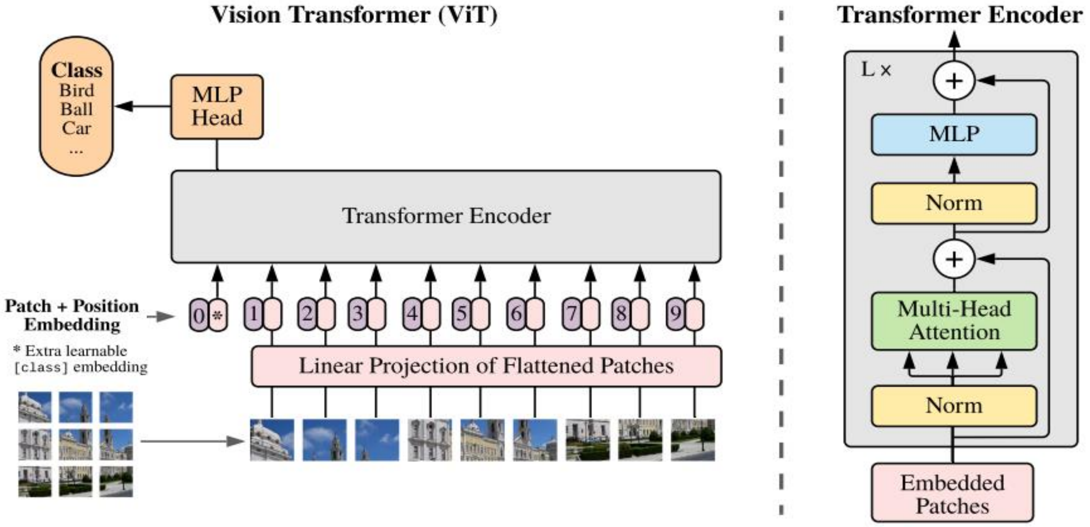
UNETR: TRANSFORMERS FOR 3D MEDICAL IMAGE SEGMENTATION

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, Daguang Xu

MONAI Bootcamp – September 22 , 2021

MOTIVATION

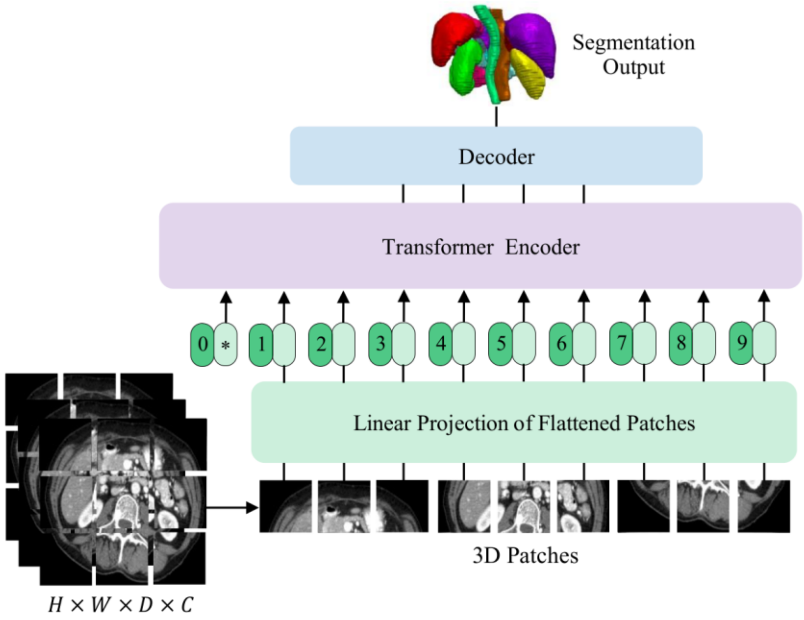
- Transformer-based models have started a revolution in NLP and computer vision due to:
  - Their exceptional capability in learning pre-text tasks
  - Scalability for large-scale training
  - Good performance in modeling long-range spatial dependencies



Vision Transformer (Dosovitskiy et al. [1])

MOTIVATION

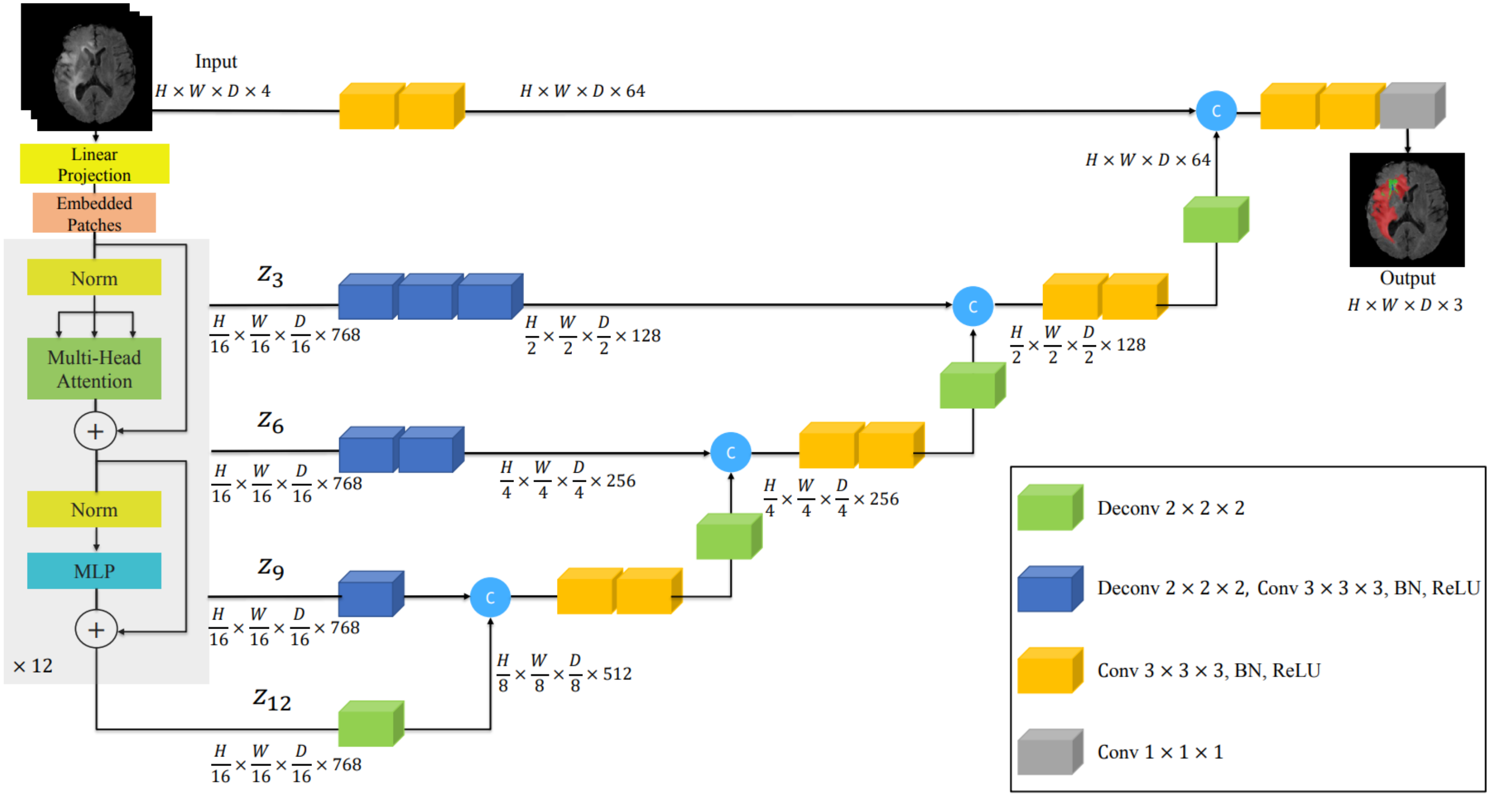
- We propose UNEt TRansformers (UNETR) which reformulates the task of 3D segmentation as 1D sequence-to-sequence prediction task:
  - Uses self-attention modules to learn weighted sum of values calculated from hidden layers
  - Achieves state-of-the-art performance on multi-organ segmentation Synapse public leaderboard



UNETR (Hatamizadeh et al. [2])

METHODOLOGY

- UNETR uses a vision transformer backbone and propose to use a CNN-based decoder in a UNET-like segmentation framework.



UNETR (Hatamizadeh et al. [2])

METHODOLOGY

- UNETR directly utilizes 3D multi-channel inputs $\mathbf{x} \in \mathbb{R}^{H\times W\times D\times C}$, divides them into non-overlapping patches $\mathbf{x}_v \in \mathbb{R}^{N\times(P^3.C)}$ with resolution P, projects the flattened patches into a K-dimentional embedding space and adds a learnable positional encoding layer to preserve the spatial information:

$$\mathbf{z}_0 = [\mathbf{x}_v^1\mathbf{E}; \mathbf{x}_v^2\mathbf{E}; ...; \mathbf{x}_v^N\mathbf{E}] + \mathbf{E}_{pos}$$

- UNETR employs transformer blocks comprising of multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers :

$$\mathbf{z}'_i = \mathrm{MSA}(\mathrm{Norm}(\mathbf{z}_{i-1})) + \mathbf{z}_{i-1}, \quad i = 1...L,$$

$$\mathbf{z}_i = \mathrm{MLP}(\mathrm{Norm}(\mathbf{z}'_i)) + \mathbf{z}'_i, \quad i = 1...L,$$

METHODOLOGY

- A MSA sublayer comprises of n parallel self-attention (SA) heads.
- The SA block, is a parameterized function that learns the mapping between a query (q) and the corresponding key (k) and value (v) representations in a sequence (z).
- The attention weights are computed by measuring the similarity between two elements in z and their key-value pairs according to

$$\mathrm{A} = \mathrm{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^{\top}}{\sqrt{K_h}}\right)$$

- Using the computed attention weights, the output of SA for values v in the sequence z and output of MSA block are computed according to

$$\mathrm{SA}(\mathbf{z}) = \mathbf{A}\mathbf{v}$$

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(\mathbf{z}); \mathrm{SA}_2(\mathbf{z}); \ldots; \mathrm{SA}_n(\mathbf{z})]\mathbf{W}_{msa}$$

QUANTITATIVE RESULTS

- UNETR model is the current state-of-the-art on BTCV public leaderboard

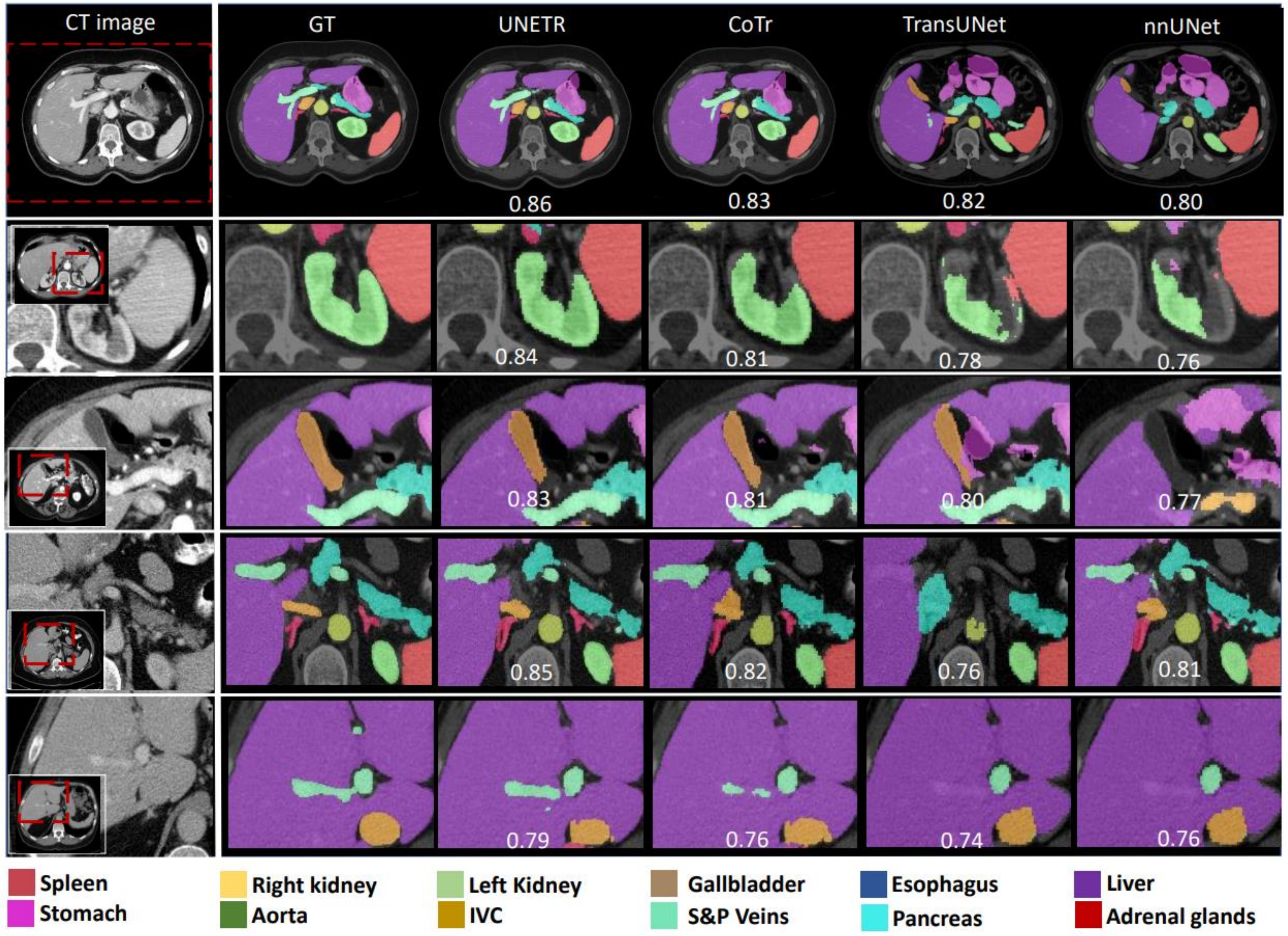
Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
SETR NUP [53]	0.931	0.890	0.897	0.652	0.760	0.952	0.809	0.867	0.745	0.717	0.719	0.620	0.796
SETR PUP [53]	0.929	0.893	0.892	0.649	0.764	0.954	0.822	0.869	0.742	0.715	0.714	0.618	0.797
SETR MLA [53]	0.930	0.889	0.894	0.650	0.762	0.953	0.819	0.872	0.739	0.720	0.716	0.614	0.796
nnUNet [21]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
ASPP [10]	0.935	0.892	0.914	0.689	0.760	0.953	0.812	0.918	0.807	0.695	0.720	0.629	0.811
TransUNet [7]	0.952	**0.927**	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
CoTr w/o CNN encoder [48]	0.941	0.894	0.909	0.705	0.723	0.948	0.815	0.876	0.784	0.723	0.671	0.623	0.801
CoTr* [48]	0.943	0.924	0.929	0.687	0.762	0.962	0.894	0.914	0.838	**0.796**	**0.783**	0.647	0.841
CoTr [48]	0.958	0.921	0.936	0.700	0.764	0.963	0.854	**0.920**	0.838	0.787	0.775	0.694	0.844
UNETR	**0.968**	0.924	**0.941**	**0.750**	**0.766**	**0.971**	**0.913**	0.890	**0.847**	0.788	0.767	**0.741**	**0.856**
RandomPatch [40]	0.963	0.912	0.921	0.749	0.760	0.962	0.870	0.889	0.846	0.786	0.762	0.712	0.844
PaNN [54]	0.966	0.927	0.952	0.732	0.791	0.973	0.891	0.914	0.850	0.805	0.802	0.652	0.854
nnUNet-v2 [21]	0.972	0.924	**0.958**	0.780	0.841	0.976	0.922	0.921	0.872	0.831	0.842	0.775	0.884
nnUNet-dys3 [21]	0.967	0.924	0.957	**0.814**	0.832	0.975	0.925	0.928	0.870	0.832	**0.849**	0.784	0.888
UNETR	**0.976**	**0.942**	0.953	**0.814**	**0.889**	**0.979**	**0.941**	**0.947**	**0.886**	**0.858**	0.823	**0.786**	**0.899**

QUANTITATIVE RESULTS

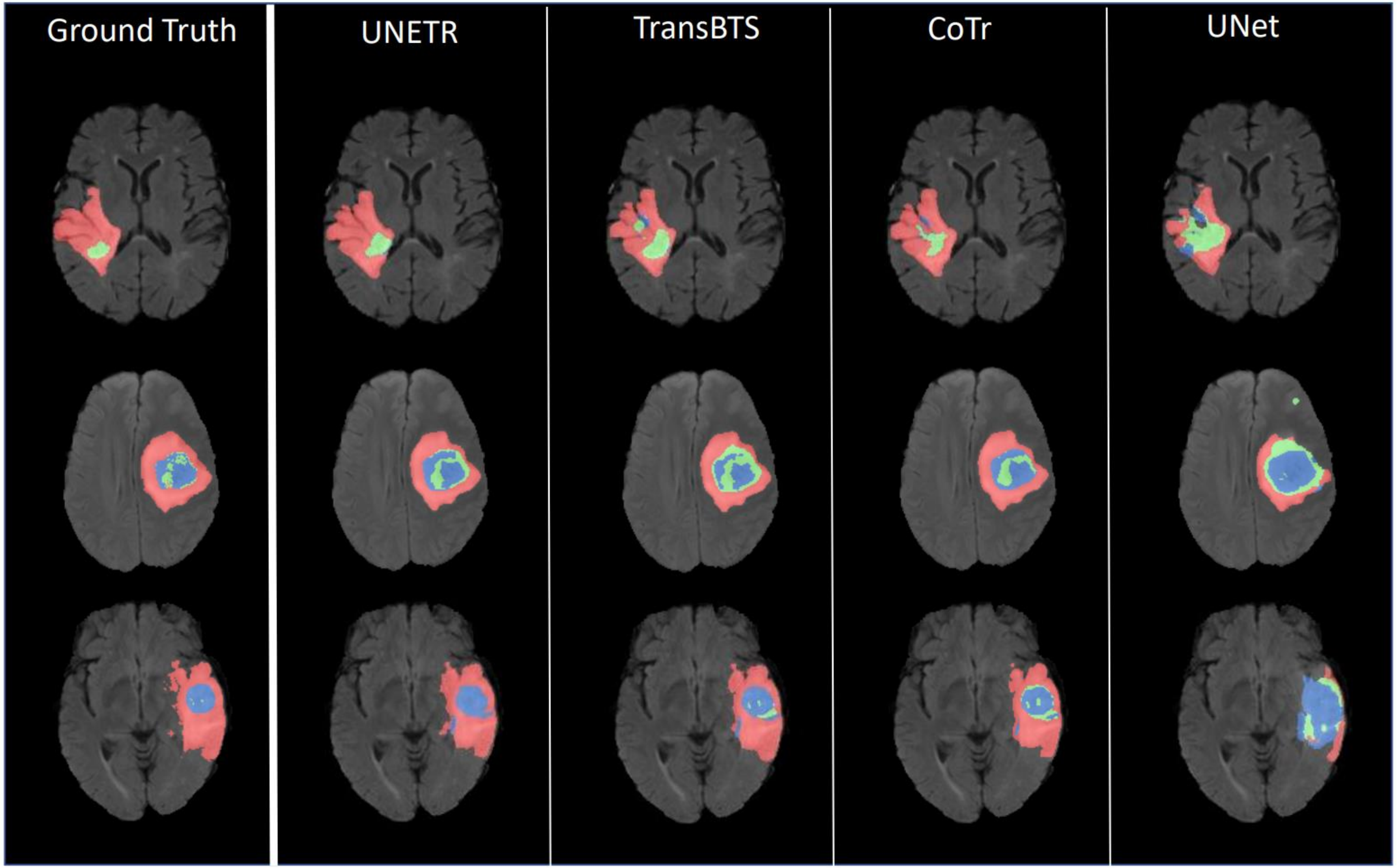
- UNETR model achives competetive performance on brain tumor and spleen segmentation task on MSD dataset.

Task/Modality	Spleen Segmentation (CT)		Brain tumor Segmentation (MRI)							
Anatomy	Spleen		WT		ET		TC		All	
Metrics	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
UNet [36]	0.953	4.087	0.766	9.205	0.561	11.122	0.665	10.243	0.664	10.190
AttUNet [34]	0.951	4.091	0.767	9.004	0.543	10.447	0.683	10.463	0.665	9.971
SETR NUP [53]	0.947	4.124	0.697	14.419	0.544	11.723	0.669	15.192	0.637	13.778
SETR PUP [53]	0.949	4.107	0.696	15.245	0.549	11.759	0.670	15.023	0.638	14.009
SETR MLA [53]	0.950	4.091	0.698	15.503	0.554	10.237	0.665	14.716	0.639	13.485
TransUNet [7]	0.950	4.031	0.706	14.027	0.542	10.421	0.684	14.501	0.644	12.983
TransBTS [44]	-	-	0.779	10.030	0.574	9.969	0.735	8.950	0.696	9.650
CoTr w/o CNN encoder [48]	0.946	4.748	0.712	11.492	0.523	9.592	0.698	12.581	0.6444	11.221
CoTr [48]	0.954	3.860	0.746	9.198	0.557	9.447	0.748	10.445	0.683	9.697
UNETR	**0.964**	**1.333**	**0.789**	**8.266**	**0.585**	**9.354**	**0.761**	**8.845**	**0.711**	**8.822**

QUALITATIVE RESULTS



QUALITATIVE RESULTS



MODEL COMPLEXITY

- Comparison of number of parameters, FLOPs and averaged inference time for various models in BTCV experiments.

Models	#Params (M)	FLOPs (G)	Inference Time (s)
nnUNet [21]	19.07	412.65	10.28
CoTr [48]	46.51	399.21	19.21
TransUNet [7]	96.07	48.34	26.97
ASPP [11]	47.92	44.87	25.47
SETR [53]	86.03	43.49	24.86
UNETR	92.58	41.19	12.08

CONCLUSION

- In this work, we have proposed a novel transformer-based segmentation network dubbed as UNETR for medical imaging semantic segmentation
- Our proposed UNETR is the current state-of-the-art on BTCV public leaderboard for the task of multi-organ semantic segmentation. UNETR also achieves competetive performance on spleen and brains segmentation tasks using MSD dataset.
- Our work paves the way for a new class of transformer-based networks for medical image segmentation.
- UNETR is currently available as part of MONAI:
  - Repository: https://monai.io/research/unetr
  - Tutorial: https://github.com/Project-MONAI/tutorials/blob/master/3d_segmentation/unetr_btcv_segmentation_3d.ipynb
  - Paper: https://arxiv.org/abs/2103.10504

REFERENCES

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020, September. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

2. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., and Xu, D., 2021, August. UNETR: Transformers for 3D Medical Image Segmentation. *arXiv preprint arXiv:2103.10504* (2021).

END OF SLIDES