CoDEx: Combining Domain Expertise for Spatial Generalization in Satellite Image Analysis

Abhishek Kuriyal[1] Elliot Vincent[1,2,3] Mathieu Aubry[1] Loïc Landrieu[1,2]

[1]LIGM, ENPC, IP Paris, Univ Gustave Eiffel, CNRS, France
[2]LASTIG, Univ Gustave Eiffel, IGN-ENSG, 94160, Saint-Mande, France
[3]Inria, ENS, CNRS, PSL Research University, France

Abstract

Global variations in terrain appearance raise a major challenge for satellite image analysis, leading to poor model performance when training on locations that differ from those encountered at test time. This remains true even with recent large global datasets. To address this challenge, we propose a novel domain-generalization framework for satellite images. Instead of trying to learn a single generalizable model, we train one expert model per training domain, while learning experts' similarity and encouraging similar experts to be consistent. A model selection module then identifies the most suitable experts for a given test sample and aggregates their predictions. Experiments on four datasets (DynamicEarthNet, MUDS, OSCD, and FMoW) demonstrate consistent gains over existing domain generalization and adaptation methods. Our code is publicly available at https://github.com/Abhishek19009/CoDEx.

1. Introduction

Earth observation data often exhibits significant spatial domain shifts, such as diverse biomes, architectures, or climate [28, 32, 36]. Consequently, trained models struggle to generalize to new conditions, especially when they differ substantially from the ones of the training data. This shortfall is particularly problematic because models then perform best where annotations are already abundant, rather than in regions where their prediction is most needed.

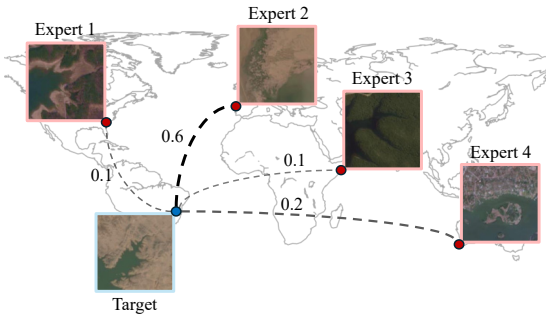


Figure 1. **Combing Domain Experts.** The appearance of satellite images can vary significantly across regions, even when they contain the same semantic features (here, coastlines). We train experts for each discrete location in the train set, and learn to aggregate the most relevant ones when confronted with an unknown domain.

Deep learning models are especially vulnerable to these domain shifts because of their capacity to learn and overfit complex data distributions. Most domain generalization strategies learn a single domain-invariant model [1, 15, 17], but real-world benchmarks (*e.g.*, WILDS [14]) show that such approaches often fail to surpass simple baselines. Domain adaptation techniques—which rely on unlabeled data from a target domain—face similar hurdles, as shown in the U-WILDS benchmark [26]. These findings underscore the difficulty of addressing domain shift for Earth observation, which remains a mostly unsolved problem despite its consid-

erable importance.

In this paper, we introduce CoDEx (COmbining Domain EXperts), a multi-expert approach to domain generalization tailored for satellite images. Instead of training a single model to perform well on all domains, we learn a specialized expert for each domain in the training set, but also a similarity between experts, which we use to encourage the consistency of similar experts to improve robustness. To aggregate experts' predictions, we then train a model selection module to weight the outputs of the domain experts for a given input image, see Fig. 1. At inference, when faced with a completely new domain, we use this selection module—without any additional fine-tuning on test data—to combine the predictions of all experts.

We validate our approach on multiple satellite image and time series datasets, including DynamicEarthNet [33], MUDS [34], OSCD-3ch. [7] and FMoW [5], and show consistent performance gains. In summary:

- We propose a novel multi-domain training strategy that enforces consistency across domain experts without relying on hand-crafted similarity metrics between domains.
- We design a model-selection module to accurately predict which domain experts will yield the most reliable predictions, allowing us to perform robust spatial generalization with minimal computational overhead.
- We show consistent qualitative and quantitative improvements over existing domain-generalization and even domain-adaption methods, on four datasets and three baselines.

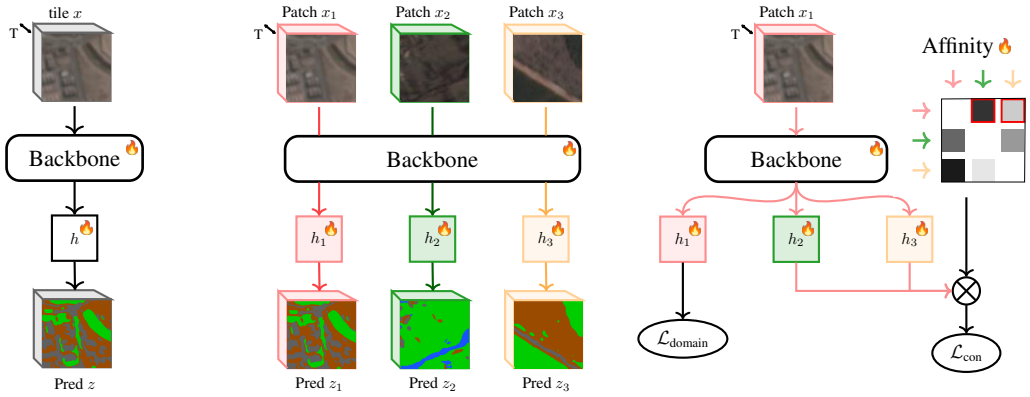
2. Related Work

In this section, we discuss existing work on unsupervised domain adaptation and domain generalization for Earth observation, and multi-experts models.

Domain Adaptation for Earth Observation. The goal of domain adaptation is to modify a model trained on annotated domains to optimize it for a target domain where only unannotated data is available. The availability of unlabeled satellite images makes such an approach particularly attractive to Earth observation [14, 19, 26, 42]. Among the most commonly used methods, feature alignment rely on moment matching [31], discriminative losses [18], or entropy minimization [4, 37] to adapt a deep learning model to a new domain. Another approach is to rely on confident predictions in the new domain, for example using pseudo-labeling [16], FixMatch [30], or Noisy Student [38]. In the case of spatial domain adaptation, spatial awareness can be incorporated into the model by leveraging geographical metadata, as demonstrated in [22]. StandardGAN [32] and StyleAugment [6] take a different approach by modifying the data distribution itself with image translation. Recently, self-supervised learning has emerged as a powerful tool for domain adaptation, where models are trained on large-scale unlabeled data in a self-supervised way and on labeled data [2, 21, 27]. In that sense, recent Earth observation foundation models like DOFA [39], pretrained on datasets covering a significant portion of the globe, can be seen as domain adaptation tools.

Domain Generalization for Earth Observation. Unlike domain adaptation, which relies on unlabeled data from the target domain, generalization approaches must learn robust representations from source data alone. The features can be trained to be invariant to spectral and temporal transformations through augmentation like color jittering, image mixing [12, 13, 24] or manipulations of time series [9, 23, 41]. This can also be achieved through the design of the features themselves, such as transformation-invariant prototypes [35], or through loss functions. For example, contrastive losses align features of spatially or temporally distant images [20], or between an image and its distorted version [44]. In contrast to these methods that focus on learning a single robust model, our approach trains and combines multiple domain-specific networks, following the idea of multi-expert models.

Multi-Experts Models. Multi-expert models aggregate predictions from multiple expert learners. Each expert may specialize in certain domains [3] or classes [43]. Experts may also be encouraged to diversify their predictions through balancing losses and different initializations [29]. Strategies for combining expert predictions at inference when given



(a) Baseline (b) Multi-Model (c) Multi-Model Training

Figure 2. **Multi-Domain Training.** We present the different multi-domain training approaches explored in this paper. In (a), we train a single model on all training domains. In (b), we train one model per training domain; all models share the same backbone network, and only see data from their domain. In (c), we add a consistency loss ensuring that the prediction of models associated to similar domains—as defined by a learnable affinity matrix—also produce accurate results. ⊗ indicates vector multiplication and 🔥 a module with tunable parameters.

an out-of-domain input vary. In particular, D^{3}G [40] aggregates predictions from domain-specific models using weights predicted as a function of domain metadata. Instead, our proposed expert selection module only relies on input features and does not require any metadata at test time.

3. Method

We propose CoDEx, a spatial generalization framework for satellite image analysis. Like domain generalization models, we handle domain shifts without retraining on target domain data or requiring domain-specific adapters. Our training set consists of satellite images or time series taken from D distinct source domains, denoted as $\mathcal{D}_1, \ldots, \mathcal{D}_D$, on which we perform segmentation or classification. We will leverage this multi-domain source data to train a model that generalizes to data from an unseen target domain from a different location.

To achieve this, we first train a set of domain experts (Sec. 3.1) and then introduce a mechanism to select and combine the most relevant model predictions for a given test input (Sec. 3.2). Implementation details are provided in Sec. 3.3.

3.1. Multi-Domain Training

We train a set of D models $\phi_1, \ldots, \phi_D$, each specialized for a specific training domain $\mathcal{D}_1, \cdots, \mathcal{D}_D$, *i.e.* domain experts. These models share the same feature extractor backbone and each maintains a small, domain-specific segmentation or classification head $h_1, \cdots h_D$. In a first training stage, visualized in Fig. 2, we train the model to produce accurate and domain-consistent predictions with the two loss functions described below.

Domain Loss. Let x be an input sample (image or time series) from domain $\mathcal{D}_d$, with corresponding ground truth label y (class or label map). We supervise the predictions of the expert ϕ_d corresponding to domain $\mathcal{D}_d$ with the following loss applied to inputs x in domain $\mathcal{D}_d$:

$$\mathcal{L}_{\text{domain}}(x) = \ell(\phi_d(x), y)\ , \tag{1}$$

where ℓ is a standard classification loss. We use the focal loss [25] in our experiments, and apply it pixel-wise for segmentation tasks.

Consistency Loss. Yao *et al*. [40] propose improving cross-domain generalization by enforcing consistency among the predictions of heads from "similar" domains. To quantify domain similarity, they construct an affinity matrix $a \in \mathbb{R}^{D \times D}$, where each

entry $a_{d,e}$ encodes the proximity between domains $\mathcal{D}_d$ and $\mathcal{D}_e$. For geospatial applications, they compute a as a combination of handcrafted and learned components. Specifically, they set the handcrafted part of $a_{d,e}$ to 1 if domains $\mathcal{D}_d$ and $\mathcal{D}_e$ are in adjacent regions and 0 otherwise. The learned part is a learned function of geographical metadata.

Instead, we propose to learn the affinity matrix a directly. More precisely, we define $a \in \mathbb{R}^{D\times D}$ as the row-wise softmax—excluding the diagonal coefficients—of a learnable parameter matrix in $\mathbb{R}^{D\times D}$. We encourage consistency between domains by minimizing the following loss for an input x in a domain $\mathcal{D}_d$:

$$\mathcal{L}_{\text{con}}(x) = \ell(\sum_{e\neq d} a_{d,e}\phi_e(x), y)\ . \tag{2}$$

This loss encourages predictions from domains $\mathcal{D}_e$ close to $\mathcal{D}_d$ according to a, to also be effective for x sampled from $\mathcal{D}_d$. Unlike $\mathcal{L}_{\text{domain}}$, which updates only the parameters of the expert expert ϕ_d for inputs in $\mathcal{D}_d$, the consistency loss $\mathcal{L}_{\text{con}}$ propagates gradients to all ϕ_e for $e \neq d$. Compared to Yao *et al*. [40], this not only provides greater flexibility but also eliminates the manual design process and hyperparameters related to the definition of the handcrafted part of a. Our ablation study also shows that it yields better results.

3.2. Domain Expert Selection

In a second stage of training, visualized in Fig. 3, we freeze all models ϕ_d, and only train an expert selection head h^{select} to predict weights to aggregate the predictions of the domain experts. For a given input x, h^{select} leverages features from x provided by the backbone and outputs a vector $\phi^{\text{select}}(x) \in \mathbb{R}^D$. We train ϕ^{select} with two losses: $\mathcal{L}_{\text{acc}}$, which encourages ϕ^{select} to predict the accuracy of experts, and $\mathcal{L}_{\text{mix}}$, which aims at maximizing the quality of a mixture. Note that this module is also exclusively trained with data from the source domains.

Accuracy Prediction Loss. We encourage the expert selection ϕ^{select} to predict the accuracy of the predictions of each domain-specific expert ϕ_d on

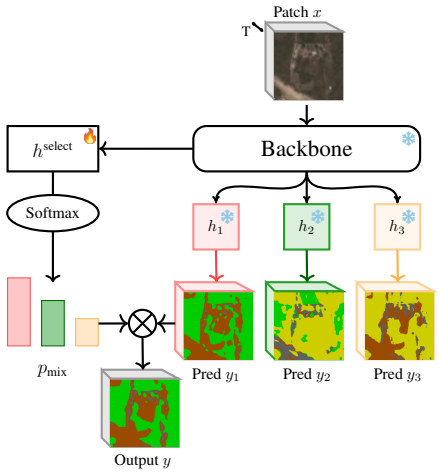


Figure 3. **Domain Expert Selection**. We freeze the models trained previously, and train a domain expert selection model to select the most relevant models for a given input sample from an unseen domain. ⊗: vector multiplication, ❄: frozen layers, 🔥: module with tunable parameters.

each sample x:

$$\mathcal{L}_{\text{acc}}(x) = \sum_d \left|\phi^{\text{select}}(x)[d] - \text{acc}\left(\phi_d(x), y\right)\right|\ , \tag{3}$$

where $\text{acc}(z, y)$ is a measure of the performance of the prediction z against the ground truth y, and $\phi^{\text{select}}(x)[d]$ is the component d of $\phi^{\text{select}}(x)$. We use pixel-wise overall accuracy for semantic segmentation tasks, and a binary value $0/1$ for classification.

Mixture Supervision Loss. We convert the predicted accuracy $\phi^{\text{select}}(x)$ into a weight vector $p_{\text{mix}}(x) \in \mathbb{R}^D_+$ using a softmax with temperature $\tau > 0$, and use it to compute a mixture of predictions from all experts $\sum_d p_{\text{mix}}(x)[d]\phi_d$, which we use at inference. We then define the loss $\mathcal{L}_{\text{mix}}$ as:

$$\mathcal{L}_{\text{mix}}(x) = \ell\left(\sum_d p_{\text{mix}}(x)[d]\phi_d(x), y\right)\ , \tag{4}$$

with $p_{\text{mix}}(x) = \text{softmax}\left(\frac{1}{\tau}\phi^{\text{select}}(x)\right)$ and ℓ the same loss as in Eqs. (1) and (2). Note that at inference, we can use the mixture prediction $\sum_d p_{\text{mix}}(x)[d]\phi_d$ without knowing the geographical location of the test sample x.

Table 1. **Datasets Characteristics**. We choose four datasets of satellite images or time series, and report their resolution, the nature of their annotations, and the number of samples and domains for the training, validation, and test sets. *We create our own custom splits for these datasets to ensure the separation of spatial domains.

	DynEarthNet	**MUDS***	**OSCD-3 ch.**	**FMoW***
Task	Segmentation	Segmentation	Change detection	Classification
Temporal Resolution	24 (monthly)	24 (monthly)	2 (image pair)	1 (monodate)
Spatial Resolution	128×128, 3-4m/pix	128×128, 3-4m/pix	75×75, 10m/pix	224×224, 0.3-1.6m/pix
Spectral Resolution	4: RGB + NIR	3: RGB	3: RGB	3: RGB
Classes	7 (land cover)	2 (land cover)	2 (change)	62 (land use)
Training	3520 (55 dom.)	2560 (40 dom.)	896 (14 dom.)	59,443 (100 dom.)
Validation	640 (10 dom.)	640 (10 dom.)	-	26,697 (50 dom.)
Test	640 (10 dom.)	640 (10 dom.)	640 (10 dom.)	32,745 (50 dom.)

3.3. Implementation Details

Our default backbone for satellite image time series (SITS) segmentation and change detection, is a MultiUTAE [8, 36], which maps SITS to time series of feature maps of the same spatial and temporal resolution as the input and has 260K parameters. The domain heads h_d are 3×3 convolutions applied to the highest-resolution feature map of the UTAE for each date for segmentation. The selection head h^{select} is also a 3×3 convolution that takes as input features from different encoder blocks concatenated along the channel dimension which are then pooled spatially and temporally. To perform change detection, we subtract the logits predicted for the two consecutive images and supervise it with the true binary change map.

For image classification, we use a larger version of MultiUTAE with 3.1M parameters as the default backbone. Both domain heads h_d and selection head h^{select} are linear layers. The domain head is applied to the pooled features, the selection head to features pooled from multiple blocks. The affinity matrix adds D^2 parameters, which is significantly smaller than the number of parameters in the encoder and heads.

As our results appear to be robust to loss weighting, we always use a weight of 1 when adding $\mathcal{L}_{\text{domain}}$ and $\mathcal{L}_{\text{con}}$ in the first stage of training, and $\mathcal{L}_{\text{acc}}$ and $\mathcal{L}_{\text{mix}}$ in the second. The temperature τ of the softmax in the mixture definition is also set to 1.

We use the Adam optimizer with a learning rate of 1×10^{-4} and coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$. Additionally, we apply a weight decay of 0.01.

4. Experiments

In Sec. 4.1, we present the four Earth observation datasets on which we evaluate our proposed approach for three tasks (classification, semantic segmentation, and change detection), with either single image or time series inputs. We then compare our approach to a wide array of domain-generalization and adaptation methods in Sec. 4.2. Finally, we ablate and analyze our approach in Sec. 4.3.

4.1. Datasets and Evaluation

Datasets. We build our benchmark for domain generalization from four standard SITS datasets, which can be naturally split into distinct spatial domains. We provide details about these dataset, DynamicEarthNet [33], MUDS [34], OSCD-3ch. [7], and FMoW [5], in Tab. 1. The domains are defined by the unique geographic locations of the images, except for FMoW for which spatial domains correspond to countries. We use train/val/test splits such that the domains are separated. We split the large spatial extent of the SITS of DynamicEarthNet, MUDS, and OSCD-3ch. into smaller SITS of spatial size 128×128, 128×128, and 75×75 respectively. For DynamicEarthNet, we only consider the images of the time series that are annotated: one image every month. For FMoW, we treat each image of the time series independently as in WILDS [14].

We evaluate model performance with the class-average intersection over union (mIoU) and overall accuracy (OA) for semantic segmentation, F1 score for change detection, and average accuracy (Avg. Acc.) for classification.

Table 2. **Quantitative Results.** We report the results of various domain generalization and adaptation methods on four datasets. Cell color represents the difference with the baseline model with the following colormap: -3 +0 +3. Oracle selects the best training head for each test input. *D^{3}G [40] is our own implementation, which we will release.

Method	DynamicEarthNet		MUDS		OSCD-3 ch.	FMoW
	mIoU	OA	mIoU	OA	F1	Avg. Acc.
Baseline [36]	37.5	73.8	63.5	95.1	46.5	51.6
Domain Adaptation: fine-tuned on the test set's input data						
StyleAugment [6]	36.6	73.4	63.1	94.7	46.3	51.3
CORAL [31]	38.3	74.6	63.3	95.1	46.6	52.0
Pseudo labels [16]	37.4	74.1	63.7	95.0	46.7	51.7
AdvEnt [37]	36.3	73.5	62.9	94.9	45.4	51.9
DANN [18]	35.8	73.9	63.3	95.1	45.9	51.8
MaxSquare [4]	38.1	74.5	64.0	95.2	47.1	52.1
Foundation Model-Based: pretrained on external data, fine-tuned on the train set						
DOFA [39]	35.8	73.7	63.2	94.3	46.2	51.3
Domain Generalization: only sees the train set						
ClassMix [24]	36.2	74.1	63.2	94.8	46.8	51.5
Contrastive Seg. [44]	37.9	73.6	63.3	95.0	46.4	53.0
D^{3}G [40]*	38.7	75.2	63.9	95.2	**48.1**	53.5
CoDEx (Ours)	**39.1**	**75.7**	**64.2**	**95.8**	47.8	**53.9**
Oracle	48.1	81.6	65.8	95.6	56.9	91.1

Competing Methods. We evaluate for each dataset a baseline approach, which trains a single model regardless of domains (see Fig. 2a). We evaluated for the first time on these datasets 6 domain-adaptation methods (StyleAugment [6], CORAL [31], Pseudo labels [16], Advent [37], DANN [18], MaxSquare [4]) and 3 domain-generalization approaches (ClassMix [24], Contrastive Seg. [44], D^{3}G [40]). Domain-adaptation methods are fine-tuned with the input data from the test set and without labels. We also evaluated a foundation model for Earth observation, DOFA [39]. As this model takes only a single image as input, we evaluate it for each image independently. Since no implementation of D^{3}G [40] is available, results are from our own implementation, which we will release. For the other methods, we modify the official code to apply it to our data.

4.2. Results

Generalization Performance. We report the performance of CoDEx and all competing approaches on all four datasets in Tab. 2. Our method consistently outperforms competing methods across all metrics, except D^{3}G, which is slightly better on OSCD-3ch. Notably, only our method and D^{3}G exceed the baseline performance for every dataset, and our approach shows slight improvement over D^{3}G on the other 3 datasets. Interestingly, generalization-focused methods often outperform domain adaptation methods, even though the latter have access to data from the target domain.

We also measure the performance of a *Domain Oracle*, which selects the best-performing expert among $\phi_1, \cdots, \phi_D$ for each test sample. This gives an upper bound on the performance one could achieve by selecting a single expert. On MUDS, the oracle's performance is only marginally higher than ours, hinting that our expert selection mechanism effectively identifies the best experts. On DynamicEarthNet and OSCD-3ch., the oracle substantially outperforms our approach, although its overall performance remains low, suggesting a considerable domain gap between training and testing. On FMoW,

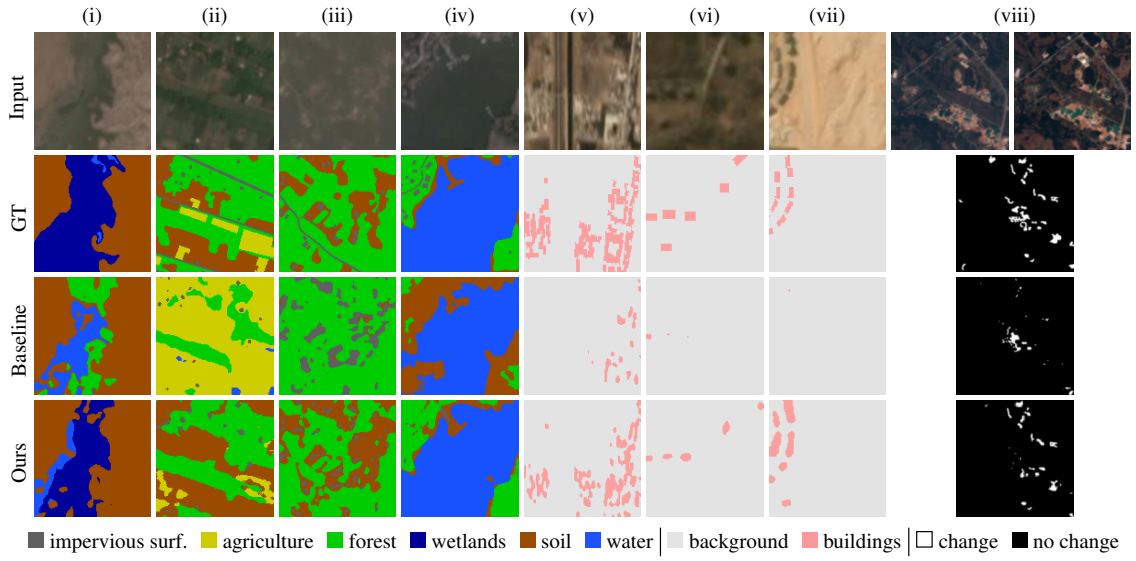
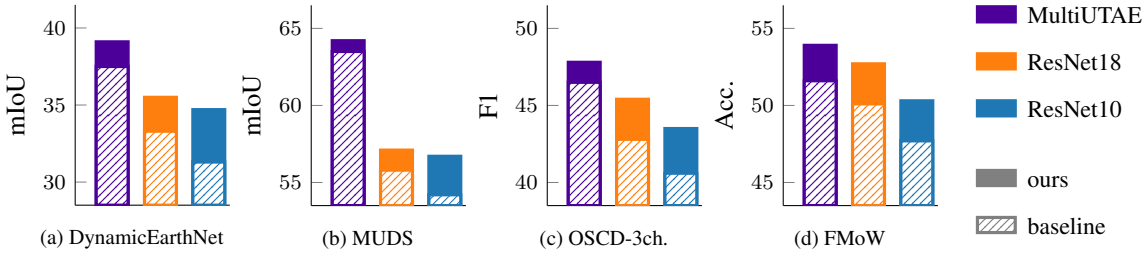


Figure 4. **Qualitative Segmentations.** We illustrate for random patches the predictions of our method and our baseline. Images from (i-iv) are selected from DynamicEarthNet, (v-vii) from MUDS, and (viii) from OSCD-3ch.



(a) DynamicEarthNet (b) MUDS (c) OSCD-3ch. (d) FMoW

Figure 5. **Impact of Backbone.** We report the performance of the baseline and our approach for all four datasets and three backbones networks.

the oracle performs extremely well. This is not surprising, as it takes the best prediction among 100 experts for a classification with 62 classes.

Qualitative Results. We illustrate in Fig. 4 the predictions from the baseline approach and our proposed method on three datasets: DynamicEarthNet (i–iv), FMoW (v–vii), and OSCD-3ch. (viii). In (i), our approach correctly resolves an ambiguous image of wetland, which might otherwise be confused with water and forest. It also provides more precise segmentation of forest and sediment which might be misclassified as agriculture (ii). For MUDS, we observe higher recall for buildings in both dense (v) and sparse (vi, vii) regions. For OSCD-3ch. (viii), our method generally improves change detection, although the task remains challenging.

4.3. Analysis

Ablations. We evaluate the contribution of different parts of our approach through ablations on DynamicEarthNet and MUDS in Tab. 3. First, rather than predicting a single mixture p_{mix} across all time steps, we remove temporal pooling of the backbone features and allow the method to choose a different model at each time step. This added flexibility reduces performance, likely because the less informative single-image features make model selection less reliable. Second, we remove either $\mathcal{L}_{\text{con}}$ or $\mathcal{L}_{\text{mix}}$, which leads to a small decrease of performance. Removing both leads to a significant drop, indicating

Table 3. **Ablation Study.** We evaluate the impact of several of our design choices.

Temp. pooling	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{mix}}$	$\mathcal{L}_{\text{acc}}$	DynEarthNet mIoU	DynEarthNet OA	MUDS mIoU	MUDS OA
✔	✔	✔	✔	**39.1**	**75.7**	**64.2**	**95.8**
✔	✔		✔	38.1	75.3	63.9	94.6
✔		✔	✔	38.2	75.1	63.6	95.3
	✔	✔	✔	37.1	74.4	63.1	94.2
✔			✔	36.5	73.6	62.9	94.4
✔	✔	✔		38.9	75.5	64.0	95.2

Table 4. **Performance comparison of different affinity matrix formulations.** We compare three approaches: fixed, D^3G-style, and fully learned weights. Results are shown for both DynamicEarthNet and MUDS datasets. The **bold** values indicate the highest performance, while the underlined values represent the second-highest.

Affinity	DynEarthNet mIoU	DynEarthNet OA	MUDS mIoU	MUDS OA
a) Handcrafted	38.8	<u>75.5</u>	63.7	95.1
b) "D^3G [40]-style"	<u>39.0</u>	75.4	<u>64.0</u>	<u>95.4</u>
c) Learned (ours)	**39.1**	**75.7**	**64.2**	**95.8**

that enforcing consistency among expert models and supervising the mixture coefficient through the final mixed predictions are both crucial and complementary. Lastly, removing $\mathcal{L}_{\text{acc}}$ and only supervising ϕ^{select} with $\mathcal{L}_{\text{mix}}$ leads to a small but consistent performance decrease, suggesting that accuracy is a useful proxy for the relevancy of a domain expert.

Consistency Loss Analysis. We examine the benefits of our fully learned domain affinity matrix in the consistency loss compared to a handcrafted baseline and the approach proposed by Yao *et al*. [40]. More precisely, we implement and compare three variations: (**a**) fixed handcrafted weights defined by row-wise softmin of the angular distances between geographical coordinates; (**b**) a learned function that takes angular distances as input, which is similar to D^3G [40]; and (**c**) a direct, unconstrained parameterization of the affinity matrix followed by a row-wise softmax normalization. We train our multi-head architecture with each weighting strategy and then perform domain expert selection, reporting the performance in Tab. 4. Our unconstrained parameterization outperforms both the fixed angular weights and the adapted D^3G weighting scheme without requiring geographic metadata, indicating that inter-domain similarity may extend beyond purely geographic distance.

Impact of Backbone. In Fig. 5, we present the performance gains achieved by our method over the baseline on all four datasets for three different backbones: MultiUTAE (our default), ResNet-10 [10], and ResNet-18 [11]. Because ResNet-10 and ResNet-18 are designed for single-image processing rather than spatio-temporal image time series (SITS), we apply them image-by-image. We observe that our method consistently improves performance across all datasets and all three backbones.

Efficiency. Employing CoDEx with 55 domains (DynamicEarthNet) introduces approximately 213K additional parameters compared to baseline model, increasing training time by roughly 14% (3.1 min *vs*. 2.7 min per epoch). At inference, the overhead remains minimal, around 6% (7.1×10^{-3} s *vs*. 6.7×10^{-3} s per 10^6 pixels processed).

5. Conclusion

We introduced CoDEx, a new framework for domain generalization in satellite imagery, addressing a challenging problem in Earth observation. Our key insight is to train multiple expert models, each specialized for a given domain of the training set, enforcing consistency among them and learning to select the most suitable expert for unseen domains. We validated our approach on four satellite image and time-series benchmarks across three tasks, outperforming ten state-of-the-art methods in both spatial domain generalization and adaptation.

Acknowledgement. This work was supported by the European Research Council (ERC project DISCOVER, number 101076028) and by ANR project sharp ANR-23-PEIA-0008 in the context of the PEPR IA. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015600.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[3] Liang Chen, Yong Zhang, Yibing Song, Zhiqiang Shen, and Lingqiao Liu. Lfme: A simple framework for learning from multiple experts in domain generalization. *Advances in Neural Information Processing Systems*, 37:102919–102947, 2025. 2

[4] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *CVPR*, 2019. 2, 6

[5] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 2, 5

[6] Sanghyuk Chun and Song Park. StyleAugment: Learning texture de-biased representations by style augmentation without pre-defined textures. *arXiv preprint arXiv:2108.10549*, 2021. 2, 6

[7] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS*, 2018. 2, 5

[8] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, 2021. 5

[9] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 2

[10] Jiaming Gong, Wei Liu, Mengjie Pei, Chengchao Wu, and Liufei Guo. ResNet10: A lightweight residual network for remote sensing image classification. In *International Conference on Measuring Technology and Mechatronics Automation*. IEEE, 2022. 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[12] Yongjun He, Jinfei Wang, Chunhua Liao, Bo Shan, and Xin Zhou. Classhyper: Classmix-based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery. *Remote Sensing*, 14(4):879, 2022. 2

[13] Svetlana Illarionova, Sergey Nesteruk, Dmitrii Shadrin, Vladimir Ignatiev, Maria Pukalchik, and Ivan Oseledets. Mixchannel: Advanced augmentation for multispectral satellite images. *Remote Sensing*, 13(11):2181, 2021. 2

[14] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 1, 2, 5

[15] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*. PMLR, 2021. 1

[16] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 2, 6

[17] Ziyue Li, Kan Ren, Xinyang Jiang, Bo Li, Haipeng Zhang, and Dongsheng Li. Domain generalization using pretrained models without fine-tuning. *arXiv preprint arXiv:2203.04600*, 2022. 1

[18] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 2018. 2, 6

[19] Xiaoyan Lu, Yanfei Zhong, Zhuo Zheng, Junjue Wang, Dingyuan Chen, and Yu Su. Global road extraction using a pseudo-label guided framework: from benchmark dataset to cross-region semi-supervised learning. *Geo-spatial Information Science*, pages 1–19, 2024. 2

[20] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 2

[21] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, pages 9414–9423, 2021. 2

[22] Valerio Marsocci, Nicolas Gonthier, Anatol Garioud, Simone Scardapane, and Clément Mallet. Geomultitasknet: remote sensing unsupervised domain adaptation using geographical coordinates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2075–2085, 2023. 2

[23] Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. Timematch: Unsupervised

cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:301–313, 2022. 2

[24] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. ClassMix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021. 2, 6

[25] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *CVPR*, 2017. 3

[26] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the WILDS benchmark for unsupervised adaptation. In *ICLR*, 2022. 1, 2

[27] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020. 2

[28] Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *CVPR*, 2024. 1

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020. 2

[31] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *ECCV Workshops*. Springer, 2016. 2, 6

[32] Onur Tasar, Yuliya Tarabalka, Alain Giros, Pierre Alliez, and Sébastien Clerc. StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In *CVPR Workshops EarthVision*, 2020. 1, 2

[33] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *CVPR*, 2022. 2, 5

[34] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *CVPR*, 2021. 2, 5

[35] Elliot Vincent, Jean Ponce, and Mathieu Aubry. Pixel-wise agricultural image time series classification: Comparisons and a deformable prototype-based approach. *arXiv preprint arXiv:2303.12533*, 2023. 2

[36] Elliot Vincent, Jean Ponce, and Mathieu Aubry. Satellite image time series semantic change detection: Novel architecture and analysis of domain shift. *arXiv preprint arXiv:2407.07616*, 2024. 1, 5, 6

[37] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2, 6

[38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[39] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *arXiv e-prints*, pages arXiv–2403, 2024. 2, 6

[40] Huaxiu Yao, Xinyu Yang, Xinyi Pan, Shengchao Liu, Pang Wei Koh, and Chelsea Finn. Improving domain generalization with domain relations. In *ICLR*, 2024. 3, 4, 6, 8

[41] Yuan Yuan, Lei Lin, Qi Xin, Zeng-Guang Zhou, and Qingshan Liu. An empirical study on data augmentation for pixel-wise satellite image time series classification and cross-year adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 2

[42] Wankang Zeng, Ming Cheng, Zhimin Yuan, Wei Dai, Youming Wu, Weiquan Liu, and Cheng Wang. Domain adaptive remote sensing image semantic segmentation with prototype guidance. *Neurocomputing*, 580:127484, 2024. 2

[43] Valerie Zermatten, Xiaolong Lu, Javiera Castillo-Navarro, Tobias Kellenberger, and Devis Tuia. Land cover mapping from multiple complementary experts under heavy class imbalance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2

[44] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *ICCV*, 2021. 2, 6