Distribution Shifts at Scale: Out-of-distribution Detection in Earth Observation

Supplementary Material

In this supplement, we first detail the datasets and models used, followed by a discussion of the introduced distribution shifts and their design rationale. Next, we evaluate the impact of various design choices, including layer selection, downsampling methods, classifiers, and surrogate label assignment hyperparameters. Finally, we present additional experimental results with detailed visualizations and performance metrics to provide deeper insights into the behavior and performance of our method.

9. Datasets and Model Details

9.1. EuroSAT

EuroSAT [12] is a scene classification dataset derived from Sentinel-2 satellite images, covering various locations across Europe. It contains 27,000 images labeled into ten land-use and land-cover classes: Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea/Lake. The images have a spatial resolution of 10 meters.

We use a ResNet50 model pre-trained on ImageNet, modified to accept 13 input channels corresponding to Sentinel-2 spectral bands. The model is fine-tuned with a learning rate of 0.0001 and a batch size of 128. Training runs for up to 100 epochs with early stopping after 5 epochs of no improvement. Input images are normalized using channel-wise mean and standard deviation statistics.

9.2. xBD

xBD [10] is a semantic segmentation dataset for building damage assessment from satellite imagery. The dataset, collected from Maxar's Open Data Program, has images with a spatial resolution below 0.8 meters. It includes pre- and post-disaster images of hurricanes, floods, wildfires, and earthquakes, making it suitable for evaluating temporal and semantic shifts.

We simplify the damage assessment task into binary segmentation by reassigning damage levels: background (0) and levels 1-2 are grouped, while levels 3-4 form a high-damage class. This minimizes concept drift and ensures a fair evaluation of distribution shifts. We train a U-Net model with a ResNet50 backbone, pre-trained on ImageNet and configured for 3 input channels. Training uses a batch size of 32, a learning rate of 0.001, and runs for up to 50 epochs with early stopping after 5 epochs of no improvement. We reserve 10% of the data for validation and normalize the input images by dividing pixel values by 255.

9.3. FTW

We follow the practices of the original study and use a U-Net model with an EfficientNet-B3 backbone for semantic segmentation on the FTW dataset. The model is configured with 8 input channels and outputs 3 classes: background, field, and field-boundary. We use class weights of [0.04, 0.08, 0.88] to address class imbalance. The learning rate is set to 0.001, and the loss function is cross-entropy. The number of filters is set to 64, and neither the backbone nor the decoder is frozen during training. We set the patience for early stopping to 100 epochs. The images are normalized by dividing pixel values by 3000.

For the classifier, we use logistic regression with a maximum number of iterations set to 500. We train the classifier with 500 ID samples and 1200 WILD samples. The number of clusters is set to 150, calculated as 0.3 times the total number of WILD samples. To reassign labels, we use an ID fraction threshold of 0.1, meaning that a cluster is assigned as OOD if ID samples comprise less than 10% of the total samples in the cluster. The values of 0.3 and 0.1 are determined based on empirical observations gathered from extensive experiments on the xBD and EuroSAT datasets.

Figure 7 provides a visual illustration of the input samples from the WILD set, where the distribution is unknown. It displays the input Sentinel-2 image pair (Window A and Window B) alongside the OOD classifier g's prediction scores and the DL model f's predictions.

9.4. Introducing Distribution Shifts to EuroSAT and xBD

The combination of EuroSAT and xBD provides a diverse testbed for evaluating distribution shifts. EuroSAT represents regional imagery at medium spatial resolution, while xBD provides global imagery at very high resolution. Their differences in acquisition times, sensor parameters, processing levels, and the tasks they cover—land-cover classification (EuroSAT) and building detection (xBD)—make them complementary. Additionally, EuroSAT focuses on patch-level classification, while xBD involves pixel-level segmentation, enabling evaluations across different problem dimensions.

To evaluate our method, we introduce two types of distribution shifts: covariate and semantic (described in Table 1). Our approach assumes that purposefully rearranging dataset splits creates measurable shifts between training and testing sets, driven by the logic of the split design.

EuroSAT Distribution Shifts. Figure 6 shows one example from each of EuroSAT's 10 classes, which differ spatially and semantically. For covariate shifts, we split the dataset by longitude at the midpoint of its spatial extent, using the western half for training and the eastern half for testing. This creates a shift based on spatial proximity.

For semantic shifts, we train the model on 9 classes and test it on the hold-out class, repeating this process for all classes. This ensures the model faces unseen scenarios during testing, providing a robust evaluation of its ability to handle semantic shifts.

xBD Distribution Shifts. Figure 5 illustrates the pre- and post-disaster image pairs in the xBD dataset. Temporal shifts arise from changes occurring between pre- and post-disaster images, while spatial and thematic shifts reflect differences in how disasters impact regions and leave varying degrees of visible marks. Using these inherent characteristics, we design covariate shift experiments for xBD.

10. Design Choices

To better understand the impact of various design choices on the performance of our OOD detection method, we conduct a series of ablation studies. Specifically, we explore four key factors: (1) the choice of layer from which to extract feature representations (Section 10.1), (2) the method used to downsample these feature maps (Section 10.2), (3) the type of binary classifier g used to distinguish between surrogate-ID and surrogate-OOD samples (Section 10.3), and (4) the selection of hyperparameters k and T for surrogate label assignment (Section 10.4).

10.1. Which Layer?

Selecting the appropriate layer for activation extraction is crucial for accurate OOD detection. Prior works have emphasized the importance of this choice. For example, ASH achieves optimal performance on later layers like the penultimate layer, as earlier layers suffer from significant performance degradation during pruning [4]. Similarly, ReAct performs best on the penultimate layer, where more distinctive patterns between ID and OOD data emerge [30]. NAP-based OOD detection further highlights the variability in layer effectiveness, dynamically selecting top-performing layers based on validation accuracy [25]. Consistent with these findings, we observe that no single layer is universally optimal across all settings.

We benchmark FPR95 scores for OOD detection across the first convolutional layer, eight randomly selected intermediate layers, and the last convolutional layer. As shown in Table 4 for the EuroSAT dataset and Table 5 for the xBD dataset, layer performance varies significantly. While late layers often perform well, early and middle layers frequently give competitive results, depending on the dataset and task. Based on these findings, we select the best-performing layer for each experiment.

For the large-scale FTW dataset, the lack of distribution shift information prevents evaluation of layer-specific performance for OOD detection. Therefore, based on the observation that many layer benchmarks perform optimally for middle layers, we select a middle convolutional layer, specifically *'decoder.blocks.0.conv1'* from the U-Net model with an EfficientNet-B3 backbone.

10.2. Which Downsampling Method?

Having identified the layer to extract internal activations from, the next step is to look into the effect of downsampling these activations, which can reduce computational complexity and noise while retaining essential features for OOD detection. We explored four methods:

1. **Mean and standard deviation (*Mean Std*)**: Computes the mean and standard deviation across the spatial dimensions (H, W) for each channel, providing two descriptive statistics per feature channel.
2. **Average pooling (*Avg Pool*)**: Global average pooling was applied, reducing the activation to a single representative value per channel by averaging all spatial values.
3. **Max pooling (*Max Pool*)**: Uses global max pooling to retain the maximum value from each spatial dimension, capturing the most prominent feature in each channel.
4. **PCA-based reduction (*PCA*)**: Applies Principal Component Analysis to reshape the activation map into a vector and projects it into a lower-dimensional space with 10 components.

We summarize the OOD detection performance across all experiments on the EuroSAT and xBD datasets under different downsampling methods in Table 6, using the FPR95 metric. Max pooling consistently achieves the best performance across the majority of experiments, making it the preferred approach. We attribute its performance to its ability to retain the most prominent features in each channel, filtering out less significant information. This focus on salient patterns likely enhances the OOD classifier's capacity to distinguish between ID and OOD samples.

10.3. Which Classifier?

The next key design choice is the selection of the binary classifier g, used to distinguish between surrogate-ID and surrogate-OOD samples based on their feature representations. The results, summarized in Table 7, report the mean performance across all experimental measurements along with the standard error of the mean to represent confidence intervals. We select Logistic Regression as it provides the best tradeoff between classification accuracy and prediction time. This balance is essential for scaling up the method, where both efficiency and accuracy are critical.

Figure 5. Examples from the xBD dataset, illustrating pre- and post-disaster images. These samples demonstrate the temporal and semantic differences between pre- and post-disaster scenes, highlighting the challenges posed by distribution shifts.

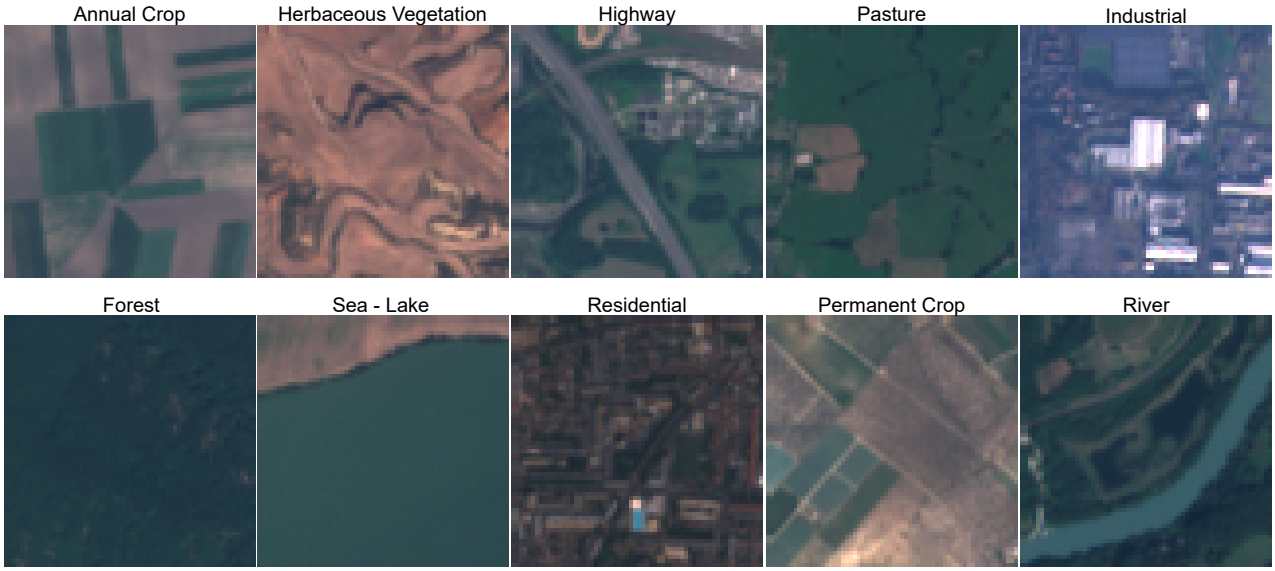

Figure 6. Examples from the EuroSAT dataset, with one sample from each class. These images highlight the spatial and semantic distinctions across classes.

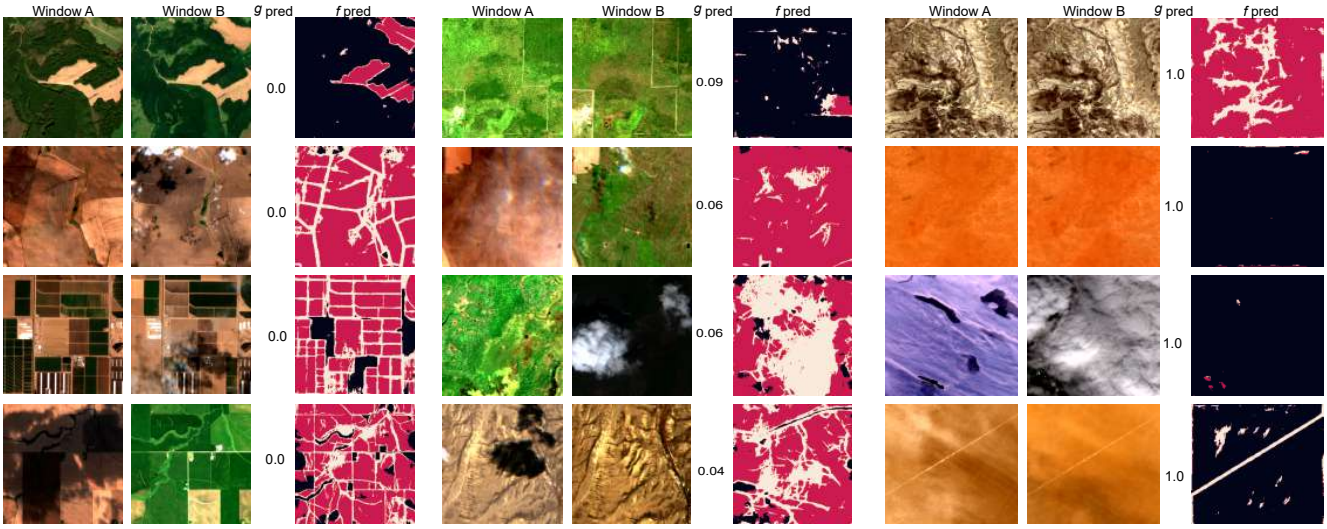

Figure 7. Deploying TARDIS over FTW dataset: The input samples are from the collected WILD set, where the distribution is unknown. The figure shows Sentinel-2 images at two different times (planting season and harvesting season — Window A and Window B). When these windows are fed together into f, the model outputs both the segmentation prediction and the OOD classifier g's prediction score.

10.4. Surrogate Label Assignment: Hyperparameter Search for k and T

TARDIS relies on a clustering-based approach in the activation space to assign surrogate ID and surrogate OOD labels. This process requires selecting two key parameters: the number of clusters (k) to segment the activation space, and the ID fraction threshold (T), which determines whether a cluster is assigned as surrogate ID or surrogate OOD. Clusters with an ID fraction above T are assigned as surrogate ID, and those below T are assigned as surrogate OOD.

The underlying assumption is that samples with similar distributions lie closer in the activation space than those from dissimilar distributions. Effectively clustering the activation space is critical, as the distributions of WILD samples are unknown during deployment, and it depends on the optimal choice of k and T.

To develop insights into selecting k and T, we conduct

controlled experiments on EuroSAT and xBD, where ID and OOD labels are known. In these experiments, we treat OOD labels as WILD and apply our clustering-based surrogate label assignment logic. By holding back the ground-truth WILD labels, we simulate real-world conditions while being able to evaluate the results against known labels.

The primary goal is to understand how to choose k and T, and whether there are patterns we can extrapolate to real-life deployment. For this, we first assign surrogate labels and calculate the ratio of k to the total number of training samples and evaluate its effect on OOD detection performance (Accuracy, FPR95, and AUROC). We plot these metrics against the ratio of k/total training samples, increasing k until the ratio reaches 1. Theoretically, OOD detection improves with more clusters as this enables finer-grained clustering of the activation space, reducing the risk of including anomalies in ID clusters.

To establish a theoretical maximum (upper-bound performance), we also evaluate OOD classification with known ID and OOD labels, bypassing the need for clustering. This oracle performance is represented by horizontal dashed lines in Figure 8 and Figure 9 (upper plots). The results for two representative experiments—one from EuroSAT and one from xBD—since all experiments show similar trends. We observe that the performance approaches the oracle boundaries when k is approximately 0.3 times the total number of training samples. While performance improves as k increases, a trade-off is required between performance and walltime as well as computational complexity. Based on this trade-off, we set k to 0.3 for all experiments, including the large-scale deployment on FTW. Furthermore, we observe that our method is not highly sensitive to T. As a result, we fix T to 0.1 for all experiments, which is the value used in this initial investigation. We use the Optuna library to implement a Bayesian-based search algorithm. The composite objective function, which we minimize to determine the optimal number of clusters and ID fraction threshold, is detailed in Section 4.

Lastly, the gradual improvement in OOD detection performance with increasing k supports our assumption that samples with similar distributions lie closer in the activation space than those with dissimilar distributions. The absence of degradation in performance further underscores the importance of activation-level clustering as a reliable proxy for domain estimation based on neighboring samples.

We set k and T as described and use t-SNE in the lower plots of Figure 8 and Figure 9 to reduce the dimensionality of the activation spaces to 2D for visualization. When ID and OOD labels are known, the t-SNE plots show that only a small fraction of labels changes from the original labels. This demonstrates the effectiveness of the surrogate label assignment process described above.

11. Further Experimental Results

In Figure 10, we show the predictions of the DL model f and the OOD classifier g, along with the ground truth class and distribution annotations for the EuroSAT experiment, where *Forest* serves as the OOD class. The model f trains on 9 classes (excluding *Forest*) and tests on *Forest*. The first row shows correct predictions by f, while the second row shows incorrect predictions. Even when f makes misclassifications, g accurately quantifies the distribution shifts in most cases. The performance of f on the test set is not directly measurable since the test uses a single unseen class. We report the performance of g as: Accuracy: 93.25%, ROC AUC: 98.86%, FPR95: 6.19%.

For xBD, we present results where f is trained on *Hurricane Matthew* (ID, Figure 11) and tested on *Mexico Earthquake* (OOD, Figure 12). Comparing the input images and masks between ID and OOD reveals that even when f performs suboptimally, g effectively quantifies the distribution shifts. The performance of f on the test set is as follows: Multi-class accuracy: 76.90%, Multi-class Jaccard index: 62.48%. We attribute f's suboptimal prediction performance to the significant distribution shift between the training (*Hurricane Matthew*) and testing (*Mexico Earthquake*) datasets, and also to the fact that we reformulate the main task of damage classification to building detection (as described in Section 9.2). The performance of g is: Accuracy: 98.06%, ROC AUC: 99.86%, FPR95: 0.00%.

Experiment	2/217	8/217	16/217	38/217	43/217	48/217	118/217	139/217	199/217	211/217
Forest	0.0625	0.01	0.00	0.00	0.00	0.00	0.0078	0.0156	0.0391	0.0391
HerbVeg	0.2857	0.22	0.3095	0.22	0.2778	0.1429	0.07	0.1032	0.2460	0.2778
Highway	0.8319	0.5462	0.6218	0.3529	0.3613	0.21	0.12	0.0840	0.1765	0.2437
Industrial	0.2406	0.01	0.0376	0.0226	0.0226	0.00	0.0150	0.0226	0.0376	0.0075
Pasture	0.1288	0.0909	0.12	0.1364	0.0985	0.03	0.0833	0.0227	0.1212	0.2273
PermCrop	0.3554	0.2975	0.3140	0.2397	0.2314	0.14	0.12	0.1322	0.2066	0.1653
Residential	0.2960	0.00	0.0160	0.0240	0.00	0.00	0.00	0.0160	0.0400	0.0480
River	0.4688	0.07	0.2031	0.03	0.0938	0.0234	0.0078	0.0078	0.0625	0.0859
SeaLake	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnnualCrop	0.2879	0.00	0.0303	0.0606	0.0530	0.02	0.0379	0.0152	0.0682	0.0758
SpatialSplit	0.3182	0.20	0.4773	0.15	0.1970	0.0909	0.2197	0.2273	0.6364	0.7652
Avg ± Stdev	0.2978 ± 0.2223	0.1313 ± 0.1729	0.1936 ± 0.2133	0.1124 ± 0.1176	0.1214 ± 0.1264	0.0597 ± 0.0741	0.0620 ± 0.0696	0.0588 ± 0.0712	0.1486 ± 0.1801	0.1760 ± 0.2185

Table 4. FPR95 scores for OOD detection for experiments on EuroSAT dataset across the first convolutional layer, eight randomly selected layers, and the last convolutional layer. Notation in the header (e.g., X/Y) refers to the 'layer number / total number of layers.' The last row, labeled 'Avg ± Stdev,' provides the mean ± standard deviation of the scores for each layer across all experiments.

Experiment	3/223	60/223	112/223	146/223	148/223	162/223	176/223	187/223	208/223	216/223
Nepal Flooding - Midwest Flooding	1.0000	0.09	0.6986	0.7534	0.6986	0.7397	0.46	0.8767	0.8356	0.8493
Santa Rosa Wildfire - Woolsey Fire	0.5000	0.36	0.36	0.7222	0.6389	0.6389	0.6944	0.6111	0.58	0.9167
Hurricane Matthew - Nepal Flooding	0.3023	0.06	0.3488	0.4419	0.4884	0.32	0.1163	0.3488	0.2791	0.4419
Hurricane Matthew - Mexico Earthquake	0.11	0.1471	0.2353	0.6471	0.4118	0.50	0.3235	0.4706	0.6471	0.9706
Portugal Wildfire (Pre-Post)	0.38	0.9583	0.3889	0.8472	0.8750	0.9167	0.9722	0.77	0.9861	0.8472
Mean ± Stdev	0.4585 ± 0.3343	0.3231 ± 0.3740	0.4063 ± 0.1735	0.6824 ± 0.1524	0.6225 ± 0.1818	0.6231 ± 0.2275	0.5133 ± 0.3316	0.6154 ± 0.2146	0.6656 ± 0.2686	0.8051 ± 0.2095

Table 5. FPR95 scores for OOD detection experiments on the xBD dataset across the first convolutional layer, eight randomly selected layers, and the last convolutional layer. Notation in the header (e.g., X/Y) refers to the 'layer number / total number of layers.' The last row, labeled 'Avg ± Stdev,' provides the mean and standard deviation of the scores for each layer across all experiments.

Experiment	Avg Pool	Mean Std	Max Pool	PCA
Forest	*0.0859*	0.4297	**0.0234**	0.8750
HerbaceousVegetation	*0.2937*	0.8651	**0.2698**	0.9921
Highway	0.7899	*0.7311*	**0.7059**	0.9412
Industrial	*0.1880*	0.2857	**0.0526**	0.9925
Pasture	**0.0909**	0.6970	*0.2576*	0.9924
PermanentCrop	**0.3058**	0.9008	*0.4215*	0.9669
Residential	*0.2640*	0.6640	**0.2160**	0.8480
River	*0.4922*	0.6172	**0.1563**	0.9922
SeaLake	**0.0000**	0.0313	**0.0000**	0.9766
AnnualCrop	*0.2500*	0.7576	**0.0455**	0.9924
SpatialSplit	*0.3182*	0.5379	**0.3030**	0.9848
Nepal Flooding - Midwest Flooding	**0.0000**	*0.6575*	0.9452	0.9726
Hurricane Matthew - Nepal Flooding	**0.0233**	0.9070	*0.5581*	0.9535
Hurricane Matthew - Mexico Earthquake	**0.0588**	0.9118	*0.5882*	1.0000
Portugal Wildfire Pre - Portugal Wildfire Post	*0.9028*	0.9861	**0.8472**	1.0000

Table 6. FPR95 scores for OOD detection across different downsampling methods. The table compares performances of average pooling, mean and standard deviation pooling, max pooling, and PCA for various experiments. Bold values indicate the best performance for each experiment, while italicized values represent the second-best performance.

Classifier	Accuracy↑	ROC_AUC↑	FPR95↓	Prediction Time (ms/sample)
KNeighbors	92.23 ± 0.81	86.85 ± 1.07	*38.79 ± 2.02*	73.00 ± 8.00
GaussianNB	84.28 ± 1.04	89.37 ± 0.91	32.89 ± 1.89	4.00 ± 1.00
DecisionTree	91.21 ± 0.93	77.43 ± 1.20	78.81 ± 2.41	**2.00 ± 1.00**
ExtraTrees	**93.30 ± 0.67**	91.29 ± 0.83	*28.53 ± 1.94*	12.00 ± 3.00
LogisticRegression	87.67 ± 1.00	*93.33 ± 0.87*	27.20 ± 1.98	*3.00 ± 1.00*
SVC	91.54 ± 0.90	**94.26 ± 0.72**	**19.87 ± 1.85**	67.00 ± 12.00
RandomForestUnbalanced	92.54 ± 0.82	91.24 ± 0.80	30.98 ± 1.95	9.00 ± 2.00
RandomForest	*92.76 ± 0.75*	91.11 ± 0.84	30.24 ± 1.91	8.00 ± 2.00
AdaBoost	92.85 ± 0.79	92.03 ± 0.82	29.84 ± 1.89	11.00 ± 3.00
GradientBoosting	92.92 ± 0.81	93.11 ± 0.85	30.47 ± 1.88	7.00 ± 2.00

Table 7. Benchmark results of classifiers g, including Accuracy, ROC AUC, FPR95, and prediction time. Values are reported as mean ± SEM over all experiments on EuroSAT and xBD. Bold indicates the best performance, and italics indicate the second-best performance. Prediction time is reported in milliseconds (ms/sample).

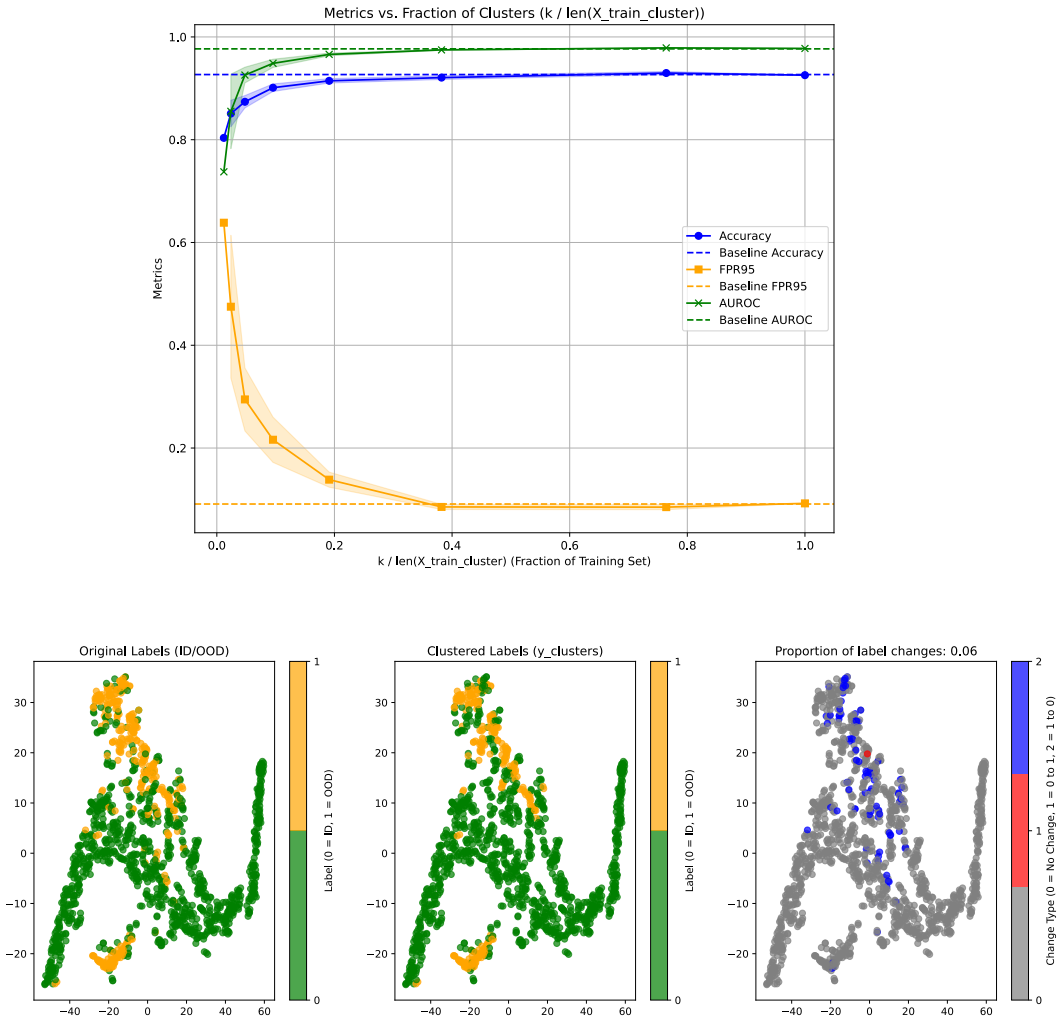

Figure 8. EuroSAT *Pasture* experiment on surrogate label assignment. The upper plot shows the performance metrics (Accuracy, FPR95, AUROC) for the oracle classifier g_{oracle} and the surrogate classifier g^* as the ratio of clusters to training samples $k/\text{len}(X_{\text{train}})$ increases. As k grows, g^* gradually improves and approaches the performance of g_{oracle}. The lower plot visualizes the feature space before and after clustering, showing how original ID and OOD labels are reassigned to surrogate ID and OOD labels based on the clustering logic.

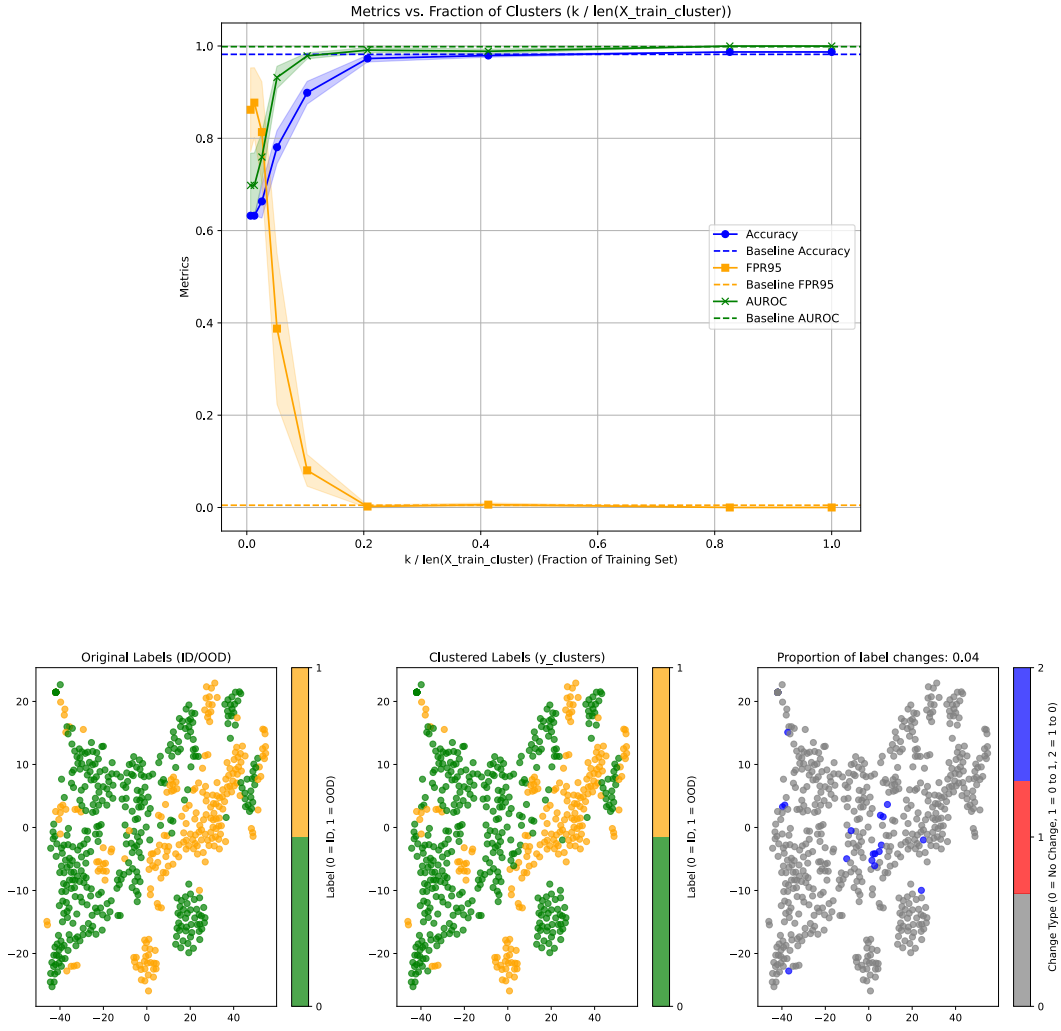

Figure 9. xBD *Nepal Flooding-Midwest Flooding* disaster experiment on surrogate label assignment. The upper plot shows the performance metrics (Accuracy, FPR95, AUROC) for the oracle classifier g_{oracle} and the surrogate classifier g^* as the ratio of clusters to training samples $k/\text{len}(X_{\text{train}})$ increases. As k grows, g^* gradually improves and approaches the performance of g_{oracle}. The lower plot visualizes the feature space before and after clustering, showing how original ID and OOD labels are reassigned to surrogate ID and OOD labels based on the clustering logic.

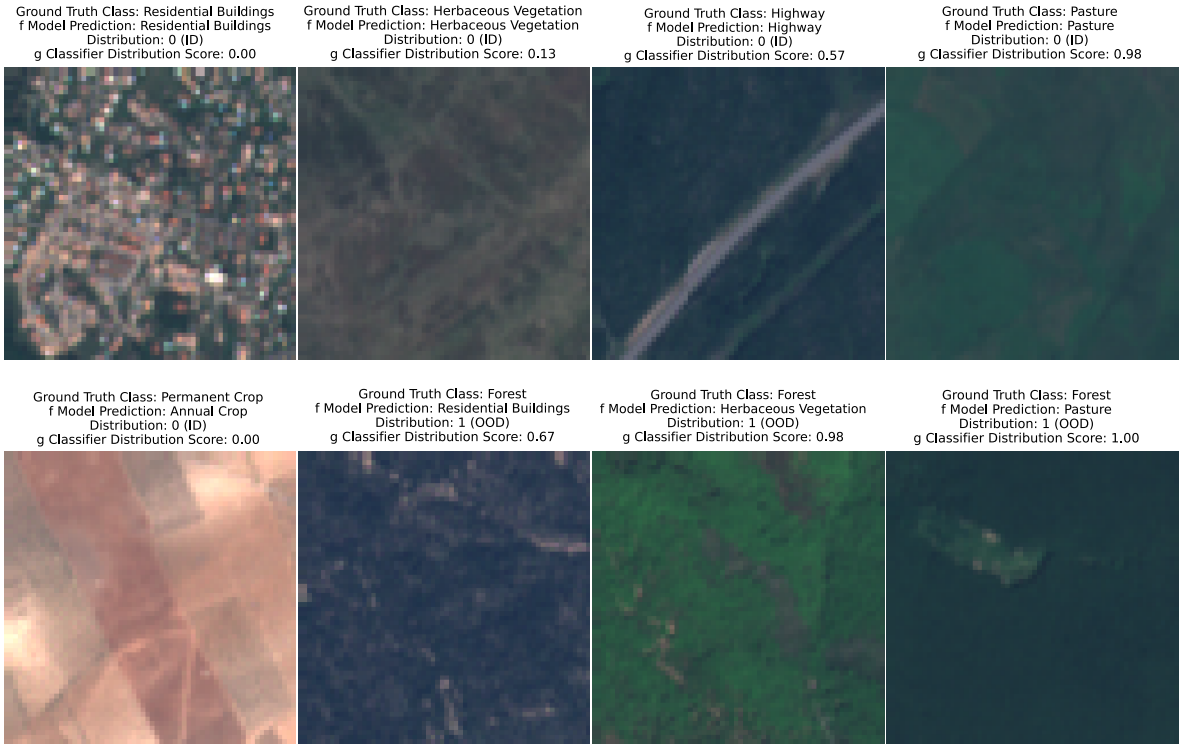

Figure 10. EuroSAT experiment with *Forest* as the OOD class. The figure shows predictions of the DL model f and the OOD classifier g, along with the ground truth class and distribution annotations. The first row represents samples where f makes correct class predictions, while the second row represents samples where f makes incorrect predictions. For each sample, we report both the ground truth distribution and the predicted distribution from g.

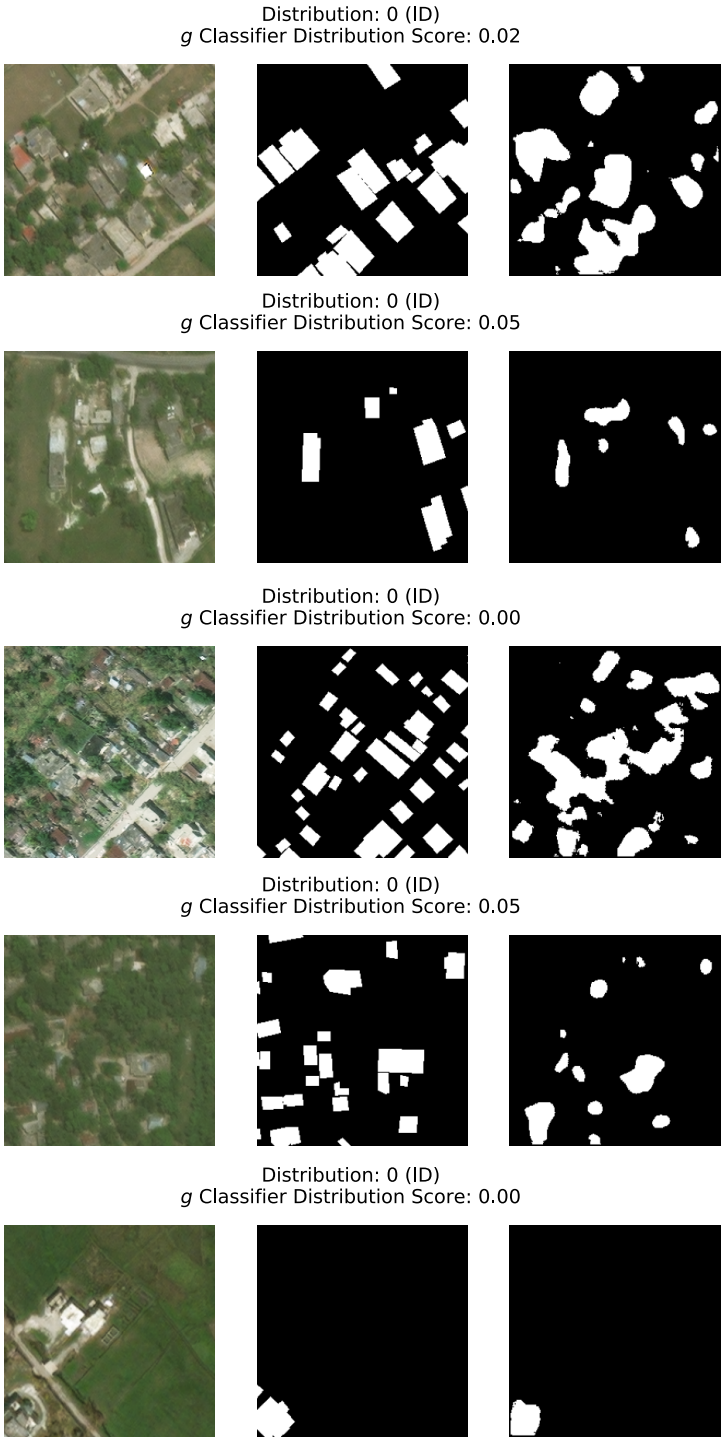

Figure 11. xBD experiment with *Hurricane Matthew* as the ID samples. The figure shows the annotations and predictions of the DL model f and the OOD classifier g. For each sample, we present f's predicted class and g's predicted distribution, along with the ground truth annotations.

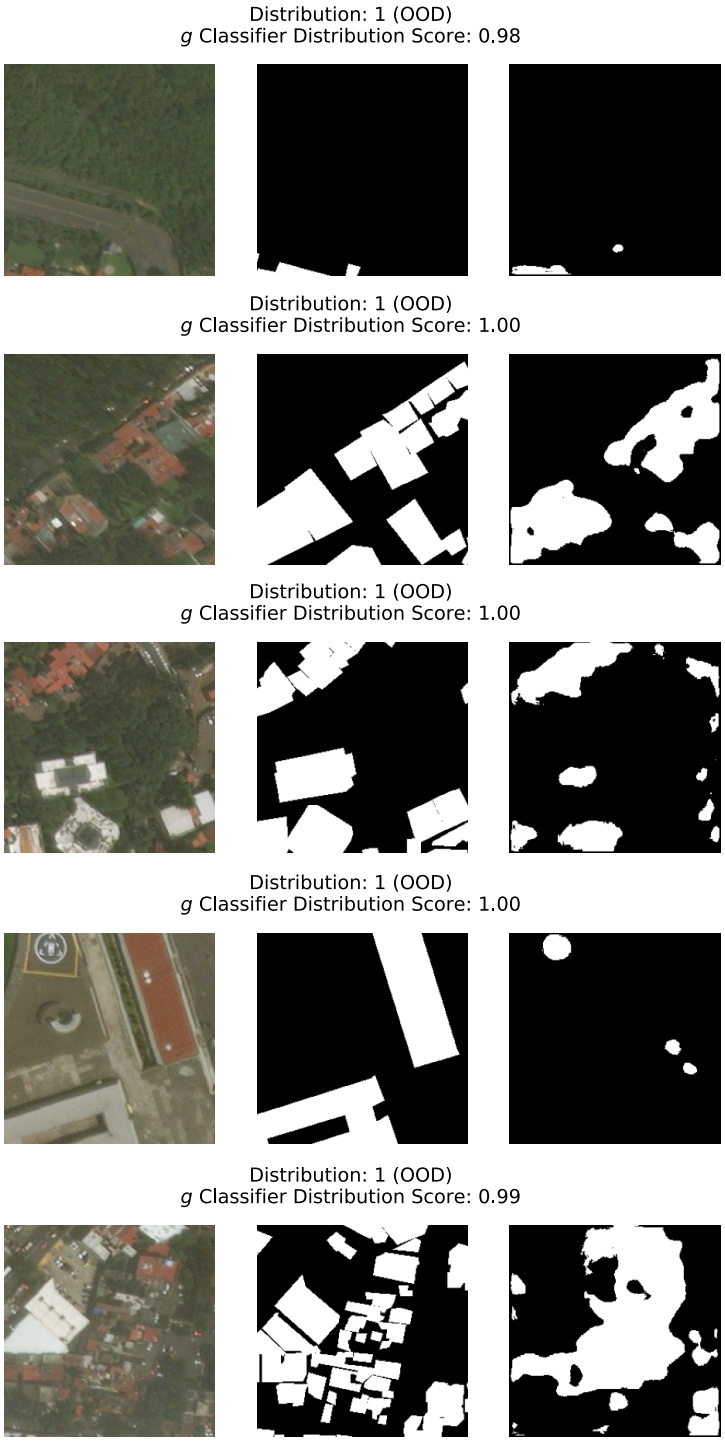

Figure 12. xBD experiment with *Mexico Earthquake* as the OOD samples. The figure shows the annotations and predictions of the DL model f and the OOD classifier g. For each sample, we present f's predicted class and g's predicted distribution, along with the ground truth annotations.