隠蔽シナリオでの深度を使用した物体追跡モデル

宇佐美大樹　鷲見和彦
青山学院大学
E-mail: usami.daiki@vss.it.aoyama.ac.jp

1 はじめに

近年，監視カメラやスマートフォンの普及に伴い，様々な環境で動画を容易に取得できるようになってきている．これにより，撮影された動画データを活用して異常行動検知や物体カウント，行動認識など，多岐にわたる応用研究が活発化している．これらのタスクを効果的に実現するためには，連続するフレーム間で同一の物体を正しく識別し続ける「複数物体追跡（MOT）」の技術が極めて重要な役割を担う．

しかしながら，MOT においては，被写体同士が重なり合う隠蔽が発生しやすく，Bounding Box の位置や形状が似通うことで追跡 ID の誤った入れ替わり（ID スイッチング）が起こることが大きな課題となっている．特に，密集した環境や複数の物体が頻繁に交差する状況では，ID スイッチングが顕著に発生し，追跡精度が顕著に低下する要因となり得る（図 1）．また，被写体が一度完全にフレーム外に出てしまったり，他の物体によって完全に視界を遮られたりした後，再度フレーム内に登場する場合には，従来の手法では元の追跡 ID を適切に再割り当てできないケースが多い．その結果，追跡が断続的になり，長期間にわたる正確な追跡の実現が困難となる．

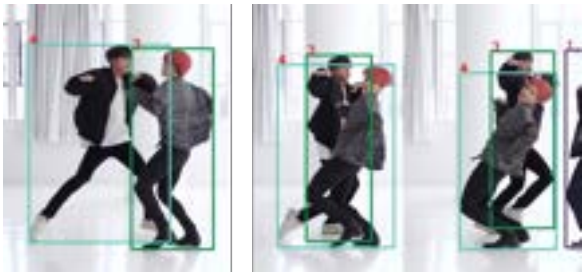

図 1: ID スイッチングによる追跡失敗例

そこで本研究では，隠蔽下での頑健な追跡を可能にする新たな手法を開発することを目的として，単眼画像深度推定モデルを活用することで，奥行き情報を推定し，2 次元の画像平面上だけでは得られない 3 次元的な特徴を追跡に導入することを提案する．深度情報を用いることで，近傍に存在する複数の物体が一時的に重なり合ったとしても，奥行きの差異を考慮して区別が可能となり，ID スイッチングの抑制が期待できる．さらに，被写体が一度完全に隠蔽され，再度フレーム内に現れた場合においても，過去の深度情報や 3 次元的な動きの履歴を活用して同一個体を再識別しやすくなる．実験の結果，提案手法の精度が向上する余地はまだあるが，この手法の有用性を示すことができた．

2 関連研究

2.1 Hybrid SORT

Hybrid SORT[1] は，Cao らの研究 [2] を基に，隠蔽や密集領域での追跡精度向上を目指した手法である．状態ベクトルに検出信頼度 c およびその変化率 $\dot{c}$ を組み込み，物体の高さ情報を反映した Height IoU（HIoU）を用いることで，隠蔽シーンでも正確な再認識を可能にしている．また，HIoU と IoU を統合した Height Modulated IoU（HMIoU）により，重なり合う物体間での ID スイッチングを抑制し，追跡の安定性を向上させている．しかし，強い隠蔽や密集領域では検出信頼度の区別が困難となる場合があり，追跡 ID の誤りが発生する課題が残る．本研究ではベースラインとして改善を目指す．

2.2 Depth Anything v2

Depth Anything v2[3] は，大規模事前学習モデルを活用した識別モデルベースの深度推定手法で，様々な環境下で高い精度と安定性を示す．マルチスケールでの特徴抽出により，大きな物体から小さな物体まで効率的に深度を推定可能であり，シンプルで高速な推論パイプラインを持つ．本研究ではこのモデルを取り入れることで，隠蔽が多発する状況でも深度情報を活用し，正確な追跡 ID の割り当てを目指す．

2.3 Segment Anything Model

Segment Anything Model(SAM)[4] は，大規模学習済みの汎用的なセグメンテーションモデルである．画像全体に対して予め学習された高品質な埋め込みを生成し，ユーザが指定するバウンディングボックスなどの「プロンプト」に応じて任意の領域を素早く切り出せる点が特徴である．本研究では，物体検出器によって得られるバウンディングボックスをプロンプトとして SAM に与え，ピクセルレベルで対象物体領域を特定する．これにより，バウンディングボックス内の物体位

置を正確に切り出したうえで深度を抽出でき，複雑なシーンでも高精度な奥行き情報の活用が可能となる．

3 提案手法

本研究では，既存の割り当てアルゴリズムにフレーム間の深度情報を組み合わせることで，2 次元の Bounding Box 情報だけでは対応が困難だったオクルージョン下での ID スイッチングを抑制する手法を提案する．具体的なフローは以下のとおりである．

1. **物体検出**
   入力画像に対して物体検出器 YOLOX[5] を用い，人物を検出する．検出結果として得られる Bounding Box および検出信頼度は，後段のマスク生成および深度解析のトリガーとして利用する．
2. **深度マップの生成**
   同じ入力画像に対して Depth Anything v2 による深度推定を行い，シーン全体の奥行き情報を取得する．
3. **人物マスク生成**
   YOLOX の検出信頼度が事前に設定した閾値を上回った人物物体検出結果のみを対象とし，SAM を用いて物体領域のマスクを生成する．
4. **平均深度の抽出**
   生成されたマスク領域と深度マップを照合し，各物体領域の平均深度を算出する．
5. **ID 割り当て**
   フレーム間の割り当てアルゴリズムに，前ステップで得られた深度情報を加味した新たな同一識別コストを導入し，同一物体の再識別を行う．
6. **情報の保存**
   各フレームで確定した検出結果と深度情報を保存し，次フレームの追跡に活用する．なお，マスク生成が行われなかった検出に関しては，深度情報を追跡情報に反映させないことで誤差の蓄積を抑制する．

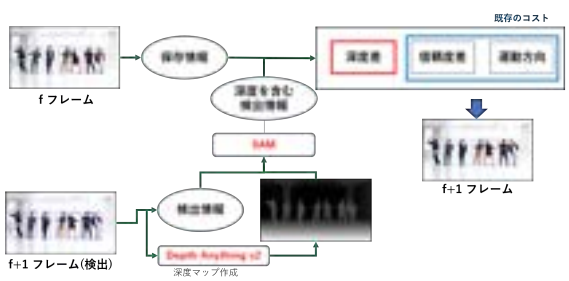

図 2: 提案手法の概略図

フローの捕捉: 本研究で用いられている物体検出器 YOLOX はベースラインとしている Hybrid SORT でも用いられており，同一モデルを使用することで追跡モデルの性能を図る．また，SAM によるマスク生成では，誤検出を防ぐために検出信頼度の閾値を設けた．これは同一物体の可能性が低い検出が近い深度を持つことでノイズ情報として強くなってしまうことを防ぐためである．(図 3)

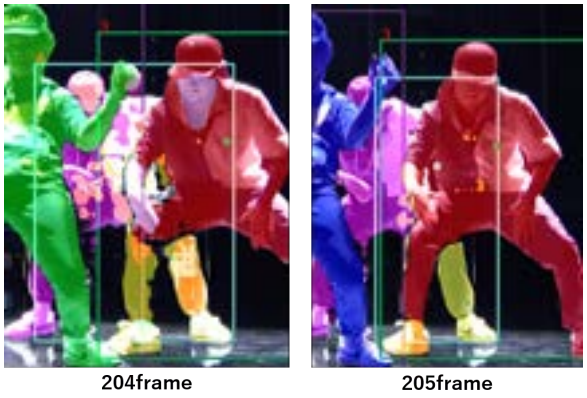

図 3: 全検出マスク生成による失敗例 - 205Frame でピンクにマスクされている人物に割り当てを行いたいが 204・205frame の低信頼度検出がノイズとなっている．

上記の手法により，2 次元 Bounding Box 情報だけではうまく扱えなかった重なり合いや複雑なシーン下においても，深度を含む 3 次元的な情報を活かして ID の混在を低減し，より安定した複数物体追跡を実現することを目指している．

4 実験

本研究ではベースライン，検出 Bounding Box の中心座標から深度を抽出するモデル，提案手法の 3 つのモデルで比較して追跡精度を検証した．

4.1 実験設定

評価では屋内，屋外，固定カメラ，動カメラの線形的な動きである人の歩行の様子を示す MOT17 train データセット [6] と，動き・外観がより複雑な DanceTrack val，test データセット [7] を用い，総合的な追跡性能を示す HOTA, 割り当ての精度を示す AssA, ID の一貫性を示す IDF1, ID スイッチングの回数を示す IDSW といった代表的な評価指標に基づいて定量評価を行っている．

4.2 予備実験: 中心座標深度の定性評価

本研究に先立ち，深度情報の有用性を示すため Bounding Box の中心座標から深度を抽出するシンプルな方法を試みた．深度の導入によってオクルージョンが軽微なシーンで深度差を利用することで ID スイッ

チングが抑えられる (図 4)，隠蔽後の再登場時に深度情報が手がかりとなる (図 5) ことから正確な深度情報の追跡に対する有用性が確認できた．

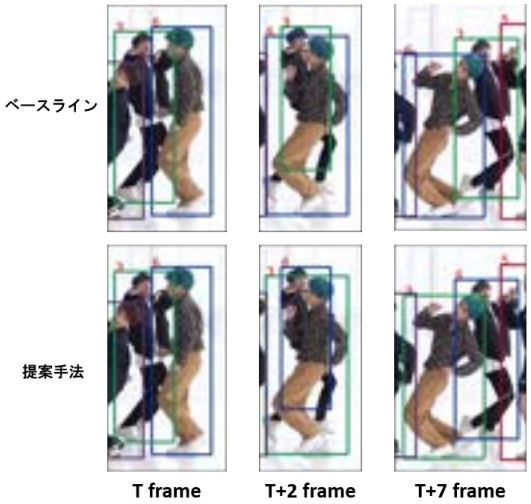

図 4: 深度が前後関係の手がかりとなる例

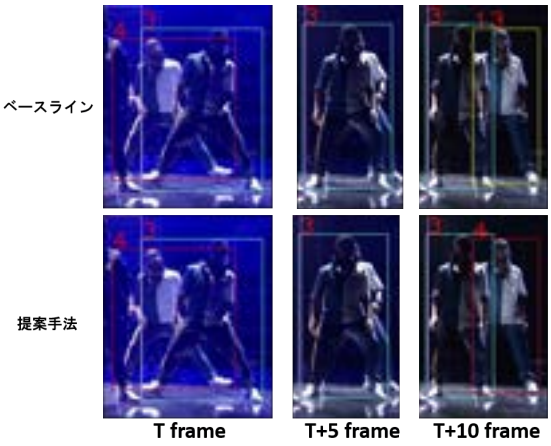

図 5: 深度が再登場時の手がかりとなる例

また，正常な深度情報の有用性が示される一方で重度な隠蔽においての前景深度抽出，Bounding Box 形状によるオブジェクト外の深度抽出による異常な深度抽出が課題として存在するため，本研究では SAM による人物マスクを用いることで正確な深度抽出を行う．

4.3 実験結果

表 1，表 2，表 3，表 4 は 2 つのデータセットで得られた定量評価結果を示す．

表 1: MOT17 train データセットでの評価

	HOTA.	AssA.	IDF1.	IDSW.
Hybrid SORT	69.51	68.03	78.58	618
中心座標深度	**69.649**	**68.505**	**79.148**	**583**
提案手法	69.202	67.725	78.252	616

表 2: MOT17 test データセットでの評価

	HOTA.	AssA.	IDF1.	IDSW.
Hybrid SORT	62.98	62.89	78.00	2223
中心座標深度	**63.27**	**63.42**	**78.18**	**1896**
提案手法	63.09	63.04	78.13	2082

表 3: Dance Track val データセットでの評価

	HOTA.	AssA.	IDF1.	IDSW.
Hybrid SORT	59.392	44.9	60.671	1721
中心座標深度	59.38	45.056	60.745	**1637**
提案手法	**60.292**	**46.229**	**61.719**	1671

表 4: Dance Track test データセットでの評価

	HOTA.	AssA.	IDF1.	IDSW.
Hybrid SORT	62.124	47.333	62.801	1566
中心座標深度	60.761	45.203	61.462	**1528**
提案手法	**63.315**	**48.909**	**64.396**	1541

表 1，表 2 から，MOT17 データセットでは中心座標深度を導入することで IDSW が大幅に減少し，他の評価指標も向上しているため深度情報を活用する有効性が確認できた一方で，提案手法は test データではベースラインと大きな性能差はなく，train データでは若干精度が低下している結果が得られた．

しかし，DanceTrack val および test データセット（表 3, 表 4）では，非線形な動きや似た外観といった複雑なシーンが多いなか，提案手法が最も高い HOTA や IDF1 を達成している．また，MOT17 データセットでは精度向上を見せていた中心座標深度は val データセットでは多くの IDSW を減少させているのにも関わらず若干の精度低下，test では大幅な精度低下が確認できたため，複雑な動きでの中心座標深度の不安定性，また SAM 導入によって複雑な動きと ID スイッチングに対してよりロバストになったことが確認できる (図 6)．

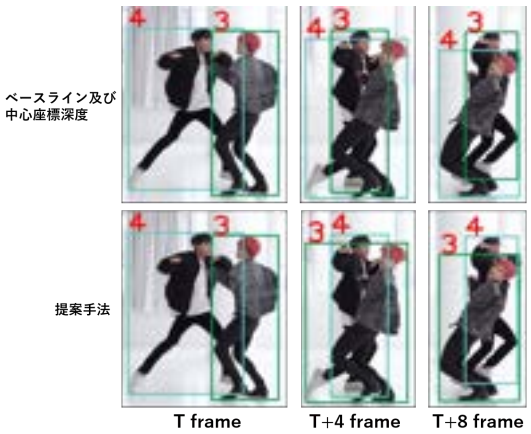

図 6: 提案手法の追跡成功例

4.4 今後の課題

実験結果を踏まえると，線形的な動きに対する性能はほぼ維持される一方，非線形的な動きに対してはロバスト性が向上していることが確認された．しかし，IDSWに関しては中心座標深度の方が圧倒的に減少しており，より正確な深度を使用したにもかかわらずMOT17データセットでは中心座標深度よりも精度の向上が見られなかった．これには，中心座標深度では全検出に対して深度抽出を行っていたのに対してノイズの強化対策として導入した抽出条件の閾値が原因になっていると考えられる．したがって，ノイズとなる検出を抑えつつ信頼度が低いにもかかわらず正しい検出結果の深度抽出ができればIDSWをより減らすことが可能であると考える．

また，重度な隠蔽において生成されるマスク領域が重なってしまう点が大きな課題として存在する (図7)．マスク領域が重なることで，隠蔽時の前景オブジェクトの深度を抽出し，完全隠蔽後の手がかり，前後関係の手がかりとして弱くなってしまう．この課題に対しては隠蔽の起こりにくい頭部検出などとの併用，Bounding Boxとマスク領域の割合を用いたマスクが信頼できるかの評価，仮割り当てを行うことで前フレームとの深度値比較を行うなどの対策が必要であると考えられる．

図 7: マスク領域の重複による深度が活かせない例

5 結論

本研究では，隠蔽が多発する環境下でも安定した複数物体追跡を実現するために，単眼画像深度推定モデルを組み込み，SAMを用いて正確に人物領域を切り出す手法を提案した．中心座標深度を用いたモデルとの比較により，深度情報自体がIDスイッチングや再登場時の対応付けの手がかりとして有効であることを確認すると同時に，オクルージョンや非線形な動きが顕著なシーンでは，提案手法のマスク抽出による深度活用がさらに有効である結果を得た．

一方で，マスク同士の重複により前景の深度が活かせなくなる問題も依然として発生しており，誤差抑制のためには低信頼度の検出を適切に扱う判断や，頭部検出など他の手法との組み合わせを検討する必要がある．今後は，こうした閾値設定の最適化や前後フレーム比較を活用した深度補正機構を導入することで，さらに安定かつ高精度な隠蔽耐性を備えた複数物体追跡手法の確立を目指す．

参考文献

[1] YANG, Mingzhan, et al. Hybrid-sort: Weak cues matter for online multi-object tracking. In: Proceedings of the AAAI conference on artificial intelligence. 2024. p. 6504-6512.

[2] CAO, Jinkun, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. p. 9686-9696.

[3] YANG, Lihe, et al. Depth anything v2. Advances in Neural Information Processing Systems, 2025, 37: 21875-21911.

[4] KIRILLOV, Alexander, et al. Segment anything. In: Proceedings of the IEEE/CVF international

[5] GE, Zheng, et al. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

[6] conference on computer vision. 2023. p. 4015-4026. MILAN, Anton, et al. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.

[7] SUN, Peize, et al. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 20993-21002.