LLMによるシーン中の物体の形容記述を用いた景観画像の印象予測

井手 海翔[1,a)]　安納 爽響[1,b)]　坪内 孝太[,2,c)]　下坂 正倫[1,d)]

概要：近年，都市景観画像に対して，人々が抱く印象を数値評価することが，都市開発や土地ブランディングに役立つとされ研究が進んでいる．既存研究では，景観画像のシーン全体をひとつの物体と仮定する，物体認識ベースの数値予測が中心であったが，シーン内の個々の物体を検出し，検出領域での画像特徴から数値を予測する手法へと進展している．この流れは，シーン全体をベースとする手法の場合，物体のクラスや局所的な特徴を捉えることが困難なために，印象の数値予測の性能が限定的であったことに起因する．一方，物体検出に特化した画像特徴は，景観の印象評価に直接関連性があるとは言いがたい．そこで本研究では，景観画像のシーン中の物体検出に加え，物体等シーン構成要素の形容に関する記述を積極的に活用し，印象予測精度の向上につながるモデルを提案する．その際，印象評価に有効な形容記述の模索の効率化のため，大規模言語モデル（Large Language Model; LLM）を活用する．クラウドソーシングで得た印象評価データセットを用いて，既存手法と提案手法の予測性能を比較することで，提案手法の有用性を示す．

1. 序論

近年，都市景観画像から印象を数値評価することが，都市開発や土地ブランディングに役立つとされ，研究が進んでいる．都市景観は，その地域の魅力やブランド価値を直接的に反映し，来訪者の誘致や都市デザインの改善に大きく寄与するためである．実際に，都市景観画像から人々が感じ取る美しさや安全性のような印象を数値予測する試みは，様々な方法で取り組まれてきた [3], [5], [10].

その中でも，シーン全体を単一の物体とみなす既存手法 [4], [12], [14], [17] がよく使われる一方で，シーン内の多様な構成要素を十分に把握できないため，予測性能が限定的という課題がある [15]．例えば，ゴミや落書きのような印象に強く影響する要素を含む景観画像において，要素の検出が困難なため予測精度が低下すると考えられる．このように，シーン構成要素のクラス情報や局所的な特徴を捉えることが困難なことから，高精度な印象予測のために工夫が求められる．

[1] 東京科学大学 情報理工学院 情報工学系
School of Computing, Department of Computer Science, Institute of Science Tokyo
[2] LINE ヤフー株式会社
LY Corporation
[a)] ide@miubiq.cs.titech.ac.jp
[b)] anno@miubiq.cs.titech.ac.jp
[c)] ktsubouc@lycorp.co.jp
[d)] simosaka@miubiq.cs.titech.ac.jp

(a)　(b)

図 1: シーンの構成要素とレイアウトは類似するが与える印象が異なる景観画像の例.

そのため，印象評価手法は，Hou ら [6] によってシーン内の個々の物体を検出し，検出領域での画像特徴から印象の数値を予測する手法へと進展するが，以下の問題点が残っている．物体検出に特化した画像特徴は，同一クラスの物体の属性の差異の考慮が困難なため，景観の印象評価に直接関連性があるとは言いがたい．同一クラスの物体の属性の差異の例として，DALL-E 3[*1]によって生成された画像である図 1 を示す．この 2 枚の景観画像は，車，道路，建物，空などのシーン構成要素と，それらの配置や大きさといったレイアウトは類似するが，与える印象が異なる景観画像の組である．一般に，人々はこれらの画像に対して，図 1(a) よりも図 1(b) の方が，清潔感があり洗練されている印象を持つと考えられる．その要因として，同じ建物で

も汚れのある古い建物と清潔な建物のように，同クラスの物体の属性の違いを認識しているからだと考えられる．しかし，既存手法では物体検出により，車や建物といったシーン構成要素のクラス及び，その視覚的な特徴を考慮するが，図 1(a) の汚れのある古い建物と，図 1(b) の清潔な建物というような，同クラスの物体の形容的な属性の違いを考慮しているとは言い切れない．そのため，既存の物体検出ベースに手法により抽出される特徴量は，印象評価に直接関連した手法とは言い難い．

そこで本研究では，形容に関する特徴に着目する．形容に関する特徴とは，既存手法の不足点であったシーン構成要素の形容表現に加え，視覚的な特徴から連想される特徴である．例えば，図 1(a) のシーンにおける，絡み合った電線のような**検出要素の状況**，画像の半分を占める建物の並びのような**構図**，さらに，人々の生活を感じさせる都市の裏路地のような**シーン全体の文脈的・背景的特徴**の利用を考える．このような形容に関する特徴をモデルに与えることで，既存手法における物体検出に特化した画像特徴の不足点を補う．

しかし，印象評価予測の高精度化に寄与する形容に関する記述は自明でない．図 1(a) の形容に関する特徴の一例として，「都市の裏路地の生活感が漂う」記述と，「雨あがりの静寂な古びた建物に囲まれた」記述などが考えられるが，どちらの表現が印象評価に有効なのか不明確である．そのため，印象評価に有効な記述の模索が必要とされる．

そこで本研究では，シーンの視覚的特徴に加えて，LLM（Large Language Model; LLM）により生成された形容に関する特徴を考慮し，より頑強な印象評価を数値予測する手法，**CityInsight モデル**を提案する．LLM を活用することで，印象評価の高精度化に寄与する表現の模索を効率的に行うことが可能になる．これは，LLM であれば，人がシーンの形容に関する特徴を記述する場合と比べて，プロンプトの調節のみで記述を自動生成，さらにテキストデータの埋め込みが容易なためである．以上のように，シーンの視覚的な特徴に加え，シーンの形容に関する特徴を明示的にモデルに与えることで，既存手法における物体検出に特化した画像特徴の不足点を補い，より精緻な印象評価の数値予測を実現する．

本研究の貢献は以下のようにまとめられる．

- 高精度な印象予測のために，物体検出に特化した画像特徴の不足点を補う，シーン構成要素の形容に関する記述を活用した手法である CityInsight モデルを提案する．
- CityInsight モデルでは，形容に関する記述の生成のために LLM を活用することで，印象評価に有効な表現の模索を効率化する．
- クラウドソーシングで得た都市景観の印象評価データセットを用いて，シーン構成要素の画像特徴に着目した既存手法と，それに加えて形容的な特徴を考慮した提案手法の予測性能を比較することで，提案手法の有用性を示す．

関連研究

Visual Urban Perception

Visual Urban Perception は，都市景観画像に対して，画像処理技術に基づき，安全性，雰囲気，美しさというような印象評価軸を設定し，印象評価を目的とする研究分野である．当該研究分野の代表的な手法として，画像全体からシーンの特徴を学習し，印象評価を行う手法 [4], [11], [17] がある．Dubey ら [4] は，クラウドソーシングにより，6 つの印象評価軸からなる印象のアノテーションが行われた景観画像に対して，各印象評価軸ごとにランク学習を行い，印象スコアを予測した．また近年の研究 [11], [17] では，景観画像に対して，複数の印象評価軸が設定されていることから，その相互作用をマルチタスク学習によって考慮している．特に，久保田ら [17] の手法は，各印象評価軸に対して 5 段階でスコアリングされたデータセットに対して，人々が感じとる印象の個人差を考慮するために，スコア分布の予測を行った．しかし，これらの手法では，シーン内の詳細な特徴を捉えることが困難である．

そこで，画像内のシーン構成要素のラベル情報を考慮し，シーンの詳細な分析による印象評価を試みた手法 [5], [15] が提案された．これらの手法では，セマンティックセグメンテーションモデルや物体検出モデルを活用し，シーンの植物の割合や，空の割合，車の個数等を特徴量として扱い，機械学習モデルによって印象予測が行われた．

しかしながら，以上のような手法では人々が都市景観画像から感じとる，印象評価過程を模倣した手法としては，不十分である．これは，シーンを詳細に解析するために，物体検出やセグメンテーションが用いられたが，同一クラスの物体間の形容的な属性の考慮が困難なためである．

Image Aesthetic Assessment

画像データに対して，人間の美的評価の定量化を試みる様な研究分野である Image Aesthetic Assessment が存在する [1], [6], [9], [16]．当該研究分野では，都市景観画像に加えて，風景写真や肖像画，抽象画等の多種多様な画像に対して，人々が感じ取る美しさの定量化を行う．Celona ら [1] の手法では，シンメトリーな配置や，3 分割法というような構図と，HDR (High Dynamic Range) やボケといった画像スタイルを画像特徴量と共にモデルに入力し，高精度な美的予測が提案されている．同様に，シーン構成要素の構図を積極的に考慮するための手法として，画像を局所領域に分割し，それらの相互関係をグラフ構造でモデル化することで画像の局所的特徴とそれらの関係性を学習する研究 [9], [16] も存在する．以上から，美的評価では芸

*1 https://openai.com/index/dall-e-3/

術性が関連するために，構図を重視した手法は高精度な予測に寄与することが分かる．

しかし，都市景観の印象評価の場合，構図の自由度は限られるため，よりシーン構成要素自体の特徴が重要となる．美的評価のうち，シーン構成要素に着目した手法として，Hou ら [6] は，物体検出を用いて，検出物体の視覚的特徴から美的評価を行い高精度な美的スコア予測を試みたが，同クラスの物体の属性性的な特徴の考慮は困難なため，都市景観の印象評価の数値予測手法としては不十分である．

2. 問題設定と既存手法の限界

2.1 問題設定

本研究では都市景観画像に対して人々が感じ取る印象の数値予測を行う．その際，感じ取る印象は複数種類あるとし，それら印象評価軸の集合を M とし，各評価軸を m とする．また，景観画像の総数を N とし，その集合を $\boldsymbol{I} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ によって定義する．それぞれの景観画像 $\boldsymbol{x}_i$ に対応する印象評価軸 m の印象スコアを $y_i^{(m)}$ と表す．以上のことから，画像に対して印象評価を行う問題は，各画像 $\boldsymbol{x}_i$ を入力として受け取り，対応する $y_i^{(m)}$ を推論する関数 $f(\cdot)$ の学習として定式化される．

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

ここで，$\mathcal{L}$，$\boldsymbol{\theta}$ 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である．

2.2 既存の印象予測手法の限界

既存手法において，画像に対する高性能な印象評価手法は多く提案されてきたが，そのほとんどは事前学習済みモデルによって得られる画像特徴に基づいた手法に留まる [4], [6], [17]．ここで，画像特徴量抽出のための事前学習済みモデルを $\boldsymbol{M}_{\text{image}}$ とすると，事前学習済みモデルから得られる画像特徴量 $\boldsymbol{v}_i$ 以下のように表せる．

$$\boldsymbol{v}_i = \boldsymbol{M}_{\text{image}}(\boldsymbol{x}_i) \tag{2}$$

また，印象予測モデルを $\boldsymbol{M}_{\text{pred}}$ とすると予測スコア $y_i^{(m)}$ は以下のように定式化される．

$$y_i^{(m)} = \boldsymbol{M}_{\text{pred}}(\boldsymbol{v}_i) \tag{3}$$

これらの手法は，式 3 の通りに，画像特徴量 $\boldsymbol{v}_i$ に基づいた学習を行っていることから，シーン構成要素の形容に関する特徴を考慮した学習が行われる保証がないため，人間の印象評価過程と乖離が生じ，印象評価手法としては不十分であると推察できる．

3. 提案手法: CityInsight

3.1 提案手法の概要

本研究では，シーン構成要素の視覚的特徴に着目した既存手法の問題の解決に取り組む．その問題とは，物体検出に特化した画像特徴は，同クラスの物体の属性的な差異の考慮が困難なため，都市景観の印象評価に直接関連し難いことであった．

そこで我々は，都市景観の高精度な印象評価のために，シーン構成要素の視覚的特徴に加えて，シーン構成要素の形容に関する特徴を明示的にモデルに与える手法，CityInsight モデルを提案する．CityInsight モデルは，景観画像と LLM によって生成されたシーン構成要素の形容に関する記述に基づき，印象スコア分布を予測する．さらに，LLM の活用によって，印象評価に有効な形容に関する記述が自明でない中，効率的な表現の模索が可能になる．

図 2 に示すように，CityInsight モデルは大きく 3 つの構造に分けられ，以下の 3 節で各構造の詳細を述べる．

- Image feature extractor: シーン全体の視覚的な特徴量抽出を行う（3.2 節）．
- Adjective feature extractor: LLM を活用したシーン構成要素の形容に関する特徴量抽出を行う（3.3 節）．
- Score distribution predictor: 景観画像の視覚的な特徴と形容に関する特徴に基づいた景観画像の印象スコア予測を行う（3.4 節）．

以上の構成により，印象評価に関わる形容に関する特徴を明示的に捉えることで，印象予測の精度向上を試みる．

3.2 画像特徴の抽出

図 2 の Image feature extractor の画像特徴量抽出について述べる．画像特徴量抽出の際，計算効率化やメモリ節約の観点から，事前学習済みの畳み込みニューラルネットワーク (CNN) を活用する．この事前学習済みモデルの出力層以外を画像特徴量抽出器として活用する．この時，式 (2) の通り，各景観画像 $\boldsymbol{x}_i \in \boldsymbol{I}$ から特徴量ベクトル $\boldsymbol{v}_i$ が得られる．

3.3 シーン構成要素の形容に関する特徴の抽出

本節では，シーン構成要素の形容に関する特徴の抽出方法ついて述べる．図 2 の Adjective feature extractor 部の通り，景観画像とプロンプトを LLM に入力することによって，景観画像の形容に関する記述を生成する．この記述の自動生成により，人が記述を生成する場合よりも，印象評価に有効な表現の効率的な模索が可能となる．そして，生成された記述を埋め込みベクトルに変換することで，形容に関する特徴を持つ特徴量ベクトルを抽出する．

*2 プロンプトとシーンの説明記述は一部省略済みのものを掲載．

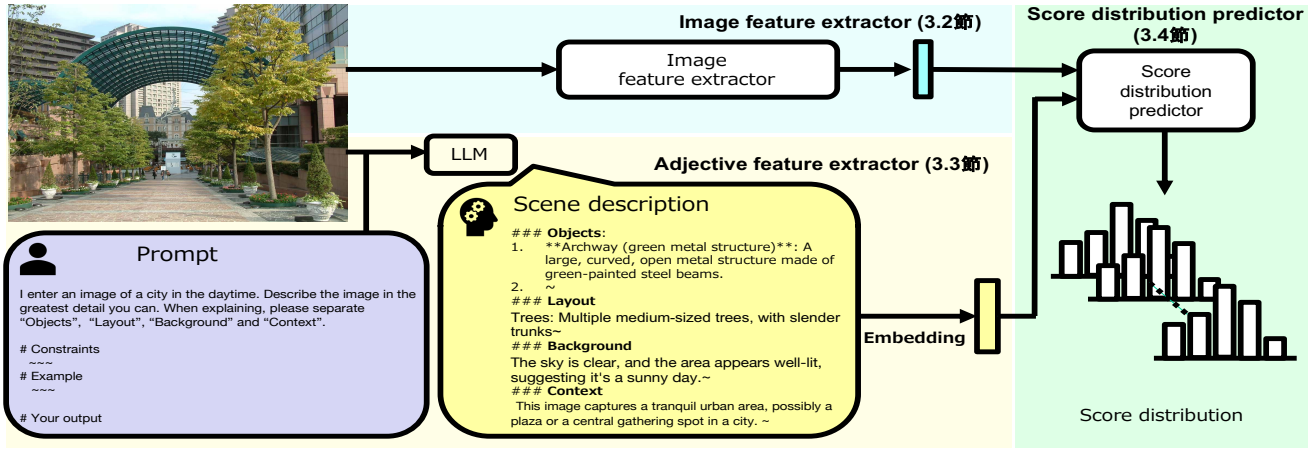

図 2: 提案手法の概要図. *2

3.3.1 形容記述生成のためのプロンプトの詳細

```
I enter an image of a city in the daytime. Describe the image in the greatest
detail you can. When explaining, please separate "Objects", "Layout as
seen from camera", "Background" and "Context".

# Constraints
- The output should be structured as follows:
  <"Objects"> : <<Nouns without adjectives> : <Describe the object in as
much detail as possible, using its materials, characteristics, texture, etc.>>
  <"Layout"> : <<Nouns without adjectives> : <Describes in detail the
positional relationships between Objects in "Objects", such as the
placement and size of objects. >
  <"Background">: <Describe background in the image. >
 <"Context">: <Describe context in the image. >

- For Objects, please list 10 in order of prominence
- If there are multiple objects of the same type, output for each object.

# Example
(出力例)~~

# Your output
```

図 3: 形容に関する記述生成のためのプロンプト例

LLM にプロンプトと景観画像を入力することで，シーンの形容に関する特徴の抽出を行う．その際，景観画像を包括的に捉えるため，本研究では **Objects**, **Layout**, **Background**, **Context** の 4 つの説明要素を定義する．それらの説明要素について，図 2 のプロンプトの通りに景観画像に関する説明を記述させる．

また，プロンプトに出力の具体例である Example も含める．これは，出力に印象評価に不要な文章を出力させず，出力形式に一貫性を持たせるためである．以下に，各説明要素と example の詳細について述べる．

Objects とは，シーン構成要素の材質，質感といった形容的な特徴の抽出のために指定する．具体的には，車，人，川ではなく光沢のある赤い車，会話している人，茶褐色に濁った川のように，シーン構成要素についての質感や形容的な特徴や，状況に関する記述を表す．また，指示文に 10 個の重要な物体について説明させるように記述する．これは，既存手法 [6] において，10 個の物体検出を行うとシーン全体の 80 %をカバーできたことを参考にした．さらに，1 つの景観画像内に複数の同クラスの物体が存在する場合，それぞれの個体について記述するように指示を行う．出力の形式は図 3 の <"Objects"> の様に，　検出要素の形容詞を除いた抽象名詞：その検出要素の説明　となる．

Layout は，**Objects** で検出した要素に対する，カメラから見た検出物体の位置関係，大きさの記述である．これは，画像内に占める割合が大きい物体や，シーンの中央に位置する物体は印象に大きく影響を与えるという仮説に基づき，説明要素として設定した．以上の出力を意図した指示文を，図 3 の <"Layout"> のように記述する．

Background は，主要な物体や人物の背後にある天気や場所といった，シーンの視覚的な舞台設定の特徴の抽出のための記述である．**Background** の説明記述例として，「多くのビルが立ち並び，ビジネス街を連想させるシーン」といった記述である．シーンの天候や場所に関する記述は，シーンを 1 つの物体と見なした場合の属性的情報に値すると考え，**Background** を説明要素とする．以上のような出力を意図した指示文を，図 3 の "Background" のように記述する．

Context とは，シーン全体における，状況や文化的な背景などの意味的な解釈とする．これは，**Objects** の個々の物体に関する状況説明や **Background** と似るが, **Context** は個々のシーン構成要素がどのように干渉し合い，シーン全体としてどのような意味を成すかを表すため役割が異なる．例えば，公園で人々がくつろいだり，遊んでいる画像について考える．この場合，**Objects** では，ベンチや，座っている人，遊具で遊んでいる人，木などについての特徴記述が生成されると想定でき，**Background** では，ビルが立ち並んでいることから，「都市の公園である」という記述が想定できる．一方で，**Context** では，家族が集まり公園でくつろぐ日常の風景から，「午後の都市公園でのリラックスした時間」といった記述を想定する．このように，**Objects** では個々の要素に着目し，**Background** ではシーン全体の環境について記述を期待し，**Context** ではシーン全体から連想される意味的解釈の記述を想定する．

Example では，出力の具体例を入力することで，LLM による出力形式を固定する．出力形式が固定でない場合，

“I will explain this image.” や “abstract noun: car” というような，形容に関する記述とは直接的に関係の無い文章が生成される．そのため，図 2 の “Scene description” のような出力を，プロンプトの “#Example” として入力する．

3.3.2 生成された形容記述の特徴量ベクトル化

プロンプトによって得られた，景観画像のシーンの形容に関する記述に対し，埋め込みベクトルに変換することで特徴量ベクトルを得る．ここで，各景観画像に対する LLM による出力を $\boldsymbol{T} = \{\boldsymbol{t}_i\}_{i=1}^{N}$ とし，記述の埋め込みのためのモデルを $\boldsymbol{M}_{\text{embedding}}$ とすると，得られる形容に関する特徴量ベクトル $\boldsymbol{u}_i$ は次のように表せる．

$$\boldsymbol{u}_i = \boldsymbol{M}_{\text{embedding}}(\boldsymbol{t}_i) \tag{4}$$

3.4 印象スコア分布の予測

図 2 の Score distribution predictor について述べる．本研究では，個人の印象の感じ方の違いを考慮するため，M 個の印象評価軸によって印象スコアを，K 段階のスコア分布で予測する．

スコア分布予測器のモデル構造は，多層パーセプトロン (Multi-layer perceptron; MLP) であり，画像特徴量 $\boldsymbol{v}_i$ と，形容に関する特徴量 $\boldsymbol{u}_i$ 入力とすると，以下のように定式化される．

$$\boldsymbol{z}_i = \boldsymbol{f}_{\text{FN}}^{\text{img}}(\boldsymbol{v}_i) \oplus \boldsymbol{f}_{\text{FN}}^{\text{txt}}(\boldsymbol{u}_i) \tag{5}$$

$\oplus$ は結合子を表し，関数 $\boldsymbol{f}_{\text{FN}}^{\text{img}}$，$\boldsymbol{f}_{\text{FN}}^{\text{txt}}$ は共に全結合層である．

また，$\boldsymbol{z}_i$ は次元削減された $\boldsymbol{v}_i$ と $\boldsymbol{v}_i$ を結合した特徴量ベクトルであり，これを式 (6) における入力とする．

$$\boldsymbol{P}_i = \text{MLP}(\boldsymbol{z}_i) \tag{6}$$

式 (6) は，$\boldsymbol{z}_i$ に基づく，MLP による各評価軸に対する印象スコア予測を表す．この時，MLP はマルチタスク学習を行うため，M 個の印象評価軸における K 段階のスコア分布である $\boldsymbol{P_i} \in \mathbb{R}^{M \times K}$ を出力する．そのため，m 番目の印象評価軸のスコア分布を $\boldsymbol{p}_i^{(m)} \in \mathbb{R}^{\text{K}}$ とすると，以下のように表せる．

$$\boldsymbol{P}_i = [\boldsymbol{p}_i^{(1)}, \boldsymbol{p}_i^{(2)}, \ldots, \boldsymbol{p}_i^{(M)}]^\top \tag{7}$$

4. 性能評価実験

景観画像に対して印象スコアがアノテーションされたデータセットを用いて，提案手法の有用性を評価する．

4.1 実験で使用するデータセット

本研究では，久保田ら [17] らの研究でも用いられたデータセットを利用する．これは，Yahoo!クラウドソーシング[*3]のプラットフォームを活用することで，2305 枚の各景観画像に対する印象のスコアの付与が実現されている．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印象の評価軸 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，久保田ら [17] の研究に基づき，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，各景観画像に対して，知覚評価軸 m に対応した 5 段階評価に基づいたスコア分布 $\boldsymbol{p}_i^{(m)} \in \mathbb{R}^5$ が紐付けられる．

4.2 実験設定

4.2.1 実験で使用するプロンプト

本研究では，印象評価に有効な形容に関する表現の模索のため，4 種類のプロンプトを設計した．ベースとなるプロンプトは，景観画像に対して **Objects**, **Background**, **Context** の 3 要素について記述させるものである．各々のプロンプトの違いは，構図に関する説明の **Layout** の有無，出力の具体例である Example の有無，シーン構成要素に関する記述の **Objects** の指示の粒度差である．**Objects** の指示の粒度差とは，図 3 の <“Objects”> の様な，**Objects** の指示文である “Describing a feature of the object.”, “Adjectival expression for that object”, “Describe the object in as much detail as possible, using its materials, characteristics, texture, etc.” の 3 通りの指示文のような，求める出力の具体性の差である．また，上記の 3 通りの指示文を記載順に，Simple Instruction，Intermediate Instruction，Detailed Instruction と呼ぶ．以下に本実験て活用したプロンプト設計について述べる．

（1）Simple Prompt without Example: このプロンプトは **Objects**, **Background**，**Context** の 3 つの説明要素に基づきシーンの説明記述を生成する．この際，**Objects** に関する出力への指示文は，粒度の荒い Simple

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

表 1: 既存手法と提案手法における印象予測性能の比較.

手法	評価指標	雰囲気	秩序	高価さ	こだわり	all
Scene Centric Model [17]	MAE ↓	0.270 ± 0.02	0.316 ± 0.04	0.309 ± 0.03	0.342 ± 0.03	0.309 ± 0.03
	相関係数↑	0.750 ± 0.02	$\mathbf{0.720 \pm 0.06}$	0.749 ± 0.05	0.690 ± 0.04	0.727 ± 0.04
Objects Centric Model [6]	MAE ↓	0.510 ± 0.01	0.460 ± 0.01	0.570 ± 0.01	0.658 ± 0.01	0.549 ± 0.01
	相関係数↑	0.281 ± 0.01	0.165 ± 0.01	0.126 ± 0.01	0.130 ± 0.01	0.176 ± 0.01
CityInsight Model	MAE ↓	$\mathbf{0.233 \pm 0.01}$	$\mathbf{0.298 \pm 0.01}$	$\mathbf{0.275 \pm 0.01}$	$\mathbf{0.305 \pm 0.01}$	$\mathbf{0.278 \pm 0.01}$
(Detailed Objects Prompt)	相関係数↑	$\mathbf{0.806 \pm 0.02}$	0.707 ± 0.03	$\mathbf{0.792 \pm 0.01}$	$\mathbf{0.731 \pm 0.01}$	$\mathbf{0.759 \pm 0.01}$

Instruction を用いる．**（2）Object Adjective Prompt**: このプロンプトは，Simple Prompt without Example に対して，出力の具体例も記述する．また，**Objects** に関する出力への指示文を，検出物体の形容詞を記述させるための Intermediate Instruction に変更することで，指示に具体性を持たせる．**（3）Detailed Objects Prompt**: このプロンプトは Object Adjective Prompt の **Objects** に関する出力への指示文を Detailed Instruction に変更したプロンプトである．**（4）Detailed Objects Prompt with Layout**: このプロンプトは，Detailed Objects Prompt に対して，説明要素である Layout を追加したものである．

これらの設計したプロンプトと景観画像を，OpenAI の LLM である GPT-4o[*4]に入力し文章を生成する．

4.2.2 生成された形容記述の埋め込みベクトル化

本研究では，生成された形容記述の埋め込みベクトル化の際，テキスト全体を埋め込みベクトルに変換することで，生成記述の特徴量化を行った．これは，シーン構成要素の形容的な表現や，構成要素間の干渉によって生まれる意味的特徴を捉えた，シーン全体の特徴をベクトルとして扱うと解釈できる．

4.2.3 特徴量の設定

モデルに与える特徴量として，Image feature extractor (3.2 節) により抽出される画像特徴量の $\boldsymbol{v}_i$ (式 (2)) と，Adjective feature extractor (3.3 節) により抽出される形容に関する特徴量の $\boldsymbol{u}_i$ (式 (4)) である．Image feature extractor では，事前学習済みモデルとして，ImageNet [2] データセットを学習した MobileNet-v1 [7] を活用し，画像特徴量 $\boldsymbol{v}_i \in \mathbb{R}^{1024}$ が得られる．また，形容に関する記述の埋め込みの際は，OpenAI の `text-embedding-ada-002` モデル[*5]を用いる．本研究ではテキスト全体に対して埋め込みを行うため，$\boldsymbol{u}_i \in \mathbb{R}^{1536}$ である特徴量ベクトルが得られる．

4.2.4 学習モデルの構造設定

MLP を活用した Score distribution predictor (3.4 節) の構造設定について述べる．式 (6) において，入力となる特徴量は，画像特徴量 $\boldsymbol{v}_i \in \mathbb{R}^{1024}$ と形容に関する特徴量 $\boldsymbol{u}_i \in \mathbb{R}^{1536}$ であり，各々全結合層に入力し次元を 512 に削減する．この時の活性化関数は ReLU [13] を用いる．これらの 2 つの特徴量ベクトルを結合し，$\boldsymbol{z}_i \in \mathbb{R}^{1024}$ を得る．また，式 (6) における MLP は 4 層で構成される．第 l 層の出力を $\boldsymbol{h}^{(l)}$ とすると，$\boldsymbol{h}^{(1)} \in \mathbb{R}^{512}$，$\boldsymbol{h}^{(2)} \in \mathbb{R}^{256}$，$\boldsymbol{h}^{(3)} \in \mathbb{R}^{128}$，$\boldsymbol{h}^{(4)} \in \mathbb{R}^{5}$ となるようにパラメータを設定する．この時，活性化関数は $l = 1, 2, 3$ 層では ReLU 関数，出力層である $l = 4$ 層では Softmax 関数を用いる．

4.2.5 モデルの学習設定と学習指標

CityInsight モデルの学習の際，景観画像データセットを 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う．その際，ハイパーパラメータはバッチサイズは 32 に設定し学習率 0.0001 の元，エポック数を 50, 80, 100, 120, 150 のいずれかに設定する．また，損失関数は KL-ダイバージェンスを使用し，最適化アルゴリズムは Adam [8] を用いる．評価指標として，平均絶対誤差 (Mean Absolute Error; MAE) と相関係数を用いる．

4.3 実験の詳細と結果

4.3.1 既存手法と提案手法における印象予測性能の比較

本実験では，以下の 2 つの既存手法と印象評価性能を比較して，CityInsight モデルの有用性を評価する．

Scene Centric Model [17]: 本研究では，画像からシーン全体の画像特徴量を抽出し，印象スコア分布を予測する久保田ら [17] の手法を Scene Centric Model と呼ぶ．本研究と同様に，マルチタスク学習により，景観画像に対して「雰囲気」・「秩序」・「こだわり」・「高価さ」の 4 つの評価軸で印象評価を行った研究のため，この手法と比較を行う．

Objects Centric Model [6]: 本実験では，シーン内の個々の物体を検出し，注意機構により，検出物体の重要度を考慮し，それらの画像特徴量から美的スコア分布を予測する手法を Objects Centric Model と呼ぶ．この手法は Hou ら [6] が提案したモデルを，メモリ節約の観点から一部の次元を削減したものである．印象評価において，物体検出をベースに画像特徴から印象評価を行う手法で最新の手法であるため，この手法と比較を行う．

表 1 に，以上のような既存手法と提案手法の景観画像に対する印象評価の比較結果を示す．all は実験で使用した各評価軸の平均値を表す．提案手法の各評価軸，all における MAE，相関係数は，既存の 2 つの手法と比較すると，秩

[*4] https://openai.com/index/hello-gpt-4o/
[*5] https://openai.com/index/new-and-improved-embedding-model/

表 2: 形容表現による印象予測性能の変化.

プロンプト	評価指標	雰囲気	秩序	高価さ	こだわり	all
Detailed Objects Prompt	MAE ↓	0.233 ± 0.01	$\mathbf{0.298 \pm 0.01}$	$\mathbf{0.275 \pm 0.01}$	$\mathbf{0.305 \pm 0.01}$	$\mathbf{0.278 \pm 0.01}$
	相関係数↑	0.806 ± 0.02	$\mathbf{0.707 \pm 0.03}$	$\mathbf{0.792 \pm 0.01}$	$\mathbf{0.731 \pm 0.01}$	$\mathbf{0.759 \pm 0.01}$
Detailed Objects Prompt with Layout	MAE ↓	0.240 ± 0.01	0.306 ± 0.01	0.287 ± 0.01	0.309 ± 0.01	0.285 ± 0.01
	相関係数↑	0.793 ± 0.02	0.687 ± 0.02	0.773 ± 0.01	0.720 ± 0.02	0.743 ± 0.01
Objects Adjective Prompt	MAE ↓	$\mathbf{0.231 \pm 0.01}$	$\mathbf{0.298 \pm 0.01}$	0.282 ± 0.01	0.315 ± 0.01	0.282 ± 0.01
	相関係数↑	$\mathbf{0.807 \pm 0.02}$	0.704 ± 0.03	0.776 ± 0.02	0.703 ± 0.02	0.747 ± 0.02
Simple Prompt without Example	MAE ↓	0.254 ± 0.01	0.319 ± 0.01	0.301 ± 0.01	0.337 ± 0.01	0.303 ± 0.01
	相関係数↑	0.760 ± 0.01	0.643 ± 0.01	0.727 ± 0.02	0.655 ± 0.02	0.696 ± 0.01

表 3: モデル構成要素の対照実験結果.

手法	評価指標	雰囲気	秩序	高価さ	こだわり	all
Image-Based Model	MAE ↓	0.270 ± 0.02	0.316 ± 0.04	0.309 ± 0.03	0.342 ± 0.03	0.309 ± 0.03
	相関係数↑	0.750 ± 0.02	0.720 ± 0.06	0.749 ± 0.05	0.690 ± 0.04	0.727 ± 0.04
Text-Based Model	MAE ↓	0.235 ± 0.02	$\mathbf{0.297 \pm 0.01}$	$\mathbf{0.274 \pm 0.01}$	0.306 ± 0.02	$\mathbf{0.278 \pm 0.01}$
(Detailed Objects Prompt)	相関係数↑	0.805 ± 0.02	$\mathbf{0.710 \pm 0.03}$	$\mathbf{0.792 \pm 0.01}$	0.725 ± 0.01	0.758 ± 0.02
CityInsight Model	MAE ↓	$\mathbf{0.233 \pm 0.01}$	0.298 ± 0.01	0.275 ± 0.01	$\mathbf{0.305 \pm 0.01}$	$\mathbf{0.278 \pm 0.01}$
(Detailed Objects Prompt)	相関係数↑	$\mathbf{0.806 \pm 0.02}$	0.707 ± 0.03	$\mathbf{0.792 \pm 0.01}$	$\mathbf{0.731 \pm 0.01}$	$\mathbf{0.759 \pm 0.01}$

序の相関係数以外は全て，大幅に改善されていることが分かる．また，Objects Centric Model の性能が著しく劣化していることがわかる．これは，本手法が個々のオブジェクトに対する画像特徴量を用いている一方で，シーン全体に対する特徴量を用いていないことが原因であると考えられる．

4.3.2 形容表現による印象予測性能の変化

印象予測に有効な形容に関する記述の評価のために，4.2.1 節の 4 種類のプロンプトによる印象スコア予測結果の比較を行った．表 2 より，Detailed Objects Prompt が雰囲気以外の印象評価軸で良い性能であることが確認できる．このプロンプトでは，説明要素 **Objects** について，物体の素材や質感まで具体的に説明させるよう指定したこのプロンプトに着目し，その他のプロンプトによる性能比較を行う．**Objects** の出力をシーン構成要素の形容詞のみになるように設定した Objects Adjective Prompt との比較では，雰囲気のみの性能が僅かに劣る．また，**Layout** を除いた Detailed Objects Prompt with Layout との比較では，Detailed Objects Prompt よりも性能が低下していることがわかる．これは出力テキスト全体を一度に埋め込みベクトル化した影響で，特徴量ベクトルの **Layout** 以外の説明要素の情報量が低下したためと考えられる．以上のことから，**Objects** に対して細かい粒度で LLM に説明させる事は性能向上に寄与すると考えられる．

また，Simple Prompt without Layout の結果に着目すると，他のプロンプトによる結果に比べて，各印象評価軸，all の全てに対して性能が悪いことが確認できる．これは，出力例をプロンプトに含めていないため，LLM から出力される記述に，“I will explain this image.” や “abstract noun: car” というような，形容に関する記述とは直接的に関係の無い文章が生成され，結果的に特徴量ベクトルが持つ情報量が減るためだと考える．

4.3.3 モデル構成要素の対照実験

LLM によるシーン構成要素の形容に関する記述を考慮した手法の有用性を調査するために，提案手法のモデル構成要素の対照実験を行う．そのため，以下の 3 つのモデルにより，印象評価性能の比較を行う．

- **Image-Based Model**：画像特徴量に基づいたスコア分布予測を行うモデル.
- **Text-Based Model**：シーン構成要素の形容に関する記述のみに基づいたスコア分布予測を行うモデル.
- **CityInsight モデル**：画像特徴量と物体の形容に関する特徴量に基づいたスコア分布予測を行うモデルであり，提案手法である．この時，4.3.2 節における実験において，最も性能の良いプロンプトである Detailed Objects Prompt を活用する.

表 3 に，モデル構成要素の対照実験の結果を示す．これを見ると，Text-Based Model よりも CityInsight モデルの方が，僅かに性能が良い事が確認できる．また，Image-Based Model を他のモデルと比較すると，どの印象評価軸を見ても，MAE，相関係数は低いことが分かる．このことから，形容に関する特徴を考慮した印象予測手法は有効であると考える．

4.3.4 CityInsight モデルによる印象スコア予測の例

図 4，5 に，CityInsight モデルによる景観画像に対する予測誤差が，小さい結果例と大きい結果例を示す．各画像右上に印象評価軸別に予測荷重平均スコアと正解荷重平均スコアを掲載した．これらは，5 段階のスコア分布における各段階の割合を重みとした，5 段階スコアの荷重平均によってスカラー値へ変換を行ったものである.

図 4 で示すように正確な印象数値予測が行えていることが確認できる一方で，図 5 の両画像では大きく予測が外れていることが分かる．特に図 5(b) の画像では，シーン構成要素自体は全体的に古く見えることから，CityInsight モデルでは各印象評価軸で低いスコアを予測している．しかし，

(a) (b)

図 4: 正解と予測の誤差が小さい例．各画像右上に各スコアを予測スコア（正解スコア）と表示．

(a) (b)

図 5: 正解と予測の誤差が大きい例．各画像右上に各スコアを予測スコア（正解スコア）と表示．

正解スコアでは日差しの入り加減や，椅子などの小物が整理され並んでいることから，雰囲気，秩序において高いスコアが付けられたと考えられる．このように，CityInsight モデルでは，シーン構成要素の詳細に着目し過ぎることから，シーン全体での印象を正確に捉えられない例の存在が確認できた．

5. 結論

本研究では，人々が都市景観画像から感じ取る印象スコア分布予測に取り組んだ．既存手法の問題点は，シーン構成要素の形容的な特徴を考慮している保証がないために，その手法によって抽出される特徴量が印象評価に直接関連しているとは言い難いことであった．そこで本研究では，CityInsight モデルを提案した．CityInsight モデルでは，画像の視覚的な特徴に加え，LLM を活用し，形容に関する特徴を明示的に取り入れることで，精度向上を実現させた．実験では，クラウドソーシングで得た印象評価データセットを用いることで，以下の 4 つのことが確認できた．(1) 既存手法との印象評価性能の比較実験（4.3.1 節）により，既存手法を上回る性能を達成した．(2) 形容に関する表現による印象評価性能の変化の評価実験（4.3.2 節）により，出力に求める説明を具体的に指定した方が，印象評価に有効であることが分かった．(3) モデル構成要素の対象実験（4.3.3 節）により，シーンに対する形容に関する特徴を明示的に活用することは，都市景観の印象評価において有効であることが分かった．(4) 一部の画像に対して，高精度な印象予測を行えなかった場合も見られた（4.3.4 節）．これに対して，印象評価に有効な表現の更なる模索や，出力された形容に関する記述に対する埋め込みベクトル化方法の工夫が考えられる．

参考文献

[1] Celona et al.: Composition and Style Attributes Guided Image Aesthetic Assessment, *IEEE Transactions on Image Processing* (2022).
[2] Deng et al.: ImageNet: A large-scale hierarchical image database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).
[3] Doersch, C. et al.: What Makes Paris Look like Paris?, *ACM Transactions on Graphics (SIGGRAPH)* (2012).
[4] Dubey, A. et al.: Deep Learning the City: Quantifying Urban Perception at a Global Scale, *Proceedings of the European Conference on Computer Vision (ECCV)* (2016).
[5] He, J. et al.: Extracting human perceptions from street view images for better assessing urban renewal potential, *Cities* (2023).
[6] Hou, J. et al.: Object-Level Attention for Aesthetic Rating Distribution Prediction, *Proceedings of the 28th ACM International Conference on Multimedia* (2020).
[7] Howard, A. G. et al.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, *arXiv preprint arXiv:1704.04861* (2017).
[8] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
[9] Liu, D. et al.: Composition-Aware Image Aesthetics Assessment, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2020).
[10] Min, W. et al.: Multi-Task Deep Relative Attribute Learning for Visual Urban Perception, *IEEE Transactions on Image Processing* (2020).
[11] Min, W. et al.: Multi-Task Deep Relative Attribute Learning for Visual Urban Perception, *IEEE Transactions on Image Processing* (2020).
[12] Moreno-Vera, F. et al.: Quantifying Urban Safety Perception on Street View Images, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2022).
[13] Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, Madison, WI, USA, Omnipress, p. 807–814 (2010).
[14] Porzi, L. et al.: Predicting and Understanding Urban Perception with Convolutional Neural Networks, *Proceedings of the 23rd ACM International Conference on Multimedia* (2015).
[15] Ramírez, T. et al.: Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety, *Landscape and Urban Planning* (2021).
[16] She, D. et al.: Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021).
[17] 久保田祐輝ほか：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，研究報告ユビキタスコンピューティングシステム (UBI) (2023).