Mixture of Geographical Experts: Disentangling Earth ❄ EurIPS

Moien Rangzan[1,2], Gregory Duveiller[1], Maha Shadaydeh[2], Markus Reichstein[1], Joachim Denzler[2]

[1]Max Planck Institute for Biogeochemistry, [2]Computer Vision Group, Friedrich Schiller University, Jena, Germany

Background and Motivation

Background

- Geographical confounders **break the covariate-shift assumption** in EO tasks, causing Domain-Invariant methods to fail to generalize.
- Domain Generalization (DG) methods rely on **manually defined domains,** with **hard boundaries,** and **need expert knowledge** for each task.

Motivation, Following the First Law of Geography*:

- **Location** can be used to learn soft overlapping domains from the data itself.
- It is also a **proxy** to condition on unobserved confounders.

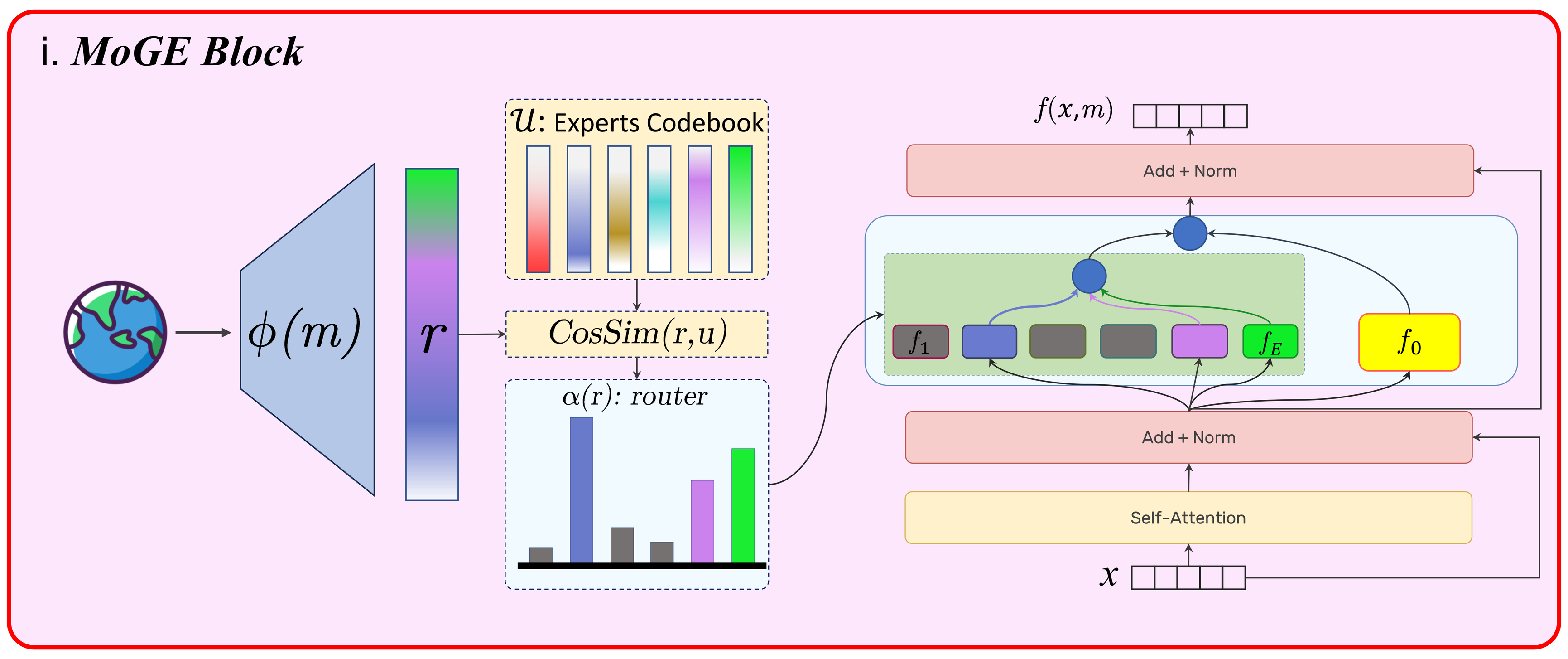
*According to Tobler: "Everything is related to everything else, but near things are more related than distant things."

Contributions

In this work we introduce MoGE, a geo-routed Mixture-of-Experts that:

- **Learns concept-consistent geographic domains** from location metadata, removing the need for hand-crafted domain definitions.
- **Disentangles global invariance from spatial variation**, yielding strong improvements over both DG and domain-specific (DS) baselines.
- Plugs into any Transformer layer **without altering pretrained weights.**
- **Provides interpretability** through explicit expert routing maps.

MoGE: Mixture of Geographical Experts



i. MoGE replaces the FFN in a Transformer encoder block with a geo-routed MoE composed of one shared invariant expert and several specialized experts selected through metadata-driven routing.

$$f(x,m) = \underbrace{\gamma\, f_0(x)}_{\text{invariant}} + \underbrace{(1-\gamma)\sum_{k=1}^{E} \alpha_k(\phi(m))\, f_k(x)}_{\text{specialized}}$$

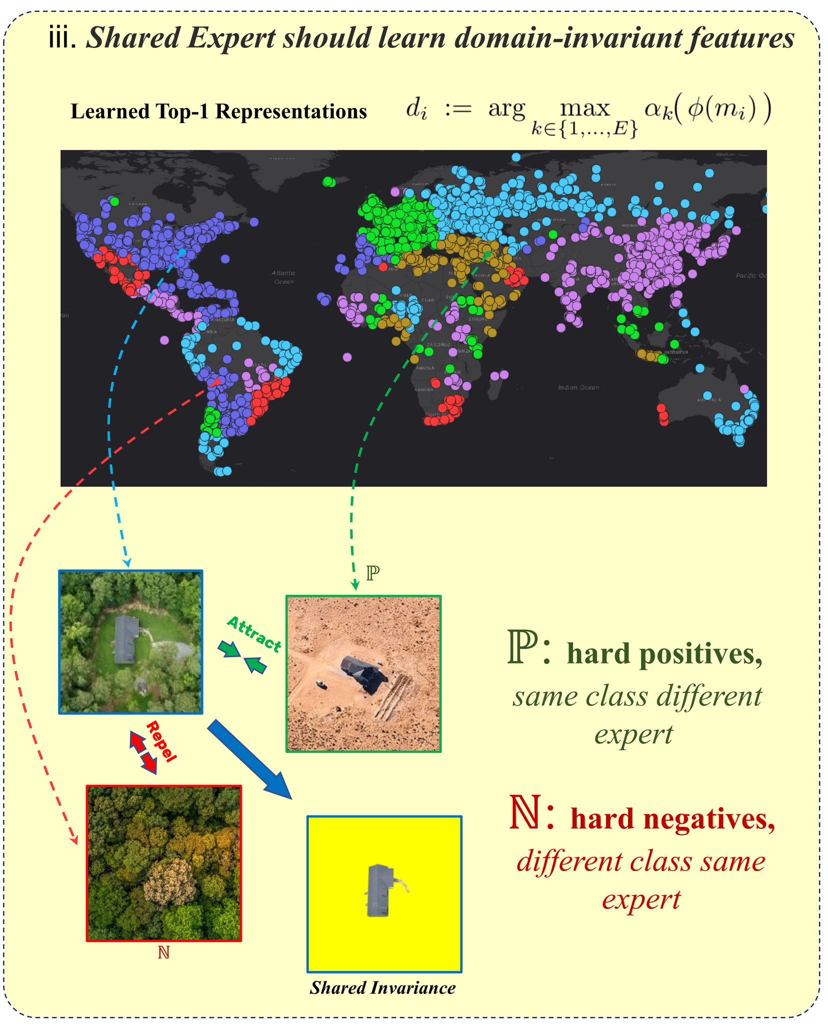
ii. To enforce **each expert to learn distinct and orthogonal representations**, we apply a CKA-based diversity loss that penalizes similarity between their outputs.

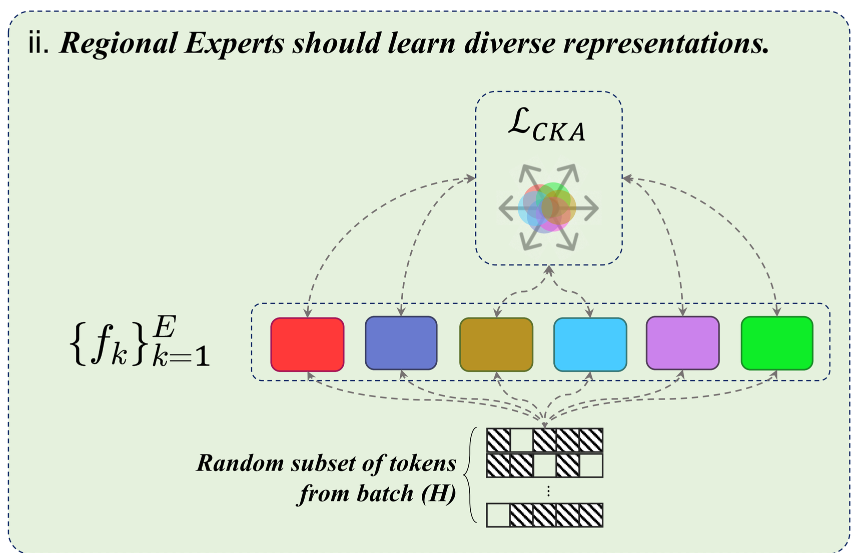
$$\mathcal{L}_{\text{CKA}} = \frac{2}{E(E-1)} \sum_{1 \le i < j \le E} \text{CKA}_{\text{lin}}(H_i, H_j) \qquad H_k = f_k(H) \text{ for } k = 1, \dots, E$$

iii. To have an invariant fallback, we force the shared expert features (f_0) to **be invariant across domains that were discovered** by our location encoding gating function.

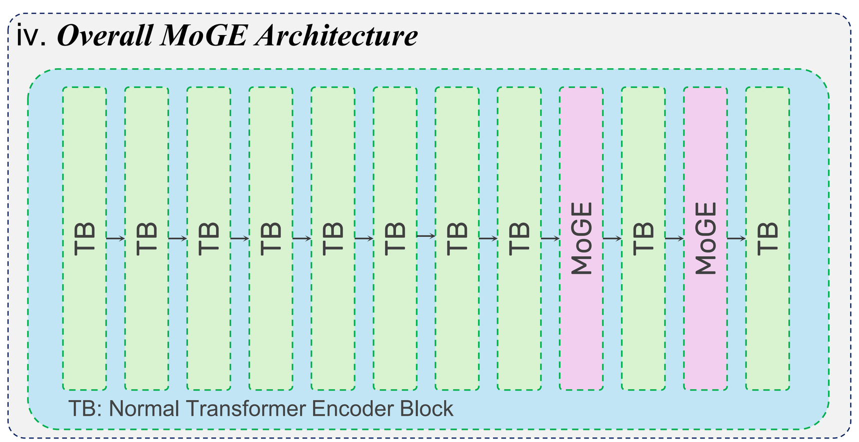
$$\mathcal{L}_{\text{SupCon}} = \frac{-1}{B} \sum_{i \in \mathcal{A}} \left[\log \sum_{p \in \mathbb{P}_i} e^{\langle q_i, p \rangle / \tau_{\text{sup}}} - \log \sum_{v \in \mathbb{P}_i \cup \mathbb{N}_i} e^{\langle q_i, v \rangle / \tau_{\text{sup}}} \right]$$

$$q_i := f_0(x_i)/\|f_0(x_i)\|_2, \quad \mathcal{A} = \{ i \text{ in batch} : |\mathbb{P}_i| \ge 1 \}, \quad B := |\mathcal{A}|$$





iv. MoGE blocks can readily replace any Transformer block. In our design, we use 10 Transformer blocks and 2 MoGE blocks in a ViT-Tiny architecture.

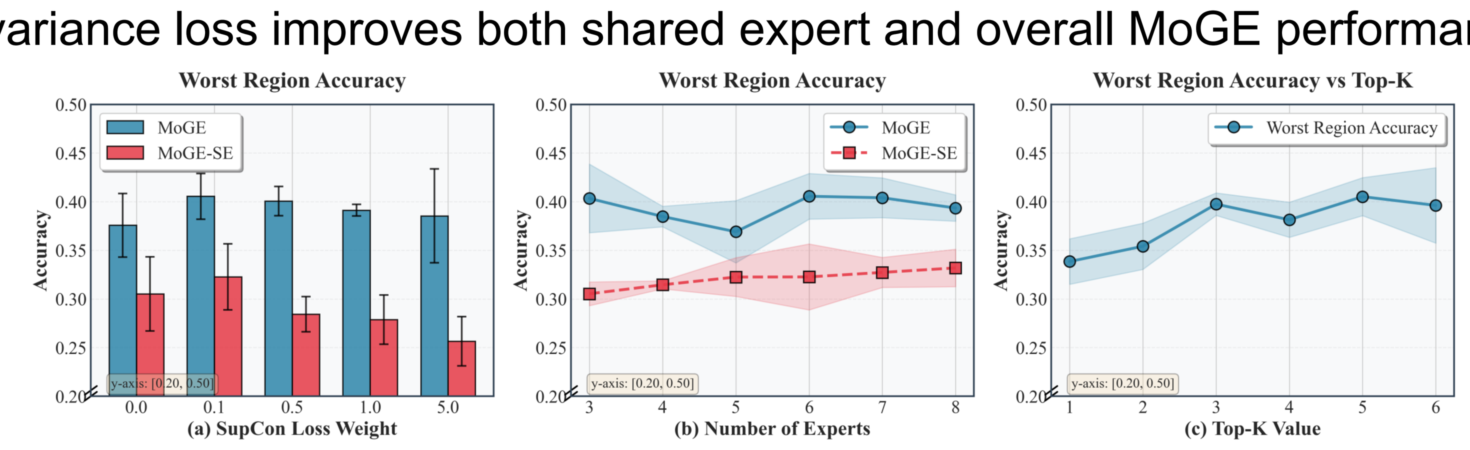


- The total MoGE loss is given below, where $\mathcal{L}_{\text{aux}}$ serves as a load-balancing term that encourages an even distribution of routing assignments.

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{aux}}\, \mathcal{L}_{\text{aux}} + \lambda_{\text{CKA}}\, \mathcal{L}_{\text{CKA}} + \lambda_{\text{SupCon}}\, \mathcal{L}_{\text{SupCon}}$$

Ablation Studies

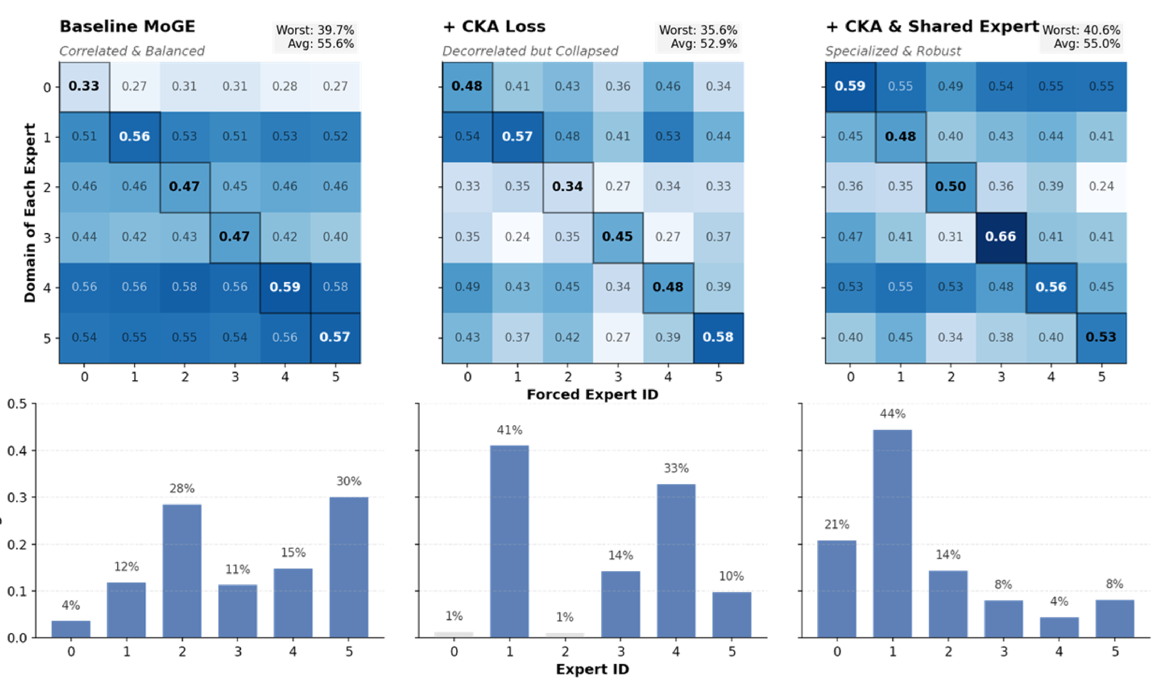
- **Ablation 1: Effect of Invariance loss, number of experts, and top-k routing.** Using 6 experts and top-3 routing yields the best results. Furthermore, using invariance loss improves both shared expert and overall MoGE performance.



- **Ablation 2: Effect of Diversity loss and shared expert.** Without the CKA loss, experts remain highly correlated. Adding CKA without the shared expert decorrelates them but leads to expert collapse and reduced performance. Using both together avoids collapse and gives the best results.

In the matrix:
Each **row** corresponds to an **expert's domain**, defined as the samples for which that expert is the top-1 route.

Each **column** shows the performance of an expert when all the samples from a given domain are forced through it.



Experimental Results

- MoGE achieves the **best** results on **FMoW** and **iWildCam** datasets.
- Generalizes to the held-out FMoW-LAO region, and improves performance.
- The **shared expert alone (MoGE-SE)** also outperforms DG methods.

(a) FMoW (with 62 classes, 5 regions as domains)

	Method	Worst Acc. (%)	Overall Acc. (%)
Domain Specific	MoGE	$\mathbf{40.56} \pm \mathbf{2.35}$	$\mathbf{54.98} \pm \mathbf{0.38}$
	ERM+LE	$\underline{35.83 \pm 0.51}$	$\underline{53.23 \pm 0.72}$
	D3G	33.38 ± 0.64	51.51 ± 0.19
	D3G+WRAP	34.60 ± 1.27	50.76 ± 0.12
	Per Region Models	26.75 ± 1.26	49.72 ± 0.26
Domain Invariant	MoGE-SE	$\mathbf{33.49} \pm \mathbf{1.78}$	$\underline{49.19 \pm 0.26}$
	ERM	32.43 ± 1.67	$\mathbf{53.69} \pm \mathbf{0.37}$
	GroupDRO	30.70 ± 0.80	49.06 ± 0.37
	IRM	25.85 ± 0.93	42.71 ± 0.41
	Fish	$\underline{33.08 \pm 0.29}$	44.99 ± 0.72
	Mixup	32.88 ± 0.63	47.25 ± 0.62
	VREx	32.48 ± 1.28	46.83 ± 0.90
	RDM	32.66 ± 0.60	47.70 ± 0.17

(c) iWildCam 182 classes

Method	F1-ood	F1-id
MoGE	N/A	**0.495**
MoGE-SE	**0.320**	N/A
ERM	0.311	0.473
GroupDRO	0.219	0.420
IRM	0.161	0.253
DeepCoral	0.290	0.472

(b) FMoW-LAO

Method	Asia Acc. (%)
MoGE+SatCLIP	**38.54**
ERM	36.27
Fish	35.25
VREx	37.03

With a pretrained SatCLIP encoder, MoGE can **generalize to the unseen Asia region**.

Discussion

- MoGE brings together the strengths of both DG and DS methods and **delivers significant improvements on both fronts**, demonstrating that learning domains directly from data can outperform handcrafted domain definitions.
- Beyond performance, **MoGE also offers clearer interpretability** compared to a naïve use of location as an input.
- Future work includes **gate-finetuning** for unseen regions and routing **distillation** to enable image-based routing.

MAX PLANCK INSTITUTE FOR BIOGEOCHEMISTRY

IMPRS gBGC

Paper and (Code soon!)

ellis

Computer Vision Group

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA