VOXELS INTERSECTING ALONG ORTHOGONAL LEVELS ATTENTION U-NET FOR INTRACEREBRAL HAEMORRHAGE SEGMENTATION IN HEAD CT

Qinghui Liu, Bradley J MacIntosh, Till Schellhorn, Karoline Skogen, Kyrre Eeg Emblem, Atle Bjørnerud

Department of Physics and Computational Radiology, Division of Radiology and Nuclear Medicine,
Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

ABSTRACT

We propose a novel and flexible attention based U-Net architecture referred to as "Voxels-Intersecting Along Orthogonal Levels Attention U-Net" (viola-Unet), for intracranial hemorrhage (ICH) segmentation task in the INSTANCE 2022 Data Challenge on non-contrast computed tomography (CT). The performance of ICH segmentation was improved by efficiently incorporating fused spatially orthogonal and cross-channel features via our proposed Viola attention plugged into the U-Net decoding branches. The viola-Unet outperformed the strong baseline nnU-Net models during both 5-fold cross validation and online validation. Our solution was the winner of the challenge validation phase in terms of all four performance metrics (i.e., DSC, HD, NSD, and RVD). The code base, pretrained weights, and docker image of the viola-Unet AI tool are publicly available at `https://github.com/samleoqh/Viola-Unet`.

Index Terms— 3D U-Net, ICH, head CT, deep learning

1. INTRODUCTION

A spontaneous intracranial hemorrhage (ICH) is the second most common subtype of stroke and a critical disease usually leading to severe disability or death [1]. Accurately estimating the volume of ICH is important in clinical diagnosis procedures for predicting hematoma progression and early mortality [2]. Non-contrast Computed Tomography (CT) is the most commonly used modality in regular clinical practice for diagnosing ICH. The hematoma volume can be calculated by radiologists manually with ABC/2 method [3] in clinical practice. However, the ABC/2 method exhibits significant volume estimation error, particularly for hemorrhages with irregular shapes as shown in Fig. 1. Hence, a fully automated segmentation method that allows accurate and rapid volume quantification of ICH is high desirable.

Deep learning (DL) algorithms have recently received increasing attention in computer-aided automatic methods for medical data analysis. The state-of-the-art medical image segmentation models tend to rely on the popular U-Net [4] architecture, an encoder-decoder convolutional neural network (CNN) based approach with end-to-end training pipeline for pixel- or voxel-wise segmentation. Several U-Net-like models have tackled ICH segmentation using head CT scans [5, 6, 7, 8] and these successes are mirrored in other brain imaging fields such as tumor segmentation of multi-modal MRI scans [9, 10]. Isensee et al., in particular, used the nnU-Net framework [11] to present a winning model for the BraTS20 challenge [12], with a self-configuring method for various DL-based biomedical image segmentation tasks. Thus, we chose nnU-Net as the strong baseline model in the current work.

In this paper, we propose a novel solution for the INSTANCE 2022 ICH segmentation challenge [13], with the goal of making a computationally efficient, accurate, and robust end-to-end 3D deep learning model. Our solution has won the challenge validation phase, with an average DSC score of 0.795.

2. METHODS

Our solution is called "viola-Unet" as it relies on Voxels in feature space that Intersect along Orthogonal Levels to provide an Attention U-Net, which is an asymmetric encoder-decoder architecture with 7-depth layers (shown in Figure 2 (a)). The number of channels at each encoder was 32, 64, 96, 128, 192, 256 and 320, while the channel-numbers at each corresponding decoder layer were 32, 64, 96, *128*, *128* and *128*. In addition, the input patch size was $3 \times 160 \times 160 \times 16$ with 2 extra scales of deep supervision outputs.

2.1. Viola attention module

Squeeze-and-Excitation (SE) networks are able to recalibrate channel-wise feature responses by explicitly modeling inter-dependencies between channels on 2D feature planes[14]. The viola-Unet attention method is similar; Fig. 2 (b) shows how the viola attention module incorporates features along orthogonal directions, an efficient way to incorporate through-plane features.
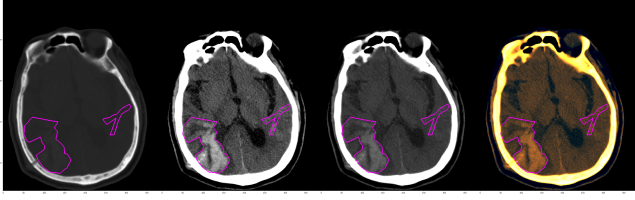
Fig. 1. An example from the INSTANCE2022 dataset. Each images shows different Hounsfield Unit (HU) windowing levels to take advance of an RGB-style 3-window data input combination (from left to right: $[-200 \sim 1300]$, $[0 \sim 100]$, $[-20 \sim 200]$, and 3-window combination), with the ICH labeled regions highlighted by the pink edge lines.

Overall Viola module is composed of three key blocks, i.e., the adaptive average pooling (AdaAvgPool) module that squeezes the input feature volume (e.g., $\mathbf{X} \in \mathbb{R}^{C \times H \times W \times D}$, where C, H, W, and D represent channel, height, width, and depth for a given feature volume.) into three latent representation spaces (e.g., $\mathbf{X}_h \in \mathbb{R}^{C \times H}$, $\mathbf{X}_w \in \mathbb{R}^{C \times W}$, and $\mathbf{X}_d \in \mathbb{R}^{C \times D}$) along each axis of the input feature patch. The customized dense dilated convolutions merging (DDCM) networks [16] fuses cross-channel and non-local contextual information on each orthogonal direction with adaptive kernel sizes (i.e., $k = [2(C//32) + 3, 1]$), dilated ratios (i.e., $dilation = [1, k, 2(k-1) + 1, 3(k-1) + 1]$) and strides (i.e, $strides = [(2,1), (2,1), (4,1), (4,1)]$). The Viola unit constructs the voxels intersecting along orthogonal level attention volume (i.e. $\mathbf{A}_{viola} \in \mathbb{R}^{C \times H \times W \times D}$) based on fused and reshaped cross-channel-direction latent representation spaces (i.e., $\mathbf{X}_h \in \mathbb{R}^{C \times H \times 1 \times 1}, \mathbf{X}_w \in \mathbb{R}^{C \times 1 \times W \times 1}$, and $\mathbf{X}_d \in \mathbb{R}^{C \times 1 \times 1 \times D}$), by the following mathematical equations[1].
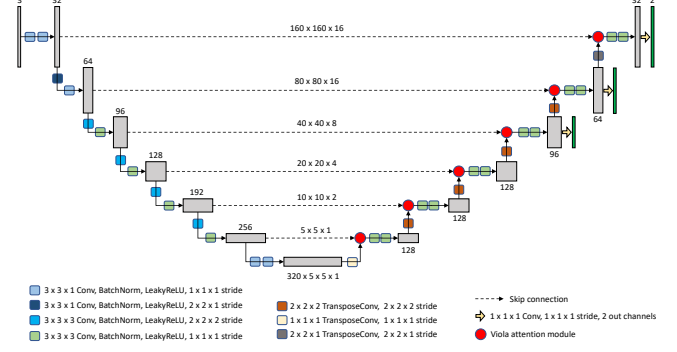
$$\tilde{\mathbf{X}}_h, \tilde{\mathbf{X}}_w, \tilde{\mathbf{X}}_d = \sigma_s \left(\mathbf{X}_h, \mathbf{X}_w, \mathbf{X}_d \right),$$
$$\hat{\mathbf{X}}_h, \hat{\mathbf{X}}_w, \hat{\mathbf{X}}_d = \sigma_{gt} \left(\mathbf{X}_h, \mathbf{X}_w, \mathbf{X}_d \right) \ , \quad (1)$$

$$\begin{aligned} \mathbf{A} = \sigma_r &\left(\tilde{\mathbf{X}}_h + \hat{\mathbf{X}}_h + \tilde{\mathbf{X}}_w + \hat{\mathbf{X}}_w + \tilde{\mathbf{X}}_d + \hat{\mathbf{X}}_d \right) \\ &+ \tilde{\mathbf{X}}_h \otimes \tilde{\mathbf{X}}_w + \tilde{\mathbf{X}}_w \otimes \tilde{\mathbf{X}}_d + \tilde{\mathbf{X}}_d \otimes \tilde{\mathbf{X}}_h \\ &+ \tilde{\mathbf{X}}_h \otimes \tilde{\mathbf{X}}_w \otimes \tilde{\mathbf{X}}_d \ , \end{aligned} \quad (2)$$
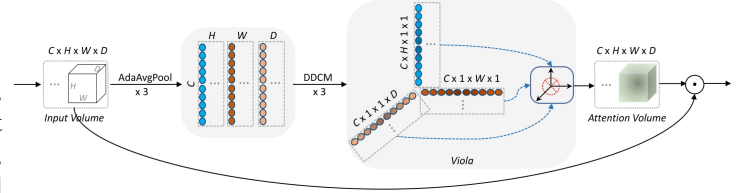
$$\mathbf{A}_{viola} = \left(\alpha + \| f_{latten} \left(\mathbf{A} \right) \|_2^{-1} \right) \mathbf{A} + \beta \ ,$$
$$\mathbf{X} = \mathbf{X} \odot \mathbf{A}_{viola} \ . \quad (3)$$

where σ_s denotes the Sigmoid activation function, σ_{gt} denotes a combination function of group normalization [17] ($G = 2$ in this work) and Tanh non-linearity, σ_r is the ReLU activation operator, $\otimes$ denotes the tensor product and

---

[1]Unless particularly specified, we use bold capital characters for matrices and tensors, lowercase and capital characters in italics for scalars and bold italics for vectors.

(a) The Viola U-Net architecture.

(b) The Viola attention module pipeline.

Fig. 2. The Viola U-Net (viola-Unet) architecture powered by the proposed Voxels Intersecting along Orthogonal Levels Attention (viola) module. Additional two output heads are only used for deep supervision [15] training. Here AdaAvgPool denotes adaptive average pooling, and DDCM denotes dense dilated convolutions' merging network [16].

$\odot$ denotes the element-wise multiplication. Furthermore, two attention coefficients are introduced: $\alpha = 0.1$ to balance the weights of attention maps and $\beta = 0.3$ to weight residual feature maps.

2.2. Architecture considerations:

The viola-Unet is flexible and configurable, i.e. strides and kernel sizes at each layer, number of features in both encoder and decoder layers, symmetric or asymmetric, the number of deep supervision outputs. In addition, we used a self-configured U-Net architecture as a baseline model from the official open source nnU-Net framework [2]. The nnU-Net had a depth of 6. The number of channels at each encoder and decoder (symmetric) level were: 32, 64, 128, 256, 320 and 320. The input path size was $1 \times 320 \times 320 \times 16$ with 5 scales of deep supervision training outputs.

---

[2]https://github.com/MIC-DKFZ/nnUNet

3. DATA, EXPERIMENTS AND RESULTS

3.1. Dataset and evaluation metrics

The INSTANCE 2022 challenge dataset [18, 13] consists of 200 non-contrast 3D head CT scans of clinically diagnosed patients with ICH of various types, such as subdural hemorrhage (SDH), epidural hemorrhage (EDH), intraventricular hemorrhage (IVH), intraparenchymal hemorrhage (IPH), and subarachnoid hemorrhage (SAH). N=100 of the publicly available cases were used for training; the remaining N=100 cases were held-out for the validation set (N=30 for the public leaderboard, and N=70 for the competitor rankings). Model performance was evaluated by four measures: Dice Similarity Coefficient (DSC), Hausdorff distance (HD), Relative absolute Volume Difference (RVD), and the Normalized Surface Dice (NSD).

3.2. Implementation and training

Our code for this study were written in PyTorch with use of the open source Monai [3] library version 0.9.0. We adopted and modified Monai's network codes to implement the proposed models (both viola-Unet and modified nnU-Net).

Guided by our empirical results, we trained all networks with randomly sampled patches of fixed size ($3 \times 160 \times 160 \times 16$) as input and a batch size of 2. Each network was trained with 5-fold cross validation for up to 72,000 steps using stochastic gradient descent (SGD) and an optimizer with Nesterov momentum of 0.99. The initial learning rate was 7×10^{-3} with applying a cosine annealing scheduler [19] to reduce the learning rate over epochs. We used a linear warm-up learning rate during the first 1000 steps. A sliding window inference method was applied to evaluate the model on the local validation set after every 200 training steps. We stored the checkpoint with the highest mean dice score on the validation set of the current fold during the training phase. Based on our training observations to achieve fast and stable convergence for each network, we applied a combination loss function of the dice loss [20] and Focal loss [21] for all our experiments.

3.2.1. self-training strategy

We also utilised self-training strategy to do semi-supervised fine-tune learning on online validation dataset. The semi-supervised learning principle with self-training algorithms is to train a model iteratively by assigning pseudo-labels to the set of unlabeled training samples in conjunction with the labeled training set [22]. In practice, we manually select the best prediction on each validation example from each submission as the pseudo-label and put them into our training set to fine-tune our models repeatedly.

---

[3] https://monai.io/.

Table 1. Average DSC for each of the 5-folds. Results for a base nnU-Net configuration are shown along with a smaller-sized version of viola-Unet (s denotes small).

Model	nnU-Net-base	viola-Unet-s
Fold 0	0.7562	**0.7786**
Fold 1	0.7345	**0.7530**
Fold 2	0.7796	**0.7990**
Fold 3	0.7555	**0.8058**
Fold 4	**0.7746**	0.7730
Mean DSC	0.7601	**0.7819** (+2.18%)

3.3. Results

Table 1 shows the average DSC scores for each 5-folds with the nnU-Net baseline models and viola-Unet models, respectively. The viola-Unet outperforms the baseline nnU-Net by a significant margin (mean DSC +2.18%). Fig. 3 shows some examples of the predictions with the Viola-Unet model on local cross-validation sets.
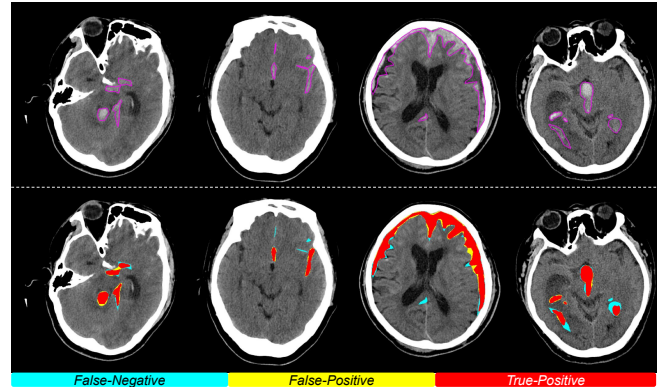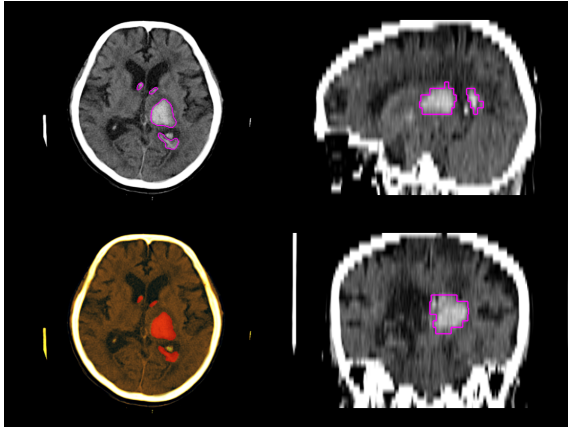

Fig. 3. Axial view segmentation results with the Viola-Unet model on local cross-validation sets. The top row represents the input images with ground-truth regions highlighted with pink edge lines, and the bottom row represents the model segmented results, where areas colored in red denote a true positive (TP), yellow a false positive (FP), and light blue a false negative (FN).
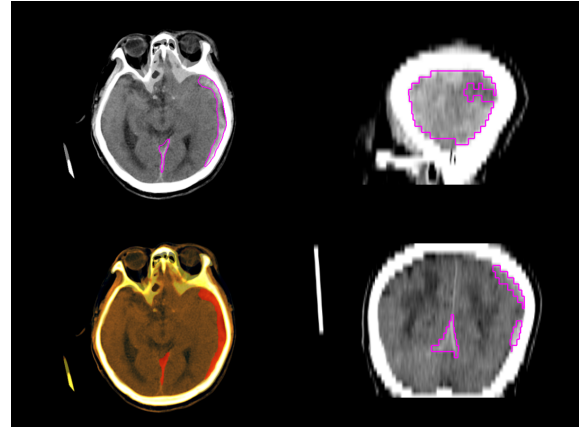
In table 2, we show the top 10 ranking scores for INSTANCE 2022 online validation phase. Our semi-supervise trained viola-Unet-l models outperformed the comparison networks on two out of four performance metrics (i.e., NSD and RVD). An ensemble model that combined viola-Unet-l and re-implemented nnU-Net-r networks had the highest performance for DSC and HD. We show two examples of segmented results from Viola-Unet models on online validation CT scans in Fig. 4.

Table 2. Top 10 ranking scores for INSTANCE 2022 online validation phase [data extracted on 7-Aug-2022]. Note that 3 submissions provided by our team scored in the top-3. A larger version of the viola-Unet (l denotes large) was fine-tuned with self-training and achieved highest validation performance for NSD and RVD scores, while an ensemble of nnU-Net-r with viola-Unet-l was top for DSC and HD scores.

Models	DSC ↑	HD ↓	NSD ↑	RVD ↓
arren	0.7435 ± 0.236	31.616 ± 33.221	0.5201 ± 0.153	0.3580 ± 0.450
asanner	0.7456 ± 0.257	21.805 ± 21.735	0.5239 ± 0.175	1.1381 ± 0.112
dongyuDylan	0.7503 ± 0.237	29.072 ± 26.121	0.5280 ± 0.165	0.2301 ± 0.218
testliver	0.7537 ± 0.236	35.843 ± 28.453	0.5289 ± 0.165	0.2208 ± 0.206
L_Lawliet	0.7640 ± 0.213	34.323 ± 29.207	0.5381 ± 0.145	0.2044 ± 0.175
yangd05	0.7645 ± 0.237	25.725 ± 23.801	0.5403 ± 0.169	0.2322 ± 0.235
amrn	0.7821 ± 0.184	32.296 ± 30.039	0.5528 ± 0.127	0.2027 ± 0.182
nnU-Net-r (our)	0.7943 ± 0.174	22.799 ± 25.423	0.5673 ± 0.129	0.1952 ± 0.182
viola-Unet-l (our)	0.7951 ± 0.171	24.038 ± 29.236	$\mathbf{0.5693} \pm 0.125$	$\mathbf{0.1941} \pm 0.179$
Ensemble (our)	$\mathbf{0.7953} \pm 0.172$	$\mathbf{21.557} \pm 25.021$	0.5681 ± 0.125	0.1980 ± 0.180

(a) DSC=0.898, HD=20.7, NSD=0.699, RVD=0.059 (b) DSC=0.736, HD=50.1, NSD=0.504, RVD=0.17

Fig. 4. Two examples of segmented results from Viola-Unet models on online validation CT scans. We present four views of the same object: axial, sagittal, RGB-axial, and coronal. In the axial, sagittal, and coronal slices, the segmented regions were highlighted with pink edge lines, while the corresponding RGB-axial slices were marked with red color.

4. CONCLUSIONS

We presented the voxels intersecting along orthogonal levels attention (Viola) module, a novel 3D attention framework that uses 3-dimensional orthogonal projections in the feature space to effectively construct fine-grained attention maps with high computational efficiency. Built upon viola module, we design a new flexible segmentation network (Viola-Unet) that can achieve high performance despite a limited training sample size in the field of biomedical imaging. On the validation dataset of the INSTANCE2022 Intracranial Hemorrhage Segmentation challenge, our Viola-Unet models outperform all other models. Importantly, the proposed network is high flexible to address different domain problems with allowing for arbitrary control of strides and kernel sizes at each layer, the number of features in both encoder and decoder layers, symmetric or asymmetric, and so on.

6. REFERENCES

[1] Sang Joon An, Tae Jung Kim, and Byung-Woo Yoon, "Epidemiology, risk factors, and clinical features of intracerebral hemorrhage: an update," *Journal of stroke*, vol. 19, no. 1, pp. 3, 2017.

[2] Joseph P Broderick, Thomas G Brott, John E Duldner, Thomas Tomsick, and Gertrude Huster, "Volume of intracerebral hemorrhage. a powerful and easy-to-use predictor of 30-day mortality.," *Stroke*, vol. 24, no. 7, pp. 987–993, 1993.

[3] R. U. Kothari, T. Brott, J. P. Broderick, W. G. Barsan, L. R. Sauerbeck, M. Zuccarello, and J. Khoury, "The ABCs of measuring intracerebral hemorrhage volumes," *Stroke*, vol. 27, no. 8, pp. 1304–1305, Aug 1996.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[5] Ali Arab, Betty Chinda, George Medvedev, William Siu, Hui Guo, Tao Gu, Sylvain Moreno, Ghassan Hamarneh, Martin Ester, and Xiaowei Song, "A fast and fully-automated deep-learning approach for accurate hemorrhage segmentation and volume quantification in non-contrast whole-head CT," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[6] Murtadha D Hssayeni, Muayad S Croock, Aymen D Salman, Hassan Falah Al-khafaji, Zakaria A Yahya, and Behnaz Ghoraani, "Intracranial hemorrhage segmentation using a deep convolutional model," *Data*, vol. 5, no. 1, pp. 14, 2020.

[7] Matthew F Sharrock, W Andrew Mould, Hasan Ali, Meghan Hildreth, Issam A Awad, Daniel F Hanley, and John Muschelli, "3D deep neural network segmentation of intracerebral hemorrhage: Development and validation for clinical trials," *Neuroinformatics*, vol. 19, no. 3, pp. 403–415, 2021.

[8] N. Yu, H. Yu, H. Li, N. Ma, C. Hu, and J. Wang, "A Robust Deep Learning Segmentation Method for Hematoma Volumetric Detection in Intracerebral Hemorrhage," *Stroke*, vol. 53, no. 1, pp. 167–176, 01 2022.

[9] Michał Futrega, Alexandre Milesi, Michal Marcinkiewicz, and Pablo Ribalta, "Optimized U-Net for Brain Tumor Segmentation," *arXiv preprint arXiv:2110.03352*, 2021.

[10] Huan Minh Luu and Sung-Hong Park, "Extending nn-UNet for brain tumor segmentation," *arXiv preprint arXiv:2112.04653*, 2021.

[11] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[12] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[13] Xiangyu Li, Kuanquan Wang, Jinbo Liu, Hongyu Wang, Mingwang Xu, and Xinjie Liang, "The 2022 Intracranial Hemorrhage Segmentation Challenge on Non-Contrast head CT (NCCT)," Mar. 2022.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[15] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan, "Deeply-supervised cnn for prostate segmentation," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 178–184.

[16] Q. Liu, M. Kampffmeyer, R. Jenssen, and A. B. Salberg, "Dense Dilated Convolutions' Merging Network for Land Cover Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6309–6320, 2020.

[17] Yuxin Wu and Kaiming He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] Xiangyu Li, Gongning Luo, Wei Wang, Kuanquan Wang, Yue Gao, and Shuo Li, "Hematoma expansion context guided intracranial hemorrhage segmentation and uncertainty estimation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1140–1151, 2021.

[19] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," *ICLR*, 2017.

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

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[22] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov, "Self-training: A survey," *arXiv preprint arXiv:2202.12040*, 2022.