変形 AR マーカの高速，高精度な 3 次元位置・姿勢推定と組み込みボードへの実装

浅野右京 †　合志武瑠 †　山内悠嗣 †
† 中部大学
E-mail: er20002-7352@sti.chubu.ac.jp, yuu@fsc.chubu.ac.jp

1 はじめに

近年，2 次元コードが普及し，キャッシュレス決済やロボットの自己位置推定など様々な用途で利用されている．しかし，2 次元コードを貼付する対象は平面であることが前提であり，円柱や非剛体などに貼り付けると変形し，認識や位置・姿勢推定が難化する．この問題を解決するために，変形した 2 次元コードを認識する手法 [1, 2, 3] が幾つか提案されている．そのうちの 1 つである従来法 [3] では，2 次元コードの一種である AR マーカを取り扱い，機械学習による変形 AR マーカの認識と 3 次元位置・姿勢推定法を提案した．従来法では，まず Single Shot Multibox Detector(SSD)[4] で変形 AR マーカの検出と ID の推定を行い，Augmented Autoencoder(AAE)[5] で検出した変形 AR マーカから背景と変形を除去した情報である潜在変数を取得する．最後に，潜在変数を予め作成したデータベースと照合することで姿勢を推定する．この従来法には 3 つの課題がある．1 つ目は検出性能であり，AR マーカの不規則かつ複雑な変形により，検出が困難な場合がある．2 つ目は姿勢推定精度である．Roll は回転角度に対して画像上の見えの変化が大きいため，ある程度の姿勢推定精度が担保されるが，Pitch と Yaw は回転角度に対しての見えの変化が小さいため姿勢推定精度が低下する．3 つ目は推定に要する時間である．先行研究では，数百次元の実数ベクトルとして表現される潜在変数とデータベースとの照合が行われるため姿勢推定に時間を要する．

そこで，本研究では 3 つの課題を解決するために次の提案と改良を行う．

- 物体検出モデルの変更
  物体検出モデルを SSD から NanoDet-Plus[6] に変更する．これにより変形 AR マーカの高精度な検出，位置推定を実現する．
- AAE の拡張
  変形を除去する AAE に連続性を考慮可能な Variational Autoencoder[7] を導入した Augmented Variational Autoencoder(AVAE) を提案する．AVAE により姿勢推定に適した潜在変数を取得する．
- 回帰による姿勢推定
  従来法では潜在変数とデータベースの照合処理に多大な計算時間を要していた．そこで，本研究では Multi-Layer Perceptron(MLP) を用いた回帰による姿勢推定に変更することで計算量を大幅に削減する．
- 2 つのモデルの交互最適化
  提案手法は，AVAE と MLP の 2 つのモデルで構成されている．AVAE と MLP を同時に最適化することが難しいため，交互に最適化するフレームワークを導入する．

また，モバイルデバイスや組み込みボード等に実装し，実用化を想定している．そこで，提案手法を NVIDIA 製の組み込みボード Jetson Orin Nano[8] 上に実装し，処理速度とリソース消費量の検証を行う．

2 関連研究

本研究は，物体検出や姿勢推定の技術を基盤とし，変形した 2 次元コードの 3 次元位置，姿勢を推定する．関連する 3 つの主要分野である 2 次元コードの認識，物体検出，姿勢推定の先行研究について述べる．

2.1 2 次元コードの認識

2 次元コードは，垂直方向と水平方向の 2 方向に配置された色のパターンで情報を表現したコードであり，カメラやスキャナで読み取ることでコードに含まれた情報を取得することが可能である．QR コード [9] は現代において最も普及しているコードの 1 つであり，白黒のドットパターンで情報を表現している．QR コードの位置を表すファインダパターンにより，あらゆる方向からでも認識が可能である．また，アライメントパターンにより歪みをある程度補正できる．AR マーカは拡張現実 (AR) のアプリケーションやロボティクス分野で使用される 2 次元コードである．AR マーカは多種多様であり，白黒かつ正方形のマーカ以外にも円形や多色なものが存在している．その中でも ARToolKit[10] をベース

にした多くのマーカが開発されており，ARTag[11] や AprilTag[12]，ArUco[13] などがある．

変形した 2 次元コードを認識する研究として，歪んだ 2 次元コードの復号手法 [1, 2] が提案されている．これらの研究は，QR コードに対して着色された補助線を引くことで歪みの補正を可能とした．また，DeepFormableTag[14] のような 2 次元コードの生成から，検出，情報の復号までの処理を全て End-to-End で学習する手法も提案されている．

2.2 物体検出

物体検出は，画像や動画内において特定の物体の位置の検出とその物体の種類を分類するタスクである．現在では深層学習を用いた物体検出手法が主流となっており，これらの手法は大きく 2 つに分類される．1 つ目は，2 段階の処理を行う物体検出手法である．R-CNN[15]，Fast R-CNN[16]，Faster R-CNN[17] などが該当し，候補領域の検出とクラス分類の 2 つの処理で構成されており，高精度な検出が可能であるものの処理時間に課題がある．2 つ目は，単一のモデルで物体検出を行う手法である．SSD[4] や YOLO[18] などが該当し，候補領域の検出と分類を同時に行うことで，処理速度が大幅に向上し，リアルタイムでの物体検出が可能である．Faster R-CNN，SSD，YOLO は，事前に定義した矩形枠であるアンカーボックスを使用している．これは，特徴マップ上に複数のサイズやアスペクト比で配置され，物体の写る領域を示すバウンディングボックスを推定する際の基準となる．また，CornerNet[19] や CenterNet[20] のようなアンカーボックスを使用しない手法が提案されている．これらの手法はアンカーフリーの検出手法と呼ばれ，物体の中心点や境界を直接推定する手法であり，アンカーボックスに関するパラメータ調整が不要かつ計算効率が高い．

近年では，Transformer[21] を導入した物体検出手法が注目されている．DETR[22] は，Transformer を初めて取り入れた End-to-End の物体検出手法であり，物体が密集した複雑なシーンにおいても高精度な検出が可能である．この手法は，大きな物体に対する検出精度は高いが，小さい物体の検出精度が低く，画像内の全ての画素間の関係性を計算するため処理コストが高い．この課題を解決した Deformable DETR[23] は，Deformable Attention により注目すべきリファレンスポイントの周辺に限定して情報を取得するため計算効率が高い．また，マルチスケールにも対応可能であり，小さな物体の検出精度を改善している．

2.3 姿勢推定

姿勢推定は，画像内の特定の物体の位置や姿勢を推定するタスクである．単一の RGB 画像から深層学習により，位置・姿勢を推定する研究が盛んに取り組まれている．PVNet[24] は，特徴点ベースの姿勢推定法であり，深層学習により画像中の物体のキーポイントを推定し，物体の 3D モデルのキーポイントとマッチングすることで姿勢を推定する．この手法は，物体のテクスチャが豊かな場合は高速かつ正確に姿勢を推定することができるが，テクスチャが乏しい場合には検出できるキーポイントが減少し精度が低下する課題がある．

Augmented Autoencoder(AAE)[5] は，テンプレートベースの姿勢推定法である．この手法は，3D モデルの様々な姿勢の特徴量をテンプレートとして作成し，物体の特徴量をテンプレートと照合することで姿勢を推定する．この手法は，テクスチャが乏しい場合においてもテンプレートの数に比例して精度を確保できる．一方で，照合処理が必要となるため処理時間はテンプレートの数に反比例する．特徴点ベースやテンプレートベースは，画像中の物体の特徴を得るために間接的に深層学習を利用することが多い．

一方で，画像から深層学習により直接姿勢を推定する手法が研究されている．SSD-6D[25] は，分類ベースの姿勢推定法であり，姿勢を離散化することで分類問題として姿勢を推定する．分類ベースの姿勢推定法は，対称性を持つ物体を適切に処理できる特徴がある．対称性を持つ物体とは，特定の軸周りの回転や球のような異なる角度でも同じ姿勢に見える物体を指す．分類ベースの姿勢推定法は，対称性を持つ物体の様々な姿勢を同じクラスとして扱うことで，視覚的に区別できない姿勢を学習可能である．しかし，姿勢を細かく離散化すると膨大なクラス姿勢数となるため学習が収束しない．一方，粗に離散する場合には正確な姿勢の推定ができない．PoseCNN[26] や DeepIM[27] は，回帰ベースの姿勢推定法であり，回帰により直接姿勢を推定する．PoseCNN は，物体のセマンティックラベリング，位置推定，姿勢推定の 3 つのタスクに分解して処理する手法である．CNN から得られた特徴マップとセマンティックラベリングから位置を推定し，特徴マップと位置から回帰により姿勢を推定する．DeepIM は，入力画像から回帰により位置と姿勢を推定し，その結果から物体の 3D モデルのレンダリングを行う．そして，レンダリング画像と入力画像を比較することで反復的に姿勢を推定する．回帰ベースは，姿勢の変化の連続性をとらえることが可能であり，学習データが不均衡である場合の影響を受けにくい．また，計算コストが低く高速に処理できる．しかし，対称性を持つ物体は正解姿勢が複数あるため，複数の正解姿勢の中間値に収束してしまう傾向がある．

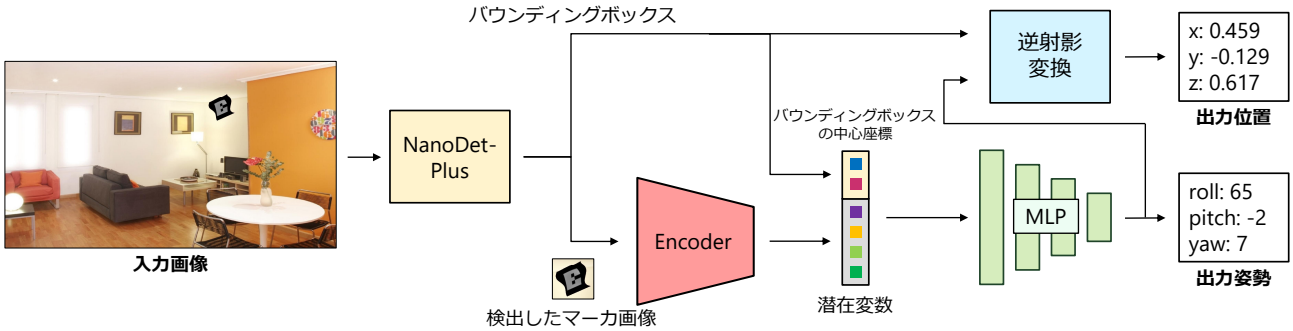

図 1　提案手法の流れ.

3　提案手法

本稿では従来法の課題を解決した変形 AR マーカの高速，高精度な 3 次元位置・姿勢推定法を提案する．図 1 に提案手法の流れを示す．まず，NanoDet-Plus[6] により画像中の変形 AR マーカのバウンディングボックスと AR マーカの種類を表す ID を推定する．次に，検出した変形 AR マーカを Augmented Variational Autoencoder(AVAE) のエンコーダに入力し，背景と変形を除去した潜在変数を取得する．そして，得られた潜在変数を Multi-Layer Perceptron(MLP) に入力することで姿勢を推定する．最後に，変形 AR マーカのバウンディングボックスと推定姿勢から逆射影変換により位置を推定する．

3.1　変形 AR マーカの検出

提案手法では，画像から変形 AR マーカを検出するために使用する物体検出モデルとして NanoDet-Plus[6] を採用する．NanoDet-Plus は，異なるスケールの物体を検出可能にする Feature Pyramid Network[28] を軽量化した Ghost-PAN や対象と非対象の物体クラスの不均衡問題を解決した損失関数 Generalized Focal Loss により軽量かつ高精度な物体検出を実現した．NanoDet-Plus は，クラス内における画像の見えの多彩さにも十分に対応しているため，複雑に変形した AR マーカの高精度な検出が期待できる．また，パラメータ数が少なく，低スペックのデバイスでも動作可能である．

3.2　変形 AR マーカの 3 次元姿勢推定

NanoDet-Plus を用いて検出された変形 AR マーカから AVAE のエンコーダと MLP により姿勢を推定する．

3.2.1　Augmented Variational Autoencoder

Augmented Autoencoder(AAE)[5] はオートエンコーダ [29] を拡張し，入力画像のノイズを除き本質的な情報のみを潜在変数に抽出することを目的としている．AAE のベースとなるオートエンコーダの学習では，入力データをエンコーダにより低次元のベクトルとして表現される潜在変数に圧縮後，デコーダにより入力画像と同じデータに復元する．これに対し AAE は，画像内の背景やオクルージョンなどのノイズを除去した上で，対象となる物体に限定して特徴を抽出することを目的としたオートエンコーダである．入力画像をノイズを含む画像，教師画像をノイズの無い画像とし，入力画像からエンコーダとデコーダにより復元される画像が教師画像に近づくように学習する．本研究では，AR マーカの背景や変形，環境の変化をノイズとして除去することで，AR マーカのパターンや姿勢などの本質的な情報のみを抽出するように学習する．

本研究では，オートエンコーダ型の AAE を Variational Autoencoder(VAE)[7] 型に拡張した Augmented Variational Autoencoder(AVAE) を用いる．VAE は，入力データの背後にある潜在的な確率分布を学習し，入力データの法則性に基づく潜在変数の獲得や新たなデータの生成を可能としたオートエンコーダである．変形 AR マーカの見えの変化と姿勢は密接な関係にある．そこで，姿勢の連続性を潜在変数により表現することを目的として AVAE を採用する．なお，AVAE のエンコーダとデコーダは ResNet-18[30] ベースとする．入力画像 $\boldsymbol{x}$ を AVAE に入力した際，復元画像 $\hat{\boldsymbol{x}}$ が教師画像 $\boldsymbol{y}$ に近づくように学習しながら，式 (2) に示すように潜在変数 $\boldsymbol{z}$ が事前分布に沿うように学習される．

$$\hat{\boldsymbol{x}} = (\Psi \circ \Phi)(\boldsymbol{x}) = \Psi(\boldsymbol{z}) \tag{1}$$

$$\boldsymbol{z} = \boldsymbol{\mu}(\boldsymbol{x}) + \boldsymbol{\sigma}(\boldsymbol{x}) \cdot \boldsymbol{\epsilon} \tag{2}$$

ここで，Φ はエンコーダ，Ψ はデコーダ，$\boldsymbol{z} \in \mathbb{R}^{256}$ である．潜在変数はエンコーダ Φ から得られた平均ベクトル $\boldsymbol{\mu}(\boldsymbol{x})$ と標準偏差ベクトル $\boldsymbol{\sigma}(\boldsymbol{x})$，及び標準正規分布に従う乱数である $\boldsymbol{\epsilon}$ を用いて Reparameterization Trick により計算される．AVAE は，背景や変形の除去を主としながら姿勢の連続性を考慮した潜在変数の取得が期待できる．

3.2.2　Multi-Layer Perceptron

潜在変数から姿勢を推定するために Multi-Layer Perceptron(MLP) を導入する．MLP は，入力層，中間層 2 層，出力層の計 4 層で構成されている．カメラで物体を撮影した場合，物体の姿勢が同一であっても画像上の

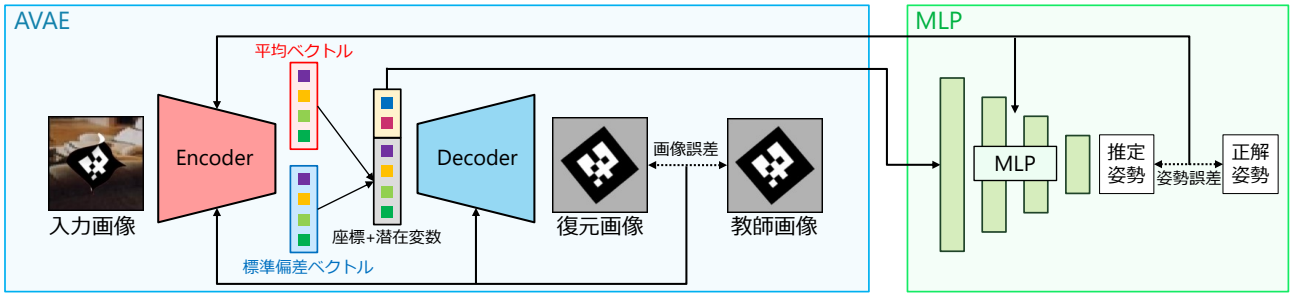

図 2　2 つのモデルの交互最適化の流れ．

位置によって見え方が大きく変化する．そのため，バウンディングボックスの中心座標を MLP に入力することで，変形 AR マーカの画像上の位置による見えの変化を考慮する．MLP への入力は，AVAE より得られる潜在変数 256 次元に，NanoDet-Plus で検出した変形 AR マーカの中心座標 2 次元を連結した計 258 次元のベクトルとする．MLP の出力は，変形 AR マーカの姿勢情報であり，Roll を 2 次元，Pitch と Yaw を各 1 次元の計 4 次元のベクトルである．本研究では変形 AR マーカの姿勢範囲を Roll は $[0\mathrm{deg}, 359\mathrm{deg}]$，Pitch と Yaw は $[-13\mathrm{deg}, 13\mathrm{deg}]$ とし，MLP から出力された値を式 (3, 4, 5) により変換する．

$$r_{angle} = \arctan2(2r_2 - 1, 2r_1 - 1) \tag{3}$$

$$p_{angle} = p \cdot (\theta_{p,max} - \theta_{p,min}) + \theta_{p,min} \tag{4}$$

$$y_{angle} = y \cdot (\theta_{y,max} - \theta_{min}) + \theta_{y,min} \tag{5}$$

ここで，MLP より出力された 2 次元の Roll を r_1 と r_2，変換した Roll を r_{angle} と表す．同様に，MLP より出力された各 1 次元の Pitch と Yaw を p，y，Pitch と Yaw の姿勢範囲を $[\theta_{p,min}, \theta_{p,max}]$，$[\theta_{y,min}, \theta_{y,max}]$，変換した Pitch と Yaw を p_{angle}，y_{angle} と表す．なお，Roll の姿勢範囲のみ $[0\mathrm{deg}, 359\mathrm{deg}]$ であるため，角度の周期性を考慮し，Roll の角度 θ を $(\cos\theta, \sin\theta)$ の 2 次元で表現している．

3.2.3　2 つのモデルの交互最適化

AVAE では変形除去，MLP では姿勢推定を行う目的でモデルが構成されており，2 つのモデルを同時に最適化することができない．そこで，提案手法では，2 つのモデルを交互に最適化する．

交互最適化の流れを図 2 に示す．AVAE を構成するエンコーダとデコーダの最適化では，まず変形 AR マーカ画像を AVAE のエンコーダに入力して潜在変数を得る．次に，潜在変数から AVAE のデコーダにより，背景や変形を除去した AR マーカ画像を復元する．そして，復元画像と教師画像の誤差 L_{AVAE} を式 (6) により計算し，AVAE のエンコーダとデコーダの重みを更新する．

$$L_{AVAE} = L_{rc} + \beta D_{KL} \tag{6}$$

$$L_{rc} = \frac{1}{2}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2 \tag{7}$$

$$D_{KL} = -\frac{1}{2}\sum_{j=1}^{d}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) \tag{8}$$

ここで L_{rc} は，復元画像 $\hat{x}$ と教師画像 x の二乗和誤差を表す．D_{KL} は，潜在変数の分布が正規乱数に沿うように制御する正則化項であり，μ が潜在変数の平均，σ が潜在変数の標準偏差を示す．また，β は正則化の強さを調整するためのハイパーパラメータである．

エンコーダと MLP の最適化では，変形 AR マーカ画像を AVAE のエンコーダに入力して潜在変数を得る．次に潜在変数を MLP に入力して姿勢を推定する．そして，推定した姿勢と正解姿勢の姿勢誤差 L_{pose} を式 (9) により計算し，AVAE のエンコーダと MLP の重みを更新する．

$$L_{pose} = \lambda\left(\frac{|\hat{r_1} - r_1| + |\hat{r_2} - r_2|}{2} + |\hat{p} - p| + |\hat{y} - y|\right) \tag{9}$$

ここで，$(\hat{r_1}, \hat{r_2})$ は Roll の推定姿勢，$\hat{p}$ は Pitch の推定姿勢，$\hat{y}$ は Yaw の推定姿勢，(r_1, r_2) は Roll の正解姿勢，p は Pitch の正解姿勢，y は Yaw の正解姿勢である．また，λ は AVAE のエンコーダに対する損失を調整するためのハイパーパラメータである．

2 つのモデルの最適化を交互に繰り返すことで，AVAE のエンコーダでノイズとなる背景，変形の情報を除去しながら，姿勢推定に必要な情報のみを表現する潜在変数の抽出が可能となる．

3.3　変形 AR マーカの 3 次元位置推定

変形 AR マーカの 3 次元位置は，NanoDet-Plus によって得られたバウンディングボックスと推定姿勢に基づいて逆射影変換により計算される．まず，事前にカメラキャリブレーションを行い，逆射影変換に必要なカメラパラメータを取得する．また，逆射影変換時の基準として，シミュレータ上のカメラパラメータと

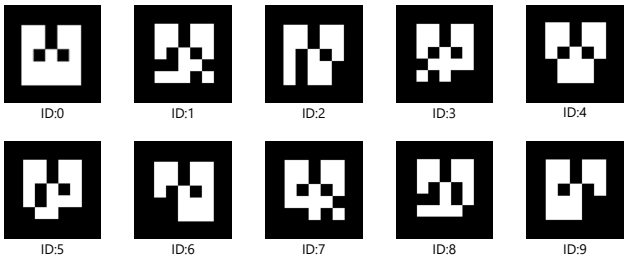

図 3　学習に使用する AR マーカ.

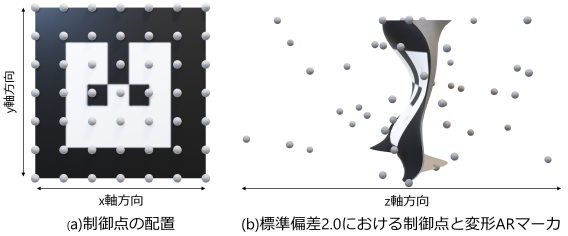

図 4　制御点の配置例と変形した AR マーカの例.

AR マーカの奥行方向の距離を 0.6[m] に設定した場合のバウンディングボックスの大きさを記録する．次に，変形 AR マーカの奥行方向の距離 $\hat{z}_{tg}$ を式 (10) より計算する．

$$\hat{z}_{tg} = z_{bs} \times \frac{bb_{bs}}{bb_{tg}} \times \frac{f_{tg}}{f_{bs}} \tag{10}$$

ここで，z_{bs} と bb_{bs}，f_{bs} はそれぞれ基準となる奥行方向の距離とバウンディングボックスの対角線の長さ，焦点距離を表す．bb_{tg} と f_{tg} は変形 AR マーカにおけるバウンディングボックスの対角線の長さと焦点距離を表す．そして，変形 AR マーカの 3 次元位置 $\hat{\boldsymbol{t}}_{tg} = (\hat{x}_{tg}, \hat{y}_{tg}, \hat{z}_{tg})$ を式 (11, 12) より求める．

$$\Delta\hat{\boldsymbol{t}} = \hat{z}_{tg}\boldsymbol{K}_{tg}^{-1}\boldsymbol{bb}_{tg,c} - z_{bs}\boldsymbol{K}_{bs}^{-1}\boldsymbol{bb}_{bs,c} \tag{11}$$

$$\hat{\boldsymbol{t}}_{tg} = \boldsymbol{t}_{bs} + \Delta\hat{\boldsymbol{t}} \tag{12}$$

ここで，$\boldsymbol{K}_{tg}$ と $\boldsymbol{bb}_{tg,c}$ は変形 AR マーカを撮影した際のカメラ行列とバウンディングボックスの中心座標，$\boldsymbol{K}_{bs}$ と $\boldsymbol{bb}_{bs,c}$ は基準となるカメラ行列とバウンディングボックスの中心座標を表す．

3.4　学習データセットの作成

変形 AR マーカの検出モデル及び，位置・姿勢推定モデルの学習には大量のデータが必要となる．実環境で変形した AR マーカを撮影しアノテーションを付与するには多大な労力を要する．そこで，本研究ではシミュレータを用いることで学習データを自動的に作成する．

AR マーカには Robot Operating System(ROS) の ar_track_alvar パッケージ [31] を用いる．使用する AR マーカを図 3 に示す．本研究には 10 種類の AR マーカを使用し，AR マーカの一辺の大きさは 50[mm] とする．AR マーカの変形はベジェ曲面により与える．ベジェ曲面はベジェ曲線を 2 次元に拡張したものであり，ベジェ曲面の制御点の位置により複雑な変形を表現できる．本研究では制御点を図 4(a) のように x 軸，y 軸方向に 7 点ずつ，計 49 点を配置する．そして，制御点の x 軸，y 軸方向の位置を固定し，z 軸方向の位置を正規乱数に従い与える．これにより，正規乱数の標準偏差により変形度合いの調節が可能となる．変形の標準偏差は，[0.2, 2.0] の範囲を 0.2 刻みで設定する．図 4(b) に変形の標準偏差を 2.0 に設定した際の制御点と変形 AR マーカの関係を示す．制御点の位置に従い，AR マーカが滑らかに変形していることが確認できる．

図 5　各標準偏差ごとの変形 AR マーカ.

変形 AR マーカの 3D モデルには，3DCG 作製ソフトウェアである Blender を利用する．作製した変形 AR マーカの例を図 5 に示す．作製した変形 AR マーカの 3D モデルを使用し，変形 AR マーカの画像をシミュレータ上で撮影する．変形 AR マーカが取りうる位置と姿勢は特定の範囲内でランダムに決定する．撮影環境の概要を図 6 に示す．カメラの位置を原点とした際に，位置の範囲は x を $[-0.25\text{m}, 0.25\text{m}]$，y を $[-0.15\text{m}, 0.15\text{m}]$，z を $[0.4\text{m}, 0.8\text{m}]$ とする．姿勢範囲は，Roll を $[0\text{deg}, 359\text{deg}]$，Pitch を $[-13\text{deg}, 13\text{deg}]$，Yaw を $[-13\text{deg}, 13\text{deg}]$ とする．

撮影には Gazebo シミュレータを使用する．この方法により撮影した 1,920 × 1,080[pixel] の画像とアノテーションデータを 1 セットとし，変形 AR マーカを検出する NanoDet-Plus の学習に使用するために 55,000 セットを用意する．2 つのモデルの交互最適化では，撮影した画像をアノテーションデータにより，変形 AR マーカを中心とした 128 × 128[pixel] の画像に切り取ったものを使用する．入力画像，入力画像に対するアノテーションデータ，教師画像の 3 つを 1 セットとして 22,000 セットを使用する．

4　評価実験

提案手法の有効性を確認するために，従来法と提案手法で変形 AR マーカの検出性能，位置推定精度，姿

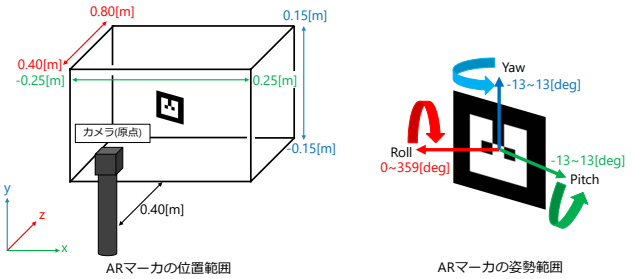

図 6　AR マーカの位置・姿勢範囲.

勢推定精度の 3 つを比較する.

4.1　変形 AR マーカの検出性能

提案手法で使用する NanoDet-Plus と従来法で使用された SSD で変形 AR マーカの検出性能を比較する．評価用画像には，シミュレータ上で撮影された 11,000 枚の変形 AR マーカ画像を使用する．評価指標として，変形 AR マーカの検出精度には mean Average Precision(mAP)，バウンディングボックスの正確性には Intersection over Union(IoU) を採用する.

変形 AR マーカの検出性能の結果を表 1 に示す．mAP では SSD の 0.80 に対し，NanoDet-Plus は 0.96 であり，mAP が 0.16 向上した．また，IoU では SSD の 0.88 に対し，NanoDet-Plus では 0.96 であり，IoU が 0.08 向上した.

表 1　各手法での検出性能.

物体検出モデル	mAP	IoU
SSD[4]	0.80	0.88
NanoDet-Plus[6]	0.96	0.96

4.2　変形 AR マーカの位置推定精度

位置の推定精度を平均絶対誤差により比較する．評価には，シミュレータ上で撮影した画像から NanoDet-Plus により検出した 11,000 枚の変形 AR マーカ画像を使用する.

表 2 に x，y，z の位置推定誤差の結果を示す．提案手法の平均誤差は 4.90[mm] であり，従来法 [3] の平均誤差 6.69[mm] と比べて 26.9[%] 減少した．これは，変形 AR マーカの検出性能が改善し，正確なバウンディングボックスの取得が可能となったためである.

表 2　各手法での位置推定誤差 [mm].

手法	x	y	z	平均
従来法 [3]	3.84	2.40	13.84	6.69
提案手法	2.51	1.51	10.68	4.90

4.3　変形 AR マーカの姿勢推定精度

姿勢の推定精度を平均絶対誤差により従来法と提案手法で比較する．提案手法においては，MLP への入力にバウンディングボックスの中心座標の有無，2 モデルの交互最適化の有無について検証する．なお，2 モデルの交互最適化をしない場合には，AVAE の最適化終了後に MLP を最適化する．評価には，シミュレータ上で撮影した画像から NanoDet-Plus により検出した 11,000 枚の変形 AR マーカ画像を使用する.

表 3 に Roll，Pitch，Yaw の姿勢推定誤差の結果を示す．提案手法は従来法の Roll の精度を保ちながら，Pitch と Yaw の精度を大きく改善できていることがわかる．提案手法の平均誤差は 1.97[deg] であり，従来法の平均誤差である 5.28[deg] と比べて 62.7[%] 減少した．また，従来法と提案手法は，Roll に対して Pitch と Yaw の姿勢推定精度が低い．Roll は平面上の回転として表現されるため，見えの変化が大きく推定が容易い．一方で，Pitch と Yaw は視点の奥行方向に回転し，見えの変化が小さく，マーカの位置によって見え方が大きく変化するため推定が難しい.

提案手法は，MLP にバウンディングボックスの中心座標を入力しない場合と比べ，平均誤差が 53.9[%] 減少した．これは，位置の違いにより変形 AR マーカの見えの変化の影響が大きく，バウンディングボックスの中心座標を MLP に入力することで見えの変化を考慮した姿勢推定が可能となったためだと考えられる．また，提案手法は，個別にモデルを学習した場合と比べ，平均誤差が 48.6[%] 減少した．モデルを交互に最適化する提案手法では，AVAE で変形を除去しながら姿勢推定に適した潜在変数を獲得するため姿勢推定精度が向上したと考えられる.

表 3　各手法での姿勢推定誤差 [deg]．位置は MLP にバウンティングボックスの中心座標を入力することを表す．個別は AVAE と MLP をそれぞれ個別に最適化することを表し，交互は 2 つのモデルを交互に最適化することを表す.

手法	位置	個別	交互	Roll	Pitch	Yaw	平均
従来法 [3]	-	-	-	0.69	7.84	7.32	5.28
提案手法			✓	1.02	5.88	5.92	4.27
提案手法	✓	✓		1.72	5.01	4.78	3.83
提案手法	✓		✓	0.83	2.61	2.48	1.97

図 7 に提案手法で学習した AVAE により復元した画像の例を示す．復元画像は教師画像と同様の見た目をしており，入力画像から背景と変形が除去されていることが分かる.

5　組み込みボードでの検証

組み込みボードを用いて，提案手法の処理速度とリソース消費量を検証し，実用性を評価する．本研究では，組み込みボードとして図 8 に示す Jetson Orin Nano を

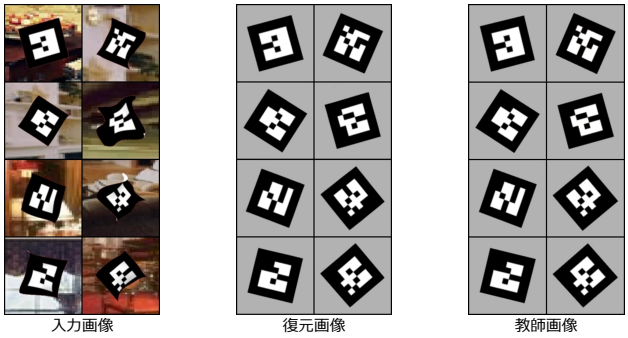

図 7　AVAE による復元画像の例.

採用する．Jetson Orin Nano は省電力で動作する小型の AI モジュールである．主なスペックを表 4 に示す．カメラには Logicool C920n HD Pro ウェブカメラ (解像度は 1,920x1,080[pixel]) を使用し，実環境で検証を行う．

表 4　Jetson Orin Nano のスペック.

CPU	6-core Arm Cortex-A78AE v8.2 64-bit CPU1.5MB L2 + 4MB L3
GPU	1024-core NVIDIA Ampere architecture GPU 32 Tensor Cores
メモリ	8GB
消費電力	7W - 15W
サイズ	100mm x 79mm x 21mm

検証の結果，提案手法はメモリ使用量が 2.0[GB] と少量ながら 10.09[fps] で動作することを確認した．使用する変形 AR マーカを図 9 に示す．変形 AR マーカの形状は 3D プリンターで作成し，AR マーカのパターンはマーカーペンで着色した．実環境の変形 AR マーカ対する位置・姿勢推定の例を図 10 に示す．変形 AR マーカの ID をワイヤーキューブの色で表し，変形 AR マーカの位置と姿勢をワイヤーキューブの位置と姿勢で表現している．デモンストレーションの例から，変形 AR マーカを高精度に位置・姿勢推定できていることが視覚的に確認できる．

6　おわりに

本稿では，変形 AR マーカの高速，高精度な 3 次元位置・姿勢推定手法の提案と組み込みボードへの実装について述べた．

従来法 [3] の課題であった精度と処理時間を解決するために，下記の 4 つを提案した．

- 物体検出モデルの変更
- AAE を拡張した AVAE の提案
- 回帰による 3 次元姿勢推定
- 2 つのモデルの交互最適化

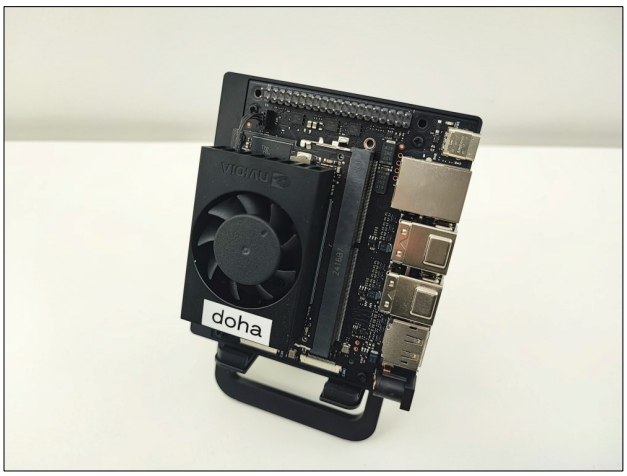

図 8　Jetson Orin Nano.

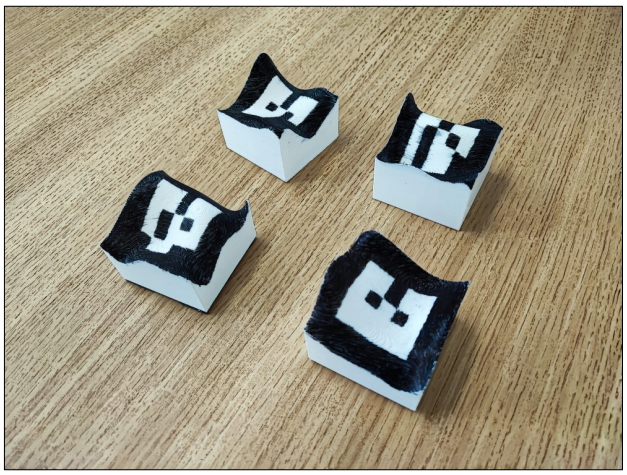

図 9　作成した変形 AR マーカ.

今後は，実環境の変形 AR マーカに対し，高速かつ高精度な位置・姿勢推定を行う予定である．

参考文献

[1] 小野智司，川上雄大，伊藤拓也，澤井陽輔，川崎洋，中山茂，“ゴミ袋に貼付された歪んだ 2 次元コードの復号”，人工知能学会全国大会論文集，2012.
[2] 小野智司，川上雄大，伊藤拓也，藤田晋輔，中山茂，川崎洋，“ゆがんだ二次元コードの復号による廃棄物認識”，人工知能学会誌，vol. 28, no. 4, pp. 575–582，2013.
[3] 榎元 洋平，山内 悠嗣，“機械学習による変形 AR マーカの 3 次元位置・姿勢推定”，動的画像処理実利用化ワークショップ，2022.
[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector”, *Proceedings of the European Conference on Computer Vision*, pp. 21–37, 2016.
[5] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, “Implicit 3d orientation learning for 6d object detection from rgb images”, *Proceedings of the European Conference on Computer Vision*, pp. 699–715, 2018.
[6] RangiLyu, “Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model.” https://github.com/RangiLyu/nanodet, 2021.
[7] D. P. Kingma, and M. Welling, “Auto-Encoding Variational Bayes”, *2nd International Conference on Learning Representations*, 2014.
[8] Jetson orin nano, https://www.nvidia.com/en-us/autonomousmachines/embedded-systems/jetson-orin.
[9] 長屋隆之，山崎知彦，原昌宏，野尻忠雄，“高速読取り対応 2 次元コード [qr コード] の開発”，全国大会講演論文集，vol. 第 52 回，no. メディア情報処理，pp. 253–254，1996.

図 10　実環境におけるデモンストレーションの例．変形 AR マーカの ID をワイヤーキューブの色で表し，変形 AR マーカの位置と姿勢をワイヤーキューブの位置と姿勢で表現している．

[10] H. Kato, and M. Billinghurst, “Marker tracking and hmd calibration for a video-based augmented reality conferencing system”, *Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality*, pp. 85–94, 1999.

[11] M. Fiala, “Artag, a fiducial marker system using digital techniques”, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 590–596 vol. 2, 2005.

[12] E. Olson, “Apriltag: A robust and flexible visual fiducial system”, *IEEE International Conference on Robotics and Automation*, pp. 3400–3407, 2011.

[13] S. Garrido-Jurado, R. Muoz-Salinas, F. Madrid-Cuevas, and M. Marn-Jimnez, “Automatic generation and detection of highly reliable fiducial markers under occlusion”, *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[14] M. B. Yaldiz, A. Meuleman, H. Jang, H. Ha, and M. H. Kim, “Deepformabletag: End-to-end generation and recognition of deformable fiducial markers”, *ACM Transactions on Graphics*, vol. 40, no. 4, 2021.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, *Conference on Computer Vision and Pattern Recognition*, 2014.

[16] R. Girshick, “Fast r-cnn”, *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks”, *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection”, *Conference on Computer Vision and Pattern Recognition*, 2016.

[19] H. Law, and J. Deng, “Cornernet: Detecting objects as paired keypoints”, *Proceedings of the European Conference on Computer Vision*, 2018.

[20] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection”, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need”, *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers”, *Proceedings of the European Conference on Computer Vision*, pp. 213–229, 2020.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection”, *International Conference on Learning Representations*, 2021.

[24] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, “Pvnet: Pixel-wise voting network for 6dof pose estimation”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[25] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, “Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again”, *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[26] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, “Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes”, *Robotics: Science and Systems*, 2018.

[27] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, “DeepIM: Deep iterative matching for 6D pose estimation”, *Proceedings of the European Conference on Computer Vision*, 2018.

[28] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[29] G. E. Hinton, and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[31] ar_track_alvar, https://wiki.ros.org/ar_track_alvar.