LocationBind: 地理空間座標を媒介とした異種都市モダリティデータの共通潜在表現学習

田村 直樹[1,a)]　寺島 青[1]　庄子 和之[2]　浦野 健太[1]　米澤 拓郎[1]　河口 信夫[1,2]

概要：都市では POI，衛星画像，人の移動履歴など多様なモダリティのデータが地理座標に紐づいて蓄積されている．都市コンピューティング分野では，これらのマルチモーダルデータを活用し，各場所の多面的な性質やモダリティ間の潜在的な対応関係を捉え，定量的な分析を可能にする枠組みが求められている．しかし，既存のアプローチは主に単一モダリティに依存しているか，あるいは画像–テキストなどの明示的なクロスモーダルペアを要するが，空間解像度の違いやデータの欠損・偏りにより，都市環境ではこうしたペアが大量に入手できない場合が多い．本論文では，地理座標を普遍的な媒介表現として活用し，異なる都市モダリティを共通の潜在空間上の表現として統合的にモデル化する LocationBind を提案する．地理座標と単一モダリティの整合に留まる既存のユニモーダル整合手法，およびモダリティ間の明示的なペアを必要とする従来のマルチモーダル手法のいずれとも異なり，LocationBind は地理座標-モダリティ（POI テキスト，衛星画像など）のペアのみを活用した学習によって，複数モダリティの共通潜在空間への整合を実現する．大阪と名古屋の 2 都市を対象とした実験では，都市の各場所の土地指標（土地利用・人口密度など）の予測や欠損モダリティの補完，クロスモーダル検索といったタスクにおいて，LocationBind が一貫してベースラインを上回り，都市の多様なモダリティを統合的にモデル化できることを示した．さらに，モダリティ横断的なパターン抽出，特定地点のマルチモーダルな特徴検索，マルチスケールな場所検索といった多様な応用分析例を示し，都市の多面的な理解を支援する汎用的な分析基盤として機能する可能性を示した．

LocationBind: Coordinate-Pivot Alignment of Heterogeneous Urban Modalities in a Unified Latent Space

Naoki Tamura[1,a)]　Haru Terashima[1]　Kazuyuki Shoji[2]　Kenta Urano[1]　Takuro Yonezawa[1]
Nobuo Kawaguchi[1,2]

1. はじめに

都市は人々の活動，交通，経済，インフラが高密度かつ複雑に集積する場である [3]．都市コンピューティングや都市データ分析分野では，こうした都市の実態を把握，モデル化，予測するために大規模な時空間データがますます活用されており，都市計画，交通最適化，防災，インフラ維持管理，環境保護，公衆衛生などへの幅広い応用が期待されている [11], [39]．同時に，都市で観測されるデータは急速に多様化しており，Points of Interest（POI），口コミ・アンケート文書，衛星画像，ストリートビュー画像，人の移動ログなどマルチモーダルなデータが地理座標に紐づき大規模に蓄積されており，オープンデータ化の試みも進行している [6], [28]．これらのデータは都市の同一地点について複数の側面からの実態把握を可能にする．一方で，これらマルチモーダルデータを統合的にモデル化し，モダリティ横断的に理解・検索・推論できる枠組みは依然として不足している．

既存の都市データモデリングの代表的な手法として，埋め込み表現学習手法が提案されている．これらの手法では地理

[1] 名古屋大学大学院 工学研究科
Graduate School of Engineering, Nagoya University
[2] 名古屋大学 未来社会創造機構
Institutes of Innovation for Future Society, Nagoya University
[a)] tam@ucl.nuee.nagoya-u.ac.jp

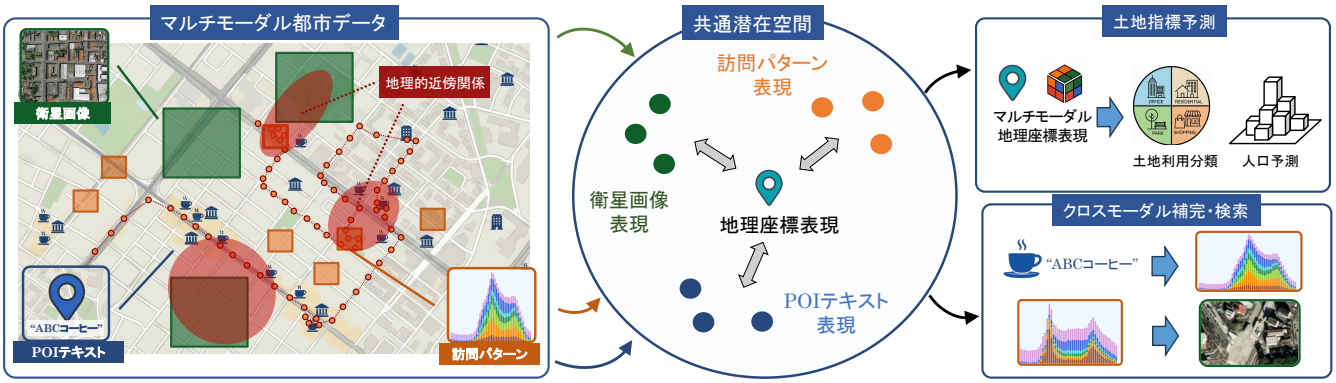


図 1: LocationBind によるマルチモーダル都市データの整合.

空間に紐づく衛星画像や人の移動データといったモダリティデータを，自己教師ありベースの学習で潜在空間上のベクトル表現として埋め込む [7], [10], [14], [16], [26], [37], [38]. 特に GeoCLIP[30] や SatCLIP[12]，CaLLiPer[32] に代表される手法では，モダリティ表現と地理座標表現の整合的な学習によって，連続的な地理座標レベルで，埋め込み表現の獲得を可能にしている．しかしこれら既存手法の多くは，ユニモーダルな観測源に依存しており，GeoCLIP や SatCLIP，CaLLiPer といった地理座標–モダリティ整合手法についても，地理座標表現と単一のモダリティ表現を 1 対 1 で整合させる枠組みに留まっている．これらの手法をモダリティごとに個別学習し表現を組み合わせるという拡張も考えられるが，各モダリティがそれぞれ独立した潜在空間を構成するため，モダリティ間で表現空間が共有されない．結果として，モダリティ横断的な検索，補完，パターンの発見といった表現の活用が困難である．加えて，特定のモダリティに依存する設計は，各座標の多面的な意味を捉えられず，さらに対象モダリティの時間的・空間的な偏りや欠損に対して脆弱になりうる．

一方，ImageBind[8] や KnowCL[16] をはじめとするマルチモーダル表現学習手法は，複数のモダリティを共通潜在空間に整合する枠組みを提供する．しかしこれらの手法は，モダリティ間の対応関係を学習するために，明示的なクロスモーダルペアや，特定の離散的な空間単位（地域ノード・メッシュ）への観測データの集約を前提とする．しかし実際の都市データでは，モダリティごとに空間解像度や観測密度が異なり，データ収集条件も地域や時間によって変動するため，任意のモダリティ組み合わせに対する安定したペアの大規模な構築は困難である．

本研究の核となる洞察は，「同一地点（あるいはその周辺）で観測される異種モダリティのデータは，その地点が持つ都市的性質を異なる側面から反映したものであり，潜在的な相互関係を共有している」という仮説である．この仮説のもとでは，モダリティ間の明示的なペアが与えられなくとも，地理座標表現という共通の媒介を介した同一の潜在空間への整合によって，モダリティ間の対応関係を間接的に獲得できる．本研究では，この仮説に基づき，都市におけるマルチモーダルデータを統合的にモデル化する枠組みとして LocationBind を提案する．LocationBind の全体像を図 1，学習アーキテクチャを図 2 に示す．LocationBind は，地理座標を共通潜在空間上の表現に変換する LocationEncoder，各モダリティ埋め込みを同一空間に写像する ModalityProjector，および共通潜在表現からモダリティ埋め込みを再構成する ModalityDecoder から構成される．地理座標を媒介とする整合学習により，明示的なクロスモーダルペアや特定の空間単位への集約を必要とせず，異なる空間解像度・観測密度を持つ複数のモダリティを単一の共通潜在空間に統合する．この設計は，既存手法では原理的に困難であったモダリティ横断的な表現の活用，すなわち欠損モダリティの補完，クロスモーダル検索，および複数モダリティを統合した地理座標表現の獲得を可能にする．さらに，応用例と定性的な分析を通じて，本フレームワークが都市を多面的に理解するための汎用的な分析基盤として機能しうることを示す．

本研究の主な貢献は以下の通りである．

- 地理座標表現を媒介として，明示的なクロスモーダルペアを必要とせずに，複数の都市モダリティを共通潜在空間に埋め込み，モダリティ横断的な表現の活用を可能にする枠組み，LocationBind を提案した．

- LocationBind がマルチモーダル情報を効果的に統合し，土地指標予測，欠損モダリティ補完，クロスモーダル検索の各タスクにおいてベースライン手法を上回る性能を示すことを確認した．

- 共有潜在空間内で異なる都市モダリティ間のパターン発見や，特定のパターンを持つ場所の検索などのアプリケーションへの活用可能性について示した．

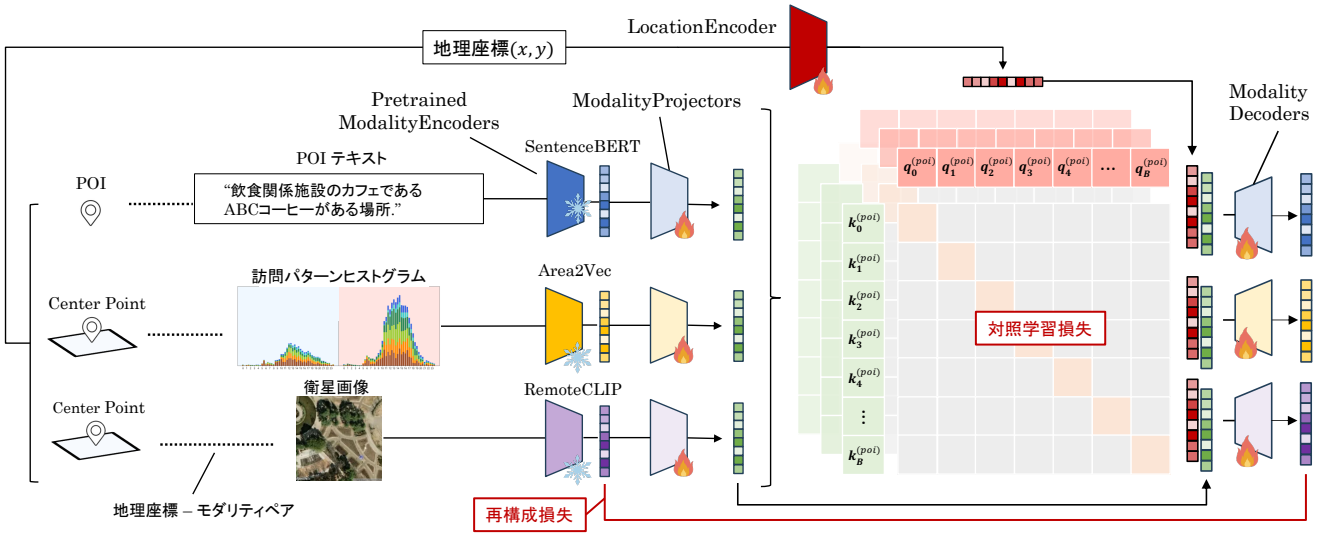


図 2: LocationBind のアーキテクチャ.

2. 関連研究

2.1 都市データのユニモーダル埋め込み

都市データの利用可能性が高まるにつれ，都市データに内在する潜在的なパターンを解明するための，さまざまな自己教師付き表現学習アプローチが提案されてきた．これらの手法の多くは，データを点やグリッドセルなどの離散的な空間に集約し，潜在空間に埋め込むものである．既存の研究では，POI [10], [35], [37]，衛星画像 [1], [4], [18], [23]，ストリートビュー画像 [14], [27]，および移動ログ [5], [7], [25], [26], [28], [31], [38] といったモダリティを扱っている．これらの単一モダリティの埋め込みは，地理空間可視化や，様々な下流タスク向けの転移可能な特徴量として利用されてきた．

さらに近年の注目すべき進展として，連続的な地理座標を埋め込み空間にマッピングする LocationEncoder [17], [19], [20] の登場が挙げられる．従来の研究では，地理情報を通常，離散的かつ粗い空間スケール（数百メートルから数キロメートル程度）でモデル化していたのに対し，LocationEncoder は，数メートルから数十メートルという微細な空間解像度で位置を表現しつつ，連続的な空間的近接性を明示的にモデル化できる．多くの LocationEncoder の重要な構成要素は，マルチスケール位置エンコーディングであり，これにより大域的および局所的な地理的関係を同時にモデル化可能になった．これに伴いいくつかの手法では，連続座標ベースの地理座標埋め込みとモダリティ固有の埋め込みとの間の対照学習を実行している．GeoCLIP [30] は，場所に紐づく画像埋め込みと地理座標埋め込みを整合させ，地理的位置特定を可能にしている．SatCLIP [12] は，衛星画像と地球規模の locationEncoder を対照学習させ，その結果得られた地理座標埋め込みを，気温予測，野生生物の認識，人口密度の推定など，地理に依存するタスクに転移している．CaLLiPer [32] は，POI テキスト埋め込みと地理座標埋め込みを整合させ，意味的な地理座標埋め込みを獲得し，これらの埋め込みを土地利用分類や社会経済的属性の予測といった下流タスクに適用している．これらの手法は地理座標表現と 1 種類のモダリティ表現を整合させる枠組みであり，本論文ではユニモーダル手法として位置付ける．

2.2 マルチモーダル表現学習

単一のモダリティに依存する表現学習には，特に観測データが疎であったり欠損していたりする場合，あるいは複数モダリティが統合された表現を通じて場所の多面的な性質を捉える必要がある場合に，本質的な限界がある．このことから，複数のモダリティを統合的にモデル化するマルチモーダル表現学習に関する広範な研究が進められている．まず都市領域外では，テキストと画像のペアを用いた対照学習が台頭しており，CLIP [22] がその代表例である．CLIP は，大規模な画像・テキストペアから共有された潜在空間を学習し，モダリティ横断的な検索やゼロショット分類を可能にする．さらに，マルチモーダル学習は 2 つのモダリティを超えて拡張され，単一の共通潜在空間上への多数のモダリティの整合が実現されつつある．ImageBind [8] は，画像をアンカーモダリティとして扱い，画像とペアで収集された音声，深度，熱信号，IMU 測定値を含む多様な観測データを画像埋め込み空間にマッピングする．このようなアンカーモダリティによる整合は，すべてのモダリティの組み合わせにわたる完全なペアリングが利用できない現実的な設定において有望であり，このような状況におけるモダリティ横断的な検索や転移を可能にする．

都市コンピューティング分野においても，これらのマル

チモーダル表現学習について検討した研究がいくつか行われている [2], [9], [13], [16], [33], [36]．代表的な例として，KnowCL では，都市の地域や POI の空間的な近接性や類似性をグラフで表した都市グラフを作成し，このグラフ表現と衛星画像やストリートビュー画像を同一の潜在空間に整合している．また MoRA [34] では，人の移動データから得られたグラフ表現をアンカーモダリティとして用い，その他のモダリティを同一潜在空間に整合している．これらの研究は，マルチモーダル統合が多くの設定において下流タスクの性能を向上させ，単一モダリティのアプローチよりも場所に対する包括的な理解を可能にすることを示唆している．

これらの既存のマルチモーダル表現学習の多くは，モダリティ間の対応関係を，明示的なペアや特定の空間単位への集約を介して整合する必要がある．しかし実際には，都市の各モダリティは空間解像度や観測密度が異なり，データ収集条件も地域や時間によってしばしば異なるため，任意のモダリティの組み合わせに対する安定したペアの構築は困難である．

3. 提案手法

LocationBind は，図 2 に示すように，地理座標とモダリティ観測のペアを用いて学習される．各学習インスタンスは，1 つの地理座標と，いずれか 1 つのモダリティから得られる観測データで構成される．本モデルでは，地理座標をエンコードする LocationEncoder，モダリティ固有埋め込みを共通潜在空間へ写像する ModalityProjector，および共通潜在空間上の表現からモダリティ固有埋め込みを再構成する ModalityDecoder を共同で学習する．

生のモダリティデータは，まず事前学習済みの ModalityEncoder によってモダリティ固有埋め込みへ変換される．これらの ModalityEncoder の重みは学習中は固定される．この設計により，大規模な ModalityEncoder の重みを最適化する必要がなくなり，座標とモダリティ観測のペアに基づく安定した整合学習が可能となる．学習では，共通潜在空間における地理座標表現とモダリティ表現の対照損失，および共通潜在空間上の表現から元のモダリティ埋め込みを再構成する損失を同時に最小化する．対照損失により，ペアとなった地理座標表現とモダリティ表現が互いに近づくように整合され，異なるサンプルの表現は引き離される．一方，再構成損失により，共通潜在空間上の表現が，元のモダリティ埋め込みを復元するために必要な情報を保持するよう促される．

3.1 ペアデータとモダリティ固有埋め込み

本研究では，地理座標とモダリティ埋め込みのペア集合を，モダリティごとに独立に構成する．モダリティ m に対するペア集合を $\mathcal{D}^{(m)} = \{(\ell_i^{(m)}, z_i^{(m)})\}_{i=1}^{N_m}$ と表す．ここで，$N_m = |\mathcal{D}^{(m)}|$ はモダリティ m におけるペア数，$\ell_i^{(m)} = (x_i, y_i)$ は平面上の地理座標，$z_i^{(m)}$ はモダリティ m に対する事前学習済み ModalityEncoder により得られる固定の埋め込みである．本研究では，POI テキスト，訪問パターンヒストグラム，衛星画像の 3 種類のモダリティを扱う．以下では，各モダリティにおけるデータ構成と，$z_i^{(m)}$ の取得方法について述べる．

POI テキスト　POI テキストは，地理空間上に存在する建物，施設，ランドマークなどのオブジェクトについて，そのカテゴリと名称を自然文として表現したものである．CaLLiPer[32] に従い，本研究では 2 段階の POI のカテゴリ名と POI 名称を日本語の一文へ整形する．例えば，大カテゴリが飲食，小カテゴリがカフェ，名称が ABC コーヒーである場合，「飲食関係施設のカフェである ABC コーヒーがある場所.」のようなテキストを構成する．この POI テキストを，日本語文書で事前学習済みの Sentence-BERT[24] によりエンコードし，POI テキスト埋め込み $z^{(\text{poi text})}$ を得る．地理座標 ℓ には，当該 POI の座標を用いる．これにより，その地点にどのような施設やランドマークが存在するかといった意味的特徴を表現できる．

訪問パターンヒストグラム　訪問パターンヒストグラムは，Area2Vec[26] に従い，エリアごとの滞在分布を集約した特徴量に基づいて構成する．具体的には，都市空間上の各エリアで観測された訪問を集計し，平日・休日の区分，24 時間の時間帯，および滞在時間長に基づいて訪問分布を作成する．本研究では，平日・休日を 2 クラス，時間帯を 24 クラス，滞在時間長を 7 クラスとし，$2 \times 24 \times 7$ 次元の頻度ベクトルを構成する．この頻度ベクトルに基づいて Area2Vec を学習し，得られた訪問パターン埋め込み $z^{(\text{visit pattern})}$ を以後の学習で固定の訪問パターン表現として用いる．地理座標 ℓ には，各エリアの中心座標を用いる．このモダリティは，各地点の周辺エリアにおける人の滞在傾向や時間的利用パターンといった行動的特徴を捉える表現として機能する．

衛星画像　衛星画像モダリティは，都市内の各地点を中心として取得した衛星画像により構成する．取得した画像を，大規模衛星画像データセットで事前学習された RemoteCLIP[15] によりエンコードし，衛星画像埋め込み $z^{(\text{satellite})}$ を得る．地理座標 ℓ には，対応する衛星画像の中心座標を用いる．これにより，建物の形状や密度，道路パターンなど，視覚的な地表構造に関する特徴を取り込むことができる．

3.2 マルチモーダル整合学習

LocationBind では，地理座標に対する LocationEn-

coder f，各モダリティ埋め込みに対する ModalityProjector $\{g^{(m)}\}$，および元のモダリティ埋め込みを再構成する ModalityDecoder $\{h^{(m)}\}$ を，対照損失と再構成損失の和により同時に最適化する．まず，LocationEncoder f は，地理座標 $\ell = (x, y)$ を共通潜在空間上の表現へと埋め込む．LocationEncoder には，モダリティごとに異なる空間解像度やデータ密度の違いを吸収しつつ，地理座標間のマルチスケールな近傍関係を潜在空間上で表現する能力が求められる．そこで本研究では，CaLLiPer[32] に倣い，Grid 位置エンコーディング（Positional Encoding，PE） [20] とニューラルネットワークを組み合わせた構造を用いる．LocationEncoder f は次式で表される．

$$
\begin{aligned}
f(x,y) &= NN(PE(x,y)) \\
PE(x,y) &= \Big[\cos\frac{x}{\alpha_0}, \sin\frac{x}{\alpha_0}, \cos\frac{y}{\alpha_0}, \sin\frac{y}{\alpha_0}, \ \cdots \\
&\qquad \cos\frac{x}{\alpha_{S-1}}, \sin\frac{x}{\alpha_{S-1}}, \cos\frac{y}{\alpha_{S-1}}, \sin\frac{y}{\alpha_{S-1}}\Big] \\
\alpha_s &= r_{\min} \cdot \left(\frac{r_{\max}}{r_{\min}}\right)^{\frac{s}{S-1}}
\end{aligned}
\tag{1}
$$

ここで，(x, y) はメートル単位の平面直角座標系における 2 次元地理座標である．NN はニューラルネットワークであり，本研究では全結合残差ネットワーク（fully connected residual network, FC-Net）を用いる．$r_{\min}$ と $r_{\max}$ はそれぞれ最小および最大の空間スケールを表し，S はスケール数を表す．これらのパラメータにより，LocationEncoder が捉える地理空間上の周波数帯域と解像度が決定される．この位置エンコーディングにより，地理座標 ℓ を，大域的な位置関係と局所的な近傍関係の双方を考慮した d 次元の表現にエンコードする．

次に，各モダリティに対する ModalityProjector $\{g^{(m)}\}$ は，モダリティごとに独立したニューラルネットワークとして実装する．各 ModalityProjector は 1 つの中間層を持つ多層パーセプトロン（Multi-Layer Perceptron, MLP）であり，モダリティ固有の表現空間の違いを吸収しながら，各モダリティ埋め込みを共通潜在空間へ写像する．具体的には，入力となるモダリティ埋め込みの次元を d_m とすると，中間層の次元を $2d_m$，出力次元を地理座標表現と同じ d 次元とする．

さらに，共通潜在空間上の表現から元のモダリティ埋め込みを再構成する ModalityDecoder $\{h^{(m)}\}$ を，モダリティごとに独立したニューラルネットワークとして実装する．各 ModalityDecoder は，ModalityProjector と対称な構造を持つ 1 つの中間層からなる MLP である．入力次元を共通潜在空間の次元 d，中間層の次元を $2d_m$，出力次元を元のモダリティ埋め込みと同一の d_m 次元とする．また，デコーダの出力には L2 正規化を適用する．

学習では，対照損失と再構成損失を同時に最小化する．まず，対照損失では，エンコードされた地理座標表現と，写像後のモダリティ表現が共通潜在空間上で整合するように学習する．モダリティ m に対する地理座標 $\ell_i^{(m)}$ とモダリティ埋め込み $z_i^{(m)}$ のペアに対し，地理座標表現を $\boldsymbol{q}_i^{(m)} = normalize(f(\ell_i^{(m)}))$，モダリティ表現を $\boldsymbol{k}_i^{(m)} = normalize(g^{(m)}(z_i^{(m)}))$ として，それぞれ L2 正規化された形で得る．ミニバッチ内の他のサンプルを負例として用い，場所からモダリティへの InfoNCE 損失 [29] を次式で定義する．

$$
\mathcal{L}_{l\to m}^{(m)} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\boldsymbol{q}_i^{(m)\top}\boldsymbol{k}_i^{(m)}/\tau\right)}{\sum_{j=1}^{B}\exp\left(\boldsymbol{q}_i^{(m)\top}\boldsymbol{k}_j^{(m)}/\tau\right)}, \tag{2}
$$

ここで，τ は温度パラメータであり，本手法では $\tau = 0.07$ に固定する．B はミニバッチサイズである．この損失により，共通潜在空間上で地理座標表現 $\boldsymbol{q}_i^{(m)}$ と対応するモダリティ表現 $\boldsymbol{k}_i^{(m)}$ が近づき，異なるサンプルに対応する表現は相対的に遠ざかる．同様に，モダリティから場所への逆方向の損失 $\mathcal{L}_{m\to l}^{(m)}$ も導入する．

加えて，共通潜在空間がモダリティ固有の情報を保持するよう促し，共通潜在空間からのモダリティ生成，すなわちモダリティ補完を可能にするため，ModalityDecoder による再構成損失を導入する．デコーダへの入力としては，地理座標表現 $\boldsymbol{q}_i^{(m)}$，モダリティ表現 $\boldsymbol{k}_i^{(m)}$，および両者の平均のいずれかを確率的に選択する．この再構成入力 $\boldsymbol{u}_i^{(m)}$ を次式で定義する．

$$
\boldsymbol{u}_i^{(m)} = \mathrm{normalize}\left(\begin{cases} (\boldsymbol{k}_i^{(m)} + \boldsymbol{q}_i^{(m)})/2 & \text{w.p. } p_{\text{both}} \\ \boldsymbol{q}_i^{(m)} & \text{w.p. } p_{\text{loc}} \\ \boldsymbol{k}_i^{(m)} & \text{w.p. } p_{\text{mod}} \end{cases}\right). \tag{3}
$$

ただし，$p_{\text{both}} + p_{\text{loc}} + p_{\text{mod}} = 1$ であり，本研究では $p_{\text{both}} = 0.5,\ p_{\text{loc}} = p_{\text{mod}} = 0.25$ と設定する．このように再構成入力を確率的に切り替えることで，デコーダは，地理座標表現単独，モダリティ表現単独，および両者の平均のいずれからでも元のモダリティ埋め込みを再構成できるように学習される．再構成損失では，元のモダリティ埋め込み $z_i^{(m)}$ を L2 正規化した $\bar{z}_i^{(m)} = z_i^{(m)}/\|z_i^{(m)}\|_2$ をターゲットとして用いる．デコーダ出力とのコサイン類似度に基づき，モダリティ m に対する再構成損失を次式で定義する．

$$
\mathcal{L}_{\text{decode}}^{(m)} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - h^{(m)}(\boldsymbol{u}_i^{(m)})^\top \bar{z}_i^{(m)}\right). \tag{4}
$$

最終的な学習損失は，全モダリティについて双方向の対照損失と再構成損失を統合し，モダリティ数で平均したものとして次式で定義する．

表 1: 各モダリティデータの統計と入力埋め込み.

モダリティ	取得元・期間	入力埋め込み（次元）	空間単位	データ量（大阪）	データ量（名古屋）
POI テキスト	Foursquare [6], 2023	SentenceBERT(768)	Point	188,390	78,224
訪問パターン	ブログウォッチャー [41], 2023	Area2Vec(64)	50m グリッド	64,880	74,202
衛星画像	Mapbox [21], 2025	RemoteCLIP(512)	250m グリッド	4,320	6,096

表 2: LocationBind の主要ハイパーパラメータ.

パラメータ	値
共通潜在空間の次元 d	128
入力次元 $(d_{\text{poi}}, d_{\text{visit}}, d_{\text{sat}})$	$(768, 64, 512)$
Grid PE $(S, r_{\min}, r_{\max})$	$(2048, 50\text{m}, 20000\text{m})$
FC-Net 隠れ次元 h	4096
Projector / Decoder 中間層	$2d_m$
ドロップアウト率	Loc. 0.1, Proj. 0.5, Dec. 0.1
λ_{decode} / τ	1.0 / 0.07
融合確率	$(0.5, 0.25, 0.25)$
Optimizer / batch / epochs	AdamW / 128 / 100
学習率	Loc. 10^{-5}, その他 10^{-4}
重み減衰	Loc. 0.1, Proj. 0.5, Dec. 0.1

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left[\frac{1}{2} \left(\mathcal{L}^{(m)}_{l \to m} + \mathcal{L}^{(m)}_{m \to l} \right) + \lambda_{\text{decode}} \mathcal{L}^{(m)}_{\text{decode}} \right], \tag{5}$$

ここで，λ_{decode} は対照損失に対する再構成損失の重み係数である．この目的関数により，対照損失を通じて共通潜在空間上で地理座標表現と各モダリティ表現が互いに近づくように整合され，地理座標表現を介して異なるモダリティの表現同士も間接的に整合される．さらに再構成損失により，共通潜在空間が各モダリティ固有の情報を保持するよう促される．

4. 実験

本論文では日本の大阪と名古屋の 2 都市地域を対象に実験を行う．両市は人口規模・都市構造の点で日本の主要大都市圏の中核を成す都市であり，都心部から郊外まで多様な都市空間構成を包含し，都市データのオープン化も進んでいることからテストベッドとして適している．この 2 都市を対象として，まず各モダリティのデータを収集し，都市ごとにモデルの学習を行う．その後学習モデルの性能を評価するため，土地指標予測，欠損モダリティ補完，クロスモーダル検索の 3 つの下流タスクを実施する．

4.1 LocationBind の学習

まず，両都市について，POI，訪問パターン，衛星画像という 3 種類のマルチモーダルデータを収集する．各モダリティにおけるデータ統計を表 1 に示す．POI データは Foursquare [6] から取得した．各 POI エントリには，POI 名称と，レベル 1 から最大レベル 5 までの階層的なカテゴリラベルが含まれ，レベル 1 のカテゴリを大カテゴリ，利用可能なラベルの中で最も細かいカテゴリを小カテゴリとして用い，POI 名称と組み合わせて日本語の POI テキストを構築する．得られた POI テキストは，日本語文書で事前学習済みの SentenceBERT [24] によりエンコードされる．

訪問パターンについては，ブログウォッチャー [41] から GPS に基づく人の滞在ログデータを取得した．本データは，同意を得たユーザについて，移動ログの抽象化後に滞在推定処理され，最終的に緯度，経度，滞在開始時刻，および滞在時間長を含むデータとして取得した．本研究では，この滞在データを 50m グリッドメッシュごとに集計し，Area2Vec を事前学習する．ただし，ユーザのプライバシーを担保するため，訪問数が 100 未満のメッシュは Area2Vec と LocationBind の事前学習の対象から除外する．

衛星画像は Mapbox API [21] を用いて取得した．各都市を 250m グリッドに分割し，各グリッドセルの重心を中心として，ズームレベル 17 の 512×512 ピクセル画像をダウンロードする．過去のデータにアクセスできなかったため，衛星画像特徴は数年では大きく変化しないと考え，本論文では執筆時点で最新の 2025 年のデータを取得した．その後，RemoteCLIP [15] の前処理に従い，画像を 224×224 ピクセルへリサイズし，RemoteCLIP ViT-B/32 により 512 次元の衛星画像埋め込みを得る．

次に，LocationBind の学習設定について述べる．本研究では，大阪と名古屋の各都市について，同一のモデル構成および学習設定を用いて都市ごとに LocationBind を学習する．主要なハイパーパラメータを表 2 に示す．学習では，地理座標表現とモダリティ表現を整合する双方向の InfoNCE 損失と，元のモダリティ埋め込みを復元する再構成損失を同時に最小化する．デコーダへの入力には，地理座標表現，モダリティ表現，および両者の平均表現を確率的に用いる．これにより，地理座標表現からのモダリティ補完と，モダリティ表現からの再構成の双方を学習できるようにする．

最適化には AdamW を用い，線形ウォームアップ後にコサイン減衰する学習率スケジューラを適用する．学習率は LocationEncoder とその他のモジュールで分けて設定し，対照損失の温度パラメータは固定する．バッチサイズ，学習エポック数，損失重み，融合確率などの主要設定は表 2 に示す．より詳細な層構成，正則化設定，およびパラメー

表 3: 土地指標予測の結果．各値は 5-fold の平均 ± 標準偏差．LUC は Macro-F1(↑)，PR および BHR は RMSE(↓).

	Osaka			Nagoya		
Model(dim)	LUC (↑)	PR (↓)	BHR (↓)	LUC (↑)	PR (↓)	BHR (↓)
Area2Vec(64)	0.372 ± 0.005	0.779 ± 0.016	0.793 ± 0.021	0.361 ± 0.003	0.809 ± 0.010	0.799 ± 0.023
RemoteCLIP(512)	0.382 ± 0.005	0.813 ± 0.017	0.787 ± 0.018	0.379 ± 0.005	0.817 ± 0.008	0.789 ± 0.017
GeoCLIP (512)	0.263 ± 0.002	0.876 ± 0.020	0.859 ± 0.020	0.223 ± 0.003	0.884 ± 0.010	0.851 ± 0.020
SatCLIP (256)	0.393 ± 0.000	0.779 ± 0.016	0.753 ± 0.018	0.369 ± 0.002	0.790 ± 0.010	0.768 ± 0.019
CaLLiPer(128)	0.367 ± 0.006	0.832 ± 0.018	0.801 ± 0.016	0.323 ± 0.003	0.869 ± 0.008	0.840 ± 0.017
MoRA(128)	0.393 ± 0.004	0.789 ± 0.015	0.767 ± 0.019	0.354 ± 0.001	0.815 ± 0.009	0.784 ± 0.017
LocationBind(128)	**0.481 ± 0.004**	**0.728 ± 0.016**	**0.706 ± 0.017**	**0.457 ± 0.004**	**0.766 ± 0.007**	**0.754 ± 0.016**

タグループの設定は付録に示す．LocationBind の学習には基本的に各都市の全座標–モダリティペアを用いる．ただし後述する欠損モダリティ実験では fold ごとに一部のモダリティデータを欠損させる．

4.2 評価タスクと指標

実験では LocationBind の定量的な評価のため，土地指標予測，欠損モダリティ補完，クロスモーダル検索の 3 つの下流タスクを定義する．まず土地指標予測タスクでは，事前学習された LocationEncoder が各地理座標の多面的な特徴を十分に捉えているか評価するため，従来研究の地理的埋め込み評価 [12], [15], [32], [34] に倣い，各 50m グリッドメッシュの中心点の地理座標表現を隠れ層を 1 つ持つ MLP に入力し，そのメッシュの土地指標予測を行う．連続的な地理座標を埋め込めない手法 [15], [26], [34] については，各離散的な空間単位で事前学習された埋め込み表現を入力として用いる．本論文では，土地利用分類 (LUC)，人口密度回帰 (PR)，建物高度回帰 (BHR) という 3 つのタスクで評価する．LUC は各地域に対応した 11 の土地利用クラスの分類タスクであり，Macro-F1 スコアで評価する．ここで扱う土地利用クラスについては，各自治体から土地利用調査データ [40], [44] を入手し，50m グリッド上で 11 のクラスに統一して再マッピングする．PR は人口密度，BHR は地域内の平均建物高度を回帰で予測するタスクであり，二乗平均平方根誤差 (RMSE) で評価する．人口密度については，国勢調査 [43]，平均建物高度は PLATEAU[42] から取得し，これらを 50m グリッドごとの人口密度と平均建物高度として集計する．またこれらの値は都市ごとに標準化する．最後に各土地指標予測については，5-fold 交差検証を行い，平均と標準偏差で評価する．

次に欠損モダリティ補完では，モダリティ $m \in \{\mathrm{poi}, \mathrm{visit}, \mathrm{sat}\}$ を欠損対象とし，他の観測済みモダリティから真の埋め込み $z^{(m)}$ を補完できるかを評価する．テスト分割には 2 種類を用いる．*Random* は全データからランダムに一部を欠損させる設定，*Region* は地理座標 (x, y) のみを特徴量とした K-Means により両都市を 5 つの隣接領域に分割し，fold k ごとに 1 領域全体を欠損させる設定である．いずれの設定においても，欠損対象モダリティ m の fold k を学習データから除外して LocationBind を再学習する．LocationBind の補完では，対象地点に地理的に最近傍な他モダリティ $m' \neq m$ の埋め込みを ModalityProjector で共通潜在空間に写像し，その平均表現を ModalityDecoder $h^{(m)}$ に入力することで $\hat{z}^{(m)}$ を得る．なお，LocationBind では地理座標のみからの補完も可能であるが，本実験では地理座標エンコーダを持たない ImageBind/MoRA との公平性を重視し，観測モダリティのみを用いる手順に統一する．評価指標には $\hat{z}_i^{(m)}$ と $z_i^{(m)}$ のコサイン類似度を用い，欠損対象モダリティごとに 5-fold の平均値と標準偏差として報告する．

最後にクロスモーダル検索では，モダリティ横断的なデータの検索タスクを行う．具体的にはモダリティごとに，そのモダリティから別のモダリティのデータを共通潜在空間上で検索する．本研究の学習データは地理座標と各モダリティのペアから構成され，POI テキスト・訪問パターン・衛星画像の間に明示的なモダリティ間ペアは含まれない．そこで評価時には，地理的対応関係に基づきモダリティ間ペアを構築し，一方のモダリティ表現をクエリとして他方を共通潜在空間上で検索する．POI–訪問パターンペアでは，各 POI の周囲 15m の訪問パターンを集計し，学習時と同一の Area2Vec 空間に埋め込んだ表現と，同 POI のテキストを Sentence-BERT でエンコードした埋め込みを 1 対 1 の正例ペアとする．POI–衛星画像ペアでは，各衛星画像タイルに対応する 250m グリッド bounding box 内の POI を正例とする．訪問パターン–衛星画像ペアでは，同 bounding box 内の 50m グリッド訪問パターン埋め込みを正例とする．衛星画像タイルが複数の POI または訪問パターンメッシュに対応する場合，各 POI テキスト，訪問パターン埋め込みを平均しペアとする．検索評価では，クエリ側および候補側の埋め込みをそれぞれ ModalityProjector により共通潜在空間へ写像し，正例が近傍に配置されているか評価する．具体的には各正例ペアについて，負例を 1000 件サンプリングし，正例と負例とのペア間の共通潜在空間上でのコサイン類似度に基づき検索を行い，R@1 および R@10 を報告する．

表 4: 欠損モダリティ補完の結果．コサイン類似度 (↑) の 5-fold の平均 ± 標準偏差．POI・Visit・Sat は欠損モダリティ．

City	Model	POI		Visit		Sat	
		random	*region*	*random*	*region*	*random*	*region*
Osaka	ImageBind	0.445 ± 0.006	0.423 ± 0.016	0.711 ± 0.008	0.706 ± 0.024	0.753 ± 0.004	0.740 ± 0.025
	MoRA	0.466 ± 0.009	0.455 ± 0.030	0.714 ± 0.010	0.709 ± 0.029	0.742 ± 0.008	0.737 ± 0.044
	LocationBind	**0.642 ± 0.000**	**0.618 ± 0.021**	**0.814 ± 0.001**	**0.798 ± 0.015**	**0.818 ± 0.003**	**0.809 ± 0.039**
Nagoya	ImageBind	0.423 ± 0.003	0.389 ± 0.007	0.634 ± 0.007	0.624 ± 0.040	0.746 ± 0.001	0.723 ± 0.019
	MoRA	0.448 ± 0.004	0.428 ± 0.018	0.667 ± 0.011	0.652 ± 0.014	0.747 ± 0.004	0.735 ± 0.017
	LocationBind	**0.599 ± 0.001**	**0.562 ± 0.008**	**0.756 ± 0.001**	**0.738 ± 0.026**	**0.825 ± 0.002**	**0.812 ± 0.020**

表 5: クロスモーダル検索の結果．各値は検索の Recall@k (↑).

	Osaka						Nagoya					
	POI–Visit		POI–Sat		Visit–Sat		POI–Visit		POI–Sat		Visit–Sat	
Model	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
ImageBind	0.004	0.032	0.093	0.265	0.059	0.210	0.004	0.035	0.207	0.449	0.037	0.154
MoRA	0.004	0.032	0.020	0.107	0.016	0.098	0.005	0.041	0.016	0.080	0.009	0.060
LocationBind	**0.007**	**0.050**	**0.331**	**0.618**	**0.614**	**0.866**	**0.011**	**0.081**	**0.391**	**0.673**	**0.450**	**0.749**

4.3 ベースライン手法

土地指標予測タスクにおいて，我々は LocationBind を代表的な地理データ埋め込み手法と比較する．まず GeoCLIP [30] は，位置エンコーダをストリートレベル写真などのジオタグ付き画像と整合する手法である．SatCLIP [12] は，衛星画像表現と組み合わせて位置エンコーダを学習させる．CaLLiPer [32] は，対照学習を用いて位置エンコーダと POI のテキスト表現を整合させる．GeoCLIP [30] については，公開されている事前学習済み重みを使用する．SatCLIP [12] および CaLLiPer [32] は地理座標表現と衛星画像表現または POI テキスト表現を整合学習する手法である．SatCLIP および CaLLiPer については，LocationBind と同一のモダリティデータ，モダリティ埋め込み，Grid+FC-Net 位置エンコーダ設計および位置エンコーディングのハイパーパラメータを用いて学習する．さらに学習で用いた Area2Vec[26]，RemoteCLIP[15] についても，各グリッドの埋め込みをベースラインとする．MoRA[34] についても，学習済みの共通潜在空間における地域表現を 50m グリッドに対応付け，土地指標予測の入力特徴量として用いる．これにより，モビリティをアンカーとするマルチモーダル表現と，地理座標をアンカーとする LocationBind の表現を比較する．

欠損モダリティ補完とクロスモーダル検索の評価においては，MoRA [34] および ImageBind [8] をベースラインとして採用する．これらは複数のモダリティを共通の潜在空間に埋め込む手法である．MoRA [34] は，グラフベースのモビリティ埋め込みをアンカー表現として使用し，グラフに紐づく衛星画像を含む複数の都市モダリティを整合する．ImageBind [8] は，画像表現をアンカーとし，画像とペアのデータに対する対照学習を用いて，画像埋め込み空間に多様なモダリティを整合する．欠損モダリティ補完実験においては，これらの手法は Decoder を持たないため，まず欠損対象モダリティの他モダリティデータを抽出する．その後，それらの共通潜在空間上の平均表現をクエリに欠損対象モダリティデータを検索し，そのインスタンスを推定値として補完する．

4.4 結果

まず土地指標予測の結果を表 3 に示す．LocationBind は両都市・全タスクで最良のスコアを達成した．特に土地利用分類（LUC）における改善が顕著であり，大阪では Macro-F1 で次点の MoRA および SatCLIP に対し，いずれも大幅な改善を実現している．人口密度（PR）および建物高度（BHR）の回帰タスクでも一貫した改善が確認でき，両都市ともベースライン中最良の SatCLIP に対し改善する傾向を示した．ベースライン手法としては，公開データの重みをそのまま用いた GeoCLIP が一貫して最低性能となり，各都市のデータによる事前学習の重要性を示している．SatCLIP は PR・BHR で相対的に強い一方，LUC では中程度にとどまっており，各土地指標に対して各モダリティが寄与する指標は一様ではないことが分かる．これに対し LocationBind は，CaLLiPer と同じ 128 次元という比較的コンパクトな表現次元でありながら全タスク・全ベースラインを上回っている．この結果から，LocationBind の地理座標エンコーダが各モダリティの表現を有効に統合し，土地利用・人口・建物高度といった異質な指標を統一的に説明する地理座標表現の獲得に寄与していると考えられる．

次に欠損モダリティ補完の結果を表 4 に示す．Location-

表 6: アブレーション実験結果．各値は大阪と名古屋の平均値．土地指標予測の LUC は Macro-F1，PR および BHR は RMSE．欠損モダリティ補完は *region* 設定でのコサイン類似度．クロスモーダル検索は R@10.

	土地指標予測			欠損モダリティ補完			クロスモーダル検索		
Model	LUC (↑)	PR (↓)	BHR (↓)	POI	Visit	Sat	POI–Visit	POI–Sat	Visit–Sat
w/o Decoder	0.462	0.748	0.734	0.417	0.688	0.745	0.065	**0.678**	**0.817**
w/o CL	**0.478**	**0.746**	**0.726**	0.492	0.562	0.749	0.013	0.032	0.020
Full Model	0.469	0.747	0.730	**0.590**	**0.768**	**0.810**	**0.066**	0.646	0.808

Bind は両都市・全モダリティ・両分割設定で最良のコサイン類似度を達成した．特に欠損対象が POI である場合の改善が顕著であり，Visit および Sat の補完でも一貫した改善が見られる．これは，他モダリティの観測のみからでも，整列された共通潜在空間を介して欠損モダリティの埋め込みを高精度で復元できることを示している．*random* と *region* を比較すると，全手法で *region* 設定の方が低いスコアおよび大きな標準偏差を示しており，各モダリティについて未観測領域の補完が，単純なランダム欠損より困難であることが確認できる．一方で，LocationBind はこの設定でもベースライン手法を上回る精度を保っている．

最後にクロスモーダル検索の結果を表 5 に示す．LocationBind は両都市・全モダリティペア・両指標で最良の Recall を達成した．特に POI–Sat および Visit–Sat における改善が顕著である．一方，POI–Visit 検索のスコアは全手法を通じて低水準に留まる．LocationBind でも大阪の R@1 は 0.007，名古屋でも 0.011 であり，ベースラインからの相対改善は明確であるものの絶対値は低い．これは，両モダリティの空間粒度が細かく（POI 単位・50m メッシュ単位）正例ペア数が膨大であり，1,000 件サンプリングの中に意味的に類似した負例が多数含まれることが考えられる．一方 Sat を含むペアでは，衛星画像が 250m メッシュという粗い粒度であるためペア数が相対的に少なく，識別が容易になる傾向があると推測される．

4.5 アブレーション実験

LocationBind の学習目的は，対照学習損失と再構成損失の 2 つから構成される．ここでは，各構成要素が下流タスクにどのように寄与しているかを明らかにするため，2 種類のアブレーションモデルを構築する．w/o Decoder は，ModalityDecoder および再構成損失を取り除き，対照学習損失のみで学習するモデルである．このモデルは，地理座標と POI テキスト・訪問パターン・衛星画像の各モダリティを共通潜在空間上で整列する設計のみを保持する．w/o Decoder は Decoder を持たないため，欠損補完時には共通潜在空間上で観測済みモダリティ表現を平均し，その近傍の欠損対象モダリティの学習サンプルとして補完値を得る．w/o CL は，対照学習損失を取り除き，地理座標から各モダリティ埋め込みを ModalityDecoder により直接復元する再構成損失のみで学習するモデルである．このモデルでは，モダリティ間の明示的な整合は行われず，LocationEncoder と各 ModalityDecoder は独立にモダリティ埋め込みを再構成するよう学習される．

表 6 に各モデルの結果を示す．最善のものを太字，次善のものを下線で強調している．土地指標予測では，w/o CL が LUC・PR・BHR の全指標で最良となり，Full Model および w/o Decoder を上回った．これは，モダリティ表現との整合を一切行わず，地理座標から各地点の局所的な属性を直接予測するよう学習されるため，地理座標表現が土地利用や人口・建物高度のような局所的指標の予測に有利な特徴を獲得していることを示唆する．一方，クロスモーダル検索では w/o Decoder が強く，POI–Sat および Visit–Sat で Full Model をわずかに上回って最良となった．対照的に w/o CL は，全ペアで R@10 が 0.05 未満に留まっており，モダリティ間整列がほとんど学習されていないことが分かる．この結果から，対照学習損失はモダリティ間表現の整合性を担保するうえで不可欠であり，再構成損失だけでは異なるモダリティ間の対応関係を共通潜在空間上で構築できないことが確認される．欠損モダリティ補完では，w/o Decoder と w/o CL の両者を上回り，Full Model が 3 モダリティすべてで最良となった．これは，欠損モダリティ補完が，他の観測済みモダリティを共通潜在空間へ整合的に写像する対照学習損失の役割と，整合された表現から欠損モダリティの埋め込みを生成する再構成損失の役割の両方を必要とするためと考えられる．

総合すると，Full Model は個別タスクで単独最良とならない指標もあるが，3 タスクすべてにわたり最良または次善の性能を一貫して維持している．これは，対照学習損失と再構成損失が相補的な役割を担っていることを示しており，両損失と構成要素の併用によって，土地指標予測・欠損モダリティ補完・クロスモーダル検索といった異質な下流タスクに対して汎用的に有効な共通潜在空間が獲得されることを示している．

4.6 都市間ゼロショット転移実験

これまでの実験では，各都市内のデータで学習したモデ

表 7: 都市間ゼロショット転移実験の結果．各値は両転移方向での平均．欠損モダリティ補完は *region* 設定でのコサイン類似度．クロスモーダル検索は R@10.

	欠損モダリティ補完			クロスモーダル検索		
Model	POI	Visit	Sat	POI–Visit	POI–Sat	Visit–Sat
ImageBind	0.404	0.672	0.731	0.018	0.029	**0.065**
MoRA	0.432	0.681	0.735	0.029	0.024	0.041
LocationBind	**0.583**	**0.769**	**0.803**	**0.039**	**0.048**	0.056

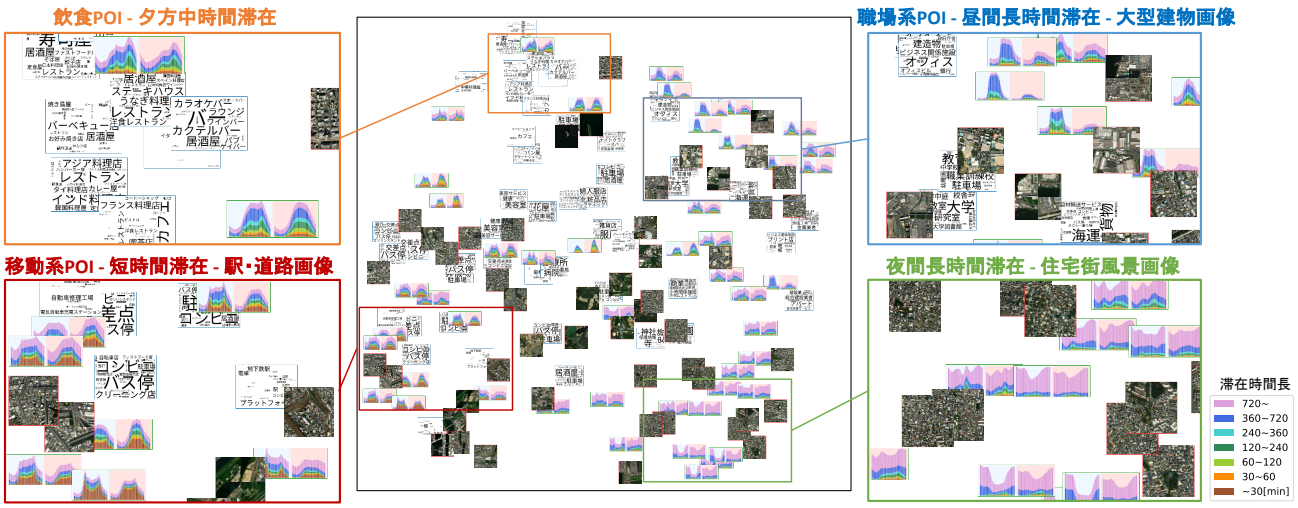


図 3: マルチモーダルクラスタリング結果の可視化 (UMAP).

ルを同一都市内で評価してきた．本節では，LocationBind により学習された共通潜在空間が単一都市に過学習しておらず，都市間で汎化する表現を獲得しているかを検証するため，都市間ゼロショット転移実験を行う．具体的には，片方の都市の全データで事前学習したモデルを，もう片方の都市の全データに対してそのまま適用し，欠損モダリティ補完とクロスモーダル検索の 2 タスクで評価する．土地指標予測については，都市固有の LocationEncoder を用いているため除外している．結果は両転移方向（大阪 → 名古屋および名古屋 → 大阪）の結果を平均して報告する．

表 7 に転移実験の結果を示す．欠損モダリティ補完では，LocationBind は 3 モダリティすべてで最良であり，都市内設定での結果とほぼ同等の性能を維持している．ベースラインの転移性能との差も明確で，都市内設定と同様の優位性が転移設定でも保たれている．これは，地理座標を媒介としてモダリティ間を整列する設計が，都市固有の細部に依らずモダリティ間の対応関係を捉えていることを示しており，LocationBind で学習される共通潜在空間の都市横断的な汎化を支持する．

一方，クロスモーダル検索では転移による性能劣化が顕著であり，特に Sat を含むペアで劣化幅が大きい．この対比は，補完と検索という 2 タスクが共通潜在空間に対して要求する性質の違いを反映していると考えられる．補完タスクは，観測済みモダリティから欠損モダリティへの局所的な対応関係を捉えていれば成立する一方で，検索タスクは 1,000 件の負例の中から正例を識別する必要があり，共通潜在空間全体の細粒度な構造に依存する可能性が高い．そのため，都市ごとの埋め込み分布の差異が検索性能に直接影響を与える．特に衛星画像は，両都市間で建物様式・密度・色調などの視覚的特徴が大きく異なるため，事前学習済みエンコーダの出力分布が都市ごとに偏り，ModalityProjector が学習時に整形した整列構造が転移先都市では崩れやすい．また，衛星画像はデータ量が他のモダリティと比較して少量であるため，学習時の都市の傾向に過学習している可能性もある．この劣化により ImageBind は Visit–Sat のペアで LocationBind を上回っている．ImageBind は画像表現をアンカーとして他モダリティを整列する構造を持ち，画像エンコーダ自体が RemoteCLIP の大規模事前学習により都市非依存な表現を獲得しているのに対し，LocationBind は共通潜在空間上の全モダリティ表現が学習対象として自由に最適化されるため，学習時の都市分布に対する過学習の影響を受けやすい構造を持つと考えられる．

5. アプリケーション

前節までの定量評価により，LocationBind が学習する共通潜在空間が下流タスクで有効に機能することを確認した．本節では，この共通潜在空間が単なる予測性能向上に

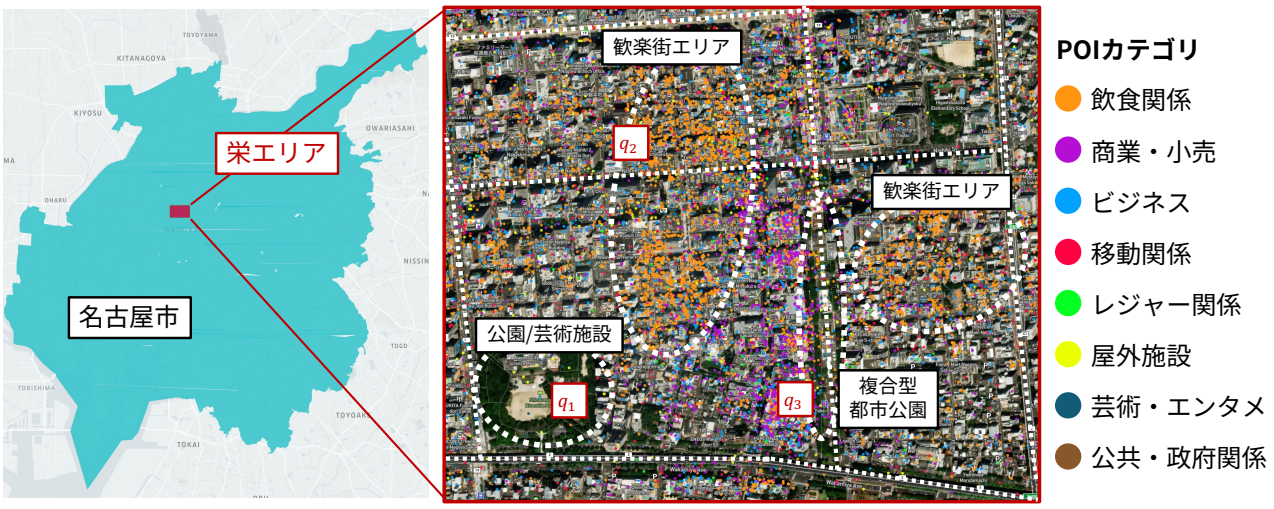


図 4: 名古屋の都心部である栄エリアの空間構成．衛星画像データに重ねて POI エントリをプロット．

とどまらず，都市の多面的な理解を支援する都市データ分析基盤としても活用可能であることを，3 つのアプリケーションを通じて示す．

5.1 マルチモーダルクラスタリング

まず，LocationBind の共通潜在空間を用いて，都市内でモダリティを跨いでどのような特徴がどのような特徴と関連し，地理的に共起するかというクロスモーダルな傾向を分析する．具体的には，名古屋市の全モダリティデータを共通潜在空間にエンコードし，モダリティごとに 50 クラスタずつ計 150 クラスタに分割する．得られた各クラスタについて，モダリティの傾向を表す可視化を生成する．POI クラスタはクラスタ内の POI テキストから構築したワードクラウド，訪問パターンクラスタは平日と休日それぞれについて時間帯ごとの訪問数を示すグラフ，衛星画像クラスタはクラスタ内から最大 4 枚の代表画像により可視化する．次に，全 150 クラスタの重心ベクトルを UMAP により 2 次元に圧縮し，対応する位置に各モダリティの可視化結果を配置した結果を図 3 に示す．

図 3 より，モダリティ横断的な共起パターンを多数観察できる．例えば，飲食系 POI と夕方にピークを持つ滞在パターンの共起，職場系 POI と昼間の長時間滞在パターンおよび大型建物の衛星画像の共起，駅・バス停・交差点などの移動関連 POI と短時間滞在パターンおよび道路を含む衛星画像の共起，夜間長時間滞在パターンと住宅街を示す衛星画像の共起などが確認できる．これらは，LocationBind の共通潜在空間がモダリティ間の意味的対応関係を保持しており，単一モダリティでは捉えられない都市のクロスモーダルな共起関係の抽出可能性を示している．

5.2 場所のマルチモーダル解釈

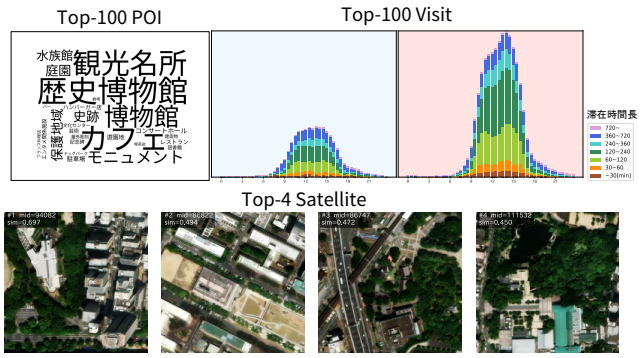
次に，LocationBind を用いた局所的な地域分析の例を示す．ここでは名古屋市の都心部である栄周辺エリアを対象とする．対象エリアの空間構成を図 4 に示す．本エリアは，中央部および東部に歓楽街・飲食店エリア，南西部に名古屋市科学館や美術館を含む公園・芸術施設エリア，中心部を南北に貫く複合型公園エリアを有し，さらに図中の直線で示す主要道路が走行する，多様な都市機能が集積する地域である．

このような複合的なエリアにおいて，LocationBind によって各地点の多面的な特徴を可視化できる．具体的には，図 4 中の 3 地点 q_1 (35.1639, 136.9010)，q_2 (35.1708, 136.9028)，q_3 (35.1652, 136.9079) の地理座標を LocationEncoder により共通潜在空間へエンコードし，名古屋市の各モダリティデータから近傍を検索する．得られた各モダリティの近傍データの可視化によって，対象地点の特徴を多面的に解釈する．ここでは，各地点の座標表現から共通潜在空間で近傍の上位 100 件の POI，100 グリッドの訪問パターン，上位 4 件の衛星画像を可視化する．

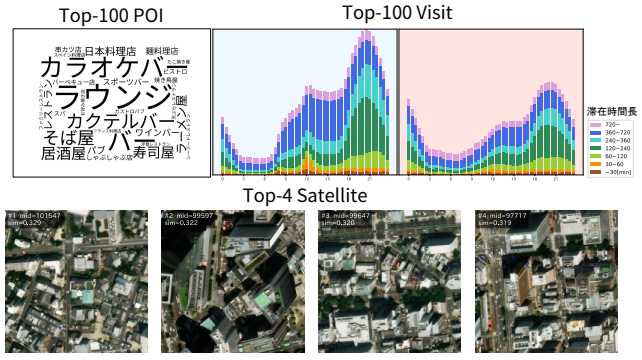
検索結果を図 5a〜図 5c に示す．q_1 は，休日昼間の滞在が多く，博物館や観光名所などの POI 特徴を持ち，やや緑地的な景観を有する観光・文化拠点的な地点として特徴づけられる．q_2 は，20 時ごろにピークを迎える滞在パターンを持ち，バー・ラウンジ・居酒屋などの POI 特徴と都心のビル群の景観を有する夜間繁華街的な傾向を示す．q_3 は，休日昼間の滞在が多く，デパートや小売店などの POI 特徴と大型モール型建物の景観を有する商業集積的な地点として解釈できる．いずれの地点においても，3 モダリティが一貫した解釈を与えており，LocationBind が地理座標からマルチモーダルな場所特徴を統合的に出力可能であり，各場所の多面的な解釈を可能にしている．

5.3 マルチスケール場所検索

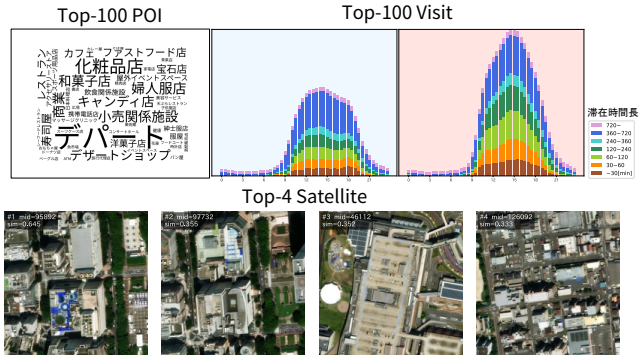
前節では地理座標から各モダリティへの検索を示した



(a) q_1 地点



(b) q_2 地点

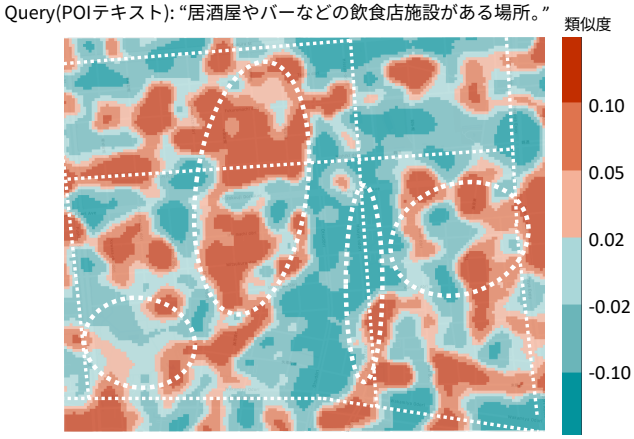


(c) q_3 地点

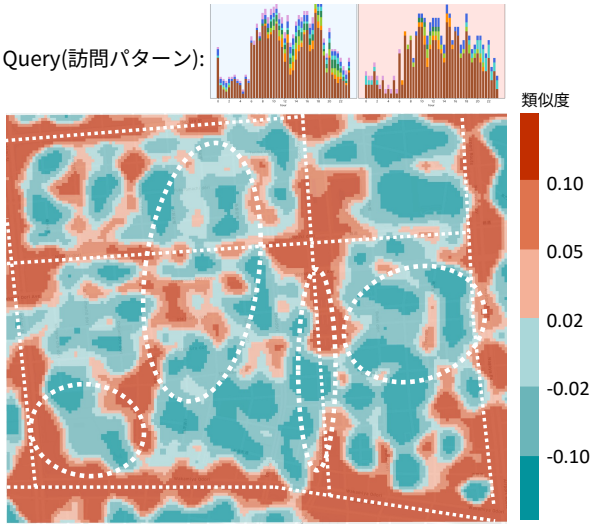
図 5: 図 4 の各クエリ地点から各モダリティを検索した結果.

が，逆方向の検索，すなわちモダリティデータをクエリとした場所の検索も可能である．本節では，図 4 の栄エリアを 10m グリッドメッシュに分割し，各メッシュの中心点を LocationEncoder でエンコードする．その後，各モダリティのデータをクエリとして，全グリッドの地理座標表現との類似度を計算し，地図上に可視化する．

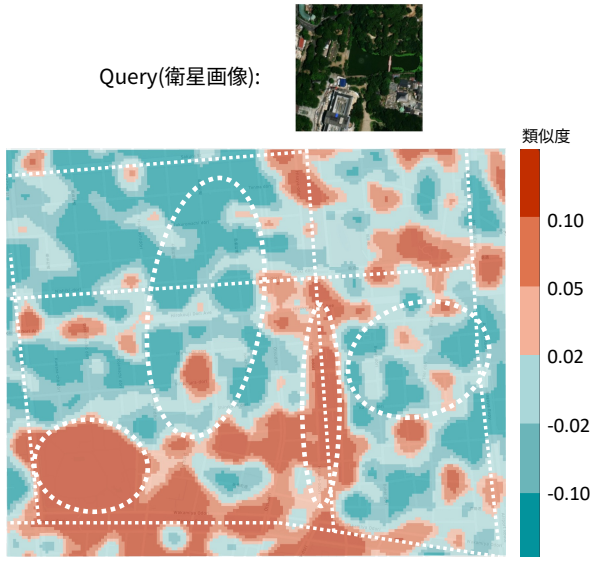
図 6a〜図 6c に，各モダリティをクエリとした場所検索の結果を示す．POI テキスト検索（図 6a）では，「居酒屋やバーなどの飲食関係施設がある場所.」というクエリに対し，対象エリア内の歓楽街エリアとの類似度が高くなっていることが確認できる．訪問パターン検索（図 6b）では，短時間滞在が支配的な訪問パターンをクエリとしており，大通り沿いや交差点周辺といった一時的な滞在が発生しや



(a) POI テキストクエリ



(b) 訪問パターンクエリ



(c) 衛星画像クエリ

図 6: 各モダリティをクエリとした場所検索の結果.

すい地点との類似度が高くなっている．衛星画像検索（図 6c）では，緑地公園の衛星画像をクエリとしており，栄エ

リア内の都市公園部分との類似度が高くなっている．

これらの結果は，各モダリティを起点とした自由な場所の検索可能性を示している．このような分析は，学習データと同じ粒度であれば従来の離散的なメッシュ埋め込みでも可能であった．一方 LocationBind では，マルチスケールな LocationEncoder により，学習時の 50m や 250m といったグリッド粒度に縛られることなく，10m グリッドレベルで場所のセマンティクスを可視化できている．これは，分析の目的に応じた任意の空間粒度での場所検索を可能にし，都市の細粒度な分析を支援する基盤となる．ただし，どの程度局所的あるいは大域的な空間粒度で分析するかは，LocationEncoder の $r_{\min}$ および $r_{\max}$ によって設定する必要がある．

以上の 3 つのアプリケーションを通じ，LocationBind が提供する共通潜在空間は，モダリティ横断的なパターン抽出，特定地点の多面的解釈，マルチスケールな場所検索という多様な分析を，単一の表現基盤の上で統合的に実現可能であることを示した．これは，LocationBind が下流タスクの精度向上にとどまらず，都市の多面的な理解を支援する汎用的な分析基盤として機能する可能性を示唆している．

6. まとめと今後の展望

本論文では，「同一地点で観測される異種モダリティは潜在的な相互関係を共有する」という仮説に基づき，地理座標を媒介として POI テキスト，訪問パターン，衛星画像を共通潜在空間に整合するマルチモーダル都市データ表現学習手法，LocationBind を提案した．LocationBind は，明示的なモダリティ間ペアを必要とせず，各モダリティと地理座標の対応関係のみを用いて学習できる点に特徴がある．対照学習によって地理座標表現と各モダリティ表現を共通潜在空間上に整合し，再構成学習により共通潜在表現から各モダリティ埋め込みを復元可能とすることで，土地指標予測，欠損モダリティ補完，クロスモーダル検索を統一的に扱える枠組みを実現した．大阪および名古屋を対象とした実験では，LocationBind により得られた地理座標表現が，土地利用分類，人口密度予測，建物高度予測の全タスクにおいて既存手法を上回る性能を示した．また，欠損モダリティ補完では，観測済みモダリティによる未観測モダリティの埋め込みの高精度な推定性能を確認した．さらに，クロスモーダル検索およびアプリケーション例を通じて，LocationBind の共通潜在空間の，異種都市モダリティ間の共起パターン抽出，特定地点の多面的解釈，任意粒度での場所検索への活用可能性を示した．

一方で，本研究にはいくつかの課題が残されている．第一に，都市間ゼロショット転移実験では，欠損モダリティ補完において高い転移性能を維持した一方，クロスモーダル検索，特に衛星画像を含む検索において性能劣化が確認された．今後は，都市間の分布差に頑健な整合学習やドメイン適応を導入し，より汎用的な都市マルチモーダル表現の獲得を目指す．第二に，本研究ではグリッドなどの面的な空間単位を，中心座標のみからエンコードしている．これはグリッドの面積が大きい場合など，中心座標とグリッドとの空間的な乖離が大きい場合に課題となるため，連続的な地理座標と離散的な空間単位を統合的に扱える LocationEncoder が必要である．また，ストリートビュー画像，口コミ文書，交通ネットワーク，環境センサデータなど多様な都市モダリティへの拡張を進め，都市計画支援，地域類似検索，災害リスク分析などの実応用への展開を検討する．

謝辞　本研究の一部は JST CREST（JPMJCR22M4），JST RISTEX（JPMJRS23K），JST ACT-X (JPMJAX24M8) に支援いただいています．また，データ提供にご協力いただきました株式会社ブログウォッチャーに感謝いたします．

参考文献

[1] Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D. and Ermon, S.: Geography-Aware Self-Supervised Learning, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10161–10170 (online), DOI: 10.1109/ICCV48922.2021.01002 (2021).

[2] Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H. and Liu, B.: Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 201, pp. 193–208 (online), DOI: https://doi.org/10.1016/j.isprsjprs.2023.05.006 (2023).

[3] Bettencourt, L. M. A.: The origins of scaling in cities, Science, Vol. 340, No. 6139, pp. 1438–1441 (online), DOI: 10.1126/science.1235823 (2013).

[4] Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B. and Ermon, S.: SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, Advances in Neural Information Processing Systems (Oh, A. H., Agarwal, A., Belgrave, D. and Cho, K., eds.), (online), available from ⟨https://openreview.net/forum?id=WBhqzpF6KYH⟩ (2022).

[5] Feng, S., Cong, G., An, B. and Chee, Y. M.: POI2Vec: Geographical Latent Representation for Predicting Future Visitors, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, No. 1 (online), DOI: 10.1609/aaai.v31i1.10500 (2017).

[6] Foursquare: Foursquare OS Places (2025). Accessed: 2025-09-01.

[7] Fu, Y., Wang, P., Du, J., Wu, L. and Li, X.: Efficient Region Embedding with Multi-View Spatial Networks: A Perspective of Locality-Constrained Spatial Autocorrelations, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, pp. 906–913 (online), DOI: 10.1609/aaai.v33i01.3301906 (2019).

[8] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A. and Misra, I.: ImageBind: One Em-

bedding Space To Bind Them All, CVPR (2023).

[9] Huang, T., Wang, Z., Sheng, H., Ng, A. Y. and Rajagopal, R.: Learning Neighborhood Representation from Multi-Modal Multi-Graph: Image, Text, Mobility Graph and Beyond, ArXiv, Vol. abs/2105.02489 (online), available from ⟨https://api.semanticscholar.org/CorpusID:233864977⟩ (2021).

[10] Huang, W., Cui, L., Chen, M., Zhang, D. and Yao, Y.: Estimating urban functional distributions with semantics preserved POI embedding, International Journal of Geographical Information Science, Vol. 36, No. 10, pp. 1905–1930 (online), DOI: 10.1080/13658816.2022.2040510 (2022).

[11] Kandt, J. and Batty, M.: Smart cities, big data and urban policy: Towards urban analytics for the long run, Cities, Vol. 109, p. 102992 (online), DOI: 10.1016/j.cities.2020.102992 (2021).

[12] Klemmer, K., Rolf, E., Robinson, C., Mackey, L. and Rußwurm, M.: SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, No. 4, pp. 4347–4355 (online), DOI: 10.1609/aaai.v39i4.32457 (2025).

[13] Li, Z., Huang, W., Zhao, K., Yang, M., Gong, Y. and Chen, M.: Urban Region Embedding via Multi-View Contrastive Prediction, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 8, pp. 8724–8732 (online), DOI: 10.1609/aaai.v38i8.28718 (2024).

[14] Liang, X., Zhao, T. and Biljecki, F.: Revealing spatio-temporal evolution of urban visual environments with street view imagery, Landscape and Urban Planning, Vol. 237, p. 104802 (online), DOI: https://doi.org/10.1016/j.landurbplan.2023.104802 (2023).

[15] Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L. and Zhou, J.: RemoteCLIP: A Vision Language Foundation Model for Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing, Vol. 62, pp. 1–16 (online), DOI: 10.1109/TGRS.2024.3390838 (2024).

[16] Liu, Y., Zhang, X., Ding, J., Xi, Y. and Li, Y.: Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction, Proceedings of the ACM Web Conference 2023, WWW '23, New York, NY, USA, Association for Computing Machinery, p. 4150–4160 (online), DOI: 10.1145/3543507.3583876 (2023).

[17] Mac Aodha, O., Cole, E. and Perona, P.: Presence-Only Geographical Priors for Fine-Grained Image Classification, ICCV (2019).

[18] Mañas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D. and Rodríguez, P.: Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9394–9403 (online), DOI: 10.1109/ICCV48922.2021.00928 (2021).

[19] Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L. and Lao, N.: Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells, The Eighth International Conference on Learning Representations, openreview (2020).

[20] Mai, G., Xuan, Y., Zuo, W., He, Y., Song, J., Ermon, S., Janowicz, K. and Lao, N.: Sphere2Vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 202, pp. 439–462 (online), DOI: https://doi.org/10.1016/j.isprsjprs.2023.06.016 (2023).

[21] Mapbox, Inc.: Mapbox Satellite Imagery, `https://www.mapbox.com/imagery` (2025). Accessed: 2025-11-20.

[22] 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.: Learning Transferable Visual Models From Natural Language Supervision, ICML (2021).

[23] Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M. and Darrell, T.: Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning, 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4065–4076 (online), DOI: 10.1109/ICCV51070.2023.00378 (2023).

[24] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks, arXiv [cs.CL] (2019).

[25] Shimizu, T., Yabe, T. and Tsubouchi, K.: Enabling Finer Grained Place Embeddings using Spatial Hierarchy from Human Mobility Trajectories, Proceedings of the 28th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '20, New York, NY, USA, Association for Computing Machinery, p. 187–190 (online), DOI: 10.1145/3397536.3422229 (2020).

[26] Shoji, K., Terashima, H., Tamura, N., Katayama, S., Urano, K., Yonezawa, T. and Kawaguchi, N.: Life Pattern Based Human Attribute Estimation Using Only Raw GPS Data, IEEE Access, (online), DOI: 10.1109/access.2025.3591811 (2025).

[27] Stalder, S., Volpi, M., Büttner, N., Law, S., Harttgen, K. and Suel, E.: Self-supervised learning unveils urban change from street-level images, Computers, Environment and Urban Systems, Vol. 112, p. 102156 (online), DOI: https://doi.org/10.1016/j.compenvurbsys.2024.102156 (2024).

[28] Tamura, N., Shoji, K., Katayama, S., Urano, K., Yonezawa, T. and Kawaguchi, N.: OpenUAS: Embeddings of Cities in Japan with Anchor Data for Cross-city Analysis of Area Usage Patterns (2024).

[29] van den Oord, A., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, NeurIPS, (online), available from ⟨https://arxiv.org/abs/1807.03748⟩ (2018).

[30] Vivanco, V., Nayak, G. K. and Shah, M.: GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization, Advances in Neural Information Processing Systems (2023).

[31] Wang, H. and Li, Z.: Region Representation Learning via Mobility Flow, Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, New York, NY, USA, Association for Computing Machinery, p. 237–246 (online), DOI: 10.1145/3132847.3133006 (2017).

[32] Wang, X., Cheng, T., Law, S., Zeng, Z., Yin, L. and Liu, J.: Multi-modal contrastive learning of urban space representations from POI data, Computers, Environment and Urban Systems, Vol. 120, p. 102299 (online), DOI: https://doi.org/10.1016/j.compenvurbsys.2025.102299 (2025).

[33] Wang, Z., Li, H. and Rajagopal, R.: Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 01, pp. 1013–1020 (2020).

[34] Wen, Y., Cai, J., Ma, Q., Li, L., Chen, X., Webster, C. and Zhou, Y.: MoRA: Mobility as the Backbone for Geospatial Representation Learning at Scale (2026).

[35] Yan, B., Janowicz, K., Mai, G. and Gao, S.: From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts, Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '17, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3139958.3140054 (2017).

[36] Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R. and Liang, Y.: UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web, Proceedings of the ACM Web Conference 2024, WWW '24, New York, NY, USA, Association for Computing Machinery, p. 4006–4017 (online), DOI: 10.1145/3589334.3645378 (2024).

[37] Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J. and Mai, K.: Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model, International Journal of Geographical Information Science, Vol. 31, No. 4, pp. 825–848 (online), DOI: 10.1080/13658816.2016.1244608 (2017).

[38] Yao, Z., Fu, Y., Liu, B., Hu, W. and Xiong, H.: Representing Urban Functions through Zone Embedding with Human Mobility Patterns, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, pp. 3919–3925 (online), DOI: 10.24963/ijcai.2018/545 (2018).

[39] Zheng, Y., Capra, L., Wolfson, O. and Yang, H.: Urban Computing: Concepts, Methodologies, and Applications, ACM Transactions on Intelligent Systems and Technology, Vol. 5, No. 3, pp. 38:1–38:55 (online), DOI: 10.1145/2629592 (2014).

[40] 愛知県建設局都市基盤部都市計画課：都市計画基礎調査 (2022). Accessed: 2026-05-08.

[41] 株式会社ブログウォッチャー：ブログウォッチャー位置情報データ, https://www.blogwatcher.co.jp/ (2026). Accessed: 2026-05-08.

[42] 国土交通省：Project PLATEAU オープンデータ, https://www.mlit.go.jp/plateau/open-data/ (2025). Accessed: 2026-01-17.

[43] 総務省統計局：令和 2 年国勢調査 小地域集計（基本単位区別 男女別人口及び世帯数）, https://www.e-stat.go.jp/stat-search/files?toukei=00200521&tstat=000001136464&tclass1=000001136472 (2022). Accessed: 2026-05-08.

[44] 大阪市計画調整局計画部都市計画課：まちづくりに関する基礎データ（メッシュデータ等）, https://www.city.osaka.lg.jp/toshikeikaku/page/0000452834.html (2025). Accessed: 2026-05-08.

付　　録

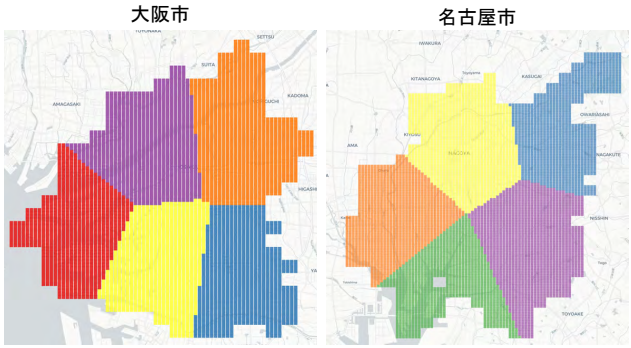


図 A·1: Region 分割のマッピング

表 A·1: 土地指標予測実験のデータセット．

都市	タスク	取得元	年	空間単位	インスタンス数
大阪	**LUC**	[44]	2022	50m-grid	81,998
	PR	[43]	2020	50m-grid	81,998
	BHR	[42]	2024	50m-grid	81,998
名古屋	**LUC**	[40]	2021	50m-grid	125,741
	PR	[43]	2020	50m-grid	125,741
	BHR	[42]	2022	50m-grid	125,741

表 A·2: LUC に用いた土地利用 11 クラス．

土地利用クラス	メッシュ数 (名古屋)	メッシュ数 (大阪)
ビジネス	3,353	8,743
工業	10,948	7,213
教育施設	5,213	3,714
文化・宗教	3,054	1,041
医療	1,693	839
商業	5,268	4,199
住宅	44,694	26,225
公園	8,460	5,336
道路・鉄道	21,977	16,476
田畑・樹林・河川	14,335	377
その他	6,746	7,835
合計	125,741	81,998

A.1　土地指標予測実験の詳細

本節では，土地指標予測実験の詳細を述べる．本実験では，各モデルから得られる 50m グリッド単位の表現を固定特徴量として用い，土地利用分類（LUC），人口密度回帰（PR），建物高度回帰（BHR）の 3 タスクを評価する．各タスクにおけるデータセットの概要を表 A·1 に示す．LUC では，各自治体の土地利用調査データを 50m グリッドへ対応付け，11 クラスの土地利用ラベルへ再マッピングしたものを用いる．土地利用ラベルの内訳を表 A·2 に示す．PR では国勢調査に基づく 50m グリッドごとの人口密度を，BHR では PLATEAU の建物データから算出した 50m グリッドごとの平均建物高度を用いる．PR および BHR

の値は都市ごとに標準化し，標準化後の値に対する RMSE を報告する．

入力特徴量については，LocationBind，SatCLIP，CaLLiPer では，各 50m グリッドの代表座標を LocationEncoder に入力して地理座標表現を得る．GeoCLIP については，公開済みの LocationEncoder に緯度経度を入力して特徴量を得る．Area2Vec，RemoteCLIP，MoRA は事前計算済みのメッシュ表現を用いる．Area2Vec は 50m グリッドの訪問パターン埋め込み，RemoteCLIP は 250m 衛星画像タイルの画像埋め込みを対応する 50m グリッドへ割り当てた表現，MoRA はモビリティグラフをアンカーとして学習された地域表現を用いる．

下流予測器には，すべてのモデルで同一構造の MLP を用いる．MLP の層構成は $d_{\text{in}} \to 128 \to d_{\text{out}}$ であり，各隠れ層では線形変換の後に Batch Normalization，ReLU，Dropout を適用する．Dropout 率は 0.3 とする．LUC では $d_{\text{out}} = 11$ とし，Cross Entropy Loss で学習して Macro-F1 を報告する．PR および BHR では $d_{\text{out}} = 1$ とし，MSE Loss で学習して RMSE を報告する．最適化には AdamW を用い，学習率は 1.0×10^{-4}，重み減衰は 1.0×10^{-1}，バッチサイズは 128，エポック数は 100 とする．

データ分割には 5-fold cross validation を用いる．各モデルには同一の fold 分割を適用し，5-fold の平均値と標準偏差を報告する．

A.2 欠損モダリティ補完実験の詳細

欠損補完では，Random と Region の 2 種類の 5-fold 分割を用いる．Random は全手法で同一の乱数シードを用い，各モダリティサンプルをランダムに 5 分割する設定であり，Region は図 A·1 のように，地理座標から K-Means で都市空間を 5 つの空間領域にクラスタリングし，各 fold で 1 領域内の欠損対象モダリティを評価対象とする設定である. クラスタリングは衛星画像に合わせて 250m グリッドメッシュ単位で行い，それらのグリッド内に位置する POI や 50m グリッド (Visit) についても同様の分割を行う．

欠損対象モダリティ m と fold k の各組について，fold k に属する m のサンプルを事前学習データから除外してモデルを学習する．一方，欠損対象ではない他モダリティは全領域のデータを観測済みモダリティとして利用する．したがって，評価対象となる欠損モダリティの正解埋め込みは，対応する fold の学習時には使用されない．これらの fold 分割および学習データ除外の設定は，ベースライン手法の ImageBind および MoRA についても同一とする．

fold$k$ における評価集合と最近傍検索型補完の候補集合をそれぞれ次式で定義する．ここで $\text{split}(i,m) \in \{1,\dots,5\}$ は，モダリティ m における 5-fold 分割でサンプル i が属する fold を表す関数である．

$$
\begin{aligned}
I^{(m)} &= \{1,2,\dots,N_m\},\\
T_{m,k} &= \{i \in I^{(m)} \mid \text{split}(i,m) = k\},\\
C_{m,k} &= I^{(m)} \setminus T_{m,k}.
\end{aligned}
\tag{A.1}
$$

ここで $I^{(m)}$ はモダリティ m の全サンプルの添字集合である．評価集合 $C_{m,k}$ は学習に用いる残りのサンプルの添字集合を表す．最近傍検索型補完では，検索候補を $C_{m,k}$ に対応するサンプルに限定し，評価 fold の正解サンプルは候補から除外する．

評価対象サンプル $i \in T_{m,k}$ の座標を $\ell_i^{(m)}$ とする．各観測済みモダリティ $m' \in \mathcal{M} \setminus \{m\}$ について，$\ell_i^{(m)}$ に最も近い m' のサンプルを取得する．

$$
j^*(i,m') = \arg\min_{j \in I^{(m')}} ||\ell_i^{(m)} - \ell_j^{(m')}||_2. \tag{A.2}
$$

本実験では最大探索半径は設定せず，各観測済みモダリティから最近傍 1 件を用いる．取得した元埋め込みを各手法の Projector で共通潜在空間へ写像し，L2 正規化後に平均する．

$$
\bar{s}_i = \text{normalize}\left(\frac{1}{|\mathcal{M}|-1}\sum_{m' \in \mathcal{M}\setminus\{m\}} g^{(m')}\left(z_{j^*(i,m')}^{(m')}\right)\right)
$$

この $\bar{s}_i$ を，欠損対象モダリティ m を推定するためのクエリ表現として用いる．公平な比較のため，LocationEncoder から得られる地理座標表現は補完入力に用いない．

LocationBind は ModalityDecoder を持つため，観測済みモダリティのみから構成した $\bar{s}_i$ を欠損対象モダリティの Decoder$h^{(m)}$ に入力する．

$$
\hat{z}_i^{(m)} = h^{(m)}(\bar{s}_i)
$$

一方，ImageBind および MoRA は本実験設定では Decoder を持たないため，欠損対象モダリティ m の学習候補集合 $C_{m,k}$ を共通潜在空間へ写像し，$\bar{s}_i$ と最も類似する候補の元埋め込みを補完結果とする．

$$
j^\dagger = \arg\max_{j \in C_{m,k}} \bar{s}_i^\top \text{normalize}\left(g^{(m)}(z_j^{(m)})\right), \qquad \hat{z}_i^{(m)} = z_{j^\dagger}^{(m)}
$$

なお，ImageBind の衛星画像アンカーのように Projector を持たないモダリティでは，L2 正規化した元埋め込みを共通潜在空間上の表現として扱う．

補完結果は，推定埋め込み $\hat{z}_i^{(m)}$ と正解埋め込み $z_i^{(m)}$ のコサイン類似度で評価する．

$$
\text{Cosine}(m,k) = \frac{1}{|T_{m,k}|}\sum_{i \in T_{m,k}} \frac{\left(\hat{z}_i^{(m)}\right)^\top z_i^{(m)}}{\left\|\hat{z}_i^{(m)}\right\|_2 \left\|z_i^{(m)}\right\|_2}. \tag{A.3}
$$

実験では，各欠損対象モダリティについて 5-fold の平均値と標準偏差を報告している．

A.3　クロスモーダル検索実験の詳細

本節では，クロスモーダル検索実験の詳細を述べる．欠損モダリティ補完実験とは異なり，クロスモーダル検索では fold ごとの除外は行わず，各都市の全データで学習した単一モデルを評価に用いる．評価対象は POI–Visit，POI–Sat，Visit–Sat の 3 ペアである．

正例ペア集合を $\mathcal{P}_{a,b} = \{(\tilde{z}_i^{(a)}, \tilde{z}_i^{(b)})\}_{i=1}^{N_{a,b}}$ とする．各ペアについて双方向の検索を行う．元埋め込みは各手法の Projector により共通潜在空間へ写像し，L2 正規化する．

$$u_i^{(a)} = \mathrm{normalize}\left(g^{(a)}(\tilde{z}_i^{(a)})\right),$$
$$v_i^{(b)} = \mathrm{normalize}\left(g^{(b)}(\tilde{z}_i^{(b)})\right)$$

ImageBind の衛星画像アンカーなど Projector を持たないモダリティでは，L2 正規化した元埋め込みを共通潜在空間上の表現として用いる．

各クエリに対して正例 1 件とランダム負例 1000 件からなる 1001 件の候補集合を作成する．負例は同一ペア集合の候補側から正例を除いてサンプリングし，すべての手法で同一の正例ペア構築方法と評価条件を用いる．各モダリティペアについては $a \to b$ と $b \to a$ の双方向検索を行い，ペア単位に集約した Recall@1 および Recall@10 を報告する．

LocationBind, ImageBind, MoRA は同一の検索ペア, 候補集合，負例サンプリング条件で比較する．LocationBind では，各モダリティの ModalityProjector 出力を用い，検索時には Decoder を使用しない．ImageBind では衛星画像をアンカーモダリティとして扱い，衛星画像は L2 正規化したアンカー表現，POI と Visit は Projector 出力を用いる．MoRA ではモビリティグラフをアンカーとして学習された共有空間を用い，検索時には POI，Visit，Sat の各モダリティ埋め込みを Projector により写像した表現のみを用いる．