InsAT: 周囲の物体や他者との関係を考慮するキーポイントベース行動認識

筒川 和樹 *1,a)　長野 紘士朗1,b)　佐藤 文彬1,c)

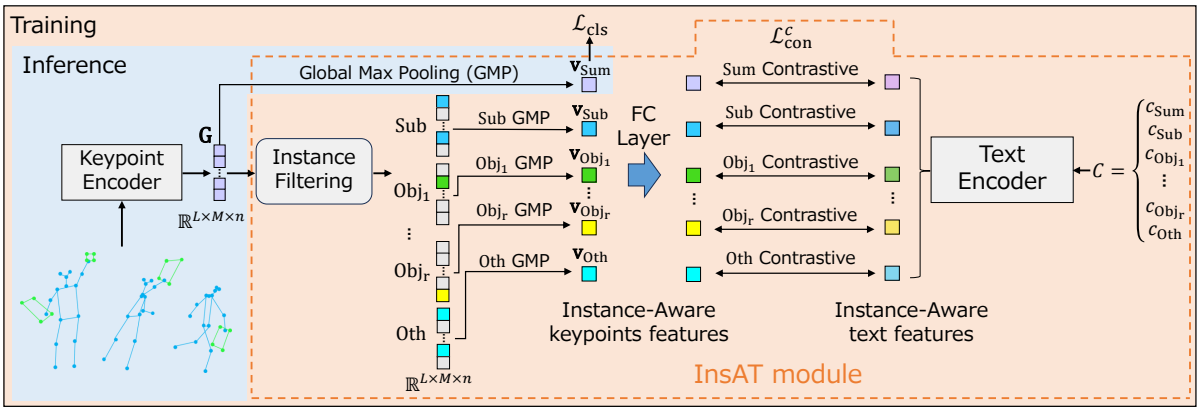

図 1: InsAT フレームワークの学習時と推論時の概要．InsAT モジュールは学習時のみ使用される．

概要

キーポイントベースの行動認識において，人物と周囲の物体や他者との関係性を明示的にモデル化する新たな学習フレームワーク **Ins**tance-**A**ware **T**raining Paradigm (InsAT) を提案する．従来の手法は，主に人物の動作情報に依存し，行動の理解に不可欠な周囲の物体や他者との関係を十分に捉えられない課題がある．例えば，「Playing tennis」では，プレーヤーの動作だけでなく，ラケットやボール，他のプレーヤーとの相互作用もその行動の理解に不可欠である．InsAT はこの課題を克服するため，行動の種類に応じて，人物の動作だけでなく周囲の物体や他者との関係性も詳細に記述する説明文を生成し，それらの説明文に対応する人物や物体のキーポイント特徴と結びつける学習により，行動を周囲の環境を考慮して効率的に学習する．具体的には，Text Encoder で得たインスタンスごとのテキスト特徴と，Keypoint Encoder で抽出したインスタンスごとの特徴を対照学習により整合させ，人物–物体および人物–人物の関係を的確に捉える．さらに，物体検出結果に過度に依存しないインスタンス特徴抽出手法を導入し，誤検出や未検出による影響を低減する．実験により，InsAT が行動の文脈的理解を深め前述の課題を解決し，最先端のキーポイントベース行動認識手法を上回る精度を達成することを示す．

1　コニカミノルタ株式会社
a)　kazuki.tsutsukawa@konicaminolta.com
b)　koshiro.nagano@konicaminolta.com
c)　fumiaki.sato1@konicaminolta.com
*　責任著者．

1. はじめに

動画中の人物の行動を認識する行動認識タスクは，ロボティクス [15,21] や監視カメラ [5,10,29,30] などのさまざまな応用分野で重要である．キーポイントベースの行動認識手法 [4,7,9,14,19,22,24,25,27,28,32,33,35,36] は，複数人姿勢推定技術 [2,23,31] や深度センサ [11,37] によって検出された低情報量のキーポイントを入力とすることで，人物やシーンの外観変化に対してロバストな特徴を持つ [9]．しかし，多くの従来手法は，主に人物の動作情報に依存しており，行動に関わる周囲の物体や他者との関係を十分に考慮できていない [34,38] 課題がある．例えば，「Playing tennis」を正しく認識するには，プレイヤーの動作だけでなく，ラケットやボール，さらにはプレイヤー同士の関係を考慮する必要がある．

本研究では，周囲の環境を考慮したキーポイント行動認識を実現するために，人物の関節と物体の外輪郭キーポイントを入力とし，人物–物体および人物–人物の関係性を明示的にモデル化する新たな学習フレームワーク **Ins**tance-**A**ware **T**raining Paradigm (InsAT) を提案する．具体的には，インスタンスごとのキーポイント特徴と，対応する人物–物体および人物–人物の関係性を表すテキスト特徴を対照学習により整合させることで，周囲の物体や他者との関係も含む豊かな文脈を学習し，前述の課題に対処する．

1.1 従来研究の課題

キーポイントベース行動認識手法において，周囲の物体との関係を考慮して認識を行うため，人物の関節点に加

えて物体のキーポイントを入力する手法が提案されている [1,6,9,12]．特に Structured Keypoint Pooling (SKP) [9] は，人物の関節点と物体検出結果から得られる物体キーポイントを統合することで，周囲の物体が関わる行動認識で有効性を示している．しかし，これらの従来手法には以下の課題がある．

課題 1．人物–物体の関係の明示的モデリングの不足

物体のキーポイントを追加することで人物の周囲の情報を補うが，どの物体がどのように行動に関与するかを明示的にモデル化していない．そのため，行動理解に不可欠な人物と物体の相互作用を明確に捉えられない．

課題 2．行動に無関係な物体情報による学習の困難性

入力には行動と無関係な物体の情報も含まれうるため，学習が困難になる可能性がある．特に，シーン内に複数の物体が存在する場合，行動に関連する物体を適切に選択できないと，学習が妨げられる．

課題 3．学習データ効率の低さ

物体との相互作用をデータドリブンに獲得するため，少量のデータでは十分な学習が難しく，精度が制限される．

1.2 提案法の概要と貢献

提案法は，Keypoint Encoder と Text Encoder の 2 つの主要なコンポーネントから構成される．Keypoint Encoder は，人物や物体の時空間情報を処理し，各インスタンスの特徴を保持しながら動画全体のグローバル特徴を抽出する．一方，Text Encoder は，行動全体や行動主体となる人物，周囲の物体や他者との関係性を詳細に記述する説明文をテキスト特徴に変換する．これらの特徴の意味的な対応関係を学習するために，グローバルおよびインスタンスごとのキーポイント特徴と対応するテキスト特徴の間で対照損失を適用する．この対応付けにより，行動全体の文脈や人物–物体，人物–人物の関係性が学習され，従来研究の課題を解消し，より精度の高い行動認識が可能となる．

本研究の主な貢献は以下のとおりである．(1) キーポイントベース行動認識において，行動ラベル以外の追加アノテーションなしに，周囲の人物や物体との関係を明示的にモデル化する学習フレームワーク (InsAT) を提案する．(2) インスタンスごとの説明文とキーポイント特徴を対応付けることで文脈を補強し，データ効率の高い学習を実現する．(3) 複数のキーポイントベース行動認識ベンチマークで，最先端手法を上回る精度を達成する．

なお，本手法の枠組みは，ゼロショット認識にも拡張可能であり，ゼロショット行動認識のための KIND が提案されている [39].

2. 提案法

従来の行動ラベルによる分類損失を用いた学習方式に，InsAT モジュールを導入することで，行動の概要に加え，人物の詳細な動作や周囲の環境を考慮した学習を行う．Keypoint Encoder は，学習と推論の両段階で，すべてのインスタンスの特徴 $\mathbf{G}$ を抽出し，Global Max Pooling (GMP) により動画全体の特徴へ集約する．学習時には，行動の概要や個別のインスタンスに関する説明文を用い，これらのテキスト特徴と対応するインスタンス特徴を，対照学習によって整合させ，行動の文脈を捉えた意味表現を獲得する．推論時には，学習時に得た各インスタンス特徴が表す文脈に適した表現を動画全体の特徴へ集約し，最終的に Fully-Connected(FC) 層を介して行動クラスを分類する．

2.1 入力と Keypoint Encoder

入力は，フレーム数 L，各フレーム内の最大インスタンス数 M (人物と物体)，各インスタンスの最大キーポイント数 K からなるキーポイント群 $\mathbf{I} \in \mathbb{R}^{L\times K\times M\times 4}$ である．各キーポイントは座標 (x, y)，検出確信度，物体カテゴリ ID の 4 次元の情報を持つ．本手法の Keypoint Encoder には，インスタンス間の特徴の関係を重視するため，[9] で検証された SKP のアーキテクチャのうち，インスタンス単位の特徴を保持する方式を用いる．この Keypoint Encoder f_{key} は，インスタンス単位の特徴 $\mathbf{G} = f_{\text{key}}(\mathbf{I}) \in \mathbb{R}^{L\times M\times n}$ を出力する．ここで，n は各インスタンス特徴の次元数である．$\mathbf{G}$ から，2.2節で説明する 4 種類の説明文に対応するグローバルおよび各インスタンスの特徴を 2.3節の方式で抽出し，GMP により動画単位の特徴に集約する．

2.2 周囲の環境を含むテキスト特徴生成

大規模言語モデル (LLM) を知識源として用いることで，人手によるラベル付けを行わず，統一基準の説明文を生成する．行動を包括的かつ詳細に記述するための指示文を，行動ラベル名とともに LLM に与え，次の 4 種類の説明文を生成する．**(1) Action Summary (Sum)**: 行動ラベルの全体像を包括的に記述する．**(2) Subject Behavior (Sub)**: 行動主体の人物の動作や姿勢に焦点を当て，身体部位の状況を詳細に記述する．**(3) Relevant Object (Obj)**: 行動ラベルと関連の深い物体を列挙し，それぞれの物体における人物との関係や役割を記述する．**(4) Relevant Others (Oth)**: 複数人が関与する行動における関係性を記述する．Text Encoder (text-embedding3-large [18]) により，これらの説明文のテキスト特徴 $\left\{\mathbf{t}_{\text{sum}}, \mathbf{t}_{\text{sub}}, \mathbf{t}_{\text{obj}_1}, ..., \mathbf{t}_{\text{obj}_r}, \mathbf{t}_{\text{oth}}\right\}$ を得る．

2.3 説明文に対応するインスタンス特徴のフィルタリング

本手法では，物体検出結果に過度に依存せず，説明文と関連の強いインスタンス特徴を抽出する Hybrid Instance Filtering (HIF) 手法を提案する（図 2）．HIF では，すべてのインスタンスの特徴 $\mathbf{G}$ にフィルタリング関数 $h_{\text{hyb}}(\cdot)$ を適用し，説明文に対応するインスタンス特徴を抽出する．インスタンス特徴を抽出する際，物体検出誤りによるノイズや未検出の影響を受けやすいため，物体検出結果に基づく Detection-based Instance Filtering (DIF) に，説明文とキーポイント特徴との関連度に基づく Semantic-based Instance Filtering (SIF) を組み合わせる．

DIF は，物体検出結果から得られた物体のカテゴリ ID に基づき，説明文が対象とする物体カテゴリ ID と一致するインスタンスの特徴 $\mathbf{S}_{\text{det}}$ を抽出する．一方，SIF は，説

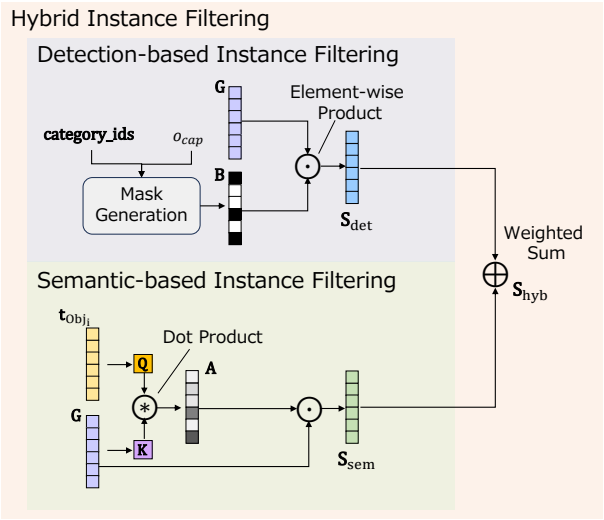

図 2: 説明文に対応するインスタンス特徴の抽出

明文のテキスト特徴との関連度に基づき，各インスタンスに重みづけを適用してインスタンス特徴 $\mathbf{S}_{\text{sem}}$ を抽出する．これらを重み付き和で統合し，説明文に対応するインスタンス特徴 $\mathbf{S}_{\text{hyb}}$ を次式で得る．

$$\mathbf{S}_{\text{hyb}} = h_{\text{hyb}}(\mathbf{G}) = \lambda \times \mathbf{S}_{\text{det}} + (1-\lambda) \times \mathbf{S}_{\text{sem}}. \quad (1)$$

ここで，入力 $\mathbf{I}$ の各キーポイントの物体カテゴリ ID を参照し，物体カテゴリ ID をインスタンスごとに格納した行列 $\mathbf{category\_ids} \in \mathbb{R}^{L\times M}$ を定義する．DIF では，$\mathbf{category\_ids}$ の各要素と，2.2節で生成する説明文が対象とする物体カテゴリ ID o_{cap} を比較してマスク行列 $\mathbf{B} \in \mathbb{R}^{L\times M}$ を生成する．$\mathbf{B}$ の各要素の値 $b(l,m)$ は，次式のように定義される．

$$b(l,m) = \begin{cases} 1, & \text{if } \mathbf{category\_ids}(l,m) = o_{cap} \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

そして，$\mathbf{B}$ と $\mathbf{G}$ の要素積により，DIF によるインスタンス特徴 $\mathbf{S}_{\text{det}}$ を得る．

$$\mathbf{S}_{\text{det}} = h_{\text{det}}(\mathbf{G}) = \mathbf{B} \odot \mathbf{G}. \quad (3)$$

DIF は特定のインスタンスに焦点を当てることができる手法だが，物体種別の誤りや未検出により，誤った特徴の選択や対象とすべき物体の情報の欠落が生じる可能性がある．

SIF は，説明文との関連度の高いインスタンスを選択することで，物体検出への依存を低減する．PURLS [38] に着想を得て，cross-attention ベースの SIF モジュールを設計する．SIF は，すべてのインスタンス特徴 $\mathbf{G}$ に対して，説明文との関連度に応じて各インスタンスに重み付けを行い，特徴を更新する．具体的には，$\mathbf{Obj}$ のテキスト特徴に基づくクエリ $\mathbf{Q}$ と，すべてのインスタンス特徴 $\mathbf{G}$ から得たキー $\mathbf{K}$ の行列積を計算し，softmax 正規化により，各インスタンスの説明文との関連度を表す行列 $\mathbf{A} \in \mathbb{R}^{L\times M}$ を計算する．この行列 $\mathbf{A}$ と $\mathbf{G}$ の要素積により，説明文に対応するインスタンス特徴 $\mathbf{S}_{\text{sem}} \in \mathbb{R}^{L\times M\times n}$ を得る．

$$\mathbf{S}_{\text{sem}} = h_{\text{sem}}(\mathbf{G}) = \mathbf{A} \odot \mathbf{G}, \quad (4)$$

ただし，$\mathbf{Q}$ は $\mathbf{t}_{\text{obj}_i}$ と $\mathbf{W_Q}$ の行列積，$\mathbf{K}$ は $\mathbf{G}$ と $\mathbf{W_K}$ の行列積により計算される．$\mathbf{W_Q} \in \mathbb{R}^h$，$\mathbf{W_K} \in \mathbb{R}^{FM\times h}$ は，それぞれ $\mathbf{Q}$ および $\mathbf{K}$ の特徴を変換する重み行列である．

2.4 インスタンス毎に詳細にモデリングする行動分類学習

図 1に示す通り，推論時は，Keypoint Encoder で抽出したすべてのインスタンスの特徴 $\mathbf{G}$ を GMP により動画全体の特徴 $\mathbf{v}_{\text{Sum}}$ に変換し行動分類を行う．学習時は，次式に示す通り，特定の行動クラスへの分類損失と，2.2節に示す，最大 4 種類の説明文と対応するインスタンスの特徴をそれぞれ整合させる対照学習損失を最小化する．

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{B} \mathcal{L}_{\text{cls},i} + \alpha \sum_{c=1}^{N_{\text{cap}}} \beta_c \mathcal{L}^c_{\text{con}}, \quad (5)$$

ただし，$\mathcal{L}_{\text{cls}}$ はクラス分類損失，B はバッチ数，$\mathcal{L}^c_{\text{con}}$ は説明文 c に対する対照損失，N_{cap} は説明文の種類数である．対照学習は [34] に基づき，Kullback-Leibler(KL) ダイバージェンス損失を用いる．説明文 c に対応するバッチ内のキーポイント特徴 $\mathbf{V}_{\text{c}}$ と対応するテキスト特徴 $\mathbf{T}_{\text{c}}$ との双方向の類似度を確率分布に変換し，次式で損失を定義する．

$$\mathcal{L}^c_{\text{con}} = \frac{1}{2}\mathbb{E}_{\mathbf{V}_c,\mathbf{T}_c\sim D}\begin{bmatrix}\text{KL}(\mathbf{P}^{v_c 2t_c}(\mathbf{V}_c), \mathbf{Y}^{v_c 2t_c}) \\ +\,\text{KL}(\mathbf{P}^{t_c 2v_c}(\mathbf{T}_c), \mathbf{Y}^{t_c 2v_c})\end{bmatrix}, \quad (6)$$

ただし，$\mathbf{P}(\mathbf{V}_c) \in \mathbb{R}^{B\times B}$ は，バッチ内の両特徴の類似度行列を sortmax 正規化して得られる確率分布である．D はデータセット全体，正解ラベル $\mathbf{Y} \in \mathbb{R}^{B\times B}$ は，ネガティブペアに対しては 0，ポジティブペアに対しては 1 の確率を持つ．ポジティブペアの説明文 c が対象とする物体が未検出の場合，対応するラベル確率 $\mathbf{Y}(i,j)$ を 0 に設定し損失項目から除外する．キーポイント特徴 $\mathbf{v}_c$ は，テキスト特徴 $\mathbf{t}_c$ との類似度計算のために，FC 層を通じて $\mathbf{t}_c$ と同じ次元数のベクトルに変換される．

3. 実験

本研究では，以下に示すデータセットを用い，Top-1 Accuracy (Acc.) により評価を行う．

Kinetics-400. YouTube の動画から収集された 400 種類の行動を含む大規模データセット [3] であり，学習，評価データはそれぞれ 250K，19K の 10 秒間の動画を含む．

Kinetics-1/10, Kinetics-1/20. 小規模データでの学習効率を検証するため，Kinetics-400 の学習データを各クラスごとに 1/10 および 1/20 にランダムサンプリングし，Kinetics-1/10 および Kinetics-1/20 データセットとする．評価データは Kinetics-400 と同じとし，ランダム性の影響を考慮して，3 種類のサンプル (split1-3) を用いる．

HMDB51. YouTube から収集された 6.7K 動画から構成される 51 種類の行動を含むデータセット [13] であり，[8,9] に基づき，学習，評価データを split1 の定義で分割する．

3.1 キーポイントベース行動認識の最先端手法との比較

表 1に示すように，提案法は Kinetics-400 および

表 1: 複数のデータセットにおけるキーポイント行動認識の最先端手法との行動分類精度の比較.

(a) Kinetics-400

Method	Acc. (%)	Object Keypoints	Keypoint Detector	COCO AP_{kp}(%)
MS-G3D [16]	45.1	-	HRNet [31]	74.6
PoseConv3D [8]	47.7	-		
SKP [9]	50.3	-		
SKP [9]	58.0	✓	PPNv2 [23]	68.5
Ours	**59.3**	✓		

(b) HMDB51

Method	Acc. (%)	Object Keypoints
PoseConv3D [8]	69.7	-
SKP [9]	70.9	-
SKP [9]	71.2	✓
Ours	**72.3**	✓

表 2: Kinetics-400 の異なる学習データ数に対する行動分類精度の比較.

Method	Ratio of scale		
	1/20	1/10	1
SKP [9]	26.3	34.8	58.0
Ours	28.0	36.9	59.3

表 3: 説明文の違いによる行動分類精度の比較．Kinetics-1/20 に対する実験結果.

Sum	**Sub**	**Obj**	**Oth**	Acc. (%)
–	–	–	–	26.2
✓	–	–	–	26.7 (↑ 0.5)
–	✓	–	–	**27.4** (↑ 1.2)
–	–	✓	–	26.8 (↑ 0.6)
–	–	–	✓	26.4 (↑ 0.2)
✓	✓	–	–	27.4 (↑ 1.2)
✓	✓	✓	–	27.6 (↑ 1.4)
✓	✓	✓	✓	**27.9** (↑ 1.7)

表 4: Kinetics-400 を用いた異なるインスタンス特徴フィルタリングの設計に対する行動分類精度の比較.

Method	Acc. (%)
DIF	58.9
SIF	59.0
Ours (HIF)	**59.2**

HMDB51 両データセットで，ベースライン（SKP）をそれぞれ 1.3%pt，1.1%pt 上回る．これらの結果より，提案法は，周囲の物体や人との関係を明示的にモデリングし，従来の学習方式と比較して人手による正解付けコストや推論時の計算コストを増加させることなく，モデルの精度を向上できることがわかる．

3.2 小規模データセットに対する提案法の有効性

Kineitcs-1/10 および Kineitcs-1/20 に対して, ベースラインとなる SKP と提案法をスクラッチ学習し，それらの結果を表 2に示す．提案法は SKP を Kinetics-1/10 で 2.1%pt，Kinetics-1/20 で 1.7%pt 上回り，Kinetics-400 と比較して小規模データでの改善幅が大きい（1.3%pt vs. 2.1%pt，1.7%pt）．また，split1〜3 の実験に対する分類精度の標準偏差は，kinetics-1/10 で 0.52% vs. 0.35%，Kinetics-1/20 で 0.53% vs. 0.13% (SKP vs. 提案法) であり，提案法はより精度のばらつきが小さい．これらの結果から，提案法は学習データ効率が高く，小規模データに対して安定した学習手法であることがわかる．

3.3 アブレーションスタディ

対照学習に用いる説明文の組み合わせ方法． 説明文のテキスト特徴とそれらに対応するキーポイント特徴に適用する対照損失の組み合わせを変えた実験結果を表 3に示す．各説明文を単独で使用した場合（上から 2〜5 行目），すべての説明文が精度向上に寄与し，特に **Sub** (↑ 1.2%pt)，次いで **Obj** (↑ 0.6%pt) の効果が大きい．さらに，説明文を順に追加した場合（2 行目および，下の 3 行）を比較すると，すべてを組み合わせた場合 (↑ 1.7%pt) が最も効果が大きい．これらの結果から，**Sub** や **Obj** のようなインスタンス単位の行動モデリングが有効であること，それらを組み合わせる方式が有効であることがわかる．

説明文に対応するインスタンス特徴のフィルタリング手法の影響． 表 4に 2.3節で述べた 3 種類のインスタンス特徴のフィルタリング手法に対する結果を示す．SIF および HIF は，物体に関する説明文のテキスト特徴である $\mathbf{t}_{\text{obj}}$ に対応するインスタンス特徴の抽出に適用し，$\mathbf{t}_{\text{obj}}$ 物体検出誤りの影響を低減する．実験の結果，DIF と SIF 手法を組み合わせる HIF が最も高い精度を達成している．この結果から，物体検出に基づくインスタンス特徴と，説明文に基づくインスタンス特徴が相補的に機能し，行動認識の精度向上に寄与することがわかる．

対照損失の重みの影響． 式 5の対照損失の重み α の影響を確認するため，$\alpha \in \{0.8, 1.0, 1.2, 1.5, 1.8\}$ の範囲で Kinetics-400 において精度比較を行った．各 α に対応する精度は 59.1%，59.2%，59.2%，59.3%，59.3% となり，α の変動による分類精度の影響は小さいことがわかる．

4. おわりに

本研究では，キーポイントベースの行動認識において，人物と物体，さらには他者との関係を明示的にモデル化する InsAT を提案した．従来手法が主に人物の動作に依存していたのに対し，InsAT は行動の種類に応じた詳細な説明文と，対応するキーポイント特徴との整合性を学習することで，文脈を考慮した行動理解を実現する．説明文の生成では LLM の知識を有効に活用することで，人手による追加のアノテーションを必要としない学習フレームワークの構築が可能となった．また，説明文に対応するキーポイント特徴の抽出においては，物体検出結果だけに依存せずに，説明文に基づいて関連インスタンスを選択する方式を導入することで，検出誤りの影響を低減した．

複数の行動認識ベンチマークにおいて InsAT は従来の最先端の精度を上回り，特に少量データ環境でも有効であることが確認された．

今後は，新たな物体カテゴリへの拡張や，人物と物体との関係性がより重要なシーンへの適用が期待される．

参考文献

[1] Aganian, D., Köhler, M., Baake, S., Eisenbach, M. and Gross, H.-M.: How Object Information Improves Skeleton-Based Human Action Recognition in Assembly Tasks, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8 (2023).

[2] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1310 (2017).

[3] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, pp. 4724–4733 (2017).

[4] Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J. and Lu, H.: Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–553 (2020).

[5] Cheng, M., Cai, K. and Li, M.: RWF-2000: An Open Large Scale Video Database for Violence Detection, *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 4183–4190 (2021).

[6] Cho, H., Kim, C., Kim, J., Lee, S., Ismayilzada, E. and Baek, S.: Transformer-based Unified Recognition of Two Hands Manipulating Objects, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4769–4778 (2023).

[7] Du, Y., Wang, W. and Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118 (2015).

[8] Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D. and Dai, B.: Revisiting Skeleton-based Action Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2969–2978 (2022).

[9] Hachiuma, R., Sato, F. and Sekii, T.: Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 22962–22971 (2023).

[10] Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H. and Farazi, M.: Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM, *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2021).

[11] Keselman, L., Woodfill, J. I., Grunnet-Jepsen, A. and Bhowmik, A.: Intel RealSense Stereoscopic Depth Cameras, *arXiv preprint arXiv:1705.05548* (2017).

[12] Kim, S., Yun, K., Park, J. and Choi, J. Y.: Skeleton-Based Action Recognition of People Handling Objects, *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 61–70 (2019).

[13] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T.: HMDB: A Large Video Database for Human Motion Recognition, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2556–2563 (2011).

[14] Lee, I., Kim, D., Kang, S. and Lee, S.: Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1012–1020 (2017).

[15] Lee, S. U., Hofmann, A. and Williams, B.: A Model-Based Human Activity Recognition for Human–Robot Collaboration, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019).

[16] Liu, Z., Zhang, H., Chen, Z., Wang, Z. and Ouyang, W.: Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 140–149 (2020).

[17] OpenAI: GPT-4o System Card, OpenAI Technical Report (2024). Reported: 2024-8-8.

[18] OpenAI: New embedding models and API updates (2024). Reported: 2024-02-14.

[19] Plizzari, C., Cannici, M. and Matteucci, M.: Spatial Temporal Transformer Network for Skeleton-Based Action Recognition, *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pp. 694–701 (2021).

[20] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision, *Proceedings of the 38th International Conference on Machine Learning (ICML)* (Meila, M. and Zhang, T., eds.), Proceedings of Machine Learning Research, Vol. 139, PMLR, pp. 8748–8763 (2021).

[21] Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A. and Maragos, P.: Multimodal Human Action Recognition in Assistive Human-Robot Interaction, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2702–2706 (2016).

[22] Sato, F., Hachiuma, R. and Sekii, T.: Prompt-Guided Zero-Shot Anomaly Action Recognition Using Pretrained Deep Skeleton Features, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6471–6480 (2023).

[23] Sekii, T.: Pose Proposal Networks, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[24] Shi, L., Zhang, Y., Cheng, J. and Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12026–12035 (2019).

[25] Si, C., Chen, W., Wang, W., Wang, L. and Tan, T.: An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1227–1236 (2019).

[26] Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y.: MPNet: Masked and Permuted Pre-training for Language Understanding, *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)* (2020).

[27] Song, S., Lan, C., Xing, J., Zeng, W. and Liu, J.: An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data, *AAAI Conference on Artificial Intelligence*, pp. 4263–4270 (2017).

[28] Song, Y.-F., Zhang, Z., Shan, C. and Wang, L.: Con-

structing Stronger and Faster Baselines for Skeleton-Based Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 2, pp. 1474–1488 (2023).

[29] Su, Y., Lin, G., Zhu, J. and Wu, Q.: Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 52–69 (2020).

[30] Sultani, W., Chen, C. and Shah, M.: Real-World Anomaly Detection in Surveillance Videos, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488 (2018).

[31] Sun, K., Xiao, B., Liu, D. and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703 (2019).

[32] Wang, Q., Peng, J., Shi, S., Liu, T., He, J. and Weng, R.: IIP-Transformer: Intra-Interpart Transformer for Skeleton-Based Action Recognition, *arXiv preprint arXiv:2110.13385* (2021).

[33] Xia, H. and Gao, X.: Multi-Scale Mixed Dense Graph Convolution Network for Skeleton-Based Action Recognition, *IEEE Access*, Vol. 9, pp. 36475–36484 (2021).

[34] Xiang, W., Li, C., Zhou, Y., Wang, B. and Zhang, L.: Generative Action Description Prompts for Skeleton-Based Action Recognition, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10242–10251 (2023).

[35] Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D. and Tang, H.: Dynamic GCN: Context-Enriched Topology Learning for Skeleton-Based Action Recognition, *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 55–63 (2020).

[36] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J. and Zheng, N.: View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2117–2126 (2017).

[37] Zhang, Z.: Microsoft Kinect Sensor and Its Effect, *IEEE MultiMedia*, Vol. 19, No. 2, pp. 4–10 (2012).

[38] Zhu, A., Ke, Q., Gong, M. and Bailey, J.: Part-aware Unified Representation of Language and Skeleton for Zero-shot Action Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[39] 筒川和樹，長野紘士朗，佐藤文彬：KIND: インスタンス単位の知識転移を用いたキーポイントベースゼロショット行動認識，第 28 回画像の認識・理解シンポジウム（MIRU2025） (2025).