

Deep learning for medical imaging

Olivier Colliot, PhD
Research Director at CNRS
Co-Head of the ARAMIS Lab –
www.aramislab.fr
PRAIRIE – Paris Artificial Intelligence
Research Institute

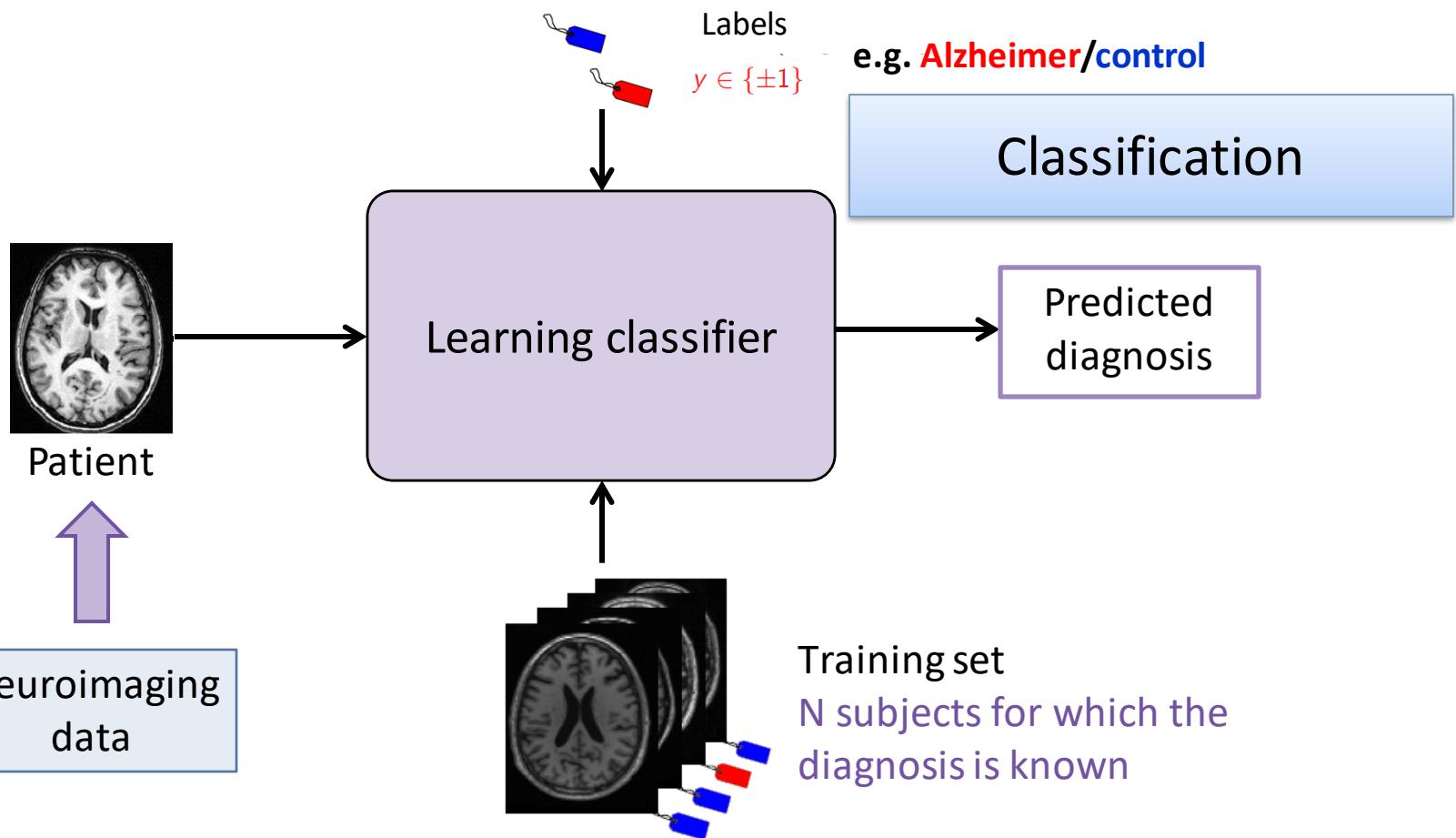
Maria Vakalopoulou, PhD
Assistant Professor at
CentraleSupélec
Mathematics and Informatics (MICS)
Office: Bouygues Building Sb.132



Master 2 - MVA

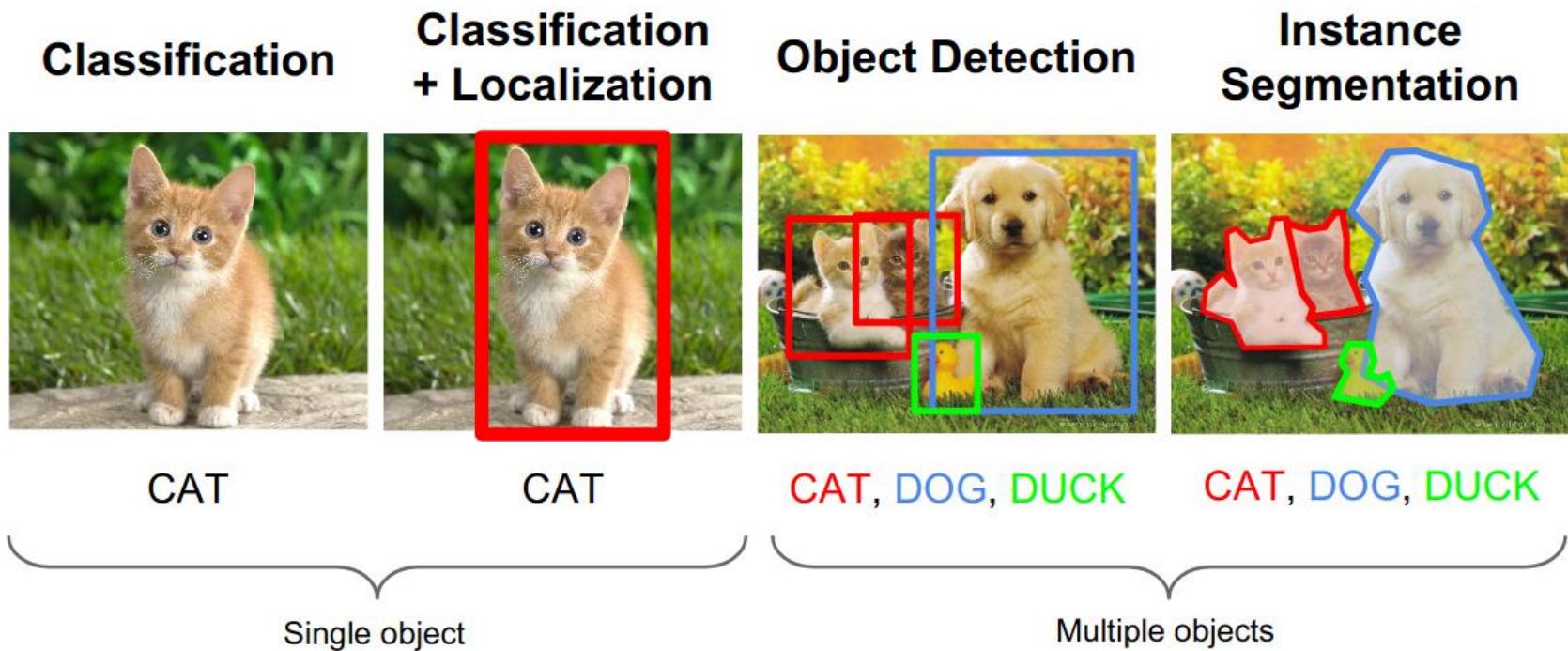
Previous Lectures

Classification



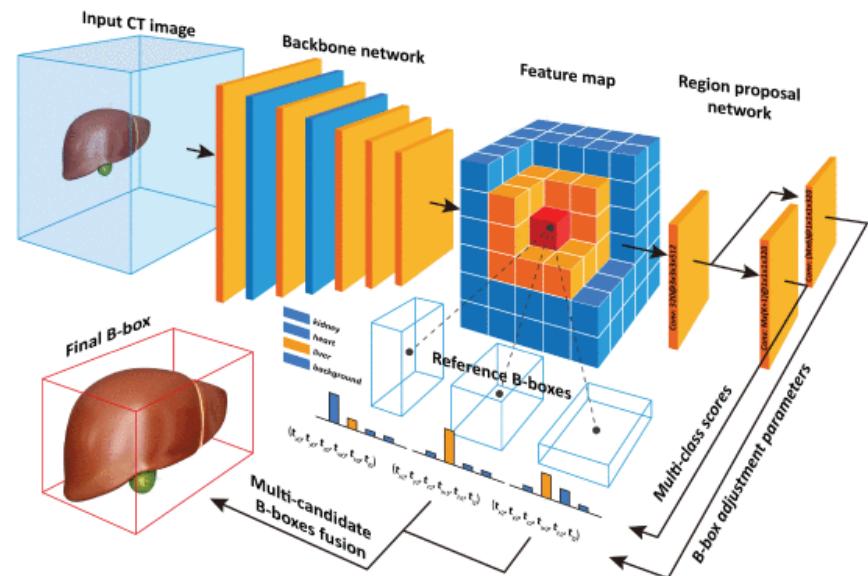
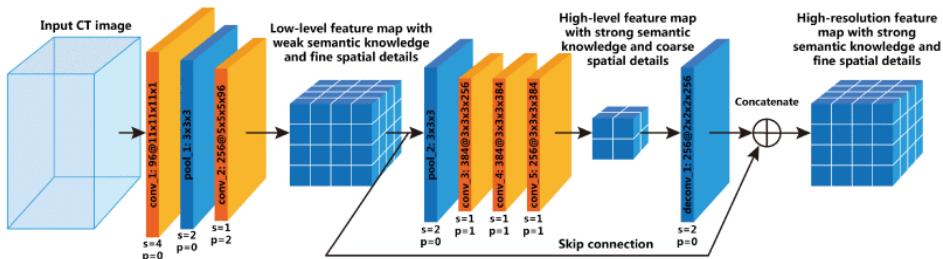
Detection/ Segmentation/ Instance Segm.

- Different problems for vision/ similar on medical



Multiple Organ Localization/ Detection in CT⁵

- 3D version of Faster R-CNN.
- Use of 3D convolutions



Journals & Magazines > IEEE Transactions on Medical ... > Volume: 38 Issue: 8 [?](#)

Efficient Multiple Organ Localization in CT Image Using 3D Region Proposal Network

Publisher: IEEE



5 Author(s) Xuanang Xu [ID](#); Fugen Zhou [ID](#); Bo Liu [ID](#); Dongshan Fu ; Xiangzhi Bai [ID](#) [View All Authors](#)

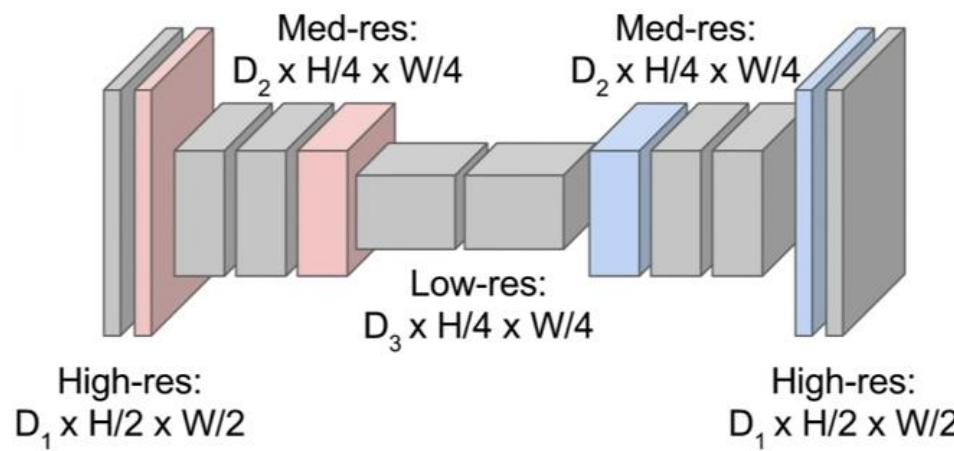
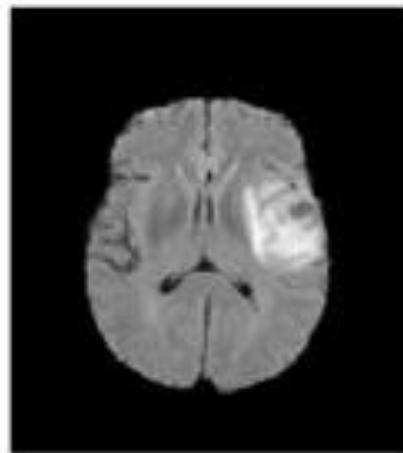
1
Paper Citation

1554
Full
Text Views



CNN-based solution for Image Denoising

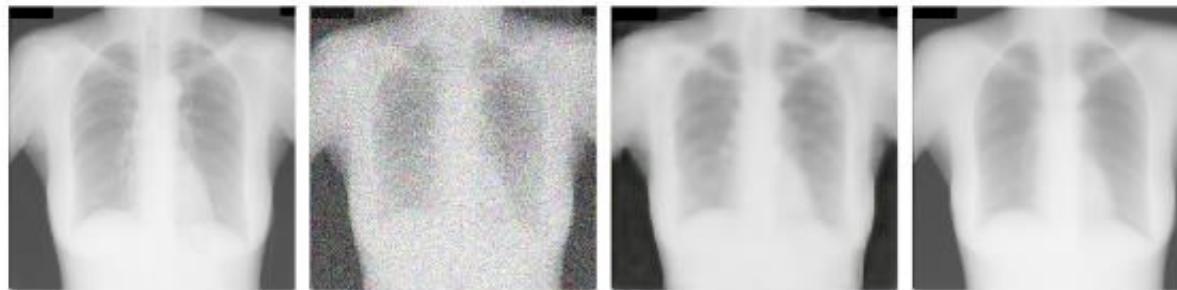
- Idea #1: Try to address the problem of denoising/reconstruction as a semantic segmentation problem
 - Make use of the FCN architectures and autoencoders



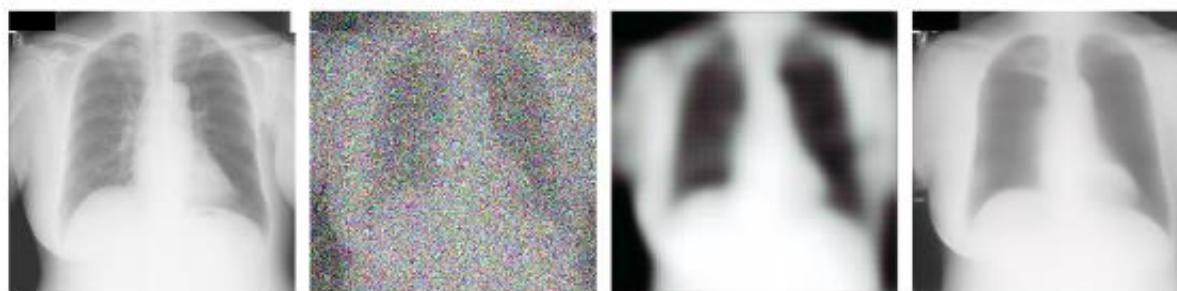
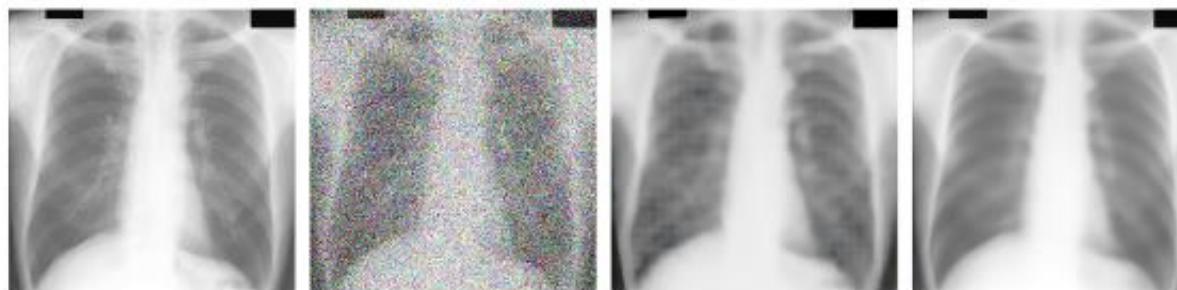
- Use Downsampling and upsampling inside the network!

What about recent powerful models?

- [Laoussy et al. 2023] Use of the Denoising Diffusion Probabilistic models for Xrays.



Results for different levels of noise
 $\sigma = 0.25, 0.5$, and 1.0 .



Input Image

Noised Image

UNet model

Diffusion

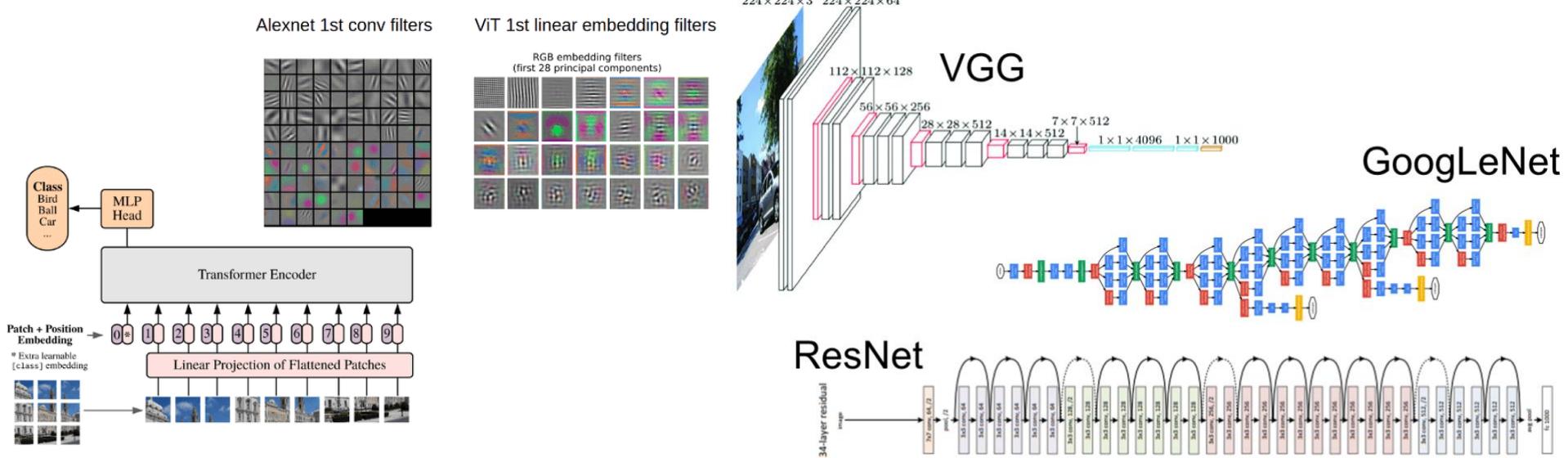
Part 6 – Foundation Models

Foundation Models

- Introduction
- Discussion of some Foundation Models
- Adaptation techniques for Foundation Models
 - Prompt-based Learning
 - Adapters
- Foundation Models in Medical Imaging

Deep Learning Models for Representation

- Deep Learning models are usually used for representing data (imaging, text, language) on different training schemes (training, fine-tuning, freezed).
- Different architectures the last 11 years based on convolutional and recently transformer architectures



Towards Large Deep Learning Models

- AI community is working on building huge and powerful models for different tasks with currently impressive results on:
 - Image generation from text
 - DALL-E: 12 billion parameters trained in 250 million text-image pairs



Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

Towards Large Deep Learning Models

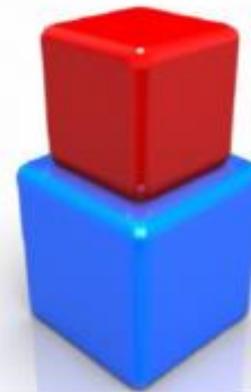
- AI community is working on building huge and powerful models for different tasks with currently impressive results on:
 - Image generation from text
 - GLIDE: 35-billion parameter text-conditional diffusion, and another 1.5 billion parameter text-conditional upsampling diffusion model.



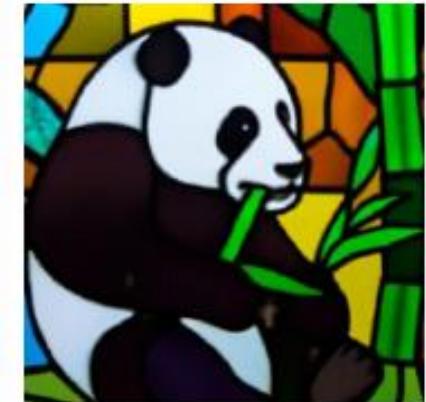
“a boat in the canals of venice”



“a painting of a fox in the style of starry night”



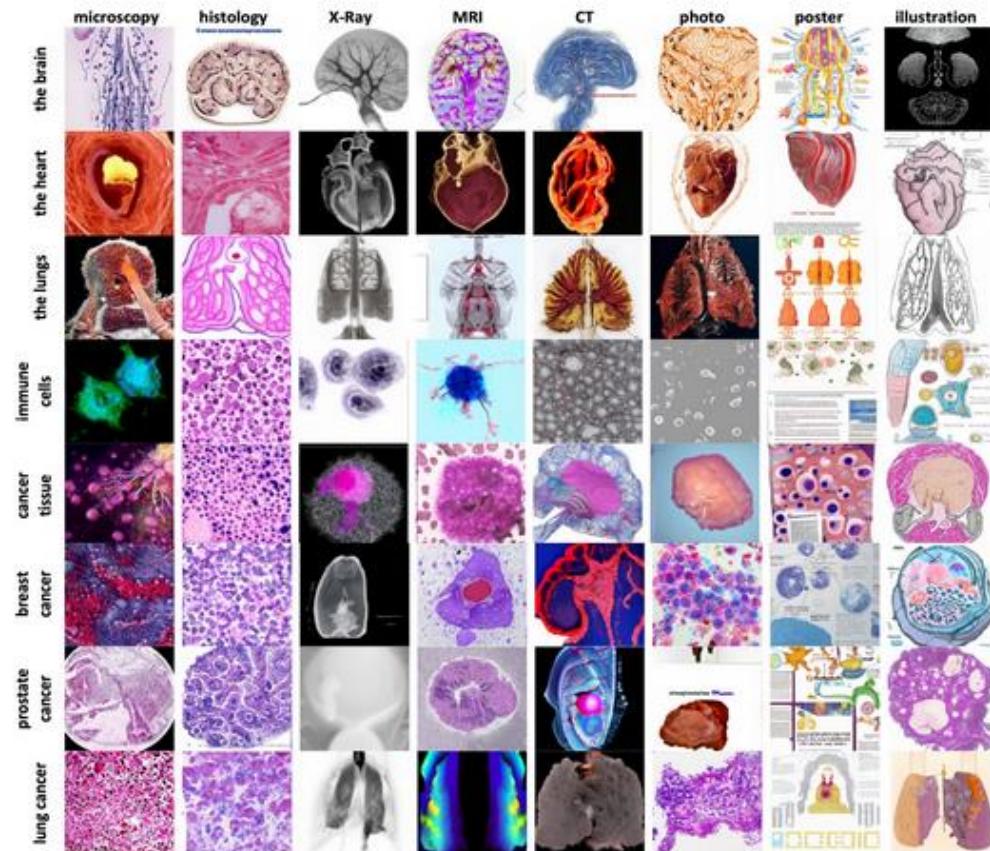
“a red cube on top of a blue cube”



“a stained glass window of a panda eating bamboo”

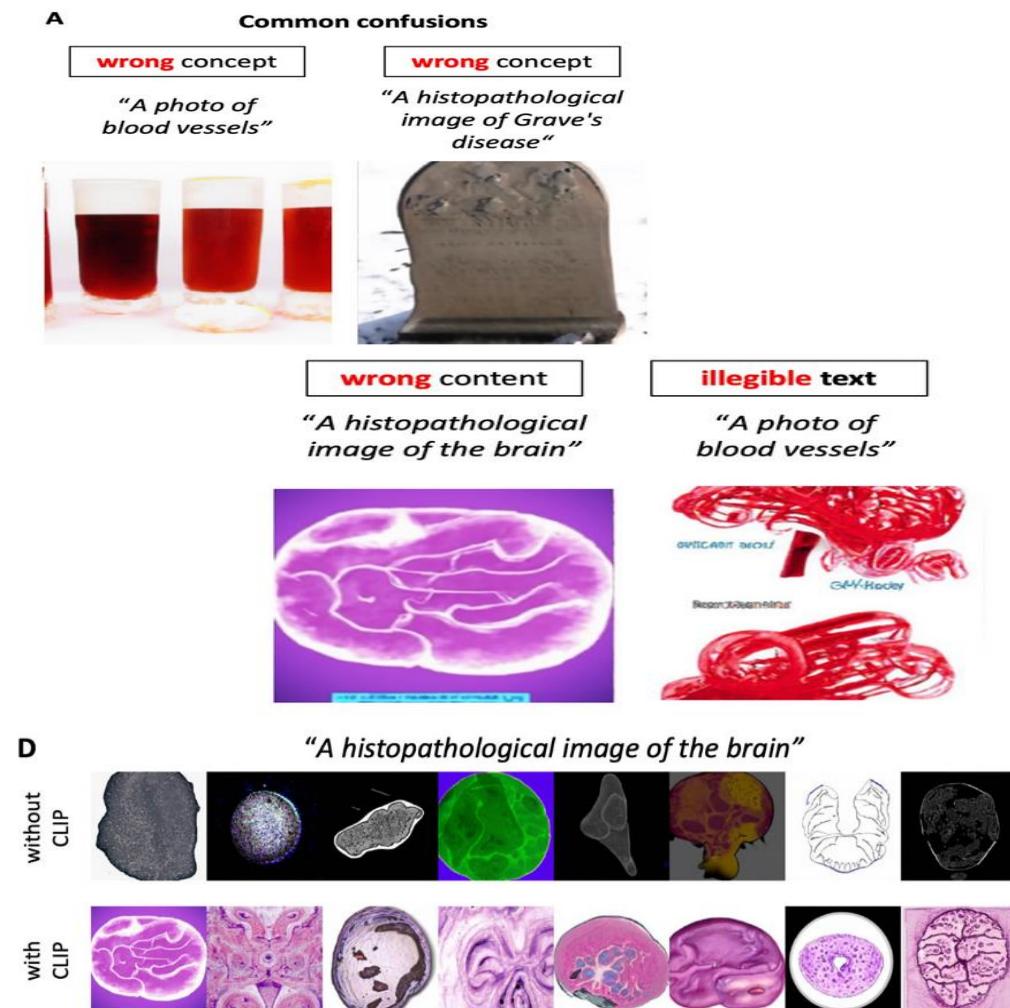
Challenges in the Medical Domain

- Diffusion models are currently very popular generative models trained on reconstruction/ denoising.
- [Kather et al. 2022] GLIDE model: a diffusion model for the problem of text-conditional synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance.
[December 2021]
- Dataset: 250 million text-image from the internet incorporating Conceptual Captions, the text-image pairs from Wikipedia and a subset of YFCC100M.



Challenges in the Medical Domain

- These models do not generalize well on medical data and particular clinical questions.
- GLIDE performance evaluation:
 - High score for: histopathology, scientific posters and scientific illustration
 - Poor score for: X-rays, MRI and CT
- Similar models could be used for exploring biological mechanisms:
 - A histology image of a patient who benefits from immunotherapy
 - An MRI image of a patient who should be treated with a statin



Towards Large Deep Learning Models

- AI community is working on building huge and powerful models for different tasks with currently impressive results on:
 - Large Language Models inferring relationships between words within the text
 - ChatGPT: 176-billion parameters trained in 570GB data from text, books, websites and more

S A 22-year-old male is presenting with delusional thinking, responding to internal stimuli, and unprovoked aggression. What is the diagnosis?

Q Based on the given symptoms, the most likely diagnosis for the 22-year-old male would be schizophrenia. Schizophrenia is a severe mental disorder characterized by a combination of symptoms that can include delusions, hallucinations, disorganized thinking, abnormal behaviors, and difficulty functioning. The presence of delusional thinking, responding to internal stimuli, and unprovoked aggression is commonly seen in schizophrenia. However, it is important to note that a proper diagnosis requires a comprehensive evaluation by a qualified mental health professional.

Galido P, Butala S, Chakerian M, et al. (April 26, 2023) A Case Study Demonstrating Applications of ChatGPT in the Clinical Management of Treatment-Resistant Schizophrenia . Cureus 15(4): e38166. doi:10.7759/cureus.38166

DL I have a tingly nose, am stuffy, and overall look ugly. What's wrong with me?

Q Based on the symptoms you mentioned, it is possible that you have a cold or flu-like illness. These symptoms are common with respiratory infections caused by viruses, and can also include a cough, fatigue, and body aches. If you're feeling unwell, it's important to rest and drink plenty of fluids. Over-the-counter cold and flu remedies may also help relieve your symptoms.

If your symptoms persist or worsen, or if you develop a fever, it is a good idea to see a doctor for a proper diagnosis and treatment. The doctor may also advise you on ways to help manage your symptoms and speed up your recovery.

<https://www.verywellhealth.com/chatgpt-in-healthcare-7107800>

Towards Large Deep Learning Models

- AI community is working on building huge and powerful models for different tasks with currently impressive results on:
 - Imaging tasks and image segmentation
 - SAM: Trained on 1B masks and from 11M licensed and privacy-preserving image

>500 masks



What is a foundation model?

- Foundation models use self-supervised learning to create labels from input data. A foundation model is a machine learning model that is trained on broad data such that it can be applied across a wide range of use cases.
- Foundation models are general purpose technologies that can support a diverse range of use cases. Building foundation models is often highly resource-intensive, with the most expensive models costing hundreds of millions of dollars to pay for the underlying data and compute required.



What is a foundation model?

- Foundation models began to materialize as the latest wave of deep learning in the late 2010s with models mainly on NLP.
- Their rise is due to:
 - The computational advances in specialized hardware and parallelism (e.g., large clusters of GPUs)
 - New developments in neural network architecture (e.g., the Transformer)
 - Increased use of training data with minimal supervision all contributed to the rise of foundation models.
- Examples of foundation models: GPT, BERT, GPT-2, T5, GPT-3, CLIP, DALL-E, Stable Diffusion, GPT-4, LLaMA, LLaMA 2, Mistral
- *In particular, 2022 was particularly influential in the history of foundation models. The releases of Stable Diffusion and ChatGPT (initially powered by the GPT-3.5 model) led to foundation models and generative AI entering widespread public discourse.*

Multimodal and Image foundation models

19

- Foundation models plays a very very important role
 - CLIP (OpenAI)
 - DINOv2 (Meta AI)
 - Segment Anything (Meta AI)

CLIP

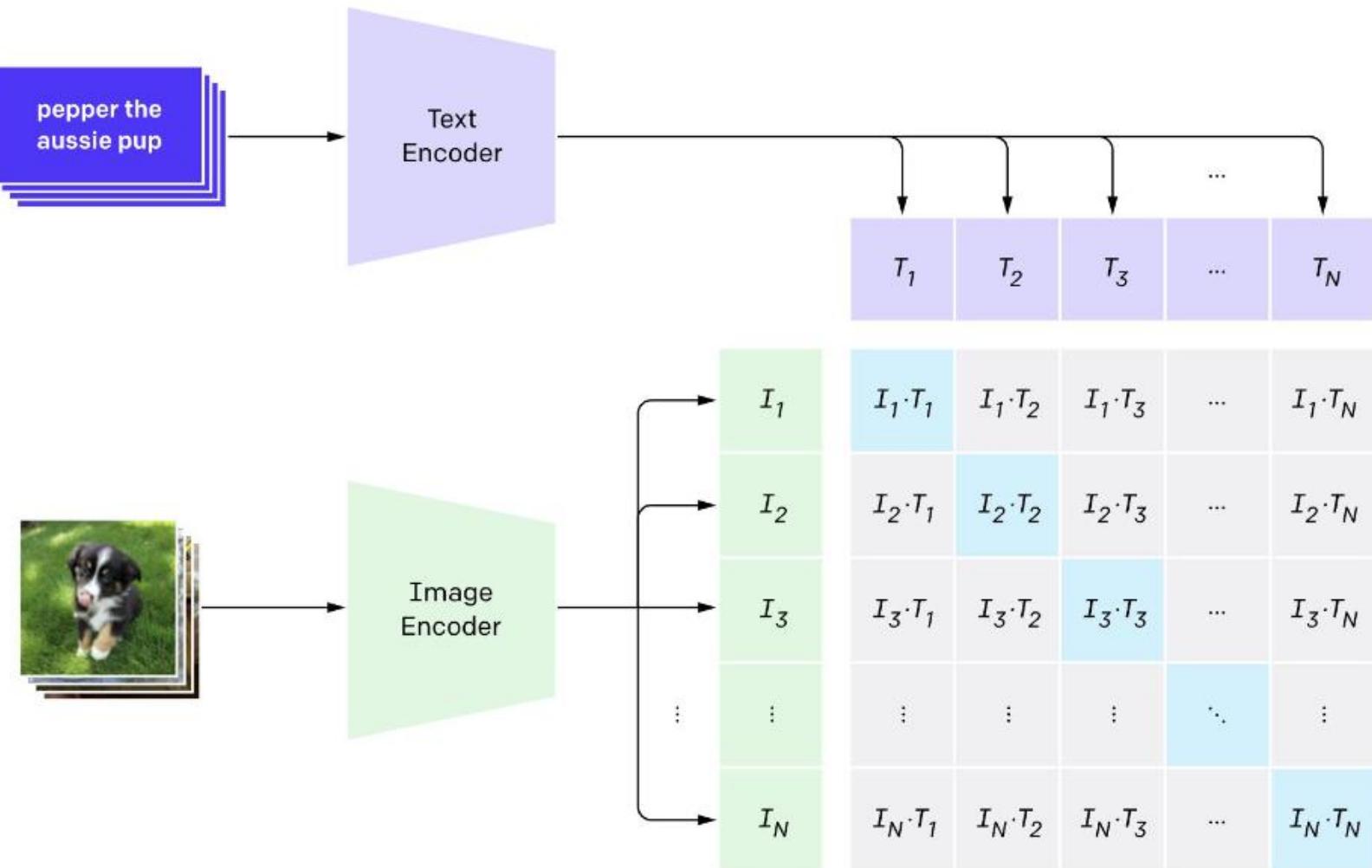
- Natural Language for Visual Reasoning (NLVR)
- Classify an image to a specific label based on its context
 - Label is a phrase or a sentence describing the image
- CLIP is a zero-shot classifier, i.e., can be used to previously unseen labels
- Trained on a highly-diversified, huge dataset: 400 million!
Images with textual descriptions
- Images, Text: transformer

CLIP

- The training's objective: "connect" image representations with text representations
 - Discover which text vector is more "appropriate" for a given image vector with contrastive learning
 - Instead of bringing together views of the same image, we are pulling together the positive image and text "views", while pulling apart texts that do not correspond to the correct image (negatives)
 - Fully supervised, i.e. labeled pairs are required
- Training: assign high similarity for fitting image-text pairs and low similarity for unfitting ones and downstream tasks

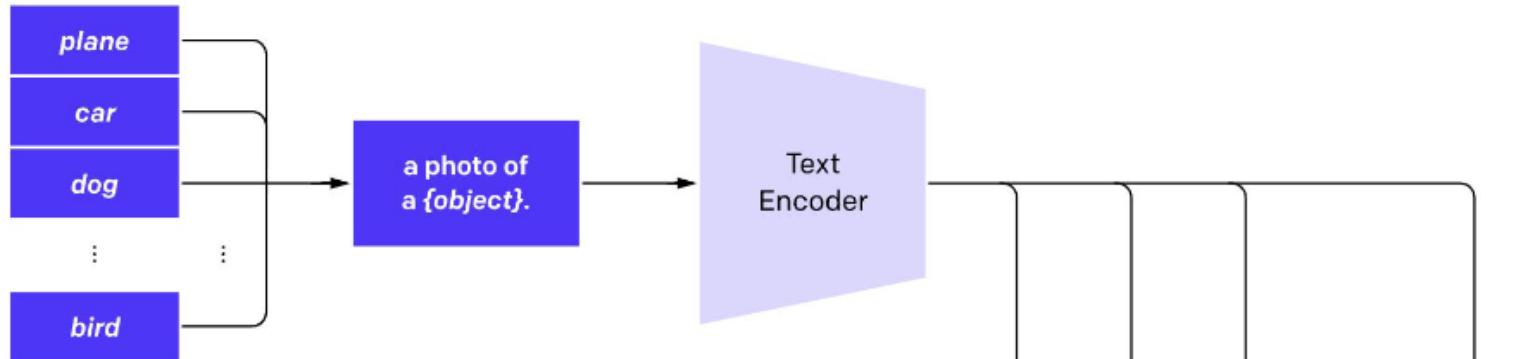
CLIP

1. Contrastive pre-training

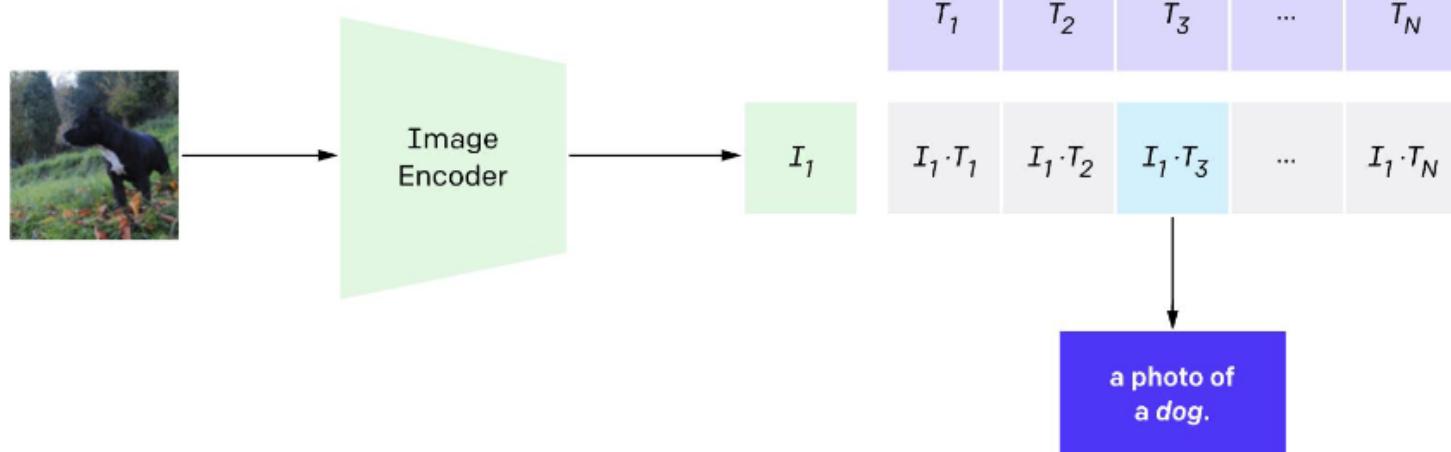


CLIP

2. Create dataset classifier from label text



3. Use for zero-shot prediction



CLIP

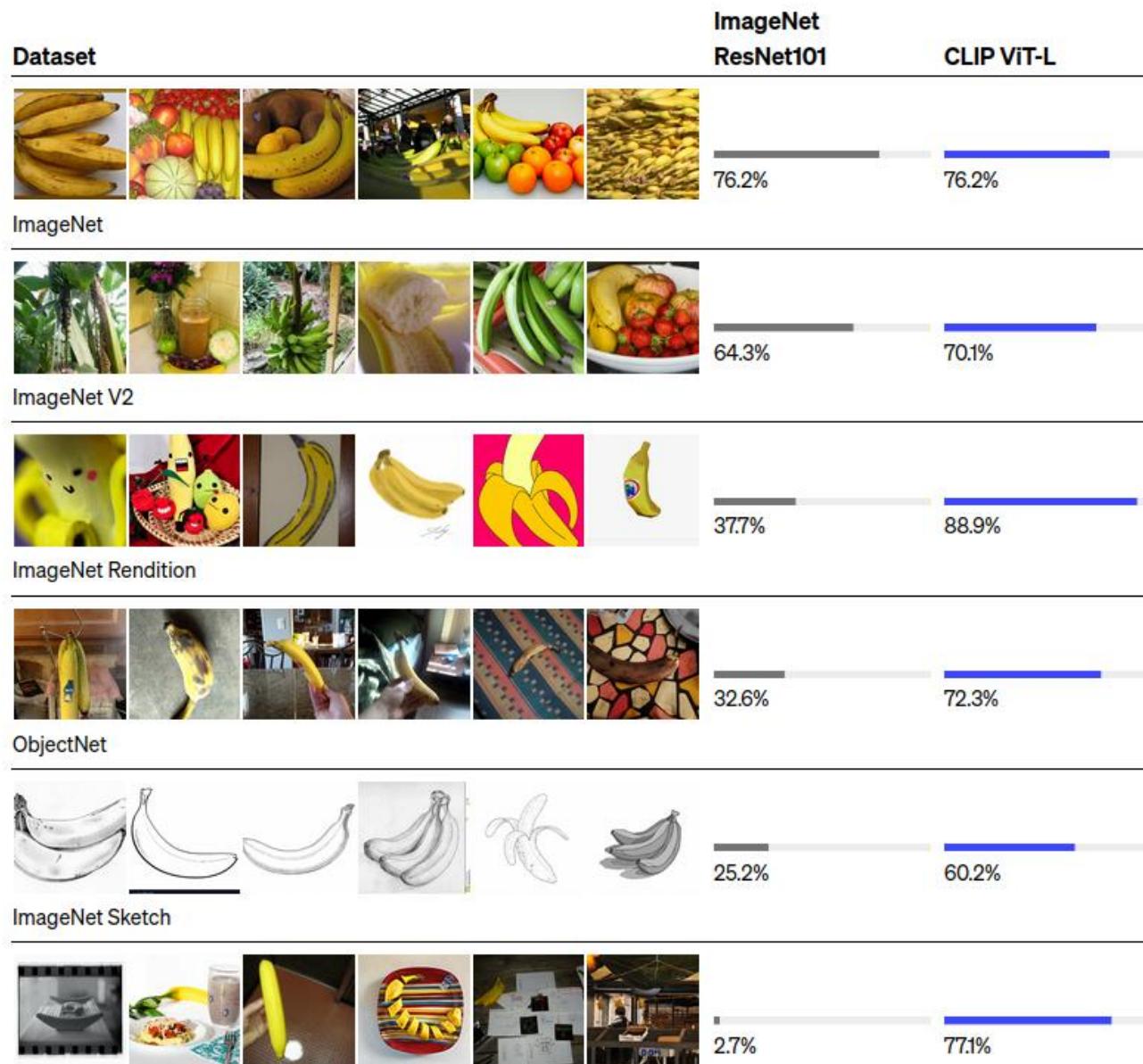
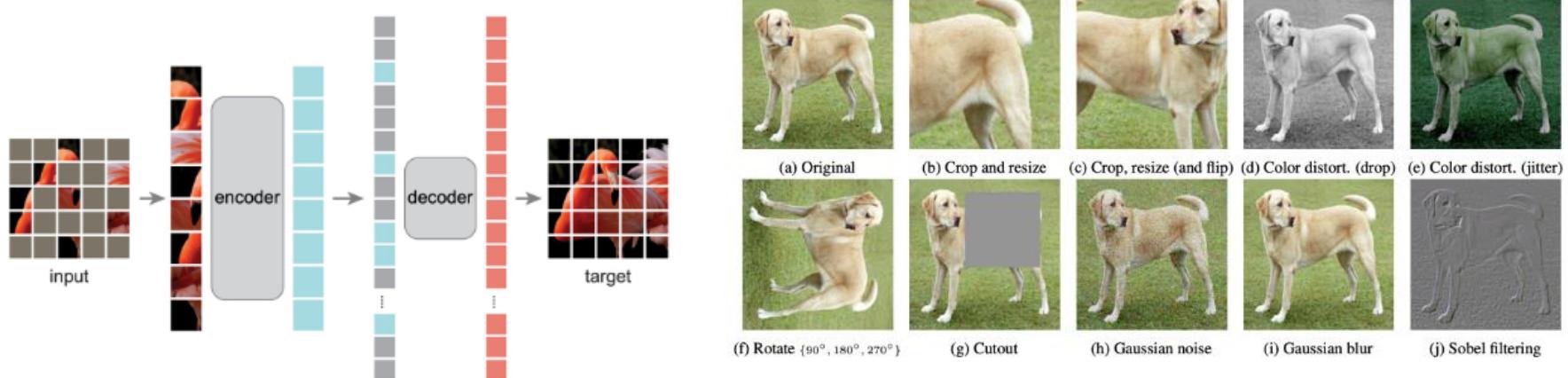


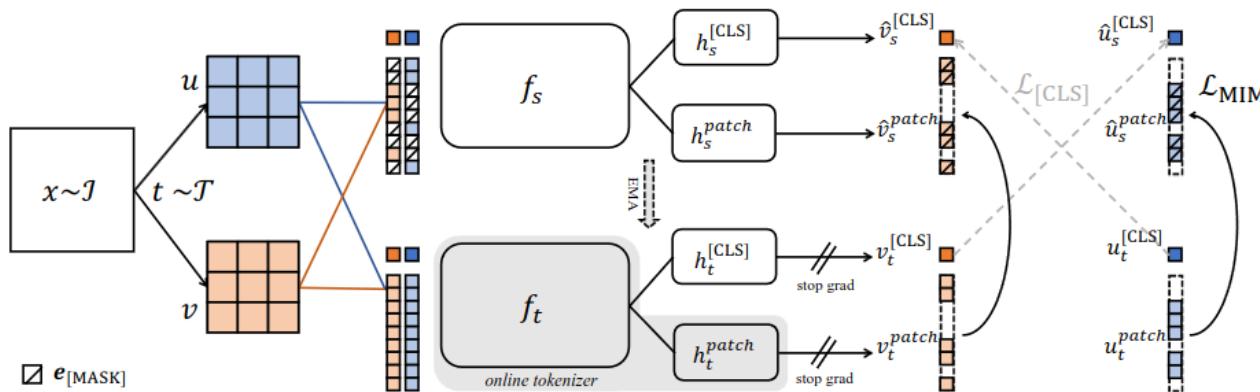
Image-based SSL pretraining

- An alternative to text-guided pretraining is self-supervised learning



He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR 2022

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICML 2020



Zhou, Jinghao, et al.
"iBOT: Image BERT Pre-training with Online Tokenizer." ICLR 2022

DINOv2

- However, despite their potential to learn general-purpose features, most of the advances in self-supervised learning were made in the context of pretraining on a small curated dataset, ImageNet-1k.
 - Some efforts on scaling these approaches beyond ImageNet-1k have been attempted.
-
- Revisit existing discriminative self-supervised approaches that learn features at both the image and patch level, such as iBOT, and reconsider some of their design choices under the lens of a larger dataset.
 - DINOv2 is around $2\times$ faster and require $3\times$ less memory than similar discriminative self-supervised methods, allowing us to leverage longer training with larger batch sizes.

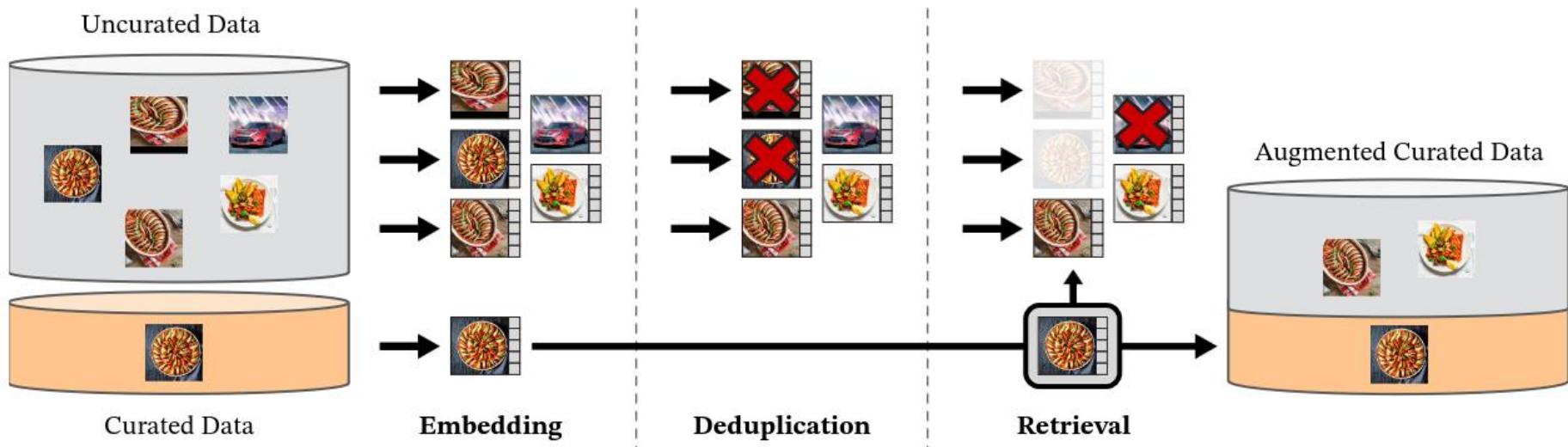
DINOv2 – Data Processing

- Assemble a curated LVD-142M dataset by retrieving, from a large pool of uncurated data, images that are close to those in several curated datasets.

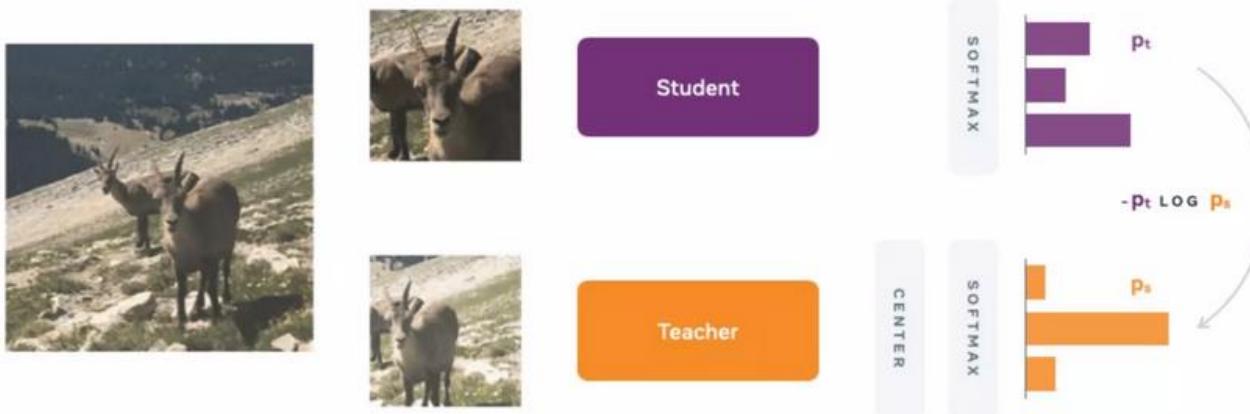
Task	Dataset / Split	Images	Retrieval	Retrieved	Final
classification	ImageNet-22k / –	14,197,086	as is	–	14,197,086
classification	ImageNet-22k / –	14,197,086	sample	56,788,344	56,788,344
classification	ImageNet-1k / train	1,281,167	sample	40,997,344	40,997,344
fine-grained classif.	Caltech 101 / train	3,030	cluster	2,630,000	1,000,000
fine-grained classif.	CUB-200-2011 / train	5,994	cluster	1,300,000	1,000,000
fine-grained classif.	DTD / train1	1,880	cluster	1,580,000	1,000,000
fine-grained classif.	FGVC-Aircraft / train	3,334	cluster	1,170,000	1,000,000
fine-grained classif.	Flowers-102 / train	1,020	cluster	1,060,000	1,000,000
fine-grained classif.	Food-101 / train	75,750	cluster	21,670,000	1,000,000
fine-grained classif.	Oxford-IIIT Pet / trainval	3,680	cluster	2,750,000	1,000,000
fine-grained classif.	Stanford Cars / train	8,144	cluster	7,220,000	1,000,000
fine-grained classif.	SUN397 / train1	19,850	cluster	18,950,000	1,000,000
fine-grained classif.	Pascal VOC 2007 / train	2,501	cluster	1,010,000	1,000,000
segmentation	ADE20K / train	20,210	cluster	20,720,000	1,000,000
segmentation	Cityscapes / train	2,975	cluster	1,390,000	1,000,000
segmentation	Pascal VOC 2012 (seg.) / trainaug	1,464	cluster	10,140,000	1,000,000
depth estimation	Mapillary SLS / train	1,434,262	as is	–	1,434,262
depth estimation	KITTI / train (Eigen)	23,158	cluster	3,700,000	1,000,000
depth estimation	NYU Depth V2 / train	24,231	cluster	10,850,000	1,000,000
depth estimation	SUN RGB-D / train	4,829	cluster	4,870,000	1,000,000
retrieval	Google Landmarks v2 / train (clean)	1,580,470	as is	–	1,580,470
retrieval	Google Landmarks v2 / train (clean)	1,580,470	sample	6,321,880	6,321,880
retrieval	AmsterTime / new	1,231	cluster	960,000	960,000
retrieval	AmsterTime / old	1,231	cluster	830,000	830,000
retrieval	Met / train	397,121	cluster	62,860,000	1,000,000
retrieval	Revisiting Oxford / base	4,993	cluster	3,680,000	1,000,000
retrieval	Revisiting Paris / base	6,322	cluster	3,660,000	1,000,000
142,109,386					

DINOv2 – Data Processing

- Assemble our curated LVD-142M dataset by retrieving, from a large pool of uncurated data, images that are close to those in several curated datasets.
- Deduplication, increasing diversity among images.
- SSL image retrieval, image embeddings using a self-supervised ViT-H/16 pretrained on ImageNet-22k and use cosine-similarity as a distance measure between images. K-means clustering with N=4 for retrieval.

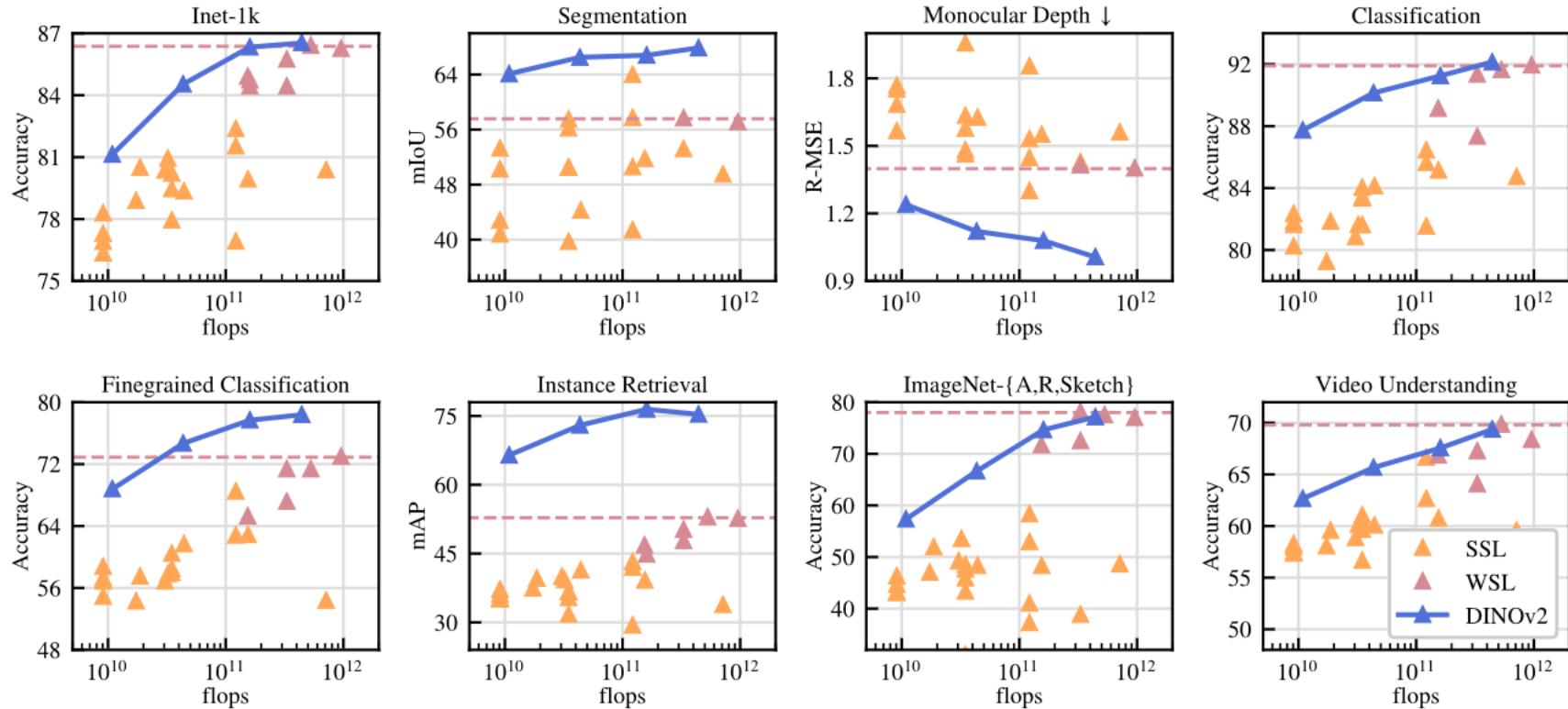


DINOv2 – Model Scaling and Stability



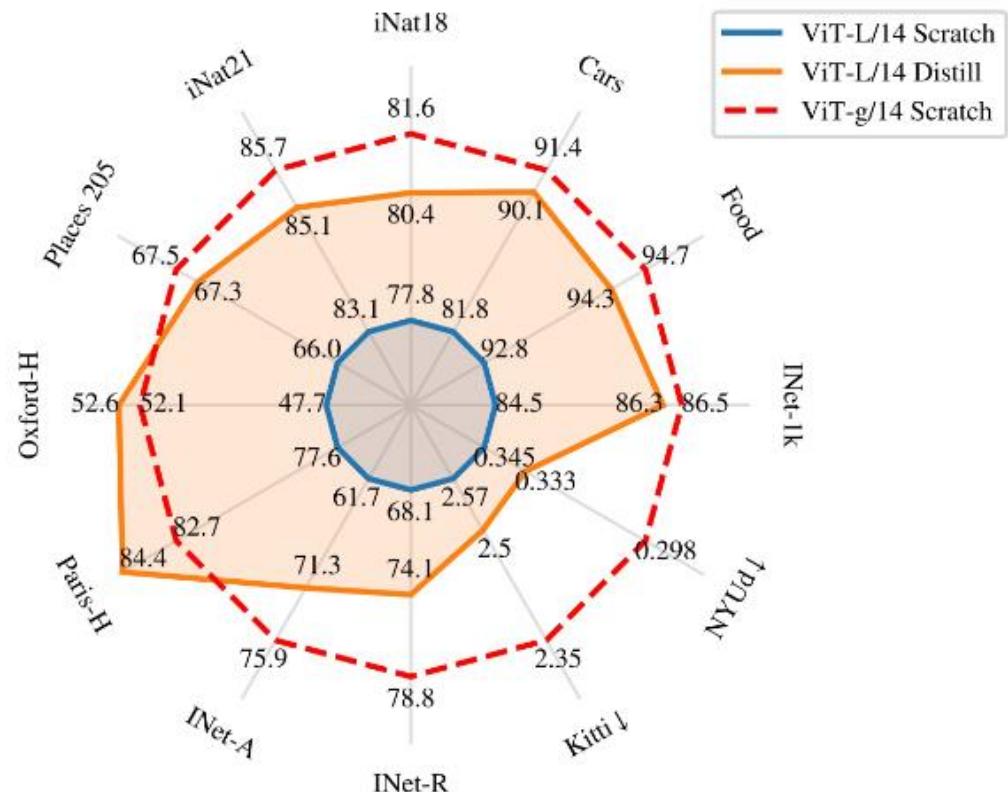
	INet-1k k-NN	INet-1k linear
iBOT	72.9	82.3
+ (our reproduction)	74.5 \uparrow 1.6	83.2 \uparrow 0.9
+ LayerScale, Stochastic Depth	75.4 \uparrow 0.9	82.0 \downarrow 1.2
+ 128k prototypes	76.6 \uparrow 1.2	81.9 \downarrow 0.1
+ KoLeo	78.9 \uparrow 2.3	82.5 \uparrow 0.6
+ SwiGLU FFN	78.7 \downarrow 0.2	83.1 \uparrow 0.6
+ Patch size 14	78.9 \uparrow 0.2	83.5 \uparrow 0.4
+ Teacher momentum 0.994	79.4 \uparrow 0.5	83.6 \uparrow 0.1
+ Tweak warmup schedules	80.5 \uparrow 1.1	83.8 \uparrow 0.2
+ Batch size 3k	81.7 \uparrow 1.2	84.7 \uparrow 0.9
+ Sinkhorn-Knopp	81.7 =	84.7 =
+ Untying heads = DINOv2	82.0 \uparrow 0.3	84.5 \downarrow 0.2

DINOv2 – Model Scaling and Stability



DINOv2 – Distillation performance

- Instead of training a family of model, distill information from larger models.
- Trained ViT-{S, B, L} from the ViT-G
- Interestingly, the ViT-L distilled works better than from scratch!

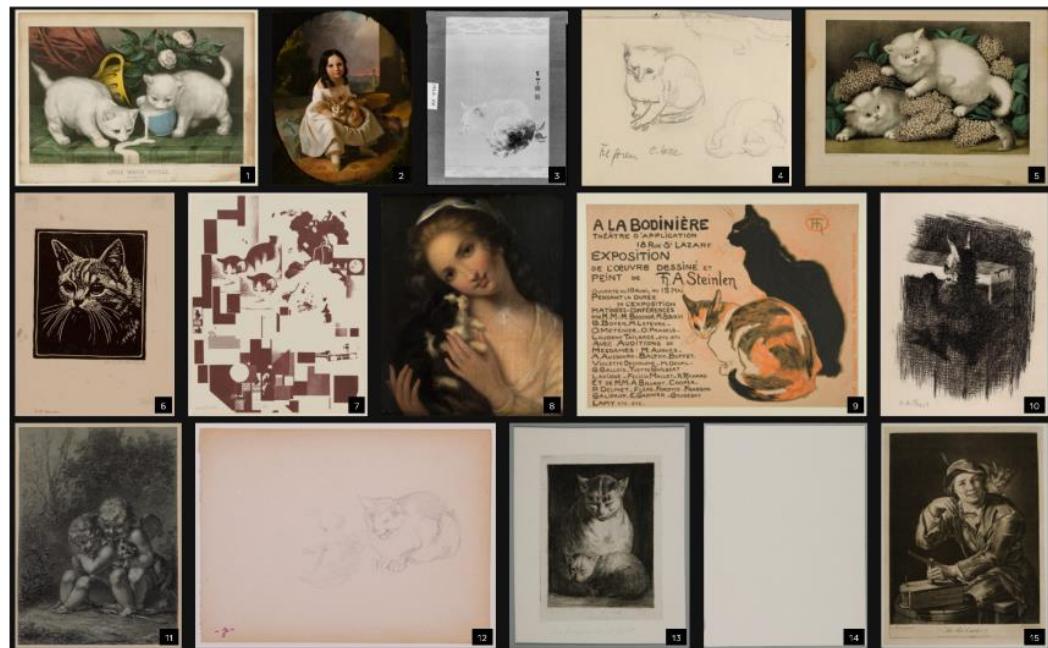


DINOv2 – Retrieval

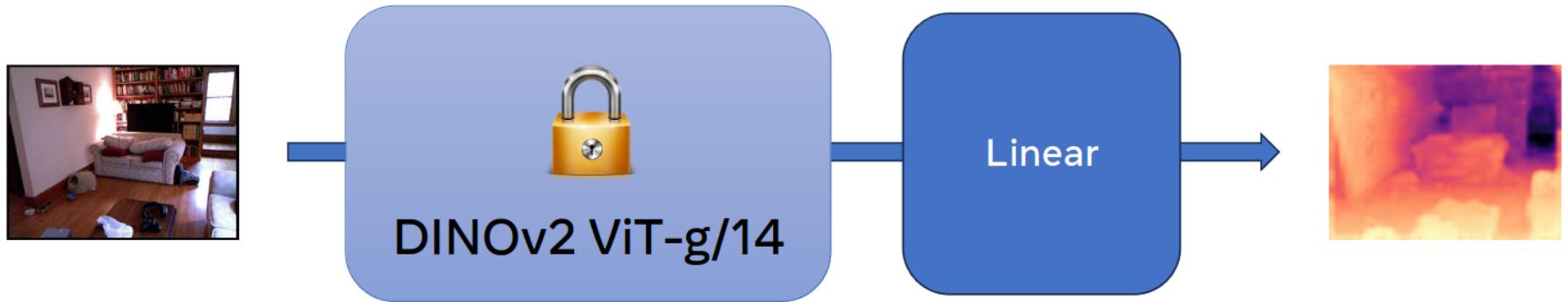
Query



Feature	Arch	Oxford		Paris		Met		AmsterTime	
		M	H	M	H	GAP	GAP-	ACC	mAP
OpenCLIP	ViT-G/14	50.7	19.7	79.2	60.2	6.5	23.9	34.4	24.6
MAE	ViT-H/14	11.7	2.2	19.9	4.7	7.5	23.5	30.5	4.2
DINO	ViT-B/8	40.1	13.7	65.3	35.3	17.1	37.7	43.9	24.6
iBOT	ViT-L/16	39.0	12.7	70.7	47.0	25.1	54.8	58.2	26.7
DINOv2	ViT-S/14	68.8	43.2	84.6	68.5	29.4	54.3	57.7	43.5
	ViT-B/14	72.9	49.5	90.3	78.6	36.7	63.5	66.1	45.6
	ViT-L/14	75.1	54.0	92.7	83.5	40.0	68.9	71.6	50.0
	ViT-g/14	73.6	52.3	92.1	82.6	36.8	73.6	76.5	46.7



DINOv2 – Dense Prediction Tasks

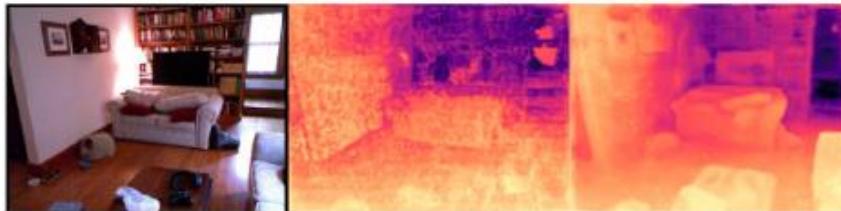


DINOv2 – Dense Prediction Tasks

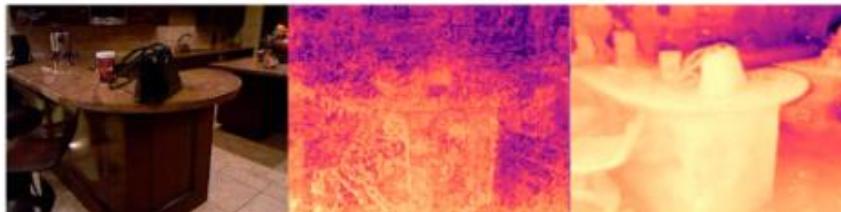
ADE20K



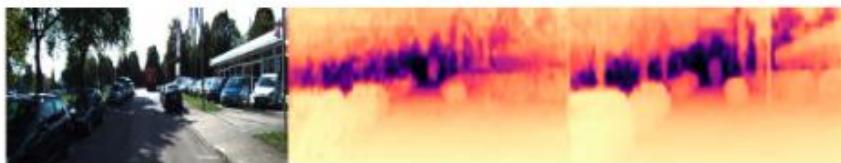
NYUd



SUN-RGBd



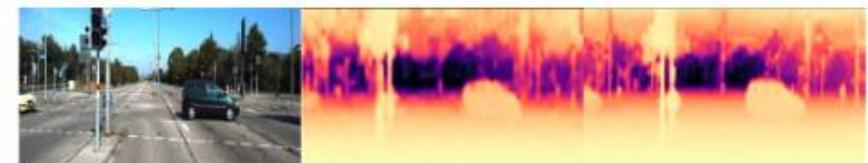
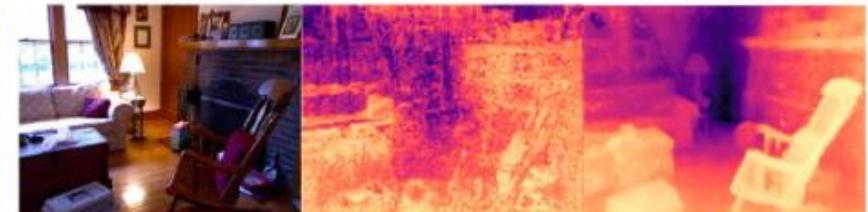
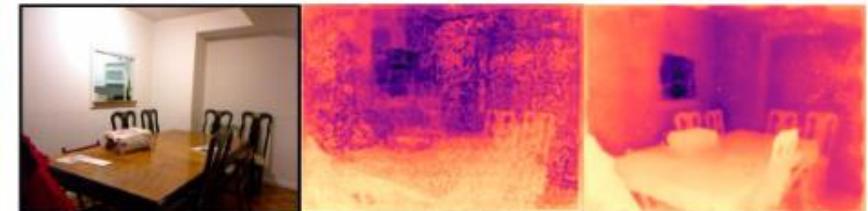
KITTI



Input

OpenCLIP-G

DINOv2-g



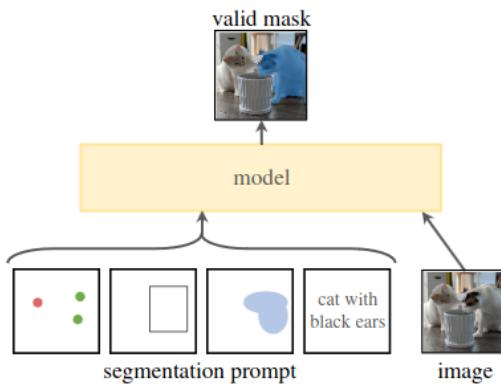
Input

OpenCLIP-G

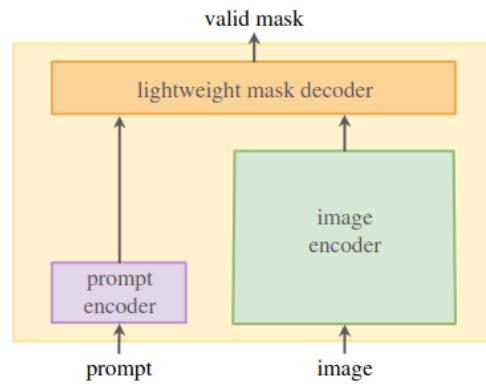
DINOv2-g

Segment Anything

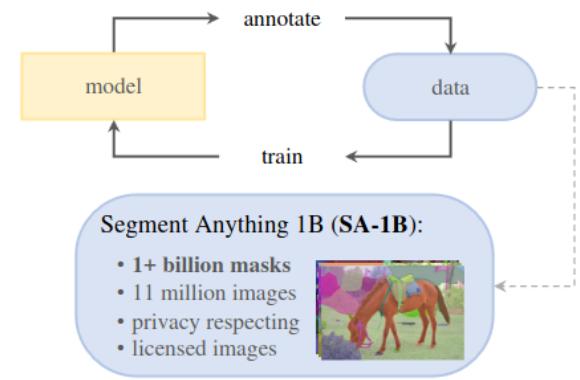
- A foundation model for segmentation by introducing three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering.



(a) **Task:** promptable segmentation

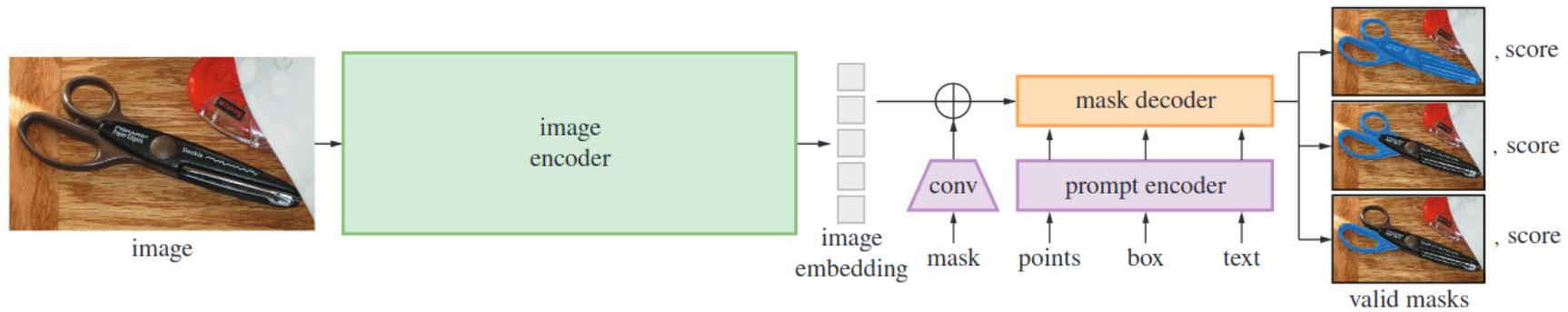


(b) **Model:** Segment Anything Model (SAM)



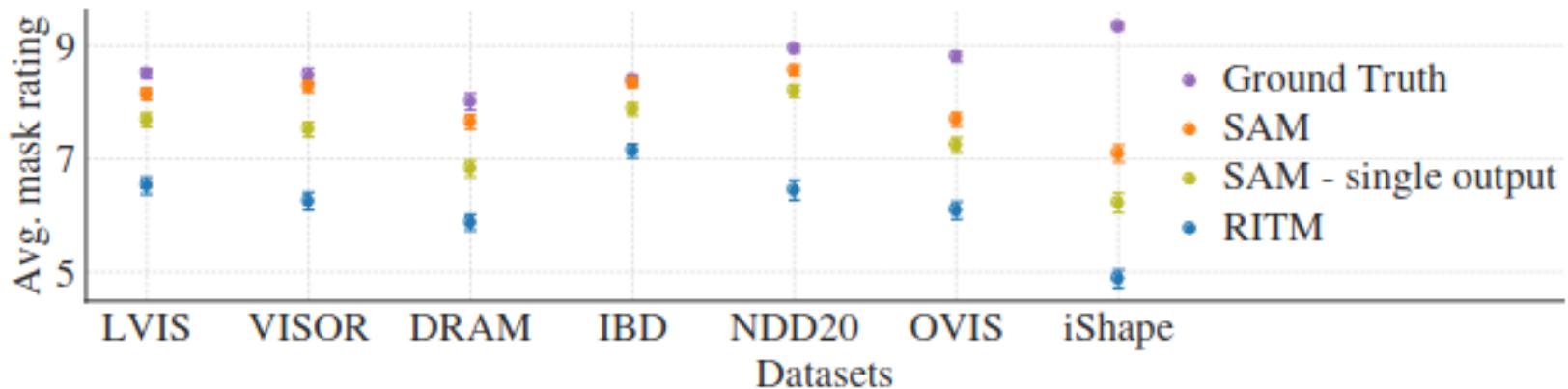
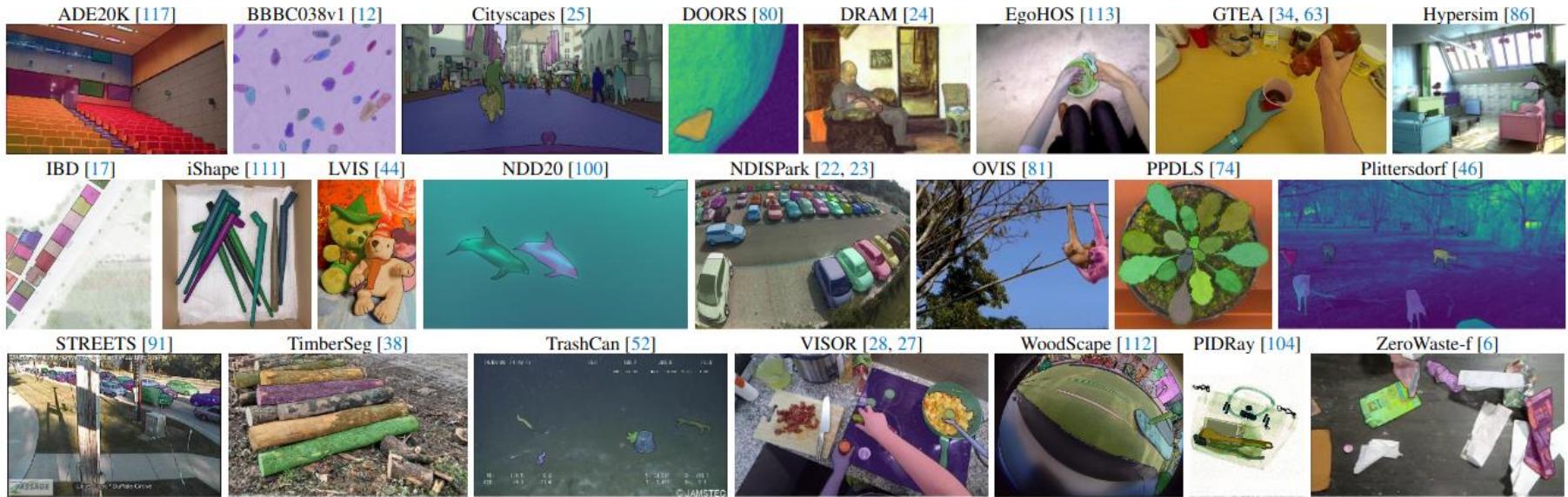
(c) **Data:** data engine (top) & dataset (bottom)

Segment Anything



- A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks.
 - Image encoder: MAE pre-trained ViT minimally adapted. (SA-1B dataset)
 - Prompt encoder: Sparse (points, boxes, text) and dense (masks). Points and boxes by positional encodings summed with learned embeddings and free-form text with an off-the-shelf text encoder from CLIP. Dense are used with convolutions.
 - Mask decoder: Maps to an output token to a mask. Prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update all embeddings.

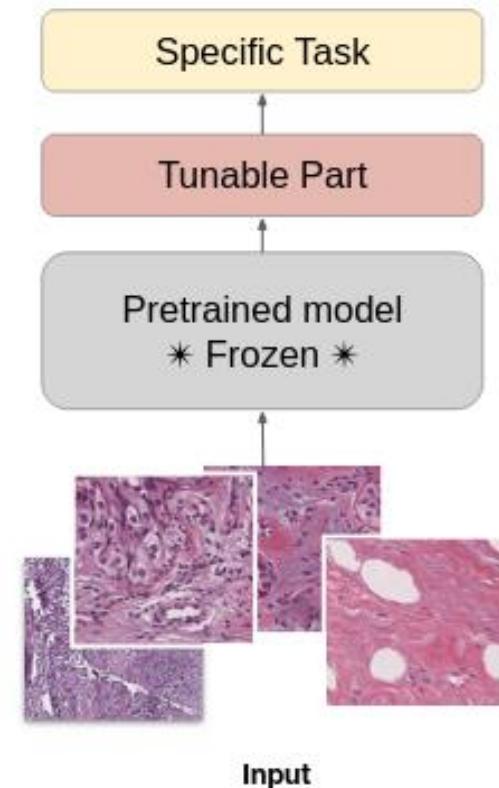
Segment Anything



How to adapt these models?

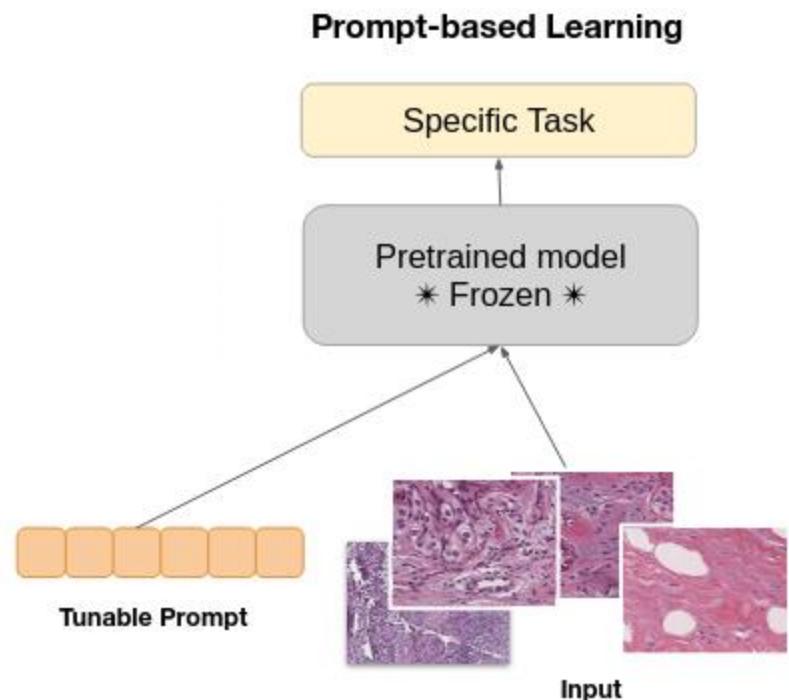
- Foundation models have a lot of capabilities however how we can easily adapt them?
 - Not easy to retrain
 - How we can get data and parameter efficient adaptation?
 - e.g. medical datasets?

Classical Fine-Tuning Learning



Prompt-based Learning?

- Prompt-based learning or prompting, started from the NLP community one year ago as an instruction by the user for the model to execute or complete
- Several advantages with respect to traditional methods:
 - Data-efficient [*also achieved with few-shot and zero-shot learning*]
 - Parameter-efficient [*also achieved with tuning*]

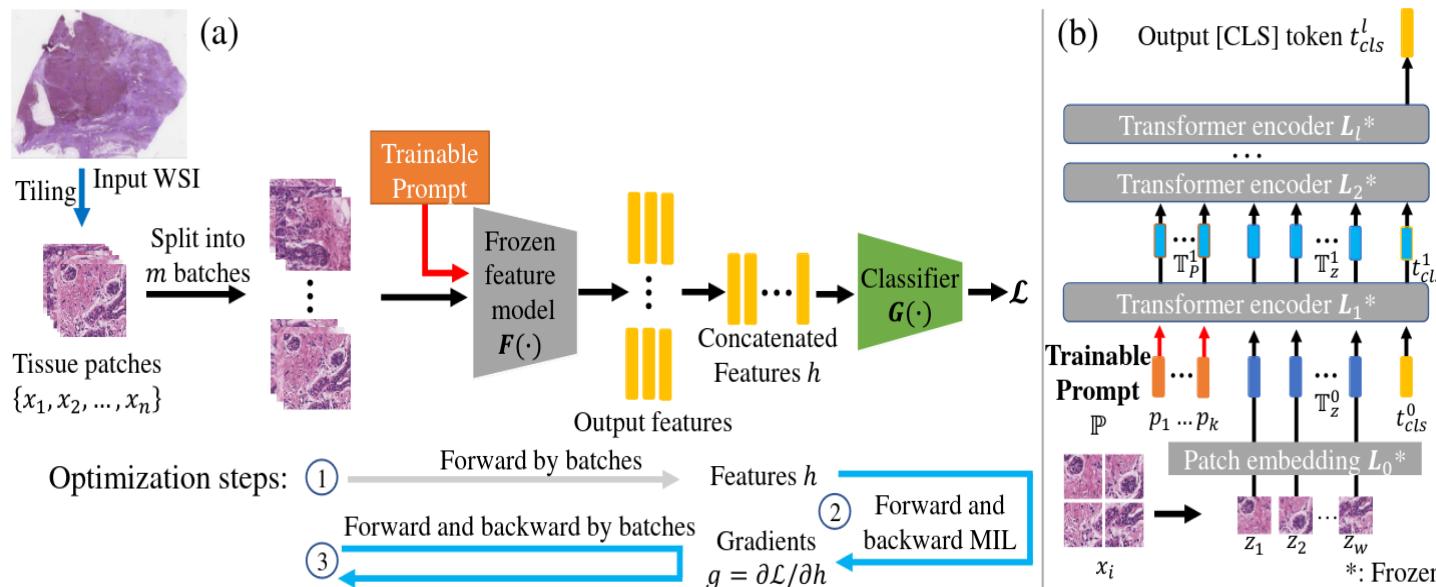


Example of prompt-based learning in MI

- Task-specific prompt for MIL Schemes
- Prompt-MIL: Calibration of the pre-trained on task-specific information using only a small fraction of tunable parameters rather than conventional fine-tuning.
 - Train only the prompt and downstream network without re-training the large backbone.
 - Better performances on limited labeled data – ideal for computational pathology.

Task-specific prompt for MIL Schemes

- Prompt-MIL framework
 - A frozen feature model $F(\cdot)$
 - A classifier to perform the task ($G(\cdot)$)
 - A trainable prompt $P(\cdot)$
- The number of the parameters is negligible with respect to the total amount of parameters of the architecture (192, less than 0.3% than the total parameters)



Task-specific prompt for MIL Schemes

- Prompt-MIL outperforms in all our experiments conventional MIL approaches as well as full fine-tuning using also much less parameters and GPU memory

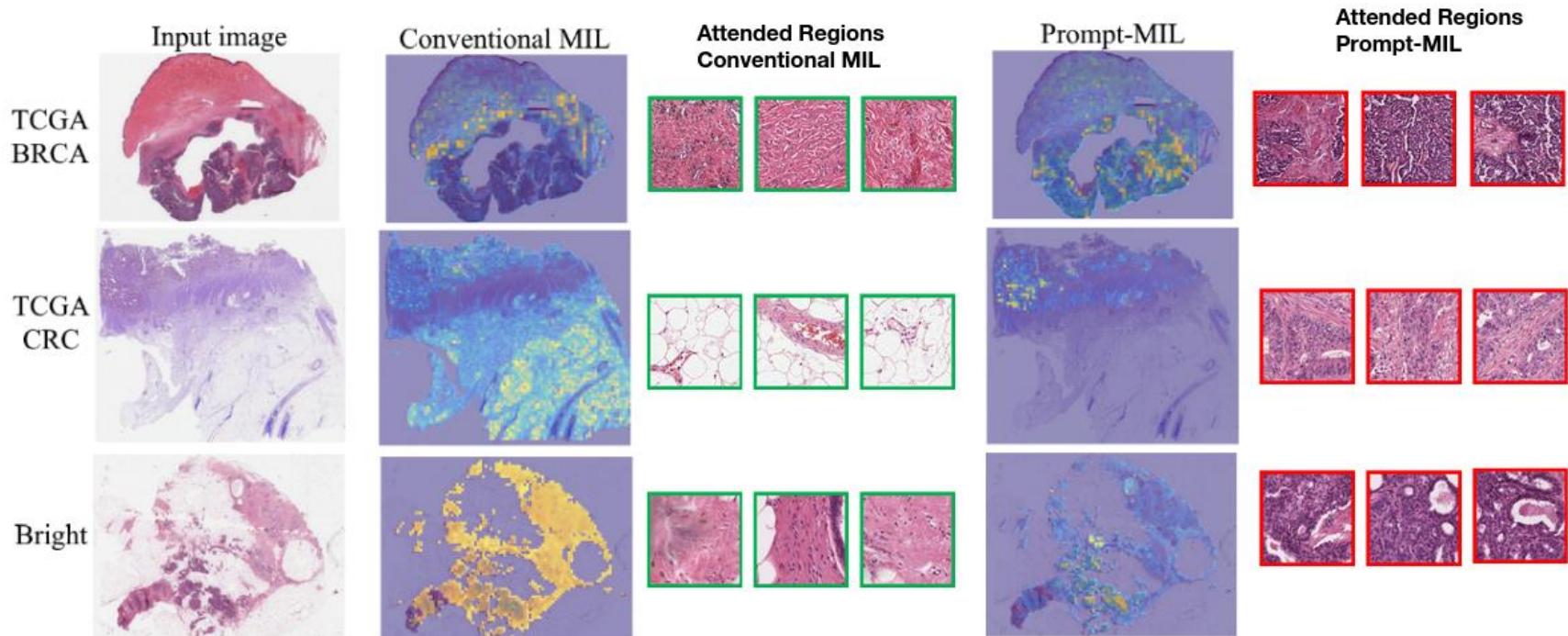
Dataset Metric	TCGA-BRCA		TCGA-CRC		BRIGHT		Num. of Parameters
	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	
Conventional MIL	92.10	96.65	73.02	69.24	62.08	80.96	70k
Full fine-tuning	88.14	93.78	74.53	56.63	56.13	75.87	5.6M
Prompt-MIL (ours)	93.47	96.89	75.47	75.45	64.58	81.31	70k+192

- Increase performance when using SSL-data specific representations trained on TCGA pan-cancer dataset

Dataset Metric	TCGA-BRCA		BRIGHT	
	Accuracy	AUROC	Accuracy	AUROC
ViT-small [27]	91.75	97.03	54.17	76.76
ViT-small w/ Prompt-MIL	92.78	97.53	57.50	78.29

Task-specific prompt for MIL Schemes

- Attention maps



Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., & Samaras, D. (2023). Prompt-MIL: Boosting Multi-Instance Learning Schemes via Task-specific Prompt Tuning. arXiv preprint arXiv:2303.12214.

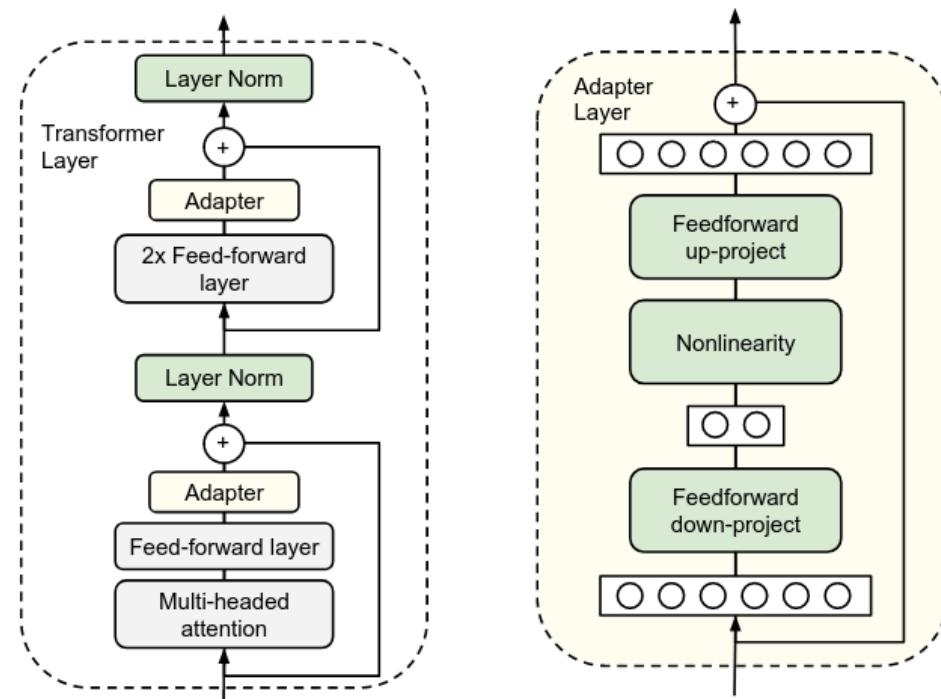
Adapters

- Houlsby et al 2019 introduced something called an adapter for NLP. Adapters serve the same purpose as fine-tuning but do it by stitching in layers to the main pre-trained model, and updating the weights of these new layers, freezing the weights of the pre-trained model.
- More efficient in terms of time and storage
- Match the performance of state-of-the-art fine-tuning methods.
- Benefits
 - Task-specific layer-wise representation learning
 - Small, scalable, shareable
 - Non-interfering composition of information



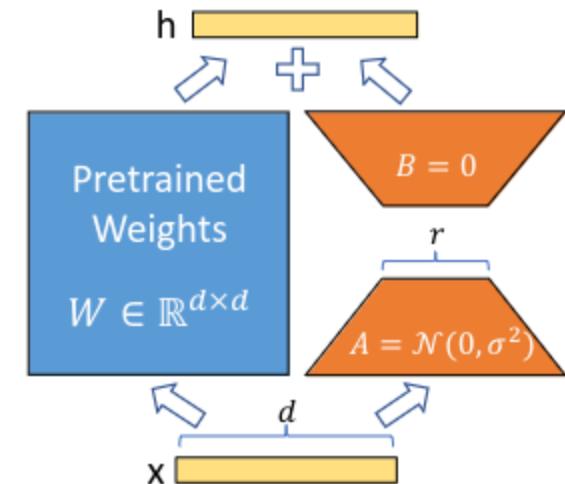
Bottleneck adapter

- Houlsby et al 2019 integrated with the Transformer.
 - Added twice to each Transformer layer. Only green layers are trained.
 - The adapter consists of a bottleneck which contains few parameters relative to the attention and feedforward layers in the original model.
 - There is a skip connection.
- On the GLUE dataset, bottleneck adapter.
 - Maintain within 0.4% of the performance of full fine-tuning
 - Adding only 3.6% parameters per task, compared to fine tuning 100% of the parameters per task.



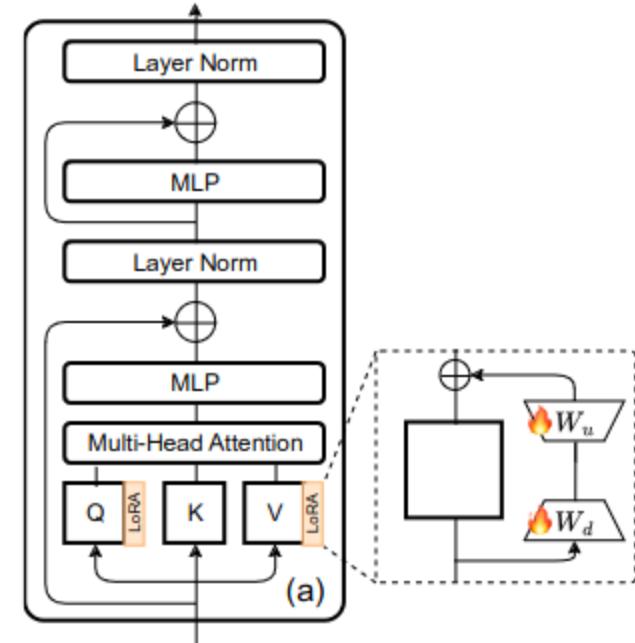
LoRA adapter

- Hu et al 2021 proposed the low-rank adaptation for NLP.
 - A pre-trained model can be shared and used to build many small LoRA modules for different tasks.
 - LoRA makes training more efficient and lowers the hardware barrier to entry by up to 3 times when using adaptive optimizers. Instead, only the injected much smaller low-rank matrices are optimized.
 - Orthogonal to many prior methods and can be combined with many of them, as prefix-training



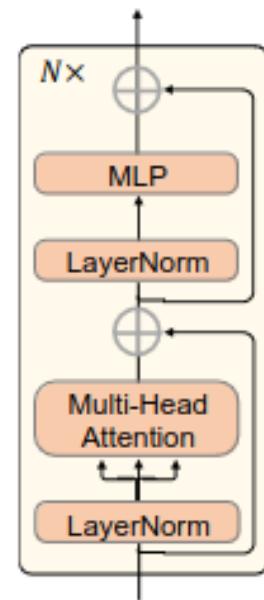
LoRA adapter

- Hu et al 2021 proposed the low-rank adaptation for NLP.
 - A pre-trained model can be shared and used to build many small LoRA modules for different tasks.
 - LoRA makes training more efficient and lowers the hardware barrier to entry by up to 3 times when using adaptive optimizers. Instead, only the injected much smaller low-rank matrices are optimized.
 - Orthogonal to many prior methods and can be combined with many of them, as prefix-training
- In the Transformer architecture, there are four weight matrices in the self-supervised module (W_q, W_k, W_v, W_o) and two in the MLP module
- There is an ablation to decide where to apply. The best performance when applied to W_q, W_v

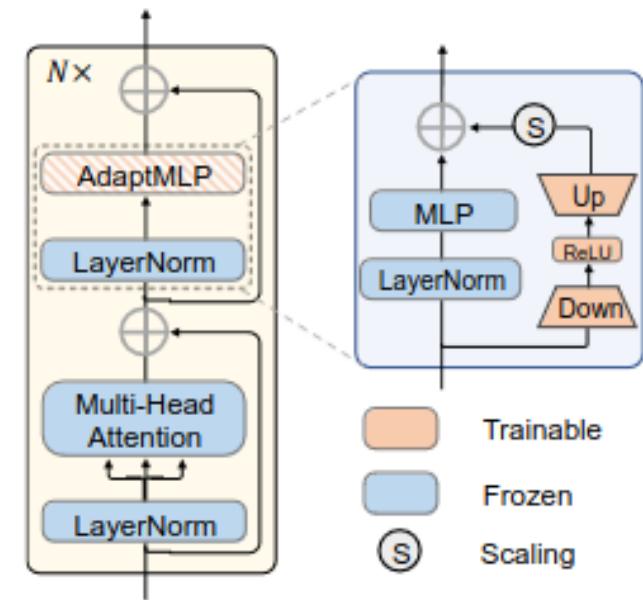


AdaptFormer adapter

- Chen et al 2022 proposed AdaptFormer, which can adapt the pre-trained ViTs into many different image and video tasks efficiently.
- Very light add less than 2% extra parameters to a ViT, while it increases the ViT's transferability without updating its original pre-trained parameters, significantly outperforming the existing fully fine tuned models.
- Plug-and-play in different Transformers and scalable to many visual tasks.



(a) Full fine-tuning.



(b) AdaptFormer fine-tuning

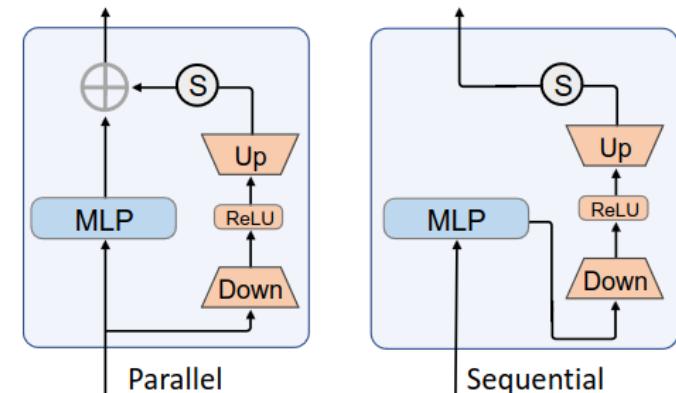
Trainable
Frozen
Scaling

AdaptFormer adapter

- Chen et al 2022 proposed AdaptFormer, which can adapt the pre-trained ViTs into many different image and video tasks efficiently.

Method	Avg. Params (M)	Image			Video	
		CIFAR-100	SVHN	Food-101	SSv2	HMDB51
Full-tuning	86.04 (100%)	85.90	97.67 [†]	90.09 [†]	53.97	46.41
Linear	0.07 (0.08%)	69.83 (-16.07)	66.91 (-30.76)	69.74 (-20.35)	29.23 (-24.74)	49.84 (+3.43)
VPT [51]	0.08 (0.09%)	82.44 (-3.46)	94.02 (-3.65)	82.98 (-7.11)	43.73 (-10.24)	52.67 (+6.26)
AdaptFormer-1	0.10 (0.12%)	83.52 (-2.38)	93.04 (-4.63)	83.64 (-6.45)	50.03 (-3.94)	51.68 (+5.27)
AdaptFormer-4	0.15 (0.17%)	84.83 (-1.07)	96.19 (-1.48)	85.42 (-4.67)	54.70 (+0.73)	51.81 (+5.40)
AdaptFormer-64	1.26 (1.46%)	85.90 (0.00)	96.89 (-0.78)	87.61 (-2.48)	59.02 (+5.05)	55.69 (+9.28)

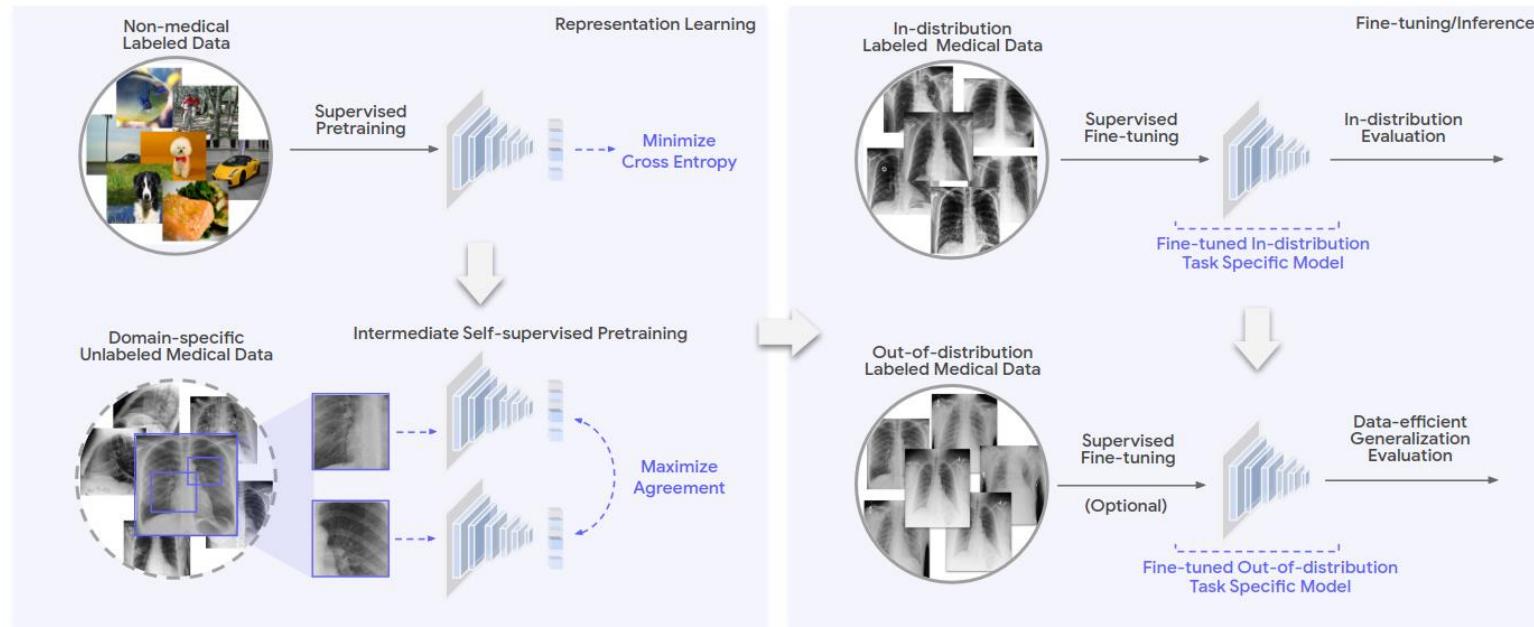
- A lot of ablations with respect to the position.
- Parallel works better than sequential
 - The parallel design maintains the original feature using an independent branch.
 - The sequential design is equivalent to adding more layers, which might cause optimization difficulty.



Foundation Models in Medical Imaging

REMEDIS model

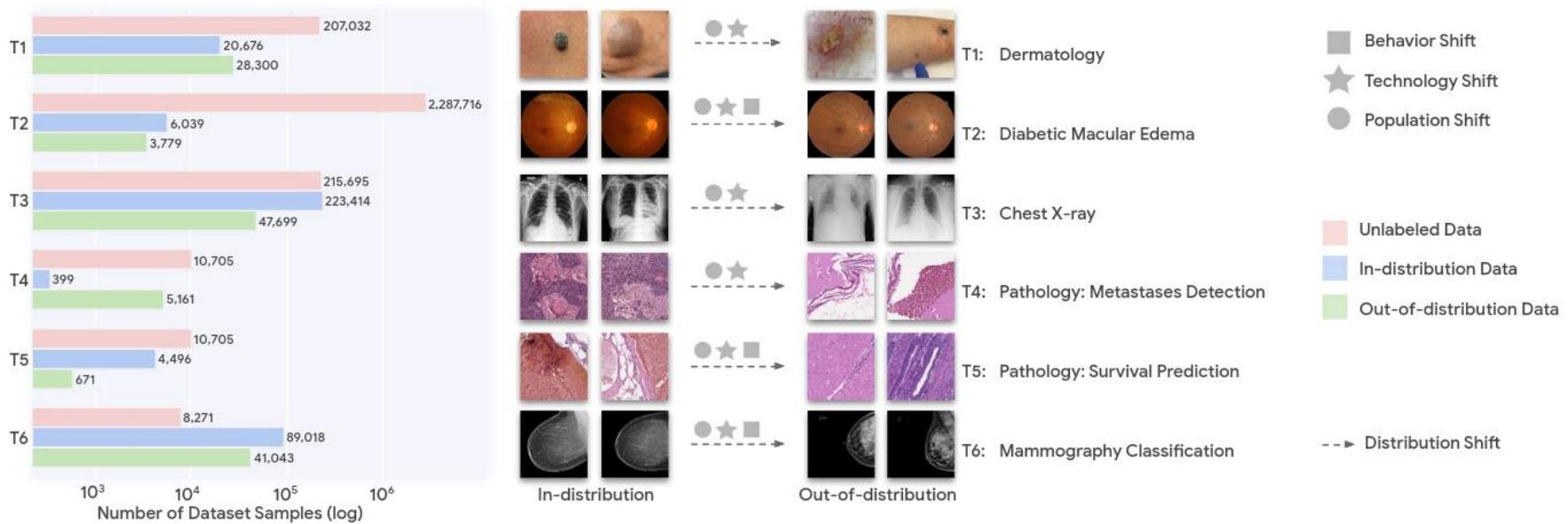
- Azizi et al 2022 proposed a unified representation learning strategy to improve robustness and data-efficiency of medical imaging AI.



- REMEDIS starts with representations initialized using large-scale natural image pretraining
- The model is adapted to the medical domain using intermediate contrastive self-supervised learning without using any labeled medical data.
- Fine-tune the model to specific downstream medical imaging AI tasks.

REMEDIS model

- Azizi et al 2022 proposed a unified representation learning strategy to improve robustness and data-efficiency of medical imaging AI.



- REMEDIS uses SimCLR contrastive self-supervised pretraining
- The encoder is further fine-tuned using annotated medical images for the domain-specific medical task.
- For each task we initialized a classifier head $g(\cdot)$ to map representations to the domain-specific label space.

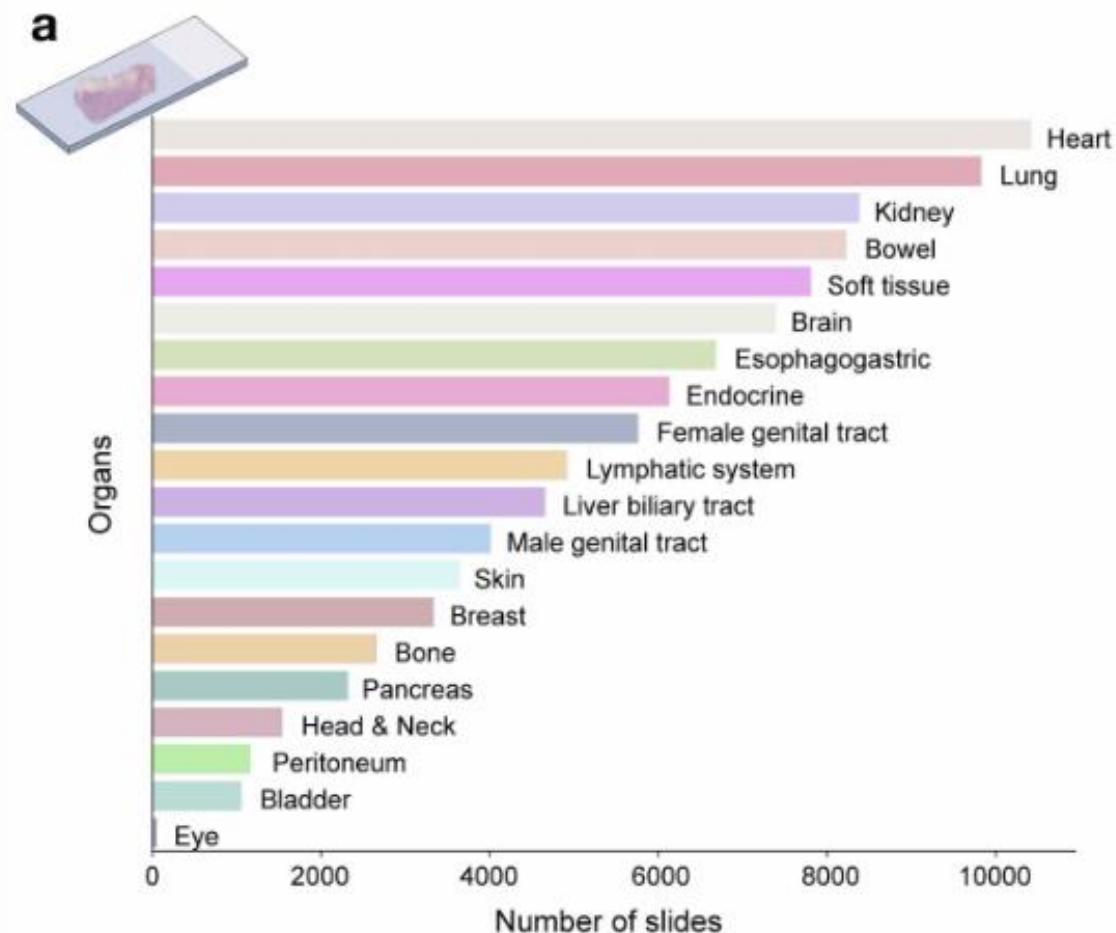
REMEDIS model

- Evaluation tasks.
 - Dermatology condition classification.
 - Chest X-ray classification
 - Pathology metastases detection
 - Pathology colorectal survival prediction

Tasks	Method	In-distribution	Out-of-dist. (0%)	Out-of-dist. (100%)
Task 1 (Top-3 Acc.)	Supervised (ImageNet)	0.900 (0.897,0.903)	0.738 (0.734,0.743)	0.839 (0.838,0.840)
	Supervised (JFT)	0.923 (0.922,0.925)*	0.755 (0.750,0.760)*	0.844 (0.842,0.845)
	REMEDIS	0.926 (0.925,0.928)	0.763 (0.760,0.769)	0.864 (0.863,0.866)
Task 2 (AUC)	Supervised (ImageNet)	0.887 (0.886,0.887)	0.685 (0.682,0.688)	0.761 (0.759,0.764)
	Supervised (JFT)	0.891 (0.889,0.892)	0.718 (0.715,0.720)	0.755 (0.750,0.761)
	REMEDIS	0.902 (0.900,0.902)	0.731 (0.727,0.736)	0.816 (0.811,0.821)
Task 3 (AUC)	Supervised (ImageNet)	0.818 (0.818,0.819)	0.786 (0.783,0.788)	0.812 (0.807,0.817)
	Supervised (JFT)	0.816 (0.815,0.816)	0.785 (0.781,0.788)	0.825 (0.824,0.826)
	REMEDIS	0.833 (0.832,0.833)	0.798 (0.796,0.800)	0.835 (0.834,0.836)
Task 4 (AUC)	Supervised (ImageNet)	0.856 (0.851,0.864)	0.757 (0.755,0.758)	0.892 (0.886,0.895)
	Supervised (JFT)	0.916 (0.916,0.917)	0.791 (0.790,0.792)	0.905 (0.897,0.911)
	REMEDIS	0.954 (0.950,0.960)	0.876 (0.876,0.876)	0.958 (0.956,0.960)
Task 5 (AUC)	Supervised (ImageNet)	0.714 (0.712,0.715)	0.649 (0.645,0.655)	0.725 (0.719,0.729)
	Supervised (JFT)	0.699 (0.698,0.699)	0.664 (0.661,0.667)	0.760 (0.757,0.763)
	REMEDIS	0.748 (0.747,0.748)	0.712 (0.710,0.714)	0.798 (0.792,0.804)
Task 6 (AUC)	Supervised (ImageNet)	0.852 (0.848,0.856)	0.700 (0.697,0.702)	0.727 (0.725,0.728)
	Supervised (JFT)	0.869 (0.866,0.872)†	0.711 (0.709,0.715)	0.734 (0.732,0.736)
	REMEDIS	0.870 (0.868,0.872)	0.725 (0.724,0.726)	0.750 (0.749,0.751)

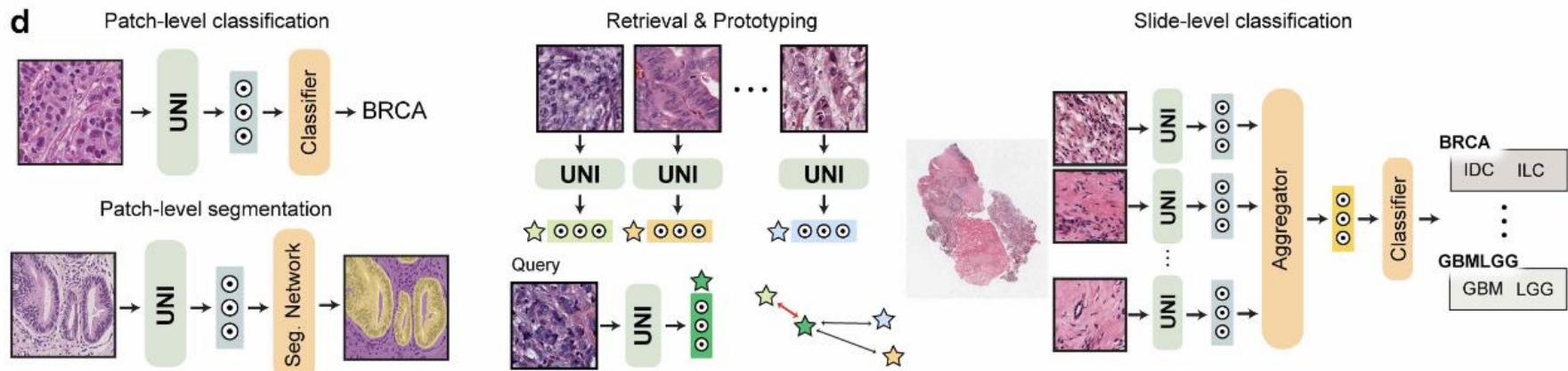
UNI model

- Chen et al. 2023 proposed a DINoV2 model trained on digital pathology.
 - Mass-100K: largest histopathology dataset: 100 million tissue patches from ~100k H&E WSIs.



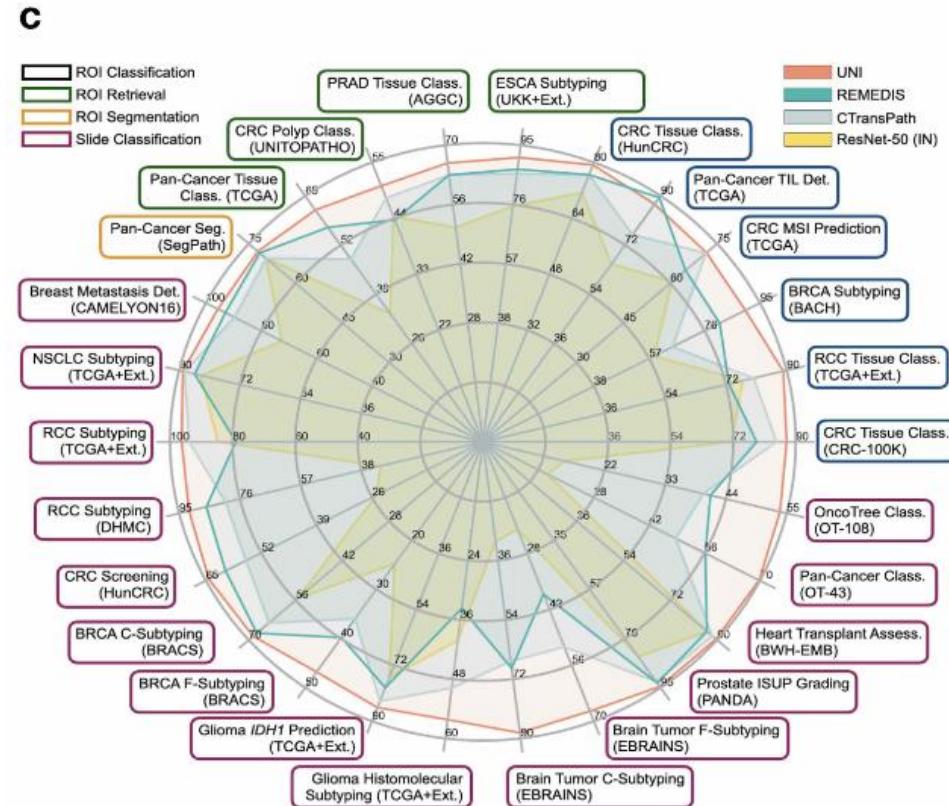
UNI model

- Chen et al. 2023 proposed a DINoV2 model trained on digital pathology.
 - Mass-100K: largest histopathology dataset: 100 million tissue patches from ~100k H&E WSIs.
 - Validated on different digital pathology tasks



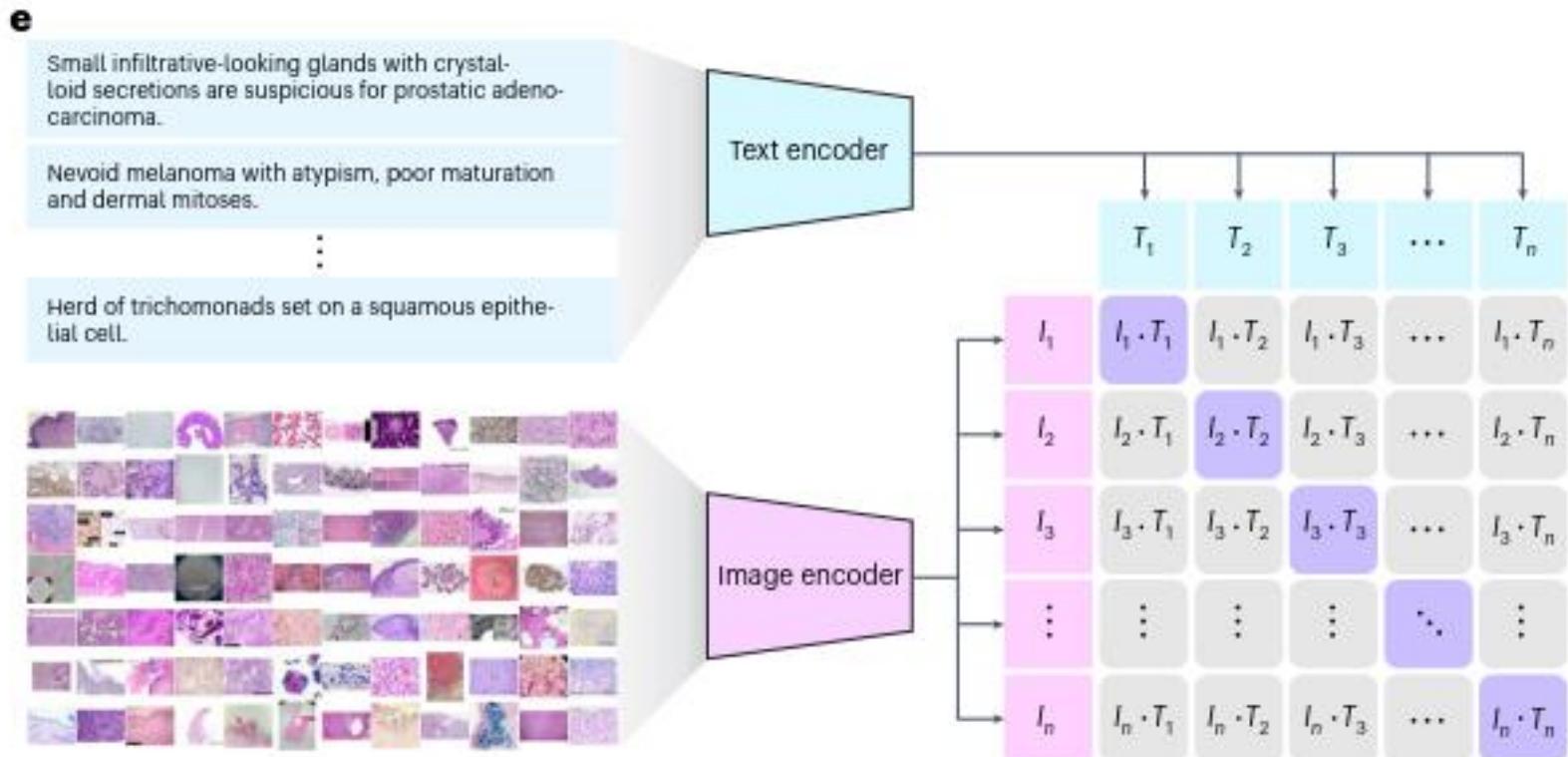
UNI model

- Chen et al. 2023 proposed a DINoV2 model trained on digital pathology.
 - Mass-100K: largest histopathology dataset: 100 million tissue patches from ~100k H&E WSIs.
 - Validated on different digital pathology tasks



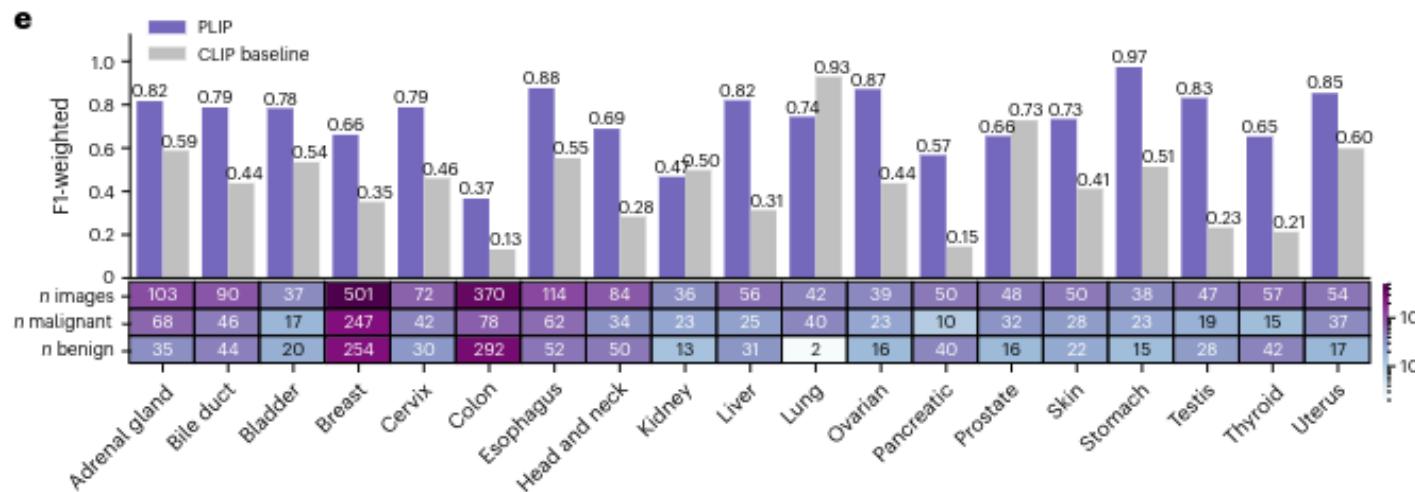
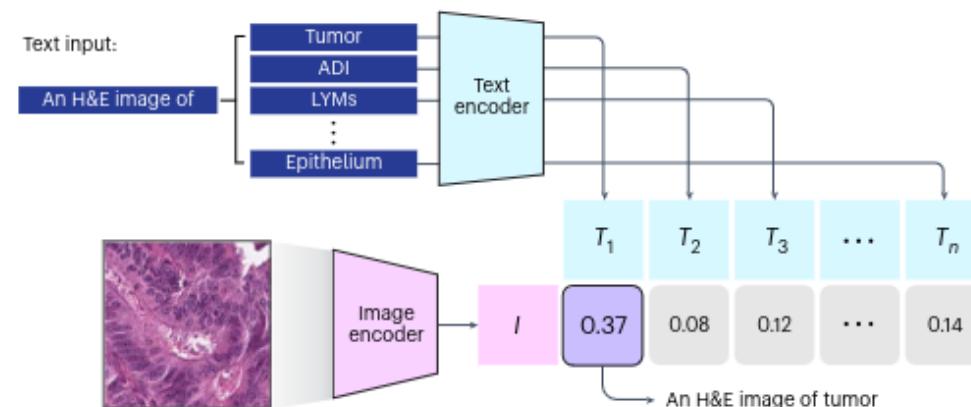
PLIP model

- Huang et al. 2023 proposed OpenPath, a large dataset of ~210k pathology images paired with natural language descriptions.
- Fine tuned a pretrained contrastive language-image pretraining (CLIP) model using contrastive learning.



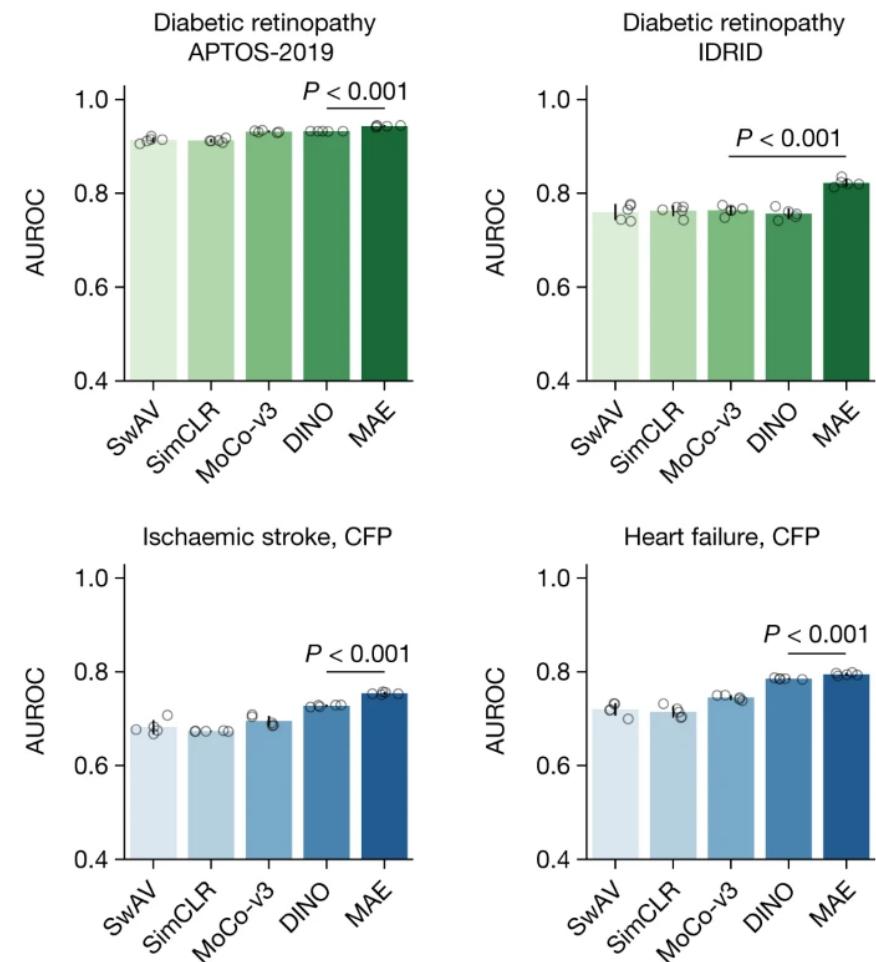
PLIP model

- Huang et al. 2023 proposed OpenPath, a large dataset of ~210k pathology images paired with natural language descriptions.
- Zero-shot capability, which enables learning new classes at scale without the need of retraining. The evaluation on four external datasets
 - Colon dataset
 - Panuke
 - DigestPath
 - WSS4LUAD



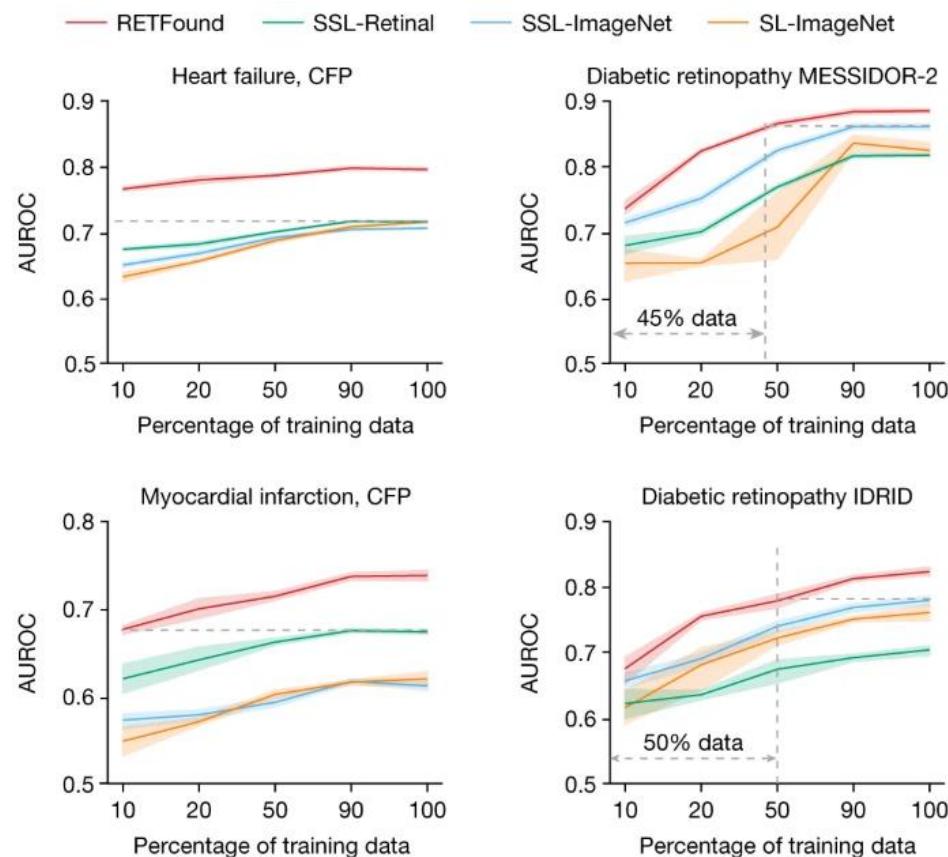
RETFound model

- Zhou et al. 2023 proposed a model trained on 1.6 million unlabelled retinal images by means of self-supervised learning and then adapted to disease detection tasks with explicit labels.
- Comparison of different SSL strategies
 - MAE perform the best



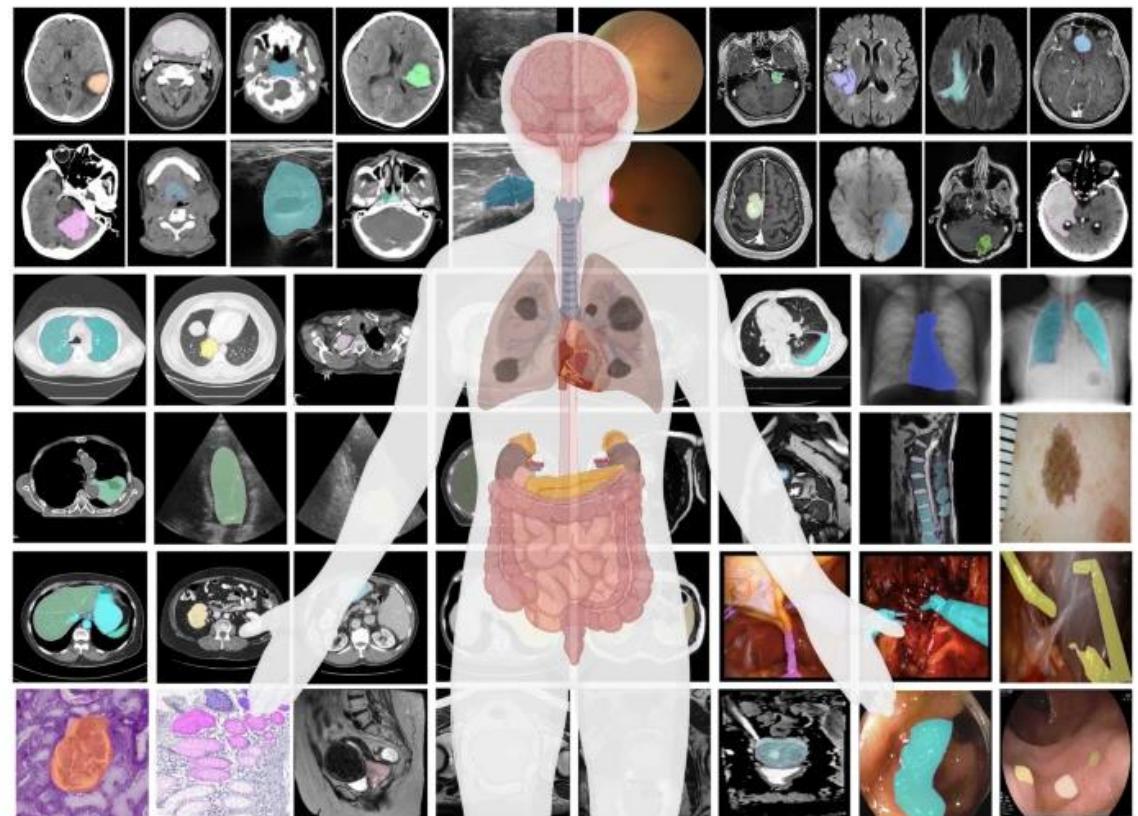
RETFound model

- Zhou et al. 2023 proposed a model trained on 1.6 million unlabelled retinal images by means of self-supervised learning and then adapted to disease detection tasks with explicit labels.
- Ocular disease diagnosis
- Ocular disease prognosis
- Systemic diseases predictions
- Label efficiency for disease detection
- Robustness to age distribution shifts



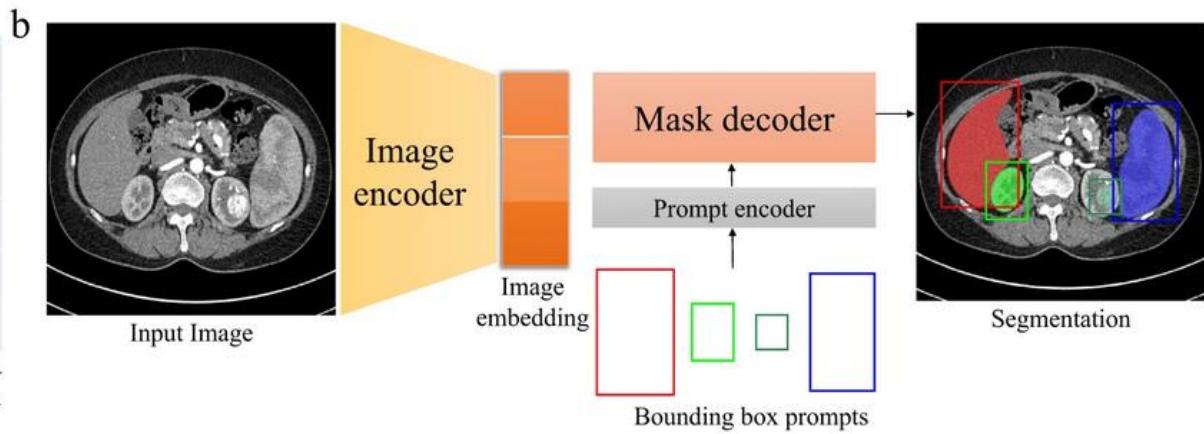
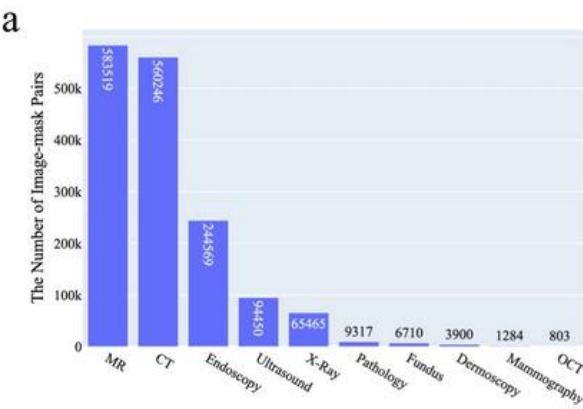
MedSAM model

- Ma et al. 2024 proposed a model enabling universal medical image segmentation.
 - 1.570.263 medical image-mask pairs
 - Covering 10 imaging modalities
 - More than 30 cancer types
- Validation
 - 86 internal validation tasks
 - 60 external validation tasks



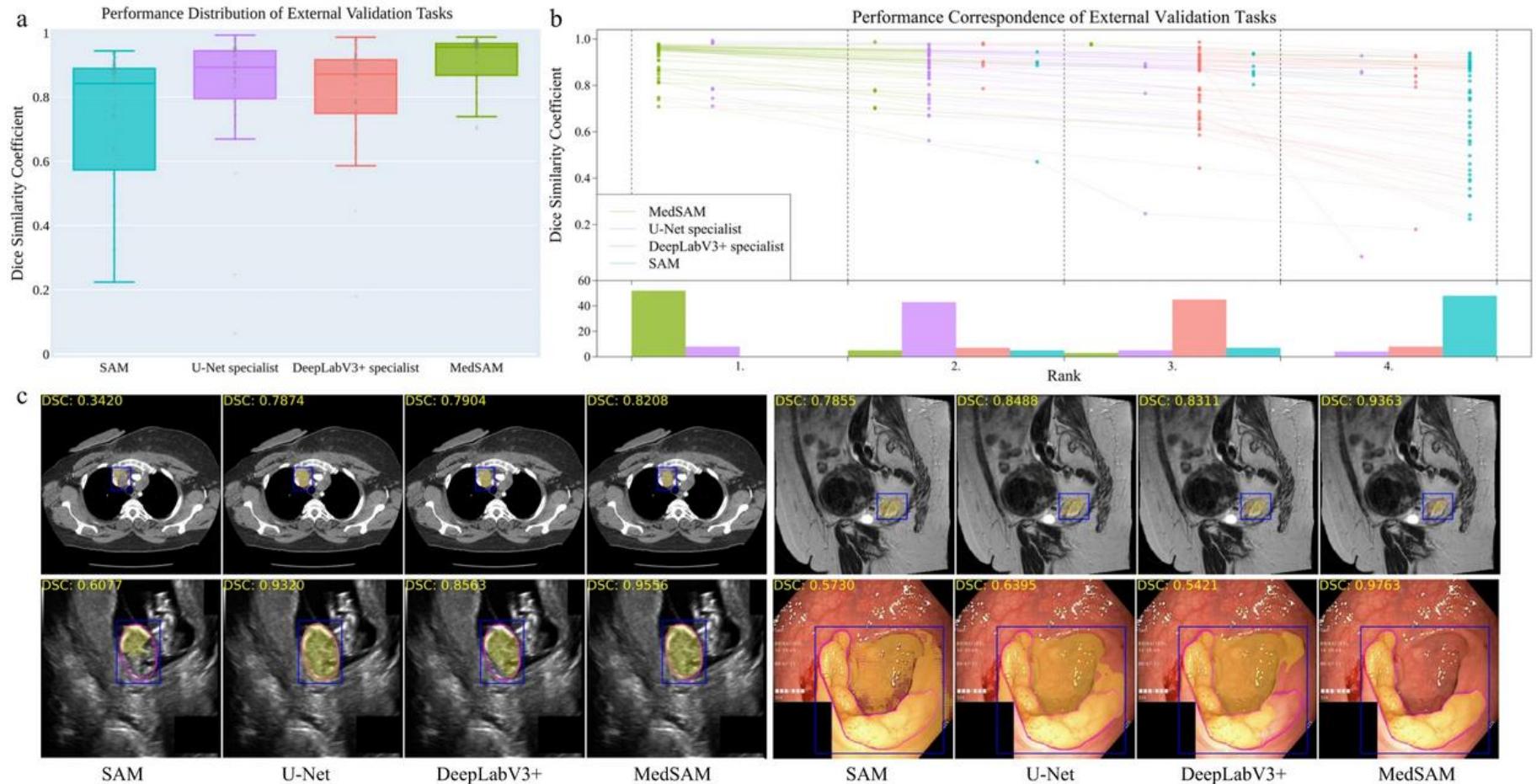
MedSAM model

- Ma et al. 2024 proposed a model enabling universal medical image segmentation.
 - 1.570.263 medical image-mask pairs
 - Covering 10 imaging modalities
 - More than 30 cancer types
- The model is developed for a promptable 2D segmentation model.
- ViT-B was used first trained with MAE modeling followed by fully supervised training on the SAM dataset.



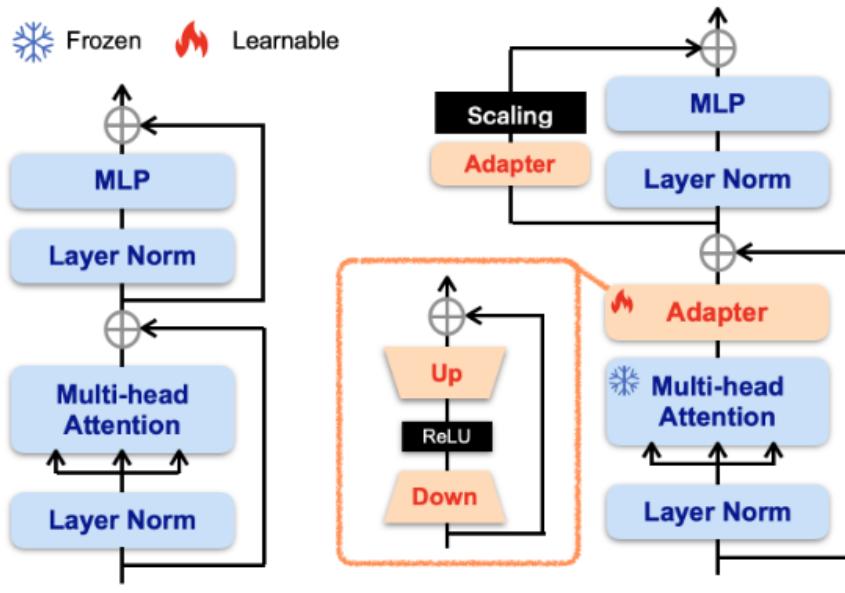
MedSAM model

- Ma et al. 2024 proposed a model enabling universal medical image segmentation.



Medical SAM adapter

- Wu et al. 2023 proposed Medical SAM Adapter (Med-SA), which incorporates domain-specific medical knowledge into the segmentation model using a light yet effective adaptation technique.

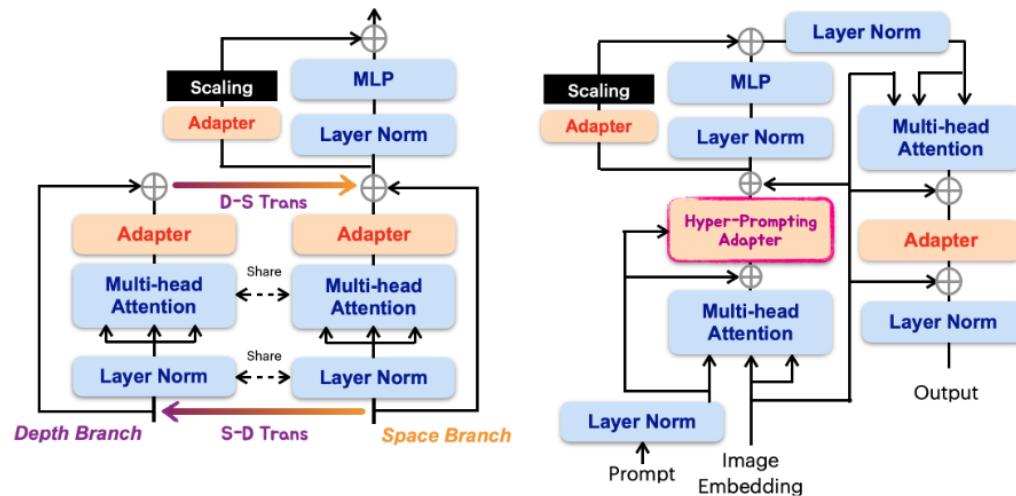


a. ViT Block in SAM

b. 2D Medical Image Adaption

Medical SAM adapter

- Wu et al. 2023 proposed Medical SAM Adapter (Med-SA), which incorporates domain-specific medical knowledge into the segmentation model using a light yet effective adaptation technique.
- Space-Depth Transpose (SD-Trans) to adapt 2D SAM to 3D medical images.
 - Depth and space branches
 - The depth branch is transposed to the original space and added to the space
- Hyper-Prompting Adapter (HyP-Adpt) to achieve prompt-conditioned adaptation.
 - Projection and reshaping operations to generate sequence of weighted maps from the prompt embedding.



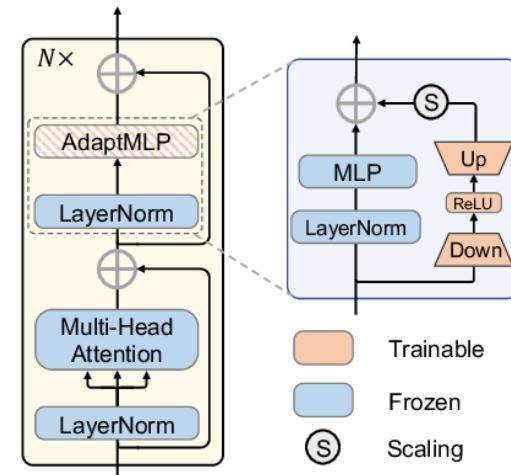
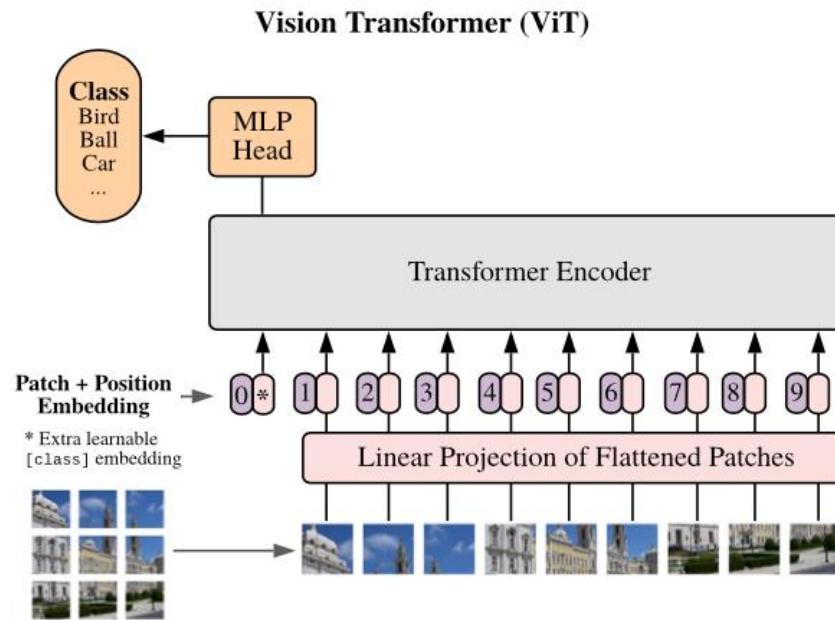
Medical SAM adapter

- Wu et al. 2023 proposed Medical SAM Adapter (Med-SA), which incorporates domain-specific medical knowledge into the segmentation model using a light yet effective adaptation technique.

Tasks	Method	In-distribution	Out-of-dist. (0%)	Out-of-dist. (100%)
Task 1 (Top-3 Acc.)	Supervised (ImageNet)	0.900 (0.897,0.903)	0.738 (0.734,0.743)	0.839 (0.838,0.840)
	Supervised (JFT)	0.923 (0.922,0.925)*	0.755 (0.750,0.760)*	0.844 (0.842,0.845)
	REMEDIS	0.926 (0.925,0.928)	0.763 (0.760,0.769)	0.864 (0.863,0.866)
Task 2 (AUC)	Supervised (ImageNet)	0.887 (0.886,0.887)	0.685 (0.682,0.688)	0.761 (0.759,0.764)
	Supervised (JFT)	0.891 (0.889,0.892)	0.718 (0.715,0.720)	0.755 (0.750,0.761)
	REMEDIS	0.902 (0.900,0.902)	0.731 (0.727,0.736)	0.816 (0.811,0.821)
Task 3 (AUC)	Supervised (ImageNet)	0.818 (0.818,0.819)	0.786 (0.783,0.788)	0.812 (0.807,0.817)
	Supervised (JFT)	0.816 (0.815,0.816)	0.785 (0.781,0.788)	0.825 (0.824,0.826)
	REMEDIS	0.833 (0.832,0.833)	0.798 (0.796,0.800)	0.835 (0.834,0.836)
Task 4 (AUC)	Supervised (ImageNet)	0.856 (0.851,0.864)	0.757 (0.755,0.758)	0.892 (0.886,0.895)
	Supervised (JFT)	0.916 (0.916,0.917)	0.791 (0.790,0.792)	0.905 (0.897,0.911)
	REMEDIS	0.954 (0.950,0.960)	0.876 (0.876,0.876)	0.958 (0.956,0.960)
Task 5 (AUC)	Supervised (ImageNet)	0.714 (0.712,0.715)	0.649 (0.645,0.655)	0.725 (0.719,0.729)
	Supervised (JFT)	0.699 (0.698,0.699)	0.664 (0.661,0.667)	0.760 (0.757,0.763)
	REMEDIS	0.748 (0.747,0.748)	0.712 (0.710,0.714)	0.798 (0.792,0.804)
Task 6 (AUC)	Supervised (ImageNet)	0.852 (0.848,0.856)	0.700 (0.697,0.702)	0.727 (0.725,0.728)
	Supervised (JFT)	0.869 (0.866,0.872)†	0.711 (0.709,0.715)	0.734 (0.732,0.736)
	REMEDIS	0.870 (0.868,0.872)	0.725 (0.724,0.726)	0.750 (0.749,0.751)

Lab Session!

- Foundation models and adapters for digital pathology



TorchVision

Facebook
Open Source

Deep learning for medical imaging

Olivier Colliot, PhD
Research Director at CNRS
Co-Head of the ARAMIS Lab –
www.aramislab.fr
PRAIRIE – Paris Artificial Intelligence
Research Institute

Maria Vakalopoulou, PhD
Assistant Professor at
CentraleSupélec
Mathematics and Informatics (MICS)
Office: Bouygues Building Sb.132



Master 2 - MVA