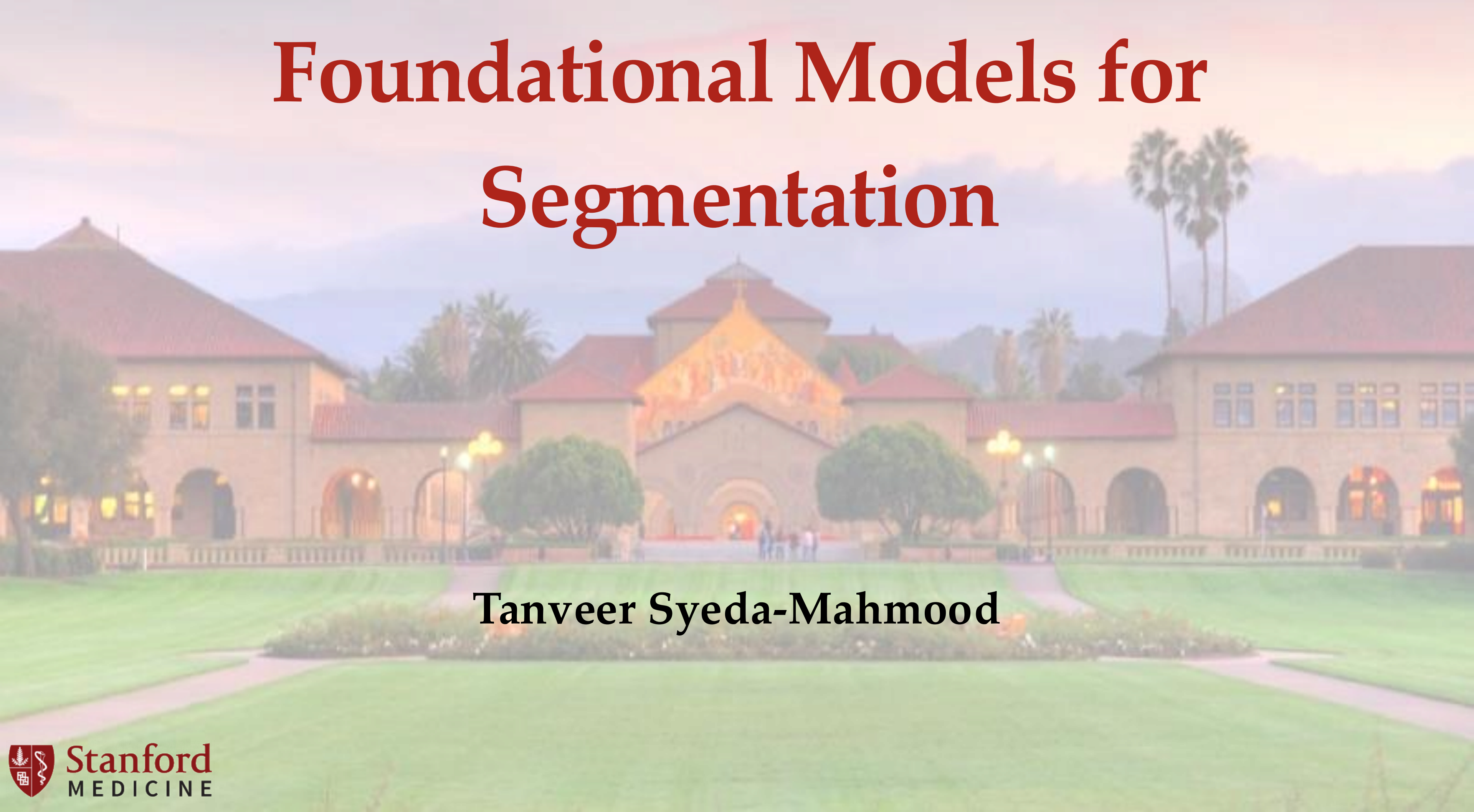


Foundational Models for Segmentation

Tanveer Syeda-Mahmood

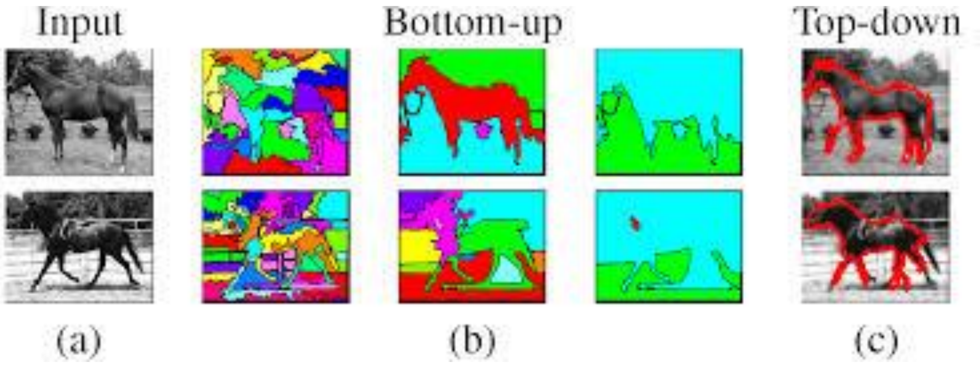
Stanford
MEDICINE

The different notions of segmentation

- Computer vision
  - Image partitioning into objects and background
- Medical imaging
  - Anatomy segmentation
  - Anomaly segmentation

Bottom-up?
Or top-down?

Supervised?
Or unsupervised?

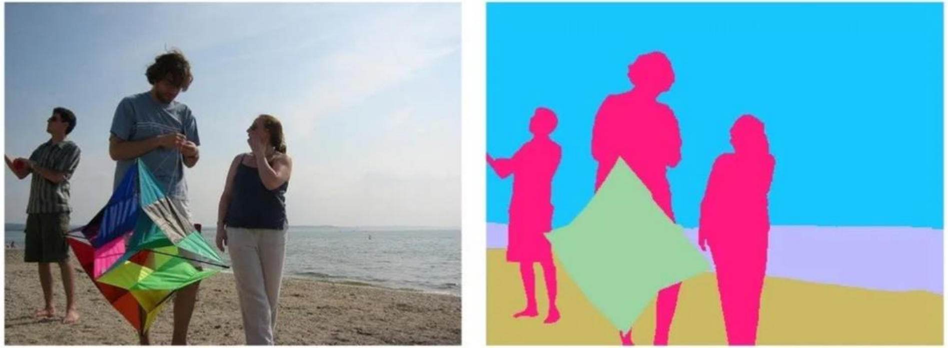


https://www.csd.uwo.ca/~oveksler/Courses/Fall2007/840/StudentPapers/LevinWeissEccv06.pdf

The different notions of segmentation

- Computer vision
  - Image partitioning into objects and background
- Medical imaging
  - Anatomy segmentation
  - Anomaly segmentation

Semantic segmentation

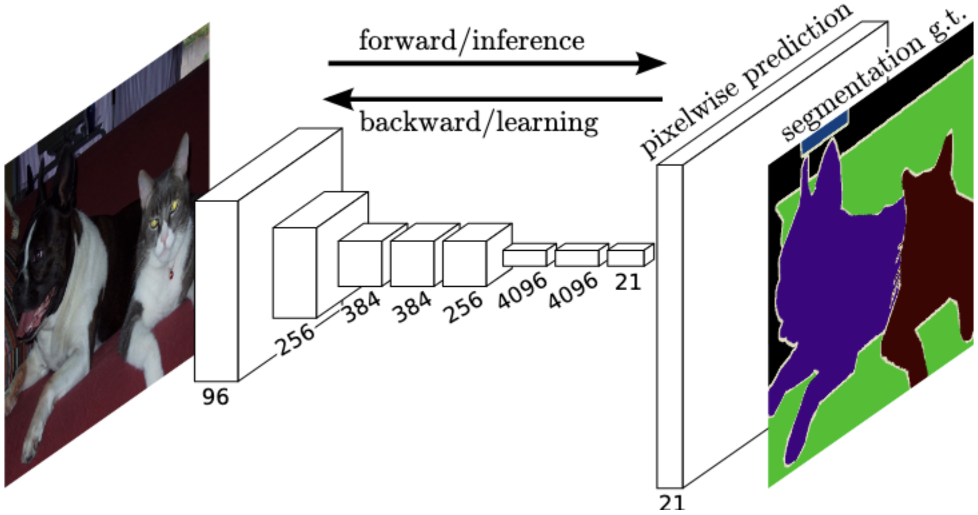


- Labels each pixel by class
- E.g. FCNet, DeepLab, PSPNet

https://medium.com/@abhishekjainindore24/semantic-vs-instance-vs-panoptic-segmentation-b1f5023da39f

Semantic segmentation example: FCN

- Fully-supervised method
- Used in many other networks for the segmentation head
- Using purely convolutional setup



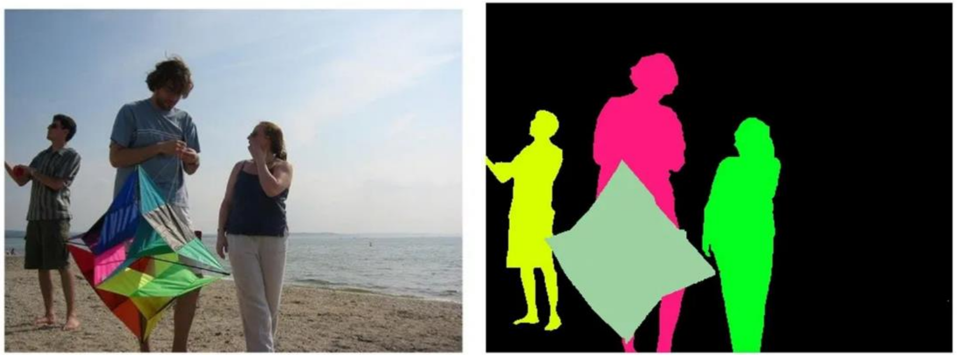
https://arxiv.org/abs/1411.4038

Fully Convolutional Networks for Semantic Segmentation

The different notions of segmentation

- Computer vision
  - Image partitioning into objects and background
- Medical imaging
  - Anatomy segmentation
  - Anomaly segmentation

Instance segmentation

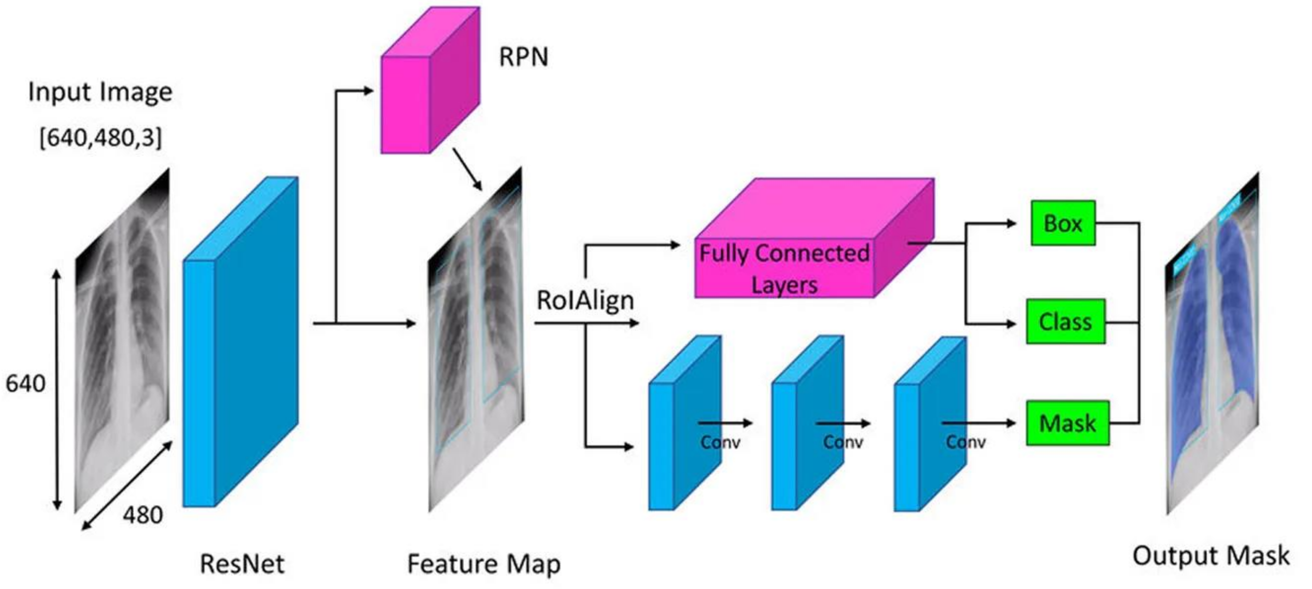


- Each region has a distinct label
- Combine object detection and semantic segmentation
- E.g. Mask R-CNN, Yolo-8

https://medium.com/@abhishekjainindore24/semantic-vs-instance-vs-panoptic-segmentation-b1f5023da39f

Instance segmentation example: Mask RCNN

- Up to 1500 class labels can be detected and labeled



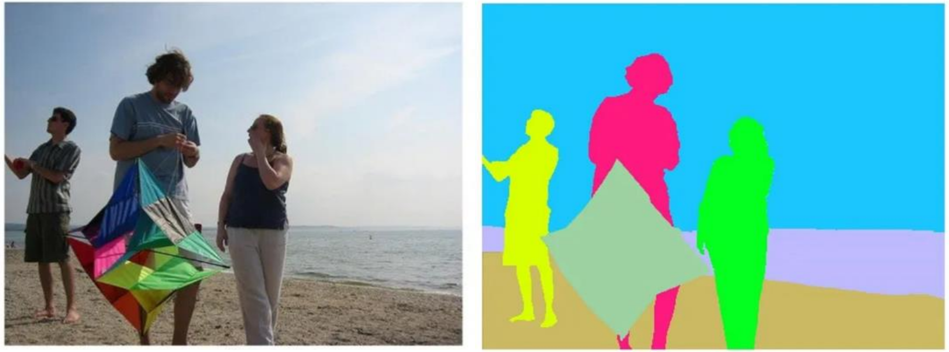
https://arxiv.org/abs/1703.06870, K.He et al. Mask R-CNN 42786 citations!

https://www.ultralytics.com/blog/what-is-mask-r-cnn-and-how-does-it-work

The different notions of segmentation

- Computer vision
  - Image partitioning into objects and background
- Medical imaging
  - Anatomy segmentation
  - Anomaly segmentation

panoptic segmentation

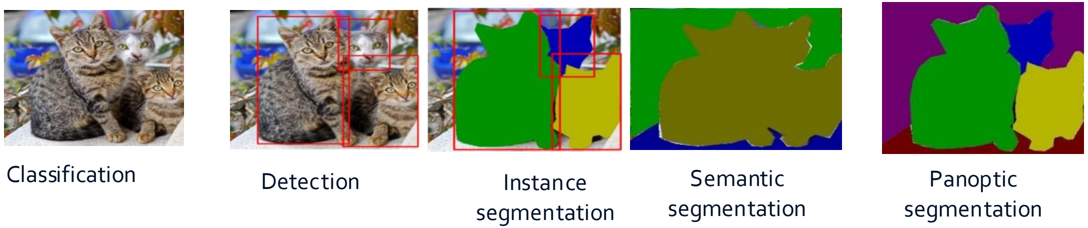


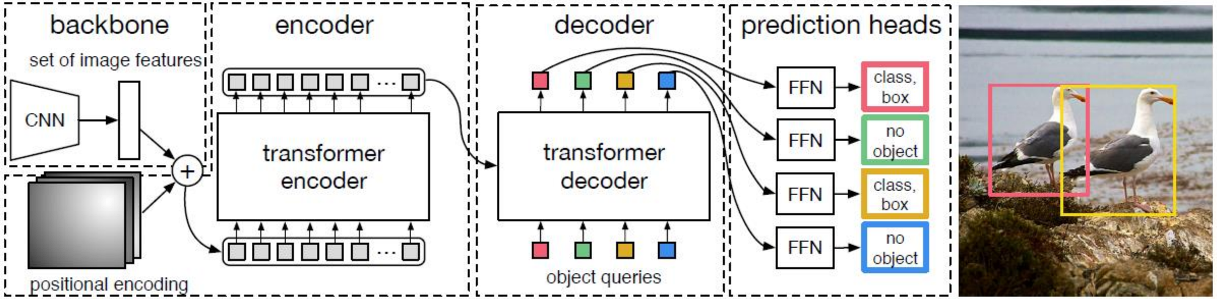
- Labels each pixel by class and also identifies different instances of the same class
- E.g. Efficient PS, Panoptic DeepLab, RS-DINO, HyperDETR, DinoV2, FCN+MaskRCN

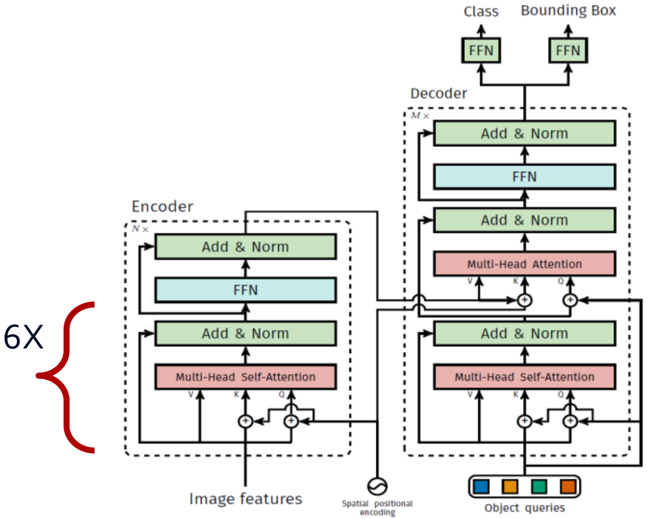
https://medium.com/@abhishekjainindore24/semantic-vs-instance-vs-panoptic-segmentation-b1f5023da39f

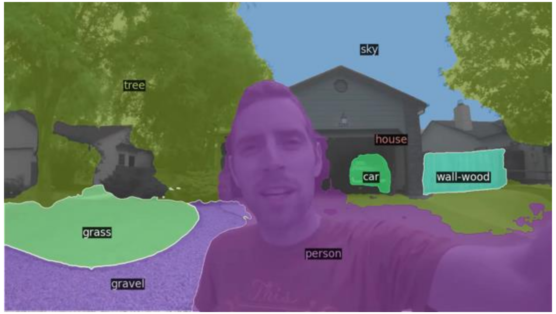
Panoptic segmentation(DETR)

- Uses the transformer approach for capturing spatial relationship between regions
- Uses feature encoding
- Class and regional predictions
- Supervision needed
- Limited object support
- Panoptic segmentation









End-to-End Object Detection with Transformers, ECCV'2020

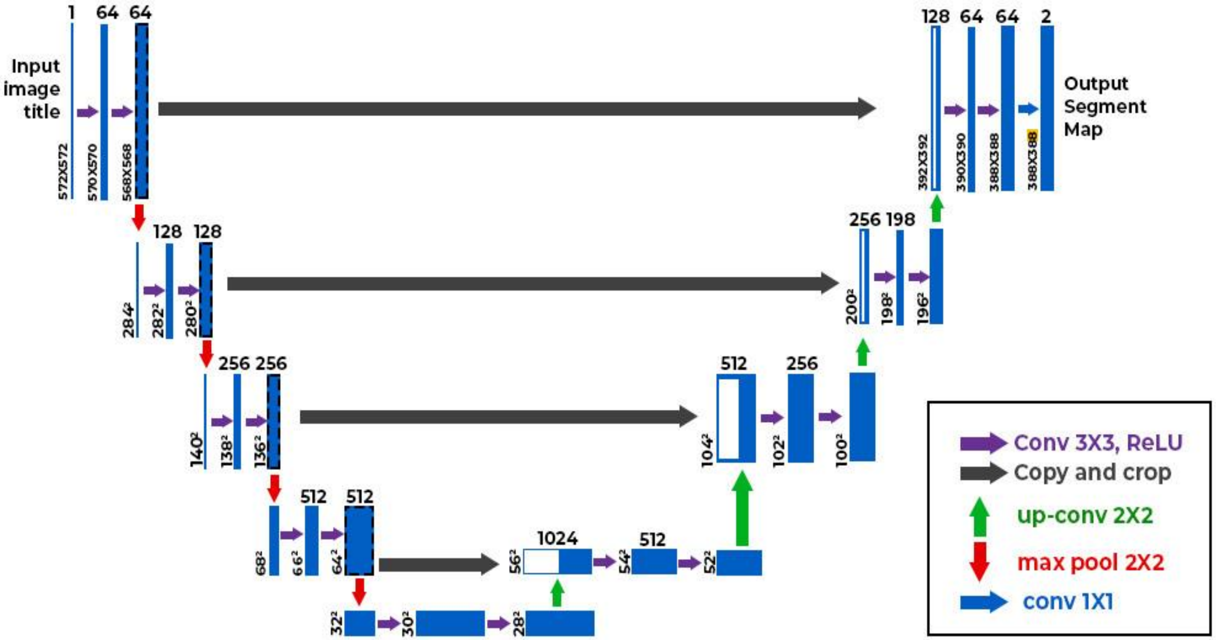
Basic architectures for segmentation FM

- Based on two major architectures:
  - CNN
  - Transformers
- Different combinations of encoders and decoders
  - U-net : Convolutional encoder and decoder
  - TransU-net: Transformer encoder and CNN decoder
  - SETR: Vision transformer and CNN decoder
  - Segformer: transformer encoder and decoder
  - CLIPSeg: CLIP encoder and transformer decoder
- Go from limited object labels to open vocabulary segmentation
- Mostly still supervised, but newer unsupervised open set models also coming back

U-net - A classic CNN-style segmentation model

- Extensively used in medical imaging
- Works for both anatomy and anamoly detection
- Top-down segmentation works best for medical images since intensities are intermingled.
- Works well when the pattern to learn is in a fixed place

- Skip connections help remember features
- Fully CNN-based

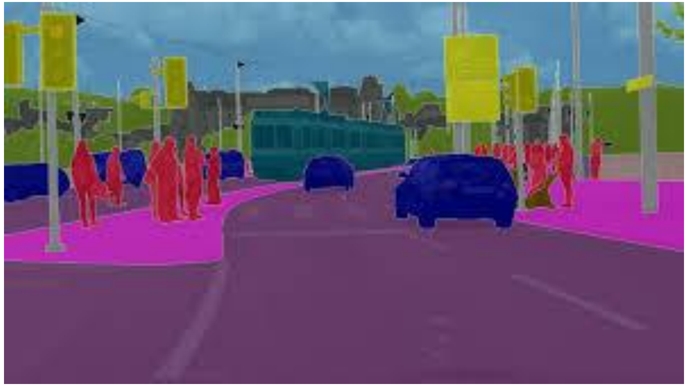


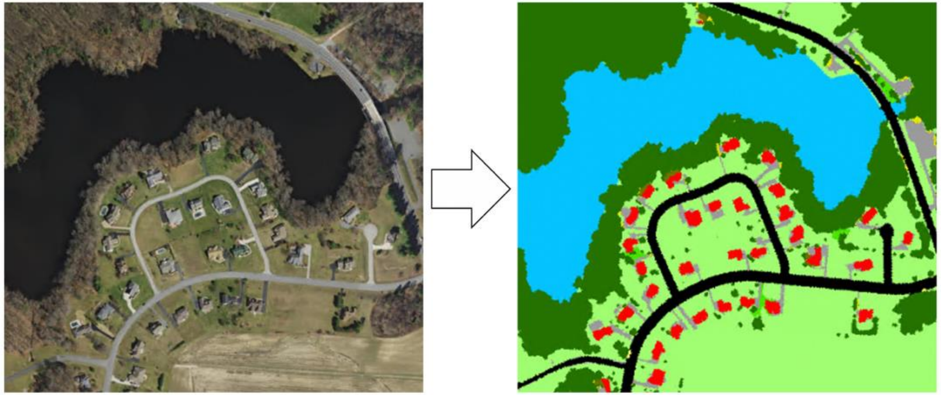
Ronnenberger: U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015

109949 citations!

U-net applications

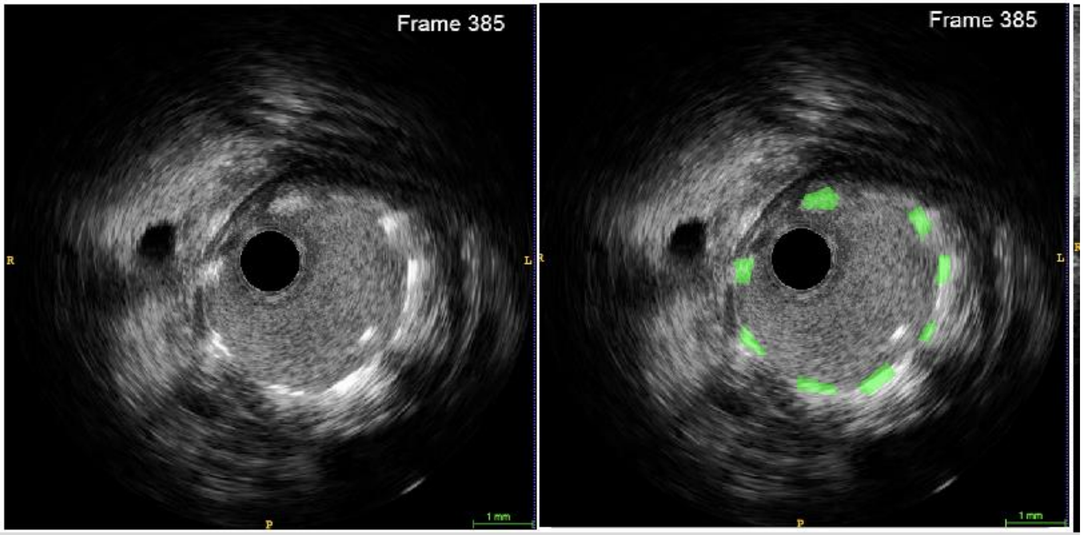
- Both medical and non-medical applications

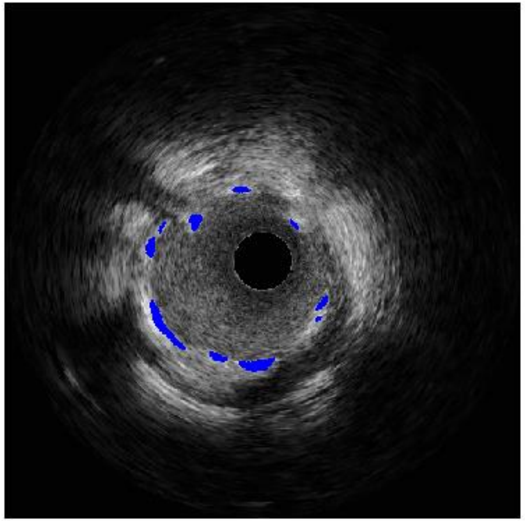




U-Net issues

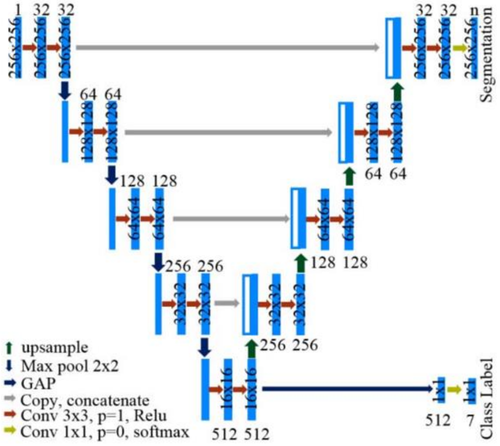
- Detections are not smooth
- Maintaining continuity is difficult
- Fragmentation can occur
- Small tumors and anatomical regions
- Stable contour detection





Making segmentation model foundational

- Train with a variety of datasets
- Adapt transformer or CNN-based encoders and decoders
- An early example of a foundational model for medical image segmentation – based on V-net



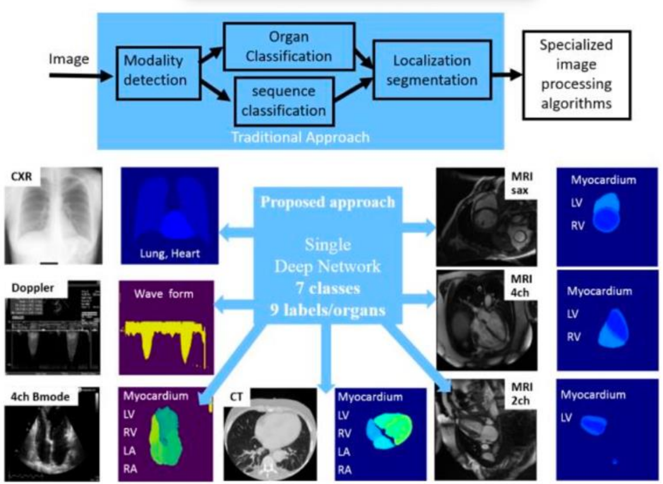


Figure 1: Top: Traditional architecture tackling one problem at a time. Bottom: our proposed network to both classify different modalities with different viewpoints (X-ray, CT, Ultra sound, 2 chamber MRI, 4 chamber MRI, short axis MRI) as well as segment different structures as Lung, heart, Doppler wave form, Myocarduim (Myo), left ventrical (LV), right ventrical (RV) ,left atrium (LA), and right atrium (RA).

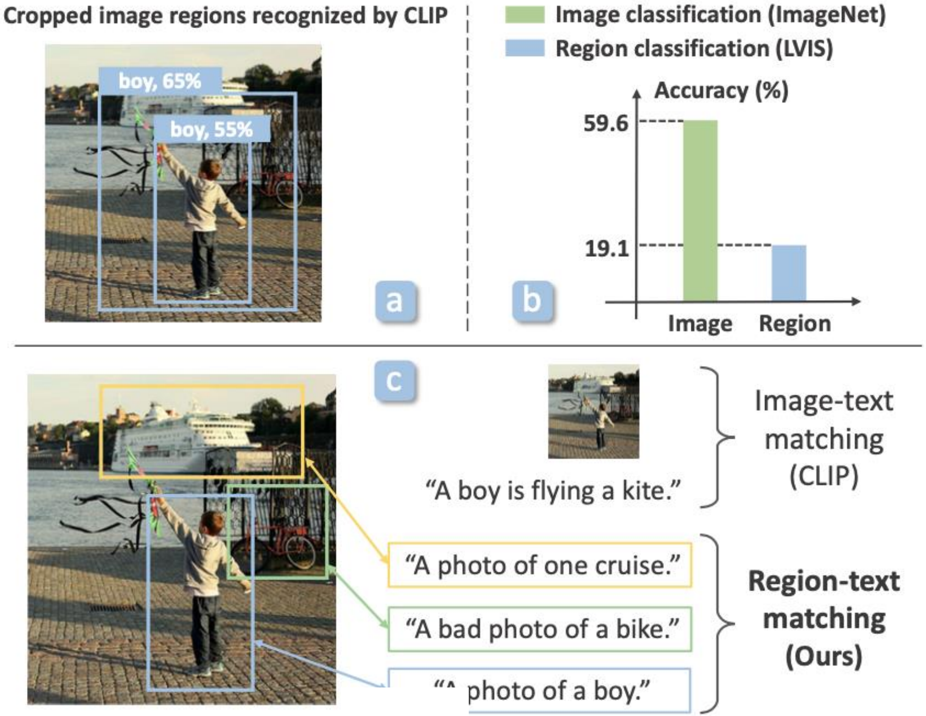
Harouni et al., "Universal multimodal deep network for classification and segmentation of medical images," in ISBI 2018.

Segmentation models derived from VLM models

- Use joint visual-textual information to aid segmentation
- Use CLIP underneath for the encoding
- Built separate decoders to aid in segmentation
    - CNN-based
    - Transformed-based decoders
- Segmentation at the level of bounding boxes
    - RegionCLIP
    - VILD
    - GLIP
- Segmentation as full regions
    - CLIPSeg

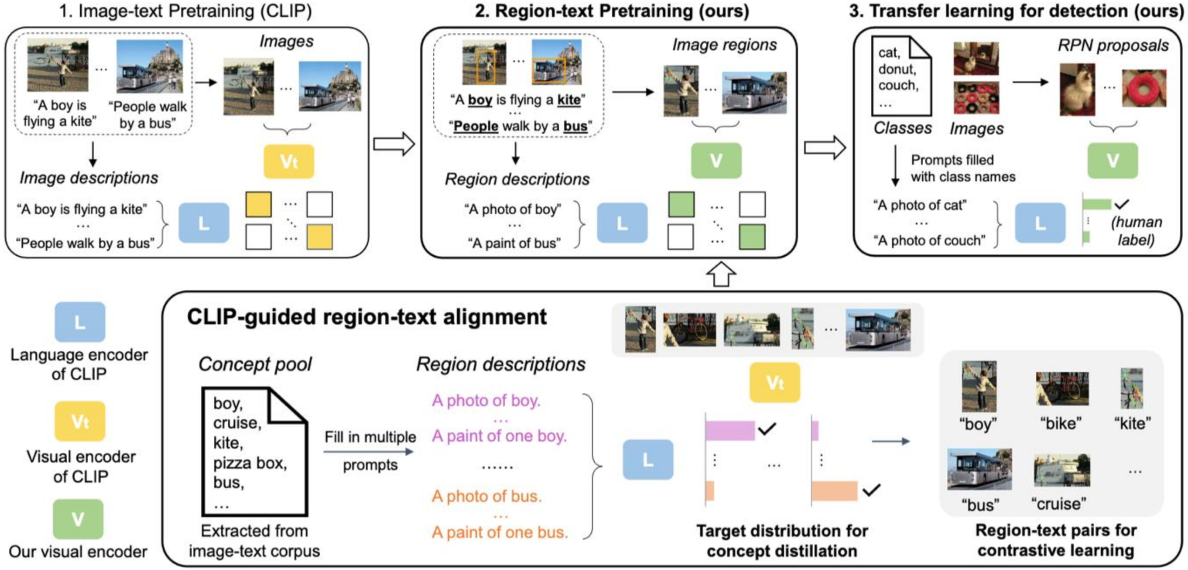
Using VLM for generalized instance segmentation

- Use region crops from RPN
- Initial CLIP image-to-text to label
- What if we do region detection and CLIP on regions fails to recognize objects


Cropped image regions recognized by CLIP
boy, 65%
boy, 55%
a
Image classification (ImageNet)
Region classification (LVIS)
Accuracy (%)
59.6
19.1
b
Image
Region
c
"A boy is flying a kite."
Image-text matching (CLIP)
"A photo of one cruise."
"A bad photo of a bike."
"A photo of a boy."
Region-text matching (Ours)

RegionCLIP – Contrastively learned region-text alignment

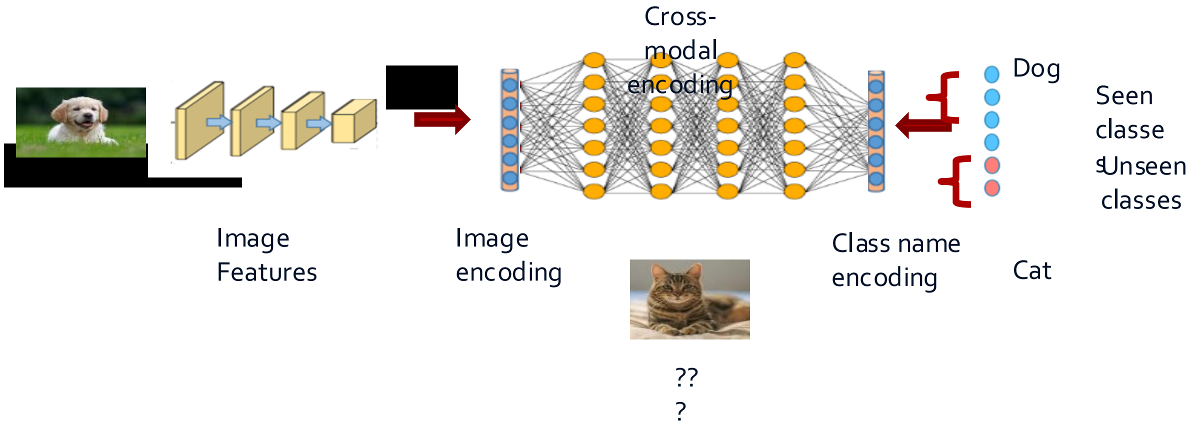
VLM – Visual Grounding



RegionCLIP – Contrastively learned region-text alignment

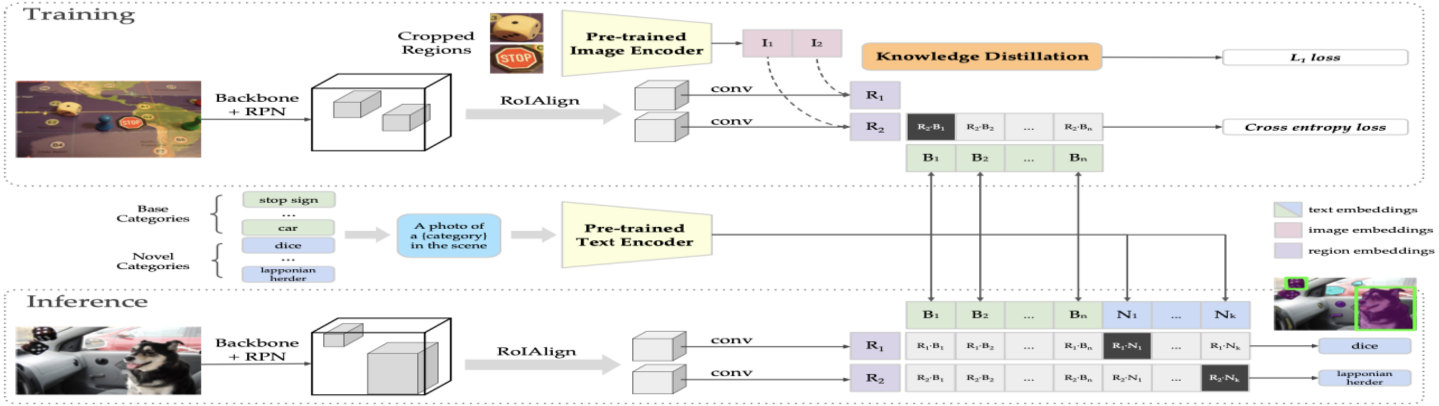
VILD architecture

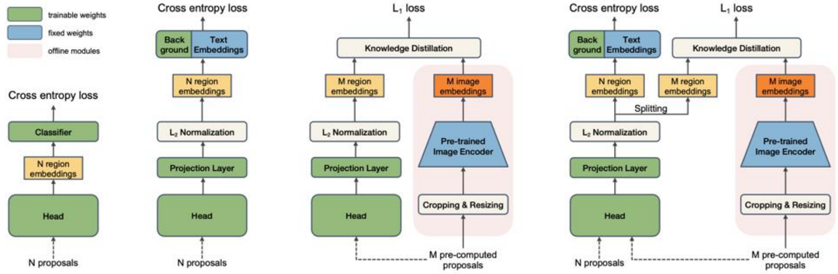
- Combines pre-trained FM (CLIP) with MaskRCNN-style RPN and zero-shot learning methods
- Distill the knowledge from the alignment of region embedding to image embeddings of cropped regions during training.


[redacted]
[redacted]

Xiuye Gu et al, ”Open vocabulary object detection vis vision and language knowledge distillation,” ICLR22

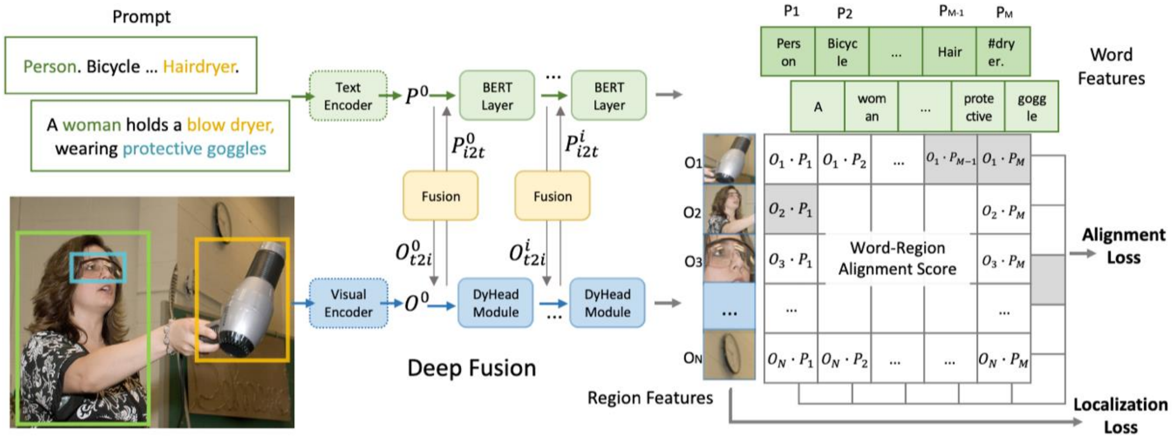
VilD – Visual Grounding

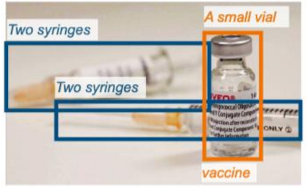




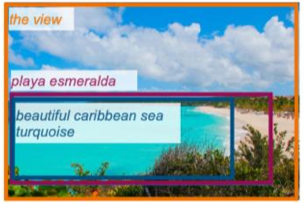
(a) Vanilla detector (b) ViLD-text (c) ViLD-image (d) ViLD

VLM – Phrase Grounding





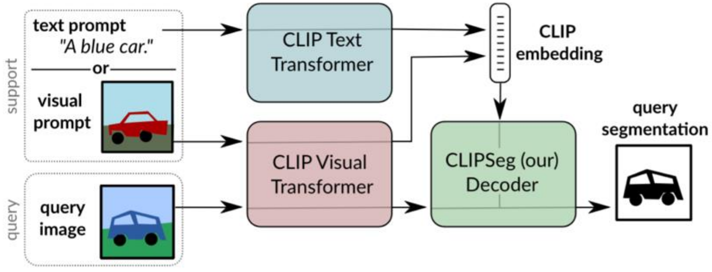
Two syringes and a small vial of vaccine.

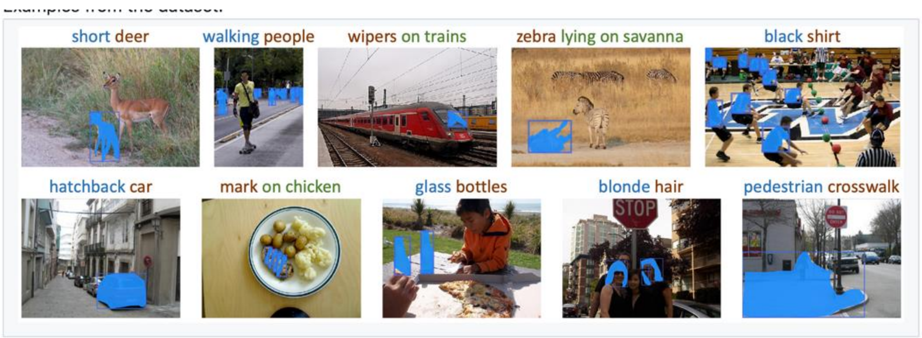


playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

- Given an image and a corresponding caption, the **Phrase Grounding** task aims to ground each entity mentioned by a noun phrase in the caption to a region in the image.
- GLIP – grounded language pre-training

CLIPSeg



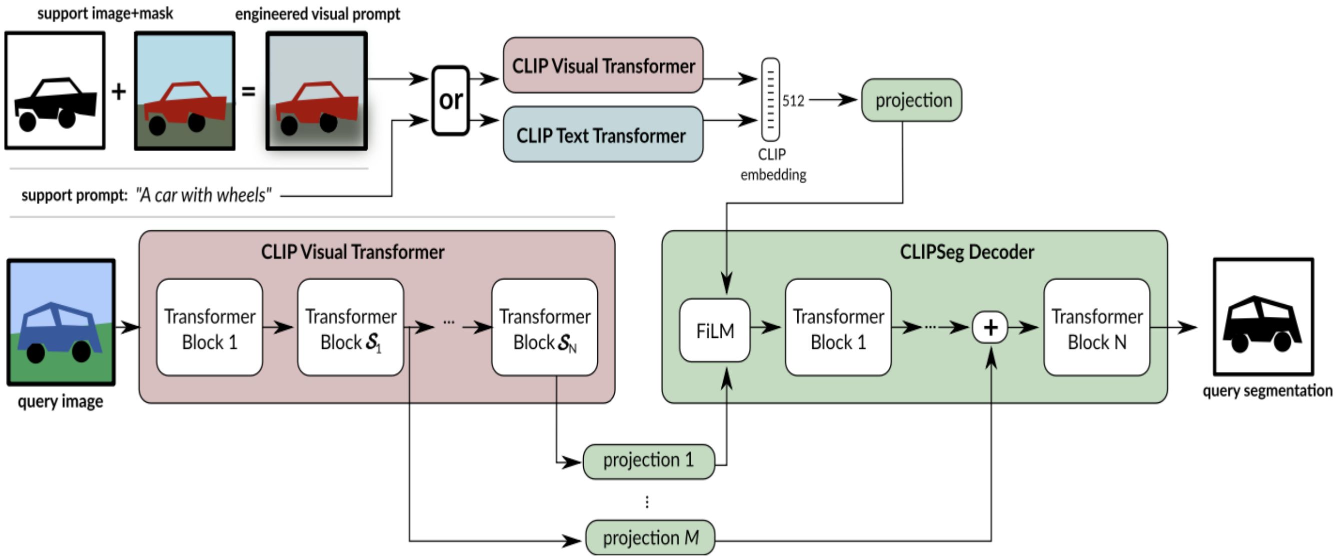


- Uses the PhrasCut dataset of regional masks derived from VisualGenome
- Combined CLIP with a lighweight decoder

CLIPSeg: Image segmentation using text and image prompts

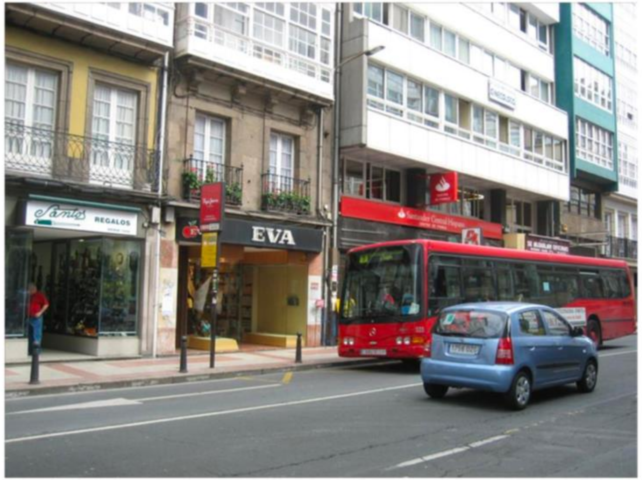
ClipSeg

- Skip connections to preserve the visual context from 3 of the layers of the VIT to the decoder as in U-net
- The decoder incoming dimension is 64

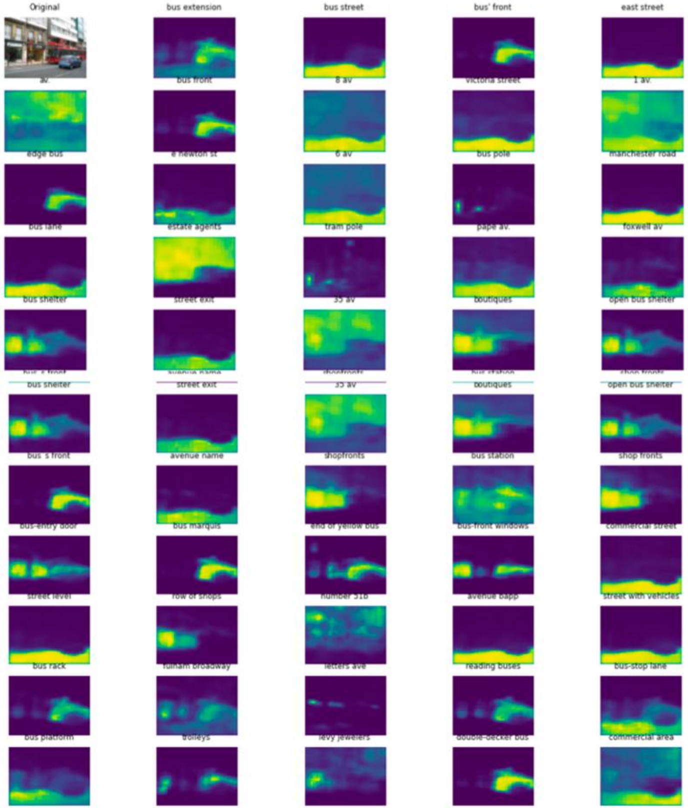


CLIPSeg results

(a)



(b)

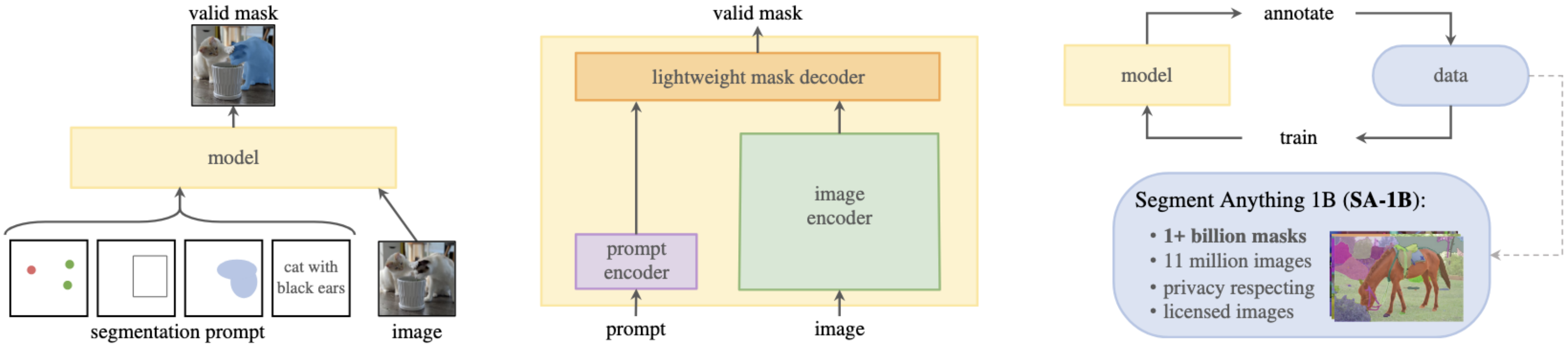


'victoria street', 'manchester road', 'street level', 'foxwell av', '6 av', 'east street', 'street with vehicles', 'bus-entry door', 'bus`s front', 'shopfronts', 'commercial street', 'bus rack', 'bus front', 'street exit', '8 av', 'estate agents', 'bus lane', '35 av', 'avenue name', 'bus marquis', 'trolleys', "bus' front", 'avenue bapp', 'row of shops', 'boutiques', 'e newton st', 'bus street', 'bus station', 'bus pole', 'edge bus', 'pape av.'

(c)

Segment Anything (SAM)

- A foundation model for image segmentation.
- A promptable model and pre-train it on a broad dataset using a task that enables powerful generalization



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

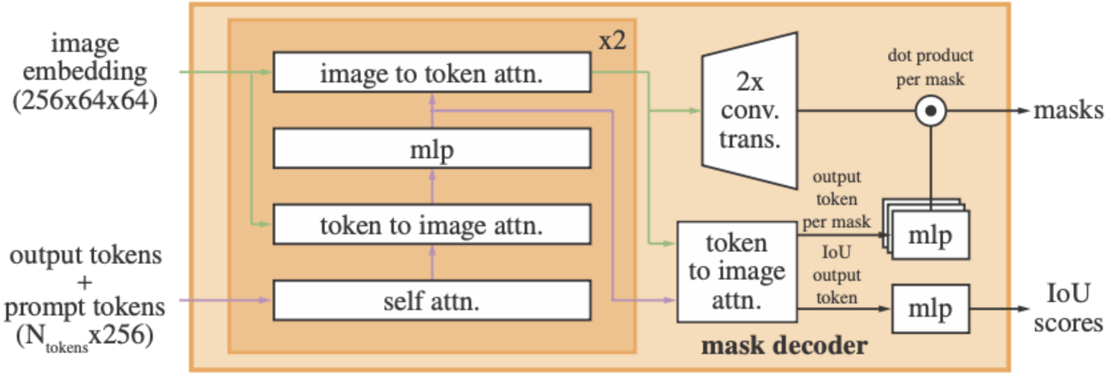
Different kinds of prompts to aid in segmentation

SAM components

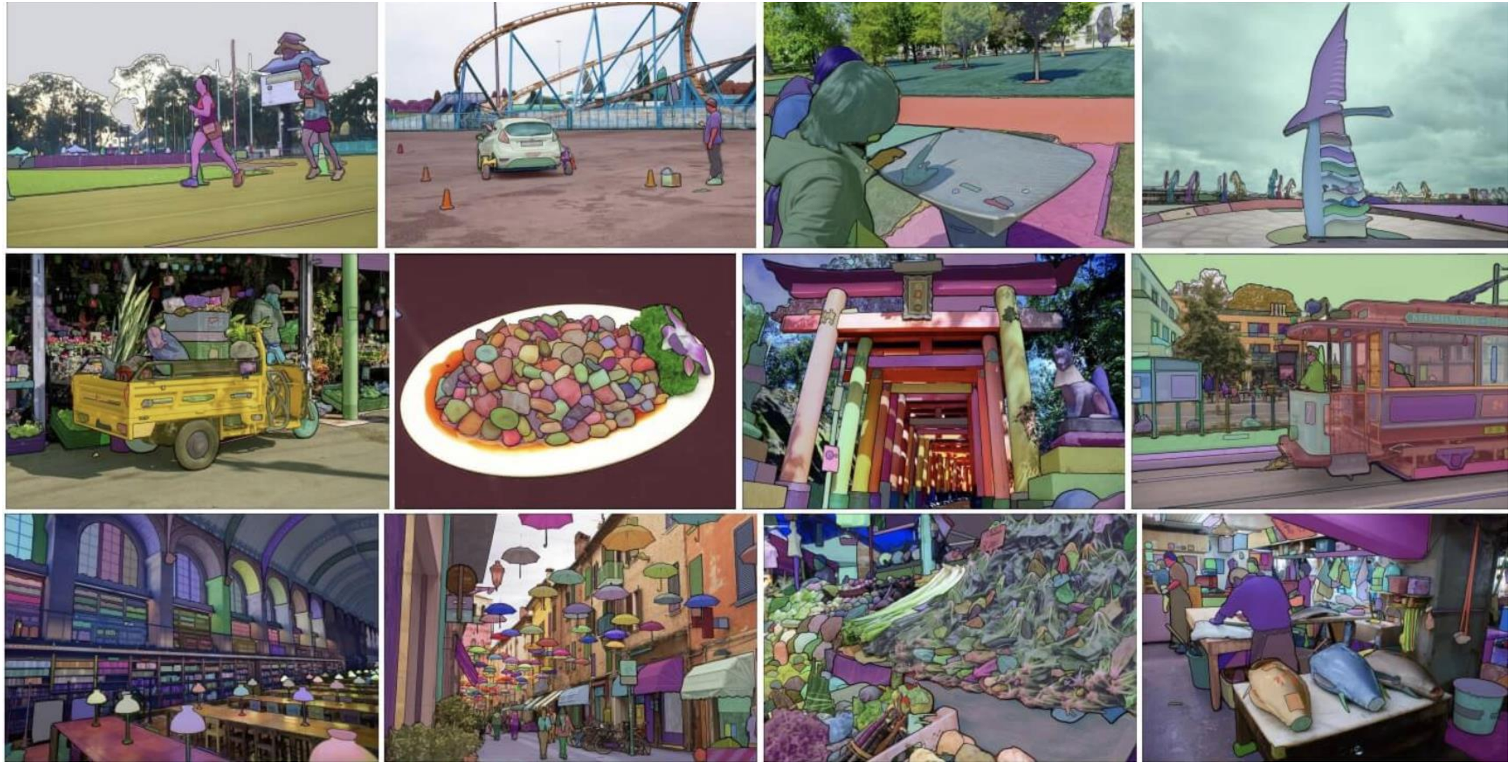
- Image Encoder:
  - Encoder portion of the Masked Auto-encoder which is a pre-trained Vision Transformer (ViT) with adaptations to process high resolution inputs,
  - input resolution of 1024×1024 obtained by rescaling the image and padding the shorter side. The image embedding is therefore 64×64x256.
- Prompt encoder:
  - Produces a 256-dimensional vector
  - Points
    - Point is represented as a sum of positional encoding of the points's location and one of the two learned embeddings to indicate either a **foreground point** or **background point**.
  - Bounding boxes:
    - Boxes are represented by an **embedding pair of corners**.
  - Text:
    - CLIP text-encoding
  - Masks:
    - Downscaled versions of masked images CNN-style and flattened to 256-dimensional vectors

Segment Anything – Mask Decoder

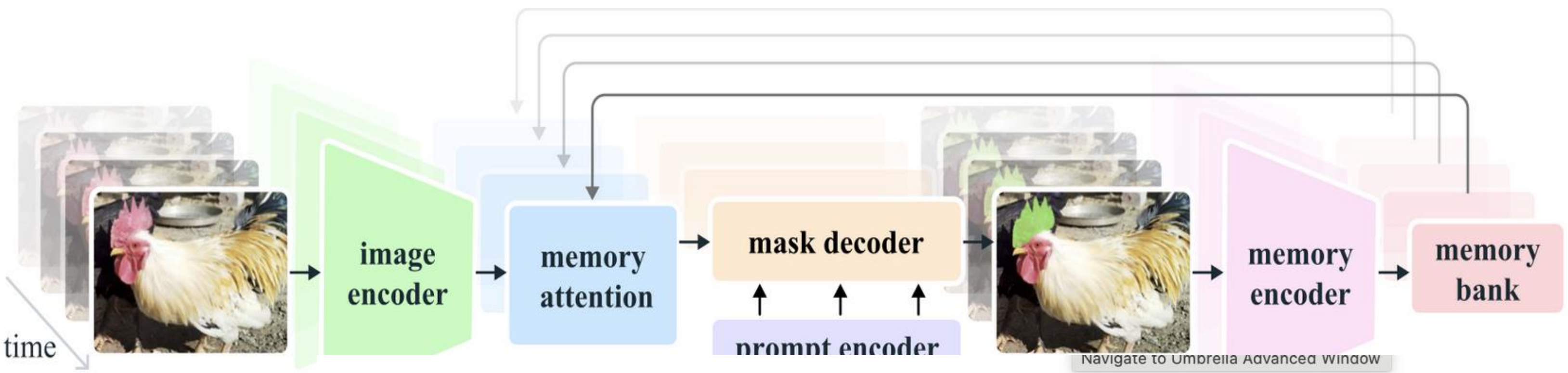
- Image embeddings and prompt embeddings are mapped to the final mask
- uses prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update *all* embeddings.
- MLP maps the output token to a dynamic linear classifier,
- which then computes the mask foreground probability at each image location.



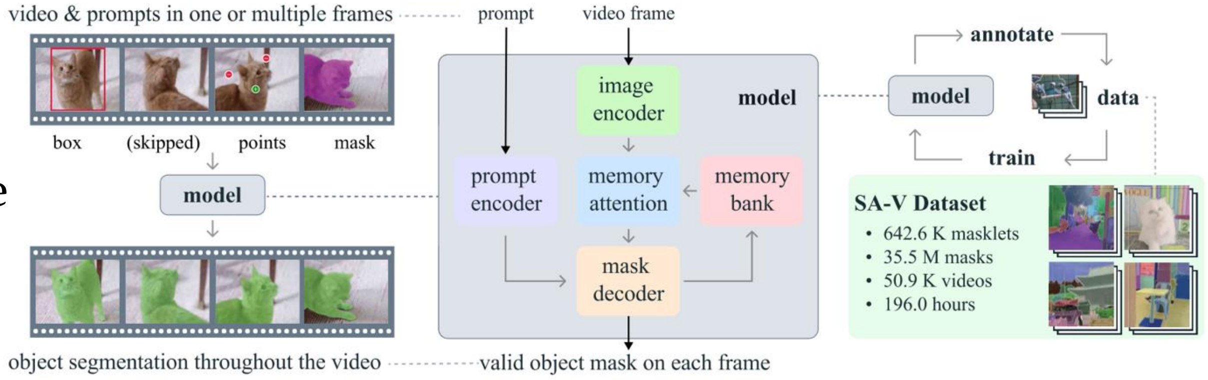
SAM results



SAM-2

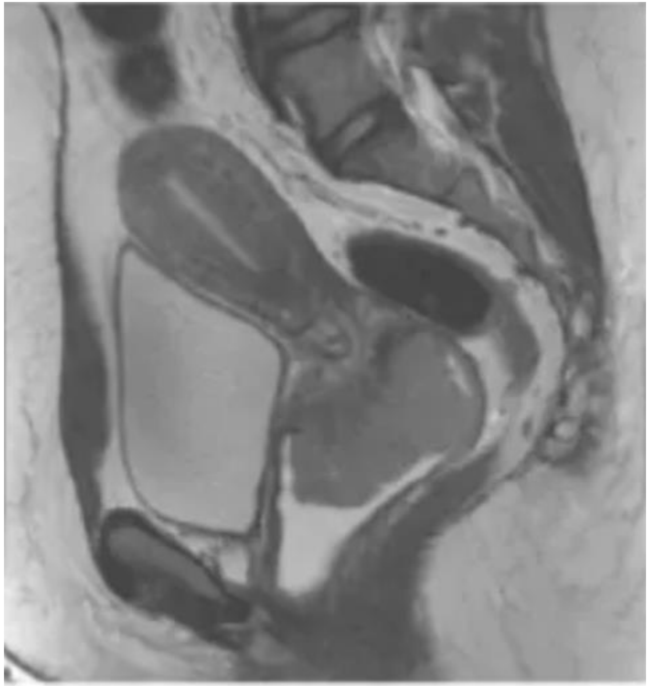


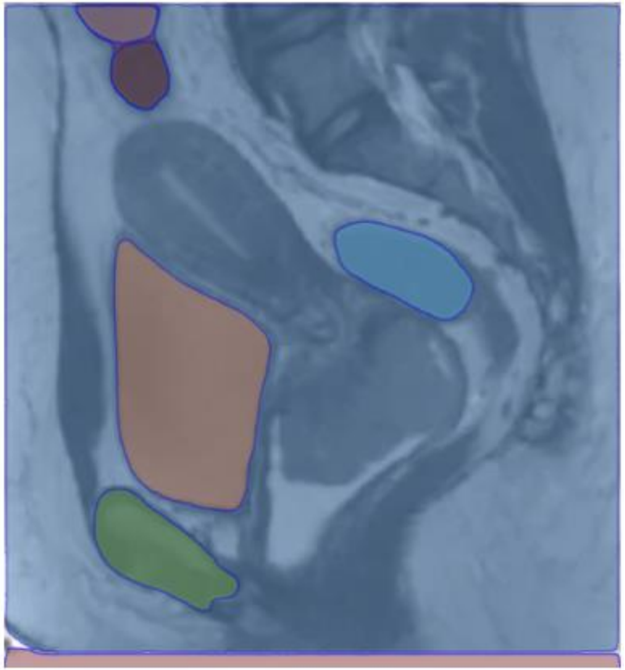
Extends to video sequences by retaining previous frame information



(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

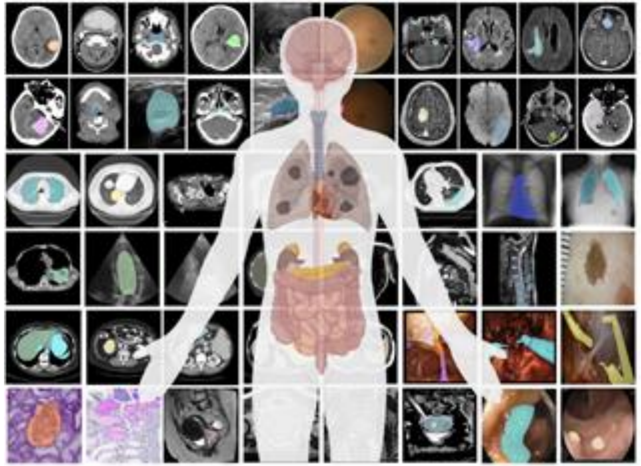
How well does SAM do on medical images?





Foundational models for segmentatio

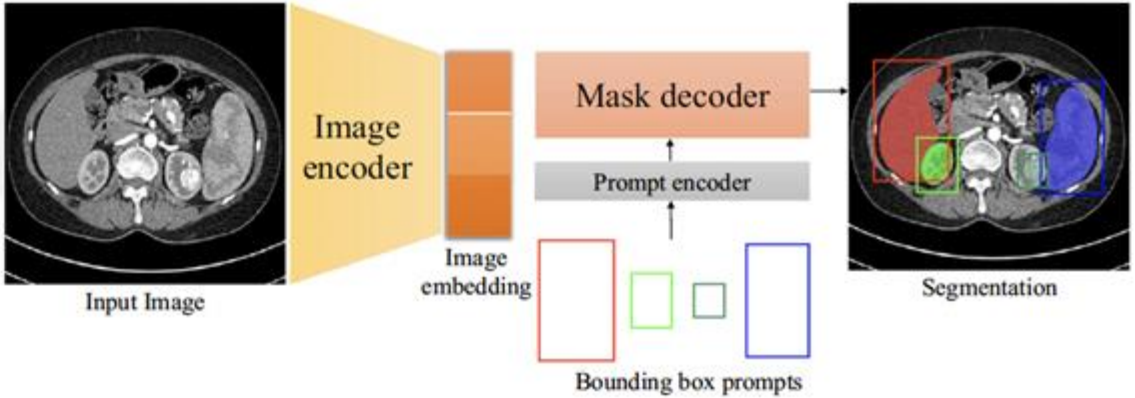
MedSAM



Trained on an huge amount of data

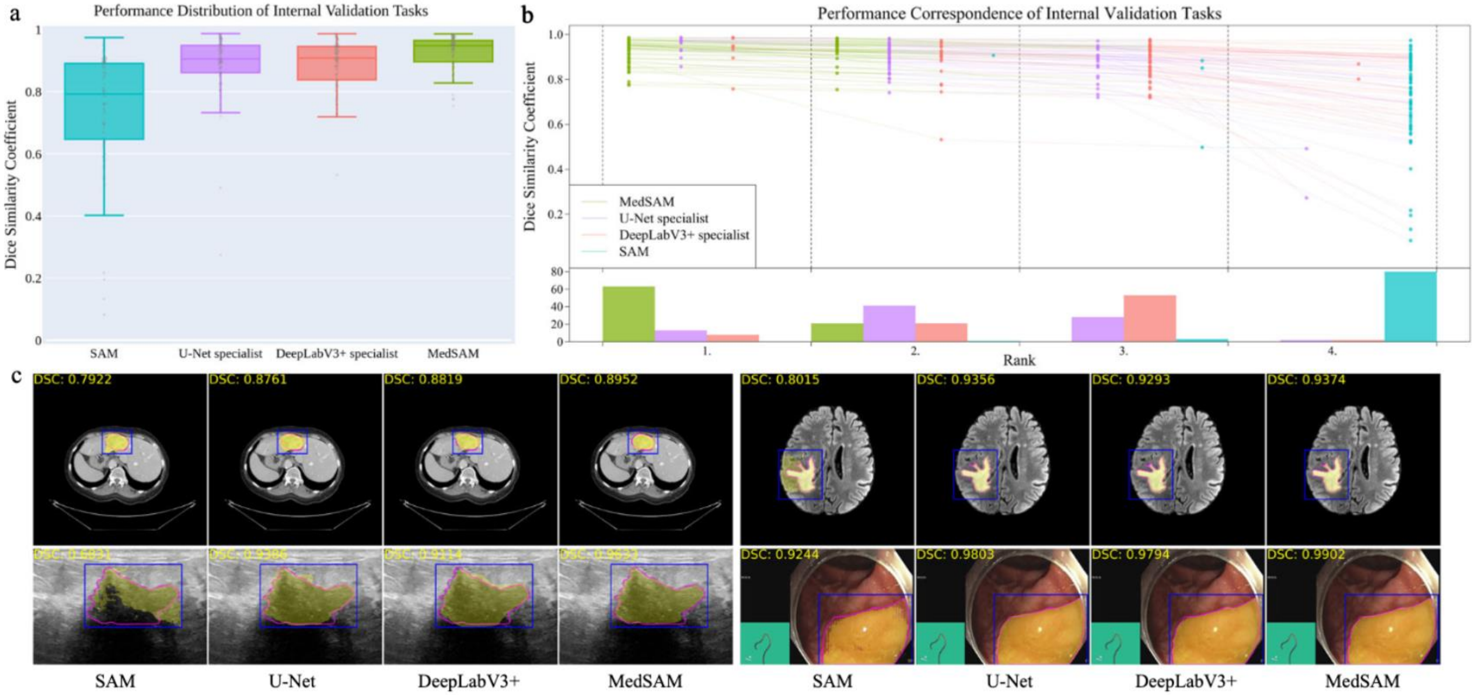
- 10 imaging modalities
- 30 cancer types
- Over 1.5M mask-image pairs

Pipeline



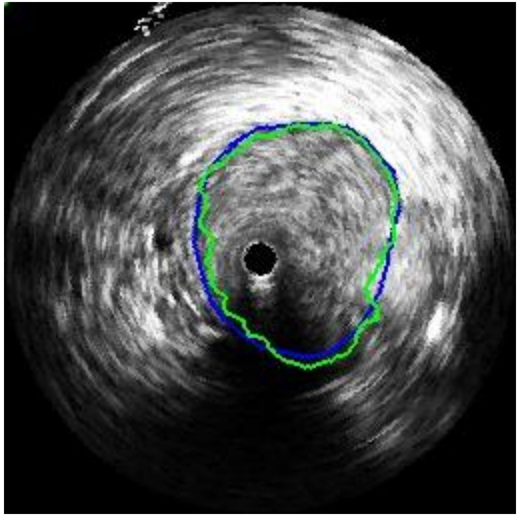
Ma et *al*. Segment Anything in medical images, Nature Communications 2024

MedSAM generalization

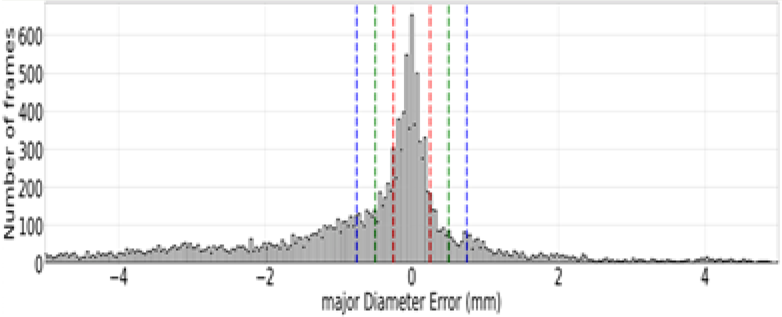


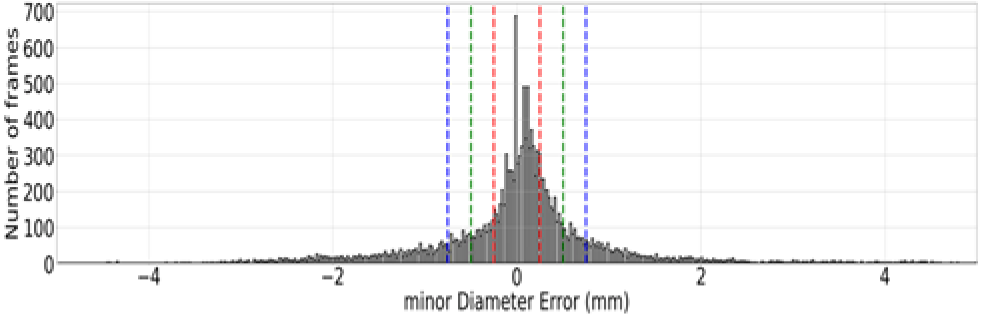
Are FMs ready for high-precision segmentation?

- Needed for accurate sizing of the stents
- Based on normal frames identification
- Estimates the maximum and minimum diameter
- Major and minor diameter errors within:
  - 0.25/0.5/0.75mm for 50/90/95% of all N1 frames.
  - 0.5/0.75mm for 50/70% of frames for N2 frames
- N2 normal are mainly used for vessel compression detection and not stent sizing
- Simple U-net will not suffice as single regions cannot be ensured
- Contours may not closely follow the lumen boundaries

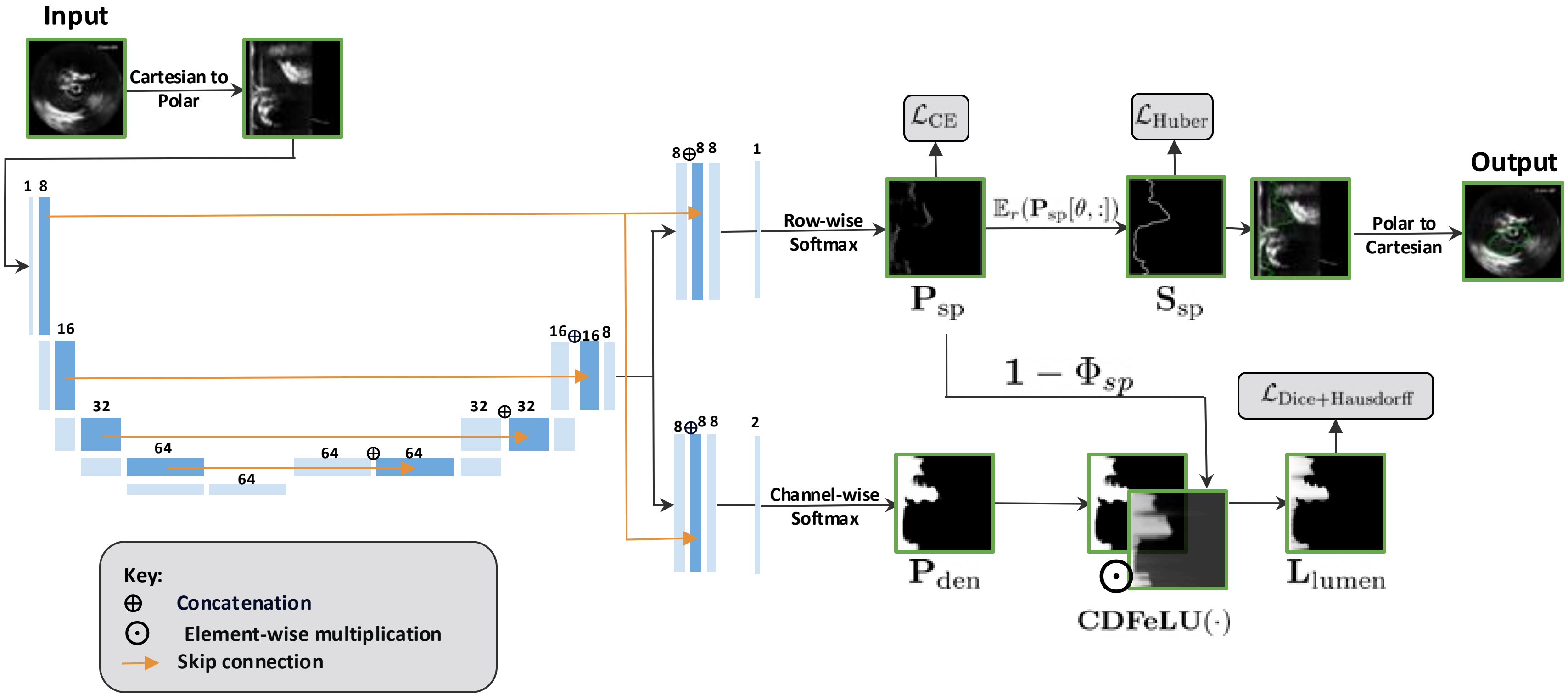


% within 0.25/0.5/0.75mm





Geo-UNet Architecture


Input
Cartesian to Polar
1 8
16
32
64
64
64 ⊕ 64
32 ⊕ 32
16⊕16 8
8⊕8 8
1
8⊕8 8
2
Row-wise Softmax
Channel-wise Softmax
$\mathcal{L}_{\mathrm{CE}}$
$\mathbf{P}_{\mathrm{sp}}$
$\mathbb{E}_r(\mathbf{P}_{\mathrm{sp}}[\theta,:])$
$\mathcal{L}_{\mathrm{Huber}}$
$\mathbf{S}_{\mathrm{sp}}$
Polar to Cartesian
Output
$1-\Phi_{sp}$
$\mathcal{L}_{\mathrm{Dice+Hausdorff}}$
$\mathbf{P}_{\mathrm{den}}$
⊙
$\mathrm{CDFeLU}(\cdot)$
$\mathbf{L}_{\mathrm{lumen}}$

Key:

⊕ Concatenation

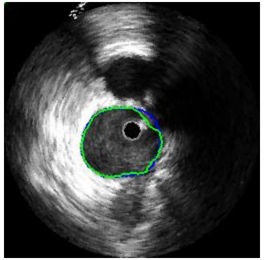
⊙ Element-wise multiplication

→ Skip connection

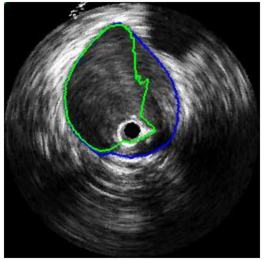
in Y. Chen et al, “Geo-UNet: A geometrically constrained neural framework for clinical grade lumen segmentation in intravascular ultrasound,” in Proc. Machine Learning for Medical Imaging (MLMI), MICCAI 2024

Results

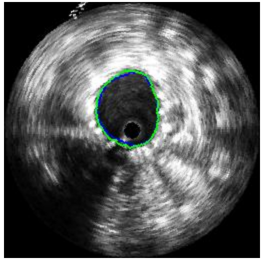
Geo-Unet++ results



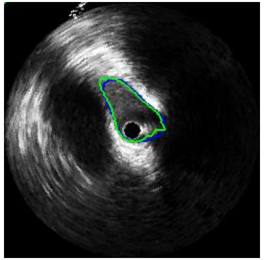
Near perfect segmentation



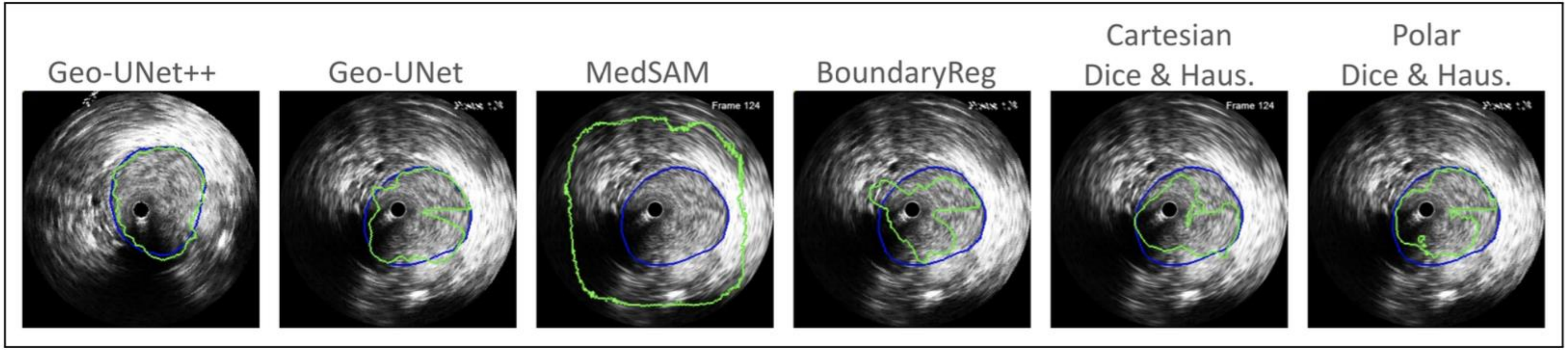
Segmentation in a case where nearby vessel is present

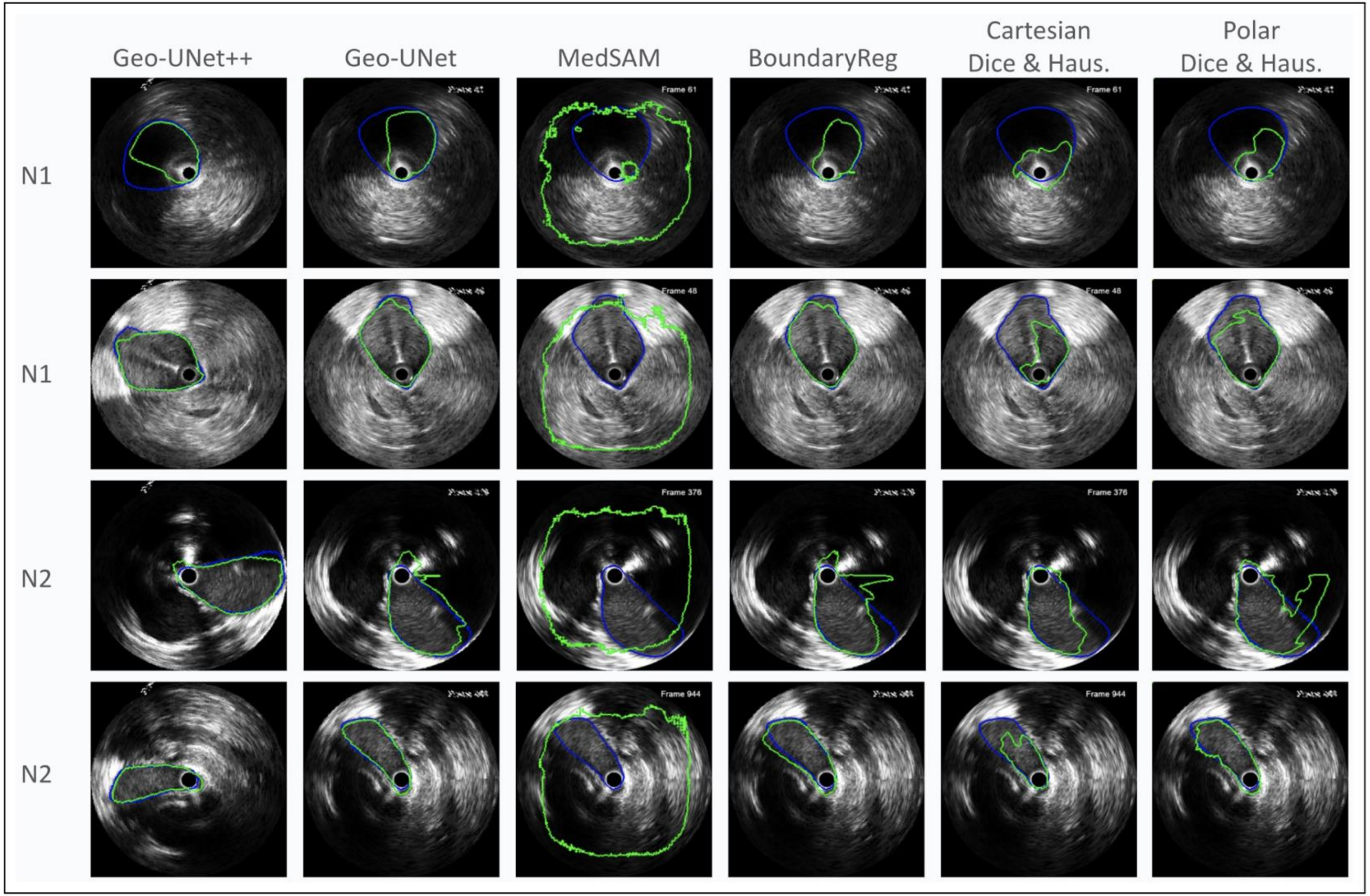


Hard case in the presence of stents

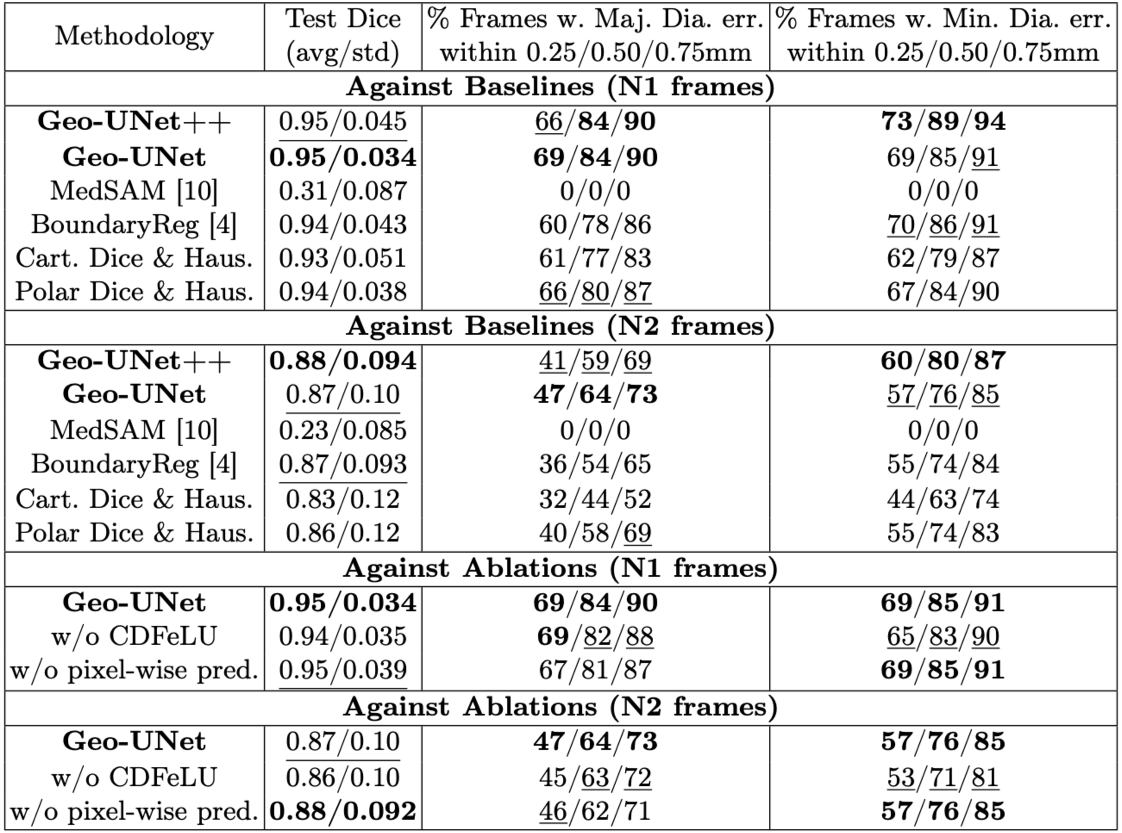


Hard case with confluence




Geo-UNet++
Geo-UNet
MedSAM
BoundaryReg
Cartesian Dice & Haus.
Polar Dice & Haus.
N1
N1
N2
N2
Frame 61
Frame 48
Frame 376
Frame 944

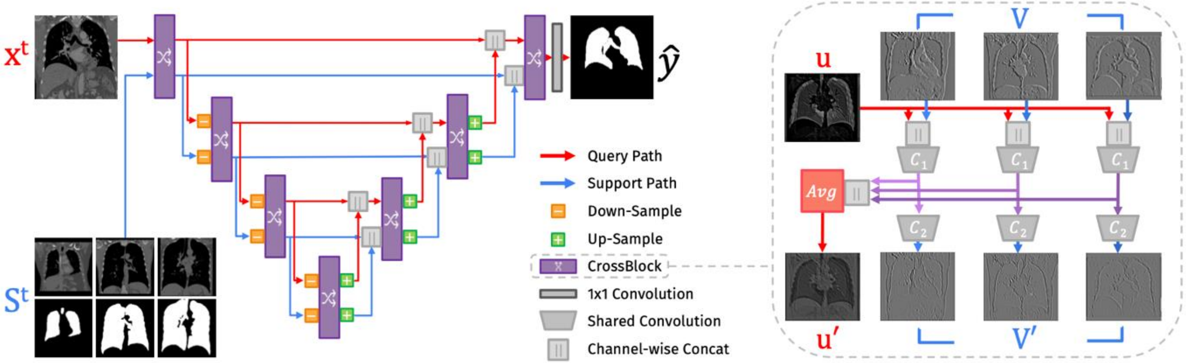
Results



Methodology	Test Dice (avg/std)	% Frames w. Maj. Dia. err. within 0.25/0.50/0.75mm	% Frames w. Min. Dia. err. within 0.25/0.50/0.75mm
Against Baselines (N1 frames)			
Geo-UNet++	0.95/0.045	66/**84**/**90**	**73**/**89**/**94**
Geo-UNet	**0.95/0.034**	**69**/**84**/**90**	69/85/91
MedSAM [10]	0.31/0.087	0/0/0	0/0/0
BoundaryReg [4]	0.94/0.043	60/78/86	70/86/91
Cart. Dice & Haus.	0.93/0.051	61/77/83	62/79/87
Polar Dice & Haus.	0.94/0.038	66/80/87	67/84/90
Against Baselines (N2 frames)			
Geo-UNet++	**0.88/0.094**	41/59/69	**60**/**80**/**87**
Geo-UNet	0.87/0.10	**47**/**64**/**73**	57/76/85
MedSAM [10]	0.23/0.085	0/0/0	0/0/0
BoundaryReg [4]	0.87/0.093	36/54/65	55/74/84
Cart. Dice & Haus.	0.83/0.12	32/44/52	44/63/74
Polar Dice & Haus.	0.86/0.12	40/58/69	55/74/83
Against Ablations (N1 frames)			
Geo-UNet	**0.95/0.034**	**69**/**84**/**90**	**69**/**85**/**91**
w/o CDFeLU	0.94/0.035	**69**/82/88	65/83/90
w/o pixel-wise pred.	0.95/0.039	67/81/87	**69**/**85**/**91**
Against Ablations (N2 frames)			
Geo-UNet	0.87/0.10	**47**/**64**/**73**	**57**/**76**/**85**
w/o CDFeLU	0.86/0.10	45/63/72	53/71/81
w/o pixel-wise pred.	**0.88/0.092**	46/62/71	**57**/**76**/**85**

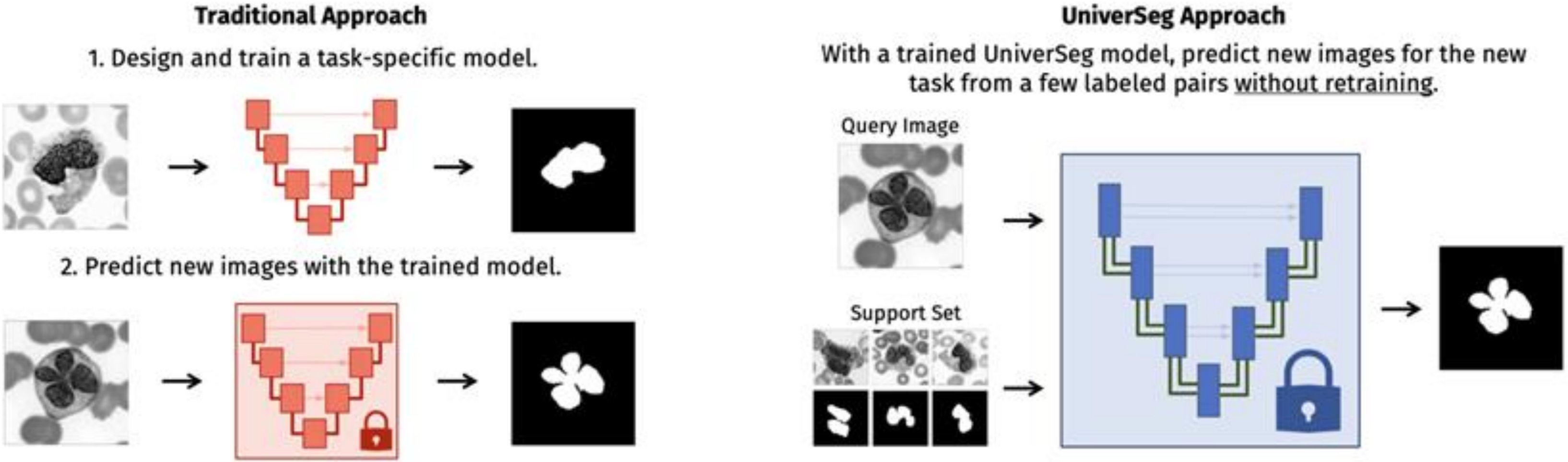
Instructional tuning of segmentation models

- Based on U-net style convolutional architecture
- Uses multi-scale cross-block features between instructional labeling sets and given image
- No retraining needed or fine-tuning needed!
- Trained on Megamedical dataset: 53 datasets, 23 medical domains, 16 modalities



Butai et al. Universal medical image segmentation, ICCV’2023

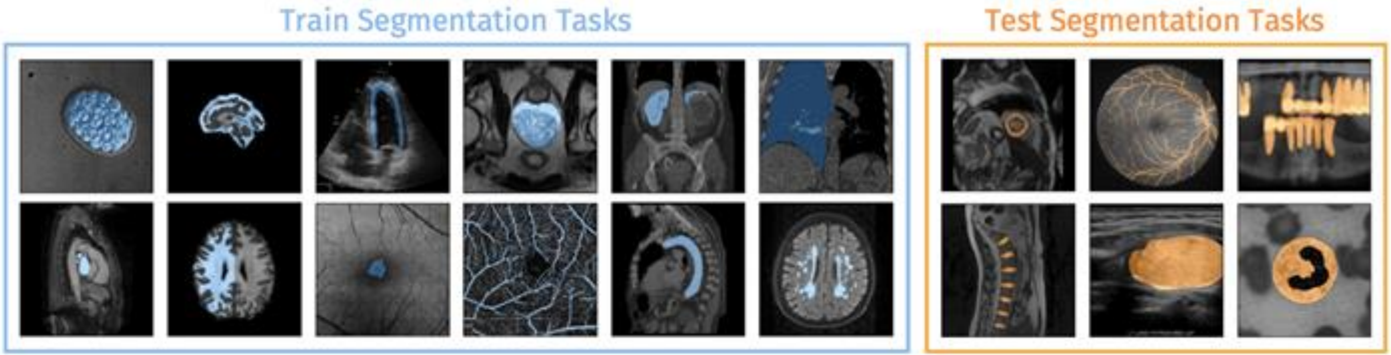
Meta learning in UniverSeg



1. Design and train a task-specific model.

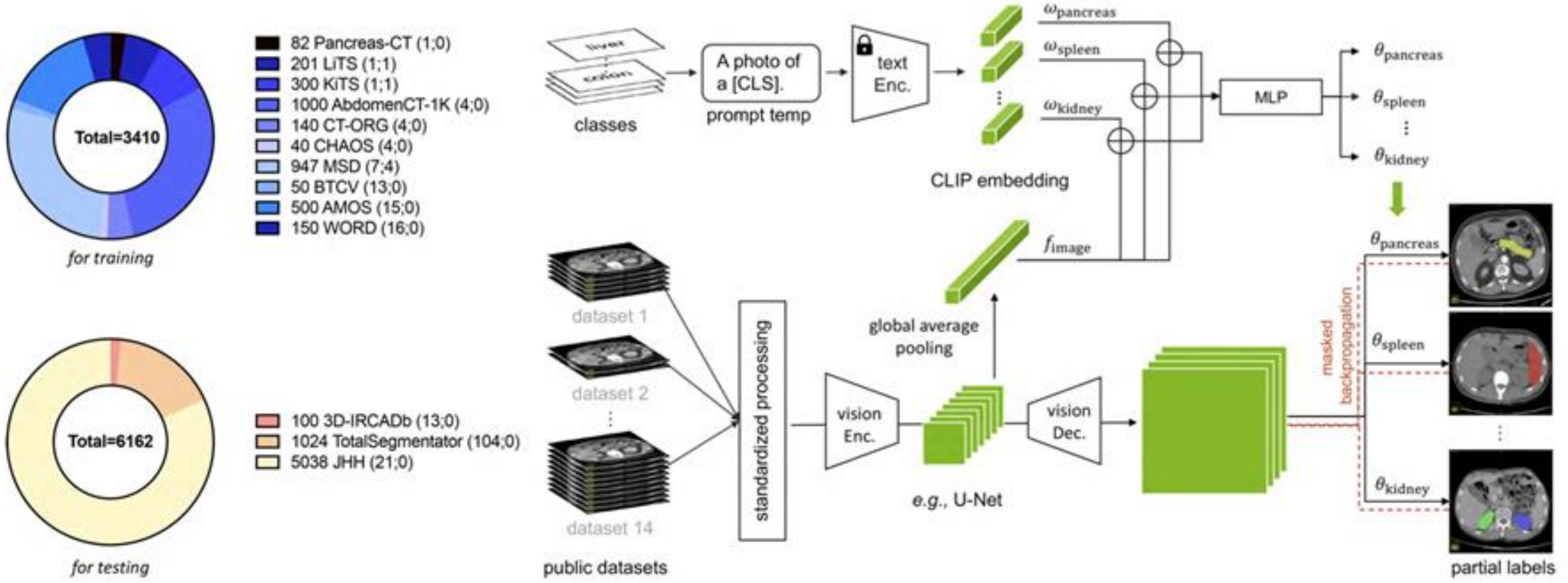
2. Predict new images with the trained model.

With a trained UniverSeg model, predict new images for the new task from a few labeled pairs without retraining.



Butoi et *al*. Universeg: Universal medical image segmentation. ICCV’23

Combining U-net and CLIP for anatomy and tumor segmentation



Slides from MICCAI 2024 tutorial

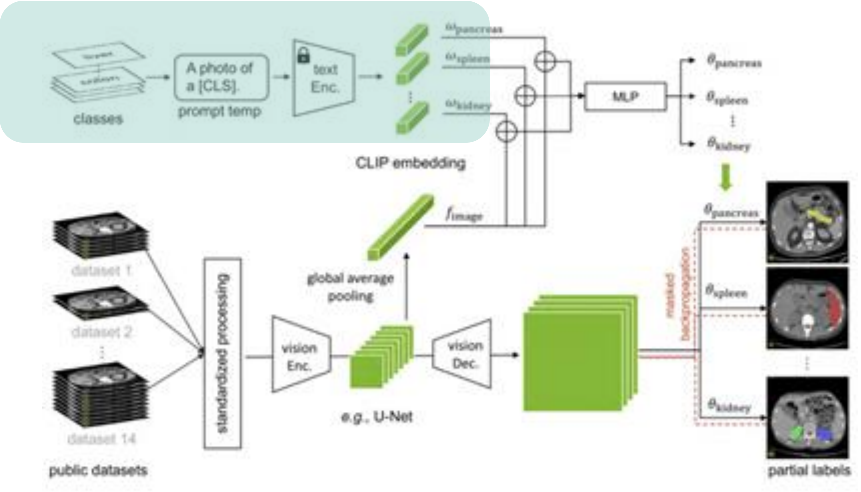
Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV’23

Foundational models for segmentatio

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

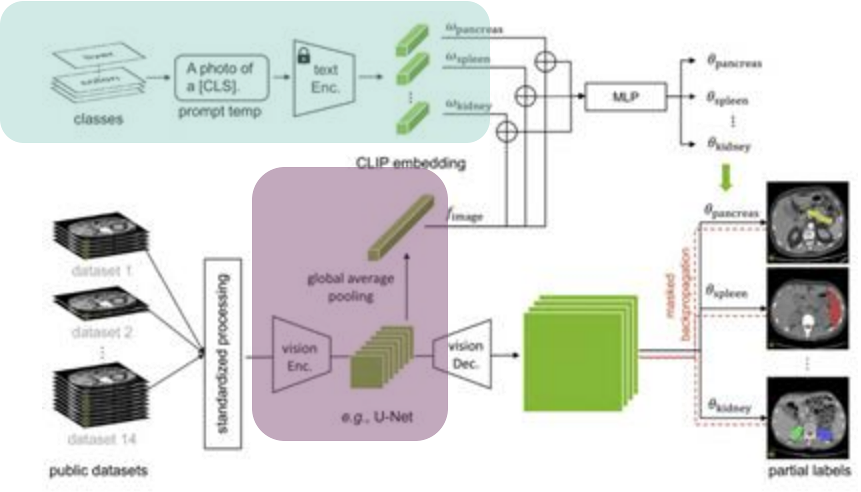
Foundational models for segmentatio

CLIP-Driven

Main idea

Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentatio

CLIP-Driven

Main idea

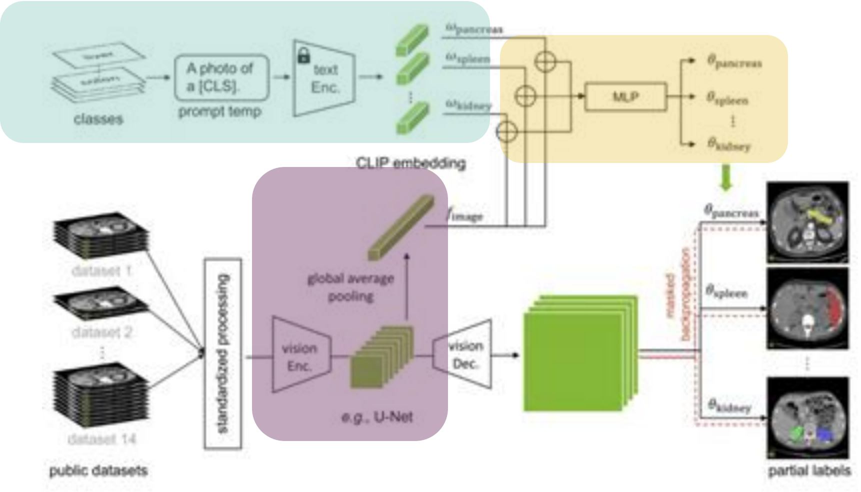
Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_k = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV’23

Foundational models for segmentatio

CLIP-Driven

Main idea

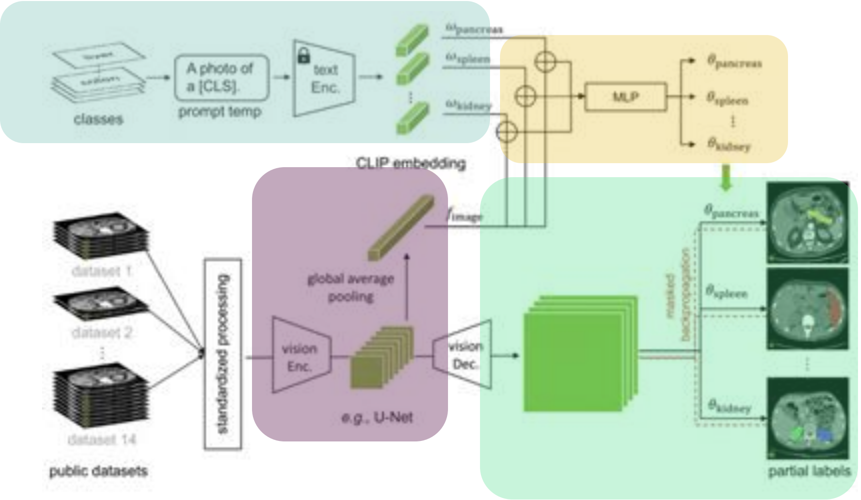
Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_{\boldsymbol{k}} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

It represents foreground class k *vs* background

Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Foundational models for segmentatio

CLIP-Driven

Main idea

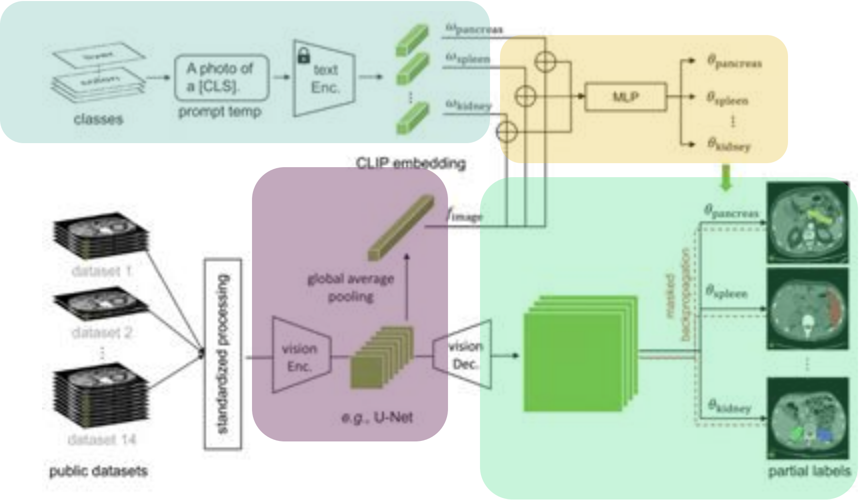
Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_k = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

It represents foreground class k *vs* background

Training loss

Binary cross-entropy per class (and terms masked for those classes not present)

$$\mathcal{L} = \sum_{k=1}^{K} \mathbf{1}_{\{k \in y\}} \cdot \text{BCE}_k$$

Liu et *al*. CLIP-Driven universal model for organ segmentation and tumor detection. ICCV'23

Summary

- Early approaches to segmentation were unsupervised and didn’t scale
- Supervised approaches had limited labels issues
- Foundational models generalized across datasets and open vocabularies
- Key architectures are still based on CNN or transformers
- Active field of research in developed generalized foundation models for segmentation with extensions to video sequences
- Gap seen with applications to medical imaging leading to some rich innovations for medical imaging adaptations
- Unsupervised segmentation approaches may still be relevant
  - STEGO: Unsupervised Semantic Segmentation by Distilling Feature Correspondences, ICLR’2022 -> Learns with no labels