全方位画像から生成した深度マップを用いた 3D 都市景観を再構築する深層学習モデルと視覚的嗜好予測への応用

Deep Learning Model to Reconstruct 3D Cityscapes by Generating Depth Maps from Omnidirectional Images and Its Application to Visual Preference Prediction

○衣川　雛[*1]，瀧澤　重志[*2]
Hina Kinugawa[*1] and Atsushi Takizawa[*2]

*1 大阪市立大学大学院　生活科学研究科　前期博士課程
Master's Student, Graduate School of Human Life Science, Osaka City University
*2 大阪市立大学大学院　生活科学研究科　教授　博士(工学)
Professor, Graduate School of Human Life Science, Osaka City University, Ph.D.

キーワード：pix2pix; ResNet; UGSCNN; 全方位画像; 深度マップ; 都市景観
Keywords: pix2pix; ResNet; UGSCNN; omnidirectional image; depth map; cityscape

1. はじめに

本研究では，複数のディープラーニング手法を用いて，一般的な RGB の画像情報に加えて，空間の奥行に対応する幾何情報を深度マップとして表現することで，両者を同じ画像解析の枠組みで扱うことができる，一人称視点の新しい空間モデリング・解析手法を開発・検証するものである．今回報告するのは，以前の Kinugawa らの研究 [1]から発展したもので，以前の研究との主な違いは，全方位画像および一般的な長方形の画像に対応する CNN の導入，生成された深度マップの精度評価，および景観の嗜好予測問題をランク学習から分類問題へと変更したことである．これらの変更により，開発しつつある方法を発展させ，より厳密に検証することを行う．

2. 提案方法

提案方法のフレームワークを Fig.1 に示す．フレームワークは大きく 2 つの部分に分かれている．はじめに，CG ベースの都市空間モデルを用意する．全方位画像と深度マップを仮想空間で捉え，ペアとなる画像を収集した．次に pix2pix[2]を使用して，全方位画像から深度マップを生成するモデルを構築する．その後，大阪府内におけるストリートビュー（SV）の都市景観画像の嗜好評価実験を行う．実験に使用した SV の全方位画像ごとに前のパートで学習した pix2pix モデルを使用して深度マップが生成され，セマンティック・セグメンテーション（SS）[3]によって，空部分のノイズをフィルタリングした後，4 チャネル RGBD 画像が生成される．最後に，深度マップの有用性を SV 画像の嗜好を予測する分類モデルの精度によって検証する．

以下では前回の研究と異なる部分を中心に述べる．

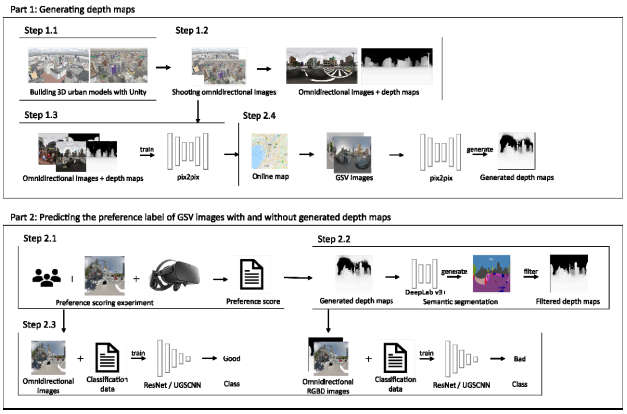


Figure 1. Framework of the proposed method.

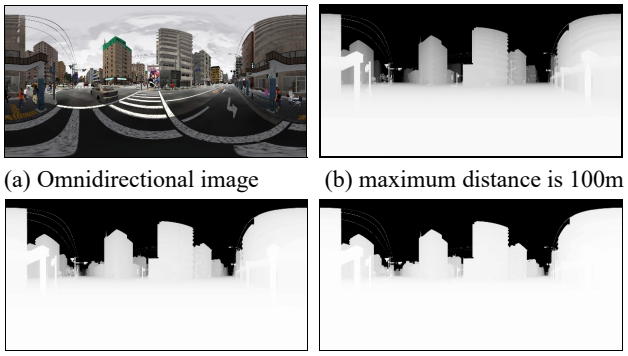
Step1.1: 3D 都市空間の構築

前回の研究と同様に，ゲームエンジン Unity を使用して対象都市の 3 次元モデルを構築・利用する．

Step1.2: 全方位画像の撮影

前回の研究から発展し，pix2pix での学習のための全方位画像のデータセットを，トレーニング・検証・テストセットに分割する際に，撮影地点の空間的自己相関を考慮してそれらのデータセットを地理的に区分した．撮影地点の数はそれぞれ 300，100，100 で，Fig.2 に示すように赤，緑，紫に色分けされている．また，GPU の深度バッファから深度マップを生成するには，深度を測定する範囲と，使用する距離変換関数を決定する必要がある．人間の感覚スケールはしばしば対数スケールで近似されているが，対数関数はいくつかのパラメーターを決定する必要があるため，単純に距離に関する線形関数を仮定する．Fig.3 に示すように，距離の最大値が 100，250，および 500m の深度マップ

Figure 2. Shooting points in each city model.



(a) Omnidirectional image (b) maximum distance is 100m

(c) maximum distance is 250m (d) maximum distance is 500m

Figure 3. Comparison of shades of depth maps.

を生成した．各最大値の 1 画素値あたりの深度は，それぞれ 0.39，0.98，1.96 m である．これらの数値を比較すると，近くのオブジェクトの解像度と遠くのオブジェクトの認識範囲に違いがあり，両者の関係はトレードオフとなる．本研究で扱う空間は主に市街地であり対象物は比較的短距離に集中する傾向がある．そこで今回は最大距離を 100m に設定して距離画像を生成する．

Step1.3: pix2pix

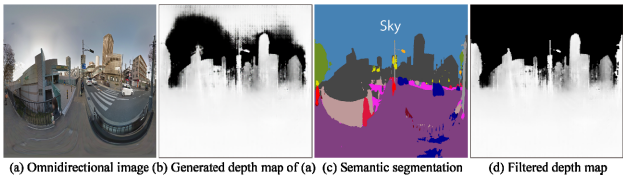
pix2pix を使用して，全方位画像を，各撮影地点で 18 度ずつカメラを回転させて 20 倍の枚数に増量したうえで，深度マップを生成するようトレーニングする．前回の研究と同様に 3 種類の空間データセット，3 種類の天候条件を組み合わせた Table1 に示す 9 つのデータセットを用意した．生成された画像の正しい画像のピクセルレベルでの誤差を，二乗平均平方根誤差（RMSE）によって評価する．

Step1.4: pix2pix を用いた深度マップの生成

合計 100 枚の SV の全方位画像を使用して，深度マップの生成と嗜好評価実験を行う．SV 画像は，前回同様にローカルエリア（寝屋川と住吉）と都市エリア（梅田と難波）の両方から，それぞれ 50 枚ずつサンプリングしたものを使用する．これらの SV 画像を pix2pix モデルに入力し，生成された深度マップを取得する．SV の深度マップはダウンロード可能だが役に立たない精度のため，生成された SV の深度マップを定量的に評価できない．そこで，生成された深度マップの精度を，筆者が視覚的に評価することにした．

Table 1. Nine models used for training.

Spatial model	Sky condition		
	a. Blue	b. Cloudy	c. Mix of a and b
1.Shibuya	M1a	M1b	M1c
2.Local city	M2a	M2b	M2c
3.Mix of 1 and 2	M3a	M3b	M3c



(a) Omnidirectional image (b) Generated depth map of (a) (c) Semantic segmentation (d) Filtered depth map

Figure 4. Filtering operation using semantic segmentation.

Step2.1: 嗜好評価実験

前回の研究と同様に｛よい ＝4，やや良い ＝3，やや悪い ＝2，悪い ＝1｝の 4 段階で嗜好評価実験を行う．被験者は，建築学を専攻する 20 人の大学生である．

Step2.2: SV 画像の RGBD 画像の作成

前回の研究と同様に，SS を使用して深度マップの空の部分のノイズをフィルタリングする（Fig.4）．そして，RGBA 形式画像の A チャンネルに深度マップの値を保存し，RGBD 画像を作成する．

Step2.3: 長方形及び全方位画像の CNN を利用して主観的な嗜好を推定する分類モデルの設定

ここまでで得られた，SV の RGB および RGBD の全方位画像を使用して，嗜好評価実験で得られた嗜好の平均値を予測する CNN モデルを構築する．前回の研究ではランク学習を使ったが，学習が相対的で不安定になりやすかったので，よりシンプルなモデルとして，回帰モデルをまずは検討した．しかし，学習したモデルの出力値が嗜好の平均値の周りに集まる傾向があり，景観のよしあしを評価するモデルとしては不十分であった．したがって今回は，2 クラスの分類問題の枠組みで嗜好をモデル化する．

分類モデルとして，一般的な矩形の画像を対象とした ResNet[4]に加え，全方位画像が本来は 360 度の空間を表現していることから，球面 CNN も導入する．球面 CNN として良好なパフォーマンスを示す UGSCNN[5]を使用し，それらの精度を比較する．ResNet は ImageNet で事前トレーニングされた ResNet-50 の出力層と入力層を問題に合わせて変更して使用する．すなわち，RGBD 画像には 4 つのチャネルがあるため，画像入力レイヤーのチャンネル数を 3 から 4 に変更する．また，クラス分類モデルに対応させるため最終層の出力ノードを 1000 から 2 つに変更する．

一方 UGSCNN は，Fig.5 に示すように正二十面体から細分割された多面体に基づく球状 CNN であり，畳み込みを隣接する頂点間で行って，細分割された立体をもとの正 20 面体図形に戻していく．今回レイヤ設定は独自に行った．

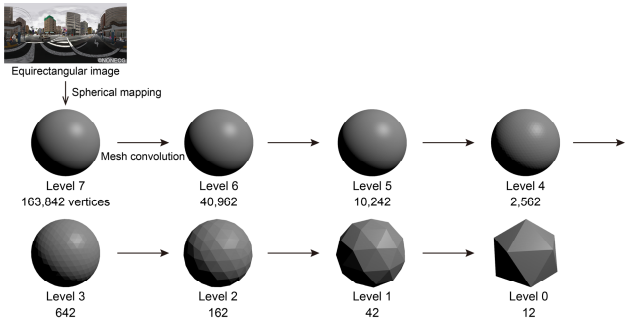


Figure 5. Mesh convolution of UGSCNN.

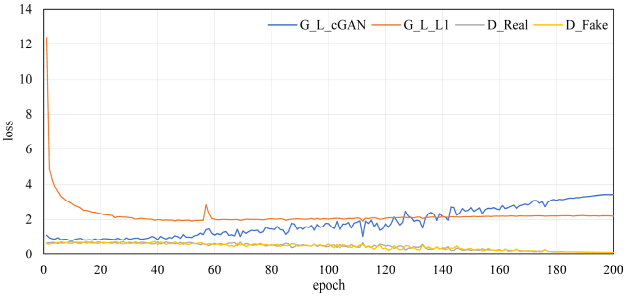


Figure 6. Example of Convergence process of loss functions of M2c.

3. 結果

3.1. pix2pix の学習結果 (Step1.3)

トレーニングと検証には、pix2pix の PyTorch 実装[6]を使用した．学習速度を向上させるために，バッチサイズを 50 に増やし，さらに，CG と実際の写真の異なるドメインで学習と推論を行うために，正規化方法をバッチ正規化からインスタンス正規化に変更して学習を行った．

Fig.6 は M2c モデルを 200 エポックまで学習した時の損失関数の変化を示している．D_Real と D_Fake はそれぞれ，実際の画像と生成された画像が入力されたときの弁別器のクロスエントロピー損失で，エポックが進むにつれて，弁別器の損失はほぼ単調に減少する．一方，生成器の L1 損失は約 50 エポックから一定の値をとる．また，エポックが進むにつれ GAN 損失は増加しているが，これは GAN の学習プロセスの一般的な傾向である．

Table 2 に，各 pix2pix によって生成されたテストデータ（各モデルとフルセットの M3c）の RMSE を示す．平均誤差は画素値で約 5，距離に換算すると約 2m であった．生成されたテストデータの深度マップの例を Fig.7 に示す．誤差が比較的低く視覚的類似性が高いため，pix2pix による距離推定は，同じドメインの画像の場合には高い汎化性能を持つと結論付けられる．

3.2. SV 画像を用いた pix2pix の学習結果 (Step1.4)

各モデルによって生成された SV 画像の深度マップの比較を Fig.8 に示す．SV の 100 画像を通して視覚的に自然な深度マップが得られたのは，M2c の場合であったので，嗜好評価実験では，このモデルの結果を利用した．

3.3.嗜好評価実験の結果 (Step2.1)

1 枚の SV 画像は 10 人の被験者で評価されるので，その平均スコアを，Step2.1 で説明した分類モデルのクラスラベルを付与するために使用した．全画像を通じた平均スコアの中央値と平均値は，それぞれ 2.4 と 2.45，標準偏差は 0.74 であった．受験者の好みは比較的多様であったが，スコアが高い画像は，建物の圧迫感が少なく，広範囲の青空や緑が含まれる画像が多い傾向があった．一方，スコアが低い画像では，アスファルトや住宅が目立つ画像が多かった．

Table 2. RMSE of test data generated by each pix2pix model.

Model	Best epoch at validation	Test data of each model		Test data of M3c	
		Mean	Std	Mean	Std
M1a	70	3.44	0.47	6.13	2.22
M1b	20	3.69	0.56	6.20	1.58
M1c	40	**3.38**	0.56	4.47	1.36
M2a	70	4.37	**0.31**	6.36	2.14
M2b	80	4.68	0.41	6.40	1.10
M2c	80	4.40	0.49	5.04	1.34
M3a	100	3.93	0.68	5.67	1.88
M3b	70	4.18	0.64	6.01	1.90
M3c	50	4.22	0.71	**4.23**	**0.71**

Figure 7. Example of a generated depth map of a test data by M2c, RMSE=4.37 (Center: original, right: generated).

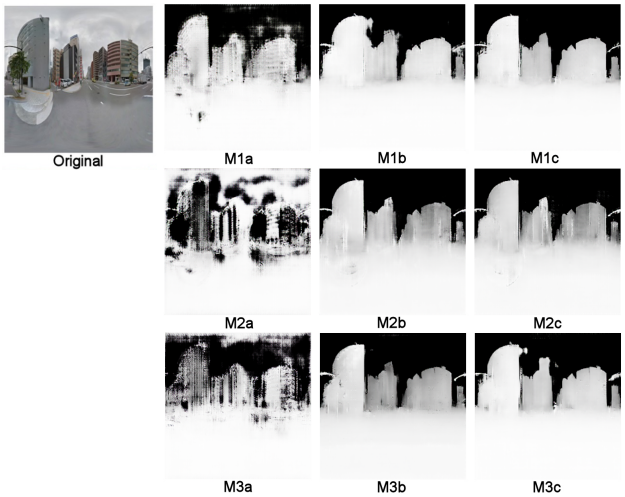


Figure 8. Comparison of depth maps of the same SV image generated by each model.

3.4 分類モデルの結果 (Step2.3)

各画像の平均スコアが 2.5 以上/未満で，それぞれ Good/Bad のクラスラベルを付けた．その結果，100 枚の SV の画像のうち 46 枚に Good を，残りの 54 枚の画像に Bad のラベルが付けられた．また，学習に用いた損失関数の値では分類性能がわかりにくいため, F1 スコアを用いて最終的な精度評価を行った．100 枚の SV 画像を前回の研究と同様に，回転と鏡像操作により 40 倍の 4000 枚に増量して，学習データを構築した．もとの 100 枚の画像に対して，学習（80 枚），検証（10），テスト（10）にそれらをランダムに分割し，それを増量した画像にも適用する操作を 5 回行った．さらに，検証とテストデータを入れ替え，最終的に，10 分割交差検証用の画像データを作成した．

Fig.9 に 10 分割交差検証による各 CNN のテストデータの F1 スコアの分布を，Table3 に F1 スコアの基礎統計値，Table4 にすべての CNN および各タイプの CNN の平均分析法の決定限界を示す. ResNet-50 は UGSCNN よりも精度が高く，すべてのモデルを比較すると RGBD を使用した ResNet-50 の精度が最も高く，統計的に有意な差がみられた．一方各 CNN で見た場合，RGBD の画像で学習したモデルの平均値と中央値は RGB の場合よりも高くなっているが，統計的な有意差はみられなかった．

4. おわりに

本研究では，最初に pix2pix による深度マップの生成において，CG 画像を生成した場合の奥行きを定量的した．その平均誤差は 1 画像あたり約 2m であり，都市スケールの空間分析という意味では誤差は許容範囲と考えられる．一方 SV 画像から生成された深度マップは目視で評価を行い，空の部分のノイズを無視しても CG 画像に比べて精度が劣ることが確認された．しかし，対象物の近距離や遠距離などの空間の質的傾向は把握できた．

次に，SV 画像の嗜好の分類モデルの結果は，RGBD 画像を使用した ResNet-50 によって最高の分類精度が達成され，全方向画像に対する UGSCNN の精度はどちらの画像でも悪かった．UGSCNN のレイヤ設定の検討が十分ではなかったことや，適切に事前学習されたモデルが用意されていないことが，その原因と考えている．各モデルで見た場合，深度マップの導入効果は統計的に有意とまでは言えなかったが，精度の上昇は見られ，街路景観の視覚的な嗜好性をモデル化するために，深度マップを考慮する必要性が示唆されたと結論付ける．

今後は異なるドメイン間での深度マップの生成精度の向上や，球面 CNN のレイヤ設定の検討が課題である．

[謝辞]

本研究の一部は科学研究費(A)(C)の補助を受けています.

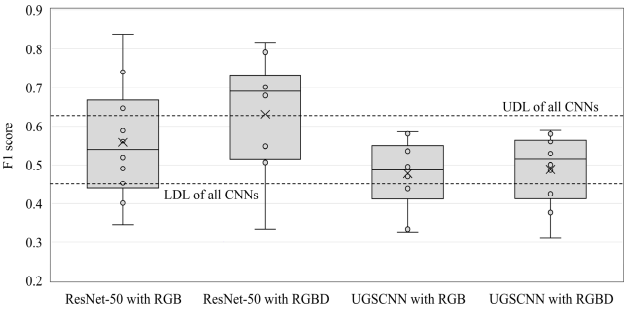


Figure 9. Distribution of F1 score of 10-fold cross validation for each CNN. X denotes mean.

Table 3. Descriptive statistics of F1 score of 10-fold cross validation for each CNN.

CNN	N	Min	Mean	Median	Max	Std
ResNet-50 with RGB	10	**0.344**	0.558	0.539	**0.837**	0.143
ResNet-50 with RGBD	10	0.333	**0.631**	**0.691**	0.817	0.143
UGSCNN with RGB	10	0.327	0.478	0.487	0.586	**0.086**
UGSCNN with RGBD	10	0.310	0.488	0.515	0.589	0.087

Table 4. Decision limit of analysis of means of F1 score for sets of CNNs, significance level = 0.05.

CNN	Lower decision limit (LDL)	Mean	Upper decision limit (UDL)
All CNNs	0.450	0.534	0.627
ResNet-50	**0.523**	**0.594**	**0.665**
UGSCNN	0.440	0.483	0.526

[参考文献]

1) Kinugawa H. and Takizawa A. (2019) Deep learning model for predicting preference of space by estimating the depth information of space using omnidirectional images, Proceedings of the 37th eCAADe and 23rd SIGraDi Conference, 2, 61-68.

2) Isola P., et al. (2017) Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 5967-5976.

3) Chen L.C., et al. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. The 15th European Conference on Computer Vision, 833-851.

4) He K., et al. (2016) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

5) Jiang C. et al. (2019) Spherical CNNs on unstructured grids, International Conference on Learning Representations 2019

6) Junyanz, CycleGAN and pix2pix in PyTorch. GitHub: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix [参照日 2020.9.30]