# Transformers for medical image segmentation


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

# Transformers everywhere since 2017

**NLP: BERT, GPT-3/4, Chat-GPT, *etc***

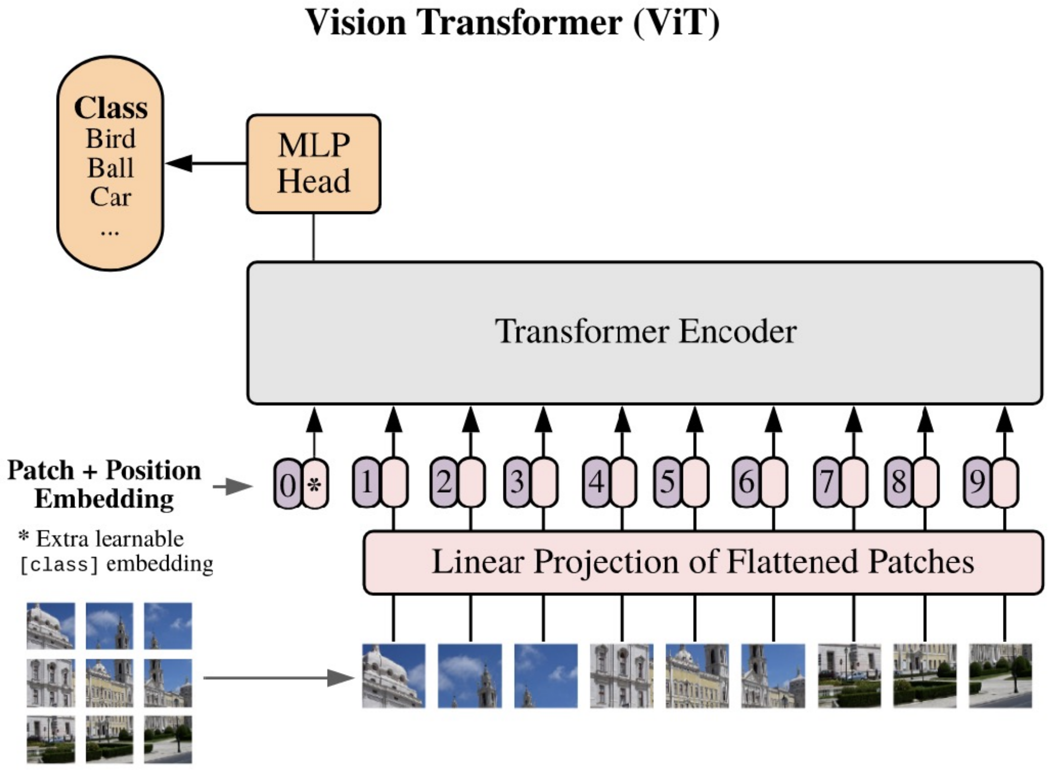
explain quantum theory to a child in the style of snoop dogg

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.

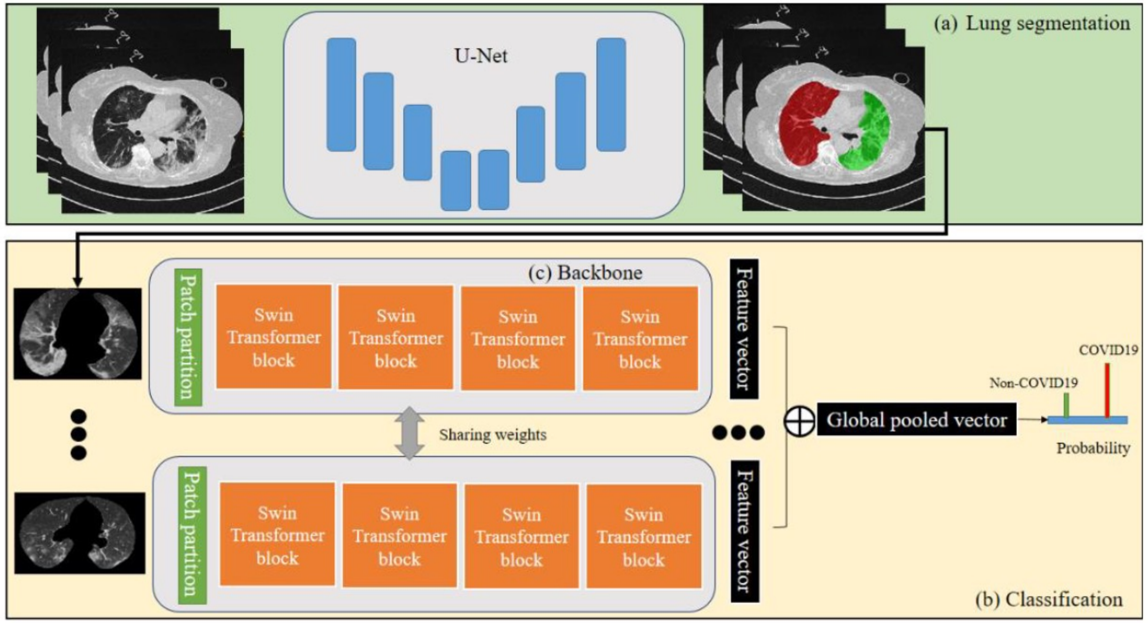
**Vision since '21: Vision Image Transformer (ViT)**



# Transformer in medical image analysis

**Used in various contexts and tasks**

- Image classification, detection, *e.g.* COVID, Semantic segmentation
- Image Registration
- Image Generation
- Im-2-im translation

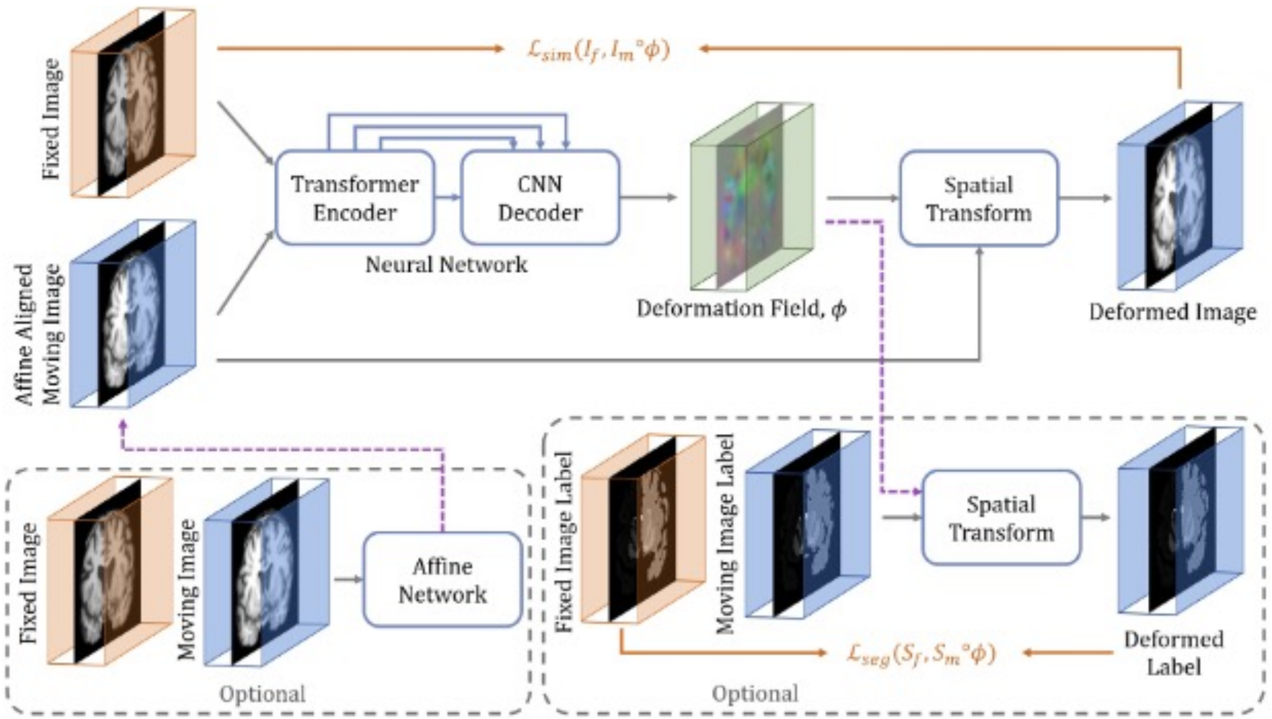



Zhang L, Wen Y. Mia-cov19d: A transformer-based framework for covid19 classification in chest cts. arXiv, 2021.

# Transformer in medical image analysis

**Used in various contexts and tasks**

- Image classification, detection, *e.g.* COVID, Semantic segmentation
- Image Registration
- Image Generation
- Im-2-im translation

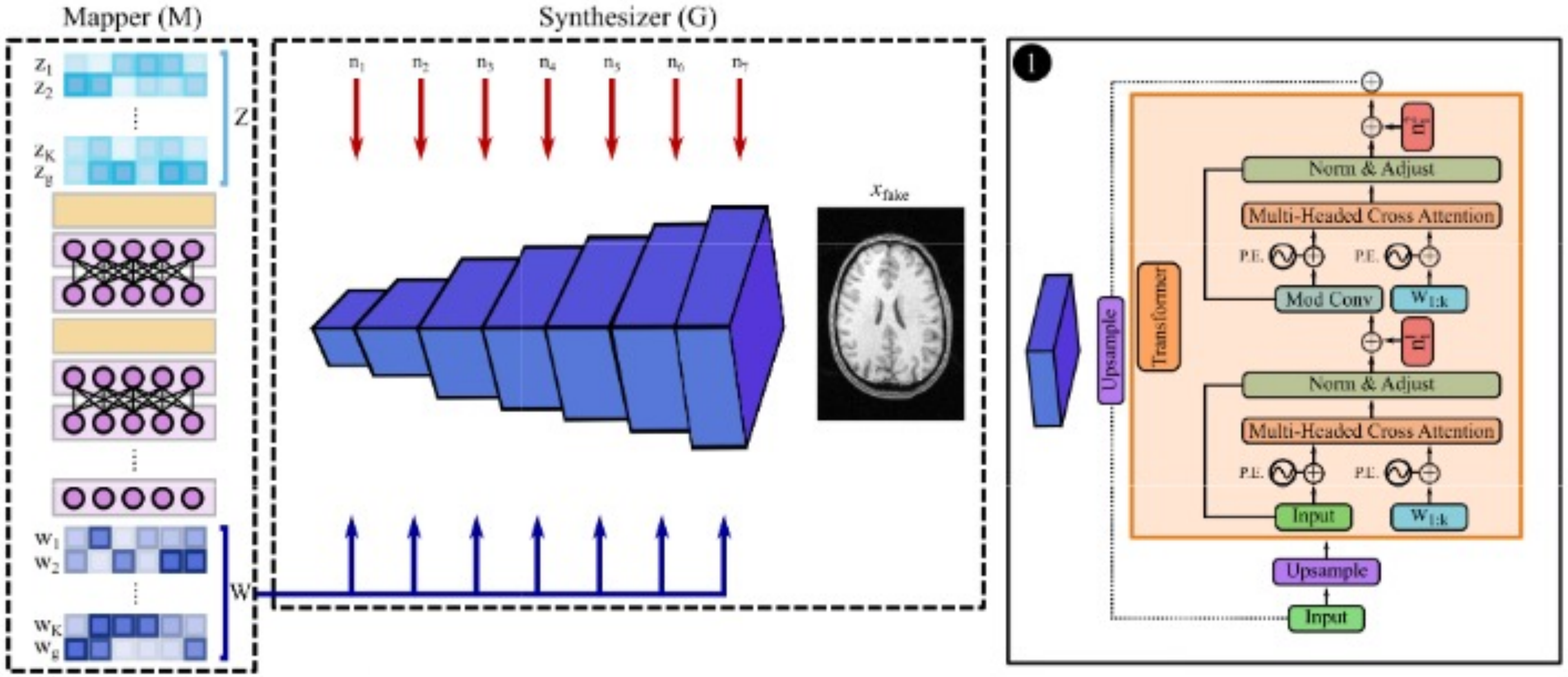



Chen J, Du Y, He Y, et al. Transmorph: Transformer for unsupervised medical image registration. Medical Image Analysis, 2022.

# Transformer in medical image analysis

**Used in various contexts and tasks**

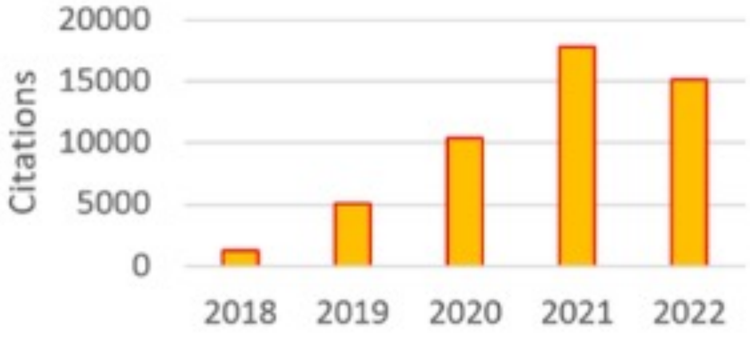
- Image classification, detection, *e.g.* COVID, semantic segmentation
- Image Registration
- Image Generation
- Im-2-im translation




Korkmaz Y, Dar SU, Yurt M, et al. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. IEEE TMI, VOL. 41, NO. 7, JULY 2022

# Focus on this talk

- Paper on transformer every day…



(a) Citations of Transformer papers in recent years

(b) Number of papers published in the last 12 months that contain "Action Recognition" + ("Transformer" OR "Attention") in their titles

- By no means exhaustive literature review


Intelligent Medicine 3 (2023) 59–78

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

ELSEVIER

Review

Transformers in medical image analysis

Kelei He [1,2,#], Chen Gan [2,#], Zhuoyuan Li [1,2,#], Islem Rekik [3,4,#], Zihao Yin [2], Wen Ji [2], Yang Gao [2,5], Qian Wang [6,*], Junfeng Zhang [1,2,*], Dinggang Shen [6,7,8,*]

# Focus on this talk

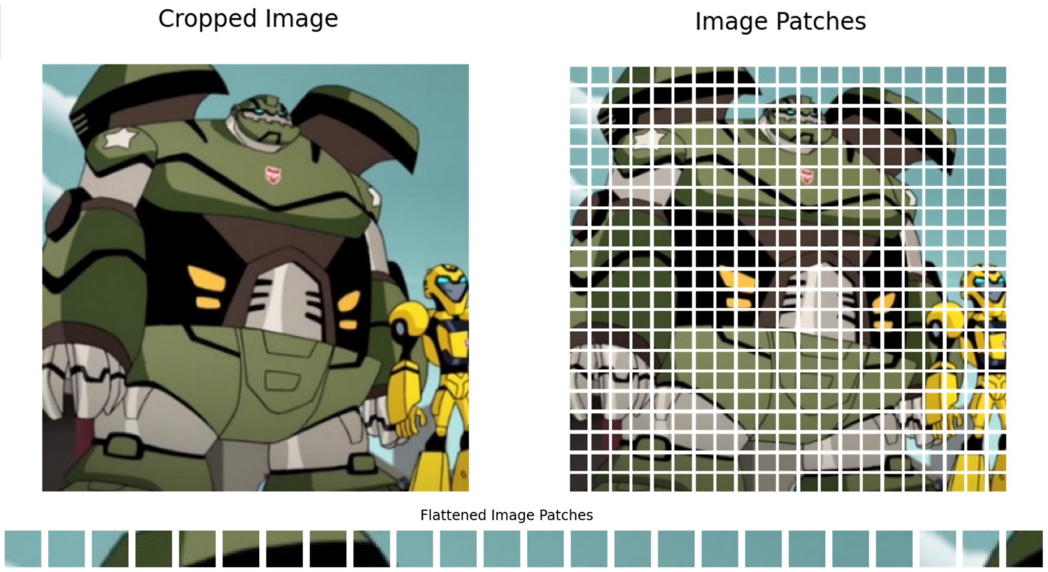
1. Transformers
2. Vision Image Transformer
3. Transformers for medical image segmentation
4. Current trend & Perspectives

Architecture: main features and processing

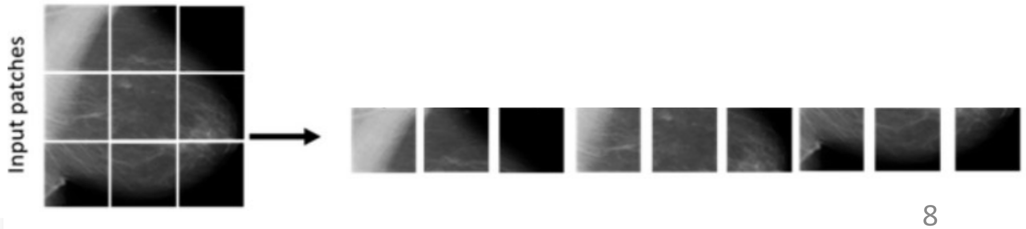
Long-range interactions
Efficient self-attention

# From sequence to set

- A sequence of elements → a **set** of tokens, no order
  - Token: primitives, elementary elements of data
    - Text: token are e.g. words
    - Image: token are e.g. patches




| Text | | Tokenization |
|---|---|---|
| | | "Hello", "I", "love", "you" |
| "Hello I love you" | ⇒ | "love", "Hello", "I", "you" |
| | | "I", "love", "Hello", "you" |

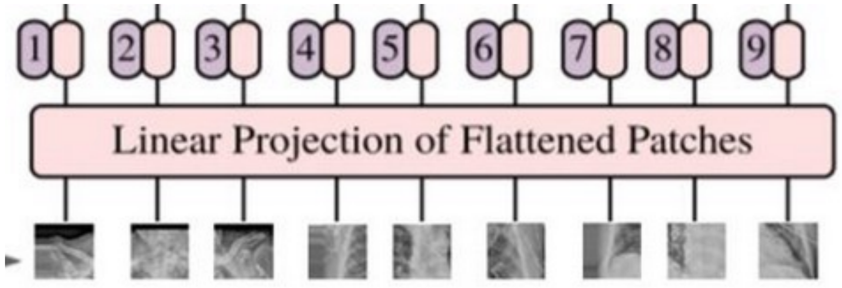


# Input embedding

- Token: input vector in $\mathbb{R}^t$
  - Word: t =|V|, V vocabulary
  - Image patch: t = $s^2$, where s is the patch size
- Input embedding: linear projection $\mathbb{R}^t \rightarrow \mathbb{R}^d$ : $e_i = x_i W^e$

$\mathbb{R}^d$

$\mathbb{R}^t$

# Positional encoding

- Sequence → set of token:
  - Permutation invariant
  - Loosing structural information from data
- Recovering structure: **positional encoding (PE)**
  - Mapping token position t to a vector $\mathbf{p}_t \in \mathbb{R}^d$
  - Seminal PE: sinusoidal

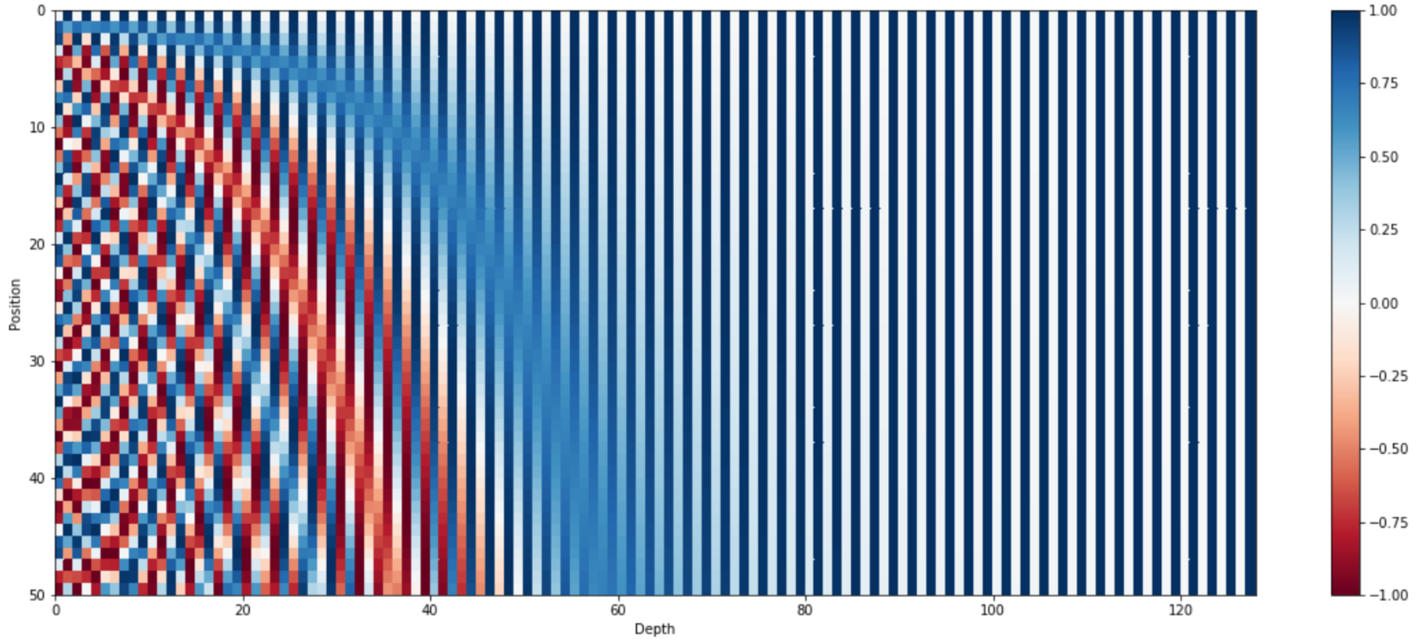
$$\vec{p}^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

$$\vec{p} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

# Sinusoidal positional encoding

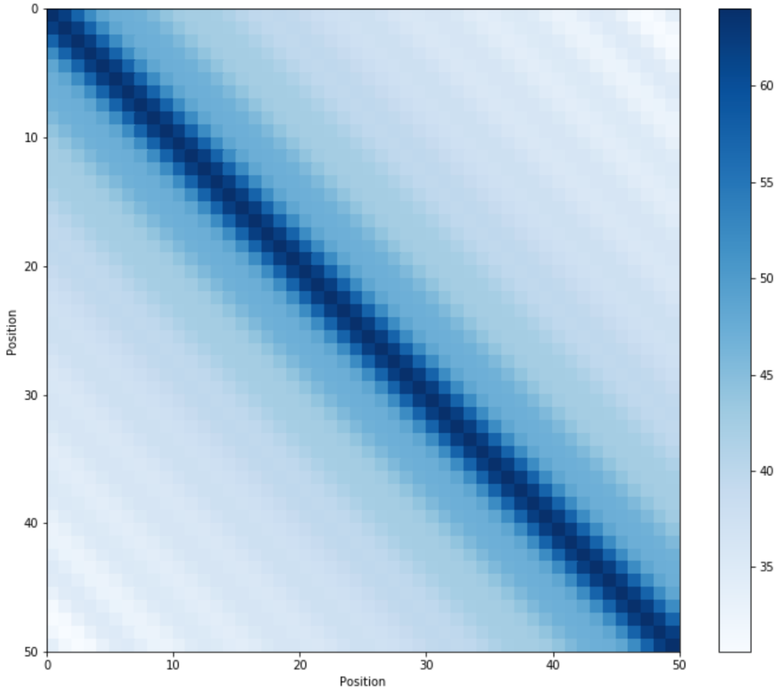
- Unique vector $\mathbf{p}_t$ for each position t
- $p_t(i) \in [-1;1]$: natural normalization




d=128, max length of token set = 50

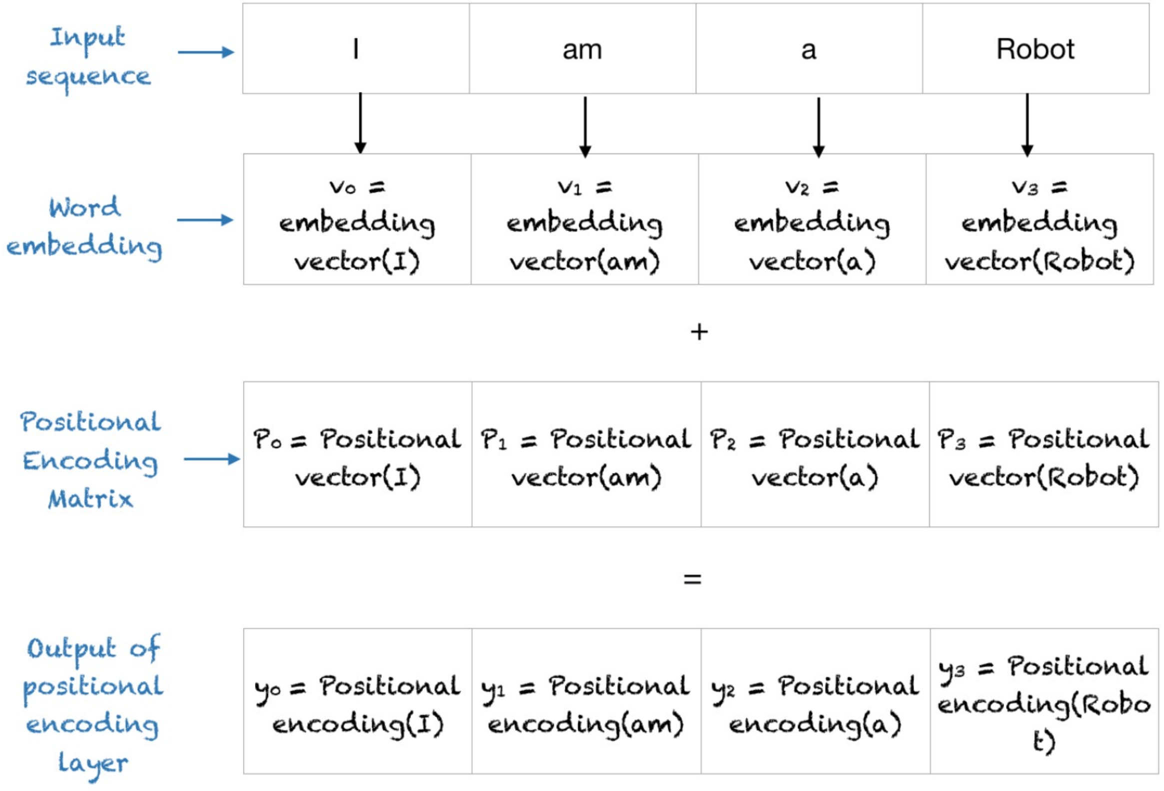
- Models relative position
- Positional similarity:

$$K = PPt$$



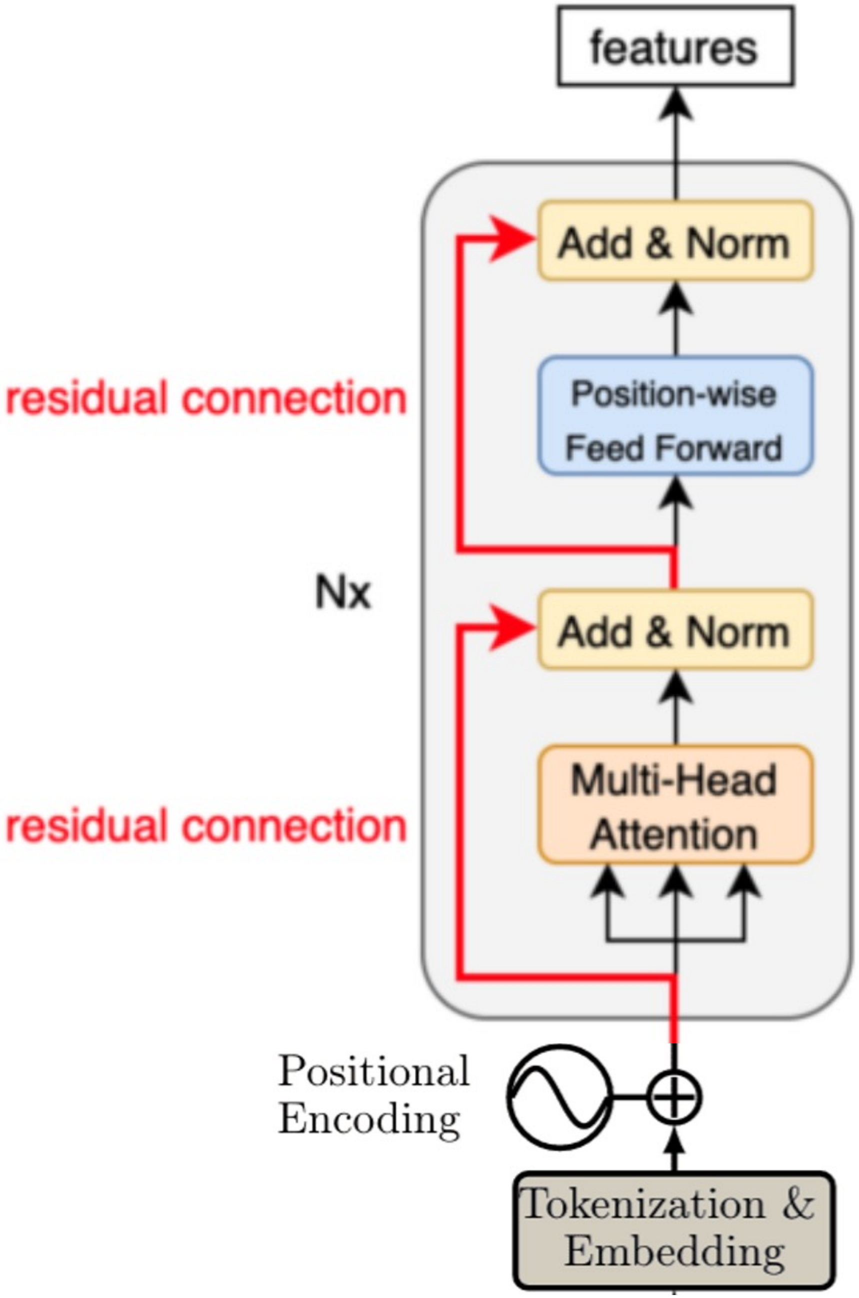
# Positional encoding

- Other possible encoding, can be learned
- Final embedding :




**=> Input of transformer!**

# Transformer [1] : the encoder

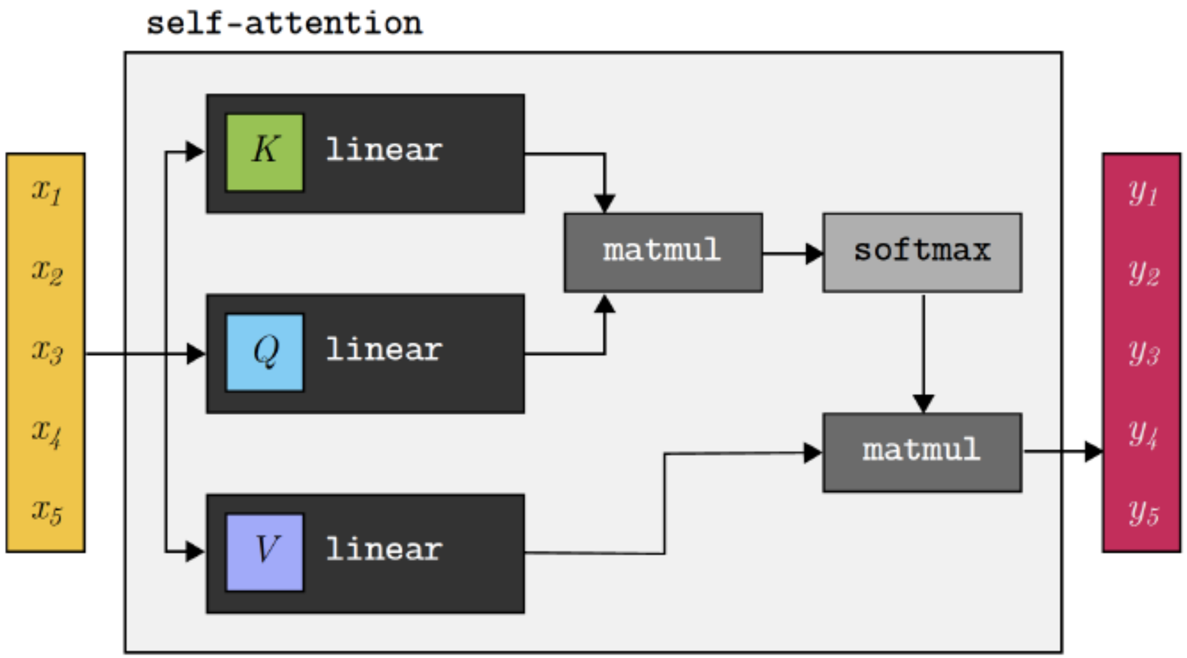


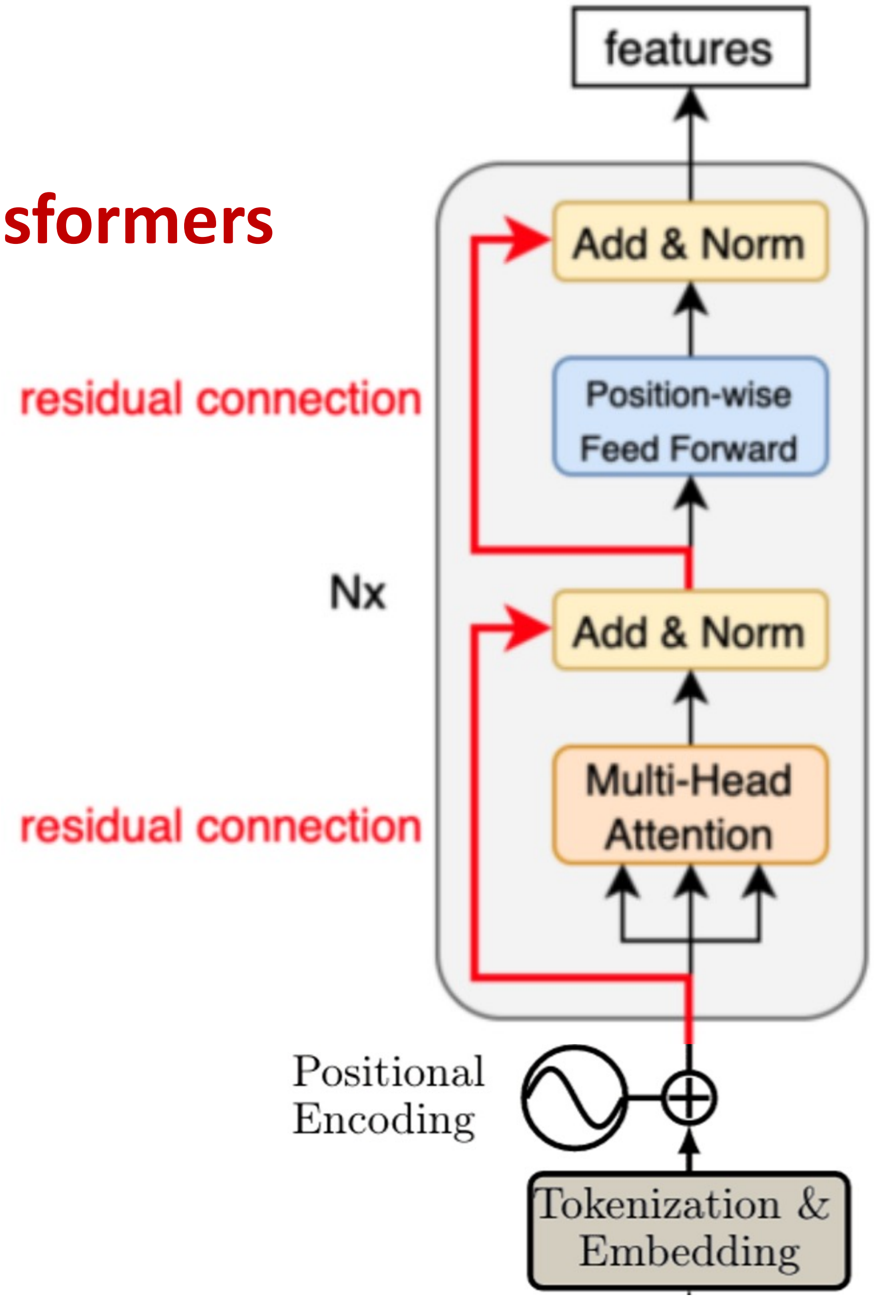

- A stack a N transformer blocks
  - Input a set of embedded tokens
  - Output: a set of re-embedded tokens

[1] Attention Is All You Need. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. NeurIPS 2017.

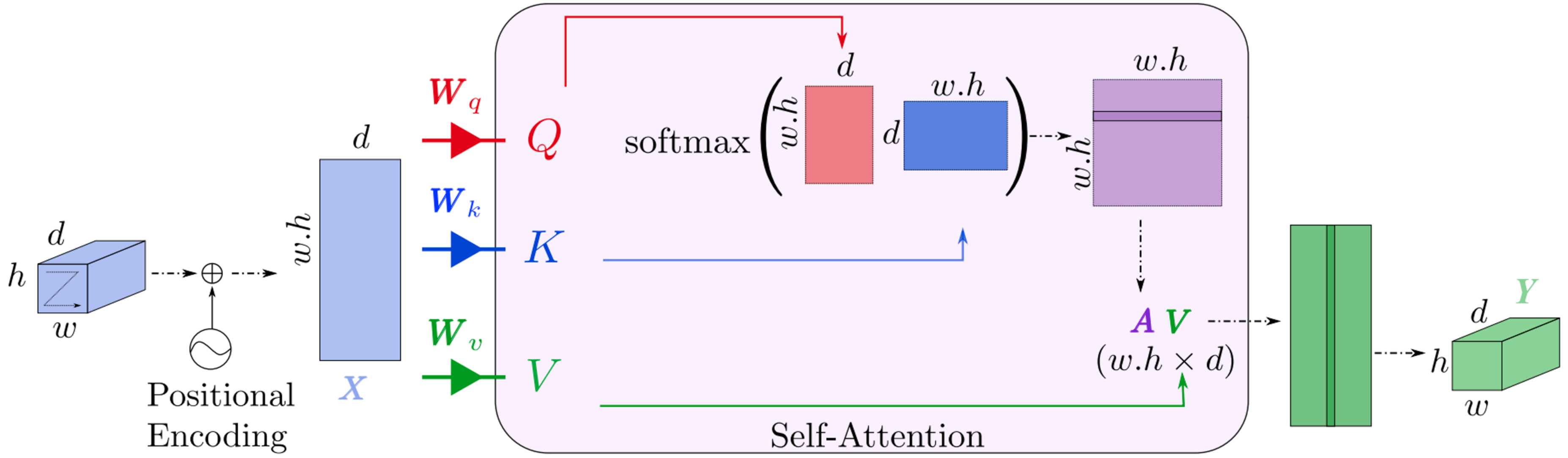
# Transformer: self attention

- **The most important and specific module in transformers**
- Project the input set into 3 sets
  - Query: sought info
  - Key:  context elements
  - Value: retrieved



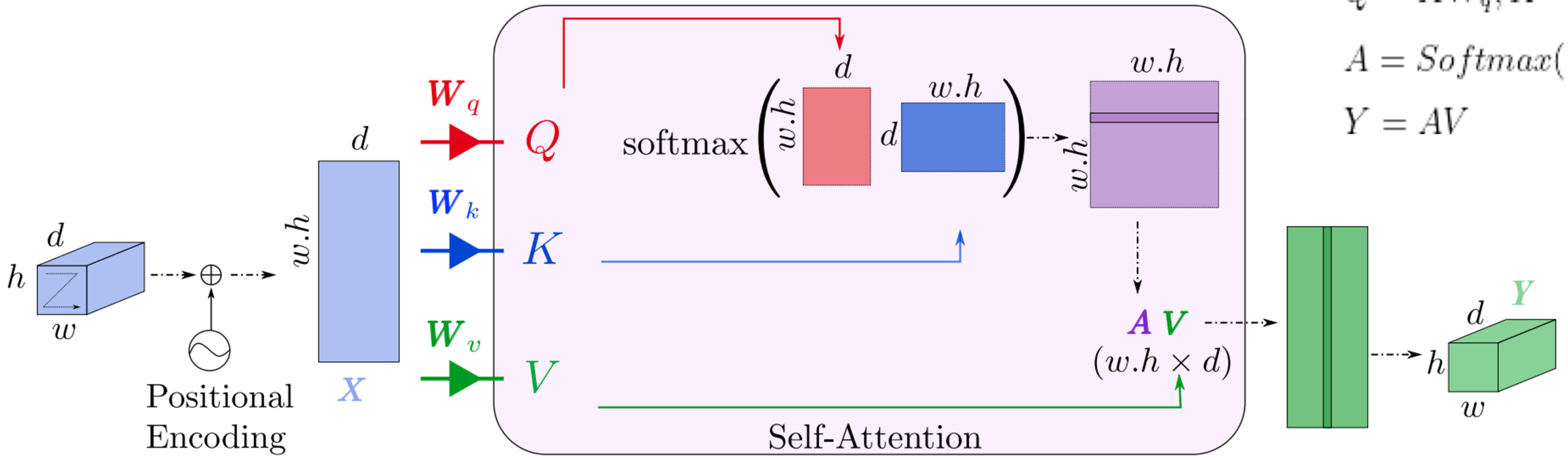



# Self-attention




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

# Self-attention: conclusion




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

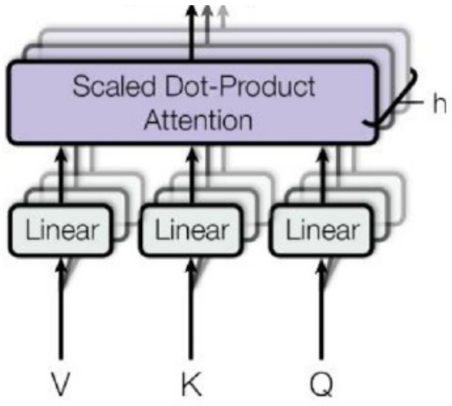
- Each token $y_i$ in Y: computed a linear combination of $v_i$
  - Enables to model **<u>global interactions</u>** between $v_i$ tokens: full contextual information
  - ≠ ConvNets in vision, interactions limited by the size of the receptive field
  - ≠ RNNs for sequence processing, interactions limited by vanishing gradients
- **Self attention: $O(N^2)$ complexity**
  - Expensive (or impossible) for large N

# Multi-headed attention

- High-level idea: multiple self-attention in parallel

- Each head: attend to different parts

- Combine the heads' outputs

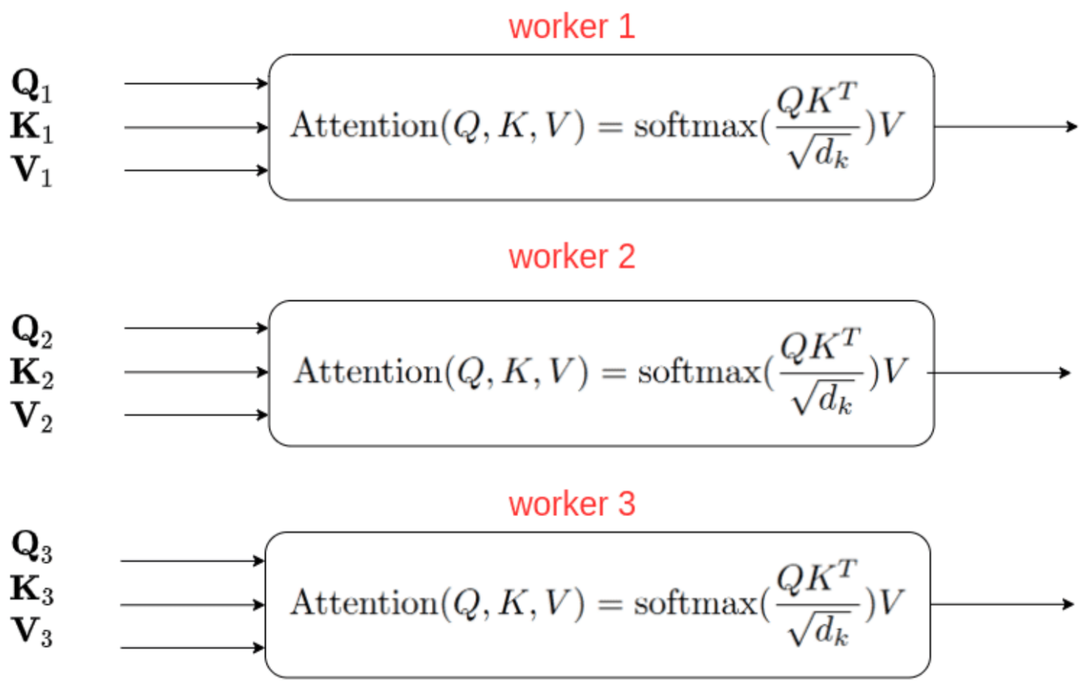


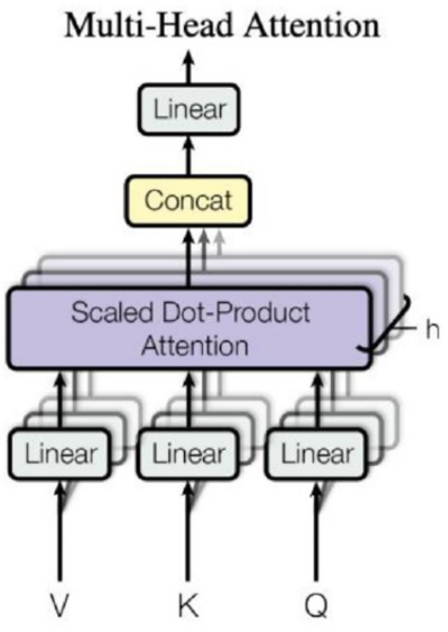

[Vaswani et al. 2017]

Wizards of the Coast, Artist: Todd Lockwood

**Credit:** Anna Goldie

# Multi-headed attention



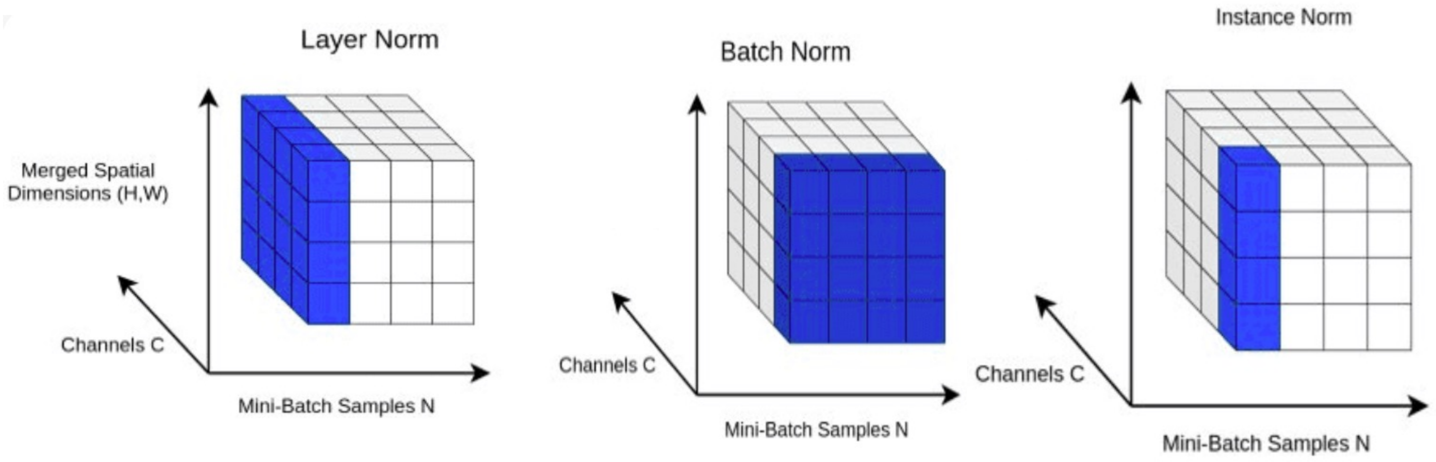




[Vaswani et al. 2017]

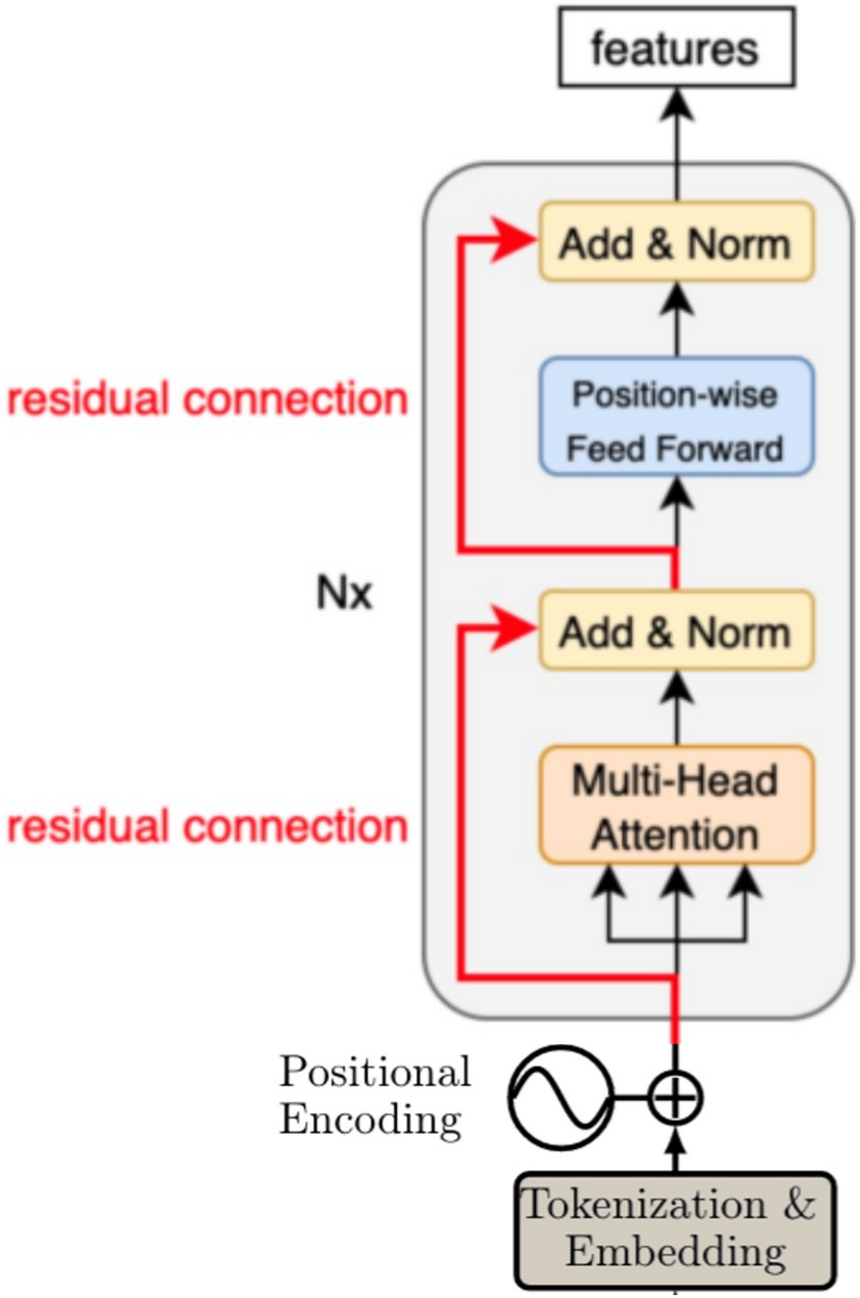
- Concatenate the heads' outputs
- Use a linear layer: desired output size

# Layer normalization

- Normalization on joint channel and spatial dimensions

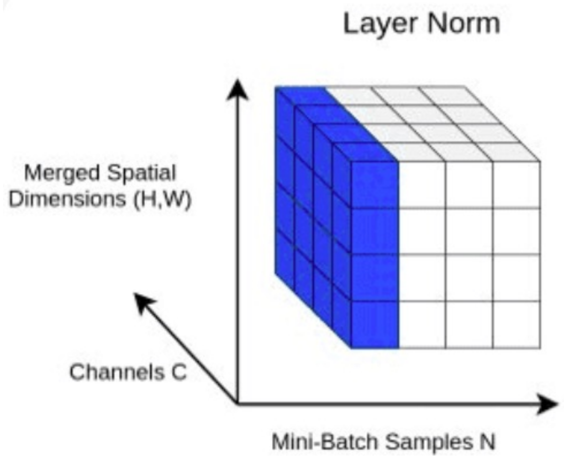



  - Stabilize training, faster convergence



# Layer normalization

- Normalization on joint channel and spatial dimensions




$$\mu_n = \frac{1}{K}\sum_{k=1}^{K} x_{nk}$$

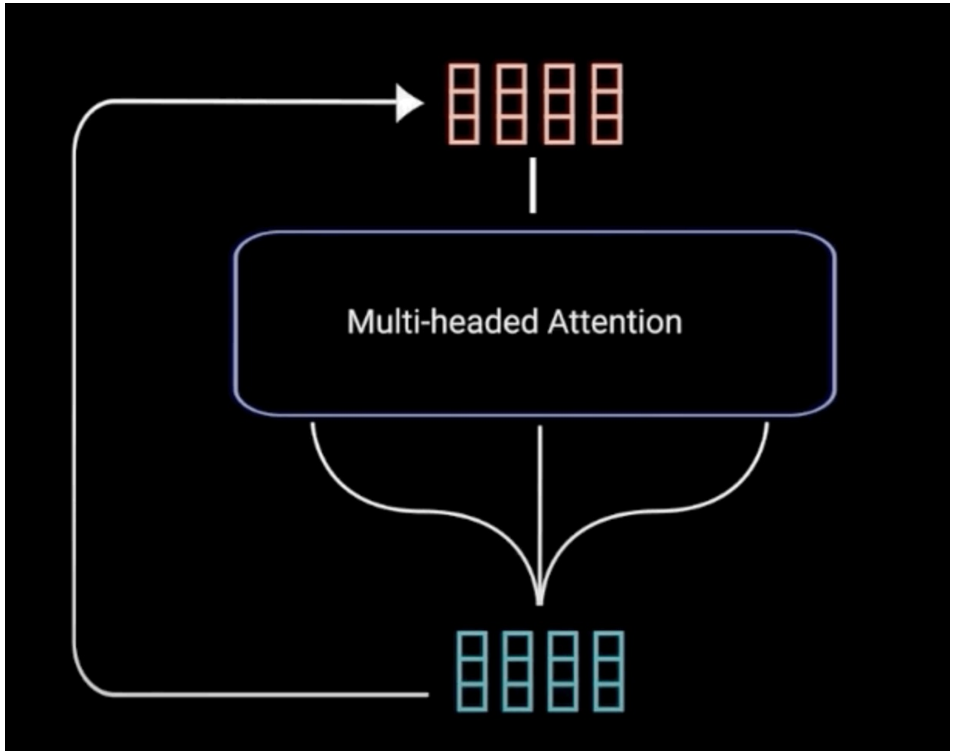
$$\sigma_n^2 = \frac{1}{K}\sum_{k=1}^{K} (x_{nk} - \mu_n)^2$$

$\beta, \gamma$ learnable parameters

$$\hat{x}_{nk} = \frac{x_{nk} - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}}, \hat{x}_{nk} \in R$$

$$\mathrm{LN}_{\gamma,\beta}(x_n) = \gamma \hat{x}_n + \beta, x_n \in R^K$$

# Layer normalization + residual connections





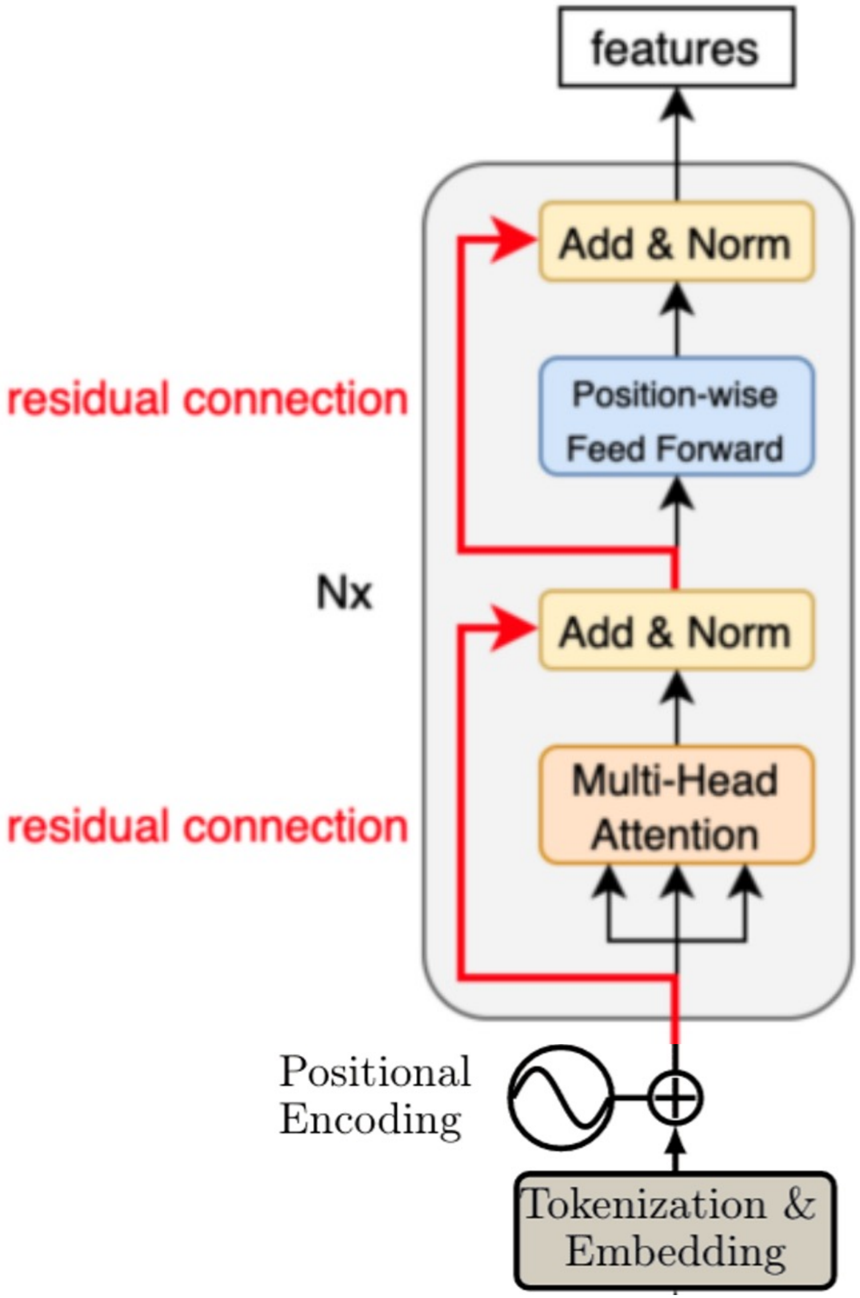
**Residual connections**

- Better gradient flow (vanishing gradients)
- Leverage input encoding, *e.g.* PE

# Feed-Forward Network (FFN)

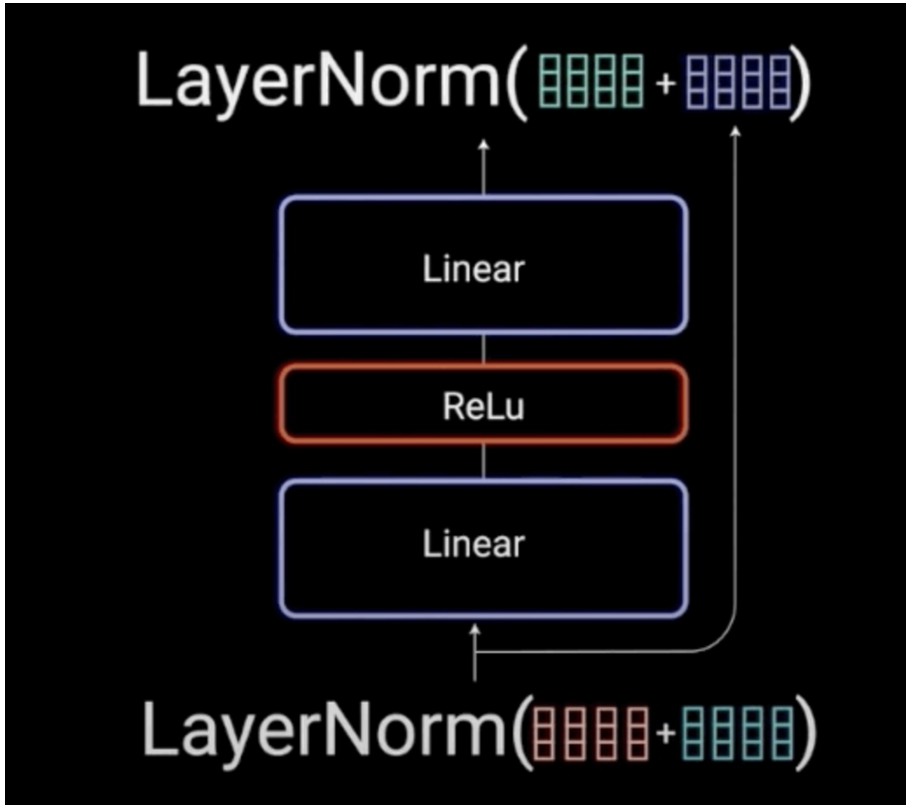
- Position-wise FFN: applied to each token separately and identically

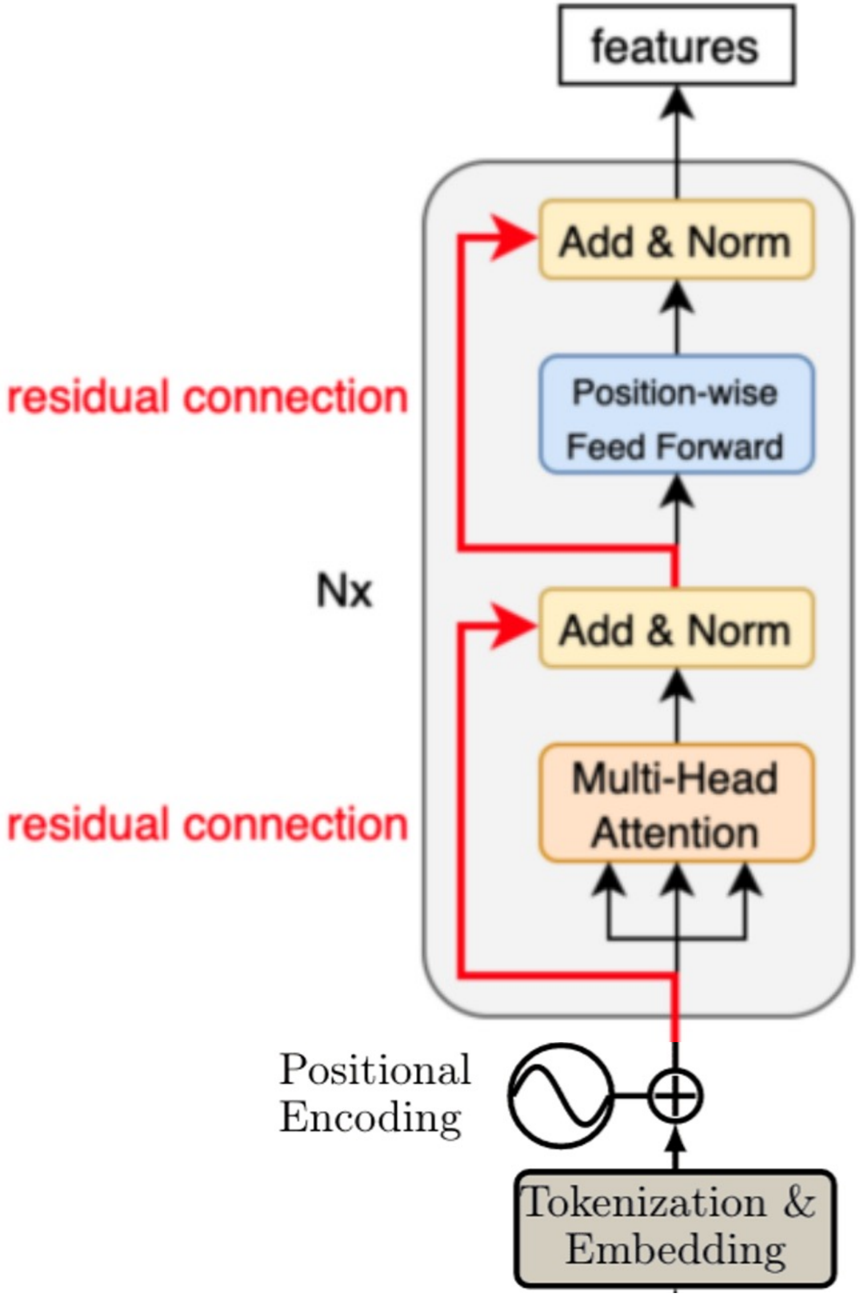
$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



# FNN + residual Layer Norm

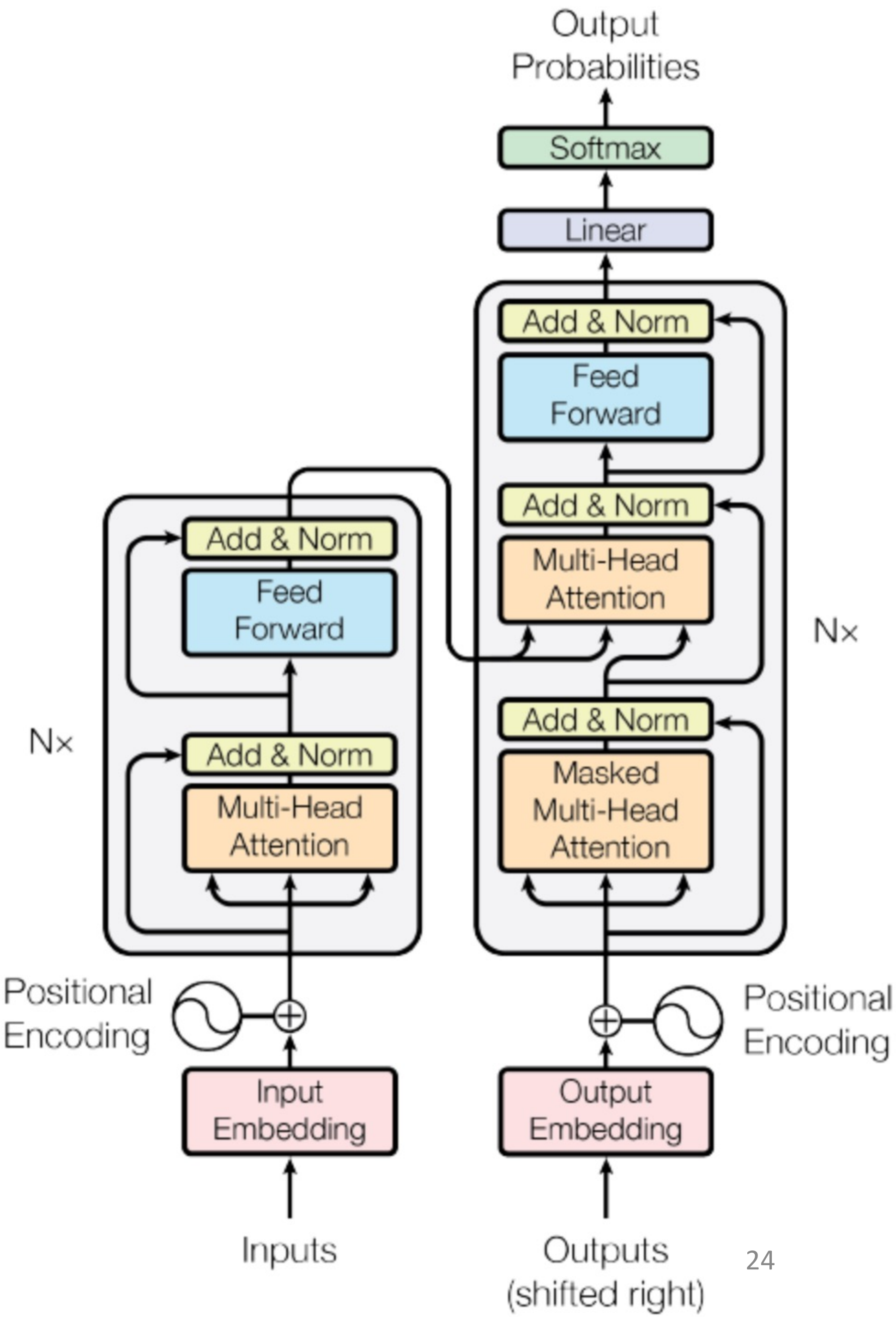
$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



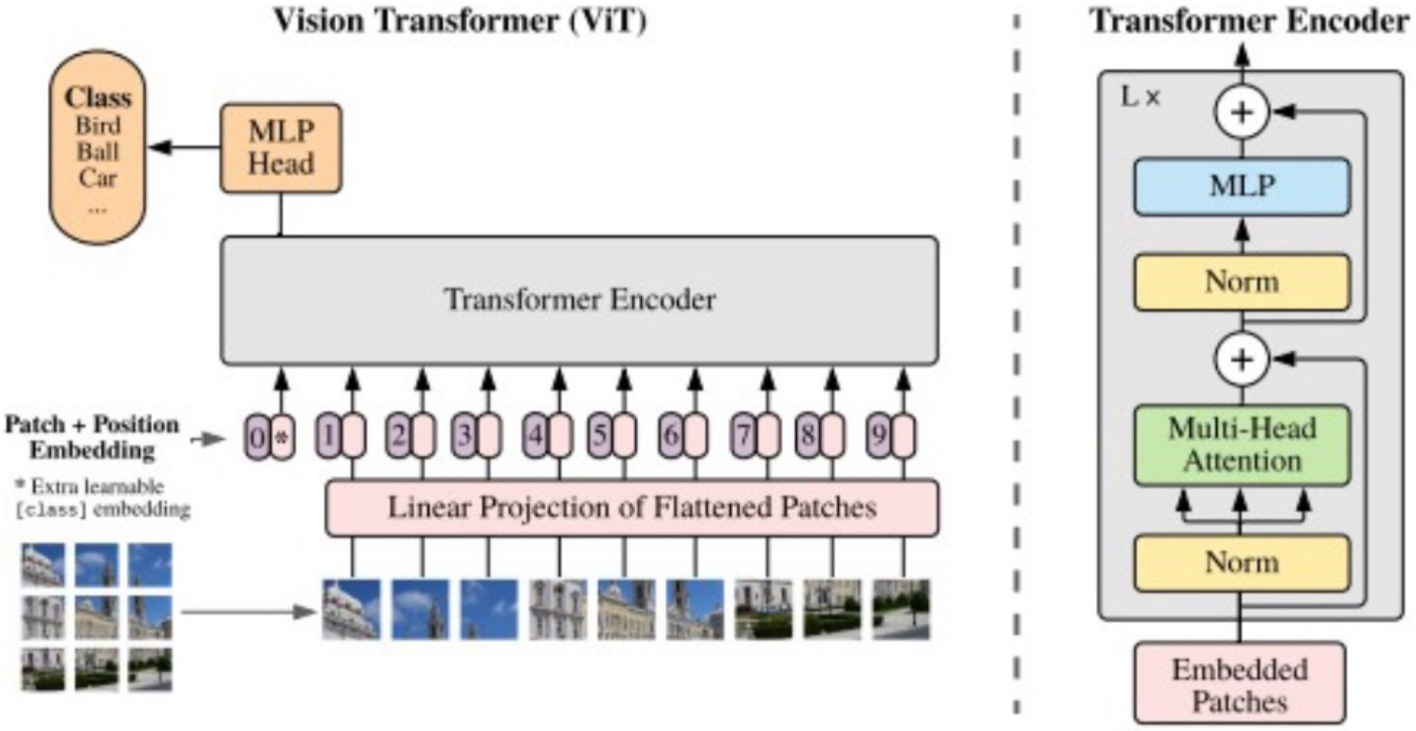
# Transformer: conclusion

- Importance of attention: global interactions between tokens
- On the other hand relaxes inductive biases
  - e.g. ConvNets translation equivariant
    - vs transformers permutation equivariant
  - More flexibility to learn adequate mapping
  - Needs more data



1. Transformers
2. Vision Image Transformer
3. Transformers for medical image segmentation
4. Current trend & Perspectives

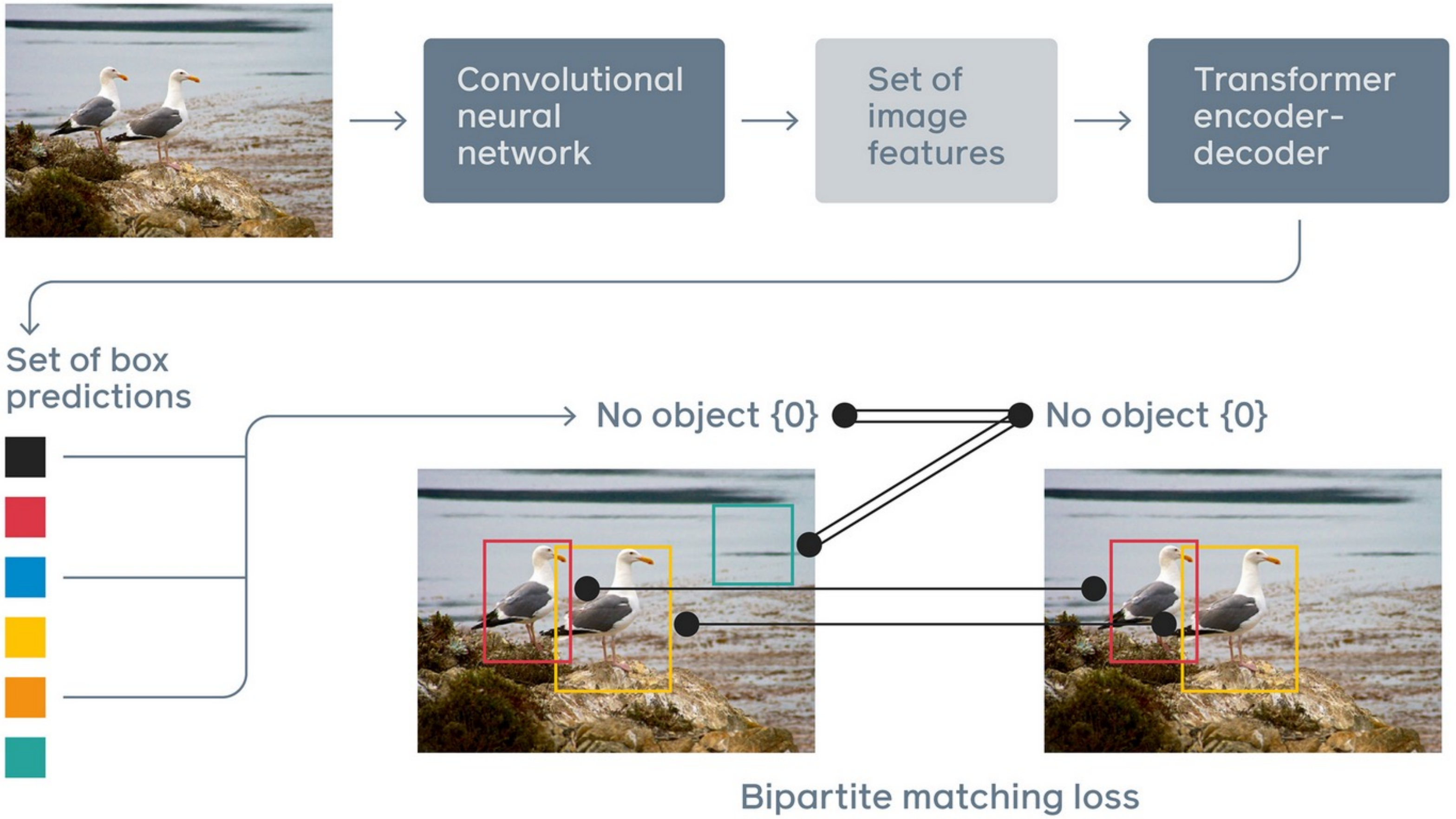
# Vision Image Transformer (ViT) [2]




- Direct application of transformer's encoder for images
- Learned on JFT ($300.10^6$ images)
- Extra learnable token: used for class prediction
  - "Learned" pooling wrt visual tokens

[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. ICLR 2020.

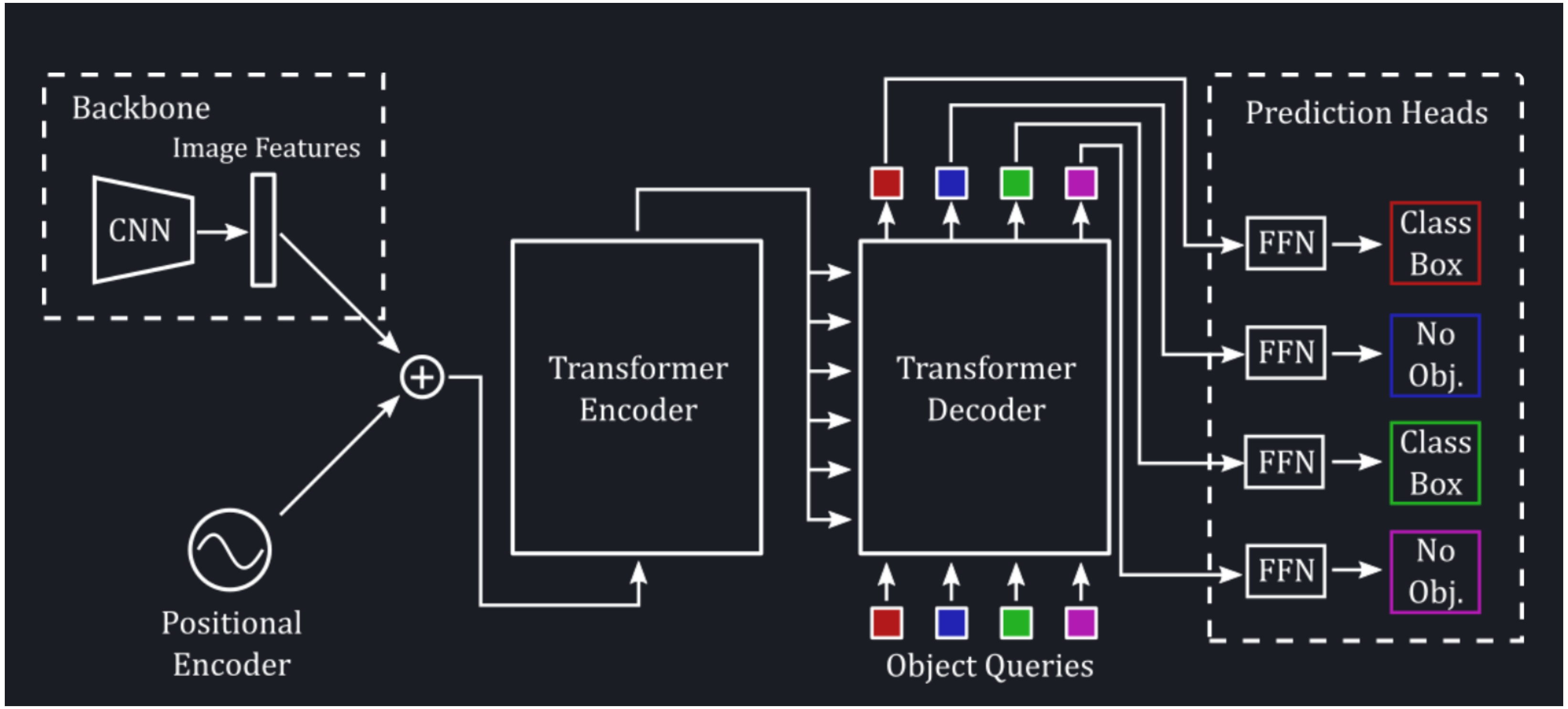
# Detection Transformer (DETR) [3]

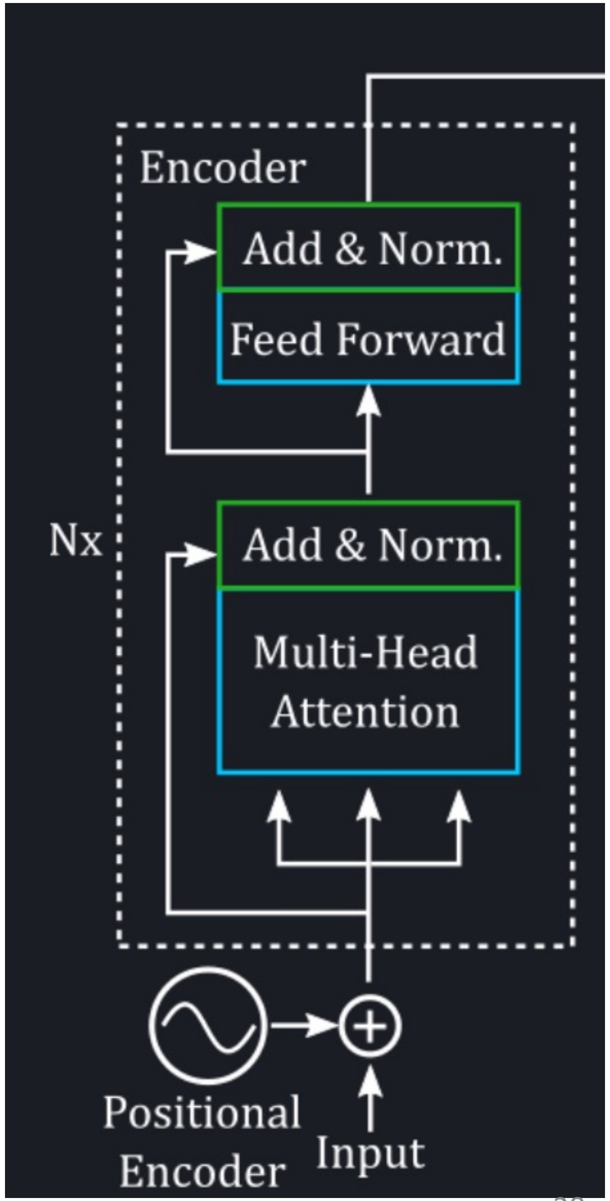



[3] End-to-End Object Detection with Transformers. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. ECCV 2020.

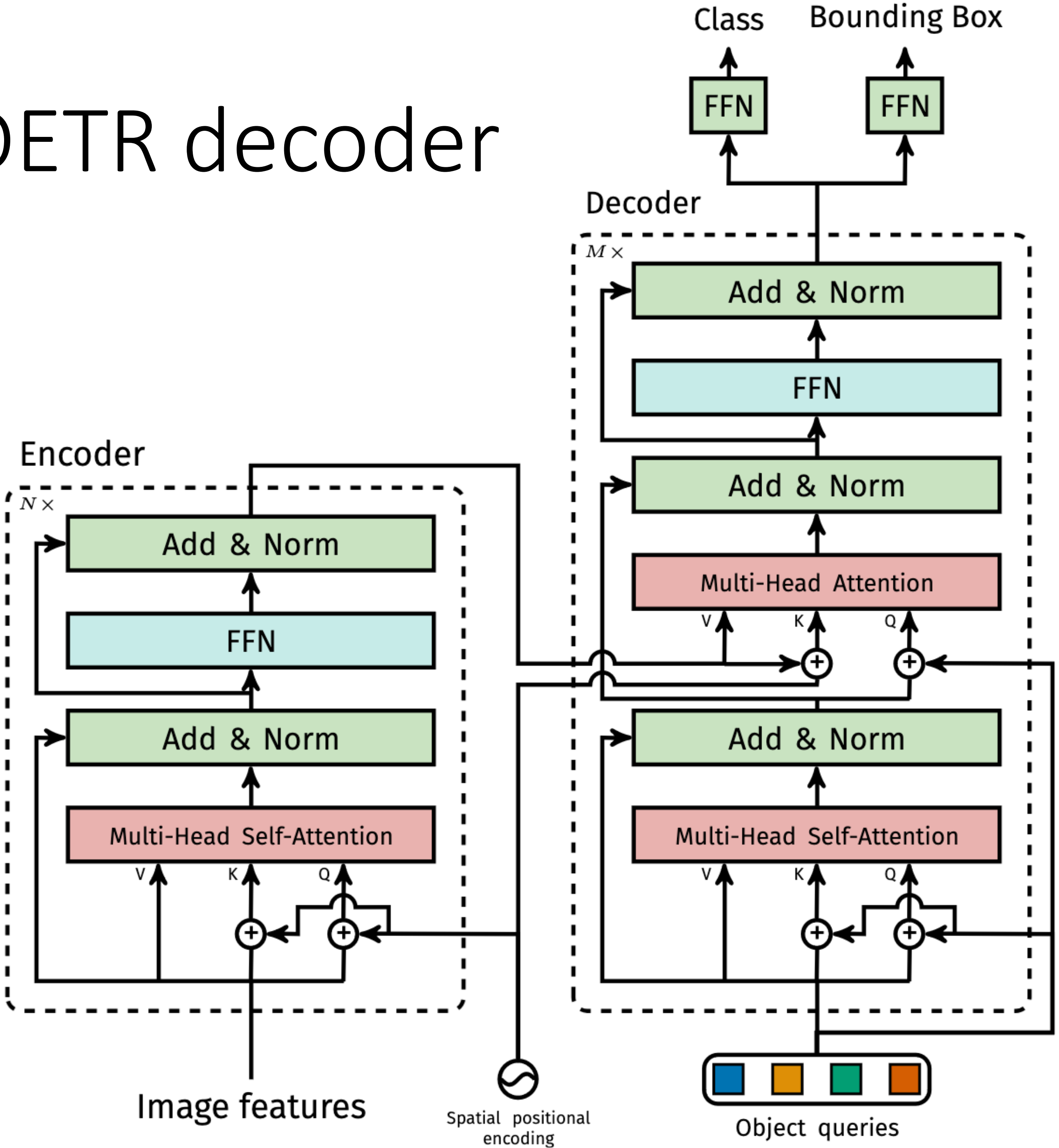
# DETR encoder

- Conv Backbone + Standard ViT with PE at each transformer layer



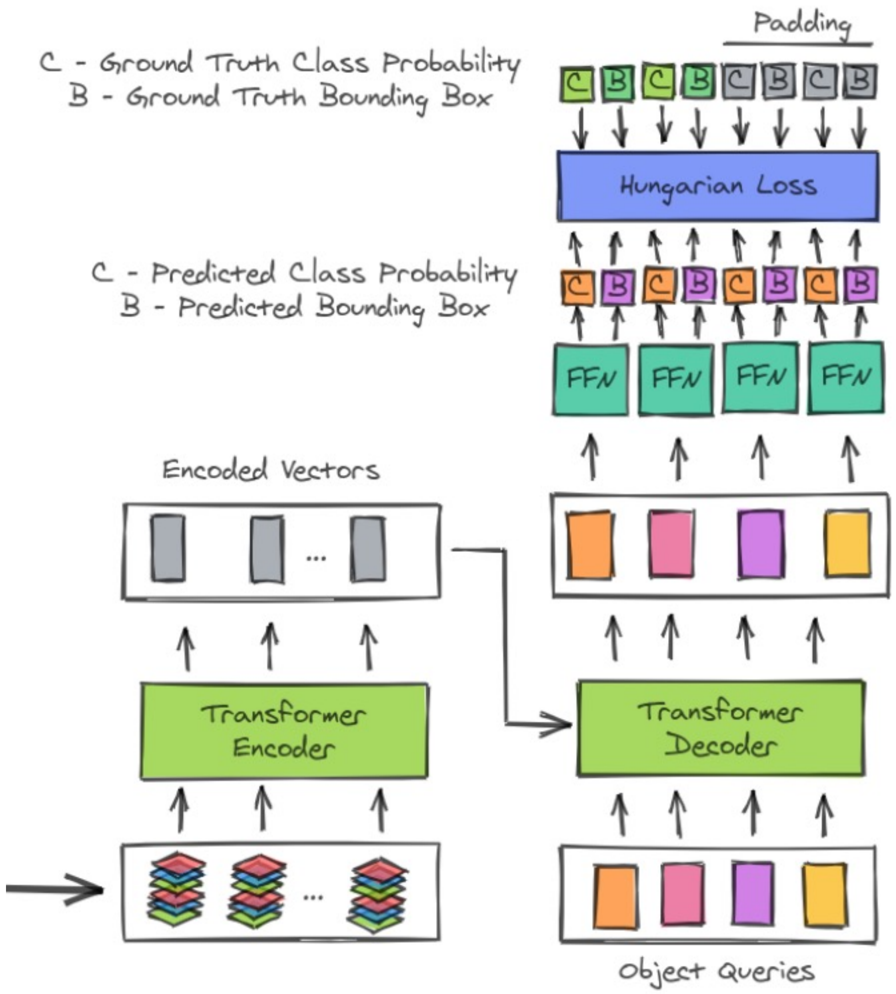
# DETR decoder




- Learned object queries (OQ 100)
- Self-attention (can be omitted at 1st decoder layer)
- Cross-attention
  - Query : OQ added
  - Key : encoder output + PE
  - Value : encoder output
- Decoder output: 2 branches
  - FFN for class prediction
    - Ø for background
  - FFN for BB prediction

$$[center_x, center_y, height, width]$$

# DETR training




- Matching between the set of prediction and set of BB in supervision
- Best match between the sets using the Hungarian algo

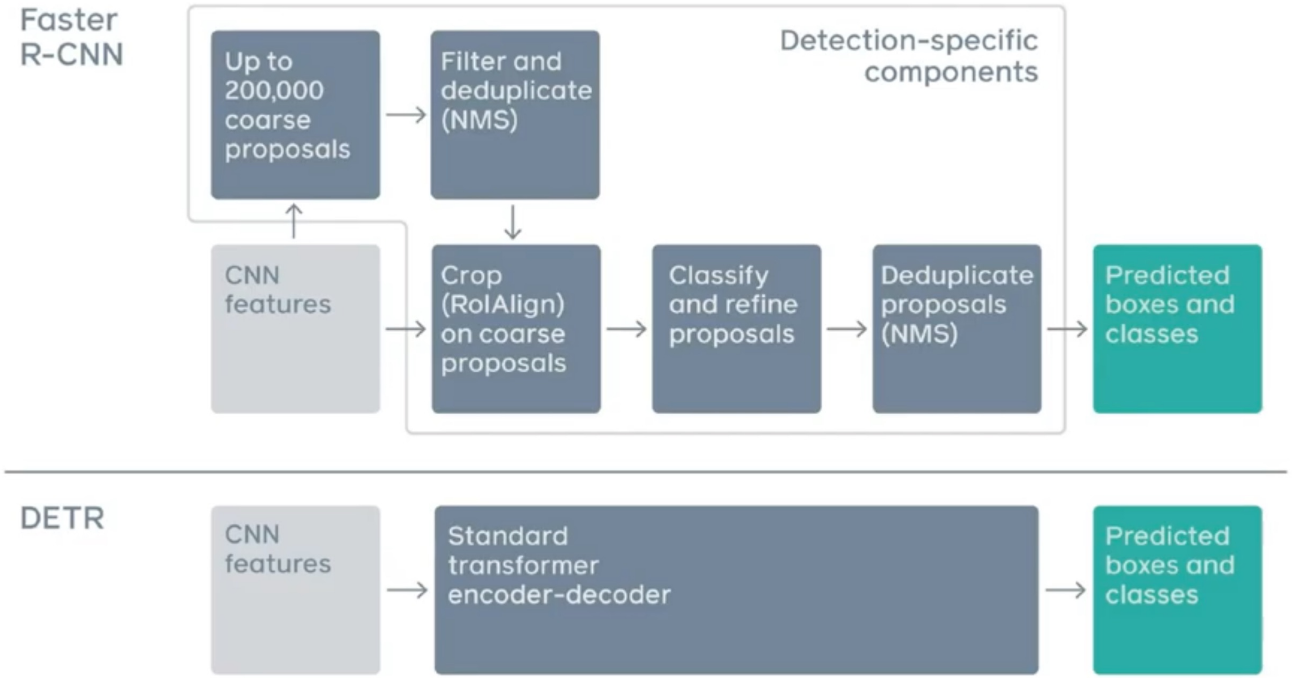
$$\mathcal{L}_{\text{Hungarian}}(y,\hat{y})=\sum_{i=1}^{N}\left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i)+\mathbb{1}_{\{c_i\neq\varnothing\}}\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\hat{\sigma}}(i))\right]$$

$$\hat{\sigma}=\arg\min_{\sigma\in\mathbb{N}}\sum_{i}^{N}-\mathbb{I}_{c_i\neq\phi}\hat{p}_{\hat{\sigma}(i)}(c_i)+\mathbb{I}_{c_i\neq\phi}L_{box}(b_i,\hat{b}_i)$$

$$L_{box}=\lambda_{iou}L_{iou}(b_i,\hat{b}_i)+\lambda_{L1}||b_i-\hat{b}_{\hat{\sigma}(i}||$$

# DETR: conclusion

- Simple model
- Works  well for large objects, less good for small objects

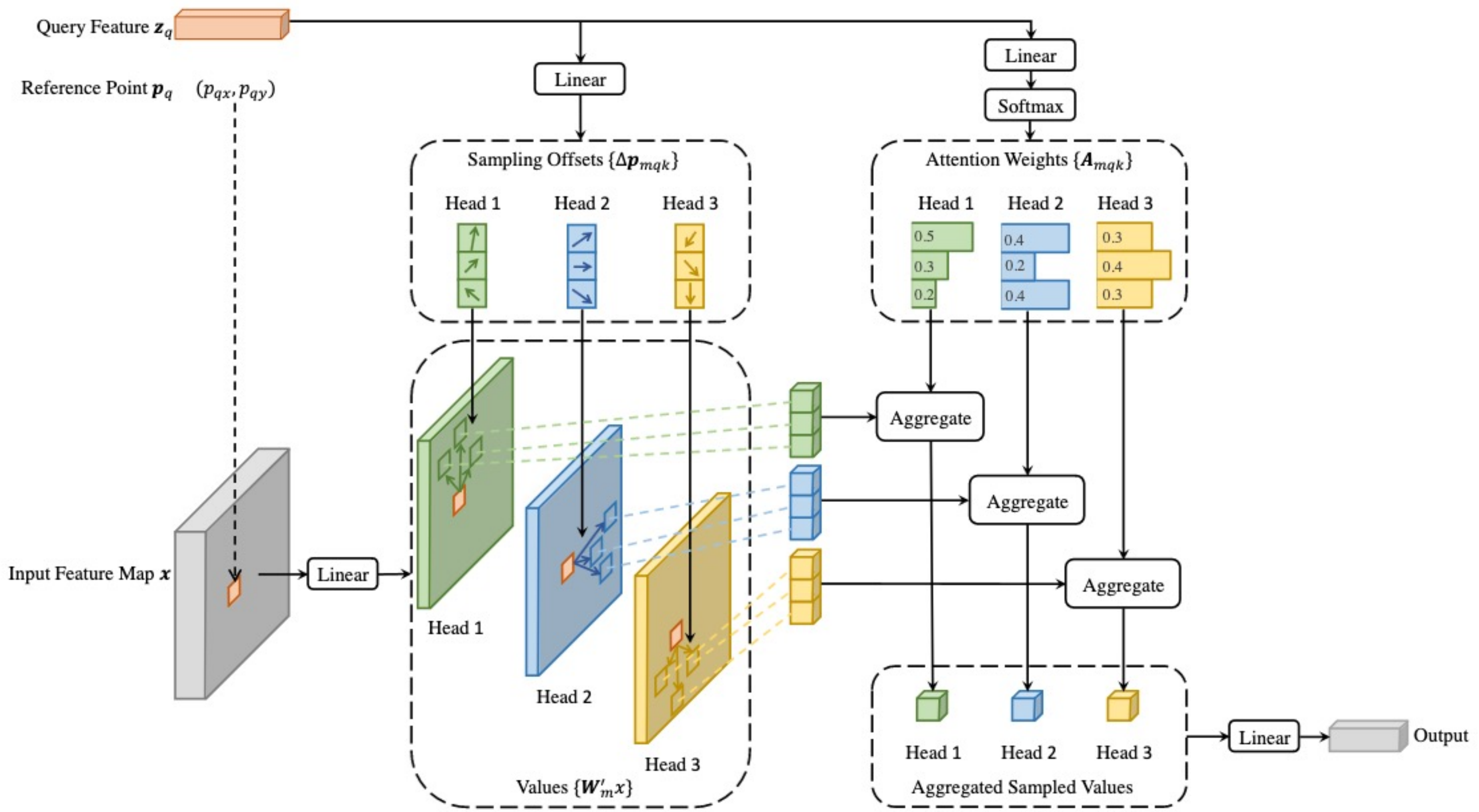


# Deformable DETR [4]

$$\mathrm{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[ \sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \Big]$$

- Deformable attention

$$\mathrm{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[ \sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \Big]$$

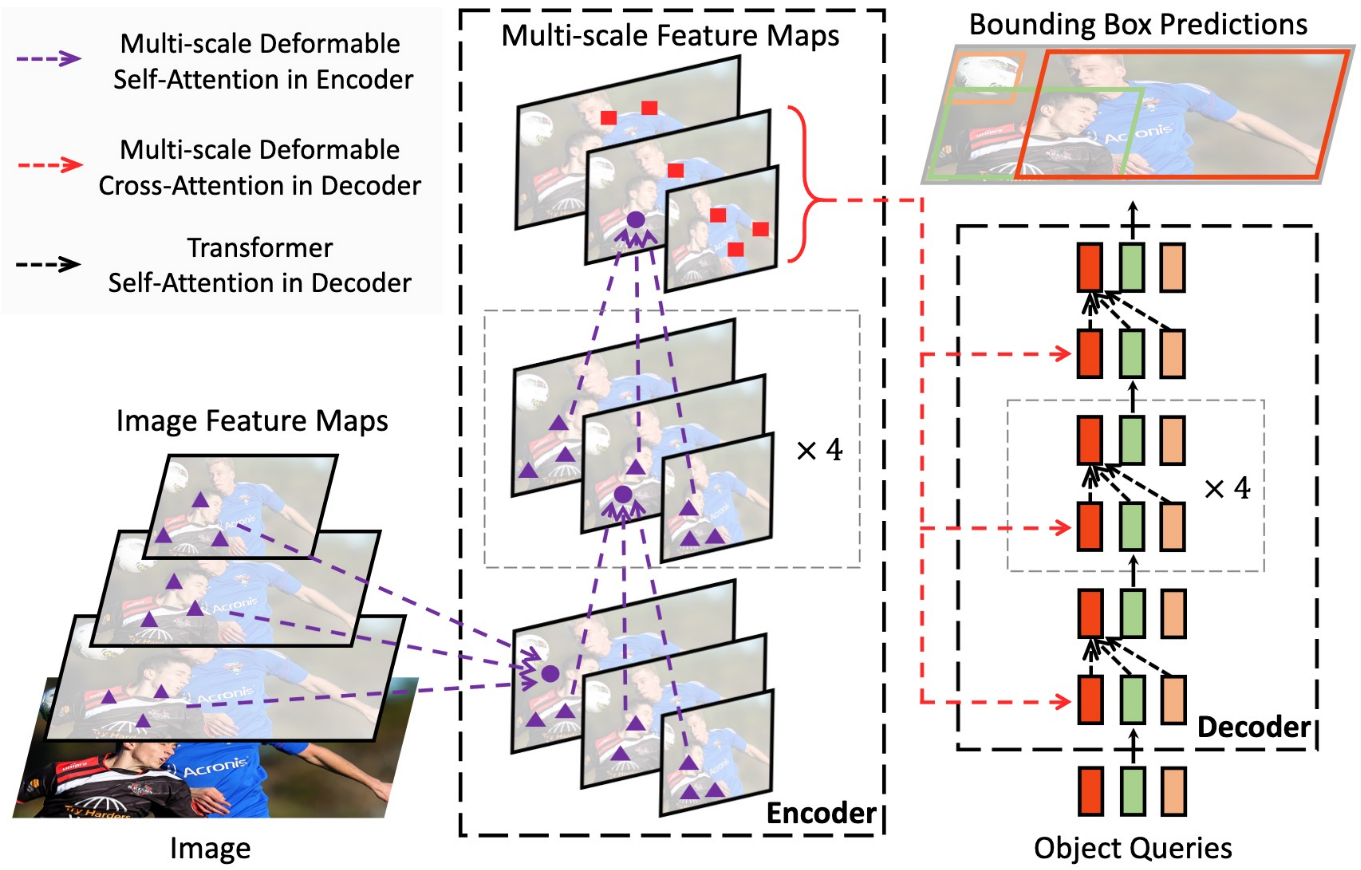



- Query: input vector from tensor. For each head, predict a 3K value
  - 2K elements for the offset for getting K Keys (here K=3)
  - K elements for getting K attention weight
- Value: for each head, weighted average of the K sampled keys
- Complexity : O(WH.K) vs O((WH)²)

[4] Deformable DETR: Deformable Transformers for End-to-End Object Detection. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. ICLR 2021.

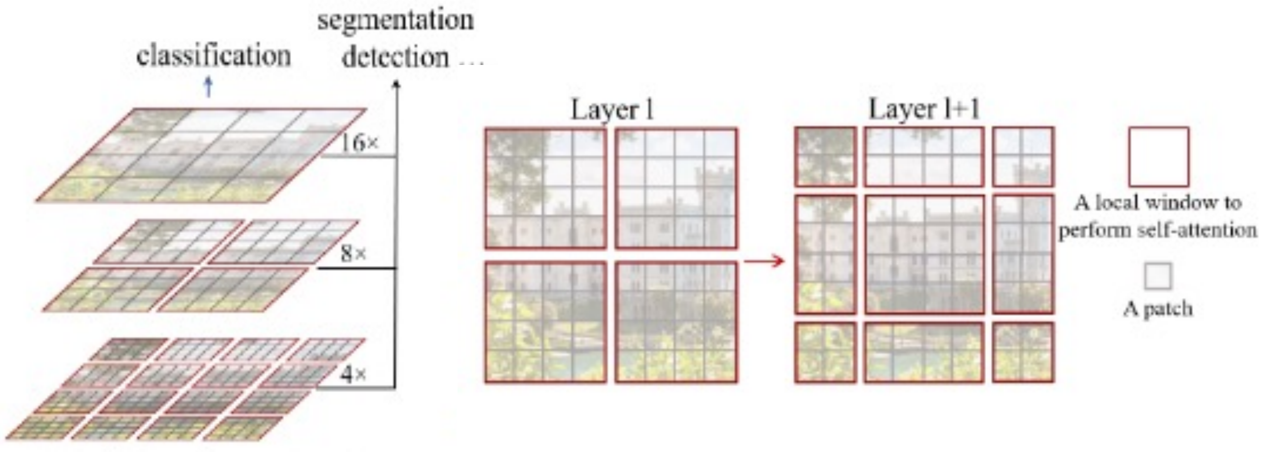
# Deformable DETR

- Applied in multi-resolution feature maps
- Improve DETR effectiveness for small objects requiring high-resolution feature maps

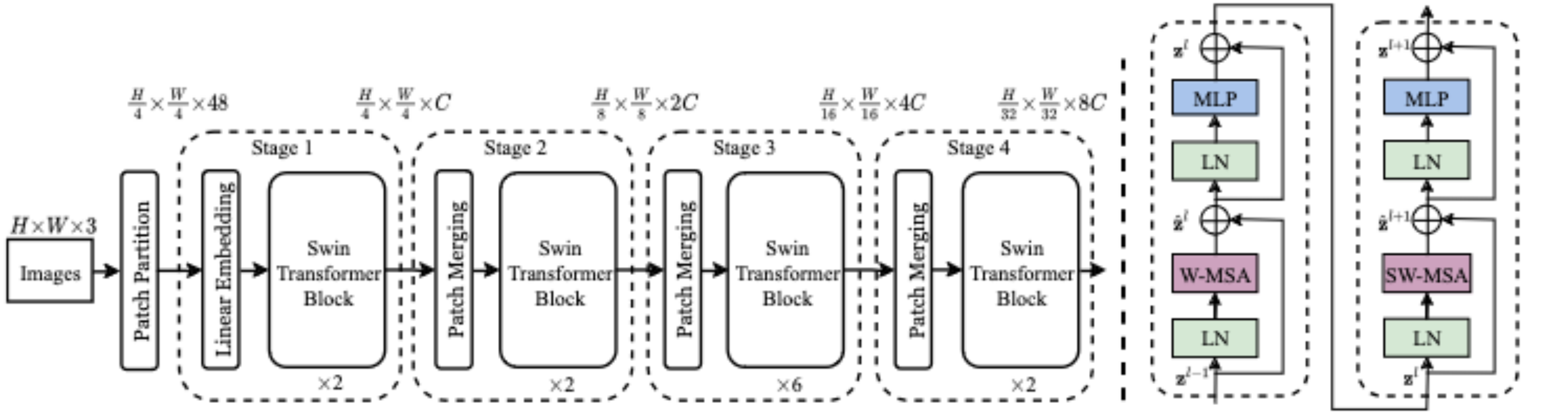


# Transformer in segmentation

- Swin-Transformer [5]
  - Multi-resolution transformer
    - Local attention in lower-layers
      - Shifted windows at layers l/l+1
    - Patch merging => larger receptive field



(a) Swin Transformer (ours)

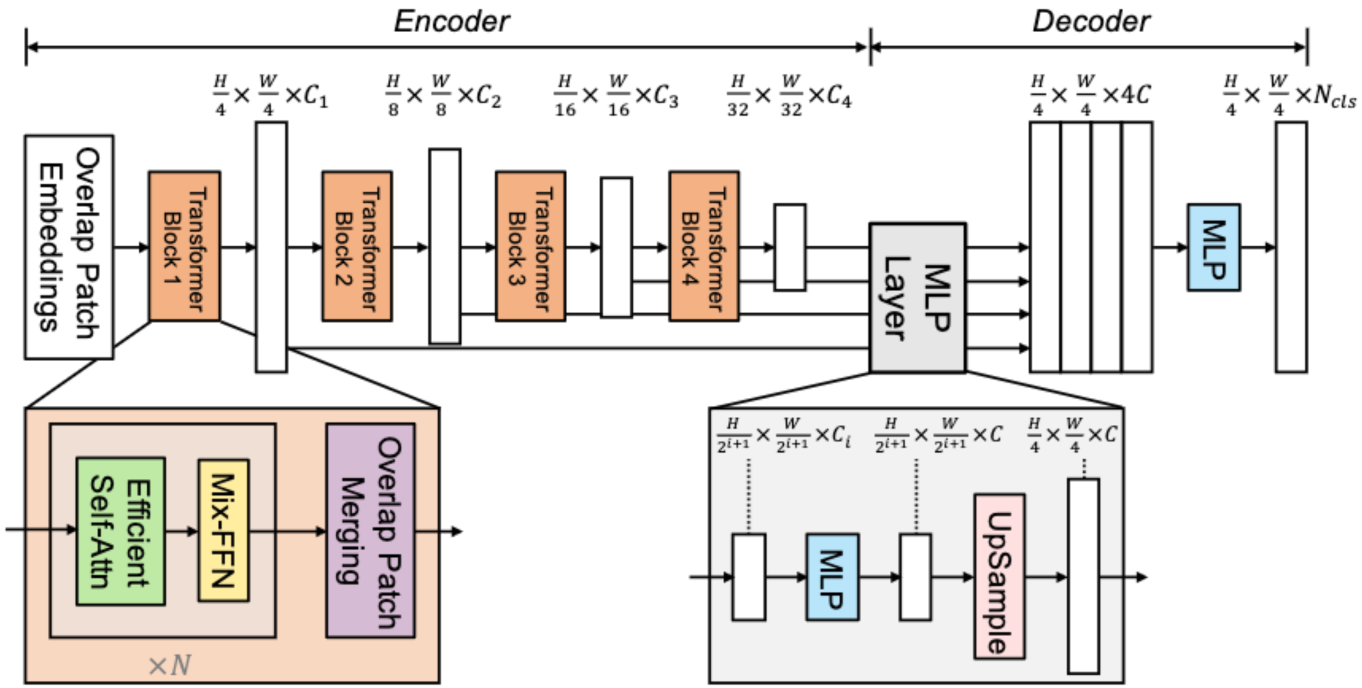


(a) Architecture (b) Two Successive Swin Transformer Blocks

[5] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. ICCV 2021

# Transformer in segmentation

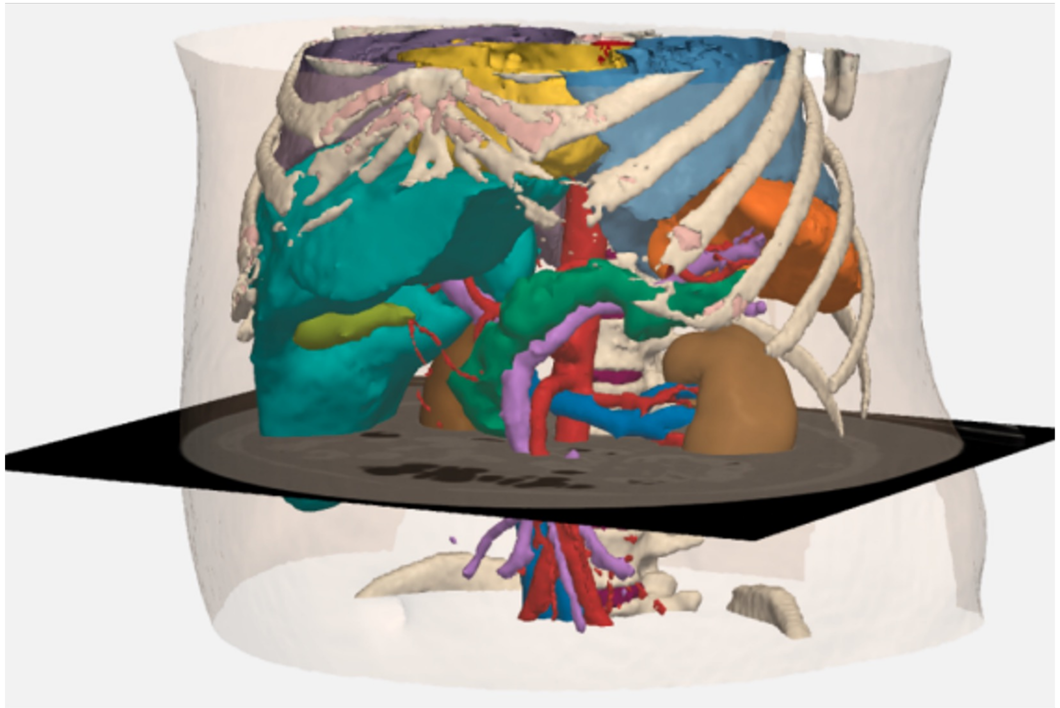
- SegFormer [6]
  - Efficient attention, at multi-scale



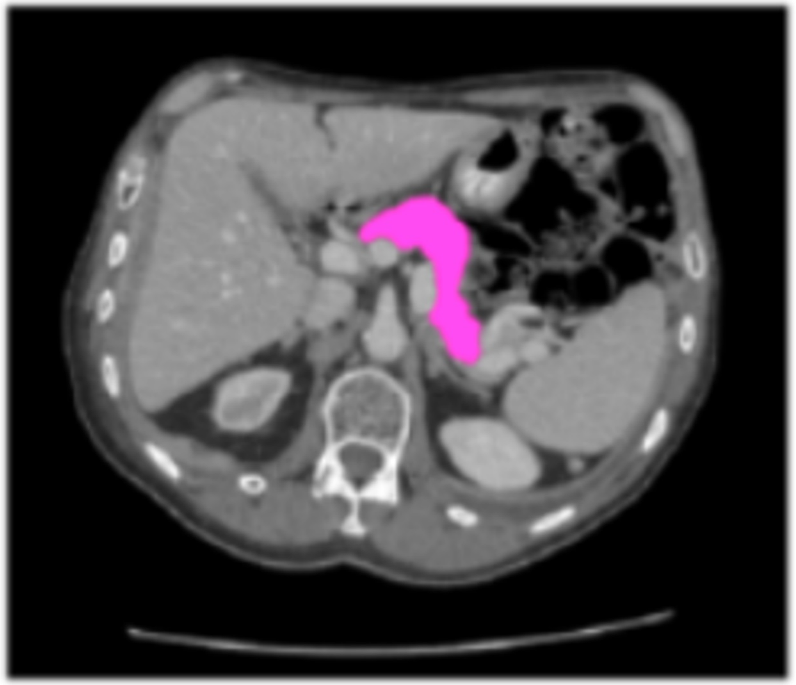

[6] SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo. NeurIPS 2021.

1. Transformers
2. Vision Image Transformer
3. Transformers for medical image segmentation
4. Current trend & Perspectives

# Context: 2D organ segmentation example



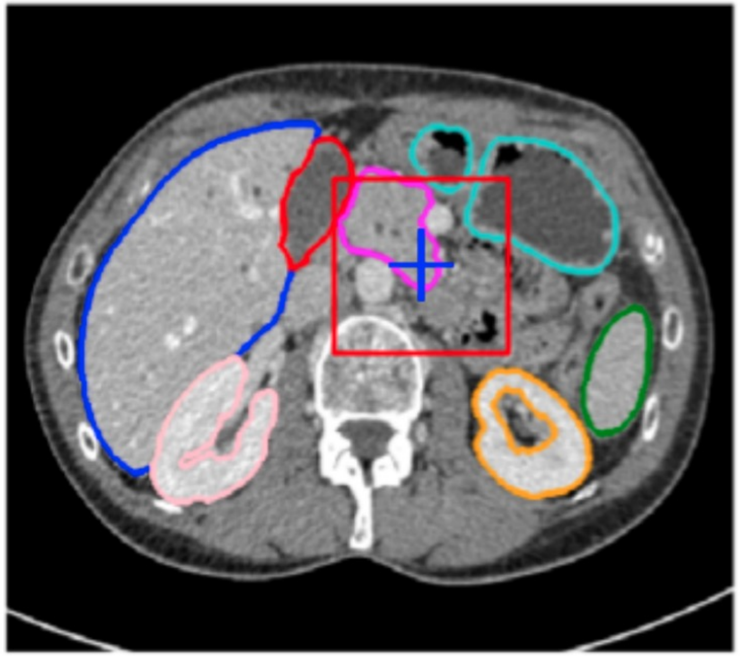
***Organs segmentation illustration***



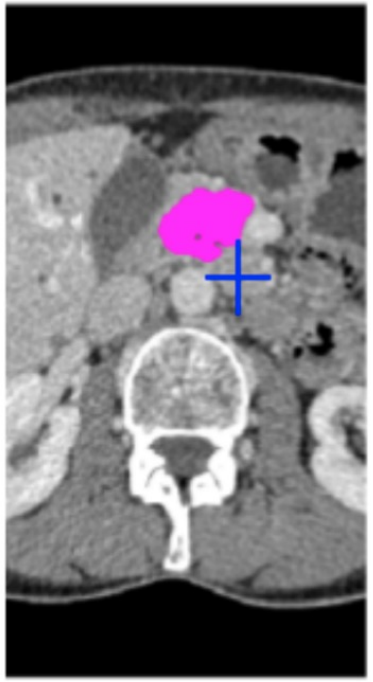
***Pancreas automatic segmentation***

# Segmentation: importance of long-range dependencies

**U-Net [A]: unable to represent full context**



a) Ground Truth



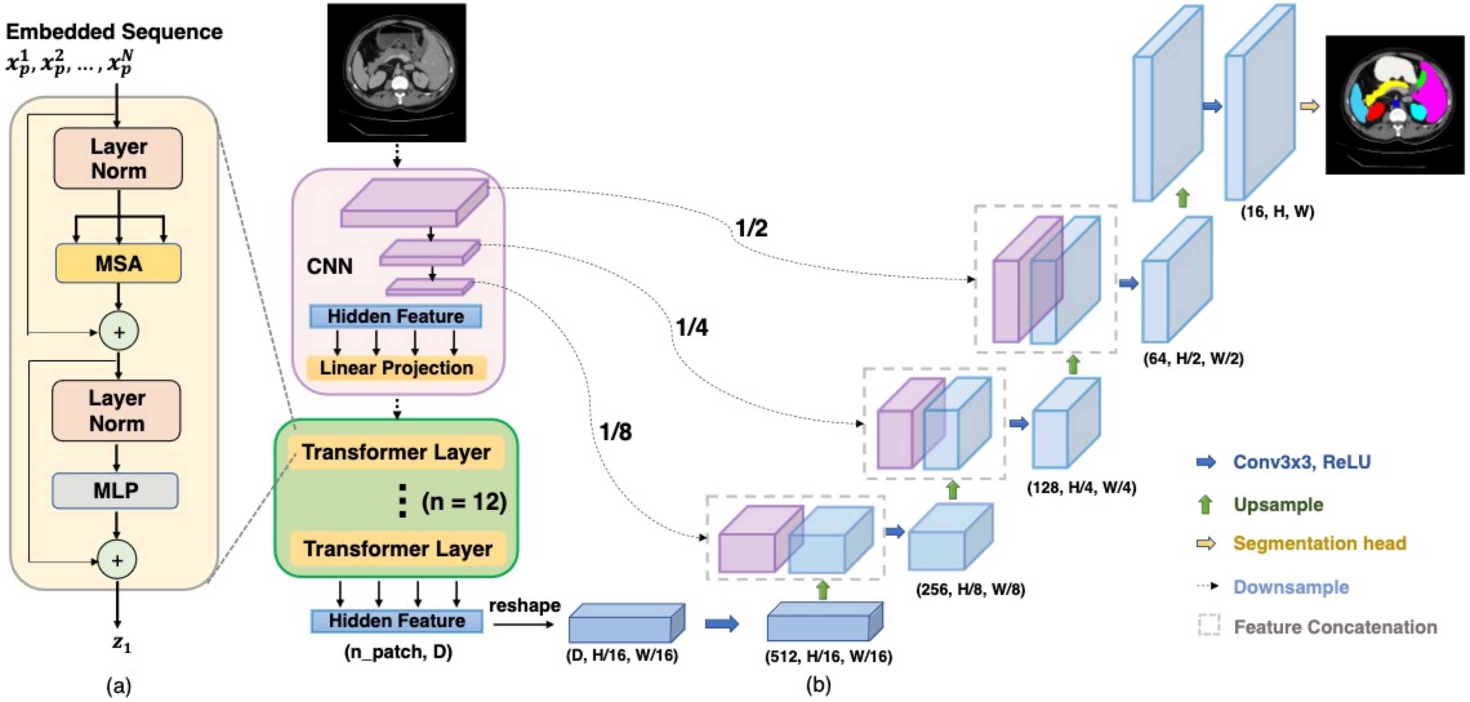
c) U-Net

***Segmentation example with U-Net's receptive field (red square)***

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

# Trans U-Net [7], U-Transformer [8]

- Seminal works for using transformers in medical image segmentation
- Adding self-attention on the bottleneck of a U-Net
  - Inspired from non-local networks [9]

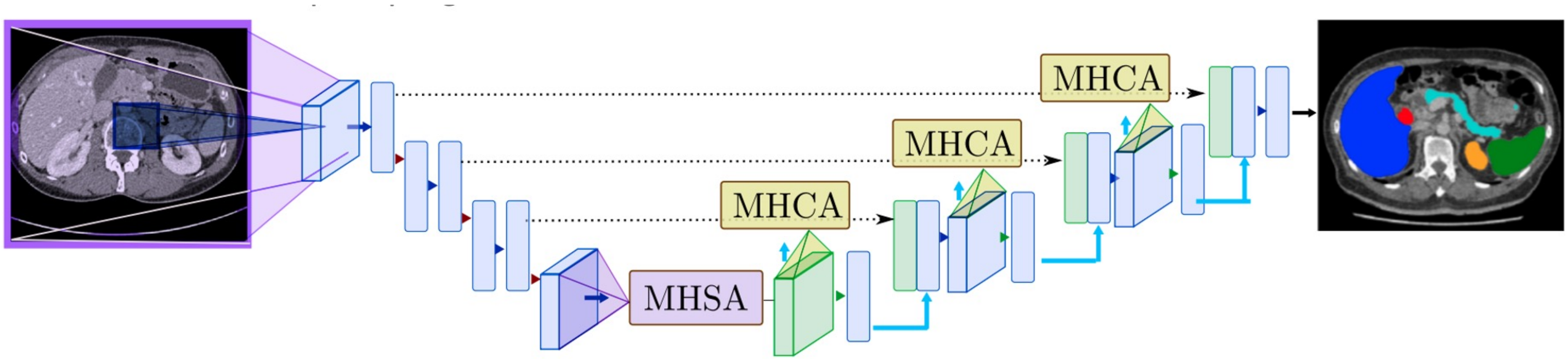



Trans U-Net architecture


[7] TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. J. Chen et.al. arXiv, Feb 2021.
[8] U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. O. Petit, N. Thome, C. Rambour, L. Soler. arXiv, March 2021.
[9] Non-local Neural Networks. X. Wang, R. Girshick, A. Gupta, K. He. CVPR 2018.

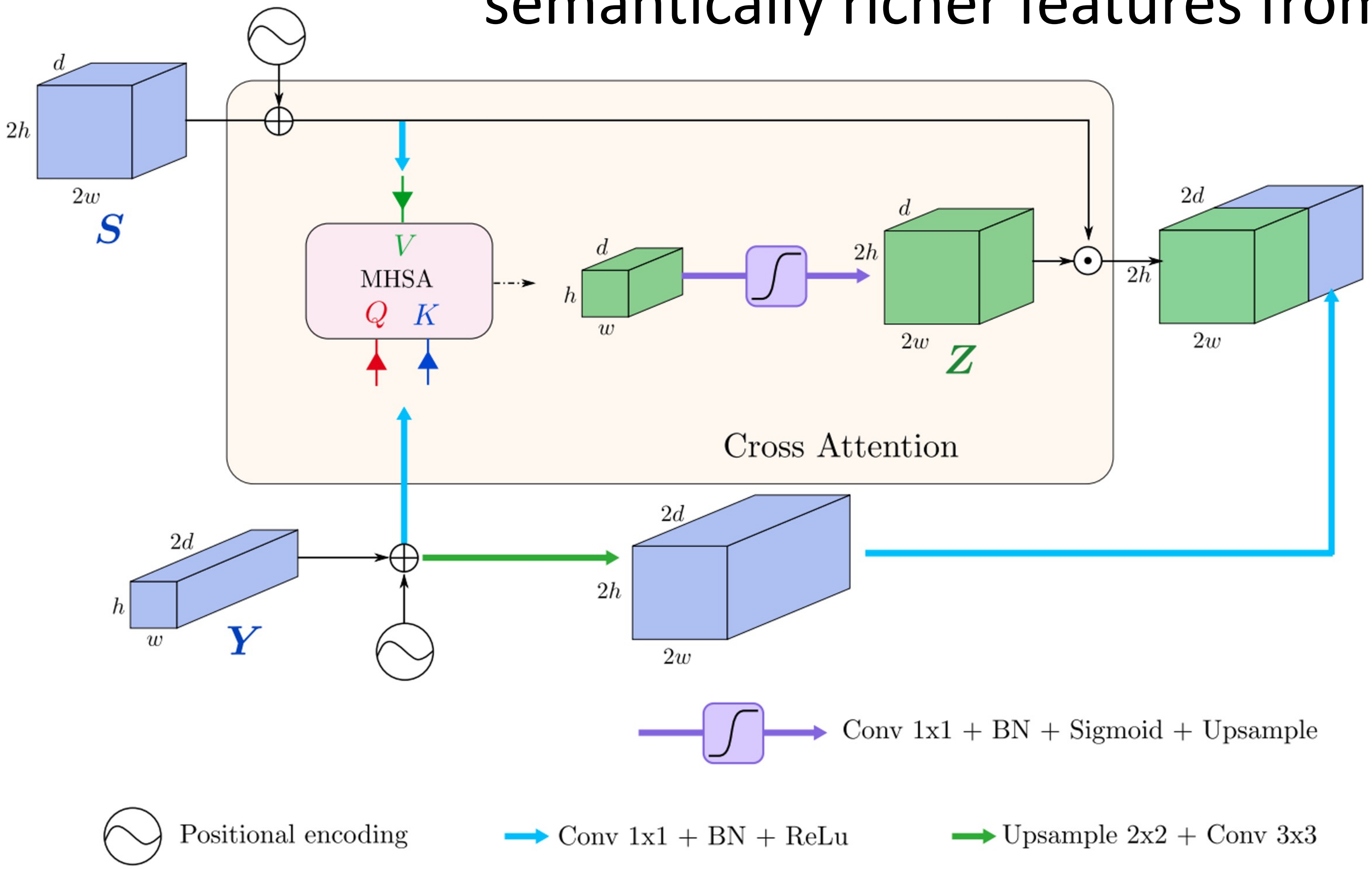
# U-Transformer [8]

- **U-Transformer**: self and cross attention in medical image segmentation
  - Self-attention in bottleneck
  - Cross attention to improve super-resolution in skip connections



# Architecture: Multi-Head Cross-Attention

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



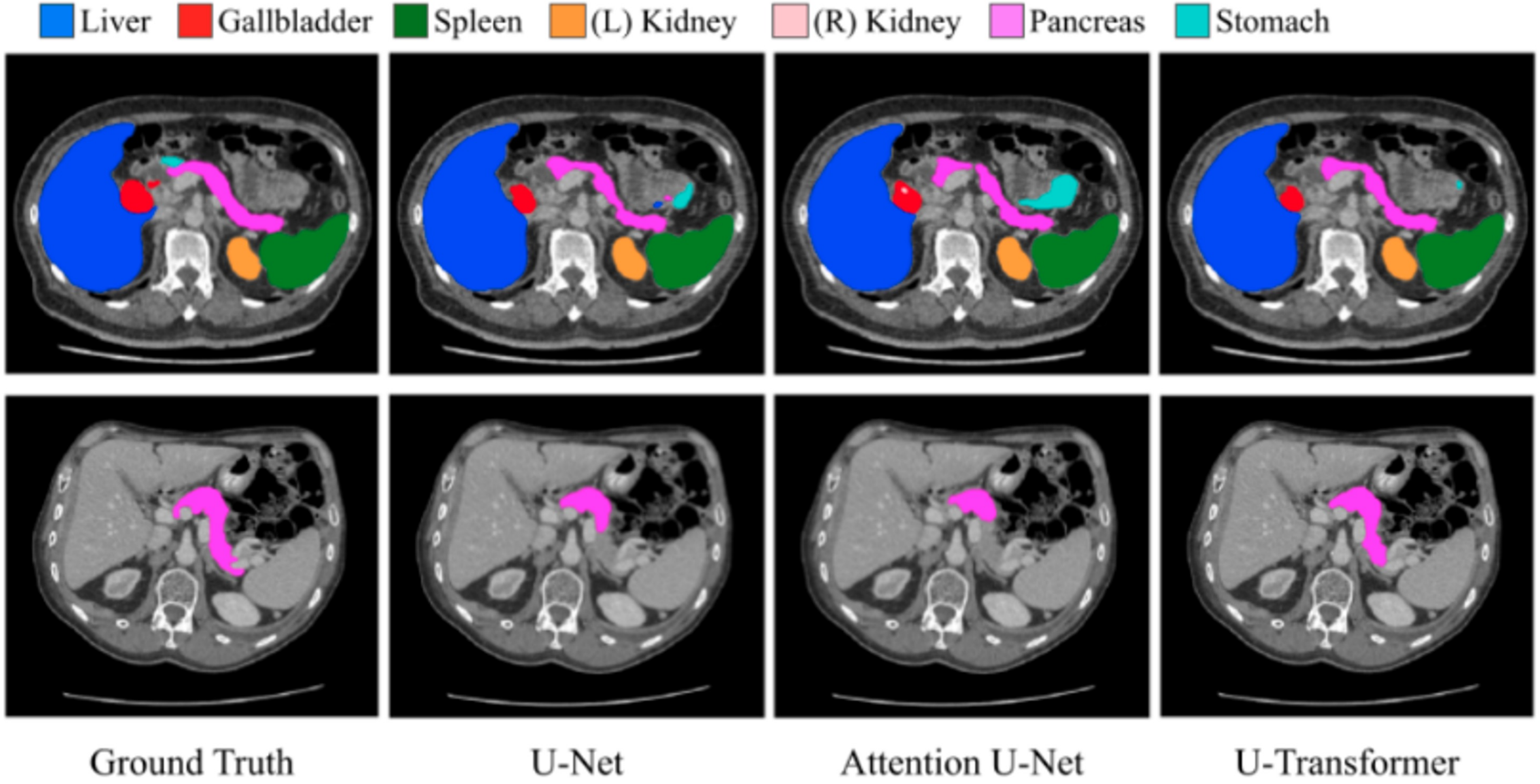

$Y$ : Semantically richer features from bottleneck
$S$ : High resolution features from skip connections

# 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) |

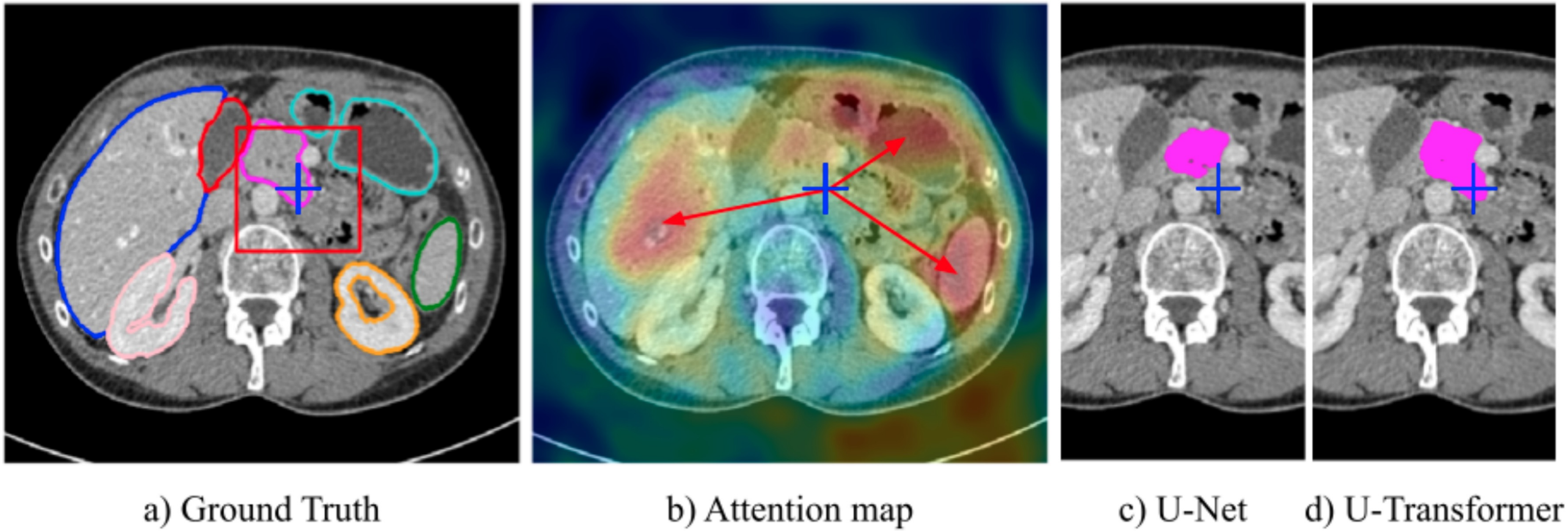
| Organ | U-Net [11] | Attn U-Net [13] | MHSA | MHCA | U-Transformer |
|---|---|---|---|---|---|
| Pancreas | 69.71 (± 3.74) | 68.65 (± 2.95) | 71.64 (± 3.01) | 71.87 (± 2.97) | **73.10** (± 2.91) |
| Gallbladder | 76.98 (± 6.60) | 76.14 (± 6.98) | 76.48 (± 6.12) | 77.36 (± 6.22) | **78.32** (± 6.12) |
| 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



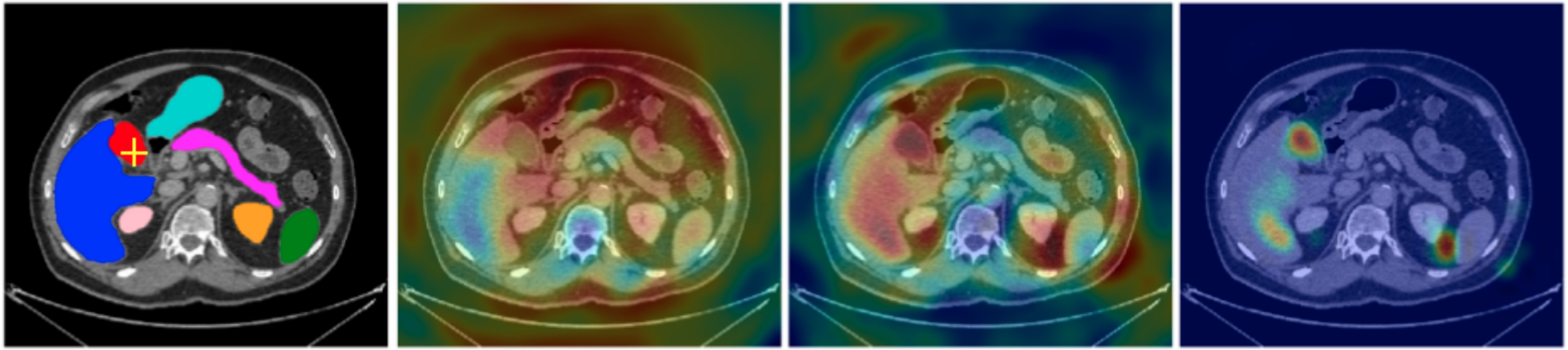
Liver
Gallbladder
Spleen
(L) Kidney
(R) Kidney
Pancreas
Stomach
Ground Truth
U-Net
Attention U-Net
U-Transformer

# Results




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

# Results

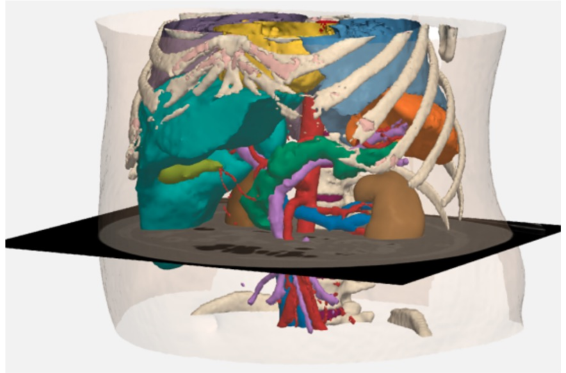


Ground Truth Cross-attn level 1 Cross-attn level 2 Cross-attn level 3

# 3D medical image segmentation

**Challenges:**

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



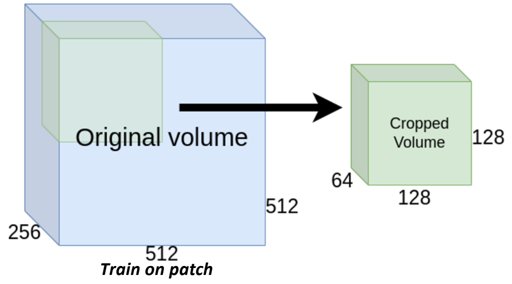
***Organs segmentation illustration***

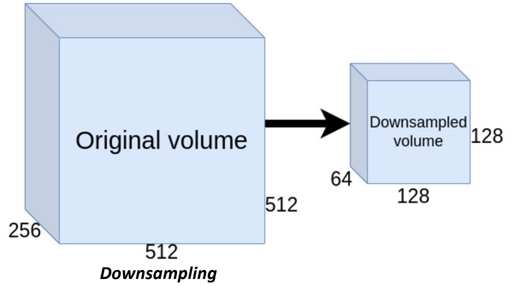
**Common strategies to reduce the memory footprint:**

- Downsampling ⇒ Drop in quality

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

⇒ No full contextual information

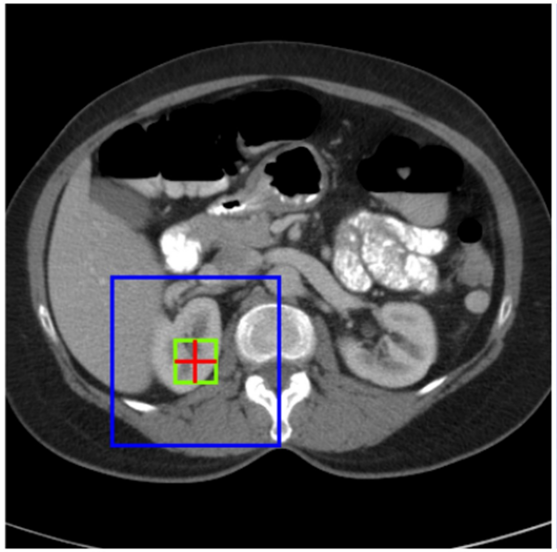


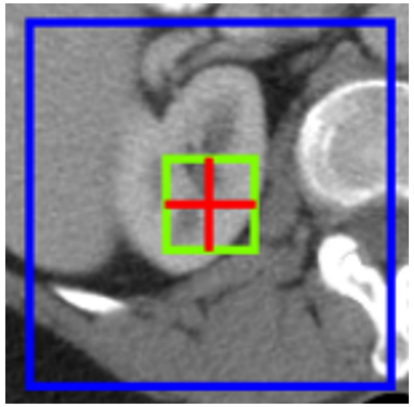
# Approaches based on patches

To keep the **full resolution,** work on patches, e.g.:

- Original image size: **512x512x256**
- Cropped patch size: **128x128x64**



*Input image 2D slice*

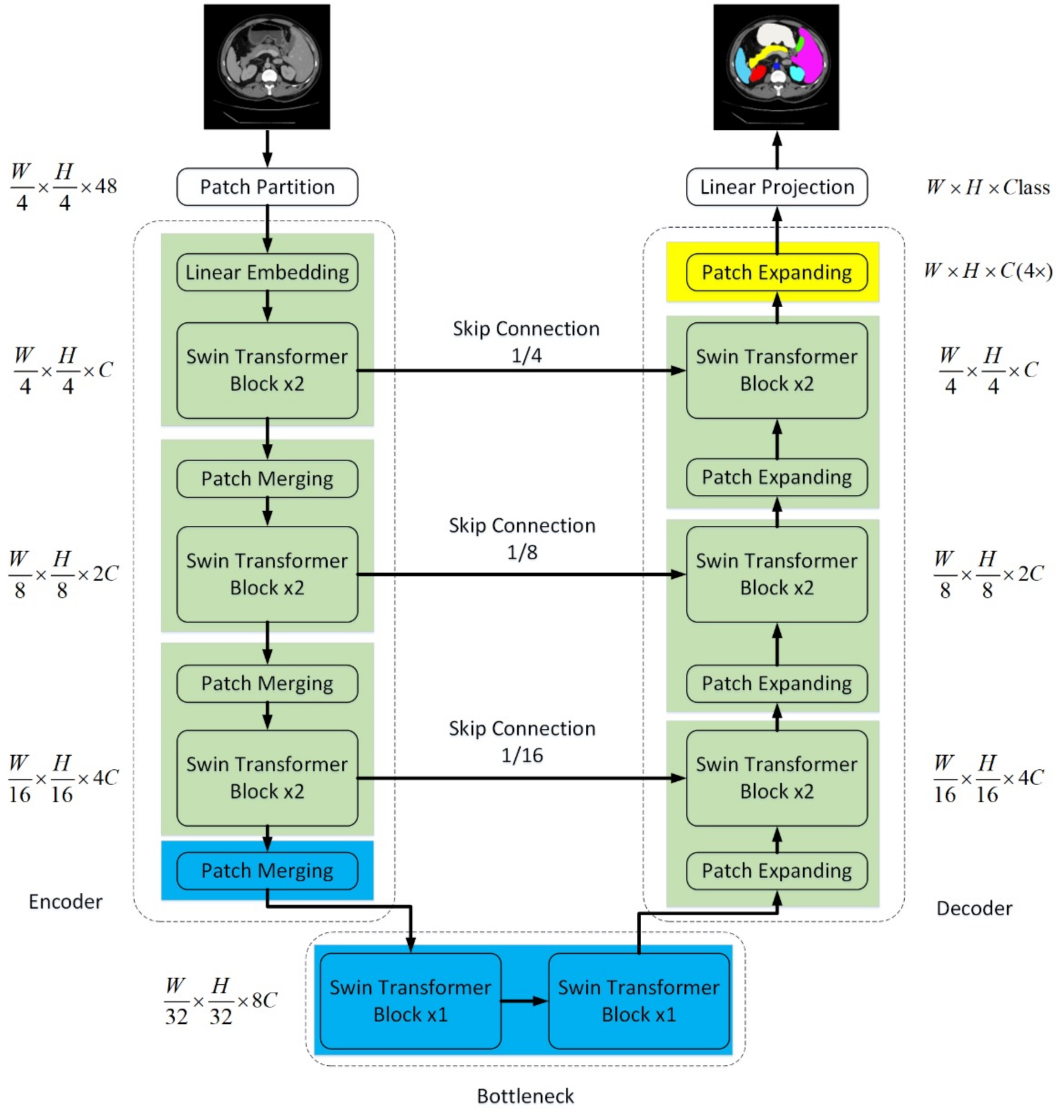


*Cropped patch 2D slice*

- **Full context lost**
- **Even on patch: full context challenging!**

# Swin-UNet [10]

- Window attention (~Swin) in a 2D multi-resolution transformer
- Patch merging: pooling

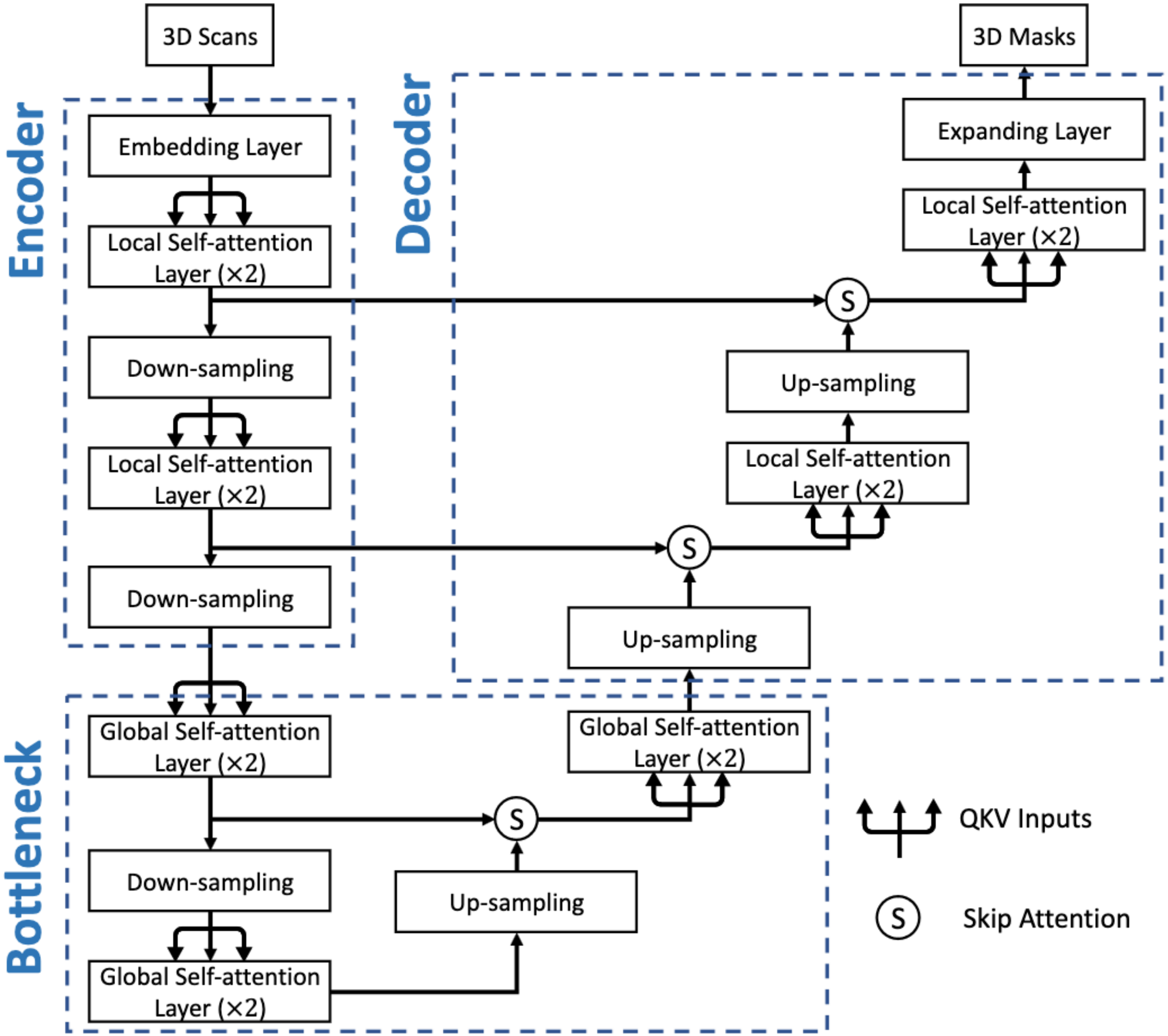





[10] Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang. Arxiv, May 2021.

# nn-Former [11]

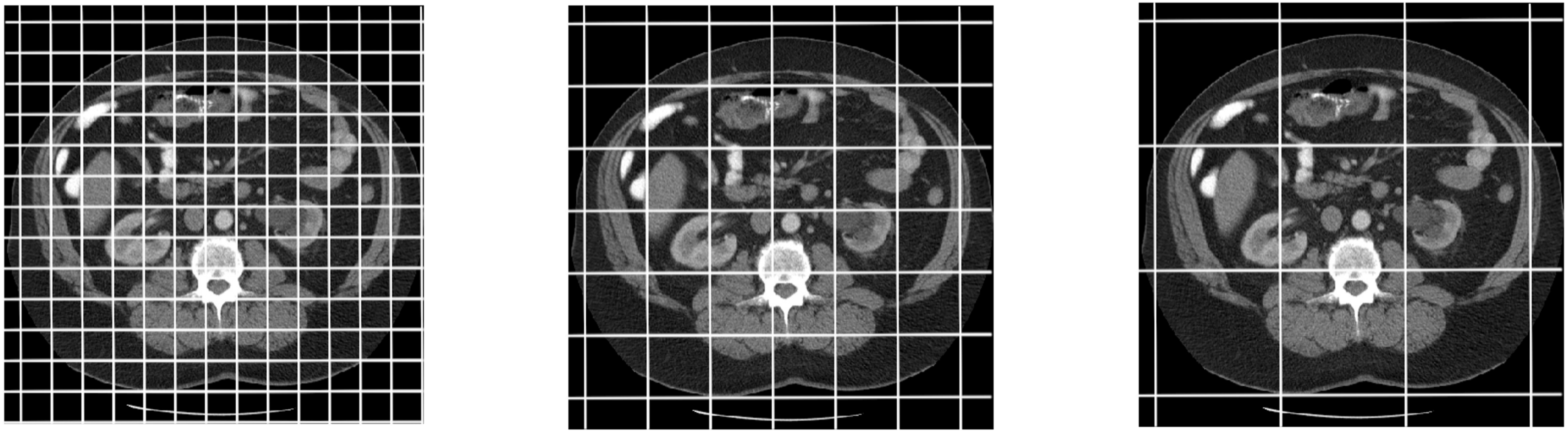
- Global self-attention in bottleneck
- Local self-attention in higher-resolution feature maps
  - ~ 3D Swin-UNet




[11] nnFormer: Interleaved Transformer for Volumetric Segmentation. H.Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu. Arxiv, September 2021.

# Multi-resolution transformers: limitations

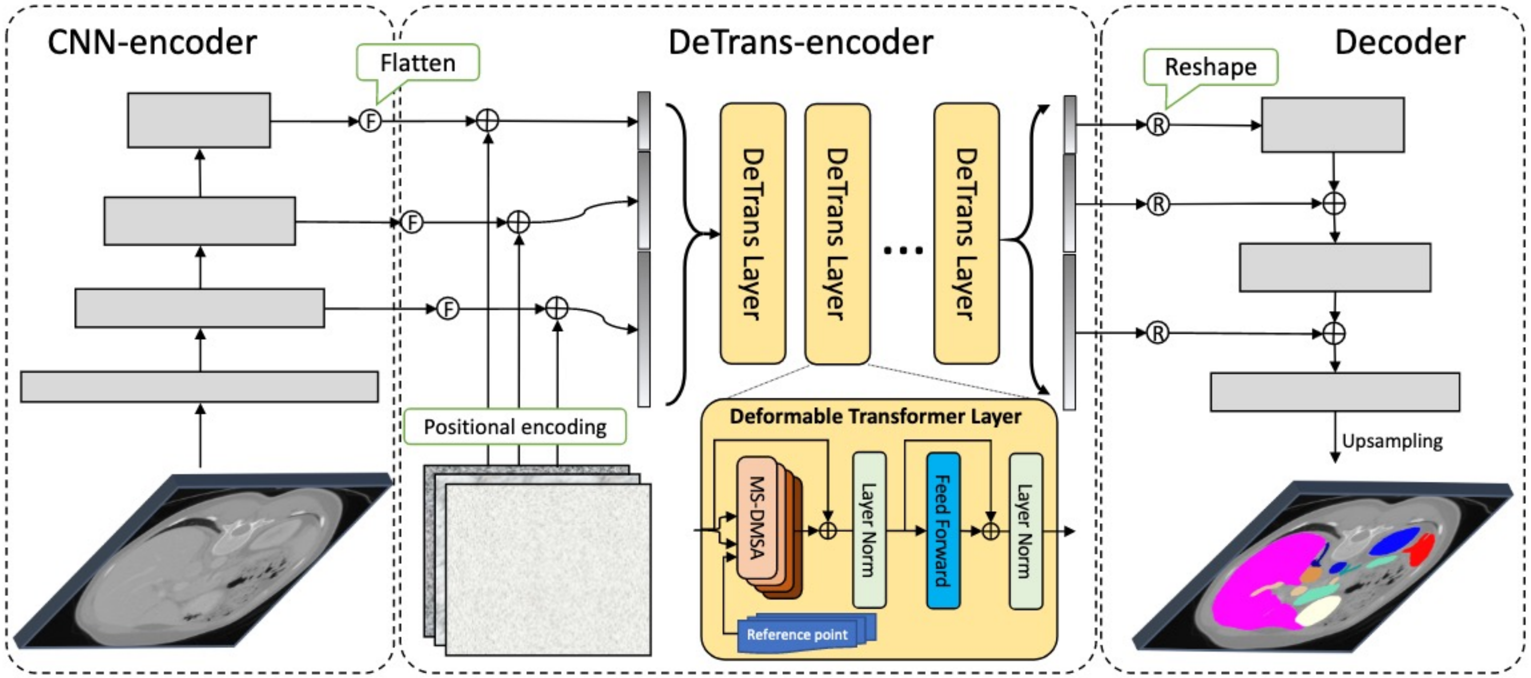
- **Windowed transformers** designed to reduce the complexity, e.g. Swin
  - **BUT:** no more long-range attention for high resolution feature maps



***Windowed input at different hierarchy levels***

# CoTR: **Co**nvolutional NN and **Tr**ansformer [12]

- **CoTr:** Conv encoder => flattened multi-scale feature
  - Deformable transformer encoder (DeTrans) in multi-res input
  - Several DeTrans layers, sent to conv decoder



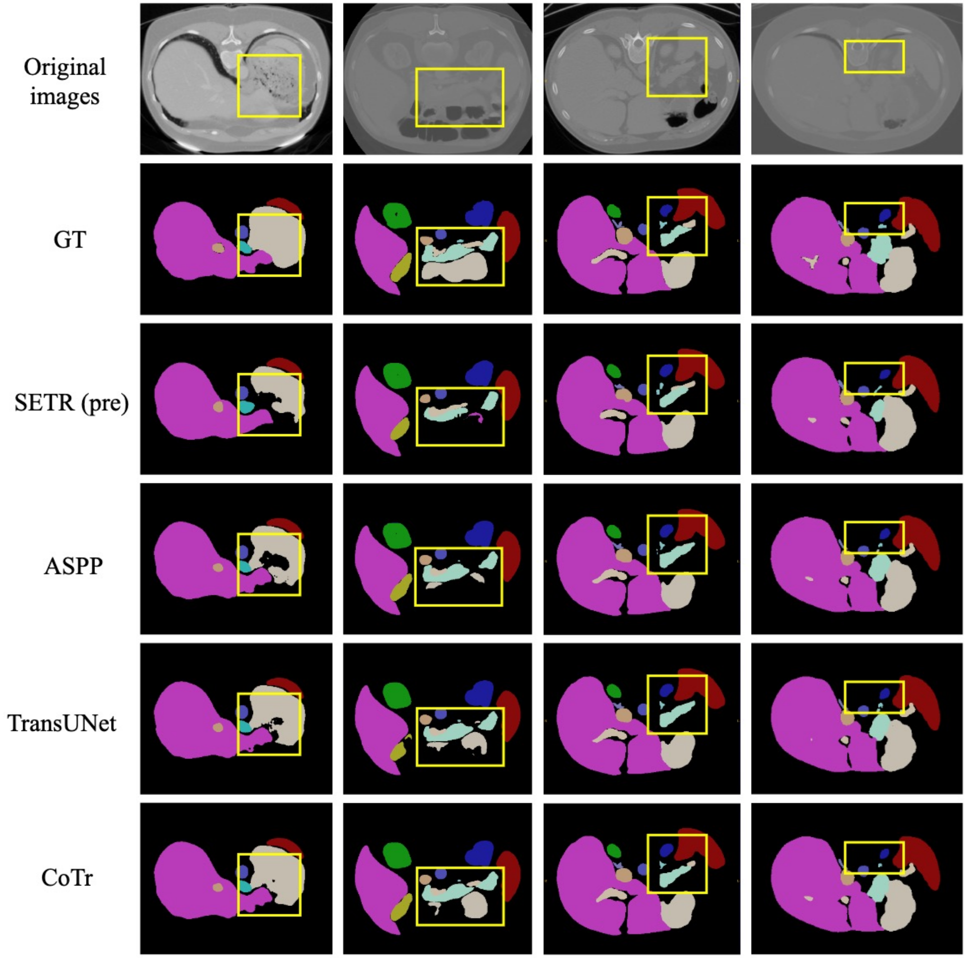

[12] CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. Y. Xie, J. Zhang, C. Shen, Y. Xia. MICCAI 2021

# CoTR: Convolutional NN and Transformer [12]

- Good performances on several datasets
- Deformable attention => reasonable to train

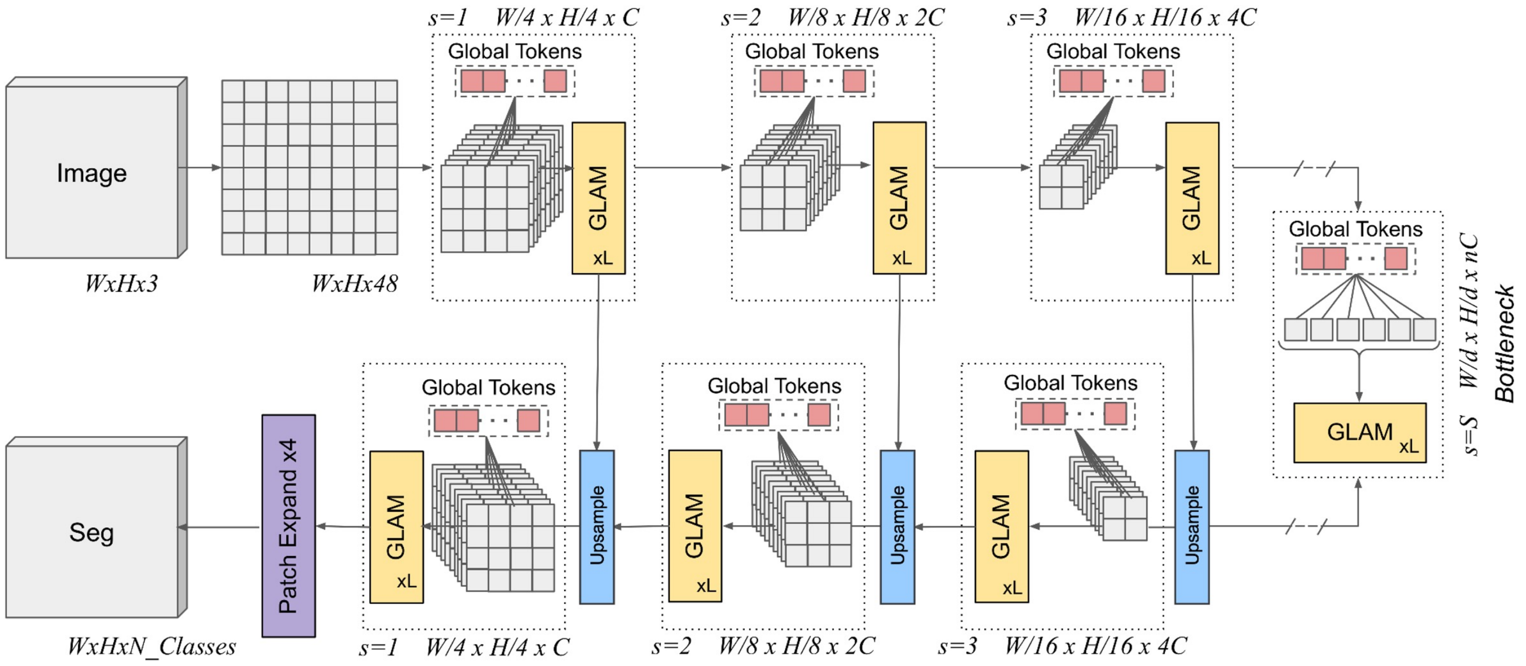
**Table 1.** Dice scores of our CoTr and several competing methods on the BCV test set. **CoTr*** and **CoTr†** are two variants of CoTr with small CNN-encoders

| Methods | Param (M) | Organs: Sp | Ki | Gb | Es | Li | St | Ao | IVC | PSV | Pa | AG | Ave |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SETR (ViT-B/16-rand) [27] | 100.5 | 95.2 | 92.3 | 55.6 | 71.3 | 96.2 | 80.2 | 89.7 | 83.9 | 68.9 | 68.7 | 60.5 | 78.4 |
| SETR (ViT-B/16-pre) [27] | 100.5 | 94.8 | 91.7 | 55.2 | 70.9 | 96.2 | 76.9 | 89.3 | 82.4 | 69.6 | 70.7 | 58.7 | 77.8 |
| CoTr w/o CNN-encoder | 21.9 | 95.2 | 92.8 | 59.2 | 72.2 | 96.3 | 81.2 | 89.9 | 85.1 | 71.9 | 73.3 | 61.0 | 79.8 |
| CoTr w/o DeTrans | 32.6 | 96.0 | 92.6 | 63.8 | 77.9 | 97.0 | 83.6 | 90.8 | 87.8 | 76.7 | 81.2 | 72.6 | 83.6 |
| APSS [5] | 45.5 | 96.5 | 93.8 | 65.6 | 78.1 | 97.1 | 84.0 | 91.1 | 87.9 | 77.0 | 82.6 | 73.9 | 84.3 |
| PP [26] | 33.9 | 96.1 | 93.1 | 64.3 | 77.4 | 97.0 | 85.3 | 90.8 | 87.4 | 77.2 | 81.9 | 72.8 | 83.9 |
| Non-local [20] | 32.8 | 96.3 | 93.7 | 64.6 | 77.9 | 97.1 | 84.1 | 90.8 | 87.7 | 77.2 | 82.1 | 73.3 | 84.1 |
| TransUnet [4] | 43.5 | 95.9 | 93.7 | 63.1 | 77.8 | 97.0 | 86.2 | 91.0 | 87.8 | 77.8 | 81.6 | 73.9 | 84.2 |
| **CoTr*** | 27.9 | 96.4 | 94.0 | 66.2 | 76.4 | 97.0 | 84.2 | 90.3 | 87.6 | 76.3 | 80.8 | 72.9 | 83.8 |
| **CoTr†** | 36.9 | 96.2 | 93.8 | 66.5 | 78.6 | 97.1 | 86.9 | 90.8 | 87.8 | 77.7 | 82.8 | 73.2 | 84.7 |
| **CoTr** | 41.9 | 96.3 | 93.9 | 66.6 | 78.0 | 97.1 | 88.2 | 91.2 | 88.0 | 78.1 | 83.1 | 74.1 | **85.0** |



# Global attention in multi-resolution transformers (GLAM) [13]

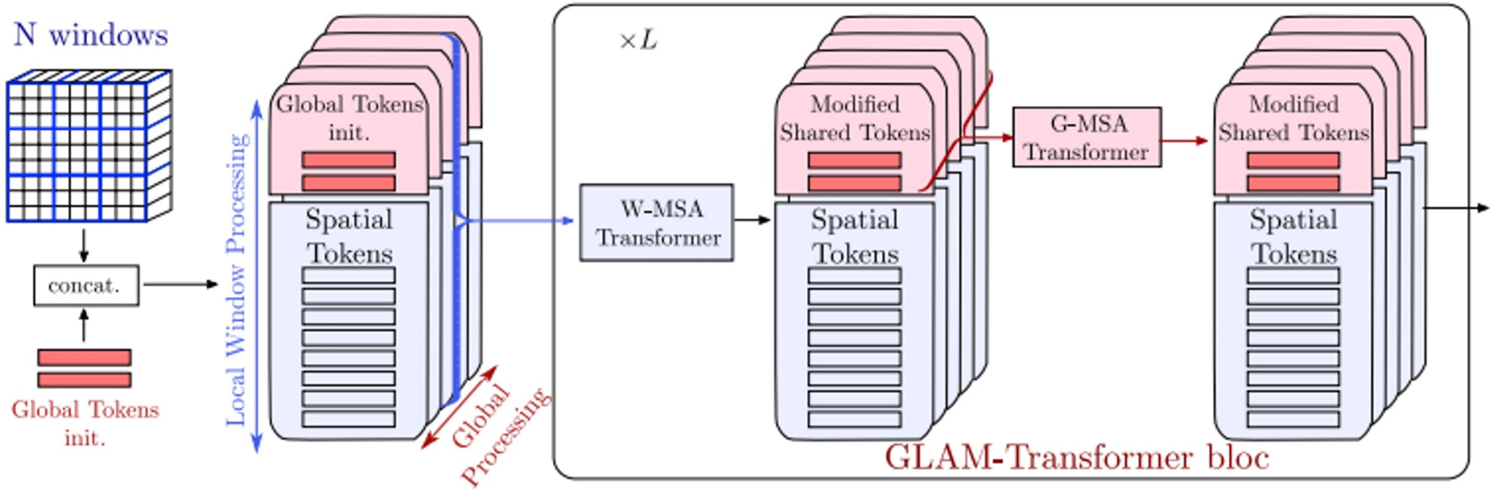
- Architecture based on hierarchical transformer (e.g. Swin, nn-Former)
  - Can also be included in any multi-resolution model (e.g. Conv)
  - GLAM Motivation: Full attention even in high-resolution features




[13] Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation. L. Themyr, C. Rambour, N. Thome, T. Collins, A. Hostettler. WACV 2023.

# GLAM block

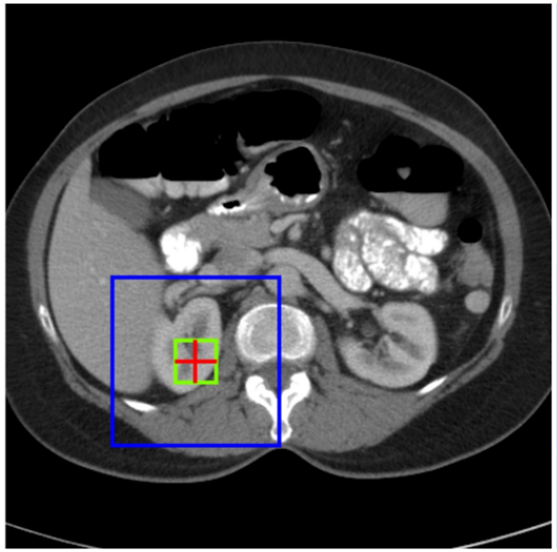
- Define learnable global tokens in each window, cf CLS in VIT
  - Window self-attention (W-MSA): attention between visual and global tokens
  - Global attention (G-MSA) between global token
- G-MSA: indirection between all visual tokens
  - Break computational complexity of full attention between visual token
  - But enables full indirect interaction between them

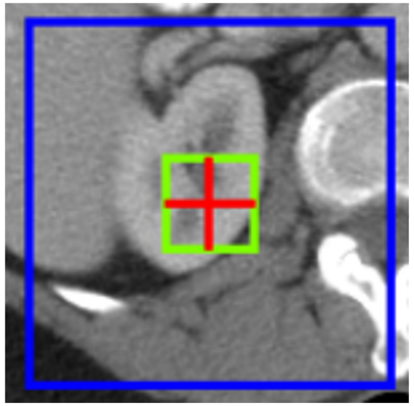


# FINE : Full resolutIoN mEmory transformer module [14]

- Extends GLAM for full context modelling in 3D segmentation
- Reminder: state-of-the-art methods based on 3D crops

- Original image size: **512x512x256**
- Cropped patch size: **128x128x64**





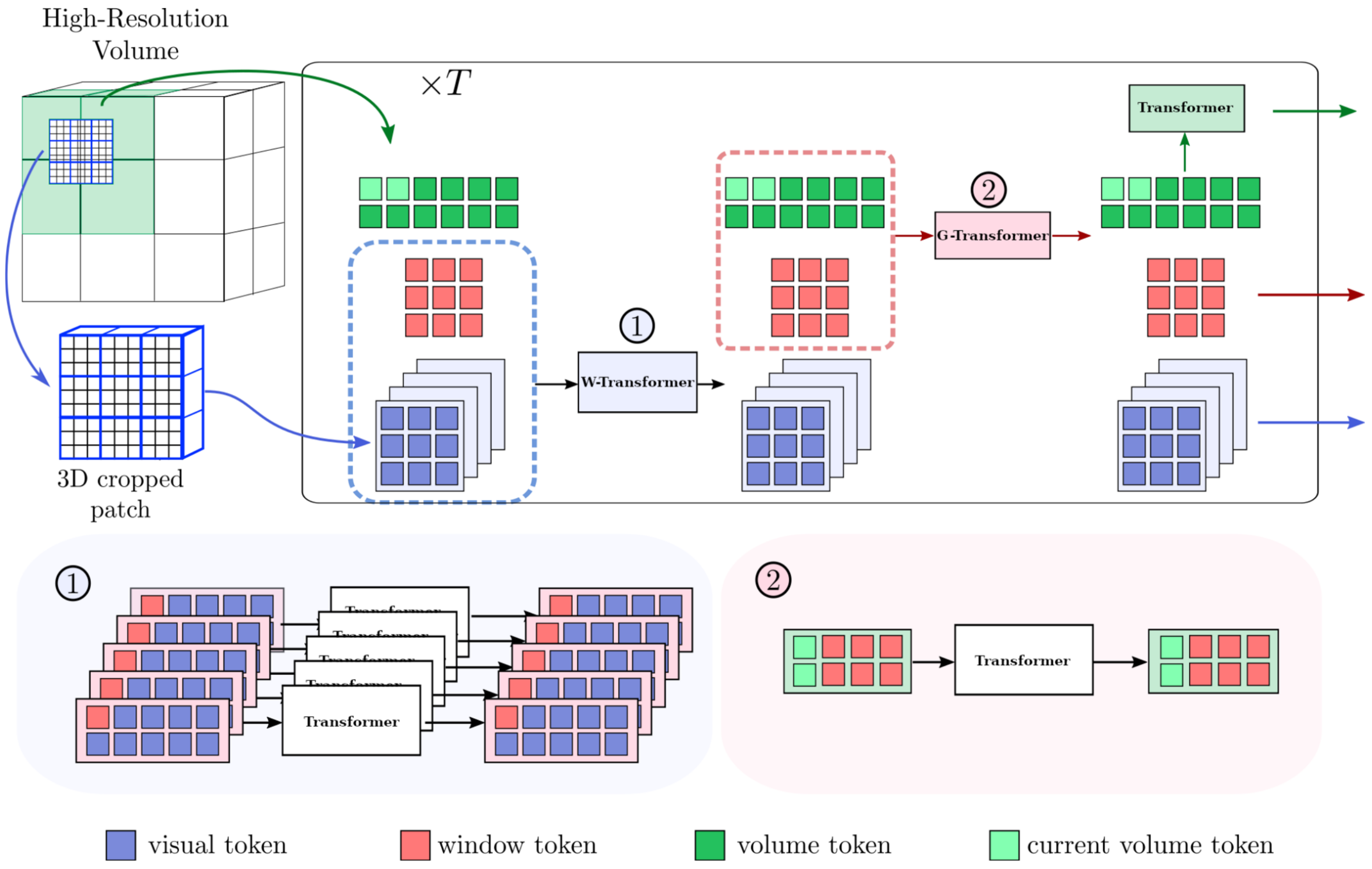
**<u>Goal:</u> learning a global representation of the full volume** from batch training with crops

[14] Memory transformers for full context and high-resolution 3D Medical Segmentation. L. Themyr, C. Rambour, N. Thome, T. Collins, A. Hostettler. MLMI workshop, MICCAI 2022.

# FINE architecture

- **2 levels of global tokens:**
    - Window tokens (red)
    - Volume tokens (green)
- W-transformer in 3D crops
- G-transformer between window and volume token

=> (indirect) **full interaction between all voxels!**



# Results

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

7 classes

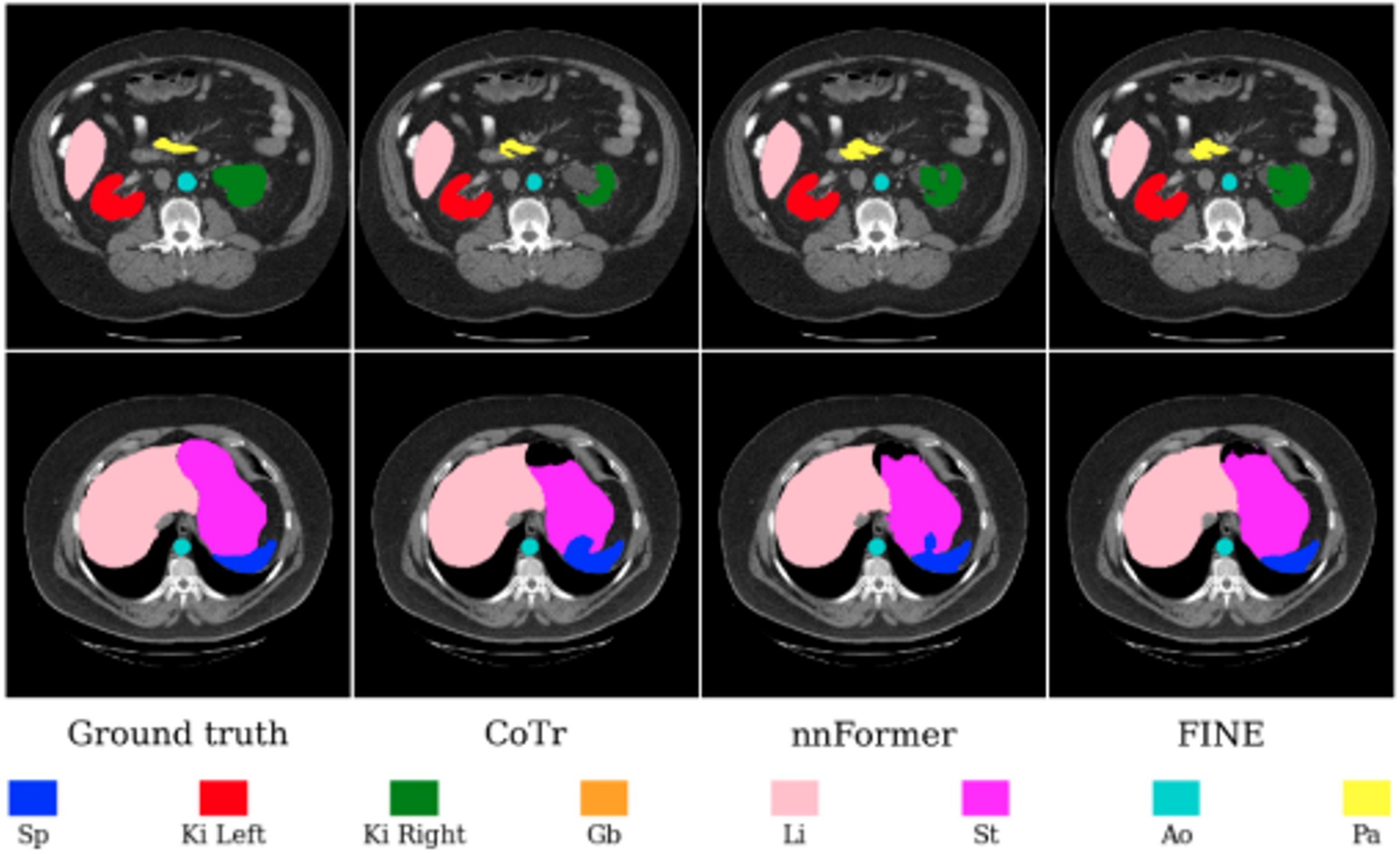
30 volumes

**Metrics :**

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

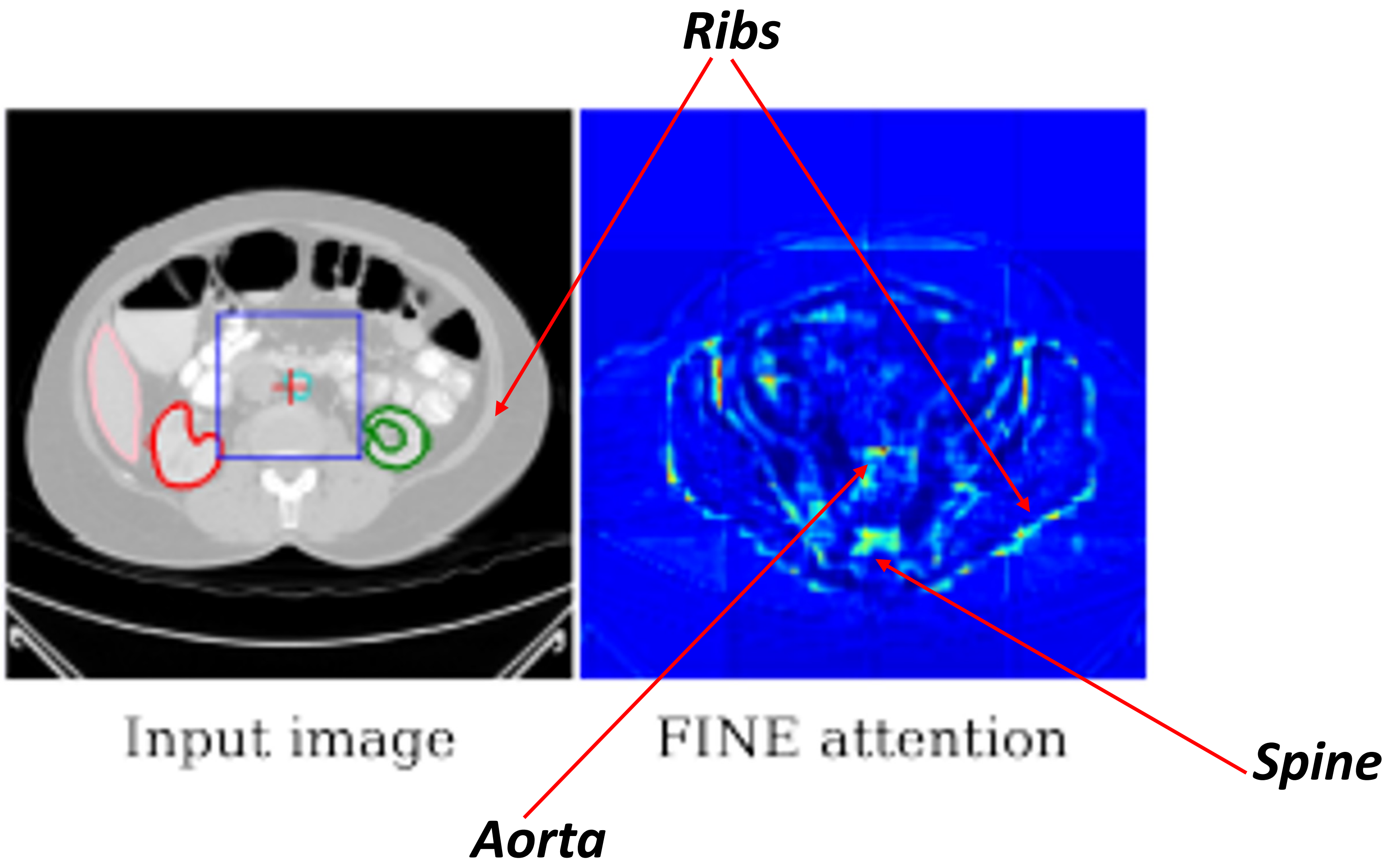
| Method | Average | | Per organ dice score (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | HD95 | DSC | Sp | Ki | Gb | Li | St | Ao | Pa |
| UNet [24] | - | 77.4 | 86.7 | 73.2 | 69.7 | 93.4 | 75.6 | 89.1 | 54.0 |
| AttUNet [19] | - | 78.3 | 87.3 | 74.6 | 68.9 | 93.6 | 75.8 | 89.6 | 58.0 |
| 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



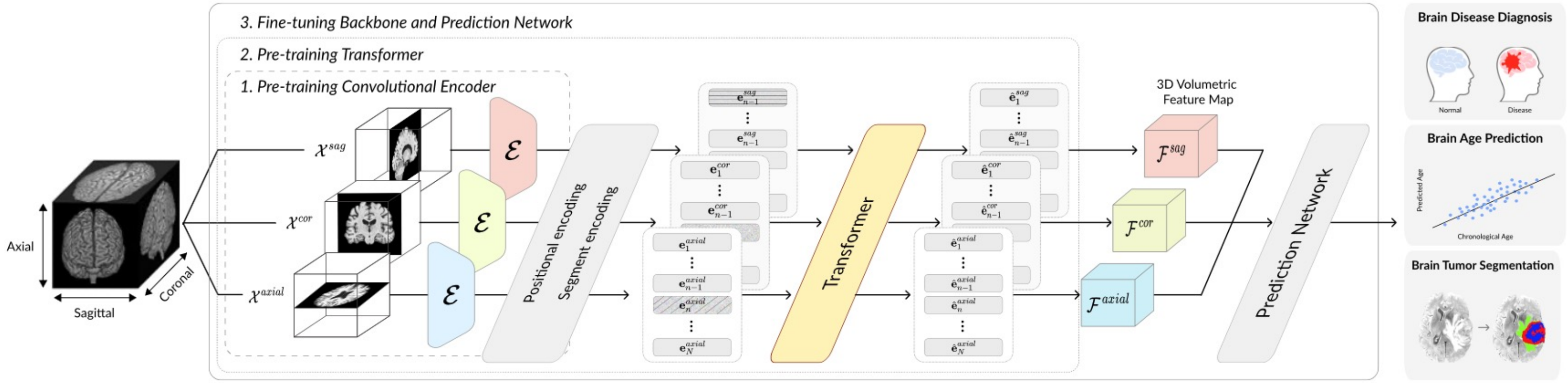
1. Transformers
2. Vision Image Transformer
3. Transformers for medical image segmentation
4. Current trend & Perspectives

# Transformer in medical image analysis

- Key feature: self-attention
  - Long-range dependencies, global context
  - Potential in image segmentation: best of both words between accurate info and full context
  - Challenge: full attention computation
- Transformer used in several medical image analysis tasks: Image Registration, Image Generation, Im-2-im translation
- Discussion and perspectives
  - Self-supervised learning
  - Multi-task learning, Multi-modal learning
  - Foundation models

# Self-supervised learning and transformers

- The way transformers have been trained in NLP: pretext task
  - Predict masked word (BERT ), next word (GPT), etc
- Several pretext tasks in vision
  - Pretext tasks (RotNet, MAE), contrastive methods (BYOL, MoCO)
- In medical image analysis: pre-train on generalist or medical images [15]

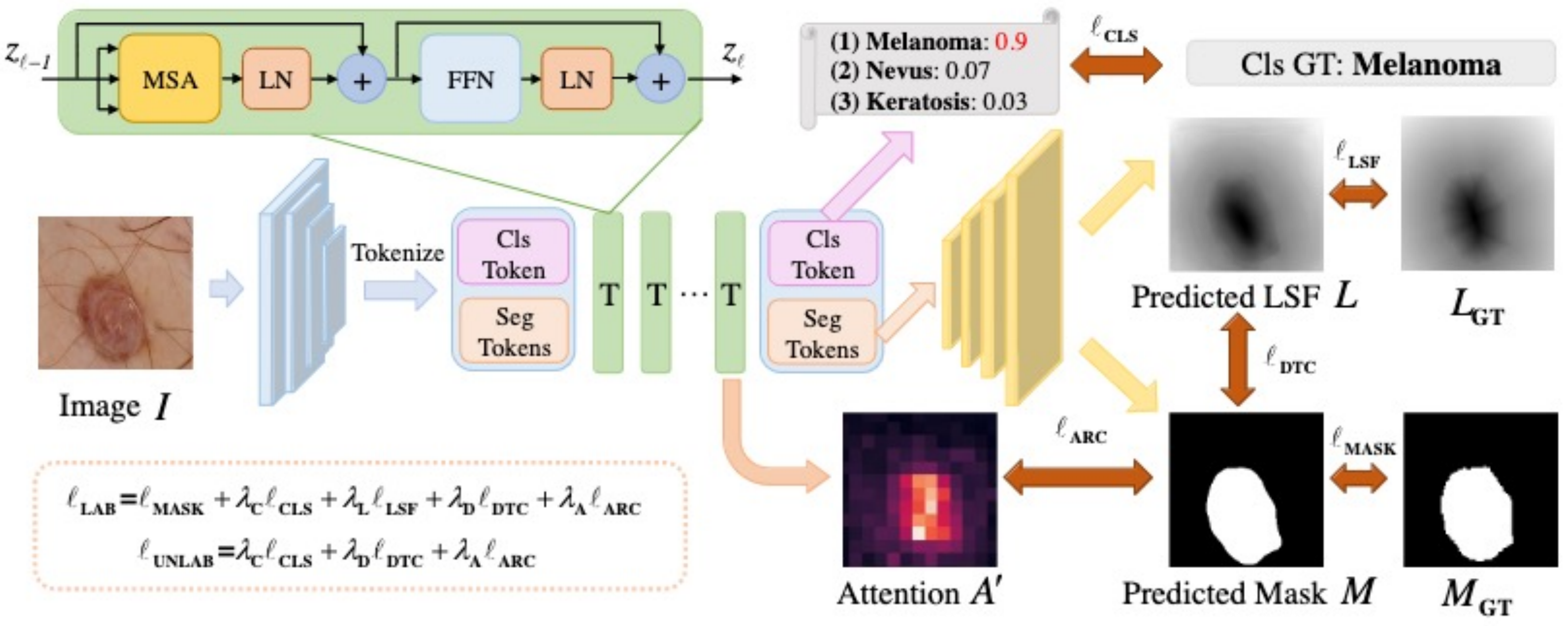



[15] Medical Transformer: Universal Brain Encoder for 3D MRI Analysis. E Jun, S Jeong, DW Heo, HI Suk. Arxiv, 2022.

- Generally leads to better OOD robustness

# Multi-task learning

- Usual to combine tasks
  - *e.g.* classification and segmentation in medical images [16]

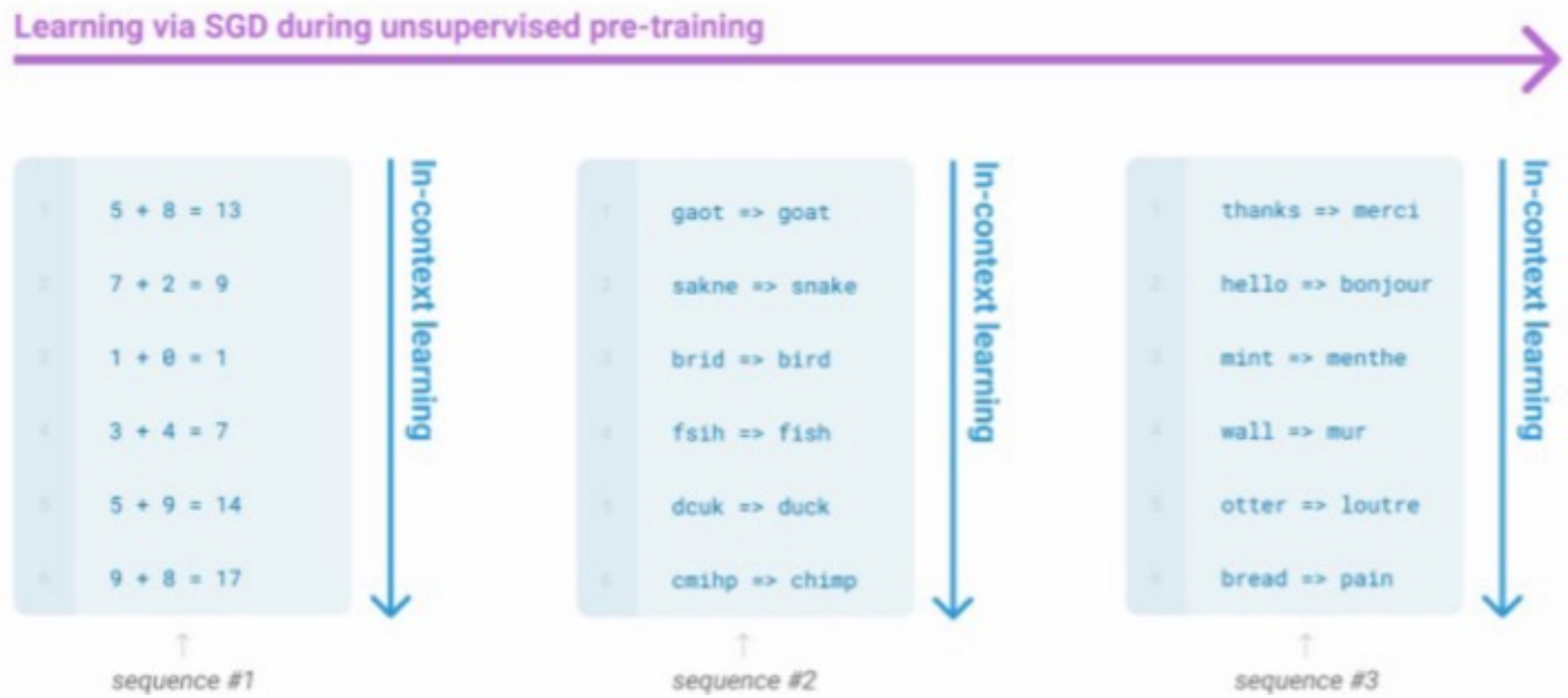



[16] MT-TransUNet: Mediating Multi-Task Tokens in Transformers for Skin Lesion Segmentation and Classification. J. Chen, J. Chen, Z. Zhou, B. Li, A. Yuille, Y. Lu. Arxiv, 2021.

# Multi-task learning & foundation models

Current trend: Train huge transformers, e.g. GPT-3/GPT-4 in NLP

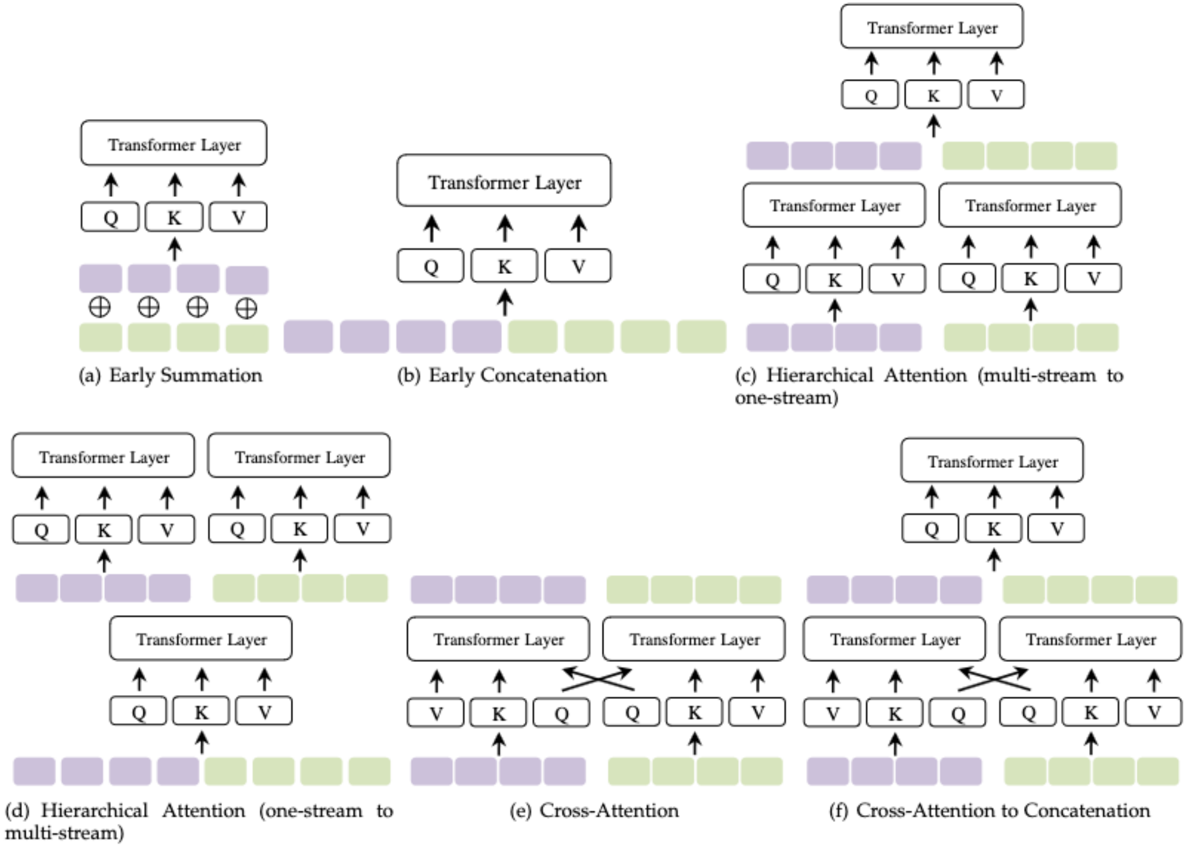
- General-purpose AI, can be fine-tuned on several tasks **=> foundation model**
- Trained on diverse datasets, predict next word
- Prompted ("in-context learning") with emerging properties
  - Can beat even model fine-tuned for the target task (*e.g.* translating to English)
  - Not fully understood



# Multi-modal learning & foundation models

Transformers naturally handle multi-modal data: **token homogeneity**

- **Different goals depending on the task [17]**
  - **Fusion:** complementarity between models
  - **Alignment:** making modalities closer
- Multi-modal: can also be used as a self-supervised signal

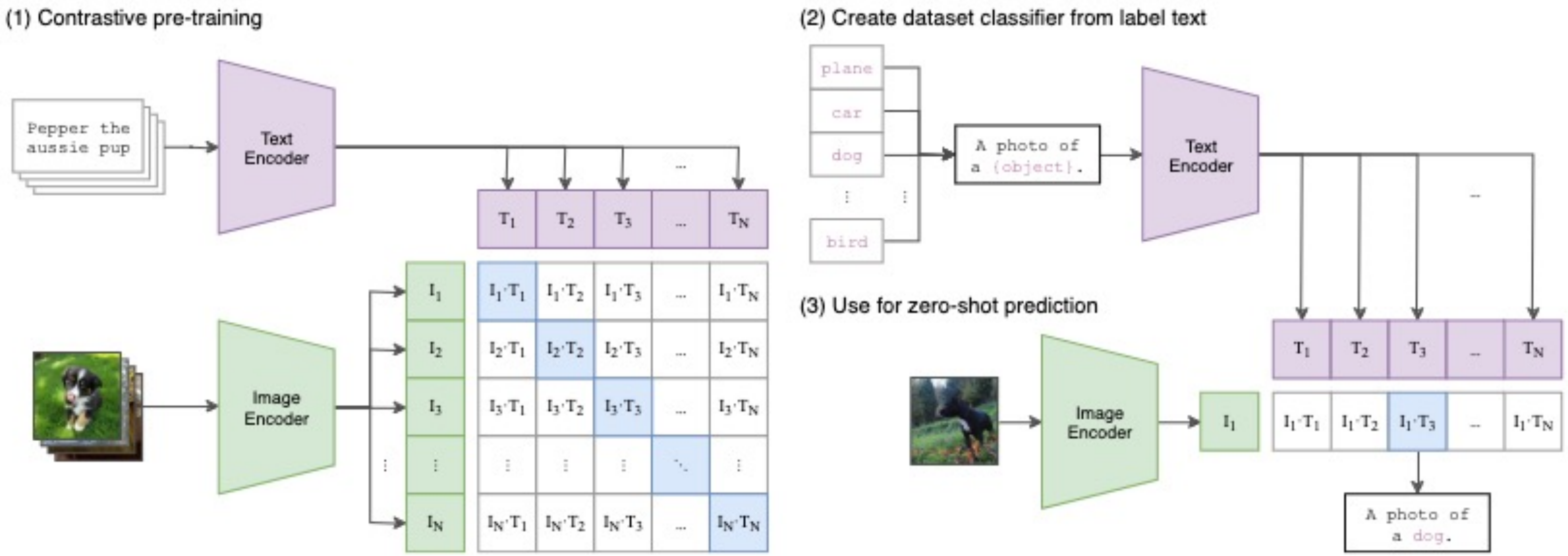



[17] Multimodal Learning with Transformers: A Survey. P. Xu, X. Zhu, D. A. Clifton. Arxiv, 2022.

# Multi-modal learning & foundation models

Multi-modal foundation models, e.g. NLP and images:

- Contrastive Language-Image Pre-training (CLIP): image/ text encoder, alignment




[17] Learning Transferable Visual Models From Natural Language Supervision. 1. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. ICML 2021

# Multi-modal learning & foundation models

## Multi-modal foundation models, e.g. NLP and images:

### DALL-E [19]: image decoder


[18] . Zero-Shot Text-to-Image Generation A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever. ICML 2020.

### Flamingo [18]: text decoder

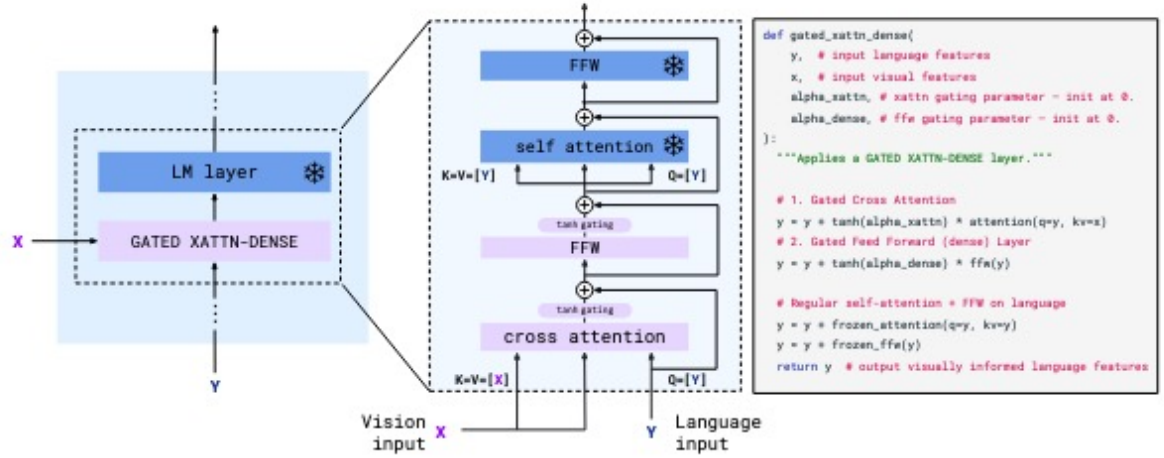



Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

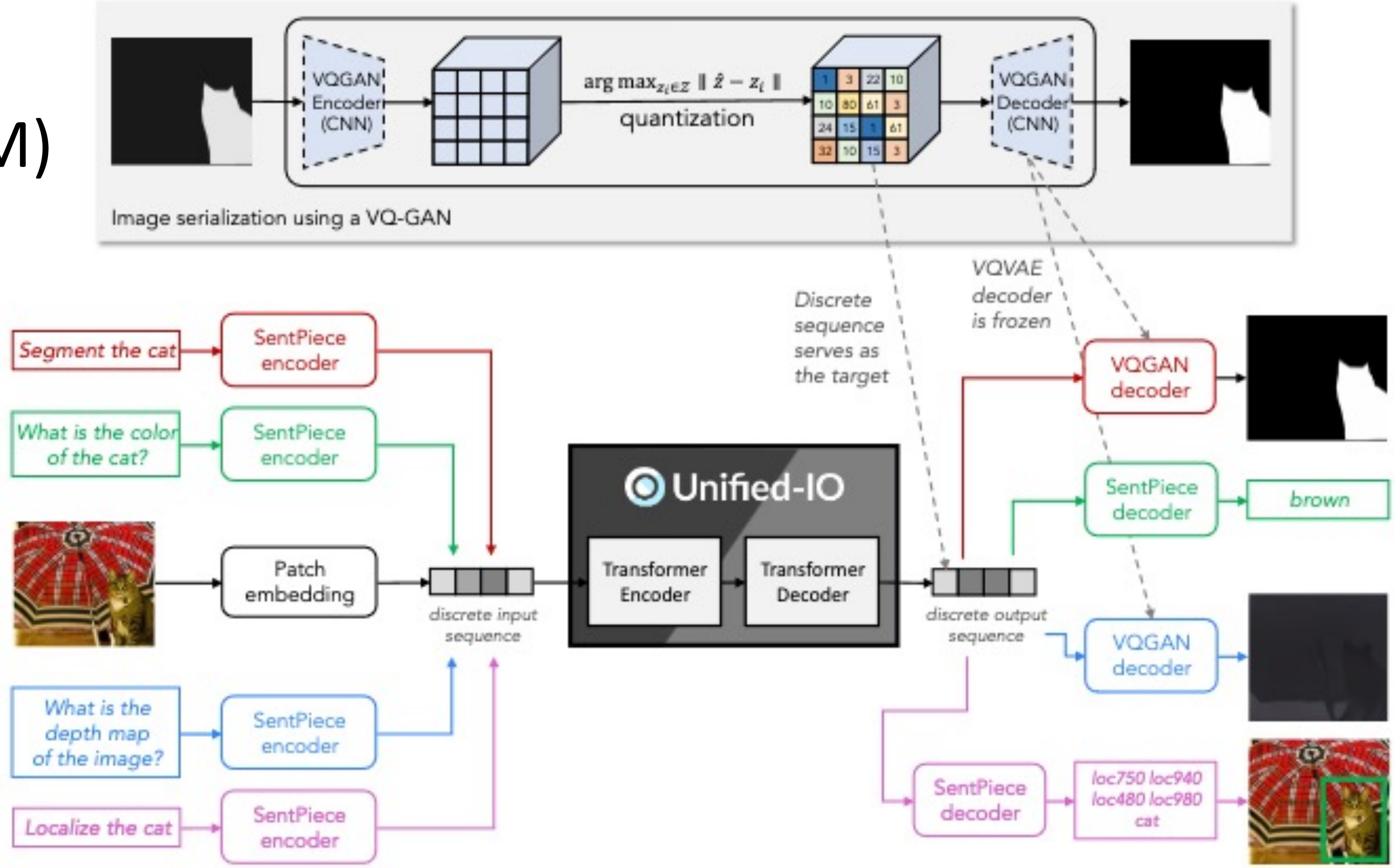
[18] Flamingo: a Visual Language Model for Few-Shot Learning. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan. . NeurIPS 2022

# Foundation models

**Combination of multi-modal and multi-task learning**

- Unified-IO [20]
- Segment Anything Model (SAM) [21]

Prompt it with interactive points and boxes.

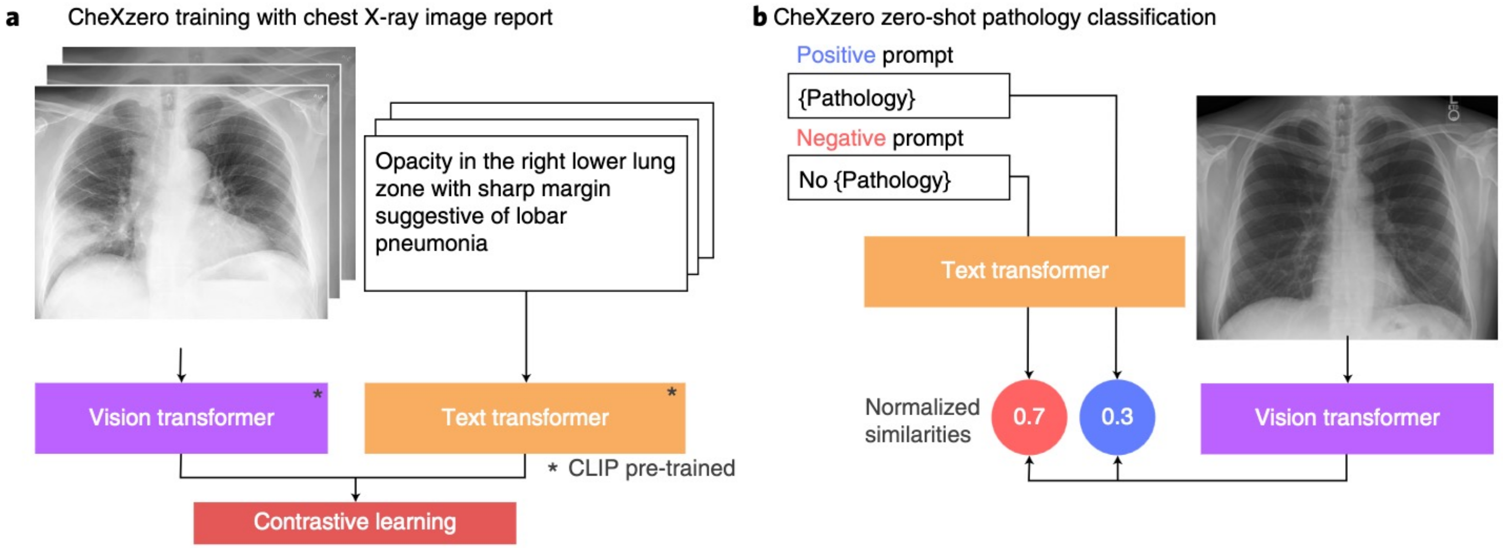



[21] Segment Anything. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.Y. Lo, P. Dollár, R. Girshick. Arxiv, 2023.

[20] Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. J. Lu, C. Clark, R. Zellers, R. Mottaghi, A. Kembhavi. ICLR 2023

# Foundation model in healthcare

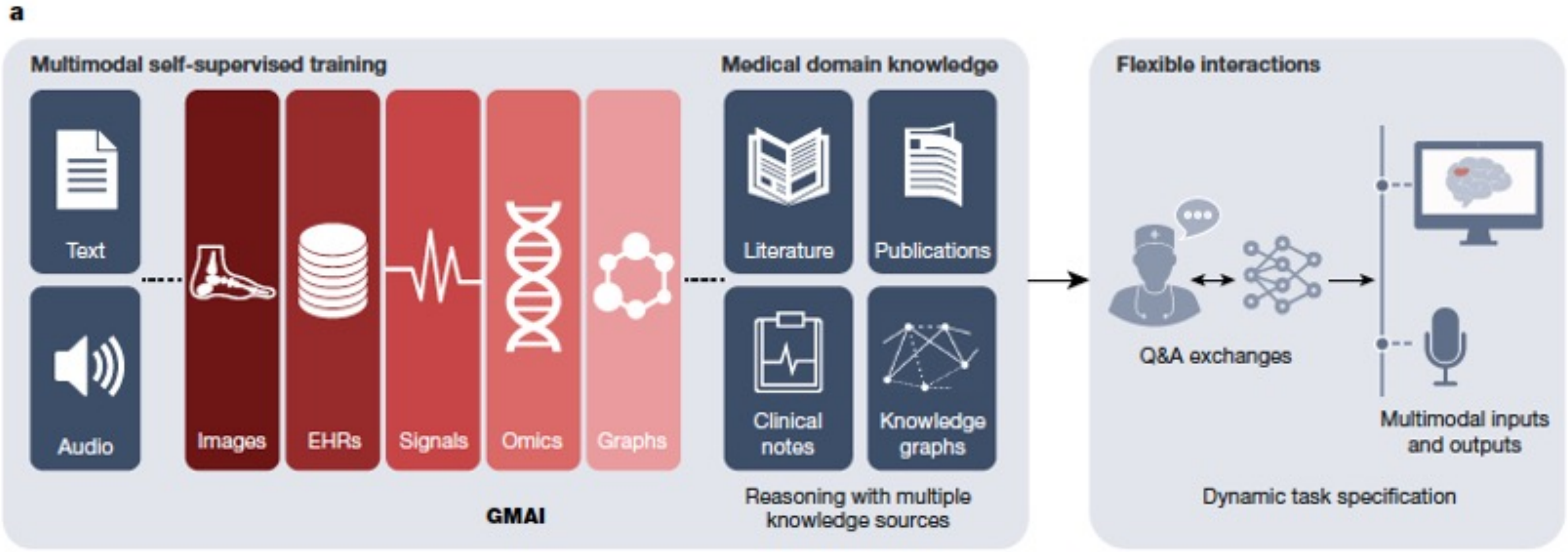
- CheXzero [22]: POC of CLIP-based model




[22] Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng (2022). E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar. Nature Biomedical Engineering volume, 2022

# Foundation models: towards generalist medical AI? [23]

- Solve more diverse and challenging tasks than current medical AI models
- Relaxing the need for labels in specific tasks.
- Potential of foundation models:
  - Flexible and dynamic interactions
  - Multi-modal inputs and outputs
  - Medical domain knowledge, more elusive?






[23] Foundation models for generalist medical artificial intelligence. M. Moor, O. Banerjee, Z.S.H. Abad, H. M. Krumholz, J. Leskovec, E.J. Topol, P. Rajpurkar. Nature volume 616, pages 259–265, 2023.

# Foundation models: towards generalist medical AI? [22]

- Important potential applications


- In-context learning for effective adaptability?

For example, a clinician might say, "Check these chest X-rays for Omicron pneumonia. Compared to the Delta variant, consider infiltrates surrounding the bronchi and blood vessels as indicative signs"[40].

# Foundation models: risks and challenges

- Access to huge-scale datasets
  - Diverse, anonymized data
  - Pre-training on generalist data?
- Robustness and certification: uncertainty, OOD detection, stability, etc
  - A general issue in deep learning, exacerbated with general-purpose AI systems
  - Crucial and especially sensible in healthcare
- Explainability, interpretability: harder or easier?
- Ethical considerations
  - Biases and fairness/discriminability
  - Privacy, informed consent, transparency

# Thank you for your attention!

Questions?