Vol.2021-NL-248 No.24
Vol.2021-CVIM-226 No.24
2021/5/21

写真の自動立体変換

長谷川　浩太郎[1,a)]　延原 章平[1,2]　西野 恒[1]

概要：本研究では一枚の写真の景色の三次元形状を復元する手法について提案する．写真一枚からの形状復元という難しい問題を，飛び出す絵本のように，平面からなる地面に対して看板状に物体を立てる簡易的な復元に置き換えることで解決する．この簡易的な復元では，写真から地面と物体を検出する必要があるが，従来の手法では地面と物体の検出を同時に行っており，人や車など，写真の大きさに対して相対的に小さな物体の検出が上手く行えなかった．本研究では，地面と物体の検出をそれぞれに特化させた既存の学習済みネットワークで行うことで検出の精度を高め，既存手法より頑健で正確な三次元形状復元を行う．

1. 写真の光景の形状復元

カメラが発明されてから写真は様々な形で人々の生活に貢献してきた．スマートデバイスが広く普及した現代では，写真をとる行為はさらに身近なものになり，生活の中に溶け込んでいる．写真は過去のある瞬間を記録することができ，その価値は鑑賞することで発揮される．我々は写真を鑑賞することで記録された過去を体験し，頭の中でその情景に飛び込むことができる．しかし，写真は空間的情報を二次元に落とし込んだものであり，写真により提供される体験はその空間的情報を経験や知識から補ったものに過ぎない．

写真鑑賞による体験を実際の体験に近づけるために，写真に写された光景（シーン）を立体的に復元する研究がコンピュータビジョン分野で行われてきた．写真に写る光景を三次元化することができれば，写真の鑑賞を，実際にその光景に自分がいるかのような体験へと変えることができる．画像から三次元形状を復元する手法として，視点の異なる画像や，光源の異なる画像を用いる手法などがあるが，付加的な情報を得ることは通常の写真では難しい．鑑賞のために身近な写真に対して復元を行うならば，一枚の写真から立体的にシーンを復元できることが望ましい．Horry らは，ユーザーがシーンの情報を手動で追加することで，一枚の画像や絵画から立体的なモデルを生成する手法を提案した [1]．Hoiem らは，複数画素領域（スーパーピクセル）に分割した画像から空，垂直成分，地面の三つの領域を検出し，単一画像から自動で立体モデルを生成する手法を提案した [2]．しかし，前者は，高い精度で立体モデルを生成できる一方で，情報を手動で追加する手間がかかり，後者は，基盤となるスーパーピクセルでの領域分割の精度が低く，人などの物体が多く写る画像ではモデルの形状が大きく歪んでしまうなどの問題点があった．

本研究では，画像の領域分割を物体の検出問題と地面の検出問題の二つに分割することで，物体が多く存在するシーンに対しても頑健に立体モデルを生成する手法を提案する．本手法ではシーンの形状を詳細には復元せず，地面を平面として復元し物体を看板状に立てることで復元を行う．地面の推定と復元，物体の検出と復元はそれぞれ個別に行う．物体の復元も各物体に対して個別に行われるため，いずれかの復元に誤りがあっても，他の物体には影響せず，シーンの復元を頑健に行うことができる．

2. 関連研究

画像からの三次元形状復元，新たな視点からの画像生成はコンピュータビジョンの主要な研究テーマであり，多くの研究が存在する．

2.1 画像からの三次元形状復元

異なるカメラで対象を撮影するときにカメラ間の位置関係が分かっている場合は，一方の画像中のピクセルの，他方の画像における存在範囲はエピポーラ線上に制限される．そのため，二画像間での対応点は，エピポーラ線上を探索することにより得られる．対応点が分かれば，カメラの位置関係およびカメラの内部パラメータと合わせ，三角測量を用いて対象の三次元位置を復元することができる．これはステレオビジョンとして知られており，画像から三次元形状を復元する手段として広く用いられている．Structure from motion（SfM）は，画像間における対応点からカメラ

[1] 京都大学大学院 情報学研究科
[2] JST さきがけ
[a)] khasegawa@vision.ist.i.kyoto-u.ac.jp

Vol.2021-NL-248 No.24
Vol.2021-CVIM-226 No.24
2021/5/21

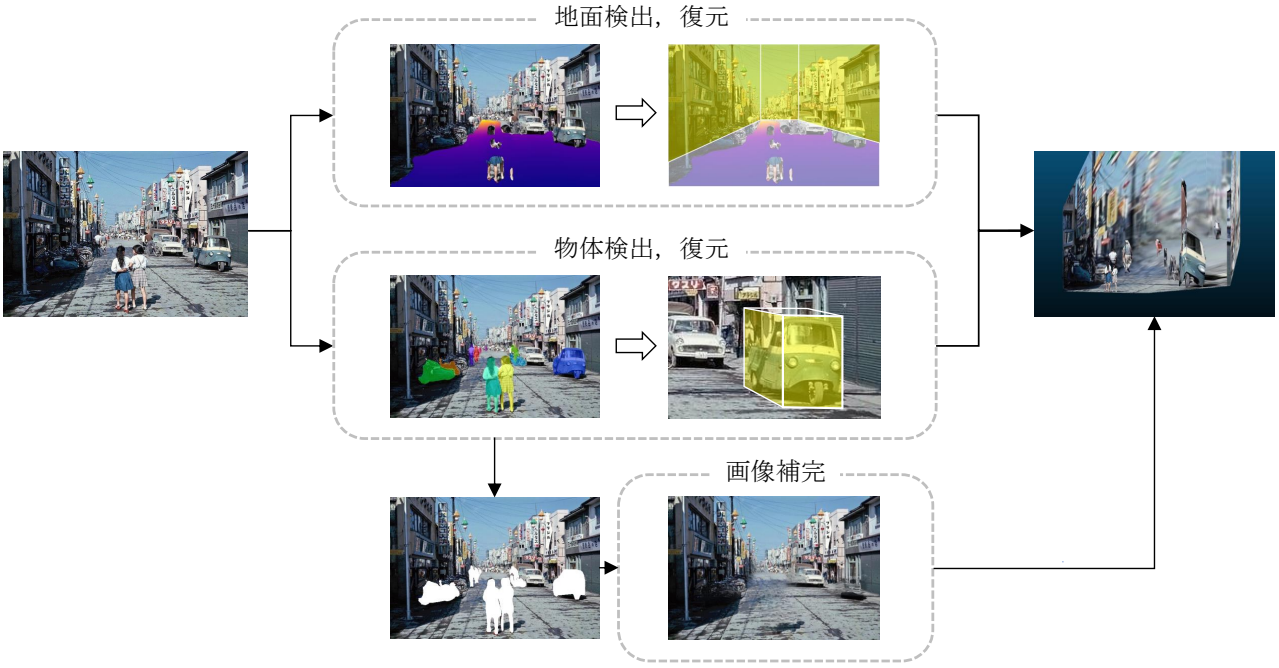

図 1 写真の自動立体変換の手順．画像は地面検出を行うネットワークと物体検出を行うネットワークに入力され，それぞれ独立に復元された後に合成される．物体検出ネットワークの出力を元にして，画像補完が行われる．

の位置関係と，対象の三次元位置を求める手法である．画像間での対応点が少なくとも五点必要であるが，カメラ間の位置関係が分からなくても三次元位置を復元することができる．

これらの手法を用いることで，画像からその三次元形状を復元することができるが，一枚の画像に対して適用することが難しい．

2.2 新視点画像生成

画像列からその中にはない、新たな視点における画像を生成する問題は盛んに研究されており，Novel View Synthesis（NVS）と呼ばれている．Martin らは，様々な視点の画像が存在する物体について，カメラパラメータが不明の多視点の画像からでも，深層学習を用いて新たな視点からの画像を生成する手法を提案した [3]．この手法は非常に高い精度で三次元形状を復元できるが，その対象が多くの写真が存在しているものに限られる．

一つのシーンに対して異なる視点の画像が複数存在する場合は，シーンに対する情報は多く，新たな視点の画像をレンダリングすることは容易になる．しかし，日常に存在する画像に写るシーンに対しては，異なる視点で画像が存在しないことも珍しくない．この問題を解決するために，一枚の画像から新たな視点の画像を生成する研究も存在する．Horry らは，一枚の写真や絵画を入力として任意視点からの画像を生成する手法を提案したが，そのためにユーザーは消失点の位置や，前景の位置を個別に与える必要がある [1]．Hoiem らは，一枚の写真のピクセルを類似した特徴をもつ小集合（スーパーピクセル）に分割，さらにそれを地面，垂直成分，空の三つの領域に分類し，地面に対して垂直成分を看板のように立てることによって，飛び出す絵本のような立体モデルを自動で生成する手法を提案した [2]．しかし，スーパーピクセルによる領域の分割は精度が低く，不自然なモデルが生成されることも多い．Hu らは，画像の特徴量をもとにグリッドをシーンの形状に合わせて変形することで，メッシュを生成し，異なる視点での真の画像との差を損失として学習させることで，一枚の写真から高い精度で新たな視点の画像を生成することを可能にした [4]．

これらの手法に対して我々は，一枚の画像を与えるだけで，自動的に，物体が多く写るシーンからでも頑健に立体モデルを生成する手法について提案する．本手法では，画像の領域分割を物体と地面の検出問題の二つに分割し個別に復元することで，頑健なシーンの形状復元を行う．

3. 提案手法

本研究では，画像の領域分割を物体と地面の検出問題の 2 つに分けて行うことで，物体が多く写るシーンからでも頑健に，飛び出す絵本のような立体モデルを生成する手法

図 2 画像のインスタンスセグメンテーション結果．インスタンスセグメンテーションでは画像中の物体が個別に検出され，それぞれに対して物体の領域を表すマスクが出力される．

を提案する．詳細に三次元形状を復元するのではなく，地面を平面と仮定し，そこに物体を垂直に立てる単純なモデルにすることで，形状復元を頑健に行う．

飛び出す絵本では，平面で表現される地面に対して，物体が看板状に立てられている．画像から地面領域とその深度を推定することができれば，透視投影変換で地面形状を復元することができる．画像中の物体を検出し地面に対する位置を推定できれば，復元された地面に物体を看板状に立てることでに形状復元を簡略化することができる．背景領域は地面領域と物体領域の補集合として得られ，復元された地面領域の縁に対して垂直に立てることで，モデルの見た目を実際のシーンに近づけることができる．手法の概要を図 1 に示す．

3.1 前景抽出

本研究のモデルでは画像中の前景を検出する必要がある。Hoiem らは，画像のスーパーピクセルを，空，垂直成分，地面の三つの領域にグループ分けしていくことで前景を垂直成分として抽出した [2]．しかし，スーパーピクセルのグループ化で得られる垂直成分領域は，実際の垂直成分領域と異なっていることが多い．特に，人や車などの小さな物体は検出されず，生成されたモデルが不自然になることがある．近年では畳み込みニューラルネットワークを用いた高精度の物体検出の手法が存在しているため，前景抽出をより高精度に行うことができる．

画像における物体検出問題はいくつか存在するが，本研究では既存のインスタンスセグメンテーション手法を用いて前景抽出を行う．インスタンスセグメンテーションとは画像中の物体をピクセルレベルで個別に検出する問題である．画像に対するインスタンスセグメンテーションの様子を図 2 に示す．

インスタンスセグメンテーションの手法としては深層学習を用いた手法が主流である．本研究では He らの手法を前景抽出に用いる．He らは物体検出の手法である Ren らのネットワーク [5] のバウンディングボックスと，クラスラベルを予測するブランチに並行して，インスタンスマスクを予測するブランチを追加することで，インスタンスセグメンテーションを実現させた [6]．Ren らのネットワークでは，入力画像と特徴マップの間にズレが生じる．物体検出を目的としているこのネットワークでは，バウンディングボックスが数ピクセルずれていたとしてもその出力の妥当性には影響を及ぼさない．しかし，ピクセル単位で物体のマスクを生成するインスタンスセグメンテーションにおいて，入力画像と特徴マップのズレは推定結果に大きな影響を及ぼす．He らは，元のピクセルと特徴マップを対応させることで，物体検出のネットワークをもとに，物体のマスクを生成することを可能にした．インスタンスセグメンテーションでは，セマンティックセグメンテーションを行ってから物体ごとにマスクを分割するという手法が一般的であった．一方で，He らは，物体検出を先に行い，それに対してマスクを生成するというアプローチをとることで，精度の高い検出を可能にした．

本研究では物体のインスタンスマスクを検出することができればその手法に関わらずモデルを生成することができる．より精度の高いインスタンスセグメンテーションの手法が提案されたときに，本研究の前景抽出の手法を差し替えるだけで，より精度高い形状復元が行えるようになる．

3.2 地面復元

画像中の地面領域のピクセルとその深度を得れば，透視投影変換によって地面領域の三次元形状を復元することができる．深度の真値は単一 RGB 画像を入力とする本研究では得ることができない．しかしながら，単一 RGB 画像からの深度推定はコンピュータビジョン分野でも広く研究されており様々な手法が存在している．Watson らは，動画により物体の別の視点からの情報を与えてネットワークを学習させることで，画像中で直接視認できる地面領域だけではなく，物体によって隠れている地面領域も推定できるモデルを提案した [7]．このモデルでは，地面領域だけでなく，その深度も同時に推定される．本研究における地面の形状復元はこのモデルの出力を用いることで行う．

透視投影変換は, 任意の空間座標系であるワールド座標系上の点と, 平面座標系である画像座標系上の点を, 対応させる変換である．透視投影変換は複数の変換の組み合わせで記述される．透視投影モデルを図 3 に示す．

カメラを原点，z 軸を光軸方向に撮った座標系をカメラ座標系という．ある点のワールド座標系での座標を $\boldsymbol{X}_{\mathrm{w}} = (X_w, Y_w, Z_w)$，カメラ座標系上での座標を $\boldsymbol{X} = (X, Y, Z)$ とすればそれらの関係は，回転行列 $\boldsymbol{R}$ と平行移動ベクトル $\boldsymbol{t}$ を用いて

Vol.2021-NL-248 No.24
Vol.2021-CVIM-226 No.24
2021/5/21

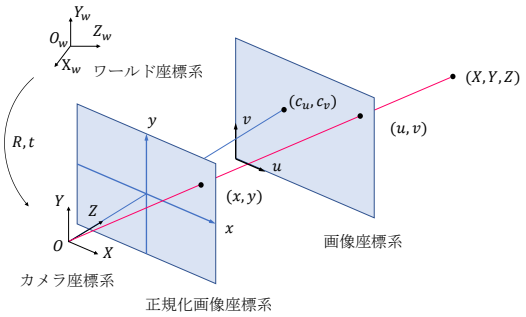

図 3 透視投影モデルの全体像．透視投影モデルによりワールド座標系とカメラ座標系の関係が記述される．

$$\boldsymbol{X} = \boldsymbol{R}\boldsymbol{X}_{\mathrm{w}} + \boldsymbol{t} \tag{1}$$

と記述できる．

画像上の位置を表す座標系として画像座標系と正規化画像座標系がある．画像座標系は画像の位置を表す一般的な座標系で多くの場合角を原点としている．正規化座標系は原点が中心に存在し焦点距離が 1 に正規化されている．カメラの焦点距離を f，横方向と縦方向の画素の物理的な間隔をそれぞれ δ_u, δ_v，カメラの光軸と画像座標系の交点の座標を (c_u, c_v) とすれば，正規化画像座標系上の点 $\boldsymbol{x} = (x, y)$ と画像座標系上の点 $\boldsymbol{m} = (u, v)$ の関係は

$$\begin{aligned} x &= \delta_u(u - c_u)/f \\ y &= \delta_v(v - c_v)/f \end{aligned} \tag{2}$$

と記述できる．また，カメラ座標系と正規化座標系の対応は

$$\begin{aligned} x &= X/Z \\ y &= Y/Z \end{aligned} \tag{3}$$

と表される．

以上の各座標間の関係は行列を用いて簡潔に記述でき，画像座標系とワールド座標系の関係は

$$\begin{aligned} [u, v, 1]^T &\sim \boldsymbol{A}(\boldsymbol{R}|\boldsymbol{t})\,[X_w, Y_w, Z_w, 1]^T \\ \boldsymbol{A} &= \begin{pmatrix} f/\delta_u & 0 & c_u \\ 0 & f/\delta_v & c_v \\ 0 & 0 & 1 \end{pmatrix} \end{aligned} \tag{4}$$

と表される．$\boldsymbol{A}$ は内部パラメータ行列と呼ばれる 3×3 の行列である．また，(4) 式の両辺が $\sim$ によって結ばれているのはその両辺が定数倍を除いて等しいことを意味する．

地面領域のピクセルとその深度が分かれば，透視投影変換によって地面領域を構成するピクセルを三次元空間中に投影しその形状を復元することができる．しかしながら，その形状は深度推定の結果によって真の地面形状と差が生じる．

本研究では透視投影変換によって得られた空間中での地面領域を平面で近似する．画像中の地面領域を透視投影変換

図 4 物体復元のイメージ図．物体をの形状を直方体に単純化して考えることで，複雑な形状の物体でも形を保ったまま頑健に復元することができる．

した三次元点群 $\boldsymbol{G}_{\mathrm{w}}$ が平面上に分布していると仮定すると，近似平面の法線ベクトルは主成分分析の第三主成分に対応する．ワールド座標上の点 $\boldsymbol{G}_{\mathrm{w}}$ の平均を $\boldsymbol{m} = (x_0, y_0, z_0)$，主成分分析の第三主成分を $\boldsymbol{w}_3 = (a, b, c)$ として，近似平面の方程式は

$$(\boldsymbol{x} - \boldsymbol{m}) \cdot \boldsymbol{w}_3^T = 0 \tag{5}$$

と表すことができる．得られた近似平面に対して $\boldsymbol{G}_{\mathrm{w}}$ の正射影をとることで，全ての点を平面上の点に変換した集合 $\boldsymbol{G}_{\mathrm{p}}$ を得る．さらに $\boldsymbol{G}_{\mathrm{p}}$ を，主成分分析で得られた基底で変換することで xz 平面に投影し $\boldsymbol{G}_{\mathrm{q}}$ を得る．$\boldsymbol{G}_{\mathrm{q}}$ に対して xz 平面上で凸包を取ることで地面の形状を復元する．

3.3 接地面推定

物体を地面に立てる位置は，物体と地面の境界により定まる．物体が地面上に存在しているとき，物体と地面の境界は地面の法線ベクトル方向と反対に位置している．法線ベクトルを $\boldsymbol{N}_{\mathrm{w}} = (N_X, N_Y, N_Z)$ とすれば，画像中での法線ベクトル $\boldsymbol{N} = (n_x, n_y)$ は透視投影変換により

$$[n_x, n_y, 1]^T \sim \boldsymbol{A}(\boldsymbol{R}|\boldsymbol{t})\,[N_X, N_Y, N_Z, 1]^T \tag{6}$$

で得られる．

物体マスクの凸包をなす頂点集合を $\boldsymbol{C}$ として $\boldsymbol{C}$ から地面と接している部分の集合 $\boldsymbol{C}_{\mathrm{g}}$ を求める．集合 $\boldsymbol{C}$ に属する全ての頂点の平均で定義される点を G とする．集合 $\boldsymbol{C}$ に存在する頂点のうち隣接する頂点の組 A,B に対して，G から直線 AB に下ろした垂線の足を H とする．このとき画像中での法線ベクトル $\boldsymbol{N}$ とベクトル $\boldsymbol{GH}$ の成す角 θ の余弦 $\cos\theta$ が小さいほどベクトルの方向は正反対に近くなる．厳密に接地面を求める代わりに $\cos\theta < thr$ となる頂点の組の集合を $\boldsymbol{C}_{\mathrm{g}}$ とする．閾値 thr は $\boldsymbol{C}_{\mathrm{g}}$ の大きさを決定し，その値が小さいほど $\boldsymbol{C}_{\mathrm{g}}$ は小さくなる．

3.4 折れ線近似

物体を看板状で立てるためには，物体が面で構成されている必要がある．実際の物体形状は看板上に立てるためには複雑であるため，本研究では全ての物体形状を直方体と仮定して形状復元を行う．この仮定により物体は二つの面で表現され，接地面は折れ線として現れる（図 4）．前節で得られた集合 $\boldsymbol{C}_{\mathrm{g}}$ に対して，折れ線を近似することで物体の接地面を決定する．

近似される折れ線の形状は直方体の仮定により制限される．制限は二つ存在し，折れ線が 2 本の線分で構成されなければならないことと，折れ線の角が $\boldsymbol{C}_{\mathrm{g}}$ の 1 点に対応していることである．以上より，接地面に対する折れ線近似は，制約問題としてラグランジュの未定乗数法を用いて解くことができる．Aronov らの手法 [8] をもとにして，折れ線を割り当てる集合を $\boldsymbol{S} = \{p_1 = (x_1, y_1), p_2 = (x_2, y_2), \ldots, p_n = (x_n, y_n)\}$ とする．ただし $x_1 < x_2 < \cdots < x_n$ である．目的はこの頂点集合 $\boldsymbol{S}$ に接点の x 座標が $x_k (1 < k < n)$ となる 1 接点の折れ線の中で，(7) 式が最も小さくなる折れ線を割り当てることである．集合 $\boldsymbol{S}$ を 2 分割し $\boldsymbol{S}_1(q) = \{p_1, p_2, \ldots, p_q\}$, $\boldsymbol{S}_2(q) = \{p_{q+1}, p_{q+2}, \ldots, p_n\}$ と置く．部分集合 $\boldsymbol{S}_1(q)$ と $\boldsymbol{S}_2(q)$ に対して 2 つの直線 $l_1 : y = a_1 x - b_1$ と $l_2 : y = a_2 x - b_2$ をそれぞれ近似する．これらの折れ線のパラメータ (a_1, b_1, a_2, b_2) は目的関数を

$$f(a_1, b_1, a_2, b_2) = \sum_{i=1}^{q} (a_1 x_1 - b_1 - y_i)^2 + \sum_{i=q+1}^{n} (a_2 x_i - b_2 - y_i)^2 \tag{7}$$

として，制約条件

$$g(a_1, b_1, a_2, b_2) = a_1 x_q - b_1 - a_2 x_q + b_2 \tag{8}$$

を加えて最小化することで求める．この問題は，ラグランジュ関数 $L(a_1, b_1, a_2, b_2)$ を

$$L(a_1, b_1, a_2, b_2) = f(a_1, b_1, a_2, b_2) - \lambda g(a_1, b_1, a_2, b_2) \tag{9}$$

としてラグランジュの未定乗数法を用いて解くことができる．(9) 式の解は接点を x_q に固定したもとでの解であるため，q を 2 から $n-1$ まで変化させ，最も (7) 式が小さくなる折れ線を集合 $\boldsymbol{S}$ の近似折れ線とする．しかし，この手法では集合 $\boldsymbol{S}$ に点が三点以上必要である．集合 $\boldsymbol{S}$ に属する点が三点未満の場合は，そのうち最も低い位置に存在する点を通り，地面に対して水平な線を接地面とする．

3.5 ホモグラフィ変換

物体の画像中での位置は前節の議論により求めることができる．しかし，実際に求めたいものは復元後の地面に対する物体の位置である．

画像中の地面領域は深度を推定することが可能であるが，地面推定の結果には誤りが存在するため，画像中の物体の接地面はかならずしも地面として推定されるわけではない．そのようなピクセルは深度が不明であり，接地面の空間中での位置は透視投影変換を用いて知ることはできない．

地面を平面と仮定した上での形状復元は，画像に写る地面を，上空から見たときの形状に変換することに等しい．視点を変えることで平面図形の形状が変化する様子は，ある四角形を別の形状の四角形に移す変換として表現でき，ホモグラフィ変換と呼ばれている．ホモグラフィ変換は，変換前の座標を $\boldsymbol{H} = (X, Y)$ ，変換後の座標系を $\boldsymbol{H'} = (X', Y')$ とすれば，

$$[X', Y', 1]^T \sim \boldsymbol{M} [X, Y, 1]^T \tag{10}$$

と表される．$\boldsymbol{M}$ はホモグラフィ行列と呼ばれる 3×3 の行列である．このホモグラフィ行列のパラメータは対応点が四点以上与えられたとき決定できる．

画像中の地面と復元後の地面は，同一の図形を異なる視点で捉えたものと考えられ，両者の関係はこのホモグラフィ変換により与えられる．ホモグラフィ行列 $\boldsymbol{M}$ を，深度が推定されている地面領域のピクセルとその復元後の点の対応から求めることで，深度が不明のピクセルに対しても，復元後の位置を計算できる．接地面と推定された折れ線も $\boldsymbol{M}$ によって復元後の地面に投影することができる．

3.6 背景領域

背景領域の推定のために地面領域と背景領域の境界を知る必要があるがこれは接地面の推定と似た手順で求めることができる．

地面領域の凸包を成す集合を $\boldsymbol{S}$ とし，$\boldsymbol{S}$ に属する頂点の平均で表される点を $\mathrm{G_s}$ とする．点 $\mathrm{G_s}$ から集合 $\boldsymbol{S}$ の隣接する二つの頂点 A，B を通る直線に対して下ろした垂線の足 $\mathrm{H_s}$ とする．画像中での法線ベクトル $\boldsymbol{N}$ とベクトル $\boldsymbol{G_s H_s}$ の成す角を θ_s として $\cos\theta_s \geq 0$ を満たすような点 A, B を結んでできる線分 AB を境界とする．頂点 A から画像中での法線ベクトル $\boldsymbol{N}$ 方向に伸ばした直線と画像端との交点を E_1，B からベクトル $\boldsymbol{N}$ 方向に伸ばした直線と画像端との交点を E_2 とし，四角形 AE_1E_2B を線分 AB に対応する背景領域とする．隣接頂点対 A と B のどちらかが画像端に位置しているとき，AE_1E_2B は三角形となるが二つの頂点が同じである四角形として扱う．

頂点，対 A,B は地面領域の凸包を構成する点であり深度を持つ．一方で E_1, E_2 は深度を持たないが，それぞれ A,B の真上に存在するため，E_1 の深度は A と E_2 の深度は B と等しいものとする．これら四点の深度とその画像座標が決定すれば，透視投影変換により復元できる．

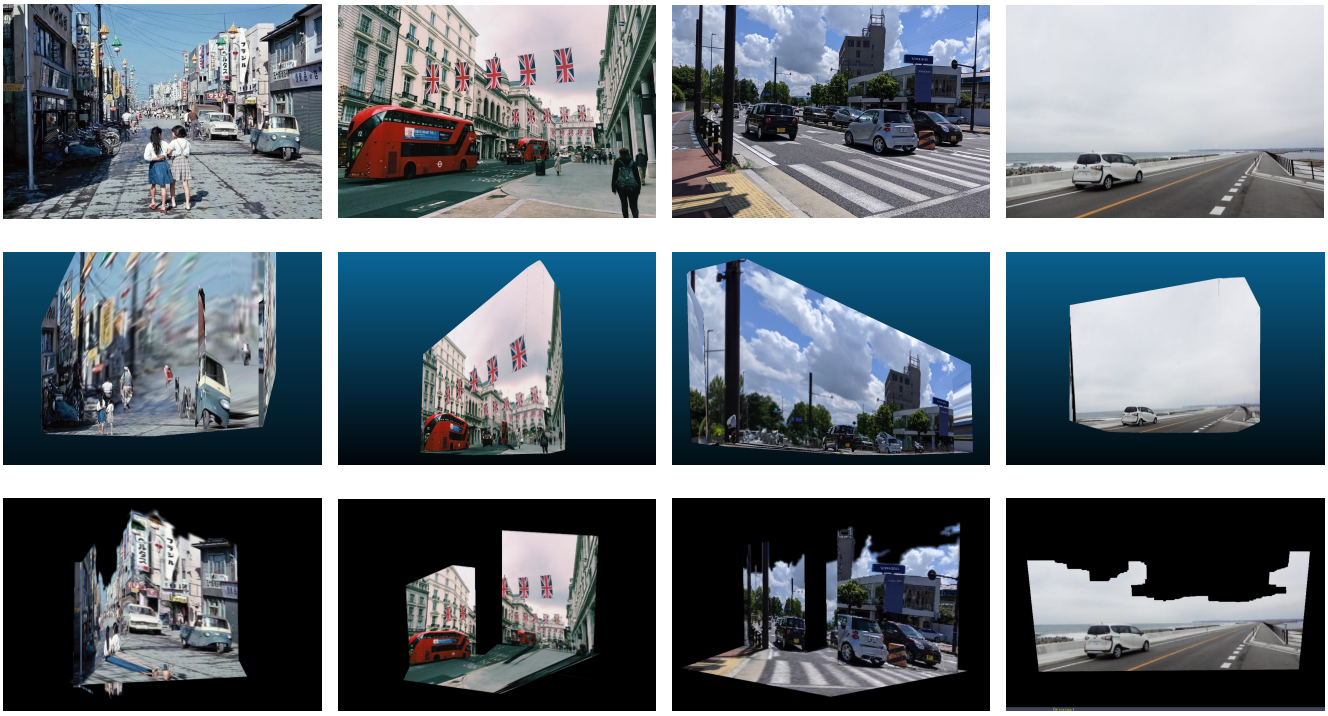

図 5 生成モデルの比較．それぞれ上段が入力画像，中段が本手法の生成モデル，下段が既存手法 [2] の生成モデル．既存手法と比べて，本手法は地面形状および，人や車などの物体を正確に復元できている．

3.7 画像補完

画像から三次元形状を復元するときの課題として, 物体によって隠れていた部分が視点を変えることによって露わになることがあげられる．しかし，地面や背景部分に元の画像をテクスチャとして貼り付けると前景が写り込んだ不確かなモデルが生成される．

これを防ぐためには，前景によって隠された地面や背景を推定する必要がある．本研究の入力は単一画像であるため真の形状を復元することはできない．そこで，隠れている場所は復元する代わりに，もっともらしく補完する．インスタンスセグメンテーションにより画像内の前景は全て検出されているものとすると，元の画像から前景部分が切り抜かれた画像を得ることができる．欠損のある画像を入力として欠損部分を周囲の画像のピクセルの情報から修復する研究は盛んに行われており，画像補完と呼ばれている．画像補完の手法は様々であるが近年は深層学習を用いた手法が多く研究されている．

Iizuka らは，学習時に画像を補完するための補完ネットワーク，補完された画像が全体として整合性のある構造になっているかどうかを識別するための大域識別ネットワーク，補完領域をパッチとして取り出しその領域のより詳細な整合性を評価する局所識別ネットワークを用いることで高い精度で画像補完を行った [9]．補完ネットワークが識別ネットワークに見分けられないように，識別ネットワークが補完ネットワークを見分けるように交互に学習することで補完ネットワークによる自然な画像補完を行う．このネットワークは，補完対象となる欠損画像と欠損領域が白色，それ以外が黒色であるマスク画像を入力とし，補完された画像を出力する

本研究ではこの Iizuka らの方法を用いて画像補完を行う．画像補完で得られる画像は前景が完全に消え去った画像であることが望ましい．しかしながら，インスタンスセグメンテーションの予測結果と画像中の物体の間にはズレが生じる．このズレにより，入力画像から予測された前景領域を切り抜いた欠損画像に物体の一部がはみ出して残ることがある．この画像をネットワークに入力すると，はみ出た物体によって画像の補完結果が悪化する．これを防ぐために，欠損画像の欠損領域を拡大し，残った前景部分を除去したものをネットワークの入力とする. この欠損領域の拡大はモルフォロジー変換で行う．

4. 評価実験

3 章で述べた手法をもとに，実際に一枚の画像からシーンの三次元形状モデルを生成し，その評価を行う．本研究において画像中の前景抽出には He らの手法 [6]，地面領域とその深度の推定には Watson らの手法 [7] を用いた．入力画像のカメラパラメータは画像から得ることはできないため，内部パラメータは $f/\delta_u = f/\delta_v = 500$，$c_u = W/2$，$c_v = H/2$ と仮定した．ただし W, H はそれぞれ入力画像の幅と高さである．また，画像の奥行きに深度を用いるため，カメラ座標系とワールド座標系は同一であり，その回転行列 $\boldsymbol{R}$ は 3×3 の単位行列に等しい．背景と地面のテク

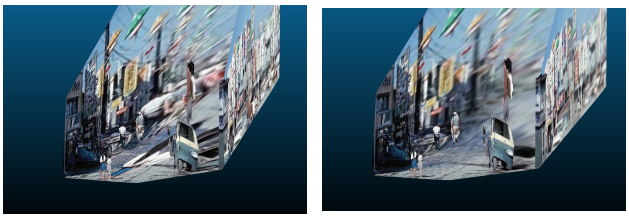

図 6 画像補完の効果．画像補完を行うことで，人や車に回り込んだ箇所でも自然な見た目となっている．

図 7 欠損領域拡大による補完結果の比較．モルフォロジー変換による欠損領域の拡大によって，画像中に残ったの物体領域が除去され，画像補完の結果が向上している．

スチャ画像のための画像補完には Iizuka らの手法 [9] を用い，モルフォロジー変換には $L \times L$ の正方形のカーネルを用いた．ただし L は画像の短辺の 40 分の 1 の大きさに設定した．

4.1 生成モデルの評価

本研究では厳密な形状復元を行うのではなく，地面に対して物体を垂直に立てることで，シーンの立体形状を復元している．モデルを新たな視点から見たときの形状と真の形状は厳密には異なっており，両者の誤差をとるなどの定量的な評価結果と，実際のモデルの見た目の良さとの相関は正になるとは限らない．生成されたモデルの評価は既存研究 [2] の生成モデルとの比較することで，定性的に行う．

4.2 既存研究とのモデル比較

画像から生成されたモデルのいくつかを図 5 に示す．図の一行目が入力画像，二行目が本研究での生成モデル，三行目が既存研究 [2] の生成モデルである．それぞれのモデルは，その形状の特徴が露わになるように元の画像と視点を変えている．

車や人などの画像中の物体に着目すると，既存手法では，物体の中間あたりで曲がっていたり，物体が二つに分かれたりしている様子が見られる．この原因はスーパーピクセルに基づいた領域分割の精度が悪く，物体を検出できていないことが原因であると考えられる．一方で，本手法では，インスタンスセグメンテーションにより高い精度で物体を検出したことで，物体復元が高い精度で行えていることが分かる．垂直成分と地面領域の境界においても，本手法の方が，実際のシーンに近いものになっていることが明らかである．

4.3 画像補完の効果

本手法では地面領域や背景領域のテクスチャ画像に対して欠損領域の補完を行った．この画像補完は自然なモデルを生成することに大きく貢献している．図 6 の一列目が画像補完を行わなかったモデル，2 列目が画像補完を行ったモデルである．補完を行わなかったモデルでは画像に写っていた人や車などの物体が大きく引き伸ばされ，モデルの印象を大きく変えている．一方で画像補完を行ったモデルは，車や人によって隠されていた地面や建物などが補完されることにより自然な見た目となっている．

また，モルフォロジー変換は画像補完による見た目の向上をさらに押し上げている．図 7 の一列目が欠損領域を拡大しなかった場合の補完結果，二列目が拡大した場合の補完結果である．拡大せずに補完した画像は，元の物体の輪郭が浮き出ており，ぼんやりと人や車が浮かんでいる．一方で，欠損領域を拡大し補完した画像は欠損部分が周囲の環境に溶け込んでいる．

4.4 不自然な生成モデル

本研究により自然な見た目のモデルを生成できることは確認できた．しかしながら，中には全体としては自然な見た目をしていても，部分的には違和感が存在していたり，全体の形状が誤っていたりするモデルも生成された．このようなモデルの誤りは，いくつかの種類に分類することができる．

4.4.1 地面上に存在しない物体

このモデルでは地面上に設置している物体を地面上に立てることでシーンの形状を復元する．しかし，インスタンスセグメンテーションによって検出される物体には地面上に存在しないものも含まれる．地面から浮いて存在している物体に接地面を推定すると，その位置は真の位置と比べて遠方に推定される．遠方の物体はホモグラフィにより引き伸ばされるため，実際にはカメラの近くに存在する物体が大きく引き伸ばされて生成モデルの見た目を大きく損なう原因となる．図 8 の左上の画像ではポールと接して空中に存在している看板を復元しようとした結果，地面領域よりも後ろ側に看板が生成されている．

4.4.2 物体同士の重なり

画像の中で複数の物体が存在している場合，重なり合うことなくそれらが存在していることよりも，重なりが生じていることの方が多い．重なりによって接地面が隠れている物体に対しても，このモデルでは接地面を割り当て，看板状に地面に立てるため，図 8 の右上の画像のように誤った形状の物体が生み出されることがある．

4.4.3 折れ線による推定

物体の形を直方体として，接地面に折れ線を推定するという手法は，車などの直方体に近い物体の立体化を補助している．一方で，接地面が全体の大きさに対して僅かであ

る人などの物体において接地面に折れ線を仮定すると，その形状が更に細くなってしまうことがある．カメラから遠くに存在する物体の接地面の領域が，実際より大きく推定されることで，図 8 の左下の画像ように人などが実際より大きく生成されることがある．

4.4.4 検出クラスの制限

前景抽出をインスタンスセグメンテーションで行う本研究では，セグメンテーションクラスに存在しない物体を検出することができない．画像中のガードレールやポール，電柱のようなクラスの存在しない物体は検出されずに残る．検出されない物体のうち，電柱や街頭など，高さのある物体は背景領域に写り込むことがある．背景領域はモデル中で遠方に位置し，テクスチャはホモグラフィ変換によって引き伸ばされるため，図 8 の右下の画像のように電柱などが背景に大きく写り込むことがある．

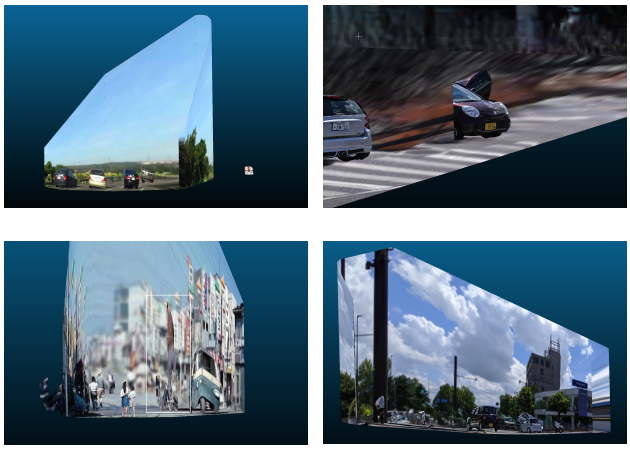

図 8 典型的な失敗例．地面から離れて存在する物体は，地面領域を超えて復元されてしまう（左上）．物体の重なりにより復元結果が悪化する（右上）．接地面が正しく推定できず物体が巨大化することがある（左下）．電柱などが検出されず背景に大きく写り込むことがある（右下）．

5. 結論

本研究では，画像の領域分割を物体と地面の検出問題の二つに分割することで，物体が多く写るシーンからでも頑健に立体モデルを生成する手法を提案し，その検証を定性的に行った．

領域分割を物体と地面の検出問題の二つに分割し，それぞれに特化した既存のネットワークを用いることで，先行研究の課題であった画像の領域分割を高い精度で行った．また，物体の形状を単純化して直方体と仮定し，地面との接地面を 1 接点の折れ線に制限することで，実際は直方体とは異なる形をしている物体でも見た目を損なうことなく復元できることが確認できた．また，前景により隠れた部分に画像補完を用いることで，モデルの見た目を向上させることができた．一方で，本研究には復元される地面を平面かつ凸に制限していたり，カメラパラメータの推定を行わず固定していたりと課題も残っている．

シーンの形状復元を簡単に，かつ頑健に行うことができる本手法により，ユーザーは好きな写真を一枚選ぶだけで，そのシーンを体験するかのように写真を鑑賞することができる．過去の写真や貴重な写真などを鑑賞することによって得られる経験をより実際の体験に近づけることができ，写真鑑賞の新たな側面を引き出すことができた．しかし，多種多様に存在する写真に対して，リアルな鑑賞体験を提供するには課題も多く存在するため，それらを解決していく必要がある．

謝辞 本研究の一部は JSPS 科研費 17K20143, 20H05951, JST さきがけ JPMJPR1858 の助成を受けたものです．

参考文献

[1] Horry, Y., Anjyo, K.-I. and Arai, K.: Tour into the picture: using a spidery mesh interface to make animation from a single image, *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 225–232 (1997).

[2] Hoiem, D., Efros, A. A. and Hebert, M.: Automatic photo pop-up, *ACM SIGGRAPH 2005 Papers*, pp. 577–584 (2005).

[3] Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A. and Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections, *arXiv preprint arXiv:2008.02268* (2020).

[4] Hu, R. and Pathak, D.: Worldsheet: Wrapping the World in a 3D Sheet for View Synthesis from a Single Image, *arXiv preprint arXiv:2012.09854* (2020).

[5] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, *arXiv preprint arXiv:1506.01497* (2015).

[6] He, K., Gkioxari, G., Dollár, P. and Girshick, R.: Mask R-CNN, *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969 (2017).

[7] Watson, J., Firman, M., Monszpart, A. and Brostow, G. J.: Footprints and free space from a single color image, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11–20 (2020).

[8] Aronov, B., Asano, T., Katoh, N., Mehlhorn, K. and Tokuyama, T.: Polyline fitting of planar points under min-sum criteria, *International journal of computational geometry & applications*, Vol. 16, No. 02n03, pp. 97–116 (2006).

[9] Iizuka, S., Simo-Serra, E. and Ishikawa, H.: Globally and locally consistent image completion, *ACM Transactions on Graphics (ToG)*, Vol. 36, No. 4, pp. 1–14 (2017).