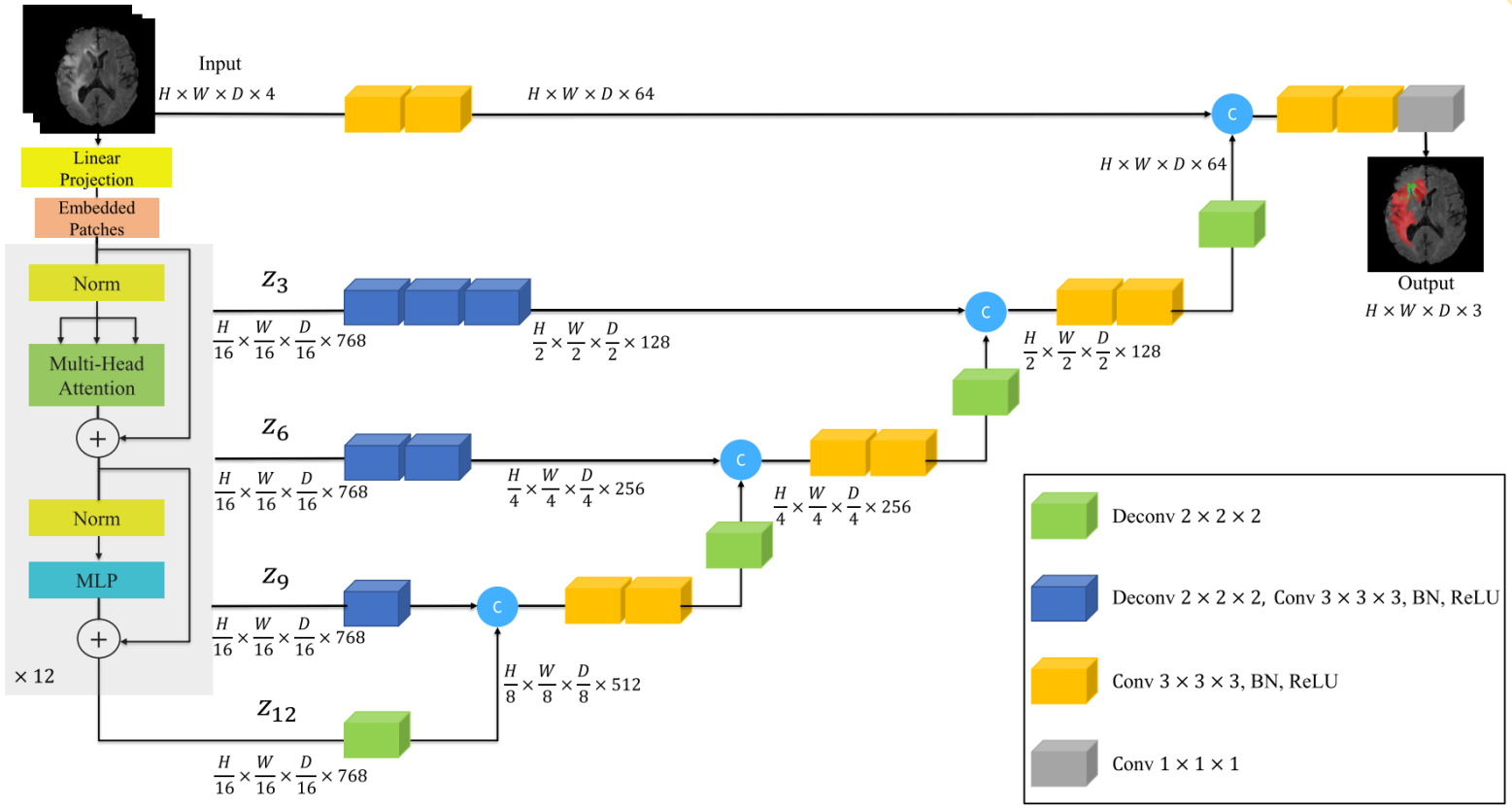
UNETR: Transformers for 3D Medical Image Segmentation

Presenter: Seungjun Lee

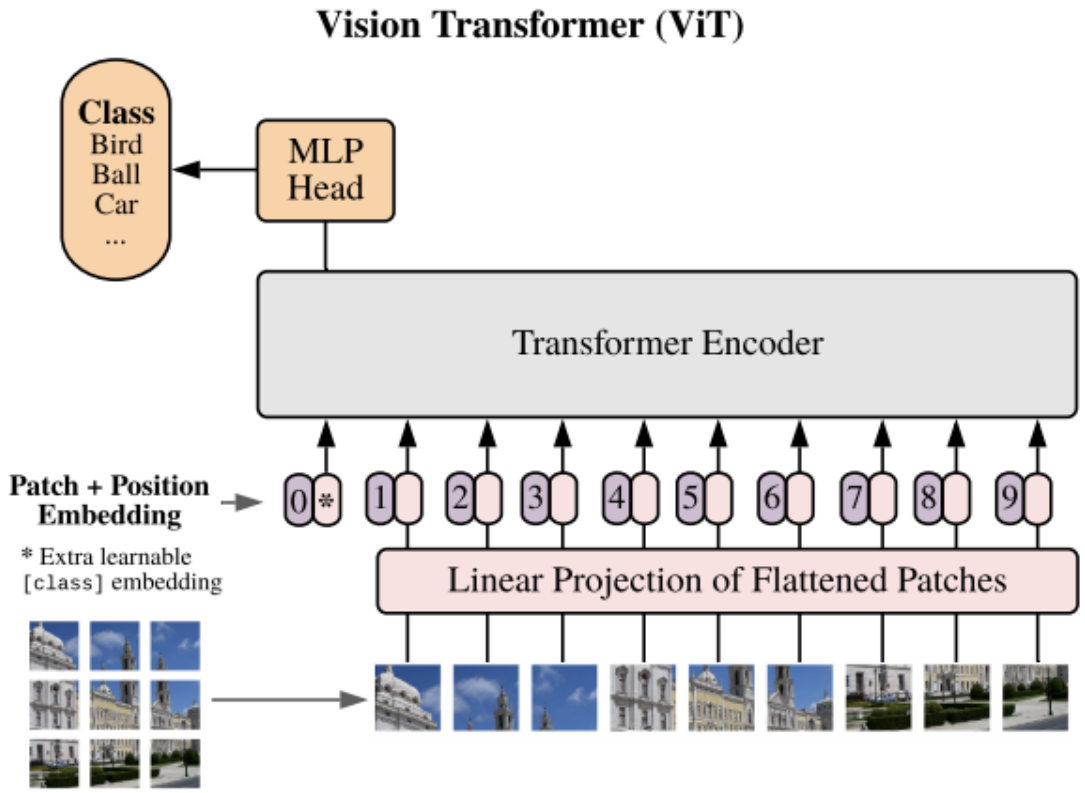
Abstract

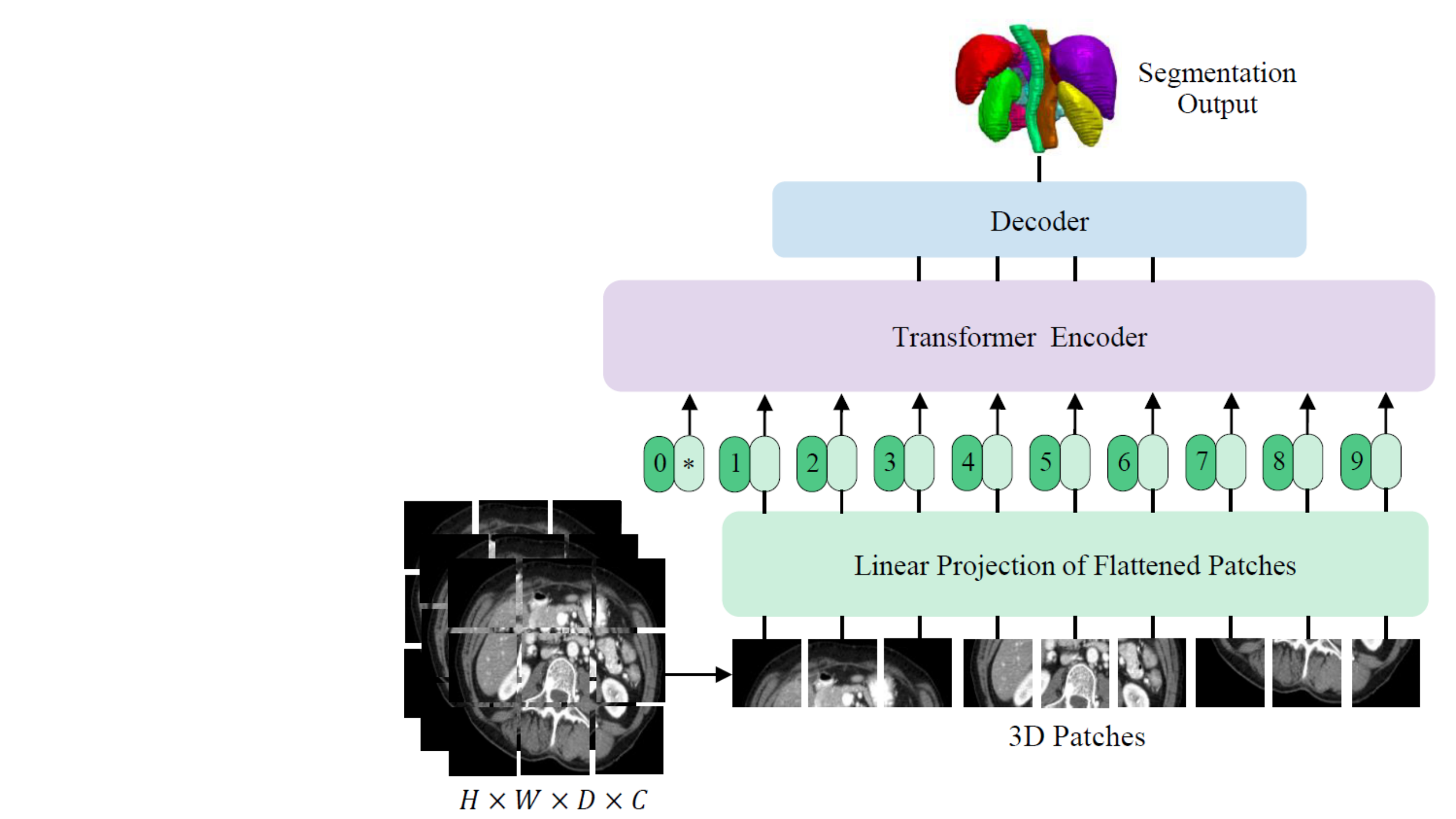
- FCNNs(Fully Convolutional Neural Networks)
  - 기존 Image segmentation 영역에서 대표적인 네트워크 구조
  - local한 receptive fields로 인한 global context와 long-range dependency에 제한적인 성능
- UNETR(UNEt TRansformers) 구조를 제안
  - Transformer를 쌓은 encoder에서 sequence representations를 학습
  - Unet처럼 encoder와 CNN-based decoder를 skip-connection으로 연결하는 구조
- 3D segmentation task: BTCV, MSD dataset
  - BTCV에서 SOTA 달성
  - MSD에서 brain tumor와 spleen segmentation에서 SOTA 달성



Transformers in Computer Vision

- Transformers as a backbone encoder
  - long-range dependencies와 global context를 학습하는 데 용이
  - Transformer는 이미지를 1차원 패치로 embedding 후, self-attention module을 통해 은닉층에서 값들의 가중합을 구함
  - Pre-text task에서 학습 후 down-stream에서 활용됨



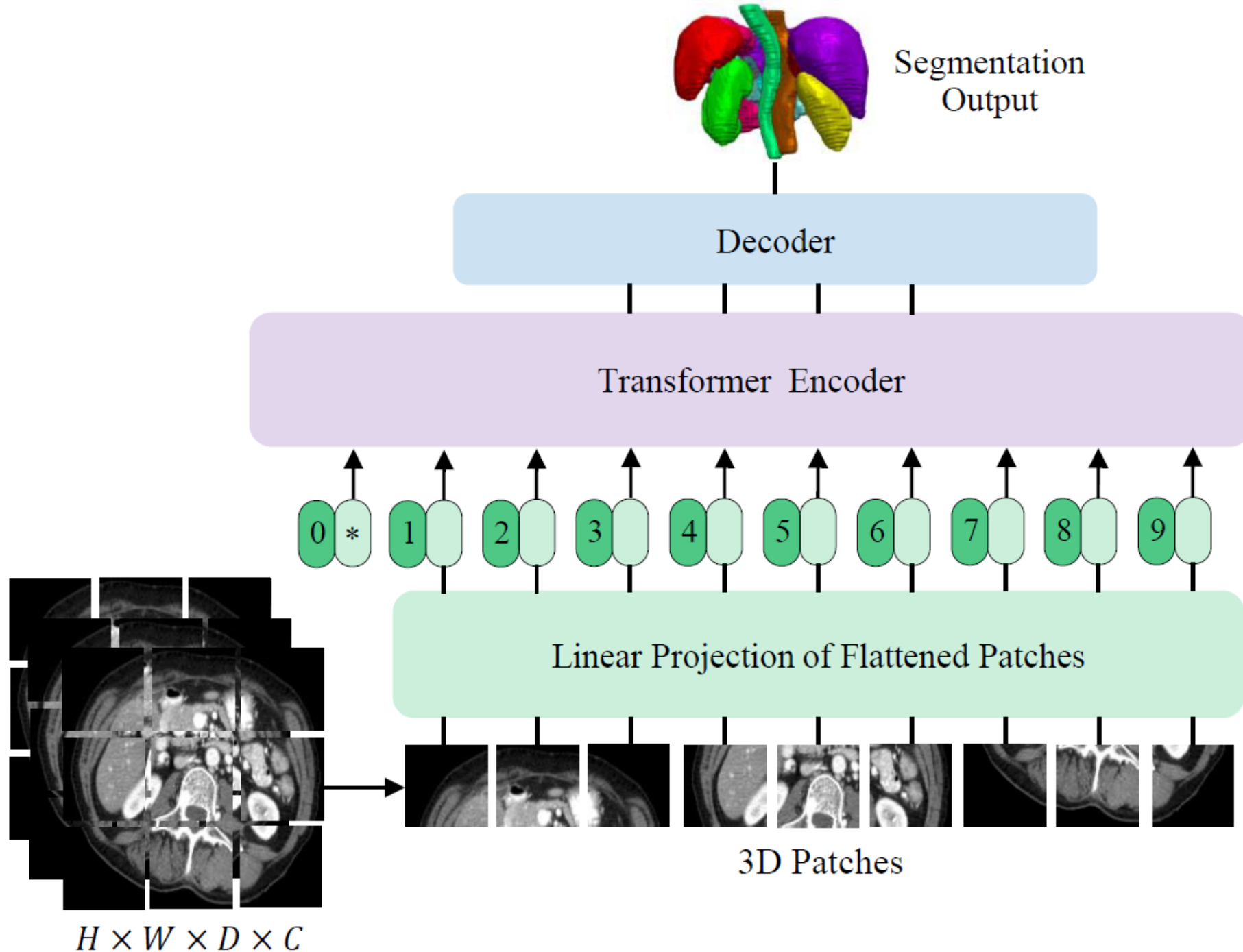

Segmentation Output
Decoder
Transformer Encoder
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
3D Patches
$H \times W \times D \times C$

Point 1.

3D segmentation

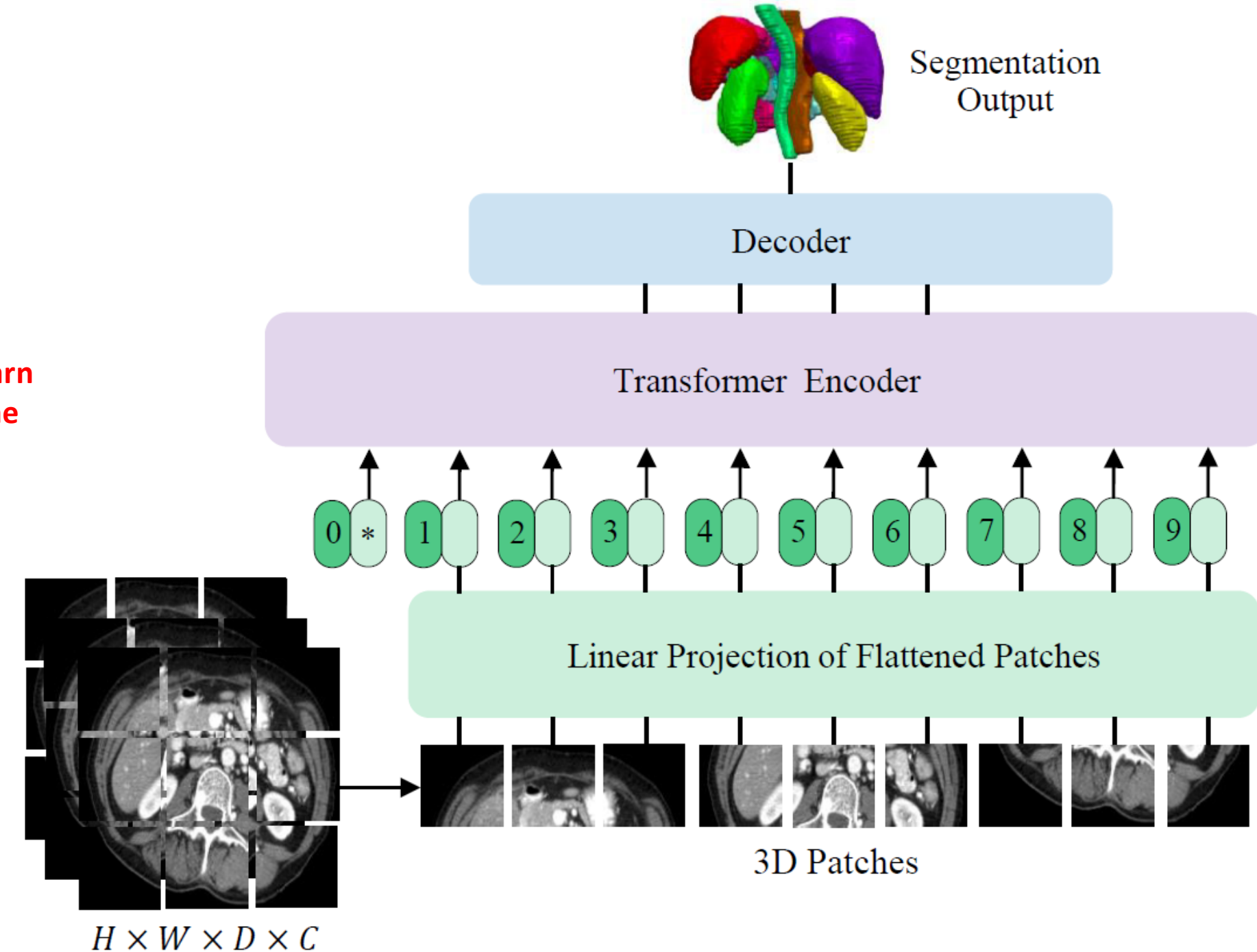
↓

1D sequence-to-sequence prediction problem



Point 2.

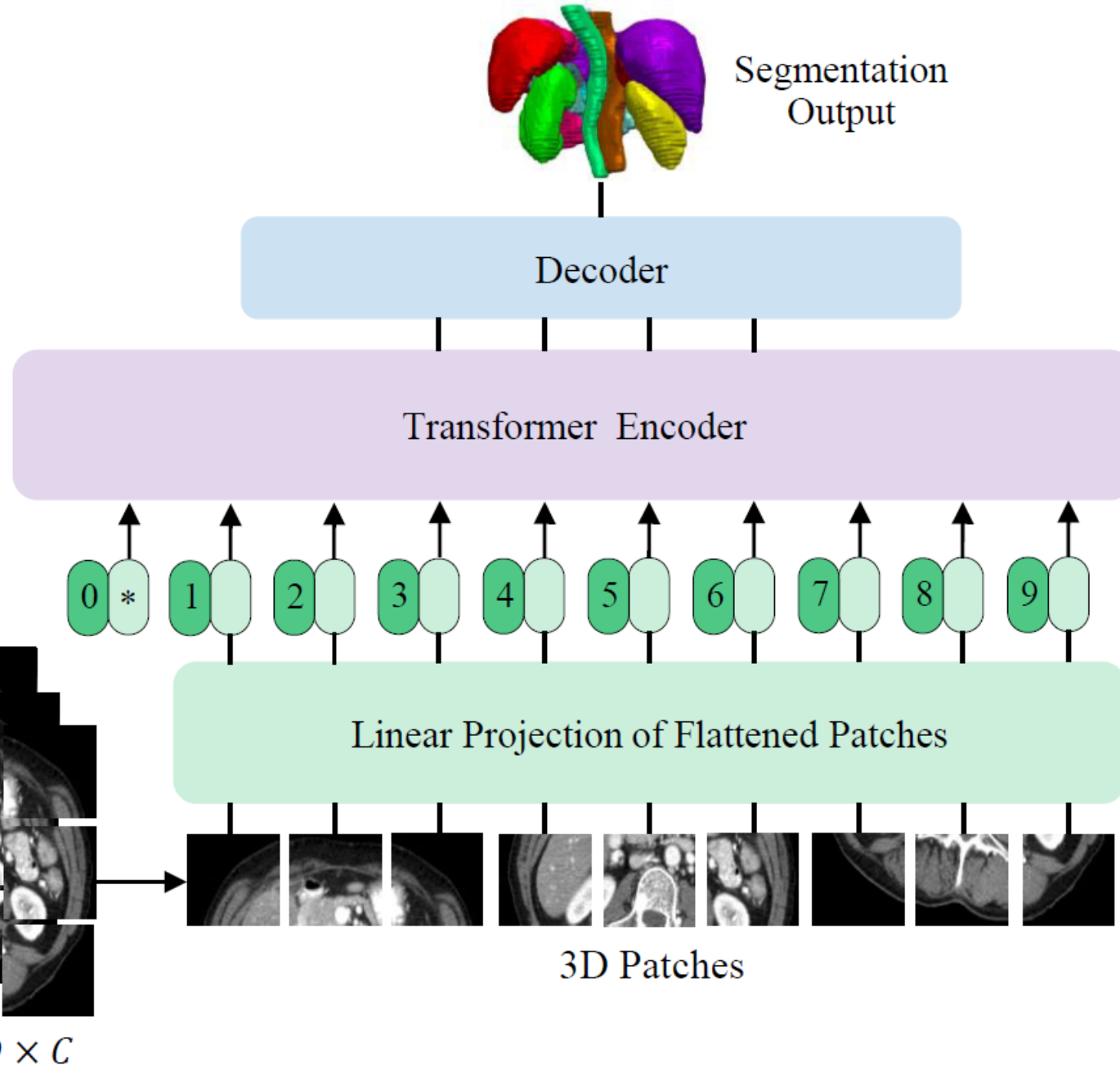
Use transformer encoder to learn contextual information from the embedded input patches



Point 3.

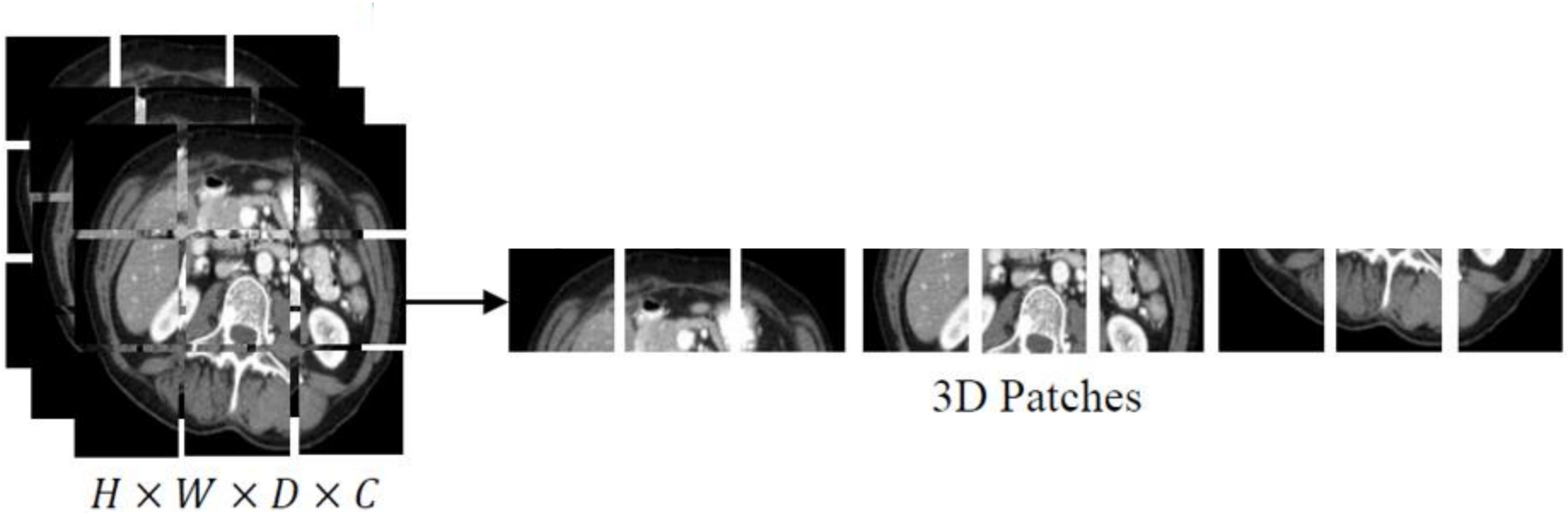
CNN-based decoder via skip-connections at multiple resolutions

-> capture both global and local dependencies



3D Input -> 3D patches

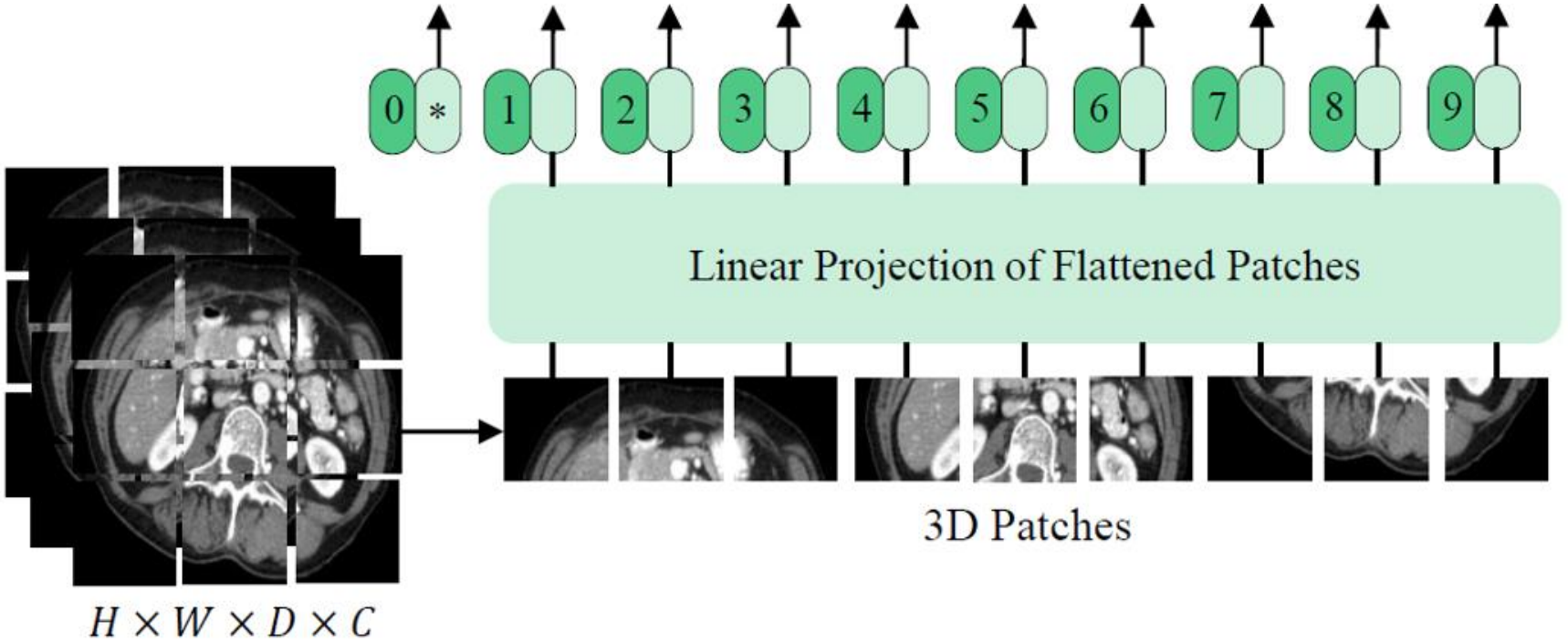
- 3D input volume $x \in \mathrm{R}^{H \times W \times D \times C}$
- Non-overlapping patches $x_v \in \mathrm{R}^{N \times (P^3 \cdot C)}$ (where $N = (H \times W \times D)/P^3$)
- P = 16

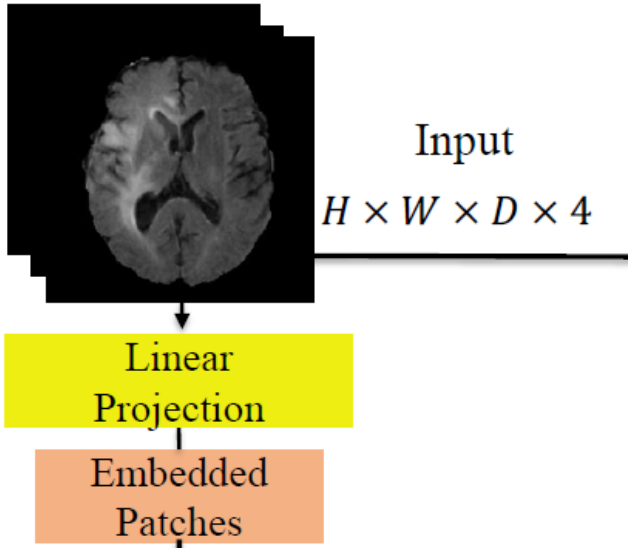


Linear Projection & Positional embedding

- Projection of x_v into a 768-dimensional embedding space using a linear layer
- The sequence is added with a 1D learnable positional embedding

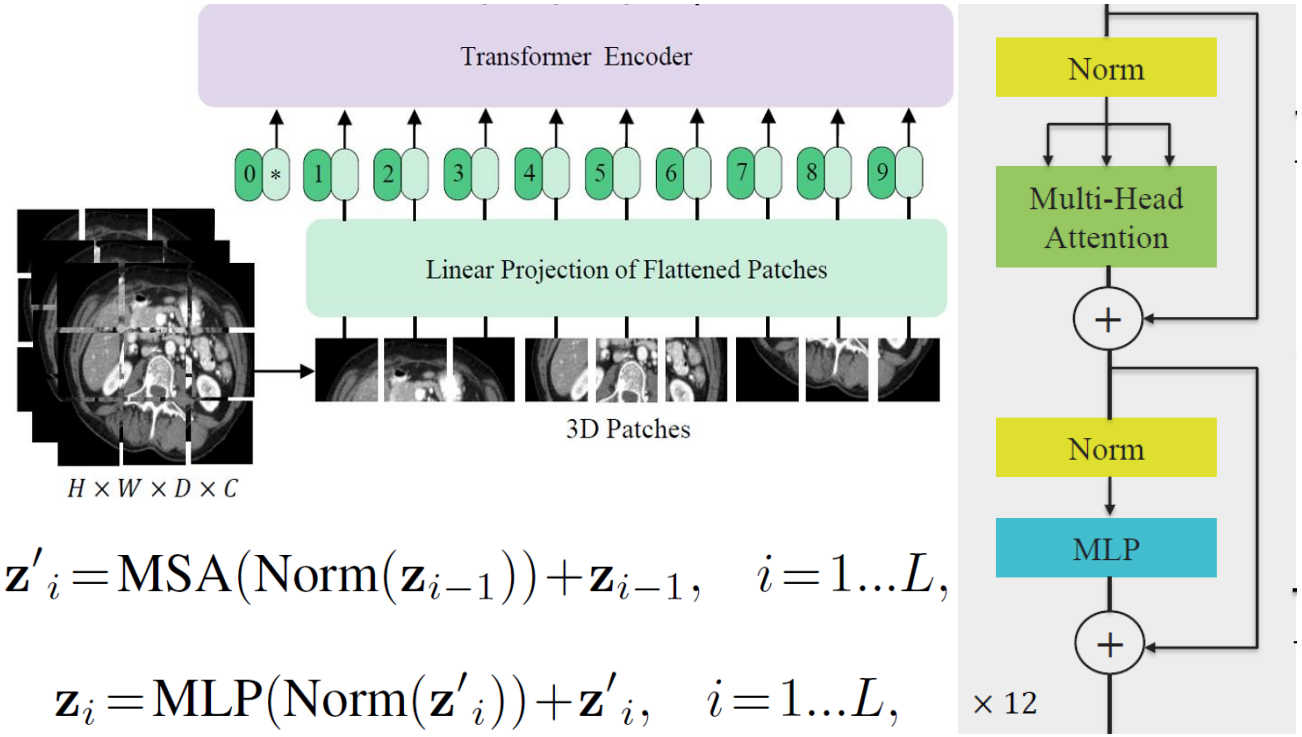
$$\mathbf{z}_0 = [\mathbf{x}_v^1\mathbf{E}; \mathbf{x}_v^2\mathbf{E}; \ldots; \mathbf{x}_v^N\mathbf{E}] + \mathbf{E}_{pos}$$





Stacked transformer

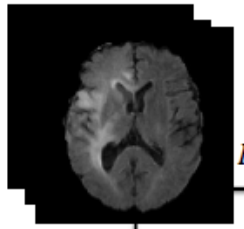
- Multi-head self-attention (MSA)
  - MSA sublayer comprises of n parallel self-attention heads
- Multilayer perceptron (MLP)
  - Two linear layers with GELU



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

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

Norm(): layer normalization
L: the number of transformer layers



Input

$H \times W \times D \times 4$

Linear Projection

Embedded Patches

Norm

Multi-Head Attention

Norm

MLP

× 12

$H \times W \times D \times 64$

Z_3

$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$

$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$

Z_6

$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$

$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$

Z_9

$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$

$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$

Z_{12}

$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$

C

$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$

C

$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$

C

$H \times W \times D \times 64$

Output

$H \times W \times D \times 3$

Deconv $2 \times 2 \times 2$

Deconv $2 \times 2 \times 2$, Conv $3 \times 3 \times 3$, BN, ReLU

Conv $3 \times 3 \times 3$, BN, ReLU

Conv $1 \times 1 \times 1$

Implementation details

- Batch size: 6
- Optimizer: AdamW
- Learning rate: 0.0001
- Iteration: 20000 (10 hours on a NVIDIA-DGX-1 server)
- Backbone: ViT-B16 L=12, K=768, P=16×16×16
  - No pretrained weights (no performance improvements)
- Augmentation: Random rotation(90°, 180°, 270°), Random flip(axial, sagittal, coronal views), Random scale, Shift intensity
- Ensemble: Five-fold cross-validation

3.2. Loss Function

Our loss function is a combination of soft dice loss [32] and cross-entropy loss, and it can be computed in a voxel-wise manner according to

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \\ -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j}. \tag{7}$$

where I is the number of voxels; J is the number of classes; $Y_{i,j}$ and $G_{i,j}$ denote the probability output and one-hot encoded ground truth for class j at voxel i, respectively.

Dataset - BTCV

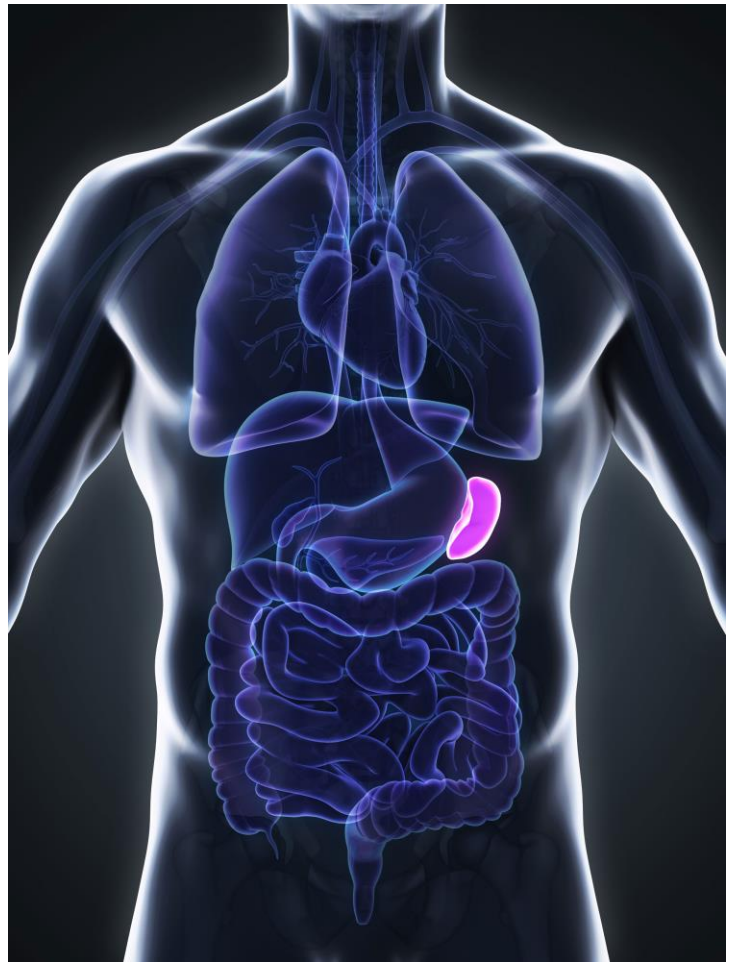
BTCV (CT): The BTCV dataset [26] consists of 30 subjects with abdominal CT scans where 13 organs were annotated by interpreters under supervision of clinical radiologists at Vanderbilt University Medical Center. Each CT scan was acquired with contrast enhancement in portal venous phase and consists of 80 to 225 slices with 512×512 pixels and slice thickness ranging from 1 to 6 mm. Each volume has been pre-processed independently by normalizing the intensities in the range of [-1000,1000] HU to [0,1]. All images are resampled into the isotropic voxel spacing of 1.0 mm during pre-processing. The multi-organ segmentation problem is formulated as a 13 class segmentation task with 1-channel input.

Dataset – MSD (MRI/CT)

MSD (MRI/CT): For the brain tumor segmentation task, the entire training set of 484 multi-modal multi-site MRI data (FLAIR, T1w, T1gd, T2w) with ground truth labels of gliomas segmentation necrotic/active tumor and oedema is utilized for model training. The voxel spacing of MRI images in this tasks is $1.0 \times 1.0 \times 1.0\ mm^3$. The voxel intensities are pre-processed with z-score normalization. The problem of brain tumor segmentation is formulated as a 3 class segmentation task with 4-channel input.

Dataset – spleen segmentation

For the spleen segmentation task, 41 CT volumes with spleen body annotation are used. The resolution/spacing of volumes in task 9 ranges from $0.613 \times 0.613 \times 1.50\ mm^3$ to $0.977 \times 0.977 \times 8.0\ mm^3$. All volumes are re-sampled into the isotropic voxel spacing of $1.0\ mm$ during pre-processing. The voxel intensities of the images are normalized to the range $[0,1]$ according to 5th and 95th percentile of overall foreground intensities. Spleen segmentation is formulated as a binary segmentation task with 1-channel input. For multi-organ and spleen segmentation tasks, we randomly sample the input images with volume sizes of $[96,96,96]$. For brain segmentation task, we randomly sample the input images with volume sizes of $[128, 128, 128]$. For all experiments, the random patches of foreground/background are sampled at ratio $1:1$.



Evaluation metrics

- Dice score and 95% Hausdorff Distance (HD)

$$\mathrm{Dice}(G,P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}, \tag{8}$$

$$\mathrm{HD}(G',P') = \max\{\max_{g' \in G'} \min_{p' \in P'} \|g' - p'\|, \max_{p' \in P'} \min_{g' \in G'} \|p' - g'\|\}. \tag{9}$$

The 95% HD uses the 95th percentile of the distances between ground truth and prediction surface point sets. As a result, the impact of a very small subset of outliers is minimized when calculating HD.

BTCV

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
SETR NUP [52]	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 [52]	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 [52]	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 [47]	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* [47]	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 [47]	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 [39]	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 [53]	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	**0.972**	**0.942**	0.954	**0.825**	**0.864**	**0.983**	**0.945**	**0.948**	**0.890**	**0.858**	0.799	**0.812**	**0.891**

Table 1. Quantitative comparisons of segmentation performance in BTCV test set. Top and bottom sections represent the benchmarks of Standard and Free Competitions respectively. Our method is compared against current state-of-the-art models. All SETR [52] baselines use ViT-B-16 [14] backbone. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, AG: adrenal gland. All results obtained from BTCV leaderboard.

MSD

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 [52]	0.947	4.124	0.697	14.419	0.544	11.723	0.669	15.192	0.637	13.778
SETR PUP [52]	0.949	4.107	0.696	15.245	0.549	11.759	0.670	15.023	0.638	14.009
SETR MLA [52]	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 [43]	-	-	0.779	10.030	0.574	9.969	0.735	8.950	0.696	9.650
CoTr w/o CNN encoder [47]	0.946	4.748	0.712	11.492	0.523	9.592	0.698	12.581	0.6444	11.221
CoTr [47]	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**

Table 2. Quantitative comparisons of the segmentation performance in brain tumor and spleen segmentation tasks of the MSD dataset. WT, ET and TC denote Whole Tumor, Enhancing tumor and Tumor Core sub-regions respectively.

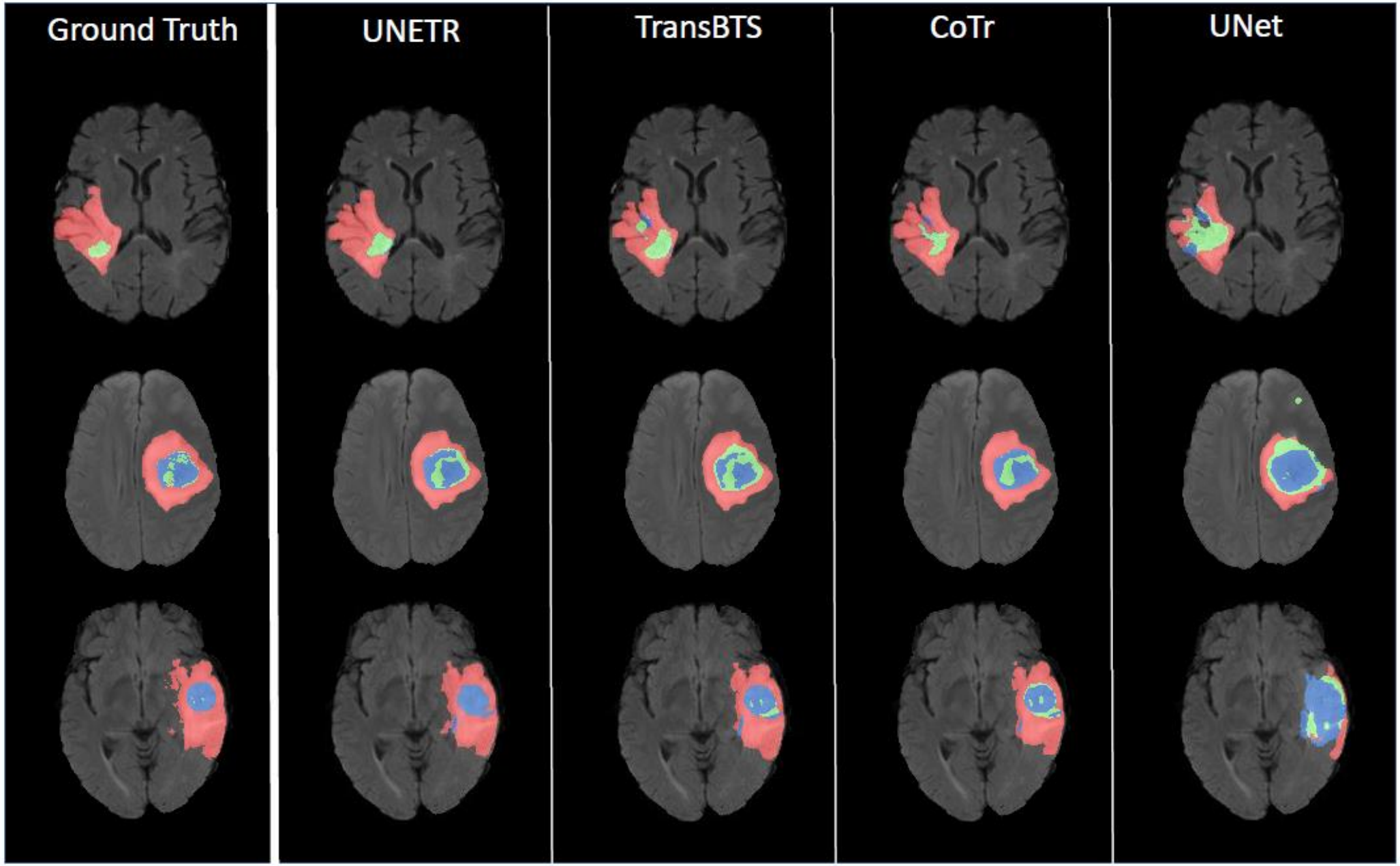


Figure 4. UNETR effectively captures the fine-grained details in segmentation outputs. The Whole Tumor (WT) encompasses a union of red, blue and green regions. The Tumor Core (TC) includes the union of red and blue regions. The Enhancing Tumor core (ET) denotes the green regions.

Ablation - Decoder

Organ	Spleen	Brain			
Decoder	Spleen	WT	ET	TC	All
NUP	0.932	0.721	0.527	0.660	0.636
PUP	0.941	0.749	0.558	0.698	0.668
MLA	0.950	0.757	0.563	0.732	0.684
UNETR	**0.964**	**0.789**	**0.585**	**0.761**	**0.711**

Table 3. Effect of the decoder architecture on segmentation performance. NUP, PUP and MLA denote Naive UpSampling, Progressive UpSampling and Multi-scale Aggregation.

Ablation – patch size

Organ	Spleen	Brain			
Resolution	Spleen	WT	ET	TC	All
32	0.953	0.776	0.579	0.756	0.703
16	**0.964**	**0.789**	**0.585**	**0.761**	**0.711**

Table 4. Effect of patch resolution on segmentation performance.

Model and computational complexity

Models	#Params (M)	FLOPs (G)	Inference Time (s)
nnUNet [21]	19.07	412.65	10.28
CoTr [47]	46.51	399.21	19.21
TransUNet [7]	96.07	48.34	26.97
ASPP [11]	47.92	44.87	25.47
SETR [52]	86.03	43.49	24.86
UNETR	92.58	41.19	12.08

Table 5. Comparison of number of parameters, FLOPs and averaged inference time for various models in BTCV experiments.

Conclusion

- 본 논문은 semantic segmentation task에서 새로운 transformer 기반의 architecture UNETR을 제안하였다.
- UNETR은 transformer 기반 encoder를 사용하여 model의 long-range dependencies를 학습하는 능력을 올렸으며 효과적으로 global contextual representation을 capture 할 수 있다.
- BTCV, MSD dataset에서 좋은 성능을 보이며 BTCV dataset에서는 SOTA를 달성하였다.
- 본 architecture는 transformer 기반의 medical image segmentation model에 새로운 foundation이 될 것이다.