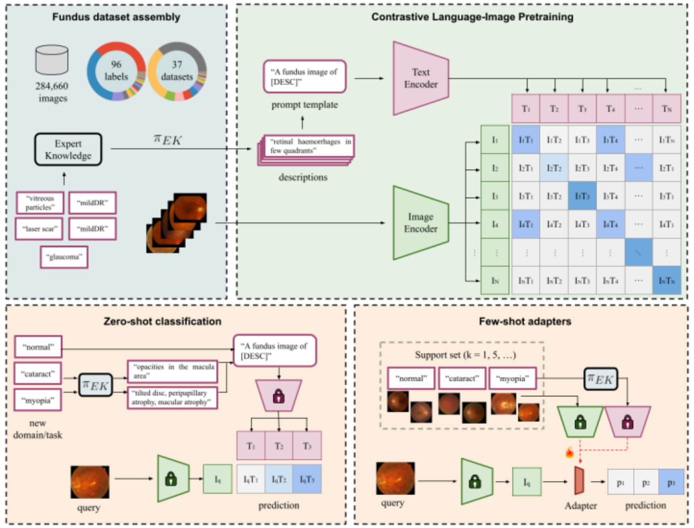
Workshop on Medical Imaging with Deep Learning
ETS Montreal

A Foundation LAnguage Image of the Retina (FLAIR): Encoding expert knowledge in text supervision

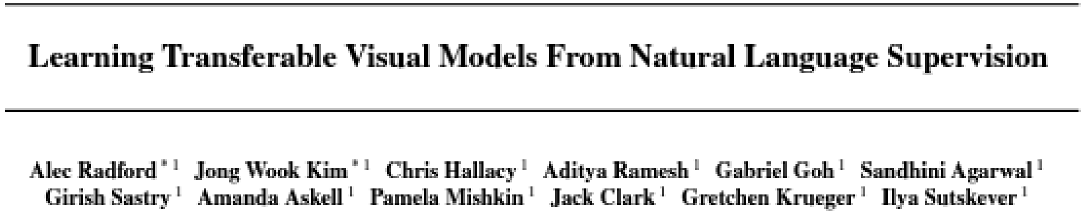
Julio Silva-Rodríguez[1], Hadi Chakor[2], Riadh Kobbi[2], Jose Dolz[1] and Ismail Ben Ayed[1]

ETS Montreal[1], DIAGNOS Inc. [2]

https://github.com/jusiro/FLAIR

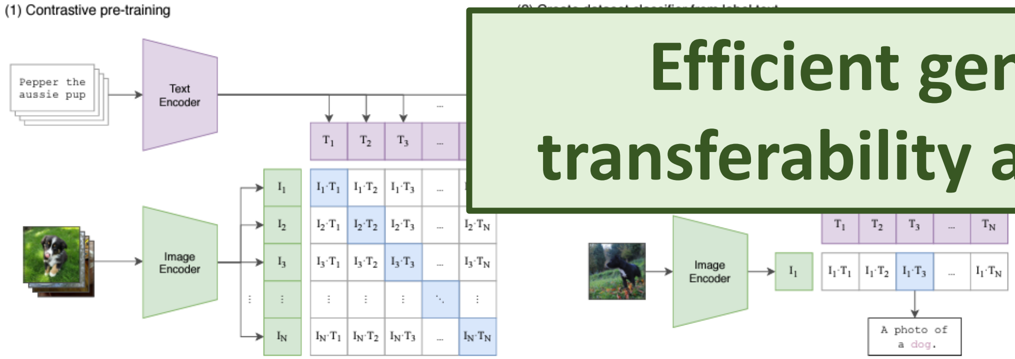


What is a *foundation model*?

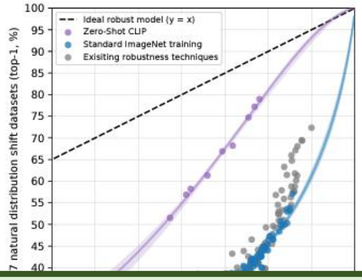


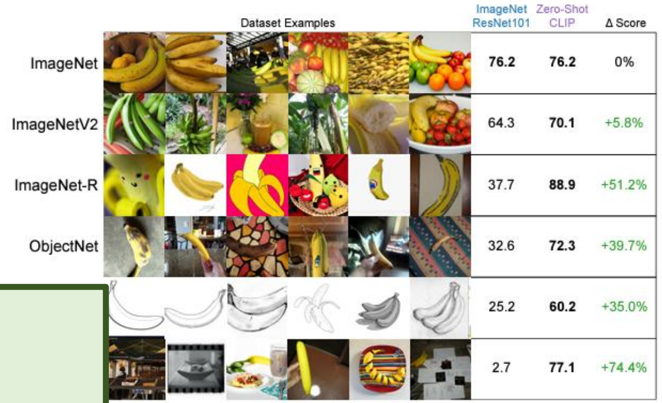
Learning Transferable Visual Models From Natural Language Supervision

Alec Radford * 1 Jong Wook Kim * 1 Chris Hallacy 1 Aditya Ramesh 1 Gabriel Goh 1 Sandhini Agarwal 1
Girish Sastry 1 Amanda Askell 1 Pamela Mishkin 1 Jack Clark 1 Gretchen Krueger 1 Ilya Sutskever 1



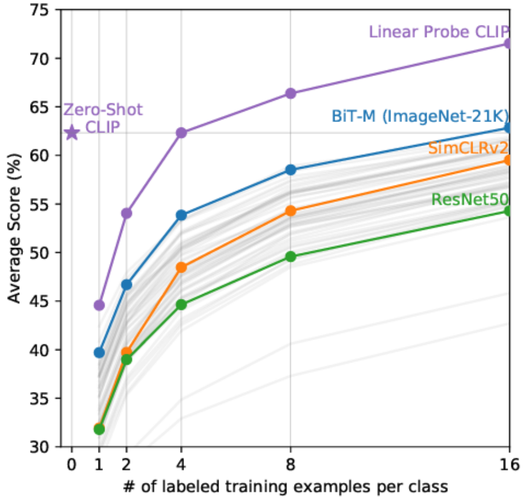
400M image-text pairs



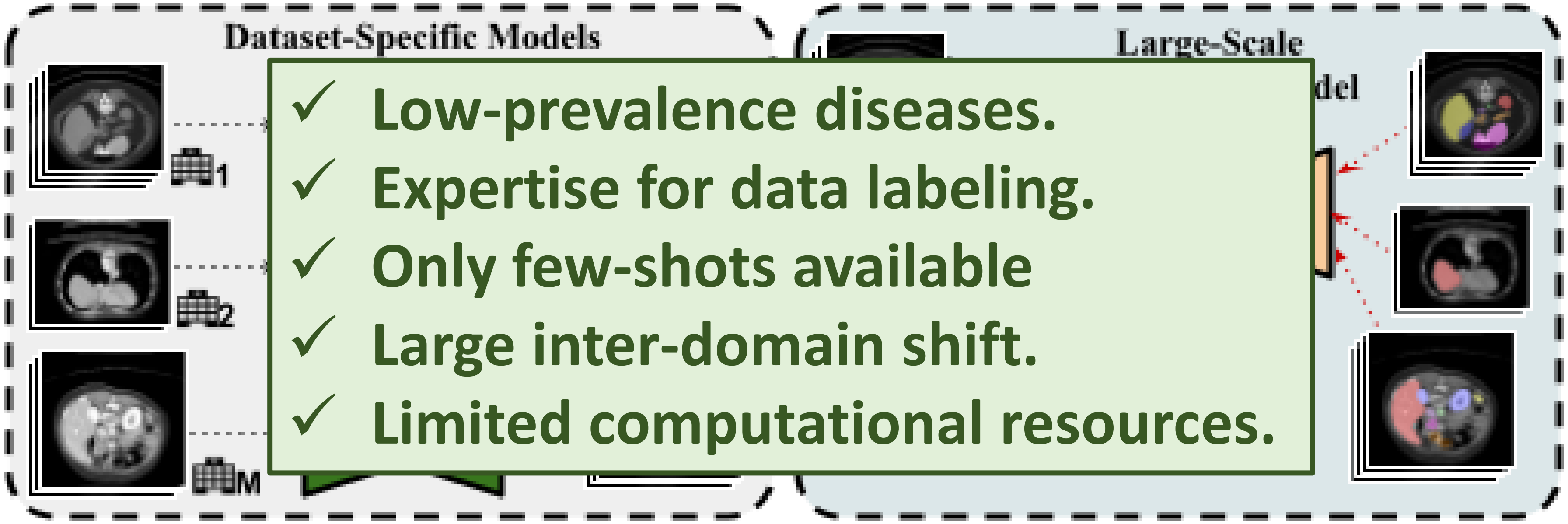


	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
	25.2	60.2	+35.0%
	2.7	77.1	+74.4%

Efficient generalization, transferability and robustness!



From *dataset-specific models* to *pretrain-and-adapt*



- ✓ Low-prevalence diseases.
- ✓ Expertise for data labeling.
- ✓ Only few-shots available
- ✓ Large inter-domain shift.
- ✓ Limited computational resources.

Generalists vs. *Specialized* Foundation Models

LARGE-SCALE DOMAIN-SPECIFIC PRETRAINING FOR BIOMEDICAL VISION-LANGUAGE PROCESSING

Sheng Zhang*, Yanbo Xu*, Naoto Usuyama*, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon
Microsoft Research

15M image-text pairs

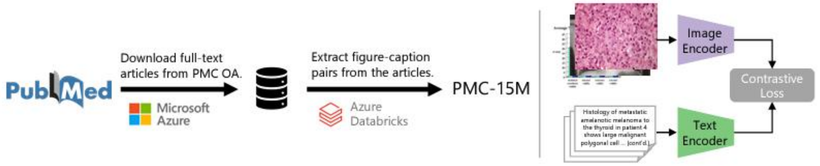
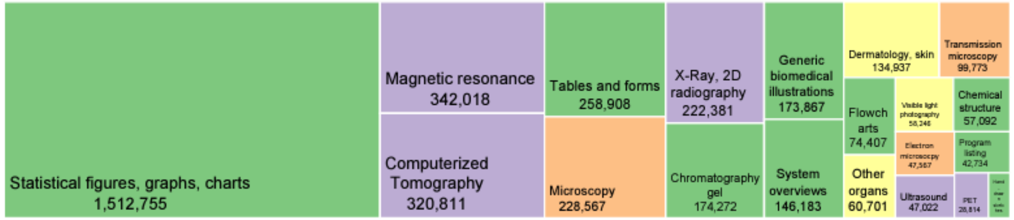


Figure 2: Overview of PMC-15M creation pipeline (left) and BiomedCLIP pretraining (right).

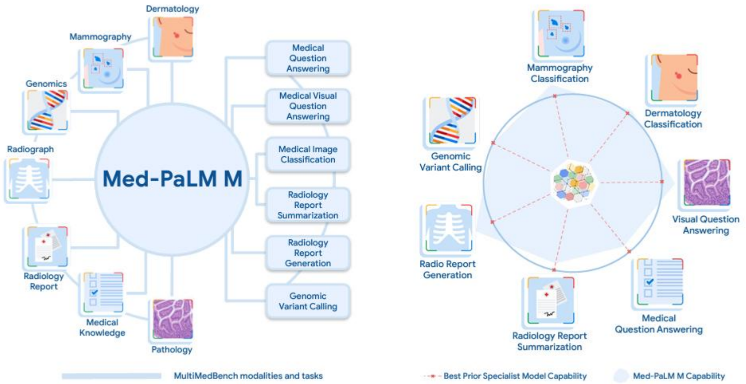


Towards Generalist Biomedical AI

Tao Tu[*, ‡, 1], Shekoofeh Azizi[*, ‡, 2],
Danny Driess[2], Mike Schaekermann[1], Mohamed Amin[1], Pi-Chuan Chang[1], Andrew Carroll[1],
Chuck Lau[1], Ryutaro Tanno[2], Ira Ktena[2], Basil Mustafa[2], Aakanksha Chowdhery[2], Yun Liu[1],
Simon Kornblith[2], David Fleet[2], Philip Mansfield[1], Sushant Prakash[1], Renee Wong[1], Sunny Virmani[1],
Christopher Semturs[1], S Sara Mahdavi[2], Bradley Green[1], Ewa Dominowska[1], Blaise Aguera y Arcas[1],
Joelle Barral[2], Dale Webster[1], Greg S. Corrado[1], Yossi Matias[1], Karan Singhal[1], Pete Florence[2],
Alan Karthikesalingam[†, ‡, 1] and Vivek Natarajan[†, ‡, 1]

[1]Google Research, [2]Google DeepMind

1M image-text pairs

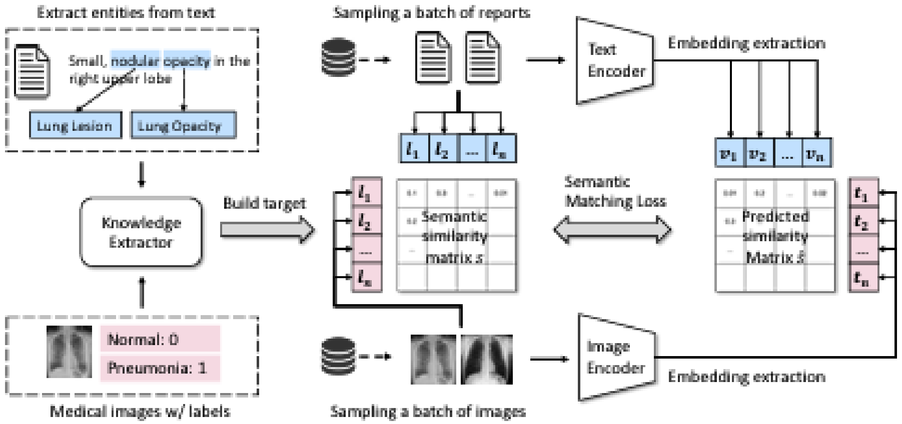


Generalists vs. ***Specialized*** Foundation Models

MedCLIP: Contrastive Learning from Unpaired Medical Images and Text

Zifeng Wang[1], Zhenbang Wu[1], Dinesh Agarwal[1,3], Jimeng Sun[1,2]
[1]Department of Computer Science, University of Illinois Urbana-Champaign
[2]Carle Illinois College of Medicine, University of Illinois Urbana-Champaign
[3]Adobe
{zifengw2, zw12, jimeng}@illinois.edu, diagarwa@adobe.com

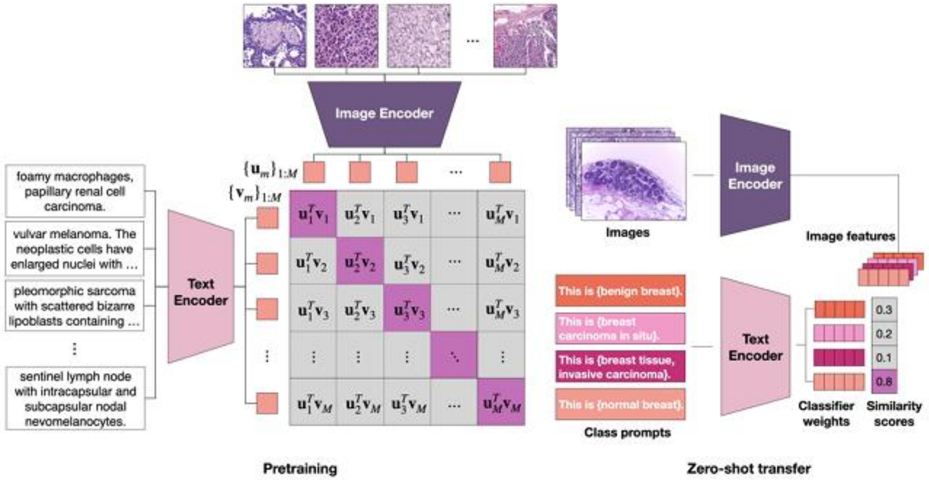
250K image-text pairs



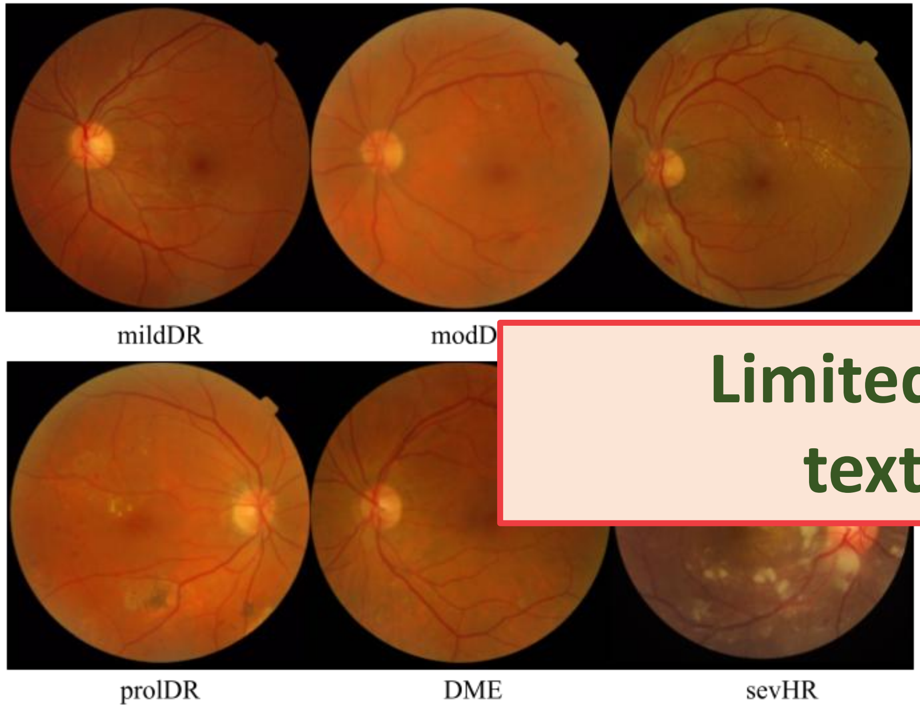
Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images

Ming Y. Lu[†,1,2,3], Bowen Chen[†,2,3], Andrew Zhang[1,2,3], Drew F.K. Williamson[2,3], Richard J. Chen[2,3], Tong Ding[2,3], Long Phi Le[2,3], Yung-Sung Chuang[1], Faisal Mahmood[2,3]
[1]Massachusetts Institute of Technology [2]Harvard University [3]Mass General Brigham
mingylu@mit.edu, bchen18@bwh.harvard.edu, faisalmahmood@bwh.harvard.edu

40K image-text pairs



Towards a foundation model for *fundus images*



Datasets	#Targets	#Images	Labels	Annotations
01.EYEPACS[1]	5	88,702	noDR, mildDR, modDR, sevDR, prolDR.	Categorical
02.MESSIDOR2 (Decencière et al., 2014; Krause et al., 2018)	9	1,748	noDR, mildDR, modDR, sevDR, prolDR, noisy, clean, DME, noDME, hEX.	Categorical
03.IDRID (Porwal et al., 2020)	10	597	MA, HE, hEX, sEX, noDR, mildDR, modDR, sevDR, prolDR, noDME, nonCSDME, DME.	Categorical
04.RFMid (Pachade et al., 2021)	46	3,200	DR, ARMD, MH, DN, MYA, BRVO, TSLN, ERM, LS, MS CSR, ODC, CRVO, TV, AH, ODP, ODE, ST, AION, PT, RT RS, CRS, EX, RPEC, RPEC, MHL, RP, CWS, CB, ODM, PRH, MNF, HR, CRAO, TD, CME, PTCR, CF, VH, MCA VS, BRAO, PLQ, HPED, CL.	Categorical
05.1000x39 (Cen et al., 2021)	39	1,000	N, TSLN, LOC, mildDR, modDR, sevDR, BRVO, CRVO, G, CRAO, RD, CSR, VKH, M, ERM, MHL, MYA, HE, OA, NP, sevHR, DSE, DD, CDA, RP, BCD, PRDB, MNF, VH, F, hEX, YWSF, CWS, TV, CB, LS, noisy, noProlDR, prolDR.	Categorical
06.DEN (Huang et al., 2021a)	-	15,708	–	Text
07.LAG (Li et al., 2019a)	2	4,854	G, noG.	Categorical
08.ODIR-5K[2]	≥7	8,000	N, DR, G, CAT, ARMD, HR, MYA.	Text
09.PAPILA (Kovalyk et al., 2022)	2	488	G, N.	Categorical
10.PARAGUAY (Castillo Benítez et al., 2021)	7	1,437	noDR, mildDR, modDR, sevDR, prolDR.	Categorical
11.STARE (Hoover, 2000; Hoover and Goldbaum, 2003)	-	397	–	Text
12.ARIA (Farnell et al., 2008)	3	143	N, ARMD, DR.	Categorical
13.FIVES (Jin et al., 2022)	6	800	noisy, clean, ARMD, DR, G, N.	Categorical
[illegible]	[illegible]	28	DR, MA.	Categorical
[illegible]	[illegible]	[illegible]590	noDR, mildDR, modDR, sevDR, prolDR.	Categorical
[illegible]	[illegible]	[illegible]200	G, N, CME, neovARMD, geoARMD, acCSR, chCSR.	Categorical
[illegible]	[illegible]	89	IrMA, neoV, ReSD, hEX, HE, sEX, MA.	Categorical
[illegible]	[illegible]	[illegible]10	noCAT, Dis.	Categorical
[illegible]	[illegible]	[illegible]00	N, G.	Categorical
[illegible]	[illegible]	[illegible]63	EX, MA.,	Categorical
[illegible]	[illegible]	[illegible]20	G, N.	Categorical
[illegible]	[illegible]	[illegible]69	EX, CWS, DN.	Categorical
[illegible]	[illegible]	81	N, G, DR, noisy.	Categorical
[illegible]	[illegible]	[illegible]50	G, noG.	Categorical
25.REFUGE (Orlando et al., 2019; Li et al., 2020)	2	1200	G, noG.	Categorical
26.ROC (Niemeijer et al., 2010)	1	100	MA.	Categorical
27.BRSET (Nakayama et al., 2023; Goldberger et al., 2000)	24	16,266	noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX, MA, AOD, AV, AM, noisy, clean, ME, S, NE, ARMD, BRVO, HR, DN, HE, RD, MYA, ICD.	Categorical
28.OIA-DDR (Li et al., 2019b)	9	13,673	noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX, MA.	Categorical
29.AIROGS (de Vente et al., 2023)	2	101,442	G, noG	Categorical
29.SYSU (Lin et al., 2020)	8	1,220	noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX.	Categorical
31.JICHI (Takahashi et al., 2017)	5	9,940	noDR, mildDR, modDR, sevDR, prolDR	Categorical
32.CHAKSU (Kumar et al., 2023)	2	1,345	G, noG	Categorical
33.DR1-2 (Pires et al., 2014)	7	1,597	N, ReSD, hEX, DN, CWS, supHE, deepHE	Categorical
34.Cataract[4]	4	601	N, G, CAT, RS	Categorical
35.ScarDat (Wei et al., 2018)	2	997	LS, noLS	Categorical
36.ACRIMA (Diaz-Pinto et al., 2019)	2	705	G, noG	Categorical
37.DeepDRiD (Liu et al., 2022)	5	2,256	noDR, mildDR, modDR, sevDR, prolDR	Categorical
	≥96	286,916		

Limited datasets with text supervision

Open-Access Datasets

Encoding *expert knowledge* in text supervision

“moderate diabetic retinopathy”

“contains few microaneurysms”

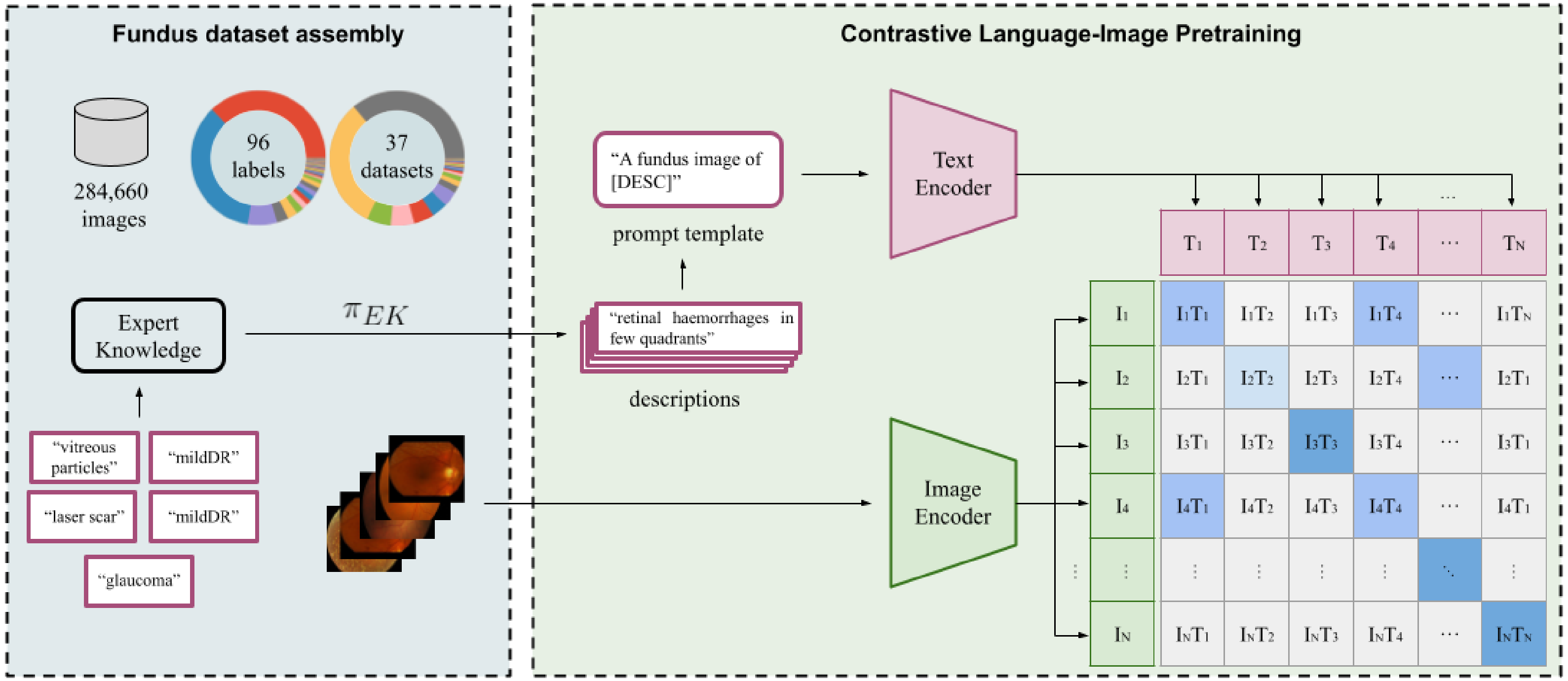
“diabetic macular edema”

“exudates near the macula center”

Category	Domain Knowledge descriptor
no diabetic retinopathy	”no relevant haemorrhages, microaneurysms or exudates” / ”no microaneurysms” / ”no referable lesions”
mild diabetic retinopathy	”few microaneurysms” / ”few hard exudates” / ”few retinal haemorrhages”
moderate diabetic retinopathy	”retinal haemorrhages in few quadrants” / ”many haemorrhages” / ”cotton wool spots”
severe diabetic retinopathy	”severe haemorrhages in all four quadrants” / ”venous beading” / ”intraretinal microvascular abnormalities”
proliferative diabetic retinopathy	”diabetic retinopathy with neovascularization at the disk” / ”neovascularization”
diabetic macular edema	”macular edema” / ”presence of exudates” / ”leakage of fluid within the central macula from microaneurysms” / ”presence of exudates within the radius of one disc diameter from the macula center”
no referable diabetic macular edema	”no apparent exudates”
hard exudates	”small white or yellowish deposits with sharp margins” / ”bright lesion”
soft exudates	”pale yellow or white areas with ill-defined edges” / ”cotton-wool spot” / ”small, whitish or grey, cloud-like, linear or serpentine, slightly elevated lesions with fimbriated edges”
microaneurysms	”small red dots”
haemorrhages	”dense, dark red, sharply outlined lesion”
non clinically significant diabetic macular edema	”presence of exudates outside the radius of one disc diameter from the macula center” / ”presence of exudates”
age-related macular degeneration	”many small drusen” / ”few medium-sized drusen” / ”large drusen”
media haze	”vitreous haze” / ”pathological opacity” / ”the obscuration of fundus details by vitreous cells and protein exudation”
drusens	”yellow deposits under the retina” / ”numerous uniform round yellow-white lesions”
pathologic myopia	”tilted disc, peripapillary atrophy, and macular atrophy. There are chorioretinal scars in the inferonasal periphery” / ”maculopahy”
branch retinal vein occlusion	”occlusion of one of the four major branch retinal veins”
tessellation	”large choroidal vessels at the posterior fundus”
epiretinal membrane	”greyish semi-translucent avascular membrane”
laser scar	”round or oval, yellowish-white with variable black pigment centrally” / ”50 to 200 micron diameter lesions”
central serous retinopathy	”subretinal fluid involving the fovea” / ”leakage”
asteroid hyalosis	”multiple sparking, yellow-white, and refractile opacities in the vitreous cavity” / ”vitreous opacities”
optic disc pallor	”pale yellow discoloration that can be segmental or generalized on optic disc”
shunt	”collateral vessels connecting the choroidal and the retinal vasculature” / ”collateral vessels of large caliber and lack of leakage”
exudates	”small white or yellowish-white deposits with sharp margins” / ”bright lesion”
macular hole	”a lesion in the macula” / ”small gap that opens at the centre of the retina”
retinitis pigmentosa	”bone spicule-shaped pigment deposits are present in the mid periphery” / ”retinal atrophy” ”the macula is preserved” / ”peripheral ring of depigmentation” / ”arteriolar attenuation and atrophy of the retinal pigmented epithelium”
cotton wool spots	”soft exudates”
glaucoma	”optic nerve abnormalities” / ”abnormal size of the optic cup” / ”anomalous size in the optic disc”
severe hypertensive retinopathy	”flame-shaped hemorrhages at the disc margin, blurred disc margins” / ”congested retinal veins, papilledema, and secondary macular exudates” / ”arterio-venous crossing changes, macular star and cotton wool spots”
no proliferative diabetic retinopathy	”diabetic retinopathy with no neovascularization” / ”no neovascularization”
hypertensive retinopathy	”possible signs of hemorrhage with blot, dot, or flame-shaped” / ”possible presence of microaneurysm, cotton-wool spot, or hard exudate” / ”arteriolar narrowing” / ”vascular wall changes” / ”optic disk edema”
intraretinal microvascular abnormalities	”shunt vessels and appear as abnormal branching or dilation of existing blood vessels (capillaries) within the retina” / ”deeper in the retina than neovascularization, has blurrier edges, is more of a burgundy than a red, does not appear on the optic disc” / ”vascular loops confined within the retina”
red small dots	”microaneurysms”
a disease	”no healthy” / ”lesions”
normal	”healthy” / ”no findings” / ”no lesion signs”

Expert Knowledge Dictionary

Vision-Language pre-training

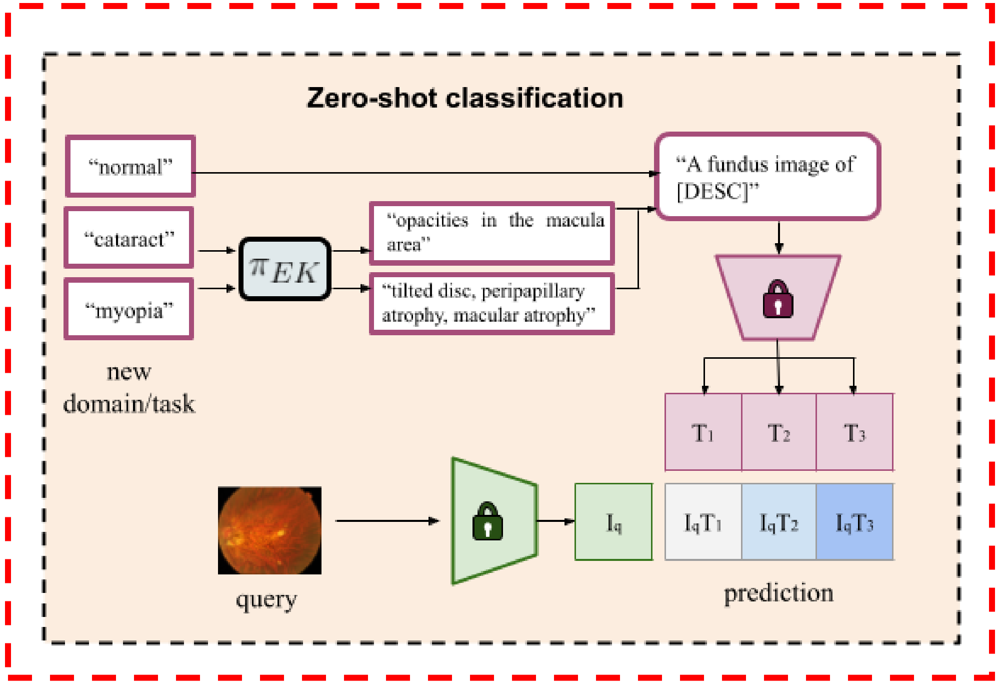


$$\mathcal{L}_{i2t}(\theta, \phi, \tau|\mathcal{B}) = -\sum_{i\in X_B} \frac{1}{|P_{\mathcal{T}_B}(i)|} \sum_{i'\in P_{\mathcal{T}_B}(i)} \log \frac{\exp(\tau \mathbf{u}_i^T \mathbf{v}_{i'})}{\sum_{j\in \mathcal{T}_B} \exp(\tau \mathbf{u}_i^T \mathbf{v}_j)} \tag{1}$$

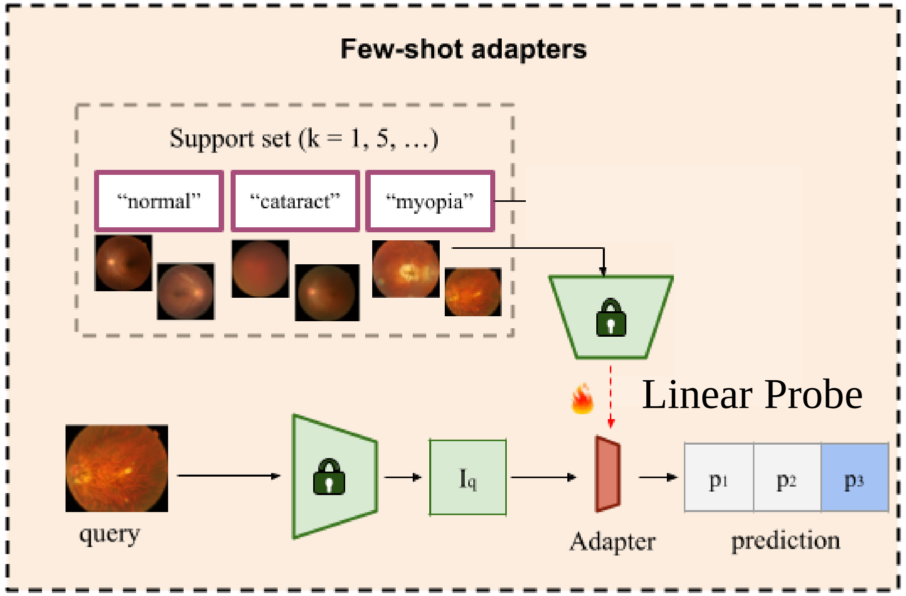
$$\mathcal{L}_{t2i}(\theta, \phi, \tau|\mathcal{B}) = -\sum_{j\in \mathcal{T}_B} \frac{1}{|P_{X_B}(j)|} \sum_{j'\in P_{X_B}(j)} \log \frac{\exp(\tau \mathbf{u}_{j'}^T \mathbf{v}_j)}{\sum_{i\in X_B} \exp(\tau \mathbf{u}_i^T \mathbf{v}_j)} \tag{2}$$

$$\min_{\theta,\phi,\tau} \sum_{\mathcal{B}} \mathcal{L}_{i2t}(\theta, \phi, \tau|\mathcal{B}) + \mathcal{L}_{t2i}(\theta, \phi, \tau|\mathcal{B})$$

Generalization and Transferability



Generalization



Transferability

Experimental setting

- **How to evaluate a vision-language foundation model?**
  - Known tasks under **domain shift**.
  - New tasks on **unseen categories**.
- **What do we want from a foundation model?**
  - **Generalization**: predictions without examples - zero-shot with prompts.
  - **Transferability**: adapting for new tasks/domains (Linear Probe).
    - Low-data regime (few shots).
    - Large-data regime (increasing data percentages).
- **What baselines to use?**
  - Other vision-language models: Vision (**CLIP**), or Generalists (**BiomedCLIP**).
  - Other pre-training strategies: adapting task-specific models (**TSM**), unsupervised pre-training (**SimCLR**).
  - Dataset-specific models (**Supervised**): **Fully-training** on the target dataset.

Dataset	#Images	Labels
Domain shift		
MESSIDOR	1448	noDR, mildDR, modDR, sevDR, prolDR.
FIVES	800	N, DR, G, ARMD.
REFUGE	1200	G, noG
Unseen categories		
20x3	60	N, RP, MHL
ODIR200x3	600	N, CAT, MYA

Results I: Generalization

1. Vision and Generalists models do not generalize to specialized domains.
2. EK prompts during training produces better pre-trained FMs.
3. EK prompts notably increses inference performance.
4. Large-scale pre-training boost performance on underepresented tasks.

Method	MESSIDOR DR grading (ACA/κ)	FIVES Diseases (ACA)	REFUGE Glaucoma (AUC)
Prior literature			
$DR_{graduate}$ (Araújo et al., 2020)	0.596/0.710	-	-
AST (Galdran et al., 2020)	0.634/0.797	-	-
$AIROGS_{lb}$ (de Vente et al., 2023)	-	-	[0.88, 0.94]
Task-specific models (TSMs)			
TSM_{DR}	0.550/**0.772**	-	-
$TSM_{Diseases}$	-	0.381	-
$TSM_{Glaucoma}$	-	-	0.904
VLP - inference w/ π_{naive}			
CLIP	0.237/0.140	0.250	0.470
BiomedCLIP	0.224/0.201	0.416	0.540
FLAIR-π_{naive}	0.545/0.662	0.732	0.899
FLAIR-π_{EK}	0.602/0.711	0.719	0.918
VLP - inference w/ π_{EK}			
CLIP	0.200/0.000	0.256	0.433
BiomedCLIP	0.207/0.188	0.415	0.624
FLAIR-π_{naive}	0.442/0.694	**0.744**	0.871
FLAIR-π_{EK}	**0.604/0.772**	0.735	**0.920**

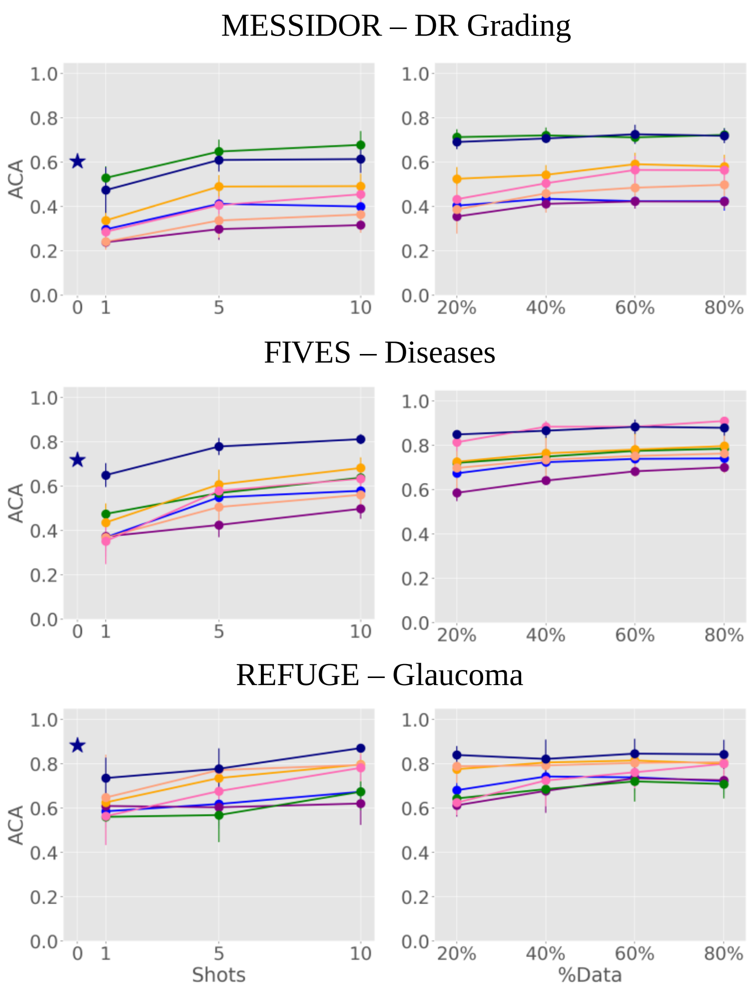
Domain Shift

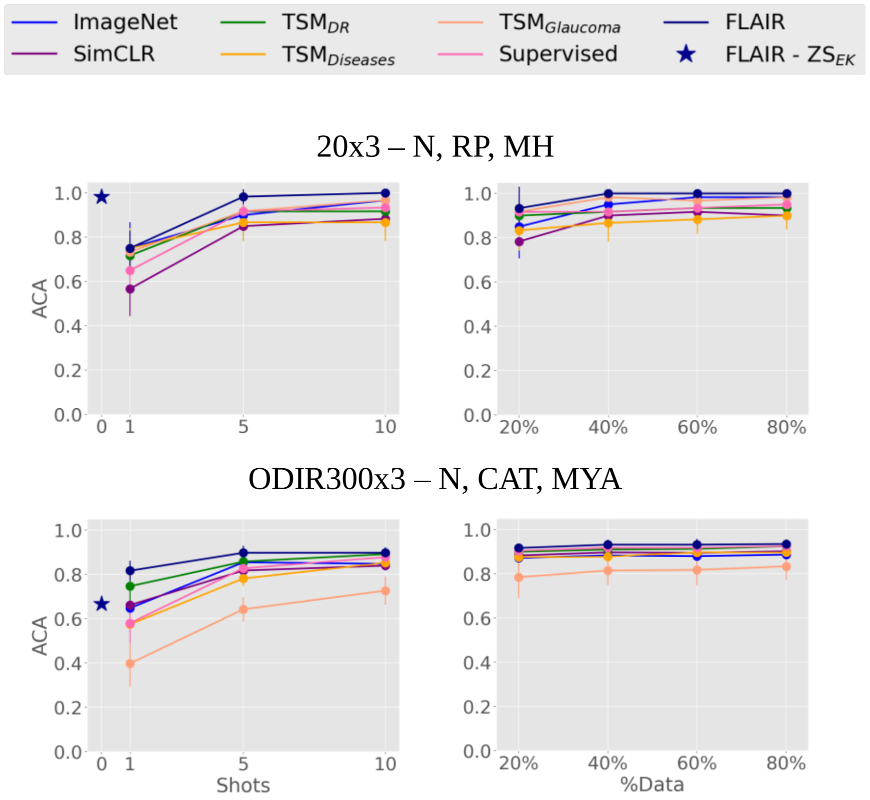
Method	20x3 N	20x3 RP	20x3 MHL	20x3 Avg.	ODIR200x3 N	ODIR200x3 CAT	ODIR200x3 MYA	ODIR200x3 Avg.
Anomaly Detection Inference (i.e. "normal/disease")								
CLIP	1.000	0.200		0.600	0.770	0.412		0.591
BiomedCLIP	0.950	0.125		0.538	0.800	0.770		**0.785**
FLAIR-π_{naive}	0.900	0.200		0.550	1.000	0.102		0.551
FLAIR-π_{EK}	0.850	0.775		**0.812**	0.985	0.350		0.668
Inference with Naive Prompts - π_{naive} (e.g. "cataract")								
CLIP	0.100	1.000	0.000	0.367	0.770	0.495	0.070	0.445
BiomedCLIP	0.900	0.950	0.400	0.750	0.765	0.920	0.495	**0.727**
FLAIR-π_{naive}	0.950	0.650	0.100	0.567	0.990	0.340	0.010	0.447
FLAIR-π_{EK}	0.950	0.600	1.000	**0.850**	0.990	0.455	0.005	0.483
Inference with Expert Knowledge Prompts - π_{EK} (e.g. "opacity in the macular area")								
CLIP	1.000	0.000	0.000	0.333	0.290	0.195	0.955	0.480
BiomedCLIP	0.400	0.800	0.650	0.617	0.125	0.695	0.930	0.583
FLAIR-π_{naive}	1.000	0.900	0.050	0.650	0.405	0.015	0.990	0.470
FLAIR-π_{EK}	1.000	0.950	1.000	**0.983**	0.760	0.765	0.475	**0.667**

Novel Classes

Results II: Transferability

5. Task-specific models (TSMs) do not transfer well to other tasks.
6. FLAIR transferability is robust to new tasks and domains.
7. FLAIR requires only few shots to outperform fully-trained dataset-specific models (Supervised).





Domain Shift

Novel Classes

Adaptation
Generalization

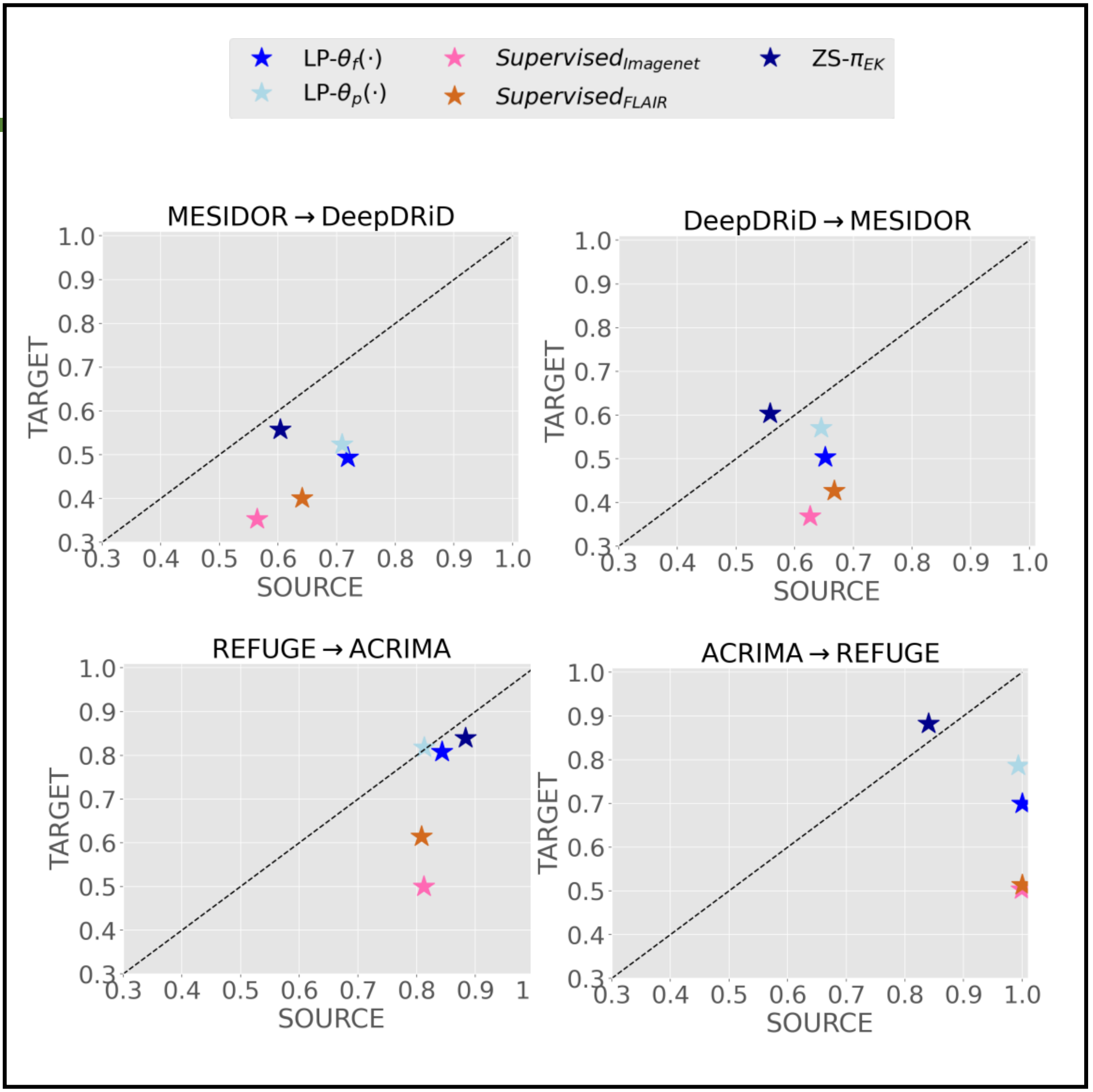
In-Depth Analysis

8 ZS or LP from FLAIR provide better ID-OOD performance.

“I don’t need foundation model, I have gathered a **large enough dataset**, and **my model performs much better!**”

Great! ☺ BUT…

- × Did you check on **OOD data**?*
- × **How much data** did you required?
- × How long **data-collection** took?
- × What if you want to predict **new categories**?
- × How **computationally expensive** is your training?



* Recommended read: Fine-tuning can distort pre-trained features and underperform OOD, ICLR 2022

What is a *foundation model*?

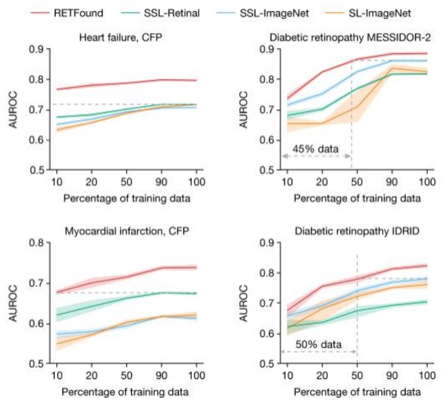
Article | Open access | Published: 13 Septembe

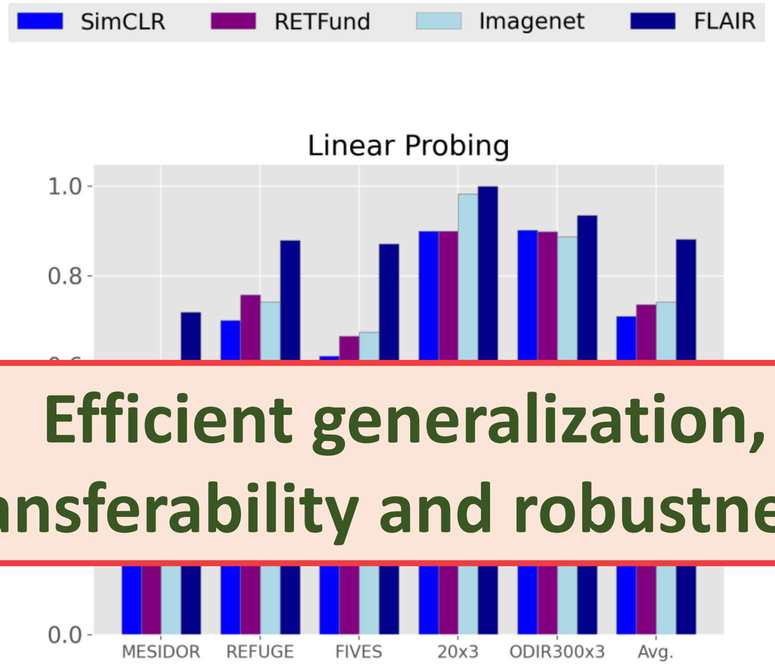
A foundation model for g detection from retinal im

Yukun Zhou, Mark A. Chia, Siegfried K. Wagn Struyven, Timing Liu, Moucheng Xu, Mateo G. L Eye & Vision Consortium, Andre Altmann, Aaro Alexander & Pearse A. Keane

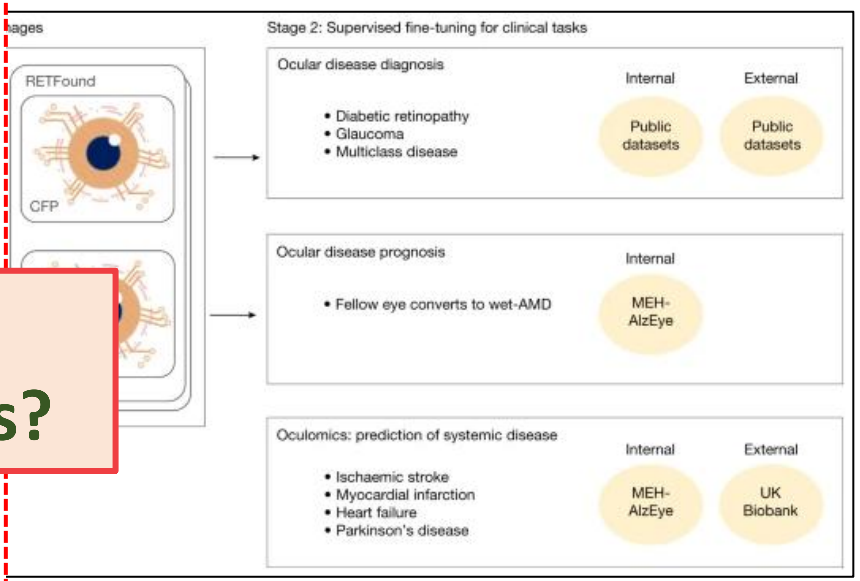
Nature **622**, 156–163 (2023) | Cite this artic

61k Accesses | **896** Altmetric | Metrics





0K fundus images!



Efficient generalization, transferability and robustness?

Take-home messages

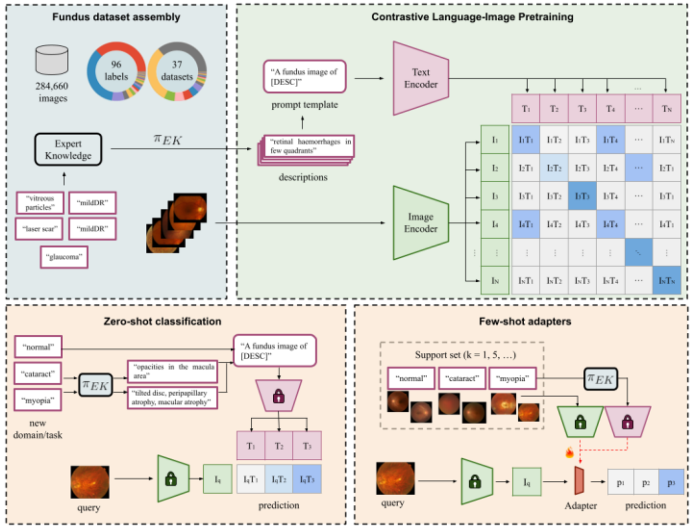
- Pretrain-and-adapt: A **paradigm change** for medical image analysis.
- **Vision-language pre-training** provides powerful foundation models.
- **Don't trust generalist models**.
- You dont have large text-supervised datasets? Try **encoding expert knowledge**!
- Potential: **Linear Probing** from FLAIR **outperforms fully-trained dataset-specific models** even for **unknown diseases**.

A Foundation LAnguage Image of the Retina (FLAIR): Encoding expert knowledge in text supervision

Julio Silva-Rodríguez[1], Hadi Chakor[2], Riadh Kobbi[2], Jose Dolz[1] and Ismail Ben Ayed[1]

ETS Montreal[1], DIAGNOS Inc.[2]

https://github.com/jusiro/FLAIR



Fonds de recherche Nature et technologies Québec

The work of J. Silva-Rodríguez was partially funded by the FRQ under the PBEEE merit scholarship

Le génie pour l'industrie

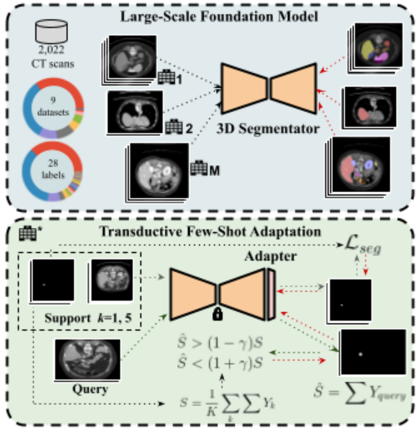
Workshop on Medical Imaging with Deep Learning
ETS Montreal

Towards foundation models and few-shot parameter efficient fine-tuning for volumetric organ segmentation

Julio Silva-Rodríguez, Jose Dolz and Ismail Ben Ayed

ETS Montreal

Best Paper Award on MICCAI MedAGI: 1st Workshop on Foundation Models



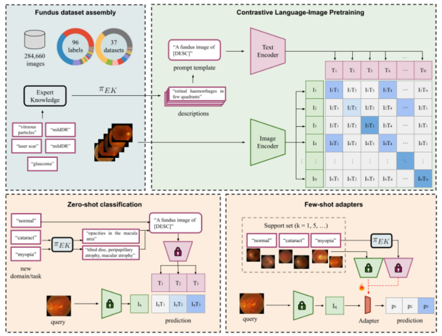
CT VOLUMES

A Foundation LAnguage Image of the Retina (FLAIR): Encoding expert knowledge in text supervision

Julio Silva-Rodríguez[1], Hadi Chakor[2], Riadh Kobbi[2], Jose Dolz[1] and Ismail Ben Ayed[1]

ETS Montreal[1], DIAGNOS Inc. [2]

Under Review



FUNDUS