Vol.2022-CVIM-230 No.5
2022/5/12

Graph Transformerによる Single-Particle Tracking

神谷 聡[†1,a)]　角山 貴昭[†2]　楠見 明弘[†2]　堀田 一弘[†1,b)]

概要：近年，免疫系の研究が盛んに行われており，粒子追跡の需要が高まっている．しかし，機械学習による Single-Particle Tracking(SPT) の研究はあまり進んでおらず，精度の悪いソフト解析に頼っているのが現状である．SPT の問題点は主に 3 つある．1 つ目は，それぞれの分子に特徴の違いがなく，特徴の差異による追跡が行えないことである．2 つ目は，分子の動きはランダムであり，移動方向の予測が難しいことである．3 つ目は，分子同士の密度が高いため，ID スイッチが発生することである．そこで本稿では，これらの問題点を解決するために分子同士の関係性を考慮する Particle Tracking by Graph Transformer(PTGT) を提案する．提案手法は 2 つのデータセットにおいて従来法よりも高い精度を達成した．

キーワード：Multi-Object Tracking，Single-particle tracking，Transformer，顕微鏡動画

1. はじめに

顕微鏡で撮影した粒子の軌道を予測することを目的とした Single-Particle Tracking(SPT) は，分子解析において非常に重要である．**図 1** に SPT におけるの問題点を示す．SPT は，1 つの画像に含まれる追跡対象の数が 100 個から 1000 個まである大規模な追跡タスクである．そのため粒子の密度が高く，ID switch が発生しやすい．また，粒子同士が重なり粒子検出を失敗することがある．粒子は白い輝点として観測され，輝点の強度は時間によってランダムに変化する．時間的に変化し続けるため各粒子は個別の特徴を持たず，特徴の違いを用いた追跡方法を用いることができない．また，粒子の運動はランダムウォークであるため運動を予測することが難しく，運動予測による追跡は精度が低下する．

本論文では，これらの問題を解決するために以下のような手法を提案する．

- 各粒子にそれぞれ異なる特徴量を付与し，識別しやすくすることにより追跡を簡単にする手法．
- グラフを用いて移動距離の制約を表現し，近くの粒子同士のみを考慮する Graph Transformer．
- 輝点同士が重なり，検出を失敗した場合や，ノイズにより誤検出する場合に追跡を補正する追跡アルゴリズム．

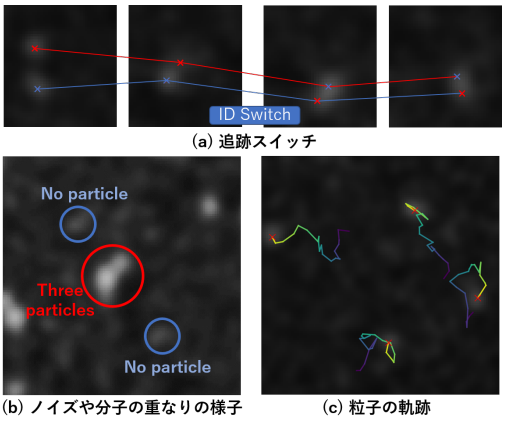

図 1　SPT における問題点

2. 関連研究

2.1 Point-Base Detector

粒子は多くの場合大きさに変化がなく，小さな輝点として観測される．Multi-Object Tracking(MOT) の追跡精度は検出精度に左右されやすいため，検出器の性能は非常に重要である．バウンディングボックスベースの検出器は数多く提案され，DPM[1] や，Faster R-CNN[2]，SDP[3] などがある．しかし，バウンディングボックスを用いた手法は，小さい物体の検出精度が低いことが知られている [3], [4]．そのため，粒子や細胞の検出には Point ベースの検出器を用いること提案されている．西村ら [5] は，U-Net[6] を用いた Point ベースの検出方法を提案し正確な位置情報を取得することに成功した．

†1　現在，名城大学
†2　現在，沖縄科学技術大学院大学
a)　180442042@ccalumni.meijo-u.ac.jp
b)　kazuhotta@meijo-u.ac.jp

2.2 Multi Object Tracking

MOT では，以下の 3 つの手法の研究が盛んに行われている．

(1) 関連問題と移動制約をグラフとしてモデル化した手法 [7], [8], [9], [10], [11].
(2) 運動を予測して追跡を行う手法 [12], [13], [14], [15], [16].
(3) 検出から追跡までのプロセスを end to end で行うモデル [17], [18], [19], [20].

１つ目の手法はグラフ理論を用いる手法である．物体の状態を基にグラフを作成するもの [8], [9] や，検出の関連問題を最適輸送問題として解くことにより追跡を行う手法 [21] などがある．その中でも移動距離の制約を用いたグラフはランダムに物体が動く場合でも使用できるため，SPT に応用することができる．2 つ目の手法は物体の移動を予測して追跡を行う手法である．人や動物などの規則性を持った動きを持つ追跡物体を追跡する場合，この手法は非常に有効であり，近くに複数の物体があったとしても移動方向の違いを利用しそれらを区別することが可能である．しかし，ランダムな移動をする SPT では移動方向の予測が難しく，精度が低下してしまう．3 つ目は物体の検出から追跡までを行う手法である．近年では，TransTrack[19],TrackFormer[22], や MOTR[20] など Transformer[23] を用いた手法が数多く提案されている．Transformer[23] は，Attention や Positional Encoding により追跡物体同士の関係や位置情報を考慮して追跡を行うことができる．

3. 提案手法

図 2 に Particle Tracking by Graph Transformer(PTGT) の概要を示す．PTGT は，物体の検出を行う 3D U-Net[24], 追跡マッチを行う Graph Transformer，検出ミスの補正を行う追跡アルゴリズムの 3 つのブロックに分けられる．まず，3D U-Net[24] により物体の位置 $\hat{p}_t = \{x_{t,i}, y_{t,i}\}_{i=0}^{\hat{N}_t}$ を検出する．(a) その後，Random Feature Assignment Module(RFAM) により検出位置に特徴量を付与し，(b)3D U-Net[24] でもう一度畳み込み処理を行う．そして，得られた特徴マップから検出位置の特徴量を抽出し，(c) 近傍付近のグラフを作成する．(d) 作成されたグラフを用いて Transformer は，マッチングマップを出力し，(e) 追跡アルゴリズムにより追跡結果を得る．(f)

3.1 Random Feature Assignment Module

Random Feature Assignment Module(RFAM) は，各物体に異なる特徴を付与しそれぞれを識別させやすくするモジュールである．まず，N_f 個の学習可能な特徴ベクトルを用意し，3D U-Net[24] の特徴マップの検出位置に用意した特徴ベクトルを付加する．追跡数の制限を無くすためにランダムに付与するが，同じベクトル同士を近くに付与すると識別が困難になる．そこで，検出位置から半径 r_f 以内には同じベクトルを付与しないようすることにより，追跡の際に特徴量の違いを用いて追跡対象を判断することができる．特徴量の付与は入力動画の 0 フレーム目と最後のフレームで行われ，その間にある画像には付与しない．付与された特徴量は 3D U-Net[24] により時系列上に伝搬されていく．3D U-Net[24] は，同じ追跡 ID を持つ物体同士は特徴が近づき，違う追跡 ID の物体同士は異なる特徴になるように学習する．また，RFAM は用意した特徴ベクトルはそれぞれ異なる特徴になるように学習していく．

3.2 Graph Transformer

Transformer[23] は，3D U-Net[24] で得られた検出位置と特徴ベクトルから物体のマッチングを行う．まず，特徴ベクトルに発生や消失の意味を持つ None token を付加する．None token はどの追跡物体とも関連しなかったことを意味し，過去の None token と関連付いた追跡物体は発生したと判断する．同様に未来の None token とマッチした追跡物体は消失したと判断する．

また，Transformer が物体の近傍周辺のみに着目するようにグラフを作成する．グラフは近い検出同士でエッジを形成するが，None token は位置に関係なくすべての検出とエッジを繋ぎ，追跡の発生，消失の可能性を残す．エッジを結ぶ検出物体は τframe 内に移動しうる最大距離 r_τ 内にある物体である．さらに，位置情報を考慮するために Positional Encording を行う．エンコーディング方法は Super Glue[25] で用いられていた Multi-Layer Perceptron(MLP) を使った方法である．また，DETR[26] のように Attention で Positional Encodings を行う．最後に，Transformer から得られた出力ベクトルを用いてマッチングマップを作成する．マッチングは 1 フレーム間の関連付け以外にも 2,3 フレーム間の場合でも行い，検出ミスや誤検出を補正する．

3.3 Time Attention

Time Attention は，Graph Transformer により作成されたグラフを用いて時系列情報を考慮する機構である．Time Attention は，過去を遡る Source-Target Attention と未来を考慮する Source-Target Attention を行う．過去を遡る Attention は，過去のベクトルと現在のベクトルで Source-Target Attention を行い，粒子が過去にどのような動きをしたかを考慮する．未来を考慮する Attention は，未来のベクトルと現在のベクトルで Source-Target Attention を行い，未来での状況を判断し現在の状態に反映する．最終的に，2 つの Source-Target Attention の結果を足し合わせ，出力とする．Time Attention は，1 フレーム差の過去と未来のベクトル変換しか行わないが，Transformer は L 回レイヤー処理を行うため，複数の時系列に情報を伝搬する．

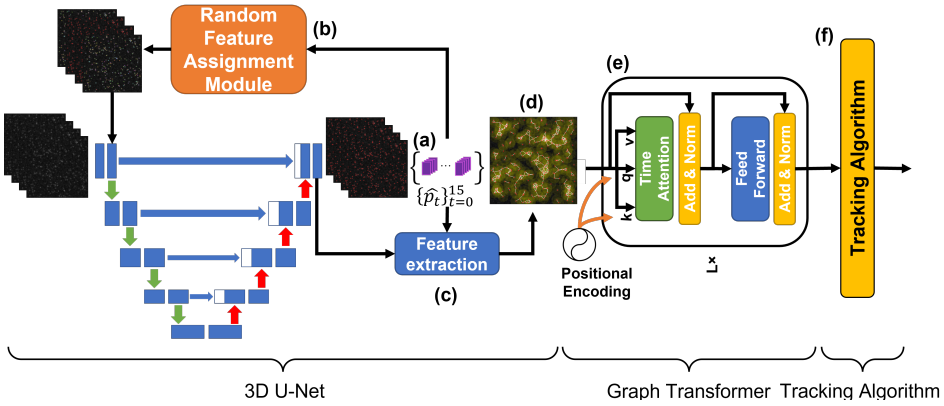

図 2 PTGT の概要

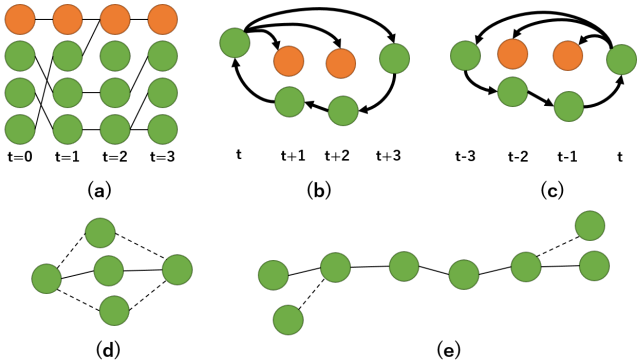

図 3 追跡アルゴリズムの概要

3.4 Tracking Algorithm

追跡アルゴリズムは，Graph Transformer で得られたマッチングマップと検出位置を用い，検出器による検出ミスと誤検出の補正を行う．**図 3** に追跡アルゴリズムの概要を示す．緑は検出された物体を表し，橙色は None token を表す．まず，1frame 差のマッチングマップを用い，図 3(a) のような追跡経路マップを作成する．物体が消失したと判断された追跡は図 3(b) のように補正される．消失したと判断された追跡は，2frame 差のマッチングマップを用い，2frame 後の検出とマッチするか検証する．マッチした場合，2frame 後の検出と時刻 $t+1$，1frame 差のマッチングマップを用いて消失した結果を上書きする．マッチしなかった場合は，3frame 差のマッチングマップを用いて 3frame 前の検出とマッチするか検証する．その後，同様の処理を行い 3frame 間の補正を行う．3frame 差の検出でもマッチしなかった場合，消失した追跡として判断される．物体が発生したと判断された追跡は，同様に過去の検出へ遡り図 3(c) のように補正を行う．

4. 評価実験

4.1 データセット

本研究では，2 つの顕微鏡シミュレーション動画像のデータセットを用いて評価実験を行った．データセットの概要を**図 4** に示す．1 つ目のデータセットは，公開されていない CD47 の顕微鏡シミュレーション動画像である．このデータセットでは粒子の密度と染色方法を設定することができ，染色方法は (a)GFP Low，(b)GFP，(c)TMR，(d)SF650 の 4 つで実験を行う．それぞれの染色方法は SNR が異なり，先述した順番は SNR が低い順となる．CD47 データセットでは長い時間動く対象が多く，粒子はランダムに運動する．

2 つめは，Particle Tracking Challenge(PTC) で公開されているデータセット [27] である．PTC データセットは 4 つのシナリオがあり，今回の実験ではその中から (e)Receptors と (f)Vesicles を用いる．また，1 つめのデータセットと同様に粒子の密度と SNR を設定することができ，$SNR = 7$ として実験を行う．Vesicles はランダムに運動するが，Receptors は直線的に移動する．このデータセットでは短い時間続く追跡が多く，1 シーケンス当たりの追跡数は Low で約 500 個，Mid では約 1500 個となる．

2 つのデータセットの画像の解像度は 512x512 であり，1 つのシーケンスは 100frame である．密度設定は 100(Low) と 300(Mid) とし，Mid を学習用データとして用いる．評価は Low と Mid の両方で行い，学習データを変更して得た平均精度で比較する．学習回数は 200epoch とし，Cosine annealing と Adam を使用した．RFAM のベクトルの数 N_f を 256，重複半径 r_f を 100 に設定し，物体が τ フレーム内に移動できる最大距離 r_τ は，学習用データを用いて計算する．

本研究では，2 つの従来手法と提案手法を比較する．PTGT は Transformer[23] を使用するため，同じく Transformer[23] を用いる Trackformer[22] と比較する．また SPT の機械学習モデルが公開されていないため，Cell Tracking を行う MPM[28] を比較対象とする．

4.2 CD47 データセットでの実験結果

まず，CD47 データセットでの実験結果を示す．**表 1** に密度設定が Low の時の追跡精度を示す．表中の IDF1，IDPr，IDRe，ID switch はそれぞれ ID F1 score，ID precision，ID Recall，Number of ID swtich を示している．また，ALL

表 1 分子密度 Low の時の CD47 データセットにおける追跡精度

Staning	Method	IDF1	IDPr	IDRe	ID switch
GFP Low	trackformer	22.60%	31.38%	17.68%	292
	MPM	11.78%	9.99%	14.40%	1863
	ours	**48.80%**	**34.77%**	**82.03%**	**105**
GFP	trackformer	22.50%	19.29%	27.20%	1095
	MPM	43.35%	44.47%	42.28%	557
	ours	**65.46%**	**51.22%**	**90.71%**	**71**
TMR	trackformer	25.75%	23.97%	28.01%	911
	MPM	59.88%	**62.18%**	57.75%	336
	ours	**66.53%**	52.49%	**91.88%**	**72**
SF650	trackformer	12.42%	8.97%	20.16%	1916
	MPM	55.24%	**57.27%**	53.36%	372
	ours	**64.19%**	49.55%	**91.74%**	**65**
ALL	trackformer	20.82%	20.90%	23.26%	1054
	MPM	42.56%	43.48%	41.95%	782
	ours	**61.24%**	**47.01%**	**89.09%**	**78**

表 2 分子密度 Low の時の CD47 データセットにおける検出精度

Staning	Method	F1	Pr	Re
GFP	trackformer	68.45%	58.44%	83.21%
	MPM	90.71%	92.87%	88.64%
	ours	**94.82%**	**96.12%**	**93.56%**
GFP Low	trackformer	43.08%	59.77%	33.69%
	MPM	70.57%	59.48%	87.05%
	ours	**89.48%**	**90.81%**	**88.21%**
TMR	trackformer	67.52%	62.65%	73.71%
	MPM	95.15%	**98.79%**	91.77%
	ours	**95.46%**	96.79%	**94.18%**
SF650	trackformer	58.99%	42.63%	95.74%
	MPM	94.49%	**97.92%**	91.30%
	ours	**94.95%**	96.26%	**93.68%**
ALL	trackformer	59.51%	55.87%	71.59%
	MPM	87.73%	87.27%	89.69%
	ours	**93.68%**	**94.99%**	**92.41%**

表 3 分子密度 Mid の時の CD47 データセットにおける追跡精度

Staning	Method	IDF1	IDPr	IDRe	ID switch
GFP Low	trackformer	13.28%	13.54%	13.04%	4533
	MPM	14.94%	15.24%	14.67%	4891
	ours	**39.15%**	**42.00%**	**36.68%**	**471**
GFP	trackformer	12.76%	10.00%	17.63%	6240
	MPM	29.64%	33.34%	26.68%	2718
	ours	**46.14%**	**47.19%**	**45.14%**	**462**
TMR	trackformer	12.69%	9.84%	17.87%	6298
	MPM	33.07%	37.36%	29.67%	2554
	ours	**46.50%**	**48.72%**	**44.48%**	**466**
SF650	trackformer	10.08%	7.40%	15.77%	7036
	MPM	31.19%	35.49%	27.82%	2582
	ours	**46.17%**	**48.33%**	**44.20%**	**432**
ALL	trackformer	12.20%	10.19%	16.08%	6027
	MPM	27.21%	30.36%	24.71%	3186
	ours	**44.49%**	**46.56%**	**42.63%**	**458**

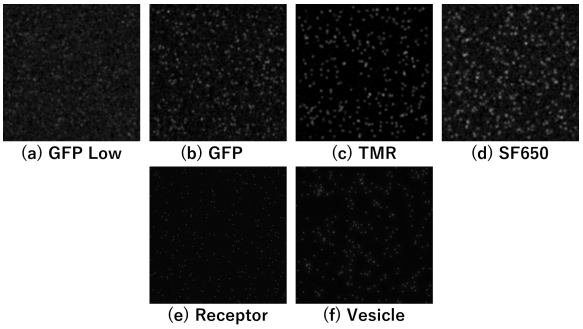

図 4 CD47 データセットの概要

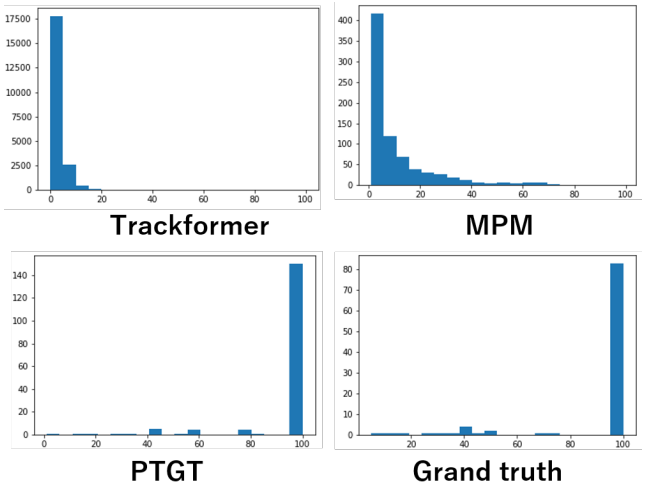

図 5 追跡の長さに関するヒストグラム

は全ての染色方法の平均の精度である．PTGT はすべての染色方法において高い精度を達成した．特に，追跡の入れ替わり回数を示す ID switch の数は著しく低下している．これは，Graph Transformer が交差する物体を注意深く考慮しているためであると考えられる．**表 2** に検出精度を示す．表 2 に示すように，染色方法が GFP Low や GFP の場合では PTGT と他の手法の検出精度の差が大きく，これにより追跡の精度が低下している可能性がある．しかし，TMR や SF650 の場合，検出精度は同じであるが PTGT は MPM[28] より追跡精度が大きく上回っている．これは PTGT が追跡のための強力な手法であることを示している．

次に，**表 3** に密度設定 Mid のときの追跡における精度を示す．密度が高い場合，追跡は非常に難しく，他の手法では ID Switch の数が増加している．しかし，PTGT は他の手法に比べて ID Switch の数が低く，全ての指標で精度が高い．GFP Low は検出が難しく従来手法では著しく精度が低下しているが，提案手法では他に比べて精度の低下は少ない．これは，追跡アルゴリズムにより検出の悪影響を軽減していることを意味している．

図 5 に密度設定が Low，染色方法 GFP での追跡が続くフレーム数に関してのヒストグラムを示す．縦軸は追跡の頻度，横軸は追跡の長さを示している．Trackformer[22] や MPM[28] は，分子の検出ミスや誤検出により追跡を失敗している．一方，PTGT のヒストグラムは正解のヒストグラムと非常に類似している．これは，PTGT が追跡アルゴリズムで追跡を補正することにより正しく追跡を行うことができることを意味している．

4.3 PTC データセットでの実験結果

汎化性を分析するために，公開データセットである PTC データセットを用いて評価を行う．**表 4** は，密度設定 Low の PTC データセットでの比較結果を示している．MPM と比較し，私たちの手法は大幅に精度を向上させている．**表 5** は検出精度を示しており，バウンディングボックスベースの検出方法である Trackformer は精度が低くなっている．point ベースの手法である U-Net や 3D U-Net を用いた MPM や PTGT は，検出精度は高い．

次に，私たちは Density 設定が Mid のときの結果を**表 6** に示す．MPM は移動予測が簡単な直線的な移動をする RECEPTOR においては高いパフォーマンスを達成しているが，ランダムに運動する VESICLE では精度が低下している．しかし PTGT は移動予測を行っていないためどちらにも対応でき，全体的な精度は高くなっている．

4.4 Ablation study

次に，提案手法の有効性を確認するために，RFAM，Time

表 4 分子密度 Low の時の PTC データセットにおける追跡精度

Molecule	Method	IDF1	IDPr	IDRe	ID switch
RECEPTOR	trackformer	40.14%	34.90%	47.31%	494
	MPM	67.13%	**69.04%**	65.31%	442
	ours	**71.92%**	68.46%	**75.98%**	**33**
VESICLE	trackformer	37.80%	29.42%	52.88%	1207
	MPM	65.64%	68.15%	63.32%	517
	ours	**78.87%**	**81.29%**	**76.59%**	**29**
ALL	trackformer	38.97%	32.16%	50.09%	851
	MPM	66.39%	68.60%	64.32%	480
	ours	**75.39%**	**74.88%**	**76.28%**	**31**

表 5 分子密度 Low の時の PTC データセットにおける検出精度

Molecule	Method	F1	Pr	Re
RECEPTOR	trackformer	45.68%	61.90%	52.53%
	MPM	88.93%	91.47%	86.53%
	ours	**92.38%**	**91.65%**	**92.02%**
VESICLE	trackformer	46.53%	83.65%	59.79%
	MPM	90.40%	**93.85%**	87.20%
	ours	**94.31%**	92.68%	**93.49%**
ALL	trackformer	46.11%	72.77%	56.16%
	MPM	89.67%	**92.66%**	86.87%
	ours	**93.35%**	92.17%	**92.75%**

表 6 分子密度 Mid の時の PTC データセットにおける追跡精度

Molecule	Method	IDF1	IDPr	IDRe	ID switch
RECEPTOR	trackformer	42.26%	37.44%	48.53%	2336
	MPM	**60.64%**	**66.29%**	55.87%	1571
	ours	57.12%	44.24%	**80.58%**	**309**
VESICLE	trackformer	35.12%	28.60%	45.49%	4655
	MPM	55.91%	**62.51%**	50.57%	2291
	ours	**60.19%**	55.41%	**65.89%**	**195**
ALL	trackformer	38.69%	33.02%	47.01%	3496
	MPM	58.27%	**64.40%**	53.22%	1931
	ours	**58.65%**	49.82%	**73.24%**	**252**

Attention，Positional Encoding のパラメータ変化による精度の違いを評価する．実験は，CD47 データセットの標準的な染色方法である GFP を用いて行う．

追跡器にとって物体の特徴の違いは非常に重要である．そのため，RFAM の各パラメータを変化さた場合の影響を分析し，**表 7** に結果を示す．Assign は RFAM の有無を示しており，Feature num と overlap range はそれぞれ N_f，r_f を示している．RFAM によって付与された特徴量は追跡器にとって，重要な情報であることを示している．重複半径 r_f が 50 の場合では間のフレームで物体が移動し，同じ特徴の物体が近づいてしまう可能性があるため精度が低下している．また，特徴ベクトルの数は追跡対象の数に対して最適な数がある．しかし特徴ベクトルの数 $N_f = 256$，粒子の密度:Mid の場合でも精度を保つことが出来る．これは，重複半径 r_f の円内にある物体の個数が N_f 以下であれば精度を保つことができることを意味している．

次に，注意機構が追跡にどのように影響しているかを分析する．**表 8** に，Attention を 4 つのタイプに変更した結果を示す．Normal はグラフを用いない同じフレームにある物体の Self Attention，Distance はグラフを用いる Self Attention，Time はグラフを用いて他の時間の物体との Source-Target Attention，Both は Distance と Time をどちらも行うモデルである．Super Glue[25] では，Self Attention と Source-Target Attention を交互に行うと精度が向上するとなっていたため Both を検証した．しかし，表 8 に示すように結果は Both よりも Time のほうがパフォーマンスが高いことを示している．これは，SPT では同じフレームの物体の関係性はあまり重要ではないことを示している．

最後に，MOT にとって重要な位置情報の処理方法についての分析を行う．Transformer など [22], [23], [29], [30] で用いられている Sin，DETR など [26], [31] の Learned，Super Glue など [25], [32], [33] の MLP の 3 つの方法で分析し，**表 9** に結果を示す．SPT では，MLP を用いた手法は高いパフォーマンスを達成している．正弦波エンコーディングは特徴が弱く，学習可能パラメータエンコーディングは分解能が低い欠点を持つ．しかし，MLP を用いた手法は両方の欠点を克服しており，物体の特徴が乏しい SPT では相性がよい．

表 7 RFAM の有効性の実験結果

Assign	Featrue num	overlap range	Low	Mid
-	-	-	64.00%	45.88%
✓	256	50	64.50%	45.80%
✓	256	100	**65.46%**	46.14%
✓	512	100	64.20%	**46.21%**

表 8 Attention の有効性の実験結果

Attention Type	Low	Mid
Normal	63.22%	45.08%
Distance	63.46%	45.45%
Time	**65.46%**	**46.14%**
Both	64.42%	45.56%

表 9 Positional Encoding の有効性の実験結果

Positional Encoding	Low	Mid
Sin	61.79%	44.35%
Learned	63.24%	45.30%
MLP	**65.46%**	**46.14%**

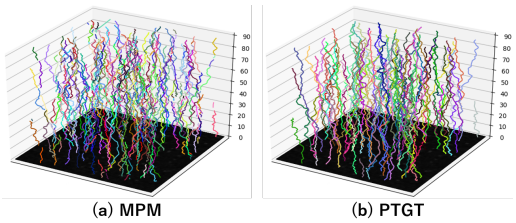

図 6 三次元描写した追跡結果

4.5 定性評価

最後に，提案手法の定性的な評価を行うために，**図 6** に MPM と PTGT の追跡結果を三次元描写した結果を示す．PTGT は他の手法とは異なり，長期的な追跡を行うために追跡アルゴリズムによって補正を行う．MPM と PTGT の結果を比較したところ，PTGT のほうが部分的な追跡が少なく，正しい追跡結果が多いことが確認できる．

しかし，この追跡補正によってミスが発生する場合がある．**図 7** に，PTGT が失敗した事例を示す．画像の中央にあるピンク色の追跡は $t = 2$ の時点で物体が消滅しているが，緑色の追跡物体に補正されている．そのため本来消失するべき物体が補正されてしまい，追跡が長くなることがある．これは，τ フレーム内に移動できる最大移動距離 r_τ を調節することで抑制できる．

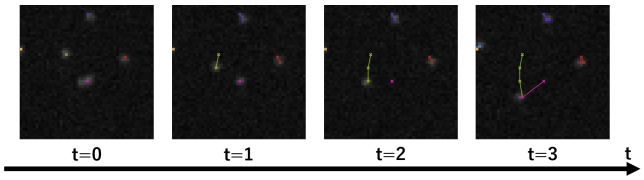

図 7 PTGT の追跡失敗例

5. おわりに

本稿では SPT のための Random Feature Assignment Module と Graph Transformer と追跡アルゴリズムを提案した．PTGT は以下の 3 つの問題点を解決した．

- 物体の検出ミスや誤検出により追跡を失敗すること．
- 分子の特徴がなく，追跡が難しいこと．
- 分子の密度が高い場合，ID Switch が頻繁に発生すること．

PTGT は追跡アルゴリズムにより長期的な追跡を可能とし，CD47 データセットにおいて最高の精度を達成した．

参考文献

[1] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D.: Object detection with discriminatively trained part-based models, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 32, No. 9, pp. 1627–1645 (2009).

[2] Ren, S., He, K., Girshick, R. and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems*, Vol. 28 (2015).

[3] Yang, F., Choi, W. and Lin, Y.: Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2129–2137 (2016).

[4] Bai, Y., Zhang, Y., Ding, M. and Ghanem, B.: Sodmtgan: Small object detection via multi-task generative adversarial network, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 206–221 (2018).

[5] Nishimura, K., Ker, D. F. E. and Bise, R.: Weakly supervised cell instance segmentation by propagating from detection response, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 649–657 (2019).

[6] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention*, Springer, pp. 234–241 (2015).

[7] Wen, L., Li, W., Yan, J., Lei, Z., Yi, D. and Li, S. Z.: Multiple target tracking based on undirected hierarchical relation hypergraph, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1282–1289 (2014).

[8] Dehghan, A., Tian, Y., Torr, P. H. and Shah, M.: Target identity-aware network flow for online multiple target tracking, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1146–1154 (2015).

[9] Li, J., Gao, X. and Jiang, T.: Graph networks for multiple object tracking, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 719–728 (2020).

[10] Tang, S., Andriluka, M., Andres, B. and Schiele, B.: Multiple people tracking by lifted multicut and person re-identification, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3539–3548 (2017).

[11] Dai, P., Weng, R., Choi, W., Zhang, C., He, Z. and Ding, W.: Learning a proposal classifier for multiple object tracking, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2452 (2021).

[12] Wojke, N., Bewley, A. and Paulus, D.: Simple online and realtime tracking with a deep association metric, *2017 IEEE international conference on image processing (ICIP)*, IEEE, pp. 3645–3649 (2017).

[13] Choi, W. and Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera, *European Conference on Computer Vision*, Springer, pp. 553–567 (2010).

[14] Zhou, X., Koltun, V. and Krähenbühl, P.: Tracking objects as points, *European Conference on Computer Vision*, Springer, pp. 474–490 (2020).

[15] Feichtenhofer, C., Pinz, A. and Zisserman, A.: Detect to track and track to detect, *Proceedings of the IEEE international conference on computer vision*, pp. 3038–3046 (2017).

[16] Bergmann, P., Meinhardt, T. and Leal-Taixe, L.: Tracking without bells and whistles, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 941–951 (2019).

[17] Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F. and Fu, Y.: Chainedtracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking, *European conference on computer vision*, Springer, pp. 145–161 (2020).

[18] Zhou, X., Koltun, V. and Krähenbühl, P.: Tracking objects as points, *European Conference on Computer Vision*, Springer, pp. 474–490 (2020).

[19] Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C. and Luo, P.: Transtrack: Multiple object tracking with transformer, *arXiv preprint arXiv:2012.15460* (2020).

[20] Zeng, F., Dong, B., Wang, T., Zhang, X. and Wei, Y.: Motr: End-to-end multiple-object tracking with transformer, *arXiv preprint arXiv:2105.03247* (2021).

[21] Brasó, G. and Leal-Taixé, L.: Learning a neural solver for multiple object tracking, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6247–6257 (2020).

[22] Meinhardt, T., Kirillov, A., Leal-Taixe, L. and Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers, *arXiv preprint arXiv:2101.02702* (2021).

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, Vol. 30 (2017).

[24] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. and Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation, *International conference on medical image computing and computer-assisted intervention*, Springer, pp. 424–432 (2016).

[25] Sarlin, P.-E., DeTone, D., Malisiewicz, T. and Rabinovich, A.: Superglue: Learning feature matching with graph neural networks, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947 (2020).

[26] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S.: End-to-end object detection

with transformers, *European conference on computer vision*, Springer, pp. 213–229 (2020).

[27] Chenouard, N., Smal, I., De Chaumont, F., Maška, M., Sbalzarini, I. F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M. et al.: Objective comparison of particle tracking methods, *Nature methods*, Vol. 11, No. 3, pp. 281–289 (2014).

[28] Hayashida, J., Nishimura, K. and Bise, R.: MPM: Joint representation of motion and position map for cell tracking, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3823–3832 (2020).

[29] Yan, B., Peng, H., Fu, J., Wang, D. and Lu, H.: Learning spatio-temporal transformer for visual tracking, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10448–10457 (2021).

[30] Yu, B., Tang, M., Zheng, L., Zhu, G., Wang, J., Feng, H., Feng, X. and Lu, H.: High-performance discriminative tracking with transformers, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9856–9865 (2021).

[31] Cao, Z., Fu, C., Ye, J., Li, B. and Li, Y.: HiFT: Hierarchical Feature Transformer for Aerial Tracking, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15457–15466 (2021).

[32] Mayer, C., Danelljan, M., Paudel, D. P. and Van Gool, L.: Learning target candidate association to keep track of what not to track, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13444–13454 (2021).

[33] DeTone, D., Malisiewicz, T. and Rabinovich, A.: Superpoint: Self-supervised interest point detection and description, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236 (2018).