A Novel Evaluation Framework for 15-Minute City Using Satellite Imagery

Chanjae Song*
Department of Industrial & Systems Engineering, KAIST
Republic of Korea
chan4535@kaist.ac.kr

Seongyeub Chu*
Graduate School of Data Science, KAIST
Republic of Korea
chseye7@kaist.ac.kr

Jongwoo Kim*
Department of Industrial & Systems Engineering, KAIST
Republic of Korea
gsds4885@kaist.ac.kr

Mun Yong Yi†
Graduate School of Data Science, KAIST
Republic of Korea
munyi@kaist.ac.kr

Abstract

The 15-minute city (15MC) is an urban planning concept that promotes sustainable and inclusive cities where residents can access essential services within a short amount of time (i.e., 15 minutes). However, evaluating 15MC compliance in hyper-dense cities remains challenging due to: (1) traditional manual assessments that are resource-intensive and difficult to scale, and (2) POI-based metrics that suffer from data unavailability and lack of spatial contexts. In this paper, we propose a novel evaluation framework that directly assesses 15MC compliance from geospatial imagery in three stages: image pre-processing, representation learning, and instance aggregation. To validate the framework, we have constructed a new dataset of 2,794 residential areas in Seoul, pairing high-resolution geospatial imagery with functional urban labels. Furthermore, we have developed a model, GeoTwin-MIL, on the basis of the proposed framework. The model includes two key components: (1) cross-modal contrastive learning that aligns satellite and map representations to capture both morphological (building density) and topological (road networks) features, enabling robust inference using only satellite images, and (2) multiple instance learning to efficiently aggregate geospatial details while detecting localized urban functions within high-resolution imagery. The experimental results obtained from various evaluation settings show that GeoTwin-MIL significantly outperforms single-modality approaches or vision baselines, validating the integrative effectiveness of the two key components and supporting the transferability of the model without POI dependencies. The code is available at https://github.com/20243439/geotwin_mil.git.

CCS Concepts

• **Information systems** → **Geographic information systems**; *Spatial-temporal systems*; Decision support systems; • **Computing methodologies** → *Image representations*.

*These authors contributed equally to this research.
†Corresponding author.

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3764182

Keywords

15-minute city, satellite imagery, multi label classification, cross-modal contrastive learning, multiple instance learning

ACM Reference Format:
Chanjae Song, Seongyeub Chu, Jongwoo Kim, and Mun Yong Yi. 2025. A Novel Evaluation Framework for 15-Minute City Using Satellite Imagery. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3748636.3764182

1 Introduction

The 15-minute city (15MC) is an urban planning concept that promotes equitable access to essential services—work, education, healthcare, commerce, and leisure—within a 15-minute walk, bike ride, or public transit trip [18, 22]. By prioritizing barrier-free environments, it improves spatial accessibility for vulnerable groups, including the elderly, children, and individuals with disabilities [19]. In response, many cities are reorganizing their urban systems in line with 15MC principles [18].

Despite its importance in urban planning, the evaluation and implementation of 15MC principles, particularly in hyper-dense environments, pose substantial challenges. Establishing a 15MC-compliant environment is based primarily on bottom-up human-led investigations, which are often time-consuming and resource-intensive [22]; in complex built forms, manual assessment often overwhelms practitioners [24]. Although various data-driven indices have been proposed using points of interest (POI) or census data, they are highly dependent on data availability and often fail to capture the spatial realities of the built environment, resulting in suboptimal accuracy, yielding surface-level, distance-based analyses [2]. These limitations underscore the need for advanced evaluation frameworks that move beyond proximity to assess functional completeness and experiential accessibility [20].

To address these challenges, we propose an automated framework for evaluating 15MC compliance using geospatial imagery, specifically satellite photographs and topographic maps, readily accessible via platforms such as Google Earth. We formulate the task as an image-based multi-label classification problem, where the presence of essential urban functions serves as target labels. Due to a lack of pairwise dataset for imagery and function labels, a custom dataset is constructed by crawling satellite and topographic images, aligning them with 15MC elements based on POI data. A variety of

neural network-based models are trained to predict the presence of 15MC-related components without relying on POI data during inference. To improve capturing high-resolution urban imagery, we integrate contrastive learning (CL) strategies and multiple instance learning (MIL), enabling patch-level focus.

We perform experiments on a manually constructed dataset and explore our approach on various image representation methods, including pre-trained CNN-based [16], CL techniques [3, 8], and MIL [13, 23, 27]. Extensive ablation studies are also conducted to assess the effectiveness of integrating MIL and contrastive learning for the representation of geospatial images in the context of the evaluation of 15MC. The results demonstrate that our method more accurately predicts 15MC-related elements by effectively leveraging salient imagery features, even with satellite-only inference. Our key contributions are as follows.

- We introduce a new benchmark dataset centered on Seoul, South Korea, one of the most hyper-dense cities in East Asia, constructed by crawling geospatial imagery from open platforms (e.g., Google Earth) and aligning each image with corresponding POI data sourced from public databases.
- We propose a novel method, GeoTwin-MIL, for evaluating 15MC compliance using satellite imagery, trained with contrastive learning and multiple instance learning strategies, enabling efficient and lightweight deployment by requiring only satellite images during inference.
- We conduct comprehensive experiments, demonstrating that our approach which incorporates contrastive learning strategy and multiple instance learning extracts more meaningful features from satellite imagery and achieves superior performance in 15MC evaluation compared to a pre-trained CNN-based and a contrastive learning-based methods.

2 Related Work

2.1 15-Minute City

The 15-minute city (15MC) is a planning paradigm that advocates for spatially distributed urban services such as living, working, and enjoying within a 15-minute radius by walking, cycling, or public transit from residents' homes [18, 22]. By promoting compact and accessible neighborhoods, the 15MC concept enhances inclusivity, particularly benefiting groups with limited mobility, including older adults, children, and individuals with disabilities, through the implementation of barrier-free urban environments [19]. As a result, municipalities worldwide are increasingly adopting this model to redesign their spatial and infrastructural layouts in alignment with 15MC principles [18]. Although the concept highlights the importance of proximity-based service provision, it does not imply the universal availability of all services within the 15-minute range. To operationalize this model, Garnier and Moreno [5] identified six core urban functions-living, supplying, working, caring, learning, and enjoying-and specified the institutional components necessary to fulfill each function. Despite its significance in enhancing urban living, the accurate and efficient evaluation of 15-minute city (15MC) completeness presents several challenges [2]. These include the need to account for public transportation and local geographic contexts, as well as the substantial time and cost required for human-based assessment.

2.2 Automated 15-Minute City Evaluation

Traditional methods of evaluating the compliance of 15MC primarily rely on POI-based indices, calculating network distances to essential services and aggregating their coverage [21]. Willberg et al. [26] demonstrated temporal variations in accessibility, which was significant for elderly populations. Network-based approaches using Urban Network Analysis (UNA) measure walkability through centrality metrics [12], while mobility-based evaluations incorporate actual movement patterns to reveal supply-demand mismatches [7, 28].

Multi-modal based methods extend these approaches by considering various transportation modes [6], and recent studies leverage large-scale check-in data for comprehensive evaluation [14]. However, these methods rely heavily on structured datasets and static POI counts, failing to capture morphological barriers, spatial discontinuities, or urban form evolution [10]. Our work addresses these limitations by directly processing satellite imagery without requiring POI data at inference time.

2.3 Image Representation Learning

Extracting meaningful representations from remote sensing imagery is a core challenge in urban analysis. CNN-based models such as ResNet [9] effectively capture local visual patterns but are inherently limited in modeling global spatial structures. To overcome this limitation and reduce reliance on labeled data, self-supervised learning (SSL) methods such as SimCLR [3] and MoCo [8] have been introduced, demonstrating strong generalized performance on geospatial imagery. In high-resolution satellite images, fine-grained spatial patterns are distributed at a local scale, and assigning a single label to the entire image can result in the loss of spatial information. Multiple instance learning has gained attention as a suitable model for such settings, as it allows for modeling localized information by dividing the image into multiple instances, learning their representations independently, and aggregating them for image-level prediction [17]. In this study, we adopt publicly available SSL with pretrained CNN encoders and design a cross-modal contrastive learning approach that jointly processes satellite and map imagery. This enables robust feature learning from high-resolution geospatial data and supports automated, large-scale evaluation of 15MC compliance across diverse urban environments.

3 Preliminary

3.1 Definition

3.1.1 Multiple Instance Learning. Multiple instance learning (MIL) is a weakly supervised learning framework in which the training data consist of a set of *bags* $\mathcal{X} = \{X^{(i)}\}_{i=1}^{N}$, where each bag is defined as $X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_{B_i}^{(i)}\}$ with $x_b^{(i)} \in \mathbb{R}^d$, and is associated with a single bag-level label $Y^{(i)} \in \{0, 1\}$. The instance-level labels $y_b^{(i)} \in \{0, 1\}$ are unobserved during training [17]. Under the standard MIL assumption, a bag is labeled positive if at least one instance is positive, and negative if all instances are negative.

3.1.2 Contrastive Learning. Contrastive learning is a self-supervised representation learning framework that aims to map similar samples closer together and dissimilar samples farther apart in the

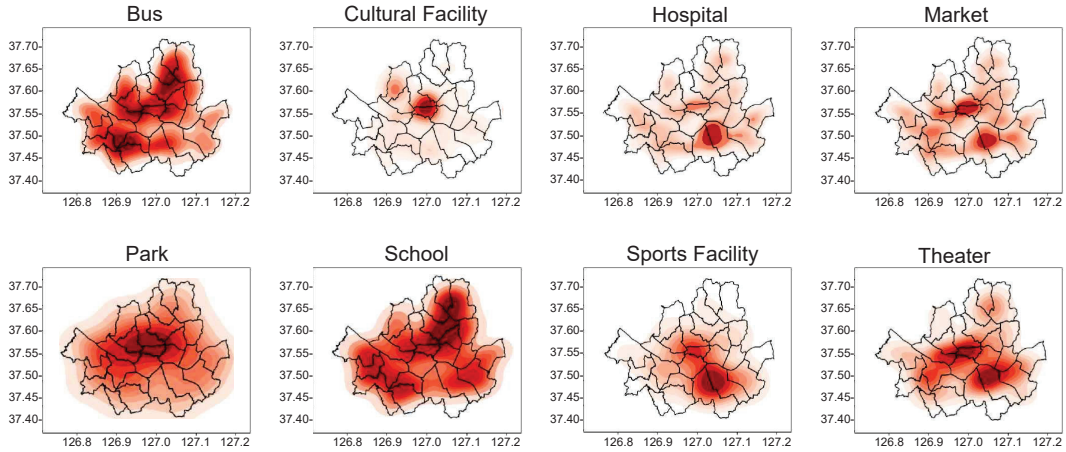


Figure 1: Kernel-density heatmaps for representative institutions overlaid on Seoul's administrative boundaries. The intensity scale indicates the spatial concentration of each institution type, revealing distinct patterns of urban service distribution across the metropolitan area.

embedding space. Given an anchor sample x, a positive sample x^+ (e.g., an augmented view of x), and one or more negative samples $\{x^-\}$, a feature encoder f_θ is trained to minimize a contrastive loss (e.g., InfoNCE) that encourages $f_\theta(x)$ and $f_\theta(x^+)$ to be more similar than $f_\theta(x^-)$. This encourages the model to learn semantically meaningful and discriminative representations without requiring explicit labels.

3.2 Spatial Bias in Urban Facility Distribution

This study aims to predict the compliance of the 15MC in Seoul, a representative hyper-dense East Asian metropolis, by assessing the presence of six social essential 15MC functions and their associated institutions organized by Garnier and Moreno [5] (see Table 1). As illustrated in Figure 1, Seoul exhibits significant intra-urban disparities in the density distribution of different institutional facilities, making manual human examination of 15MC compliance significantly costly. In this section, we propose a method that leverages geospatial imagery to address the aforementioned challenge by automatically predicting the presence of 15MC-related institutions. Furthermore, we illustrate how we synthesized a new benchmark dataset for training the model used in our framework.

Table 1: Essential functions and corresponding institutions for 15MC.

Social Essential Functions	Institutions
Living	Police station, Shared accommodation, Park
Supplying	Market, Bakery, Post office
Working	Warehouse, Bicycle rental station, Bus stop
Caring	Hospital, Sports facility, Swimming pool, Pharmacy
Learning	Kindergarten, School
Enjoying	Theater, Library, Bookstore, Museum, Cafe, Restaurant, Playground

3.3 15MC Evaluation Framework

Our proposed framework evaluates 15-minute city (15MC) compliance in three decoupled stages. **(i) Image Pre-processing**: A high-resolution satellite, map, or fused geospatial tile is divided into fixed, non-overlapping patches and intensity-normalized, producing an ordered set suitable for batch processing. **(ii) Representation Learning**: A generic encoder projects each patch to an embedding, distilling salient morphological or spatial cues **(iii) Instance Aggregation**: Instance embeddings are pooled or attended to a region-level descriptor, after which a multi-label classifier jointly predicts (a) six essential urban functions and (b) 22 institution categories, listed on Table 1. This modular separation of pre-processing, representation, aggregation and inference allows any feature encoder or aggregation strategy to be plugged in without altering the downstream 15MC assessment.

For this study we instantiate stages (ii) and (iii) with a lightweight pairing of cross-modal contrastive learning and TransMIL aggregation, referred to as **GeoTwin-MIL**. Section 4 and Figure 2 details the network architecture, training logic, and schedule of GeoTwin-MIL.

4 Method

4.1 Dataset Synthesis

4.1.1 Image cropping based on 15MC. To the best of our knowledge, no publicly available dataset pairs geospatial imagery with labels explicitly related to the 15MC concept. To address this issue, we developed a benchmark dataset that integrates high-resolution satellite imagery and topographic maps with labels specifically tailored for the evaluation of 15MC in Seoul, South Korea. Given that the 15MC concept emphasizes proximity to essential facilities surrounding residential areas, we selected apartment complexes, Seoul's predominant housing type, as reference points. To systematically evaluate the performance of our proposed framework, apartment complexes were classified according to household count into three groups: *small* (under 500 households), *medium* (500–1,000 households), and *large* (over 1,000 households). This stratification

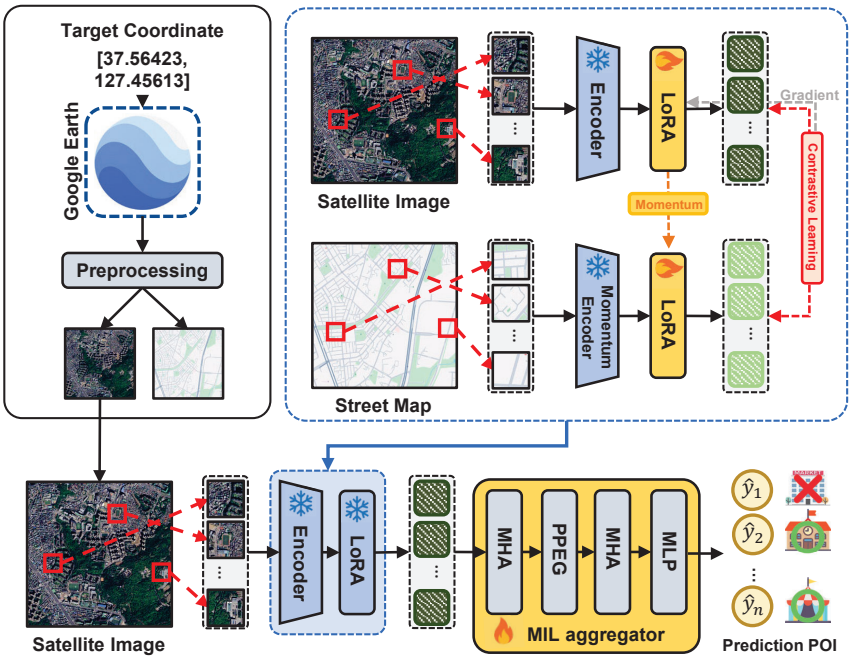


Figure 2: The overall architecture of GeoTwin-MIL: Cross-modal contrastive learning is first performed between satellite imagery and street map data using momentum encoders with LoRA adaptation, followed by TransMIL aggregation with multi-head attention and positional encoding for urban function prediction.

ensures typological diversity and comprehensive spatial representativeness in Seoul's residential areas, encompassing a total of 2,794 complexes, which consists of 408 large, 642 medium, and 1,744 small complexes.

For each selected complex, a satellite image and the corresponding topographic map were constructed covering a spatial area of 2km × 2km, precisely centered on the complex coordinates (latitude and longitude). This spatial scale ensures coverage of an approximate 1 km radius, reflecting the distance typically covered within a 15-minute walk, a key measure within the 15MC concept. The detailed process is as follows. Due to restrictions on direct access to satellite-imagery APIs, we implemented an automated acquisition method using Google Earth [1] controlled via Selenium WebDriver [2] using the Python programming language. Our methodology involves a two-step image capture process designed to overcome viewport limitations in Google Earth. Initially, the location of the target apartment complex is pinpointed using its geographic coordinates, and a uniform altitude of 1, 500m is maintained to ensure consistent resolution of the ground sample. Subsequently, two overlapping images are captured by vertically shifting the viewpoint 400m (upward and downward) from the reference point. The overlapping regions of these images serve as alignment references for seamless stitching, a procedure that is applied consistently to both satellite and topographic images. The resulting geospatial images are therefore uniformly sized at 2km × 2km and resampled to high-resolution PNG files of 4, 096 × 4, 096 pixels, ensuring consistent spatial resolution and comprehensive coverage throughout the data set. Consequently, we collected 2,794 regional images in total. The details of the process are outlined in Figure 3 and Algorithm 1.

Algorithm 1: Cropping process of geospatial images

Input: Latitude of base location ($\mathcal{LA}$), Longitude of base location ($\mathcal{LO}$), Distance adjustment factor ($d = 0.00113$)
Output: Combined satellite image
Step 1: Initialize WebDriver and Image Storage
Initialize WebDriver ⟶ `driver`
Initialize empty list `images` for storing captured images
Step 2: Capture Images at Adjusted Latitudes
for $i \in \{-1, 1\}$ **do**
    `adjusted_latitude` ⟵ $\mathcal{LA} + i \times d \times 4$
    `url` ⟵ `GenerateURL(adjusted_latitude,` $\mathcal{LO}$`)`
    `image` ⟵ `driver.GetIMAGE(url)`
    `cropped_image` ⟵ `Crop(image)`
    Append `cropped_image` to `images`
`final_image` ⟵ `MergeIMAGE(images)`
return `final_image driver.quit()`

4.1.2 Labeling institutions based on POI data. This part outlines the process of labeling institutions within previously cropped region-specific satellite tiles used to evaluate the compliance of 15MC. We started by reviewing previous research to identify the necessary conditions to meet the 15MC criteria. Building on the foundational study of Garnier and Moreno [5], we adopt the six essential social functions proposed as the core dimensions of the 15MC. Based on

[1] https://earth.google.com/web/
[2] https://selenium-python.readthedocs.io/getting-started.html

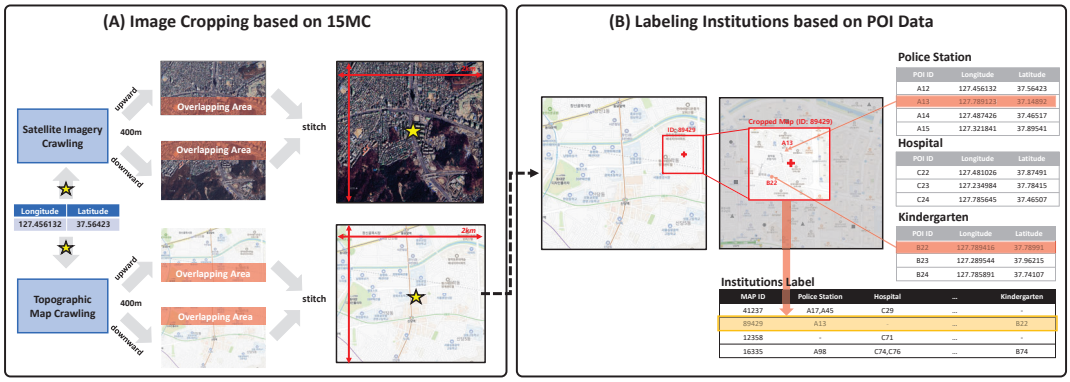


Figure 3: Dataset Synthesis: (A) Depicts the process of crawling and cropping satellite imagery. (B) Illustrates the procedure for locating institutions on the satellite imagery and assigning labels based on POI data.

these functions, we curated a corresponding set of institutional types and adapted them to reflect the urban characteristics of Seoul. Through this adaptation process, we finalized a list of 22 institutional categories. For each region, we labeled the presence of the institutions using publicly available POI datasets. Categories with insufficient POI coverage (e.g., office buildings) were excluded from the final inventory. Broad or ambiguous categories (e.g., leisure facilities) were further disaggregated into more specific types, such as museums and playgrounds, to enhance semantic clarity. A comprehensive mapping of functions and their corresponding institution types is provided in Table 1.

The construction of labels follows a hierarchical structure of two levels, namely the function level and the institution level: (i) At the function level, labels represent the presence or absence of six essential social functions, detailed in Table 1. Each region is encoded as a 6-dimensional multi-hot vector, where each element is set to 1 if the corresponding function is satisfied and 0 otherwise. (ii) At the institution level, the labels denote the existence of 22 specific institution types associated with the six functions. Similarly, each region is represented as a 22-dimensional multi-hot vector, with binary values indicating whether each institution is present (1) or not (0). The final dataset consists of triples ⟨image, function, institution⟩, where function $\in [0,1]^6$ and institution $\in [0,1]^{22}$. This hierarchical labeling strategy facilitates a comprehensive evaluation of both high-level functional accessibility and fine-grained institutional presence through multi-label classification tasks for 15-minute city (15MC) assessment. The systematic procedure ensures consistent spatial representation and reliable ground truth labeling in all 2,794 residential areas in our curated dataset. The labeling procedure is provided in Algorithm 2.

Algorithm 2: Labeling process of functions and institutions within a boundary

Input: Bag of POI datasets $\mathcal{B}$, Base latitude ($\mathcal{LA}$), Base longitude ($\mathcal{LO}$)
Output: Presence of each function and institution within a region

Step 1: Obtain the coordinates of a target boundary
top_left ← CalculateCoordinate($\mathcal{LA}$, $\mathcal{LO}$, offset_top, offset_left)
top_right ← CalculateCoordinate($\mathcal{LA}$, $\mathcal{LO}$, offset_top, offset_right)
bottom_left ← CalculateCoordinate($\mathcal{LA}$, $\mathcal{LO}$, offset_bottom, offset_left)
bottom_right ← CalculateCoordinate($\mathcal{LA}$, $\mathcal{LO}$, offset_bottom, offset_right)
Step 2: Locate the target boundary
boundary ← {top_left, top_right, bottom_left, bottom_right}
Step 3: Filter functions and institutions within the target boundary
functions_list = [*living*, *supplying*, *working*, *caring*, *learning*, *enjoying*]
functions_within_boundary = $[0] \times 6$
institutions_within_boundary = $[0] \times 22$
foreach *POI dataset in* $\mathcal{B}$ **do**
  foreach *index* i *and institution* $inst$ *in POI dataset* **do**
    if $inst$ *within* boundary **then**
      institutions_within_boundary$[i] = 1$
  foreach *index* f *and function* $func$ *in* functions_list **do**
    if *all* $inst$ *of* $func$ *exists* **then**
      functions_within_boundary$[f] = 1$
return functions_within_boundary and institutions_within_boundary

4.2 Patch-Based Input Slicing

To support fine-grained spatial reasoning and downstream MIL, we divide each high-resolution input image into fixed-size, non-overlapping patches prior to training. Each satellite image yields a set of patches $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$, and each corresponding map image yields $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$, where $h_k, m_k \in \mathbb{R}^d$ denotes the d-dimensional feature vectors obtained by independently encoding each patch through a pretrained encoder.

This patch-level representation allows the model to localize functional attributes (e.g., parks, facilities) that may only occupy a portion of the image. To aggregate these localized features into a coherent region-level prediction, we adopt an MIL framework.

4.3 Patch-Based Cross-Modal Pretraining

To build a modality-aligned encoder that integrates satellite and map features, we perform a cross-modal contrastive pre-training based on MoCo [8]. Satellite patches ($h \in \mathcal{H}$) contain fine-grained morphological cues (e.g., building shapes, densities), while map patches ($m \in \mathcal{M}$) emphasize topological structures (e.g., road layouts, zoning), which are often occluded in aerial views. This stage aims to align these complementary views and learn spatial representations that are robust across modalities.

Each training instance consists of a triplet (h, m^+, m^-), where h is a satellite patch, m^+ is a map patch from the same location (positive),

and m^- is a patch from a different region (negative). All patches are passed through a shared ResNet-50 backbone. The query encoder f_θ processes the satellite anchor, while the momentum encoder $f_{\theta'}$, updated via the exponential moving average (EMA), encodes the map patches.

$$q = f_\theta(h), \quad k^+ = f_{\theta'}(m^+), \quad k^- = f_{\theta'}(m^-). \tag{1}$$

The momentum encoder parameters are updated as:

$$\theta' \leftarrow \alpha\theta' + (1-\alpha)\theta, \tag{2}$$

where α is the momentum coefficient.

To enable parameter-efficient learning, only the LoRA modules [11] within the query encoder are updated during training, while the ResNet-50 backbone remains frozen.

We use the InfoNCE objective to contrast positive and negative pairs.

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(\text{sim}(q,k^+)/\tau\right)}{\sum_{k^-\in\mathcal{K}^-} \exp\left(\text{sim}(q,k^-)/\tau\right)}, \tag{3}$$

where $\mathcal{K}^-$ is the set of negative keys from the current mini-batch, and τ controls the temperature of the softmax. Lower τ sharpens the focus on hard negatives, promoting stronger alignment.

Since gradients are propagated only through the query encoder and its LoRA modules, the momentum encoder serves as a stable, non-trainable target. This pretraining phase yields an encoder that effectively captures and aligns local morphological and global topological patterns, serving as the foundation for downstream spatial reasoning tasks such as MIL.

4.4 Patch-Based Feature Aggregation

After contrastive pretraining, we reuse the query encoder f_θ to transform each satellite patch $h_k \in \mathcal{H}$ into a latent embedding $z_k \in \mathbb{R}^d$:

$$z_k = f_\theta(h_k), \quad \forall k \in \{1,2,\ldots,K\}, \tag{4}$$

yielding the full encoded bag $\mathbf{Z} = \{z_1, z_2, \ldots, z_K\} \in \mathbb{R}^{K\times d}$, which captures fine-grained morphological cues extracted from satellite patches.

To aggregate these patch-level embeddings into a holistic region-level representation suitable for downstream classification, we adopt the TransMIL [23] model, designed to handle weakly-supervised learning scenarios, where only bag-level (region-level) labels are available, and instance-level annotations are unknown.

First, each feature vector z_k is spatially contextualized using the Pyramid Position Encoding Generator (PPEG), which reshapes the patch sequence into a pseudo-2D layout and applies multi-scale convolutional filters to encode local spatial dependencies:

$$\tilde{\mathbf{Z}} = \text{PPEG}(\mathbf{Z}), \tag{5}$$

where $\tilde{\mathbf{Z}} \in \mathbb{R}^{K\times d}$ contains position-aware patch embeddings.

Next, the model applies a stack of Transformer layers to model complex dependencies between spatially-aware patch tokens. Specifically, a Multi-Head Self-Attention (MHA) mechanism is used to capture both short- and long-range interactions.

$$\mathbf{H}^{(1)} = \text{MHA}(\tilde{\mathbf{Z}}), \quad \mathbf{H}^{(2)} = \text{MHA}(\mathbf{H}^{(1)}), \tag{6}$$

where $\mathbf{H}^{(1)}, \mathbf{H}^{(2)} \in \mathbb{R}^{K\times d}$ denote the intermediate feature maps refined through attention.

To derive a fixed-size region-level descriptor from the sequence, a special [CLS] token is used, whose embedding is updated along with other tokens and finally extracted as the region representation. This vector is normalized and used as input for the classification layer.

$$F = \text{LN}(\mathbf{H}^{(2)}_{[\text{CLS}]}), \tag{7}$$

where $F \in \mathbb{R}^d$ encodes the aggregated evidence throughout the region.

Finally, the model predicts the probability of each functional class via a sigmoid-activated linear classifier.

$$\hat{y}_n = \sigma(W_n F + b_n), \quad \forall n \in \{1,\ldots,C\}, \tag{8}$$

where C is the number of predefined social function categories (e.g., Living, Supplying) or institution-level categories, and $\hat{y}_n \in [0,1]$ denotes the predicted likelihood that the region supports the class n.

This architecture allows the model to exploit both morphological patterns and spatial relationships between patches, making it well-suited for tasks where the presence of target semantics is localized and weakly annotated.

5 Experiments

We investigate the effectiveness of various image representation methods, specifically, a pre-trained convolution-based approach, a contrastive learning strategy, and a multiple instance learning strategy on a custom dataset that we constructed. Furthermore, we examine the incremental performance gains achieved by combining two or more of the aforementioned methods. The evaluation is conducted at two levels: function-level and institution-level. More precisely, we assess the alignment between model predictions and the actual presence of socially essential functions and their corresponding institutions across urban regions. We further dissect model behavior through ablation studies isolating each architectural component, benchmark data efficiency in few-shot settings, and probe robustness under distribution shifts via intra and cross city transfers.

5.1 Baselines

We systematically evaluate various approaches for 15MC compliance prediction from geospatial imagery. For feature extraction, we adopt and examine ResNet-50 [9] pretrained on ImageNet as a standard baseline and MoCo [8] representing self-supervised contrastive learning framework. For aggregating different types of image features, we compare non-learnable methods including pooling operations (min/mean/max) [16], Hadamard product [15], and concatenation [4], as well as learnable attention mechanisms including gated attention [1] and cross attention [25]. Finally, for the classification, we explore fine-tuning (FT), where a linear classification head is fully fine-tuned on pretrained features, versus multiple instance learning framework that preserves spatial structure through bag-level aggregation. This comprehensive evaluation enables us to identify the most effective combination for visual urban function assessment.

5.2 Implementation Details

Satellite–map patch pairs are resized to 224×224 pixels and normalized using ImageNet statistics. For single-modality contrastive learning, data augmentation includes random horizontal and vertical flips with probability 0.5, and color jitter applied on-the-fly during training. For cross-modal contrastive learning, we use the MoCo v3 architecture pretrained with the ResNet-50 encoder from `r-50-1000ep.pth.tar`, using momentum coefficient $m = 0.999$, the queue size of 65,536 negative keys, and the temperature $\tau = 0.2$. We freeze the backbone encoder parameters and update only rank-4 LoRA adapters, introducing approximately 0.34M trainable parameters. Training continues for 50 epochs using AdamW optimizer with learning rate $\eta = 1 \times 10^{-3}$ and weight decay $\lambda = 1 \times 10^{-4}$, implementing a global batch size of 1,024 through gradient accumulation every 32 steps, with a mini-batch size of 32 at each step. The learning rate schedule includes linear warm-up during the first epoch followed by cosine decay.

For downstream classification, precomputed patch embeddings are loaded slide-wise and linearly projected to 512 dimensions; a learnable `[CLS]` token is prepended to each sequence. We employ the original TransMIL architecture, two NystromAttention layers with 8 heads and 17 landmarks and positional PPEG encoding, changing only the dropout rate to 0.1. The MIL model is optimized using Adam optimizer with learning rate $\eta = 1 \times 10^{-3}$ and weight decay $\lambda = 1 \times 10^{-5}$, batch size of 32 slides for up to 50 epochs with early stopping patience of 10. Data are stratified by apartment-complex category and split into train/validation/test sets with an 8:1:1 ratio, with all results averaged over three random seeds $\{1, 17, 42\}$ and experiments conducted on a single NVIDIA RTX 4070 using PyTorch 2.4 and CUDA 12.5.

5.3 Evaluation Protocol

We operationalize 15-minute city (15MC) compliance as two complementary multi-label classification problems:
(1) Function level: 6 binary labels that indicate the presence of *living, supplying, working, caring, learning, enjoying*. A function label y_k^{func} is positive *iff* every institution linked to that function appears within the image,

$$y_k^{\text{func}} = \prod_{i \in \mathcal{I}_k} y_{k,i}^{\text{inst}}, \quad k \in \{1, \ldots, 6\}, \tag{9}$$

where $\mathcal{I}_k$ denotes the set of associated institutions listed in Table 1.
(2) Institutional level: 22 independent binary labels that predict the presence of each specific institution (parks, hospitals, schools, etc.).

Both tasks are evaluated using *macro-averaged* the F1 score and the AUC. Macro-F1 offers a robust measure under severe label imbalance by equally weighting classes, while AUC complements it by assessing ranking quality across all decision thresholds.

6 Results

6.1 Dataset Analysis

Table 2 presents the general statistics of the institutions labeled associated with each regional image, which were cropped and automatically annotated as described in Section 4.1. As shown in Table 2, there is a substantial variation in the average number of institutions, and the standard deviations indicate considerable disparities in the number of institutions between regions.

Table 2: Descriptive Statistics of Institution

# Images	2794			
Category	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Police	2.280	1.383	0	11
Accommodation	21.726	29.026	0	289
Park	1.082	0.966	0	5
Market	361.586	197.165	18	1377
Bakery	40.093	21.981	0	152
Post Office	1.999	1.311	0	9
Warehouse	0.232	0.630	0	6
Bus	112.690	39.078	18	226
Bike	27.545	9.754	1	71
Hospital	225.625	170.286	0	1482
Pharmacy	58.715	29.384	0	199
Sports	0.879	1.393	0	9
Swimming	1.521	1.535	0	11
Kindergarten	10.111	4.577	1	25
School	13.614	5.419	0	34
Theater	6.595	8.197	0	39
Library	2.367	1.598	0	9
Bookstore	5.641	4.875	0	51
Museum	0.804	1.747	0	26
Cafe	142.377	83.800	5	570
Restaurant	1186.295	722.098	29	4421
Playground	6.566	4.571	0	25

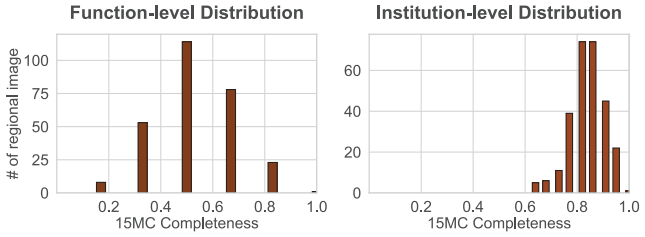


Figure 4: Distribution of 15-Minute City (15MC) Completeness Scores.This histogram illustrates the number of residential areas corresponding to each completeness level (%) based on the presence of essential social functions, reflecting the spatial distribution of functional accessibility within the city.

To further examine the overall distribution of 15MC compliance based on the presence of essential social functions and the corresponding institutions, Figure 4 illustrates the completeness of functions and institutions in the collected geospatial image dataset. As shown, the distribution at the function level resembles a normal distribution. However, when observed at the more granular institution level, the distribution exhibits noticeable skewness. This skewness can be attributed to the characteristics of hyper-dense cities like Seoul, where certain institutions such as bus stops, schools, and theaters are disproportionately abundant. In contrast, several other institution types are sparsely and unevenly distributed, appearing infrequently between regions, as visualized in Figure 1.

Table 3: Performance comparison for urban function (F) and institution (I) classification using different image types with or without contrastive learning (CL) and multiple instance learning (MIL). The best performances are written in bold, and the second best performances are underlined.

Input	CL	MIL	F-F1 (↑)	F-AUC (↑)	I-F1 (↑)	I-AUC (↑)
sat	✗	✗	0.609 (0.032)	0.770 (0.004)	0.890 (0.009)	0.792 (0.005)
map	✗	✗	0.557 (0.021)	0.730 (0.008)	0.866 (0.009)	0.759 (0.003)
sat	✗	✓	0.686 (0.016)	0.764 (0.029)	0.915 (0.010)	0.833 (0.020)
map	✗	✓	0.685 (0.029)	0.766 (0.018)	0.913 (0.006)	0.841 (0.011)
sat	✓	✗	0.654 (0.006)	0.782 (0.005)	0.894 (0.006)	0.792 (0.001)
map	✓	✗	0.606 (0.003)	0.672 (0.006)	0.875 (0.003)	0.710 (0.002)
sat	✓	✓	0.751 (0.025)	0.846 (0.009)	0.924 (0.004)	0.856 (0.012)
map	✓	✓	0.752 (0.021)	0.844 (0.008)	0.921 (0.004)	0.856 (0.006)
sat+map	✓	✗	0.711 (0.007)	0.790 (0.028)	0.900 (0.002)	0.774 (0.001)
sat+map	✓	✓	**0.784** (0.021)	**0.871** (0.008)	**0.945** (0.002)	**0.915** (0.008)

6.2 Main Results

Table 3 presents a comprehensive evaluation of various approaches for predicting 15-minute city compliance from geospatial imagery in function-level and institution-level. We observed that leveraging a pre-trained CNN-based image representation model on either satellite imagery or topographic maps yields moderate performance in predicting the presence of functions or institutions, achieving over 60% on both F1-score and AUC, except for the F1-score at the function level. Furthermore, incorporating MIL substantially improves the performance across all imagery types (e.g., satellite image, topographic map) by enabling the model to focus on informative regions within high-resolution urban imagery. CL further enhances feature representations, particularly when combined with MIL. In particular, the CL method using augmented data based on a single image type, either satellite imagery or topographic maps, achieves competitive performance compared to the use of both image types in the CL setting without MIL. Consequently, GeoTWin-MIL, the full integration of both image types with the CL and the MIL achieves the highest performance across all metrics, confirming that the combination of satellite visual features and topographic map-based geometric information through advanced learning strategies is optimal for urban function assessment.

6.3 Robustness Comparison

Observing GeoTwin-MIL's cross-modal superiority, we investigate its robustness under various training conditions. Figure 5 examines the robustness of different geospatial image configurations for the classification of urban functions and institutions in batch sizes. When CL is applied to single image types with augmentation, each image type shows distinct performance: satellite augmentation benefits from larger batch sizes by generally improving performance, while map augmentation shows the opposite trend with overall degradation. This indicates that relying on a single type of image can be significantly affected by batch size, which impacts the contrastive learning (CL) strategy [3]. Interestingly, GeoTwin-MIL successfully leverages both modalities, achieving not only the highest performance, but also remarkable stability across all batch sizes. This batch size invariance demonstrates that the proper fusion strategy of complementary image types creates more robust representations, making the approach particularly suitable for practical 15MC evaluation systems.

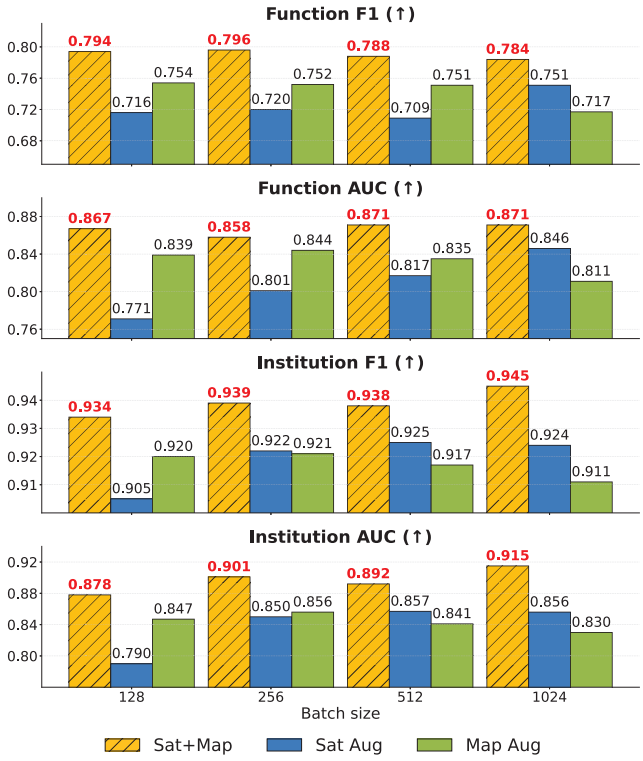


Figure 5: Performance comparison of models leveraging contrastive learning and multiple instance learning on different image types across batch-size. *Sat+Map* denotes joint satellite–map inputs, *Sat Aug* denotes satellite-only inputs with augmentation, and *Map Aug* denotes map-only inputs with augmentation.

6.4 Aggregation Strategies Comparison

The success of cross-modality fusion naturally raises questions about optimal aggregation strategies. Consequently, we examine various aggregation strategies to combine satellite and map characteristics in 15MC prediction as shown in Table 4. Non-learnable methods, particularly max pooling, achieve competitive performance, suggesting that simple feature selection can be effective for this task. Learnable attention mechanisms improve over most non-learnable baselines but remain still comparable to max pooling, suggesting that merely using attention strategies fails to sufficiently capture geospatial information. The contrastive learning approach achieves substantial gains across all metrics representing that selectively aligning meaningful features from both modality, which are satellite imagery and topographic maps, leads to the optimal performance.

Table 4: Comparison of input aggregation methods for combining satellite and map patches. Non-learnable methods apply fixed operations, whereas learnable methods rely on attention mechanisms. CL denotes our proposed cross-modal contrastive learning approach. The best performances are written in bold, and the second best performances are underlined.

Method	Learnable	F-F1 (↑)	F-AUC (↑)	I-F1 (↑)	I-AUC (↑)
Min Pooling	✗	0.673	0.749	0.924	0.862
Mean Pooling	✗	0.681	0.758	0.924	0.873
Max Pooling	✗	0.725	0.815	0.922	0.866
Hadamard Product	✗	0.720	0.816	0.926	0.867
Concatenation	✗	0.693	0.759	0.926	0.863
Gated Attention	✓	0.721	0.767	0.928	0.858
Cross Attention	✓	0.730	0.830	0.936	0.849
CL (Ours)	✓	**0.784**	**0.871**	**0.945**	**0.915**

6.5 Few-Shot Performance Analysis

To examine the performance of the model under a data scarcity condition that reflects the reality of urban systems where data collection is one of the challenging tasks, we conducted a few-shot learning experiment with limited training data shown in Figure 6. Although the ResNet50-FT baseline shows reasonable ability and improves steadily with more shots, its performance gains are relatively small compared to methods that incorporate contrastive learning or MIL, indicating that additional components provide substantial benefits in few-shot settings. GeoTwin-FT, which adds cross-modal contrastive learning while involving linear classification, improves moderately across all shots, demonstrating that a better alignment enhances the performance even with limited data. MoCo-MIL, which combines single-modal contrastive learning with MIL aggregation, shows an interesting pattern: Despite various prior studies suggesting that MIL requires abundant data [13, 23, 27], it consistently outperforms GeoTwin-FT, indicating that spatial aggregation enables effective information capture even with a few examples. GeoTwin-MIL integrates both innovations, cross-modal representations, and MIL aggregation, achieving the best performance with significant improvements from 1 to 4 shots and near-optimal results at 16 shots. Component ablation reveals that while cross-modal learning provides better features and MIL enables spatial reasoning, their combination yields synergistic benefits essential for data-efficient evaluation of 15MC.

6.6 15MC Completeness Alignment Analysis

Beyond classification accuracy, the ultimate goal of this study is not only accurately predicting each label's presence, but also predicting 15-minute city completeness scores. Figure 7 validates that our visual evaluation model GeoTwin-MIL accurately predicts the compliance of the city in 15 minutes, translating multi-label classifications into meaningful urban accessibility scores. The strong alignment between predicted and true distributions across both tasks confirms that our approach captures real urban patterns. This distributional accuracy demonstrates that beyond individual label

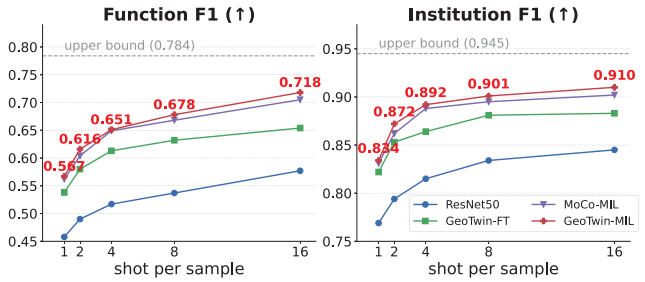


Figure 6: Few-shot learning performance evaluation across varying training samples per class (1-16 shots) for function-level and institution-level predictions. The upper bounds are established on GeoTwin-MIL with full training data.

predictions, our model, especially GeoTwin-MIL reliably aggregates urban elements into 15MC completeness metrics, enabling automated accessibility evaluation without costly manual surveys.

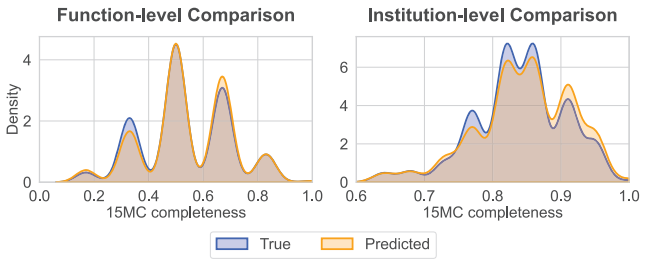


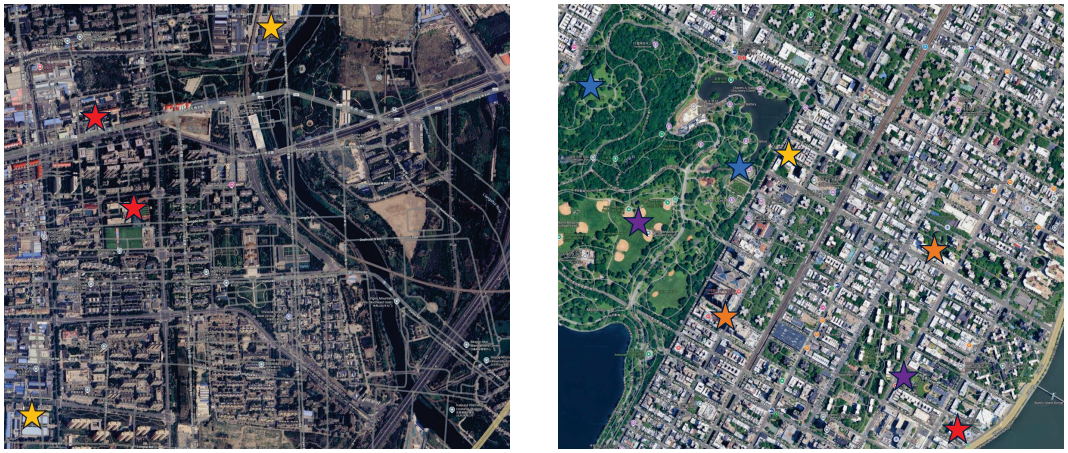
Figure 7: Probability-density comparison of 15-minute city completeness based on our method. (a) Function-level distributions and (b) institution-level distributions, where the blue curve denotes observed values and the orange curve denotes model predictions.

6.7 Transfer Performance Analysis

Table 5: Cross-group transfer performance on our best performance method GeoTwin-MIL."Train & Valid" denotes the group used for training (and validation if same), "Test" the held-out group. The best performances are written in bold, and the second best performances are underlined.

Train & Valid	Test	Transfer	F-F1 (↑)	F-AUC (↑)	I-F1 (↑)	I-AUC (↑)
small	small	✗	0.671	0.771	**0.906**	0.771
small	medium	✓	0.667	0.749	0.905	0.768
small	large	✓	0.661	0.736	0.898	0.754
medium	small	✓	0.715	0.737	0.902	0.760
medium	medium	✗	**0.726**	**0.810**	0.899	**0.800**
medium	large	✓	0.711	0.784	0.904	0.773
large	small	✓	0.673	0.667	0.896	0.738
large	medium	✓	0.686	0.773	0.897	0.767
large	large	✗	0.679	0.787	0.906	0.770

Real-world deployment of the proposed framework requires robustness to distribution shifts, which we evaluate under two transfer



(A) Case Study for Beijing 39°48'11"N 116°27'06"E

True Label: [0, 0, 1, 0, 1, 0]
(Working, Learning)

Predicted Label: [0, 1, 1, 0, 1, 0]
(Supplying, Working, Learning)

(B) Case Study for NewYork 40°47'34"N 73°57'07"W

True Label: [1, 1, 1, 1, 1, 0]
(Living, Supplying, Working, Caring, Learning)

Predicted Label: [1, 1, 1, 0, 1, 0]
(Living, Supplying, Working, Learning)

Figure 8: Cross-city transfer: GeoTwin-MIL trained on Seoul applied to (A) Beijing and (B) New York. Colored markers denote facilities satisfying function labels—park (Living), market (Supplying), school (Learning), warehouse (Working), sports facility (Caring). Label order: [Living, Supplying, Working, Caring, Learning, Enjoying]. Red text indicates prediction errors.

scenarios. Within Seoul, we examine performance across apartment complex sizes in Table 5. Although in-domain performance is the highest, we also consider cross-group settings in which the model is trained on one of three size groups (small, medium, large) and tested on another (e.g., trained on large, tested on medium). Results show a clearer degradation in *function*-level metrics when training and testing scales diverge, indicating greater functional diversity across complex sizes. By contrast, *institution*-level performance remains stable across transfers (I-F1 $\approx$ 0.90), suggesting consistent spatial placement patterns regardless of scale. Medium-sized complexes emerge as the most robust training source, achieving strong cross-scale performance, likely because they capture intermediate urban characteristics present in both extremes.

To verify the generalizability of the proposed method beyond Seoul, South Korea, we evaluated its performance on two additional urban systems: Beijing, China, and New York, USA. Since data limitations prevent quantitative evaluation, we conducted case studies to demonstrate GeoTwin-MIL's applicability in other cities. Figure 8 shows the predictions for Beijing and New York using our Seoul-trained model without any fine-tuning. Although trained exclusively on Seoul data, the model captures general patterns of urban function in different geographic and cultural contexts. In Beijing, it correctly identifies the function of "Working" and "Learning" while overestimating the presence of "supplying". In New York's diverse neighborhood, the model accurately detects most functions, demonstrating that the learned representations generalize well beyond the city used for training. Although city-specific fine-tuning would improve the model's 15MC evaluation performance, these results validate that GeoTwin-MIL learns fundamental urban patterns transferable across different cities, offering practical value for regions with limited labeled data.

7 Conclusion

This study presents a novel framework that evaluates 15-minute city (15MC) compliance directly from satellite imagery, bypassing the limitations of POI-based applications. We construct an annotated Seoul dataset and systematically assess combinations of feature extraction, spatial aggregation, and multi-label classification. Our experiments show that the proposed model, GeoTwin-MIL, effectively captures both fine-grained institutional presence and broader functional completeness in diverse hyper-dense urban settings, validating the integrative effectiveness of cross-modal contrastive learning to bridge morphological-topological features and multiple instance learning for fine-grained aggregation.

Overall, the study findings support the validity of the proposed framework and indicate a scalable path toward near-real-time urban-function monitoring that can complement or potentially replace conventional approaches. Our present implementation relies on off-the-shelf contrastive learning and multiple-instance learning components, and systematically benchmarking state-of-the-art variants to quantify their task-level impact remains an important direction for future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2022-NR068758).

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A Gonzalez. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1701.01369* (2017).

[2] Thiago Carvalho, Steven Farber, Kevin Manaugh, and Ahmed El-Geneidy. 2025. Assessing the readiness for 15-minute cities: a literature review on performance metrics and implementation challenges worldwide. *Transport Reviews* (2025), 1–27.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[4] Brendan Duke and Graham W Taylor. 2018. Generalized Hadamard-Product Fusion Operators for Visual Question Answering. *arXiv preprint arXiv:1803.09374* (2018).

[5] M. Garnier and C. Moreno. 2022. *15-Minute City White Paper.* Technical Report. KRIHS. 36–39 pages. p. 154.

[6] Karst T Geurs, Lissy La Paix, and Sander Van Weperen. 2016. A multi-modal network approach to model public transport accessibility impacts of bicycle-train integration policies. *European transport research review* 8 (2016), 1–15.

[7] Eduardo Graells-Garrido, Feliu Serra-Burriel, Francisco Rowe, Fernando M Cucchietti, and Patricio Reyes. 2021. A city of cities: Measuring how 15-minutes urban accessibility shapes human mobility in Barcelona. *PloS one* 16, 5 (2021), e0250080.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Zongze He and Xiang Zhang. 2025. Towards More Reliable Measures for "Perceived Urban Diversity" Using Point of Interest (POI) and Geo-Tagged Photos. *ISPRS International Journal of Geo-Information* 14, 2 (2025), 91.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[12] Aline Nourma Iksanti. 2021. Walkability design study using urban network analysis in Tanah abang station area Jakarta. In *ARTEPOLIS 8-the 8th Biannual International Conference (ARTEPOLIS 2020)*. Atlantis Press, 87–95.

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*. PMLR, 2127–2136.

[14] Tanhua Jin, Kailai Wang, Yanan Xin, Jian Shi, Ye Hong, and Frank Witlox. 2024. Is a 15-minute city within reach? Measuring multimodal accessibility and carbon footprint in 12 major American cities. *Land Use Policy* 142 (2024), 107180.

[15] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. *arXiv preprint arXiv:1610.04325* (2017).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[17] Oded Maron and Tomás Lozano-Pérez. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems* 10 (1997).

[18] Carlos Moreno, Zaheer Allam, Didier Chabaud, Catherine Gall, and Florent Pratlong. 2021. Introducing the "15-Minute City": Sustainability, resilience and place identity in future post-pandemic cities. *Smart cities* 4, 1 (2021), 93–111.

[19] Carlos Moreno and Hyung Joon Park. 2024. *Busan Bets on the 15-Minute City.* Retrieved April 12, 2025 from https://nextcity.org/features/busan-south-korea-tech-hub-15-minute-city-carlos-moreno-park-hyung-joon Next City Feature Article.

[20] Kostas Mouratidis. 2024. Time to challenge the 15-minute city: Seven pitfalls for sustainability, equity, livability, and spatial analysis. *Cities* 153 (2024), 105274.

[21] Beatrice Olivari, Piergiorgio Cipriano, Maurizio Napolitano, and Luca Giovannini. 2023. Are Italian cities already 15-minute? Presenting the Next Proximity Index: A novel and scalable way to measure it, based on open data. *Journal of Urban Mobility* 4 (2023), 100057.

[22] Georgia Pozoukidou and Zoi Chatziyiannaki. 2021. 15-Minute City: Decomposing the new urban planning eutopia. *Sustainability* 13, 2 (2021), 928.

[23] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* 34 (2021), 2136–2147.

[24] Lu Song, Xuesong Kong, and Peng Cheng. 2024. Supply-demand matching assessment of the public service facilities in 15-minute community life circle based on residents' behaviors. *Cities* 144 (2024), 104637.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[26] Elias Willberg, Christoph Fink, and Tuuli Toivonen. 2023. The 15-minute city for all?–Measuring individual and temporal variations in walking accessibility. *Journal of Transport Geography* 106 (2023), 103521.

[27] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18802–18812.

[28] Shanqi Zhang, Feng Zhen, Yu Kong, Tashi Lobsang, and Sicong Zou. 2023. Towards a 15-minute city: A network-based evaluation framework. *Environment and Planning B: Urban Analytics and City Science* 50, 2 (2023), 500–514.