アノテーションコスト削減を目的とした0-shotセグメンテーションの実用性評価：製造現場における基盤モデルの比較

① 背景

製造業における自動化の進展に伴い、3DセンサとAIを組み合わせたピッキングシステムの導入が加速している。特に、インスタンスセグメンテーションを用いた3Dピッキングは、複雑な形状や重なり合った物体の認識において高い精度を発揮する技術として注目されている。しかし、従来のセグメンテーションモデルは教師あり学習に依存しており、学習データのアノテーションには膨大な時間とコストがかかる。たとえば、COCOデータセットでは1枚の画像に対するインスタンスセグメンテーションのアノテーションに平均で約240秒を要し、10万枚の画像に対しては約6,666時間以上の作業が必要とされる。 95 pageに記載あり1。

このような課題に対し、Meta社が提案したSegment Anything Model（SAM）は、プロンプトに基づく0-shotセグメンテーションを可能とする基盤モデルとして注目されている。SAMは、事前学習済みの大規模データを活用し、追加の学習なしに任意の物体をセグメンテーションできる点で、アノテーションコストの削減と柔軟な運用の両立が期待されている 2。

さらに、SAMの軽量版であるnanoSAM、テキストと視覚情報を統合するGroundingDINO、精度向上を図ったSAM2など、実運用を見据えた多様な基盤モデルが登場しており、それぞれの特性を活かした応用が可能となってきている 3。

② 目的

本研究では、京都製作所における3DセンサとAIを組み合わせたピッキングシステムの高度化を目的として、SAMを起点に、GroundingDINO、nanoSAMといった複数の基盤モデルを比較・検証する。特に、アノテーション作業の削減、セグメンテーション精度、モデルの軽量性および推論速度といった観点から、製造現場における実運用への適性を評価し、最適なモデル選定の指針を示すことを目的とする。

③ 材料と方法

3.1 データセット

3Dピッキングとして引合のあるワークを対象（何かを具体化する）に、1,000枚の物体画像をデータ拡張にて収集した。各物体は異なる、配置パターンで撮影され、現実的なピッキング環境を模倣した。

3.2 モデル構成

以下の3つの基盤モデルを比較対象とした：

- **SAM**：Meta社が開発した0-shotセグメンテーションモデル 2。
- **GroundingDINO**：テキストプロンプトに基づく物体検出を可能とするモデル 3。
- **nanoSAM**：軽量化されたSAMで、エッジデバイスでのリアルタイム処理を想定。

3.3 評価指標

- **mAP（Mean Average Precision）**：検出精度の総合的な指標。

- **推論時間**：1画像あたりの平均処理時間。
- **IoU（Intersection over Union）**：セグメンテーションマスクの重なり具合を評価。
- **アノテーション削減率**：従来の手動アノテーションと比較した削減割合。

3.4 実験手順

1. 各モデルに対して、同一のデータセットを用いてセグメンテーションを実行。
2. 出力マスクをGround Truthと比較し、IoUおよびmAPを算出。
3. 推論時間を測定し、リアルタイム性を評価。
4. アノテーション作業時間を記録し、削減効果を算出。

④ 結果

モデル	IoU	mAP	推論時間（秒）	アノテーション削減率
SAM	.*	.*	.*	***%
GroundingDINO	.*	.*	.*	***%
nanoSAM	.*	.*	.*	***%
Mask R-CNN（従来）	0.96	0.75	0.031	0%

- nanoSAMは最も高速な推論を実現した。
- SAMはバランスの取れた性能を示し、
- GroundingDINOはテキストプロンプトによる柔軟な操作性→ただし、テキストプロンプトの指定したものがないと誤検出が発生。

⑤ 考察

⑥ 結論