令和 2 年度修士論文

個体毎の領域分割を用いた遮蔽に強い複数物体追跡

宮崎大学大学院　工学研究科　工学専攻

機械・情報系コース　情報システム工学分野

学籍番号　T1903290

竹井　奏一郎

指導教員　椋木　雅之　教授

令和 3 年 1 月 25 日

概要

本研究では、物体検出・セグメンテーションの手法としてインスタンス・セグメンテーションを用いた遮蔽に強い複数物体追跡を行う。

複数物体追跡を困難にする要因として、追跡対象同士の重なりによる遮蔽問題が挙げられる。従来の物体追跡では、追跡対象がどのようなものであるかという事前知識を用いず、追跡開始時に指定された対象の存在する領域の情報のみを用いて追跡対象を検出し追跡する。しかし、遮蔽が生じることで追跡対象の見え方が変わるため、追跡の失敗に繋がる。特に類似性が高い個体同士が隣接・重なるように位置する場合、個体同士の境界が曖昧となり追跡を行うための検出に影響を及ぼす。検出の精度が低くなると追跡対象の位置を正確に把握できないため、結果的に追跡の精度が下がる。このことから追跡を正しく行う際は類似性の高い個体同士が密接に存在している状況でも対象を正しく検出する必要がある。

この問題を解決するために、本研究では、検出性能の高いインスタンス・セグメンテーションを用いることで遮蔽問題が発生しても検出を行えるように改善する。次に検出された領域を時間方向に対応付けることで追跡を実現する。この際、検出漏れによって追跡対象を一旦見失っても、再度検出できた時点で対応付けをして、追跡を継続する。

提案手法と 7 つの従来の追跡手法の結果を比較評価した。従来手法は追跡対象の大きさの変化や類似度の高い個体同士による遮蔽や隣接が生じた場合に、追跡に失敗していた。これに対し、提案手法はいずれの場合に対しても従来手法と同等以上の結果を出し、遮蔽の生じる複数物体追跡の精度が改善した。

内容

概要

1. 序論 1

2. 物体追跡問題 2

2.1. 牛を対象とした複数物体追跡問題 2
2.2. 従来手法による物体追跡 2
2.3. 物体追跡における遮蔽問題 3
2.4. インスタンス・セグメンテーション 5
2.5. Mask R-CNN 6

3. 個体毎の領域分割を用いた物体追跡 9

3.1. 個体毎の領域分割を用いた物体追跡の処理手順 9
3.2. 追跡対象数の指定 9
3.3. インスタンス・セグメンテーションによる検出 9
3.4. 候補の組み合わせ作成 10
3.5. 追跡対象と作成した候補の重なりの計算 11
3.6. 追跡対象と候補領域の対応付け 12

4. 評価実験 14

4.1. 実験設定 14
4.2. 実験データ 15
4.3. 評価指標 17
4.4. 実験結果 19

5. 結論 41

謝辞 42

参考文献 43

1. 序論

畜産農業において家畜として飼育されている牛は、伝染病を持っている可能性があるなどの要注意な個体が存在する。伝染病管理のために牛を個体毎に行動追跡する技術は重要である。一般に牛の個体管理には耳についているタグが用いられるが、タグの汚れや顔の向きによって情報を得られない場合がある。発信機等を牛に装着して行動履歴を取得する方法もあるが、機器の装着で牛に負担がかかる、比較的広範囲の飼育場を牛が移動する場合、受信設備等の負担がかかるといった問題がある。ビデオカメラで広範囲の飼育場を観測し、映像内の牛の行動を追跡できれば、牛の個体管理に役立つ。そのために、動画像中の物体追跡技術を利用する。

従来の物体追跡では、追跡対象がどのようなものかという事前知識は用いず、追跡開始時に指定された追跡対象の存在する枠（バウンディングボックス）の情報のみから、バウンディングボックスと類似した画像領域を時間方向に追跡していた。この場合、牛の個体管理の対象とするシーンの多くでは類似性の高い対象が隣接して存在するため、対象を間違えやすいという問題がある。また、複数頭の牛が狭い範囲に存在しており、牛同士での遮蔽が発生しバウンディングボックスが重なるため、追跡時の個体の分離が困難であるといった問題がある。

一方、近年の深層学習の発展により物体検出技術が著しく向上している。インスタンス・セグメンテーションでは画素単位での個体検出・分離が可能になっている。

本研究ではインスタンス・セグメンテーションを用いて動画像中から牛を個体毎に分離して検出し、時間方向に対応付けをして追跡を行う。インスタンス・セグメンテーションを用いることによって、遮蔽が発生し対象の一部のみしか見えないシーンでも個体の分離が可能となる。また、牛の動きはそれ程速くない特性を利用して、フレーム間での領域の重なり情報を利用し対応付けを行う。インスタンス・セグメンテーションでは、検出漏れや一個体が複数に分割されて検出される場合がある。これに対しては、対応が取れない場合は検出漏れとみなして、次フレームでの対応付けを行う。また検出結果の複数の組み合わせも対応付けの候補とすることで対処する。

本論文の構成は以下の通りである。第 2 章では物体追跡問題の定義、従来の追跡手法、物体追跡を困難にする遮蔽の問題について述べる。第 3 章ではインスタンス・セグメンテーションを用いることで遮蔽の問題に対応した提案手法について述べる。第 4 章では評価実験について述べ、第 5 章で本研究の結論と今後の課題について述べる。

2. 物体追跡問題

2.1. 牛を対象とした複数物体追跡問題

物体追跡とは、入力された一連の動画像から、指定した物体が画像上でどのように移動したかを推定する問題である。本研究では、特に飼育環境での牛の映像を対象とする。牛を追跡対象とする場合、牛同士が密集して行動することが多いという特徴がある。密集することで牛同士の重なりにより遮蔽が生じるという問題がある。また、牛同士は見た目の類似性が高いという特徴がある。類似性の高い物体を複数追跡すると、追跡中の対象の取り違えや複数頭を融合して追跡してしまう問題が生じる。遮蔽による追跡の困難性、類似性の高い複数物体の追跡は一般的な物体追跡問題にも通じる重要な課題であり、これらの問題を解決する必要がある。

本研究では、最新の物体検出技術であるインスタンス・セグメンテーションを用いてこれらの問題に対処する。また、今回追跡対象とする牛は動きがそれ程速くないため、牛の移動速度の遅さという特徴も利用しこれらの問題を解決する。

2.2. 従来手法による物体追跡

物体追跡では一般に、現在フレームで追跡している追跡対象の位置が既知であるとする。この時、追跡対象が次フレームにおいて、どのように変化したかを順方向に逐次的に調べる。毎フレームでこの処理を繰り返すことで現在フレームの追跡対象の位置から、次フレームでの追跡対象の位置を推定する。

本研究では、OpenCV[1]のライブラリに含まれている以下の物体追跡の手法を従来手法として使用する。

- **Boosting** [2]

  AdaBoost アルゴリズムを用いた追跡手法である。新しい画像を読み込むと前の画像の追跡対象の近傍領域から負のサンプルを生成して、追跡対象ではない例として識別器を学習させる。その後、学習した識別器のスコアから追跡対象の新たな場所を検出する。この処理を繰り返して追跡を行う。

- **MIL（Multiple Instance Learning）**[3]

  Boosting 法と考え方が似ている手法である。Boositng 法と違う点として、追跡対象の近傍領域から正と負のサンプルが混在する集合を与える。正のサンプルが含まれている集合を負のサンプルも含めて識別器に学習させる。その後、学習した識別器のスコアから追跡対象の新たな場所を検出する。追跡対象の近傍領域のサンプルを含めることで周辺を加味した追跡を行う。

- **KCF（Kernelized Correlation Filter）**[4]

  先頭フレームで追跡したい物体をテンプレートとして指定し、追跡し続けながらテンプレートの学習をする手法である。Boosting や MIL は学習を行うためのサンプルを

与えるのにランダムで選んでいたが、KCF では一つのサンプルを少しずつずらしたものを大量に生成することで、疑似的に学習画像を増やすという特徴がある。フーリエ変換を用いることでメモリと計算量を削減したため、多くのサンプルを追加することが容易となった。サンプルが多ければ識別器の学習も多くできる。OpenCV の中でも精度の良い追跡手法だとされている。また高速な追跡手法である。

- MedianFlow[5]

  グリッドの中央値により追跡する手法である。追跡対象のバウンディングボックスをグリッド上に分割し、各グリッドの点を追跡し、エラーの大きいもの破棄しながら残った点の中央値を用いてバウンディングボックスの座標を更新して追跡する。遮蔽の発生や高速な動きに弱い特徴を持つ。

- TLD（Tracking Learning Detection）[6]

  追跡対象の物体の学習、検出、追跡を毎フレーム行う手法である。最初に追跡対象のバウンディングボックスを手動で与え、学習結果に基づき検出を行った後、MedianFlow 法で追跡を行う。学習を行う際、偽陽性と偽陰性を分けて学習することで、追跡対象の見た目の変化に対応し、変化を学習することで長期的な追跡を目指した手法である[7]。

- MOSSE（Minimum Output Sum of Squared Error）[8]

  単フレームを使用し初期化時に相関フィルタを生成して、トラッキングを行う手法である。照明や拡大縮小、非剛体の変形に対して堅牢である。実装が簡単である。

- CSRT（Channel and Spatial Reliability Tracker）[9]

  チャネルと空間信頼性の概念を用いてフレームから選択された領域を識別相関フィルタによる調整を受けながら追跡する手法である。

これらの従来手法は、追跡する対象を先頭フレームで指定して追跡を行う。本研究では、牛を追跡対象とするが、従来の物体追跡手法では「牛」が映像上でどのような見え方をするかといった事前知識は用いず、先頭フレームで指定した領域と類似した画像領域を時間方向に対応付けている。

2.3. 物体追跡における遮蔽問題

物体追跡を困難にする問題の一つとして遮蔽の問題がある。遮蔽とは、対象物体が移動し何らかの物体の後方に入った場合、対象物体の前方に存在するものによって一部または全体像が見えなくなることである。物体は向きの変更や他の物体の後ろへ移動するといった様々な要因によって遮蔽が生じ、全体像の一部が見えなくなり、見え方が変わってしまう。2.2 節で述べた従来の追跡手法は、先頭フレームにおいて追跡対象の領域を与え、その情報を元に検出を行い追跡する手法であった。この場合、先頭フレームで与えられた情報のみだと、一部が見えなくなるといった遮蔽が生じた際、追跡する対象と見え方が変わってしまうため、対象をうまく検出できず見失ってしまう。また、物体追跡を行う際、追跡対象と類似

性の高い領域が近くに存在した場合、検出対象を間違えやすいという問題がある。従来手法による物体検出は対象物体の境界に関する知識も利用していないため、類似している個体同士の隣接や重なりが生じると、個体毎の境界の検出の結果が悪くなる。また、牛の追跡では牛同士の遮蔽が生じやすい。この場合、類似した追跡対象が隣接、または重なった状態で画像中に現れる。そのため、個々の牛個体の取り違えが生じがちとなり、追跡の精度の低下につながる。

物体追跡の別のアプローチとして検出に基づく追跡（Tracking by Detection）がある。このアプローチでは、追跡対象がどのような見え方をするかという事前知識に基づき、各フレームから対象物体を検出する。検出した追跡対象を時間方向に対応付けることで追跡を行う。対象物体の見え方の変化に対応できるので、遮蔽が生じても追跡を続けることが出来る。また、個体同士が重なっても、それらを区別して個体毎に検出することができる。一方で、物体検出の精度が高くなければ、追跡対象を検出できず、追跡対象を見失ってしまう。

近年、物体検出の方法として、ディープラーニングを用いた物体検出が主流となっている。ディープラーニングを用いた物体検出の精度は高く、遮蔽が生じることで一部が見えない物体も検出が可能となっている。また、複数個体が隣接していても、個体毎に分離して検出できるようになっている。遮蔽に対応した物体検出を物体追跡に用いれば、遮蔽により一部が見えないような対象でも物体追跡が可能になる。

ディープラーニングを用いた物体検出には、Faster R-CNN などの物体検出と画像分類を同時に行う手法が存在する。しかし、ディープラーニングを用いた物体検出はバウンディングボックスの領域に対する検出のため、対象と関係のない領域がバウンディングボックス内に存在する。バウンディングボックスより正確な対象の存在する領域を用いて追跡を行うことで、遮蔽問題を解決した精度の高い追跡が実現できると考えられる。

本研究では、遮蔽による対象の一部が見えなくなる問題と類似性の高い個体同士の重なりによる物体検出の困難性を解決し追跡するため、インスタンス・セグメンテーションを用いて検出を行う。

2.4. インスタンス・セグメンテーション

インスタンス・セグメンテーションとは、入力された画像内の個体毎の領域を検出し、同時にその個体領域の種類（クラス）を推測する処理である。通常の物体検出では、個体を含むバウンディングボックスで領域を検出するのに対し、インスタンス・セグメンテーションでは個体毎の領域をピクセルレベルで認識する。

インスタンス・セグメンテーションを行った結果の例を図 1 に示す。図中の各色で塗りつぶされた領域が個体の存在する範囲である。それを囲む緑色の枠がバウンディングボックスである。この例のように、インスタンス・セグメンテーションでは、同一クラスの複数物体が隣接していても、個体毎の領域を検出することが出来ている。また、灰色で塗りつぶされている中央の人物のように遮蔽が発生している物体に対しても検出を高い精度で行うことができる。さらに、中央の灰色で塗りつぶされた人物をバウンディングボックスで検出すると隣の緑色で塗りつぶされた人物と大きく重なるが、インスタンス・セグメンテーションの結果では、このような重なりは生じない。

遮蔽問題が発生して一部が見えない対象や類似性の高い個体が近くに存在している場合でも個体毎により正確な検出を行うため、本研究では物体検出に検出結果ごとのセグメンテーションが付加されたインスタンス・セグメンテーション（Instance Segmentation）を用いる。ピクセルレベルで物体の領域の場所を示すことで、バウンディングボックスと比べ、より正確な領域を用いることができ、追跡の精度の向上が期待できる。本研究では、個体の存在する範囲をピクセル単位で示した領域をマスク領域と呼称する。

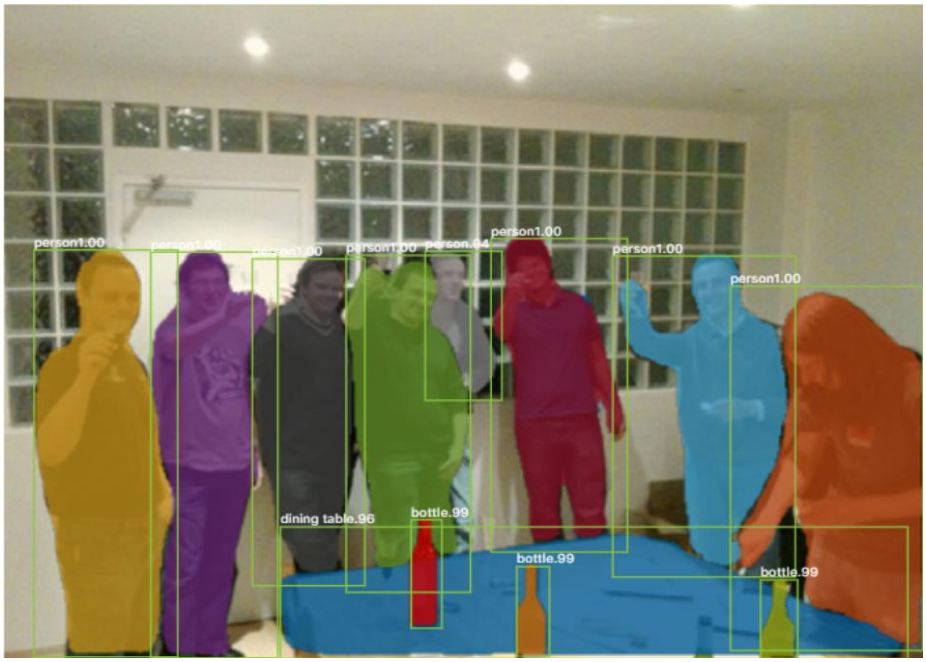

図 1　インスタンス・セグメンテーションの出力例　[10]より引用

2.5. Mask R-CNN

本研究ではインスタンス・セグメンテーションを行うための手法として Mask R-CNN[10]を採用する。Mask R-CNN は、入力された画像に対して物体検出と画像分類を行う Faster R-CNN[11]という手法にセグメンテーションタスクを追加した手法である。本章では Mask R-CNN の構成の大部分を占める Faster R-CNN の構成と Mask R-CNN の実装のために変更した部分について解説する。Faster R-CNN と Mask R-CNN の構成を図 2、図 3 に示す。

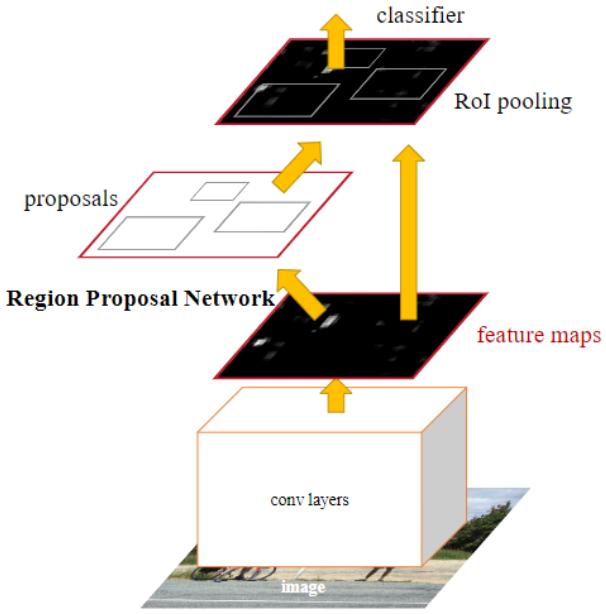

Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

図 2　Faster R-CNN のネットワーク構成図　[11]より引用

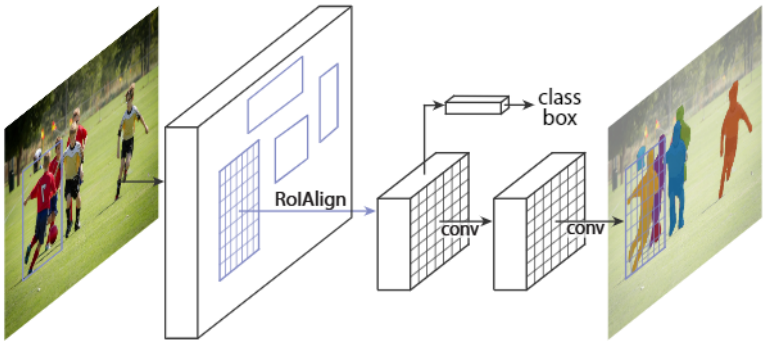

Figure 1. The **Mask R-CNN** framework for instance segmentation.

図 3　FasterR-CNN の構成に加えられた Mask R-CNN のネットワーク変更部分　[10]より引用

- Faster R-CNN

Faster R-CNN は、入力画像に対しオブジェクトの存在する可能性のある領域の検出と、その検出領域のオブジェクトのクラス分類を同時に行う手法である。

Faster R-CNN は以下の処理によって検出と分類を行う。

① 入力画像から CNN（Convolutional Neural Network）を通して特徴量マップを抽出
② 特徴量マップから RPN（Region Proposal Network）を通してオブジェクトの候補領域を提案
③ 特徴量マップに対しプーリングし、入力を固定長の長さに変換し候補領域の特徴を取得
④ 取得した特徴から候補領域に対して画像分類
⑤ 候補領域に対してバウンディングボックスを予測

- RPN（Region Proposal Network）

Faster R-CNN では特徴量マップを RPN というネットワークに読み込ませ処理することで物体の候補領域を得る。Faster R-CNN の前身となる Fast R-CNN では、入力した画像に対し、外部のアルゴリズムを使用して候補領域を提案していたため低速であった。Faster R-CNN では候補領域の提案に RPN 用い、画像の入力から物体検出までを End-to-End で行うことで高速かつ精度改善を実現した。

RPN を用いた候補領域の提案には、アンカーボックスと呼ばれる特定の高さ・幅で定義されたバウンディングボックス群を用いる。まず、特徴量マップにアンカーボックスをタイル状に配置する。次に各アンカーボックスに対して物体が存在する確率を計算する。物体が存在する確率の高いアンカーボックスを採用し候補領域の提案を行う。同じ物体に対して複数の候補領域の提案がある場合は非極大値の抑制をして、最も評価の高い候補領域以外を破棄し最終的な候補領域を提案する。

Fast R-CNN で行われる候補領域の提案との違いは、特徴量マップを用いて候補領域の提案を行うことである。Fast R-CNN では候補領域の提案を行うためにバウンディングボックスを画像に対して走査していた。RPN を用いると、この走査が必要ないため高速となる。

- RoI Pooling（Region of Interesting Pooling）

RoI Pooling は、特徴量マップから提案された候補領域内から固定長の特徴量マップを抽出するために行う処理である。元画像の提案された候補領域を特徴量マップに重ねるとわずかなズレが生じるため、候補領域を特徴量マップのグリッドに合うよう丸め込みを行ってからプーリングを行う。この丸め込みにより実際の候補領域と特徴量マップの出力に差が出る。しかし、Faster R-CNN ではバウンディングボックスによる物体検出とそのクラスの分類を行うため、この丸め込まれた誤差は検出結果に大きく影響しない。

- Mask R-CNN と Faster R-CNN の構成の違い

Mask R-CNN は、Faster R-CNN の処理である物体検出と画像分類に加え、提案された領域内に対してピクセル単位でセグメンテーションを行う処理を加えた手法である。Mask R-CNN と Faster R-CNN の違いとしてセグメンテーションの処理の追加の他に、RoI Pooling を RoI Align に変更した点がある。以下では Mask R-CNN での変更点について述べる。

- mask branch の追加

Mask R-CNN ではセグメンテーションを行うために、Faster R-CNN に mask branch というネットワークを追加する。特徴量マップを用いて候補領域の各ピクセルに対して、どのクラスの物体が存在しているかを求める処理を行う。候補領域内に存在する物体がどのクラスであるか Faster R-CNN によりクラス分類されているので、分類されたクラスの存在確率の高いピクセルがマスク領域として得られる。この処理は各提案領域に対して行われるので、同クラスの物体が予測されても、候補領域が異なれば別個体であると判断される。このことから、インスタンス・セグメンテーションは、類似している複数の個体が、隣接や重なりにより全体像が見えていなくても別個体として出力できる。

- RoI Align（Region of Interesting Align）

Mask R-CNN では RoI Pooling の代わりに RoI Align を用いる。

Faster R-CNN の RoI Pooling は、提案された候補領域を特徴量マップに重ねるとわずかなズレが生じるため、候補領域を特徴量マップのグリッドに合うよう丸め込みを行ってからプーリングを行う。しかし、Mask R-CNN で行う処理の一つであるセグメンテーションは、提案された領域の各ピクセルに対しクラスを割り当てるので、クラスを与える際のピクセルのズレは許されない問題となる。したがって、Mask R-CNN では候補領域に対してズレの生じないプーリングを行う必要がある。

RoI Pooling では特徴量マップに候補領域を割り当ててから等分割したのに対し、RoI Align は提案された領域をそのまま等分割してプーリングを行う。これにより、位置合わせを行わずプーリングを行うため、サブピクセルレベルのズレは生じなくなる。

3. 個体毎の領域分割を用いた物体追跡

3.1. 個体毎の領域分割を用いた物体追跡の処理手順

提案手法の処理手順を以下に示す。

① 動画中の追跡対象の個体数と先頭フレームでの個体毎の位置の指定
② インスタンス・セグメンテーションによる候補領域の検出
③ 候補領域の組み合わせの作成
④ 現在フレームと次フレームにおける、追跡対象と組み合わせた候補の重なりの算出
⑤ 追跡対象と候補領域の対応付け
⑥ ③～⑤の繰り返し

②のインスタンス・セグメンテーションでは一つの個体が複数の候補領域に過剰に分割される誤りや、一部の個体が検出されない検出漏れが生じることがある。候補領域が過剰に分割される誤りに対処するため、③で複数の候補領域を組み合わせた候補を作成し、対応付けに利用する。また検出漏れに対応するために、⑤で追跡を継続するための処理をする。④では追跡対象の対応付けをするための要素として、追跡対象と候補の重なりの大きさを求める。以下で各処理の詳細を述べる。

3.2. 追跡対象数の指定

本研究では、追跡対象の個体数と追跡を開始する先頭フレームでの個体毎の位置を与える。先頭フレームの個体毎の位置は、それぞれの個体が存在する領域をマスク領域として与えて示す。

3.3. インスタンス・セグメンテーションによる検出

Mask R-CNN を利用したインスタンス・セグメンテーションを行う。インスタンス・セグメンテーションの検出結果であるマスク領域は個体毎に出力する。

インスタンス・セグメンテーションを実行するためのプログラム[12]は、GitHub で公開されているものを使用した。

3.4. 候補の組み合わせ作成

本研究では追跡する追跡対象の対応付けを取る際、単一のマスク領域同士だけでなく、複数のマスク領域の和を取ったものも追跡対象の候補とする。インスタンス・セグメンテーションによる検出は個体領域の分割が発生してしまい、本来 1 つの個体が存在する領域に対し複数の領域が出力されることがある。このような場合、同フレームで出力された候補領域を組み合わせた領域の和集合を考えることによって、より個体に対応する領域を作成する。

図を用いて説明する（図 4）。左側の牛 α は一頭の牛の領域に対し正常に検出できている。それに対して、インスタンス・セグメンテーションによる検出は、右側の牛のように一頭の牛に対して a,b のように領域の分割が生じた出力をする場合がある。この場合、α に相当する理想的な検出結果は a+b である。このような状況に対応するために、同フレームで出力された追跡対象の候補領域は、検出された単一の個体の候補領域のみでなく、a、b、c を任意個の組み合わせを全通り分考えることで、a+b の領域を候補として含めることが出来る。

このことから、複数の候補領域を組み合わせた候補領域を全て生成して、対応付けの候補とする。時刻tでP_t個の候補領域$M_t = \{m_1, \cdots, m_{P_t}\}$が得られた時、$M_t$の要素の任意個の組み合わせ（べき集合$2^{M_t}$の要素）を対応付けの候補とする。べき集合$2^{M_t}$は空要素$\varepsilon$も含む。$\varepsilon$は候補領域が存在しないことを表す。

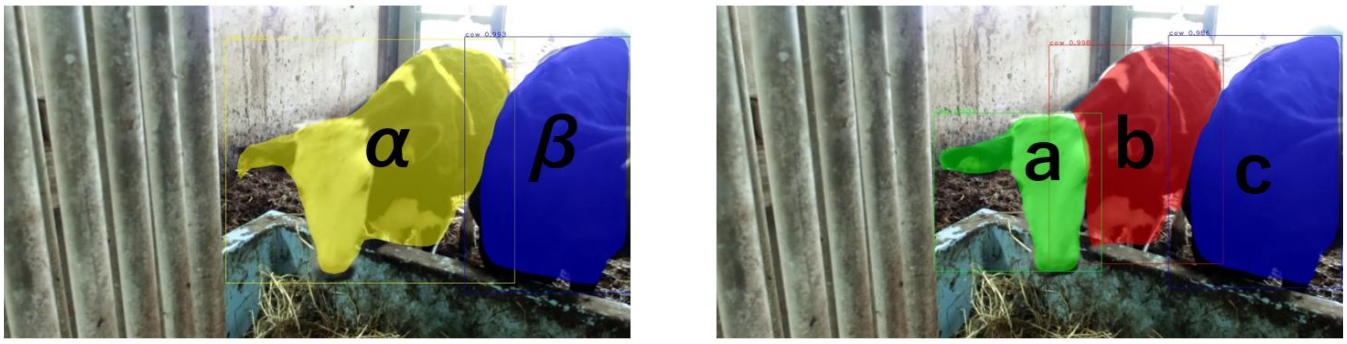

図 4　インスタンス・セグメンテーションの正常な検出例（左）と分割が生じた例（右）

3.5. 追跡対象と組み合わせ候補の重なりの計算

現在追跡中の対象に当てはまる候補が存在するかを判断するために、3.4 節で作成した組み合わせ候補を使用する。

現在フレームの既知の追跡対象と次フレームの追跡対象の候補の間で、類似しているものを対応付けて追跡を行う。本研究では類似度の評価として、現在フレームと次フレームにおける牛の存在する領域同士の重なりの大きさを利用した。牛の動きは遅いことから、1 フレームであれば同じ場所からほとんど移動していないと考えられるため、連続する 2 フレームにおける同一個体の牛の重なりの割合は大きいと想定できる。また、遮蔽により画像上の見え方や見えている部分が変わっても同様の条件で考えられるため、重なりの大きさを利用した対応付けにより安定した評価が可能である。

重なりの大きさの評価には IoU（Intersection over Union）を用いる（図 5)。追跡中の対象領域と候補領域の重複領域と全体領域を作成し、全体領域の画素数に対する重複領域の画素数の割合を IoU 値とする。IoU 値は 0 から 1 の値を取り、1 の時最も重なりが大きい（類似度が高い)。

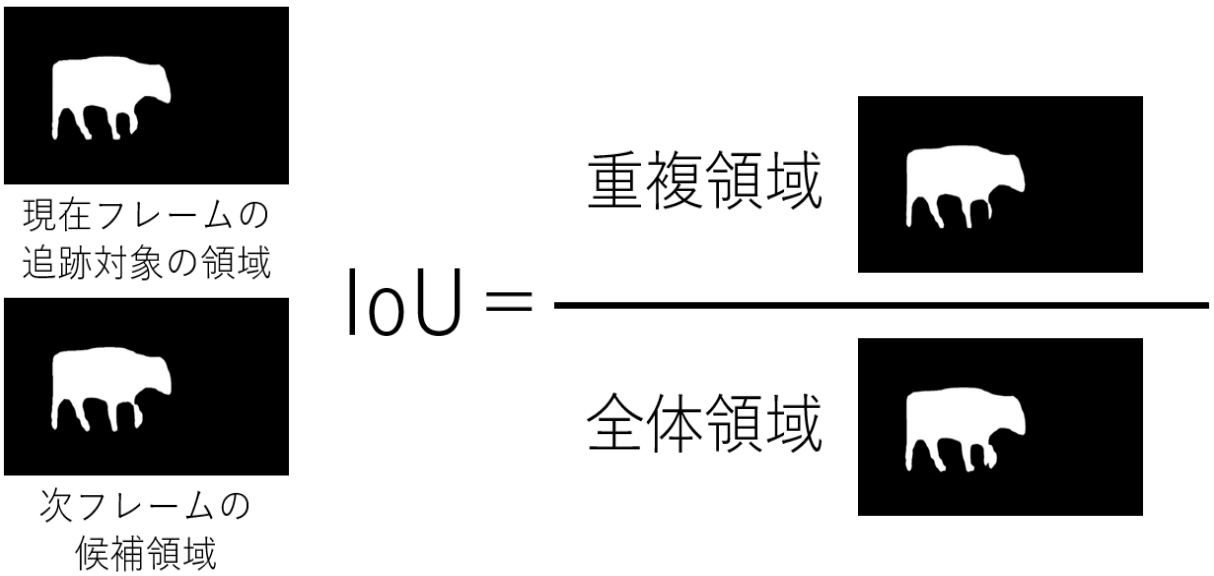

図 5　IoU の概念図

3.6. 追跡対象と候補領域の対応付け

現在フレームでの追跡対象の領域と次フレームから検出・組み合わせた候補領域との重なりに基づき、両者の対応づけを行う。インスタンス・セグメンテーションによる検出は、個体同士に重なりがあっても個体毎の検出が行えるが、時には検出漏れが発生することがある。特に本研究で取り扱う牛のような類似性が高い追跡対象が隣接、または重なった状態で画像中に現れると、複数の個体を同一の個体として検出してしまい片方の個体の検出漏れとなる場合がある。検出漏れが生じてしまうと追跡対象を見失ってしまうが、インスタンス・セグメンテーションは検出漏れ後も検出を引き続き行うことは可能である。

対象物体の検出漏れが発生した場合、2 フレーム間での対象物体の候補の対応が取れないため、追跡が出来なくなってしまう。検出漏れが発生した後に再度検出した場合、別個体とみなされ、個体間の対応付けが難しい。本研究では検出漏れが発生し、次フレームにおいて追跡候補となり得る閾値を持つ領域が存在しない場合、その追跡中の対象を保持して更に次のフレームで対応付く候補がないかを同様に調べる。今回対象とする家畜牛は移動速度が遅く、数フレーム程度であれば画像内での位置が変化しないため、数フレーム間隔が開いた場合でも対応付けが可能となる。

図 6 に例を示す。フレーム t における追跡対象 α に対応する候補の領域はフレーム t+1 に存在しない。この場合、追跡対象 α は t+1 において t の位置から移動していないと仮定する。フレーム t+1 の追跡対象とフレーム t+2 の候補を比較し閾値を上回れば、フレーム t の α とフレーム t+2 の α は同一物体と判定する。

図 6　追跡対象の対応付けの概念図

現在フレームの追跡対象の領域（追跡領域）に対する、次フレームの対応付けの各候補（追跡候補）との重なり（IoU 値）は 3.5 節で計算が完了している。この計算結果を用いて対応付けを行うために以下の処理を行う。

追跡対象の数を N とし、各追跡対象を$i = 1{\sim}N$で表す。現在の時刻を t とし、フレーム t での追跡領域を$A_{i(i=1\sim N)}$とする。フレーム t+1 におけるA_iの追跡候補を

$$C(A_i) \in 2^{M_{t+1}} \tag{3.1}$$

とする。また、A_iと$C(A_i)$の重なりの大きさ（IoU 値）を$iou\left(A_i, C(A_i)\right)$とする。

同フレームで候補領域を重複して使用しない、つまり t+1 において$C(A_i)$と$C\left(A_j\right)$の一致部分がないよう、下式の条件

$$C(A_i) \cap C\left(A_j\right) = \emptyset \quad (i, j \in 1{\sim}N, i \neq j)$$

を満たし、かつ、全ての追跡対象についての重なりの和が最大となる下式の条件

$$\sum_i iou\left(A_i, C(A_i)\right) \rightarrow max$$

を満たす追跡候補領域の組み合わせ

$$\{C(A_i) | i = 1{\sim}N\} \tag{3.2}$$

を求め、t+1 における追跡領域とする。

この時、

$$C(A_i) = \varepsilon \text{（空要素）}$$

になった追跡対象はフレーム t+1 での対応する候補がないので

$$C(A_i) = A_i$$

のままとする。

以上の処理を繰り返すことで追跡対象の対応付けを行う。

処理時間短縮のために、実際の処理では式（3.1）において、候補領域$m_c \in M_{t+1}$の内、$iou(A_i, m_c) > T_c$を満たすものの組み合わせのみを組み合わせ候補とする。また、式（3.2）の対応付けを求める際は、追跡領域と組み合わせ候補領域のすべての組み合わせを生成すると計算量が大きくなる場合があるため、$iou\left(A_i, C(A_i)\right)$の上位$C_c$個のみを組み合わせの候補とする。次章の実験では、$T_c = 0.3$、$C_c$ =上位 10 個とする。

4. 評価実験

4.1. 実験設定

提案手法の追跡精度を評価するために、牛を撮影した映像に提案手法を適用する。また、比較手法として 2.2 節で述べた OpenCV に実装されている 7 つの手法を用いて、同様に映像中の牛を追跡し、結果を比較する。

提案手法・従来手法共に、先頭フレームにおける追跡対象の位置は正解データを用いて与える。提案手法に対しては、正解データのマスク領域を先頭フレームにおける対象位置として追跡を行う。従来手法は正解データのマスク領域を覆うバウンディングボックスを与える。

また、提案手法でインスタンス・セグメンテーションに用いた Mask R-CNN は、牛のクラスを含めた 20 種類のクラスの識別ができるよう学習済みである。

4.2. 実験データ

実験には三種類の映像を用いた。表 1 に各映像の概要を表す。いずれの映像もフレームレートは 30fps、画像サイズは 1980×1280 ピクセルである。以下、各映像の特徴について説明する。

- C123（図 7）

  畜舎内で撮影された。後方には光源が存在する。2 頭の黒毛牛が横並びとなっている。フレーム数は 216 フレームである。

- M358（図 8）

  畜舎内で撮影された。常時、褐毛の 2 頭の牛同士による重なりが発生している。また、牛を囲うために存在する柵による遮蔽が発生し、手前に存在する牛の脚が遮蔽によって途切れている。フレーム数は 343 フレームである。

- M724（図 9）

  放牧に向かう際の屋外で撮影された。奥から手前へ向かって進行する 7 頭のホルスタイン種の牛がそれぞれ重なり合い、遮蔽が発生している。個体毎の類似性が高く、かつ対象が他の 2 映像より遠方に存在する。全体像が良く見える個体がいる一方、動画の最初から最後まで体の一部のみしか映っていない個体など様々な牛が存在する。フレーム数は 182 フレームである。

図 7　C123 の映像例

図 8　M358 の映像例

図 9　M724 の映像例

表 1　各動画における牛の情報

動画名	個体数	フレーム数	撮影環境	移動状態	牛の色
C123	2	216	牛舎内	あまり移動しない	白と黒
M358	2	343	牛舎内	あまり移動しない	茶
M724	7	182	屋外	大きく移動する	白と黒

- **正解データ**

評価を行うために上の 3 種類の映像に対し、2 フレームに 1 フレームの割合で正解データを作成した。フレームが違っても個体毎の答えの判別がつくよう色を指定してピクセル単位で塗りつぶしを行い作成した（図 10、図 11）。

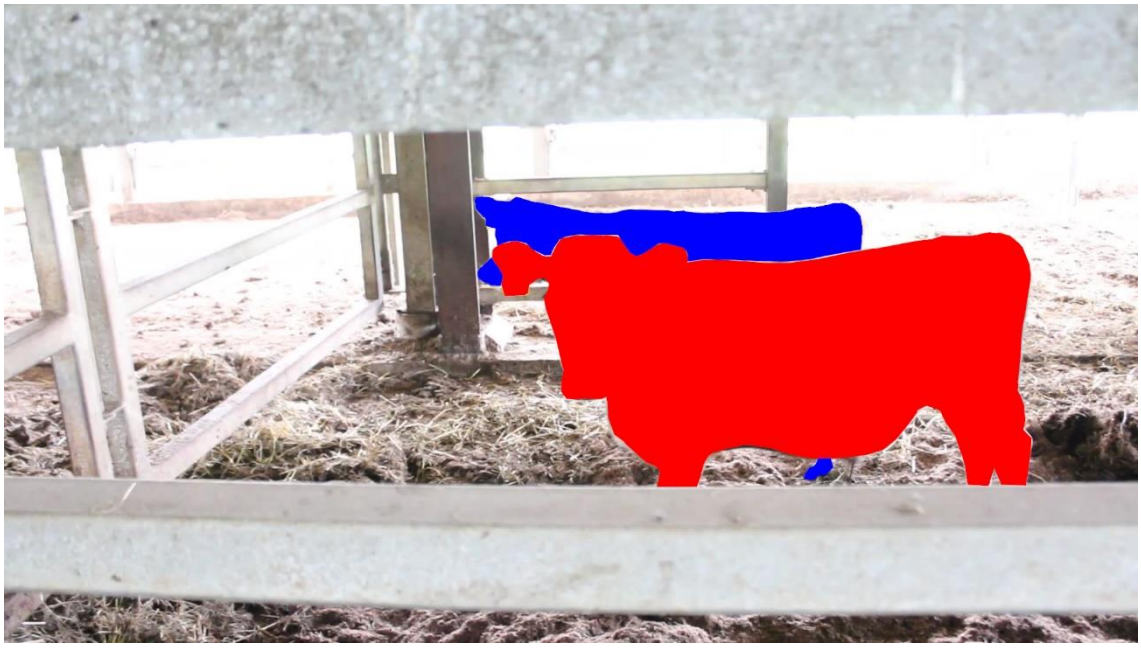

図 10　正解データにおけるマスク領域の塗りつぶし例（M358）

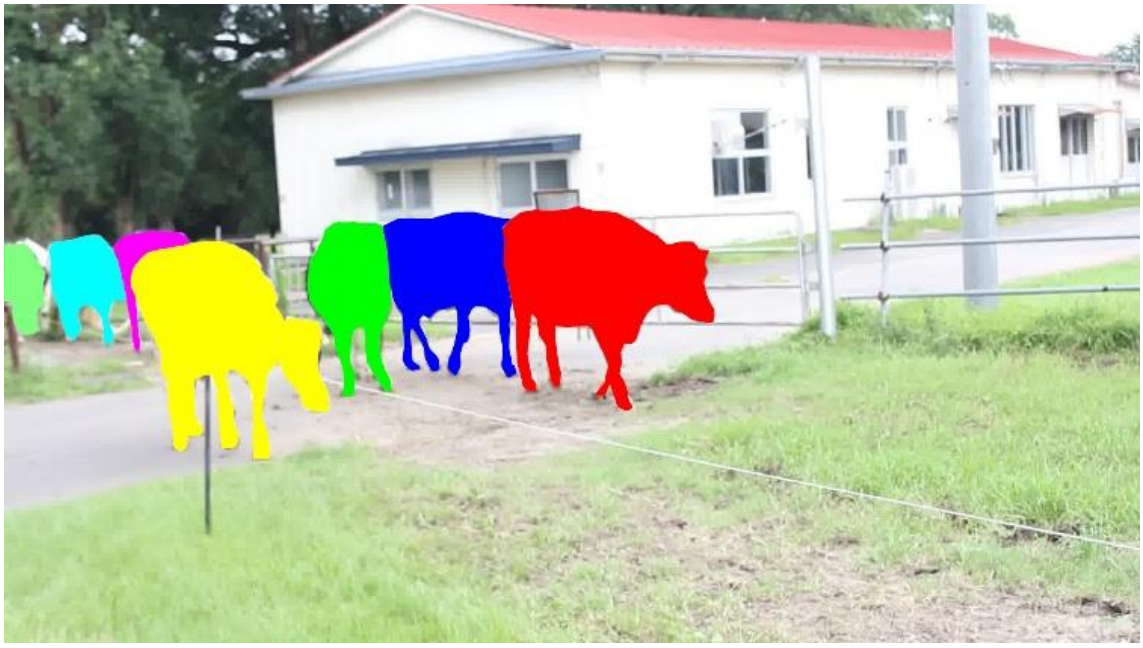

図 11　正解データにおけるマスク領域の塗りつぶし例（M724）

4.3. 評価指標

- Intersection over Union (IoU)

牛の追跡結果と正解データを比較評価するための指標として Intersection over Union(IoU)（図 12）を用いる。IoU は正解データの領域（正解領域）と追跡結果の領域（追跡領域）の一致度を示すものである。正解領域と追跡領域を合わせた領域に対する重複領域の割合で一致度を示す。

IoU の計算はバウンディングボックスに対して行う。提案手法の追跡領域と正解領域は物体領域を塗りつぶしたマスク領域であるのに対し、従来手法の追跡領域はバウンディングボックス（矩形）のため、そのまま IoU を取ると従来手法は不利となる。提案手法では追跡領域（マスク領域）を囲むバウンディングボックスを作成する。また正解データに対しても正解領域（マスク領域）を囲むバウンディングボックスを作成して IoU を計算するものとする。

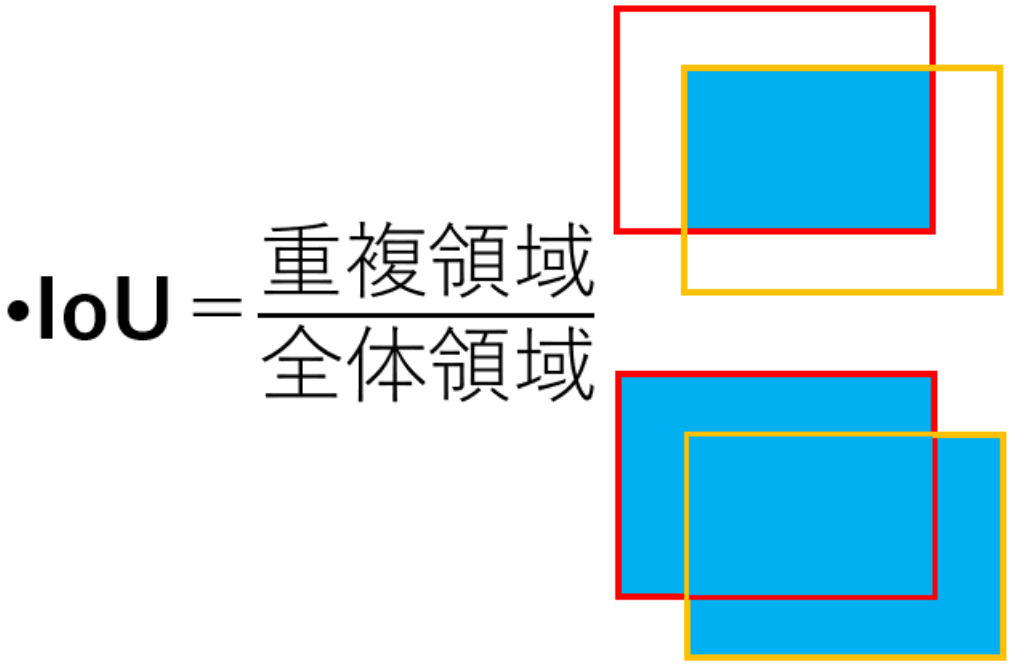

図 12　Intersection over Union(IoU)の概念図（赤枠：正解領域、橙枠：追跡領域）

- **重心位置の平均誤差**

正解領域と追跡領域の一致度を評価するもう一つの指標として重心位置の平均誤差を用いる（図 13）。正解領域の重心位置と追跡領域の重心位置とのユークリッド距離を求め、追跡全体での平均誤差を計算する。提案手法の追跡領域の重心位置と正解領域の重心位置は、マスク領域の重心位置とする。従来手法の追跡領域の重心位置はバウンディングボックスの重心位置となる。バウンディングボックスの重心位置はバウンディングボックスの中心位置となる。各手法の動画全体の平均 IoU と併せ、散布図で提案手法の評価を行う。

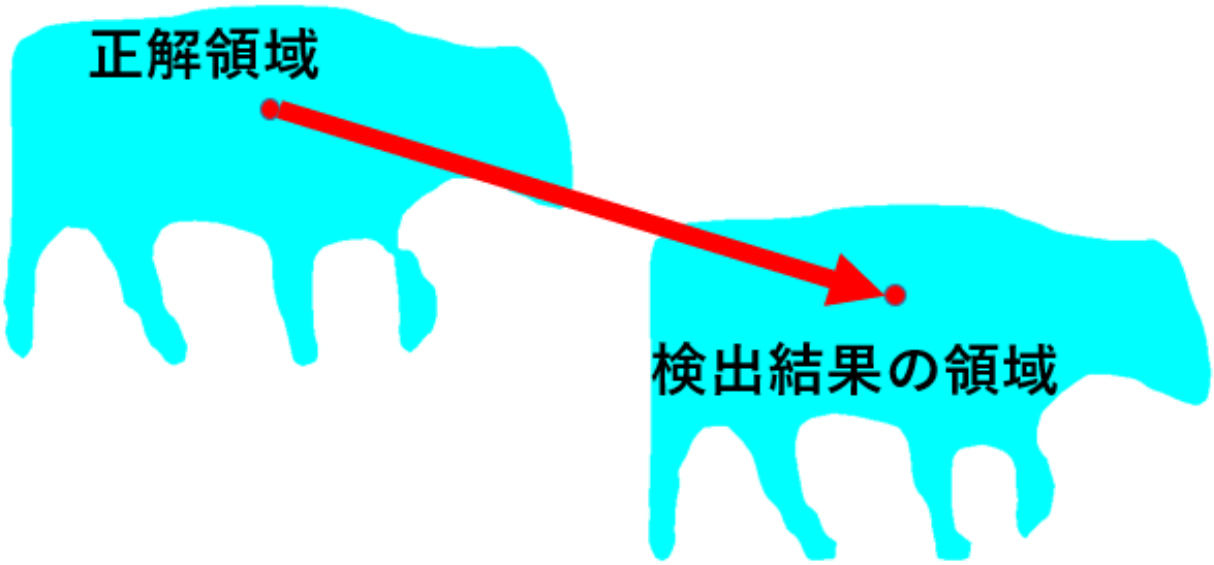

図 13　重心位置の平均誤差

4.4. 実験結果

- **C123 の実験結果**

C123 の動画例を図 14、実験結果を図 15、図 16 に示す。C123 で左の牛を tr00、右の牛を tr01 とする。図 15、図 16 は縦軸が全フレームに対する平均 IoU、横軸が重心位置の平均誤差の散布図である。平均 IoU が 1 に近いほど、また、重心位置の平均誤差が 0 に近いほど、全体の検出結果が良く高い精度で追跡をできている。提案手法は tr00、tr01 の両頭に対し、平均 IoU が約 0.8 と追跡手法として高い性能を示した。また、提案手法は従来手法と比較し平均 IoU、平均誤差の精度が良く、提案手法が有効であることを示した。

図 14　C123 の動画例（左の牛：tr00、右の牛：tr01）

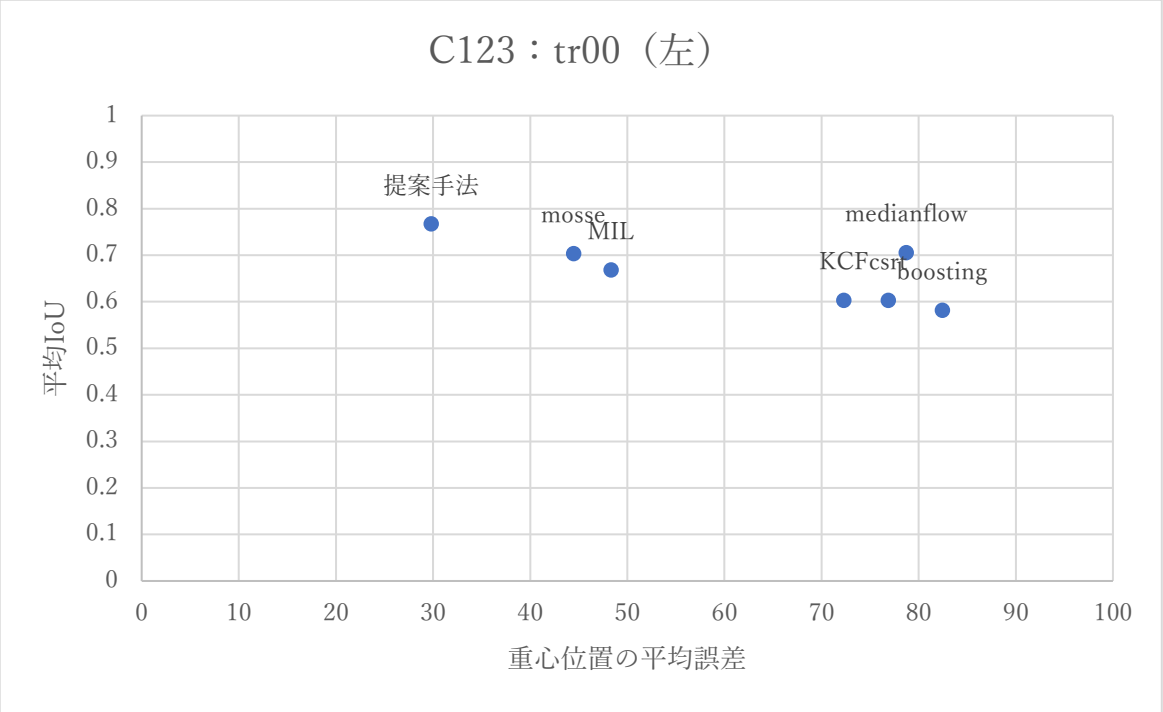

図 15　tr00（C123）の手法別散布図

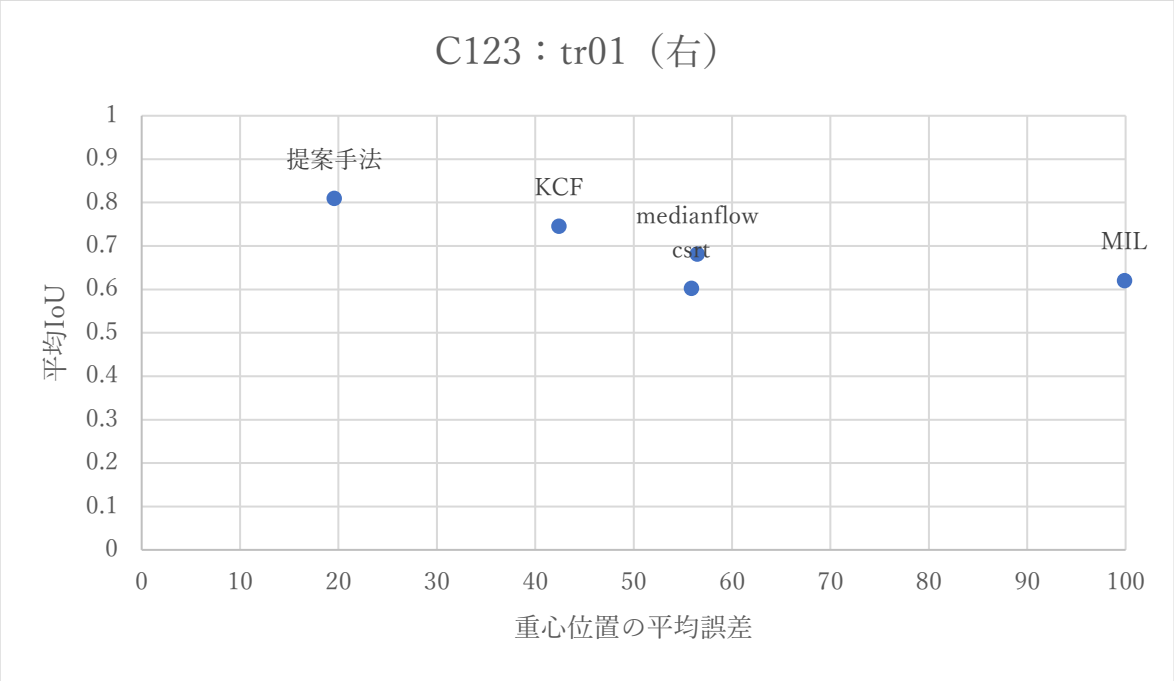

図 16　tr01（C123）の手法別散布図

- **C123 の IoU 値の推移**

図 17、図 18 に IoU 値のフレーム毎の推移を示す。図の横軸はフレーム、縦軸は IoU 値を示している。時間方向に数字の小さい方から順に IoU 値の推移を見ることで、追跡結果の確認ができる。

提案手法は tr00、tr01 共に他手法に比べ IoU 値が高いが、局所的に IoU 値が著しく下がっているフレームが存在する。これはインスタンス・セグメンテーションによる検出結果が悪かったフレームである。tr00 においては 50～60 のフレームの検出結果が悪いため従来手法の結果に比べ IoU 値が特に低くなっている。しかし、インスタンス・セグメンテーションは対象を見失っても同クラスの検出は可能なため、全体での IoU 値は従来手法に比べ高いものとなっている。

図 18 において、従来手法の Boosting 法は 80 フレーム周辺で IoU 値が急激に下がっている。tr01 は 80 フレーム周辺において、tr01 の牛が左から右に振り向き、頭部が画面外に出た。Boosting 法による tr01 を対象とした追跡は、頭部が見えなくなったフレームから隣接している tr00 に ID スイッチしたため、IoU 値が急激に下がった。また、そのまま隣接している tr00 の追跡を続けたため IoU 値は低く推移した。

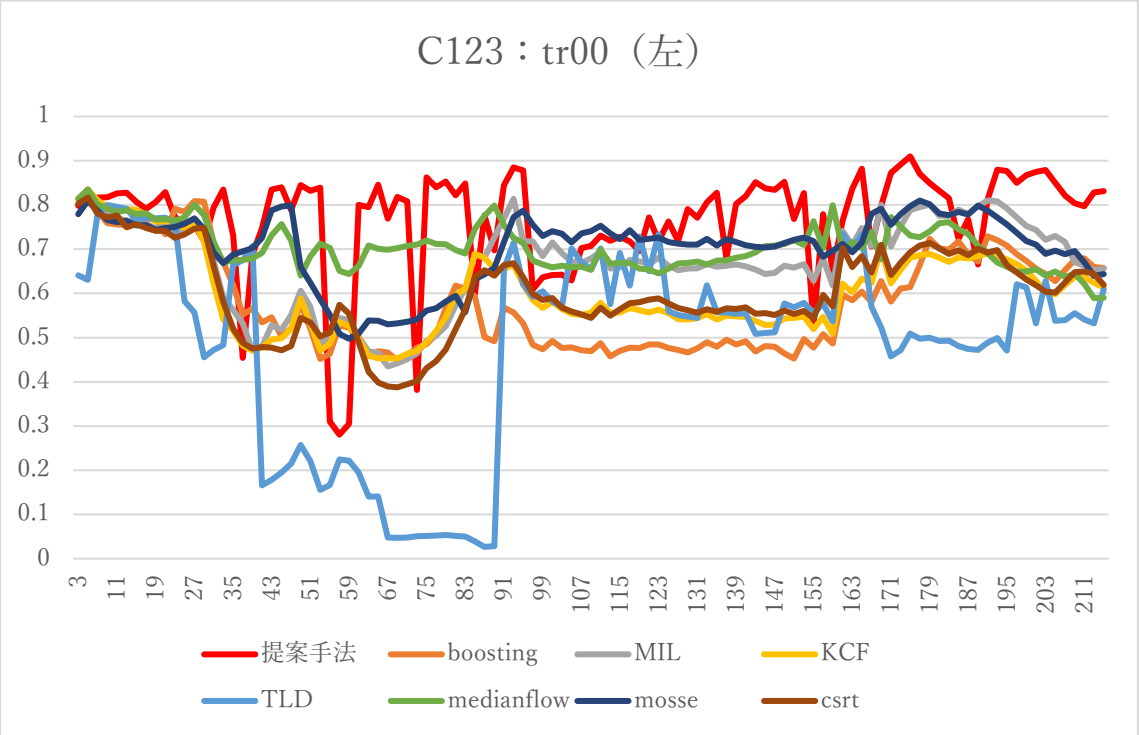

図 17　tr00（C123）の手法別 IoU の推移

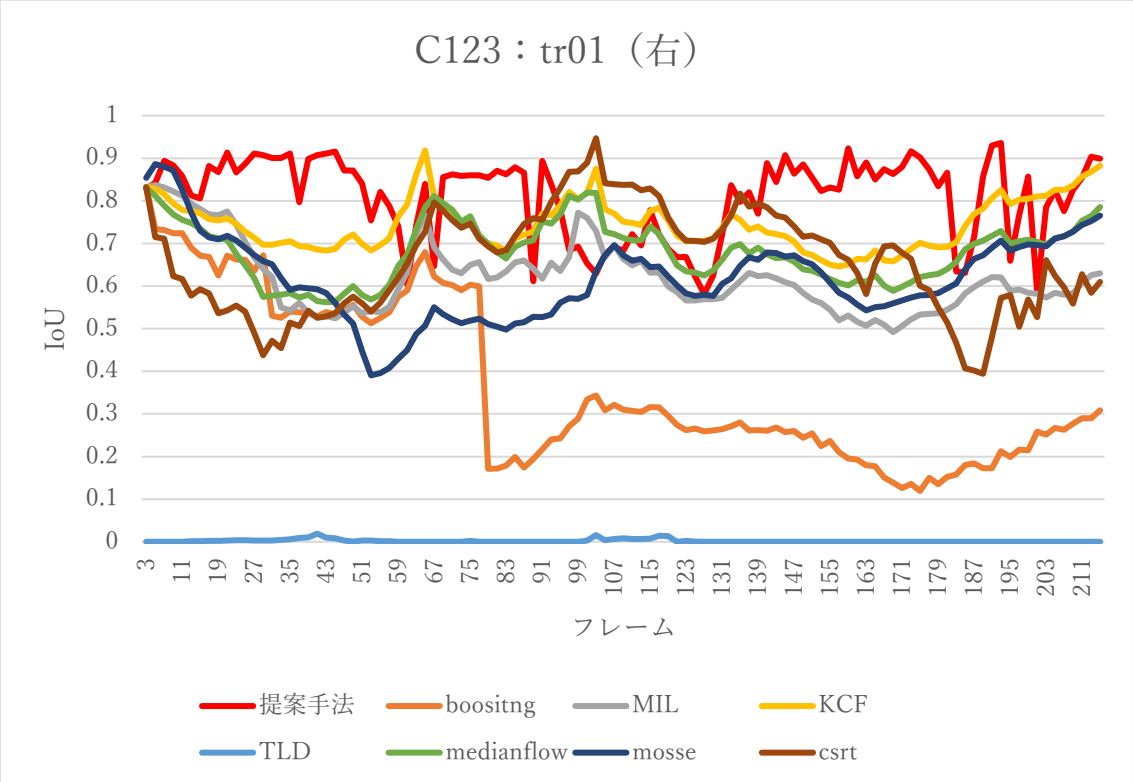

図 18　tr01（C123）の手法別 IoU の推移

- **M358 の実験結果**

M358 の動画例を図 19、実験結果を図 21、図 22 に示す。M358 では前方の牛を tr00、後方の牛を tr01 とする。提案手法の平均 IoU は従来手法よりやや優れている結果となった。tr00 において、多くの従来手法同士は散布図の位置が近い結果となった。tr01 においても同様の傾向となっているが、提案手法含め平均 IoU・重心位置の平均誤差がともに tr00 と比較して低い値となっている。また、提案手法は散布図に表示されている手法の中で最も重心位置の平均誤差が大きいが平均 IoU は最も良いという結果となった。

全体の平均 IoU が低い原因として、遮蔽による正解領域の分離が挙げられる。遮蔽によって tr01 の脚の一部が tr00 の腹部の下に現れているフレームが多く存在する(図 20)。本研究の評価方法として、正解領域が複数に分離していた場合、そのすべてを含むバウンディングボックスを生成し評価に用いるため、実際の追跡対象の正解領域と関係のない情報を多く含み評価をすることになる。そのため、各手法は平均 IoU が低くなった。このことから、遮蔽が生じることにより正解領域が分離している追跡対象も考慮した評価方法の導入が必要である。

また、tr01 の結果が tr00 に比べ低かったことの原因として、類似性の高い個体同士の重なりによる遮蔽が挙げられる。tr01 は遮蔽により見える部分が少ないため、検出を上手くできず、追跡を正確に行えていないと推測される。提案手法は従来手法と比較し、より正確な対象の存在する領域を用いて追跡を行うことで、遮蔽問題を解決した精度の高い追跡をしようとした。tr01 の分離した正解領域を含むバウンディングボックスの重心位置は、tr00 の正解領域を間に挟むため重心位置が tr00 側に寄る。従来手法は類似度の高い個体同士が隣接している場合、似ている領域を追跡してしまう可能性がある。今回は tr01 でなく類似している tr00 の領域を含めて追跡してしまったことから、重心位置は提案手法より近く、平均 IoU は低い結果になった。

図 19　M358 の動画例（手前の牛：tr00、後方の牛：tr01）

図 20　図 19 の画像の正解例（tr00 の正解領域：赤色、tr01 の正解領域：青色）

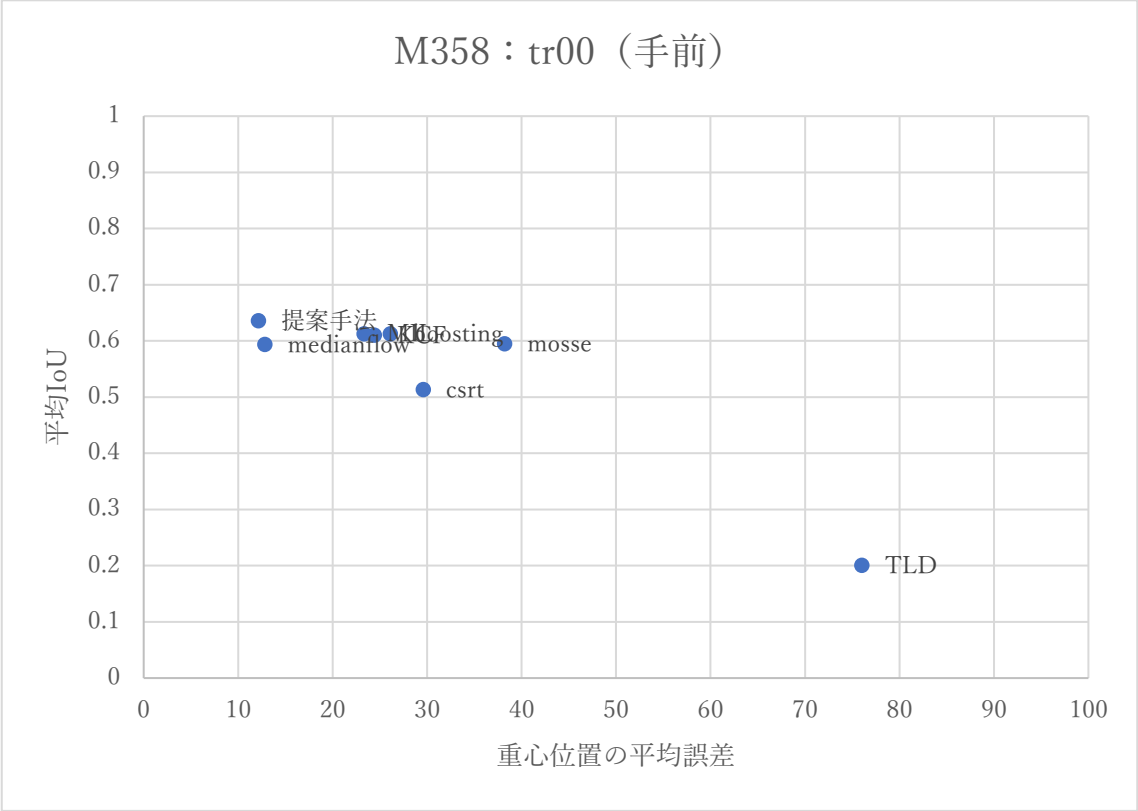

図 21　tr00（M358）の手法別散布図

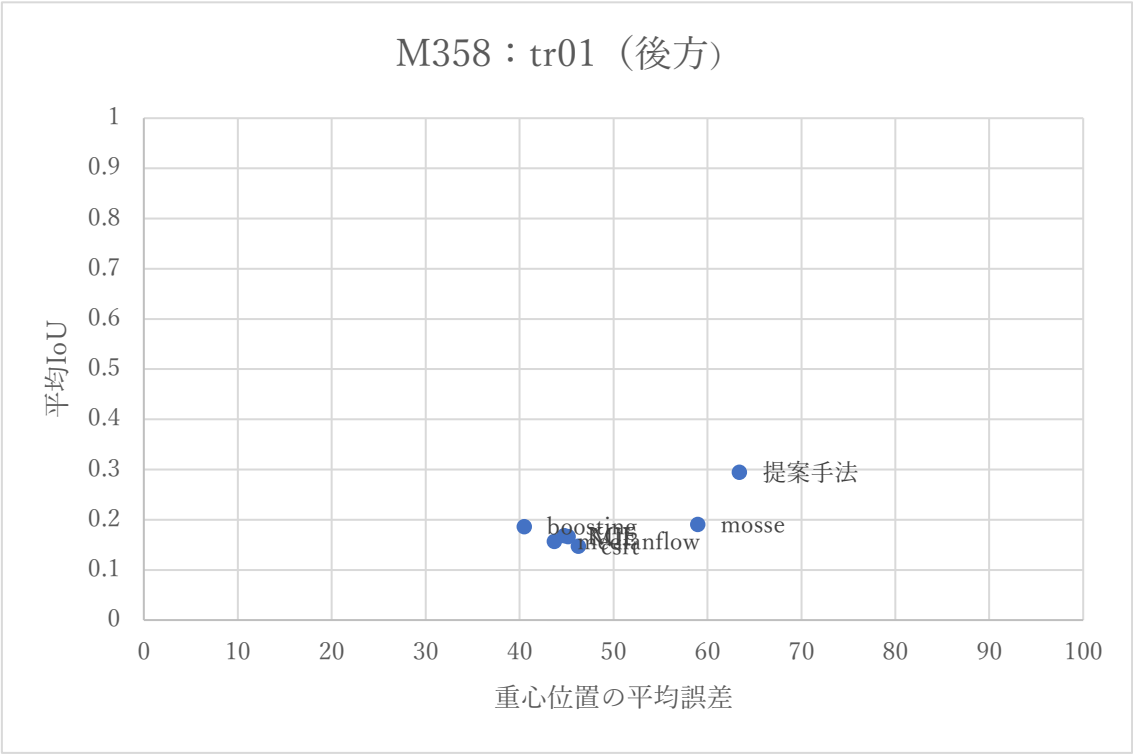

図 22　tr01（M358）の手法別散布図

- **M358 の IoU 値の推移**

他の動画に比べ全体の平均 IoU が低い tr01 の牛について注目する（図 23)。従来手法の多くは散布図で類似した傾向が出ており、フレーム毎の IoU 値の推移をグラフで確認するとほぼ横ばいであった。これは前述の通りバウンディングボックスに対象物体以外の領域を多く含んでいるためである。それに対し、提案手法は従来手法より多くのフレームにおいて IoU 値が高く 0.8 以上となるフレームも存在した。

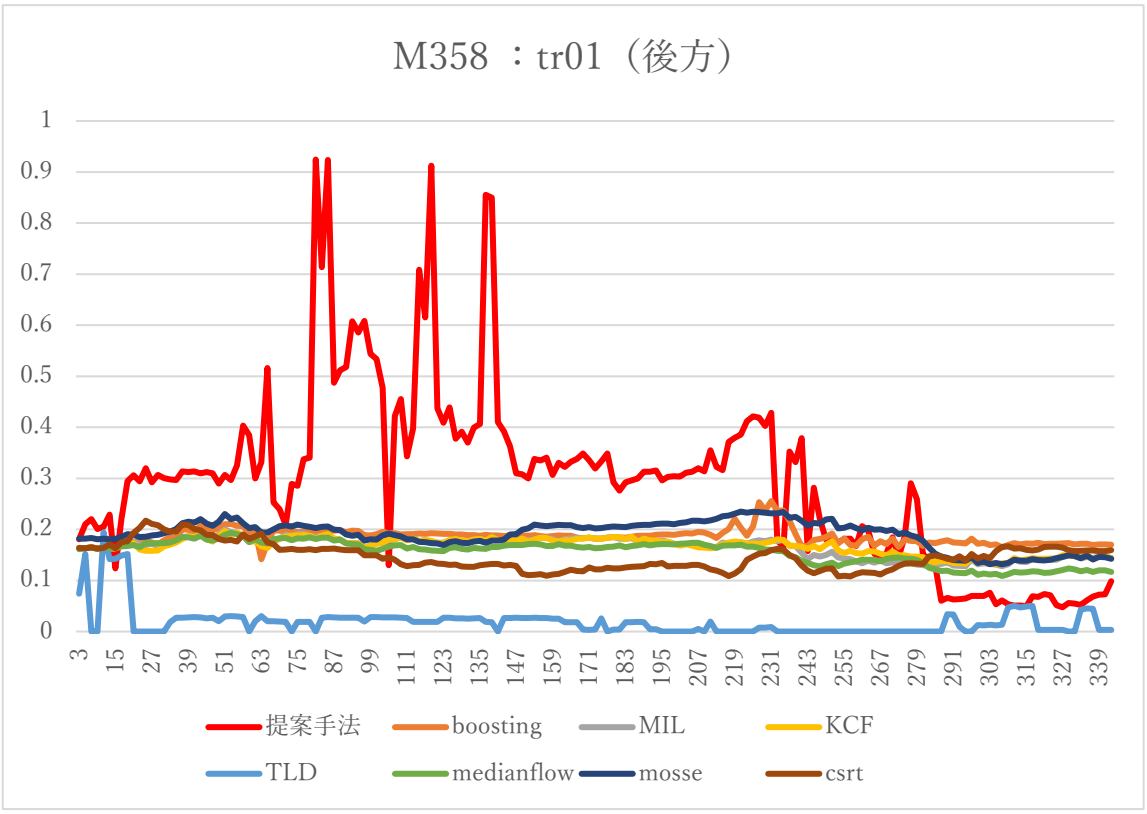

図 23　tr01（M358）の手法別 IoU の推移

- **M724 の実験結果**

M724 の動画例を図 24 に、実験結果の一部を図 25～28 に示す。M724 には 7 頭の牛が存在する。先頭フレームにおける位置を基準に画面右から順に tr00、tr01、…、tr06 とする。従来手法と比較して提案手法の結果が良好だったものは 7 頭中 4 頭（tr00,tr01,tr04,tr06）という結果となった。特に tr00（図 25)、tr01（図 26）においては平均 IoU・平均誤差共に従来手法と比較して高い性能を示した。tr06（図 27）の平均 IoU は少し低いが、従来手法を上回る結果となった。tr04 も同様の傾向であった。tr04,tr06 の 2 頭の牛は共に遮蔽によって常時体の一部のみが見える追跡対象となっている。類似性の高い別個体の後方に存在し遮蔽が発生していたことから、従来手法は遮蔽に対応できず十分な精度が出なかった。提案手法はこれらの遮蔽に対応して追跡をできた事から従来手法より高い性能を示した。

tr05（図 28）は従来手法と同等程度の結果となった。tr05 は画像内における存在する領域が小さく、かつ進行方向の変化が少ない。また、遮蔽の発生前後において、存在する領域があまり変化していない。これらの 2 点から、見た目の特徴が大きく変化しなかったため、従来手法は提案手法と同等の高い IoU 値をとなった。tr02、tr03 も同様に、従来手法と同等程度の結果となった。

図 24　M724 の動画例（右から順に tr00、…、tr06）

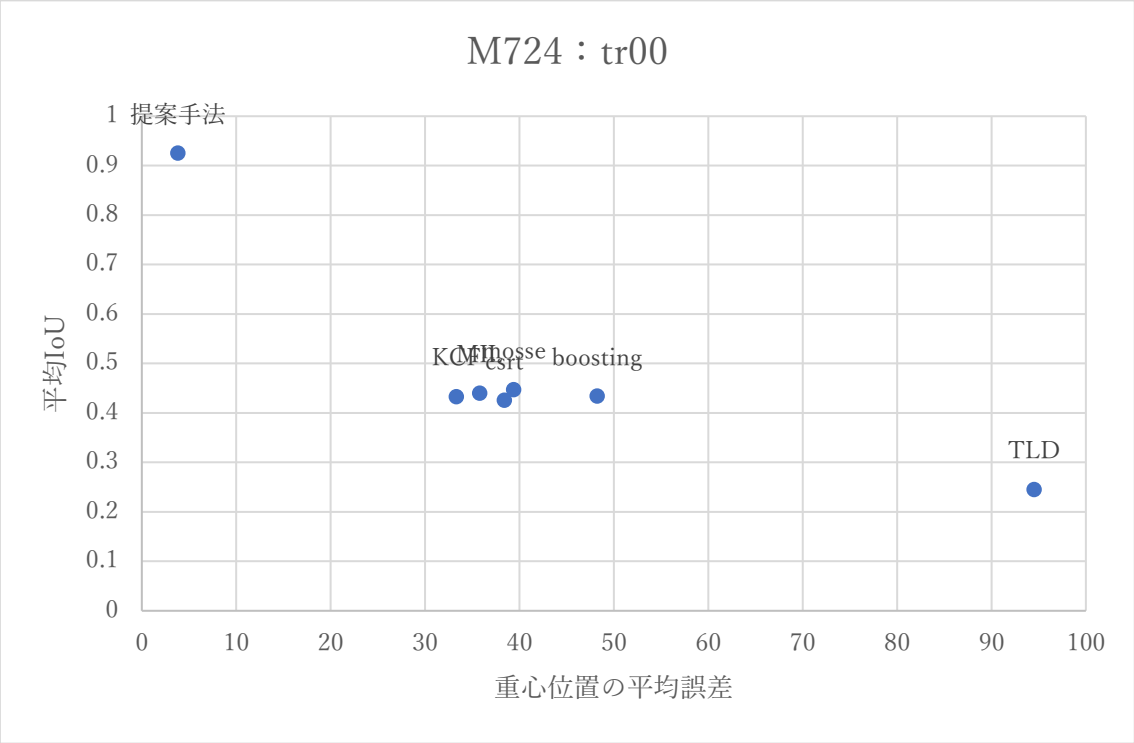

図 25　tr00（M724）の手法別散布図

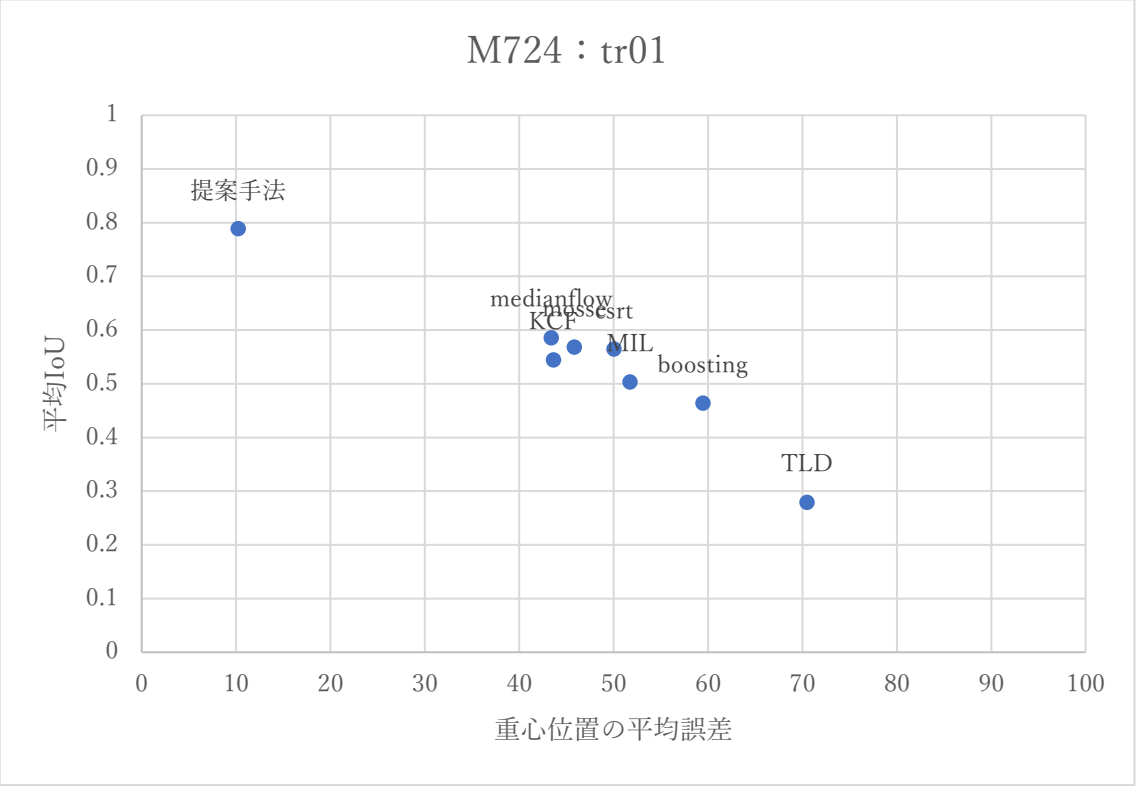

図 26　tr01（M724）の手法別散布図

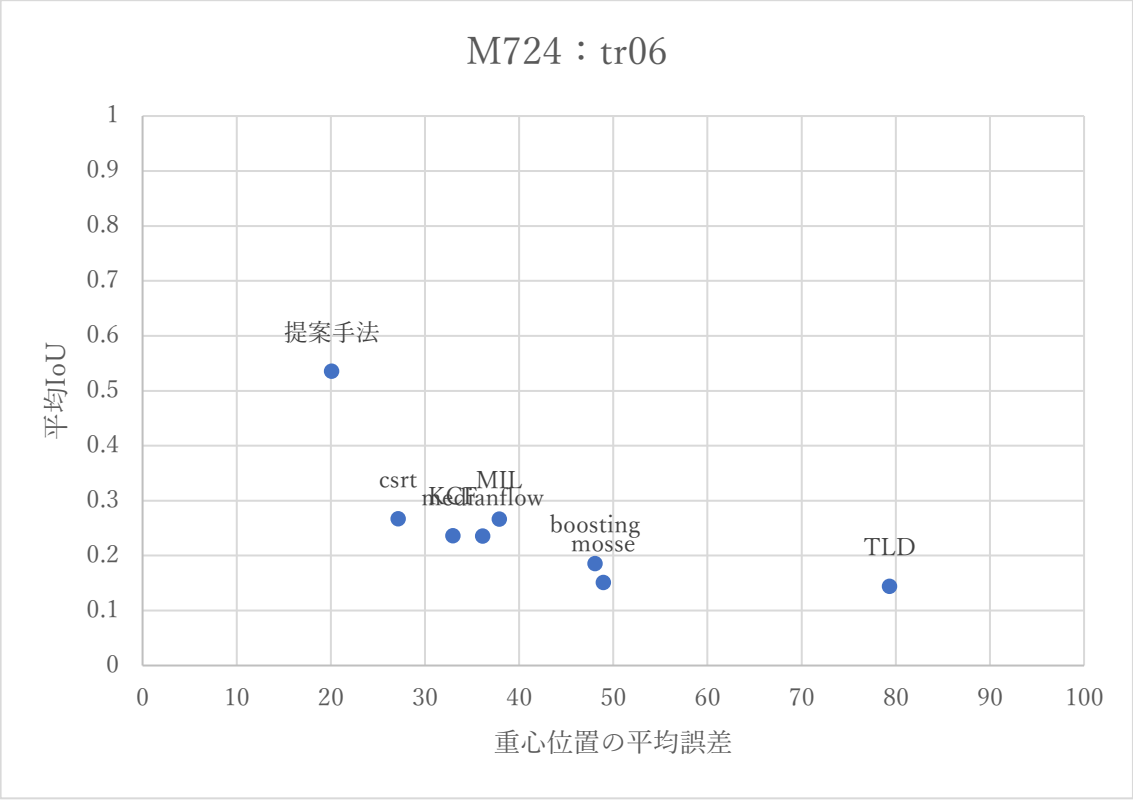

図 27　tr06（M724）の手法別散布図

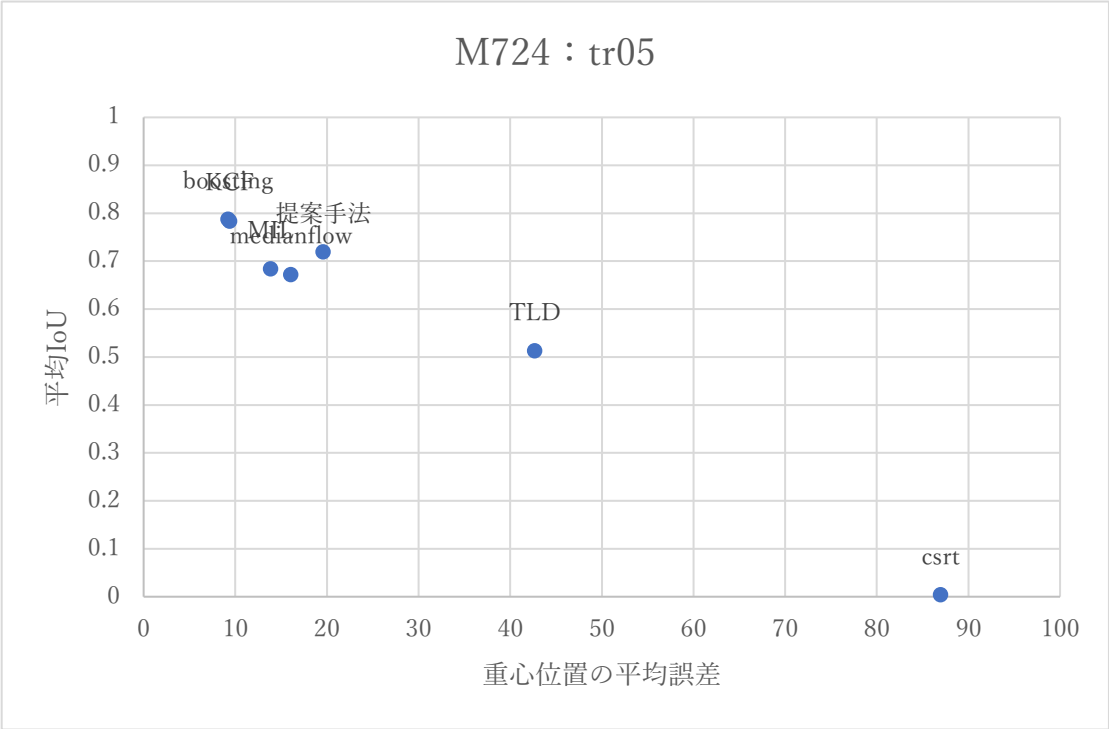

図 28　tr05（M724）の手法別散布図

- **M724 の IoU 値の推移**

図 29 に tr00 の IoU 値の推移を示す。提案手法は tr00 に対してほとんどのフレームにおいて、IoU 値が 0.9 を超える高い検出精度を示した。一方従来手法は、454 フレーム周辺から IoU 値が 0.5 近くで推移し追跡が十分な精度でできていない。tr00 は 454 フレーム周辺から向きを変え大きさの変化が生じていた。従来手法は大きさの変化に追従できなかったため、図 29 のように IoU 値が下がり続けた。

図 30 に tr01 の IoU 値の推移を示す。提案手法は、動画開始からしばらくは従来手法と同等の IoU 値で推移しているが、途中から IoU 値が 0.8 を超える場面が多く見られた。対して従来手法は開始時から最後まで 0.4～0.7 程度で推移した。動画開始時は tr01 に対し遮蔽が発生しているが、遮蔽が解消される際に提案手法では個体毎の検出が出来ていたために高い IoU 値を継続して出せた。一方すべての従来手法は、470 フレーム周辺から 490 フレーム周辺にかけて、IoU 値が下がり続けた。tr01 は 470 フレーム周辺から 490 フレーム周辺にかけて、tr00 と tr02 により生じた遮蔽の後ろで移動し、体の見える部分が増えた。従来手法は新たに見えた体の一部を tr01 の追跡対象に含めなかったため、IoU 値が下がった。

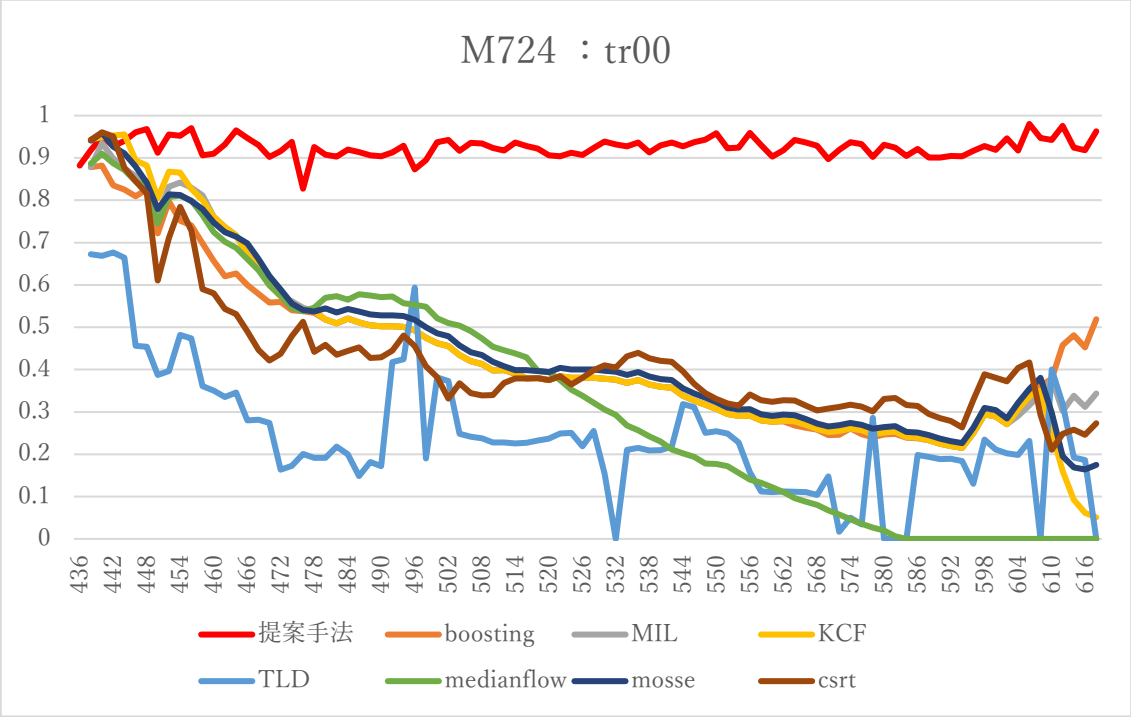

図 29　tr00（M724）の手法別 IoU の推移

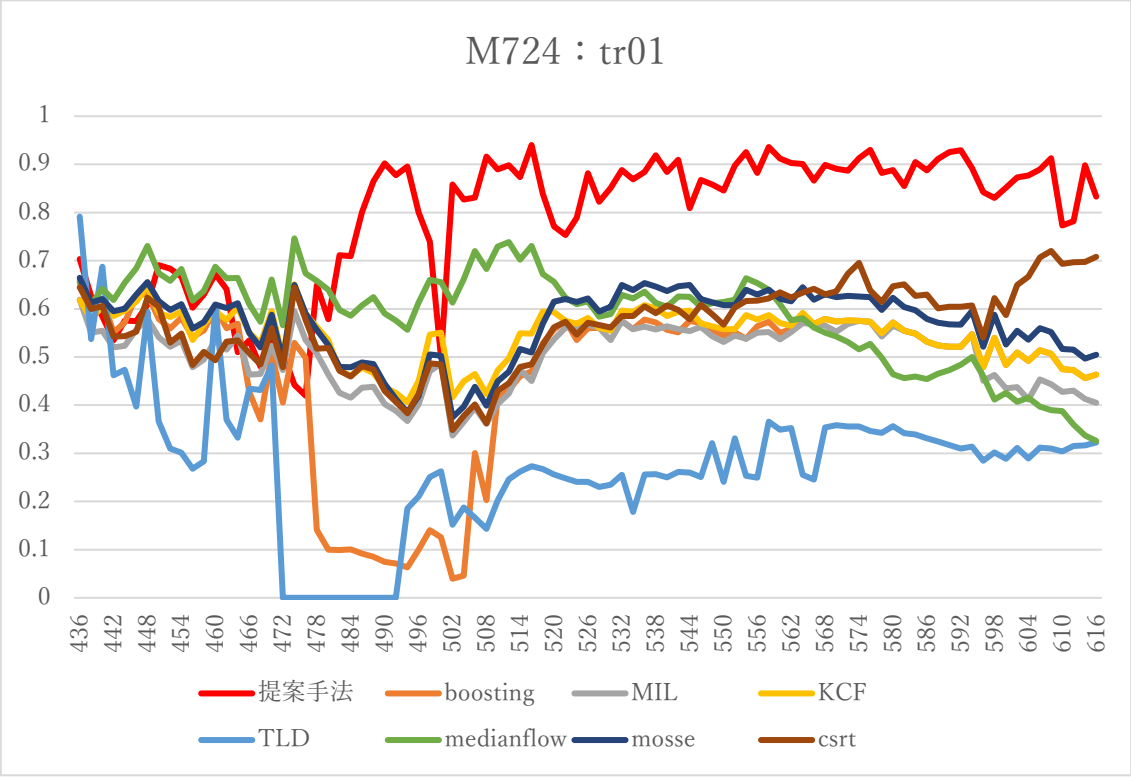

図 30　tr01（M724）の手法別 IoU の推移

次に従来手法と同等の結果となった tr05 について注目する（図 31）。tr05 はどちらも途中までは従来手法と比較して高い、もしくは同等レベルの IoU 値で推移していたが、592 フレームあたりから急激に低い値となっていた。また、従来手法よりやや実験評価が上回った tr06（図 32）についても tr05 と同様の傾向がみられた。これは実験に使った動画の 592 フレームあたりから手ブレが発生していたためである。本研究で対象としている牛は移動速度が遅く、数フレーム程度では大きく移動しないという前提があったため、手ブレの発生により疑似的に高速に移動したことから、フレーム間で重なりの大きい候補領域がなく追跡に失敗した。

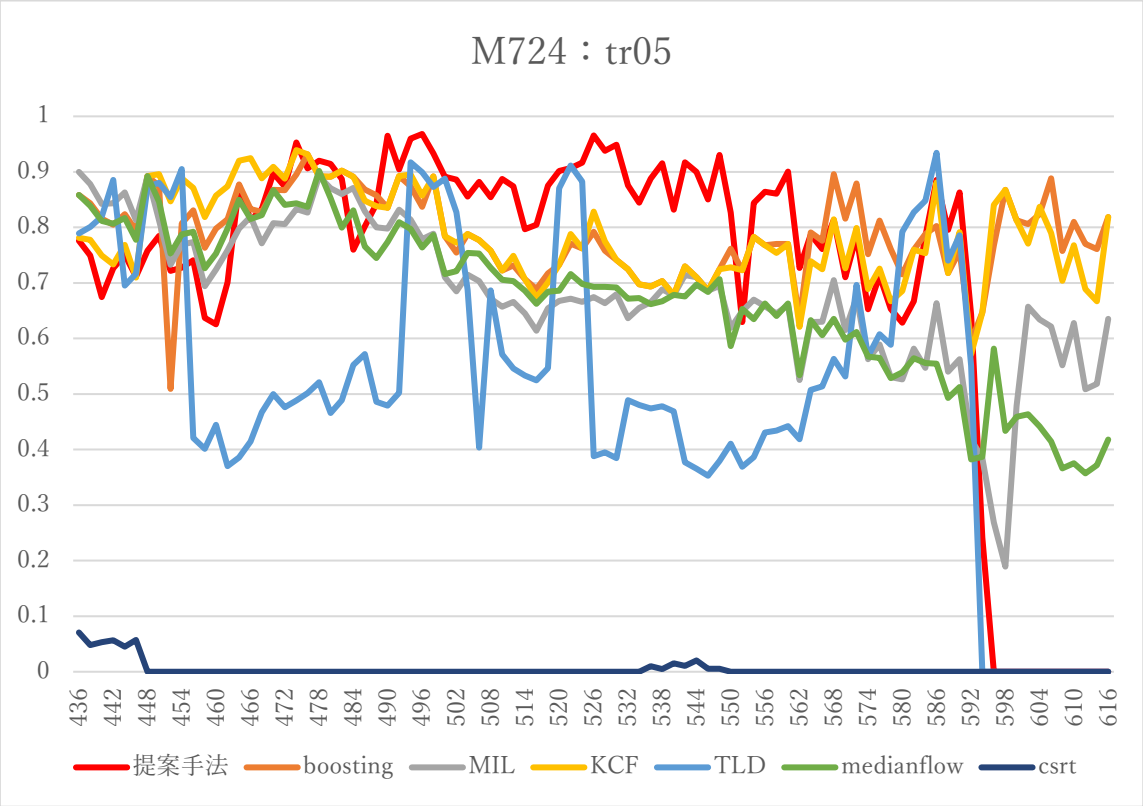

図 31　tr05（M724）の手法別 IoU の推移

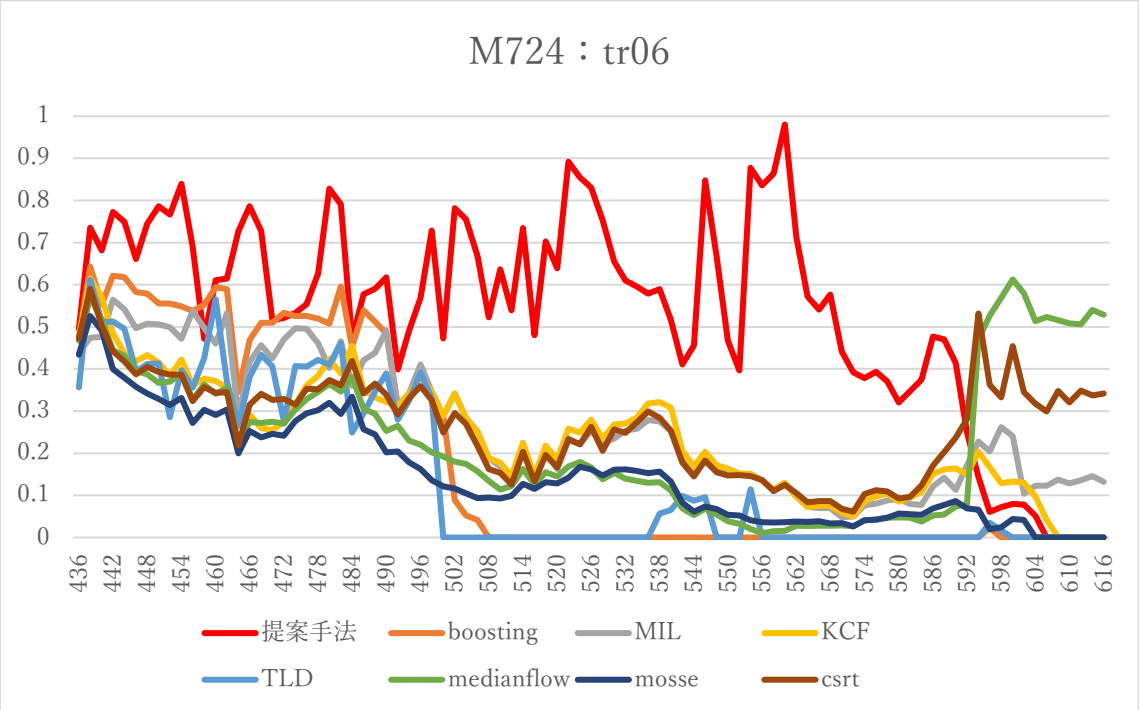

図 32　tr06（M724）の手法別 IoU の推移

図 33 にインスタンス・セグメンテーションの出力結果と追跡結果の成功例（500 フレーム周辺）の画像を示す。また、図 34 に 592 フレーム周辺、図 35 に 600 フレーム周辺におけるインスタンス・セグメンテーションの出力結果と追跡結果の画像を示す。左側がインスタンス・セグメンテーションの出力、右側が追跡領域である。右側の追跡領域の内、同じで色で表されているマスク領域は同一の個体として追跡している領域である。右列の追跡結果の画像において、tr00 は青色、tr01 は黄色、tr02 は紫色、tr03 は緑色、tr04 は水色、tr05 は赤色、tr06 は黄土色で示されている。

図 33 において、tr03 によって遮蔽が生じ一部しか見えていない tr04 や画像の端で見切れている tr06 等すべての牛を追跡できている。左列のインスタンス・セグメンテーションの結果を見るとすべての牛に対して検出ができている。tr03 はインスタンス・セグメンテーションの結果を見ると頭と胴体で領域が分かれてしまっているが、3.4 節で作成した組み合わせの候補を用いて 3.6 節の追跡対象と候補領域の対応付けを行うことで追跡できている。

インスタンス・セグメンテーションの出力は手ブレの発生する 592 フレーム前後や手ブレの収まった 600 フレームより後ろも tr05 に該当する牛の多くは出力できている。tr05 を追跡している領域（赤色）は手ブレの発生後（図 35 右列の上から 1 番目と 2 番目）に tr06 の追跡へと変わったことが確認できる。これは手ブレの発生前後において、画像内における tr05、tr06 の追跡領域の位置が近いことが原因である。tr05 の IoU 値の推移が下がった理由は追跡していた領域が別の追跡対象と認識したためである。一方、IoU 値が常時高かった tr00、tr01（青色と黄色）は、隣接する類似個体が存在しなかったため、手ブレが生じても継続して追跡できた。

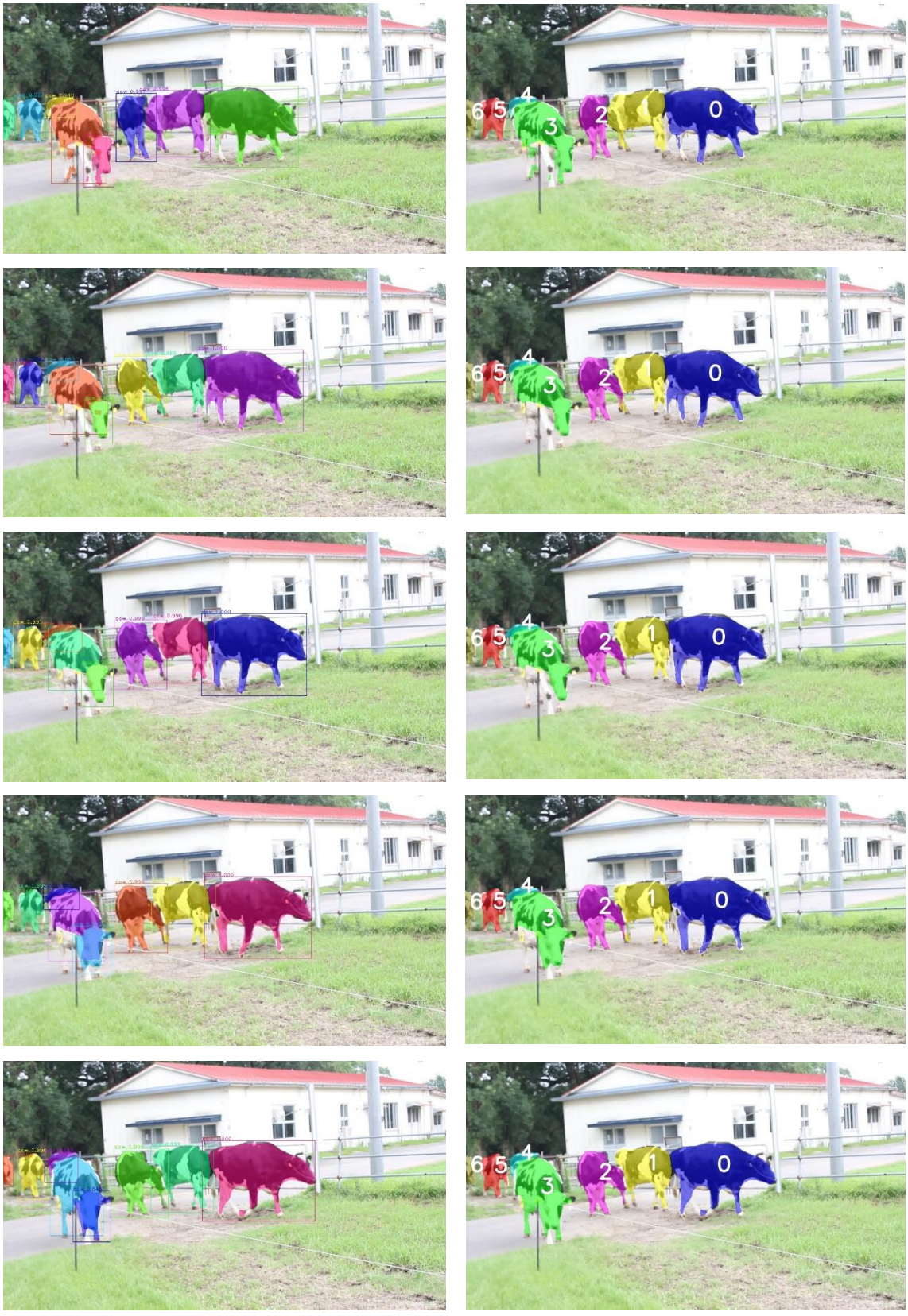

図 33　インスタンス・セグメンテーションの出力と追跡結果（500 フレーム周辺）

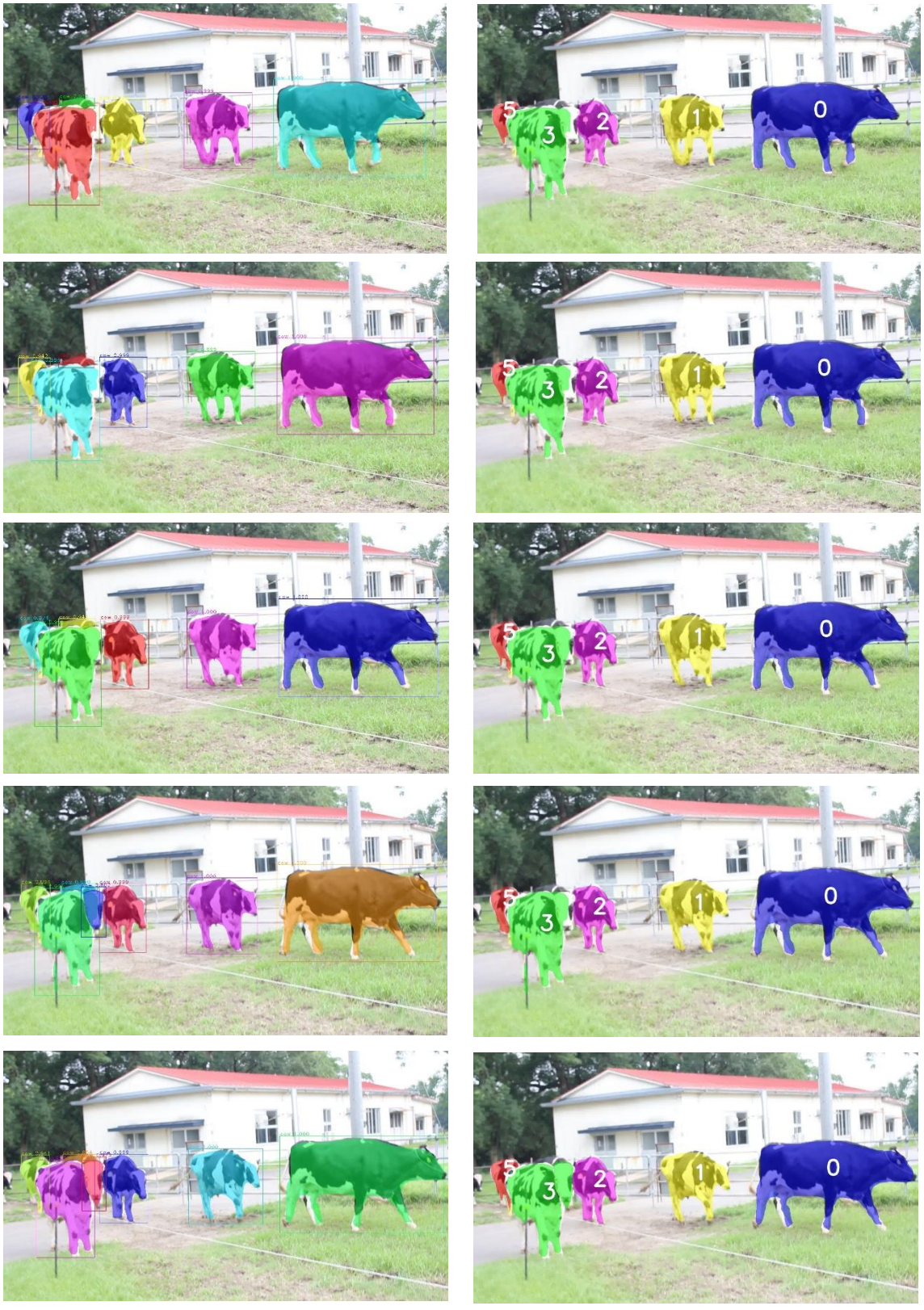

図 34　インスタンス・セグメンテーションの出力と追跡結果（592 フレーム周辺）

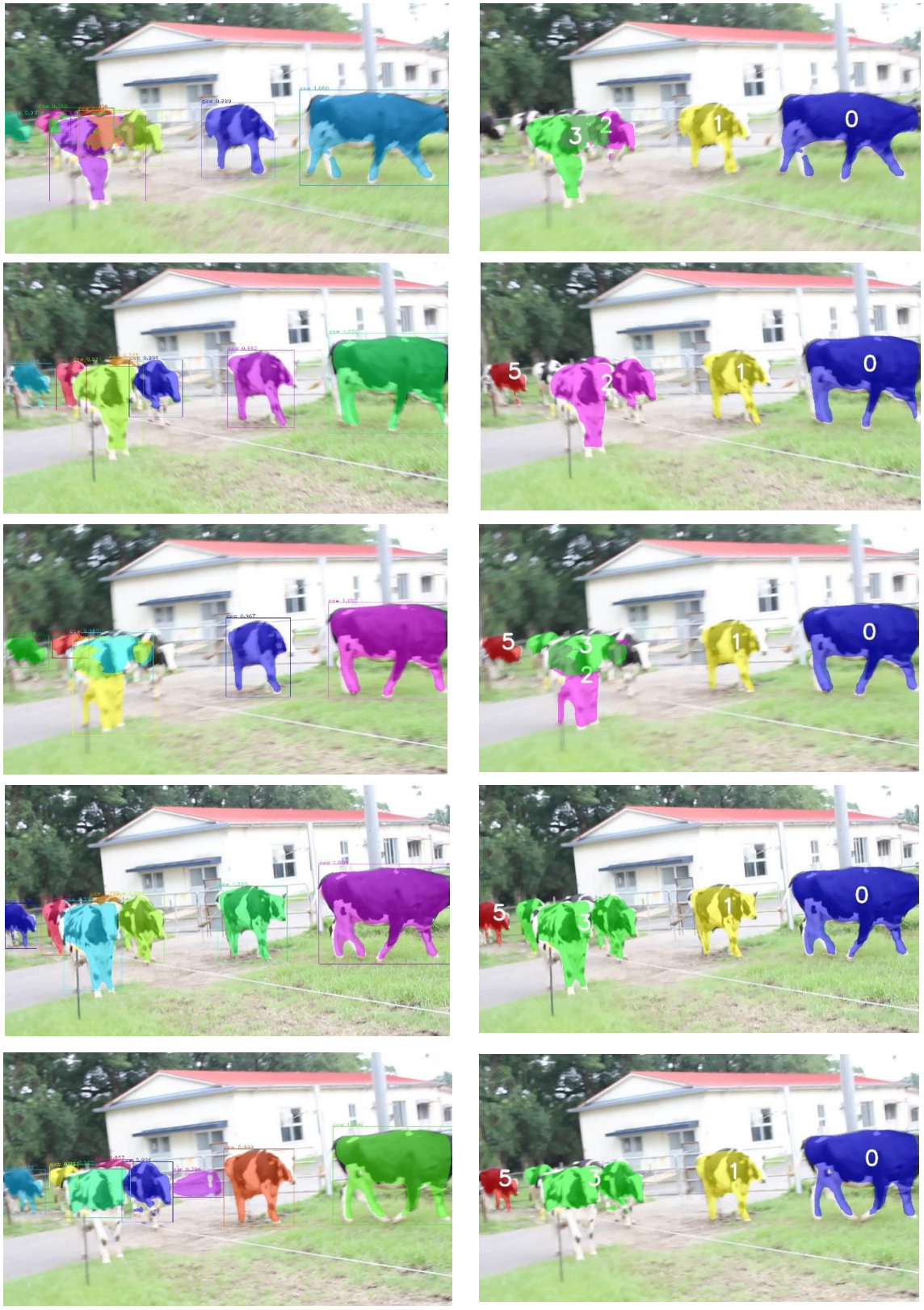

図 35　インスタンス・セグメンテーションの出力と追跡結果（600 フレーム周辺）

5. 結論

本研究では類似性の高い物体を対象に、個体毎の領域分割を用いて物体検出を行った情報を元に遮蔽に強い追跡を提案し、従来の物体追跡手法との比較実験により有効性を示した。

複数物体追跡における検出の精度は、追跡を行うための重要な要素である。検出精度が低いと、追跡する対象物体同士による遮蔽問題が発生した時、対象を上手く検出できず物体を見失う可能性がある。また、類似性の高い個体同士が隣接、または重なるようにして存在している場合、個体同士の重なりにより遮蔽が発生する可能性がある。従来の追跡手法では対象物体の見た目の変化に十分対応できず、追跡対象と類似性の高い領域によって遮蔽が発生すると個体同士の境界が曖昧となり検出の精度を下げてしまう要因となり得た。本研究では、インスタンス・セグメンテーションを用いて、従来手法より正確に追跡対象を検出した結果を利用することで、追跡問題における遮蔽の問題に対応した。

今後の課題として、提案手法において、検出結果としてマスク領域ではなくバウンディングボックスを利用した場合との追跡結果の比較が挙げられる。本研究では、インスタンス・セグメンテーションで得られるマスク領域を利用することで、従来の物体検出で得られるバウンディングボックスより正確な位置が求まり、追跡精度も向上すると想定したが、この点については評価実験を行っていない。追跡時にバウンディングボックスを用いた場合との比較実験により、この点を評価する必要がある。

また、今回は従来手法と条件を合わせるため、追跡結果の評価の際には、提案手法の追跡領域と正解領域を囲むバウンディングボックス同士の重なりの割合（IoU 値）を用いた。追跡対象に遮蔽が発生している状況であれば、正解領域が複数に分かれるためバウンディングボックスのサイズは大きくなる。検出結果が一方の領域を検出できていない場合、バウンディングボックスの重なりの割合は減少するが、追跡領域と正解領域の重なりをマスクレベルで求めることで、検出精度をより正確に評価できる。そのため、実験結果をマスクレベルで評価することが課題として挙げられる。

さらに、家畜牛と異なる、より移動の速い物体を対象とした類似個体同士での物体追跡への対応も課題である。本研究では家畜牛という移動速度が比較的遅い物体を対象に追跡を行ったが、検出漏れが発生した後の対応付けは、追跡対象の速度によって結果が大きく変わると推測される。一般の物体追跡問題に提案手法を適用する場合、移動速度の速い追跡対象にも対応できるよう、領域の重なり以外の類似性の評価基準も導入する等、改良が必要である。

謝辞

本研究を進めるにあたり、ご指導いただいた指導教員である椋木雅之教授に深く感謝をいたします。研究を進める中、同じところで詰まりがちである私に対して何度も細やかに指導をして頂けたこと、多忙な中たくさんの質問に快く回答いただけたこと、論文の添削や構成の指南といった様々な面でお力添えいただきました。椋木研究室に所属した約 4 年間は充実したものとなりました。本当にありがとうございます。

また、お忙しい中副査を務めていただきました坂本准教授、油田准教授に感謝いたします。

また、研究室内での相談や助言をくださった先輩、同級生、後輩の皆様に感謝いたします。

参考文献

[1]OpenCV ライブラリポータルサイト　http://opencv.org/
[2] Helmut Grabner, Michael Grabner, Horst Bischof, “Real-Time Tracking via n-line Boosting”, BMVC, 2006
[3] Boris Babenko, Ming-Hsuan Yang, Serge Belongie, “Visual Tracking with Online Multiple Instance Learning”, CVPR, 2009
[4] João F. Henriques, Rui Caseiro, Pedro Martins, Jorge Batista, “High-Speed Tracking with Kernelized Correlation Filters”, TPAMI37 (3), pp.583-596, 2014
[5] Zdenek Kalal, Krystian Mikolajczyk, Jiri Matas, “Forward-Backward Error: Automatic Detection of Tracking Failures”, ICPR, 2010
[6] Zdenek Kalal, Krystian Mikolajczyk, Jiri Matas, “Tracking-Learning-Detection”, TPAMI 34(7), 1409-1422, 2012
[7] 藤本雄一郎，青砥隆仁，浦西友樹，大倉史生，小枝正直，中島悠太，山本豪志朗, ”OpenCV3 プログラミングブック”, マイナビ出版, 2015
[8] David S. Bolme, J. Ross Beveridge, Bruce A. Draper, Yui Man Lui, “Visual Object Tracking using Adaptive Correlation Filters”, CVPR, 2010
[9] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, Matej Kristan, “Discriminative Correlation Filter Tracker with Channel and Spatial Reliability”, IJCV, 2018
[10] Kaiming He,Georgia Gkioxari,Piotr Dollár,Ross Girshick, “Mask R-CNN”, CVPR, 2017
[11] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, CVPR ,2016
[12] Mask R-CNN プログラム　https://github.com/akTwelve/Mask_RCNN