Mixture of Geographical Experts: Disentangling Earth

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

Joachim Denzler[1]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Max Planck Institute for Biogeochemistry, Jena, Germany
{mrangzan, gduveiller, mreichstein}@bgc-jena.mpg.de
{maha.shadaydeh, joachim.denzler}@uni-jena.de

Abstract

Most domain generalization techniques assume there exists a stable predictive relationship from input features to labels across domains, an assumption that breaks in many Earth observation tasks, where stable signals are weak or geographically confined. We introduce a sparse, geo-routed Mixture of Geographical Experts (MoGE) that *explicitly disentangles* global invariance from spatial variation. A shared invariant expert captures features that hold everywhere, while metadata-driven routing activates a subset of geo-specialized experts forced to learn region-specific cues. This separation lets experts self-organize into continuous, concept-consistent regions, discovering domains rather than handcrafting them, while the invariant path remains robust across space. MoGE's factorization yields strong performance on generalization benchmarks.

1 Introduction

Building globally reliable models from Earth observation data remains difficult. Distribution shifts across regions, sensors, seasons, and years break systems trained under i.i.d. assumptions [1, 2]. These shifts challenge the common assumption that a single model can hold everywhere [3].

The prevailing answer in machine learning has been to enforce *invariance* [4, 5]: learn a backbone [6, 7] or head [8, 9] that discards location-linked cues as spurious [10], so that a single valid classifier $P(Y|X)$ applies globally. This assumption, also known as the *covariate-shift assumption* [11], does not hold in many remote-sensing problems, where the stable signal is geographically confined. Enforcing global invariance prunes informative, region-specific evidence; in practice, such methods often underperform even Empirical Risk Minimization (ERM) on real-world tasks [2].

This shortcoming of invariant methods can be traced back to the concept shift [11]: causal mechanisms change across regions, where an attribute relation with the label depends on *where you are* or merely our observations are not enough to capture all the influencing attributes [12]. A cluster of trees may signal a natural forest in one ecoregion and a plantation in another; places of worship [2] vary dramatically across cultures. When semantics change with geography, insisting on a single $P(Y|X)$ is restrictive [13]. At the opposite extreme, handcrafting domains, such as continents, countries, or climate zones, and then training separate models for those domains reduces the effective data per model, blurs cross-boundary continuity, and makes performance hinge on arbitrary partitions.

However, geography is not arbitrary. Ecoregions, climate, economies, and cultural practices, broadly, vary smoothly in space [14]; nearby places tend to be more alike than distant ones [15, 16]. The recent emergence of location encoders (LE) [17–19] captures this continuity. Nonetheless, these encodings are typically injected into models as extra inputs or priors [20], leaving the models to internally learn this marginalization, reducing interpretability and generalization.

Many spatial problems can be modeled as a *family of overlapping conditionals* $\{P_d(Y|X)\}_{d=1}^{E}$ rather than a single global one [21]; local distributions that share structure yet differ subtly. However,

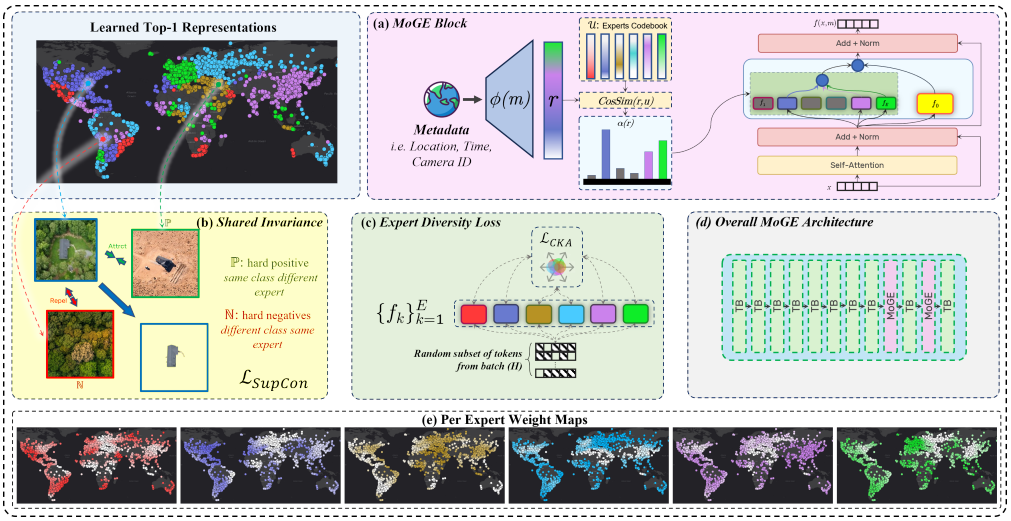

Figure 1: Overview of the MoGE architecture and discovered continuous domains. (a) Architecture of a MoGE block. (b) Invariance loss that uses discovered domains for hard positive/negative mining. (c) Expert diversity loss to force orthogonality among experts. (d) Placement of MoGE blocks inside a ViT backbone. (e) Example of discovered continuous domains on FMoW.

naïvely modeling each leads to redundant, highly correlated models. We address this by explicitly factorizing shared and local structure: learning a global expert to capture invariant patterns, and sparse, geo-specialized experts to model fine-grained variation where it matters.

We operationalize our view with the Mixture-of-Geographical-Experts, **MoGE**, a sparse Mixture-of-Experts (MoE) [22] transformer [23], with image-level routing, *driven by geo-metadata* instead of the image, selecting a sparse set of specialists while always retaining a shared [24, 25], invariant expert. Classical MoEs divide complex functions by routing on the input x, implicitly assuming that local changes in $p(y \mid x)$ correlate with x [26], which can also be addressed by increasing the model's complexity [27–29]. Our key departure is to *route on metadata* m (e.g., geo-coordinates, time, sensor), acknowledging that concept shift is often exogenous to data itself [12, 30, 31]. This induces a top-k convex mixture over E latent experts that model a family of *overlapping conditionals* $\{P_d(Y \mid X)\}_{d=1}^{E}$, where each d, unlike other domain adaptive methods [21, 32] corresponds not to hand-crafted "domains" but to **concept-consistent regions** that emerge from data. The result is a clean disentanglement: a global pathway that captures what is *invariant* everywhere, plus sparse specialists that preserve *variant*, place-specific cues, yielding models that capture spatial continuity.

2 Method Overview

2.1 Problem Setup and Method

Problem setup. We observe triples (x, m, y), where $x \in \mathcal{X}$ is an image, $m \in \mathcal{M}$ represents metadata (e.g., latitude/longitude, sensor ID, or timestamp), and $y \in \mathcal{Y}$ is the target label. The goal is to learn a predictor $f : \mathcal{X} \times \mathcal{M} \to \mathcal{Y}$ that generalizes across spatial and domain shifts.

Method. A metadata encoder $\phi : \mathcal{M} \to \mathbb{R}^{D_r}$ maps m to a context vector $r = \phi(m)$. MoGE swaps the feed-forward-networks (FFN) in selected ViT blocks [33] with a MoE [34]; let $f_k(x) \in \mathbb{R}^D$ denote the features produced by expert k.

We instantiate one *shared* expert $f_0 : \mathcal{X} \to \mathbb{R}^D$ (always active) and E *specialized* experts $\{f_k\}_{k=1}^{E}$ implemented as parallel FFN branches inside a transformer block.

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

where $\gamma \in [0, 1]$ is a learned scalar that balances invariant vs. specialized pathways, and α_k is the cosine router: each expert has a code $u_k \in \mathbb{R}^{D_r}$, and $\alpha_k(r) = \left[\mathrm{softmax}_{\text{top-}K}\left(s(r)/\tau_{\text{route}}\right)\right]_k$ where $s(r)$ is a vector with entries $s_k(r) = \langle r/\|r\|,\, u_k/\|u_k\| \rangle$, and τ_{route} is a temperature.

(A) Load balancing. To prevent expert collapse, we include the MoE load-balancing auxiliary loss following [34]. Per-expert *importance* and *load* are computed (with small gate noise), yielding the auxiliary loss term $\mathcal{L}_{\text{aux}}$. The full definition is in Appendix A.

(B) Expert diversity via Centered Kernel Alignment (CKA) with token forcing. Homogeneous representation among experts has been a long-standing problem with MoEs [35]. To encourage specialists to learn diverse features, inspired by [36], we use a linear CKA not just as a metric of similarity between expert representations, but also as a penalty on a random subset of tokens that are forced through every expert. Concretely, at each step, we sample t tokens $H \in \mathbb{R}^{t \times D}$ from the MoE input and pass them through all expert FFNs to obtain $H_k = f_k(H)$ for $k = 1, \ldots, E$. After column-centering $\tilde{H}_k = H_k - \mathbf{1}\mu_k^\top$ with $\mu_k = \frac{1}{t} H_k^\top \mathbf{1}$, we measure pairwise linear CKA, and then uniformly average the similarity across all unordered expert pairs:

$$\mathcal{L}_{\text{CKA}} = \frac{2}{E(E-1)} \sum_{1 \le i < j \le E} \text{CKA}_{\text{lin}}(H_i, H_j), \quad \text{CKA}_{\text{lin}}(H_i, H_j) = \frac{\|\tilde{H}_i^\top \tilde{H}_j\|_F^2}{\sqrt{\|\tilde{H}_i^\top \tilde{H}_i\|_F^2}\sqrt{\|\tilde{H}_j^\top \tilde{H}_j\|_F^2}}. \tag{2}$$

This penalty discourages redundant specialists while the token-forcing step ensures every expert receives gradient signal regardless of routing sparsity.

(C) Invariance in the shared expert via supervised contrast. We want f_0 to encode class-relevant features that hold across discovered regions. For each minibatch sample (x_i, m_i, y_i), let

$$d_i \;:=\; \arg\max_{k \in \{1,\ldots,E\}} \alpha_k\big(\,\phi(m_i)\,\big), \qquad \text{and} \qquad q_i \;:=\; f_0(x_i)/\|f_0(x_i)\|_2.$$

To pull together same-class samples from different domains and push apart different-class samples within the same domain, we first define positive and negative sets for each anchor q_i:

$$\mathbb{P}_i \;=\; \{\, q_j \;:\; y_j = y_i \text{ and } d_j \neq d_i \,\}, \qquad \mathbb{N}_i \;=\; \{\, q_j \;:\; y_j \neq y_i \text{ and } d_j = d_i \,\}.$$

Let $\mathcal{A} = \{i : \mathbb{P}_i \neq \emptyset \text{ and } \mathbb{N}_i \neq \emptyset\}$ be a random subset of valid anchors in a batch and $B := |\mathcal{A}|$. With temperature τ_{sup}, the supervised contrastive loss is

$$\mathcal{L}_{\text{SupCon}} = \frac{-1}{B} \sum_{i \in \mathcal{A}} \left[\log \sum_{p \in \mathbb{P}_i} e^{\langle q_i, p \rangle / \tau_{\text{sup}}} - \log \sum_{v \in \mathbb{P}_i \cup \mathbb{N}_i} e^{\langle q_i, v \rangle / \tau_{\text{sup}}} \right]. \tag{3}$$

Training objective. We train MoGE by minimizing the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{aux}}\,\mathcal{L}_{\text{aux}} + \lambda_{\text{CKA}}\,\mathcal{L}_{\text{CKA}} + \lambda_{\text{SupCon}}\,\mathcal{L}_{\text{SupCon}}\,, \tag{4}$$

where the λs are hyperparameters balancing the losses; Appendix A provides the definition of $\mathcal{L}_{\text{task}}$.

2.2 Generalizing Beyond the Metadata Distribution

MoGE targets global generalization under the idealized assumption of full metadata coverage. In practice, coverage is often sparse, patchy, or missing by region, so metadata encoders may not extrapolate to unseen coordinates. Nonetheless, MoGE's disentangled design permits simple, robust adaptations. We outline two complementary strategies supported by MoGE:

(i) Invariant fallback (shared expert only). Disable specialist routing and use f_0 alone: $f(x, m) = f_0(x)$, and train a new classifier on top of frozen invariant features. This removes dependence on potentially shifted metadata at the cost of region-specific cues, yielding a safe lower bound.

(ii) Plug-in pretrained location encoders. Use a frozen, pretrained location encoder (e.g., SatCLIP [18]) instead of training ϕ, improving zero-shot generalization to unseen coordinates via structured geospatial embeddings, assuming pretraining's semantic alignment with the downstream task.

2.3 Implementation Details

We use ViT-Tiny (DeiT ImageNet-pretrained [37]) with 224×224 inputs, 10 Transformer blocks, and 2 MoGE blocks with 6 experts and top-3 routing; for iWildCam [2], we use ViT-S384/32 with 448×448 inputs (to match WILDS [2] official input size) and the same layout (interpolated positional encoding to 448). The location encoder is a randomly initialized SatCLIP [18] with Legendre polynomials of degree 10, while for iWildCam we embed discrete location IDs. Models are selected by validation worst-group accuracy over 50 epochs, and hyperparameters are tuned with DomainBed random search [5]. We use Adam [38] optimizer. Ablation studies can be found in Appendix A.

Table 1: Results on WILDS FMoW, FMoW-Left-Out Asia, and iWildCam using official WILDS [2] metrics. MoGE-SE uses only the shared expert. FMoW: worst-group and overall accuracy (%); iWildCam: F1-macro scores.

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

	Method	Worst Acc. (%)	Overall Acc. (%)
Domain Specific	MoGE	**40.56 ± 2.35**	**54.98 ± 0.38**
	ERM+LE	35.83 ± 0.51	53.23 ± 0.72
	D3G [21]	33.38 ± 0.64	51.51 ± 0.19
	D3G+WRAP [31]	34.60 ± 1.27	50.76 ± 0.12
	Per Region Models	26.75 ± 1.26	49.72 ± 0.26
Domain Invariant	MoGE-SE	**33.49 ± 1.78**	49.19 ± 0.26
	ERM	32.43 ± 1.67	**53.69 ± 0.37**
	GroupDRO [39]	30.70 ± 0.80	49.06 ± 0.37
	IRM [7]	25.85 ± 0.93	42.71 ± 0.41
	Fish [40]	33.08 ± 0.29	44.99 ± 0.72
	Mixup [41]	32.88 ± 0.63	47.25 ± 0.62
	VREx [42]	32.48 ± 1.28	46.83 ± 0.90
	RDM [43]	32.66 ± 0.60	47.70 ± 0.17

(b) FMoW-LAO

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

(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 [44]	0.290	0.472

2.4 Datasets

Functional Map of the World (FMoW): We evaluate MoGE on the WILDS-FMoW dataset [2], a benchmark for global land use classification from satellite imagery with temporal domain shift.

FMoW-Left-Out Asia (FMoW-LAO): We make a custom split of FMoW by holding out all Asian samples during training, to probe MoGE's ability to generalize to a completely unseen continent.

iWildCam: To assess MoGE on a task with minimal concept shift, we use the iWildCam2020-WILDS dataset [2], a large-scale benchmark for species recognition from camera-trap images.

3 Results and Discussion

On FMoW, we observe that MoGE outperforms all baselines by a large margin, with a $\sim$8% absolute gain in worst-group accuracy over ERM (Table 1), and 4.7% gain over ERM+LE, which combines ERM with location encodings with a 2-layer MLP following [21]. It should be noted that all Domain Generalization (DG) models have access to group labels during training, whereas MoGE learns the domains without requiring fused or explicit domain knowledge, relying solely on location metadata. On FMoW-LAO, MoGE+SatCLIP outperforms ERM and DG methods on the held-out continent, showing that MoGE can generalize to unseen regions even with a pretrained LE. On iWildCam, where we expect less concept shift, MoGE still improves over ERM by $\sim$2.2% in F1-id, and in the F1-ood setting, using only the shared expert (MoGE-SE) outperforms ERM by a small margin. The shared-expert ablation (MoGE-Shared Expert) performs marginally better than ERM, while other DG methods underperform ERM, consistent with prior findings [2]. More importantly, we don't observe the typical tradeoff between the worst group and overall accuracy [45].

4 Conclusion and Future Work

We introduced MoGE to leverage often-underutilized geospatial metadata for better generalization, addressing the limitations of global invariance without requiring domain experts to manually partition domains. By allowing concept-consistent regions to emerge directly from data, MoGE outperforms domain-specific baselines and offers a superior mechanism for location conditioning: unlike standard injection methods, our routing strategy preserves the backbone architecture, enabling the seamless integration of pretrained weights. Crucially, the explicit disentanglement yields a standalone shared expert that operates independently of location metadata, outperforming other DG methods on unseen domains and providing a robust fallback. Finally, MoGE's modular design opens a promising path for adaptation in non-stationary environments by simply fine-tuning the gating mechanism on new target distributions, which should be explored in future work.

Acknowledgments and Disclosure of Funding

Maha Shadaydeh is funded by the ERC Synergy Grant: Understanding and Modelling of the Earth System with Machine Learning (USMILE).

References

[1] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4205–4230, 2021.

[2] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 5637–5664, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/koh21a.html

[3] M. Ludwig, A. Moreno-Martinez, N. Hölzel, E. Pebesma, and H. Meyer, "Assessing and improving the transferability of current global spatial prediction models," *Global Ecology and Biogeography*, vol. 32, no. 3, pp. 356–368, 2023.

[4] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is Hard: A Closer Look at Subpopulation Shift," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 39 584–39 622, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v202/yang23s.html

[5] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[6] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Ré, "Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations," Dec. 2024, arXiv:2203.01517 [cs]. [Online]. Available: http://arxiv.org/abs/2203.01517

[7] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," Mar. 2020, arXiv:1907.02893 [stat]. [Online]. Available: http://arxiv.org/abs/1907.02893

[8] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson, "On Feature Learning in the Presence of Spurious Correlations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 516–38 532, Dec. 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/fb64a552feda3d981dbe43527a80a07e-Abstract-Conference.html

[9] J. Blunk, P. Bodesheim, and J. Denzler, "Adaptive model selection for expanded post hoc debiasing and mitigating varying degrees of spurious correlations," in *Computer Analysis of Images and Patterns*, M. Castrillón-Santana, C. M. Travieso-González, O. Deniz Suarez, D. Freire-Obregón, D. Hernández-Sosa, J. Lorenzo-Navarro, and O. J. Santana, Eds. Cham: Springer Nature Switzerland, 2026, pp. 101–111.

[10] I. Asaad, M. Shadaydeh, and J. Denzler, "Gradient extrapolation for debiased representation learning," in *International Conference on Computer Vision (ICCV)*, 2025, inproceedings, (accepted).

[11] K. P. Murphy, *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[12] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, p. 113924, 2024.

[13] J. Monteiro, X. Gibert, J. Feng, V. Dumoulin, and D.-S. Lee, "Domain conditional predictors for domain adaptation," in *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*. PMLR, 2021, pp. 193–220.

[14] G. W. Leibniz, *Monadology*, 1714, section 56 includes the phrase *natura non facit saltus*. [Online]. Available: https://www.gutenberg.org/ebooks/62106

[15] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, no. Supplement, pp. 234–240, 1970, introduces "Tobler's First Law of Geography". [Online]. Available: https://doi.org/10.2307/143141

[16] W. Tobler, "On the first law of geography: A reply," *Annals of the association of American geographers*, vol. 94, no. 2, pp. 304–310, 2004.

[17] D. G. Shatwell, I. R. Dave, S. Sirnam, and M. Shah, "Geotimeclip: Unveiling the when and where of images," in *International Conference on Learning Representations (ICLR)*, 2025, submitted to ICLR 2025. [Online]. Available: https://openreview.net/forum?id=z9UABOHCZc

[18] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4347–4355.

[19] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 498–23 515.

[20] N. Wu, Q. Cao, Z. Wang, Z. Liu, Y. Qi, J. Zhang, J. Ni, X. A. Yao, H. Ma, L. Mu, S. Ermon, T. Ganu, A. Nambi, N. Lao, and G. Mai, "Torchspatial: A location encoding framework and benchmark for spatial representation learning," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: https://openreview.net/forum?id=DERtzUdhkk

[21] H. Yao, X. Yang, X. Pan, S. Liu, P. W. Koh, and C. Finn, "Improving Domain Generalization with Domain Relations," Mar. 2024, arXiv:2302.02609 [cs]. [Online]. Available: http://arxiv.org/abs/2302.02609

[22] S. Mu and S. Lin, "A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications," *arXiv preprint arXiv:2503.07137*, 2025.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] Meta, "The llama 4 herd: The beginning of a new era of natively multimodal intelligence," https://ai.meta.com/blog/llama-4-multimodal-intelligence/, Apr. 2025, accessed: 2025-10-09.

[25] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.

[26] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[27] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[28] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[29] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[30] K. Tsai, S. R. Pfohl, O. Salaudeen, N. Chiou, M. Kusner, A. D'Amour, S. Koyejo, and A. Gretton, "Proxy methods for domain adaptation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 3961–3969.

[31] R. Crasto, "Robustness to geographic distribution shift using location encoders," *arXiv preprint arXiv:2503.02036*, 2025.

[32] A. Kuriyal, E. Vincent, M. Aubry, and L. Landrieu, "CoDEx: Combining Domain Expertise for Spatial Generalization in Satellite Image Analysis," Apr. 2025, arXiv:2504.19737 [cs]. [Online]. Available: http://arxiv.org/abs/2504.19737

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu, "Sparse mixture-of-experts are domain generalizable learners," *arXiv preprint arXiv:2206.04046*, 2022.

[35] B. Liu, L. Ding, L. Shen, K. Peng, Y. Cao, D. Cheng, and D. Tao, "Diversifying the mixture-of-experts representation for language models with orthogonal optimizer," *arXiv preprint arXiv:2310.09762*, 2023.

[36] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[38] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization," Apr. 2020, arXiv:1911.08731 [cs]. [Online]. Available: http://arxiv.org/abs/1911.08731

[40] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," *arXiv preprint arXiv:2104.09937*, 2021.

[41] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, "Improve unsupervised domain adaptation with mixup training," *arXiv preprint arXiv:2001.00677*, 2020.

[42] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International conference on machine learning*. PMLR, 2021, pp. 5815–5826.

[43] T. Nguyen, K. Do, B. Duong, and T. Nguyen, "Domain generalisation via risk distribution matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2790–2799.

[44] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[45] Y. Zong, Y. Yang, and T. Hospedales, "Medfair: benchmarking fairness for medical imaging," *arXiv preprint arXiv:2210.01725*, 2022.

A Supplementary Material

In this section, we provide additional ablations to analyze the main design choices in MoGE.

Effect of SupCon loss, number of experts, and Top-k. Figure 2 summarizes three ablations on FMoW. Figure 2(a) varies the weight of the supervised contrastive loss on the shared expert. Removing SupCon ($\lambda_{\text{SupCon}} = 0$) leads to a clear drop in worst-region accuracy, confirming that explicitly enforcing invariance on shared expert features across discovered regions is crucial. Performance peaks around a weight of 0.1, after which larger weights start to hurt performance, indicating that an overly strong invariance constraint begins to conflict with the classification loss.

Figure 2(b) studies the number of experts under fixed top-3 routing. Using E=6 experts yields the best worst-region accuracy, while surprisingly E=3 experts performs competitively. Increasing the number of experts beyond 6 does not further improve overall performance and slightly degrades worst-region accuracy. Interestingly, the performance of the shared expert alone (MoGE-SE) continues to improve with more experts. This suggests that additional specialists help MoGE discern minority regions more cleanly, which in turn supports learning stronger invariant features in the shared pathway.

Figure 2(c) evaluates top-k routing without a shared expert. Top-5 routing achieves the highest accuracy but comes with higher variance and increased computational cost. Top-1 routing, in contrast, leads to the weakest performance, showing that complete sparsity hinders experts from learning from complementary regions and supports the idea of overlapping conditionals. Based on this trade-off between accuracy, variance, and efficiency, we use top-3 routing in all main experiments.

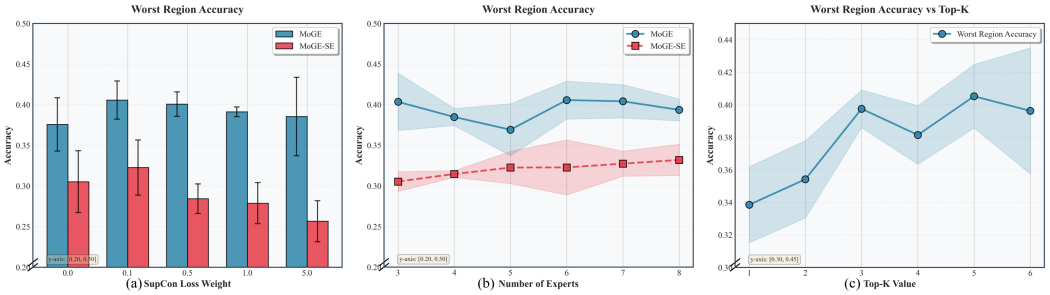

Figure 2: Ablations on FMoW: (a) SupCon loss weight on the shared expert, (b) number of experts with fixed top-3 routing, and (c) top-k routing without a shared expert.

Effect of CKA diversity loss and shared expert. Figure 3 analyzes how the CKA-based diversity loss and the shared expert together shape specialization and robustness. The top row shows, for each configuration, the matrix of per-domain accuracies when forcing each expert to handle a given domain (columns: forced expert ID; rows: domain of the test sample). The bottom row reports the routing frequency of each expert.

The baseline MoGE (left) is trained without CKA loss and without a shared expert. It attains competitive performance, but the performance matrix exhibits small row-wise variance: all experts behave similarly across domains, indicating highly correlated specialists. Routing usage is relatively balanced, confirming that the auxiliary load-balancing loss is effective, but does not by itself enforce diversity.

Adding the CKA loss (middle) decorrelates experts and sharpens the diagonal of the performance matrix, meaning that each expert becomes more specialized to a subset of domains. However, the overall worst-region performance drops and routing usage collapses to a small subset of experts. In practice, the orthogonality constraint pulls experts too far from the true data distribution, overriding the load-balancing loss and causing expert collapse. However, we can see that the collapsed expert still delivers moderate performance, which is due to the fact that it was still used as a top-2 or top-3 expert during routing, thus receiving some gradient signal, indicating low importance but high load in its auxiliary loss.

Finally, enabling both the CKA loss and the shared expert (right) yields the best worst-region performance. The diagonal of the performance matrix is strongly pronounced, indicating clear

domain-expert specialization, while off-diagonal entries drop compared to the baseline. Routing usage is again well spread across experts, showing that the shared expert stabilizes training and prevents collapse by absorbing globally useful features, allowing the specialized experts to focus on complementary, region-specific cues.

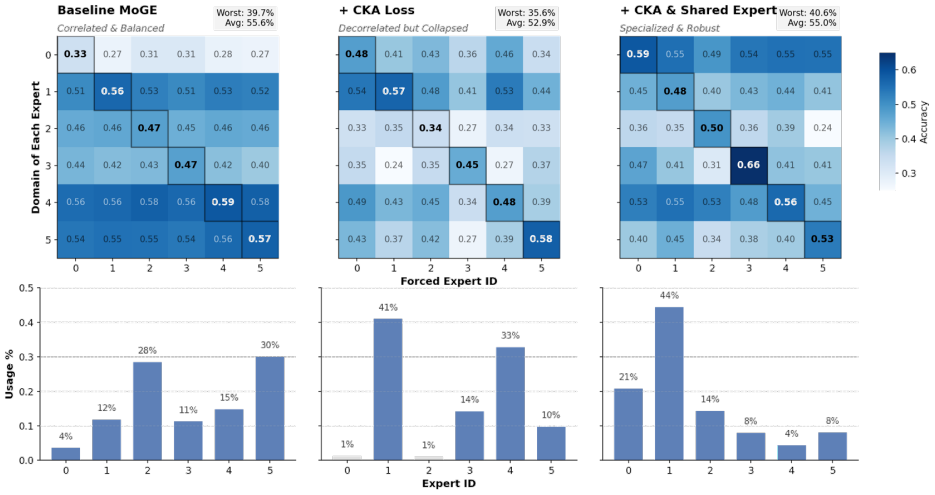

Figure 3: Ablation of the CKA diversity loss and the shared expert on FMoW. **Top:** per-domain accuracy when forcing each expert. Each row corresponds to an expert domain (test samples that have the highest routing score for that expert), and each column to the samples routed to a specific expert. **Bottom:** routing frequency of each expert. From left to right: baseline MoGE (no CKA, no shared expert), MoGE with CKA only, and full MoGE with both CKA and shared expert.

Task loss. For completeness, we define the task loss used in Eq. (4) of the main text.

$$\mathcal{L}_{\text{task}} = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i \mid x_i, m_i).$$

Load-balancing auxiliary loss. Expert collapse is when a mixture-of-experts model routes most tokens to only a few experts, leaving others underused and effectively wasted. This can be mitigated by encouraging balanced expert utilization through an auxiliary loss.

Consider a batch $\mathcal{B} = \{(x_i, m_i)\}_{i=1}^{N}$, with metadata embeddings $r_i = \phi(m_i)$ and cosine router weights $\alpha_k(r_i)$ as in Eq. (1).

Importance. The importance of expert k is the total routing mass it receives:

$$\text{imp}_k(\mathcal{B}) = \sum_{i=1}^{N} \alpha_k(r_i), \qquad \mathcal{L}_{\text{imp}}(\mathcal{B}) = \left(\frac{\text{STD}(\text{imp}(\mathcal{B}))}{\text{MEAN}(\text{imp}(\mathcal{B}))}\right)^2.$$

Load. We add small Gaussian noise with std. σ to the logits $s_k(r_i)$. Let $\eta_K^{(i)}$ be the K-th largest entry of $s(r_i)$; the probability that expert k is in the top-K for sample i is approximated by

$$p_k(r_i) = 1 - \Phi\left(\frac{\eta_K^{(i)} - s_k(r_i)}{\sigma}\right),$$

where Φ is the standard normal CDF. The expected load and its loss are

$$\text{load}_k(\mathcal{B}) = \sum_{i=1}^{N} p_k(r_i), \qquad \mathcal{L}_{\text{load}}(\mathcal{B}) = \left(\frac{\text{STD}(\text{load}(\mathcal{B}))}{\text{MEAN}(\text{load}(\mathcal{B}))}\right)^2.$$

The auxiliary term used in Eq. (4) is

$$\mathcal{L}_{\text{aux}}(\mathcal{B}) = \tfrac{1}{2}\big(\mathcal{L}_{\text{imp}}(\mathcal{B}) + \mathcal{L}_{\text{load}}(\mathcal{B})\big),$$