Co-Heterogeneous and Adaptive Segmentation from Multi-Source and Multi-Phase CT Imaging Data: A Study on Pathological Liver and Lesion Segmentation - Supplementary material

Ashwin Raju[1,2], Chi-Tung Cheng[3], Yuankai Huo[1], Jinzheng Cai[1], Junzhou Huang[2], Jing Xiao[4], Le Lu[1], ChienHung Liao[3], and Adam P. Harrison[1]

[1] PAII Inc., Bethesda MD, USA
[2] The University of Texas at Arlington, Arlington TX, USA
[3] Chang Gung Memorial Hospital, Linkou, Taiwan, ROC
[4] PingAn Technology, Shenzhen, China

1 Methodology Figures

Figures 1 and 2 depict the progressive holistically nested network (PHNN) architecture and our holes-based pseudo-labeling, respectively.

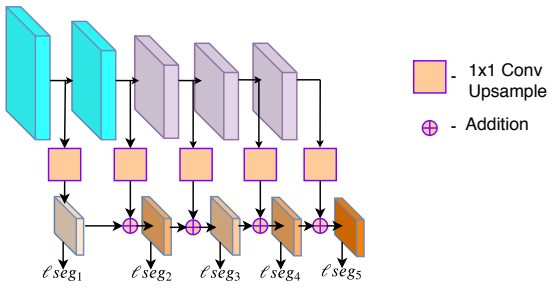

Fig. 1. PHNN architecture. Here we use the V-phase pathway coloring from Figure 3. At each backbone stage, deeply supervised predictions and losses are calculated. Similar to residual-style connections [1], each stage's predictions are built off the prior one's using addition.

2 Implementation details

2.1 Network training

We first initialize co-heterogenous and adaptive segmentation (CHASe) with the weights trained on supervised venous phase data from the public datasets. To train this segmentation network, we use the Adam optimizer [2] with an initial

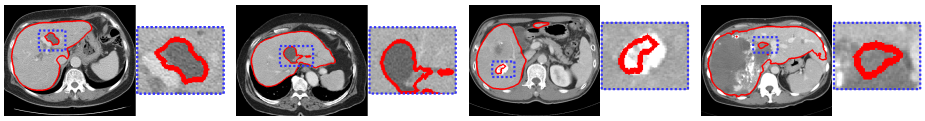

Fig. 2. Hole-Based Pseudo-Labelling. 3D holes greater than 100 voxels are extracted as lesion pseudo-masks missed by the prediction. Regions outside the hole are ignored. The third liver slice from left shows a TACE-treated lesion, which is not seen in public datasets.

learning rate of 3×10^{-4} and values of 0.9 and 0.99 for the β_1 and β_2 hyperparamters, respectively. We reduce learning rateswhen the validation accuracy does not improve for 10 epochs using a factor of 0.1.

To train CHASe, we use the stochastic gradient descent (SGD) optimizer with an initial learning rate of 1×10^{-5} and a momentum of 0.9. We reduce the learning rate when the validation loss does not reduce for 10 epochs using a factor of 0.1.

For training the discriminator, we use the Adam optimizer with an initial learning rate of 3×10^{-4} and reduce the learning rate with a polynomial decay schedule with a power of 0.9 as specified in [3].

We augment the dataset in both source and target domain by performing random rotation, random elastic deformation, gamma correction and random scaling.

3 Additional Results

Table 2 shows the performance of different models on the test dataset using all 15 possible combinations of phases during inference. For H-DenseUNet, Baseline, Co-training, which do not naturally accept multi-channel inputs, we perform majority voting across the appropriate single-phase predictions.

Table 1. Data distribution for $\mathcal{D}_\ell$. Each dataset shows whether it contains only healthy liver or pathological liver and the number of volumes.

Dataset ($\mathcal{D}_\ell$)	**Total**	**Healthy**	**Pathological liver**
LiTS	130		✓
CHAOS	40	✓	
3D-IRCADb	20		✓
Gibson	35		✓
Sliver07	20	✓	

Table 2. Combination of views. Mean DSCs are tabulated across different combinations of contrast phases used for input. The number of samples are indicated in parentheses. ✓ signifies the presence of a phase and ✖ represents the absence of a phase.

Multi-phases (90)				Models					
Non-contrast	Arterial	Venous	Delay	H-DenseUNet	Baseline	Co-training	Co-hetero	Cohetero w ADA	CHASe
✖	✖	✖	✓	85.7	86.4	92.9	93.8	94.0	**94.3**
✖	✖	✓	✖	90.9	90.7	93.7	94.5	94.9	**95.0**
✖	✖	✓	✓	90.5	90.9	93.8	94.7	**94.9**	94.8
✖	✓	✖	✖	90.8	91.1	93.6	94.1	94.3	**94.6**
✖	✓	✖	✓	91.1	91.3	93.1	94.6	94.9	**95.1**
✖	✓	✓	✖	91.9	91.8	92.9	94.8	94.8	**95.0**
✖	✓	✓	✓	91.4	91.6	93.5	95.0	**95.2**	**95.2**
✓	✖	✖	✖	85.6	85.9	92.4	93.5	93.8	**94.0**
✓	✖	✖	✓	90.4	90.7	92.6	93.8	94.0	**94.1**
✓	✖	✓	✖	91.1	91.8	93.4	94.9	95.1	**95.2**
✓	✖	✓	✓	91.2	92.0	94.1	94.8	95.0	**95.4**
✓	✓	✖	✖	90.9	91.6	93.7	94.9	94.8	**95.0**
✓	✓	✖	✓	91.6	91.4	94.3	95.0	95.0	**95.2**
✓	✓	✓	✖	91.5	91.9	94.4	95.0	95.1	**95.3**
✓	✓	✓	✓	91.6	92.1	94.5	95.1	95.4	**95.7**

Table 3. Pathological Liver Segmentation. Mean DSC and ASSD results on the Anonymized PACS dataset are tabulated across different contrast phase inputs. For "All", all available phases in the CT study are used as input. Number of samples are indicated in parentheses. The segmentation model is trained with VGG16 backbone.

Models	NC (96)		A (98)		V (97)		D (98)		All (100)	
	DSC	ASSD	DSC	ASSD	DSC	ASSD	DSC	ASSD	DSC	ASSD
HDenseUNet	85.2	3.25	90.1	2.19	90.7	2.61	85.2	2.91	89.9	2.59
Baseline	85.1	2.81	90.1	1.33	90.2	1.21	86.9	2.03	90.9	1.25
Baseline w pseudo	87.4	1.47	90.3	1.37	90.8	1.13	91.1	1.12	91.7	1.23
Baseline w ADA	88.3	1.38	91.2	1.08	91.1	1.12	92.1	0.99	92.4	1.01
Co-training	91.8	1.03	92.5	1.01	92.9	0.95	92.5	1.02	93.8	0.99
Co-hetero	93.1	0.95	93.3	0.95	94.0	0.80	93.1	1.06	94.6	0.73
Co-hetero w ADA	93.4	0.89	93.6	0.85	**94.3**	0.74	93.6	0.91	94.7	0.73
CHASe	**93.7**	**0.82**	**93.8**	**0.83**	94.2	**0.73**	**93.8**	**0.87**	**95.0**	**0.70**

Table 3 provides the ablation study results when VGG16 is used as backbone. As can be seen, the results exhibit identical trends as when using a ResNet50-based DeepLabv2 backbone, except that absolute numbers are slightly worse. Nonetheless, even with an older backbone CHASe is able to provide excellent results.

Figure 3 depicts a box-and-whisker plot of the lesion DSC scores on the public dataset. As can be seen, all components of CHASe contribute to higher performance. Although the mean scores of CHASe were lower when using the holes-based pseudo-labeling (see main text), the figure demonstrates that the median values are higher, with a tighter spread of quartile values.

Figure 4 depicts additional qualitative results demonstrating the visual improvements provided by CHASe.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
3. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)

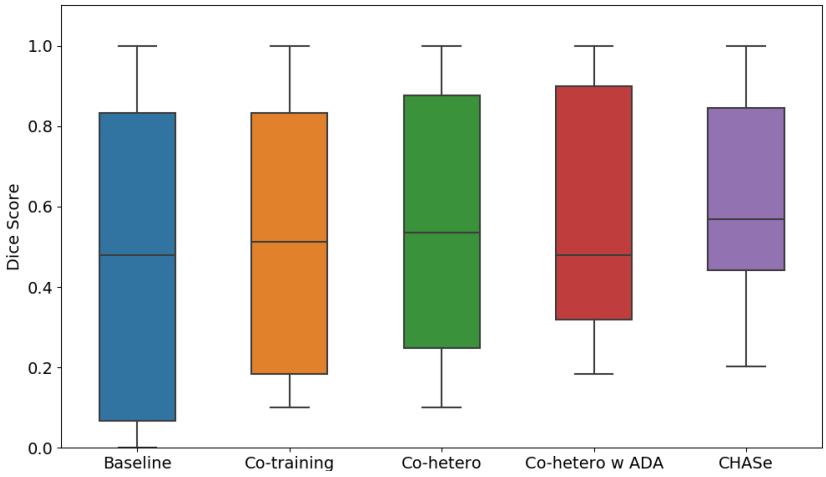

Fig. 3. Box-and-whisker plots of lesion scores on the public dataset. DSCs of 1.0 and near 0.0 are possible, as many studies had no lesions present. If the model did not predict any lesions, it yielded perfect DSCs. Conversely, predictions of any lesion when none are present penalize scores very heavily.

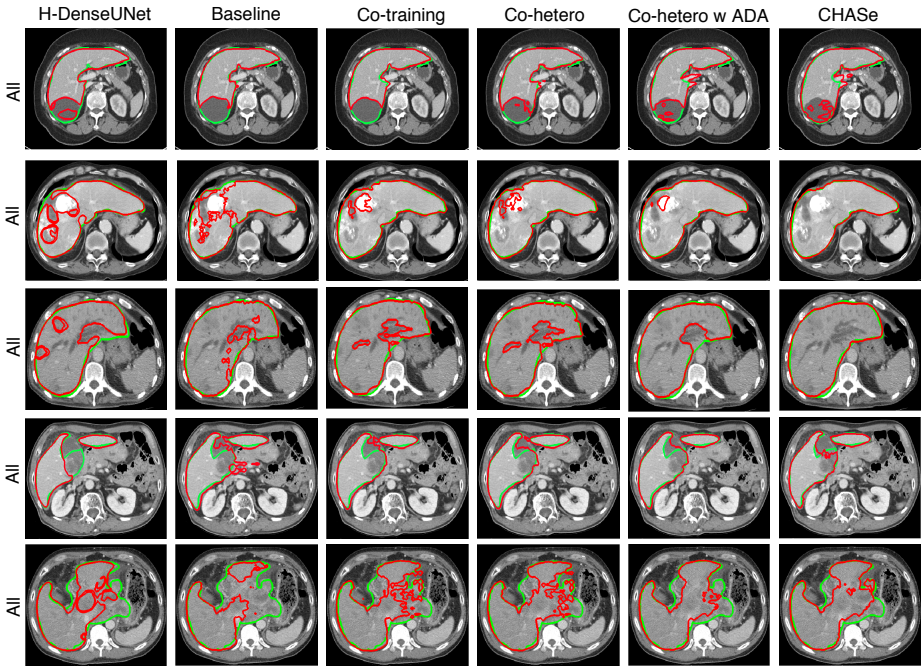

Fig. 4. Qualitative results. Green and **red** curves depict the ground truth and segmentation predictions, respectively. All predictions executed with all phases used as input. The first and last rows depict failure cases, where the latter is an extremely challenging case with an extraordinarily large lesion occupying much of the liver space. CHASe still manages to provide superior results compared to the alternatives. The second row demonstrates CHASe's ability to account for TACE-treated lesions, which are not present in public datasets. The fourth row depicts another highly challenging case, where the gallbladder is difficult to distinguish from a lesion. As can be seen, CHASe is the only model able to successfully differentiate these two structures.