複数の深層学習手法による街路の全方位の深度推定と都市景観評価への応用

瀧澤 重志*・衣川 雛**

Depth Estimation of Streets in All Directions by Multiple Deep Learning Techniques and Its Application to Cityscape

Atsushi Takizawa*, Hina Kinugawa**

Abstract: In this study, using plural deep learning techniques, a new space analysis method of the first-person viewpoint which can express geometric information of space as depth map of omnidirectional in addition to general RGB omnidirectional image and treat both in the frame of the same image analysis is developed. Concretely, the depth map is generated using pix2pixHD, after photographing a large number of omnidirectional images of pedestrian viewpoints in three-dimensional city models with high reality by using a game engine, and applying a style of real picture to them. Next, streets in the southern part of Osaka City are photographed by omnidirectional camera, and examinee experiment on some spaciousness is carried out, and CNN models which predict experimental result using general ResNet from input image including estimated depth map are learned, and validity and possibility of the proposed model are verified.

Keywords: 全方位画像，深度マップ，都市景観，Tobler 図法，pix2pixHD，WCT2，ResNet

１．はじめに

街路空間の景観の印象を評価する研究は以前から広く行われている．こうした印象評価の実験では，Virtual Reality（VR）を用いることが次第に増加しているが, 一般には景観の実写画像が利用されている．景観画像はインターネットを媒介として，高解像度な画像や利用可能なエリアが増加している．印象評価に空間の画像解析手法を用いる場合，従来は事前に画像からなにかしら意味のある計算可能な特徴を抽出する必要があった．例えば，天空率，緑視率，色彩分布といった評価指標が用いられることが一般的に用いられている．しかし，空間の印象評価は主観性が大きいので，それに影響を与える要素をすべて明示的に定義することは困難だと考えられる．

しかし近年，畳み込みニューラルネットワーク（CNN）が急速に進歩している，CNN は画像から特徴量を自動的に学習するため，都市の空間分析を含む多様な分野にますます適用されることが考えられる。たとえば Liu ら（2017）による研究では，専門家によって行われた中国の都市空間における印象評価実験の評価値を, CNN を使用してストリートビューの画像から直接予測する方法を提案している．この例の他にも，近年 CNN を使用した景観研究が増えつつあるが，これらの研究では全方位ではなく標準の画角の画像を使用している．しかし空間は視点から全方位に広がっており，開放感やサイズなどの幾何学的特徴も空間の評価には重要である．

従来，これら空間の幾何学的特徴は Space Syntax（Hillar and Hanson, 1984）と呼ばれるジャンルの空間分析の中で研究されてきた．その中で，isovist（Benedikt, 1979）は，空間の局所的な特徴を表現する基本的かつ重要なモデルである．isovist は，視点から見た可視領域を表し、2 次元の場合は多角形を，3 次元の場合は多面体を示す。Batty（2001）は，2 次元 isovist の平均距離や面積等のさまざまな空間特徴量を提案した．しかし isovist は複雑な形状になりやすく，形状特徴量を明示的に用いるアプローチは，従来の画像解析と同様の限界を有している．

ゲームなどで 3 次元空間の奥行情報を画像データとして記録されている情報を深度マップと呼ぶが，これを全方位でマッピングしたものは，3D isovist と近似的に同等の情報量を持つ. この事実に基づいて，Furuta and Takizawa（2017）は，ゲームエンジンの Unity を使用して構築された仮想の都市空間で, 多数

の全方位画像とその深度マップをリアルタイムで撮影した．そして VR を使用して行った嗜好調査実験から得た評価値を，先に撮影した画像を入力データとして，CNN を用いて予測するモデルを構築し，深度マップの有用性を確認した．さらに Kinugawa and Takizawa (2020) は，取得が困難な実空間の深度マップを pix2pix と呼ばれる深層学習の方法で推定し，ストリートビューの画像を対象として，球面型の CNN も含めて，嗜好予測を行った．

本研究は筆者らによるこれらの研究を発展させたものである．具体的には，大阪市郊外の全方位画像を歩道位置で撮影し，その結果に基づいて，CG 画像で取得する深度マップのスケーリングを行う．加えて，WCT2（Jaejun et al., 2019）と DeepLab v3+（Chen et al., 2018）と呼ばれる深層学習の手法を用いて CG 画像を実写に近づける．そしてそれらの画像から，pix2pix を改良した pix2pixHD (Ting-Chun et al., 2018) によって，より精度の高い深度マップの生成を試みる．さらに，調査地点の 3 つの空間に関連する因子を，40 人に増やした被験者にアンケートし，その平均値を，RGB，深度マップ，セマンティック・セグメンテーション（SS）画像などを組み合わせて，多チャンネルの入力画像とし，一般的な CNN である ResNet（He et al., 2016）を用いて回帰問題として学習・推定し，提案手法の有用性を検証する．

以降，実空間の全方位画像撮影，深度マップの学習，深度マップの学習，印象評価実験，印象評価値推定モデルの学習の順で説明する．

２．実空間の全方位画像撮影

2.1 撮影準備と実施

後述する印象評価の実験用に，大阪市住吉区の街路で全方位画像の撮影を行った．この地域は都心部から離れた住宅地が卓越する地域であるが，高級住宅地から狭小な住宅地域まであり，さらに大通り，高架，大きな公園，繁華街，学校，路地，路面電車など多様な空間構成がみられ，今回の印象評価実験が空間であることを考慮すると適切だと判断した．

全方位カメラには Richo THETA Z1 を使用した．撮影対象が屋外の全方位で明暗差の激しいことを考慮し，撮影モードを手持ち HDR とした．画像は最高画質・解像度の JPEG で出力した．同時に，撮影地点の位置を，Drogger GNSS DG-PRO1RWS を用いて RTK 測量で記録し，スマートフォン経由で EXIF データとして画像に記録した．RTK 測量の基準局には神戸市東灘区（JP-RJBE10）を設定し，Fix 時の理論測定誤差は最大 1.5cm であった．図 1 に示すように，市販のリュックのポケットに一脚を 2 本固定し，全方位カメラと測量機材を装着した．全方位カメラの高さは，リュックを撮影者が背負ったときに，頭が干渉しないよう，日本の Google Street View のカメラ高さの地上 2.05m となるよう調整した．

撮影日は，2020 年 11 月 5，9，18，19 日で，いずれも晴れが卓越する日であった．徒歩にて適当な間隔をあけながら合計 415 枚の画像を撮影した．撮影後に画像の EXIF データから，撮影地点の経緯度情報を抽出し，撮影地点を GIS でマッピングした（図 2）．画像の多様性を最大化するため，撮影地点間のユークリッド距離を最大化する数理計画問題を構成し，415 枚の画像の中から 200 枚の画像を選択した．結果として，撮影地点間の最小距離は 53m となった．最後に，200 枚の画像を 1,024×512px に縮小し，後の画像処理に備えた．

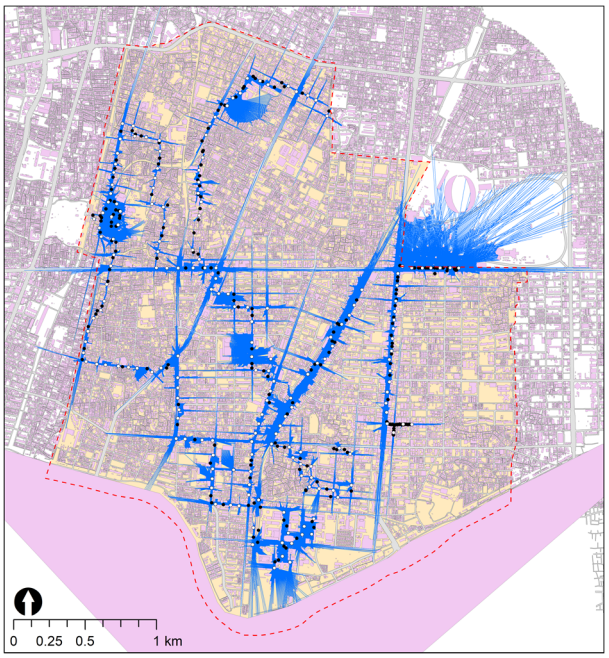


図 2 全方位画像の撮影地点と視線の分布（〇使用した撮影地点，●未使用の撮影地点，青線：視線）

2.2 深度マップの奥行の設定方法

CG で深度マップを生成するために，GPU の深度バッファの情報を用いる．この情報を画像として表現するには，最大距離のカットオフ値を設定する必要がある．素朴に考えると，距離 0 からカットオフ距離までを均等に分割し，距離を離散化して画素値として表現することになる．図 3 はカットオフ距離を 100，250，500m として，画素値を線形にスケーリングして作成した全方位の深度マップである．以後この方法を「線形（スケール）」と呼ぶ．本研究の深度マップでは、白色は距離が直近，黒色は距離がカットオフ距離以上であることを示す．

一般的な画像の 1 チャンネルあたりの情報量は 256 階調しかないため，カットオフの距離を長くすると距離の分解能が粗くなる．先の各距離での 1 画素値あたりの距離は，それぞれ 0.39，0.98，1.96m である．距離の分解能と，どこまで遠いオブジェクトを区別するかには，トレードオフの関係がある．今回用いる 200 の撮影地点を GIS でプロットした際に, 位置の誤差で, 建物の輪郭線ポリゴン（大阪市, 2020）内に入っていた 15 地点を除いた 185 の各観測点で，長さ 1,000m の視線を 1 度間隔で水平方向全周に生成する．そして，それらの視線と建物ポリゴンの線分交差判定を行い, 図 2 に示す 2D isovist を作成した．視線の長さの中央値は 21.9m，平均値は 49.4m であった. 図 4 に示すように, 視線の長さは，100m 以内が 88%，250m 以内が 97%，500m 以内が 99%であり，それらはさらに$N(3.08, 1.29)$の対数正規分布に従うことが確認できた．本研究で作成する深度マップは 3D isovist に対応するので，この視線の長さの分布は異なるが，水平方向の視線長さの分布状態は，空間の印象評価するうえで重要と思われる．そこでこの結果を踏まえ，距離の分解能と奥行のバランスを考慮し，本研究では線形モデルのカットオフ距離を 100m として，GPU で深度マップをリアルタイムにモノクロ画像として記録する．

しかし今説明したように，視線の距離分布は一様ではないことや，人間の感覚尺度は一般的に対数的であることから，理想的には距離が近いほど距離の解像度が高く，距離が遠くなるほど，値としては大

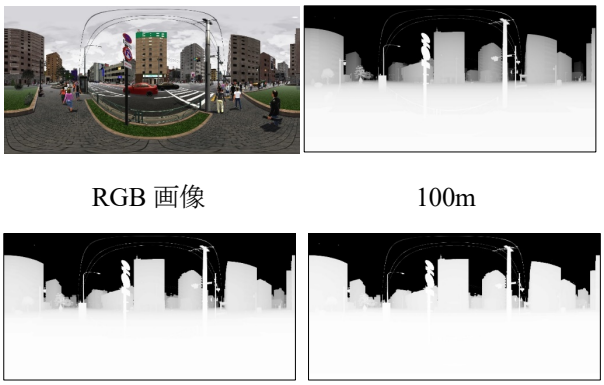


図 3 カットオフ距離の違いによる深度マップ

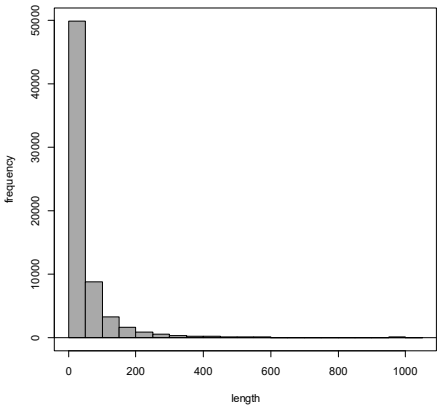


図 4 185 撮影地点の視線の長さのヒストグラム

雑把に把握されつつ，それなりに遠くまで見通せるような認識が自然だと考えられる．視線の長さが対数正規分布に従うことから，この確率密度関数の累積分布関数を，図 5 に示すように縦に 256 倍して画素値に対応させた曲線を描く．次に，この曲線の縦軸の範囲を 256 個の等区間に分割し，累積分布関数を介して，横軸の距離の区間に対応させ，この対応する距離で深度マップを作成する．すなわち，視線の長さの出現確率が同程度になるように，距離の区間を変えるのである．この時のカットオフ距離は約 625m である．ただし Unity のシェーダーで計算する場合，累積分布関数はシェーダーの関数として用意されていなかったため，それを近似する 5 次の多項式で代用した（図 5 の青線）．この方法を以後「非線形（スケール）」と呼ぶことにする．図 6 に二つの距離スケールの違いを示す．非線形のほうが近距離側の解像度が高く, かつ奥の建物まで認識されている．

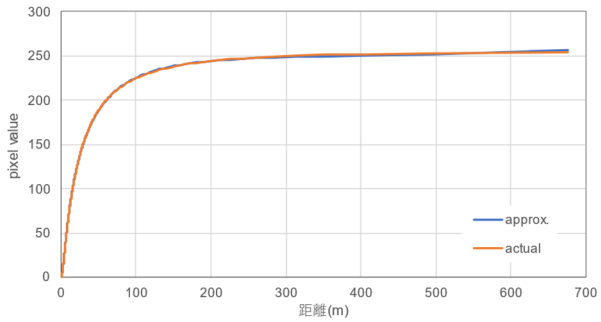
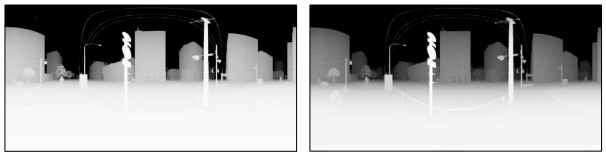


図 5 縦軸を 256 倍した視線の距離分布の累積分布関数と近似曲線

元画像



線形　　非線形

図 6 深度マップの 2 つの距離スケールの比較

3．深度マップの学習

3.1 CG の都市モデルの準備

pix2pixHD により全方位の RGB 画像から深度マップを生成する学習で使用するために，3 次元の都市空間モデルを準備する．既往研究と同様に，日本の都市空間を高いリアリティで再現した渋谷モデルと，郊外都市モデル 2 つの都市モデルを，海外の 3D モデル販売サイト（NoneCG）から購入し，Unity にインポートした（図 7）．元のモデルの範囲は，深度マップを撮影するのに十分な広さではなかったため，各モデルを複数回コピーし，モデルの領域を周囲に拡張した．さらに，RGB 画像は人や車が写り込んだ状態で撮影するために，それらのモデルも複数個を空間内に配置した．加えて，実際の街並みの写真は，天候，季節，時間などによって大きく異なる．特に空の状態は深度マップの生成に大きく影響するため，Tenkoku Dynamic Sky アセットを使用し，晴と曇の 2 つの空の状態を設定した．

渋谷　　郊外

図 7 二つの都市モデルとそれらの撮影地点（上段：全景，中断：配置図全体，下段：撮影地点の拡大図）

3.2 CG 空間内での全方位画像の撮影

次に，CGモデルの全方位画像の撮影を行う．図8に示すように，各都市モデルの配置図の画像をGISに入力し，空間の中心部の道路上で500地点の撮影箇所をランダムに設定した．次に，それらの撮影地点をpix2pixHDの学習，検証，テストの3つに分割した．数はそれぞれ300，100，100で，赤，緑，紫に色分けされている（図7）．隣接する撮影地点までの距離が数mと近く，撮影地点の位置を考慮せずに画像をランダムに分割した場合，各データセットに類似した画像が混在することになり，意味のあるテストにならない可能性があるため，上記のように，空間でデータセットを分割する方式とした．

カメラの高さを2.05mに設定し，全方位画像の撮影のために，かつてUnityのアセットとして販売されていたSpherical Image Camを用いて，各撮影地点について，同一角度で1枚ずつ撮影した．線形，非線形の各深度スケールで，1,024×512pxの正距円筒図法で画像を撮影・記録した．なお，RGB画像は，実写の

状況を考慮して，歩行者と自動車が入った状態で撮影したが，深度マップは空間の状態だけを把握したいため，歩行者と自動車のモデルを除外して撮影した．

3.3 スタイル変換の適用

pix2pixHD を使う目的は，実写の全方位の RGB 画像を入力し，その深度マップを生成することであるが，その学習では CG の全方位の RGB 画像を用いる．したがって，CG の画像の雰囲気を実写画像に近づけて学習を行ったほうが，生成される深度マップの品質が高くなると予想される．そこで，WCT2 というスタイル変換方法を使用して，CG のスタイル変換を行う．WCT2 では変換すべき（CG）画像 1 枚について，参照する 1 枚の（実写）画像を用意する．そして，両画像に対してセマンティック・セグメンテーション（SS）を適用し，同種のオブジェクトに対して同種のスタイルを適用する．本研究では SS として，CityScapes データセット（Cordts et al., 2016）で学習させた DeepLab v3+を用いた．実写画像には 2 章で撮影した 200 枚の画像を使い，それらも SS しておく．各都市モデルの CG 画像も同様に SS しておき，各 CG 画像の構成要素に最も類似した実写画像を 1 枚選んで，WCT2 を適用した．WCT2 では変換方法がいくつか選べるが，option_unpool=cat5 と transfer_at_decoder とした．ただし，空が曇の画像は色味が不自然になったため，晴の CG 画像でのみ変換を行った．

3.4 その他の前処理

実写の全方位画像には撮影者と撮影機材が画像の下部に写り込んでいるが，CG 画像にはそれが無い．この実写画像の状況に合わせるため，実写画像 200 枚から状態の異なる 10 枚を選び，撮影者と撮影機器部分のみをマスク処理で抽出し，各 CG 画像に対して，それら 10 枚からランダムに 1 枚を選んで重ねる処理を行った（図 9 右下）．その処理を行った後，各 CG 画像と対応する深度マップをカメラの鉛直軸に対して 22.5 度ずつ回転させて撮影したのと同等の画像を得るために，正距円筒図法として保存されている全方位画像を縦に 16 等分し，その左端のブロックを右端に異動させる操作を 16 回行い，

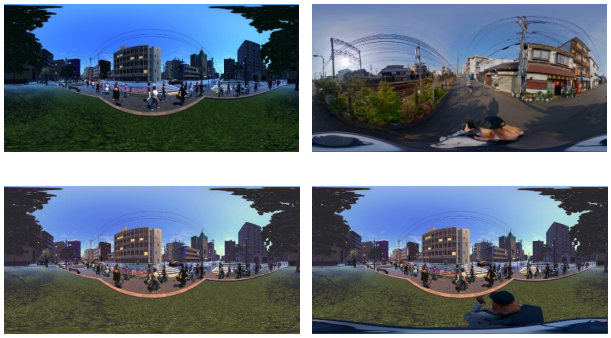


図 8 WCT2 によるスタイル変換例（上段左：元の CG 画像，上段右：スタイル画像，下段左：スタイル変換した CG 画像）と撮影者の重ね合わせ（下段右）

一か所の撮影地点について 16 枚の鉛直軸で回転させた全方位画像に増やした．このようにして，各都市モデルの CG の全方位画像に対して，学習データ 4,800 枚，学習と検証データをそれぞれ 1,600 枚の RGB と深度マップのペア画像を作成した．

3.5 pix2pixHD の学習と精度評価

都市モデル=｛渋谷，地方都市｝，空=｛晴，曇｝，深度スケール=｛線形，非線形｝で合計 2^3=8 通りの組み合わせで，前述した画像のデータセットを構成し，pix2pixHD の学習を行った．pix2pixHD はオリジナルの pytorch 実装を用い，デフォルトから変化させたパラメータは，エポック数=120（100(constant)+20(decay)）に，バッチサイズ=10 の二つである．GPU に GeForce RTX 3090 を 1 枚用いて，10 エポックごとにモデルを保存した．

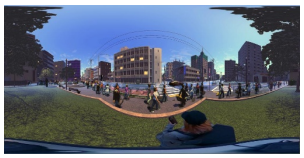
pix2pixHD によって生成された深度マップがどの程度の精度を有するかを評価するために，生成された深度マップと正解の深度マップのピクセルレベルでの誤差を，RMSE によって評価する．ここで，X, Y, Zをそれぞれ順序付けられた入力画像集合，出力（正解）画像集合，ノイズ画像の集合，nをデータセット内の画像枚数，mを 1 枚の画像のピクセル数とする．生成された画像$G(x \in X, z \in Z)$および，対応する出力画像$y \in Y$に関する RMSE は，

$$RMSE(x,y,z) = \sqrt{\frac{1}{m}\|y - G(x,y)\|^2}$$

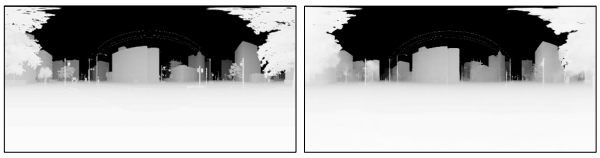
で表される．すべての画像に対する平均 RMSE は

表 1 各データセットの最良モデルの平均 RMSE

データセット	最良エポック	平均 RMSE
渋谷_曇_線形	90	4.43
渋谷_曇_非線形	110	5.19
渋谷_晴_線形	120	4.08
渋谷_晴_非線形	120	4.77
郊外_曇_線形	120	4.43
郊外_曇_非線形	80	3.95
郊外_晴_線形	90	**3.92**
郊外_晴_非線形	100	4.62



元画像



元の深度マップ　　推定深度マップ

図 9 郊外_晴_線形 90 エポックの深度マップの例

$$\overline{RMSE(X,Y,Z)} = \frac{1}{n} \sum_{x \in X, y \in Y, z \in Z} RMSE(x,y,z)$$

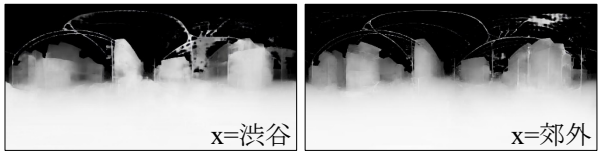
で定義される.

各 pix2pixHD モデルによって検証データで最良の平均 RMSE を示したエポックのモデルで，テストデータの平均 RMSE を求めた結果を表 1 に示す．郊外_晴_線形で学習させたモデルの平均RMSEが最も低かった．そのモデルでの，深度マップの生成例を図 9 に示す．見ての通り視覚的類似性が高いため，pix2pixHD による距離推定は，同じドメインの画像の場合に高い汎化性能を持つと結論付けられる．

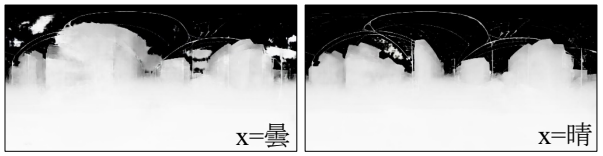
3.6 実写の全方位画像の深度マップの生成と比較

異なる CG のデータセットで学習された pix2pixHD モデルによって，実写の全方位画像から生成した深度マップがどのように異なるかを，図 10 に示す．それぞれ，都市モデル，空，深度スケールそれぞれの違いを比較するために，他の条件は揃えている．まず都市モデルの違いを見ると，表 1 でも

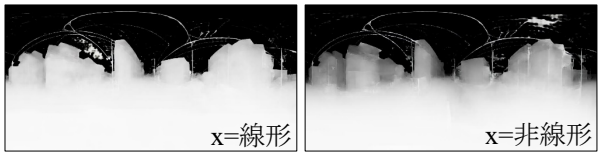
元画像



都市モデルの違い（x_晴_非線形）



空の違い（郊外_x_線形）



深度スケールの違い（郊外_晴_x）

図 10 異なるモデルで生成した深度マップの比較

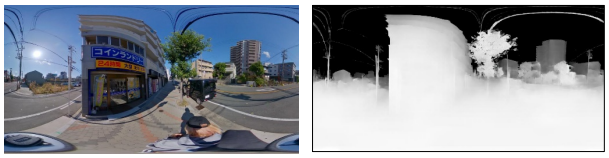


図 11 最終モデルでの実写画像の推定深度マップ例

示された通り，渋谷モデルよりも郊外モデルのほうが建物の輪郭がはっきりしているなど，視覚的に優れた結果が見られた. 次に, 空の違いを比較すると，曇のモデルで空部分の誤推定が目立つ傾向があった. 深度スケールでは，線形は短距離を示す白色部分が比較的建物の奥の領域まで占めていてグラデーションに乏しいが，非線形のほうがグラデーションが豊かで，より望ましいといえる.

以上より実写の深度画像を推定するモデルとして，郊外_晴_非線形モデルを採用した．最終的に実写の RGB 画像から深度マップを推定するために，このモデルで，500 地点の画像すべてを学習データとして pix2pixHD を一から学習しなおし，比較の結果，20

エポック時点でのモデルを最終モデルとした．図 11 に最終モデルによる実写画像の推定深度マップを示す．全体的に，既往研究と比較して，深度マップの精度が視覚的に向上していることが確認できた．

４．印象評価実験

全方位画像をパノラマで投影可能な Google フォトと，回答用に Google フォームを用いて，オンラインで印象評価実験を行った．評価する画像は，先の大阪市住吉区で撮影した 200 枚の画像を用いた．既往研究では場所の好みを 4 段階で評価したが，晴れといった写真自体の質が，好みに影響していることが推測される結果となり，深度画像を適用した意義があまり見られなかった．そこで今回は，積田（1993）の研究を参考にして，空間性に関わる因子として，①立体性因子（平面的－立体的）②統一性因子（ばらばら－統一的）③開放性因子（閉鎖的－開放的）の 3 項目を，それぞれ 4 段階で評価した．

被験者は建築系の学生 20 名とそれ以外の様々な職種の社会人 20 名とした．1 枚の全方位画像を，学生 5+社会人 5 人の合計 10 人で評価するよう，各被験者にランダムに画像を割り当て，かつ割り当てのパタンがすべての画像で異なるようにして画像を提示し，被験者 1 人あたり 50 枚の画像を評価させた．各画像の各因子について 10 人の評価値の平均値をとり，その値の画像 200 枚についての基礎統計を表 2 に示す．中央値と平均値は 2.5~2.9 の範囲にあり，どちらかに偏った因子は見られなかった．

図 12 に，因子それぞれの評価値が低・中・高の画像の例を示す．立体感と開放感は逆の関係があるように思われるが，立体感は建物の軒先など細かな建物の凹凸の影響を受けている傾向がある．一方，統一感が無い場合は，建物の広告など表層的な特徴をとらえている傾向が読み取れる．

表 2 印象評価値の平均値の画像 200 枚の基礎統計

因子	Min	Median	Mean	Max	Std
立体感	1.7	2.9	2.8	3.7	0.42
統一感	1.5	2.5	2.5	3.9	0.45
開放感	1.3	2.8	2.7	3.9	0.58

立体感　1.7（平面的）　2.9　3.7（立体的）

統一感　1.5（ばらばら）　2.9　3.9（統一的）

開放感　1.3（閉鎖的）　2.8　3.9（開放的）

図 12 各因子の評価値が低（左），中（中央），高（右）の画像例

５．印象評価値推定モデルの学習

5.1 各種画像の準備

次章で印象評価値を推定するモデルを CNN で学習させるために，2 章で撮影した 200 地点の実写の全方位 RGB 画像をもとに，推定深度マップ，SS 画像，グレースケール画像を用意する．まず推定深度マップは, 3.6 で示した学習済みモデルを用いて実写画像から生成した．今回の推定深度マップは改良の結果，空部分も比較的良い精度が得られているが，雲が多いとノイズが増える傾向があったため，既往研究と同様に，WCT2 の時と同様の方法で実写画像の SS 画像を生成し，推定深度マップの空部分の画素値を最大距離とするフィルタリングを行った．

さらにこの時生成した SS 画像を新たな入力画像とする．元の SS 画像は RGB 画像として出力されるが，チャンネル数を 1 チャンネルにするために，情報の整理を行う．CityScapes データセットでの空間構成要素は 20 種類あるが，これらの中で，表 3 に示す 11 種類の固定的な空間構成要素のみを使用し, 残りはその他とした．画素値はその他を除き，画像の下部から上部に向かって出現しやすい順に，降順で[0,255]の範囲で，ほぼ等間隔の整数値に設定した．

グレースケール画像は，まず実写の RGB 画像のガンマ補正を解除し，その画像を CIE XYZ 色空間（BT. 709）に変換し，再度ガンマ補正を行ったうえで，Y に対応する重みを画素値とする 1ch の画像として作成した．

最近の CNN には全方位画像に対応したものも提案されているが，筆者らの既往研究（Kinugawa and Takizawa, 2020）により, 一般的な矩形画像用の CNN を使った場合よりも推定精度が低下したため，本研究でも一般的な CNN を用いている．既往研究では正距円筒図法で表された全方位画像を, CNN の入力形式である正方形にリサイズして用いていたが，正距円筒図法には，画像の上下のオブジェクトになるほど，面積が過大になってしまう問題がある．本研究の場合，そこは空や道路などが占めているだけで情報量に乏しい領域といえる．そこで今回は，正積円筒図法の一種であるトブラー図法により，正距円筒図法→球面→正方形平面として画像を変換する．変換を行った一組の入力画像の例を図 14 に示す．最上段は，参考までに正距円筒図法を正方形にリサイズした画像である．トブラー図法の方が目線付近にあるオブジェクトが画像中に占める面積が多い．

最後に，各画像を 224×224px にリサイズし，鉛直軸で 22.5 度ずつ回転させるとともに，それを左右に反転させて画像を増やし，各種の画像について，それぞれ 32 枚×200 地点=6,400 枚の画像を生成した．

表 3 1ch の SS 画像で採用する構成要素と画素値

構成要素	画素値	構成要素	画素値
Road	255	Traffic light	116
Sidewalk	232	Vegetation	93
Fence	209	Wall	70
Terrain	185	Building	46
Pole	162	Sky	23
Traffic sign	139	その他	0

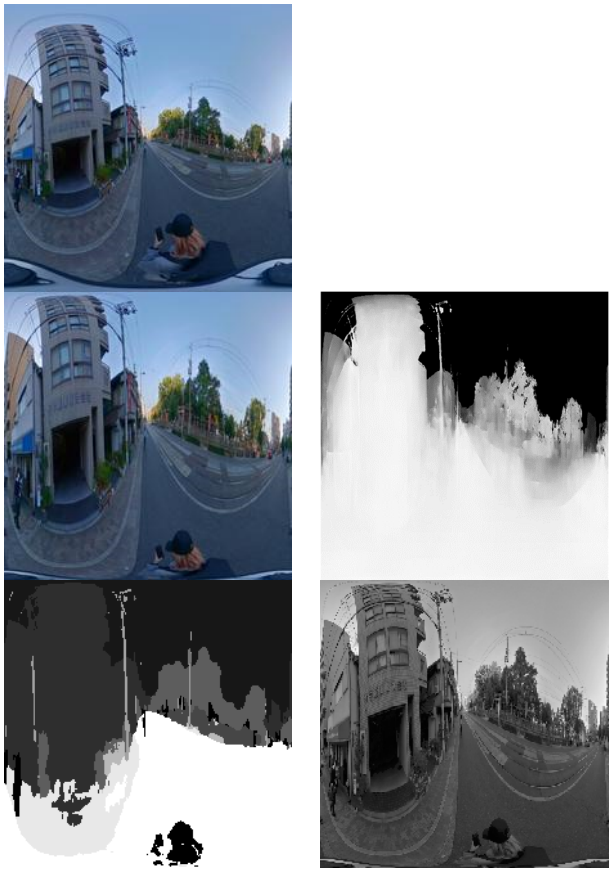


図 14 入力画像セット : RGB 画像（中段左），深度マップ（中段右），SS 画像（下段左），グレースケール画像（下段右），※上段は正距円筒図法のリサイズ

5.2 CNN の設定と学習

準備した画像を用いて，印象評価実験で得られた各因子の印象評価値を予測する CNN モデルを構築する．印象評価値には個人の嗜好等によるばらつきが見られたが，平均的な印象を推定する為，外れ値を削除した．各画像の各因子について，10 人の評価値の外れ値の下側境界を，$Q_{1/4} - IQR \times 1.5$，上側境界を，$Q_{3/4} + IQR \times 1.5$に設定し，この範囲外の値を外れ値として, 平均印象評価値を再計算して用いた．

CNN には，標準的な ResNet-18 と ResNet-50 を使用し，Torchvision の実装で，ImageNet により事前学習した重みを初期値として学習を行う．ただし入力画像のチャンネル数が 1~5ch と変化するので，それらに合わせて入力層を修正している．さらに，印象評価値を推定する回帰問題となるので，損失関数を MSE 誤差に変更している．ResNet の主なハイパーパラメータを以下に示す．総エポック数=50，バッチサイズ=16，最適化手法=Momentum SGD，モーメンタム=0.9，初期学習率=0.001，学習率の低減方法=総エポック数の 1/3 経過毎に 1/2 に逓減. 入力画像は，r，g，b（以上 RGB），d（深度），s（SS），y（グレースケール）の 6 種類の画像チャンネルを，表 4 の組み合わせで用いた．まずは単純な ResNet-18 で学習を行い，そのテスト結果から精度が期待できる組み合わせのみを ResNet-50 で学習させた．

200 地点の画像のうち 160 地点の画像を学習データに，残り 40 地点を 20+20 の地点に分け，それぞれ検証データとテストデータの役割を入れ替え，疑似的な 10-fold の交差検証で精度評価を行った. データを分ける際は，元の 200 枚の画像の印象評価値の分散にできるだけ類似するように，それぞれの地点を各データセットに分割した．

5.3 テスト結果

モデルの精度評価には，平均二乗和誤差（MSE）と決定係数（R2）を用いた．検証データによって，MSE+(1-R2)の値が最も小さいエポックのモデルを使用してテストデータの精度評価を行った．その結果を表 5 に示す．チャンネルの末尾に 50 とついているのは，ResNet-50 の結果である．立体感では d モデルが，統一感では rgbds50 モデルが，開放感では dys モデルが最も優れている．総合的には dys と rgbds50 がベストモデルとなった（図 15）．特に dys モデルは 34 層の ResNet であったが，5ch の 50 層の ResNet モデルと同等の精度を示している．

表 4 CNN に入力する画像チャンネルの組み合わせ

CNN	1ch	2ch	3ch	4ch	5ch
ResNet-18	r g b d s	ds	rgb dys	rgbd rgbs	rgbds
ResNet-50		ds	rgb dys	rgbs	rgbds

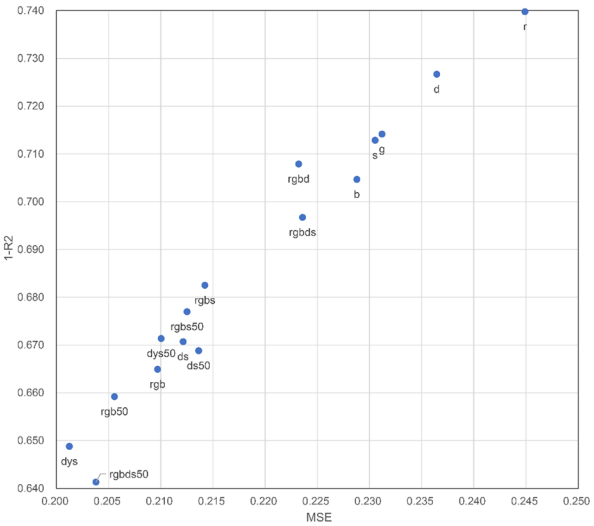


図 15 総合評価の各モデルの MSE と 1-R2 の散布図

６．まとめ

いくつかの改良を加えることで，実写の深度画像から従来よりも良い品質の深度マップを得ることができた．さらに深度マップや SS 画像など，加工した画像を組み合わせて入力データとして，空間性に関連する印象評価値を推定する CNN の学習を行い，汎化性能が高いモデルを学習することができた．

謝辞

本研究は科学研究費補助金基盤 C（20K04872）の補助を受けています．

表 5 各モデルの各印象評価値のテストデータによる平均精度

モデル	立体感		統一感		開放感		総合	
	MSE	R2	MSE	R2	MSE	R2	MSE	R2
r	0.233	0.158	0.228	0.211	0.274	0.412	0.245	0.260
g	0.231	0.124	0.244	0.193	0.218	0.541	0.231	0.286
b	0.235	0.138	0.243	0.214	0.208	0.533	0.229	0.295
d	0.206	**0.198**	0.226	0.238	0.278	0.384	0.236	0.273
s	0.229	0.130	0.248	0.185	0.215	0.546	0.231	0.287
ds	0.220	0.150	0.220	0.261	0.196	0.577	0.212	0.329
rgb	**0.202**	0.183	0.213	0.265	0.213	0.557	0.210	0.335
rgbd	0.222	0.129	0.220	0.233	0.228	0.515	0.223	0.292
rgbs	0.223	0.144	0.218	0.246	0.202	0.562	0.214	0.317
rgbds	0.247	0.106	0.225	0.229	0.198	0.575	0.224	0.303
dys	0.208	0.174	0.217	0.259	**0.178**	**0.621**	**0.201**	0.351
ds50	0.212	0.188	0.227	0.248	0.202	0.557	0.214	0.331
rgb50	0.208	0.173	0.218	0.260	0.190	0.589	0.206	0.341
rgbs50	0.229	0.133	0.217	0.258	0.192	0.578	0.213	0.323
rgbds50	0.217	0.169	**0.202**	**0.311**	0.192	0.596	0.204	**0.359**
dys50	0.218	0.152	0.223	0.243	0.189	0.591	0.210	0.329

参考文献

Liu L., Silva E.A., Wu C. and Wang H. (2017) A machine learning-based method for the large-scale evaluation of the urban environment. Computers, Environment and Urban Systems 65, 113-125.

Hillier B. and Hanson J. (1989) The Social Logic of Space. Cambridge University Press.

Benedikt M. (1979) To take hold of space: isovists and isovist fields. Environment and Planning B 6, 47-65.

Batty M. (2001) Exploring isovist fields: space and shape in architectural and urban morphology. Environment and Planning B: Planning and Design 28, 123-150.

Takizawa A. and Furuta A. (2017) 3D spatial analysis method with first-person viewpoint by deep convolutional neural network with omnidirectional RGB and depth images. The 35th Education and research in Computer Aided Architectural Design in Europe, 693-702

Takizawa A. and Kinugawa H. (2020). Deep learning model to reconstruct 3D cityscapes by generating depth maps from omnidirectional images and its application to visual preference prediction. Design Science, 6, E28.

Jaejun Y., Youngjung U., Sanghyuk C., Byeongkyu K and Jung-Woo H. (2019) Photorealistic Style Transfer via Wavelet Transforms, International Conference on Computer Vision.

Chen L.C., Zhu Y., Papandreou G., Schroff F. and Adam H. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. The 15th European Conference on Computer Vision, 833-851.

Ting-Chun W., Ming-Yu L., Jun-Yan Z., Andrew T., Jan K. and Bryan C. (2018) High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. arXiv 1711.11585.

He K., Zhang X., Ren S. and Sun J. (2016) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

大阪市 (2020)【R01 年度】大阪市地形図（構造化データ _ESRI Shapefile ）, https://www.geospatial.jp/ckan/dataset/r01-esri-shapefile.

Cordts M. et al. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.

積田洋 (1993) 都市的オープンスペースの空間意識と物理的構成との相関に関する研究. 日本建築学会計画系論文報告集, 451, 145-154.

* 正会員 大阪市立大学生活科学研究科（Osaka City University）
〒558-8585 大阪府大阪市住吉区杉本 3-3-138 E-mail：takizawa@osaka-cu.ac.jp

** 非会員 大阪市立大学生活科学研究科（Osaka City University）