Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation (supplemental material)

Jintao Ren

1. Metrics

The Dice similarity coefficient (Dice) describes the spatial overlapping proportion between prediction P and the ground truth G. Dice is defined as:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

P and G voxel points can be represented by Boolean using true positive (TP), false positive (FP), and false negative (FN). The size of P is TP+FP, the size of G is TP+FN. Thus, Dice could also be defined as:

$$\text{Dice} = 2 \times \frac{1}{1/\text{ recall } + 1/\text{ precision}} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

$$\text{Recall } = \frac{TP}{TP + FP}, Precision = \frac{TP}{TP + FN} \tag{3}$$

~~Where TP, TN, FP and FN indicate true positive, true negative, false positive and false negative voxel points, respectively.~~ Hausdorff distance is defined as:

$$HD(X,Y) = d_H(S_1, S_2) = \max\left\{ \sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y) \right\} \tag{4}$$

Where S_1 and S_2 denote the contour voxel sets of prediction and ground truth. Where d(x,y) indicates the Euclidean distance between voxel x and voxel y from two surfaces. The extreme long-distances were excluded by the 95th percentile of HD (HD95).

2. Data acquisition

CT and PET were obtained by FDG-PET/CT scanner with various resolution spacing. CT has a spatial resolution of $1.17 \times 1.17 \times 3mm^3$ in most cases, whereas PET has a spatial resolution of $2 \times 2 \times 2mm^3$. T1-weighted (in-phase mDIXON) and T2-weighted MRI sequences were used. The resolution spacing for T1 and T2 sequences is $0.69 \times 3 \times 0.69mm^3$ and $0.93 \times 0.93 \times 4mm^3$, respectively. We deformably registered MR images to CT scans using Elastix[1]. We adapted the registration parameters from [2]. We registered the T1 and T2 image simultaneously, using a pyramid scheme with 3 resolutions, and downsampling of 4, 2, and 1. Registration metric was a combination of$AdvancedMattesMutualInformation$ and a $TransformRigidityPenalty$, with weights of 1 and 40. Each resolution consisted of 600 iterations, using 5000 random spatial samples, and a $finalgridspacing$ of 10 mm.

3. Image sampling and data augmentation

The Head and Neck tumor segmentation dataset has a skewed/unbalanced foreground to background class ratio. The median image size is $268 \times 268 \times 226mm = 16,232,224mm^3$, while the median foreground GTVs size is $28,329mm^3$, according to our data. As a result, the class unbalances problem must be considered when choosing image sampling, augmentation, and loss function for deep learning.

The choice of batch size, patch size, and patch numbers must all be taken into account as a chain for image sampling. Before these steps, the z-score was employed to standardize each modality image for each patient, which ensures voxels in each modality has a similar data distribution. To fit the GPU memory while ensuring a smooth gradient descent, we employ a batch size of 2 and a patch size of $128 \times 128 \times 64 \times N$, where N refers to the number of modalities.

We used sliding window patch extraction over region-of-interest(ROI) extraction to avoid additional workflow. Each patient requires multiple patches when using the sliding window method. We first used the slide window to extract patches that just covered all regions from the image and then randomly over-sampled more patches over the GTV regions. We decided on 64 as the total number of patches for each patient since it worked the best among several candidates.

To avoid the over-fitting, we used random rotations from -30 to +30 degrees over the axial axis, axial rotations (90, 180, 270 degrees), and flipping over the axial, sagittal, or coronal axes for data augmentation.

4. Residual 3D UNet

We employed the Residual 3D UNet[3,4] to conduct our segmentation task. Like other state-of-the-art segmentation architecture[5,6], it obeys an encoder-decoder structure. In all convolutional blocks, 3D convolutions with residual connections, ReLU activation functions, and batch normalization are used. We chose three down-sampling and three up-sampling processes based on the size of our image, which helps to lower the parameters and mitigate the over-fitting problem. It has 64 layers of feature maps after the first convolution block, and 512 layers after the bottleneck block. 3D transposed convolution is used for upsampling, and 3D max-pooling is used for downsampling. Element-wise addition is used to create both short and lengthy residual/skip connections. (Figure 1).

According to our findings, 3D Residual UNet outperformed 3D Dense UNet and 3D UNet with our data. Three times dowsampling/upsampling is preferred to four times dowsampling/upsampling because it reduces generalization error and propagation time while maintaining the same level of accuracy.

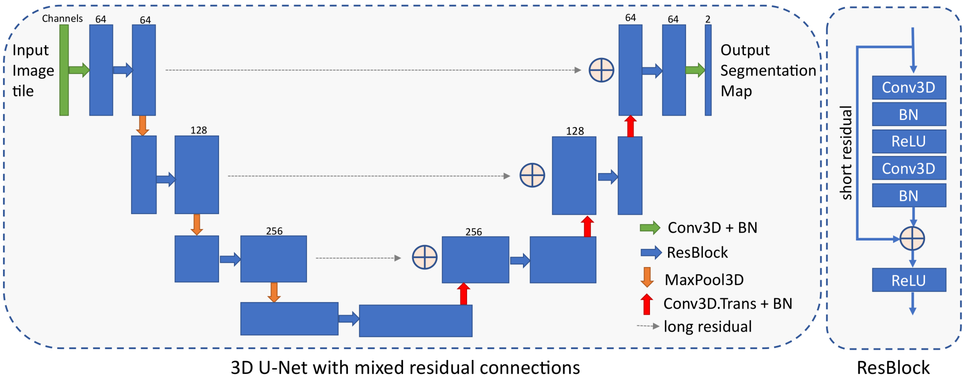

Figure 1. Network architecture plot. For input image tiles, different multi-modality images were consisted in the channel dimension along with three spatial dimensions.

5. Loss function

It is common in medical image segmentation, for the target region to encompass only a small portion of the image. This usually resulting in a network with a high bias towards background predictions[7] . Several attempts have been made to solve this unbalanced class problem previously[7–10]. Among them, Dice loss[7] is the most common one, it alleviates the imbalance problem by directly optimizing Dice Similarity Coefficient using overlaps between prediction and ground truth. Thus, small targets have more weight on the loss calculation. In this task, however, we observed severe oscillation in the loss curve when training while only optimizing Dice loss. The high GTV size variance may have contributed to this unstable situation. The focal loss[8] is a variant of cross entropy(CE), and it skews weights for the hard examples by reducing the loss assigned to well-classified examples.

We employ a hybrid dual loss function as our loss function in this study: $\mathcal{L}_{Focal} + \mathcal{L}_{Dice}$. The hybrid loss function combines the benefits of both loss functions. The focal part skews the weight for challenging samples, while the Dice part is sensitive to small objects[11], the training procedure is also much more stable when combined[12].

The focal loss is computed as:

$$\mathcal{L}_{Focal} = -\frac{1}{N}\sum_{n=1}^{N} y_n \left(1-\hat{y}_n\right)^2 \log\left(\hat{y}_n\right) \tag{5}$$

where $\hat{y} \in [0,1]$ is the predicted output, $y \in 0,1$ is the ground truth mask. N is the total number of voxels in the images. The Dice loss is computed as:

$$\mathcal{L}_{Dice} = 1 - 2\frac{\sum_n \hat{y}_n y_n + \epsilon}{\sum_n \hat{y}_n + \sum_n y_n + \epsilon} \tag{6}$$

The ϵ is used to avoiding the dividing by 0 numerical problem.

Based on validation set Dice Similarity Coefficient on CT-PET-MRI, we experienced performance differences between the choice of loss functions as $Dice + Focal > Dice + CE > Dice > CE$.

6. Train and inference

Tensorflow 1.15 was used to train the models. we choose Adam as the optimizer with a learning rate of $3e-5$, $beta_1 = 0.9$, $beta_2 = 0.999$, and a batch size of 2. The test was carried on on an Nvidia Tesla V100 GPU with 16GB of RAM. The maximum number of training epochs was set to 60. We saved all of the trained model weights after each epoch while training. The best trained weights were chosen with highest average Dice on validation set, which is usually around 40 epochs.

For test inference, we used the sliding window to extract 32 patches with overlap for each image combination. ~~To inference the test set, we used the sliding window method to extract patches. In order to reduce the impact from model bias, we extracted 32 sub-volumes around the image at test time.~~ The mean values of overlapping voxels were obtained to aggregate an output map. We use threshold of 0.5 ~~majority vote~~ to obtain a whole 3D binary GTV prediction afterward. Predicting a GTV segmentation map from CT-PET-MRI took about 60 seconds per patient (size of $268 \times 268 \times 226$), while other combinations used less time.

To ensure consistency, a variety of random initiated training processes were performed to observe the findings and conclusions. We analyzed all of the results for each modality combination and chose one of several ”best-performing” models to report on. The model random led occasional situations did not contribute to the paper conclusion. However, we did not perform data shuffle procedures in a systematic manner for all combinations.

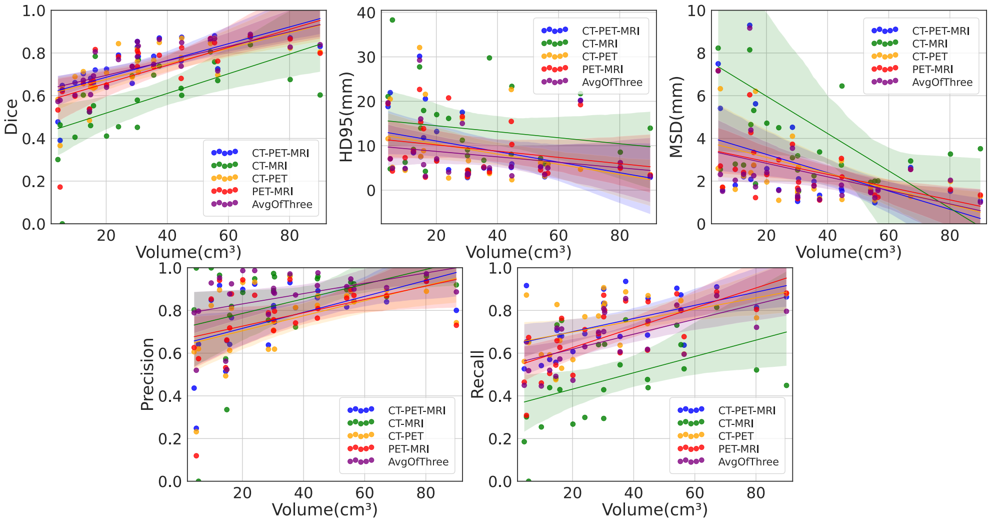

Figure 2. Linear correlations between different metrics and volumes of ground truth GTV for test set.

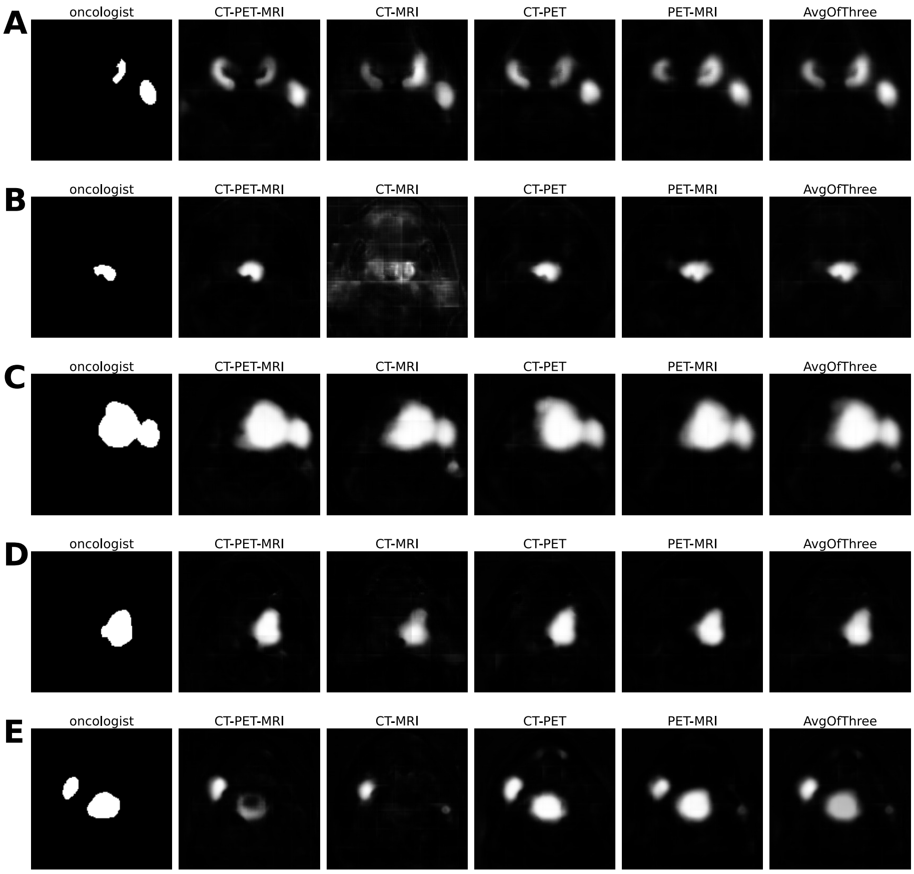

Figure 3. Five different patients' oncologist delineated binary mask; CT-PET-MRI, CT-PET, PET-MRI, CT-MRI and "Average of three" softmax output maps.

7. Result supplement

7.1. GTV volumes and evaluation scores

As mentioned in the article, there is correlation between the actual GTV volumes and prediction evaluation scores (Figure 2).

7.2. Case study probability maps

The Figure 3 shows experts delineation binary mask, CT-PET-MRI, CT-PET, PET-MRI, CT-MRI, and "average of three" probability maps of the five patients mentioned in the main article.

References

[1] Klein S, Staring M, Murphy K, et al. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging. 2009;29(1):196–205.

[2] Fortunati V, Verhaart RF, Angeloni F, et al. Feasibility of multimodal deformable registration for head and neck tumor treatment planning. International Journal of Radiation Oncology* Biology* Physics. 2014;90(1):85–93.

[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention; Springer; 2015. p. 234–241.

[4] Yu L, Yang X, Chen H, et al. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In: Thirty-first AAAI conference on artificial intelligence; 2017.

[5] Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention; Springer; 2016. p. 424–432.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[7] Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV); IEEE; 2016. p. 565–571.

[8] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2980–2988.

[9] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging; Springer; 2017. p. 379–387.

[10] Berrada L, Zisserman A, Kumar MP. Smooth loss functions for deep top-k classification. arXiv preprint arXiv:180207595. 2018;.

[11] Zhu W, Huang Y, Zeng L, et al. Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Medical physics. 2019;46(2):576–589.

[12] Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. Medical Image Analysis. 2021;:102035.