Generalizable Slum Detection from Satellite Imagery with Mixture-of-Experts

Sumin Lee[*1,2], **Sungwon Park**[*1,2], **Jeasurk Yang**[†1], **Jihee Kim**[†2], **Meeyoung Cha**[†1,2]

[1] Max Planck Institute for Security and Privacy (MPI-SP)
[2] Korea Advanced Institute of Science and Technology (KAIST)

Abstract

Satellite-based slum segmentation holds significant promise in generating global estimates of urban poverty. However, the morphological heterogeneity of informal settlements presents a major challenge, hindering the ability of models trained on specific regions to generalize effectively to unseen locations. To address this, we introduce a large-scale high-resolution dataset and propose **GRAM** (Generalized Region-Aware Mixture-of-Experts), a two-phase test-time adaptation framework that enables robust slum segmentation without requiring labeled data from target regions. We compile a million-scale satellite imagery dataset from 12 cities across four continents for source training. Using this dataset, the model employs a Mixture-of-Experts architecture to capture region-specific slum characteristics while learning universal features through a shared backbone. During adaptation, prediction consistency across experts filters out unreliable pseudo-labels, allowing the model to generalize effectively to previously unseen regions. GRAM outperforms state-of-the-art baselines in low-resource settings such as African cities, offering a scalable and label-efficient solution for global slum mapping and data-driven urban planning.

Datasets — https://github.com/DS4H-GIS/GRAM

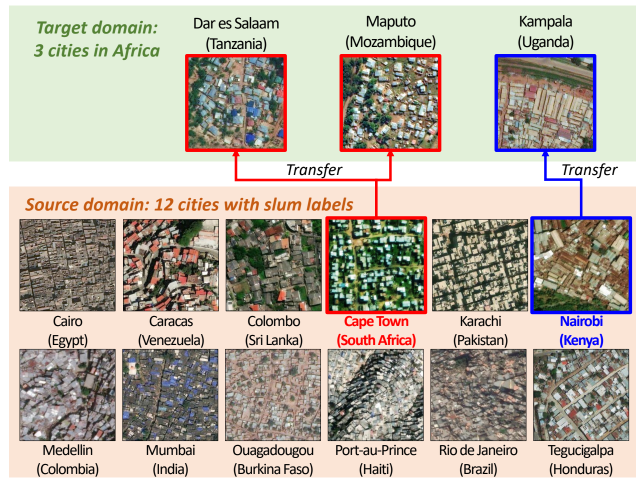

Figure 1: Our source domain spans 12 cities across four continents; the target domain covers three African cities. Morphological similarities guide the model to prioritize Cape Town features for detecting slums in Dar es Salaam and Maputo (red), and Nairobi features for Kampala (blue).

Introduction

In 2003, UN-Habitat introduced a globally standardized definition to describe deprived urban settlements—commonly referred to as *slums*—marking a pivotal shift in how urban poverty was conceptualized and measured worldwide (UN-Habitat 2003). This universal framework, formally adopted by the United Nations, was established to identify and quantify urban deprivation. It defines slums as settlements lacking one or more basic living conditions, such as durable housing, sufficient living space, access to safe water and sanitation, and secure tenure. This standardized definition enabled the production of comparable global estimates of slum populations, shaping major policy agendas like the Sustainable Development Goals (SDGs) and supporting cross-national research efforts (UN-Habitat 2025).

Despite continued efforts, the worsening socioeconomic conditions in slums continue to present substantial challenges (Kohli et al. 2012). Measuring these conditions is complex, as the physical form and living conditions vary widely—not only across continents but also within individual countries—reflecting differences in geography, governance, cultural norms, and historical trajectories of urbanization (Taubenböck and Kraff 2014). These localized differences make it difficult to apply standardized criteria for identification and comparison, thereby complicating conventional data collection methods (Yang et al. 2025a). A key limitation of traditional approaches to identifying slum settlements—such as surveys, censuses, and field assessments—is their variability in definitions and implementation. These methods differ across countries in how they conceptualize "slum," the spatial scales they employ, and the degree to which political and institutional factors shape data collection. They are also resource-intensive to conduct. As a result, global estimates remain uneven, and cross-national comparisons are constrained by divergent defini-

[*]Equal contribution to this work.
[†]Co-corresponding authors.

tions, methodologies, and practices on the ground (Simon 2011).

Recent research has increasingly turned to satellite data and deep learning to measure urban deprivation (Kuffer, Pfeffer, and Sliuzas 2016). High-resolution remote sensing data provide a globally accessible, frequently updated, and non-intrusive source of information, while convolutional neural networks and other computer vision techniques have shown promise in detecting spatial patterns associated with slum conditions (Kit and Lüdeke 2013; Wurm et al. 2019; Duque, Patino, and Betancourt 2017). Although promising, developing a computational approach is non-trivial due to visual heterogeneity (Verma, Jana, and Ramamritham 2019; Stark et al. 2020). Models trained on imagery from one country fail to generalize to other regions, as architectural styles, building materials, and spatial organization of slums differ (Figure 1). As a result, cross-country and regional applications suffer from limited transferability and unreliable performance (Stark et al. 2024).

To transfer knowledge from labeled source regions to unlabeled targets, we propose **GRAM**, a test-time adaptation (TTA) framework for slum segmentation. Unlike conventional methods that require annotated samples for targets, GRAM dynamically adapts a model pre-trained on source domains to the distributional characteristics of the test imagery for cross-regional generalization. To implement this idea, we construct a new slum segmentation dataset from very-high-resolution (VHR) satellite imagery of 12 cities in four continents, each representing distinct morphological, architectural, and socio-spatial patterns of informal settlements. These cities encompass diverse urban forms and regional contexts, offering a strong foundation for learning transferable representations. Using this multi-continent dataset and adaptation at inference time, we show how GRAM bridges the gap between universal slum definitions and region-specific visual characteristics.

The framework comprises two distinct training components. The first is the source model training phase, where Mixture-of-Experts (MoE) layers are integrated into the segmentation backbone. This architecture organizes the model into expert groups specialized in learning region-specific slum characteristics, while a shared backbone captures universal features across all regions. The second component is the target adaptation phase. For each unlabeled target image, a classifier identifies the source region with the most similar visual characteristics. The pseudo-label generated from that corresponding expert is used as a primary prediction. The model then assesses the image's reliability by evaluating the consistency of this reference against the predictions from all other experts. Images with the highest cross-expert agreement, measured by a "stability score," are selected as reliable, effectively filtering out noise from uncertain pseudo-labels before the model is fine-tuned.

We validate GRAM using a diverse benchmark of slum segmentation tasks in three African cities with distinct urban morphologies and socio-spatial patterns. Our model consistently outperforms state-of-the-art baselines in low-resource settings, demonstrating its potential as a scalable and label-efficient solution for slum monitoring. These findings highlight the practical value of TTA for enabling more inclusive, data-driven urban development policies.

Related Work

Satellite Data

Satellite imagery provides a scalable and efficient alternative to traditional ground-based surveys of slums (Ahn et al. 2023; Han et al. 2020a,b). Early approaches relied on proxy indicators, such as nighttime light intensity, to infer socioeconomic conditions from medium-resolution data (Jean et al. 2016; Park et al. 2022). With advances in deep learning and VHR imagery, subsequent research shifted toward pixel-level segmentation. For example, convolutional neural networks (CNNs) have been adapted for semantic segmentation of slum areas (Lumban-Gaol, Rizaldy, and Amadi 2023). Comparative studies have assessed various deep learning models in capturing the distinctive morphological characteristics of informal settlements, such as high-density and irregularly shaped structures (Gadiraju et al. 2018; Leonita et al. 2018). Transfer learning (Verma, Jana, and Ramamritham 2019; Wurm et al. 2019), semi-supervised learning (Rehman et al. 2022; Lin et al. 2024), and foundation model (Zhang et al. 2024) have also been applied, enabling models trained on large-scale datasets to adapt to slum segmentation tasks with limited labels.

In terms of data quality, both high- and medium-resolution imagery are being used for slum detection. One study explored the trade-offs between spatial detail and data accessibility (Verma, Jana, and Ramamritham 2019). Low-resolution multispectral data have also helped extend coverage in areas lacking VHR imagery (Gram-Hansen et al. 2019). Additionally, machine learning–based slum mapping supports urban improvement efforts and enables monitoring of informal settlement dynamics over time (Leonita et al. 2018; Maiya and Babu 2018; Yang et al. 2025b). Despite recent advances, segmenting slum settlements remains difficult due to significant variation in urban morphology across geographies. The scarcity of large-scale, pixel-level annotated datasets limits the development of fully supervised models. This constraint continues to drive research into developing new approaches including TTA that leverages abundant unlabeled satellite imagery to address data gaps.

Test-Time Adaptation (TTA)

Unsupervised domain adaptation improves model generalization by adapting a model trained on labeled data to a different, unlabeled domain with similar structure but differing data distribution. TTA builds on this idea under the stricter constraint that the source data is inaccessible during adaptation. Instead, TTA methods adapt a pre-trained source model *during inference* using only the test sample data (Ahn et al. 2025; Liang, He, and Tan 2025).

Two widely used strategies include batch normalization (BN) adaptation and entropy minimization. In BN-based approaches (Nado et al. 2020), the running statistics of the BN layers, originally computed during training, are updated using statistics from the target data, without requiring gradient computation (Ioffe and Szegedy 2015). This provides a lightweight yet effective adaptation mechanism. Entropy-based methods aim to reduce the task uncertainty by minimizing the entropy of predictions on target samples. A typical approach is to fine-tune parameters, such as the affine components of BN layers, through backpropagation at test time (Wang et al. 2021). While both strategies perform well on static domains, they face challenges in dynamic environments where target distributions shift over time or multiple domains are encountered sequentially. Under such conditions, performance tends to degrade due to accumulated errors and catastrophic forgetting. Recent work has explored continual adaptation to enhance model robustness across evolving and diverse target domains (Wang et al. 2022; Lee, Yoon, and Hwang 2024).

Data

For detecting slums, we construct a diverse labeled dataset of 12 cities listed below to cover a broad representation of urban morphologies and socio-economic conditions:

- **Africa (4 cities)**: Cairo in Egypt, Cape Town in South Africa, Nairobi in Kenya, Ouagadougou in Burkina Faso
- **Asia (3 cities)**: Colombo in Sri Lanka, Karachi in Pakistan, and Mumbai in India
- **South America (3 cities)**: Caracas in Venezuela, Medellín in Colombia, and Rio de Janeiro in Brazil
- **Central America (2 cities)**: Port-au-Prince in Haiti, and Tegucigalpa in Honduras

Satellite imagery of these selected urban centers, sourced from the ESRI World Imagery Wayback[1], is preprocessed into uniform 256×256-pixel tiles at zoom level 16, corresponding to an approximate spatial resolution of 1.2 meters per pixel, subject to latitudinal variation.

Ground-truth annotations are from the Atlas of Informality (Samper 2025), a global mapping initiative that documents the growth and transformation of informal settlements. We supplemented the labels with city-specific datasets (University of Edinburgh 2020; Earth Observation for Sustainable Development 2018; Slum Rehabilitation Authority 2016; Alcaldia de Medellin 2010; Prefeitura da Cidade do Rio de Janeiro 2021), each of which provides spatial delineations of informal settlements. Based on this resource, binary masks are manually generated by geography experts, where pixels corresponding to informal housing areas are assigned a value of 1, and all remaining pixels are set to 0. Each annotation is spatially aligned with its corresponding satellite tile, resulting in a 1×256×256 label image. An illustrative example is provided in Figure 2.

Using the manually labeled subset, we initially train a semi-supervised segmentation model based on the ST++ architecture (Yang et al. 2022). The model is trained on the full dataset comprising 2,714,489 image tiles, of which 86,752 have ground-truth annotations. This approach facilitates the integration of both labeled and unlabeled data to enhance segmentation performance under limited supervision. Subsequently, we use the model's predictions to generate pseudo-labels across the dataset. These pseudo-labels are then used to supervise the training of a fully supervised baseline model, enabling the use of a broader and more heterogeneous image set.

[1] https://livingatlas.arcgis.com/wayback/

Figure 2: Example of labeled satellite imagery used for training. (a) Original satellite image tile. (b) Corresponding binary mask overlaid in red, indicating slum areas (label=1), while black regions denote non-slum areas (label=0).

To test generalizability, we hold out a separate set of cities from the training process and use them exclusively for TTA and evaluation. This evaluation set comprises Dar es Salaam (Tanzania), Kampala (Uganda), and Maputo (Mozambique), totaling 529,633 image tiles, of which 17,536 are manually annotated using the AoI Database. The dataset details can be found in the Appendix.

To facilitate further research in slum detection and urban analysis, we release the dataset utilized in this study publicly available. This release includes both manually annotated segmentation labels and GRAM-derived masks across our 12 training and 3 testing cities. As of 2025, these 15 cities have an aggregated population of 120,174,837, with 215,148 informal settlement polygons identified by GRAM. This extensive coverage enables robust, cross-regional analyses of urban informality at scale. Each image tile has a spatial resolution of 10 meters per pixel at 256×256 pixels, with corresponding geospatial coordinates aligned to the satellite imagery.

Method

Overview. Let $\mathcal{D}_s$ denote the source training dataset and $\mathcal{D}_t$ the target test dataset. Each sample in the source dataset is a triplet $\{x, y, d\} \in \mathcal{D}_s$, where $x \in \mathbb{R}^{3\times H\times W}$ is a satellite image, $y \in \{0,1\}^{H\times W}$ is the corresponding ground truth slum segmentation mask, and $d \in \{1, \ldots, D\}$ is the region label, with D denoting the total number of source regions. Our goal is to develop a TTA framework that enables a source-trained model to generalize effectively to unseen target region images $x \in \mathcal{D}_t$.

Slum segmentation poses substantial difficulties due to the high variability in visual characteristics across geographies, cultures, and imaging conditions. This heterogeneity hampers the ability of a single model to generalize across domains. We propose GRAM, a two-stage training framework leveraging a MoE architecture to enhance adaptability and robustness in diverse environments, as outlined below:

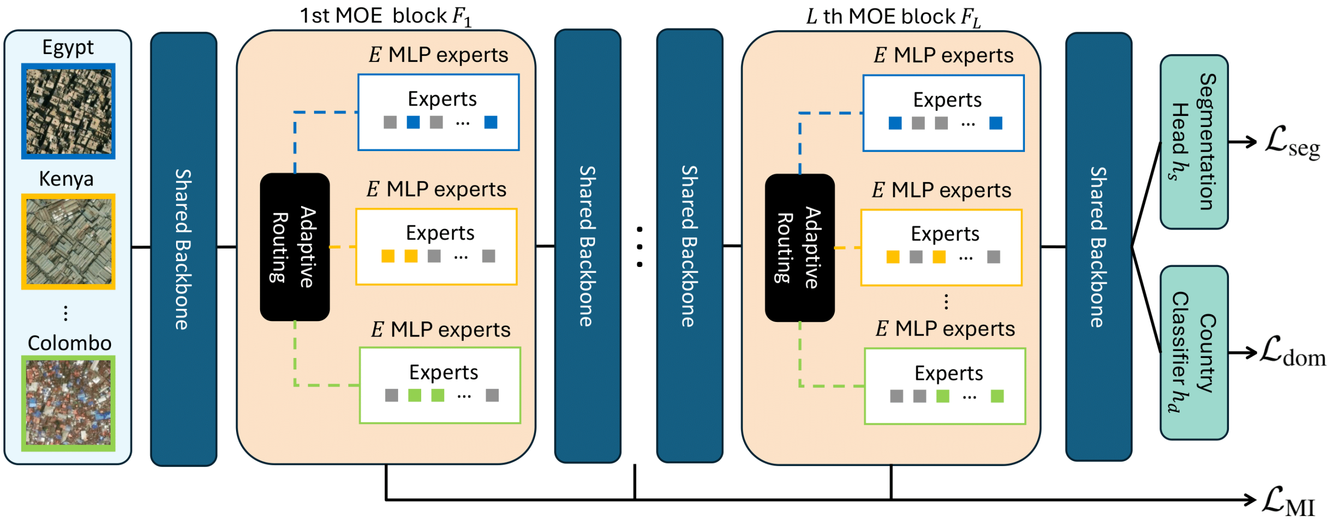

Figure 3: Overview of the Mixture-of-Experts (MoE) architecture in GRAM. The diagram illustrates the integration of lightweight MoE blocks $\mathcal{F}$ into the transformer encoder, with region-specific gating networks g_d dynamically routing token features z to a top-k subset of expert adapters $\{\mathcal{E}_e\}_{e=1}^{E}$.

- **Step 1.** Train the segmentation model f_θ using modular expert routing and region-aware learning on the multi-region slum dataset $\mathcal{D}_s$.
- **Step 2.** Adapt the model f_θ on $\mathcal{D}_t$ using pseudo-labels generated with routing guided by an external classifier h_ψ, filtering out unreliable labels via cross-region prediction consistency, and fine-tuning on a high-stability subset $\bar{\mathcal{D}}_t$.

Step 1. Source Training with Mixture-of-Experts

While slums often share common morphological features—such as dense roof coverage and spatial layout—they also display region-specific variations in roof materials and construction styles. To disentangle these localized features from globally shared representations, we integrate lightweight Mixture-of-Experts (MoE) blocks $\mathcal{F}$ into L intermediate layers of the transformer encoder, followed by a segmentation head h_s (i.e., $f_\theta = h_s \circ \mathcal{F}_L \circ \cdots \circ \mathcal{F}_1$), as illustrated in Figure 3. Each MoE block $\mathcal{F}$ contains a set of lightweight MLP expert adapters $\{\mathcal{E}_e\}_{e=1}^{E}$, where E is the number of experts. These experts capture region-specific slum characteristics, while the shared transformer backbone learns generalizable representations across diverse geographical contexts.

Adaptive Expert Routing within MoE Blocks To enable region-specific specialization, we initialize a lightweight gating network g_d for each source region d. Given a token feature z extracted from an input image $x \in \mathcal{D}_s$, each MoE block dynamically selects a top-k subset of experts for z using a noisy top-k routing strategy. The gating network for region d computes logits $g_d(z) \in \mathbb{R}^E$, where each element indicates the relevance of an expert for processing z. To promote diversity and prevent overconfident expert selection (Chen et al. 2023), we add Gaussian noise to the logits:

$$\tilde{g}_d(z) = g_d(z) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

After injecting noise, we select the top-k experts with the highest scores. Then, a softmax is applied over just these top-k scores to produce a normalized set of weights α, which determine how much each selected expert contributes:

$$\boldsymbol{\alpha} = \text{Softmax}\left(\tilde{g}_d(z)_{\text{top-}k}\right), \tag{2}$$

$$\text{MoE}(z) = \sum_{e \in \text{top-}k(\tilde{g}_d(z))} \alpha_e \cdot \mathcal{E}_e(z), \tag{3}$$

where $\mathcal{E}_e(z)$ is the output of expert e-th expert and α_e reflects how much that expert should contribute for token z. This routing mechanism allows the model to flexibly adapt to different region by selectively activating experts that are most relevant to the token's region-specific context.

Region-Aware Regularization In a MoE setup, the model learns to route tokens to different experts. Without proper regularization, experts may converge to similar behaviors, undermining specialization. To encourage diversity, we introduce a mutual information (MI)-based regularization term that promotes distinct expert selections across regions.

During training, we estimate the joint distribution $P^l(d, e)$ for l-th MoE layer, representing the frequency of expert e being selected for samples from region d in that layer. The mutual information between domains and experts for each layer is:

$$I^l(d; e) = \sum_{d=1}^{D} \sum_{e=1}^{E} P^l(d, e) \log \left(\frac{P^l(d, e)}{P^l(d) P^l(e)} \right). \tag{4}$$

By maximizing $I^l(d; e)$ for each layer, we encourage a strong dependence between the region and the selected experts, ensuring that different regions activate distinct sets of experts across the MoE blocks. This fosters expert specialization and prevents mode collapse, where experts learn redundant representations. To achieve this, we minimize the loss $\mathcal{L}_{MI} = -\sum_{L=1}^{E} I^l(d;e)$, which effectively maximizes the mutual information across all layers and promotes the desired domain-expert alignment.

Additionally, to complement the MI regularization and further promote expert specialization, we incorporate a lightweight region classifier h_d into the training framework. This classifier operates on intermediate features extracted from the shared backbone, encouraging the model to learn domain-discriminative representations that enhance the dependence between cities and expert selections. It is trained to predict the region label using standard cross-entropy loss $\mathcal{H}$:

$$\mathcal{L}_{\text{dom}} = \frac{1}{|\mathcal{D}_s|} \sum_{(x,d)\in\mathcal{D}_s} \mathcal{H}(d, h_d(x)), \tag{5}$$

where $h_d(x)$ denotes the predicted region probabilities for the sample x. This auxiliary supervision synergizes with the MI term by improving the quality of features used in routing, thereby preventing mode collapse and fostering region-specific expertise in the MoE layers.

Training Objectives For segmentation supervision, we apply pixel-wise cross-entropy loss between the predicted segmentation map $f_\theta(x)$, and the ground truth y. The segmentation loss over the entire source dataset $\mathcal{D}_s$ is

$$\mathcal{L}_{\text{seg}} = \frac{1}{|\mathcal{D}_s|} \sum_{(x,y,d)\in\mathcal{D}_s} \mathcal{H}_p(y, f_\theta(x,d)), \tag{6}$$

where the cross entropy loss $\mathcal{H}_p$ is averaged over all pixels and classes. Finally, the overall training objective combines segmentation accuracy, expert diversity, and domain awareness:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{MI}} \cdot \mathcal{L}_{\text{MI}} + \lambda_{\text{dom}} \cdot \mathcal{L}_{\text{dom}}$$

where λ_{MI} and λ_{dom} are hyperparameters controlling the influence of each regularization component.

Step 2. Target Adaptation with Pseudo Label

Given the trained source model f_θ, we perform TTA on the target dataset $\mathcal{D}_t$ by generating pseudo-labels for unlabeled target images. An external region classifier h_ψ predicts the source region index most similar to the target region, which is used to guide the routing in f_θ predicts segmentation masks for samples in $\mathcal{D}_t$. The model is refined using self-training with these pseudo-labeled samples, minimizing a segmentation loss. However, the variability in slum characteristics introduces significant distribution shifts relative to the source data, potentially leading to unreliable pseudo-labels from the source model trained on a different distribution. To address this, we propose an image-level selection strategy that filters out the unreliable pseudo label via consistency of predictions across region-routed experts, followed by self-training to refine the model by minimizing a segmentation loss on the selected pseudo-labeled samples (Park et al. 2021; Yang et al. 2022).

Adaptive Self-Training via Cross-Region Prediction Consistency For a target sample $x \in \mathcal{D}_t$, we first infer the target region index d_t using the external classifier h_ψ. Using this index d_t, we generate a pseudo-labeled dataset $\bar{\mathcal{D}}_t$ as follows:

$$d_t = \arg\max_d [h_\psi(x)]_d, \tag{7}$$

$$\bar{\mathcal{D}}_t = \{(x, \bar{y}_{d_t}) \mid x \in \mathcal{D}_t, \bar{y}_{d_t} = \arg\max_c f_\theta(x, d_t)\}, \tag{8}$$

where $h_\psi(x)$ outputs the predicted probabilities for source region indices, and where $\bar{y}_d = \arg\max_c f_\theta(x,d)$ is the pseudo-mask generated by routing through region index d.

Next, we assess the uncertainty of each image by evaluating the consistency of predictions across different region-specific routings. The stability score s for an image x is computed as the mean intersection-over-union (mIoU) between the pseudo-label $\bar{y}_{d_t}$ generated with d_t and the pseudo-masks obtained from routing through all other source region indices $d \in \{1,\ldots,D\} \setminus \{d_t\}$. The stability score is defined as:

$$s(x) = \sum_{d \neq d_t} \text{mIoU}(\bar{y}_{d_t}, \bar{y}_d). \tag{9}$$

Finally, we construct the adaptive target dataset $\bar{\mathcal{D}}_t$ by selecting the most reliable images with the highest stability scores, constituting a fraction ρ_s of the target dataset (i.e., $|\bar{\mathcal{D}}_t| = \rho_s \cdot |\mathcal{D}_t|$). The model is fine-tuned on $\bar{\mathcal{D}}_t$ using self-training, minimizing a pixel-wise cross-entropy loss:

$$\mathcal{L}_{\text{target}} = \frac{1}{|\bar{\mathcal{D}}_t|} \sum_{(x,\bar{y})\in\bar{\mathcal{D}}t} \mathcal{H}_p(\bar{y}, f_\theta(x, d_t)). \tag{10}$$

Experiment

We evaluate our model in terms of its generalizability for slum segmentation across domains, focusing on three African cities as unseen target regions.

Performance Evaluation

Implementation details Adaptation is performed in a fully unsupervised setting, without access to any ground truth labels from the target domain. We use the SegFormer (Xie et al. 2021) backbone and train all models under identical settings using SGD with a learning rate of 0.0001 and momentum of 0.99. In our experiments, we set $\rho_s = 0.5$, $E = 12$, and $k = 2$. Please refer to the Appendix for additional details on training details. We release the code of our model to facilitate broader adoption and greater impact within the research community: https://github.com/DS4H-GIS/GRAM

Result We compare its performance against several state-of-the-art TTA methods. The compared baselines include: (1) Vanilla Source: A standard segmentation model without MoE; (2) MoE Source: An MoE-based model without TTA; (3) SHOT (Liang, Hu, and Feng 2020): Aligns target features to a frozen source classifier via information maximization and self-supervised learning; (4) TENT (Wang

Method	Dar es Salaam (Tanzania)				Kampala (Uganda)				Maputo (Mozambique)			
	mIoU	F1-score	Precision	Recall	mIoU	F1-score	Precision	Recall	mIoU	F1-score	Precision	Recall
Vanilla Source	0.681	0.792	0.746	0.890	0.716	0.814	0.805	0.824	0.800	0.888	0.877	0.902
MoE Source	0.806	0.885	0.895	0.876	0.800	0.881	0.831	**0.956**	0.900	0.947	**0.942**	0.953
SHOT	0.712	0.813	0.847	0.786	0.713	0.810	0.846	0.782	0.813	0.895	0.890	0.901
TENT	0.691	0.800	0.755	0.889	0.716	0.814	0.809	0.819	0.802	0.889	0.878	0.903
CoTTA	0.762	0.853	0.857	0.850	0.821	0.900	0.894	0.908	0.821	0.900	0.895	0.905
SAR	0.700	0.807	0.764	0.889	0.748	0.843	0.829	0.859	0.804	0.890	0.881	0.900
BeCoTTA	0.741	0.836	0.900	0.793	0.844	0.911	**0.965**	0.866	0.904	0.949	0.938	0.965
GRAM	**0.859**	**0.921**	**0.911**	**0.931**	**0.870**	**0.927**	0.932	0.921	**0.907**	**0.951**	0.939	**0.966**

Table 1: Comparison of various metrics (mIoU, F1-score, Precision, Recall) across baseline and proposed methods in three regions: Dar es Salaam (Tanzania), Kampala (Uganda), and Maputo (Mozambique).

Component	mIoU	F1-score	Precision	Recall
w/o $\mathcal{L}_{dom}$	0.836	0.906	0.876	**0.944**
w/o $\mathcal{L}_{MI}$	0.734	0.823	0.898	0.786
No Filtering	0.818	0.893	0.910	0.878
Confidence Filtering	0.463	0.501	0.821	0.514
Temporal Consistency	0.837	0.907	0.884	0.933
Full Component	**0.859**	**0.921**	**0.911**	0.931

Table 2: Average performance across slum and non-slum classes for various ablation settings on Dar es Salaam

et al. 2021): Adapts the model during inference by minimizing the prediction entropy by updating the batch normalization statistics; (5) CoTTA (Wang et al. 2022): A mean-teacher-based approach that incorporates stochastic weight restoration to mitigate error accumulation over time; (6) SAR (Niu et al. 2023): Improves robustness by optimizing the sharpness of the entropy surface; (7) BeCoTTA (Lee, Yoon, and Hwang 2024): Extends continual TTA with a MoE adapter architecture. We modify its target adaptation mechanism by incorporating our cross-region consistency sampling, which better leverages region-specific experts and filters unreliable pseudo-labels during target adaptation.

Table 1 displays the performance with respect to mIoU, F1-score, precision, and recall. Our model consistently outperforms all evaluated baselines across multiple cities. The superior performance of the MoE Source over the Vanilla Source demonstrates that the MoE architecture enhances generalizability during the source training process. Although BeCoTTA outperforms other baselines, it inadequately captures target prediction reliability. These results highlight our model's robustness and effectiveness in slum detection, particularly under severe cross-regional domain shifts. While conservative approaches (e.g., entropy regularization) suffice for limited shifts, they struggle in slum segmentation due to significant source-target divergences. In contrast, our self-training framework utilize pseudo-labels filtered by cross-expert consistency, enabling aggressive adaptation, noise robustness, and enhanced generalization in diverse urban contexts.

Ablation Study To evaluate the contribution of each component within our framework, we conduct an ablation study on the Dar es Salaam region. Table 2 summarizes the average performance across slum and non-slum classes under different ablation settings. We evaluate five configurations: the first two examine modifications to source training, while the remaining three focus on the target adaptation phase.

Removing either the domain alignment loss $\mathcal{L}_{\text{dom}}$ or the mutual information loss $\mathcal{L}_{\text{MI}}$ results in a performance decline, confirming the necessity of both components. The impact is particularly pronounced for $\mathcal{L}_{\text{MI}}$, whose removal leads to a significant drop in mIoU (0.734) and F1-score (0.823), underscoring its central role in promoting expert consistency and preserving informative predictions.

We evaluate the impact of pseudo-label selection by comparing three strategies: no filtering, confidence-based filtering, and temporal consistency filtering. Omitting filtering results in moderate degradation, indicating that noisy pseudo-labels impair adaptation. Confidence-based filtering performs the worst (mIoU: 0.463), likely due to overconfidence under domain shift. Temporal consistency filtering enhances robustness by leveraging temporal agreement across source training checkpoints. (mIoU: 0.837, F1: 0.907). Our full method achieves the best performance (mIoU: 0.859, F1: 0.921), validating the effectiveness of combining mutual information regularization, domain alignment, and consistency-aware pseudo-labeling. See the Appendix for additional baseline comparisons and component analysis.

Discussion

Region-Aware Routing Mirrors Visual Similarity

Starting from the data complexity of capturing heterogeneous geographical patterns of slums, our method leverages the Mixture-of-Experts architecture for region-specific specialization in slum segmentation. During target adaptation, we utilize the output of an external region classifier h_ψ to identify the source region most visually similar to the target region. Target samples are then dynamically routed through the corresponding expert set to facilitate adaptation.

To validate the effectiveness of this region-aware routing, we compute the Jaccard similarity between the image sets

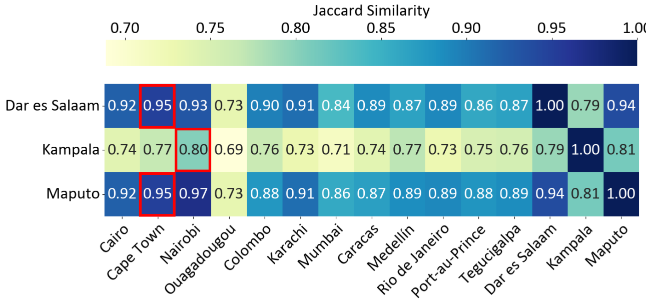

Figure 4: Jaccard similarity between the image sets of the three test cities (Dar es Salaam, Kampala, and Maputo) and those in the training set. Higher values indicate greater visual similarity in slum characteristics across cities. Red boxes denote the predictions made by the region classifier.

of source and target regions using visual features extracted from a pretrained CLIP model (Radford et al. 2021). Specifically, we extract features from all images in the source and target datasets, apply K-means clustering in the feature space to assign discrete cluster labels, and calculate the Jaccard similarity between the label sets for regions A and B. The results are visualized in Figure 4.

Region pairs predicted by the classifier (highlighted in red boxes—for example, Dar es Salaam and Cape Town) consistently exhibit higher similarity scores, indicated by the darker shades within the boxes. This correlation further supports the visual coherence of the classifier's predictions. These matched pairs also reflect real-world geographic and socio-spatial patterns. For instance, Cape Town, Dar es Salaam, and Maputo are coastal port cities, while Nairobi and Kampala are neighboring inland cities in East Africa. We can further associate the specific patterns of informal settlements with the examples shown in Figure 1: where Cape Town, Dar es Salaam, and Maputo share square-like and light-gray features, whereas those in Nairobi and Kampala tend to be more rectangular and rusty brown. Importantly, the classifier is trained without explicit geographic information, suggesting that these associations are derived solely from the visual characteristics of slum regions.

Slum Tracking Can Guide Policy

One of the goals of this work was to find a generalizable framework to help compute slum-related statistics for diverse regions where official data are scarce or unavailable. By applying our model to multi-temporal satellite imagery, stakeholders can now produce quantitative baseline estimates and monitor the evolution of informal settlements over time. For instance, our analysis revealed divergent trends in the three target cities: In Kampala, slum areas increased slightly from 8.4% in 2015 to 8.6% in 2023 (Figure 5); Maputo experienced a sharp increase, from 35.3% in 2016 to 41.2% in 2023; while Dar es Salaam experienced a gradual decrease from 17.3% in 2015 to 12.6% in 2022.

These divergent outcomes are informative given that all three countries have experienced a comparable pace of economic growth. The ability to computationally track such

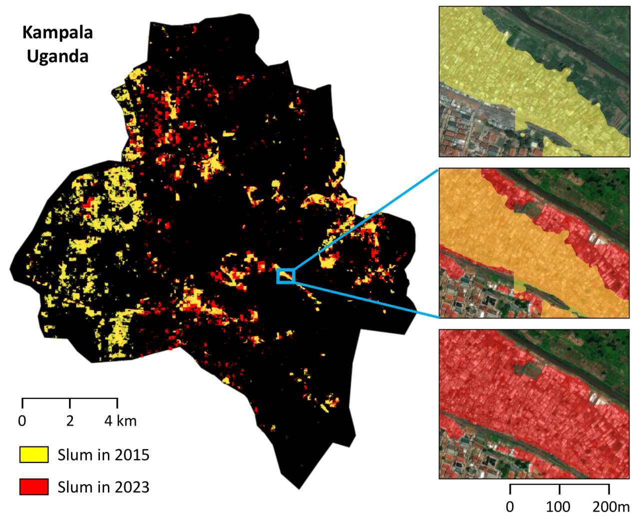

Figure 5: Slum segmentation results in Kampala in 2015 (yellow) and 2023 (red). Over the eight-year period, the slum ratio in the city increased from 8.4% to 8.6%.

nuanced, on-the-ground changes that broad economic indicators often overlook is tremendously valuable for urban policy making. Moreover, because the framework is label-efficient and required substantially little human effort compared to previous work, it can be scaled across a broader set of regions without requiring new, resource-intensive ground-truth data collection.

This capability is a critical advantage, especially in low- and middle-income settings, where field surveys can be costly, politically sensitive, or logistically challenging. The ability to accurately identify areas of greatest need allows for a more strategic and targeted allocation of resources, which is paramount for effective data-driven interventions. Furthermore, by publicly releasing a computation method to assess the long-term impact of urban policies, the framework ultimately empowers local decision-makers with the actionable evidence needed to address urban poverty, even in the absence of traditional census data.

Conclusion

We present GRAM, a TTA framework for slum detection in unseen regions, designed to tackle domain shifts in cross-regional settings. By integrating an MoE architecture with adaptive routing and cross-country prediction consistency, GRAM captures both shared and region-specific features while filtering unreliable pseudo-labels during self-training. Experiments on three African cities show that GRAM outperforms state-of-the-art baselines, demonstrating its effectiveness as a scalable, label-efficient solution for global slum monitoring. This approach holds promise in supporting inclusive, data-driven urban policy and can be extended to broader geographic contexts in future work.

Acknowledgments

JKim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-NR068758 and RS-2025-00563196). We also thank the anonymous reviewers for their valuable comments and suggestions.

References

Ahn, D.; Yang, J.; Cha, M.; Yang, H.; Kim, J.; Park, S.; Han, S.; Lee, E.; Lee, S.; and Park, S. 2023. A human-machine collaborative approach measures economic development using satellite imagery. *Nature Communications*, 14(1): 6811.

Ahn, K.; Han, S.; Park, S.; Kim, J.; Park, S.; and Cha, M. 2025. Generalizable disaster damage assessment via change detection with vision foundation model. In *proc. of the 39th AAAI Conference on Artificial Intelligence*, volume 39, 27784–27792.

Alcaldia de Medellin. 2010. Informal Settlement Map. https://datacatalog.worldbank.org/search/dataset/0039832/Karachi--Pakistan----Informal-Settlements--ESA-EO4SD-Urban-.

Chen, Z.; Shen, Y.; Ding, M.; Chen, Z.; Zhao, H.; Learned-Miller, E. G.; and Gan, C. 2023. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *proc. of the 2023 Computer Vision and Pattern Recognition*, 11828–11837.

Duque, J. C.; Patino, J. E.; and Betancourt, A. 2017. Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, 9(9): 895.

Earth Observation for Sustainable Development. 2018. Service Operations Report - Karachi. https://datacatalog.worldbank.org/search/dataset/0039832/Karachi--Pakistan----Informal-Settlements--ESA-EO4SD-Urban-.

Gadiraju, K. K.; Vatsavai, R. R.; Kaza, N.; Wibbels, E.; and Krishna, A. 2018. Machine learning approaches for slum detection using very high resolution satellite images. In *proc. of the 2018 IEEE International Conference on Data Mining Workshops*, 1397–1404.

Gram-Hansen, B. J.; Helber, P.; Varatharajan, I.; Azam, F.; Coca-Castro, A.; Kopackova, V.; and Bilinski, P. 2019. Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. In *proc. of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368.

Han, S.; Ahn, D.; Cha, H.; Yang, J.; Park, S.; and Cha, M. 2020a. Lightweight and robust representation of economic scales from satellite imagery. In *proc. of the 34th AAAI Conference on Artificial Intelligence*, volume 34, 428–436.

Han, S.; Ahn, D.; Park, S.; Yang, J.; Lee, S.; Kim, J.; Yang, H.; Park, S.; and Cha, M. 2020b. Learning to score economic development from satellite imagery. In *proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2970–2979.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *proc. of the 32nd International Conference on International Conference on Machine Learning*, 448–456.

Jean, N.; Burke, M.; Xie, M.; Davis, W.; Lobell, D.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Kit, O.; and Lüdeke, M. 2013. Automated detection of slum area change in Hyderabad, India using multitemporal satellite imagery. *ISPRS journal of photogrammetry and remote sensing*, 83: 130–137.

Kohli, D.; Sliuzas, R.; Kerle, N.; and Stein, A. 2012. An ontology of slums for image-based classification. *Computers, environment and urban systems*, 36(2): 154–163.

Kuffer, M.; Pfeffer, K.; and Sliuzas, R. 2016. Slums from Space—15 Years of Slum Mapping Using Remote Sensing. *Remote Sensing*, 8(6): 455.

Lee, D.; Yoon, J.; and Hwang, S. J. 2024. BECoTTA: input-dependent online blending of experts for continual test-time adaptation. In *proc. of the 41st International Conference on Machine Learning*, 27072–27093.

Leonita, G.; Kuffer, M.; Sliuzas, R.; and Persello, C. 2018. Machine learning-based slum mapping in support of slum upgrading Programs: The case of Bandung City, Indonesia. *Remote Sensing*, 10(10).

Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1): 31–64.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *proc. of the 37th International Conference on Machine Learning*, volume 119, 6028–6039. PMLR.

Lin, Y.; Zhang, X.; Liu, Y.; Han, Z.; Liao, Q.; and Li, Y. 2024. Long-term Detection and Monitory of Chinese Urban Village Using Satellite Imagery. In *proc. of the 33rd International Joint Conferences on Artificial Intelligence*, 7349–7357.

Lumban-Gaol, Y. A.; Rizaldy, A.; and Amadi, M. 2023. Comparison of deep learning architectures for the semantic segmentation of slum areas from satellite images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1/W2-2023: 1439–1444.

Maiya, S. R.; and Babu, S. C. 2018. Slum segmentation and change detection: A deep learning approach. arXiv:1811.07896.

Nado, Z.; Padhy, S.; Sculley, D.; D'Amour, A.; Lakshminarayanan, B.; and Snoek, J. 2020. Evaluating Prediction-Time Batch Normalization for Robustness under Covariate Shift. arXiv:2006.10963.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *proc. of the 2023 International Conference on Learning Representations*.

Park, S.; Han, S.; Ahn, D.; Kim, J.; Yang, J.; Lee, S.; Hong, S.; Kim, J.; Park, S.; Yang, H.; et al. 2022. Learning economic indicators by aggregating multi-level geospatial information. In *proc. of the 36th AAAI Conference on Artificial Intelligence*, volume 36, 12053–12061.

Park, S.; Han, S.; Kim, S.; Kim, D.; Park, S.; Hong, S.; and Cha, M. 2021. Improving unsupervised image clustering with robust learning. In *proc. of the 2021 Computer Vision and Pattern Recognition*, 12278–12287.

Prefeitura da Cidade do Rio de Janeiro. 2021. Limite de Favelas 2019. https://www.data.rio/datasets/limite-favelas-2019/explore.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *proc. of the 38th International Conference on Machine Learning*.

Rehman, M. F. U.; Aftab, I.; Sultani, W.; and Ali, M. 2022. Mapping Temporary Slums From Satellite Imagery Using a Semi-Supervised Approach. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Samper, J. 2025. Atlas of Informality. https://www.atlasofinformality.com/. Accessed: 2025-11-13.

Simon, D. 2011. Situating slums: Discourse, scale and place. *City*, 15(6): 674–685.

Slum Rehabilitation Authority. 2016. Georeferenced City Survey Plan of Mumbai City & Suburban District Showing Area Boundaries of Slum Clusters. https://sra.gov.in/page/innerpage/gis-mis-slum-data.php.

Stark, T.; Wurm, M.; Zhu, X. X.; and Taubenböck, H. 2020. Satellite-based mapping of urban poverty with transfer-learned slum morphologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 5251–5263.

Stark, T.; Wurm, M.; Zhu, X. X.; and Taubenböck, H. 2024. Quantifying uncertainty in slum detection: advancing transfer-learning with limited data in noisy urban environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.*, 17: 4552–4565.

Taubenböck, H.; and Kraff, N. 2014. The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data. *Journal of Housing and the Built Environment*, 29(1): 15–38.

UN-Habitat. 2003. *The challenge of slums: global report on human settlements, 2003*. Routledge.

UN-Habitat. 2025. Urban Indicators Database. https://data.unhabitat.org/. Accessed: 2025-11-13.

University of Edinburgh. 2020. Dwelling outline - Informal Settlements of Cape Town, 2018.

Verma, D.; Jana, A.; and Ramamritham, K. 2019. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. *Habitat International*, 88: 101981.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *proc. of the 2021 International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *proc. of the 2022 Computer Vision and Pattern Recognition*.

Wurm, M.; Stark, T.; Zhu, X. X.; Weigand, M.; and Taubenöck, H. 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150: 59–69.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. In *proc. of the 35th International Conference on Neural Information Processing Systems*.

Yang, J.; Lee, S.; Park, S.; Ahn, D.; and Cha, M. 2025a. Towards consistent and robust slum detection using multi-year satellite data. In *proc. of IEEE Joint Urban Remote Sensing Event*, 1–4.

Yang, J.; Lee, S.; Park, S.; Lee, M.; and Cha, M. 2025b. AI-based Ger detection reveals post-pandemic delay in informal housing progress in Mongolia. *npj Urban Sustainability*, 5(1): 78.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation. In *proc. of 2022 Computer Vision and Pattern Recognition*.

Zhang, X.; Liu, Y.; Lin, Y.; Liao, Q.; and Li, Y. 2024. UV-SAM: Adapting Segment Anything Model for Urban Village Identification. In *proc. of the 38th AAAI Conference on Artificial Intelligence*, volume 38, 22520–22528.