Dynamic deformable attention (DDANet) for semantic segmentation

Journal:	*IEEE Journal of Biomedical and Health Informatics*
Manuscript ID	Draft
Manuscript Type:	Special Issue on AI-driven Informatics, Sensing, Imaging and Big Data Analytics for Fighting the COVID-19 Pandemic
Date Submitted by the Author:	n/a
Complete List of Authors:	Rajamani, Kumar; Universität zu Lübeck, Institute of Medical Informatics; Siebert, Hanna; Universität zu Lübeck, Institute of Medical Informatics Heinrich, Mattias; Universität zu Lübeck, Institut of Medical Informatics

Dynamic deformable attention (DDANet) for semantic segmentation

Kumar T. Rajamani, Hanna Siebert, and Mattias P. Heinrich,

***Abstract*— Deep learning based medical image segmentation is an important step within diagnosis, which relies strongly on capturing sufficient spatial context without requiring too complex models that are hard to train with limited labelled data. Training data is in particular scarce for segmenting infection regions of CT images of COVID-19 patients. Attention models help gather contextual information within deep networks and benefit semantic segmentation tasks. The recent criss-cross-attention module aims to approximate global self-attention while remaining memory and time efficient by separating horizontal and vertical self-similarity computations. However, capturing attention from all non-local locations can adversely impact the accuracy of semantic segmentation networks. We propose a new Dynamic Deformable Attention Network (DDANet) that enables a more accurate contextual information computation in a similarly efficient way. Our novel technique is based on a deformable criss-cross attention block that learns both attention coefficients and attention offsets in a continuous way. A deep segmentation network (in our case a U-Net [1]) that employs this attention mechanism is able to capture attention from pertinent non-local locations and also improves the performance on semantic segmentation tasks compared to criss-cross attention within a U-Net on a challenging COVID-19 lesion segmentation task. Our validation experiments show that the performance gain of the recursively applied dynamic deformable attention blocks comes from their ability to capture dynamic and precise (wider) attention context. Our DDANet achieves Dice scores of 73.4% and 61.3% for Ground-Glass-Opacity and Consolidation lesions for COVID-19 segmentation and improves the accuracy by 4.9% points compared to a baseline U-Net.**

***Index Terms*— Attention Mechanism, CCNet, COVID-19, Criss-Cross Attention, Deformable Attention, Segmentation, U-Net**

I. Introduction

THE coronavirus COVID-19 pandemic is having a global impact affecting 213 countries so far. The cases world wide as reported on Worldometers [2] is about 16,482,271 as of end July 2020. Many of the countries have steadily flattened the curve by stringent social distancing measures. In the last several months of managing this pandemic globally, several screening options have become main stream from Nucleic Acid Amplification Tests (NAAT) assay tests, serological tests, and radiological imaging (X-rays, CT). Recent studies have also demonstrated that lack of taste and smell is a new indicator for this virus [3].

The gold-standard for COVID-19 diagnosis is currently using reverse-transcription polymerase chain reaction (RT-PCR) testing [4]. It has been observed that RT-PCR also has several vital limitations. The most pertinent of this limitation is that the test is not universally available. To further compound the drawbacks, the turnaround times for this test is currently lengthy and the sensitivities vary. Some studies have even pointed out that that sensitivity of this test is largely insufficient [4]. To mitigate some of the challenges in rapid screening given the large incidence rate of this virus and limited testing facility, radiological imaging complements and supports immensely stratify therapy options for more severe cases of COVID-19.

Radiological imaging equipment, such as X-ray, are more easily accessible to clinicians and also provide huge assistance for diagnosis of COVID-19. CT imaging and Chest radiographs (CXR) are two of the currently used radiological imaging modalities for COVID-19 screening. Lung CT can detect certain characteristic manifestations associated with COVID-19. Several studies [5] [4] have demonstrated that CT is more sensitive to detect COVID-19, with 97-98%, compared to 71% for RT-PCR [4]. CXR might have lesser scope in the first stages of the disease as the changes are not evident on CXR. Studies have shown [6], [7] that CXR may even present normal in early or mild disease , as demonstrated in Figure 1 [8]. CT is hence preferred for early stage screening and is also generally better than X-rays as it enables three dimensional views of the lung.

The typical signs of COVID-19 infection observed in CT slices are ground glass opacities (GGO), which occur in the early stages and pulmomary consolidation, which occur in later stages. Detection of these regions in CT slices gives vital information to the clinicians and helps in combating COVID-19. Manual detection is laborious, highly time consuming, tedious and error prone. It has to be pointed out that COVID-19 associated abnormalities, such as ground glass opacities and consolidations, are not characteristic for only COVID-19 but can occur in other forms of pneumonia.

Deep Learning plays a vital role in processing these medical images and correctly diagnosing patients with COVID-19. In regular clinical workflow, while assessing the risks for

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment.

Kumar T Rajamani is with the Institute of Medical Informatics, University of Lübeck, Germany (e-mail: kumar.rajamani@uni-luebeck.de).

Hanna Siebert is with the Institute of Medical Informatics, University of Lübeck, Germany (e-mail: siebert@imi.uni-luebeck.de).

Mattias P Heinrich is with the Institute of Medical Informatics, University of Lübeck, Germany (e-mail: heinrich@imi.uni-luebeck.de).

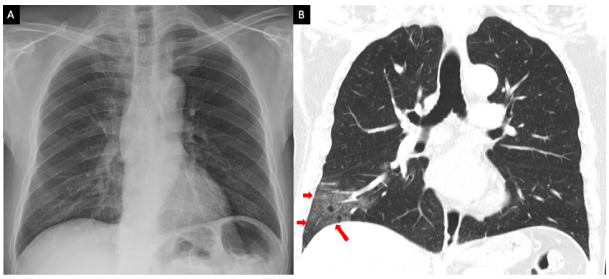


Fig. 1. Comparison of chest radiograph (A) and CT thorax coronal image (B). The ground glass opacities in the right lower lobe periphery on the CT (red arrows) are not visible on the chest radiograph, which was taken 1 hour apart from the first study. Image courtesy - Ming-Yen et al [9]

progression or worsening, the images need to be segmented and quantified. Deep learning based algorithms are able to automatically segment images when trained on manually segmented lesion labels. Several researchers have already established the efficacy of such algorithms on COVID-19 images. One of these early works was DenseUNet proposed by [10] to segment the lesions, lungs and lobes in 3D. They compute percentage of opacity and lung severity scores and report this on entire lung and lobe-wise. The algorithm was trained on 613 manually delineated CT scans (160 COVID-19, 172 viral pneumonia, and 296 ILD). They report Pearson Correlation coefficients between prediction and ground truth above 0.95 for all the four categories. In CovidENet [11] propose a combination of a 2D slice-based and 3D patch-based ensemble architectures, trained on 23423 slices. Their finding was that CovidENet performed equally well as trained radiologists, with a Dice coefficient of 0.7.

For the diagnosis of lung diseases, CT scans have been the preferred modality, and this has therefore been actively utilized in managing COVID-19 [12] [13] [14]. AI in medical imaging has largely aided in automating the diagnosis of COVID-19 from medical images [15] [16]. A detailed review of AI in Diagnosis of COVID-19 has been presented by Shi et al [12]. They broadly group AI based automated assistance for image acquisition, accurate segmentation of organs and infections and for clinical decision making. Under the segmentation approaches they have comprehensively covered nearly all the research that has happened so far in the automated segmentation of lung regions and lesion regions in CT and Xray images.

A variant of inception network was proposed by Wang [16] for classifying COVID-19 from healthy. U-Net++ architecture [17] has been effectively put to use for COVID-19 diagnosis, which worked better than expert radiologists. In the realm of segmentation based methods, [18] and [19] use VB-net [20] to segment of lung and infection regions in CT images. Chaganti et. al [10] use DenseUNet to 3D segment the lung and the lesions.

Fan et al. [21] have reported in their paper the list of public COVID-19 imaging datasets. As mentioned in their paper, there is only one dataset which provides segmentation labels [22]. From this public database [22], we have combined the first dataset of 100 sparsely selected axial CT slices from over 40 patients with a dense set of slices from 9 patients CT scan

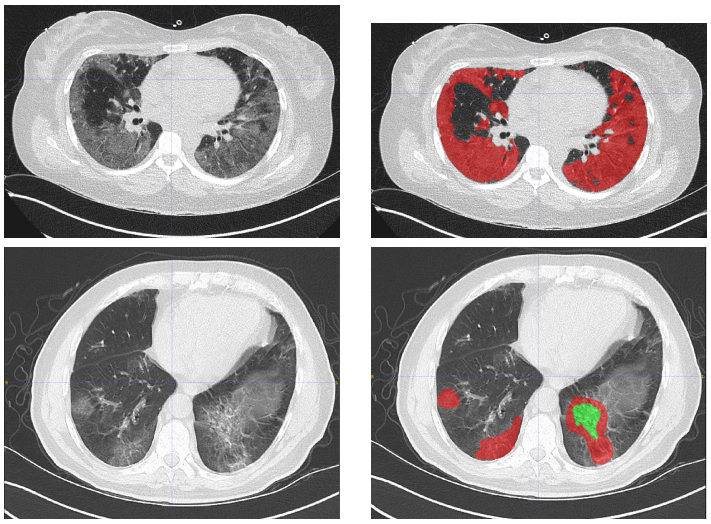


Fig. 2. Sample slice from one of the dataset and the corresponding Ground-glass opacity lesion (GGO) marking in first row and GGO and Consolidation lesion marking in second row. Dataset from website [22]

and use this larger datasets for our studies. A few exemplary slices are demonstrated below to get a visual impression of how the Ground-glass opacity lesions and Consolidation lesions manifests itself in Figure 2.

In this paper, we propose Dynamic Deformable Attention Network (DDANet), a novel deep network for COVID-19 infection segmentation in 2D CT slices. Our inspiration for this network is the recent success of self attention mechanisms and sparse deformable convolutions [23]. Attention blocks do not have to be regularly structured, this opens the novel research area that motivates our investigation of spatially-adaptive attention filters. In this work we generalize criss-cross attention [24] for semantic segmentation tasks. We enhance the criss-cross attention and propose a novel deformable attention in which both the attention filter offsets and coefficients are learnt in a continuous, differentiable space. We carried out extensive experiments of our novel algorithm on a large publicly available COVID-19 dataset. Our proposed DDANet achieves very good lesion segmentation and outperforms most cutting-edge segmentation models reported so far on Ground-glass opacity and consolidation lesions. The proposed solution greatly enhances the performance of the baseline U-Net architecture [25]. The baseline U-Net we have employed in our work is from Oktay et al. [25], which has a well-proven strong baseline. Our novel adaptation of the criss-cross attention module is generic and can also be easily plugged into any state-of-art segmentation architecture. These results demonstrate that our proposed DDANet can be effectively used in image segmentation in general and COVID-19 automated image analysis in particular and can greatly aid in clinical workflow handling of these images.

In summary, our main contributions in our work are:

- We propose a novel deformable attention module in which sparse attention filter offsets are learnt in a continous differentiable space and can capture contextual information in an efficient way
- We demonstrate that employing this new deformable

attention mechanism within the U-Net architecture [25] [1] achieves superior performance of lung infection segmentation compared to conventional U-Nets or U-Nets with criss-cross attention [24]

- The DDANet reaches state-of-the-art segmentation performance of 73.4% and 61.3% for Ground-Glass-Opacity and Consolidation lesions, on a large publicly available CT COVID-19 infection dataset in a three-fold cross validation on GGO and consolidation labels.

II. Related work

We discuss two areas of research that are related to our work - Semantic segmentation and Attention mechanism, specifically the criss-cross attention module.

A. Semantic Segmentation

Semantic segmentation has steadily progressed in the last few years evolving from Fully Convolutional Network (FCN) [26], to the use of dilated convolutions [27] and extensive adaptation of encoder decoder architectures - U-Net [28], Attention U-Net [25] [1], nnU-Net [29], DeepLabv3+ [30], Semantic Prediction Guidance (SPGNet) [31], Discriminative Feature Network (DFN) [32], RefineNet [33] and Multi-Scale Context Intertwining (MSCI) [34]. To detect objects of various scale, the convolution operator has been enhanced using Deformable Convolution [35] [1] and Scale adaptive convolutions [36]. Graphical models have also been employed effectively for the task of semantic segmentation [37] [27].

Attention models initially gathered a lot of traction after the successful introduction of transformer models in Natural Language Processing (NLP) domain [38]. It has been demonstrated that NLP models perform better when the encoder and decoder are connected through attention blocks.

Attention mechanism have subsequently been utilized in computer vision tasks to capture long-range dependencies. The earlier approaches have tried to augment convolutional models with content-based interactions [30] [39] [1]. The seminal work in attention mechanisms was non-local means [40], which was then followed by self-attention [39]. These have helped achieve better performance on computer vision tasks like image classification and semantic segmentation. Attention-gates have also shown promising results when incorporated into U-Nets for 3D medical segmentation [1]. There have also been successful experiments of building pure self-attention vision models [41].

Non-Local Networks [40] enable full-image context information by utilizing self-attention which helps reference features from any position to perceive the features of all other positions. The drawback of Non-Local network is the large time and space complexity ($\mathcal{O}(H \times W) \times (H \times W)$)) to measure every pixel-pair relation, and also requiring large GPU memory to train such models.

CCNet [24] elegantly solves the complexity issue by using consecutive sparse attention. With two criss-cross attention modules, CCNet captures contextual information from all pixels with far less time and space complexity.

B. Criss-Cross Attention Module

The criss-cross attention module (CCA) proposed by Huang et al. [24] aggregates contextual information in horizontal and vertical directions for each pixel. The input image $\mathbf{X}$ is passed through convolutional neural network (CNN) to generate the feature maps $\mathbf{H}$ of reduced dimension. The CCA module comprises of three convolutional layers applied on $\mathbf{H} \in \mathbb{R}^{C \times H \times W}$ with 1×1 as kernel size.

First, the local representation feature maps $\mathbf{H}$ are fed into two convolutional layers in order to obtain two feature maps - query $\mathbf{Q}$ and key $\mathbf{K}$ with the same reduced number of feature channels C'. By extracting feature vectors at each position u from $\mathbf{Q}$, a vector $\mathbf{Q_u} \in \mathbb{R}^{C'}$ is generated. From $\mathbf{K}$ feature vectors in the same row and column as u are collected in $\mathbf{\Omega_u} \in \mathbb{R}^{(H+W-1) \times C'}$ with elements $\mathbf{\Omega_{i,u}} \in \mathbb{R}^{C'}$.

Attention maps $\mathbf{A} \in \mathbb{R}^{(H+W-1) \times H \times W}$ are obtained by applying the affinity operation $d_{i,u} = \mathbf{Q_u} {\mathbf{\Omega_u}}^T$ with $d_{i,u} \in \mathbf{D}$ being the degree of correlation between feature $\mathbf{Q_u}$ and $\mathbf{\Omega_{i,u}}$, $i = [1, ..., |\mathbf{\Omega_u}|]$, $\mathbf{D} \in \mathbb{R}^{(H+W-1) \times H \times W}$ followed by a softmax layer on $\mathbf{D}$ over the channel dimension.

The third convolutional layer applied on $\mathbf{H}$ generates Value $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ for feature adaption. Therefore, a feature vector $\mathbf{V_u} \in \mathbb{R}^{C}$ and a set $\mathbf{\Phi_u} \in \mathbb{R}^{(H+W-1) \times C}$ are extracted at each position u in the spatial dimension of $\mathbf{V}$.

The contextual information is aggregated by

$$\mathbf{H'_u} = \sum_{\mathbf{i} \in |\mathbf{\Phi_u}|} \mathbf{A_{i,u}} \mathbf{\Phi_{i,u}} + \mathbf{H_u} \tag{1}$$

with $\mathbf{H'_u}$ being a feature vector in the module's output feature maps $\mathbf{H'} \in \mathbb{R}^{C \times H \times W}$ at position u and $\mathbf{A_i, u}$ being a scalar value at channel i and position u in $\mathbf{A}$. Finally, the contextual information is weighted with a learnable scalar γ and added to the feature map $\mathbf{H}$.

CCNet [24] was shown to enable improvements in computer vision semantic segmentation tasks on Cityscapes, ADE20K datasets. Tang et al. [42] have successfully employed criss cross attention in medical organ segmentation (lung segmentation). In their XLSor paper [42] they used a pretrained ResNet101 replacing the last two down-sampling layers with dilated convolution operation.

The aim of our work is two-fold. First, we evaluate whether criss-cross attention can be employed within a U-Net [1] to improve medical image lesion segmentation for labelled data which is relatively small, a common scenario currently for COVID-19. Second, we incorporate our novel adaptation of this attention model and extend it with a dynamic deformable attention mechanism where the attention filter offsets are learnt in a continuous differentiable space. We strongly believe that the deformable attention module that automatically adapt their layout is an important step to get better insight into the computation mechanism of attention modules. We have discovered in our work that capturing attention from all non-local locations does negatively impact the accuracy of semantic segmentation networks. Capturing only the necessary and essential non-local contextual information in a smart and data driven way yields far more promising segmentation results. We also demonstrate that having the attention offsets learnable enables the network

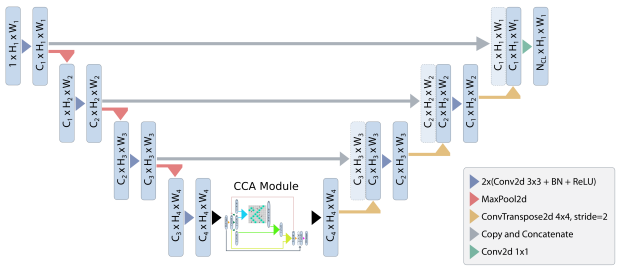


Fig. 3. A block diagram of the proposed Deformable Attention Net (DDANet). Input image is progressively filtered and downsampled by factor 2 at each scale in the encoding part. The deformable criss-cross attention is inserted as an extension of the U-Net's bottleneck in order to capture contextual information from only the necessary and meaningful non-local contextual information in smart and efficient way.

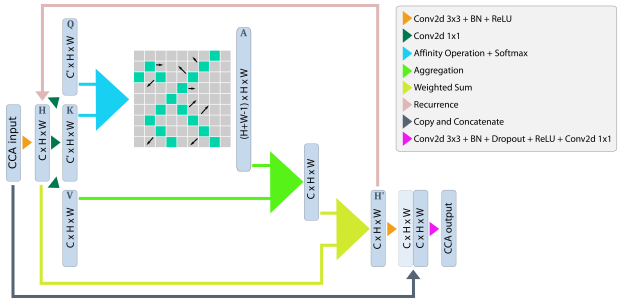


Fig. 4. A block diagram of the proposed deformable criss-cross attention module. In our deformable criss-cross, we have the $\mathbf{H}+\mathbf{W}-\mathbf{1}$ learnable attention offset parameters for each of the criss-cross locations. Differentiable bilinear interpolation is used to sample the attention values for the query, key and value feature maps from the learnt positions of deformed criss-cross offset locations.

to smartly decide on its own the locations where to obtain non-local attention from for improved results.

III. Methods

In this section we explain the details of the proposed network architectures. The basic idea of all variants presented is that an attention module is integrated within the U-Net architecture [1] as an extension of the U-Net's bottleneck in order to capture contextual information from only the necessary and meaningful non-local contextual information in smart and efficient way. Our models utilize the approach of the criss-cross attention module proposed by [24] and modify it to enhance the segmentation performance on COVID-19 datasets.

A. Network Architecture

The architecture of our model combines the concepts of U-Net [28] and CCNet [24]. A block diagram of the proposed Deformable Attention Net (DDANet) is shown in Figure 3.

We use a U-Net structure from Oktay et al. [25] [1], adapting it slightly by reducing one downsampling (and corresponding upsampling path), to best process our image dimension (256*256). It consists of three blocks in the downsampling path and three blocks in the upsampling block. Each block consists of $2\times$(Batch Normalization - 2D Convolution (kernel size 3×3, stride 1, padding 1) - ReLU). The last block consists of a 2D convolution with kernel size 1×1. For downsampling, max pooling is applied in the downsampling path to halve the spatial dimension of the feature maps after each block. In the upsampling path ConvTranspose2d is used to double the size of the spatial dimension of the concatenated feature maps. The number of feature channels is increased $1-64-128-256-512$ in the downsampling path and decreased again accordingly in the upsampling path. The U-Net's last layer outputs a number of feature channels matching the number of label classes for semantic segmentation.

The local representation feature maps $\mathbf{H}$ being output from the U-Net's last block within the downsampling path serve as input of reduced dimension to the criss-cross module. The attention module is inserted in the bottleneck, as the feature maps are of reduced dimension, and hence the attention maps have smaller, more manageable time and space complexity. In the orginial CCNet [24], the following attention module gathers contextual information in the criss-cross path of each pixel leading to feature maps $\mathbf{H}'$. In our proposed novel DDANet, the pattern is dynamic and learnable, and hence a dynamic deformable criss-cross path is used to obtain the attention feature maps $\mathbf{DH}'$. These feature maps are again passed through the dynamic deformable attention module again which results in feature maps $\mathbf{DH}''$ capturing attention information from the most relavant locations from the whole-image at each of its positions. The contextual features $\mathbf{DH}''$ obtained after passing $R = 2$ loops through the attention module are concatenated with the feature maps $\mathbf{X}$ and merged by a convolutional layer. The resulting feature maps are then passed through the U-Net's upsampling path.

We implement the following modifications of the criss-cross attention module: Deformable CCA module with $R = 2$ loops, $\mathbf{X} + \gamma\mathbf{DH}''$

Differentiable attention sampling : Consider a classical criss-cross attention operation which gathers non-local information on a feature map of Height H and width W. The initial shape of the criss-cross pattern is a cross as the orginial CCNet [24] which aggregates contextual information for each pixel in its criss-cross path. We have realized the baseline criss-cross attention by first initializing statically defined locations in a 2D flow field (sampling grid), of size $H * W$. The attention filter offsets for the vertical direction is defined as the locations where the x coordinates matches a tensor of length H equally spaced points between -1 and 1. Similarly, the attention filter offsets for the horizontal direction is defined as the locations where the y coordinates match tensor of length W equally spaced points between -1 and 1. These vertical and horizontal offsets help to compute the attention along a cross pattern at $\mathbf{H}+\mathbf{W}$ non-local locations.

To make the attention map differentiable, we compute displacement for the horizontal and vertical offsets. For computing the displacement for each of the horizontal and vertical locations we use H+W random locations sampled from a standard normal distribution. We distribute these displacement locations smoothly by convolving them three times with

gaussian kernel with a kernel size 5. We then use spatial transformer network to sample the attention values from the offset locations coupled with the displacements. To obtain the attention output for inputs on a discrete grid, we use differentiable bilinear interpolation. This makes our attention sampling differentiable and the attention locations are dynamic and deformable.

We realized our dynamic deformable attention mechanism by the differentiable attention sampling described above which deforms the criss-cross pattern. In our deformable attention implementation, we have included $\mathbf{H} + \mathbf{W}$ learnable attention offset parameters in our deep neural network definition. These are the learnt displacements for each of the criss-cross locations. The learnt displacement vector ($\mathbf{x}$ and $\mathbf{y}$ displacement) for each of the criss-cross locations is used to displace the horizontal and veritcal offsets, while sampling the attention maps. For the second recurrence, a second set of different $\mathbf{H} + \mathbf{W}$ learnable attention parameters is used for determining the displacements.

We use differentiable bilinear interpolation to differentiably sample the attention values for the query, key and value feature maps from the deformed and dynamically learnt positions of criss-cross offset locations. Hence the attention filter offsets for each of the original criss-cross pattern are learnt in continuous differentiable space. The proposed deformable criss-cross attention is depicted in the CCA-Module in Figure 4. As depicted in the figure, the criss-cross pattern is learnt and dynamically deformed to best capture the most relevant non-local information.

The infection class in COVID-19 data is generally under represented as compared to the background class especially in early stages of the disease. This leads to a large class imbalance problem. As found in several studies, Ground-glass opacities generally precede consolidations lesions. This progression of the lesion development in COVID-19 leads to the another scenario of class-imbalance. In some patients only one of the lesions is largely present and the second lesion is highly under-represented (less than 10% of the total infection labels. This also leads to a second category of class-imbalance. To address all of these class-imbalance issues, especially present in COVID-19 lesion segmentation scenarios, we propose to use the inverse class-weighted cross-entropy loss. The weights are computed to be inversely proportional to the square root of class frequency. Given a sample with class label y, this inverse class-weighted cross-entropy loss can be expressed as

$$\mathrm{CE}(\boldsymbol{z}, y) = w_y \left(-\log \left(\frac{\exp(z_y)}{\sum_{j=1}^{C} \exp(z_j)} \right) \right) \tag{2}$$

with C being the total number of classes and $\boldsymbol{z}$ the output from the model for all classes. The weighting factor

$$w_y = \frac{\sqrt{\frac{1}{z_y}}}{\frac{1}{C} \sum_{j=1}^{C} \sqrt{\frac{1}{z_j}}} \tag{3}$$

is determined with help of the inverse square root of the number of samples in each label class to address the problem of training from imbalanced data. The training and validation sets also have different distributions, hence we have computed the inverse weighting separately for the train and validation sets. We have also used learning rate finder [43] to find the optimal learning rate, and a 1cycle learning rate policy scheduler, where the maximum learning rate was also determined using the learning rate finder.

IV. Experimental Setup and Results

We have used the publically available COVID-19 CT segmentation dataset [22]. We have taken the 100 axial CT images from different COVID-19 patients. This first collection of data is from the Italian Society of Medical and Interventional Radiology. We have also utilized the second dataset of axial volumetric CTs of nine patients from Radiopaedia. This second dataset with whole volumes having both positive (373 positive) and negative slices (455 negative slices). We perform experiments with a 3-fold cross validation on this combined dataset consisting of 471 two-dimensional axial lung CT images with segmentations for ground glass opacities (GGO) and consolidation lesions. Each fold comprises data acquired from three different patients plus one third of images from the 100 slice CT stack taken from more than 40 different patients. The CT images are cropped and rescaled to a size of 256×256. During training, we perform random affine deformations for data augmentation.

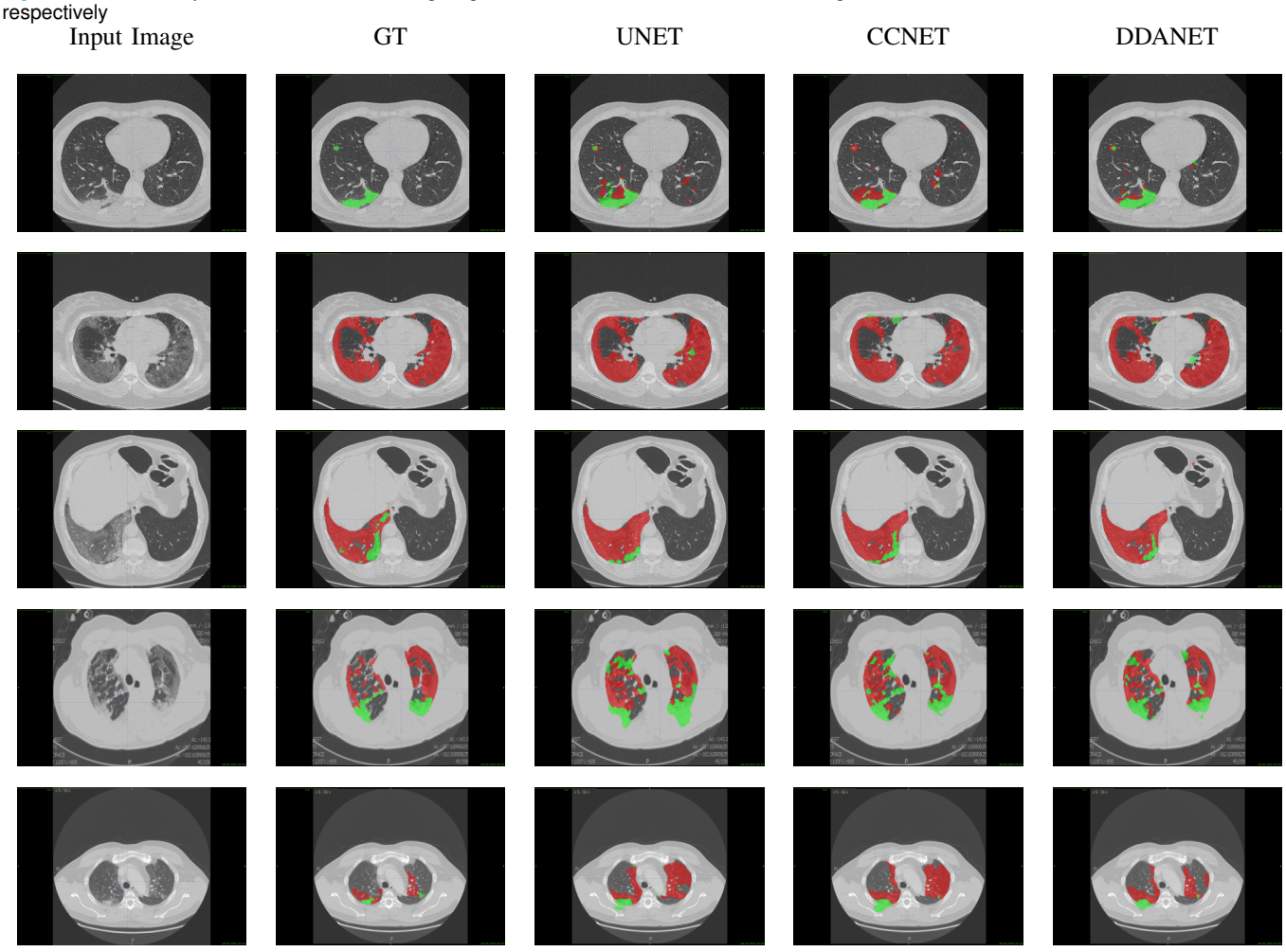
Training is performed for 500 epochs using the Adam optimizer and an initial learning rate of 0.002. We further use a cyclic learning rate with an upper boundary of 0.005 and a class-weighted cross-entropy loss to address the problem of training from imbalanced data.

For the infection region experiments and multi-class labeling we compared our model with two cutting-edge models: U-Net [25] and Criss-Cross Attention [24]. The number of trainable parameter for the U-Net [25] is 611K. For the U-Net incorporated with the criss cross attention the parameter count is 847K. Our proposed variant of modified CCNet has slightly more parameters at 849K. We have used four widely adopted metrics, *i.e.*, Dice similarity coefficient, Sensitivity (Sen.), Specificity (Spec.) and Mean Absolute Error *(MAE)*. If we denote the final prediction as $\mathbf{F_p}$ and the object-level segmentation ground-truth as $\mathbf{G}$, then the Mean Absolute Error which measures the pixel-wise error between final prediction and ground truth is defined as

$$\mathrm{MAE} = \frac{1}{w \times h} \sum_{x}^{w} \sum_{y}^{h} \mid F_p(x, y) - G(x, y)) \mid \tag{4}$$

We have adopted a similar approach to Fan et al. [21] and present first the results of our proposed DDANet on detecting lung infections. Our network is trained on multi-class lung infection (GGO and consolidation) and during evaluation we combine these multiple classes into one infection label. We present our 3-fold cross-validation studies results in Table **I**, which is averaged over multiple runs that we have conducted. We have also included the results from Fan et al. [21] in each of our experiments. It has to be noted that Inf-Net

Fig. 5. Visual comparison of multi-class lung segmentation results, where the red and green labels indicate the GGO and Consolidation, respectively



was only trained with the first dataset which is smaller (100 axial slices) and Semi-Inf-Net was trained with pseudo labels from unlabelled CT images. As captured in the Table I, our proposed DDANet achieves the best Dice scores in each of the folds. The best Dice score obtained is **0.814** and least mean absolute error (MAE) is **0.0185**. We have also captured the average infection segmentation performance of our network in the same Table I. Our proposed DDANet has the best infection segmetation performance in average with the average Dice score of **0.791**). In terms of Dice, our proposed DDANet out-performs the cutting-edge U-Net model [25] by **1.91**% on average infection segmentation.

We have also include the infection segmentation performance of our DDANet on each of the Patients in the supplementary materials. In each of the patients, our proposed DDANet had the best Dice score and the minimum MAE. The average across all the patients is captured in Table II. In terms of Dice, our DDANet method achieves the best competitive performance of **0.7789** averaged across all the patients. It outperforms the baseline best U-Net model Dice by **3.658**% on infection segmentation.

We have included the fold-wise performance of our DDANet on multi-class labeling in the supplement section. We have captured the average multi-label segmentation performance of our network in Table III. We have also compared our results with the results from Inf-Net by Fan et al. [21]. Our baseline U-Net [25] and proposed DDANet has far less trainable parameters at (**611K**) and (**849K**) as compared to **33M** in Inf-Net [21]. Our proposed DDANet has the best multi-label segmetation performance also in average with the best Dice score of **0.734**) for GGO lesions and best Dice score of **0.613**) for Consolidation lesions. Our proposed DDANet has average best dice score of **0.673** for detecting COVID-19 lesions. In terms of Dice, our proposed DDANet out-performs the cutting-edge U-Net model [25] by **4.90**% on average multi-label segmentation. We have increased the trainable parameters in our proposed DDANet only by a neglible amount of 2450 (or 0.3%) in comparison to the original model with criss-cross attention.

We have also captured the multi-label segmentation performance of our DDANet on each of the Patients in the supplementary materials. In terms of Dice, our DDANet method achieves the best competitive performance of **0.702** for GGO lesion and **0.681** for Consolidation lesion averaged across all the patients. In average the proposed DDANet outperforms the baseline best U-Net model Dice by **2.86**% on GGO ,

TABLE I
PERFORMANCE (AVERAGED) OF INFECTION REGIONS ON COVID-19 DATASETS. WE HAVE SPLIT OUR DATA INTO THREE FOLDS, AND THE RESULTS HERE ARE AVERAGED OVER MULTIPLE RUNS FOR EACH FOLD. THESE ARE QUANTITATIVE RESULTS OF INFECTION REGIONS COMPUTED FOLD-WISE AND WE REPORT 3D DICE-SCORES

Model	Fold	Dice	Sen.	Spec.	MAE
Inf-Net [21]		0.682	0.692	0.943	0.082
Semi-Inf-Net [21]		0.739	0.725	0.960	0.064
UNET	0.800	0.776	0.879	0.989	0.0208
	0.787		0.887	0.985	0.0274
	0.740		0.823	0.984	0.0331
+CCA	0.809	0.781	0.876	0.990	0.0192
	0.798		0.888	0.986	0.0258
	0.735		0.850	0.981	0.0357
DDANet	**0.814**	**0.791**	0.889	0.989	**0.0185**
	0.808		0.872	0.988	**0.0240**
	0.750		0.825	0.985	**0.0318**

TABLE II
PERFORMANCE (AVERAGED) ON NINE REAL CT PATIENT DATA. THESE ARE QUANTITATIVE RESULTS OF INFECTION REGIONS COMPUTED PATIENT-WISE AND WE REPORT 3D DICE-SCORES. THE BEST RESULTS ARE SHOWN IN BLUE FONT AND THE GAIN WITH RESPECT TO BASELINE UNET IS SHOWN IN GREEN.

Model	Dice	Sen.	Spec.	MAE	% Gain
Inf-Net [21]	0.579	0.87	0.974	0.047	
Semi-Inf-Net [21]	0.597	0.865	0.977	0.033	
UNET	0.7515	0.8811	0.9904	0.0149	
+CCA	0.7633	**0.8934**	0.9908	0.0143	**1.5819**
DDANet	**0.7789**	0.8840	**0.9915**	**0.0135**	**3.658**

4.73% on Consolidation and in average **3.52**% on multi-label segmentation. The distribution of the GGO and Consolidation lesions are not even among the different patient scans. Some patients had predominantly only GGO (Patient-8) while other patients had predominantly Consolidation (Patient-3). This skew in distribution impacts the segmentation dice scores significantly, when the lesions are minimally represented in the patients.

V. DISCUSSION

COVID-19 lesion segmentation is a very challenging problem. One of the major challenge is the regional manifestation of lesions especially in the early stages of the disease, and this can be very hard to get good segmentation in those high class-imbalance scenarios. A similar challenge arises when one of the lesion classes is majorly represented and the other class is highly under-represented which makes it very difficult. This also is a challenging scenario of skewed class-imbalance and gets very hard to get good segmentation in this context as well. The third challenge is the very limited availability of large public datasets, which has been the case until recently. Slowly a number of COVID-19 datasets are made publically available and this scenario could change quite dramatically in the future. This would then enable further research into more compelling algorithms to address this challenging problem.

Our proposed deformable attention is only one of the potential ways to realize learnable attention mechanisms that are smarter elegant and have better performance than earlier proposed criss-cross attention or non-local methods. There are lots of research possibilities to make this even better. There is no requirement or limitation to gather attention from **H+W** locations as we are currently computing. We have currently computed it that way to make it comparable to criss-cross attention. The attention could be gathered from lesser or more locations. One of the next research problems could be to explore what could be the optimal or minimal number of non-local attention that needs to be gathered to get the best results. It would also be interesting to establish theoretical upper and lower bounds for number of locations to get non-local attention and its impact on performance. Our work opens up all these and more possible research directions and can be the trigger for more fundamental work on learnable attention mechanisms.

TABLE III
QUANTITATIVE RESULTS OF GROUND-GLASS OPACITIES AND CONSOLIDATION. THE RESULTS ARE AVERAGED ACROSS MULTIPLE FOLDS AND MULTIPLE RUNS. THE BEST RESULTS ARE SHOWN IN BLUE FONT.

Model	GGO	Consol.	Avg	%Gain	#Params
Semi-Inf-Net+FCN8s	0.646	0.301	0.474		33.1M
Semi-Inf-Net+MC	0.624	0.458	0.541		33.1M
UNet	0.717	0.566	0.641		611.7 K
+CCA	0.723	0.596	0.660	**2.84**	847.3K
DDANet	**0.734**	**0.613**	**0.673**	**4.90**	849.7K

VI. CONCLUSION

In this paper, we have proposed a novel adaptation to the criss-cross attention module with deformable criss-cross attention. This has been incorporated into the U-Net framework (DDANet) to improve the segmentation of lesion regions in COVID-19 CT scans. Our extensive experiments have demonstrated that both adapting the U-Net with a straightforward incorporation of the CCNet module and also extending this CCNet with multilpe recurrent application does not yield substantial improvements in segmentation quality. Our novel solution and smart combination of adapted dynamic deformable spatial attention have shown to be a working combination yielding superior and promising results. This solution has immense potential in better aiding clinicians with state-of-art infection segmentation models. For our future studies, we plan to apply explore its adaptation in ResNet like architectures for 2D and once more labelled 3D scans become available the module can easily be adapted to 3D V-Net architectures. We will make our source-code and trained models publicly available.

REFERENCES

[1] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, pp. 197 – 207, 2019.
[2] worldometers.info, "Covid-19 coronavirus pandemic." [Online]. Available: https://www.worldometers.info/coronavirus/
[3] C. Menni, A. M. Valdes, M. B. Freidin, and et al., "Real-time tracking of self-reported symptoms to predict potential covid-19." in *Nature Medicine*, 2020.

[4] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, "Sensitivity of chest ct for covid-19: Comparison to rt-pcr," *Radiology*, 2020.

[5] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: A report of 1014 cases," *Radiology*, 2020.

[6] H. Y. F. Wong, H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. Chung, E. Y. P. Lee, E. Y. F. Wan, F. N. I. Hung, T. P. W. Lam, M. Kuo, and M.-Y. Ng, "Frequency and distribution of chest radiographic findings in covid-19 positive patients," *Radiology*, 2020.

[7] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, C. K.-M. Hui, K.-y. Yuen, and M. D. Kuo, "Imaging profile of the covid-19 infection: Radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, 2020.

[8] icometrix.com, "https://icometrix.com/resources/the-role-of-imaging-ai-and-ct-in-covid-19."

[9] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, C. K.-M. Hui, K.-y. Yuen, and M. D. Kuo, "Imaging profile of the covid-19 infection: Radiologic findings and literature review," *Radiology Cardiothoracic Imaging*, 2020.

[10] S. Chaganti, A. Balachandran, and et. al., "Quantification of tomographic patterns associated with covid-19 from chest ct," *arxiv*, 2020.

[11] G. Chassagnon, M. Vakalopoulou, and et. al., "Ai-driven ct-based quantification, staging and short-term outcome prediction of covid-19 pneumonia," *medRxiv*, 2020.

[12] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," *IEEE Reviews in Biomedical Engineering*, 2020.

[13] Y. Oh, S. Park, and J. C. Ye, "Deep learning covid-19 features on cxr using limited training data sets," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020.

[14] D. Dong, Z. Tang, S. Wang, H. Hui, L. Gong, Y. Lu, Z. Xue, H. Liao, F. Chen, F. Yang, R. Jin, K. Wang, Z. Liu, J. Wei, W. Mu, H. Zhang, J. Jiang, J. Tian, and H. Li, "The role of imaging in the detection and management of covid-19: a review," *IEEE Reviews in Biomedical Engineering*, pp. 1–1, 2020.

[15] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, L. Yu, and H. Yu, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020.

[16] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, and B. Xu, "A deep learning algorithm using ct images to screen for corona virus disease (covid-19)," 2020.

[17] Z. Zhou, M. M. Rahman 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*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 3–11.

[18] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, J. Liu, and D. Shen, "Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images," 2020.

[19] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. Gao, H. Sui, and D. Shen, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification," 2020.

[20] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, and et al, "Lung infection quantification of covid-19 in ct images with deep learning," 2020.

[21] D. Fan, T. Zhou, G. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[22] MedicalSegmentation.com, "Covid-19 ct segmentation dataset." [Online]. Available: http://medicalsegmentation.com/covid19/

[23] M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical Image Analysis*, 2019.

[24] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018. [Online]. Available: http://arxiv.org/abs/1804.03999

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR)*, 2016.

[28] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[29] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "Abstract: nnu-net: Self-adapting framework for u-net-based medical image segmentation," in *Bildverarbeitung für die Medizin 2019*, H. Handels, T. M. Deserno, A. Maier, K. H. Maier-Hein, C. Palm, and T. Tolxdorff, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2019, pp. 22–22.

[30] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," in *Advances in Neural Information Processing Systems 31*, 2018.

[31] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi, "Spgnet: Semantic prediction guidance for scene parsing," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5168–5177. [Online]. Available: https://doi.org/10.1109/CVPR.2017.549

[34] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[36] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[37] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2015.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[39] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[40] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[41] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing Systems 32*, 2019.

[42] Y. Tang, Y. Tang, J. Xiao, and R. Summers, "Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation," 04 2019.

[43] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

Supplementary Materials: Dynamic deformable attention (DDANet) for semantic segmentation

Kumar T. Rajamani, Hanna Siebert, and Mattias P. Heinrich,

I. Introduction

The Supplementary Material section provides the following further details of the main paper topics:

1) Enlarged depiction of the block diagram of our proposed Deformable Attention Net (DDANet) is shown in Figure 1;
2) Enlarged depiction of the block diagram of the proposed deformable criss-cross attention module is shown in Figure 2;
3) One Exemplary Slice Showing Difference between UNet+CCNet (state-of-art) and Proposed DDANet.
4) The infection segmentation performance of our DDANet on each of the Patients in Table **I**;
5) The multi-label segmentation (GGO and Consolidation) performance of our DDANet on each of the Patients through 3D dice scores in Table **II**;
6) The multi-label segmentation (GGO and consolidation), on each of the folds through 3D dice scores in Table **III**. The results are shown across three folds.

We first present the expanded view of the block diagram of the proposed Deformable Attention Net (DDANet) is shown in Figure **1**. The input image is progressively filtered by two consecutive convolution blocks. The number of activation maps or feature channels is increased in the second convolution block. The number of feature channels is progressively increased $1 - 64 - 128 - 256 - 512$ in the downsampling path. This double convolution block is then followed by maxpooling lyer. The maxpool layer downsamples the activation maps by factor 2 at each scale in the encoding part. The deformable criss-cross attention is inserted as an extension of the U-Net's bottleneck in order to capture contextual information from only the necessary and meaningful non-local contextual information in smart and efficient way.

In the upsampling path ConvTranspose2d is used to double the size of the spatial dimension of the concatenated feature maps. The number of feature channels is decreased $512 - 256 - 128 - 64 - N_{CL}$ in the upsampling path. The U-Net's last layer outputs a number of feature channels matching the number of label classes for semantic segmentation.

We next present the expanded view of the block diagram of the proposed deformable criss-cross attention module in Figure **2**. In our deformable criss-cross, we have the $\mathbf{H} + \mathbf{W} - \mathbf{1}$ learnable attention offset parameters for each of the criss-cross locations. Differentiable bilinear interpolation is used to sample the attention values for the query, key and value feature maps from the learnt positions of deformed criss-cross offset locations.

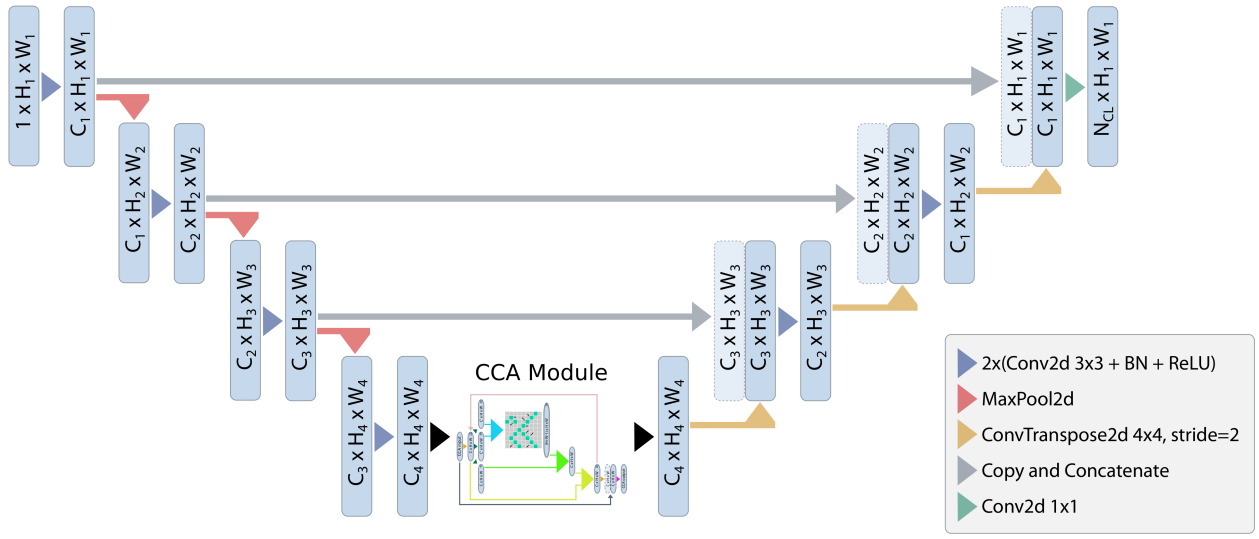


Fig. 1. A block diagram of the proposed Deformable Attention Net (DDANet). Input image is progressively filtered and downsampled by factor 2 at each scale in the encoding part. The deformable criss-cross attention is inserted as an extension of the U-Net's bottleneck in order to capture contextual information from only the necessary and meaningful non-local contextual information in smart and efficient way.

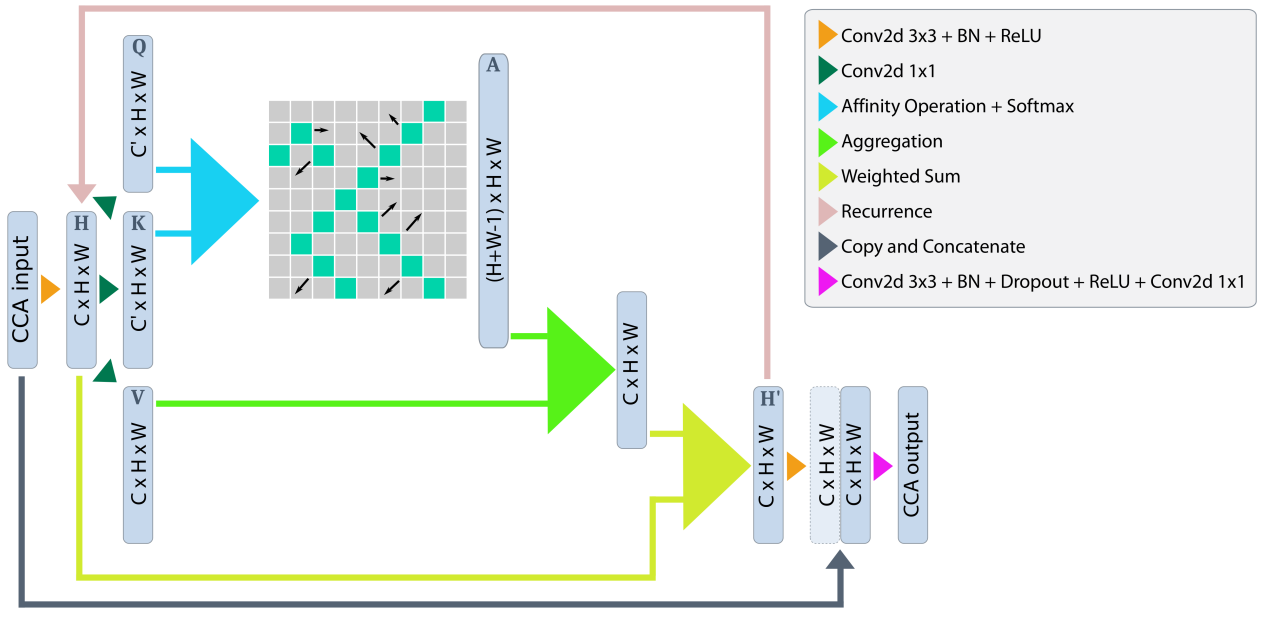


Fig. 2. A block diagram of the proposed deformable criss-cross attention module. In our deformable criss-cross, we have the $\mathbf{H} + \mathbf{W} - \mathbf{1}$ learnable attention offset parameters for each of the criss-cross locations. Differentiable bilinear interpolation is used to sample the attention values for the query, key and value feature maps from the learnt positions of deformed criss-cross offset locations.

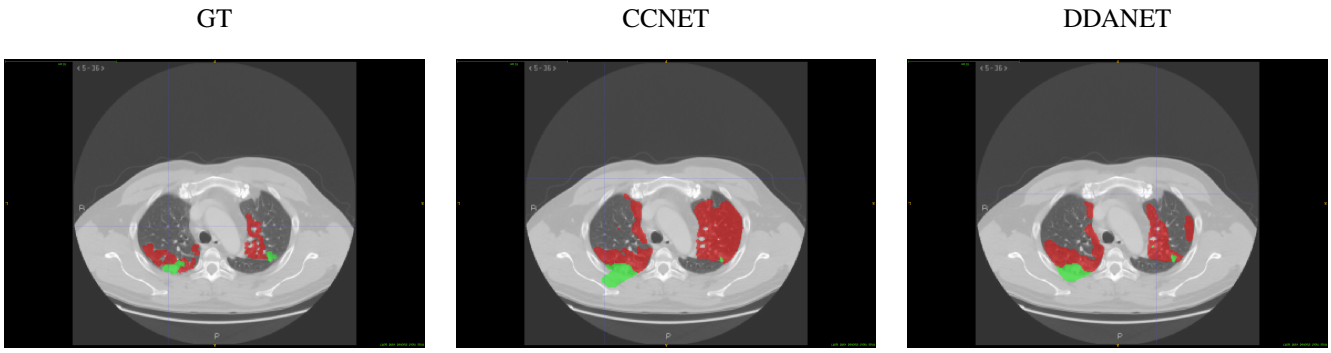


Fig. 3. One exemplary slice showing difference between UNet+CCNet (state-of-art) and Proposed DDANet. As is clearly evident UNet+CCNet segmentations leaks into the background when the contrast between structures is smaller and hence it generate spurious segmentations whereas our proposed DDANet has lesser of such leaky effects and has superior performance.

In the Figure **3**, we demonstrate one exemplary slice showing Difference between UNet+CCNet (state-of-art) and Proposed DDANet. As is clearly evident UNet+CCNet segmentations leaks into the background when the contrast between structures is smaller and hence it generate spurious segmentations whereas our proposed DDANet has lesser of such leaky effects and has superior performance.

TABLE I

PERFORMANCE ON NINE REAL CT PATIENT DATA. THESE ARE QUANTITATIVE RESULTS OF INFECTION REGIONS COMPUTED PATIENT-WISE AND WE REPORT 3D DICE-SCORES

Pat-1	Dice	Sen.	Spec.	MAE.
UNet	0.85	0.92	0.99	0.017
+CCA	0.86	0.90	0.99	0.016
DDAN	**0.87**	0.89	0.99	0.014

Pat-2	Dice	Sen.	Spec.	MAE.
UNet	0.79	0.92	0.99	0.009
+CCA	0.81	0.88	0.99	0.007
DDAN	**0.85**	0.88	0.99	0.006

Pat-3	Dice	Sen.	Spec.	MAE.
UNet	0.63	0.86	0.99	0.016
+CCA	0.71	0.96	0.99	0.013
DDAN	**0.72**	0.97	0.992	0.013

Pat-4	Dice	Sen.	Spec.	MAE.
UNet	0.76	0.87	0.98	0.025
+CCA	0.76	0.91	0.98	0.027
DDAN	**0.76**	0.88	0.98	0.026

Pat-5	Dice	Sen.	Spec.	MAE.
UNet	0.63	0.85	0.99	0.021
+CCA	0.64	0.88	0.98	0.021
DDAN	**0.64**	0.88	0.98	0.021

Pat-7	Dice	Sen.	Spec.	MAE.
UNet	0.84	0.91	0.99	0.002
+CCA	0.79	0.96	0.99	0.003
DDAN	**0.83**	0.94	0.99	0.002

Pat-8	Dice	Sen.	Spec.	MAE.
UNet	0.66	0.80	0.99	0.003
+CCA	0.69	0.78	0.99	0.003
DDAN	**0.70**	0.72	0.99	0.002

Pat-9	Dice	Sen.	Spec.	MAE.
UNet	0.86	0.91	0.98	0.027
+CCA	0.86	0.88	0.99	0.024
DDAN	**0.87**	0.92	0.98	0.022

Avg Pat	Dice	Sen.	Spec.	MAE.
UNet	0.75	0.88	0.99	0.015
+CCA	0.76	0.89	0.99	0.014
DDAN	**0.78**	0.88	0.99	0.014

We have also captured the infection segmentation performance of our DDANet on each of the Patients in Table I. We have skipped using one Patient (Patient-6) from the dataset, as that had only one slice with infection and only 85 voxels of infection marked in that slice against the total 167M voxels. In each of the patients, our proposed DDANet is having the best Dice score and the minimum MAE. In terms of Dice, our DDANet method achieves the best competitive performance of **0.78** and MAE of **0.014** for Infection segmentation averaged across all the patients.

TABLE II

PERFORMANCE ON NINE REAL CT PATIENT DATA. THESE ARE QUANTITATIVE RESULTS OF MULTI-LABEL REGIONS COMPUTED PATIENT-WISE AND WE REPORT 3D DICE-SCORES

Patient-1	GGO	Consolidation
UNet	0.8133	0.5233
UNet+CCA	0.7999	0.5837
DDANet	**0.8248**	**0.5973**

Patient-2	GGO	Consolidation
UNet	0.3225	0.861
UNet+CCA	0.3378	**0.867**
DDANet	**0.396**	0.861

Patient-3	GGO	Consolidation
UNet	NA	0.7663
UNet+CCA	NA	0.7942
DDANet	NA	**0.8053**

Patient-4	GGO	Consolidation
UNet	0.7352	NA
UNet+CCA	0.7359	NA
DDANet	**0.7432**	NA

Patient-5	GGO	Consolidation
UNet	0.5599	0.4517
UNet+CCA	**0.5694**	**0.4696**
DDANet	0.5555	0.4616

Patient-7	GGO	Consolidation
UNet	0.8403	NA
UNet+CCA	0.7973	NA
DDANet	**0.8268**	NA

Patient-8	GGO	Consolidation
UNet	0.6533	NA
UNet+CCA	0.6843	NA
DDANet	**0.695**	NA

Patient-9	GGO	Consolidation
UNet	0.8529	NA
UNet+CCA	0.8613	NA
DDANet	**0.8726**	**NA**

Mean-Ac. Pat.	GGO	Consolidation
UNet	0.683	0.651
UNet+CCA	0.684	0.679
DDANet	**0.702**	**0.681**

We have also captured the multi-label segmentation performance of our DDANet on each of the Patients through 3D dice scores in Table **II**. We have again skipped Patient-6 (due to low lesion representation). The average across all the patients is also captured in the same table in the last block. In six out of the eight patients, our proposed DDANet had the best Dice score for both GGO and Consolidation lesion. In terms of Dice, our DDANet method achieves the best competitive performance of **0.702** for GGO lesion and **0.681** for Consolidation lesion averaged across all the patients.

In average the proposed DDANet outperforms the baseline best UNet model Dice by **2.86**% on GGO , **4.73**% on Consolidation and in average **3.52**% on multi-label segmentation. The distribution of the GGO and Consolidation lesions are not even among the different patient scans. Some patients had predominantly only GGO (Patient-8) while other patients had predominantly Consolidation (Patient-3). This skew in distribution impacts the segmentation dice scores significantly, when the lesions are minimally represented in the patients. We have not taken into consideration those labels in some of the patients when the representation is lower than 10% of the overall lesion distribution as the dice scores gets impacted due to this skewed distribution.

TABLE III

QUANTITATIVE RESULTS OF GROUND-GLASS OPACITIES AND CONSOLIDATION. THE RESULTS ARE SHOWN ACROSS THREE FOLDS AND AVERAGED OVER MULTIPLE RUNS. THE BEST RESULTS ARE SHOWN IN BLUE FONT AND THE GAIN WITH RESPECT TO BASELINE UNET IS SHOWN IN GREEN.

Model	Fold	GGO	%Gain	Consol.	%Gain
Semi-Inf-Net+FCN8s		0.646		0.301	
Semi-Inf-Net+MC		0.624		0.458	
UNET	fold0	0.7687		0.6799	
	fold1	0.7225		0.5699	
	fold2	0.659		0.4485	
+CCA	fold0	0.7809		0.7153	
	fold1	0.7254	0.89	0.6055	5.3
	fold2	0.6631		0.4676	
DDANet	fold0	**0.787**		**0.733**	
	fold1	**0.738**	**2.38**	**0.6085**	**8.22**
	fold2	**0.675**		**0.4967**	

We have capture the performance of our DDANet on multi-class labeling. We present our 3-fold cross-validation studies results in Table **III**, which is averaged over multiple runs that we have conducted. We have also included the results from Fan et al. [1] in each of our experiments. As captured in the Table **III**, our proposed DDANet achieves the best Dice scores in each of the folds. The Best Dice score achieved for GGO is **0.787** and best Dice score for Consolidation is **0.733**. Our proposed DDANet outperforms the cutting-edge UNet model, in terms of Dice, by **2.38**% in GGO lesion and **8.22**% in Consolidation lesion segmentation in average. Our proposed deformable criss-cross attention is able to segment GGO and consolidation lesions far better than the state-of-art models or baseline UNet models.

REFERENCES

[1] D. Fan, T. Zhou, G. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.