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.

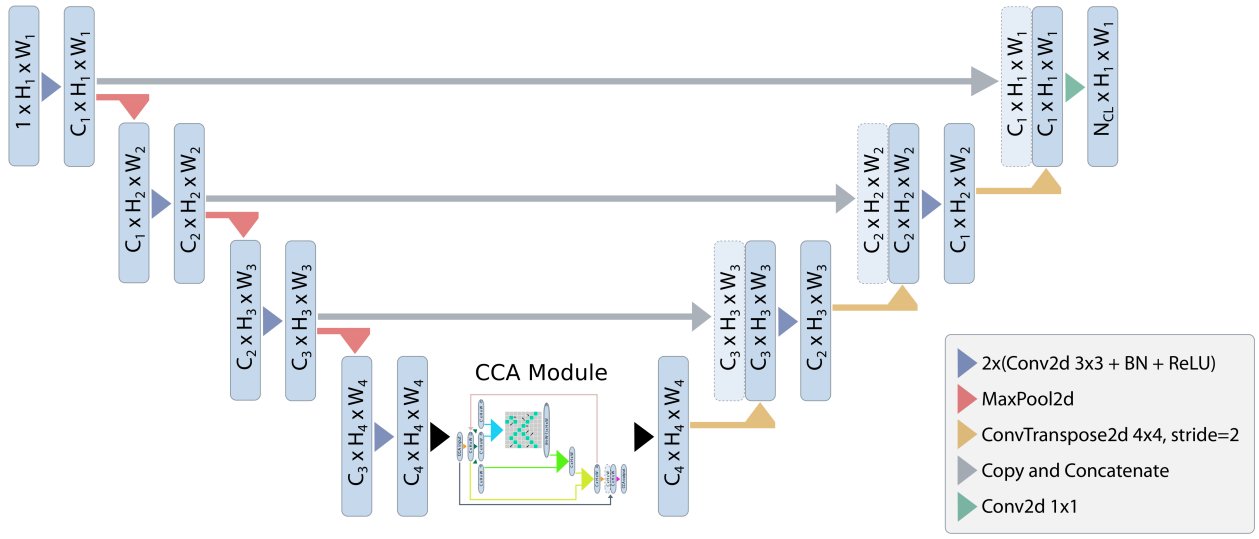

Fig. 1. A block diagram of the proposed Deformable Attention Net (DDANet). Input image is progressively filtered and downsampled by factor 2 at each scale in the encoding part. The deformable criss-cross attention is inserted as an extension of the U-Net's bottleneck in order to capture contextual information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

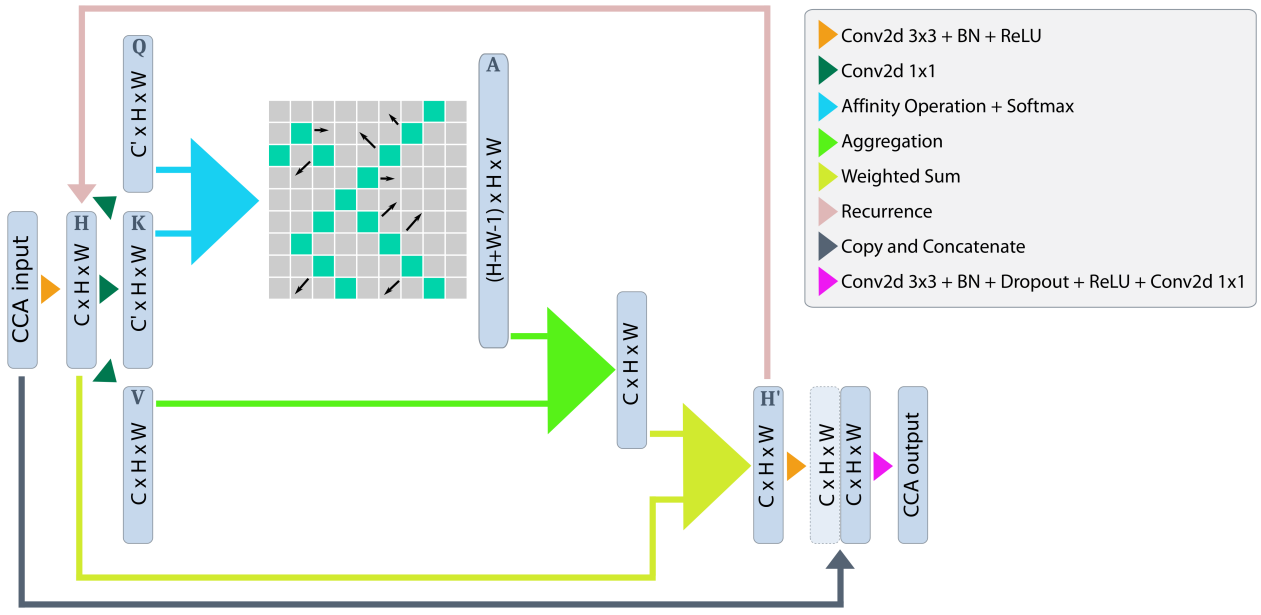

Fig. 2. A block diagram of the proposed deformable criss-cross attention module. In our deformable criss-cross, we have the $\mathbf{H} + \mathbf{W} - \mathbf{1}$ learnable attention offset parameters for each of the criss-cross locations. Differentiable bilinear interpolation is used to sample the attention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

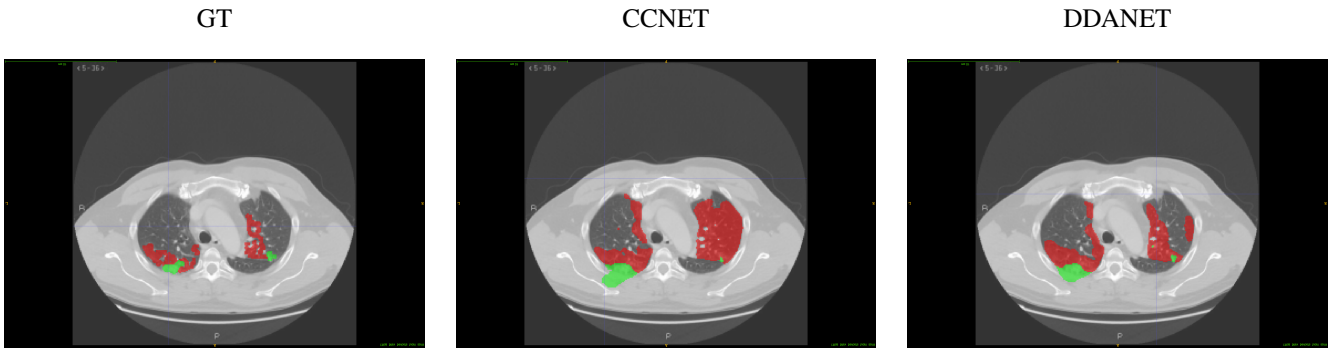

Fig. 3. One exemplary slice showing difference between UNet+CCNet (state-of-art) and Proposed DDANet. As is clearly evident UNet+CCNet segmentations leaks into the background when the contrast between structures is smaller and hence it generate spurious segmentations whereas our proposed DDANet has lesser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