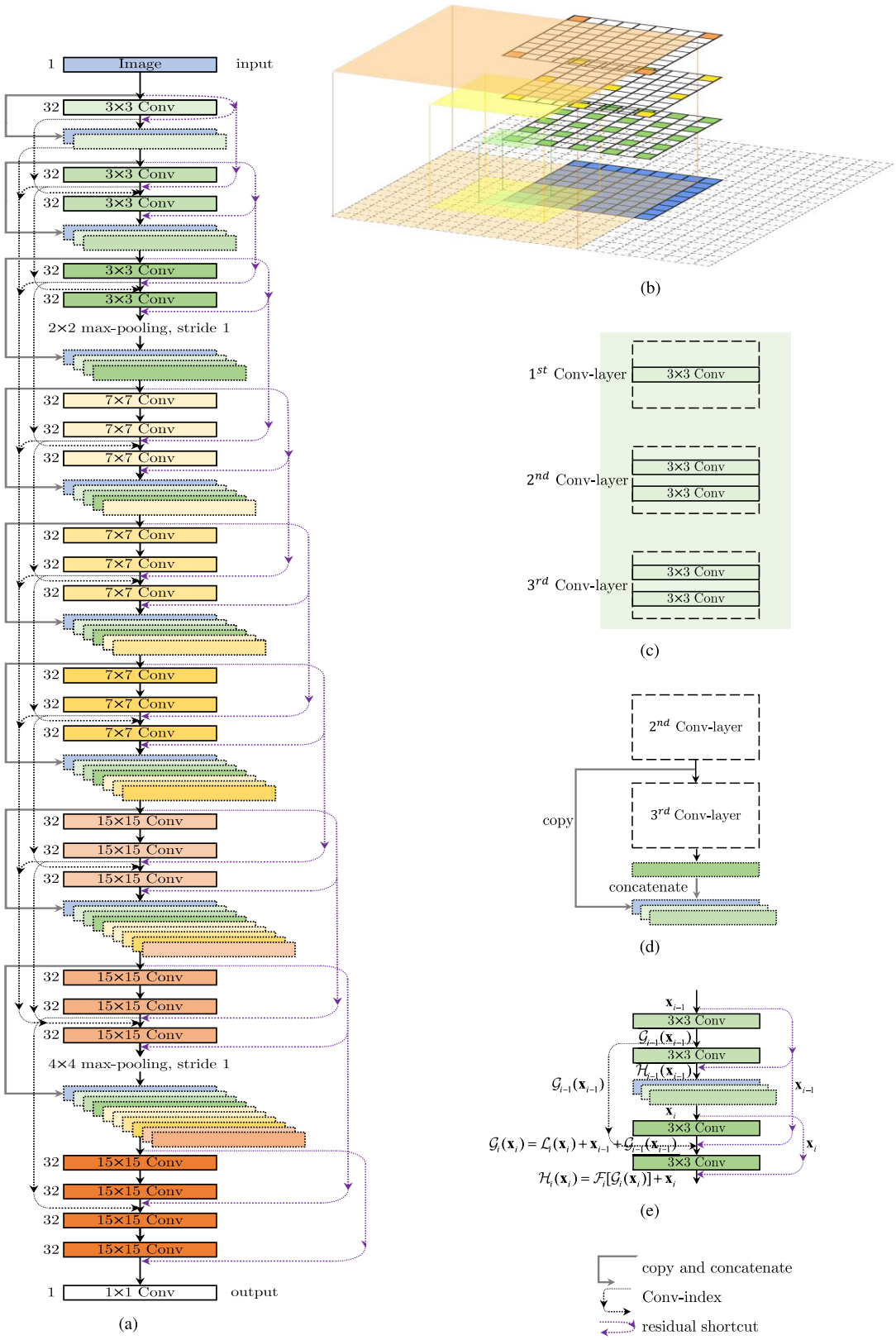

FIGURE 2. (a) The INet architecture. The blue box indicates the input image. Each green, yellow, and orange box corresponds to a multichannel feature map produced by 3 × 3, 7 × 7, and 15 × 15 Convs, respectively, with stride 1 × 1. The number of channels is denoted on left of the box. (b) The theoretical receptive field of 3 × 3 Convs with stride 2 × 2 that the proposed method intends to follow. (c) The Conv-subnetwork. (d) The concatenation operation. (e) The customized residual shortcuts and the convolutional index example of INet.

maps of a Conv-layer, expressed by $\mathcal{H}(\mathbf{x}_i)$, concatenate the input feature maps of the Conv-layer, represented by $\mathbf{x}_i$, to become the input to the next Conv-layer, indicated by $\mathbf{x}_{i+1}$. Denoting the input image as $\mathbf{x}_0$, then we get:

$$\mathbf{x}_{i+1} = \text{concat}[\underbrace{\mathbf{x}_0, \mathcal{H}(\mathbf{x}_0), \mathcal{H}(\mathbf{x}_1), \ldots, \mathcal{H}(\mathbf{x}_{i-1})}_{\mathbf{x}_i}, \mathcal{H}(\mathbf{x}_i)]. \quad (2)$$

In addition to fusing multilevel semantics, reusing the input feature maps of all preceding layers helps the different kernels extract features from one input feature map. The inputs to the 3×3 and 7×7 Conv-subnetworks will be used to extract features by kernels of three and two different sizes, respectively. Previous work on cortex-like mechanisms [34] and spatial pyramid pooling [35] have shown that using kernels of different sizes to extract features at different scales can enable the fusion of features to obtain a better representation of the image. In our proposed architecture, we extract features from each image by stacked Inception-like modules [4], which contain 3×3, 7×7, and 15×15 Convs in parallel. Inception-ResNet [36] shows that residual shortcuts can accelerate the training of CNNs with Inception-like modules. Because INet has fixed-size and fixed-channel feature maps, we introduce multiple shortcuts across the Conv-layers to accelerate the training process.

1) CUSTOMIZED RESIDUAL SHORTCUT

As shown on the righthand side of Fig. 2(e) and denoting the desired underlying mapping of the second-last Conv and last Conv of a Conv-layer as $\mathcal{G}(\mathbf{x}_i)$ and $\mathcal{H}[\mathcal{G}_i(\mathbf{x}_i)]$, respectively, we let the stacked Convs fit another mapping: $\mathcal{L}_i(\mathbf{x}_i) = \mathcal{G}_i(\mathbf{x}_i) - \mathbf{x}_{i-1}$ and $\mathcal{F}_i[\mathcal{G}_i(\mathbf{x}_i)] = \mathcal{H}_i[\mathcal{G}_i(\mathbf{x}_i)] - \mathbf{x}_i$, respectively. If $\mathcal{L}_i(\mathbf{x}_i)$ saturates (i.e., $\mathcal{G}_i(\mathbf{x}_i) = \mathbf{x}_{i-1}$), INet tries to optimize $\mathcal{F}_i(\mathbf{x}_{i-1}) = \mathcal{H}_i(\mathbf{x}_{i-1}) - \mathbf{x}_i$ rather than skip the last Conv and perform identity mapping as the original residual shortcut (i.e., $\mathcal{H}_i[\mathcal{G}_i(\mathbf{x}_i)] = \mathbf{x}_i$).

2) CONVOLUTIONAL INDEX

As shown on the left-hand side of Fig. 2(e), the convolutional index (Conv-index, $\mathcal{G}_{i-1}(\mathbf{x}_{i-1})$) enables INet to skip the intermediate Convs between the second-last Conv of the 2^{nd} Conv-layer and the last Conv of 3^{rd} Conv-layer. We can consider the concatenation of the feature maps as giving equal importance to all preceding Conv-layers in INet. The Conv-index is then putting a larger weight on the output feature maps of the previous Conv-layer, which contains the highest-level semantics. In the extreme, Conv-index lets INet remove the concatenation of feature maps. Formally, in this paper, we consider a Conv-layer to be defined as:

$$\begin{aligned} \mathcal{G}_i(\mathbf{x}_i) &= \mathcal{L}_i(\mathbf{x}_i) + \mathbf{x}_{i-1} + \mathcal{G}_{i-1}(\mathbf{x}_{i-1}) \\ \mathcal{H}_i(\mathbf{x}_i) &= \mathcal{F}_i[\mathcal{G}_i(\mathbf{x}_i)] + \mathbf{x}_i \end{aligned} \quad (3)$$

3) OVERLAPPING MAX-POOLING

Encoder–decoder networks not only implement max-pooling for downsampling but also for extracting the sharpest features of images by indicating the contrast of adjacent regions. Nonoverlapping max-pooling discards information about the position of the maximal value, leading to coordinate transform problems and spatial information loss [37]. Therefore, INet adopts two max-poolings of stride 1. They overlap and preserve the positions of features by coarse coding [38].

We had tried to adopt max-pooling at the end of each Conv-subnetwork, like max-pooling at the Conv of the encoder. However, when the last layer (before the output layer) is max-pooled, the network loses many nonmaximal features, and the maximal value may not be the most valuable. Therefore, we merge the max-pooling at the 2^{nd} and 3^{rd} Conv-subnetworks to the end of the second-last Conv-layer, and add one more Conv to the end of the last Conv-layer. When the Conv at the end of the last Conv-layer extracts the maximal value, it performs similarly to max-pooling. Max-pooling would extract values from a larger portion of an input image as downsampling reduces the size of the feature maps. In addition to the 2×2 max-pooling in the encoder–decoder network, the pool size of the second max-pooling of INet is 4×4.

III. EXPERIMENTS

A. BASELINE MODELS

Compared to the commonly used baseline model, UNet, the proposed INet has additional residual shortcuts, and therefore the learning task of the optimizer becomes easier [39] (We denote INet without customized residual shortcuts and the Conv-index as Plain-INet). As a baseline model, we adapted the original U-Net to include residual shortcuts (ResUNet), thereby ensuring that any improvements achieved by INet would not be attributed only to its implementation of residual shortcuts. The Atrous convolution (dilated convolution) is widely used to increase the receptive field while avoiding the downsampling operations. We selected DeepLabV3 [32] for comparison to better support our claim that the ''gridding issue'' leads to worse segmentation performance. INet is compared with two recent methods named HRNet and MS-NAS [40], respectively. HRNet maintains high-resolution representations through the whole process like INet to improve segmentation results. MS-NAS is a multi-scale neural network architecture search framework for biomedical image segmentation. We also added residual shortcuts to the original DenseUNet [41], giving Res-DenseUNet as another baseline model to compare with INet equipped with dense connections (DenseINet). This was to test the idea that INet can serve as an alternative backbone architecture for biomedical image segmentation. The input of each Conv of DenseINet is the concatenation of the feature maps for all preceding layers in the same Conv-layer. Table 1 gives the details of the models used in our experiments. Most CNNs, including UNet and DenseUNet, use downsampling. This reduces the size of feature maps by a factor of 4 (for a stride of 2). To maintain the same number of features, models have to augment the number of channels before pooling to compensate reduced resolution. The opposite situation

TABLE 1. Proposed models against alternative models based on UNet.

Method	Parameters	The Number of Channels								
ResUNet (baseline)	8,984,961	1^{st}: 32	2^{nd}:64	3^{rd}: 128	4^{th}:256	5^{th}:512	6^{th}:256	7^{th}:128	8^{th}:64	9^{th}:32
INet (proposed)	7,467,425	1^{st}: 32	2^{nd}:32	3^{rd}: 32	4^{th}:32	5^{th}:32	6^{th}:32	7^{th}:32	8^{th}:32	9^{th}:32
ResDenseUNet (baseline)	16,038,625	1^{st}: 32	2^{nd}:64	3^{rd}: 128	4^{th}:256	5^{th}:512	6^{th}:256	7^{th}:128	8^{th}:64	9^{th}:32
DenseINet (proposed)	10,011,041	1^{st}: 32	2^{nd}:32	3^{rd}: 32	4^{th}:32	5^{th}:32	6^{th}:32	7^{th}:32	8^{th}:32	9^{th}:32

occurs in the decoder, which reduces the number of channels in steps. Because INet and DenseINet have no down- and upsampling operations, they keep the number of channels unchanged at 32.

B. DATASETS

To demonstrate the generalization of the proposed method in the segmentation of biomedical images, we carried out experiments on nine public imaging datasets: three brain (BraTS2013, LMS, BTD) one heart MRI [42]–[44], one liver one spleen CT [44], [45], one lung X-ray [9], one colon endoscopic [46], and one nerve ultrasonic [47].

1) MRI

Glioma constitutes 80% of all malignant primary brain tumors in adults [48]. There are two main groups of gliomas following the classification of the World Health Organization: high- (HGG) and low-grade glioma (LGG), which reflect their differences in patient survival. The LGG cases lack the vascularity of HGG and are visually bland. We conducted comprehensive experiments on three brain MRI images datasets to analyze the effectiveness of the proposed INet for segmenting brain MRI images. The BraTS2013 dataset contains 20 HGG and 10 LGG cases with four MRI modalities: T1w, T1C, T2w, and FLAIR. We trained the models with the 20 HGG cases and tested the performance when segmenting LGG cases. The LMS [42] dataset contains 3,929 FLAIR images from 110 LGG cases. The BTD [43] dataset contains 3,064 T1C images from 233 patients with three kinds of brain tumor: meningioma (708 slices), glioma (1,426 slices), and pituitary tumor (930 slices).

We also tested the performance of our proposed methods when segmenting heart MRI images. Segmentation of the left atrium (LA) is essential for atrial fibrillation ablation guidance, fibrosis quantification, and biophysical modeling. The heart MRI images dataset in [44] includes MRI images from 30 cases covering the entire heart and acquired during a single cardiac phase with masks for the left atrium appendage, the mitral plane, and the portal vein end points.

2) CT IMAGES

The liver CT dataset in [45] contains images of primary tumors, secondary tumors, and metastases. Liver metastases are cancerous tumors that have spread to livers from other parts of body and are more common than primary liver cancers. The spleen is also involved in many different types of pathologic disorder. Recent studies have found that the correlation between hepatic and splenic hypertrophy [49] and between liver and spleen are both critical to maintaining the reticuloendothelial system [50]. This suggests that they may share regulatory pathways. To test this conjecture, we compared all models with respect to segmenting spleen CT images from patients undergoing chemotherapy treatment for liver metastases [44]. The liver and spleen CT datasets each contained contrast-enhanced CT images of 40 randomly chosen cases.

3) X-RAYS

Lung image segmentation is the first step in lung X-ray analysis and plays a vital role in diagnosing lung diseases such as tuberculosis and corona-virus-related pneumonia. The lung X-ray dataset in [9] contained 138 X-rays and corresponding masks.

4) ENDOSCOPIC IMAGES

Colonoscopy is the gold standard for colorectal polyp and cancer screening. Colorectal cancer arises from adenomatous polyps developing in glandular tissues of colonic mucosa. Adenomatous polyps can become malignant over time and spread to both adjacent and distant organs, where they are ultimately responsible for complications and possible death. To segment polyps from endoscopy images, we used a colon endoscopy image dataset with 612 polyp images and their corresponding segmentation masks [46].

5) ULTRASOUND IMAGES

The nerve ultrasound image dataset [47] contains 5,638 nerve ultrasound images with corresponding masks. Regional anesthesia is one of the most frequently undertaken tasks in hospitals in the world. Ultrasound-guided regional anesthesia is a rapidly growing alternative to general anesthesia, following advances in ultrasound imaging technology. Nerve segmentation of ultrasound images is therefore of great clinical significance because any errors in the anesthetic provision may cause lethal damage to the corresponding region of the body or side effects with respect to the rest of body.

C. IMPLEMENTATION DETAILS

The proposed network was implemented using the Keras framework [51] and trained on a NVIDIA Tesla K80 GPU, with an He-Normal initializer, an ADAM optimizer, batch normalization, and a batch size of 16. Weight values decay by a factor of $1 \times e^{-4}$ if validation loss has not improved after four epochs and training stops once validation loss has not improved after ten epochs. The loss function was a combination of binary-cross-entropy and Dice coefficient loss, which

TABLE 2. Segmentation results for the BraTS2013 dataset.

	T1wimages Dice			T1Cimages Dice			T2wimages Dice			FLAIR images Dice		
UNet	39.37			42.22			69.52			72.64		
ResUNet	41.48			46.61			72.60			74.16		
DeepLabV3	39.21			42.63			64.75			64.35		
HRNet	43.21			44.51			72.63			74.28		
MS-NAS	41.10			43.61			71.02			71.82		
Plain-INet	49.96			53.75			70.95			75.39		
INet	**51.21**			**57.98**			**75.55**			**76.14**		
ResDenseUNet	42.53			47.33			80.79			80.15		
DenseINet	**56.16**			**59.83**			**82.62**			**83.61**		
	TPR	TNR	HD95	TPR	TNR	HD95	TPR	TNR	HD95	TPR	TNR	HD95
UNet	38.34	98.96	35.60	37.79	99.24	33.22	76.84	99.27	33.31	77.83	99.30	30.13
ResUNet	45.62	98.76	32.22	45.56	99.29	**29.82**	77.96	**99.40**	29.41	77.30	**99.63**	**23.66**
DeepLabV3	45.55	98.48	37.51	52.20	98.52	33.87	72.73	99.15	**21.86**	**82.03**	98.88	24.55
HRNet	39.85	**99.64**	34.24	46.79	99.17	29.76	78.77	99.27	31.22	77.26	99.60	27.84
MS-NAS	42.55	98.65	33.33	43.90	99.24	29.44	78.71	99.13	37.51	76.54	99.55	31.28
Plain-INet	**52.98**	99.14	35.68	56.37	99.26	39.28	77.07	99.35	36.31	80.10	99.39	26.02
INet	50.24	99.42	**30.75**	**58.30**	**99.36**	35.10	**80.23**	99.35	26.08	81.04	99.55	24.91
ResDenseUNet	47.05	98.55	35.78	49.78	99.21	35.62	81.30	99.66	23.00	78.60	**99.81**	**10.98**
DenseINet	**59.54**	**99.26**	**32.60**	**58.78**	**99.49**	**34.51**	**83.04**	**99.67**	**16.95**	**84.56**	99.74	15.18

TABLE 3. Segmentation results for the BTD dataset.

	Dice	TPR	TNR	HD95
UNet	74.60	68.13	99.81	4.47
ResUNet	75.36	69.36	**99.84**	4.19
DeepLabV3	71.85	69.12	99.73	3.87
HRNet	69.36	63.34	99.79	5.39
MS-NAS	74.86	71.80	99.78	4.43
Plain-INet	76.20	71.83	99.82	4.18
INet	**78.14**	**76.63**	99.79	**3.76**
ResDenseUNet	74.68	70.38	**99.80**	4.21
DenseINet	**77.55**	**75.40**	99.79	**3.93**

TABLE 4. Segmentation results for the LMS dataset.

	Dice	TPR	TNR	HD95
UNet	53.57	58.84	99.64	17.46
ResUNet	59.16	60.37	99.71	11.40
DeepLabV3	59.11	61.79	99.61	12.38
HRNet	62.42	56.43	**99.83**	**8.37**
MS-NAS	62.88	60.57	99.77	9.70
Plain-INet	59.13	65.99	99.55	13.03
INet	**65.98**	**67.52**	99.72	8.63
ResDenseUNet	59.02	58.58	**99.77**	12.10
DenseINet	**68.08**	**69.91**	99.70	**8.55**

TABLE 5. Segmentation results for the Heart dataset.

	Dice	TPR	TNR	HD95
UNet	80.50	79.66	99.96	4.17
ResUNet	85.09	83.02	99.97	1.83
DeepLabV3	65.84	71.14	99.90	3.58
HRNet	75.19	71.59	99.96	2.54
MS-NAS	82.84	83.95	99.95	2.32
Plain-INet	84.42	81.27	99.97	1.83
INet	**87.34**	**85.24**	**99.98**	**1.60**
ResDenseUNet	72.26	66.78	99.97	3.39
DenseINet	**88.15**	**86.63**	**99.98**	**1.50**

TABLE 6. Segmentation results for the Liver dataset.

	Dice	TPR	TNR	HD95
UNet	87.07	86.33	99.79	2.85
ResUNet	91.06	92.06	99.83	1.52
DeepLabV3	87.32	89.62	99.72	2.16
HRNet	91.76	90.73	99.87	**1.19**
MS-NAS	91.48	89.73	**99.88**	1.39
Plain-INet	91.37	93.04	99.78	2.00
INet	**91.82**	**93.46**	99.83	1.21
ResDenseUNet	78.89	72.94	99.81	6.13
DenseINet	**92.79**	**91.96**	**99.90**	**1.00**

is widely used for training UNet and its variants [52]. The loss is as follows:

$$L = -\frac{1}{N}\sum_{k=1}^{N}\left(y_k \log + (1-y_k)\log(1-t_k) + \frac{2y_k \cdot t_k}{y_k + t_k}\right), \tag{4}$$

where y_k is predicted by the network, t_k denotes the ground truth, and N indicates the batch size. It was observed that the binary-cross-entropy loss optimized for pixel-level accuracy whereas the dice loss helped in improving the segmentation quality [53].

All datasets were divided randomly into training (50%), validation (25%), and test (25%) sets. All images were resized to 128 × 128. Data augmentation was performed, including zooming and changing the brightness levels. We used a Dice coefficient (Dice %), sensitivity (TPR %), specificity (TNR %), and a 95% Hausdorff distance (HD95) to validate the performance. In Tables 2 to 10, the best scores are shown in bold.

IV. RESULTS AND DISCUSSION

A. OBJECTWISE AND PIXELWISE LEARNING

The results given in Tables 2 to 10 show that the proposed method nearly always outperformed the baseline models while ResUNet outperformed the original U-Net and DeepLabV3. Therefore, we focus on comparing INet with ResUNet. For the T1C and FLAIR images in BraTS2013,

TABLE 7. Segmentation results for the Spleen dataset.

	Dice	TPR	TNR	HD95
UNet	86.62	84.18	99.96	2.44
ResUNet	88.96	86.23	**99.99**	1.73
DeepLabV3	74.80	77.83	98.81	3.22
HRNet	88.89	89.95	99.96	1.83
MS-NAS	89.40	87.29	**99.99**	1.78
Plain-INet	88.12	87.06	99.96	1.84
INet	**92.72**	**91.67**	**99.99**	**1.06**
ResDenseUNet	75.34	74.55	**99.97**	3.39
DenseINet	**89.38**	**94.61**	99.96	**1.27**

TABLE 8. Segmentation results for the Lung dataset.

	Dice	TPR	TNR	HD95
UNet	83.66	76.29	**99.91**	4.47
ResUNet	90.47	88.18	99.80	1.41
DeepLabV3	86.39	84.47	99.84	2.44
HRNet	88.39	84.32	99.90	2.12
MS-NAS	88.40	85.57	99.88	2.24
Plain-INet	90.88	90.51	99.87	1.71
INet	**91.50**	**91.75**	99.87	**1.24**
ResDenseUNet	95.95	97.10	98.28	0.00
DenseINet	**97.01**	**98.10**	**98.64**	**0.00**

TABLE 9. Segmentation results for the Colon dataset.

	Dice	TPR	TNR	HD95
UNet	75.68	72.90	97.72	4.25
ResUNet	81.95	79.41	98.57	3.00
DeepLabV3	78.22	81.97	97.61	**2.84**
HRNet	76.18	74.63	**98.69**	4.04
MS-NAS	**82.47**	82.65	98.23	3.98
Plain-INet	81.40	77.98	98.61	3.24
INet	82.37	**83.69**	98.06	2.97
ResDenseUNet	79.00	76.80	**98.23**	3.32
DenseINet	**81.86**	**83.00**	98.03	**2.95**

TABLE 10. Segmentation results for the Nerve dataset.

	Dice	TPR	TNR	HD95
UNet	55.08	55.14	99.63	16.05
ResUNet	57.58	57.82	99.46	10.36
DeepLabV3	55.84	54.67	99.66	8.62
HRNet	61.22	61.17	99.55	**6.99**
MS-NAS	**63.17**	60.02	99.62	7.00
Plain-INet	60.64	56.08	**99.67**	7.68
INet	61.14	**61.27**	99.51	7.46
ResDenseUNet	54.65	55.94	**99.39**	14.11
DenseINet	**60.96**	**65.90**	99.37	**9.08**

INet achieved larger Dice coefficients but also a larger HD95 than those for ResUNet. We can note that T1C uses a paramagnetic contrast agent (usually gadolinium-based) to improve the contrast in areas that are affected by hemorrhage. The contrast agent highlights tumor core regions. FLAIR imaging keeps the abnormalities bright while attenuating normal-CSF areas, making the differentiation between CSF and tumor much easier. Therefore, the color contrast in the T1C and FLAIR images is higher than that in the T1w and T2w images, respectively. ResUNet has more max-pooling operations to extract the sharpest features, giving the best lower-level representation of an image. ResUNet can therefore achieve smaller HD95s, which are more sensitive to segmented boundaries.

Even though INet learns less information from the boundaries, it returned a better performance in terms of the Dice coefficients, which are more sensitive to the overlapping between ground truth and predicted masks. Furthermore, for the T1w and T1C images in BraTS2013 where whole tumor regions are not highlighted, INet results in significantly larger Dice scores than ResUNet (Mann-Whitney U test, $p = 0.00$ and $p = 0.00$). These results confirm that the performance improvement with INet can be attributed to its combined multilevel semantics and maintaining spatial information. For the LMS and BTD datasets, both INet and ResUNet exhibited low HD95s but INet performed significantly better with respect to HD95s and Dice coefficients. For the LMS dataset, INet's results were lower in HD95 that those of ResUNet (Mann-Whitney U test, $p = 0.08$). Because of the lower contrast and smaller size, segmenting low-grade glioma (LGG) MRI images is more complex than for high-grade glioma (HGG) [48]. Recent methods have combined three-dimensional UNet with a conditional random field [54] to supply the spatial information. These methods improve the segmentation results for LGG MRI images, but are constrained to three-dimensional MRI images. In contrast, INet can maintain spatial information in two-dimensional MRI images, with our experiments demonstrating that INet is better at segmenting LGG MRI images than ResUNet, irrespective of whether the corresponding training is with HGG MRI images or LGG MRI images.

B. ROBUSTNESS AND INTERPRETABILITY

The results in Table 8 show that ResUNet achieved similar performance with INet. However, we found that INet is more robust against artifacts and noise. Acquisition or preprocessing artifacts and various types of noise in biomedical images make distinguishing target objects from the background more challenging. For example, Figs. 3(b) and 3(d) show that ResUNet classified tissue areas of low signal intensity within target objects as healthy tissues, even though doing so would scatter a continuous segmented region. Fig. 3(f) depicts a more serious problem, i.e., all target objects are lungs and located near the middle of the images in the Lung dataset. We observe that ResUNet's segmentation considered the tissue with an area of low signal intensity surrounded by a mass with high signal intensity as a target object. As a result, ResUNet classified the shadow on the arms as part of the lung. Even though such outliers do not lower the values of segmentation results, they would impair further biomedical image analysis. Because ResUNet did not recognize lungs properly, it obtained a high discrimination score at the cost of a low interpretability of black-box representations. In contrast, a representation learned by INet contains information not only about surrounding elements but also the relationship between the features and the whole image. The spatial information improves not only the robustness but also the trustworthiness of INet for subsequent biomedical image analysis.

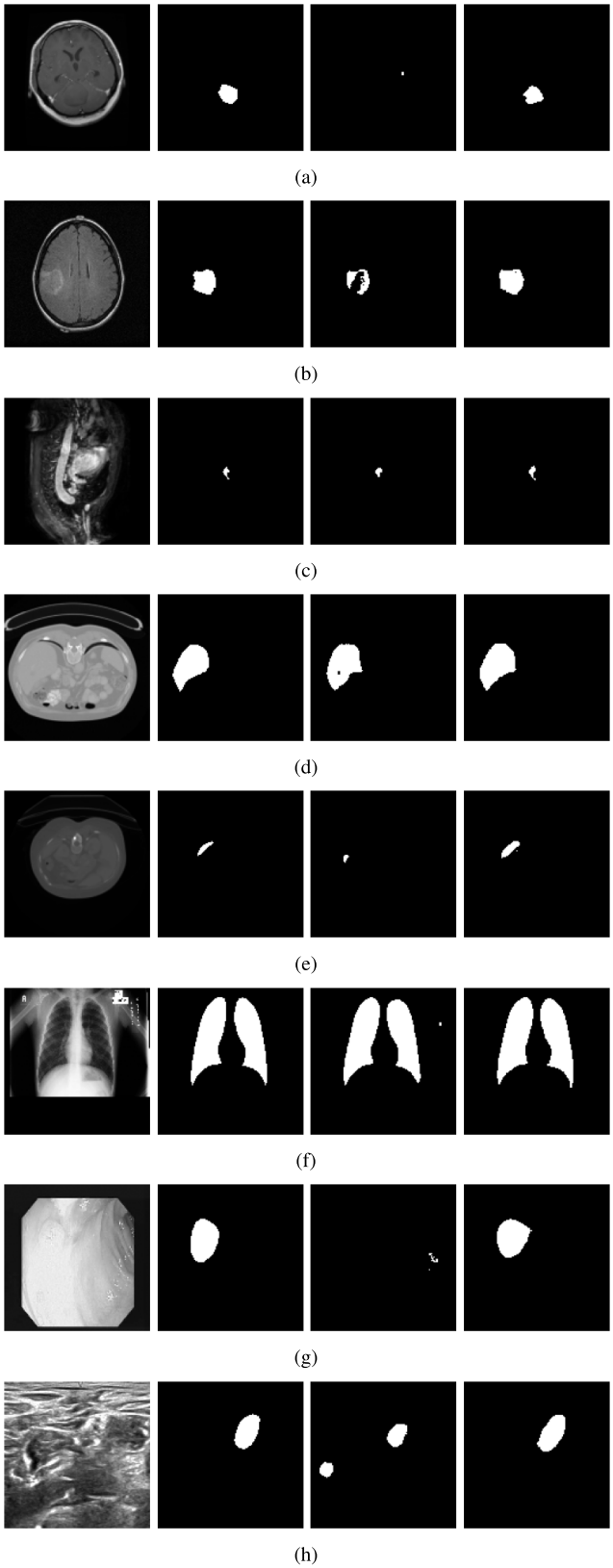

FIGURE 3. Qualitative results for (a) BTD, (b) LMS, (c) Heart, (d) Liver, (e) Spleen, (f) Lung, (g) Colon, and (h) Nerve test sets. From left to right: input image, ground truth, ResUNet, and INet.

C. SENSITIVITY AND SPECIFICITY

The results in Table 9 show that INet returned larger Dice coefficients (Mann-Whitney U test, $p = 0.79$) and higher sensitivities (Mann-Whitney U test, $p = 0.00$) than ResUNet for the Colon dataset. However, ResUNet outperformed INet with respect to specificity (Mann-Whitney U test, $p = 0.00$). Endoscopes have a narrow and poorly illuminated field of view, which often leads to the overexposure of near objects and the underexposure of distant structures. As shown in Fig. 3(g), a polyp within the overexposure area at the left had inconspicuous boundaries. ResUNet categorized the whole area as negative and turned to search the correctly exposed area. In contrast, lacking clear boundaries did not deter INet from searching more widely. INet persisted and finally found an adenomatous polyp. We can also observe this phenomenon of a sensitivity increase accompanied by a specificity decrease in the segmentation results for the BTD dataset. Fig. 3(a) depicts an example of qualitative results for the BTD dataset. A tumor is gray and not apparent, but a lateral ventricle filled with CSF is dark. ResUNet classified the tumor as healthy brain tissues and segmented the lateral ventricle with low signal intensity. However, INet analyzes objectwise and therefore tries to identify images by other means when the boundary of the tumor appears to be missing. We consider that this characteristic enabled INet to outperform ResUNet in the segmentation of the T1w and T1C images in BraTS2013 containing tumors with no distinct edges, such as Figs. 1(a) and 1(b).

For clinical purposes, we should guarantee a high sensitivity with a reasonable specificity. Furthermore, for people who have already complained symptoms, the nonrecognition of cancerous areas could lead to delayed treatment possible with worse outcomes. Therefore, a high sensitivity is usually desired in medical diagnosis, even at the cost of a slight decrease in specificity. In such circumstances, INet would be preferable to UNet as a backbone architecture for biomedical image segmentation. For example, high-sensitivity colorectal polyp detection is more valuable than accurate segmentation during colonoscopy procedures for the early screening of colorectal cancer.

D. SPATIAL INFORMATION AND SEMANTICS

1) DENSE CONNECTIONS

In terms of HD95, INet outperformed ResUNet except for the T1C and FLAIR images in BraTS2013, but dense connections enabled DenseINet to outperform ResDenseUNet when segmenting T1C images (Mann-Whitney U test, $p = 0.85$). As shown in Fig. 3, T1C imaging does not always highlight tumor cores, whereas FLAIR images consistently show the boundary of the whole tumor. Dense connections help a network consider all preceding layers in the same Conv-layer (the original INet considers the output feature maps of all preceding Conv-layers). This strengthens the relationship

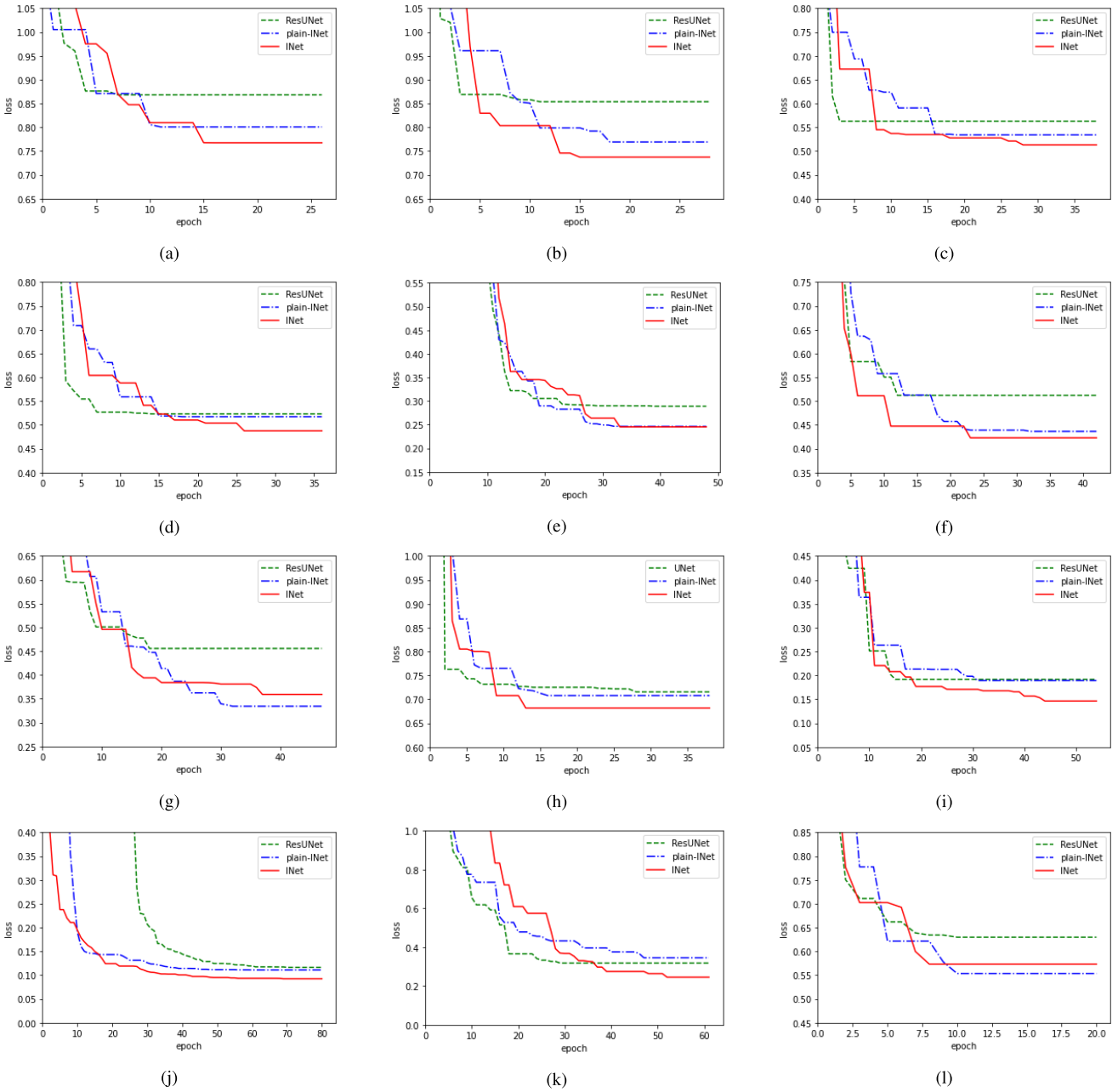

FIGURE 4. Progress of the validation loss with the number of epochs when training on (a) T1w, (b) T1C, (c) T2w, and (d) FLAIR image of the BraTS2013, (e) BTD, (f) LMS, (g) Heart, (h) Liver, (i) Spleen, (j) Lung, (k) Colon, and (l) Nerve dataset. **Green:** ResUNet. **Blue:** Plain-INet. **Red:** INet.

between the part and the whole within a layer. Therefore, DenseINet was better to recognize tumors whose color contrast did not make boundaries stand out.

2) MULTIPLE SHORTCUTS

Fig. 4 compares INet to ResUNet and Plain-INet with respect to validation loss. Two results stand out. First, although INet was easy to optimize, its counterpart Plain-INet exhibited a higher validation loss than that of ResUNet when Plain-INet was trained on the Colon dataset. Second, INet outperformed Plain-INet except when segmenting heart MR and nerve ultrasound images. Plain-INet outperforming ResUNet demonstrates that spatial information and multilevel semantics can affect the results of segmenting MR, endoscopy, and ultrasound images. INet performed better than Plain-INet. This indicates that customized residual shortcuts and the Conv-index improve the training process for tasks involving the segmentation of LGG MR, CT, X-ray, and endoscopy images.

The results in Tables 6 and 7 demonstrate that all INet and UNet based networks could achieve similar scores when segmenting liver and spleen images. For example, INet achieved around 92% Dice for the Liver and Spleen datasets. Dense connections enable networks to consider all preceding

layers within the same Conv-layer. However, the results in Tables 6 and 7 demonstrate that ResDenseUNet performed less well than ResUNet when segmenting liver and spleen images. This indicates that the low-level semantics of liver and spleen images offer little benefit to the segmentation task. Furthermore, Figs. 4(h) and 4(i) show that Plain-INet had similar validation losses to those of ResUNet because they both involve the output feature maps of all preceding Conv-layers. Recent studies have found that the correlation between hepatic and splenic hypertrophy [49] and between liver and spleen are both critical to maintain the reticuloendothelial system [50]. This suggests that they may share regulatory pathways, and our results also demonstrate that liver and spleen CT images share common features.

E. INet AND THE STATE-OF-THE-ART METHODS

INet outperformed HRNet and MS-NAS except when segmenting colon endoscopy and nerve ultrasound images in terms of the Dice coefficients. As shown in Tables 9 and Tables 10, INet returned the second largest and the third largest Dice coefficients, but the highest sensitivities for the Colon and Nerve datasets. Because INet, HRNet and MS-NAS all maintain high-resolution feature maps through the whole process, we argue that INet achieved high sensitivities is attributed to the large kernels (larger than 3×3). The computational complexity of a single Conv-layer [55] can be measured by:

$$\mathrm{O}\left(M^2 \cdot K^2 \cdot C_{in} \cdot C_{out}\right) \tag{5}$$

where M is the side length of each convolution kernel's output feature map. K is the side length of each convolution kernel Kernel. C_{in} and C_{out} are the number of input and output channels, respectively. INet uses large kernels and thus requires more running time. When models are trained using the same terminal condition in Keras with NVIDIA Tesla K80 GPU, the time cost of INet is 1.67, 4.55, 5.28 and 5.71 times compared with those of HRNet, DeepLabv3, ResUNet and UNet, respectively. Since the present works' primary goal is not for segmentation in real-time, a high sensitivity is preferred, even at the cost of a little more runtime. The Dice scores of HRNet for all test datasets except for the nerve dataset are less than those of INet and DenseINet. The INet also uses fewer parameters than HRNet (7.5M *vs.* 28.5M). As for the MS-NAS, it requires notably more search times to train thousands of candidate models for a number of epochs throughout the search.

Given the same computational complexity, stacking small filters, typically 3×3, is more efficient than using a large kernel. However, one of recent trends [4], [35] in network architecture design is concatenating parallel kernels of different sizes. Using kernels of different sizes to extract features at different scales can enable the fusion of features to obtain a better representation of image. Furthermore, in the field of semantic segmentation, where we need to perform dense per-pixel prediction, a large kernel is crucial to relieve the contradiction between classification and localization [25]. Instead of concatenating parallel kernels of different sizes, INet stacks kernels of different sizes, and thus avoids down-sampling operations. INet extracts features by kernels of different sizes by concatenating the output feature maps of all preceding Conv-layers. The size of each Conv-layer's kernels is derived from effective receptive fields. In contrast, the architectures of Inception-like modules (e.g., the number of parallel paths, the number of stacked kernels each path, and the size of kernels) lack interpretability.

In addition, we explain why we did not compare the proposed method with state-of-the-art systems for each dataset. Specialized state-of-the-art models for biomedical image segmentation exist and are all based on the UNet backbone architecture [52], [56]. However, this paper presents INet as an alternative to UNet as a backbone architecture. Therefore, we focused on comparing INet directly with UNet and present DenseINet as an example of INet also being adaptable in the same way as is UNet. INet aims to maintain the same spatial resolution between the parts and the whole. In this respect, even though INet did not achieve state-of-the-art performance in all indexes for every task, it demonstrated higher consistencies among feature maps than the encoder–decoder networks.

Validation methods may also cause the difference of index values. We trained and tested the proposed method through randomly dividing cases into training (50%), validation (25%), and test (25%) sets, i.e., the images from a specific case only belong to one of the three sets. However, some other investigations only used training (80%) and test (20%) sets. In one example [42], a Dice score of 82% was achieved by Plain-UNet for the LMS dataset. In our experiments, we tested Plain-UNet with the LMS dataset and achieved a similar score (Plain-UNet and INet achieved Dice scores of 82.40% and 87.73%, respectively) when splitting all images into training (80%) and test (20%) sets.

V. CONCLUSION

This paper presents INet as a backbone architecture for biomedical image segmentation. INet expands the receptive fields by gradually increasing the kernel sizes of Conv-layers to retain spatial information. INet also fuses the multilevel semantics by concatenating the feature maps of all preceding layers and improves the training process by adding multiple shortcuts. This paper also presents a variant of INet (called DenseINet) that is equipped with dense connections. We have tested our models against alternative models based on UNet for nine distinct biomedical image applications. Because INet and DenseINet have no down- and upsampling operations, they maintain the same number of features by keeping the number of channels unchanged at 32. INet and DenseINet require 16.9% and 37.6% fewer parameters than ResUNet and ResDenseUNet, respectively, and achieve consistent performance improvements. Its dense connections help DenseINet outperform INet with respect to LGG, heart, liver, and nerve segmentation. INet outperforms an atrous

convolution-based method named DeepLabV3, which inserts ''holes''(zeros) between pixels in convolutional kernels to enlarges the receptive fields but also has a so-called ''gridding issue.'' INet outperforms two recent methods named HRNet and MS-NAS. HRNet and MS-NAS maintain high-resolution feature maps through using kernels no larger than 3×3. The experimental comparison shows that large kernels in INet are more feasible for extracting features in biomedical image segmentation.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[3] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[5] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[8] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, ''Deep learning for brain MRI segmentation: State of the art and future directions,'' *J. Digit. Imag.*, vol. 30, no. 4, pp. 449–459, Aug. 2017.

[9] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, ''Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration,'' *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[10] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, ''Improved automatic detection and segmentation of cell nuclei in histopathology images,'' *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, Apr. 2010.

[11] B. S. Lin, K. Michael, S. Kalra, and H. R. Tizhoosh, ''Skin lesion segmentation: U-Nets versus clustering,'' in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.

[12] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, ''Unsupervised learning of invariant feature hierarchies with applications to object recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[13] A. K. Ray and T. Acharya, *Information Technology: Principles and Applications*. New Delhi, India: PHI Learning Pvt. Ltd., 2004.

[14] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, ''SegNet: A deep convolutional encoder-decoder architecture for image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, ''Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),'' in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[17] J. Yang, H. Veeraraghavan, S. G. Armato, K. Farahani, J. S. Kirby, J. Kalpathy-Kramer, W. van Elmpt, A. Dekker, X. Han, X. Feng, P. Aljabar, B. Oliveira, B. van der Heyden, L. Zamdborg, D. Lam, M. Gooding, and G. C. Sharp, ''Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017,'' *Med. Phys.*, vol. 45, no. 10, pp. 4568–4581, Oct. 2018.

[18] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, ''Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology,'' in *Proc. 5th IEEE Int. Symp. Biomed. Imaging, From Nano Macro*, May 2008, pp. 284–287.

[19] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, ''Benign and malignant breast tumors classification based on region growing and CNN segmentation,'' *Expert Syst. Appl.*, vol. 42, no. 3, pp. 990–1002, Feb. 2015.

[20] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[22] A. Newell, K. Yang, and J. Deng, ''Stacked hourglass networks for human pose estimation,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[23] J. Wang, Z. Wei, T. Zhang, and W. Zeng, ''Deeply-fused nets,'' 2016, *arXiv:1605.07716*. [Online]. Available: http://arxiv.org/abs/1605.07716

[24] K. Sun, B. Xiao, D. Liu, and J. Wang, ''Deep high-resolution representation learning for human pose estimation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703. [Online]. Available: https://github.com/leoxiaobin/deep-high-resolution-net.pytorch

[25] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, ''Large kernel matters—-Improve semantic segmentation by global convolutional network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.

[26] T. Kong, A. Yao, Y. Chen, and F. Sun, ''HyperNet: Towards accurate region proposal generation and joint object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[27] W. Liu, A. Rabinovich, and A. C. Berg, ''ParseNet: Looking wider to see better,'' 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579

[28] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, ''Inside-outside Net: Detecting objects in context with skip pooling and recurrent neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[29] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, ''H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes,'' *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[30] S. Bakas *et al.*, ''Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge,'' 2018, *arXiv:1811.02629*. [Online]. Available: http://arxiv.org/abs/1811.02629

[31] V. Dumoulin and F. Visin, ''A guide to convolution arithmetic for deep learning,'' 2016, *arXiv:1603.07285*. [Online]. Available: http://arxiv.org/abs/1603.07285

[32] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[33] W. Luo, Y. Li, R. Urtasun, and R. Zemel, ''Understanding the effective receptive field in deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[34] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, ''Robust object recognition with cortex-like mechanisms,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[35] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, ''Inception-v4, inception-ResNet and the impact of residual connections on learning,'' in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–12.

[37] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, ''An intriguing failing of convolutional neural networks and the coordconv solution,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9605–9616.

[38] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, ''Distributed representations,'' in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, vol. 1. Cambridge, MA, USA: MIT Press, 1986.

[39] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[40] B. Demirkan, S. Bozkurt, A. Şavk, K. Cellat, F. Gülbağca, M. S. Nas, M. H. Alma, and F. Sen, "Composites of bimetallic platinum-cobalt alloy nanoparticles and reduced graphene oxide for electrochemical determination of ascorbic acid, dopamine, and uric acid," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.
[41] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19. [Online]. Available: https://github.com/SimJeg/FC-DenseNet
[42] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Comput. Biol. Med.*, vol. 109, pp. 218–225, Jun. 2019.
[43] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PLoS ONE*, vol. 10, no. 10, Oct. 2015, Art. no. e0140381.
[44] A. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*. [Online]. Available: http://arxiv.org/abs/1902.09063
[45] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*. [Online]. Available: http://arxiv.org/abs/1901.04056
[46] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.
[47] H. Health. (2016). *Ultrasound Nerve Segmentation*. Kaggle. [Online]. Available: https://www.kaggle.com/c/ultrasound-nerve-segmentation/data
[48] A. Messali, R. Villacorta, and J. W. Hay, "A review of the economic burden of glioblastoma and the cost effectiveness of pharmacologic treatments," *PharmacoEconomics*, vol. 32, no. 12, pp. 1201–1212, Dec. 2014.
[49] H. Ando, M. Nagino, T. Arai, H. Nishio, and Y. Nimura, "Changes in splenic vol. during, liver regeneration," *World J. Surg.*, vol. 28, no. 10, pp. 977–981, 2004.
[50] G. Petrovai, S. Truant, C. Langlois, A. F. Bouras, S. Lemaire, D. Buob, E. Leteurtre, E. Boleslawski, and F.-R. Pruvot, "Mechanisms of splenic hypertrophy following hepatic resection," *HPB*, vol. 15, no. 12, pp. 919–927, Dec. 2013.
[51] F. Chollet. (2015). *Keras, Github*. GitHub Repository. [Online]. Available: https://github.com/fchollet/keras
[52] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11. [Online]. Available: https://github.com/MrGiovanni/UNetPlusPlus
[53] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Med. Image Anal.*, vol. 51, pp. 21–45, Jan. 2019.
[54] Z. Li, Y. Wang, J. Yu, Z. Shi, Y. Guo, L. Chen, and Y. Mao, "Low-grade glioma segmentation based on CNN with fully connected CRF," *J. Healthcare Eng.*, vol. 2017, pp. 1–12, Mar. 2017.
[55] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5353–5360.
[56] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020. [Online]. Available: https://github.com/nibtehaz/MultiResUNet

WEIHAO WENG (Graduate Student Member, IEEE) was born in Zhuhai, Guangdong, China, in 1995. He is currently pursuing the master's degree with The University of Aizu. His current research interests include machine learning and computer vision.

XIN ZHU (Senior Member, IEEE) received the bachelor's and master's degrees in biomedical engineering from Tianjin University, China, in 2000 and 2002, respectively, and the Ph.D. degree in computer science and engineering from The University of Aizu, Aizu-Wakamatsu, Japan, in 2006. From 2006 to 2009, he worked as a Postdoctoral Researcher with the Biomedical Information Technology Laboratory, The University of Aizu, where he was an Associate Professor, in 2009. He is currently a Senior Associate Professor with the Biomedical Information Engineering Laboratory, The University of Aizu. He is also a Research Leader with the Center for Advanced Information Science and Technology, The University of Aizu. His research interests include biomedical signal processing, cardiac modeling and simulation, biomedical image processing and analysis, and healthcare.