【調査報告】

射影変換深層学習による物体認識の調査

大関和夫・上條浩一

Research of Recognizing Objects with Projective Transform Deep-Learning

Kazuo Ohzeki and Koichi Kamijo

Abstract: An object with a fixed shape, such as a car (rigid body), looks different in shape because the camera looks in different directions. Therefore, even a rigid body undergoes a large change in shape. This is the effect of parallax distortion on object recognition. In vehicle recognition in autonomous driving, shape recognition (identification) of a vehicle is input from a vehicle on the road. In the existing research by the authors, we have examined deep learning by putting vehicle images with different shapes at the shooting position into the same learning data class. Consider a method that uses the same shape as learning data for the same object. Equivariant learning based on a single shape is currently being researched for 'only congruent transformations' of rotation and translation. We are studying the possibility of 'projective transformation learning' that extends this to 'cases with perspective distortion' caused by shape changes seen from different directions. In this research report, we analyze papers related to the above research and attempt to explore future directions.

Keywords: Shape Space, Rotation, Equivariant, CNN, Riemann

１．まえがき

深層学習では、複数の種類の物体の画像を多数集めて学習データとすれば、各物体を識別する認識器を生成できる。しかしその識別率を向上させる手法として、データ量を増やすことが重要とされており、畳み込み演算とプーリングという主たる演算の改良は進んでいない。本稿では、畳み込みニューラルネットワーク (CNN) を用いる深層学習の改良を目指す調査を行う。自動運転における車両認識においては、車両の形状認識（識別）が通常道路上にある車両を入力とするため、一般的な物体が任意の３D 空間内の位置にあるのに比べ、変形量が少ない点に着目する。

CNN を用いた深層学習の高性能化として、最近見られるのは、物体が回転したときの変形に対する研究がある。それらの中で、人間や動物のように骨格や羽根の動きで形状が大きく変化するものも、もともとは同一物体であり、人間には認識が可能な物である。これを定式化したものとして、Kendall の形状空間があり、物体の 3D 空間での移動や回転、動作による形状変化を統合して認識しようとしている。いずれも非線形な変形であるため線形な空間に投影するための個別的な処理が必要で、困難度は高い。

2．射影歪を含む物体の認識

射影歪を含む物体認識として、本報告では、自動車のような剛体であって、カメラが見込む方向が異なるため、形状が異なって取得される場合を主に想定している。これに対し、筆者等の既存研究では、図 1 のような撮影状況で異なる形状の車両をまとめて学習データの 1 種のクラスとして深層学習するという検討を行ってきた。各クラス 100 枚、合計 1000 枚の画像で、精度約 0.99 程度を達成している。より高い精度を達成するためには、定点カメラ撮影を行いデータ数を増強することが考えられる。一方、混在する形状を増やすと曖昧な識別に繋がる可能性があるため、同一の物体に対し 1 種の形状を学習する方式がある。単一形状を基にした同変学習は、現在回転と平行移動という合同な変換のみに対応するものだが、これを視点の異なる方向から見た形状変化が発生した射影歪にまで拡張する射影変換学習への可能性を検討している。本調査報告は、上記のような研究に関する論文を分析し、今後の方向を探る試みを行っている。

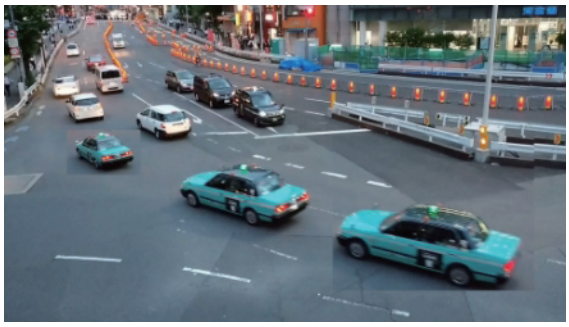

図 1　交差点でのカーブ状の道路での車両形状の変化
（東京都新宿区西新宿交差点）（同一車両を 3 つの位置で合成したもの）

CNN を用いた深層学習の論文は、米国 IEEE の CVPR、ICCV、ECCV などの国際会議、EU の Springer、Elsevier など論文誌が先端的である。本報告では、主に CVPR の採択論文を参照し、調査を行う。まず 2022 年の論文は、2000 件以上あり [1]、国内でも web ページ「定

表 1　IEEE CVF の発表に現れるキーワードのランキング [1] より

順位	キーワード	割合
1	Transformer	6%
2	Object Detection	5%
3	Self Supervis	4%
4	Adversarial	3%
5	Attention	3%
6	Few Shot	2%
7	Semi Supervised	2%
8	Weakly Supervised	2%
9	Contrastive Learning	2%
10	Point Cloud	2%

点観測」[1] や各種勉強会で検討が実施されている。2022 年のキーワードで多いものは、表 1 に示すように、Transformer（転移学習）、Object Detection（物体検出）、Self Supervis（自己教師あり学習）等が上位にある。

2.1 Rotation があるもの

タイトルに Rotation がある文献について調べる。これは、物体認識で、位置が変化する場合に、平行移動、回転が形状変化に影響があると考えられるからである。深層学習の畳み込み演算とプーリングは、小さい平行移動や回転というアフィン変換された入力に対しニューラルネットワークの係数が追従する事ができる。しかし、平行移動量が大きい場合や、回転角の精度要求が高い場合は、追従できない。また、模様のある背景がある場合は物体だけが回転する場合と、背景ごと回転する場合があり、処理が変わる。

Bao ら [3] は、深層学習の外観ベースの視線推定で驚くべきパフォーマンスが達成されて来たが、ターゲット・ドメイン データの不足とターゲット・ラベルの欠如により、視線推定アルゴリズムを一般化することは依然として困難であることを述べている。 この論文では、視線推定における回転一貫性プロパティを利用し、教師なしドメイン適応にて、12.2% から 30.5% のゲインを得ている。既知の視線の顔画像から回転させた複数画像を用意し、学習を行う。ここで、回転に対し一貫性があることを見出している。そこで、新規の入力顔画像を回転させたものとドメイン適合により得た視線を結果として出力する。

Bökman ら [4] は、2D 点群データの回転等分散性（同変性）に関し、任意の連続回転同変関数および順列不変関数を近似できる関数の特定のセットについて説明。この結果に基づいて、2D 点群を処理するための新しいニューラルネットワークアーキテクチャを提案しこれらの対称性を示す関数を近似するための普遍性を証明した。また、同様の等分散特性を維持しながら、2D-2D 対応のセットを indata として受け入れるようにアーキテクチャを拡張する方法も示している。

同変性の簡単な例として、夜空の北斗七星から、北極星への方向を決定するタスクにおいて 入力は、ある 2D 座標フレーム内の北斗七星の位置のセットで、点群プロセッサの回転同変性は、夜空 (または観測者) が回転する場合、決定された方向が回転することを意味する。

ここで、同変とは、equivariant の訳で、２D 上の平行移動・回転等の移動で、図形の合同性が保たれることを意味している。3D 空間での回転や奥行方向の移動は、形状の変形を伴うため、同変とはならない。同変は、変形の中でも形状変化がない最も基本的な場合であるが、CNN との組み合わせ、計算量の爆発などから、CNN と同変の研究が論文に多数採択されているものと考えられる。

基本問題 (a = 0°) ではうまく機能しない。（著者はパラメータ不足のためではないか、と述べているが、違和感が大である。）回転同変および順列不変のニューラルネットワークアーキテクチャに基づくタスクを学習するための基本的なフレームワークを提示している。このアーキテクチャが実際に普遍的であることを証明した。アーキテクチャを変更するいくつかの方法、特に、対応問題に現れる点群のペアに拡張する方法と効率的な計算を実行する方法について述べている。 制限については、フレームワークは 2 次元でのみ適用される点である。30~60 度の回転では従来方法より優位であるが、30 度以下では同等で

ある。また 60 度以上は非対応。

Feng ら [5] は、航空画像からの弱教師付きオブジェクト検出 (WSOD：weakly supervised object detection) のタスクに関し、長年にわたり、まだ調査されていない難しい問題の一つと述べている。 通常の CNN に基づいて構築された既存の主要な WSOD アプローチは、対応する制約なしでオブジェクトの回転に取り組むように本質的に設計されていないため回転に敏感な（不安定な）オブジェクト検出器が生成されてくる問題を指摘している。この論文では、新しい弱い教師あり学習の分野で回転不変の空中物体検出ネットワーク (RINet) を構築している。方向付けられたオブジェクトに対して自然に回転を認識できるようになっている。具体的には、提案方式の RINet は最初に、予測されたインスタンスから回転されたインスタンスへのラベルの伝播を行い、回転一貫性のある教師データを生成する

RINet は、さまざまな回転知覚を備えたオンライン検出器の改良で実装されている。トレーニング中に、回転一貫性のある教師データ（タグ）を生成し、その間、予測されたインスタンス ラベルを異なる回転知覚ブランチ間で補完的な方法で結合することにより、すべての可能なインスタンスを追跡する。包括的な実験により、提案された RINet が既存のすべての WSOD メソッドよりも優れており、新しい最先端の結果が得られたことが実証されたと述べている。

Chen ら [6] は、回転不変性（RI）を 3D 深層学習手法に導入する際の最近の進歩に関し、入力において 3D 座標を置き換えることに着目している。この手法は、入力 RI 機能によって失われたグローバル情報を復元することにある。最新の研究では、追加のブロックや複雑なグローバル表現を発生させようとしているが、計算時間がかかる問題があった。この論文では、グローバルな情報損失は未調査のポーズ情報損失の問題に起因することを指摘している。一般的な畳み込みレイヤーは RI 機能間の相対的なポーズをキャプチャできないため、グローバル情報が深いネットワークで階層的に集約されるのを妨げていると考え、相対ポーズに基づいてカーネルを動的に適応させる Pose-ware Rotation Invarint Convolution（つまり PaRI-Conv）を開発した。具体的には、各 PaRI-Conv レイヤーで、軽量の Augmented Point Pair Feature (APPF) が設計され、RI 相対ポーズ情報を完全にエンコードする。 次に、因数分解された動的カーネルを合成する。これは、APPF から学習できる共有基底行列と姿勢認識対角行列に分解することで、計算コストとメモリ負荷を削減している。形状の分類とパーツのセグメンテーション・タスクに関する広範な実験により、PaRI-Conv が最先端の RI メソッドを凌駕すると同時に、よりコンパクトで効率的であることが示された。法線を安定軸として座標を入力として直接取得することにより、PaRI-Conv は 93.8% と従来例より、0.1 ポイント高い総合精度（総合精度、overall accuracy; OA は全画素の何割が正しく分類されたかを示す精度評価指標）を達成している。

Seo ら [7] は、対称性を検出する時、対称性パターンが任意の方向から生じる問題があるが画像から対称パターンを発見することは、画像の反射と回転に応じて一貫して変化する同変特徴表現が有効であると見出している。対称性検出のためのグループ同変畳み込みネットワーク (group-equivariant convolutional network) EquiSym を提案している。これは、反射と回転の二面体グループに関して同変特徴マップを活用している。提案したネットワークは、二面体同変層（dihedrally-equivariant layers）で構築され、反射軸または回転中

心の空間マップを出力するようにトレーニングされる。実験では、提案方式が LDRS と DENDI というテスト用に対称変換、回転変換して自作したデータセットに対して最先端を達成することを示している。

入力画像がグループ等価エンコーダー Enc に渡された後、反射と回転のグループ等価デコーダーを介し、反射と回転の 2 系統の CNN で検出が行われる。

Yu ら [8] は、点群データから、方向性、回転、対称性を見出し、物体検出をする。方向と物体切り出しを統合処理したが、精度は、統合せずに単独に処理したほうが良かった。統合したシステム EON を開発したところが特徴で、まだ性能的には進展していない。

回転の同変性は、最近、3D ディープ ラーニング分野で注目されているが回転対称性が独自の空間サポートを持っているという事実を無視し、グローバルな入力回転に関する同変性に焦点を当てている。本論文では、シーンの動きとは無関係に、オブジェクトの境界ボックスがオブジェクトの姿勢に関して同変である必要がある 3D オブジェクト検出を行う。これは、オブジェクトレベルの回転の同変性と呼ばれる新しい望ましい特性を示唆していると述べている。

オブジェクトレベルの回転の同変性を 3D オブジェクト検出器に組み込むには、クロスオブジェクトコンテキスト情報をモデル化しながら、ローカルオブジェクトレベルの空間サポートを使用して同変特徴を抽出するメカニズムが必要で、その目的のために、オブジェクトレベルの同変性を達成するための回転同変サスペンション設計を備えた同変オブジェクト検出ネットワーク (EON) を提案している。EON は、VoteNet や PointRCNN などの最新のポイントクラウドオブジェクト検出器に適用でき、シーンスケールの入力でオブジェクトの回転対称性を活用できる。屋内シーンと自動運転データセットの両方での実験では、EON 設計を既存の最先端の 3D オブジェクト検出器にプラグインすることで大幅な改善が得られることが示されている。

システムは 3 つのモジュール、バックボーンが点群を処理して高密度のフィーチャセットにする「シードフィーチャ抽出」、空間領域を要約する「領域コンテキスト集約」各地域フィーチャから候補を提案する「OBB（ 指向性バウンディングボックス）を生成」から成る。この構成を回転同変サスペンション設計と呼んでいる。

Kang ら [9] は、回転した画像を回転する前の画像と同じクラスに属す様にして、データ増強を行い、Leave-One-Out 画像分類と回転画像識別の結合確率を最大化することにより、新しい損失関数を提案している。他の最先端の損失関数と比較し、有効性を確認して

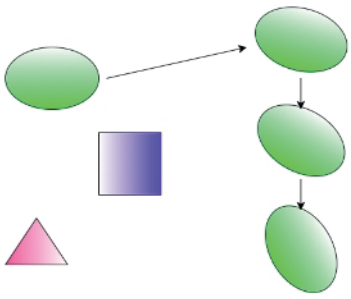

図 2　画像から物体を取り出し、回転した画像を生成し、同じクラスに属する様にして、データ増強（埋め込み）する説明図。文献 [9]Kang の Fig.1 参照。

いる

Zand ら [10] は、現在のオブジェクト検出方法は、もともと軸に沿ったバウンディングボックス検出に対応するように設計されているため、自由に回転するオブジェクトを最もよく表す方向ボックスを正確に特定できない。提案手法は、CNN ベースのアプローチで、アンカーボックスなどの外部リソースを必要とせずに、複数のスケールレベルで潜在的なピクセル情報を使用する。この方法は、グリッドセル位置でのターゲットオブジェクトのフィーチャの正確な位置と方向をエンコードする。境界ボックスの位置と寸法を回帰する既存の方法とは異なり、提案された方法は、回転した物体の枠と成るバウンディングボックスを含む情報を学習することができる。さらに、回転不変の特徴表現が各スケールに適用され、同様の特徴を共有するためにトレーニングサンプルの面内回転の 360 度の範囲をカバーするという正則化制約を課している。xView(Lam らの衛星から見た地上の物体のデータベース)および DOTA（Xia らの Dataset for Object deTection in Aerial images というデータセット）での評価は、提案された方法が既存の最先端の方法よりも一様にパフォーマンスを向上させることを示している。

Zhang ら [11] は、リモートセンシング画像（RSI）の背景の複雑さを課題として取り上げ、機能と最適化の観点からフォアグラウンド (前景) 領域の情報を活用することによる回転精密検出器「フォアグラウンド・リファインメント・ネットワーク（ForRDet)」を提案している。粗密 2 段階方式で、粗い段階では前景コンテキスト表現を集約し、詳細段階で特徴マップ上の前景領域の識別を改善する前景関係モジュール (FRL) を開発。粗い段階（第一段階）から前景アンカーの分類信頼度とローカリゼーション精度を統合する前景アンカーの再重み付け（FRW）損失を設計し、それらの寄与を動的に調整している。前景アンカー強調部、回転オブジェクト検出 DOTA、HRSC2016、および UCAS-AOD の 3 つの公開データセットに関する実験結果は、提案された方法の有効性を示している。

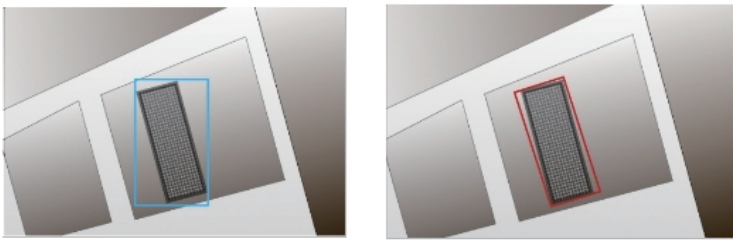

図 3　リモートセンシング画像に対する、前景の物体（建物）の切り出し（アンカー）の様子。左：青線の枠が初期の粗い状態、右が、赤線の枠が詳細段階で、回転した枠の近似が得られている。[11]Zhang の Fig.3 参照。

Cohen ら [12] は、Group equivariant Convolutional Neural Networks (G-CNN) について発表している。対称性を利用してサンプルの複雑さを軽減する畳み込みニューラル ネットワークの自然な一般化を行う。G-CNN は、通常の畳み込み層よりも大幅に高度な重み共有ができる新しいタイプの層である G- 畳み込みを使用している。G 畳み込みは、パラメーターの数を増やすことなく、ネットワークの表現力を高められる。グループ畳み込み層は使いやすく、平行移動、反射、回転によって生成される離散グループの計算オーバーヘッドを

無視して実装できる。反転と小さな変換による拡張は、p4 と p4m-CNN の結果を一貫して改善した。(p4 とは、90 度の回転で 4 要素の群ができていることを指す）

G-CNN は、CIFAR10 およびローテーションされた MNIST で最先端の結果を達成している。回転を組み込んでいるが、離散的ものに限定され、4 種の回転（90 度単位）の例が示されている。

表 2　Rotation 比較

番号	著者	概要	性能
[3]	Bao	視線推定、回転したサンプルを学習 : 回転した学習サンプルを多数生成するのが大変	+12~31%
[4]	Bökman	2D 上での合同図形の回転(同変)と CNN の融合:点群データなので、背景は考慮不要	30~60 度回転で優位
[5]	Feng	航空画像、回転一貫性のある教師データ（タグ）を生成 : 回転不変の情報を CNN に入力	WSOD の範囲で優れる
[6]	R. Chen	回転不変の CNN で、ポーズ情報を法線から取得し入力に加える：ポーズ情報を CNN に与える	精度 93.8% は従来最高の 93.7% を上回る
[7]	Seo	反射 (鏡映) と回転を CNN に組み込んだグループ同変畳み込みネットワー EquiSym を開発：対称性と回転角度の 2 面性をグループ化した二面体同変層を開発	最先端と同等
[8]	Yu	3D 検出にオブジェクトレベルの同変性を活用する見込みがある：点群データの方向データを CNN に送る	途上だが、動作例などが有益
[9]	Kang	回転した画像を回転する前の画像と同じクラスに属す様にする：回転画像のデータ増強	新しい損失関数で実現
[10]	Zand	回転した物体から位置と方向を検出し学習する：枠情報は送らず画像から検出する	パフォーマンスが向上
[11]	Zhang	前景を基に回転精密検出を行い、傾いた領域検出を行っている。:コースファイン方式で、背景から注目物体（前景) を分離	
[12]	Cohen	4 種の回転を組み入れ、群論に対応している	理論面で先行

3.　Kendall の形状空間に関する研究

回転や平行移動による物体の視点から見た形状の変化は、剛体の場合に加え、生物のような動いて変形する場合も対象となる。Kendall は物体の認識を生物が形を変えるところまで考慮した。全体を形状空間（Shape Space）と呼び、端点である 3 次元の特徴点（Landmark）の動きを定義してる。この理論の解説として、Klingenberg の解説論文 [13] がある。

Paskin ら [14] は、人体、動物などの柔軟に変形する物体形状の同一視、識別に、物体の特徴点（Landmark）の点群の分布や、接続関係を表現した shape space（形状空間）の枠組みを用い、2D の特徴点から 3D 形状の推定を行った。3D 形状は、2D 画像よりもはるかに多くの情報を提供できる。ただし、3D 形状の取得は、2D 画像の取得と比較して非常に困難または不可能な場合があり、2D 画像から 3D 形状を導出する必要がある。一般に、これは数学的に不適切な設定の問題だが、事前情報を使用して問題の定式化を制約することで解決できる場合がある。単一の単眼 2D 画像から 3D 形状を再構築するために、

Kendall の形状空間に基づく新しいアプローチを提案している。この研究は、絶滅危惧種であるウバザメの摂食行動を研究するためのアプリケーションに始まった。ウバザメは巨大なサイズと移動性のために 3D 形状データを取得することがほぼ不可能であり、摂食行動と生態学の理解を妨げていた。一方、これらの動物の摂食位置の 2D 画像はすぐに入手できる。人間の棒モデルとサメの頭の骨格の両方で、提案アプローチを最先端の形状ベースのアプローチと比較した。トレーニング形状の小さなセットを使用して、Kendall 形状空間アプローチが以前の方法よりも大幅に堅牢であり、妥当な形状になることを示した。

これは、標本がまれであり、したがって利用できるトレーニング形状がほとんどないアプリケーションに好適である。

一般に、オブジェクトの形状は、回転、並進、およびスケーリングに対して不変であると見なされているため、商構造を持つ非線形空間で値をとる。特に、ランドマーク（特徴点）ベースの表現の場合、この概念はよく知られている Kendall 形状空間を生み出していく形状空間の非ユークリッドの性質は、線形性の仮定に反するだけでなく、グローバルなベクトル空間構造がないため、式

$$X = \sum_{j=1}^{n} c_j B_j, \qquad (2)$$

の代数式を直接適用することを妨げる。それにもかかわらず、(計算）微分幾何学のフレームワークは、再構成問題内で使用するための ASM（*active shape model*）アプローチを一般化するための豊富なツールセットを提供してくれる。

結果として得られる幾何学的モデルは、制約が自然にエンコードされるため、3D 形状の効率的かつ一貫性のある推定を提供している。以前のアプローチはより低い再投影エラーを達成したが、対応する 3D 形状は、ドメイン固有のさらなる正則化の必要性を示す非生理学的な歪みのような欠陥を示した。たとえば ASM の凸緩和は、基底形状が対称であるにもかかわらず、強い非対称性を発達させていた。対照的に、この Paskin らの幾何学的方法は、ターゲット形状がトレーニング分布と大きく異なる困難な状況でも、高度に生理学的な結果をもたらした。補助的な正則化なしで妥当な補間を行った。これは、推定された形状が、トレーニングデータがまたがる Kendall の形状空間の部分空間に常にあるためである。

三角形の頂点である特徴点と推定三角形の形状との関係が Paskin ら文献 [14] の Fig.1 に示されている。ウバザメ (basking shark) のような変形する形状である生物の口を開いた 2D 画像に対する 3D 形状推定結果を、上から見た図、横から見た図を示し、従来の ActiveShapeModel(ASM) による手法より、Kendall の形状空間による 3D モデルが優れていることを述べている。Paskin らの文献 [14]Fig 7 参照。

表 3　Shape Space

番号	著者	概要	性能
[13]	Klingenberg	Kendall の形状区間を歩く感覚で解説	解説のみ
[14]	Paskin	Kendall の手法で、2D 特徴点から 3D 再構成を行った	再構成が従来より良い

5. 同変学習による変形と CNN の融合

Luo ら [15] は、同変理論は、長年の関心事だが複雑すぎるなどの問題があったとの観点で、メッセージパッシング（グラフニューラルネットワーク）スキームに基づき、点群分析の同変量を達成するためのフレームワークを提案している。各ポイントの相対位置をポイントクラウド全体のグローバル・ポーズから切り離すために、各ポイントに方向を導入することで、同変プロパティを取得できることがわかった。したがって、各ポイントの向きを学習するモジュールを使用して、現在のメッセージ・パッシング・ネットワークを拡張した。ポイントの近隣からの情報を集約する前に、ネットワークはポイントの学習した方向に基づいて近隣の座標を変換している。提案されたフレームワークの同変量を示す正式な証明を示した。経験的に、提案された方法だが点群解析と物理モデリングタスクの両方で競争力があることを示している。

Luo らの提案手法では、最初に点群の各点の向きを学習する。情報を集約する前に隣接点の相対座標を投影することにより、モデルはグローバルな回転をうまく分離し、同変性を達成する。回転した点群データから、飛行機の他、室内のテーブル、椅子、ランプなどの識別を行い、各物体ごとに方向を検出し、基準位置、不変形状を確保している。Luo らの文献 [15]Fig.1 参照。

・方向の視覚化

Liu ら [15] は更に、点群分類モデルから学習した方向の 2 例（Airplane, Chair）を示している。学習した方向がグローバル構造と関連していることがわかる。たとえば、学習により、飛行機の翼と尾翼の最初の方向ベクトルは、翼と尾翼が伸びる方向に向かい、2 番目の方向ベクトルは翼に法線方向になっている。さらに、椅子の両前脚の最初の方向ベクトルは一貫して正面方向を向いており、学習した方向が構造上の類似性を認識していることを示している。

この論文では、同変点集合分析のスキームを紹介している。中心的な要素は、情報を集約するときに、点ごとの向きを学習し、学習した向きを使用して近隣座標を投影することだ。広範な実験により、モデルの有効性と一般性を実証している。

提案された方法では、ポイントごとの方向を推定するために追加のサブネットワークが必要となる。これは無視できない計算量のオーバーヘッドであり、向きに対する学習のコストを削減することが重要となる。この方法のもう 1 つの懸念は、ノイズ耐性だ。学習した向きがノイズの多いデータに対してどの程度脆弱であるかは不明で、これは現実世界の設定では避けられないので、今後の課題としている。

コードは https://github.com/luost26/Equivariant-OrientedMP で入手できる。

Musallam ら [16] は、姿勢推定において 3D ジオメトリベースの方法では end-to-end を達成することはまだできていないこと、さらに、絶対姿勢回帰は、画像検索との関連性が高いことが述べられている。その結果、従来の CNN によって学習された統計的特徴には、この本質的に幾何学的なタスクを確実に解決するのに十分な幾何学的情報が含まれていないという仮説を出している。この論文では、平行移動と回転の同変 CNN がカメラの動きの表現を特徴空間に直接誘導する方法を示している。次に、この幾何学的特性により、イメージプレーン保存変換のグループ全体でトレーニングデータを暗黙的に拡張できること

を示した。したがって、データ集約型の中間表現を学習するよりも、同変特徴を直接学習する方が好ましいと主張している。

包括的な実験的検証により、軽量モデルが標準データセットでの既存のモデルよりも優れていることが実証された。

Musallam らの文献 [16] の Fig.1 に、このアプローチを示されている。方法は、カメラの平面運動 R、t を直接エンコードするジオメトリ認識機能を抽出するために、並進および回転同変畳み込みニューラルネットワークを採用している。またカメラが動いている間、提案された特徴抽出器 F の同変量は、明示的な画像 ($\phi^{(II)}_{R,t}$) と特徴 ($\phi^{(F)}_{R,t}$) の変化の関係をつなげる。このプロパティは、絶対姿勢回帰の問題に対するより効率的なソリューションを提案するために活用される。

・この論文の貢献として、下記が挙げられている。

(1) 同変 CNN が SE(2) にある平面カメラの動きの表現を特徴空間に直接誘導する方法の定式化。（セクション 4.1）（SE(2) は 2 次元の特殊 Euclidean 群）

(2) SE(3) にあるカメラの動きを復元するために、SE(2) の同変特徴をどのように活用できるかについて、直感的な説明の提供。（セクション 4.2、SE(3) は 3 次元の特殊 Euclidean 群）

(3) E-PoseNet と呼ばれる軽量の同変ポーズ回帰モデルの導入。（セクション 5）

(4) E-PoseNet 広範な実験的評価は、標準データセットでの既存の APR (Absolute Pose Regression) メソッドと比較して、その競争力のあるパフォーマンス。（セクション 6）

この論文には、以下の解説が述べられている。深層学習における同変特徴に付いて。コンピュータービジョンには、手作りの同変機能の設計に関する豊富な歴史がある（たとえば、Scale-Invariant Feature Transform (SIFT)、Oriented フィルター、Steerable フィルター、Rotation-equivariant Fields of Experts (R-FoE)、リー群ベースのフィルター）。深層学習の文献では、畳み込み層は画像シフトに対して同変であることが証明されているが、最大プーリング層は入力画像の小さなシフトに対してのみ不変である。

畳み込み層は本質的に変換と同変だが、CNN によって正確にエンコードされていない入力に関する大量の空間情報がある。より具体的には、ローカルおよびグローバルプーリングが CNN に追加された場合、変換情報を回復不能にし、前述の同変条件が破棄される。

最近の調査では、CNN の多くのニューロンが、同じ基本機能のわずかに変換された（回転などの）バージョンを学習することが示されている。これらは、曲線検出器、高低周波数検出器、線検出器など、初期の視覚で特に一般的だ。

G-CNN をより広い変換グループに拡張する試みが行われている。マラットらは事前定義されたウェーブレットによる散乱変換を使用して、CNN を SE(2) と同変になるように拡張した。ベッカーズらは、また、Bsplines を介して CNN を SE(2) グループと同変になるように拡張した。

Cohen らは、90 度の回転と反転を介して p4m 離散グループと等価なグループ畳み込みネットワークを提案し、分類タスクに対するグループ畳み込みの有効性を実証した。（前記でも紹介済）

最近では、HaiweiChen による 3D 点群解、空中物体検出 (Jiaming)、2D 追跡 (Siamese) などのさまざまなコンピューター・ビジョン・タスクを解決するために、同変機能の使用が探究されている。Esteves らはオブジェクトの相対的な向きを推定するために、2D 画像か

ら球状 CNN 潜在空間への投影と埋め込みを使用することを提案した。同様に、Zhang らは、全方向ローカリゼーションにおけるカメラ姿勢推定の学習に球状 CNN を使用することを提案している。ただし、著者の Musallam らの知る限り、単一の 2D 入力画像の APR のコンテキストでは、同変量の機能はまだ明示的に活用されていないと述べている。

APR パイプラインの特徴抽出部分に同変操作を導入することに重点を置いているが、次の段階（つまり、埋め込み、回帰）には同じプロパティがないため、パイプライン全体の同変が成立しない。提案された APR モデルのもう 1 つの制限は、従来の CNN モデルと比較して、同変 CNN モデルに必要な計算時間が長いことだ。これはトレーニング中のみであり、推論時間は少ない。

この論文では、入力画像に関するより多くの幾何学的情報をエンコードするために、同変な特徴を活用するカメラ ポーズ回帰の問題の新しい方向性を提示した。SE(2)-equivariant 特徴抽出器を使用することにより、モデルは屋外と屋内の両方のベンチマークで既存の方法よりも優れたパフォーマンスを発揮できた。さらに、幾何学的推論に使用される深層学習モデルの同変特性は、絶対姿勢回帰の可能性に到達するための有望な方向性を提供すると結論付けている。

Atzmon ら [17] は、形状空間学習のタスクには、形状のトレーニングセットを、適切な一般化特性を持つ潜在表現空間に対してマッピングすることに着目している。実世界の物体の形状には対称性があり、形状の本質を変えない変換として定義できる。形状空間学習に対称性を組み込む自然な方法は、形状空間へのマッピング（エンコーダー）と形状空間からのマッピング（デコーダー）が対称性に対して同変であることを活かすことである。

この研究論文では、次の 2 つの貢献を導入することにより、エンコーダとデコーダに同変を組み込むためのフレームワークを提示している。

(i) 最近のフレーム平均化 (FA) フレームワークを適応させて、一般的で効率的で最大限に表現力のある同変オートエンコーダーを構築する。トレーニングでは、新しい損失の導入を必要とせずに、標準のオートエンコーダー再構成損失のみでよい。

(ii) 関節のある人体など、形状のさまざまな部分に適用された区分的なユークリッド運動に同変なオートエンコーダーを構築する。

著者 Atzmon らの知る限り、この論文は最初の完全に区分的なユークリッドの同変オートエンコーダーの構築であると述べている。フレームワークのトレーニングは簡単である。標準的な再構成損失を使用し、新しい損失を導入する必要はない。著者らのアーキテクチャは、標準（バックボーン）アーキテクチャで構築されており、適切なフレーム平均化を行って同変性を実現しているとのことである。

暗黙的なニューラル表現を使用した剛体形状データセットと、メッシュベースのニューラル ネットワークを使用した多関節形状データセットの両方でフレームワークをテストすると、目に見えないテスト形状への最先端の一般化が示され、関連するベースラインが大幅に改善された。特に、この方法は、目に見えない多関節ポーズへの一般化において大幅な改善を示している。

Atzmon らの文献 [17] の Fig_1 に区分ユークリッドの説明がなされている。メッシュ→メッシュでは、各部分の同変符号化（エンコード）には、同じ φ バックボーンが使用される。同様に、各部分の潜在コードの同変復号（デコード）には、同じ ψ バックボーンが使

用される。最後に、最終的な予測は、各パーツの同変出力メッシュの加重合計である。

形状空間を学習することは、目に見えないテスト形状にうまく一般化する入力トレーニング形状のコレクションに対する潜在的な表現を見つけるタスクである。これは多くの場合、オートエンコーダーフレームワーク、つまりエンコーダー$\Phi : X \to Z$　内で行われ、X の入力形状 (何らかの 3D 表現) を潜在空間 Z にマッピングすることである。デコーダ $\Psi : Z \to Y$、

Z の潜在的な表現を形状 Y に (おそらく X 以外の 3D 表現で) マッピングする。多くの一物体に対する入力は対称性を有す。つまり、形の本質を変えない変形である。たとえば、ユークリッド・モーション (回転、反射、平行移動) を家具などの剛体オブジェクトに適用すると、同等のバージョンのオブジェクトが生成される。同様に、動物や人間などの同じ多関節体が、空間でさまざまなポーズを取ることができる。形状空間学習に対称性を組み込む方法は、潜在空間（中間層）へのマッピングを定義することである。つまり、エンコーダー、および潜在空間からのマッピング（デコーダー）は、関連する対称性に対して同変になる。つまり、入力形状に対称性を適用してからそれをエンコードすると、元の形状の潜在コードに適用されるのと同じ対称性が得られ、同様に、変換された潜在コードから形状を再構築すると（不要な情報が除かれて）もとの形状になる。

モデルが単一の形状を学習した場合、モデルはすべての対称バージョンに完全に一般化できる。残念ながら、グローバル ユークリッド モーションのおそらくより単純な設定であっても、表現力と効率性を兼ね備えた同変ニューラル ネットワークを構築することは依然として課題です。ユークリッド運動同変関数に普遍的であることが知られている唯一のアーキテクチャは、テンソル フィールド ネットワークとグループ平均化であり、どちらも計算量が多く、メモリ集約的です。Vector Neurons などの他のアーキテクチャは、計算上は効率的ですが、普遍的であるとは知られていない、と述べられている。

形状空間学習のコンテキストで、構成によって対称性をエンコーダー・デコーダーに組み込むための一般的な方法論を導入した。フレーム平均化を使用して、表現力豊かで効率的な同変オートエンコーダーを構築する方法を示した。グローバルおよび区分的なユークリッド モーションの場合、メッシュ→メッシュ、点群→暗黙のシナリオのフレームワークをインスタンス化した。すべての実験で最先端の定量的および定性的な結果を達成した。

この方法にはいくつかの制限がある。まず、メッシュ→メッシュの場合、固定接続とスキニング ウェイトを使用する。

区分的なユークリッドのケースを暗黙的な表現に一般化すること、複数のオブジェクトを含む大規模なシーンを処理すること、またはスキニングの重みを学習することは、将来の課題としている。点群データの学習の様子が文献 [17] の Fig.2 等に示されている。

入力点群、暗黙の学習に他の教師データを使用していない。テディベア (747 ポイント クラウド)、ボトル (296 ポイント クラウド)、スーツケース (480 ポイント ラウド)、バナナ (197 ポイント クラウド) の 4 つのオブジェクトカテゴリを使用している。70% から 30% の分割に基づいてランダムにセットを分割した。VAE (Variational Auto-Encoder) は変分オートエンコーダーを示す。この VAE の VectorNeurons バージョンを、VN と表示。区分的ユークリッド人体の動きに対する再構成例が、文献 [17] の Fig.3 等に示されている。AE、ArapReg(As-Rigid-As-Possible Regularization）よりは優れ、正解の例とはあまり変わ

らない。

D.Chen ら [18] は、医用画像に対し、end-to-end（最初から最後まで全自動：部分研究ではない）の剛体（内蔵等）の部分画像から完全な画像への同変再構成を発表している。ノイズを加えた耐性評価で、ノイズの性質を与える必要がある。この同変イメージングでは、同変にしたいのはネットワークではない。

5.　剛体回転による変形をと CNN に組み込む方式

これまでの調査で、回転や平行移動対称性のある物体に対し、同一物体に対して、視点が異なることで、形状が変化し、CNN の学習が不完全になることが知られており、その対策が鋭意なされている様子が分かった。物体の回転は容易な処理で認識できそうな感があるが、CNN への組み込みを行うには、大変な工夫を行う必要がある。剛体や可塑性のある人体などに対する研究がそれぞれ行われている。本章では、自動運転における車両認識・検出に於いて剛体をターゲットとした、回転による変形で、射影歪を含む場合について、参考となる文献を調査する。

自動運転では、車両の認識（識別）はかなり の精度が達成されているが、天候、夜間などの広い環境での完全な認識には至っていない。そのため、性能を更に向上させる必要がある。学習データの増強だけではなく、NN の構成の改良も必要である。元々単一の形状の物体が、視点位置により別の形状になることに対応する、形状空間の同変学習が回転や平行移動という線形だが回転する多種の学習データを与えないと学習しないことから、学習データを増強するか、学習ネットワークで回転に対応するかという 2 つの方向性を抽出できた。同変学習には点群データに対してではあるが、実装プログラムが公開されており、3D 回転による射影変換の要素を追加することが可能になっている。

Marcos ら [19] は、テクスチャを考慮する場合、回転、平行移動、スケーリングなどの特定の変換に対して不変な外観記述子を使用した。これは、ほとんどの場合、カメラに対する相対的な位置に関係なく、マテリアルを識別しなければならないためである。多くのコンピューター ビジョンの問題が任意の向きの画像を考慮していることを考えると、回転不変性は特に重要である。取り上げた例は、リモート センシング画像と顕微鏡画像だ。

バイキュービック補間による標準の画像回転を実装し、R は合計数です。

各回転グループで考慮される角度。一部の回転でフィルターの外側にあるピクセルの影

表 4　同変 CNN 学習

番号	著者	概要	性能
[15]	Luo	点群データ、同じ物体が 3D 多方向に回転している時、その方向を CNN で検出できた。	ノイズ耐性は不明
[16]	Musallam	従来技術のレビューが豊富。平行移動と回転の同変 CNN がカメラの動きの表現を特徴空間に直接誘導する方法を開発し、回転のデータ増強で CNN 学習を進めた。	
[17]	Atzmon	剛体と人体等の可塑性動きの単一物体の同変学習する。Enc-Dec で対称の変形を自動生成する。	目に見えない部分の一般化で改善
[18]	D. Chen	医用画像、end-to-end での耐ノイズ同変学習	

響を無視することが重要です。したがって、正方形フィルターに外接する円内のピクセルのみを使用する。

この畳み込み層の後、向きの次元に適用される max-pooling 操作を適用し、同じ回転グループのアクティベーション全体で最大値を求める。これにより、ローカル回転に対する不変性が保証され、したがって入力画像のグローバル回転に対する同変性が保証さる。図 4 に、この不変性がどのように得られるかを示す。バックワード パスでは、標準の max-pooling の動作と非常によく似た、最大のアクティベーションを生成する角度の勾配が渡される。向きの最大プーリングをアクティブにする角度のインデックスは、フォワードパス中に記録され、フィルターの重みが更新される。

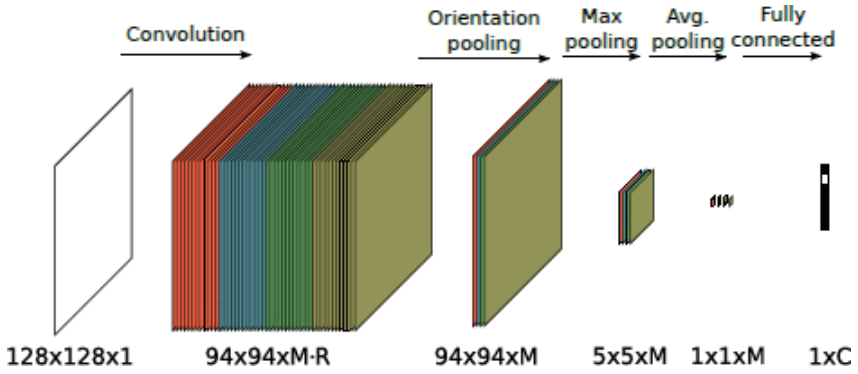

図 4　回転フィルターを有する CNN（Marcos らの文献 [19] の Fig2 より引用）

モデルは各トレーニングイメージの複数の場所（最大プーリングウィンドウごとに一つずつ）から学習できる。M は回転グループの総数（つまり、固有の標準フィルターの総数）であり、R は各グループ内で考慮される個別の向きの数で、C は分類問題のクラス数である。CNN パラメーターは畳み込みフィルターで重み減衰を使用している（ただし、バイアスでは使用していない）。Weight Decay は、小さな大きさのパラメーターを優先する正則化である。これにより、誤った大きなパラメーターが排除され、収束が促進される。

この論文では、標準的な畳み込みニューラル ネットワーク (CNN) の定式化を使用して、回転不変の回転可能なフィルターを明示的に学習する戦略を提案している。テクスチャ分類の識別フィルターバンクを学習するときに、回転不変性を明示的に考慮することの利点を示している。これらの結果は、各フィルターがグループ内の他のフィルターの回転バージョンになるように、浅い CNN の最初のレイヤーでフィルターの各グループの重みを結び付けることによって達成した。これらの回転可能なフィルターの高い表現力と、それに続く学習するパラメーターの数の削減により、回転不変テクスチャ分類のベンチマークで最先端技術を満たすのに十分なパフォーマンスの向上が実現した。

さらに、提案された方法論が、特に小さなトレーニングセットのシナリオで、標準のデータ拡張 CNN よりも大幅に優れていることを示している。

5. まとめ

以上、自動運転の車両検出において、車両の形状がカメラの位置の違いによる視点の変化で形状が 3D 的に変わり、射影歪を保証する必要がある状態での方式の検討を行った。一つは、各種の回転をする画像をデータ増強で用意し、学習を深めるやり方、もう一つは剛体の同変性を利用する同変学習により CNN の内部で対応させるやり方がある。更に大きくは、形状区間での変形の一般化で対処する研究が多くなっていることが分かった。回転対称性を CNN に埋め込む手法では、依然として固有性が高く、汎用性が達成されていない。そこで、データ増強と、剛体の回転や変形の程度が小さい対象に限定して、CNN の改造を進めていくことが堅実な方向と考えられた。

図 5 は車両の画像を入力し、3D のワイヤーフレームをマッチングさせ、水平回転を与えて、形状変化した画像を得る例を示している。これは、データ増強にも使用可能であるまた、図 6 に示すように学習ネットワークに図 5 を埋め込んで使うことへの拡張については今後行っていく。

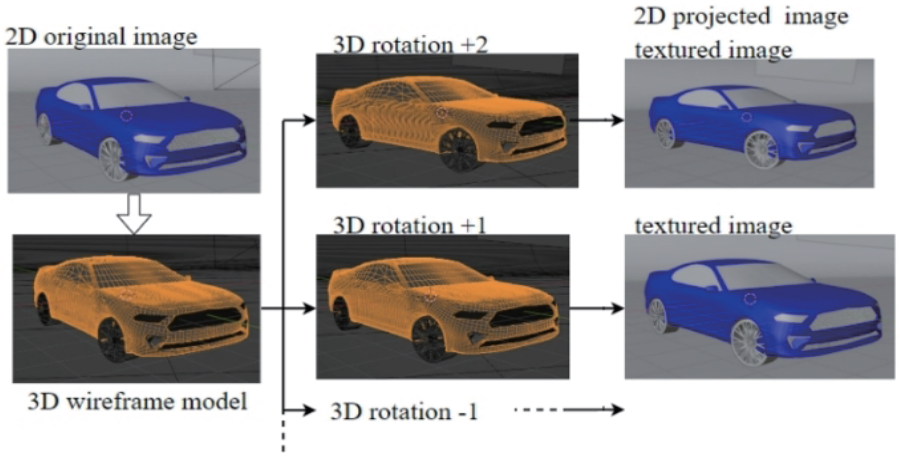

図 5　入力画像から 3D データを生成する構成

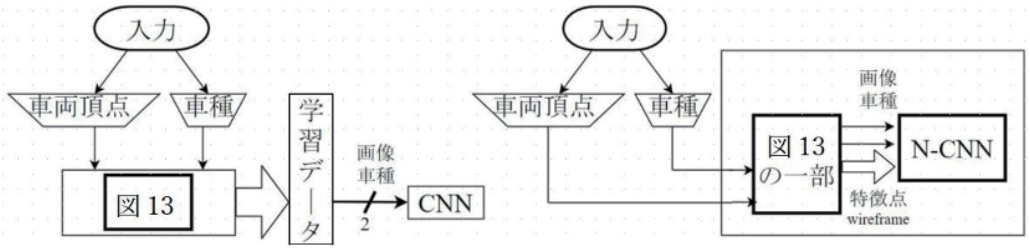

図 6(a) 学習データ数の回転による増強　　(b)CNN での回転の学習では特徴点をタグに追加

参考文献

[1] Research Port(リサーチポート)「CVPR2022」ResearchPort トップカンファレンス定点観測 vol.3　https://research-p.com/column/561

[2] Open Access versions, provided by the Computer Vision Foundation https://openaccess.thecvf.com/CVPR2022?day=all

[3] Y. Bao, Y. Liu, H. Wang and F. Lu, "Generalizing Gaze Estimation with Rotation Consistency," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4197-4206.

[4] G. Bökman, F. Kahla and A. Flinth, "ZZ-Net: A Universal Rotation Equivariant Architecture for 2D Point Clouds," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10966-10975.

[5] X. Feng, X. Yao, G. Cheng and J. Han, "Weakly Supervised Rotation-Invariant Aerial Object Detection Network," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14126-14135.

[6] R. Chen and Y. Cong, "The Devil is in the Pose: Ambiguity-free 3D Rotation-invariant Learning via Pose-aware Convolution," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7462-7471.

[7] A. Seo, B. Kim, S. Kwak and M. Cho, "Reflection and Rotation Symmetry Detection via Equivariant Learning," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9529-9538.

[8] H. -X. Yu, J. Wu and L. Yi, "Rotationally Equivariant 3D Object Detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1446-1454.

[9] J. Kang, R. Fernandez-Beltran, Z. Wang, X. Sun, J. Ni and A. Plaza, "Rotation-Invariant Deep Embedding for Remote Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022, Art no. 5509713.

[10] M. Zand, A. Etemad and M. Greenspan, "Oriented Bounding Boxes for Small and Freely Rotated Objects," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp.1-15, 2022, Art no. 4701715.

[11] T. Zhang et al., "Foreground Refinement Network for Rotated Object Detection in Remote Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022, Art no. 5610013,.

[12] Taco S. Cohen, and Max Welling, “Group Equivariant Convolutional Networks”, Proceedings of the 33rd International Conference on International Conference on Machine Learning ICML'16 - Volume 48 June 2016 Pages 2990–2999.

[13] Klingenberg, C.P. Walking on Kendall’s Shape Space: Understanding Shape Spaces and Their Coordinate Systems. Evol Biol 47, 334–352 (2020). https://doi.org/10.1007/s11692-020-09513-x

[14] Paskin, M., Baum, D., Dean, M.N., von Tycowicz, C. (2022). A Kendall Shape Space Approach to 3D Shape Estimation from 2D Landmarks. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13662. Springer, Cham.

[15] S. Luo et al., "Equivariant Point Cloud Analysis via Learning Orientations for Message Passing," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18910-18919.

[16] M. A. Musallam, V. Gaudillière, M. O. Del Castillo, K. Al Ismaeil and D. Aouada, "Leveraging Equivariant Features for Absolute Pose Regression," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6866-6876.

[17] M. Atzmon, K. Nagano, S. Fidler, S. Khamis and Y. Lipman, "Frame Averaging for Equivariant Shape Space Learning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 621-631.

[18] D. Chen, J. Tachella and M. E. Davies, "Robust Equivariant Imaging: a fully unsupervised framework

for learning to image from noisy and partial measurements," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5637-5646.

[19] D. Marcos, M. Volpi and D. Tuia, "Learning rotation invariant convolutional filters for texture classification," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2012-2017.

大関和夫　東京国際工科専門職大学　工科学部　情報工学科　教授
上條浩一　東京国際工科専門職大学　工科学部　情報工学科　教授