Transformers for medical image segmentation

SCAI-IDS Workshop 2023: AI and Medicine

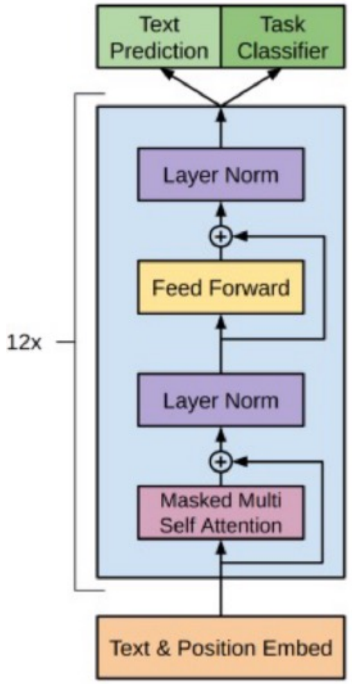
THOME Nicolas – Prof. at SORBONNE University
ISIR Lab, MLIA TEAM

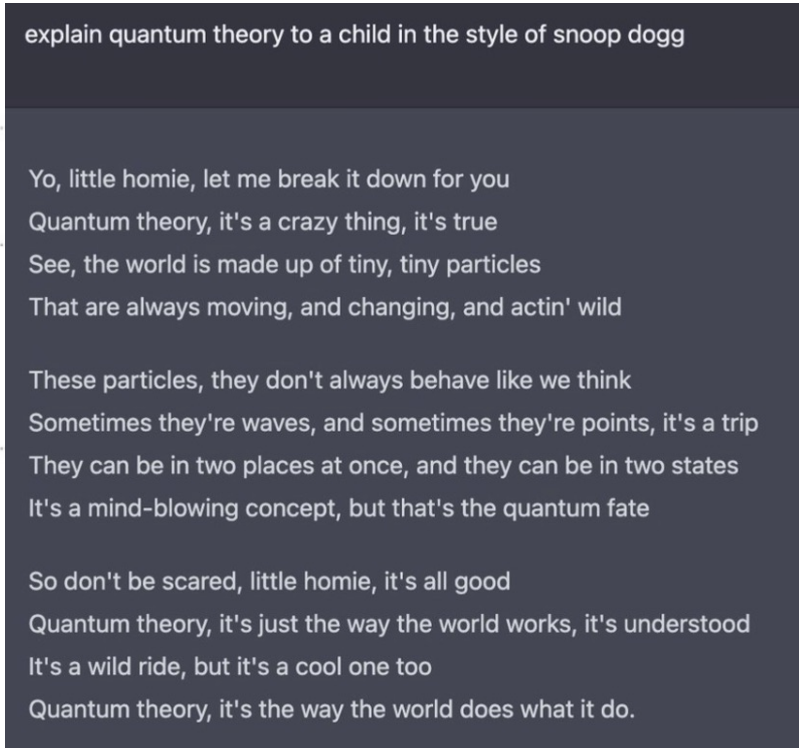
Loïc Themyr, Olivier Petit, Clément Rambour – Cnam Paris
Toby Collins, Alex Hostettler, IRCAD Strasbourg & Africa
Luc Soler, Visible Patient Inc

1. Context
2. U-Net Transformer
3. GLAM

Transformers everywhere since 2017

NLP: BERT, GPT3, Chat-CPT, etc





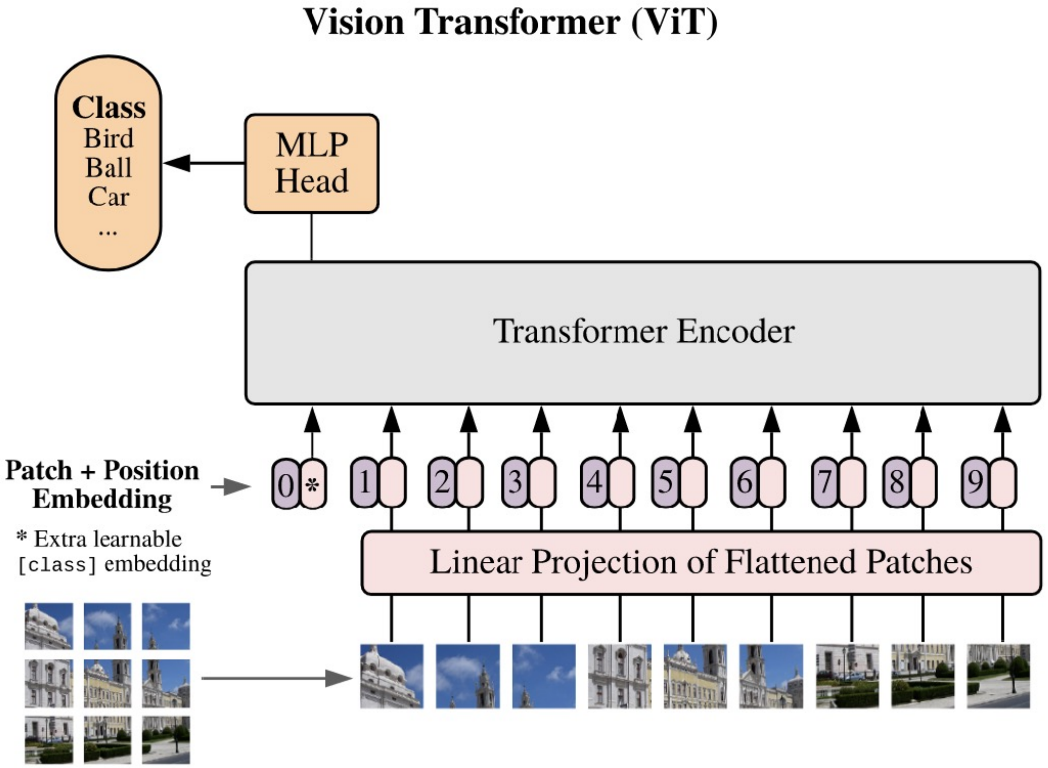
explain quantum theory to a child in the style of snoop dogg

Yo, little homie, let me break it down for you
Quantum theory, it's a crazy thing, it's true
See, the world is made up of tiny, tiny particles
That are always moving, and changing, and actin' wild

These particles, they don't always behave like we think
Sometimes they're waves, and sometimes they're points, it's a trip
They can be in two places at once, and they can be in two states
It's a mind-blowing concept, but that's the quantum fate

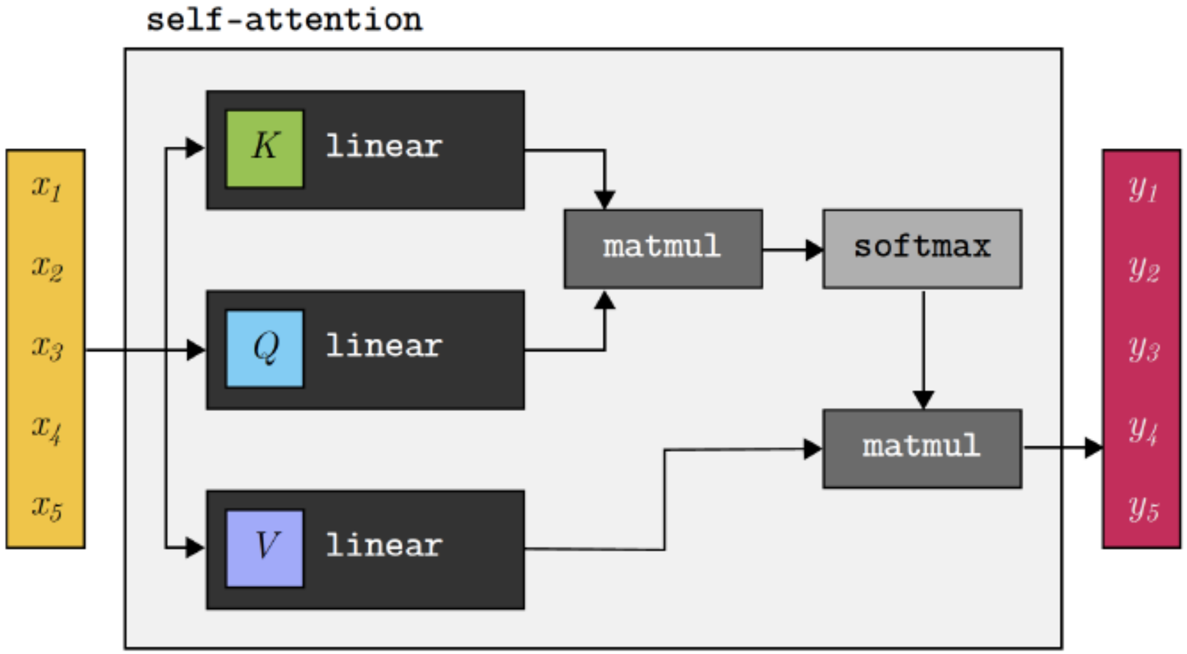
So don't be scared, little homie, it's all good
Quantum theory, it's just the way the world works, it's understood
It's a wild ride, but it's a cool one too
Quantum theory, it's the way the world does what it do.

Vision since ’21: Vision Image Transformer (ViT)



Transformers: self-attention

Key element: self-attention



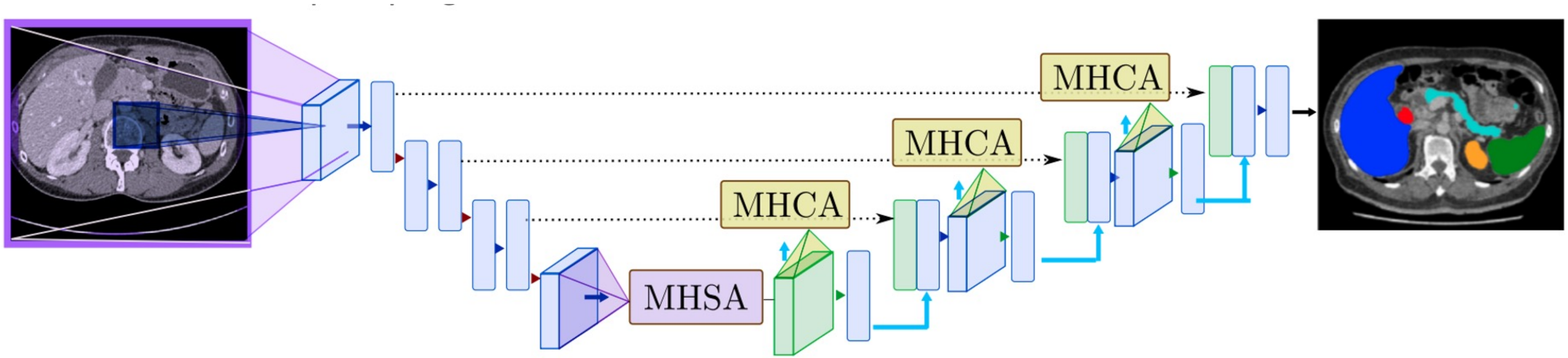
- Each «token» re-embedded wrt all token

  ⇒ Global interactions

- Self attention: $O(N^2)$ complexity
  - Expensive (or impossible) for large N

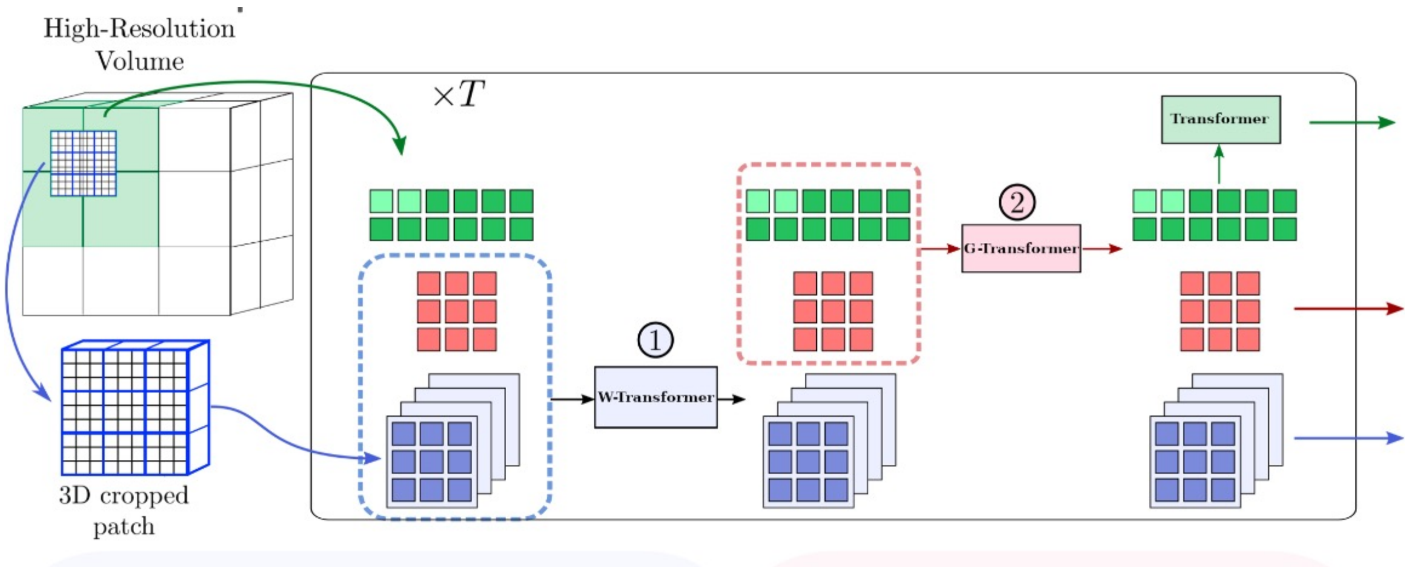
Transformers for medical image segmentation

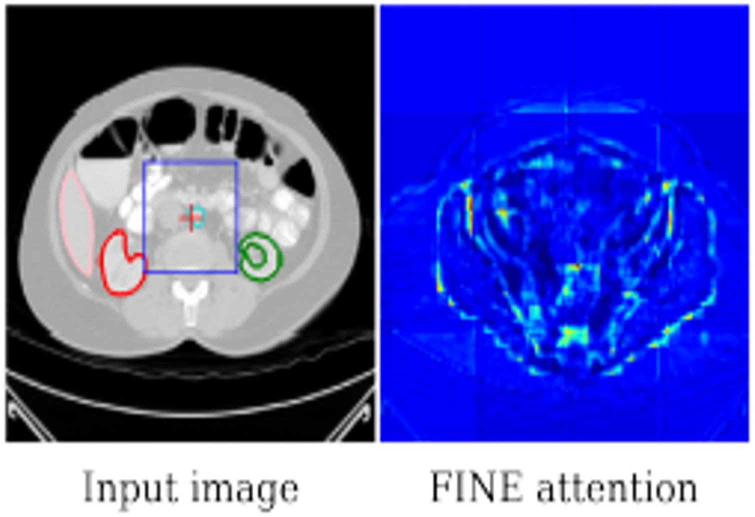
- **Which features for medical image segmentation?**
  - How to include long-range dependencies in U-shaped architectures
  - **U-Transformer** [MLMI'21]: self and cross attention in medical image segmentation



Transformers for medical image segmentation

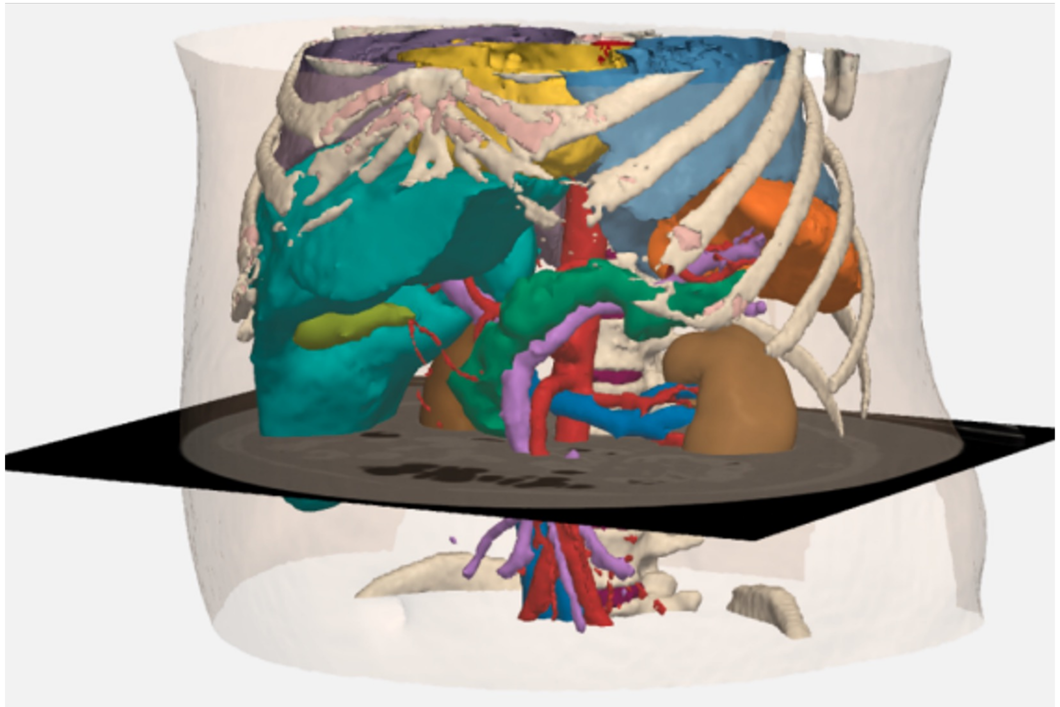
- Full transformer architecture: more and more competitive in segmentation
- **How to adapt them to the** medical **domain's specicities?**
  - 3D inputs: bottleneck in computational complexity
  - Ex: 512^3 input volume : impossible to get tractable attention!
- **Full attention in 3D transformers with indirections (GLAM/FINE)** [WACV'23, MLMI'22]



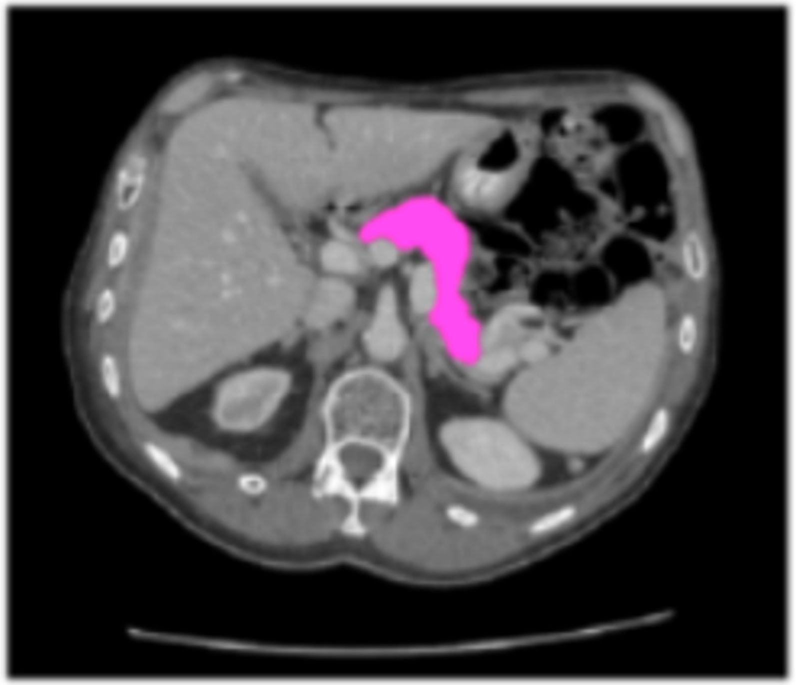


1. Context
2. U-Net Transformer
3. Full attention in 3D transformers

Context



Organs segmentation illustration

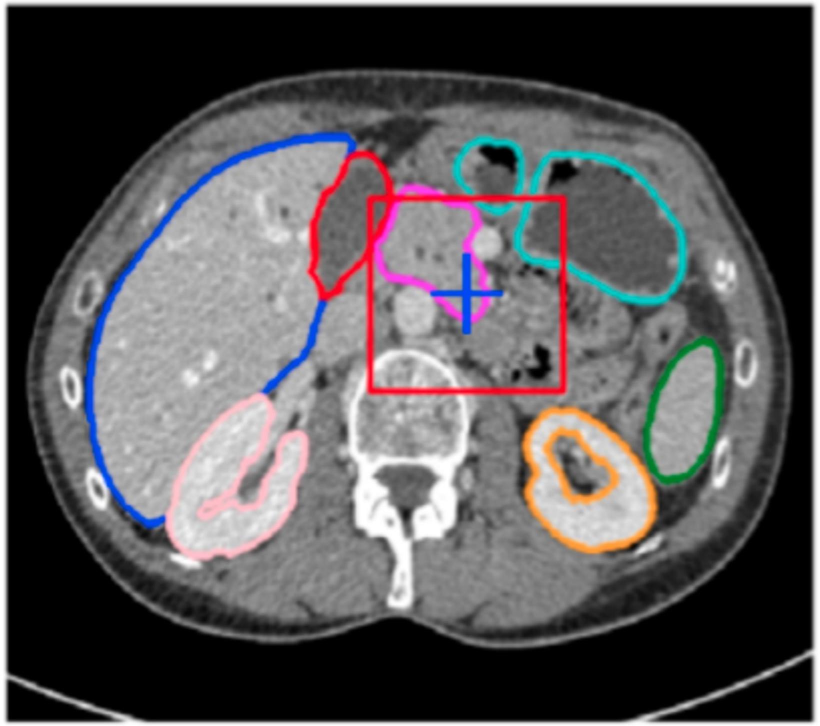


Pancreas automatic segmentation

Context

Problems:

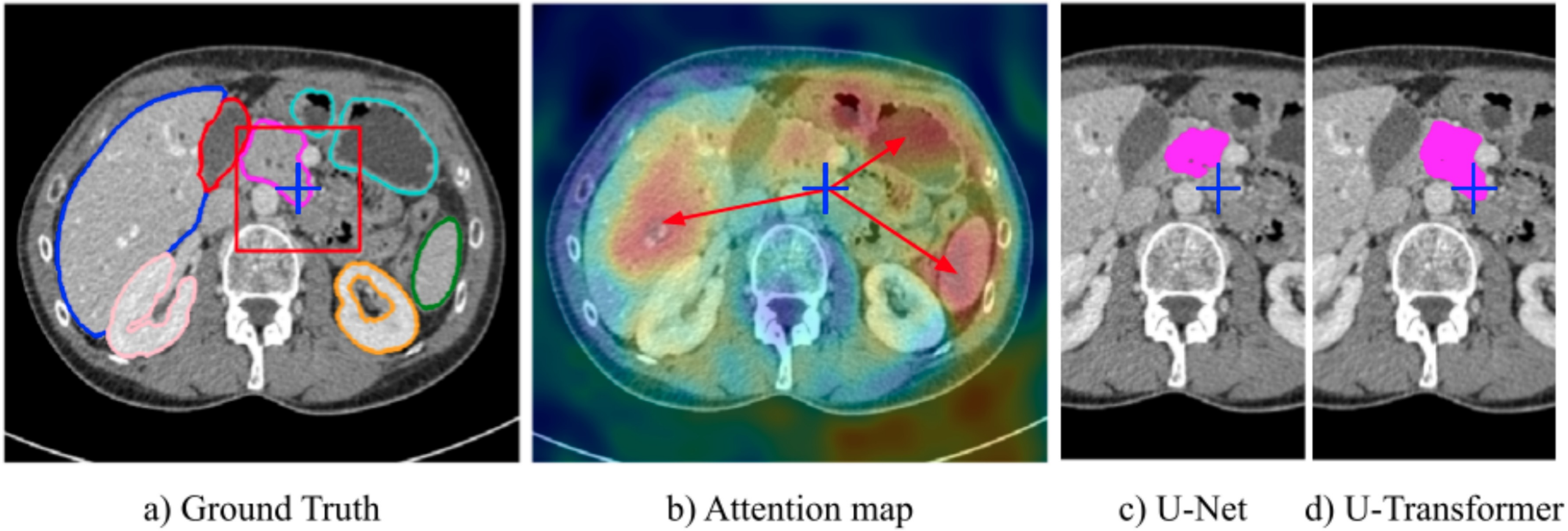
- High image resolution (512x512 pixels)
- Importance of global context in medical images segmentation
- Limitations of ConvNets receptive field



Abdominal CT-scan exemple showing receptieve field's limitations

U-Net Transformer

Hybridation between U-Net [A] and Transformers

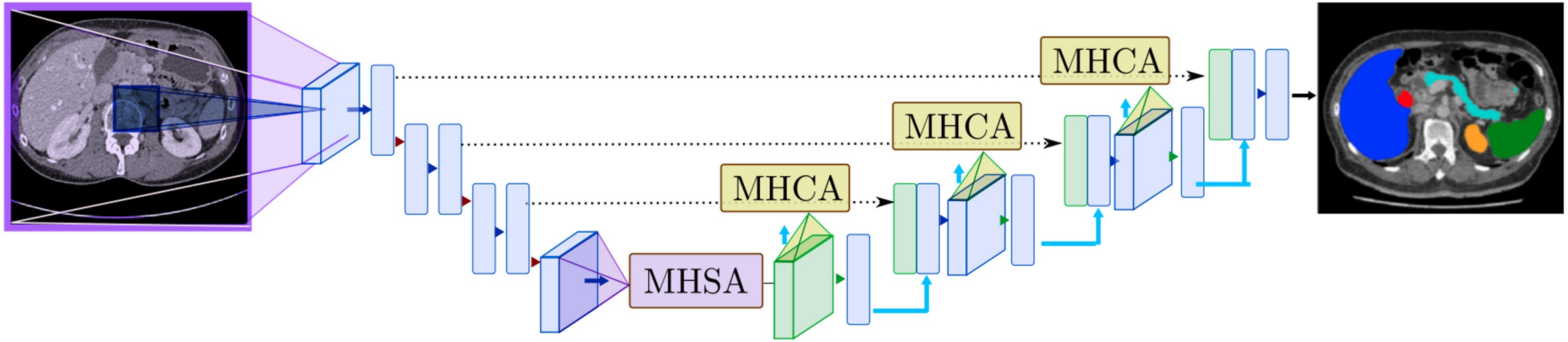


Segmentation example with U-Net's receptive field (red square) and U-Transformer's attention map.

[A] O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomedical image segmentation, 2015.

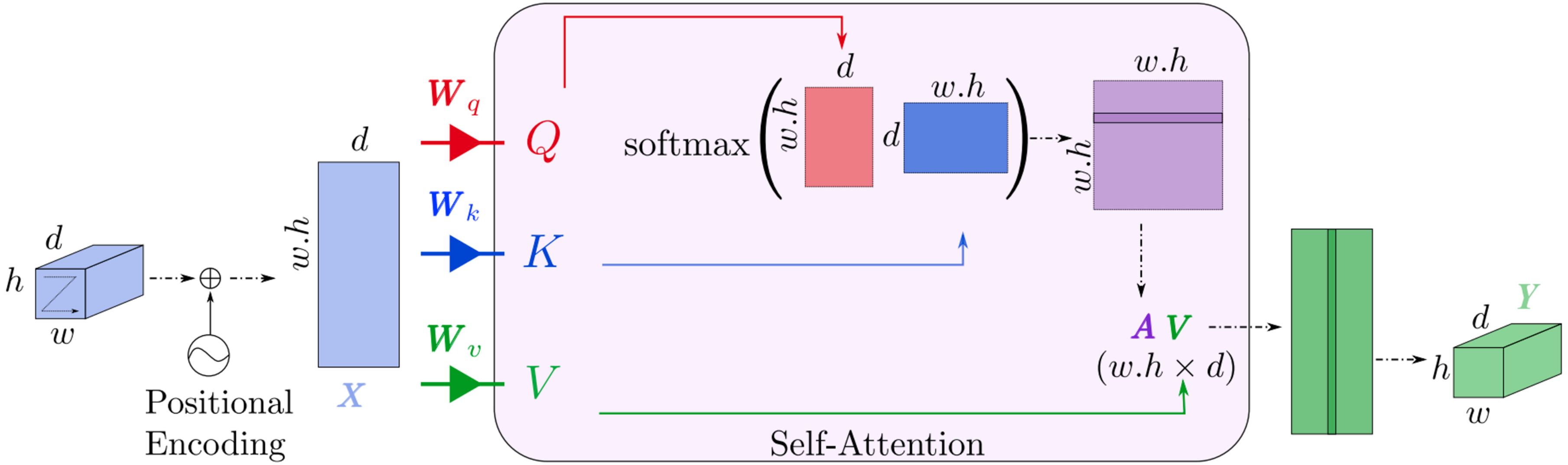
Architecture: U-Net Transformer

- Multi-Head Self-Attention (MHSA) in bottleneck (control complexity)
- Multi-Head Cross-Attention (MHCA) as an upsampling module



- Max-pooling (by 2)
- (Conv 3x3x3 + BN + ReLU)(x2)
- Conv 1x1 + BN + ReLu
- Upsampling (by 2) + Conv 3x3

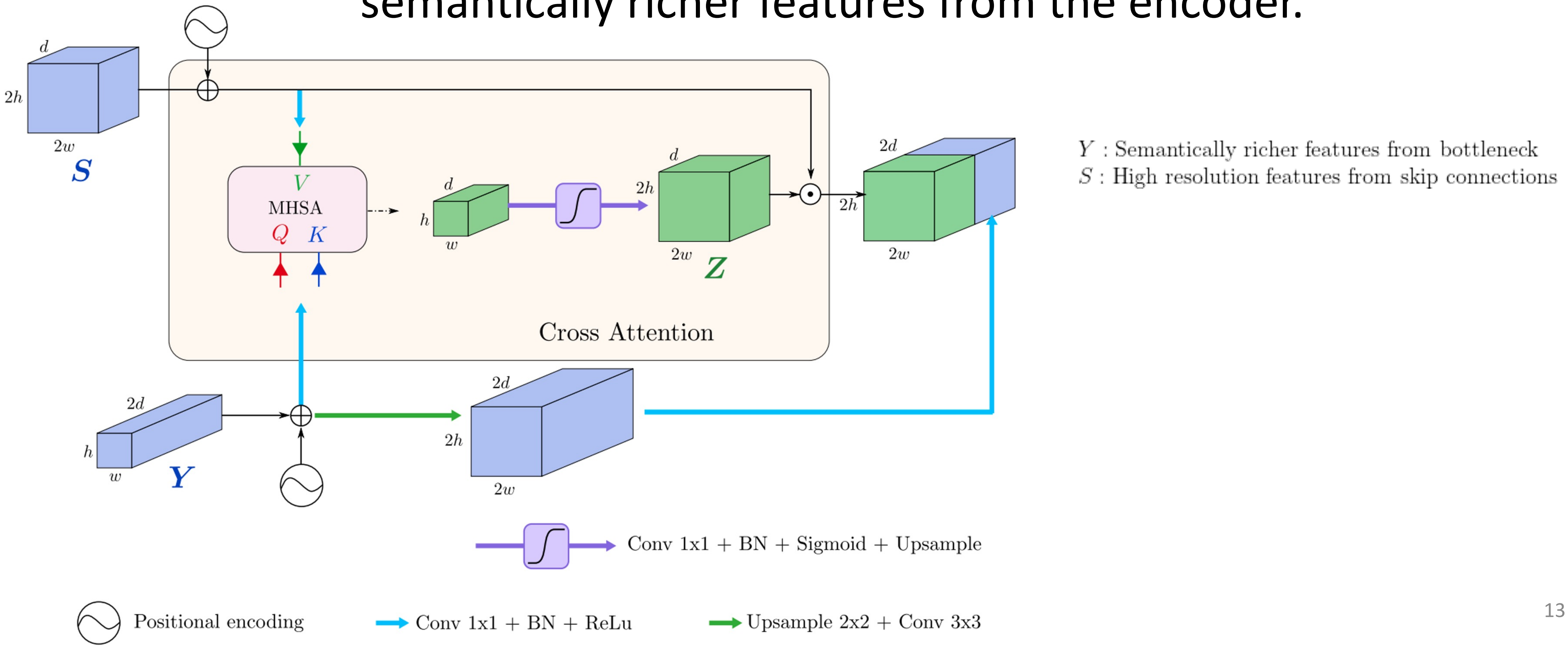
Architecture: Multi-Head Self-Attention



$$X \in \mathbb{R}^{wh \times d}, W_q \in \mathbb{R}^{d \times d}, W_k \in \mathbb{R}^{d \times d}, W_v \in \mathbb{R}^{d \times d}$$
$$Q = XW_q, K = XW_k, V = XW_v$$
$$A = Softmax(\frac{QK^T}{\sqrt{d}})$$
$$Y = AV$$

Architecture: Multi-Head Cross-Attention

MHCA : Filter high resolution features based on semantically richer features from the encoder.



Results

Experiments:

TCIA Pancreas dataset:

- Pancreas segmentation,
- 82 CT-scans
- 181~466 slices
- Size 512x512 pixels

Internal Multi-Organ dataset (IMO):

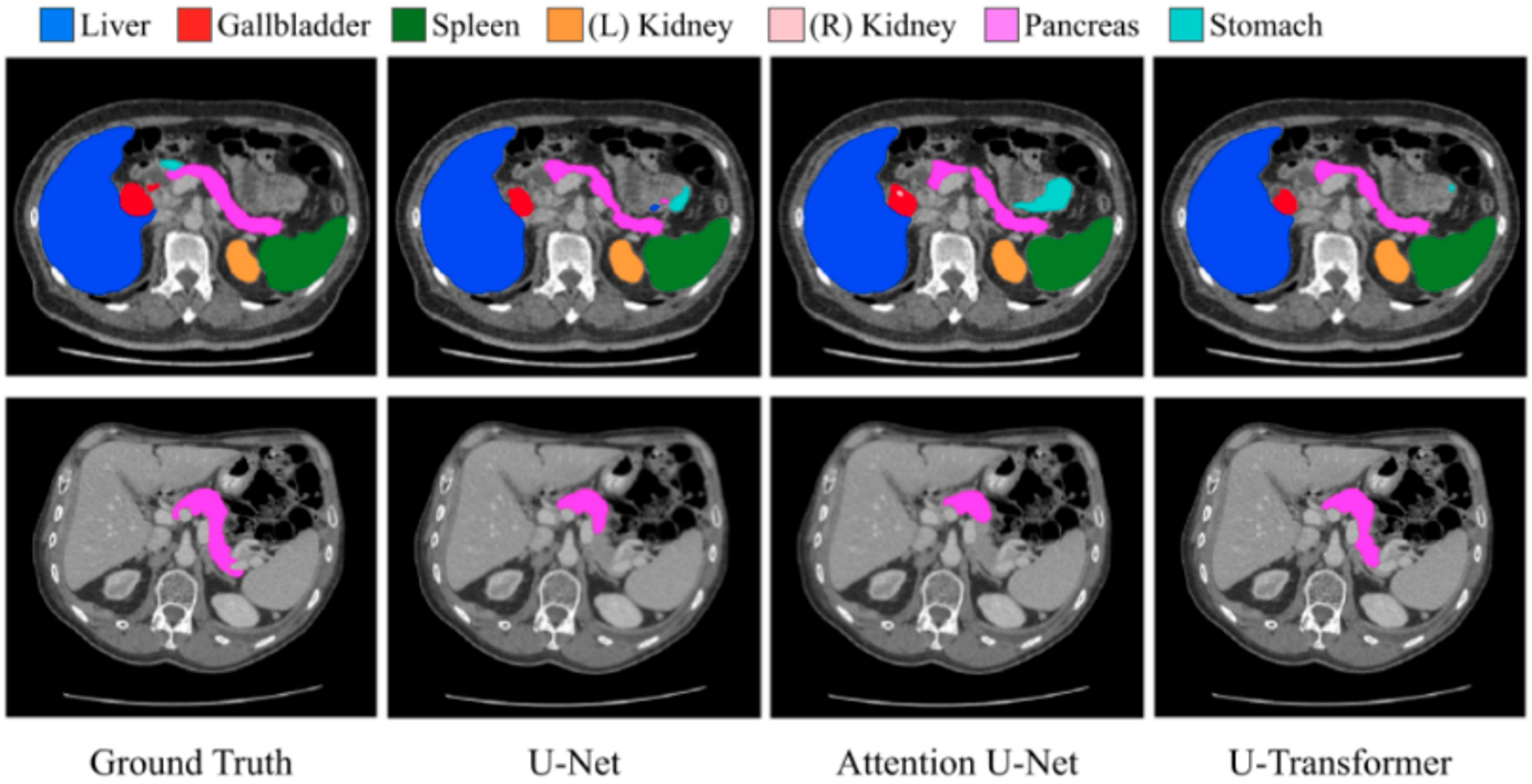
- 7 classes: liver, gallbladder, pancreas, spleen, right and left kidneys, stomach
- 85 CT-scans
- 57~500 slices
- Size 512x512 pixels

Results

Dataset	U-Net [11]	Attn U-Net [9]	MHSA	MHCA	U-Transformer
TCIA	76.13 (± 0.94)	76.82 (± 1.26)	77.71 (± 1.31)	77.84 (± 2.59)	**78.50** (± 1.92)
IMO	86.78 (± 1.72)	86.45 (± 1.69)	87.29 (± 1.34)	87.38 (± 1.53)	**88.08** (± 1.37)

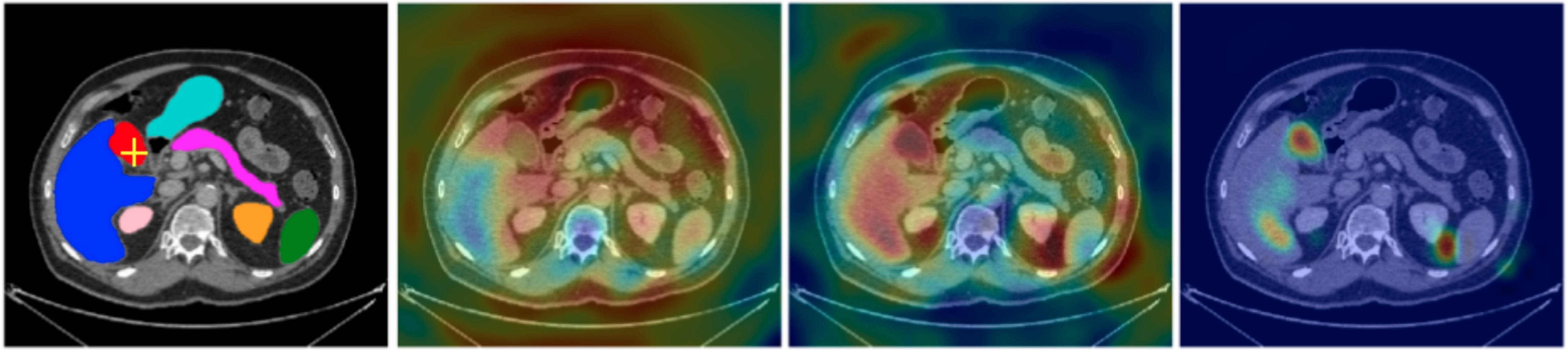
Organ	U-Net [11]	Attn U-Net [13]	MHSA	MHCA	U-Transformer
Pancreas	69.71 (± 3.74)	68.65 (± 2.95)	71.64 (± 3.01)	71.87 (± 2.97)	**73.10** (± 2.91)
Gallbladder	76.98 (± 6.60)	76.14 (± 6.98)	76.48 (± 6.12)	77.36 (± 6.22)	**78.32** (± 6.12)
Stomach	83.51 (± 4.49)	82.73 (± 4.62)	84.83 (± 3.79)	84.42 (± 4.35)	**85.73** (± 3.99)
Kidney(R)	92.36 (± 0.45)	92.88 (± 1.79)	92.91 (± 1.84)	92.98 (± 1.70)	**93.32** (± 1.74)
Kidney(L)	93.06 (± 1.68)	92.89 (± 0.64)	92.95 (± 1.30)	92.82 (± 1.06)	**93.31** (± 1.08)
Spleen	95.43 (± 1.76)	95.46 (± 1.95)	95.43 (± 2.16)	95.41 (± 2.21)	**95.74** (± 2.07)
Liver	96.40 (± 0.72)	96.41 (± 0.52)	96.82 (± 0.34)	96.79 (± 0.29)	**97.03** (± 0.31)

Results



Liver
Gallbladder
Spleen
(L) Kidney
(R) Kidney
Pancreas
Stomach
Ground Truth
U-Net
Attention U-Net
U-Transformer

Results



Ground Truth Cross-attn level 1 Cross-attn level 2 Cross-attn level 3

1. Context
2. U-Net Transformer
3. Full attention in 3D transformers

Context

Inherent problem to high-resolution volumes segmentation:

- Size of the input
- Large memory requirements
- 180Gb for U-Net with image size 512x512x256

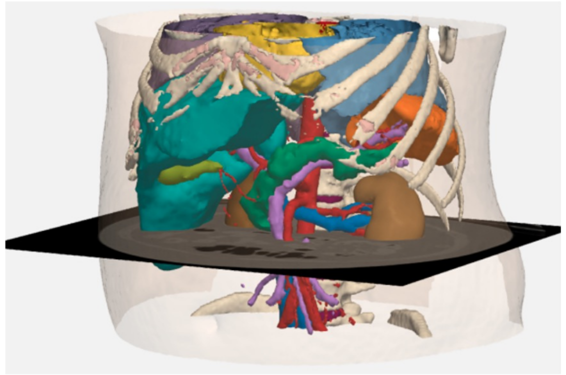
Common strategies to reduce the memory footprint:

- Limited model size
- Train on 2D slices
- Train on patches

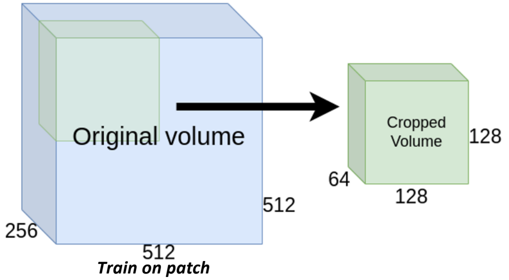
⇒ No long-range information

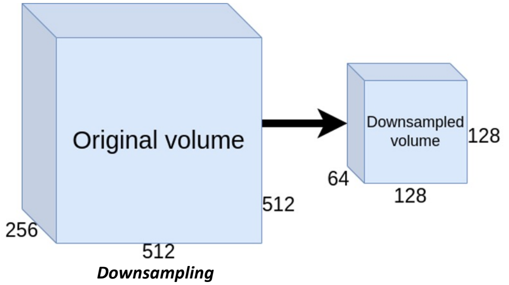
- Downsampling

⇒ Drop in quality



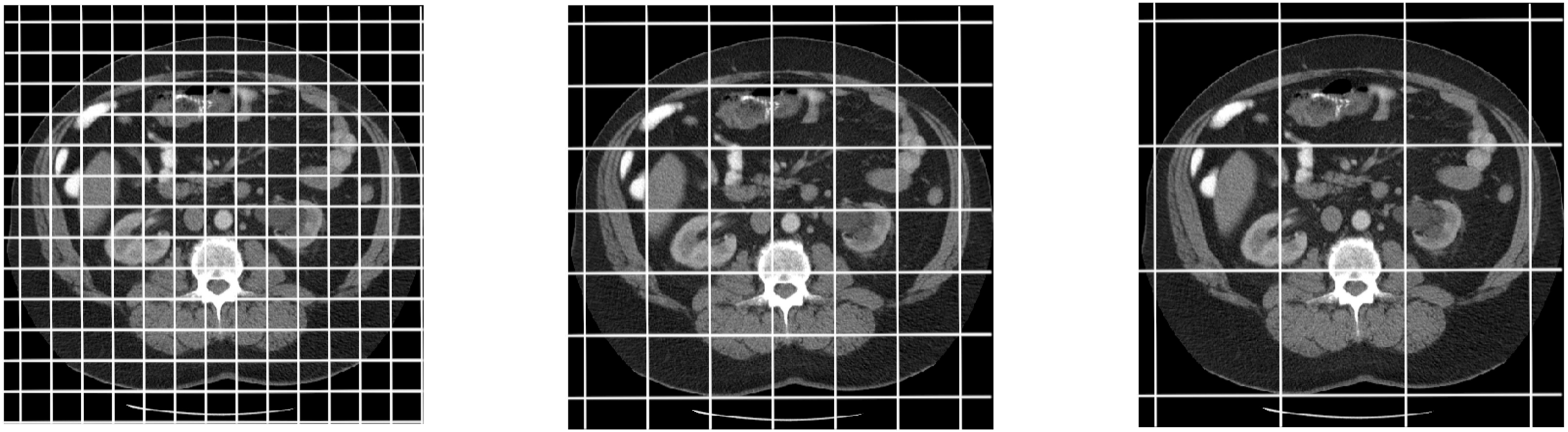
Organs segmentation illustration





Transformers for medical image segmentation

- **Transformers** became SOTA for **image segmentation [B]**
  - **BUT: not possible to have full attention at high-recolution feature maps**
- **Windowed transformers [C,D]** designed to reduce the complexity
  - **BUT:** no more long-range attention for high resolution feature maps



Windowed input at different hierarchy levels

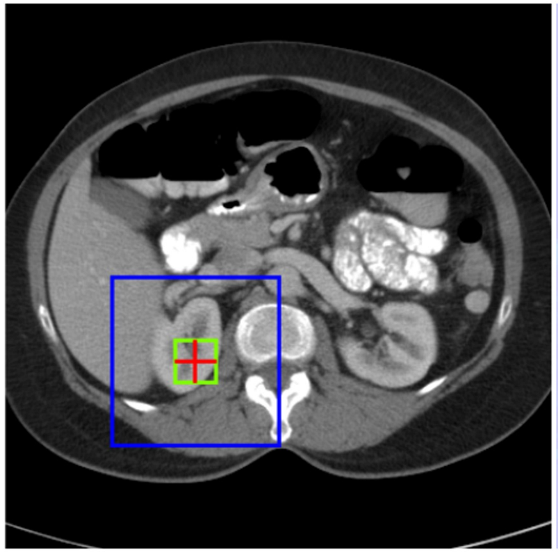
[B] Y. Xie, J. Zhang, C. Shen,, D., Lu, T., Luo, P., Shao, L.: Pyramid vision and Y. Xia. Cotr : Efficiently bridging CNN and transformer for 3d medical image segmentation.CoRR, abs/2103.03024, 2021.
[C] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer : Hierarchical vision transformer using shifted windows.CoRR, abs/2103.14030, 2021.
[D] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE ICCV (2021)

Motivation

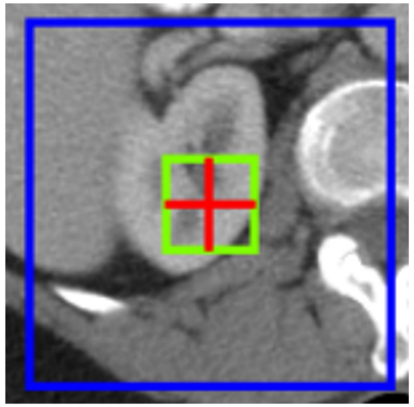
To keep the **full resolution** we work on patches:

Original image size: **512x512x256**

Cropped patch size: **128x128x64**



Input image 2D slice



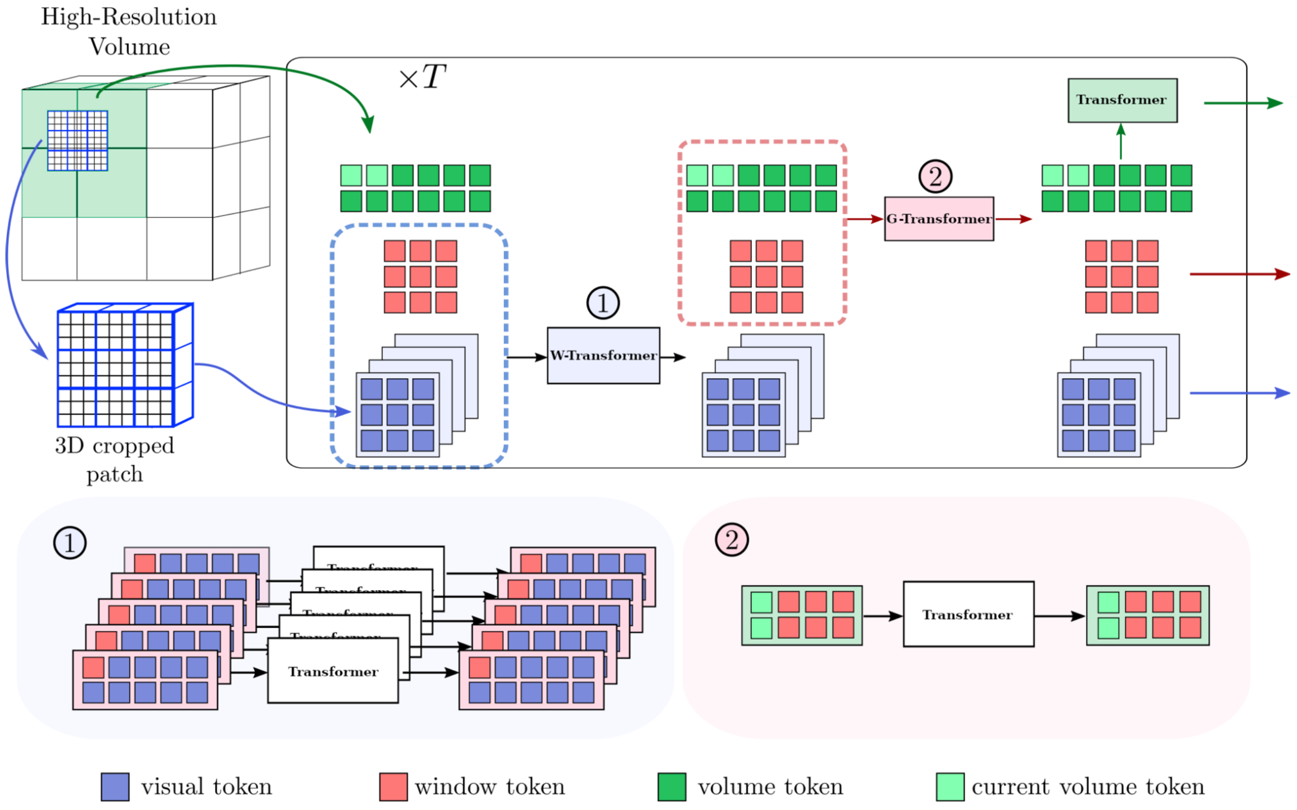
Cropped patch 2D slice

Goal: learning a global representation of the full volume from batch training with crops

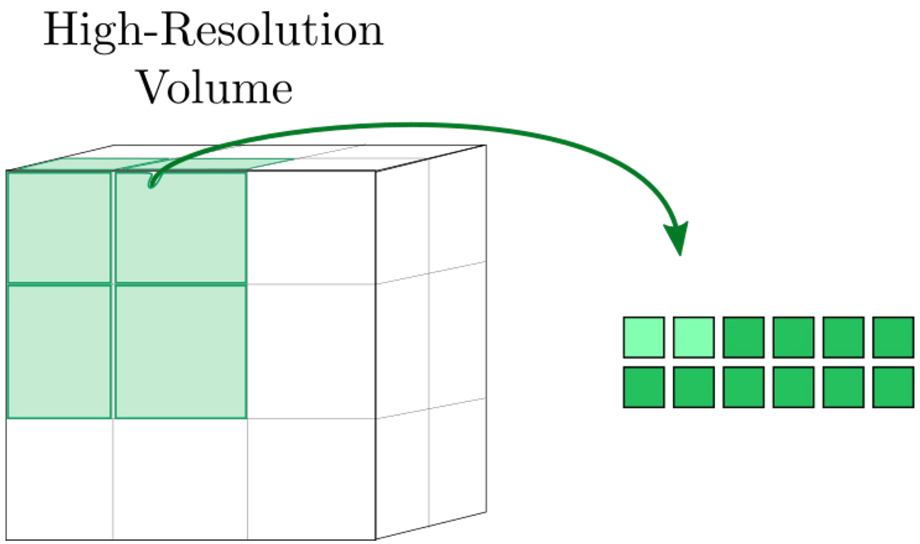
FINE : Full resolutIoN mEmory transformer module

- Global representation embedded in **multiple level of memory tokens**
- **Visual tokens:** high-resolution 3D features (4x4x4)

- **Window token:** information at the window scale
- **Volume token: learned** representation of the full-size volume

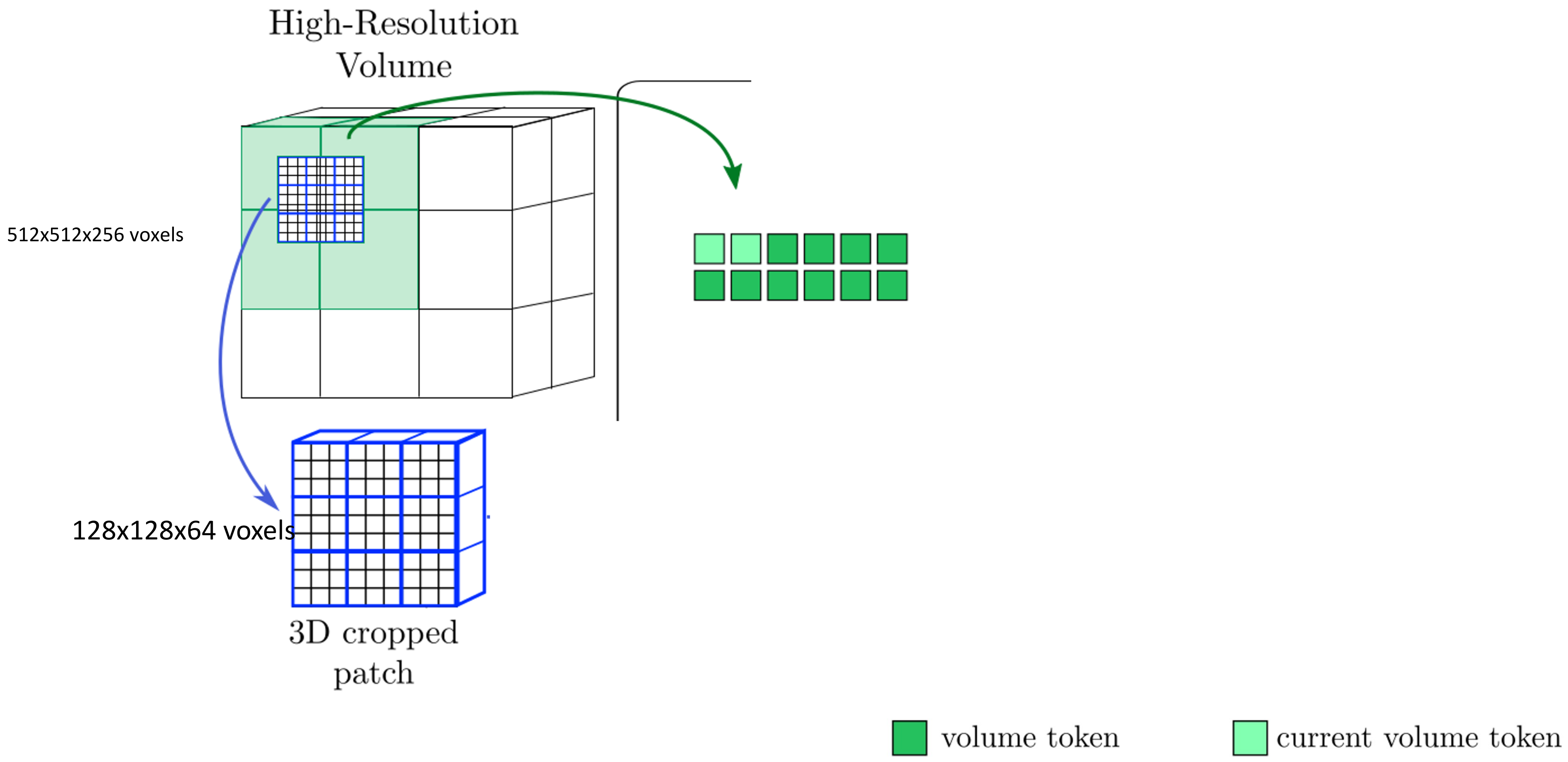


Volume tokens



volume token   current volume token

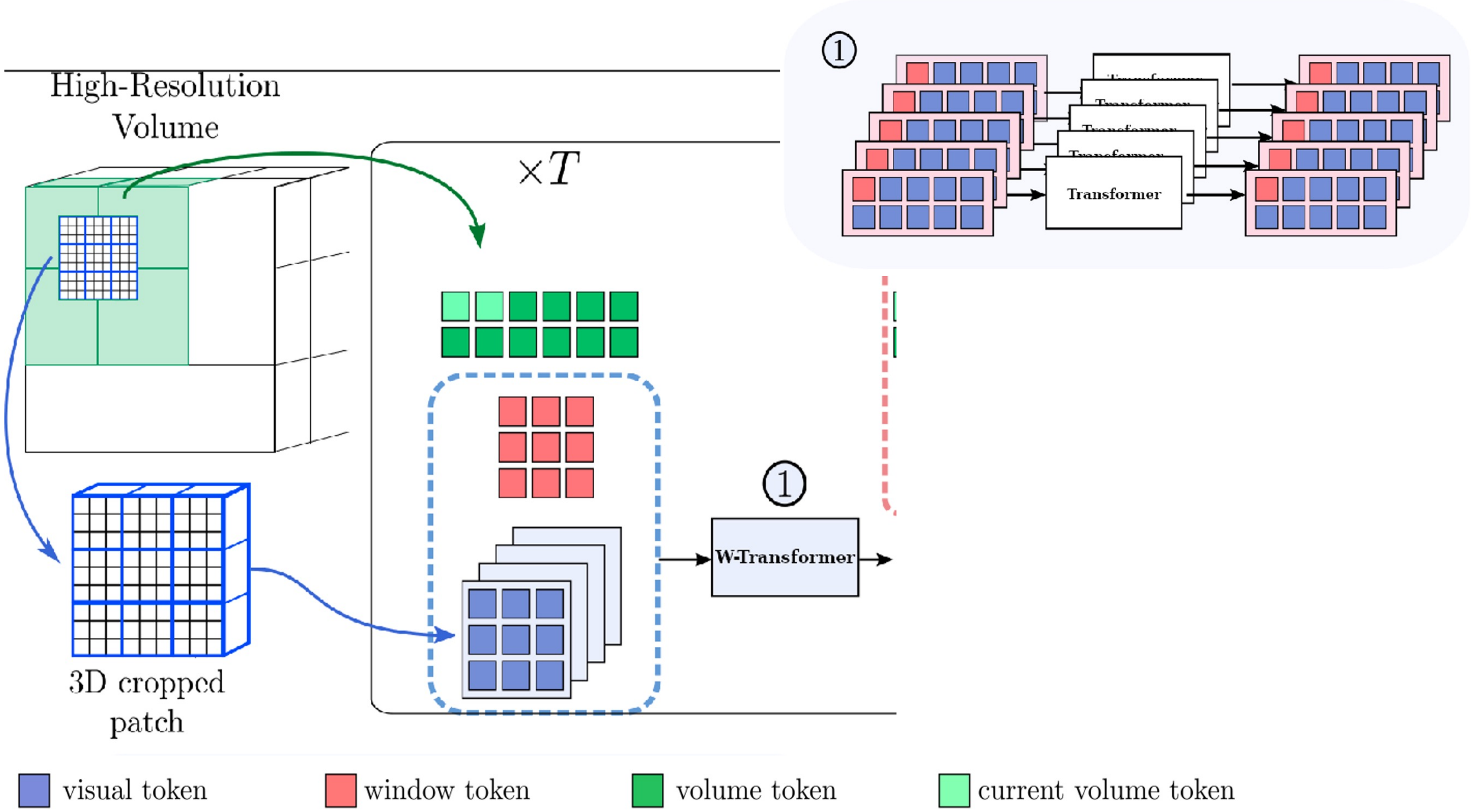
Cropped patch



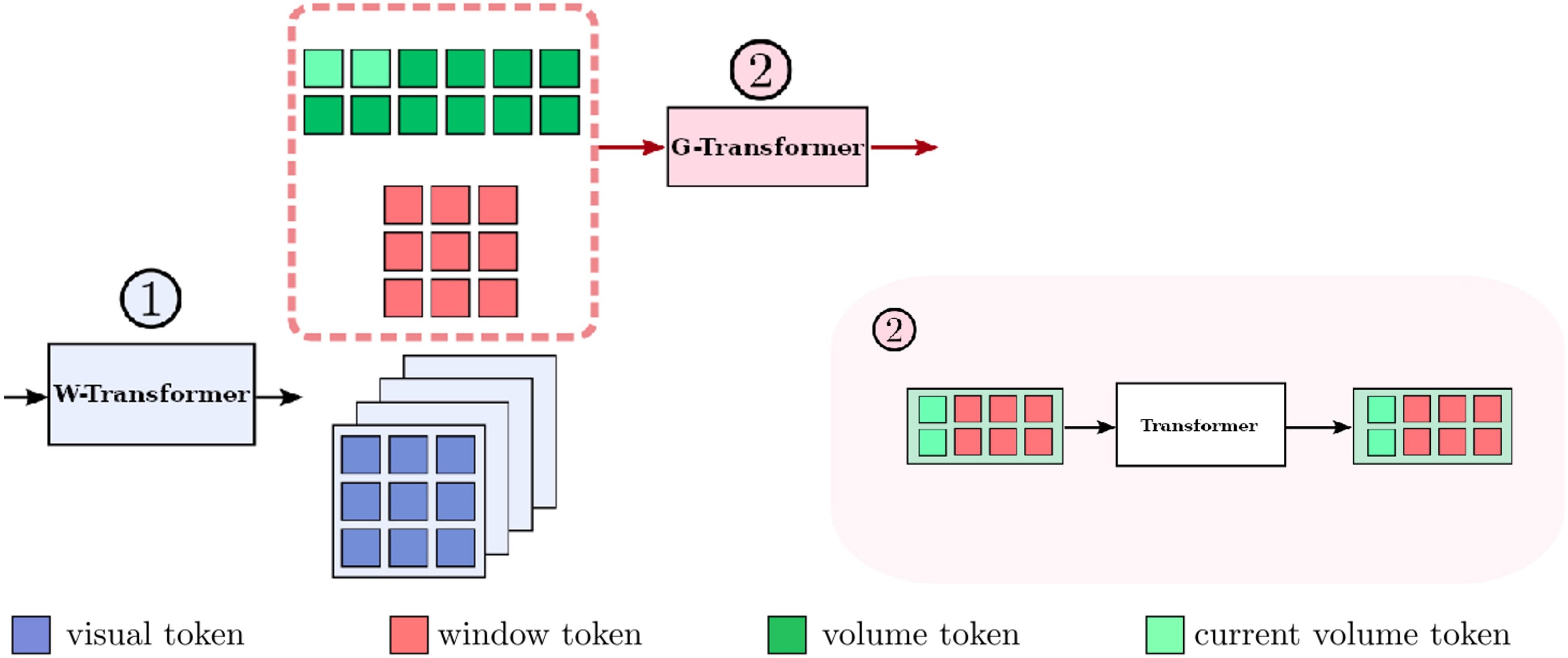
Local window attention

512x512x256 voxels

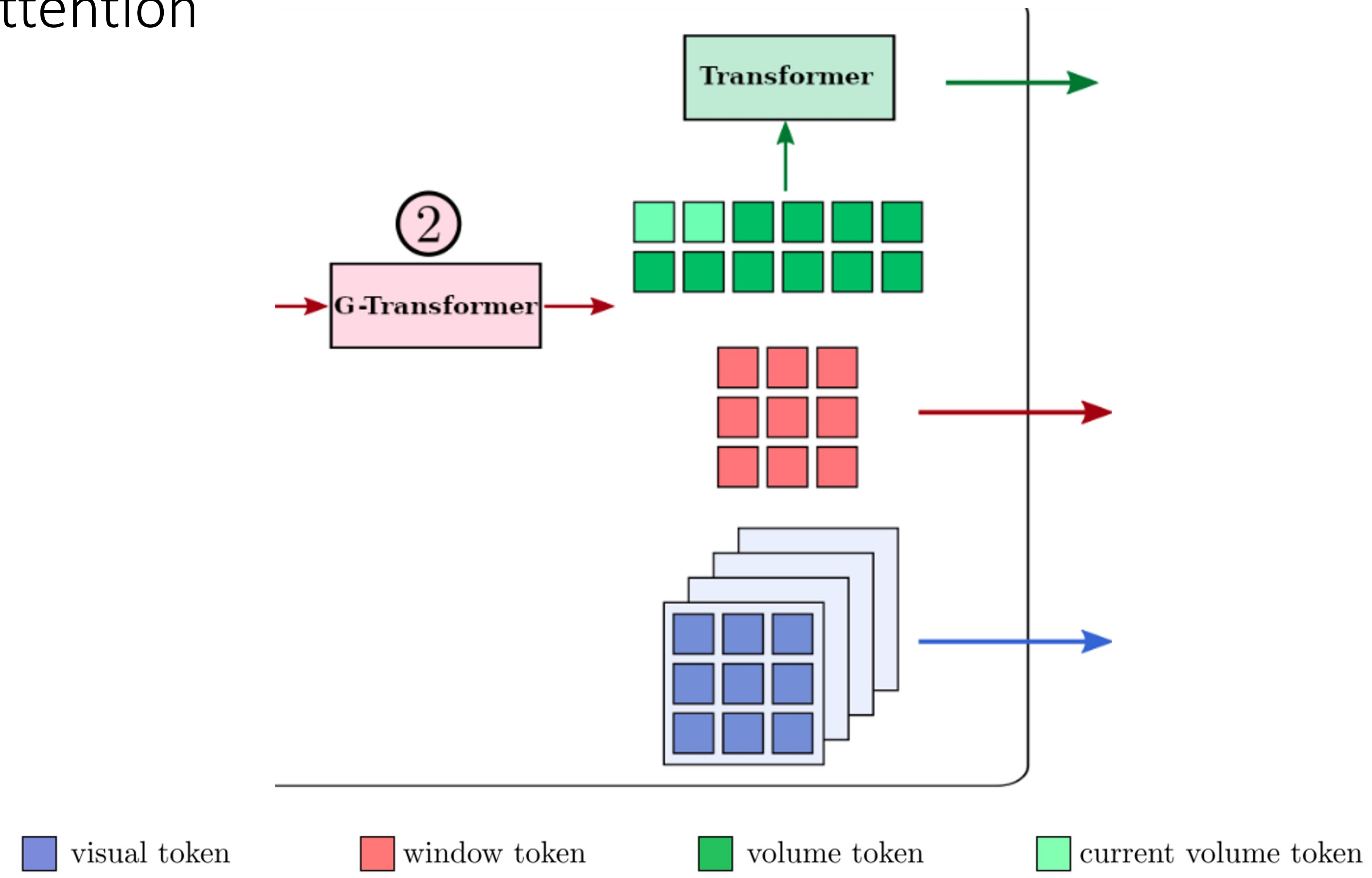
128x128x64 voxels



Cross-window attention



Volume attention



Results

Synapse BCV [17] : CT scans Abdominal multi-organs segmentation

7 classes

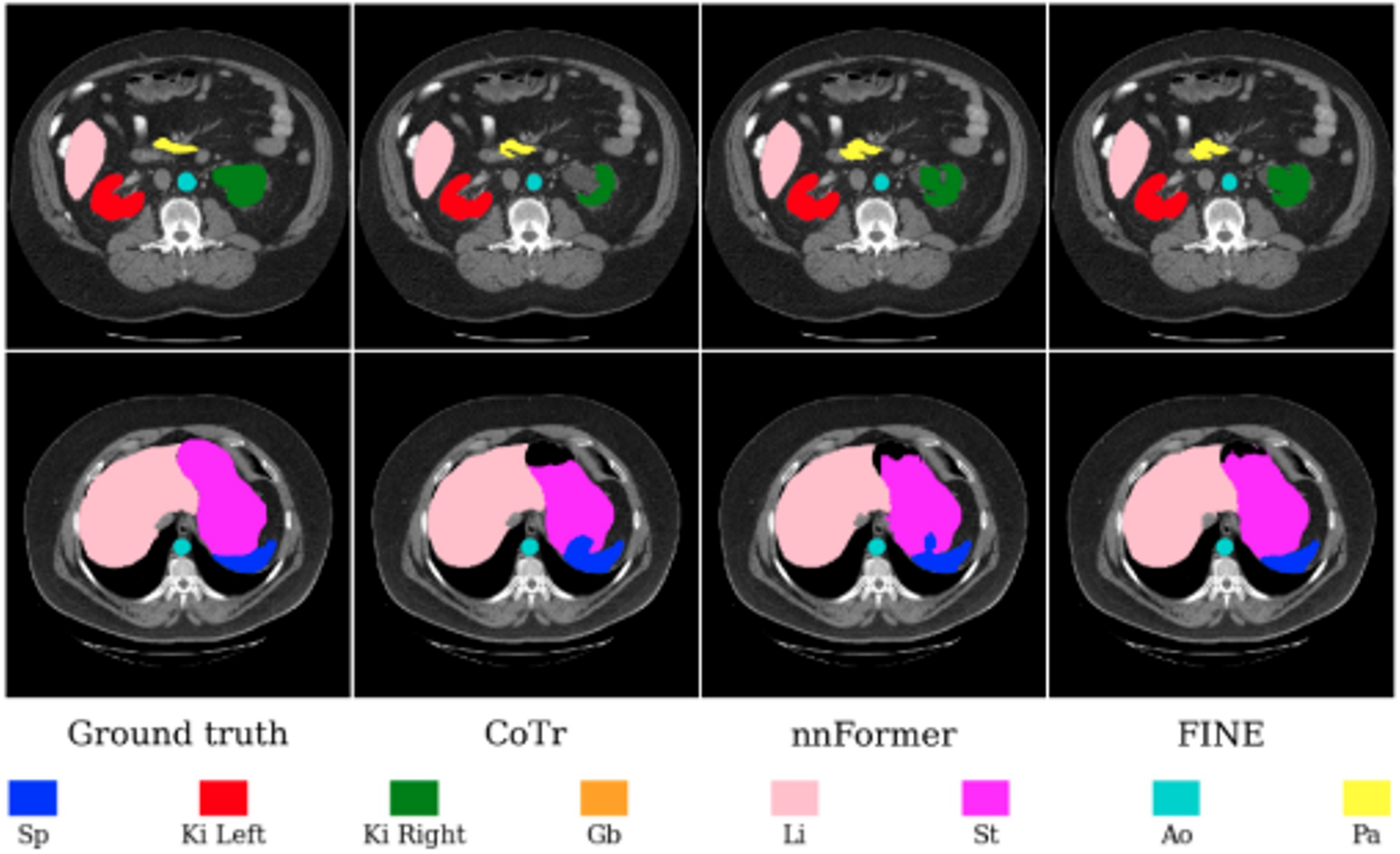
30 volumes

Metrics :

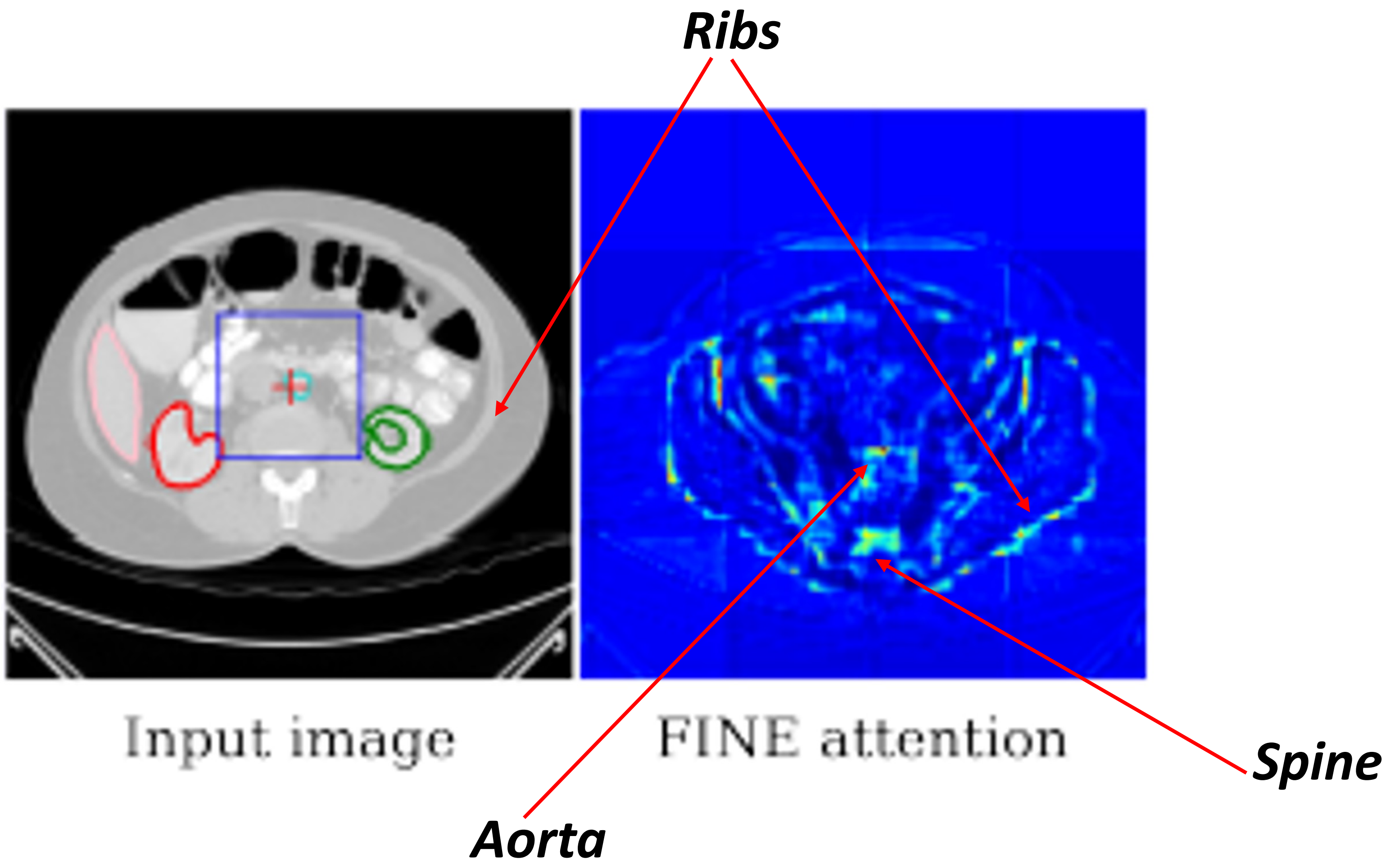
- Dice score in % (DSC)
- 95% Hausdorff distance in mm (HD95)

Method	Average		Per organ dice score (%)						
	HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
UNet [24]	-	77.4	86.7	73.2	69.7	93.4	75.6	89.1	54.0
AttUNet [19]	-	78.3	87.3	74.6	68.9	93.6	75.8	89.6	58.0
VNet [18]	-	67.4	80.6	78.9	51.9	87.8	57.0	75.3	40.0
Swin-UNet [3]	21.6	78.8	90.7	81.4	66.5	94.3	76.6	85.5	56.6
nnUNet [10]	10.5	87.0	91.9	86.9	**71.8**	**97.2**	85.3	**93.0**	**83.0**
TransUNet [4]	31.7	84.3	88.8	84.9	72.0	95.5	84.2	90.7	74.0
UNETR [8]	23.0	78.8	87.8	85.2	60.6	94.5	74.0	90.0	59.2
CoTr* [31]	11.1	85.7	93.4	86.7	66.8	96.6	83.0	92.6	80.6
nnFormer [33]	9.9	86.6	90.5	86.4	70.2	96.8	86.8	92.0	83.3
FINE*	**9.2**	**87.1**	**95.5**	**87.4**	66.5	97.0	**89.5**	91.3	82.5

Results


Ground truth
CoTr
nnFormer
FINE
Sp
Ki Left
Ki Right
Gb
Li
St
Ao
Pa

Results



Conclusion

U-Net Transformer : MHSA + MHCA

- Combination of powerful image segmentation models (U-Net) with long-range interaction model (Transformers)
- MHSA + MCASLong range interaction and spatial dependencies

FINE :

- Generic module for any windowed transformers
- Able to model **long-range** interaction beyond cropped input patch
- Perspective: hybrid FINE module in Conv architectures (nnU-Net)

Thank you