景観画像と地理的特性を考慮した都市における雰囲気の定量化

久保田 祐輝[1,a)]　安納 爽響[1,b)]　坪内 孝太[,2,c)]　下坂 正倫[1,d)]

概要：近年，景観画像に基づき都市に対する人々の知覚を定量化する試みが注目されている．都市に対して人々が抱く知覚の正確な定量化が実現されれば，都市開発の方針策定や来訪者の誘致をはじめとした様々な課題への応用が見込まれる．しかしながら，これまでの取り組みは特定の景観画像に対する知覚の精緻な定量化を実現しているにとどまり，同一の解析地点においても，景観の撮影方向によって得られる結果が大きく異なるという課題が存在した．本研究ではこの課題に取り組み，都市各地点における雰囲気の良し悪しを正確に定量化するための枠組み，Geospatial CityScouter を提案する．提案する枠組みでは，都市各地点から複数の方向より景観画像を撮影し，各景観の重要度を画像特徴と地理的特性に基づき考慮したうえで，雰囲気の良し悪しの推論を実現する．多数の人々の意見を反映した，都市景観から感じとる知覚のデータを用いた実験により，提案する枠組みが都市各地点における雰囲気の高精度な定量化を実現することを示す．

1. 序論

近年，都市各地から撮影された景観画像に基づき，都市の特性を定量化する試みが盛んに取り組まれている [5]．割れ窓理論 [17] で提唱されているように，都市の景観は周辺の犯罪状況や近隣住民の健康状態とも関係することが指摘されており，これまでも画像処理技術を活用し，景観画像と近隣の犯罪状況の解析を試みた研究 [4], [11] や，景観画像と近隣住民の活動量の多さとの関連性の解析を試みた研究 [15] などが存在する．中でも，景観画像に基づき都市に対して人々が感じる美しさや安全性といった知覚を定量化する試みが盛んであり，これまで数多くの研究で取り組まれてきた [2], [10], [16]．都市景観から感じとる印象や知覚の定量化が実現されれば，人々に好まれる都市の開発や，外部からの来訪者の誘致を目的とした計画の策定などへの応用が期待される．

しかしながら，これまでの取り組みは景観の画像に対する知覚の定量化を実現した一方で，実際に人々が都市現地を来訪した際に感じとる知覚の定量化としては不十分である．これは，先行研究における取り組みは特定の景観画像に対する知覚の精緻な定量化を目的としており，同一の解析地点においても，景観撮影時の視点によって得られる結果が大きく変化してしまうためである．図 1 に Google Street View API[*1] より同一の緯度経度から撮影時の視点のみを変更して取得した 4 枚の景観画像を示す．なお，図に示されている角度は道路と水平でかつ北側を向いた方向を 0 度として時計回りに変化させた値である．

図に示されている 4 枚の景観画像からは，同一の緯度経度の撮影地点においても，撮影時の視点によって景観画像の内容が大きく異なることが分かる．すなわち，実際の都市に対する知覚の定量化を実現するには，特定の景観画像に対するスコアの精密な推論のみでは不十分であり，景観の視点をモデリング時に考慮することが必要不可欠になると考えられる．また，人間の都市に対する知覚は，都市の道路状況や近隣の建造物などの地理的特性による影響を受けることが指摘されている [19]．すなわち，都市に対する人間の知覚の定量化には，景観画像の解析に限らず景観の撮影地点自体の性質の考慮が重要となる．

本研究ではこれらの観点を考慮し，都市現地における人間の知覚を，景観の画像と都市の地理的特性の双方を活用することで定量化する枠組み，Geospatial CityScouter を提案する．より詳細には，まず都市各地において複数の視点から景観画像を取得し，各視点の重要度を景観の画像特

[1] 東京工業大学 情報理工学院 情報工学系
Department of Com-puter Science, School of Engineering, Tokyo Institute of Technology
[2] Yahoo! JAPAN 研究所
Yahoo! JAPAN Research
[a)] kubota@miubiq.cs.titech.ac.jp
[b)] anno@miubiq.cs.titech.ac.jp
[c)] ktsubouc@yahoo-corp.jp
[d)] simosaka@miubiq.cs.titech.ac.jp

[*1] https://developers.google.com/maps/documentation/streetview

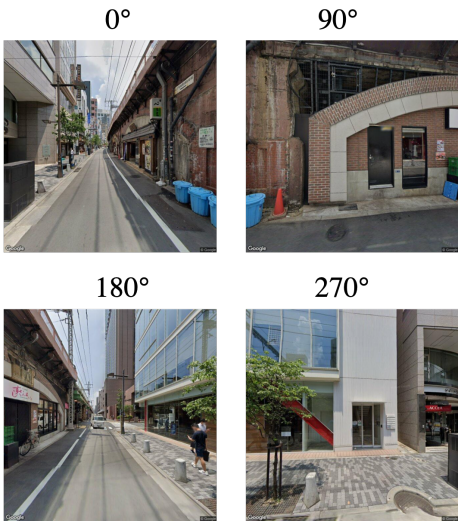


図 1: 同一の緯度経度の撮影地点において撮影時の視点を変更した際に取得される景観画像の例.

性と地理的特性に基づき考慮した上で，都市現地から人々が感じとる知覚の定量化を実現する．提案手法では，景観画像の特性と地理的性質の交互作用をモデリングすることで，画像処理技術のみでは困難な，都市の雰囲気の高精度な推論を実現する．本研究では，人間の知覚の中でも「雰囲気の良し悪し」に着目し，人々が感じとる雰囲気を都市各地点において定量化することを実現する．

本研究の貢献は以下のようにまとめられる．

- 本研究では，都市各地から実際に人々が来訪した際に感じとる雰囲気の良し悪しを，景観の画像に基づいた特徴量及び撮影地点の地理的特性を考慮して定量化する枠組みを提案する．
- 多数の人々より収集した，景観に対する知覚のデータに基づく実験により，提案する枠組みが，人間が都市現地から感じとる雰囲気の高精度な定量化を実現することを示す．

関連研究

Visual Urban Perception

都市の景観画像の解析に着目した研究として，画像処理技術に基づき，景観画像に対する人々の知覚の定量化を目的とする Visual Uraban Perception と呼ばれる研究分野が存在する [2], [10], [16]．当該分野において，Dubey ら [3] はベンチマークとなるデータセットである Place Pulse 2.0 を提案した．Place Pulse2.0 では Google Street View より取得された景観画像に対し，クラウドソーシングを活用することで合計 6 つの知覚評価軸からなるスコアのアノテーションを行っている．各景観画像は 2 枚で一つのペアとして扱われ，クラウドソーシングの参加者は二者択一の形式で，より特定の知覚項目を強く認知する景観画像を選択する．この Place Pulse 2.0 のデータセットに対し，画像に付与されたスコアの精密な定量化が盛んに取り組まれている．Porzi ら [16] は画像が二者択一の形式でアノテーションされていることを考慮し，ランク学習の枠組みでスコアを推論するモデルを提唱した．また，データセットにおいて 6 つの知覚の評価軸が存在することを考慮し，複数の評価軸間における交互作用をマルチタスク学習によって学習することを試みた研究 [5], [13] も存在する．

しかしながら，これらの研究は画像に対して付与されたスコアを精度良く推論することを実現する一方で，人々が実際の都市現地から感じとる知覚の定量化としては不十分である．これは，先行研究においては景観画像を取得する際の視点が考慮されておらず，同一の撮影点においても，得られる結果が景観撮影時の視点によって依存するためである．

Image Aesthetic Assessment

画像データに対する人間の知覚の定量化を試みた研究分野としては Image Aesthetic Assessment が存在する．当研究分野では景観画像に限らず多種多様な画像に対して，人間が感じとる美しさの定量化を試みている．Image Aesthetic Assessment においては AVA [14] データセットがベンチマークとして使用されることが多く，当データセットでは画像から感じとれる美しさの度合いが，クラウドソーシングを介して 10 段階で評価されている．これまで，AVA データセットに対し，画像が美しいか否かを二値分類で推論した研究 [12]，美しさのスコアの回帰推定を試みた研究 [6]，スコアの分布自体の推論を試みた研究 [7] など，数多くの研究でスコアの推論が取り組まれてきた．

これらの取り組みにより画像の美しさを高精度に推論することが可能となった一方で，当該研究分野は画像に付与されたスコアの高精度な推論を目的としており，都市などの現実の空間に対する知覚の評価は想定されていない．当該研究分野は，画像に対する知覚推論の精緻化を目的としているという観点から，前述した Visual Urban Perception と類似した問題に取り組んでいると解釈できる．

2. 問題設定と先行研究の限界

2.1 問題設定

本研究では，まず都市の景観画像に対して人々が感じとる知覚の定量化を実現する．次に，都市各地から複数の視点より景観画像を取得し，各視点の景観に対する知覚の数値，及び撮影地点の地理的特性に基づき，都市現地から感じとれる知覚の推論を実現する．各段階における問題設定は以下のように定式化される．

2.1.1 景観画像に対する知覚の定量化

本研究では，景観画像の総数を N によって表し，画像全体の集合を $\boldsymbol{I} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ によって定義する．定量化の対象である人々の知覚は複数存在するとし，それら知覚の評価軸の集合を M，各知覚評価軸を $m \in M$ によって表す．また，各景観画像 $\boldsymbol{x}_i$ における知覚評価軸 m の定量化された

スコアを $v_i^{(m)}$ によって定義する．すなわち，景観画像に対する知覚の定量化を行う問題は，各画像 $\boldsymbol{x}_i$ を入力として受け取り，対応する $v_i^{(m)}$ を推論する関数 $f(\cdot)$ の学習として定式化される．

$$\underset{\boldsymbol{\theta}_f}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{m \in M} \mathcal{L}_f \left(v_i^{(m)},\ f(\boldsymbol{x}_i; \boldsymbol{\theta}_f) \right). \tag{1}$$

ここで，$\mathcal{L}_f$，$\boldsymbol{\theta}_f$ は各々関数 $f(\cdot)$ の学習に用いられる損失関数とパラメータを意味する．

2.1.2 都市に対する知覚の定量化

次に，都市現地における知覚の定量化の問題設計に関して述べる．本研究では知覚の解析対象地の総数を L によって表し，各解析対象地を $l = 1, \ldots, L$ によって定義する．解析対象地点 l において取得された景観画像の集合を $\boldsymbol{I}_l = \{\boldsymbol{x}_{l,d}\}_{d=1}^{D}$ と定める．ここで，D は各解析地点において取得される景観画像の総数である．また，解析対象地 l における各景観画像 $\boldsymbol{x}_{l,d}$ について，前述した関数 $f(\cdot)$ を適用することで，知覚評価軸 m に関して得られた推論値を $\hat{v}_{l,d}^{(m)}$ とする．さらに，同一地点における全ての景観画像に対する推論値 $\{\hat{v}_{l,d}^{(m)}\}_{d=1}^{D}$ をまとめたベクトルを $\hat{\boldsymbol{v}}_l^{(m)} \in \mathbb{R}^D$ によって定義する．地点 l における知覚評価軸 m の定量化されたスコアを $y_l^{(m)}$ とすれば，都市に対する知覚の定量化を行う問題は，$\hat{\boldsymbol{v}}_l^{(m)}$ を入力として受け取り，$y_l^{(m)}$ を推論する関数 $g(\cdot)$ の学習として以下のように定式化される．

$$\underset{\boldsymbol{\theta}_g}{\operatorname{argmin}} \sum_{l=1}^{L} \sum_{m \in M} \mathcal{L}_g \left(y_l^{(m)},\ g\left(\hat{\boldsymbol{v}}_l^{(m)}; \boldsymbol{\theta}_g \right) \right). \tag{2}$$

ここで，$\mathcal{L}_g$，$\boldsymbol{\theta}_g$ は各々関数 $g(\cdot)$ の学習に用いられる損失関数とパラメータを意味する．

2.2 先行研究における知覚の定量化手法の限界

先行研究においては主に画像に対する知覚の定量化，すなわち関数 $f(\cdot)$ の学習の精緻化に対する試みは多数存在する一方で [5], [7], [13]，関数 $g(\cdot)$ に関しては単純な手法に留まるのが現状である．例えば，Dubey ら [3] はランク学習の枠組みにより関数 $f(\cdot)$ を学習した後に，都市各地点 l から取得された単一の画像に対する $f(\cdot)$ の推論値によって都市現地における知覚を評価している．解析対象地 l における知覚評価軸 m の推論値を $\hat{y}_l^{(m)}$ によって表せば，この手続きは以下のように定式化される．

$$\hat{y}_l^{(m)} = x \in_R \{\hat{v}_{l,d}^{(m)}\}_{d=1}^{D}, \tag{3}$$

ここで，$\in_R$ は集合内から要素を無作為に抽出する演算を意味する．

また，市区町村や道路ネットワークなど，特定の解析対象領域内において取得された全ての画像に関する推論値の算術平均を導出した取り組みも存在し [15], [18]，この処理は以下のように定式化される．

$$\hat{y}_l^{(m)} = \frac{1}{D} \sum_{d=1}^{D} \hat{v}_{l,d}^{(m)}. \tag{4}$$

これらの手法は，人間が都市現地から感じとる知覚が，ある特定の視点の景観によって表される，もしくは解析対象地内の全ての景観が等しく重要であると仮定していると解釈することができ，現実の環境のモデリングとしては不適切な仮定をおいていることが分かる．

3. 都市の雰囲気の定量化に向けた提案手法: Geospatial CityScouter

3.1 提案する枠組みの概要

本研究で提案する枠組みである Geospatial CityScouter の概要図を図 2 に示す．Geospatial CityScouter は，景観の画像に対する雰囲気の定量化モデルと，都市現地における雰囲気の定量化を実現する 2 つのモデルを含んでいる．提案する枠組みでは，まず特定の解析対象地点 l において複数の視点より景観画像を取得する．続いて，取得された各景観画像に対し，関数 $f(\cdot)$ に相当する，画像の雰囲気を定量化するモデルを適用することで，各視点の景観における雰囲気の推論値 $\hat{v}_{l,d}$ を得る．次に，l における全ての視点の景観画像に対する雰囲気の推論値 $\{\hat{v}_{l,d}\}_{d=1}^{D}$ と，解析対象地点 l の地理的特性に基づき，各景観の視点の重要度を考慮した上で，都市現地における雰囲気の定量化を実現する．この手続きは 2.1 節で述べた関数 $g(\cdot)$ に相当する．以下に各手続きの詳細に関して述べる．

3.2 景観画像に対する雰囲気の定量化手法：DIPNet

本研究では，景観画像に対する雰囲気の定量化を実現するモデルを **D**istribution-based **I**mage **P**erception Learning Network（DIPNet）と称する．DIPNet は，事前学習された畳み込みニューラルネットワーク（CNN）を活用することで，入力画像から画像特徴量を抽出する．次に，抽出された画像特徴量を各知覚評価軸 $m \in M$ ごとに設けられた 2 層の全結合層へ入力することで，各知覚評価軸に対応するスコアの推論を行う．

また，DIPNet では景観画像に対する雰囲気の良し悪しを単一の数値として評価するのではなく，複数段階の評価に基づいたスコアの分布により学習する．スコアを分布として学習，推論することにより，単一の数値による学習と比較して知覚の個人差による影響を考慮したモデリングを実現する．

3.2.1 雰囲気のスコア分布によるアノテーション

本節では，DIPNet が学習時に活用する，雰囲気スコアの分布によるアノテーションの実現方法に関して述べる．本研究では，各景観画像 $\boldsymbol{x}_i \in \boldsymbol{I}$ に関する知覚が K 段階の評価に基づいているとし，知覚のスコアは各評価段階に対するユーザーからの投票により算出されるとする．画像 $\boldsymbol{x}_i$

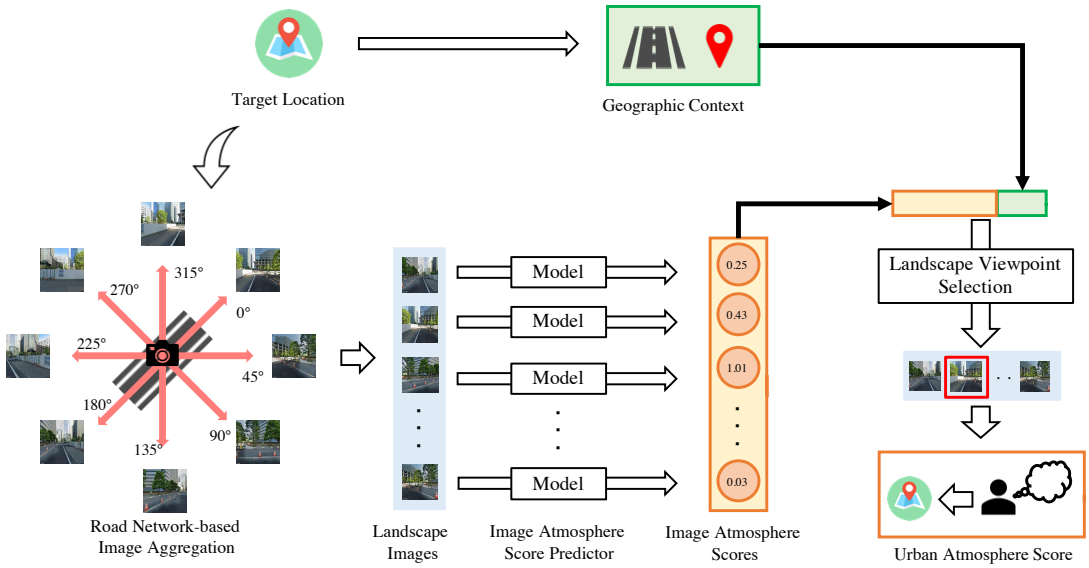


図 2: 本研究で提案する Geospatial CityScouter の概要図.

の知覚評価軸 m に対し，k 番目の評価段階に投票を行ったユーザーの総数を $u_i^{(m)}(k)$ とすれば，ユーザからの得票率に基づいたスコア分布ベクトル $\boldsymbol{v}_i^{(m)} \in \mathbb{R}^K$ は以下のように算出される．

$$\boldsymbol{v}_i^{(m)} = \left\{ p_i^{(m)}(1), \ldots, p_i^{(m)}(k), \ldots, p_i^{(m)}(K) \right\} \in \mathbb{R}^K,$$
$$p_i^{(m)}(k) = \frac{u_i^{(m)}(k)}{\sum_{j=1}^{K} u_i^{(m)}(j)}. \tag{5}$$

ここで，$p_i^{(m)}(k)$ は正規化された k 番目の評価段階に対するユーザーからの投票率である．

3.3 都市に対する雰囲気の定量化手法: GeoBiR

本節では，Geospatial CityScouter の中で，都市現地から人間が感じとる雰囲気を定量化する関数 $g(\cdot)$ に対応する機能の詳細に関して述べる．本研究では，この都市現地に対する雰囲気の定量化を実現するためのモデルを Geospatial Context-based Viewpoint-aware Bilinear Regression（GeoBiR）と称する．GeoBiR は，3.2 節で述べた DIPNet によって得られた景観画像に対する雰囲気の推論値と，景観画像の撮影地点における地理的特性のコンテクスト特徴量を入力として受け取る．続いて，これらの特徴量の交互作用に基づき，景観画像のうちどの景観をより注視するかを考慮した上で都市から感じとれる雰囲気の値を推論する．

3.3.1 雰囲気スコアと地理的コンテクストの特徴量設計

本節では，GeoBiR が入力として受け取る，雰囲気のスコアと地理的コンテクストの特徴量の設計方法に関して述べる．まず，解析対象地点 l より取得された景観画像の集合 $\boldsymbol{I}_l = \{\boldsymbol{x}_{l,d}\}_{d=1}^{D}$ に対し， 3.2 節で述べた景観画像に対する雰囲気の定量化モデル DIPNet を適用することで，各画像に対応した推論値の集合 $\{\hat{\boldsymbol{v}}_{l,d}^{(m)}\}_{d=1}^{D}$ を得る．なお，DIPNet はスコアを分布ベクトル $\hat{\boldsymbol{v}}_{l,d}^{(m)} \in \mathbb{R}^K$ として出力するため，スコア分布を分布の各評価段階に数値を配分した荷重平均によってスカラー値へ変換を行う．具体的には，各評価段階 k と対応した数値の集合 $S = \{s(1), \ldots, s(K)\}$ を導入し，k 番目の評価段階に対する推論値を $\hat{p}_{l,d}^{(m)}(k)$ と表せば，スカラーの推論値 $\hat{v}_{l,d}^{(m)}$ への変換は以下のように定式化される．

$$\hat{v}_{l,d}^{(m)} = \sum_{k=1}^{K} s(k)\, \hat{p}_{l,d}^{(m)}(k). \tag{6}$$

スカラーの推論値の集合 $\{\hat{v}_{l,d}^{(m)}\}_{d=1}^{D}$ をまとめたベクトルを $\hat{\boldsymbol{v}}_l^{(m)} \in \mathbb{R}^D$ によって表せば，GeoBiR の入力となる各景観画像に対応した雰囲気の推論値のベクトルが得られる．

続いて，GeoBiR では解析対象地点における地理的なコンテクスト情報を道路幅に基づき考慮する．本研究では，道路の幅の広さに基づき事前に C 個のカテゴリを定義し，解析対象地点 l における道路幅がどのカテゴリに属するかをベクトル $\boldsymbol{g}_l \in \mathbb{R}^C$ によって表現する．コンテクストベクトル $\boldsymbol{g}_l$ は l における道路幅が属するカテゴリに対応する要素のみが 1 を取り，他の値を 0 としたダミー変数に基づくベクトルである．

3.3.2 雰囲気スコアと地理的コンテクストの交互作用に基づく都市の雰囲気の推論

GeoBiR では，景観画像の雰囲気スコアベクトル $\hat{\boldsymbol{v}}_l^{(m)}$ と，地理的コンテクストベクトル $\boldsymbol{g}_l$ 間における双線形回帰を行うことで，両方の特徴量間における交互作用を考慮した推論を実現する．具体的には，本手法による推論は以下のように定式化される．

$$\hat{y}_l^{(m)} = \boldsymbol{g}_l^{\top} \boldsymbol{W} \hat{\boldsymbol{v}}_l^{(m)} + b, \tag{7}$$

ここで，$\boldsymbol{W} \in \mathbb{R}^{C \times D}$， b は各々モデルによって学習される

パラメータ行列，バイアス項を表す．

4. 性能評価実験

4.1 景観画像に対する雰囲気の定量化手法の性能評価実験

4.1.1 実験に使用するデータセット

本研究では，まず手動で景観画像を撮影することで，合計 2,305 枚の景観画像のデータセットを作成している．次に，Yahoo!クラウドソーシング[*2]のプラットフォームを活用することで各景観画像に対する雰囲気のスコアの付与を実現する．Yahoo!クラウドソーシングは，アンケート調査などの簡易な作業を不特定多数の Yahoo!JAPAN ユーザーに対して依頼できるプラットフォームである．当サービスにおいては，各ユーザーにマスク処理が施された ID を割り当てることで，回答者個人に関する情報は一切開示されないように対処している．

実際に景観画像に対するアノテーションを行う際には，不特定多数のユーザーから各雰囲気の評価軸 m に関して，景観画像からどの程度その要素を知覚するかを $K = 5$ 段階評価に基づいた回答を収集する．例えば，雰囲気の良し悪しに関する調査であった場合，ユーザーは調査対象の景観画像と共に「とても雰囲気が悪い」,「雰囲気が悪い」,「どちらでもない」,「雰囲気が良い」,「とても雰囲気が良い」という 5 つの選択肢が提示される．ユーザーはこれらの選択肢の中から最も直感に合うと感じた選択肢を単一選択形式により回答を行い，各評価段階に対するユーザーの回答率に基づきアノテーションを行う．本研究では，知覚の評価軸の集合 M を，景観から感じとる「雰囲気」・「秩序」・「こだわり」・「高価さ」の 4 つの評価軸として定めた．最終的に，データセットに含まれる合計 2,305 枚の各景観画像に対して，知覚評価軸 m に対応した $K = 5$ 段階評価に基づいたスコア分布 $\boldsymbol{v}_i^{(m)} \in \mathbb{R}^5$ が紐付けられる．

4.1.2 比較手法

比較手法として，知覚スコアの分布 $\boldsymbol{v}_i^{(m)}$ ではなく，平均化されたスカラー値 $v_i^{(m)}$ による学習を行う回帰推定モデルを用いる．本手法を，スコア分布を考慮しないモデルとして Regression-based Image Perception Learning Network（RIPNet）と称する．RIPNet では，式 6 で示した変換と同様の処理に基づき，スコア分布をスカラー値 $v_i^{(m)}$ へ変換した値を正解ラベルとして与え，モデルの学習を行う．

4.1.3 モデルの学習・評価設定

本研究では，DIPNet，RIPNet において画像の特徴量抽出器として，ImageNet [1] によって事前学習された MobileNet [8] を活用する．両モデルを学習する際は，バッチ数は 32 に設定し，学習率 0.0001 のもと 50 エポック学習を繰り返した．また，DIPNet の学習時に用いる損失関数として KL-ダイバージェンスを，RIPNet には平均二乗誤差を使用し，最適化アルゴリズムには Adam [9] を用いる．画像データセット全体を 5 分割交差検証法により，学習用データと検証用データへ分割した上でモデルの学習を行い，モデルの推論性能は平均絶対誤差（ Mean Absolute Error, 以降 MAE と略記）と相関係数により評価する．なお，スコアを分布として学習・推論する DIPNet については，式 6 に基づき真値と推論値を各々スカラー値に変換した上で性能評価を行う．

4.1.4 景観画像に対する雰囲気の定量化手法の実験結果

本節では景観画像に対する雰囲気の定量化手法の性能評価実験の結果に関して述べる．表 1 に DIPNet と比較手法である RIPNet の各知覚評価軸に対する推論精度の比較を示す．表で示した実験結果から，全ての知覚評価軸に関して，DIPNet が MAE，相関係数共に RIPNet を上回っていることが確認できる．従って，DIPNet においてスコアの分布ベクトルに基づいた学習を行うことの有用性，すなわち知覚の個人差を考慮したモデリングを行うことの有用性が示された．

4.2 都市に対する雰囲気の定量化手法の性能評価実験

4.2.1 実験に使用するデータセット

都市の雰囲気スコアデータ

本研究では，都市現地から感じとる雰囲気の良し悪しに関する正解データを，360 度画像ビューワーを活用したアンケート調査により収集する．具体的には，まず特定の各解析対象地 l において Google Street View により 360 度のパノラマ画像の取得を行う．続いて，得られたパノラマ画像を 360 度画像ビューワーを介して被験者に Web ブラウザ上で閲覧してもらい，感じとった雰囲気の良し悪しをアンケートにより回答してもらう．360 度画像を活用することで，単一の景観画像に基づく評価と比較して，より臨場感が得られる形式で雰囲気を評価することを実現し，実際に都市現地を来訪した場合と近い状況を再現する．

また，雰囲気のスコアの評価方法は 4.1.1 節で述べた景観画像に対する雰囲気のアノテーションと同一の基準を採用する．すなわち，知覚の評価軸は合計 4 つからなり，各景観を各々の評価軸の観点から $K = 5$ 段階評価に基づき回答してもらう．また，各々の評価段階に対して同様に $S = \{-2, -1, 0, 1, 2\}$ のスコアを割り振ったうえで，各評価段階の得票率に基づいた加重平均により雰囲気の真値 $y_l^{(m)}$ を算出する．本研究では，雰囲気の解析対象地点として，日本全国から著名な都市を中心に $L = 110$ ヶ所を選出した．

道路情報データ

3.3 節で述べたように，本研究では解析対象地点における地理的特性として道路幅の情報を活用する．解析対象地点の道路情報を取得するために，本研究では国土地理院より

[*2] https://crowdsourcing.yahoo.co.jp/

表 1: 景観画像の各知覚評価軸に関する推論精度の比較.

手法	評価指標	雰囲気	秩序	高価さ	こだわり
RIPNet	MAE ↓	0.315 ± 0.02	0.328 ± 0.02	0.358 ± 0.04	0.324 ± 0.03
	相関係数↑	0.727 ± 0.04	0.703 ± 0.06	0.722 ± 0.04	0.682 ± 0.04
DIPNet	MAE ↓	$\mathbf{0.255 \pm 0.01}$	$\mathbf{0.289 \pm 0.03}$	$\mathbf{0.302 \pm 0.02}$	$\mathbf{0.319 \pm 0.01}$
	相関係数↑	$\mathbf{0.755 \pm 0.02}$	$\mathbf{0.728 \pm 0.05}$	$\mathbf{0.799 \pm 0.05}$	$\mathbf{0.694 \pm 0.03}$

提供されている道路中心線データ[*3]を活用する．本データでは日本全国の道路中心線の位置情報，及び道路の種別や管理団体などの補足情報が提供されており，その中の道路幅情報を参照する．データにおいて，道路幅情報は「3m 未満」,「3m 以上 5.5m 未満」,「5.5m 以上 13m 未満」,「13m 以上 19.5m 未満」,「19.5m 以上」の $C = 5$ つのカテゴリのうちどれに該当するかが情報として与えられている．

4.2.2 学習用データの作成

4.2.1 節で述べた正解データの作成方法に基づき，各解析地点 l に関して雰囲気の真値である $y_l^{(m)}$ が得られる．本研究では，解析対象地点 l において撮影された景観画像の集合 $\boldsymbol{I}_l = \{\boldsymbol{x}_{l,d}\}_{d=1}^{D}$ を，l の位置情報と付近の道路中心線データに基づき取得する．具体的には，各解析地点が位置する道路との方向を考慮した上で，景観撮影時の水平角度を 45 度ずつ変更させ，合計 8 方向から $D = 8$ 枚の景観画像を Google Street View API より取得する．景観画像を取得する際は，まず解析対象地が位置する道路の方向と水平かつ北画を向いた方向を 0 度して定め，時計回りに 45 度ずつ角度を変化させる．これらの景観画像の集合 $\boldsymbol{I}_l$ の各要素に対して 3.2 節で述べた DIPNet を適用することで，解析対象地点 l において推論された雰囲気スコアのベクトル $\hat{\boldsymbol{v}}_l^{(m)} = [\hat{v}_{l,1}^{(m)}, \hat{v}_{l,2}^{(m)}, \ldots, \hat{v}_{l,D}^{(m)}]^\top \in \mathbb{R}^8$ を得る．

4.2.3 比較手法

都市各地点におけるの雰囲気のスコア $y_l^{(m)}$ を推論するための比較手法として以下のものを用いる．

Random

本手法では解析対象地点内の画像集合 $\boldsymbol{I}_l$ に対する推論値のうち，無作為に抽出した値を対象地点 l の推論値 $\hat{y}_l^{(m)}$ とする手法である．本手法は Dubey らの研究 [3] において，各地点 l から単一の景観画像を取得し，スコアの推論を行っている手続きに倣ったものであり，式 3 により定式化される．

算術平均（Average All）

本手法は解析対象地点において得られた景観画像の推論値，すなわち $\hat{\boldsymbol{v}}_l^{(m)}$ 全体の算術平均を推論値とする手法であり，この手続きは式 4 により定式化される．本手法は先行研究 [15], [18] において，行政区画など特定の領域内の全ての景観の平均を算出した手続きを参照している．

道路と垂直方向の考慮（Perpendicular）

本手法は各対象地点において，道路と垂直方向に所得された景観画像に対するスコアの算術平均を推論値とする．道路と垂直方向に景観画像を取得するという手続きは，先行研究 [4] において建築物を中心に写すことを目的に提唱されたものである．本研究においては，道路と垂直方向，すなわち 90 度と 270 度の方向から撮影された景観画像に対するスコアの算術平均を取ることでこの手法を実現する．

線形回帰（Linear）

本手法では各景観画像に対する定量化された雰囲気スコア $\hat{\boldsymbol{v}}_l^{(m)}$ のみを入力として受け取り，それらの線形回帰による推論を実現する．すなわち，本手法では撮影手地点における地理的特性は考慮されない．

地理的特性を考慮した線形回帰（Geographical Context Linear）

本手法では景観画像の雰囲気のスコアベクトル $\hat{\boldsymbol{v}}_l^{(m)}$ に加えて，解析対象地点の地理的特性に基づくコンテクストベクトル $\boldsymbol{g}_l$ を結合したベクトルを入力として受け取り，線形回帰を行う．本手法では解析対象地の地理的特性を考慮することが可能であるが，単純な特徴ベクトルの結合に基づいているため，特徴量間の交互作用は考慮できない．

4.2.4 モデルの学習・評価設定

パラメータを学習するモデルについては，全ての手法においてバッチ数を 8 に設定し，学習率 0.001 のもと 100 エポック学習を行う．また，最適化アルゴリズムには Adam [9] を，損失関数としては平均二乗誤差を使用する．全てのモデルの学習および評価は，解析対象地点に対する 5 分割交差検証法により分割された学習用データと検証用データにより行う．また，評価指標としては MAE と相関係数を使用する．

4.2.5 都市に対する雰囲気の定量化手法の推論精度

表 2 に各手法による都市に対する雰囲気の推論精度の比較を掲載する．実験結果から，解析対象地点に依存しない固定の基準で推論を行う Random，Average All，Perpendicular と比較して，提案する GeoBiR が「こだわり」以外の全ての知覚評価軸において優れた推論性能を示しており，特に MAE において大きく性能を改善することに成功していることが分かる．すなわち，解析対象地点に依存しない固定的な基準で景観画像を取得する方針や，単純に取得された景観画像のスコアの算術平均を取る手続きでは，都市から感じとる知覚を精度良く定量化することは困難で

[*3] https://github.com/gsi-cyberjapan/vector-tile-experiment

表 2: 都市の各知覚評価軸に関する推論精度の比較.

手法	評価指標	雰囲気	秩序	高価さ	こだわり
Random [3]	MAE ↓	0.742 ± 0.04	0.727 ± 0.02	0.641 ± 0.07	0.429 ± 0.05
	相関係数↑	0.475 ± 0.17	0.636 ± 0.16	0.499 ± 0.16	0.320 ± 0.15
Average All [15], [18]	MAE ↓	0.551 ± 0.03	0.491 ± 0.03	0.504 ± 0.08	0.484 ± 0.05
	相関係数↑	0.764 ± 0.11	0.823 ± 0.08	0.738 ± 0.12	**0.594 ± 0.19**
Perpendicular [4]	MAE ↓	0.754 ± 0.04	0.743 ± 0.02	0.638 ± 0.07	0.429 ± 0.05
	相関係数↑	0.621 ± 0.17	0.703 ± 0.10	0.631 ± 0.17	0.406 ± 0.20
Linear	MAE ↓	0.301 ± 0.05	0.324 ± 0.06	0.425 ± 0.09	0.328 ± 0.05
	相関係数↑	0.756 ± 0.10	0.806 ± 0.09	0.738 ± 0.12	0.586 ± 0.18
Geographical Context Linear	MAE ↓	0.311 ± 0.05	0.322 ± 0.05	0.436 ± 0.08	**0.321 ± 0.04**
	相関係数↑	0.781 ± 0.04	0.814 ± 0.09	0.715 ± 0.13	0.565 ± 0.17
GeoBiR (Proposed)	MAE ↓	**0.294 ± 0.05**	**0.297 ± 0.05**	**0.401 ± 0.06**	0.366 ± 0.03
	相関係数↑	**0.774 ± 0.09**	**0.849 ± 0.06**	**0.751 ± 0.14**	0.474 ± 0.13

あることが確認できる.

また，雰囲気スコアのみを考慮する Linear や，地理的特性に基づく特徴ベクトルを単純に結合する Geographical Context Linear と比較しても GeoBiR が総合的に優れた性能を示すことから，雰囲気スコアと地理的特性の交互作用を考慮することの有用性が確認できる．特に，Linear と地理的特性を考慮する Geographical Context Linear を比較しても，両者の推論精度に大きな差異は見受けられないことからも，単純に地理的特性を考慮するのではなく，雰囲気のスコアとの相互関係をモデリングすることの重要性が示唆された.

4.2.6 提案する枠組みによる応用可能性の定性的評価

本節では，提案する枠組みを活用した応用可能性に関して述べる．図 3 に浅草二丁目より取得された景観画像に対し，GeoBiR により推論された雰囲気に対応するスコアを可視化した結果を示す．図において，推論されたスコアが正の値を取る地点を青色の円で，負の値を取る地点を赤色の円で示しており，円の大きさは推論されたスコアの絶対値の大きさに対応している．図における可視化結果から，都市各地点において感じとれる雰囲気が空間的にどのように分布しているのかを確認することが可能となる．例えば，地図上の東側にスコアが正に大きい領域が，西側にスコアの絶対値が小さい領域，もしくはスコアが負の領域が集中していることが確認できる．すなわち，浅草二丁目は同じ丁目内でも東部と西部によって感じられる知覚が大きく異なっており，実際に東部には浅草寺に代表される神社仏閣が，西部には商店街や住宅街が立地しており，東部と西部における土地の機能的側面の差異が反映された結果であると考えられる.

また，GeoBiR により学習された重みを解析することでモデル推論の解釈を行うことが可能である．具体的には，道路幅のコンテクストがダミー変数で表されていることから，GeoBiR による推論はパラメーター行列 $\boldsymbol{W}$ の特定の行と，入力された各視点の雰囲気ベクトル $\hat{\boldsymbol{v}}_l^{(m)}$ 間の重み

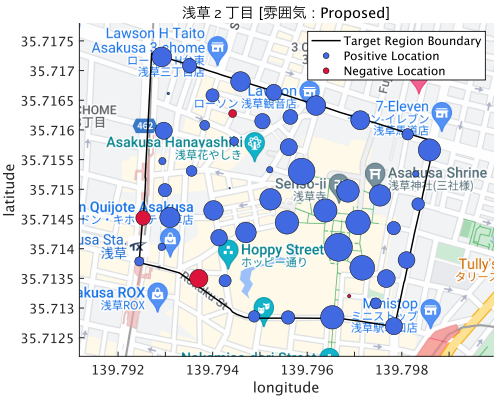
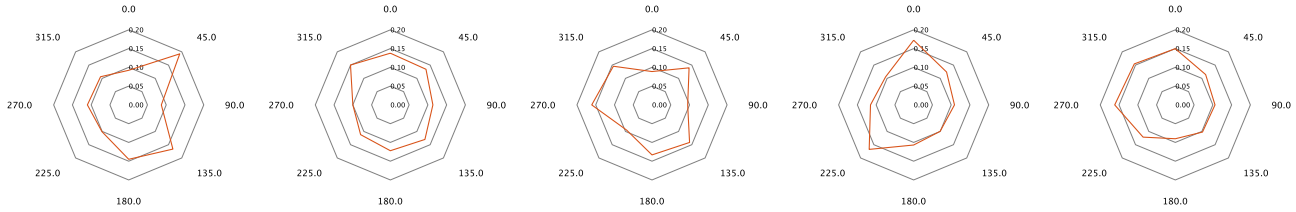


図 3: 浅草二丁目において GeoBiR により推論されたスコアを地図上へ可視化した図.

付き和により実現される．すなわち，学習されたパラメーター行列 $\boldsymbol{W}$ の各行の値を取得することで，GeoBiR が特定の道路幅カテゴリが入力された際に，8 つの景観の視点のうち，どの視点を重視する傾向にあるかを解析することが可能である．図 4 に，各道路幅カテゴリに関して学習された重みをレーダーチャートとして可視化した結果を示す．図では 0 度から 315 度の 8 方向の各視点に対する正規化された重みが示されている．可視化結果から，道路幅が異なるとモデルが重要視する視点の傾向も大きく変化していることが確認できる．全ての道路幅カテゴリにおいて共通して極端に大きい，もしくは小さい重みが与えられている視点は存在せず，道路幅のコンテクストによって GeoBiR が重要視する景観の視点を適切に変更していることが分かる．また，図 4a で示した道路幅が「3m 未満」のカテゴリにおいて学習された重みが，他の道路幅カテゴリの結果と比較すると特定の視点に偏った結果であることが確認できる．この結果は，人間が知覚を感じとる際に特定の視点から受ける影響の強さと，当該地点の地理的特性の間における関連性の存在を示唆しており，雰囲気のスコアと地理的特性間の交互作用を考慮するモデリングの重要性が確認できる.



(a) 3m 未満 **(b)** 3m 以上 5.5m 未満 **(c)** 5.5m 以上 13m 未満 **(d)** 13m 以上 19.5m 未満 **(e)** 19.5m 以上

図 4: 各道路幅カテゴリに対して学習された GeoBiR のパラメーターの可視化結果

5. 結論

本研究では，都市各地点における雰囲気の良し悪しを定量化するための枠組み，Geospatial CityScouter を提案した．提案する枠組みでは，都市各地点から複数の視点より景観画像を撮影し，各景観の重要度を画像特徴と地理的特性に基づき考慮することで，都市における雰囲気の定量化を実現した．多数の人々の意見を反映した，都市景観から感じとる知覚のデータを用いた性能評価実験により，DIPNet が景観画像に対する雰囲気の高精度な定量化を実現すること，及び GeoBiR が既存手法を上回る精度で都市の雰囲気の定量化を実現することを示した．

将来課題として，考慮する地理的特性のさらなる多様化や，各地点ごとに重視される景観の視点の解析を実現する手法の考案などが挙げられる．

参考文献

[1] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).

[2] Doersch, C., Singh, S., Gupta, A., Sivic, J. and Efros, A. A.: What Makes Paris Look like Paris?, *ACM Transactions on Graphics (SIGGRAPH)*, Vol. 31, No. 4, pp. 101:1–101:9 (2012).

[3] Dubey, A., Naik, N., Parikh, D., Raskar, R. and Hidalgo, C. A.: Deep Learning the City: Quantifying Urban Perception at a Global Scale, *Proceedings of the European Conference on Computer Vision (ECCV)* (2016).

[4] Fu, K., Chen, Z. and Lu, C.-T.: StreetNet: Preference Learning with Convolutional Neural Network on Urban Crime Perception, *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2018).

[5] Guan, W., Chen, Z., Feng, F., Liu, W. and Nie, L.: Urban Perception: Sensing Cities via a Deep Interactive Multi-Task Learning Framework, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 17, pp. 1–20 (2021).

[6] Hosu, V., Goldlucke, B. and Saupe, D.: Effective Aesthetics Prediction With Multi-Level Spatially Pooled Features, *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9367–9375 (2019).

[7] Hou, J., Yang, S. and Lin, W.: Object-Level Attention for Aesthetic Rating Distribution Prediction, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 816–824 (2020).

[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, *arXiv preprint arXiv:1704.04861* (2017).

[9] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[10] Liu, L., Silva, E. A., Wu, C. and Wang, H.: A machine learning-based method for the large-scale evaluation of the qualities of the urban environment, *Computers, Environment and Urban Systems*, Vol. 65, pp. 113–125 (2017).

[11] Liu, X., Chen, Q., Zhu, L., Xu, Y. and Lin, L.: Place-Centric Visual Urban Perception with Deep Multi-Instance Regression, *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 19–27 (2017).

[12] Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J. Z.: Rating Image Aesthetics Using Deep Learning, *IEEE Transactions on Multimedia*, Vol. 17, No. 11, pp. 2021–2034 (2015).

[13] Min, W., Mei, S., Liu, L., Wang, Y. and Jiang, S.: Multi-Task Deep Relative Attribute Learning for Visual Urban Perception, *IEEE Transactions on Image Processing*, Vol. 29, pp. 657–669 (2020).

[14] Murray, N., Marchesotti, L. and Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis, *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2408–2415 (2012).

[15] Nadai, M. D., Vieriu, R. L., Zen, G., Dragicevic, S., Naik, N., Caraviello, M., Hidalgo, C. A., Sebe, N. and Lepri, B.: Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life, *Proceedings of the 24th ACM international conference on Multimedia* (2016).

[16] Porzi, L., Rota Bulò, S., Lepri, B. and Ricci, E.: Predicting and Understanding Urban Perception with Convolutional Neural Networks, *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 139–148 (2015).

[17] Wilson, J. Q. and Kelling, G. L.: *Broken windows*, Critical issues in policing: Contemporary readings (1982).

[18] Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H. and Ratti, C.: Measuring human perceptions of a large-scale urban region using machine learning, *Landscape and Urban Planning*, Vol. 180, pp. 148–160 (2018).

[19] Zhang, L., Zhang, R. and Yin, B.: The impact of the built-up environment of streets on pedestrian activities in the historical area, *Alexandria Engineering Journal*, Vol. 60, No. 1, pp. 285–300 (2021).