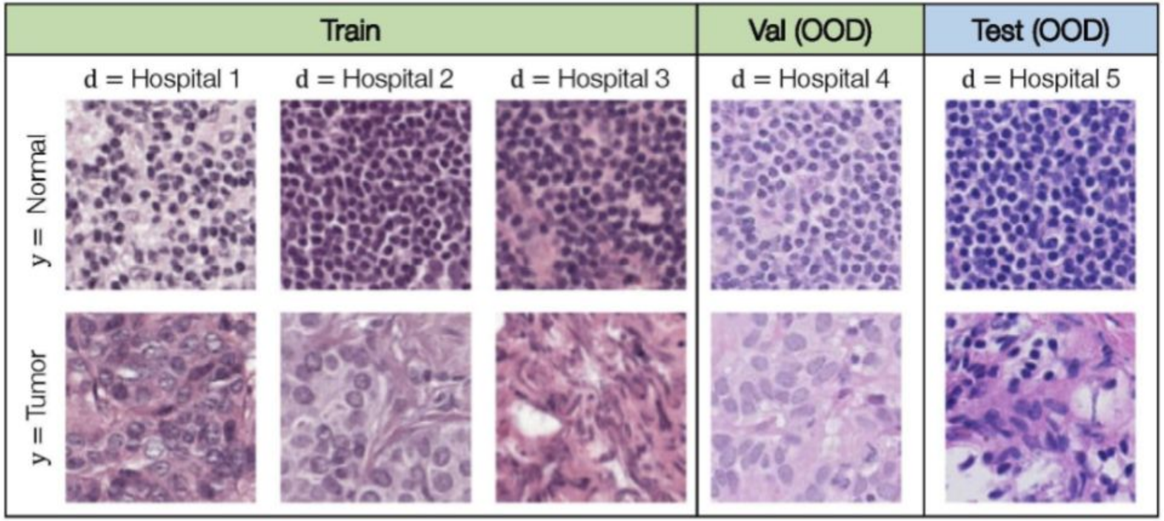
Learning Invariant Representations with a Nonparametric Nadaraya-Watson (NW) Head

AQ Wang, MR Sabuncu *et. al.*, Cornell University

NeurIPS 2023

Domain generalization and Robustness

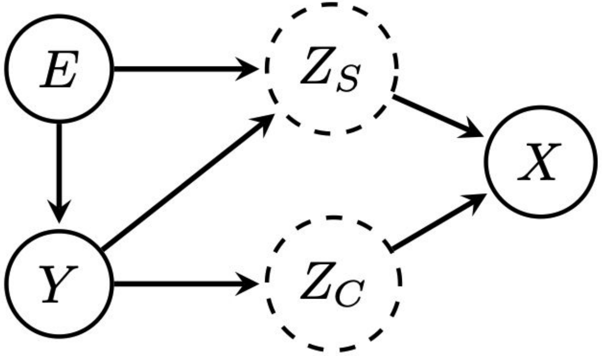


- Varying features: color, staining or markings
- Invariant features: structure or dots

Literature on domain generalization

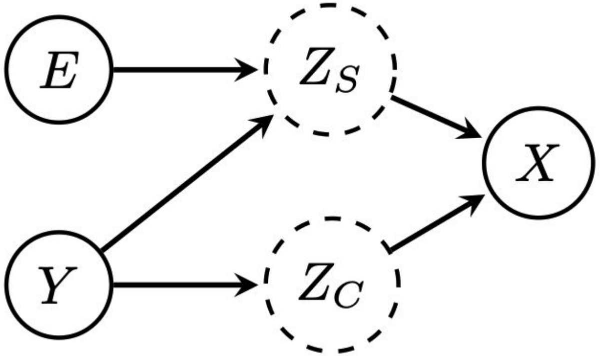
- Data augmentation or synthetically generation
  - LISA (mixup-style)
- Align features
  - layer: CORAL
  - distribution: distance metric or adversarial loss
  - gradients : FISH
- Invariant representations
  - employ causal perspective to define appropriate constraints
    - Invariant causal prediction (ICP)
    - Invariant Risk Minimization (IRM)

Causal DAG (Directed Acyclic Graph)



- E: environment
- (X, Y): image and class label
- Z_S : style latent representation (spurious)
- Z_C : content latent representation (causal)

d-separation of Causal DAG



- Invariance constraint by condition of Z_C, such that Y is not influenced by E

$$Y \perp\!\!\!\perp E \mid Z_C \iff \underbrace{P_e(Y \mid Z_C) = P_{e'}(Y \mid Z_C)}_{\text{invariance constraint}} \quad \forall e, e' \in E$$

Learning Invariant Representation across Environments

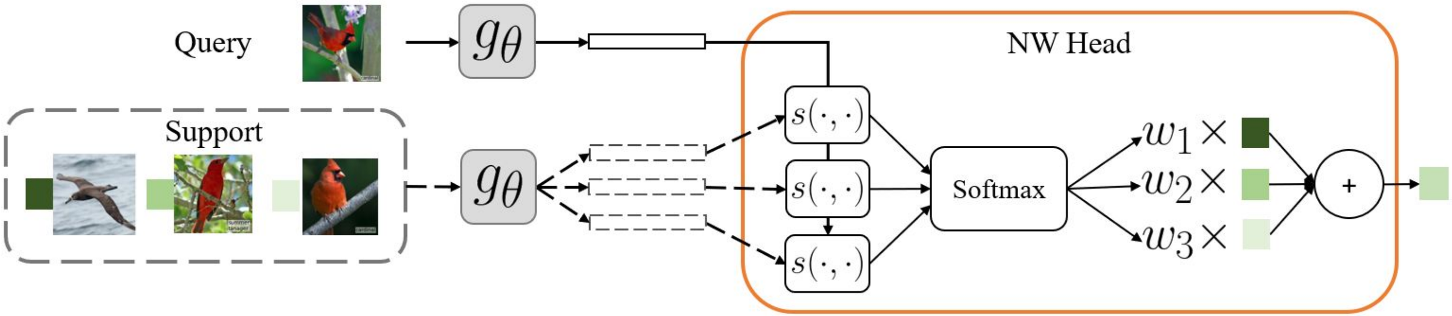
- Goal is to learn an estimator $\hat{P}(Y \mid \varphi(X))$

$$\underset{\varphi}{\operatorname{argmax}} \sum_{e} \hat{P}_e(Y \mid \varphi(X))$$

$$\text{s.t. } \hat{P}_e(Y \mid \varphi(X)) = \hat{P}_{e'}(Y \mid \varphi(X)) \; \forall e, e' \in E.$$

- Propose a novel estimator using **Nadaraya-Watson head**

Nadaraya-Watson (NW) Head



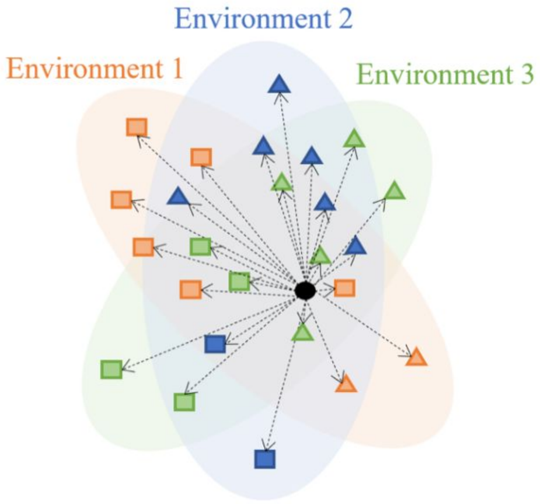
- NW prediction: $$f(x, \mathcal{S}) = \sum_{i=1}^{N_s} w(x, x_i)\vec{y}_i,$$

$$w(x, x_i) = \frac{\exp\{-\|\varphi(x) - \varphi(x_i)\|_2\}}{\sum_{j=1}^{N_s} \exp\{-\|(\varphi(x) - \varphi(x_j))\|_2\}}.$$

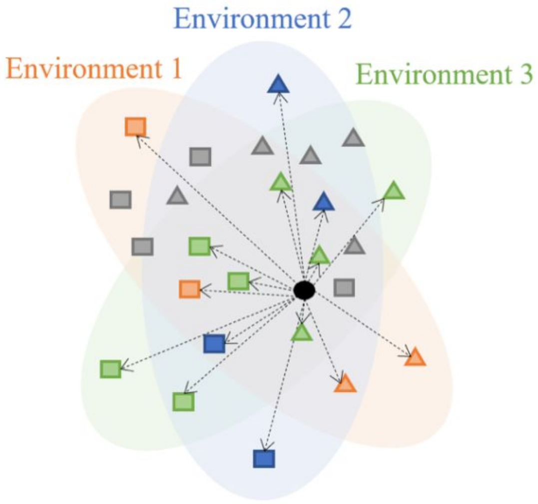
- Soft version of NN classifier

Method: NW Head for Invariant Prediction

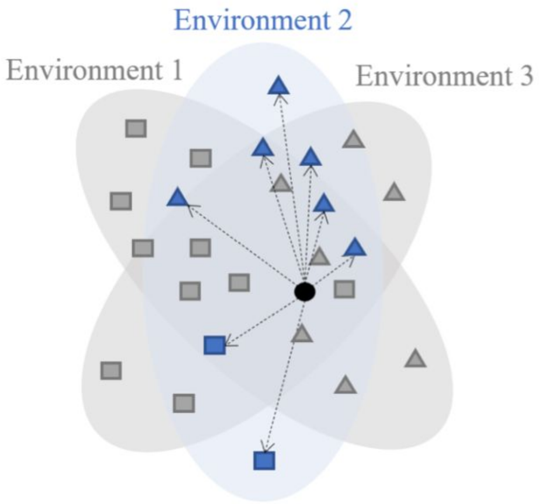
- Support set can be manipulated during training
- Restrict the support set to from a single environment



Unconditional support set



Balanced support set



Conditioning support set on an environment

NW Head for Invariant Prediction

- Objective is to constrained maximum likelihood:

$$\operatorname*{argmax}_{\varphi} \sum_{e} \hat{P}_e(Y \mid \varphi(X))$$

$$\text{s.t. } \hat{P}_e(Y \mid \varphi(X)) = \hat{P}_{e'}(Y \mid \varphi(X)) \ \forall e, e' \in E.$$

- Replace NW head with conditional support sets:

$$\operatorname*{argmin}_{\varphi} \sum_{i=1}^{N} L(f_\varphi(x_i, \mathcal{S}^B_{e_i}), y_i)$$

$$\text{s.t. } f_\varphi(x_i, \mathcal{S}^B_e) = f_\varphi(x_i, \mathcal{S}^B_{e'}), \ \ \forall i \in \{1, ..., N\}, \forall e, e' \in E,$$

NW Head for Invariant Prediction

1. Explicit via Lagrangian:

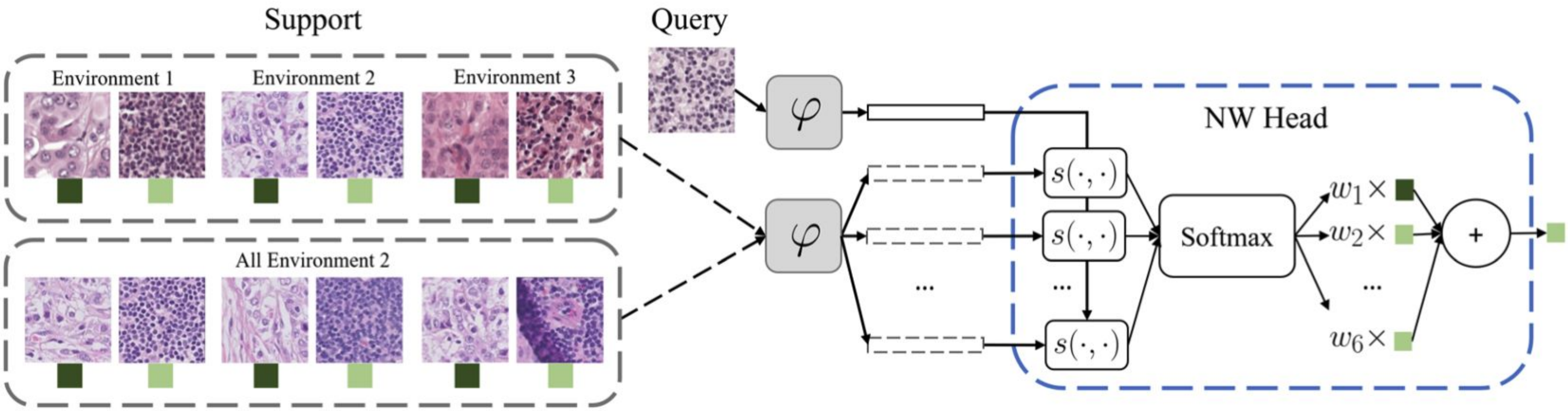
$$\underset{\varphi}{\operatorname{argmin}} \sum_{i=1}^{N} L(f_\varphi(x_i, \mathcal{S}_{e_i}^B), y_i) + \lambda \sum_{e,e' \in E} \sum_{i=1}^{N} \| f_\varphi(x_i, \mathcal{S}_e^B) - f_\varphi(x_i, \mathcal{S}_{e'}^B) \|_2^2.$$

2. Implicit via sampling environment-specific support sets during training

$$\underset{\varphi}{\operatorname{argmin}} \sum_{e \in E} \sum_{i=1}^{N} L(f_\varphi(x_i, \mathcal{S}_e^B), y_i).$$

No invariance regularizer and hyperparameter to tune!!

NW Head for Invariant Prediction



Experimental settings

- Datasets

Dataset	*# Classes*	*Env*	*# Envs*	*Architecture*	*Metric*
Camelyon-17	2	Hospital	3	DenseNet-121	Average acc.
ISIC	2	Hospital	3	ResNet-50	F1-score
FMoW	62	Region	5	DenseNet-121	Worst-region acc.

- Inference mode with different support sets
    - **Random:** sample randomly with each class represented k times
    - **Full:** entire balanced training set
    - **Ensemble:** based each environment
    - **Cluster:** k cluster centroids per class
    - **Probe**: add linear classifier

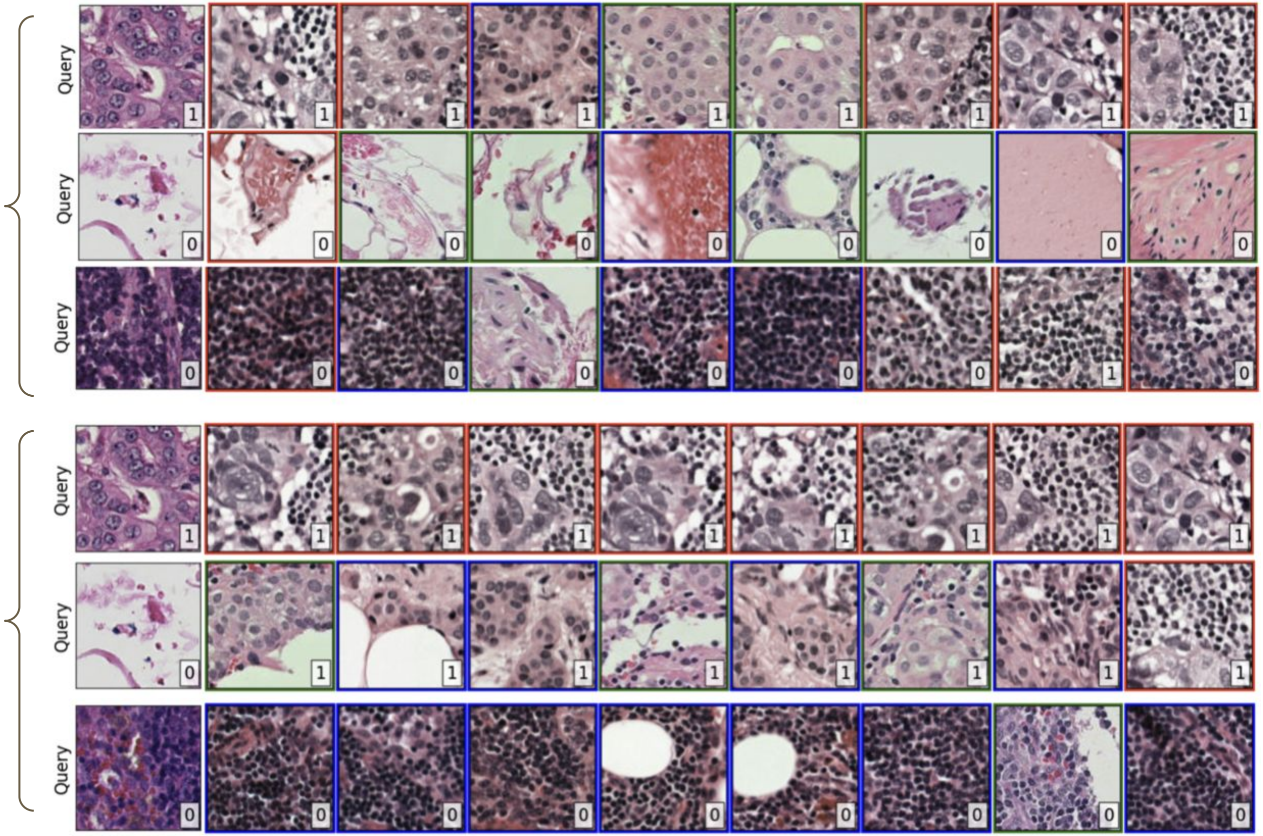
NW head generalizes to new distributions

	Algorithm	*Camelyon-17*	*ISIC*	*FMoW*
	ERM [52]	$70.3_{\pm 6.4}$	$58.2_{\pm 2.9}$	$32.6_{\pm 1.6}$
	IRM [1]	$70.9_{\pm 6.8}$	$57.9_{\pm 1.0}$	$31.3_{\pm 1.2}$
	CORAL [48]	$72.4_{\pm 4.4}$	$59.1_{\pm 2.2}$	$31.7_{\pm 1.0}$
	Fish [46]	$74.7_{\pm 7e\text{-}2}$	$64.4_{\pm 1.7}$	$34.6_{\pm 0.0}$
	LISA [64]	$77.1_{\pm 6.5}$	$64.8_{\pm 2.3}$	$35.5_{\pm 1.8}$
	CLOvE [54]	$\underline{79.9}_{\pm 3.9}$	$66.2_{\pm 2.2}$	$\mathbf{40.1}_{\pm 0.6}$
w/o conditional support	NW^{B}, Random	$71.7_{\pm 5.3}$	$56.7_{\pm 1.4}$	$31.1_{\pm 0.8}$
	NW^{B}, Full	$72.0_{\pm 6.7}$	$61.9_{\pm 3.5}$	$31.6_{\pm 0.9}$
	NW^{B}, Cluster	$70.6_{\pm 6.9}$	$61.4_{\pm 2.3}$	$31.3_{\pm 0.9}$
	NW^{B}, Ensemble	$71.9_{\pm 6.0}$	$63.9_{\pm 3.8}$	$32.2_{\pm 1.0}$
	NW^{B}, Probe	$69.2_{\pm 7.4}$	$59.7_{\pm 2.5}$	$29.9_{\pm 1.5}$
w/ conditional support	$\mathrm{NW}^{\mathrm{B}}_{\mathrm{e}}$, Random	$74.8_{\pm 8.4}$ / $75.3_{\pm 3.2}$	$57.5_{\pm 1.9}$ / $55.0_{\pm 0.9}$	$31.2_{\pm 0.7}$ / $30.9_{\pm 0.5}$
	$\mathrm{NW}^{\mathrm{B}}_{\mathrm{e}}$, Full	$\mathbf{80.0}_{\pm 2.7}$ / $79.7_{\pm 1.9}$	$69.6_{\pm 2.3}$ / $70.0_{\pm 1.0}$	$35.0_{\pm 0.7}$ / $34.6_{\pm 0.4}$
	$\mathrm{NW}^{\mathrm{B}}_{\mathrm{e}}$, Cluster	$78.6_{\pm 2.5}$ / $79.0_{\pm 1.4}$	$\mathbf{71.1}_{\pm 1.7}$ / $\underline{71.0}_{\pm 1.0}$	$33.9_{\pm 0.6}$ / $34.0_{\pm 0.3}$
	$\mathrm{NW}^{\mathrm{B}}_{\mathrm{e}}$, Ensemble	$79.5_{\pm 2.6}$ / $79.6_{\pm 1.9}$	$69.5_{\pm 2.2}$ / $69.8_{\pm 0.8}$	$37.8_{\pm 0.9}$ / $\underline{38.2}_{\pm 0.4}$
	$\mathrm{NW}^{\mathrm{B}}_{\mathrm{e}}$, Probe	$75.3_{\pm 7.3}$ / $75.8_{\pm 8.3}$	$61.4_{\pm 3.1}$ / $63.4_{\pm 2.8}$	$33.9_{\pm 1.5}$ / $32.7_{\pm 1.4}$

Interpretability of NW Head

w/ conditional support

w/o conditional support



Histogram of nearest neighbours

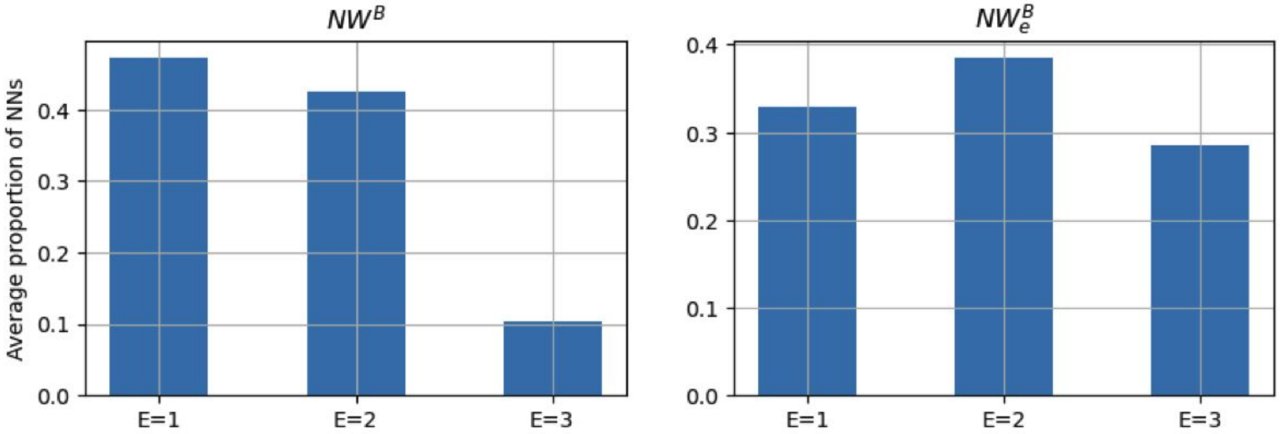


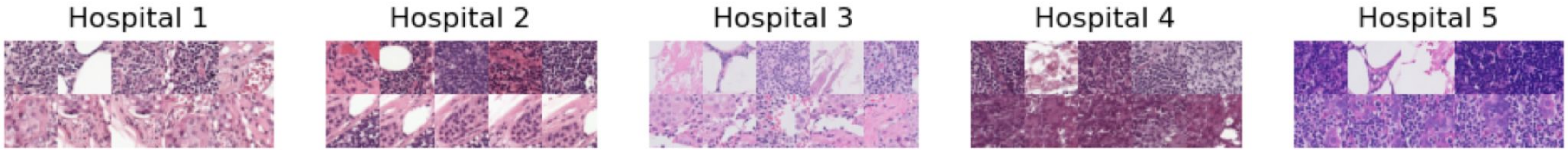
Figure 9: Normalized histogram of the environments from which the top 20 nearest neighbors originate in the training dataset for Camelyon-17, averaged over all queries in the test set. We observe a more balanced proportion for $\mathrm{NW}_\mathrm{e}^\mathrm{B}$, indicating that the model relies more evenly across all 3 environments to make its prediction, and further suggesting that representations are more invariant than NW^B.

Conclusion

- Motivated domain generalization from causal perspective
- Propose a nonparametric invariant representation
- NW head enables interpretability by nearest neighbors
- Need support sets for inference
- Computationally expensive
- Future works
    - Regression task
    - Adaption to test domain given additional information (edge $E \rightarrow Y$)

Extra slides

Dataset



(a) Camelyon-17

(b) ISIC

(c) FMoW