KiU-Net: Overcomplete Convolutional Architectures for Biomedical Image and Volumetric Segmentation

Jeya Maria Jose Valanarasu, *Student Member, IEEE*, Vishwanath A. Sindagi, *Student Member, IEEE*, Ilker Hacihaliloglu, *Member, IEEE* ,and Vishal M. Patel, *Senior Member, IEEE*

Abstract—Most methods for medical image segmentation use U-Net or its variants as they have been successful in most of the applications. After a detailed analysis of these “traditional” encoder-decoder based approaches, we observed that they perform poorly in detecting smaller structures and are unable to segment boundary regions precisely. This is in spite of the fact that these approaches propagate low-level features to the output through skip connections. This issue can be attributed to the increase in receptive field size as we go deeper into the encoder. The extra focus on learning high level features causes the U-Net based approaches to learn less information about low-level features which are crucial for detecting small structures. To overcome this issue, we propose using an overcomplete convolutional architecture where we project our input image into a higher dimension such that we constrain the receptive field from increasing in the deep layers of the network. We design a new architecture for image segmentation- KiU-Net which has two branches: (1) an overcomplete convolutional network Kite-Net which learns to capture fine details and accurate edges of the input, and (2) U-Net which learns high level features. Furthermore, we also propose KiU-Net 3D which is a 3D convolutional architecture for volumetric segmentation. We perform a detailed study of KiU-Net by performing experiments on five different datasets covering various image modalities like ultrasound (US), magnetic resonance imaging (MRI), computed tomography (CT), microscopic and fundus images. The proposed method achieves a better performance as compared to all the recent methods with an additional benefit of fewer parameters and faster convergence. Additionally, we also demonstrate that the extensions of KiU-Net based on residual blocks and dense blocks result in further performance improvements. The implementation of KiU-Net can be found here: https://github.com/jeya-maria-jose/KiU-Net-pytorch

Index Terms—Brain Anatomy Segmentation, Brain Tumor Segmentation, Liver Segmentation, Gland Segmentation, Nerve Segmentation, Medical Image Segmentation, Deep Learning, Overcomplete Representations.

1 Introduction

Medical image segmentation plays a pivotal role in computer-aided diagnosis systems which are helpful in making clinical decisions. Segmenting a region of interest like an organ or lesion from a medical image or a scan is critical as it contains details like the volume, shape and location of the region of interest. Automatic methods proposed for medical image segmentation help in aiding radiologists for making fast and labor-less annotations. Early methods were based on traditional pattern recognition techniques like statistical modeling and edge detection filters. Later, machine learning approaches using hand-crafted features based on the modality and type of segmentation task were developed. Recently, the state of the art methods for medical image segmentation for most modalities like magnetic resonance imaging (MRI), computed tomography (CT) and ultrasound (US) are based on deep learning. As convolutional neural networks (CNNs) extract data-specific features which are rich in quality and effective in representing the image and the region of interest, deep learning reduces the hassle of extracting manual features from the image.

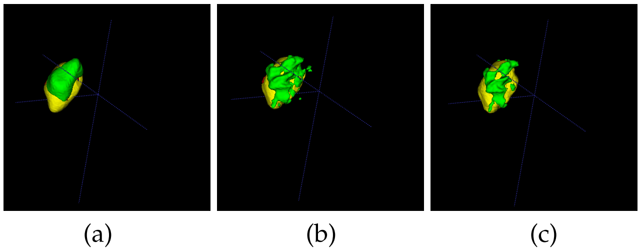


Fig. 1. A sample brain tumor segmentation prediction using (a) U-Net 3D, (b) KiU-Net 3D, and (c) Ground-Truth for BraTS dataset. KiU-Net 3D results in better segmentation of fine details when compared to U-Net 3D as it focuses more on low-level information effectively.

Most of the architectures developed for semantic segmentation in both computer vision and medical image analysis are encoder-decoder type convolutional networks. Seg-Net [1] was the first such type of network that was widely recognized. In the encoder block of Seg-Net, every convolutional layer is followed by a max-pooling layer which causes the input image to be projected onto a lower dimension similar to an undercomplete auto-encoder. The receptive field size of the filters increases with the depth

- *Jeya Maria Jose Valanarasu, Vishwanath A. Sindagi and Vishal M. Patel are with the Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, e-mail: (jvalana1,vishwanathsindagi,vpatel36)@jhu.edu*
- *Ilker Hacihaliloglu is with Rutgers, The State University of New Jersey, NJ, USA, e-mail: ilker.hac@rutgers.edu*

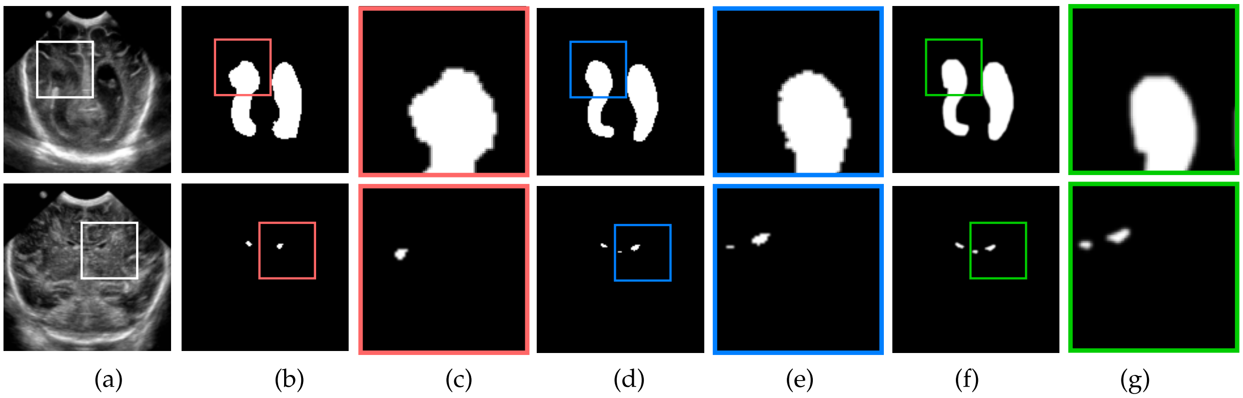


Fig. 2. (a) Input B-Mode Ultrasound Image. Predictions from (b) U-Net, (d) KiU-Net (ours), (f) Ground Truth. (c),(e) and (g) are the zoomed in patches from (b),(d) and (f) respectively. The boxes in the original images correspond to the zoomed in portion for the zoomed images. Our method segments small anatomy and edges better than U-Net.

of the network thereby enabling it to extract high-level features in the deeper layers. The initial layers of the encoder extract low-level information like edges and small anatomical structures while the deeper layers extract high-level information like objects (in the case of vision datasets) and organs/lesions (in the case of medical imaging datasets). A major breakthrough in medical image segmentation was brought by U-Net [2] where skip connections were introduced between the encoder and decoder to improve the training and quality of the features used in predicting the segmentation. U-Net has became the backbone of almost all the leading methods for medical image segmentation in recent years. Subsequently, many more networks were proposed which built on top of U-Net architectures. U-Net++ [3], [4] proposed using nested and dense skip connection for further reducing the semantic gap between the feature maps of the encoder and decoder. UNet3+ [5] proposed using full-scale skip connections where skip connections are made between different scales. 3D U-Net [6] and V-Net [7] were proposed as extensions of U-Net for volumetric segmentation in 3D medical scans. In other extensions of U-Net like Res-UNet [8] and Dense-UNet [9], the convolutional blocks in encoder and decoder consisted of residual connections [10] and dense blocks [11] respectively. It can be noted that all the above extensions of U-Net used the same encoder-decoder architecture and their contributions were either in skip connections, using better convolutional layer connections or in applications.

The main problem with the above family of networks is that they lack focus in extracting features for segmentation of small structures. As the networks are built to be more deeper, more high-level features get extracted. Even though the skip connections facilitate transmission of local features to the decoder, from our experiments we observed that they still fail at segmenting small anatomical landmarks with blurred boundaries. Although U-Net and its variants are good at segmenting large structures, they fail when the segmentation masks are small or have noisy boundaries which can be seen in Fig 1 and 2. As mentioned earlier, U-net and its variants belong to undercomplete convolutional architectures which is what causes the network to focus on high-level features. To this end, we propose using overcomplete convolutional architectures for segmentation. We call our overcomplete architecture Kite-Net (Ki-Net) which transforms the input to higher dimensions (in the spatial sense). Note that Kite-Net does not follow the traditional encoder-decoder style of architecture, where the inputs are mapped to lower dimensional embeddings (in the spatial sense). Compared to the use of max-pooling layers in the traditional encoder and upsampling layers in the traditional decoder, Kite-Net has upsampling layers in the encoder and max-pooling layers in the decoder. This ensures that the receptive field size of filters in the deep layers of the network does not increase like in U-Net. This ensures that the Kite-Net is able to extract fine details of boundaries as well as small structures even in the deeper layers. Although Kite-Net extracts high quality low-level features, the lack of filters extracting high-level features makes Kite-Net not perform on par with U-Net when the dataset consists of both small and large structure annotations. Hence, we propose a multi-branch network, KiU-Net, where one branch is overcomplete (Ki-Net) and another is undercomplete (U-Net). Furthermore, we propose to effectively combine the features across the two branches using a novel cross-residual fusion strategy which results in efficient learning of KiU-Net.

Note that, we presented a preliminary version of this work in [12] at MICCAI 2020 where we:

- Proposed an overcomplete convolutional architectures (Kite-Net) for segmentation and studied how it is different from undercomplete architectures like U-Net.
- Proposed a novel architecture KiU-Net which combines feature maps of both under-complete and overcomplete deep networks such that the network learns to segment both small and large segmentation masks effectively.
- Evaluated the performance of the proposed method for brain anatomy segmentation from 2D US scans where we achieved significantly better performance as compared to recent methods [2], [13], [14], [15].

In the current work, we propose additional improvements over our earlier work. Specifically, we make the following contributions:

- KiU-Net 3D which is an extension of KiU-Net for volumetric segmentation. Here, we use a 3D

convolution-based overcomplete network for efficient extraction of low-level information.
- Res-KiUNet and Dense-KiUNet architectures where we use residual connections and dense blocks respectively for improving the learning of the network.
- We evaluate the performance of the proposed architecture for volumetric and image segmentation across 5 datasets: Brain Tumor Segmentation (BraTS), Liver Tumor Segmentation (LiTS), Gland Segmentation (GlaS), Retinal Images vessel Tree Extraction (RITE) and Brain Anatomy segmentation. These datasets individually correspond to 5 different modalities: ultrasound (US), magnetic resonance imaging (MRI), computed tomography (CT), microscopic and fundus images. With these additional experiments on multiple datasets, we demonstrate that the proposed method generalizes well to different modalities.

2 Related Work

In this section, we briefly review the deep learning works proposed for medical image segmentation. We mainly focus on methods that deal with datasets which we conduct our experiments on.

For brain ventricle segmentation from US scans, methods based on U-Net and residual architectures have been investigated [16]. Wang et al. [14] proposed a PSP-net based method for this task. In [13], the authors explored using uncertainty to further improve the predictions. For brain tumor segmentation for MRI scans, a lot of methods have been proposed based on 2D U-Net and its variations [17], [18], [19], [20]. Many other methods based on Res-Net [10], pixel-net [21] and PSP-net [22] have been proposed for brain tumor segmentation in [23]. 3D convolution based methods have been proved to be better for segmentation of brain tumor when compared to training 2D convolution networks on individual 2D slices of MRI scans and then combining them back together to get 3D segmentation. So, 3D U-Net based methods have been proposed in many recent works for brain tumor segmentation. Most of the top performing methods in the BraTS challenge used 3D U-Net as their backbone network [14], [24], [25], [26]. Out of these methods, regularization on a 3D U-Net backbone architecture [27] gives the best performance for BraTS test dataset (2018). A simple 3D U-Net with just proper hyper parameter tuning is a close second [28]. U-Net based methods have also been adopted for kidney tumor segmentation from CT scans [29], [30], [31], cell/nuclei segmentation from microscopic images [32], [33], [34], [35] and retina nerve segmentation from fundus images [36], [37].

LiTS challenge which deals with liver segmentation and liver lesion segmentation has been one of the leading medical image segmentation datasets in recent times. Several works [9], [38], [39], [40] that perform well for this task are also based on U-Net. There are many more deep learning works for medical image segmentation which can be found in review papers for medical image segmentation [41], [42]. Unlike the above methods which are encoder-decoder based convolutional (undercomplete) networks designed for specific applications, in this work we propose a new architecture which uses overcomplete representations for segmentation. Also, we show that our proposed network is not application specific but a generic solution that can be used for segmentation in any modality.

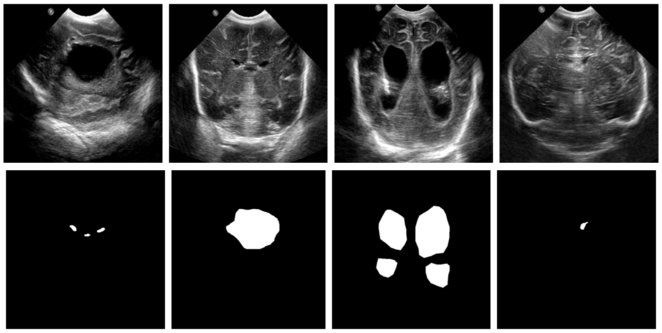


Fig. 3. Sample images from Brain US data. *Top row:* B-Mode US scan. *Bottom row:* Ground truth. The structures present in the dataset are of varying sizes.

3 Method

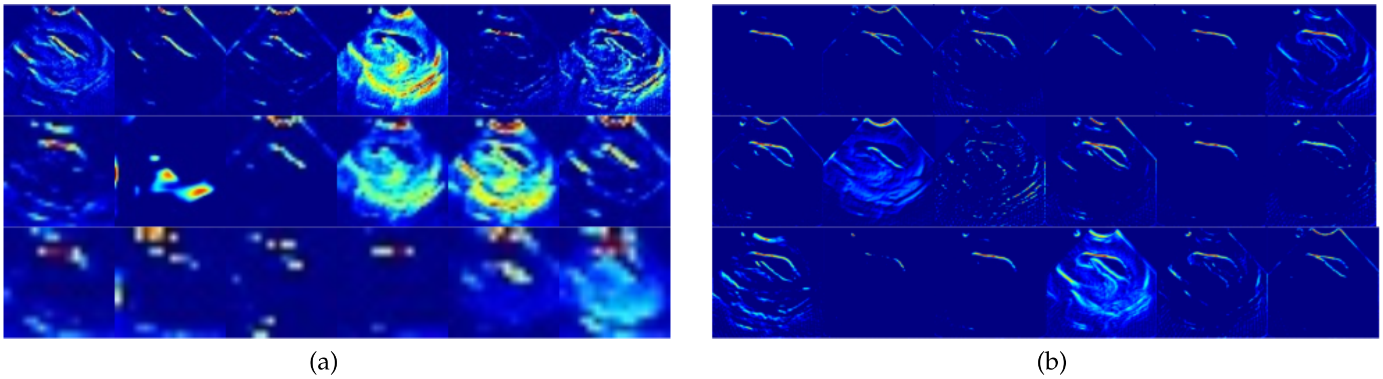
In this section, we first discuss the issues with the U-Net family of architectures and motivate why we propose using overcomplete representations. Later, we describe the proposed architectures in detail.

3.1 KiU-Net

3.1.1 Issues with traditional encoder-decoder networks

In the dataset that we collected for Brain Anatomy Segmentation from US images, the segmentation masks are heterogeneous in terms of the size of the structures. For example it can be seen from Fig 3 that while one image has a large segmentation label, the other one has tiny segmentation mask. U-Net and its variants yield relatively good performance for this dataset as seen in [13], [14]. However, in our experiments we observed that these methods fail to detect tiny structures in most of the cases. This does not cause much decrement in terms of the overall dice accuracy for the prediction since the datasets predominantly contain images with large structures. However, it is crucial to detect tiny structures with a high precision since it plays an important role in diagnosis. Furthermore, even for the large structures, U-Net based methods result in erroneous boundaries especially when the boundaries are blurry as seen in Fig 2 (b),(c).

In order to clearly illustrate these observations, we evaluated U-Net on the Brain Anatomy Segmentation from US images dataset and the results are shown in Fig 2. It can be observed from the bottom row of this figure (Fig 2 (b),(c)) that U-Net fails to detect the tiny structures. Further, the first row demonstrates that in the case of large structures, although U-Net produces an overall good prediction, it is unable to accurately segment out the boundaries. Additionally, we also made similar observations when U-Net based 3D architecture was used for volumetric segmentation of lesion. Specifically, the predictions from U-Net are blurred as it fails to segment the surface perfectly especially when


(a)

(b)

Fig. 4. Feature maps extracted from the encoder of (a) U-Net (b) Kite-Net. Top row: Feature maps extracted from the first layer. Middle row: Feature maps extracted from the second layer. Bottom row: Feature maps extracted from the third layer. It can be observed that while the features extracted from U-Net are coarse and focus on high-level features in the deeper layers, Kite-Net extracts fine details like edges in its deep layers. Also, the feature maps of Kite-Net in the deeper layers are of a higher resolution as we perform upsampling in the feature space.

the surface of the tumor is not smooth and has a high curvature (see Fig 1(a)).

To gain a further understanding of why U-Net based models are unable to segment small structures and boundaries accurately, we analyze the network architecture in detail. In each convolutional block of the encoder, the input set of features to that block get downsampled due to max-pooling layer. This makes sure that the encoder projects the input image to a lower dimension in spatial sense. This combination of convolution layer and max-pooling layer in the encoder causes the receptive field of the filters in deep layers of encoder to increase. With an increased receptive field, the deeper layers focus on high level features and thus are unable to extract features for segmenting small masks or fine edges. The only convolution filters that capture low-level information are the first few layers. Also, it can be noted that in a traditional "encoder-decoder" architecture, the network is designed such that the number of filters increases as we go deeper in the network. That is, for a typical convnet, more than 95% of the filters present in the network focus on extracting high level information. For example, the traditional UNet architecture has a 5 layer deep encoder with 64, 128, 256, 512 and 1024 filters. From Fig 4 (a), we can observe that only the first layer captures low level information like edges and small anatomy. So, effectively the number of filters that capture low level information becomes $\frac{64}{64+128+256+512+1024} = 3.22\%$ which implies 96.78% of the filters work on learning high-level features.

3.1.2 Overcomplete Networks

In signal processing, overcomplete representations [43] were first explored for making dictionaries such that the number of basis functions can be more than the number of input signal samples. This enables a higher flexibility for capturing structure in the data. In [43], the authors show that overcomplete bases work as a better approximators of any underlying statistical distribution of a data. It has also been widely used for reconstruction of signals under the presence of noise and for source separation in a mixture of signals. It is mainly popular for these tasks because of its greater robustness in the presence of noise when compared to undercomplete representations. For denoising autoencoders [44], models with overcomplete hidden layer expression were observed to perform better as they are more useful feature detectors. The authors note that the proposed idea of denoising autoencoders in fact improved the feature detecting ability of an overcomplete fully connected network and it performs better than the standard bottleneck architectures for that task. Overcomplete representations have also been backed in the field of neuroscience where it is solves the problem of maintaining a stable sensory percept in the absence of time-invariant persistent activity [45].

The idea of overcomplete networks (in the spatial sense) in the convnet-deep learning era has been unexplored. To this end, we propose Kite-Net which is an overcomplete version of U-Net. In Kite-Net, the encoder projects the input image into a spatially higher dimension. This is achieved by incorporating bilinear upsampling layers in the encoder. This form of the encoder constrains the receptive field from increasing like in U-Net as we carefully select the kernel size of the filters and upsampling coefficient such that the deep layers learn to extract fine details and features to segment small masks effectively. Furthermore, in the decoder, each conv block has a conv layer followed by a max-pooling layer. In Fig 6, we can see how the receptive field is constrained in our proposed Kite-net when compared to U-Net.

To analyze this in detail, let I be the input image, F_1 and F_2 be the feature maps extracted from first and second conv blocks respectively. Let the initial receptive field of the conv filter be $k \times k$ on the image. In an undercomplete network, the receptive field size change due to max-pooling layer is dependent on two variables: pooling coefficient and stride of the pooling filter. For convenience, the pooling coefficient and stride is both set as 2 in our network. Considering this configuration, the receptive field of conv block 2 (to which F_1 is forwarded) on the input image would be $2 \times k \times 2 \times k$. Similarly, the receptive field of conv block 3 (to which F_2 is forwarded) would be $4 \times k \times 4 \times k$. This increase in receptive field can be generalized for an i^{th} layer in an undercomplete network as follows:

$$RF(w.r.t\ I) = 2^{2*(i-1)} \times k \times k.$$

In comparison, the proposed overcomplete network has an upsampling layer with a coefficient 2 in the conv blocks replacing the max-pooling layer. As the upsampling layer actually works opposite to that of max-pooling layer, the

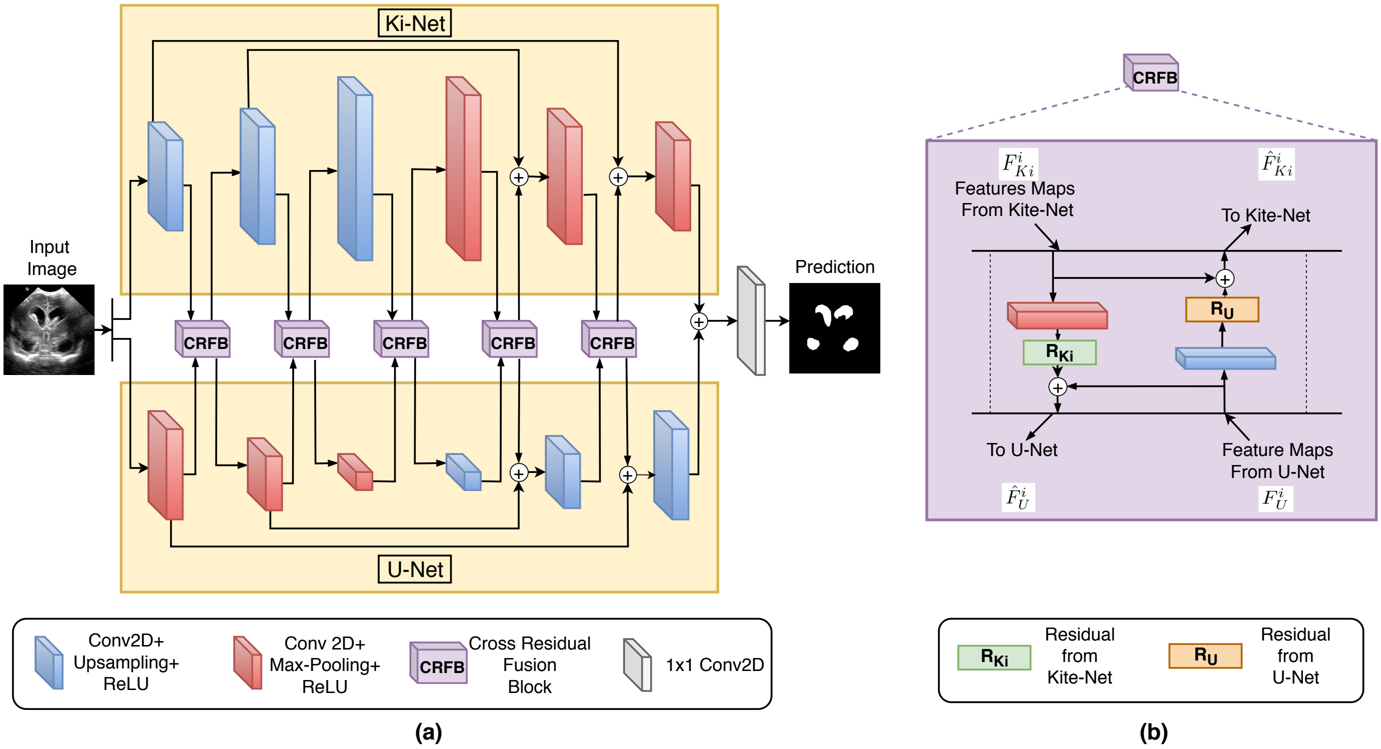


Fig. 5. (a) Architecture details of KiU-Net for 2D image segmentation. In KiU-Net, the input image is forwarded to the two branches of KiU-Net: Kite-Net and U-Net which have CRFB blocks connecting them at each level. The feature maps from the last layer of both the branches are added and passed through 1×1 2D conv to get the prediction. In CRFB, residual features of Kite-Net are learned and added to the features of U-Net to forward the complementary features and vice-versa. (b) Details of Cross Residual Fusion Block (CRFB).

receptive field of conv bock 2 on the input image now would be $\frac{1}{2} \times k \times \frac{1}{2} \times k$. Similarly, the receptive field of conv block 3 now would be $\frac{1}{4} \times k \times \frac{1}{4} \times k$. This increase in receptive field can be generalized for i^{th} layer in the overcomplete network as follows:

$$RF(w.r.t\ I) = (\frac{1}{2})^{2*(i-1)} \times k \times k.$$

Note that the above calculations are based on a couple of assumptions. We assume that the pooling coefficient and stride are both set as 2 in both overcomplete and undercomplete network. Also, we consider that the receptive field change caused by the conv layer in both undercomplete and overcomplete networks would be the same and do not consider in our calculations. This can be justified as we have maintained the conv kernel size to 3×3 with stride 1 and padding 1 throughout our network and this setting does not actually affect the receptive as much as max-pooling or upsampling layer does. The above explanations are illustrated in Fig 6.

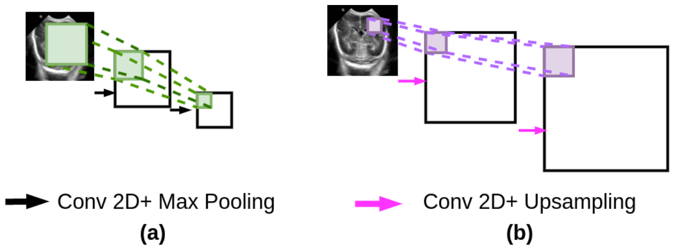


Fig. 6. Effect on receptive field change due to architecture type as seen in (a) U-Net (b) Kite-Net. The receptive field of Kite-Net is constrained when compared to U-Net in the deeper layers.

In Fig 4, we visualize the feature maps of both U-Net and Kite-Net. The first row corresponds to the features extracted from the first layer of encoder, second row corresponds to the second layer and third row corresponds to the third layer. From this, it can be observed that Kite-Net extracts more finer details over smaller regions when compared to U-Net. Although Kite-Net learns low-level features better than U-Net, it does not learn any high-level features. Due to this, Kite-Net is unable to segment out any large masks present in the input image. To overcome this, we propose KiU-Net, which efficiently combines both Kite-Net and U-Net while achieving the best of both networks.

3.1.3 Architecture Details

KiU-Net is a two-branch network where one branch is Kite-Net and the other is U-Net. This network exploits the low-level fine edge capturing power of Kite-Net as well the high-level shape capturing power of U-Net. KiU-Net has been illustrated in Fig 5.(a). The input image is forwarded to both Kite-Net and U-Net in parallel. Each conv block in the encoder of Kite-Net has a conv 2D layer followed by a bilinear upsampling layer with coeffecient of two and ReLU activation. Each conv block in the encoder of U-Net has a conv 2D layer followed by a max-pooling layer with kernel size of two and ReLU activation. In each conv block in the decoder of Kite-Net, we have a conv 2D layer followed by a max-pooling layer with kernel size of two and ReLU

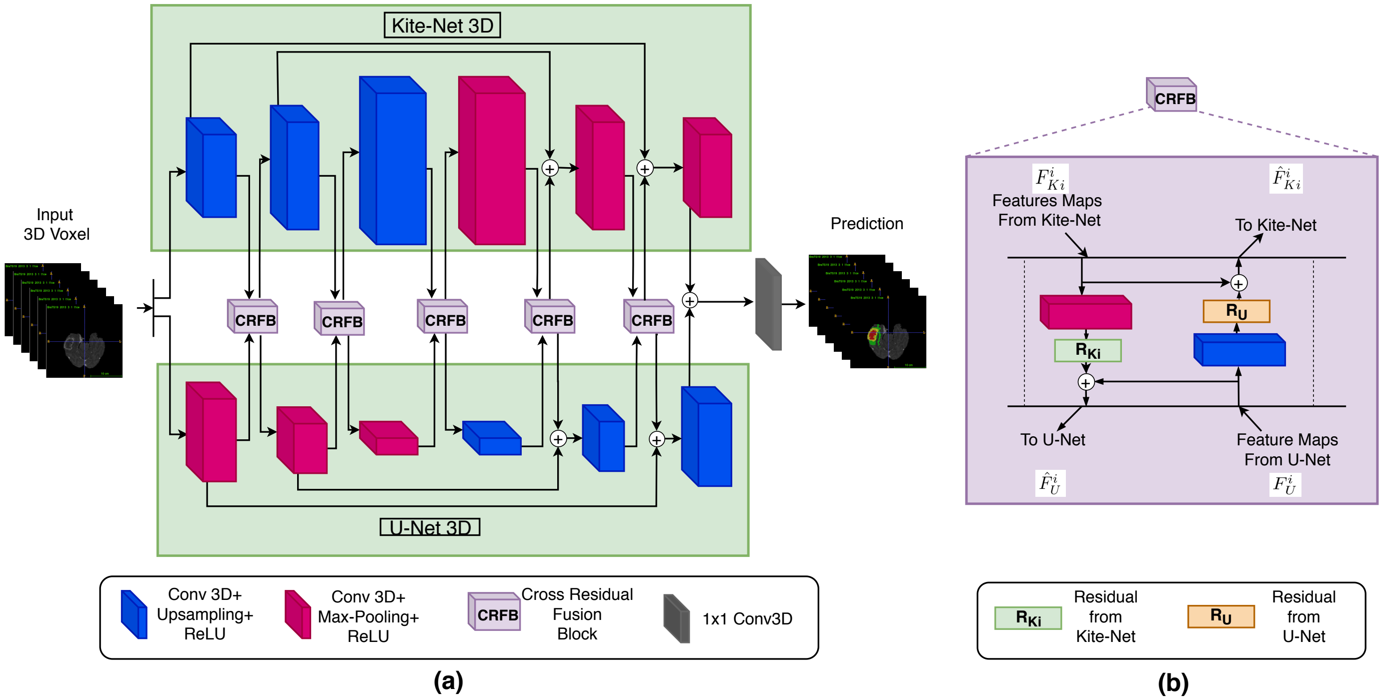


Fig. 7. (a) Architecture details of KiU-Net 3D for 3D volumetric segmentation. (b) Details of Cross Residual Fusion Block (CRFB) for KiU-Net 3D. In KiU-Net 3D, the input 3D voxel is forwarded to the two branches of KiU-Net 3D: Kite-Net 3D and U-Net 3D which have 3D CRFB blocks connecting them at each level. The feature maps from the last layer of both the branches are added and passed through 1×1 3D conv to get the prediction. In CRFB, the residual features of Kite-Net 3D are learned and added to the features of U-Net 3D to forward the complementary features to U-Net and vice-versa.

activation. Similarly for U-Net, we have a conv 2D layer followed by a bilinear upsampling layer with coeffecient of two and ReLU activation in its decoder. We have 3 such conv blocks in encoder and decoder in both Kite-Net and U-Net branches of KiU-Net. We then add the output feature maps of Kite-Net and U-Net and forward them through a 1×1 conv 2D layer to obtain the prediction. All the conv layers in our network (except for the last layer) are of 3×3 size with stride 1 and padding 1.

3.1.4 Cross residual feature block (CRFB)

In order to further exploit the capacity of the two networks, we propose to combine the features of the two networks at multiple scales through a novel cross residual feature block (CRFB). That is, at each level in the encoder and decoder of KiU-Net, we combine the respective features using a CRFB. As we know that the features learned by U-Net and Kite-Net are different from each other, this characteristic can be used to further improve the training of the individual networks. So, we try to learn the complementary features from both the networks which will further improve the quality of features learned by the individual networks.

The CRFB block is illustrated in Fig 5.(b). Denoting the features maps from U-Net as F_U^i and F_{Ki}^i as the feature maps from Ki-Net after i^{th} block in KiU-Net, the cross-residual features R_U^i and R_{Ki}^i are first extracted using a conv block. The conv block that F_{Ki}^i is forwarded to has a combination of conv 2D layer and max-pooling layer. The conv block that F_{Ui}^i is forwarded through has a combination of conv 2D layer and upsampling layer. These cross-residual features are then added to the original features F_U^i and F_{Ki}^i to obtain the complementary features $\hat{F}_U^i$ and $\hat{F}_{Ki}^i$, $\hat{F}_U^i = F_U^i + R_{Ki}^i$ and $\hat{F}_{Ki}^i = F_{Ki}^i + R_U^i$. With this kind of complementary feature extraction from the two networks, we observe a considerable improvement in the segmentation performance of the network.

3.2 KiU-Net 3D

With the success of KiU-Net for 2D segmentation, we explore its usage for volumetric segmentation from 3D medical scans. Specifically, we propose KiU-Net 3D which is a two-branch network similar to KiU-Net but contains Kite-Net 3D and U-Net 3D as its branches. The input will be a scan of shape $H \times W \times S$ where H, W are the height and width and S is the number of slices of 2D images in the scan. So, in our proposed KiU-Net 3D, we use 3D convolutions in the place of 2D convolutions to work on the voxel space of the scans.

In the encoder of Kite-Net 3D branch, every conv block has a conv 3D layer followed by a trilinear upsampling layer with coeffecient of two and ReLU activation. In decoder, every conv block has a conv 3D layer followed by a 3D max-pooling layer with coeffecient of two and ReLU activation. Similarly, in the encoder of U-Net 3D branch, every conv block has a conv 3D layer followed by a 3D max-pooling layer with coefficient of two and ReLU activation. In decoder, every conv block has a conv 3D layer followed by a trilinear upsampling layer with coefficient of two and ReLU activation. We have CRFB block across each layer in KiU-Net 3D similar to KiU-Net. The difference in the CRFB block architecture of KiU-Net 3D is that we have conv 3D layers and trilinear upsampling instead of conv 2D layers and bilinear upsampling like in KiU-Net. The output of both the branches are then added and forwarded to $1 \times 1 \times 1$ conv 3D layer to get the prediction voxel. All the conv layers in our network (except for the last layer) have $3 \times 3 \times 3$ kernel sizes with stride 1 and padding 1. KiU-Net 3D is illustrated in Fig 7.

4 Experiments

In this section, we describe the experimental settings and the datasets that we use for 2D medical image segmentation and 3D medical volumetric segmentation to evaluate and compare the proposed KiU-Net and KiU-Net 3D networks respectively.

4.1 KiU-Net

4.1.1 Datasets

Brain Anatomy Segmentation (US): Intraventricular hemorrhage (IVH) which results in the enlargement of brain ventricles is one of the main causes of preterm brain injury. The main imaging modality used for diagnosis of brain disorders in preterm neonates is cranial US because of its safety and cost-effectiveness. Also, absence of septum pellucidum is an important biomarker for septo-optic dysplasia diagnosis. Automatic segmentation of brain ventricles and septum pellucidum from these US scans is essential for accurate diagnosis and prognosis of these ailments. After obtaining institutional review board (IRB) approval, US scans were collected from 20 different premature neonates (age < 1 year). The total number of images collected were 1629 with annotations out of which 1300 were allocated for training and 329 for testing. Before processing, each image was resized to 128×128 resolution.

Gland Segmentation (Microscopic): Accurate segmentation of glands is important to obtain reliable morphological statistics. Histology images in the form of Haematoxylin and Eosin (H&E) stained slides are generally used for gland segmentation [46]. Gland Segmentation (GLAS) dataset contains a total of 165 images out of which 85 are taken for training and 80 for testing. We pre-process the images by resizing them to 128×128 resolution.

Retinal Nerve Segmentation (Fundus): Extraction of arteries and veins on retinal fundus images are essential for delineation of morphological attributes of retinal blood vessels, such as length, width, patterns and angles. These attributes are then utilized for the diagnosis and treatment of various ophthalmologic diseases such as diabetes, hypertension, arteriosclerosis and chorodial neovascularization [47]. To perform retinal nerve segmentation, we use Retinal Images vessel Tree Extraction (RITE) dataset [48] which is a subset of the DRIVE dataset. RITE dataset contains 40 images split into 20 for training and 20 for testing. We pre-process the images by resizing them to 128×128 resolution.

4.1.2 Training and Implementation

As the purpose of these experiments are to demonstrate the effectiveness of proposed architecture, we do not use any application specific loss function or metric loss functions in our experiments. We use a binary cross-entropy loss between the prediction and ground truth to train the network. The cross-entropy loss is defined as follows:

$$\mathcal{L}_{CE(p,\hat{p})} = -\frac{1}{wh}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(p(x,y)\log(\hat{p}(x,y)))+ \\ (1-p(x,y))log(1-\hat{p}(x,y)),$$

where w and h are the dimensions of image, $p(x,y)$ corresponds to the image and $\hat{p}(x,y)$ denotes the output prediction at a specific pixel location (x,y). We set the batch-size as 1 and learning rate as 0.001. We use Adam optimizer for training. We use these hyperparameters uniformly across all the experiments. We train our networks for 300 epochs or until convergence depending on the dataset.

4.2 KiU-Net 3D

For volumetric segmentation, we use two widely used public datasets: BraTS challenge dataset and LiTS challenge dataset for our experiments.

4.2.1 Datasets

Brain Tumor Segmentation (MRI): Segmentation of the tumor is important for volume estimation of gliomas which is essential for planning clinical treatment. Brain Tumor Segmentation (BraTS) challenge has a curated collection of MRI scans with expert annotations of brain tumor. It contains multimodal MRI scans of confirmed cases of glioblastoma (GBM/HGG) and low grade glioma (LGG). The modalities present in these scans are native (T1), post-contrast T1 (T1ce), T2-weighted (T2) and T2 attenuated inversion recovery (T2-FLAIR) [49], [50], [51]. The annotations are provided for four classes - enhancing tumor, peritumoral edema and the necrotic and non-enhancing tumor core. We used the 2019 version of the BraTS challenge training data for training our the baselines and our proposed. It has a total of 335 MRI scans. For quantitative comparisons, we use the validation dataset provided by BraTS 2019 challenge which has 125 scans. For qualitative comparisons, we use the new 33 scans added to the training set of BraTS 2020 challenge dataset as the validation datasets do not come with ground truth annotations. Each MRI scan contains 155 slices each of dimensions 255×255.

Liver Segmentation (CT): Accurate segmentation of liver and liver lesions from abdominal CT scans are essential for diagnosis, treatment and surgical planning. Liver Tumor Segmentation Challenge (LiTS) dataset [52] contains contrast-enhanced abdominal CT scans along with annotations of liver and liver lesions. The training set contains 130 CT scans and the testing set contains 70 CT scans. For our experiments, we focus on liver segmentation.

4.2.2 Training and Implementation

For training the KiU-Net 3D, we use the cross entropy loss between the prediction and the input scan. Since, the scan can be viewed as a 3D voxel, the cross entropy loss can be formulated as follows:

TABLE 1
Comparison of quantitative metrics for Brain US, GLAS and RITE datasets between KiU-Net, U-Net++ [3], U-Net [2] and Seg-Net [1].

Dataset	Metrics	Seg-Net [1]	U-Net [2]	U-Net++ [3]	KiU-Net
Brain US	Dice	0.8279	0.8537	0.8659	**0.8943**
	Jaccard	0.7502	0.7931	0.7995	**0.8326**
GLAS	Dice	0.7861	0.7976	0.8005	**0.8325**
	Jaccard	0.6596	0.6763	0.6893	**0.7278**
RITE	Dice	0.5223	0.5524	0.5410	**0.7517**
	Jaccard	0.3914	0.3111	0.3724	**0.6037**

TABLE 2
Comparison of quantitative metrics for brain tumor segmentation in BraTS dataset.

Type	Network	Dice-ET	Dice-WT	Dice-TC	Hausdorff95-ET	Hausdorff95-WT	Hausdorff95-TC
Slice wise	Seg-Net [1]	0.4994	0.7611	0.6887	65.6867	20.3247	20.0050
	U-Net [2]	0.5264	0.8083	0.7032	17.5458	13.9467	19.2653
	KiU-Net	0.6637	0.8612	0.7061	9.4176	12.7896	13.0401
Voxel wise	Seg-Net 3D [1]	0.5599	0.8062	0.7073	10.0037	10.4584	10.7513
	U-Net 3D [6]	0.6711	0.8448	0.7059	10.2162	13.0005	15.0856
	KiU-Net 3D	**0.7321**	**0.8760**	**0.7392**	**6.3228**	**8.9424**	**9.8929**

TABLE 3
Comparison of quantitative metrics for liver segmentation in LiTS dataset.

Type	Network	Dice	Jacard	VOE	FNR	FPR	ASSD	MSD
Slice-wise	Seg-Net [1]	0.7656	0.6265	0.3734	0.0673	0.3061	7.6310	45.3610
	U-Net [2]	0.7723	0.6350	0.3649	0.0688	0.0688	7.6968	48.7151
	KiU-Net	0.8035	0.6412	0.2956	0.0605	0.2045	6.8452	42.5421
Voxel-wise	Seg-Net 3D [1]	0.8789	0.7904	0.2091	**0.0428**	0.1666	3.9452	42.0544
	U-Net 3D [6]	0.9346	0.8828	0.1171	0.0622	0.0549	1.9450	33.8872
	KiU-Net 3D	**0.9423**	**0.8946**	**0.1053**	0.0548	**0.0505**	**1.7711**	**29.9831**

$$\mathcal{L}_{CE(p,\hat{p})} = -\sum_{c=0}^{C-1} \frac{1}{whl} \sum_{z=0}^{l-1} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} p(z,x,y) \log(\hat{p}(z,x,y)),$$

where w, h are the dimensions of each slice while l is the number of slices in the scan, $p(z,x,y)$ corresponds to the input 3D scan and $\hat{p}(z,x,y)$ denotes the output prediction at a specific pixel location (z,x,y) and C corresponds to the number of classes found in the dataset. The above loss formulation suits the multi-class segmentation framework of BraTS dataset where $C = 4$. For both the datasets, we use a learning rate of 0.0001 with batch size 1 and use Adam optimizer. We do voxel-wise training (similar to patch-wise training on 2D) for the BraTS dataset with voxel shapes $128 \times 128 \times 50$.

5 Results

In this section, we discuss the results and outcomes of the experiments that we conducted using KiU-Net and KiU-Net 3D. First, we discuss the quantitative evaluations where the proposed method is compared with other recent approaches using metrics that are widely used for medical image segmentation. Next, we provide qualitative results where we visualize sample predictions to analyze why the proposed method's performance is superior as compared to the other approaches.

5.1 Quantitative Results

5.1.1 KiU-Net

Following existing approaches like [4], we use Dice Index (F1-score) and Jaccard Similarity score (mIOU) for evaluating and comparing the proposed method (KiU-Net) on the medical image segmentation datasets:

$$Dice = \frac{2TP}{2TP + FP + FN},$$

$$Jaccard = \frac{TP}{TP + FP + FN},$$

where TP, FP and FN correspond to the number of pixels that are true positives, false positives and false negatives respectively of the prediction when compared with the ground truth. Table 1 shows the results for the experiments on all the 3 datasets for image segmentation. As it can be observed, we compare KiU-Net with some of the widely used backbone architectures for medical image segmentation like Seg-Net, U-Net and U-Net++. KiU-Net not only performs better than the other networks but also achieves a significant boost in terms of performance.

5.1.2 KiU-Net 3D

For 3D volumetric segmentation, we adopt the performance metrics used in the BraTS challenge and LiTS challenge. For brain tumor segmentation from MRI scans, we report the Dice accuracy of enhancing tumor (ET), whole tumor (WT) and tumor core (TC). We also report the Hausdorff distance for all these three classes. More details about these metrics for brain tumor segmentation can be found in [51]. Similarly for liver segmentation in LiTS dataset, we report metrics such as Dice score, Jaccard index, volume overlap error (VOE), false negative rate (FNR), false positive rate (FPR), average symmetric surface distance (ASSD) and maximum symmetric surface distance (MSD). The details of these metrics can be found in [52]. For both these datasets we compare KiU-Net 3D with U-Net 3D and Seg-Net 3D.

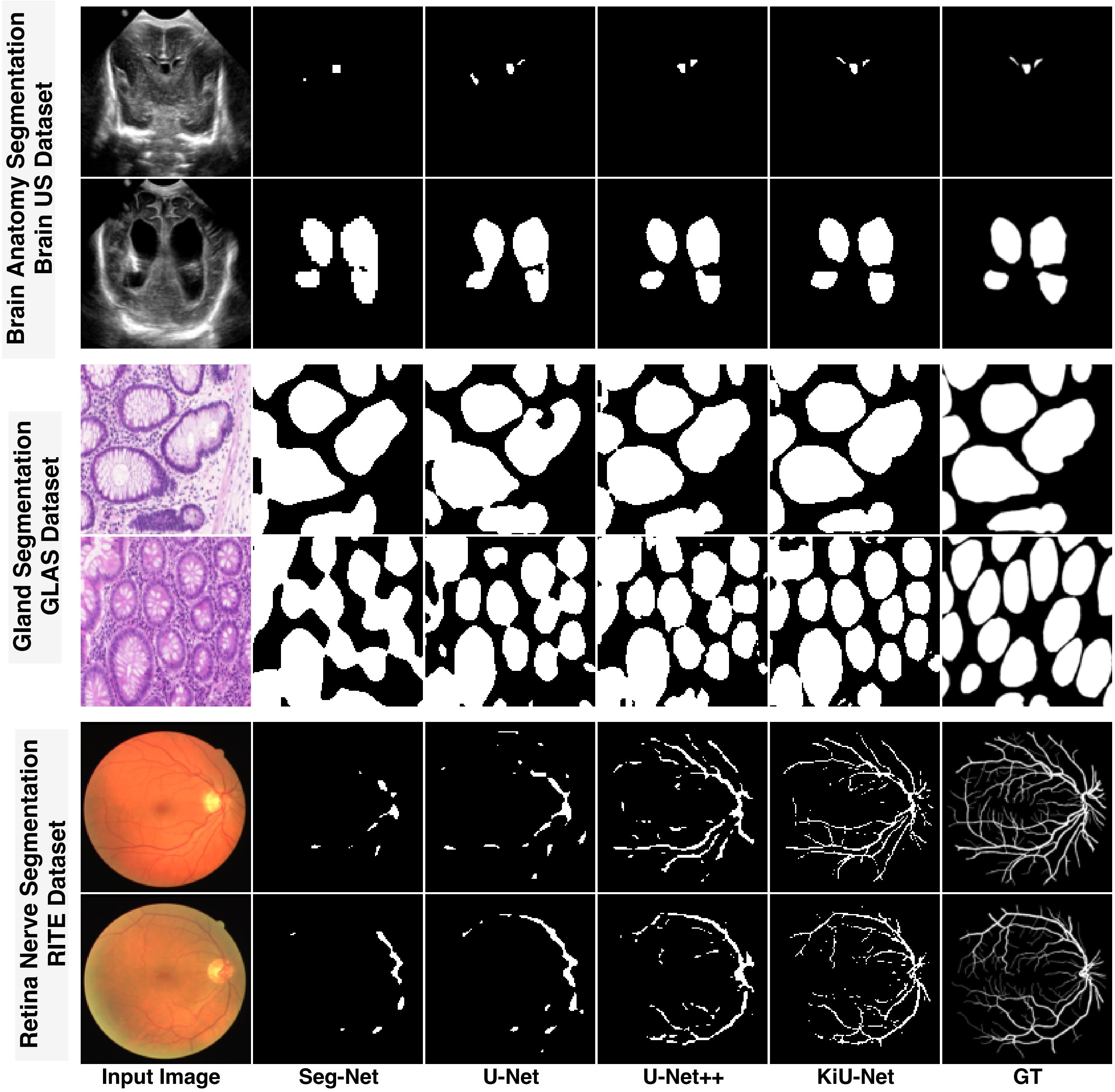


Fig. 8. Comparison of qualitative results between SegNet, UNet , UNet++ and KiU-Net for Brain anatomy segmentation using brain US dataset, gland segmentation using GLAS dataset and retina nerve segmentation using RITE dataset.

Additionally, we also conduct experiments with slice-wise training of the scans using 2D versions of all these networks. Tables 2 and 3 report the quantitative metrics comparison for BraTS dataset and LiTS dataset respectively. We observe that KiU-Net outperforms other methods. Note that we have used the same pipeline for all these experiments for a fair comparison. Further, since the primary goal of these experiments is to show that KiU-Net can act as a better backbone network architecture, we do not perform any pre-processing or post-processing which can improve the performance further irrespective of the network architecture.

5.2 Qualitative Results

5.2.1 *KiU-Net*

Fig 8 illustrates the predictions of KiU-Net along with Seg-Net, U-Net and U-Net++ for all 3 medical image segmentation datasets we used for evaluation. In the first row, we can observe that KiU-Net is able to segment the small ventricles accurately and the other "traditional" networks fail to do so. Similarly from second, third and fourth rows, we can observe that the predictions of KiU-net are able to segment out the edges significantly better when compared to the other networks. In the predictions of RITE dataset which has very low number of images to train, our network performs

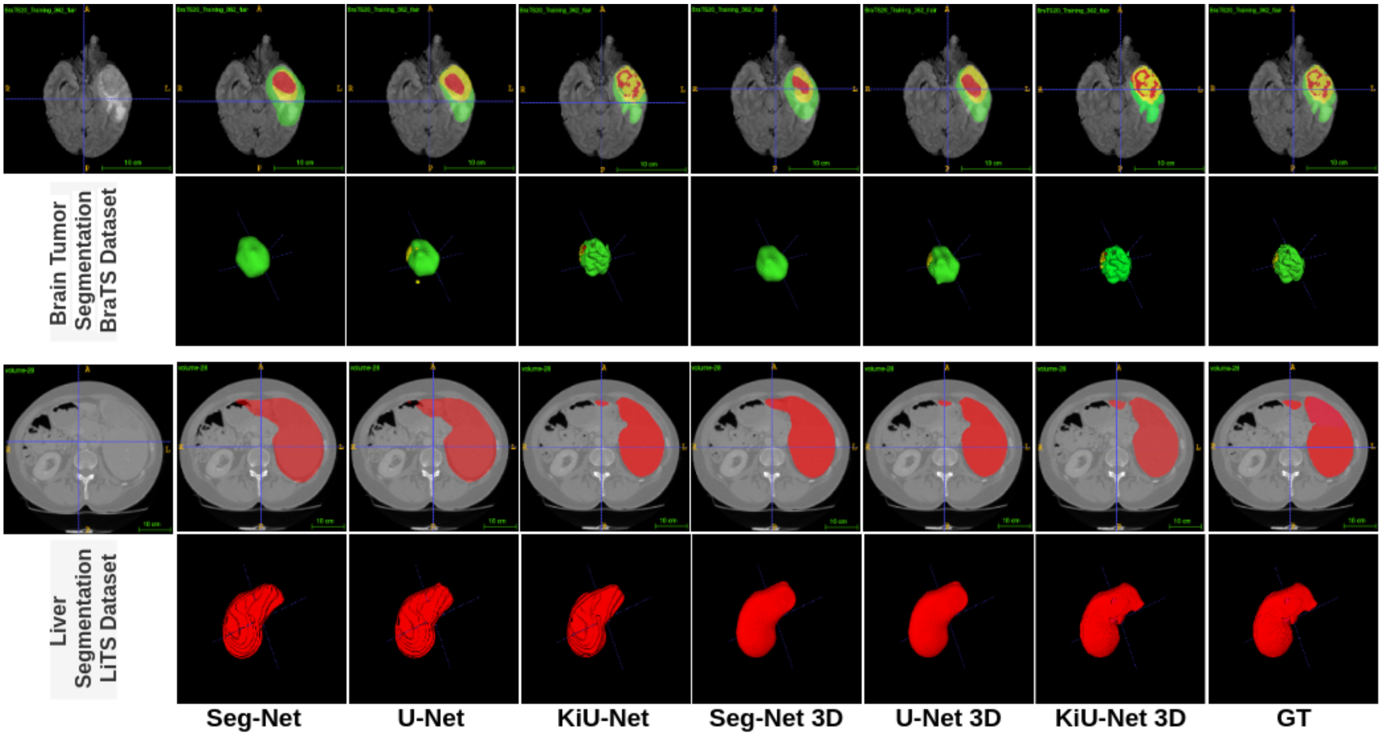

Brain Tumor Segmentation BraTS Dataset
Liver Segmentation LiTS Dataset
Seg-Net
U-Net
KiU-Net
Seg-Net 3D
U-Net 3D
KiU-Net 3D
GT

Fig. 9. Comparison of qualitative results between SegNet, UNet , KiU-Net, SegNet 3D, UNet 3D and KiU-Net 3D for brain tumor segmentation using BraTS dataset and Liver segmentation using LiTS dataset. The first and third row correspond to the prediction from a 2D slice in the MRI brain scan and abdominal CT scan respectively. The second and third rows visualize the 3D volume segmentation predictions.

reasonably well as compared to others learns, thus demonstrating that low level features for nerve segmentation are used effectively in our network.

5.2.2 KiU-Net 3D

Fig 9 illustrates the predictions of KiU-Net 3D for volumetric segmentation experiments. The first two rows correspond to the results for BraTS dataset and the bottom two rows correspond to the results for LiTS dataset. Note that the first row and third rows correspond to segmentation prediction of a single slice of the scan where as the second and fourth row correspond to the 3D segmentation prediction of the scan. The 3D segmentation results are visualized using ITKSnap [53] where each scan prediction consists of 155 2D images in the case of BraTS dataset and 48 2D images in the case of LiTS dataset. From visualizations of the brain tumor segmentation task, it can be observed that KiU-Net is able to segment the surface and edges of the tumor significantly better than any other network and is more closer to the ground truth. For BraTS dataset, the red regions correspond to tumor core, yellow regions correspond to non-enhancing tumor and the green regions correspond to edema. From the second row for BraTS dataset, it can be observed that the tumor surface of the ground truth has sharp edges. While all the other methods smooth out these edges, KiU-Net predicts the sharp edges of the surface of the tumor more precisely. While these results are focused on demonstrating the superiority of KiU-Net in segmenting small lesions, the results on the LiTS dataset show that the proposed method is equally effective in segmenting larger regions. From the bottom two rows of Fig 9, we can observe that KiU-Net performs better than other networks in segmenting large masks as well. Based on this, we would like to point out that even though KiU-Net focuses more on low-level features when compared to UNet or UNet++, its performance is on-par/better as compared to them while segmenting large masks as well. We also observe that performing 2D segmentation on individual 2D images and then combining them to form a 3D scan does not work well. This can be observed from the first 3 images from the last row where the surfaces are not smooth while the ground truth looks smooth.

6 Discussions

As we propose a generic solution to image and volumetric segmentation, we believe it is important to study the computational complexity and convergence trends of the proposed network. In this section, we provide these details and compare the proposed method with other approaches in terms of number of parameters and rate of convergence.

6.1 Number of Parameters

Seg-Net, U-Net and U-Net++ have a 5 layer deep encoder and decoder. The number of filters in each block of these networks increase gradually as we go deeper in the network. For example, U-Net uses this sequence of filters for its 5 layers - 64, 128, 256, 512 and 1024. Although KiU-Net is a multi-branch network, we limit the complexity of our network by using fewer layers and filters. Specifically, we use a 3 layer deep network with 32, 64 and 128 respectively as the number of filters. Due to this, the number of parameters in our network is significantly fewer as compared

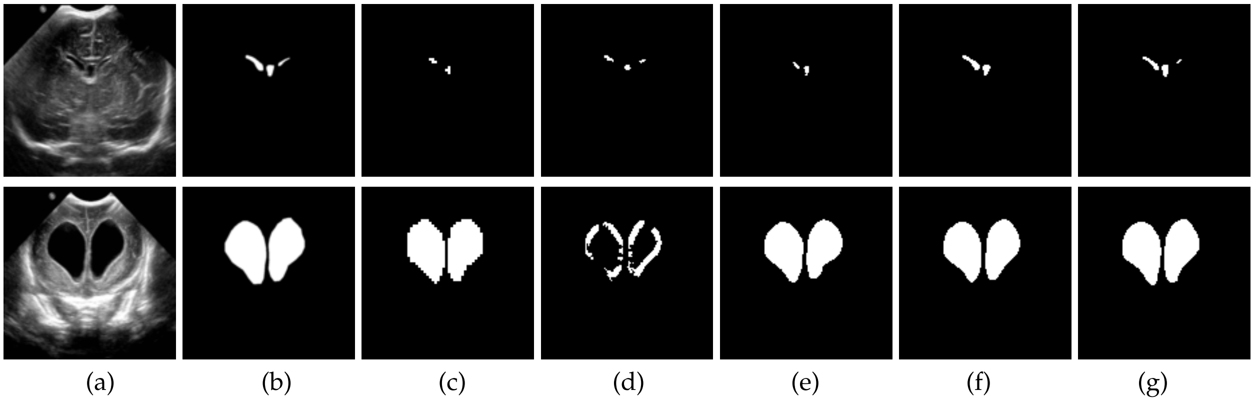


Fig. 10. Qualitative results of ablation study on test images. (a) B-Mode input US image. (b) Ground Truth annotation. Prediction of segmentation masks by (c) UC - Under-complete architecture (d) OC - Over-complete architecture (e) UC + SK (under-complete architecture with skip connections) (f) UC + OC with SK (combined architecture with skip connections) (g) KiU-Net (ours). The change in quality of the predictions can be observed after each addition to the network.

TABLE 4
Comparison of number of parameters

Network	Seg-Net [1]	U-Net [2]	U-Net++ [3]	KiU-Net
No. of Parameters	12.5M	3.1M	9.0M	**0.29M**

TABLE 5
Ablation Study

Metrics	UC	OC	UC+SK	OC+SK	UC+OC+SK	KiU-Net
Dice	0.82	0.56	0.85	0.60	0.86	**0.89**
Jaccard	0.75	0.43	0.79	0.47	0.78	**0.83**

to other methods. In Table 4, we tabulate the number of parameters for KiU-Net and other recent networks and it can be observed that KiU-Net has ∼10× lesser parameters as compared to U-Net and ∼40× fewer parameters as compared to SegNet. Further, it is important to note that the other approaches have have resulted in higher complexity in an attempt to improve the performance, however, our approach is able to obtain better performance while have significantly fewer parameters.

6.2 Convergence

The convergence of loss function is an important characteristic associated with a network. A faster convergence means is always beneficial as it results in significantly lower training complexity. Fig 11 compares the convergence trends for SegNet, U-Net, U-Net++ and KiU-Net when trained on GLAS dataset. It can be observed that KiU-Net converges faster when compared to other networks. Similar trends were observed while training the network on other datasets as well.

6.3 Ablation Study

We conduct an ablation study to analyze the effectiveness of different blocks in the proposed method (KiU-Net). For these experiments, we use the brain anatomy segmentation US dataset. We start with the basic undercomplete ("traditional") encoder-decoder convolutional architecture (UC) and overcomplete convolutional architecture (OC). Note that these networks do not contain any skip connections (SK). Next, we add skip connections to the UC and OC baselines. These networks are basically U-Net (UC+skip connections) and Kite-Net (OC+skip connections). We then fuse both these networks by adding the feature map output of both the networks at the end. This is in fact KiU-Net without the CRFB block. Finally, we show the performance of our proposed architecture - KiU-Net. Table 5 shows the results of all these ablation experiments. It can be observed that the performance improves with addition of each block to the network. Please note that the performances of OC and Kite-Net are lower because these predictions contain only the edges of the masks and do not contain any high-level information. Fig 10 illustrates the qualitative improvements after adding each major block.

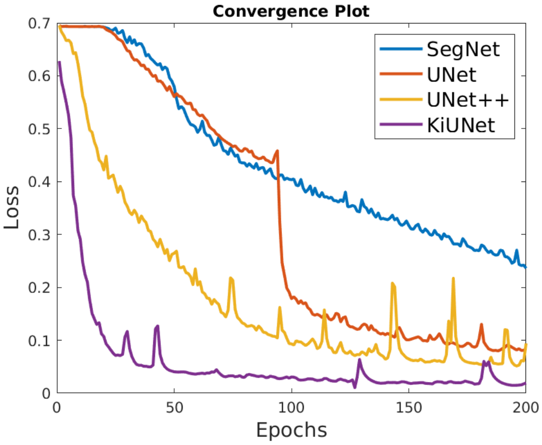


Fig. 11. Comparison of convergence of loss function between KiU-Net, UNet++ [3], UNet [2] and SegNet [1]. The convergence of KiU-Net is faster when compared to all other methods.

6.4 Further Improvements

As it is clear from the above discussions that KiU-Net is a good backbone architecture for both image and volumetric segmentation, we experiment with other variants of KiU-Net which result in further improvements. In this section, we describe these variants and present the detailed results corresponding to each of these variants.

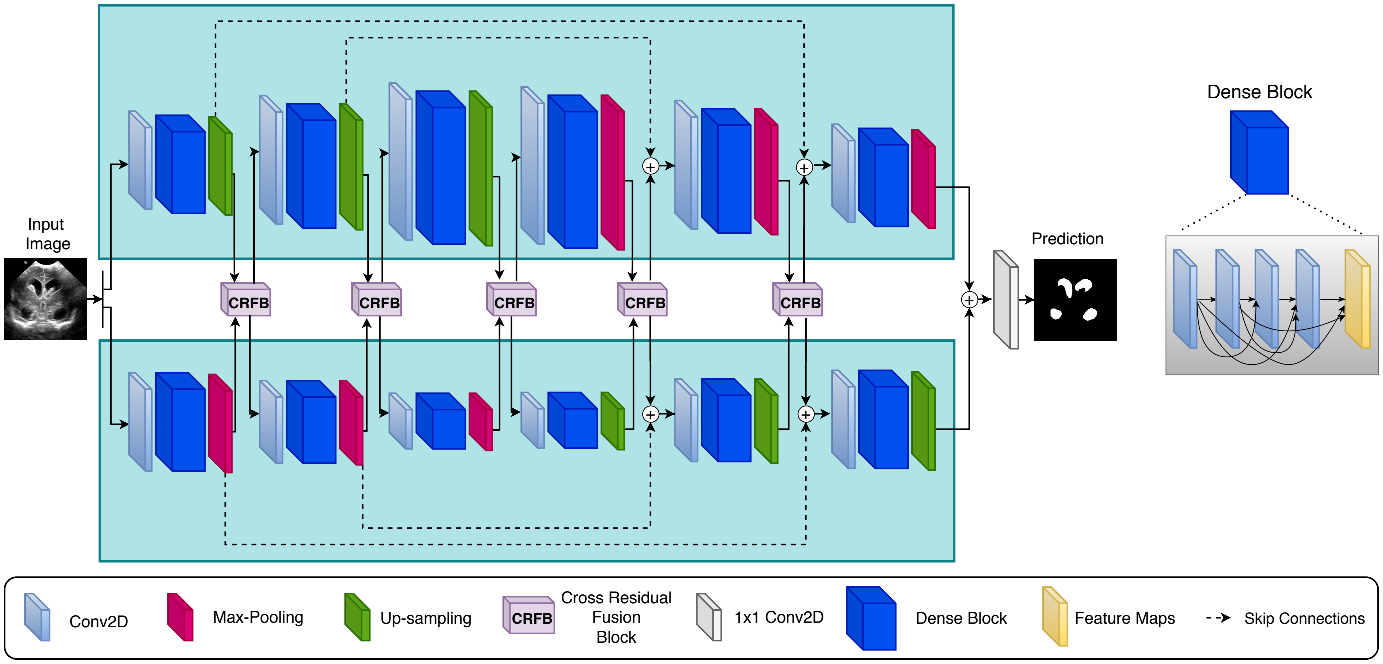


Fig. 12. Details of Dense-KiU-Net architecture. The input image is forwarded to the two branches of Dense-KiUNet where each branch has dense blocks at each level. The feature maps are added at the last layer and forwarded through a 1×1 conv 2D layer to get the prediction . In the right side of the figure, dense block architecture has been visualized.

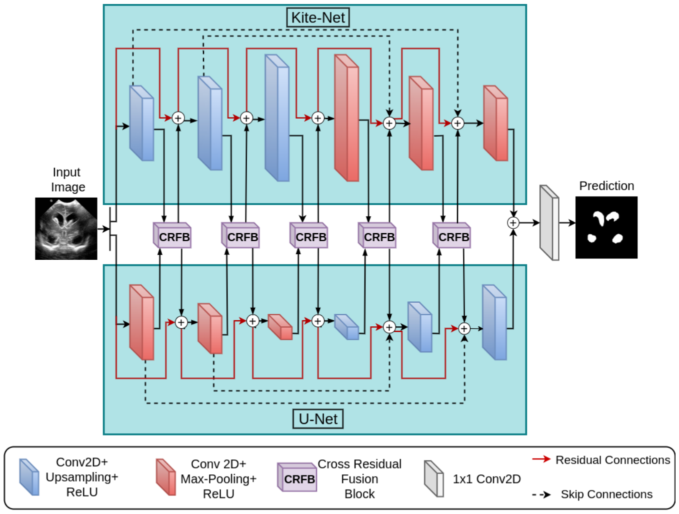


Fig. 13. Details of Res-KiU-Net architecture. The input image is forwarded to the two branches of Res-KiUNet where each branch has residual connections at each level. The feature maps are added at the last layer and passed through a 1×1 conv 2D layer to get the prediction.

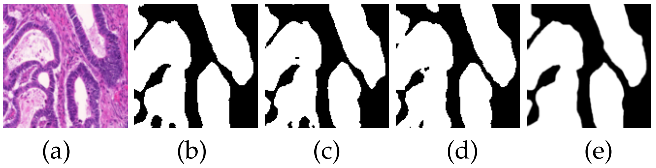


Fig. 14. (a) Input Image, Prediction using (b) KiU-Net (c) Res-KiUNet (d) Dense KiU-Net (e) Ground Truth for GLAS dataset. It can be observed that the predictions of Res-KiUNet and Dense-KiUNet are better in terms of quality when compared to KiU-Net.

TABLE 6
Comparison of performance metrics for Res-KiUNet and Dense KiUNet using GLAS dataset.

Performance Metrics	KiU-Net	Res-KiUNet	Dense-KiUNet
Dice	0.8325	0.8385	**0.8431**
Jaccard	0.7278	0.7303	**0.7422**

6.4.1 *Res-KiUNet*

In Res-KiUNet, we employ residual connections in both the branches of KiU-Net. We use residual learning in every conv block at each level in both the encoder and decoder part of both branches. If x and y are the input and output of each conv block ($F()$) of our network, the residual connection can be formulated as follows:

$$y = F(x) + x.$$

We illustrate the architecture details of Res-KiUNet in Fig 13 where the residual connections are denoted using red arrows. Residual connections are helpful in efficient learning of the network since we can propagate gradients to initial layers faster and thus solving the problem of vanishing gradients.

6.4.2 *Dense-KiUNet*

In Dense-KiUNet, we employ dense blocks after every conv layer in both the branches. We use a dense block of 4 conv layers where the input consists of k feature maps. Each conv layer outputs $k/4$ feature maps which is concatenated with the input to all the next conv layers. The output of all these conv layers are then concatenated to obtain k output feature maps. This is added with the input and sent to the next layer in the network. The output of the dense block is forwarded to a max-pooling layer in the encoder of undercomplete branch and in the decoder of overcomplete branch. In the encoder of overcomplete branch and in the decoder of undercomplete branch, the output of the dense block is forwarded to an upsampling layer. Fig 12 illustrates the architecture details of Dense-KiUNet and the dense block we have used.

To evaluate both Res-KiUNet and Dense KiUNet, we conduct experiments on the GlaS dataset and report the dice and Jaccard metrics in Table 6. Further, we visualize the results of these experiments in Fig 14. It can be observed that Dense-KiUNet and Res-KiUNet provide further improvements in performance as compared to KiU-Net.

7 Conclusion

In this paper, we proposed KiU-Net and KiU-Net 3D for image and volumetric segmentation, respectively. These are two-branch networks consisting of an undercomplete and an overcomplete autoencoder. Our novelty lies in proposing overcomplete convolutional architecture (Kite-Net) for learning small masks and finer details of surfaces and edges more precisely. The two branches are effectively fused using a novel cross-residual feature fusion method that results effective training. Further, we experiment with different variants of the proposed method like Res-KiUNet and Dense-KiUNet. We conduct extensive experiments for image and volumetric segmentation on 5 datasets spanning over 5 different modalities. We demonstrate that the proposed method performs significantly better when compared to the recent segmentation methods. Furthermore, we also show that our network comes with additional benefits such as lower model complexity and faster convergence.

Appendix A
Architecture Details

Tables 7 and 8 show the configuration of the Ki-Net and U-Net branch of our KiU Network. H and W are 128 each for the ultrasound train and test images.

TABLE 7
Configuration of the Ki-Net branch of KiU-Net.

Block name	Layer	Kernel size/Scale Factor	Filters	Padding	Input size	Output size
Encoder	Conv1	3×3	32	1	$1 \times H \times W$	$32 \times H \times W$
	Upsampling	2×2	-	-	$32 \times H \times W$	$32 \times 2H \times 2W$
	ReLU	-	-	-	$32 \times 2H \times 2W$	$32 \times 2H \times 2W$
	Conv2	3×3	64	1	$32 \times 2H \times 2W$	$64 \times 2H \times 2W$
	Upsampling	2×2	-	-	$64 \times 2H \times 2W$	$64 \times 4H \times 4W$
	ReLU	-	-	-	$64 \times 4H \times 4W$	$64 \times 4H \times 4W$
	Conv3	3×3	128	1	$64 \times 4H \times 4W$	$128 \times 4H \times 4W$
	Upsampling	2×2	-	-	$128 \times 4H \times 4W$	$2C \times 8H \times 8W$
	ReLU	-	-	-	$128 \times 8H \times 8W$	$128 \times 8H \times 8W$
Decoder	Conv1	3×3	128	1	$128 \times 8H \times 8W$	$128 \times 8H \times 8W$
	Max-Pooling	2×2	-	-	$128 \times 8H \times 8W$	$128 \times 4H \times 4W$
	ReLU	-	-	-	$128 \times 4H \times 4W$	$128 \times 4H \times 4W$
	Conv2	3×3	64	1	$128 \times 4H \times 4W$	$128 \times 4H \times 4W$
	Max-Pooling	2×2	-	-	$64 \times 4H \times 4W$	$64 \times 2H \times 2W$
	ReLU	-	-	-	$64 \times 2H \times 2W$	$64 \times 2H \times 2W$
	Conv3	3×3	32	1	$64 \times 2H \times 2W$	$32 \times 2H \times 2W$
	Max-Pooling	2×2	-	-	$32 \times 2H \times 2W$	$32 \times H \times W$
	ReLU	-	-	-	$32 \times H \times W$	$32 \times H \times W$

TABLE 8
Configuration of the U-Net branch of KiU-Net.

Block name	Layer	Kernel size/Scale Factor	Filters	Padding	Input size	Output size
Encoder	Conv1	3×3	32	1	$1 \times H \times W$	$32 \times H \times W$
	MaxPooling	2×2	-	-	$32 \times H \times W$	$32 \times H/2 \times W/2$
	ReLU	-	-	-	$32 \times H/2 \times W/2$	$32 \times H/2 \times W/2$
	Conv2	3×3	64	1	$32 \times H/2 \times W/2$	$64 \times H/2 \times W/2$
	MaxPooling	2×2	-	-	$64 \times H/2 \times W/2$	$64 \times H/4 \times W/4$
	ReLU	-	-	-	$64 \times H/4 \times W/4$	$64 \times H/4 \times W/4$
	Conv3	3×3	128	1	$64 \times H/4 \times W/4$	$128 \times H/4 \times W/4$
	MaxPooling	2×2	-	-	$128 \times H/4 \times W/4$	$128 \times H/8 \times W/8$
	ReLU	-	-	-	$128 \times H/8 \times W/8$	$128 \times H/8 \times W/8$
Decoder	Conv1	3×3	128	1	$512 \times H/32 \times W/32$	$128 \times H/32 \times W/32$
	Upsampling	2×2	-	-	$128 \times H/32 \times W/32$	$128 \times H/16 \times W/16$
	ReLU	-	-	-	$128 \times H/16 \times W/16$	$128 \times H/16 \times W/16$
	Conv2	3×3	64	1	$128 \times H/16 \times W/16$	$64 \times H/16 \times W/16$
	Upsampling	2×2	-	-	$64 \times H/16 \times W/16$	$64 \times H/8 \times W/8$
	ReLU	-	-	-	$64 \times H/8 \times W/8$	$64 \times H/8 \times W/8$
	Conv3	3×3	32	1	$64 \times H/8 \times W/8$	$32 \times H/8 \times W/8$
	Upsampling	2×2	-	-	$32 \times H/8 \times W/8$	$32 \times H/4 \times W/4$
	ReLU	-	-	-	$32 \times H/4 \times W/4$	$32 \times H/4 \times W/4$

Acknowledgments

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[3] 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*, pp. 3–11, Springer, 2018.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[5] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[8] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331, IEEE, 2018.

[9] 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 transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[12] J. M. Jose, V. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using overcomplete representations," *arXiv preprint arXiv:2006.04878*, 2020.

[13] J. M. J. Valanarasu, R. Yasarla, P. Wang, I. Hacihaliloglu, and V. M. Patel, "Learning to segment brain anatomy from 2d ultrasound with less data," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[14] P. Wang, N. G. Cuccolo, R. Tyagi, I. Hacihaliloglu, and V. M. Patel, "Automatic real-time cnn-based neonatal brain ventricles segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 716–719, IEEE, 2018.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[16] M. Martin, B. Sciolla, M. Sdika, X. Wang, P. Quetin, and P. Delachartre, "Automatic segmentation of the cerebral ventricle in neonates using deep learning with 3d reconstructed freehand

ultrasound imaging," in *2018 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2018.

[17] Y.-T. Weng, H.-W. Chan, and T.-Y. Huang, "Automatic segmentation of brain tumor from 3d mr images using segnet, u-net, and psp-net," in *International MICCAI Brainlesion Workshop*, pp. 226–233, Springer, 2019.

[18] L. Fang and H. He, "Three pathways u-net for brain tumor segmentation," in *Pre-conference proceedings of the 7th medical image computing and computer-assisted interventions (MICCAI) BraTS Challenge*, vol. 2018, pp. 119–126, 2018.

[19] N. Fridman, "Brain tumor detection and segmentation using deep learning u-net on multi modal mri," in *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*, pp. 135–143, 2018.

[20] A. Kermi, I. Mahmoudi, and M. T. Khadir, "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes," in *International MICCAI Brainlesion Workshop*, pp. 37–48, Springer, 2018.

[21] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan, "Pixelnet: Representation of the pixels, by the pixels, and for the pixels," *arXiv preprint arXiv:1702.06506*, 2017.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

[23] M. Islam, V. J. M. Jose, and H. Ren, "Glioma prognosis: segmentation of the tumor and survival prediction using shape, geometric and clinical information," in *International MICCAI Brainlesion Workshop*, pp. 142–153, Springer, 2018.

[24] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3d-unet: separable 3d u-net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, pp. 358–368, Springer, 2018.

[25] U. Baid, S. Talbar, S. Rane, S. Gupta, M. H. Thakur, A. Moiyadi, S. Thakur, and A. Mahajan, "Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas," in *International MICCAI Brainlesion Workshop*, pp. 369–379, Springer, 2018.

[26] M. Islam, V. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3d attention unet," in *International MICCAI Brainlesion Workshop*, pp. 262–272, Springer, 2019.

[27] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*, pp. 311–320, Springer, 2018.

[28] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *International MICCAI Brainlesion Workshop*, pp. 234–244, Springer, 2018.

[29] G. Yang, G. Li, T. Pan, Y. Kong, J. Wu, H. Shu, L. Luo, J.-L. Dillenseger, J.-L. Coatrieux, L. Tang, *et al.*, "Automatic segmentation of kidney and renal tumor in ct images based on 3d fully convolutional neural network with pyramid pooling module," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3790–3795, IEEE, 2018.

[30] Q. Tao, Z. Wu, I. Cheong, J. Yang, Z. Ge, G. Lin, and J. Cai, "3d unet-based kidney and kidney tumer segmentation with attentive feature learning," 2019.

[31] C. Wang, Y. He, X. Qi, Z. Zhao, G. Yang, X. Zhu, S. Zhang, J.-L. Dillenseger, and J.-L. Coatrieux, "Bisc-unet: A fine segmentation framework for kidney and renal tumor," 2019.

[32] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images," *Ieee Access*, vol. 7, pp. 21420–21428, 2019.

[33] H. Hu, Y. Zheng, Q. Zhou, J. Xiao, S. Chen, and Q. Guan, "Mc-unet: Multi-scale convolution unet for bladder cancer cell segmentation in phase-contrast microscopy images," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1197–1199, IEEE, 2019.

[34] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, *et al.*, "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.

[35] S. E. A. Raza, L. Cheung, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Mimo-net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 337–340, IEEE, 2017.

[36] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[37] W. Liu, Y. Sun, and Q. Ji, "Mdan-unet: Multi-scale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images," *Algorithms*, vol. 13, no. 3, p. 60, 2020.

[38] S. Cheon, E. Yang, W. J. Lee, and J.-H. Kim, "Cai-unet for segmentation of liver lesion in ct image," in *Medical Imaging 2020: Image Processing*, vol. 11313, p. 1131325, International Society for Optics and Photonics, 2020.

[39] J. Tian, L. Liu, Z. Shi, and F. Xu, "Automatic couinaud segmentation from ct volumes on liver using glc-unet," in *International Workshop on Machine Learning in Medical Imaging*, pp. 274–282, Springer, 2019.

[40] Q. Jin, Z. Meng, C. Sun, L. Wei, and R. Su, "Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans," *arXiv preprint arXiv:1811.01328*, 2018.

[41] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[42] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[43] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[45] S. Druckmann and D. B. Chklovskii, "Over-complete representations on recurrent neural networks can support persistent percepts," in *Advances in Neural Information Processing Systems*, pp. 541–549, 2010.

[46] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, vol. 35, pp. 489–502, 2017.

[47] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[48] Q. Hu, M. D. Abràmoff, and M. K. Garvin, "Automated separation of binary overlapping trees in low-contrast color retinal images," in *International conference on medical image computing and computer-assisted intervention*, pp. 436–443, Springer, 2013.

[49] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[50] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170117, 2017.

[51] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[52] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[53] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.