Is the problem of medical image segmentation a thing of the past ?

by
Olivier Bernard
Professor – University of Lyon (INSA), France

July 10, 2024

CREATIS; CNRS (UMR 5220); INSERM (U1294); INSA Lyon; Université de Lyon, France

Resume

AI methods in cardiac image analysis

Acquisition

Ultrafast cardiac imaging

Convolutional NN
Realistic simulations

Image quantification

Segmentation
Tissue motion / blood flow
Uncertainty modeling
Domain adaptation

Convolutional NN
Variational Auto-Encoders
Physics informed NN
Diffusion networks

Population representation

Multi-modal fusion
Heterogenous data integration

Transformers

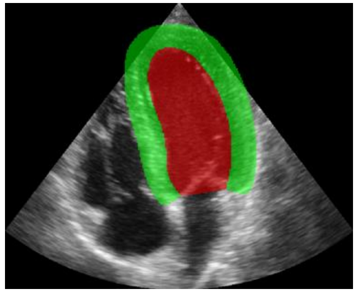
Etiology classification
Hypertension characterization

Robust estimation of existing / new biomarkers

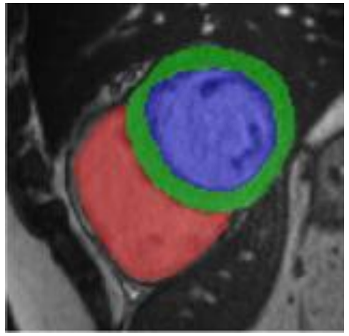
Cardiac imaging

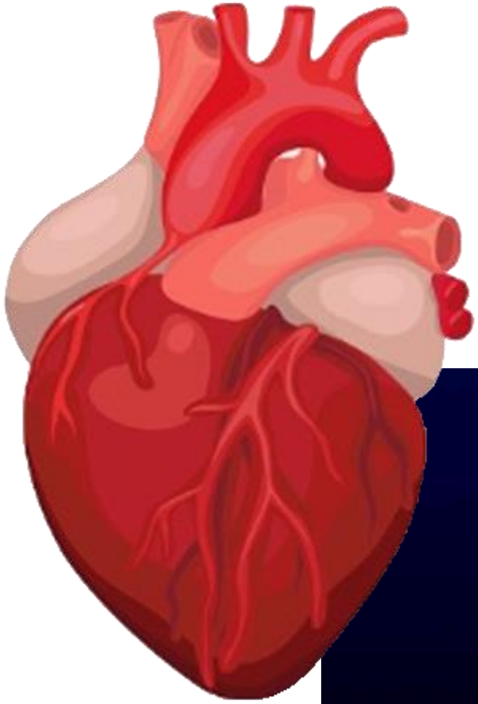
Quantification of clinical indices to diagnose cardiac pathologies

Echocardiography

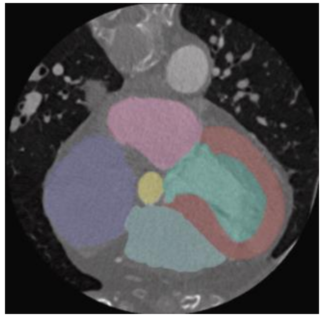


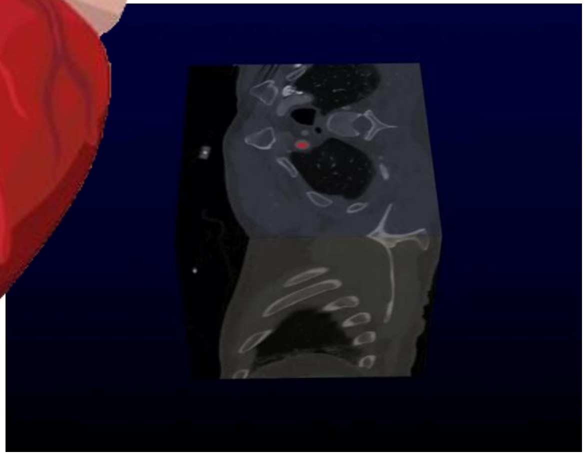
MRI





CT



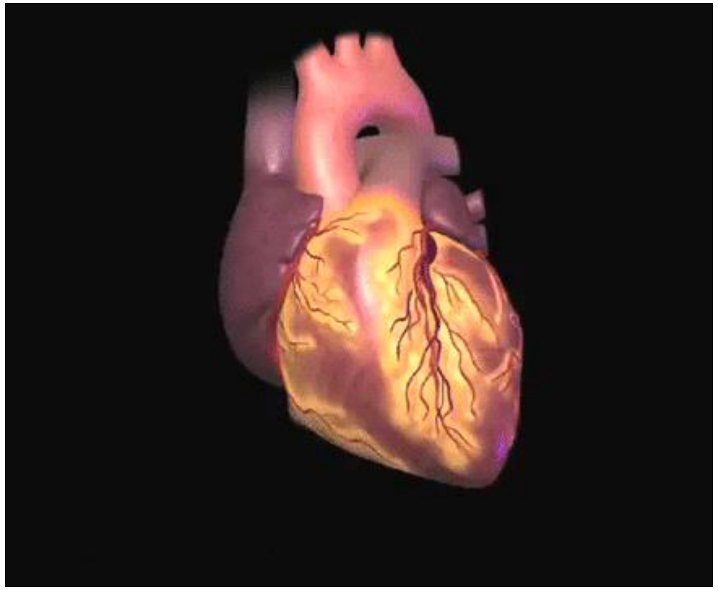


From the inHEART company website

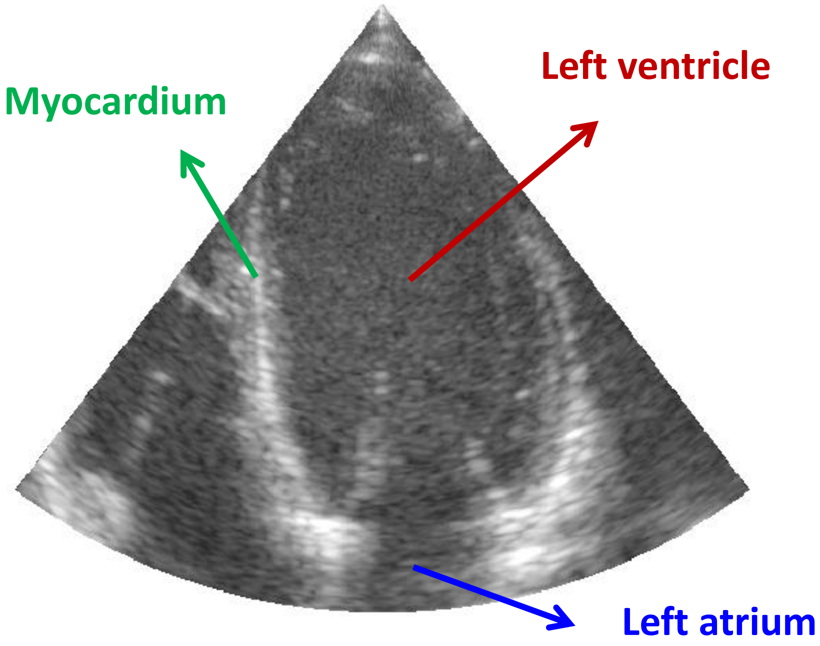
- ✓ High annotation costs
- ✓ Inter/intra expert variability
- ✓ Acquisition variabilities
- ✓ Acquisition artifacts

Quantification of clinical indices to diagnose cardiac pathologies

- Anatomical imaging



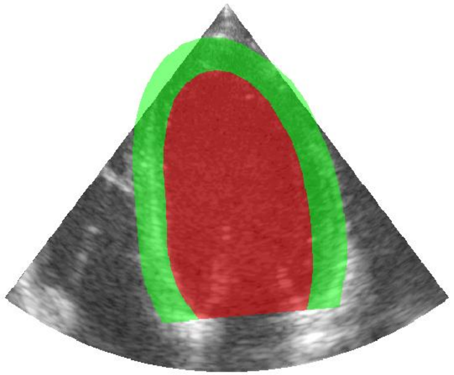
Source: GE Healthcare web site



Conventional ultrasound acquisition in clinical routine

Echocardiographic imaging

Quantification of clinical indices to diagnose cardiac pathologies



Automatic delineation of anatomical structures

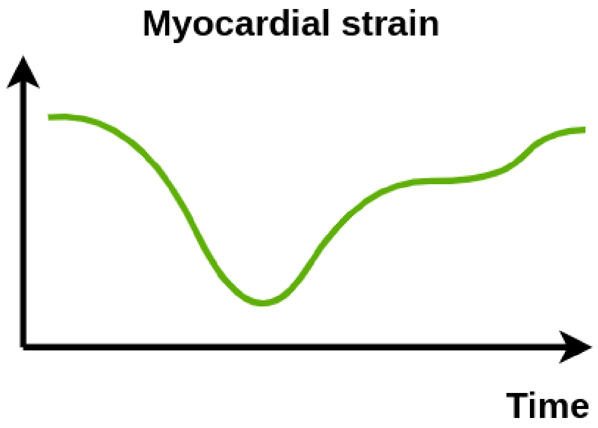
- Scalar descriptors
- Time-series descriptors

Scalar descriptors

- Myocardial mass
- Left ventricle ejection fraction

Time-series descriptors

- Left ventricle area
- Global longitudinal strain



Challenges

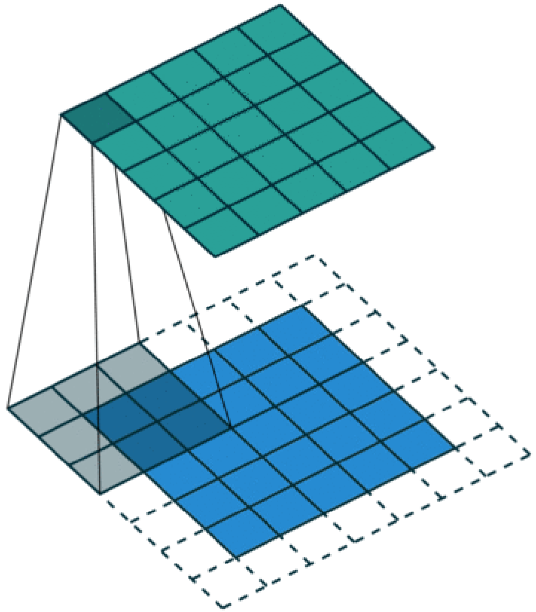
- How to make the measurements extracted from medical images automatic, reliable and accurate ?
- How to make these measurements reproducible at different centers, in different countries, whatever the expert ?

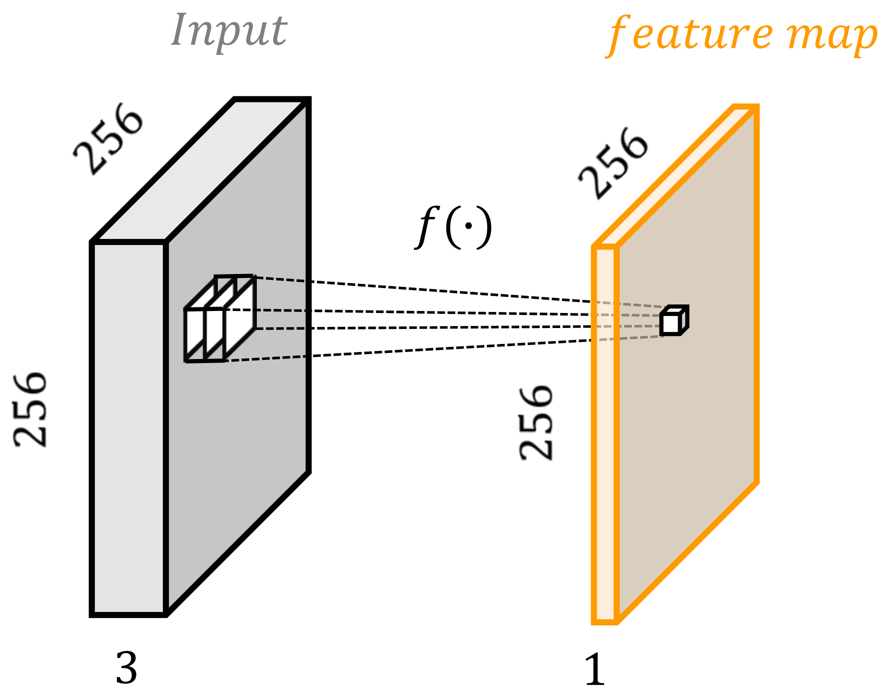
Deep learning families

Convolutional Neural Network

Convolutional layer

- ✓ Create relevant information called *feature map* (convolution + non linear function)
- ✓ Parameters that are learned during training



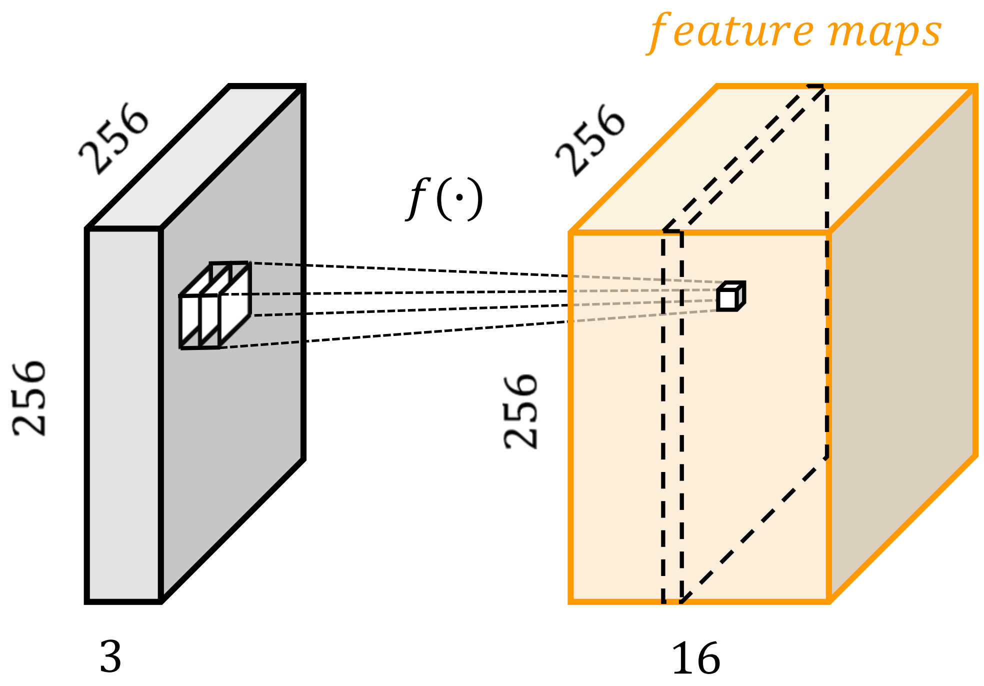


Filter of size 3×3

$$\# \, param = 3 \times 3 \times 3 + 1 = 28$$

Convolutional layer

- ✓ Create relevant information called *feature map* (convolution + non linear function)
- ✓ Parameters that are learned during training

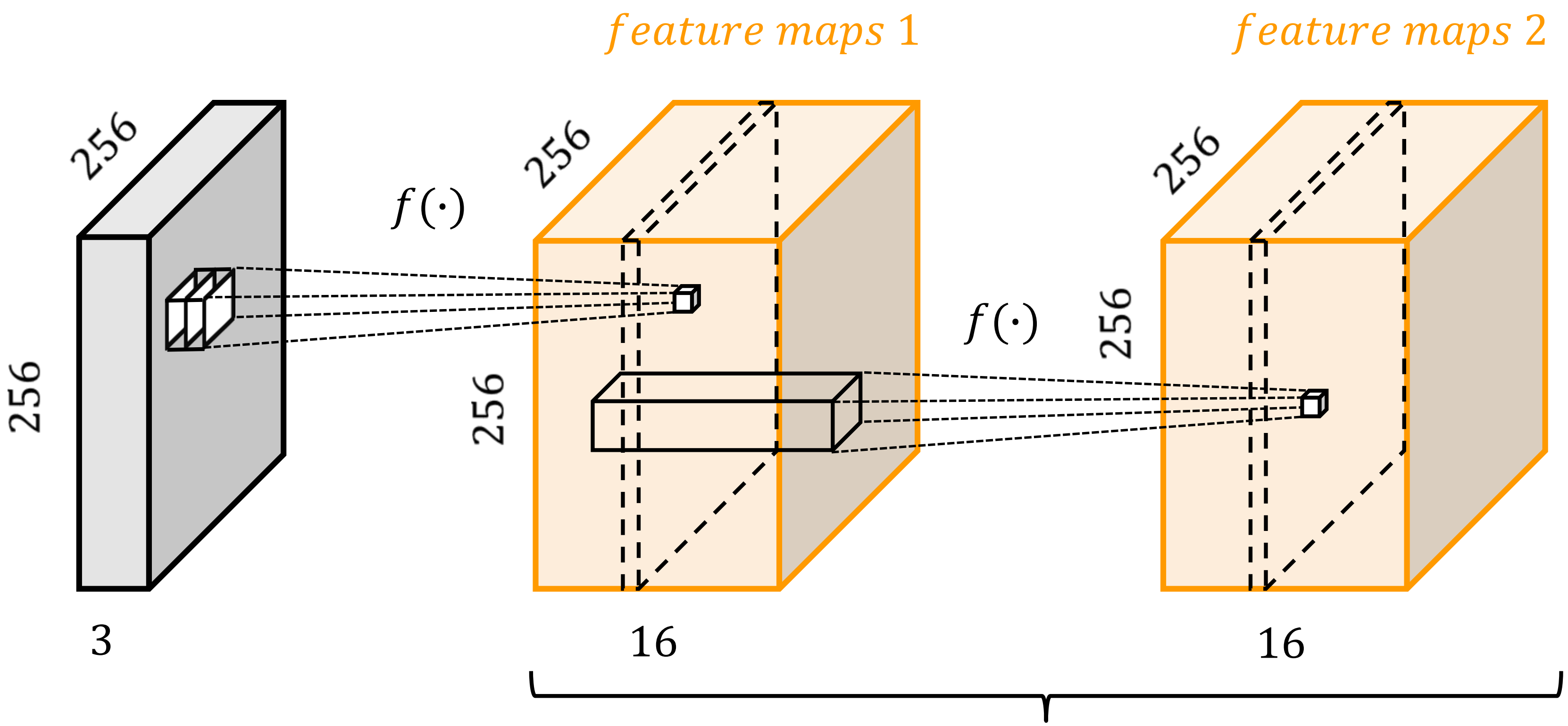


Filter of size 3×3

$$\# param = 16 \times (3 \times 3 \times 3 + 1) \\ = 448$$

Convolutional layer

- ✓ Create relevant information called *feature map* (convolution + non linear function)
- ✓ Parameters that are learned during training

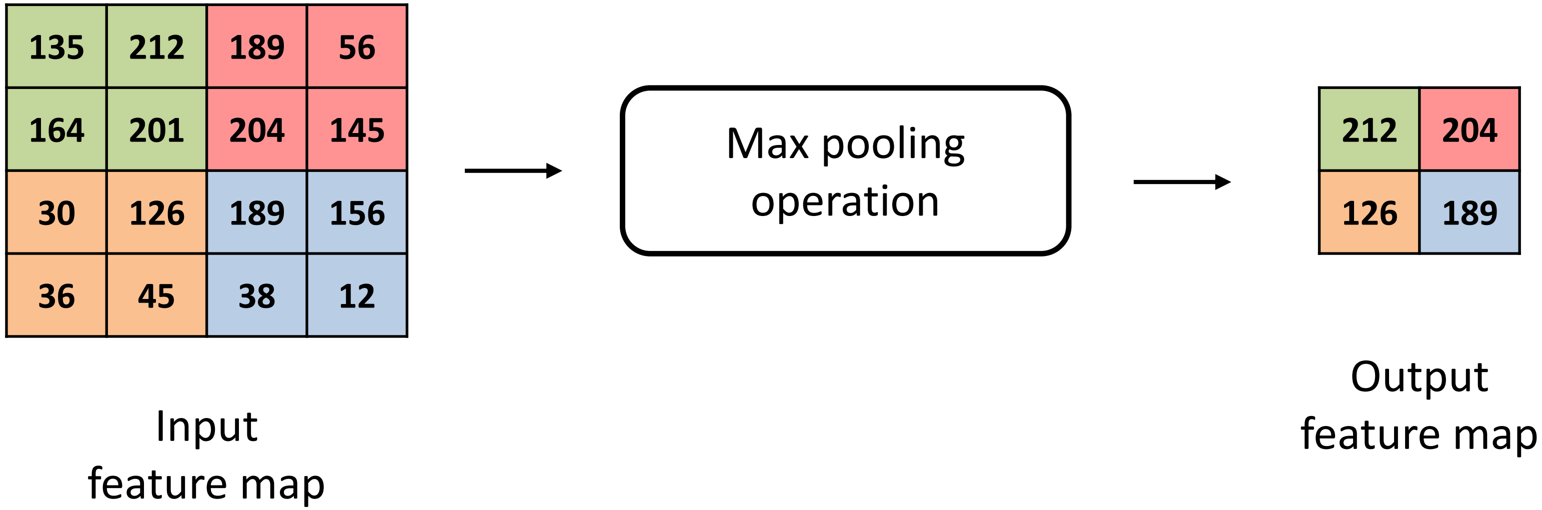


Filter of size 3×3

$\# \, param$
$= 16 \times (3 \times 3 \times 16 + 1)$
$= 2{,}320$

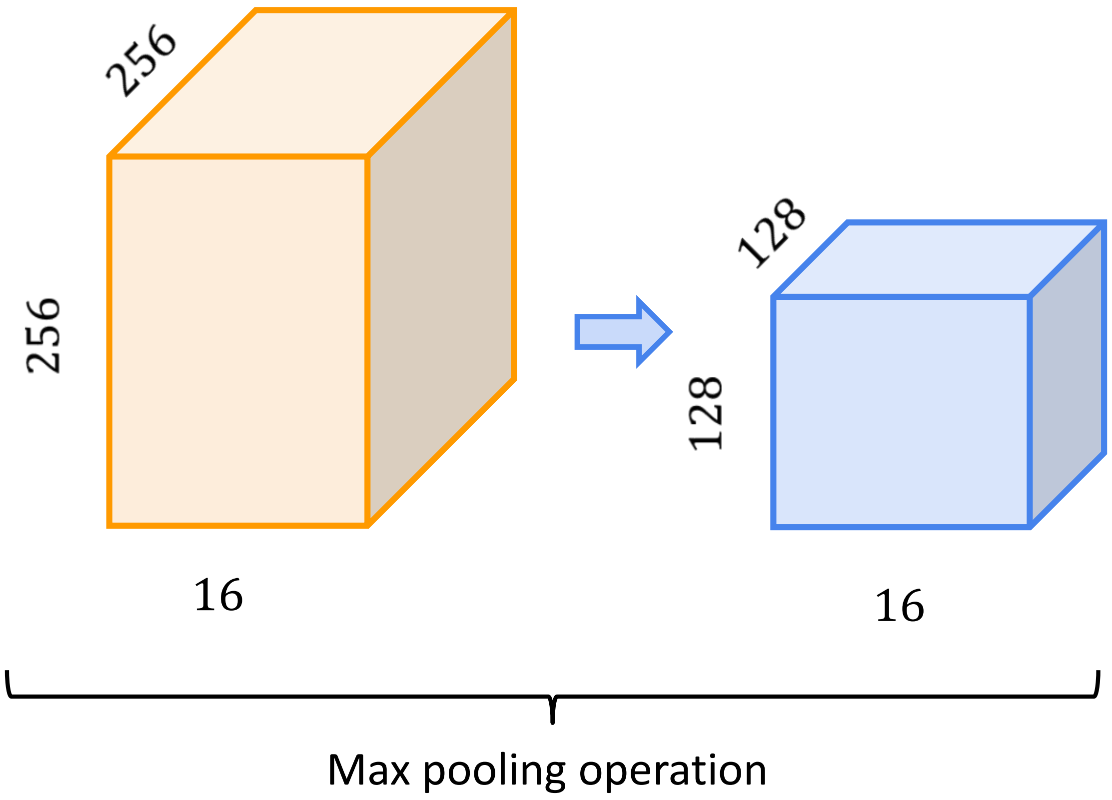
Pooling operation

- ✓ Concentrate information into lower dimensional space
- ✓ Applied individually to each feature map



Pooling operation

- ✓ Concentrate information into lower dimensional space
- ✓ Applied individually to each feature map



No parameter to train

Image encoding

- ✓ Learning to encode relevant information
- ✓ Projection to a lower dimensional space

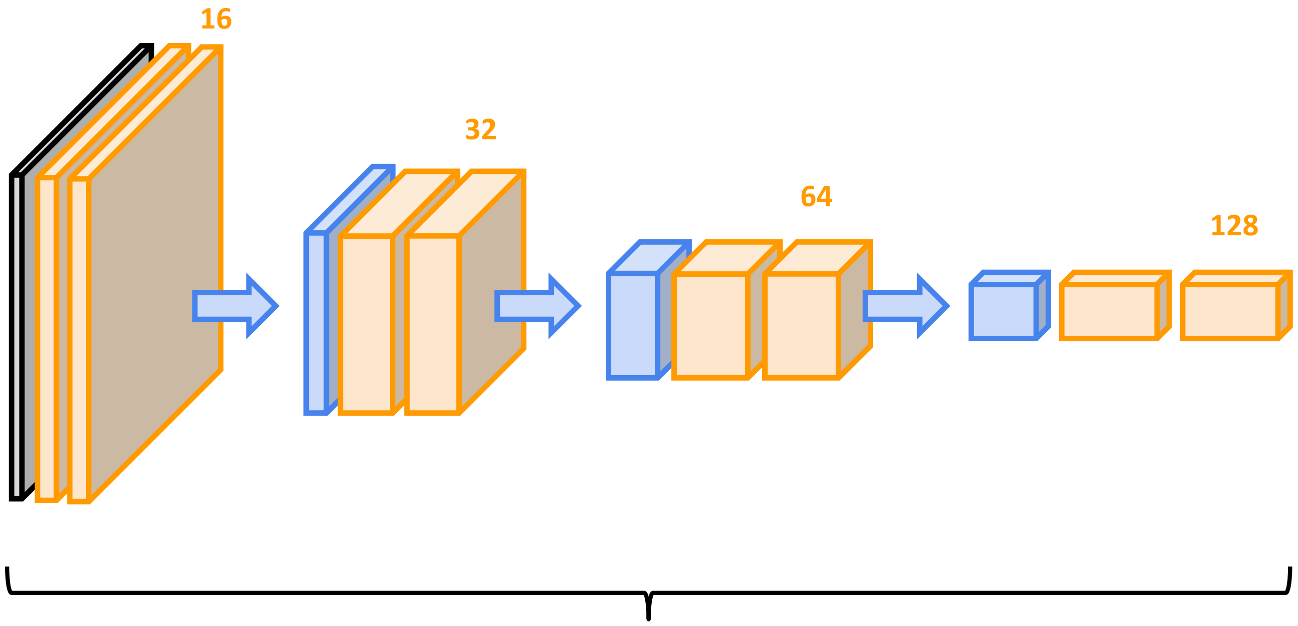


Image encoding

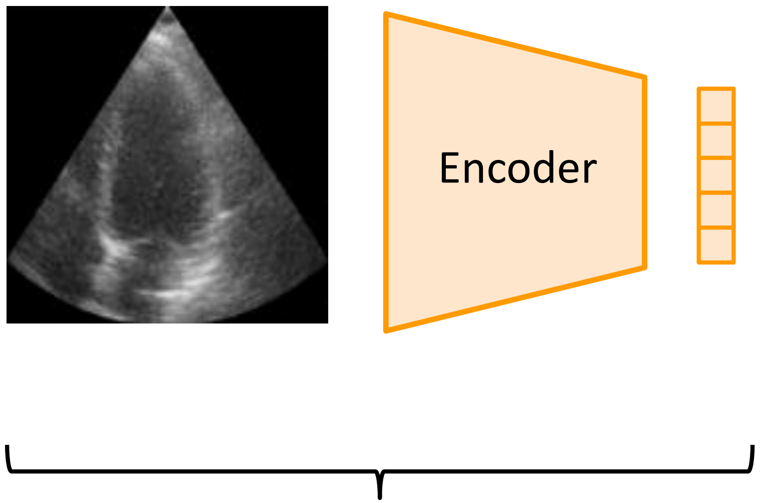
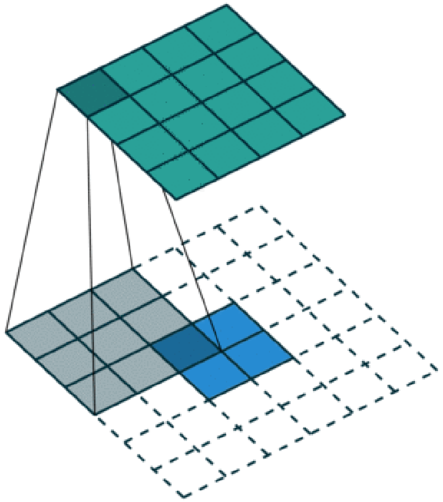
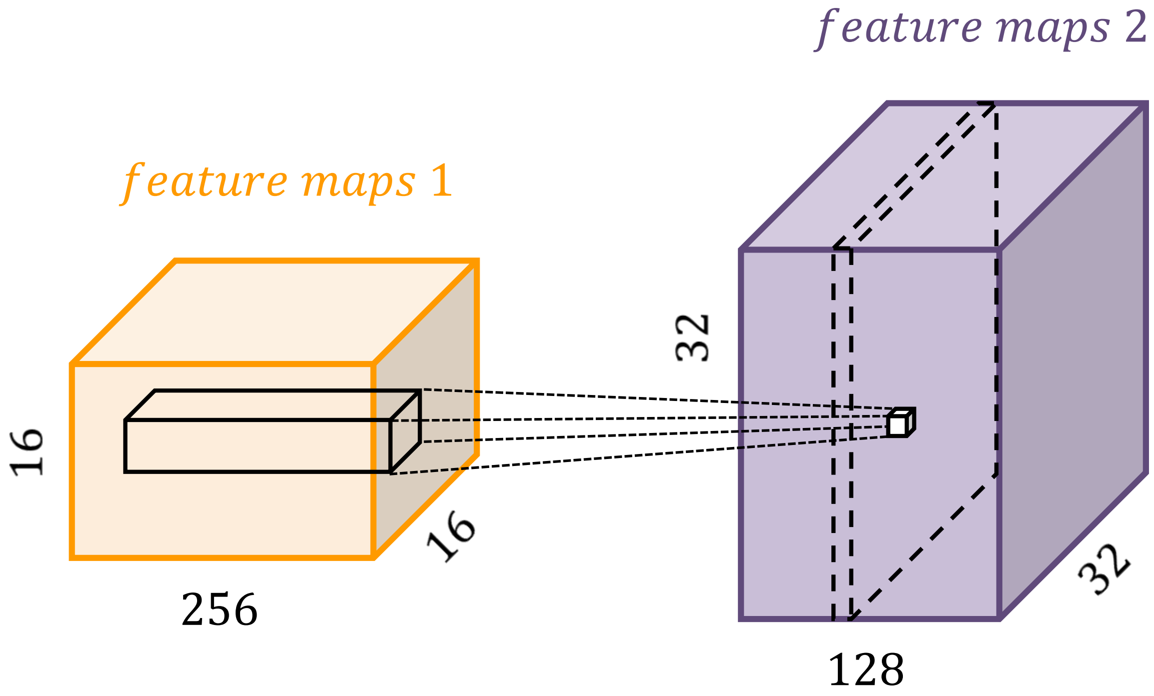


Image encoding

Deconvolutional layer

- ✓ Propagate relevant information to the input dimension space
- ✓ Parameters that are learned during training

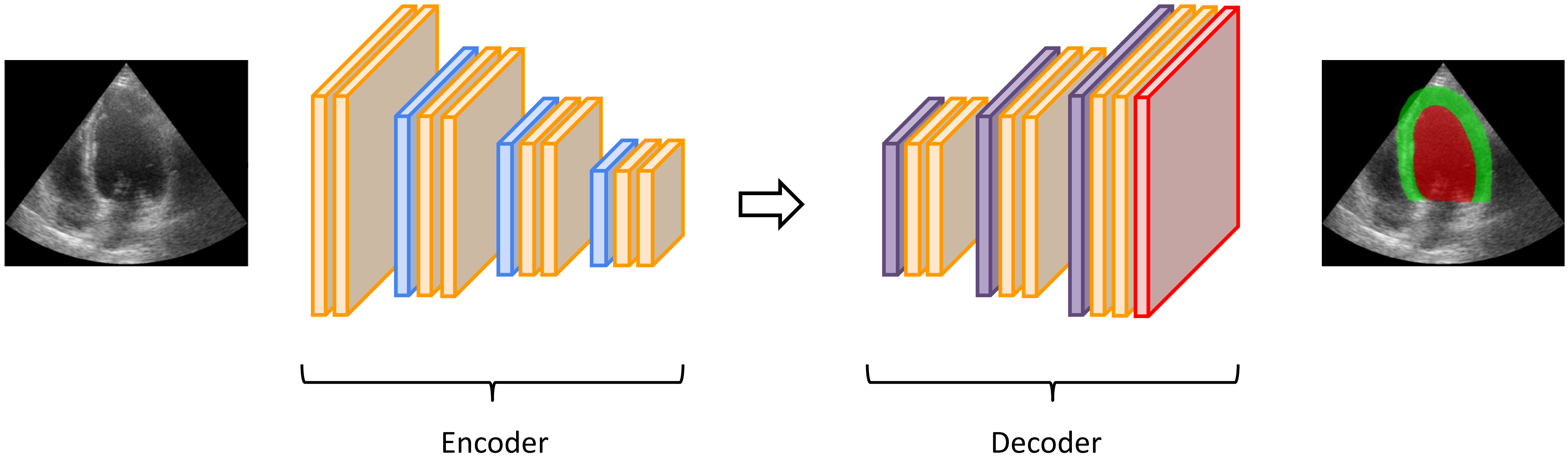




Filter of size 3×3

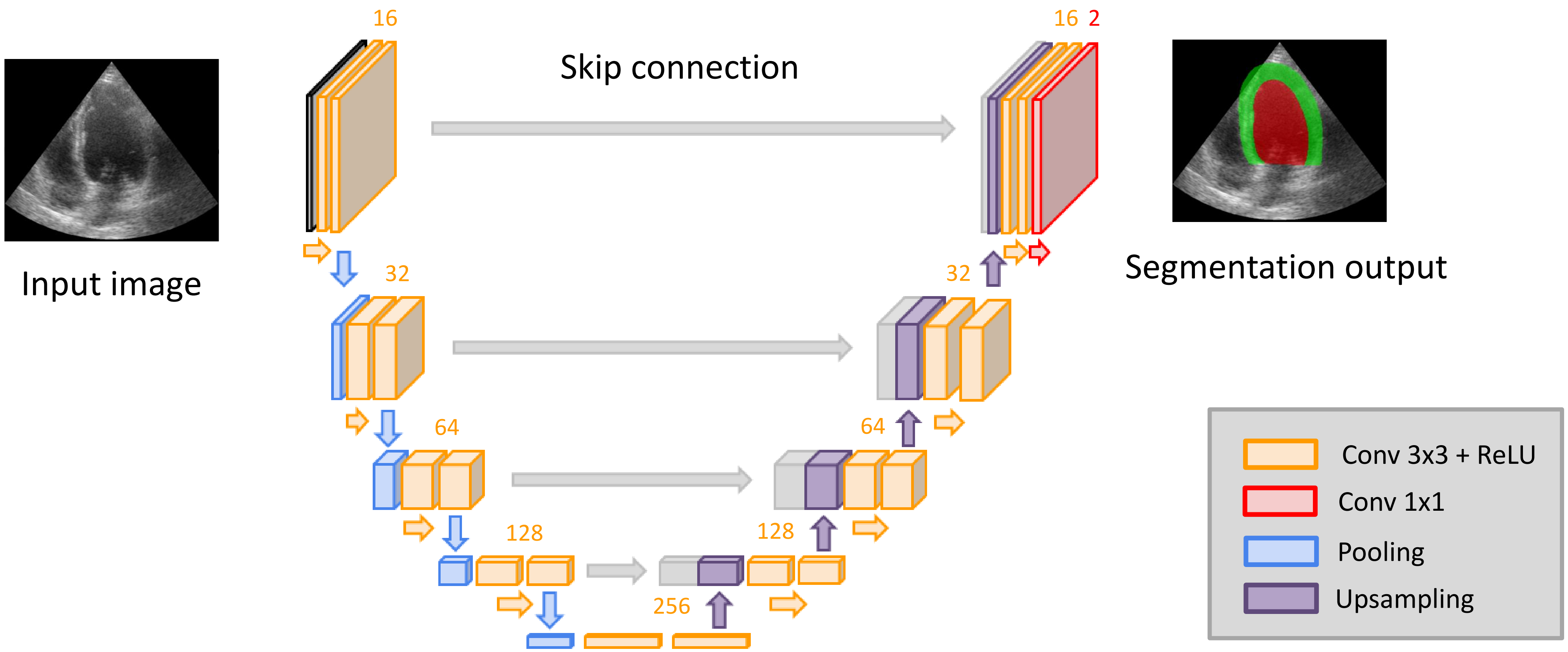
$$\# \, param = 128 \times (3 \times 3 \times 256 + 1) = 295{,}040$$

Encoder-decoder architectures



U-Net architecture

✓ Between 3 M to 40 M of parameters to train



Deep learning families

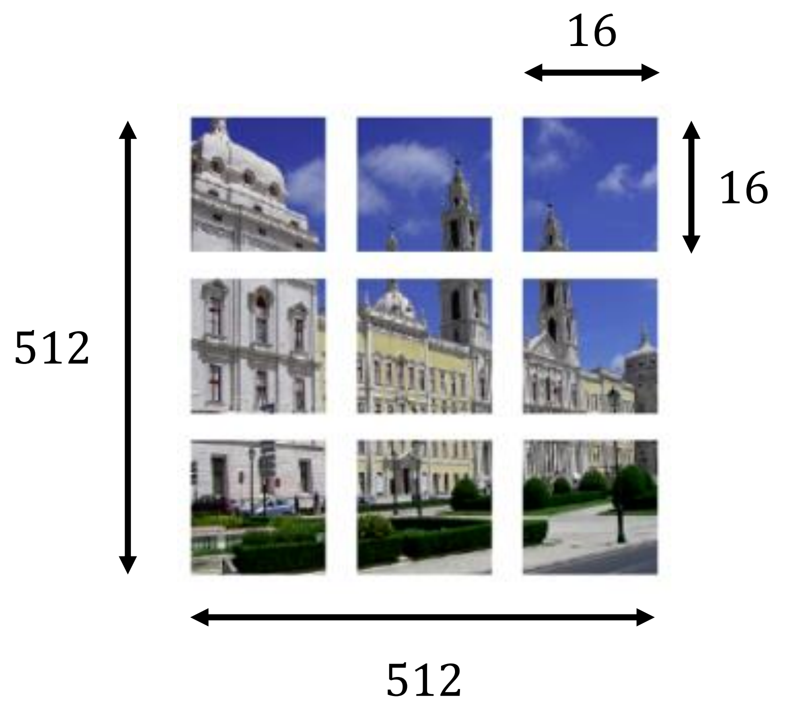
Transformers

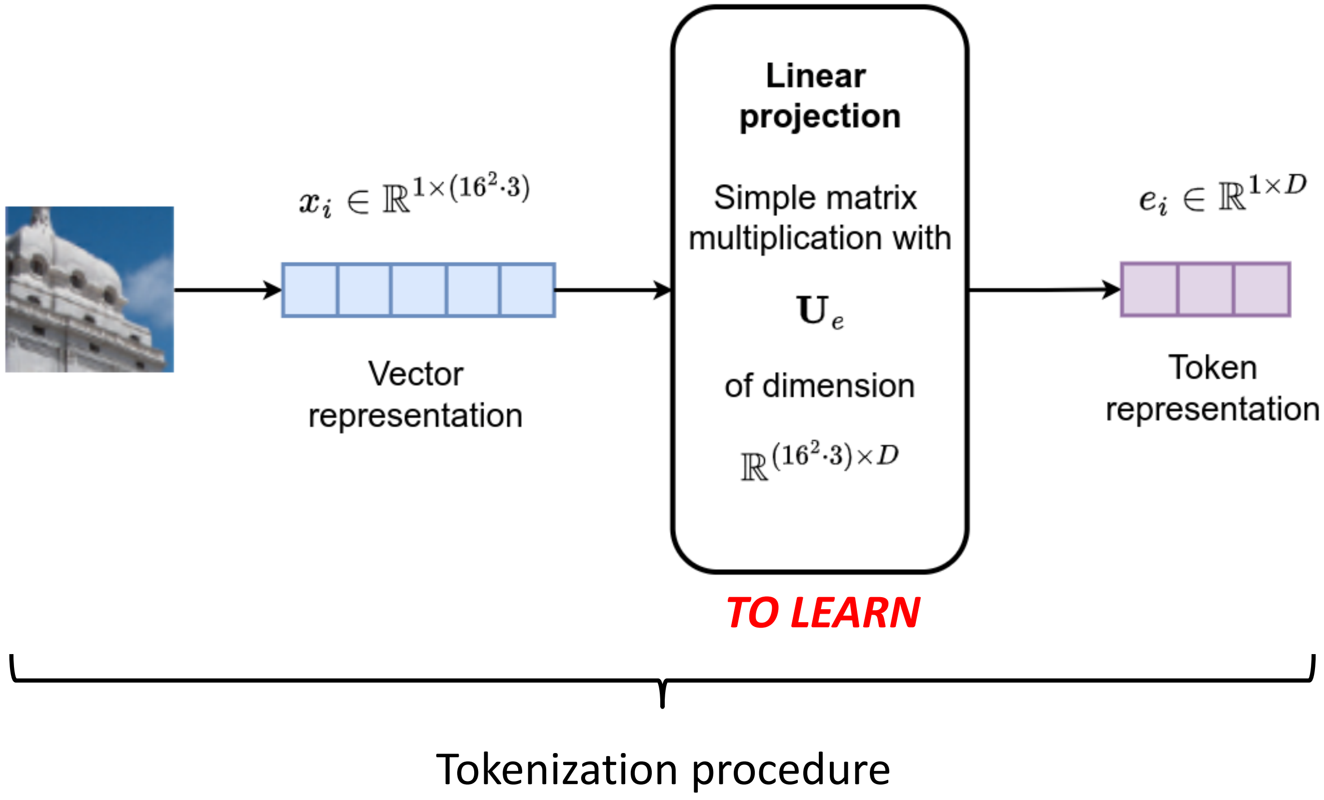
Tokenization procedure

✓ Efficient representation of an image

$D = 768$

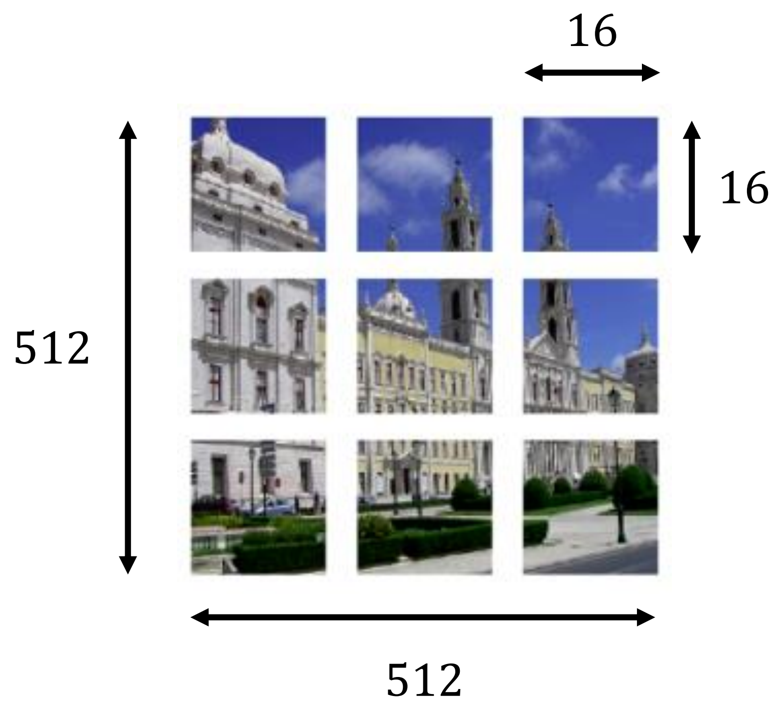
$\# \, param$
$= 16 \times 16 \times 3 \times 768$
$= 589{,}824$



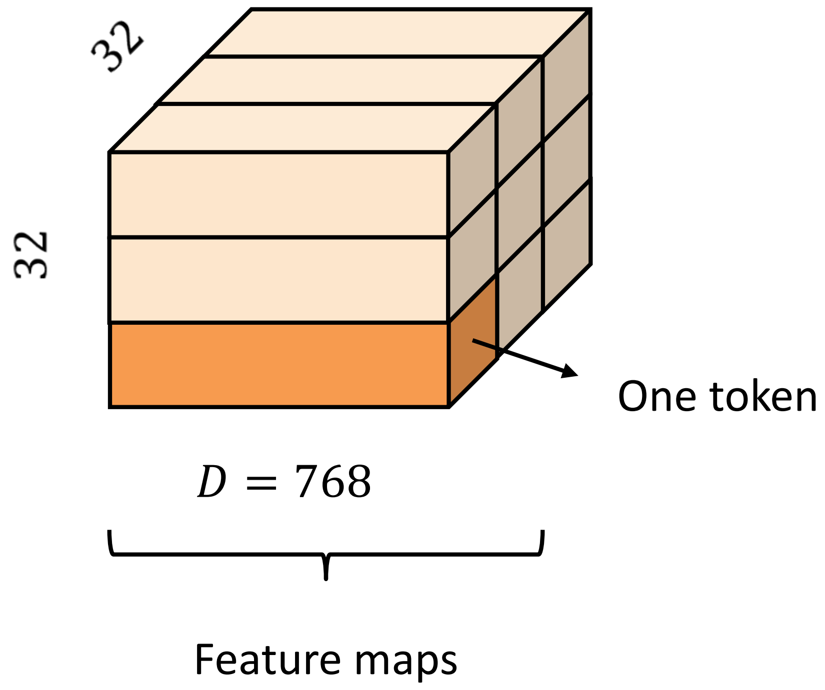


Tokenization procedure

✓ Representation of an image into a lower dimensional space

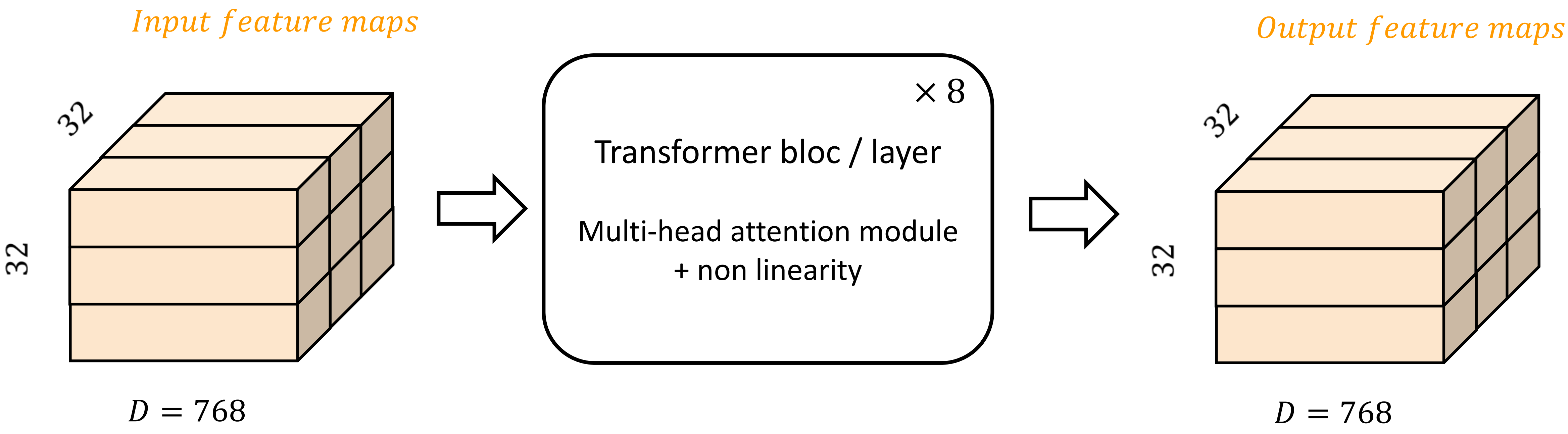


Tokenization procedure applied on each 16×16 patch



Transformer blocs / layers

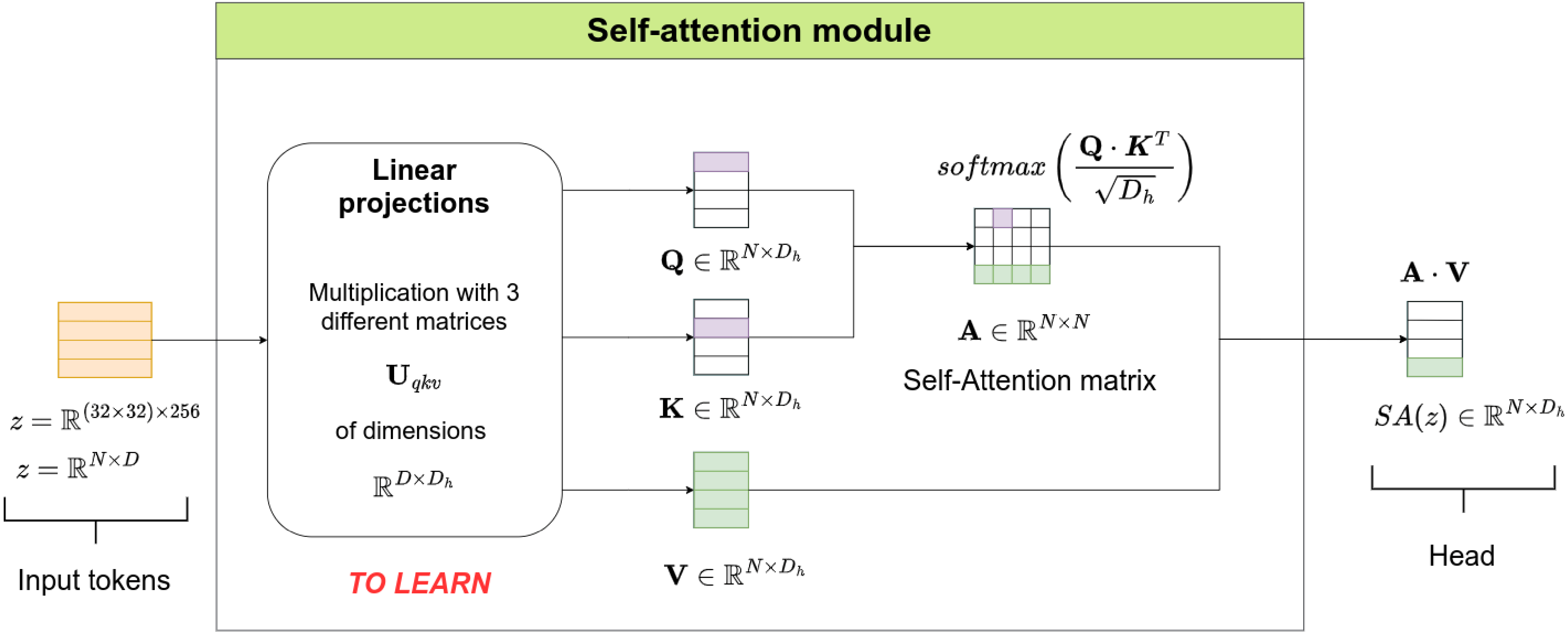
- ✓ Create relevant information (attention + non linear function)
- ✓ Parameters that are learned during training



Self-attention module

$$D = 768,\ D_h = 64$$

$$\#\, param = 3 \times 768 \times 64 = 147{,}456$$



Multi-head attention module

$$D = 768,\ D_h = 64,\ k = 12$$

$$\#\, param = 12 \times 3 \times 768 \times 64 + 768 * 768 = 2{,}359{,}296$$

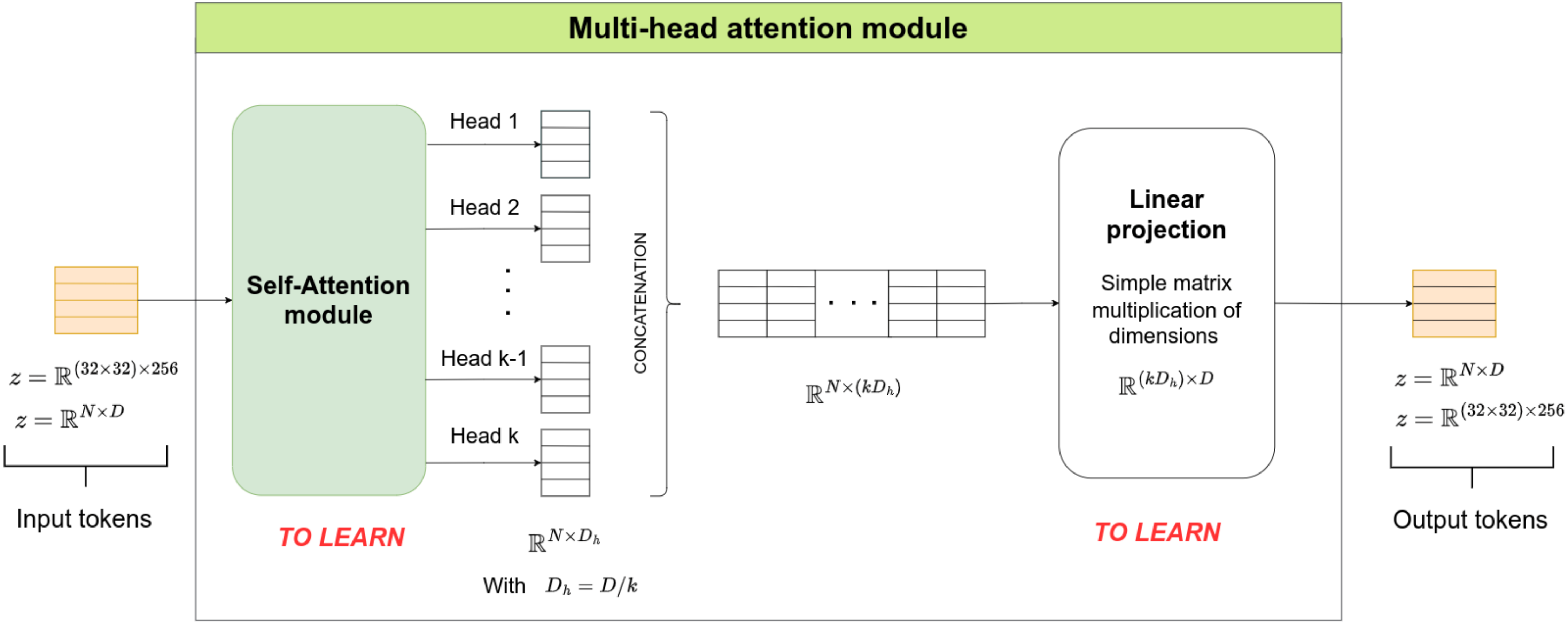


Image encoding

- ✓ Learning to encode relevant information
- ✓ Projection to a lower dimensional space

$$D = 768,\ D_h = 64,\ k = 12,\ N_{blocs} = 12$$

$$\begin{aligned}\#\,param \\ &= 12 \times (12 \times 3 \times 768 \times 64 + 768 * 768) \\ &+ 12 \times \ (2 \times 768 \times 3072) \\ &= 84{,}934{,}656\end{aligned}$$

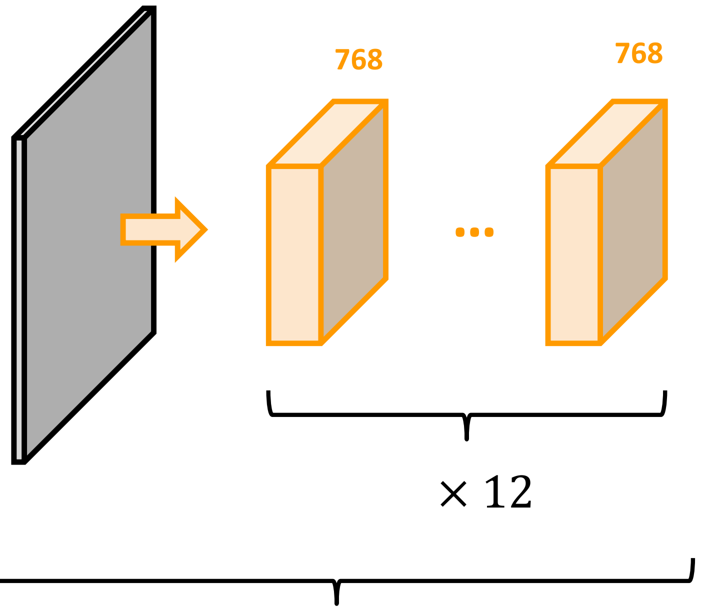


Image encoding

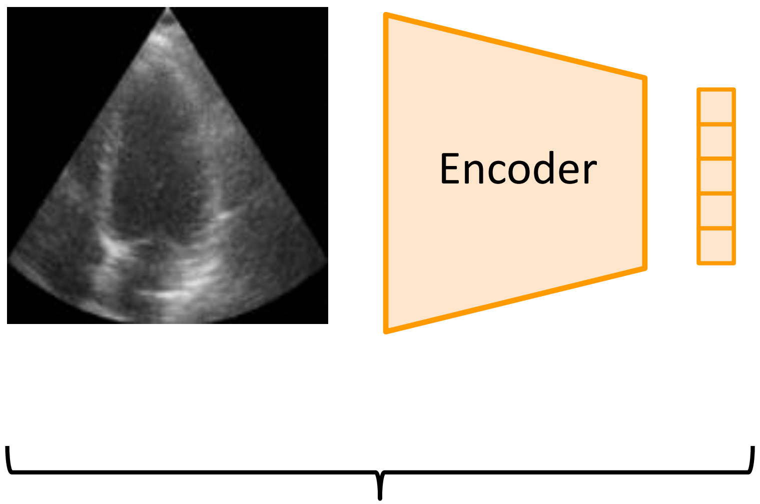
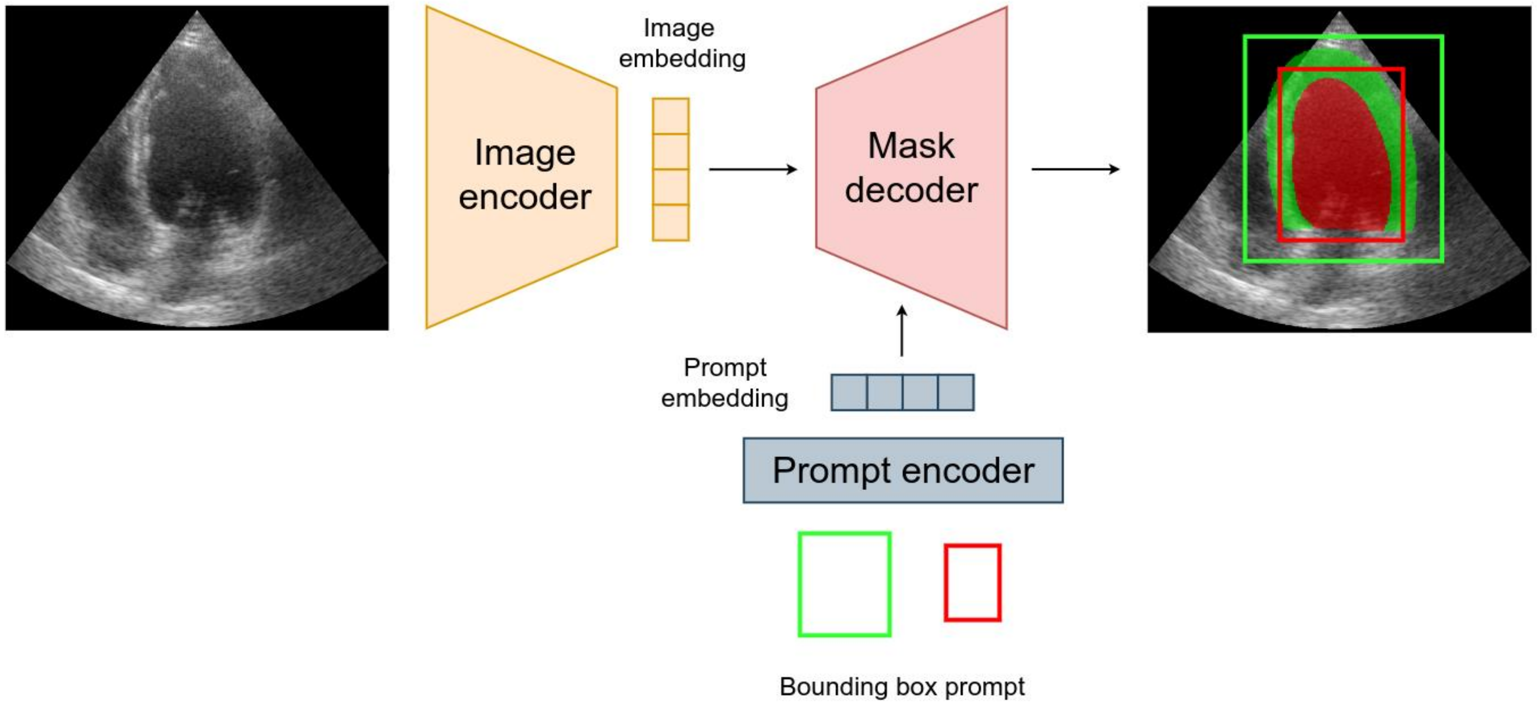


Image encoding

Foundation models

- ✓ 91 M of parameters to train

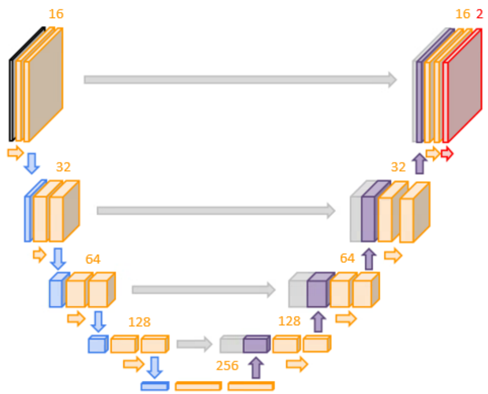


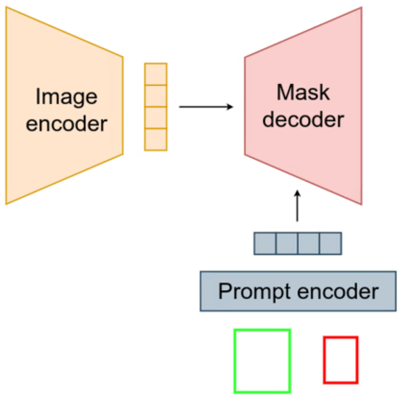
Segmentation of echocardiographic images

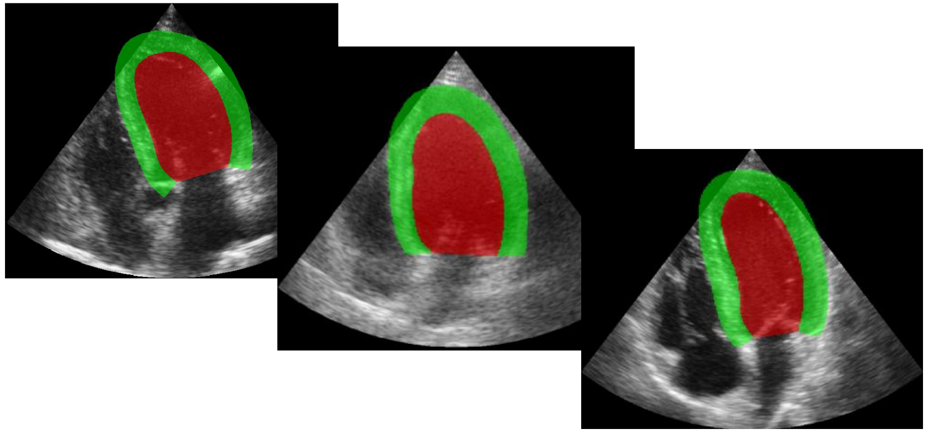
[Leclerc et al., IEEE TMI 2019]

The two key ingredients

- ✓ Deep learning solution with the proper complexity
- ✓ Database with good quality annotations







Echocardiographic datasets

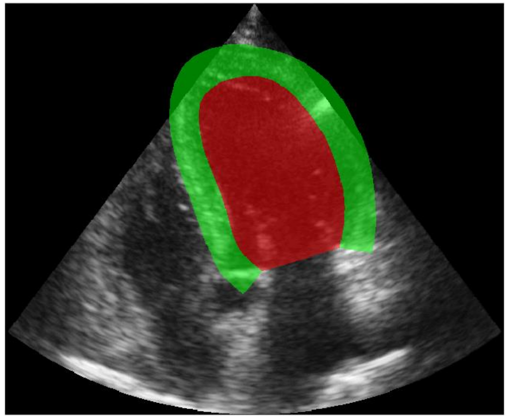
2D Public Echocardiographic Datasets										
Name	**Year**	**Nb. Subjects**	**Ground truth**				**Views**		**Characteristics**	
			LV_{endo}	***LV_{epi}***	***LA***	***Full cardiac cycle***	***A2C***	***A4C***	***Multi-Center***	***Multi-Vendor***
CAMUS	2019	500	✓	✓	✓	✗	✓	✓	✗	✗
EchoNet	2019	10,036	✓	✗	✗	✗	✗	✓	✗	-
HMC-QU	2021	292	✓	✓	✗	✗	✓	✓	✗	✓
TED	2022	98	✓	✓	✗	✓	✗	✓	✗	✗

Echocardiographic datasets

2D Public Echocardiographic Datasets										
Name	Year	Nb. Subjects	Ground truth				Views		Characteristics	
			LV_{endo}	*LV_{epi}*	*LA*	*Full cardiac cycle*	*A2C*	*A4C*	*Multi-Center*	*Multi-Vendor*
CAMUS	2019	500	✓	✓	✓	✘	✓	✓	✘	✘
EchoNet	2019	10,036	✓	✘	✘	✘	✘	✓	✘	-
HMC-QU	2021	292	✓	✓	✘	✘	✓	✓	✘	✓
TED	2022	98	✓	✓	✘	✓	✘	✓	✘	✘

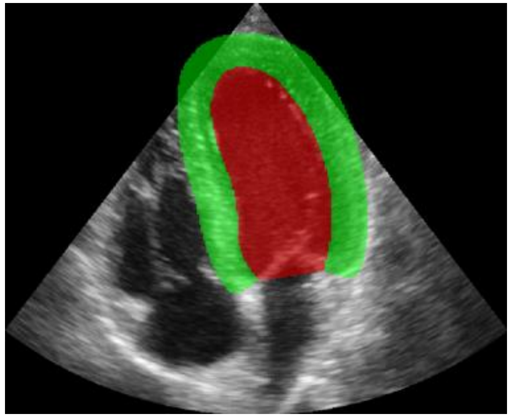
CAMUS

- ✓ Center 1
- ✓ Annotator 1
- ✓ Vendor 1
- ✓ 500 patients
- ✓ Image annotations



TED

- ✓ Center 1
- ✓ Annotator 1
- ✓ Vendor 1
- ✓ 98 patients
- ✓ Sequence annotations



CAMUS - Performance

- ✓ Geometric accuracy

(CS:CAMUS)

Methods	Dice		Hausdorff (mm)	
	ED	ES	ED	ES
Intra-obs.	.945	.930	4.6	4.5
2D nnU-Net	**.952**	**.935**	**4.3**	**4.2**
CLAS	.947	.929	4.6	4.6
GUDU	.946	.929	4.7	4.7

- ✓ Clinical accuracy

(CS:CAMUS)

Methods	EF		Volume ED		Volume ES	
	Corr.	MAE (%)	Corr.	MAE (ml)	Corr.	MAE (ml)
Intra-obs.	.896	4.7	.978	6.5	.981	4.5
2D nnU-Net	.857	4.7	**.977**	**5.9**	**.987**	**4.0**
CLAS	**.926**	**4.0**	.958	7.7	.979	4.4
GUDU	.897	**4.0**	**.977**	6.7	.981	4.6

CAMUS - Conclusions

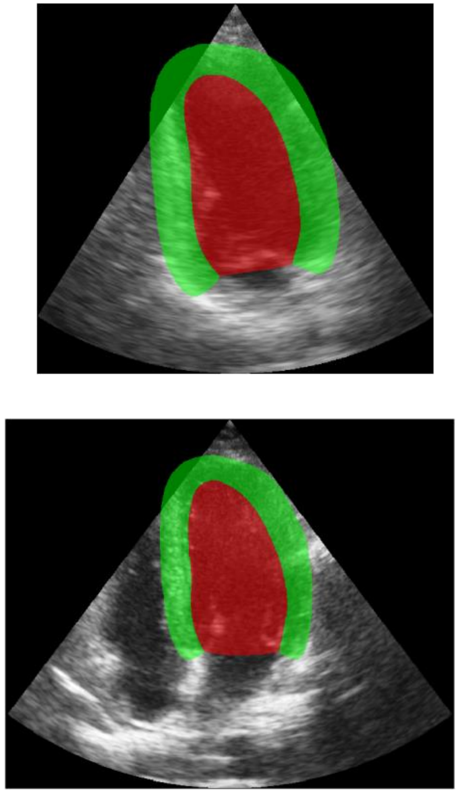
What are the conclusions of the pilot CAMUS's story ?

✓ nnU-Net produces:

- accurate scores from a controlled dataset
- within the intra-expert variability

✓ Has the potential to replace the expert's hand !

How can these results be generalized to large-scale datasets involving data from multiple centers, multiple vendors and multiple experts?



nnU-Net predictions

Two tendencies

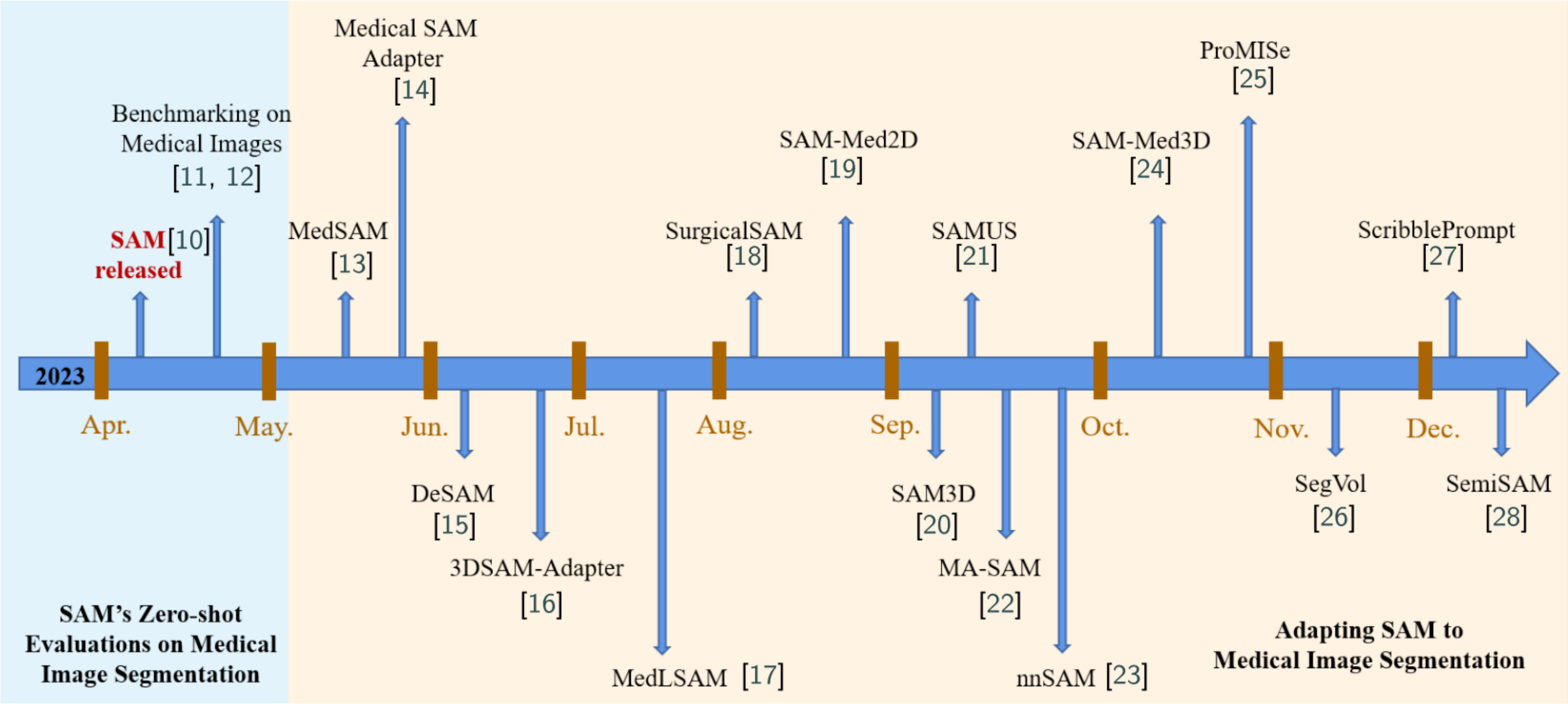
✓ Foundation models

- Learning from large scale datasets with different modalities, organs, views, …

✓ Domain adaptation

- Efficient transfer from a source dataset (CAMUS) to a target dataset

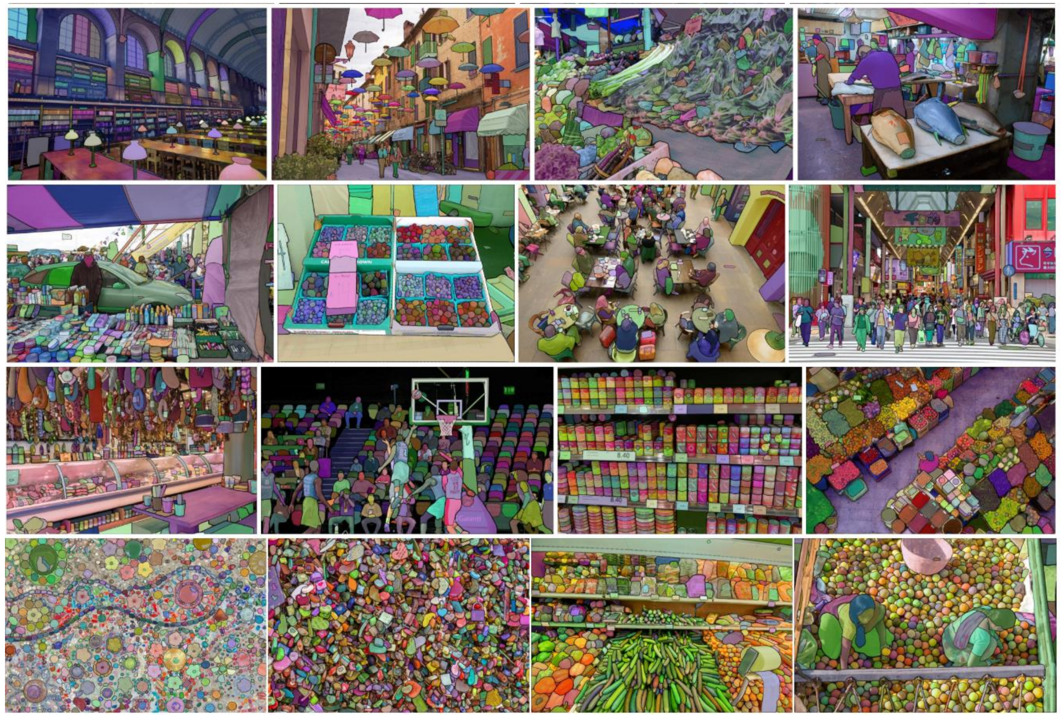
Brief chronology for Segment Anything (SAM) models



From [Zhang et al., CIBM, 2024]

Large scale datasets

- ✓ SAM dataset
- ✓ Licensing private dataset accessible for research purposes

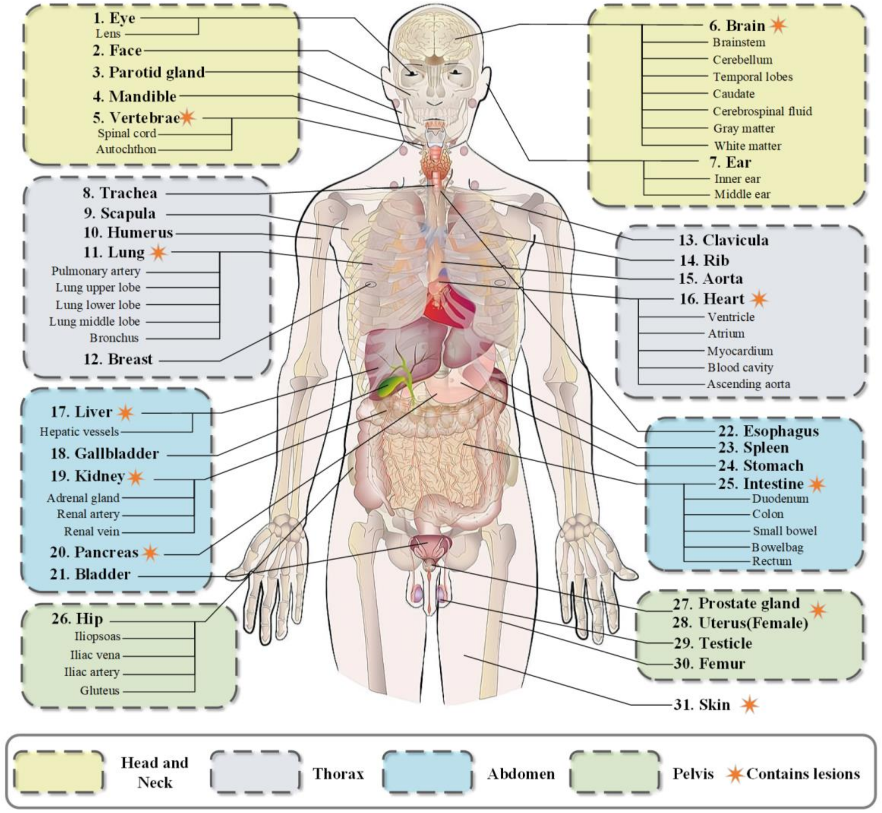


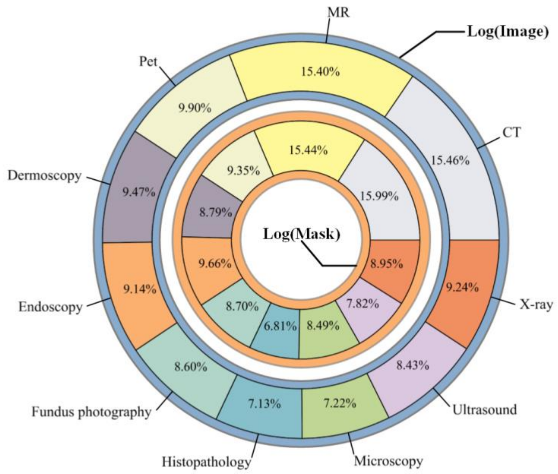
From [Kirillov et al., Arxiv, 2023]

- 11 M images / 1 B masks
- 2D images
- Natural scene images
- Shortest side 1500 px

Large scale datasets

- ✓ SAM-Med2D dataset
- ✓ Collating from publicly available medical datasets + private datasets



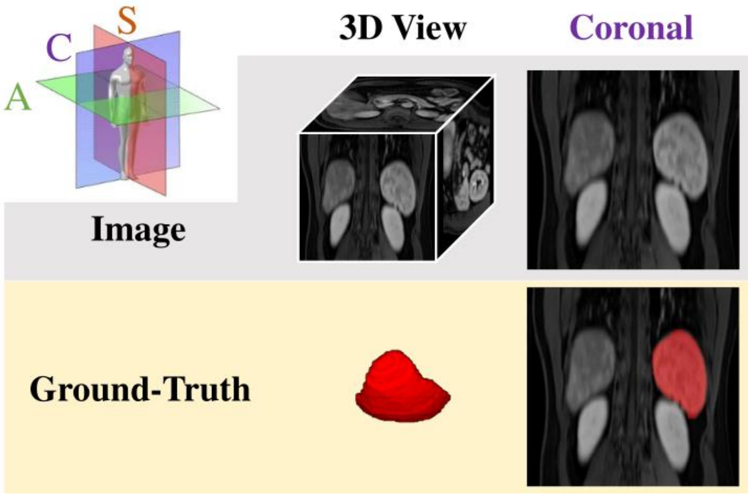


From [Cheng et al., Nature, 2024]

- 4.6 M images / 19.7 M masks
- 2D images
- 10 imaging modalities
- 31 major organs
- 15% of CT images
- $256 \times 256 \times 3$ image size
- Image intensity homogenization

Large scale datasets

- ✓ SAM-Med3D dataset
- ✓ Collating from publicly available medical datasets + private datasets

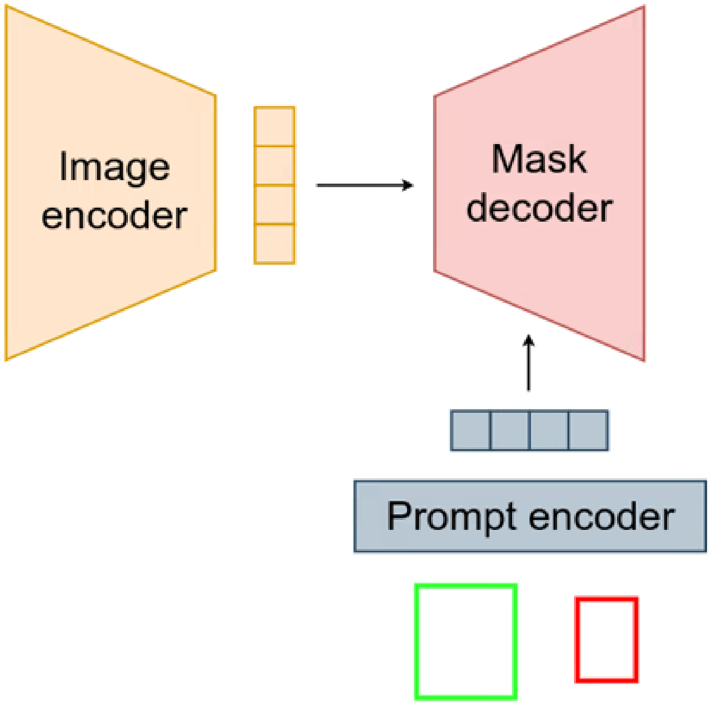


From [Wang et al., Arxiv, 2024]

- 21 K images / 131 K masks
- 3D images
- 27 imaging modalities (among CT)
- 7 anatomical structures
- $128 \times 128 \times 128$ patch size

AI architecture

✓ Transformer model with high complexity
✓ More than 91 M of parameters



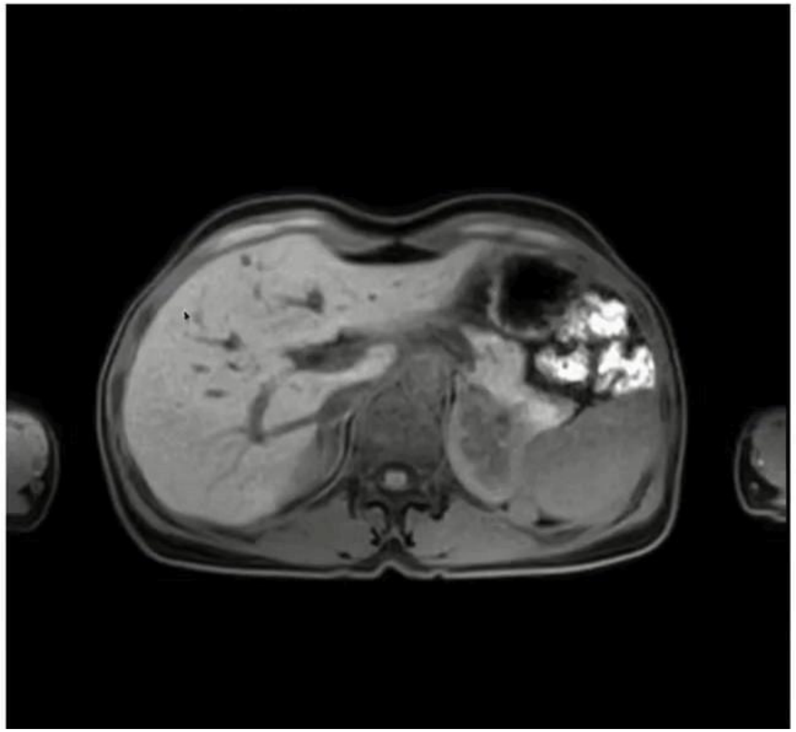
Training strategies

- Pre-training from SAM dataset
- Fine-tuning on SAM-Med datasets

Architecture choices

- Freeze prompt encoder while fine-tuning image encoder and mask decoder
- Freeze image encoder while introducing learnable adapter layer, fine-tuning the prompt encoder and mask decoder

Performance illustration



https://github.com/bowang-lab/MedSAM

Two tendencies

✓ Foundation models

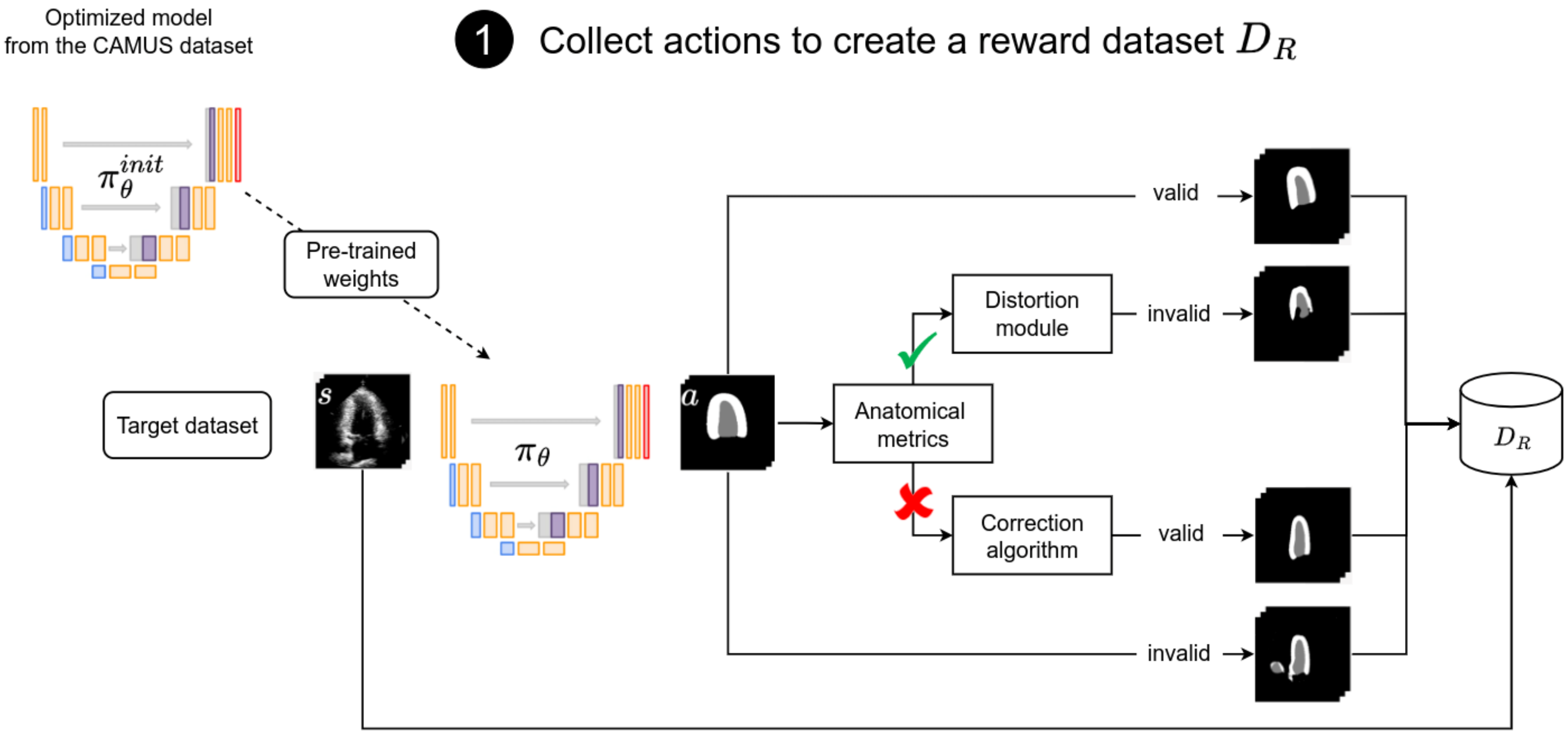
- Learning from large scale datasets with different modalities, organs, views, …

✓ Domain adaptation

- Efficient transfer from a source dataset (CAMUS) to a target dataset

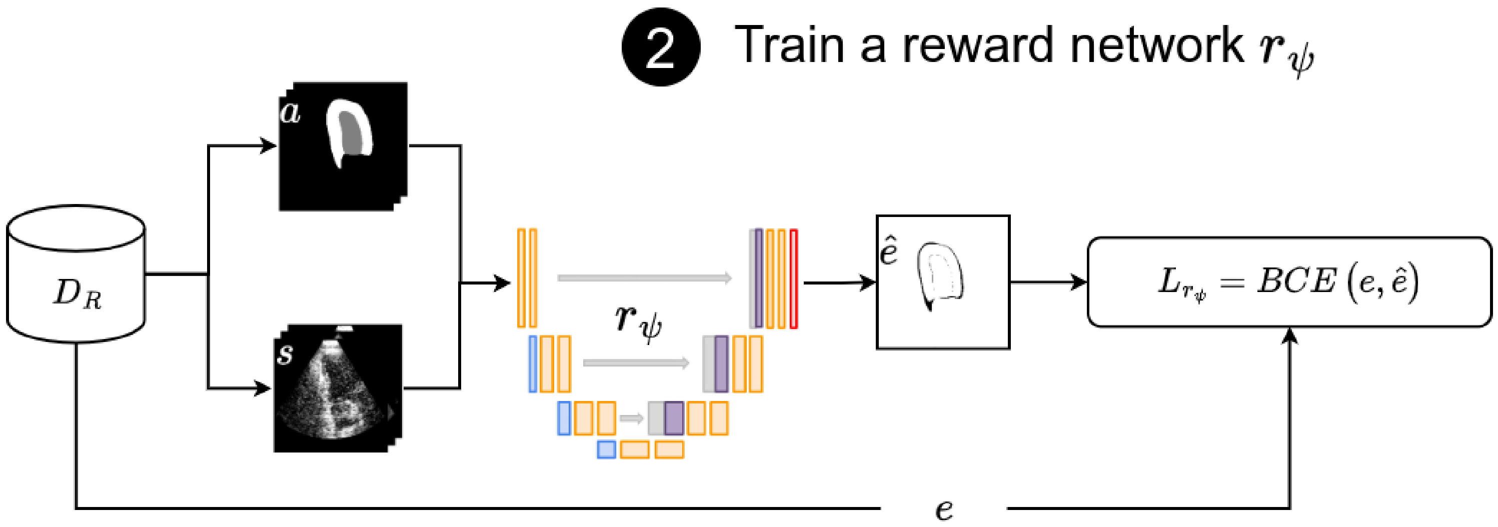
Inspired from reinforcement learning

✓ Update nnU-Net weights to fit with the target dataset



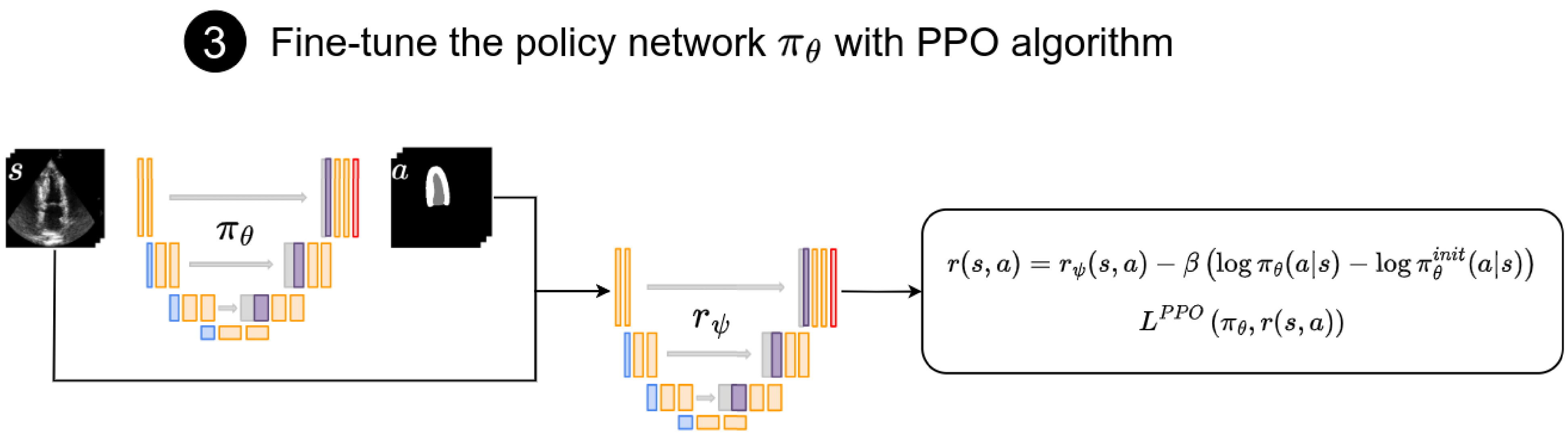
Inspired from reinforcement learning

✓ Update nnU-Net weights to fit with the target dataset



Inspired from reinforcement learning

✓ Update nnU-Net weights to fit with the target dataset



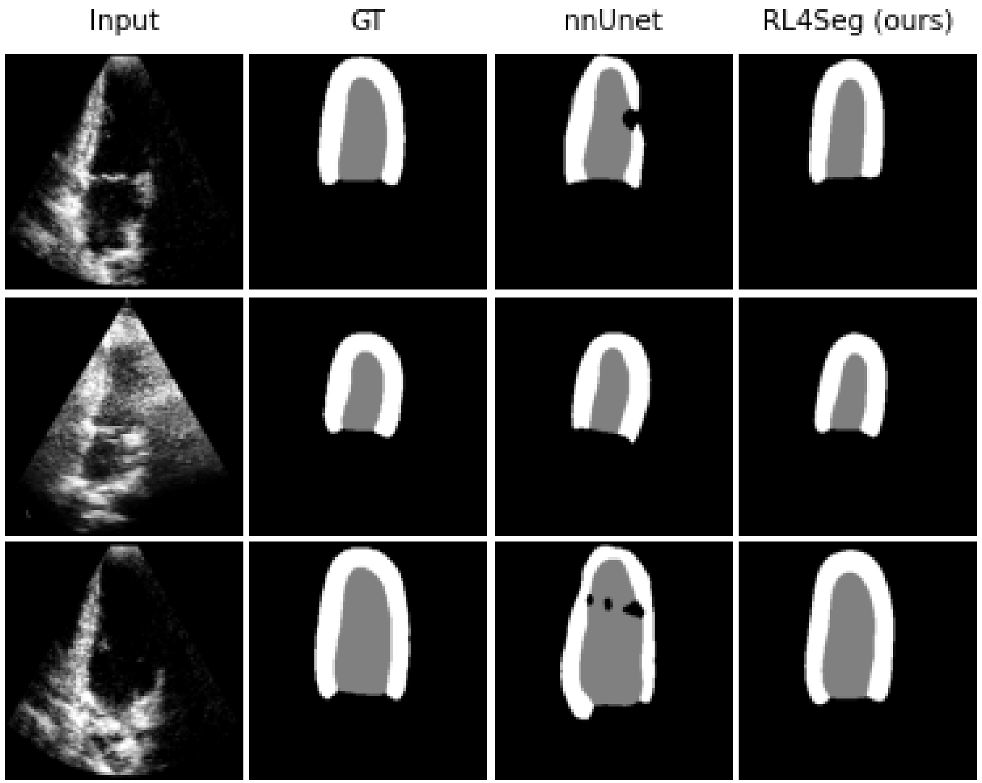
Preliminary results

✓ Scores computed from 220 patients from the target dataset

Method	Dice (%) ↑			Hausdorff (mm) ↓			Anatomical Validity (%) ↑
	ENDO	EPI	Avg.	ENDO	EPI	Avg.	
$\mathcal{D}_S$ intra-expert var.	94.4	95.4	94.9	4.3	5.0	4.6	100
nnUnet	91.0	94.6	92.8	6.3	7.8	7.1	95.0
RL4Seg (ours)	**91.9**	**94.7**	**93.3**	**4.9**	**5.6**	**5.3**	**98.9**

Preliminary results

✓ Scores computed from 220 patients from the target dataset



Conclusions & Perspectives

Conclusions & perspectives

- Conclusions
  - ✓ AI methods have already revolutionized medical image segmentation
  - ✓ Pilot studies have shown that such methods can faithfully reproduce the hand of an expert

- Perspectives
  - ✓ Intensive studies on the generalization of AI model to large scale dataset
  - ✓ We are undoubtedly witnessing the resolution of the segmentation problem in medical imaging!

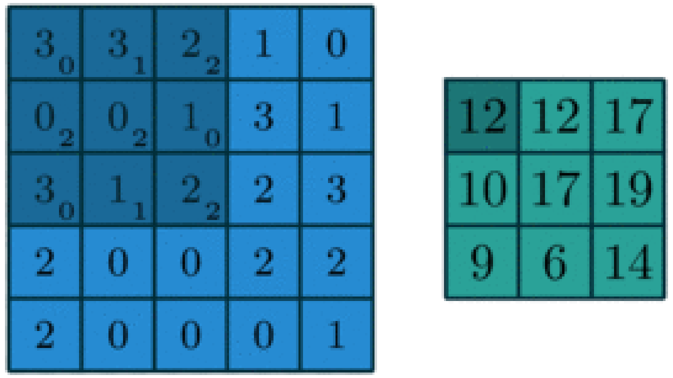
Thanks

Appendices

Convolution reminder

$$
\begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} * \begin{pmatrix} 3 & 3 & 2 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 \\ 3 & 1 & 2 & 2 & 3 \\ 2 & 0 & 0 & 2 & 2 \\ 2 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 12 & 12 & 17 \\ 10 & 17 & 19 \\ 9 & 6 & 14 \end{pmatrix}
$$

Filter Image Output

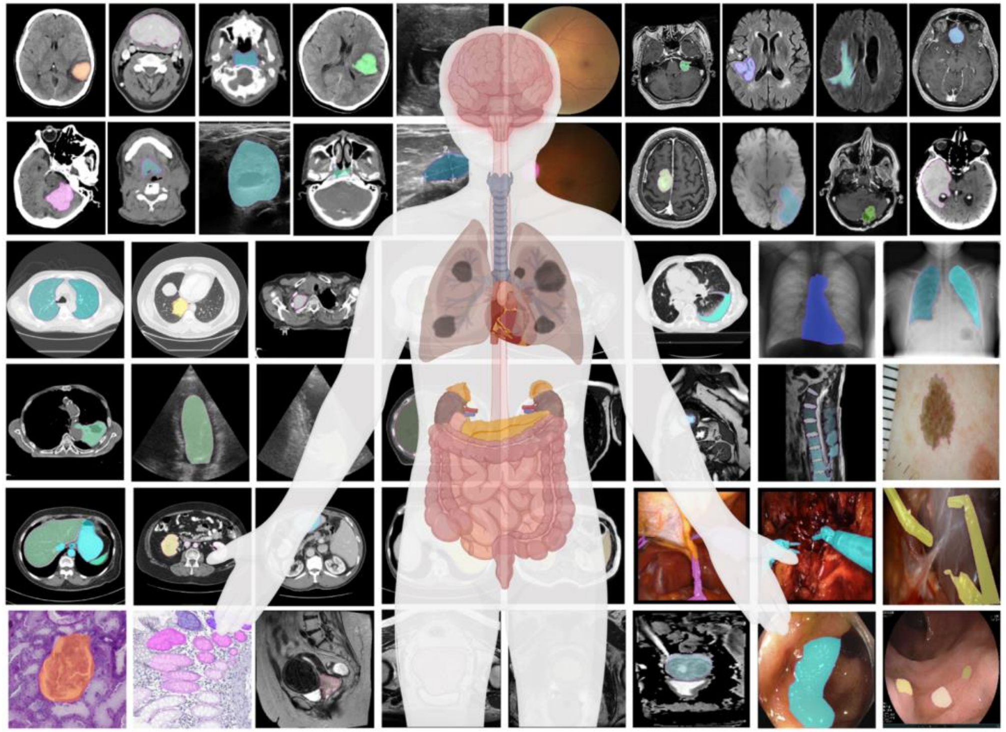


$$
O = I * h \qquad o[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} i[u,v]\, h[i-u, j-v]
$$

Foundation models for medical image segmentation

Large scale datasets

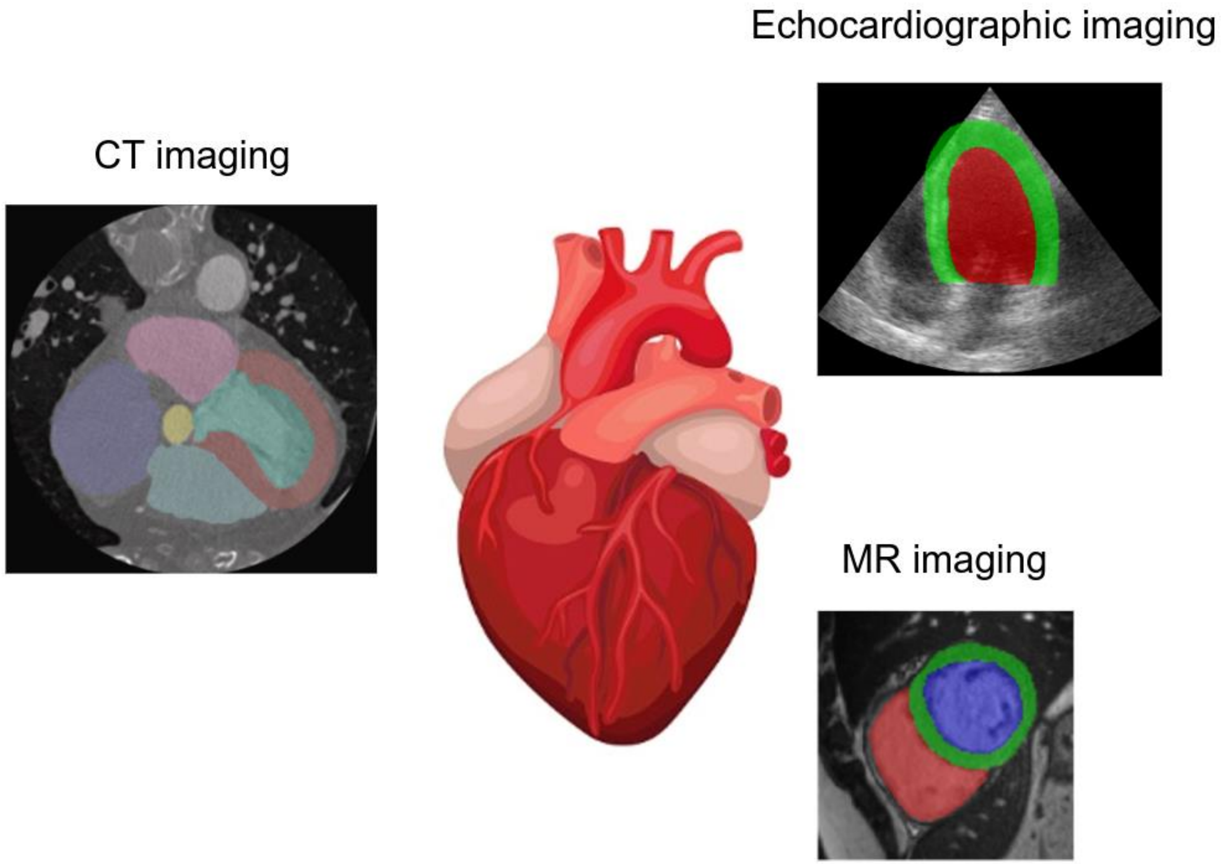
- ✓ MedSAM dataset
- ✓ Collating from publicly available medical datasets



Taken from [Ma et al., Nature, 2024]

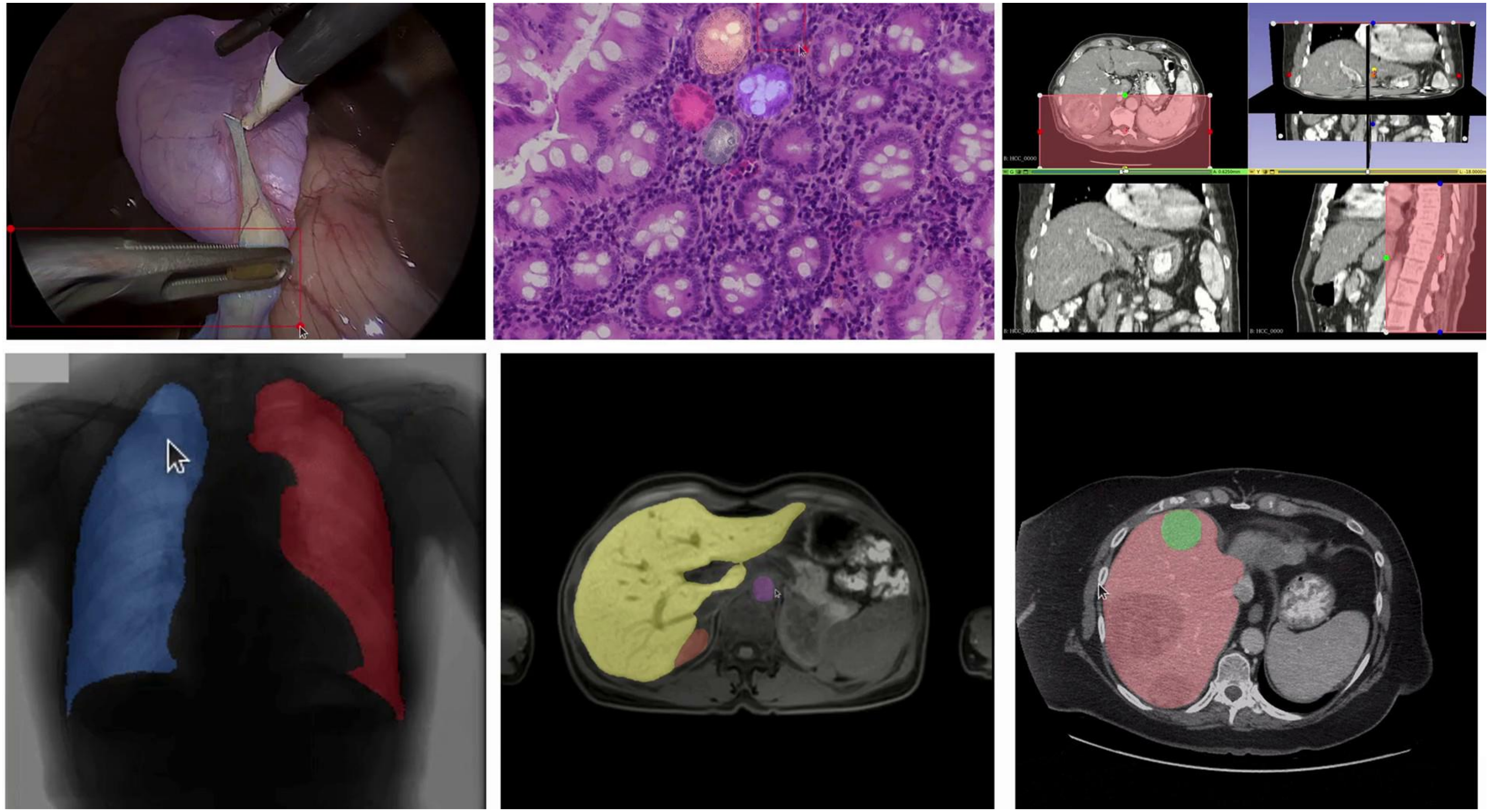
- 1.5 M image-mask pairs
- 2D images
- 10 imaging modalities
- 30 cancer types
- 24% of CT images
- $1024 \times 1024 \times 3$ image size
- Image intensity homogenization

Quantification of clinical indices to diagnose cardiac pathologies



- ✓ High annotation costs
- ✓ Inter/intra expert variability
- ✓ Acquisition variabilities
- ✓ Acquisition artifacts

Performance illustration



https://github.com/bowang-lab/MedSAM