KiU-Net: Towards Accurate Segmentation of Biomedical Images using Over-complete Representations

Jeya Maria Jose Valanarasu[1], Vishwanath A. Sindagi[1], Ilker Hacihaliloglu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD, USA
[2] Rutgers, The State University of New Jersey, NJ, USA

Abstract. Due to its excellent performance, U-Net is the most widely used backbone architecture for biomedical image segmentation in the recent years. However, in our studies, we observe that there is a considerable performance drop in the case of detecting smaller anatomical structures with blurred noisy boundaries. We analyze this issue in detail, and address it by proposing an over-complete architecture (Ki-Net) which involves projecting the data onto higher dimensions (in the spatial sense). This network, when augmented with U-Net, results in significant improvements in the case of segmenting small anatomical landmarks and blurred noisy boundaries while obtaining better overall performance. Furthermore, the proposed network has additional benefits like faster convergence and fewer number of parameters. We evaluate the proposed method on the task of brain anatomy segmentation from 2D Ultrasound (US) of preterm neonates, and achieve an improvement of around 4% in terms of the DICE accuracy and Jaccard index as compared to the standard-U-Net, while outperforming the recent best methods by 2%. Code: https://github.com/jeya-maria-jose/KiU-Net-pytorch

Keywords: over-complete representations, ultrasound, brain, deep learning, segmentation, preterm neonate

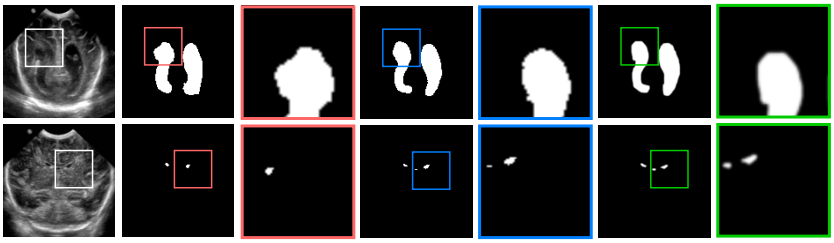
1 Introduction

Preterm birth is among the leading public health problems in the USA and Europe [14]. The reported annual cost of care for preterm neonates exceeds $18 billion dollars every year in the USA alone [14]. Although, advancements made in neonatal care have increased the survival rates, majority of these infants are at risk for adverse neuro-developmental outcomes. Among the different types of preterm brain injury, intraventricular hemorrhage (IVH) remains the most common cause of acquired hydrocephalus resulting in the enlargement of ventricles. On the other hand, absence of septum pellucidum is used as a biomarker for the diagnosis of other brain disorders such as septo-optic dysplasia. Cranial ultrasound (US) remains the main imaging modality used to diagnose brain disorders in preterm neonates due to its real-time, safe, and cost effective imaging

capabilities. Current clinical evaluation involves qualitative investigation of the collected US scans or quantitative manual measurement of landmarks such as ventricular index (VI), anterior horn width (AHW), frontal and temporal horn ratio (FTHR) [4]. Qualitative evaluation is subjective and manual measurement involves intra and inter-user variability errors. The diagnostic accuracy is further affected by the unclear boundary of the ventricles, due to build up of bleeding pressure, or sub-optimal orientation of the transducer during imaging. Additionally, shading artifacts causes incomplete boundaries in the acquired US data. Depending on the bleeding extend, the shape of the ventricle varies for different subjects. Finally, manual measurement is also problematic for normal preterm neonates without any brain injury due to very small ventricle size and blurred boundaries. Similar problems are also faced for identifying septum pellucidum due to its small size and unclear boundary. In order to overcome these challenges, precise and automatic segmentation of ventricles and septum pellucidum is critical for accurate diagnosis and prognosis.

Several groups have proposed semi-automatic and fully automatic methods for segmentation of ventricles from 2D/3D US scans. Methods based on traditional medical image analysis are time consuming or not robust enough to the previously mentioned challenging scan conditions [2, 19, 16]. The reported DICE similarity coefficient values were 70.8% [2], 80% [19], and 76.5% [16]. The reported computation times were 54 minutes for [16]. The other methods did not report any computation time. Most recently, methods based on deep learning were also investigated by various groups to improve the robustness and computation time of segmentation [13, 21, 20]. Since the introduction of U-Net [17] in 2015, it has been the leading deep learning-based network of any method that deals with biomedical image segmentation [3, 15, 23, 12, 8, 7, 24]. In [13], a U-Net architecture was used for segmentation of ventricles.

Based on the observations that the existing approaches do not achieve optimal performance (especially in the case of segmenting out small anatomical structure), we analyze this issue in detail. Specifically, we conducted experiments with the standard U-Net architecture which is a leading backbone in several segmentation algorithms. In spite of the skip connections that enable the propagation of information from shallower layers to deeper layers, the network is unable to capture finer details (see Fig. 1) for the following reasons. The standard encoder-decoder architecture of U-Net belongs to the family of under-complete convolutional autoencoders, where the dimensionality of data is reduced near the bottleneck. The initial few blocks of the encoder learn low-level features of the data while the later blocks learn the high-level features. Eventually, the encoder learns to map the data to lower dimensionality (in the spatial sense). The increasing receptive field size over the depth of the network, constrains the network to focus more on the higher-level features. However, it is important to note that tiny structures require smaller receptive fields. In the case of standard U-Net, even with skip connections, the smallest receptive field is limited by that of the first layer. Hence, under-complete architectures are essentially limited in their abilities to capture finer details.



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. (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. It can be seen that our proposed network captures edges and small masks better than U-Net.

Considering the aforementioned drawback of under-complete representations, we resort to over-complete architectures where the data is projected onto a higher dimension in the intermediate layers. In the literature, over-complete representations have been shown to be more robust and stable, especially in the presence of noise [11]. However, such architectures have been relatively unexplored for segmentation tasks in both the computer vision and medical imaging communities [6]. In this paper, we explore the use of such an over-complete network for segmentation to address the issue of lack of smaller receptive field in the standard U-Net. We refer to the over-complete network as Kite-Net (Ki-Net) as it's shape is similar to that of a kite. In the following sections, we show how the information learned by Ki-Net actually helps in capturing finer shape structures and edges better than the generic under-complete networks. Furthermore, we propose to effectively combine the benefits of the proposed Ki-Net with that of the standard U-Net using a novel cross-scale fusion strategy. We show that this novel network (KiU-Net) achieves state-of-the-art performance on the brain anatomy segmentation task from US images when compared with the latest methods.

In summary, this paper (1) explores over-complete deep networks (Ki-Net) for the task of segmentation, (2) proposes a novel architecture (KiU-Net) combining the features of both under-complete and over-complete deep networks which captures finer details better than the standard encoder-decoder architecture of U-Net thus aiding in precise segmentation, and (3) achieves faster convergence and better performance metrics than recent methods for segmentation.

2 Proposed Method

Over-complete representations: As illustrated in Fig 2, the receptive field of the filters in a generic "encoder-decoder" architecture increases as we go deeper in the network. This increase in receptive field size can be attributed to two reasons: (i) every conv layer filter gathers information from a surrounding window, and (ii) the use of max-pooling layer after every conv layer. The max-pooling layers essentially double the receptive field size after every conv layer. The in-

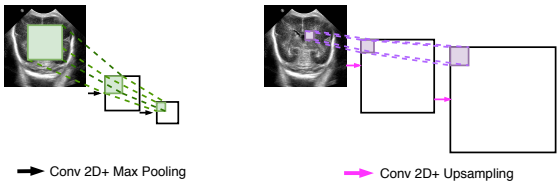


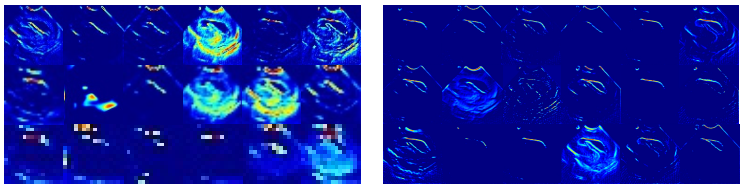
Fig. 2. Effect of architecture type on receptive field. (a) U-Net: Each location in the intermediate layers focuses on a much larger region in the input. (b) Ki-Net: Each location in the intermediate layers focuses on a much smaller region in the input.

creasing receptive field reasons is critical for CNNs to learn high-level features like objects, shapes or blobs. However, a side effect of this is that it reduces the focus of the filters. That is, except the first layer, filters in the other layers have reduced abilities to learn features that correspond to fine details like edges and their texture. This causes any network with the standard under-complete architecture to not produce sharp predictions around the edges in tasks like segmentation.

To overcome this issue, we propose Ki-Net which is over-complete in the spatial sense. That is, the spatial dimensions of the intermediate layers is more than that of the input data. We achieve this by employing an upsampling layer after every conv layer in the encoder. Furthermore, we employ a max pooling layer after every conv layer in the decoder in order to reduce the dimensionality back to that of the input. This forces the over-complete conv architecture to behave differently than the standard under-complete conv architecture. The filters in this type of architecture learn finer low-level features due to the decreasing size of receptive field even as we go deeper in the encoder network.

Fig 2(a) illustrates how the receptive field is large for U-Net. Fig 2(b) illustrates how the use of over-complete architecture like Ki-Net restricts the receptive field size to a smaller region. Hence, by constricting the receptive field size, we force the filters in the deeper layers to learn very fine edges as it tries to focus heavily on smaller regions. To illustrate this, we show how the filters of encoder fire in a Ki-Net when compared to U-Net in Fig 3. It can be observed that the filters in U-Net become smaller as we go deeper and fire across high-level shapes where as the filters become bigger as we go deeper in Ki-Net and the features captured are fine edges across all layers with an increased resolution.

KiU-Net: As we have established that our proposed Ki-Net has better abilities to captures edges compared to U-Net, we combine it with the standard U-Net in order to improve the overall segmentation accuracy as Ki-Net if used separately will only capture the edges. The combined network, KiU-Net, exploits the low-level fine edges capturing feature maps of Ki-Net as well as the high-level shape capturing feature maps of U-Net. We propose using a parallel network architecture where one branch is a Ki-Net and the other a U-Net as seen in Figure 4(a). The input image is forwarded through both the branches simultaneously. In both the branches, we have 3 layers of conv blocks in the encoder as well as the decoder. Each conv block in the encoder of Ki-Net branch consists of a 2D



(a) (b)

Fig. 3. Visualization of filter responses for (a) U-Net, and (b) Ki-Net. Top row: Feature maps from the first layer of encoder. Middle row: Feature maps from the second layer of encoder. Bottom row: Feature maps from the third layer of encoder. By restricting the receptive field, Ki-Net is able to focus on edges and smaller regions.

conv layer followed by a bilinear interpolation with a scale factor of 2 and ReLU non-linearity. Similarly, each conv block in the decoder of Ki-Net branch consists of a 2D conv layer followed by a max-pooling layer with a pooling coefficient of two. In addition, we use skip connections between the blocks of encoder and decoder similar to U-Net to enhance the localization. In the U-Net branch, we adopt the "encoder-decoder" architecture of a U-Net.

In order to augment the two networks, one can perform simple concatenation of features at the final layer. However, this may not be necessarily optimal. Instead, we combine the feature maps at each block and this results in better convergence as the flow of gradients during back propagation is across both the branches at each block level [18]. Furthermore, in order to combine the features at each block level more effectively, we propose a cross residual fusion block (CRFB). This block extracts complementary features from both network branches and forwards to both of them respectively. Specifically, the CRFB consists of residual connections, followed by a set of conv layers (see Fig. 4 (b)). In order to combine the feature maps from the two networks F_U^i (U-Net) and F_{Ki}^i (Ki-Net) after the i^{th} block, cross-residual features R_U^i and R_{Ki}^i are first estimated through a set of conv layers. These cross-residual feature are then added to the original features F_U^i (U-Net) and F_{Ki}^i to obtain the complementary features $\hat{F}_U^i$ and $\hat{F}_{Ki}^i$, *i.e*, $\hat{F}_U^i = F_U^i + R_{Ki}^i$ and $\hat{F}_{Ki}^i = F_{Ki}^i + R_U^i$. This strategy is more effective compared to simple feature fusion schemes like addition or concatenation. Finally, the features from decoder in both the branches are added and forwarded through 1×1 conv layer to produce the final segmentation mask. The complete details of the network such as the kernel size, number of filters, etc. are included in supplementary material.

We train the network using pixel-wise binary cross entropy loss between the prediction and ground-truth. The loss function between the prediction p and the ground truth $\hat{p}$ 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, $p(x,y)$ and $\hat{p}(x,y)$ denote the output at a specific location (x,y) of the prediction and ground truth, respectively.

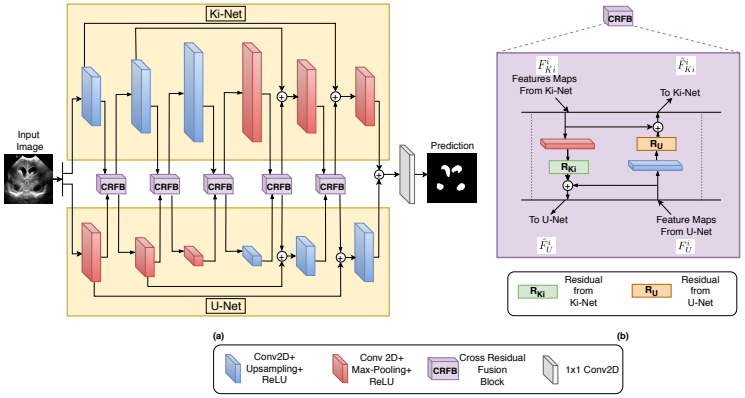


Fig. 4. (a) An overview of the proposed KiU-Net architecture. (b) Cross Residual Fusion block architecture.

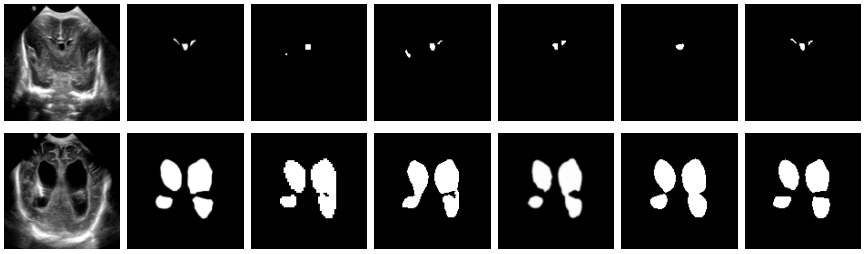
3 Experiments and results

Dataset acquisition and details: After obtaining institutional review board (IRB) approval, US scans were collected from 20 different premature neonates (age < 1 year). The dataset contains subjects with IVH as well as healthy ones. The US scans were collected using a Philips US machine (Philips iE33) with a C8-5 broadband curved array transducer using coronal and sagittal scan planes. Imaging depth and resolution varied between 6-8 cm and 0.1-0.15 mm, respectively. Ventricles and septum pellecudi were manually segmented by an expert ultrasonographer. A total of 1629 images with annotations were obtained in total. The scans were randomly divided into 1300 images for training and 329 images for testing. This process was repeated 3 times. During random split the training and testing data did not include scans from the same patient. Before processing the resolution of each image was changed to 128×128.

Implementation details: KiU-Net is trained using cross-entropy loss $\mathcal{L}_{CE(p,\hat{p})}$ with the Adam optimizer [10] and a batch-size of 1. The learning rate was set equal to 0.001. The network was built in PyTorch framework and trained using Nvidia-RTX 2080Ti GPUs. The network was trained for a total of 100 epochs.

Comparison with recent methods: Since the main focus of this work is to augment the U-Net architecture with additional capabilities, we compare our method with U-Net and other recent methods. Table 1 shows that the proposed method performs better than other recent methods like Seg-Net [1], pix2pix [9], and Wang et al. [21]. Seg-Net [1] has been most recently investigated for segmentation of kidneys from US data [22], pix2pix [9] has been used for multi-task organ segmentation from chest x-ray radiography[5], and Wang *et al.* [21] has been previously used for segmentation of ventricles from brain US data. We run the experiments 3 times for different random folds of training and testing data and report the mean metrics with the variance.

It can be observed that the proposed method achieves an improvement of 4% in DICE accuracy with respect to U-Net and a 2% improvement with respect to state-of-the-art [21] (see Table 1). Fig. 5 illustrates the prediction of segmentation



(a) (b) (c) (d) (e) (f) (g)

Fig. 5. Qualitative results on sample test images. (a) B-mode input US image. (b) Ground truth. (c) Seg-Net [1]. (d) U-Net [17] (e) pix2pix [9]. (f) Wang et al. [21]. (g) KiU-Net (ours).

masks using different methods along with the input and ground truth. From the first row in Fig. 5, we can observe that KiU-Net (our method) is able to predict even very small masks precisely, whereas all the other methods fail. Similarly, from the second row we can observe that our network detects the edges better than other methods. This demonstrates that the intuition of constricting the receptive field size by following the over-complete representation served its purpose as the smaller masks are not missed in our method. Additionally, it may be noted that the proposed method performs well irrespective of the size of the anatomy structures. Furthermore, the proposed network has the following additional benefits. First, it uses much fewer number of parameters in comparison to the other methods. Note that U-Net used in KiU-Net has less number of blocks and filters compared to the original U-Net as in [17], thus resulting in less number of parameters. Second, it converges much faster compared to the standard U-net (see Fig. 6). Its inference time is 8 ms for one test image.

Table 1. Comparison of results. Proposed method outperforms existing approaches.

Method	DICE Acc (%)	Jaccard Idx (%)	Parameters
Seg-Net [1]	82.79 ± 0.320	75.02 ± 0.570	12.5M
U-Net [17]	85.37 ± 0.002	79.31 ± 0.065	3.1M
pix2pix [9]	85.46 ± 0.022	77.45 ± 0.56	54.4M
Wang et.al[21]	87.47 ± 0.080	80.51 ± 0.190	6.1M
KiU-Net (ours)	**89.43 ± 0.013**	**83.26 ± 0.047**	**0.29M**

Ablation study: We study the performance of each block's contribution to our KiU-Net by conducting a detailed ablation study. The results are shown in Fig 7. We start with the standard under-complete architecture (UC) and the over-complete architecture (OC). It can be noted here that the performance of OC is lesser than UC because even though OC captures the edges properly it does not capture most high level features like UC. Then, we show that fusing both the networks (OC+UC) just by combining the feature maps at the final layer helps in improving the performance. This is followed by an experiment where we use skip connections (SK). It may be noted that UC with SK is basically the U-Net. Finally, we incorporate the cross residual fusion block (CRFB) at each block

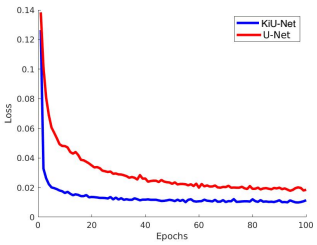


Fig. 6. Comparison of convergence of the loss between KiU-Net and U-Net.

Method	DICE	Jaccard
UC	82.79	75.02
OC	56.04	43.97
OC+UC	84.80	76.48
UC with SK	85.37	79.31
OC with SK	60.38	47.86
OC+UC with SK	86.24	78.11
KiU-Net (ours)	**89.43**	**83.26**

Fig. 7. Ablation study.

level in our KiU-Net, resulting in further improvements which demonstrates the effectiveness of our novel cross fusion strategy. Fig 8 illustrates the qualitative improvements after adding each major block.

(a) (b) (c) (d) (e) (f) (g)

Fig. 8. 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)

More results on different datasets can be found in supplementary material.

4 Conclusion

We proposed a novel network called KiU-Net which is constructed by augmenting the standard under-complete architecture based U-Net with an over-complete structure (Ki-Net). The purpose of Ki-Net is to specifically capture fine edges and small anatomical structures which are typically missed out in the other methods. Further, we incorporate a new fusion strategy that is based on cross-scale residual blocks which results in a more effective use of information from the two networks. The proposed network has additional benefits like it uses much fewer number of parameters and results in faster convergence. The proposed method achieves better performance as compared to recent methods on a relatively complex dataset which has both small and big segmentation masks.

Acknowledgement

This work was supported by the NSF grant 1910141.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Boucher, M.A., Lippé, S., Damphousse, A., El-Jalbout, R., Kadoury, S.: Dilatation of lateral ventricles with brain volumes in infants with 3d transfontanelle us. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 557–565. Springer (2018)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 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)
4. El-Dib, M., Massaro, A.N., Bulas, D., Aly, H.: Neuroimaging and neurodevelopmental outcome of premature infants. American journal of perinatology **27**(10), 803–818 (2010)
5. Eslami, M., Tabarestani, S., Albarqouni, S., Adeli, E., Navab, N., Adjouadi, M.: Image to images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. arXiv preprint arXiv:1906.10089 (2019)
6. Haque, I.R.I., Neubert, J.: Deep learning approaches to biomedical image segmentation. Informatics in Medicine Unlocked **18**, 100297 (2020)
7. Islam, M., Vaidyanathan, N.R., Jose, V.J.M., Ren, H.: Ischemic stroke lesion segmentation using adversarial learning. In: International MICCAI Brainlesion Workshop. pp. 292–300. Springer (2018)
8. Islam, M., Vibashan, V., Jose, V.J.M., Wijethilake, N., Utkarsh, U., Ren, H.: Brain tumor segmentation and survival prediction using 3d attention unet. In: International MICCAI Brainlesion Workshop. pp. 262–272. Springer (2019)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lewicki, M.S., Sejnowski, T.J.: Learning overcomplete representations. Neural computation **12**(2), 337–365 (2000)
12. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
13. Martin, M., Sciolla, B., Sdika, M., Wang, X., Quetin, P., Delachartre, P.: 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)
14. Ment, L.R., Hirtz, D., Hüppi, P.S.: Imaging biomarkers of outcome in the developing preterm brain. The Lancet Neurology **8**(11), 1042–1055 (2009)
15. Milletari, F., Navab, N., Ahmadi, S.A.: 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)

16. Qiu, W., Chen, Y., Kishimoto, J., de Ribaupierre, S., Chiu, B., Fenster, A., Yuan, J.: Automatic segmentation approach to extracting neonatal cerebral ventricles from 3d ultrasound images. Medical image analysis **35**, 181–191 (2017)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1002–1012 (2019)
19. Tabrizi, P.R., Obeid, R., Cerrolaza, J.J., Penn, A., Mansoor, A., Linguraru, M.G.: Automatic segmentation of neonatal ventricles from cranial ultrasound for prediction of intraventricular hemorrhage outcome. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3136–3139. IEEE (2018)
20. Valanarasu, J.M.J., Yasarla, R., Wang, P., Hacihaliloglu, I., Patel, V.M.: Learning to segment brain anatomy from 2d ultrasound with less data. IEEE Journal of Selected Topics in Signal Processing (2020)
21. Wang, P., Cuccolo, N.G., Tyagi, R., Hacihaliloglu, I., Patel, V.M.: 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)
22. Yin, S., Peng, Q., Li, H., Zhang, Z., You, X., Fischer, K., Furth, S.L., Tasian, G.E., Fan, Y.: Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks. Medical image analysis **60**, 101602 (2020)
23. Zhao, N., Tong, N., Ruan, D., Sheng, K.: Fully automated pancreas segmentation with two-stage 3d convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 201–209. Springer (2019)
24. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)