

Latent Domain Modeling Improves Robustness to Geographic Shifts

Supplementary Material

1. Dataset Details

In this section, for each dataset used in this paper, we report the total size and domain distribution within each split.

WILDS-FMoW We use the default WILDS training set and OOD val and test sets with continent domain labels:

Group	Train		Val		Test	
	Count	%	Count	%	Count	%
Asia	17,809	23%	4,121	21%	4,963	22%
Europe	34,816	45%	7,732	39%	5,858	26%
Africa	1,582	2%	803	4%	2,593	12%
Americas	20,973	27%	6,562	33%	8,024	36%
Oceania	1,641	2%	693	3%	666	3%
Other	42	0%	4	0%	4	0%
Total:	76,863	100%	19,915	100%	22,108	100%

Table 1. Continent distribution in WILDS-FMoW train, OOD validation, and OOD test sets.

WILDS-PovertyMap We use the default WILDS training set and OOD val and test sets with urban/rural domain labels:

Group	Train		Val		Test	
	Count	%	Count	%	Count	%
Rural	6,410	66%	2,647	67%	2,455	62%
Urban	3,368	34%	1,316	33%	1,473	37%
Total:	9,778	100%	3,963	100%	3,928	100%

Table 2. Urban/Rural distribution in WILDS-PovertyMap train, OOD validation, and OOD test sets.

iNat-Biomes iNat-Biomes is a subset of the TorchSpatial-iNat2018 [?] dataset. About 10% of datapoints were removed because they did not correspond to any of the 14 biome classes. The biome distribution across data splits in iNat-Biomes is shown in Tab. 4

TorchSpatial-YFCC The train, validation, and test splits are unmodified from the TorchSpatial YFCC dataset [?]:

Group	Train		Val		Test	
	Train	%	Val	%	Test	%
Rural	33,360	50%	2,156	49%	9,044	51%
Urban	33,379	50%	2,293	51%	8,754	49%
Total:	66,739	100%	4,449	100%	17,798	100%

Table 3. Urban/Rural distribution in TorchSpatial-YFCC train, validation, and test sets.

2. Model Details and Sizes

In this section we report model architecture details and sizes for each location encoder. In Tabs. 6 to 9 we report total model sizes for each combination of location encoder and fusion method. In Tab. 10 we report domain predictor sizes.

2.1. Location Encoders

We train a ResNet consisting of 4 residual blocks similar to $\text{NN}^{\text{wrap}}()$ in [?] on top of each location featurization method (i.e., WRAP, GeoCLIP, RFF, SatCLIP). The ResNet consists of 1) a linear layer with ReLU activation, 2) 4 residual blocks, each consisting of 2 linear layers with ReLU activation, followed by a skip connection. The output dimension of the ResNet is 256. For WRAP and RFF, the Fourier features are directly input to the ResNet. For GeoCLIP and SatCLIP, the features output from the pre-trained encoders are input to the ResNet and kept frozen. The total size of each location encoder are shown in Table 5.

2.2. Fusion Methods

In Tabs. 6 to 9 we report the total model size for each combination of dataset, fusion method, and location encoder. These results exclude the domain predictor sizes for each dataset which are reported in Sec. 2.3. Note that the image-only base model used for ERM is the same as what is used for the IRM, CORAL, and GroupDRO results reported in [?]. Overall, these fusion methods do not incur a significant increase in model size compared to the image-only base model.

2.3. Domain Predictors

Domain prediction applied to the location encoder is a lightweight addition to any image-location fusion method. For each dataset, we use a single linear layer for the domain predictor. The size of the domain predictor for different datasets is shown in Table 10. Since the location encoder output dimension is always 256, these sizes are constant across all choices of location encoder.

Group	Train		Val		Test	
	Count	%	Count	%	Count	%
1. Tropical & Subtropical Moist Broadleaf Forests	13,521	3%	1,530	7%	1,530	7%
2. Tropical & Subtropical Dry Broadleaf Forests	10,204	3%	872	4%	872	4%
3. Tropical & Subtropical Coniferous Forests	7,222	2%	557	3%	557	3%
4. Temperate Broadleaf & Mixed Forests	125,749	31%	6,817	31%	6,817	31%
5. Temperate Conifer Forests	41,480	10%	2,243	10%	2,243	10%
6. Boreal Forests/Taiga	1,237	0%	113	1%	113	1%
7. Tropical & Subtropical Grasslands, Savannas & Shrublands	16,742	4%	1,140	5%	1,140	5%
8. Temperate Grasslands, Savannas & Shrublands	81,881	20%	137	1%	137	1%
9. Flooded Grasslands & Savannas	2,163	1%	69	0%	69	0%
10. Montane Grasslands & Shrublands	1,279	0%	3,430	16%	3,430	16%
11. Tundra	762	0%	151	1%	151	1%
12. Mediterranean Forests, Woodlands & Scrub	70,319	17%	2,786	13%	2,786	13%
13. Deserts & Xeric Shrublands	29,631	7%	1,790	8%	1,790	8%
14. Mangroves	2,437	1%	149	1%	149	1%
Total:	404,627	100%	21,784	100%	21,784	100%

Table 4. Biome distributions in iNat-Biomes. As with TorchSpatial-iNat2018, the test and validation sets are the same.

	Total Params	Trainable Params	ResNet Input Dim.
WRAP	527.6 K	527.6 K	4
GeoCLIP	10.11 M	657.7 K	512
RFF	658.18 K	657.7 K	512
SatCLIP	1.81 M	592.1 K	256

Table 5. Size of each location encoder.

	Method	Total Params	Trainable Params
None	ERM	427.99 M	427.99 M
	D ³ G	428.26 M	428.26 M
WRAP	Concat	428.54 M	428.54 M
	Geo Priors	428.54 M	428.54 M
	FiLM	429.31 M	429.31 M
	D ³ G	428.86 M	428.86 M
GeoCLIP	Concat	438.11 M	428.67 M
	Geo Priors	438.11 M	428.67 M
	FiLM	438.89 M	429.44 M
	D ³ G	438.44 M	428.99 M

Table 6. Size of each model trained on WILDS-FMoW for different combinations of location encoder and fusion method. These sizes exclude the domain predictor.

3. Training Hyperparameters

We report training hyperparameters and design choices specific to each dataset (e.g. batch size, number of epochs) and specific to each model component and fusion method.

	Method	Total Params	Trainable Params
None	ERM	11.20 M	11.20 M
	D ³ G	11.23 M	11.23 M
WRAP	Concat	11.73 M	11.73 M
	Geo Priors	—	—
	FiLM	12.12 M	12.12 M
	D ³ G	11.83 M	11.83 M
GeoCLIP	Concat	21.31 M	11.86 M
	Geo Priors	—	—
	FiLM	21.70 M	12.25 M
	D ³ G	21.40 M	11.96 M

Table 7. Size of each model trained on WILDS-PovertyMap for different combinations of location encoder and fusion method. These sizes exclude the domain predictor.

WILDS-FMoW Each model is trained for 5 epochs and we select the model checkpoint with lowest validation loss across epochs. We train with batch size 16 and use the Adam optimizer with initial learning rate 10^{-4} that decays by a factor of 0.96 each epoch. The CLIP ViT-L/14 backbone is finetuned with the same optimizer and learning rate schedule but with initial learning rate 10^{-5} . We use gradient accumulation to achieve an effective batch size of 64. Finally, we normalize all images using ImageNet mean and standard deviation, and apply random horizontal flip.

WILDS-PovertyMap We follow a training setup identical to [?], where each model is trained for 200 epochs

	Method	Total Params	Trainable Params
None	ERM	16.68 M	16.68 M
	D ³ G	233.59 M	233.59 M
WRAP	Concat	19.29 M	19.29 M
	Geo Priors	2.62 M	2.62 M
	FiLM	21.94 M	21.94 M
	D ³ G	234.19 M	234.19 M
GeoCLIP	Concat	28.87 M	19.42 M
	Geo Priors	12.20 M	2.75 M
	FiLM	31.51 M	22.07 M
	D ³ G	243.77 M	234.32 M

Table 8. Size of each model trained on iNat-Biomes for different combinations of location encoder and fusion method. These sizes exclude the domain predictor.

	Method	Total Params	Trainable Params
None	ERM	204.90 K	204.90 K
	D ³ G	443.34 K	443.34 K
WRAP	Concat	758.12 K	758.12 K
	Geo Priors	553.32 K	553.32 K
	FiLM	5.46 M	5.46 M
	D ³ G	1.04 M	1.04 M
GeoCLIP	Concat	10.34 M	888.16 K
	Geo Priors	10.13 M	683.36 K
	FiLM	15.04 M	5.59 M
	D ³ G	10.62 M	1.17 M

Table 9. Size of each model trained on TorchSpatial-YFCC for different combinations of location encoder and fusion method. These sizes exclude the domain predictor.

FMoW	PovertyMap	iNat-Biomes	YFCC
1.5 K	0.5 K	3.6 K	0.5 K

Table 10. Domain predictor model sizes (i.e., number of parameters) for each dataset.

and we select the model checkpoint with highest validation r across epochs. We train with batch size 64 using Adam with initial learning rate 10^{-3} that decays by a factor of 0.96 each epoch. As in [?], we apply random horizontal and vertical flip and color jitter to each image.

iNat-Biomes Each model is trained for 50 epochs and we select the model checkpoint with highest validation accuracy across epochs. We train with batch size of 1024 using Adam with initial learning rate 10^{-3} that decays by a factor of 0.96 each epoch. We use the image predictions and

2048-dimensional frozen image features from [?] exactly as-is, without modification.

TorchSpatial-YFCC Each model is trained for 100 epochs and we select the model checkpoint with highest validation accuracy across epochs. We train with batch size of 2048 using Adam with initial learning rate 10^{-4} that decays by a factor of 0.96 each epoch. As with iNat-Biomes, we use the image predictions and 2048-dimensional frozen image features from [?] exactly as-is, without modification.

D³G There are two training hyperparameters used in D³G, denoted λ and β in the original paper [?]. D³G has two loss terms: the task prediction loss (denoted \mathcal{L}_{pred} in the original paper) computed from the prediction head for the current domain, and the consistency loss (\mathcal{L}_{rel}) that is computed from all other prediction heads whose predictions are weighted by their relation to the current domain. The final loss is of the form: $\mathcal{L}_{pred} + \lambda\mathcal{L}_{rel}$. The β hyperparameter is used to average fixed and learned domain relations (see [?] for more details). We use a value of $\lambda = 0.5$ and $\beta = 0.8$ for all experiments.

Domain Prediction The linear layer we use for domain prediction in all experiments is trained with Adam and step decay learning rate schedule (decayed by a factor of 0.96 per epoch) but with initial learning rate 0.1 times the initial learning rate of the prediction head. We tuned the domain prediction weight α on a subset of models, experimenting with values in $\{0.001, 0.01, 0.1, 0.2\}$, and used $\alpha = 0.2$ for most other experiments as it generally performed best.

4. Additional Results

Additional Main Results Plots We include the two scatter plots showing overall average against worst group performance that were missing from ?? due to space constraints. They are shown in Figures 1a and 1b. These additional plots are consistent with those in the main section: we observe that our instances of our proposed framework outperform all other methods on worst-group performance, and that these instances lead to significant gains in overall average performance as well compared with other baselines.

Full Ablation Results We report full ablation results (i.e. including overall average and worst group performance) for the auxiliary domain prediction loss in Tab. 11 and for the location encoder in Tab. 12.

Full RFF/SatCLIP Results In Tab. 13 we report full results (i.e. including overall average and worst group performance) on PovertyMap and iNat-Biomes using additional location encoders: RFF and SatCLIP.

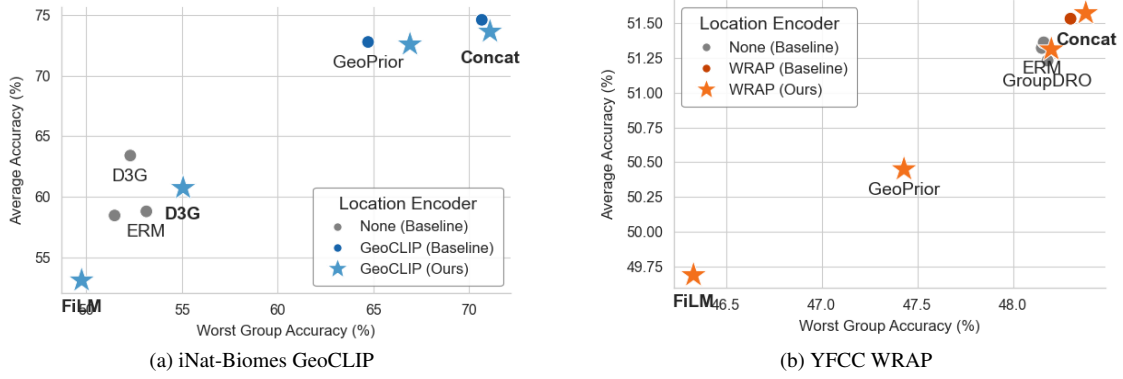


Figure 1. Overall average accuracy vs. worst group accuracy. Each result is averaged over 3 random seeds.

Method	Loc. Enc.	FMoW		PovertyMap		iNat-Biomes		YFCC	
		Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Concat	WRAP	66.9 (0.3)	48.1 (1.6)	0.78 (0.02)	0.48 (0.04)	71.7 (0.1)	68.4 (0.3)	51.5 (0.0)	48.3 (0.1)
	GeoCLIP	66.0 (0.8)	52.8 (1.0)	0.80 (0.02)	0.53 (0.03)	74.6 (0.1)	70.7 (0.6)	52.4 (0.0)	49.2 (0.1)
Geo Priors	WRAP	65.7 (0.2)	48.1 (1.7)	—	—	72.5 (0.0)	65.2 (0.2)	50.5 (0.0)	47.4 (0.0)
	GeoCLIP	66.8 (0.1)	49.6 (0.8)	—	—	72.8 (0.0)	64.7 (0.6)	50.9 (0.0)	47.9 (0.0)
FiLM	WRAP	66.4 (0.2)	49.4 (2.0)	0.79 (0.02)	0.50 (0.02)	59.6 (0.7)	52.3 (1.1)	49.6 (0.1)	46.2 (0.0)
	GeoCLIP	65.52 (0.5)	49.9 (1.7)	0.82 (0.01)	0.57 (0.02)	48.3 (2.3)	43.9 (3.7)	49.6 (0.1)	46.3 (0.1)
D ³ G	WRAP	66.4 (0.4)	54.2 (1.6)	0.78 (0.02)	0.46 (0.02)	66.1 (0.8)	46.8 (8.4)	51.4 (0.0)	48.3 (0.1)
	GeoCLIP	70.0 (0.8)	54.0 (0.9)	0.78 (0.02)	0.45 (0.02)	61.9 (6.3)	53.6 (10.1)	51.4 (0.0)	48.2 (0.0)

Table 11. Ablation results for the auxiliary domain prediction loss. The table shows both overall average and worst-group performance results when no domain prediction loss is applied (i.e. $\alpha = 0$) across different combinations of fusion method and location encoder.

Method	Loc. Enc.	FMoW		PovertyMap		iNat-Biomes		YFCC	
		Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Concat	None	66.2 (0.2)	50.9 (0.1)	0.78 (0.02)	0.40 (0.02)	57.4 (0.2)	49.9 (4.7)	51.5 (0.0)	48.3 (0.0)
	WRAP	66.8 (0.4)	49.1 (1.6)	0.80 (0.02)	0.52 (0.04)	73.1 (0.0)	70.2 (0.2)	51.6 (0.1)	48.4 (0.2)
	GeoCLIP	66.2 (0.2)	52.0 (0.3)	0.80 (0.02)	0.54 (0.02)	73.6 (0.0)	71.1 (0.1)	52.4 (0.0)	49.2 (0.0)
Geo Priors	None	65.5 (0.6)	50.0 (1.4)	—	—	59.1 (2.5)	43.7 (8.83)	50.2 (0.0)	47.1 (0.1)
	WRAP	65.7 (0.2)	49.1 (2.0)	—	—	72.6 (0.03)	65.5 (0.2)	50.5 (0.0)	47.4 (0.1)
	GeoCLIP	66.8 (0.1)	50.9 (1.1)	—	—	72.6 (0.03)	66.9 (0.5)	50.9 (0.0)	47.9 (0.0)
FiLM	None	64.5 (0.2)	50.0 (2.0)	0.80 (0.02)	0.43 (0.03)	56.8 (1.0)	52.1 (1.4)	50.2 (0.1)	46.9 (0.1)
	WRAP	66.9 (0.5)	51.8 (2.1)	0.80 (0.02)	0.52 (0.04)	61.5 (1.2)	56.0 (1.1)	49.7 (0.0)	46.3 (0.1)
	GeoCLIP	66.9 (0.8)	51.7 (2.0)	0.82 (0.02)	0.57 (0.03)	53.1 (1.5)	49.8 (1.6)	49.7 (0.2)	46.2 (0.2)
D ³ G	None	66.7 (0.2)	51.7 (1.2)	0.79 (0.02)	0.46 (0.04)	63.2 (1.4)	44.9 (8.2)	51.4 (0.0)	48.3 (0.1)
	WRAP	66.9 (0.2)	53.1 (0.7)	0.79 (0.02)	0.47 (0.04)	65.2 (1.6)	47.0 (1.4)	51.3 (0.1)	48.2 (0.2)
	GeoCLIP	66.6 (0.1)	55.8 (0.7)	0.79 (0.02)	0.46 (0.04)	60.7 (2.7)	55.1 (0.7)	51.3 (0.0)	48.2 (0.1)

Table 12. Ablation results for the choice of location encoder, showing both overall average and worst-group performance.

	Loc. Enc.	Method	PovertyMap		iNat-Biomes	
			Avg	Worst	Avg	Worst
Baselines	RFF	Concat	0.80 (0.02)	0.53 (0.05)	<u>74.4 (0.1)</u>	<u>70.5 (0.2)</u>
	RFF	Geo Priors	—	—	72.6 (0.0)	65.7 (0.8)
	SatCLIP	Concat	0.79 (0.01)	0.48 (0.03)	72.6 (0.1)	68.6 (0.5)
	SatCLIP	Geo Priors	—	—	71.4 (0.1)	64.0 (0.2)
Ours (all w/ DP)	RFF	Concat	0.79 (0.02)	0.54 (0.03)	75.1 (0.0)	71.8 (0.2)
	RFF	Geo Priors	—	—	72.7 (0.0)	65.5 (0.6)
	RFF	FiLM	<u>0.81 (0.02)</u>	0.56 (0.03)	61.2 (1.8)	55.8 (0.6)
	RFF	D ³ G	0.79 (0.02)	0.46 (0.04)	66.8 (0.7)	54.7 (7.0)
	SatCLIP	Concat	0.80 (0.02)	0.50 (0.03)	72.6 (0.0)	69.0 (0.4)
	SatCLIP	Geo Priors	—	—	71.5 (0.1)	64.2 (0.4)
	SatCLIP	FiLM	0.82 (0.02)	<u>0.55 (0.05)</u>	69.5 (0.1)	64.3 (0.6)
	SatCLIP	D ³ G	0.79 (0.02)	0.48 (0.04)	65.9 (0.7)	62.6 (0.5)

Table 13. Full results on PovertyMap and iNat-Biomes with RFF and SatCLIP location encoders.