Comparative study of transformer robustness for multiple particle tracking without clutter

Piyush Mishra, Philippe Roudot

▶ **To cite this version:**

Piyush Mishra, Philippe Roudot. Comparative study of transformer robustness for multiple particle tracking without clutter. EUSIPCO, Aug 2024, Lyon, France. ⟨hal-04619330⟩

HAL Id: hal-04619330

https://hal.science/hal-04619330v1

Submitted on 2 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Comparative study of transformer robustness for multiple particle tracking without clutter

Piyush Mishra, Philippe Roudot
Aix Marseille Univ, CNRS, Centrale Med, I2M, Institut Fresnel, Turing Centre for Living Systems
Marseille, France

***Abstract*—The tracking of multiple particles in lengthy image sequences is challenged by the stochastic nature of displacements, particles detection errors, and the combinatorial explosion of all possible trajectories. As such, extensive work has focused on the modeling of noisy trajectories to try and predict the most likely trajectory-to-measurements associations. Recently, transformers have been shown to significantly accelerate the evaluation of probabilistic models for the system dynamics and detection clutter generated from false positives. However, little work has focused on clutter-free scenarios with multiple particles moving erratically, where the challenge resides not in the model complexity, but in the combinatorial burden of considering all possible trajectory-to-measurements associations. This is a common occurrence in fluorescence microscopy at low framerate. This paper offers a proof-of-concept study of the benefit of the transformer architecture in such scenarios through the simulation of two-particle-systems undergoing Brownian motion. Specifically, we designed a transformer for this estimation task and compared it with the Multiple Hypothesis Tracker (MHT), the optimal estimator when all trajectory-to-measurement associations can be computed. We first show increased robustness of the transformer against erratic displacements over long sequences, with significantly lower computational complexity than MHT. Then, we show that while the transformer requires very little training to significantly outperform MHT on long sequences, it cannot match the theoretically optimal performances of MHT on short sequences even with extensive training. Hence, our work motivates the broader application of transformers in high-SNR sequences and opens the way to the development of frugal methods thanks to the combination of both statistical and neural network frameworks for particle tracking.**

***Index Terms*—Multiple Particle Tracking, Transformers, Multiple Hypothesis Tracking.**

I. Introduction

The tracking of multiple objects detected in image sequences is a fundamental problem that can be encountered across many imaging modalities. Here, we focus on the challenging case of dense populations of particles undergoing random walks and imaged with limited framerate. Fluorescence microscopy [1] imaging is a good example of this scenario. Indeed, these data exhibit many diffusing molecules imaged as diffraction-limited spots with identical appearance, while phototoxicity limits temporal sampling. Under reasonable assumptions on object displacements and detector noise, an optimal estimator for the set of trajectories involves the temporally iterative estimation of hidden parameters (true position, speed, etc.) from all possible combinations of trajectories-to-measurements (or trajectories-to-detections) associations. Since the number of possible trajectories increases exponentially with the number of objects and frames [2], associations are typically gated through their likelihood, leading to suboptimal solutions that impact the interpretation of physical quantities like the duration of biological processes [3]. As such, improving the robustness of tracking approaches has remained an active field.

Recent works have focused on transformer-based neural networks for this task, especially in the context of noisy images with frequent detection errors [4, 5, 6, 7]. These works have exhibited high computational efficiency and sometimes even higher accuracy than conventional approaches based on stochastic filtering [4]. In this study, we seek to test if this computational efficiency could be beneficial when using simpler modeling. Here, performance limits stem not from complex modeling but from the abundance of trajectories-to-measurements association hypotheses due to erratic displacements, as evidenced in high SNR scenarios where the motion is considered as perfectly known with linear transition. In turn, our goal is to test the robustness of transformers against increasingly noisy dynamics or sequence length.

Section II formulates the problem background and briefly reviews pertinent literature. Section III then establishes the different methods used to simulate motions, the estimators used for the system parameters and evaluation metrics. Section IV compares the robustness and efficiency of transformers and conventional filtering methods across varying sequence lengths and Brownian motion scales. Finally, Section V synthesizes this work and briefly outlines avenues for future work, especially in the field of bioimaging and frugal machine learning.

II. Background

A. Problem Formulation and Notations

Let us consider the hidden state $\mathbf{x}_t \in \mathbb{R}^D$ of a particle at time t, and a set of N particles $\mathbf{X} = \{\mathbf{X}^p\}_{p=0:N-1}$ where $\mathbf{X}^p = \{\mathbf{x}^p_{t_0}, ..., \mathbf{x}^p_{t_0+l}\}$ is a sequence of positions describing an unknown dynamic process, $t_0 \in \mathbb{N}$ denoting its time of birth and $l \in \mathbb{N}$ denoting its lifetime. We also denote $\mathbf{Z} = \{\mathbf{Z}_t\}_{t=0:T-1}$ as the union of unlabelled measurements that come from the particles of interest in $\mathbf{X}$ after imaging and object detection process. Without loss of generality, the

The project leading to this publication has received funding from France 2030, the French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University - A*MIDEX. Correspondences should be addressed to Philippe Roudot (philippe.roudot@univ-amu.fr).

following Bayesian filtering equation provides an iterative framework to solve this problem:

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) = p(\mathbf{Z}_t|\mathbf{X}_t) \int p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) d\mathbf{X} \tag{1}$$

with the assumptions that $\mathbf{X}_t$ is Markovian with known transition probability $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ and *a priori* $p(\mathbf{X}_0)$, and that $\mathbf{Z}_t$ only depends on $\mathbf{X}_t$. The likelihood function $p(\mathbf{Z}_t|\mathbf{X}_t)$ quantifies the probability of observing the measurements $\mathbf{Z}_t$ given the current state $\mathbf{X}_t$. The trajectory set $\mathbf{X}$ is updated recursively over time as new measurements in $\mathbf{Z}$ arrive, and the posterior distribution $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$ is computed. Of note, estimating $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$ requires the evaluation of all possible combination of association between $\mathbf{Z}$ and each trajectory in $\mathbf{X}$, leading to exponentially increasing costs.

B. Related Work

Within the last decades, efforts have focused either on the likelihood of trajectory candidates taken independently (using stochastic filtering or supervised machine learning techniques) or discrete optimization schemes for the selection of the best possible set of trajectories [8] with many efforts dedicated to bioimaging [9]. More recently, advances in deep learning have demonstrated great potential, especially in scenarios with low SNR. Indeed, RNNs have been used to learn complex biodynamic models using simulated data [10, 11]. The role of RNNs is to estimate the parameters of individual tracks while the selection of the best possible set of trajectories at each time step is carried out by a dedicated optimization algorithm. These approaches perform particularly well under low SNR conditions where noise-induced clutter and transient mis-detections challenge the typical assumptions of linearity and Gaussian noise made by conventional models designed for scalability. They have, however, shown little to no improvement under high SNR conditions, where the engineering of multiple motion modeling performs best [12]. Consequently, RNNs excel in estimating parameters associated with particle imaging amidst clutter but offer limited utility in high SNR scenarios, where the emphasis is on evaluating numerous potential assignments rather than individual trajectory parameter estimation. More recently, transformers have also been applied to the same task in moderate noise level [5], focusing particularly in complex molecular dynamics, such as the simulation of microtubule polymerization. These works also exclude the combinatorial aspect from the network design.

In the broader field of sensor tracking, transformers have been adapted to assess the likelihood of combined, instead of individual, trajectory parameters and their different combinations, thus eliminating the necessity for the global optimization step [4, 6]. The scenario under study mimicked submarine target tracking, where clutter level is typically higher than in bioimaging with fewer targets, accentuating the importance of measurement selection. In [4], the network was compared against stochastic filtering approaches that estimate the parameters associated with both targets and clutter. Transformers exhibited comparable accuracy while boasting a remarkable speedup of 10^5 times. Leveraging the attention layer, the spatial context is used to select the best set of measurements to be associated with a given trajectory. This work highlights the potential of transformers to improve tracking performances beyond the estimation of individual trajectory parameters, it also raises questions on the performance of transformers in high-SNR tracking challenges.

III. Methods

A. Simulated data generation

We simulate two particles undergoing Brownian motion as

$$\mathbf{y}_t^p = \mathbf{y}_{t-1}^p + \epsilon_t^p + \delta^p, \tag{2}$$

for $\mathbf{y}_t^p \in \mathbb{R}^2 \ \forall p \in \{0 : N-1\}$, where $\epsilon_t^p \sim \mathcal{N}(0, \mu^2 \mathbf{I}_2)$ represents the random component (we call μ the scale of the process noise) and $\delta^p \in \mathbb{R}^2$ is the drift. To simulate an increase in the measurement noise, we incorporate an additive term $\omega^p \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ to the above equation, where we call σ the scale of this noise.

$$\mathbf{z}_t^p = \mathbf{y}_t^p + \omega^p \tag{3}$$

This information $\mathbf{z}_t^p$ is referred to as the measurement of the p^{th} particle at time t. The resulting sequences are stored in two matrices $\mathbf{X}$ and $\mathbf{Z}$ for the ground truth and the measurements respectively, of shape (N, T) for each Cartesian dimension, where $N = 2$ is the number of particles in the system and T is the lifetime of the particles. We, thus, assume that all particles have equal lifetime ($l = T$). All elements within $\mathbf{Z}$ are shuffled independently for each frame to prevent the transformer from using the measurement order.

B. Multiple Hypothesis Tracking

Here, we briefly introduce MHT starting from the presentation of the Kalman filter. Let's consider a particle to be at state $\mathbf{x}_t$ at a time instant t such that $\mathbf{x}_t = \begin{pmatrix} x_t, & y_t, & dx_t, & dy_t \end{pmatrix}^T$, the Kalman filter uses a state transition equation that governs the changes in the particle hidden state, and an observation equation that maps true states to observed measurements, as follows:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t + \eta \\ \mathbf{z}_{t+1} = \mathbf{H}\mathbf{x}_{t+1} + \nu \end{cases}, \tag{4}$$

where $\mathbf{F} \in \mathbb{R}^{4\times4}$ is the transition matrix, $\mathbf{H} \in \mathbb{R}^{2\times4}$ is the observation matrix and where $\eta \sim \mathcal{N}(0, \eta'^2 \mathbf{I}_2)$ and $\nu \sim \mathcal{N}(0, \nu'^2 \mathbf{I}_2)$ are the random components representing the process and measurement noises used for inference. The Kalman filter is an iterative approach to predict and update the particle state every time a new measurement is added to the trajectory. It is considered to be an optimal estimator of the particle state in the sense that it minimizes the mean square error if the conditions of equation 4 are respected [8].

In a multiple particle tracking context, a single measurement must be selected among many at time $t+1$ to update the state estimated at time t. To do so, MHT uses one Kalman filter per particle, and considers all possible combinations of measurements observed within a specified time window, known as the lookback window, which we denote as K_{MHT}.

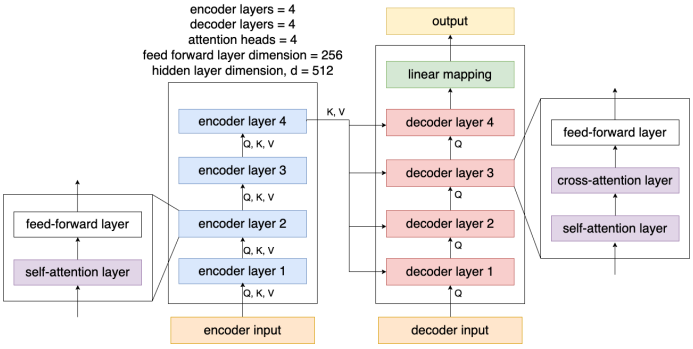

Fig. 1. Encoder-decoder-based transformer used for this study. Each encoder layer contains a self-attention and feed-forward layer; each decoder layer contains a self-attention, cross-attention and feed-forward layer.

Each hypothesis is assessed using the Mahalanobis distance that accounts for the probability of predicted state. The most likely hypothesis, i.e. the hypothesis with the least combined Mahalanobis distance across all targets, is then selected and propagated to the next time step. In the case where the sequence length is inferior to the lookback window, MHT is the optimal estimator of the set of states.

C. Attention and Transformer

A transformer is composed of an encoder, learning the latent space of the sequence, and a decoder, predicting the next state. The encoder input is the set of elements of $\mathbf{Z}$ truncated to K_E elements in temporal order (with the temporal information being passed into the encoder along with each detection coordinate) and the decoder output is the set of corresponding elements in $\mathbf{X}$. The decoder input is the set of elements from the previous K_D decoder output elements. Akin to MHT, we refer to the lookback window size of the transformer encoder as K_E and that of the transformer decoder as K_D. A key difference in transformers compared to other neural networks is the attention layer. Attention is used to weigh the importance of different input elements when computing the hidden representations for each layer. Consider an input $\mathbf{A}$. The three input matrices $\mathbf{Q}$ (query), $\mathbf{K}$ (key) and $\mathbf{V}$ (value) are defined as follows:

$$\mathbf{Q} = \mathbf{W}_Q\mathbf{A}, \mathbf{K} = \mathbf{W}_K\mathbf{A}, \mathbf{V} = \mathbf{W}_V\mathbf{A} \tag{5}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are randomized weight matrices that are updated as the model learns.

$$attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{6}$$

The attention mechanism is used inside an encoder-decoder-based transformer architecture [13] as shown in Fig. 1. Note that $\mathbf{A}$ for calculating the 3 input matrices can vary: for self-attention, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are sampled from the encoder input whereas for cross-attention $\mathbf{Q}$ is sampled from the decoder input, and $\mathbf{K}$ and $\mathbf{V}$ are sampled from the encoder output (Fig. 1).

D. Performance evaluation

Matching of estimated and real particle positions is achieved by minimizing the sum of Euclidean distances for both the particles at each frame. A true positive (TP) indicates a correctly predicted link between two positions. A false positive (FP) represents such a link that is erroneously predicted, where the model suggests a transition that does not exist. A false negative (FN) is the case where said model fails to predict such a link that actually exists. We evaluate performance in tracking through the Jaccard coefficient (JC) given by $JC = \frac{TP}{TP+FP+FN}$.

E. Time Complexity

The expected time complexity of MHT in terms of K_{MHT} is $O(N!^{K_{MHT}})$ [2], where N is the number of particles being tracked. Considering $K_E = K_D = K$, the time complexity of a transformer is expected to follow $O(Kd^2 + K^2d)$; $O(K^2d)$ comes from its reliance on self-attention and $O(Kd^2)$ comes from the usage of parallel multi-heads [13], where d is the hidden layer dimension. In the results section, we measure time complexity by counting FLOPs (Floating-Point Operation) using [14] and [15] for the transformer and the MHT implementations respectively.

IV. Results

We now present a quantitative comparison of our transformer architecture and the MHT algorithm using the simulations described above. In Section IV-A we show that transformers exhibit heightened robustness to increasing noise levels when applied over long sequence, maintaining superior performances when both methods are mis-parameterized. We also demonstrate in Section IV-B that MHT outperforms transformers over short sequences where MHT is optimal. Albeit transformer performance improves with an increased number of experiments, it never matches that of MHT. Furthermore, we present that transformers demonstrate superior computational efficiency compared to MHT for extended lookback windows in Section IV-C. Finally, we examine in Section IV-D the yield of transformer training across increasing sequence lengths, indicating that while extensive training is required to approach MHT performances in the regime of optimality (short sequence), little training is required for transformer to show excellent performance over long sequences.

A. Robustness to increasing noise in long sequences

We first sought to test the standard scenario where MHT performances are limited by the number of trajectory-to-measurement combinations to consider. To do so, we consider the task of inferring the state of a 2-particle-system with motion as described in Section III-A with $T = 150$ using an MHT estimator equipped with a history length set to $K_{MHT} = 1$. The transformer applied for that same estimation task is equipped with a decoder of the same lookback window, $K_D = 1$, and a encoder of length $K_E = T$. We also set $K_D = K_{MHT} = 8$ to test if increasing the length of measurement history affects the difference in performance.

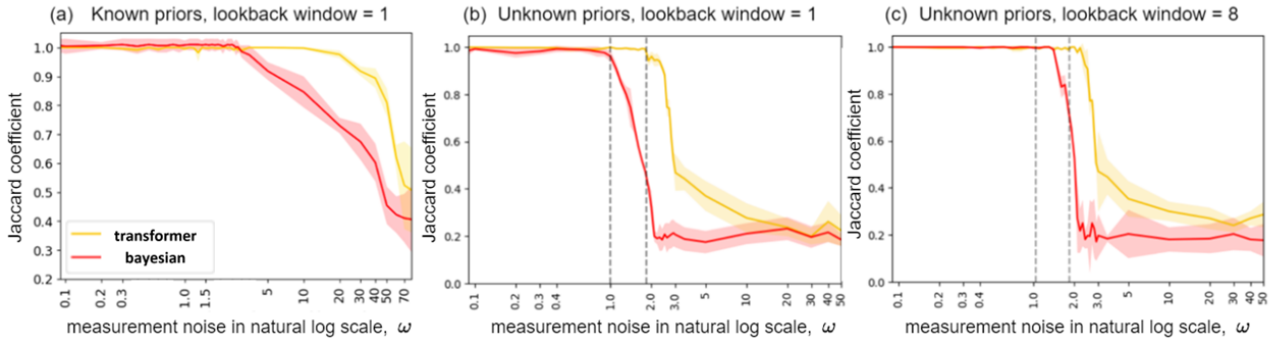

Fig. 2. Tracking performances as measured through the Jaccard coefficient applied on frame-to-frame link against varying simulated measurement noise σ and process noise $\eta' = 1.5$. Bayesian approach (MHT) and transformer are respectively parameterized and trained using (a) known measurement noise ν', with $K_D = K_{MHT} = 1, K_E = T = 150$ (b) a fixed prior at $\nu' = 1$ with $K_D = K_{MHT} = 1, K_E = T = 150$ (c) Same as (b) but $K_D = K_{MHT} = 8$, the dashed lines show breakpoints as seen in (b)

Fig. 2(a) illustrates the results when both MHT and transformers are parameterized optimally with exact *a priori* information for the MHT parameters and for the training data used with the transformer. The transformer has been trained using 6 experiments (1 experiment is a single 2-particle-simulation, 1 experiment per batch, and 6 batches). In this experiment, the transformer shows a significantly higher robustness when compared against the MHT approach. We then sought to test if this difference in robustness was still present in case of misparameterization, by fixing the prior on measurement noise for MHT to $\nu' = 1$ as well as using the same training dataset for the transformer, with $\sigma = 1$ for all experiments. Thus, in Fig. 2(b), we can see that performance of both the methods drop, and that the robustness of the transformer remains higher.
Together, this data suggests that transformers provide high robustness when provided with the same a priori information than classic Bayesian filtering approaches, even when this a priori does not exactly match the data. In Fig. 2(c), the setup is the same except that $K_{MHT} = K_D = 8$. We observe higher robustness of both approaches while the difference in their respective performances remains significant.

B. Robustness to increasing noise in short sequences

We then sought to benchmark the transformer against a scenario where MHT can provide the optimal solution. To do so, we simulate a 2-particle system with shorter lifetime ($T = 8$), filtered with an MHT equipped with a lookback window of the same length, $K_{MHT} = 8$. Thus, this MHT can compute all possible trajectory hypotheses. Similarly, we designed a transformer that mirrors these characteristic with $K_E = 8$ and $K_D = 8$. In the remainder of the paper, we denote transformer$_{T,i}$ a transformer trained in this context, i.e., with T time steps and i number of experiments per batch of training. Note that inferences may be carried on sequences of size $\leq T$ for transformer$_{T,i}$.
In Fig. 3, we see that the transformer$_{8,1}$ performs significantly worse than MHT. As we increase the number of experiments, however, we see transformer performance approaches the optimal estimator of MHT, but does not reach it. Fig. 4 also

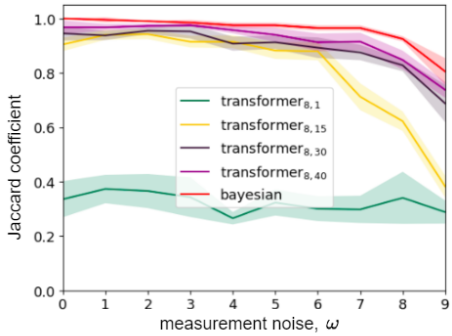

Fig. 3. Tracking performances against varying simulated measurement noise σ and process noise $\eta' = 1.5$ and increasing number of simulation used for training in conditions where the Bayesian technique is optimal. "transformer$_{T,i}$" refers to a transformer trained on a sequence of T time steps and i experiments per batch. Other parameters: $\sigma = 1.2$ (known prior); $K_D = K_{MHT} = K_E = T = 8$.

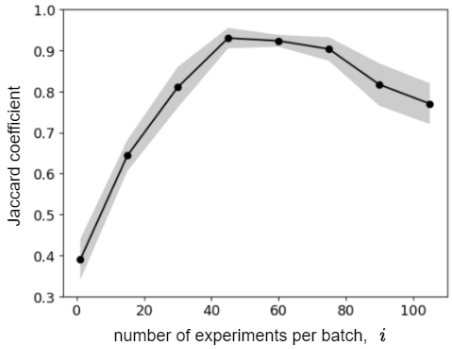

Fig. 4. Tracking performances against increasing number of simulations used for training for a sequence of fixed size 8; $\sigma = 7.0$

shows that more experiments do not help the transformer in approaching the optimal estimator further: as i increases, the growth in Jaccard coefficient becomes that of decreasing returns. This behavior of LLM has been previously reported in [16]. Interestingly, Fig. 3 also shows that the increase in Jaccard coefficient as we increase i is also non-uniform.

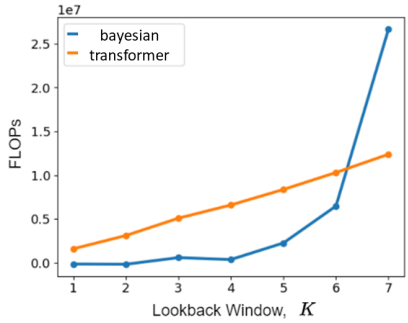

Fig. 5. Computational cost for both MHT and transformer as a function of increasing lookback windows; $\sigma = 7.0$.

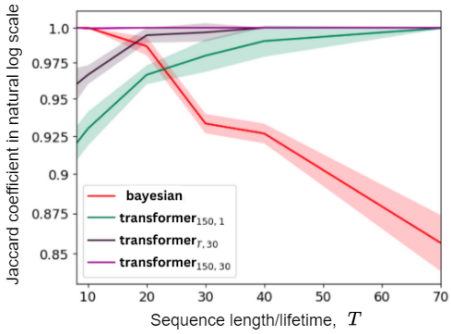

Fig. 6. Tracking performances against increasing sequence length, comparing different number and length of simulations used for training; $\sigma = 7.0$

C. Measurement of computational times

We measured FLOPs for both the transformer and MHT in sequence of length 150. As shown in Fig. 5, the computational cost of MHT is directly linked to the exponential increase in the number of trajectory-to-measurement combinations associated to longer lookback windows. Conversely, while the transformer model initially exhibits higher computational load than MHT for shorter lookback windows, its computational growth approximates a linear trajectory, thereby rendering it more computationally efficient than MHT for longer lookback windows. This is in line with the projections in Section III-E. Indeed in our use case $K_E < d$, making the transformer closer to a time complexity of $O(Kd^2)$.

D. Robustness to increasing sequence length

Additionally, we aimed to determine the effect of the training protocol on experiments with varying particle lifetimes ($8 \leq T \leq 150$). In Fig. 6, we compare the performances of transformer$_{T,30}$, transformer$_{150,30}$, and transformer$_{150,1}$ with that of MHT ($K_{MHT} = K_D = 8, K_E = T$). Transformer$_{150,30}$ achieves the best performances overall, while the performance of MHT diminishes as T increases. Although transformer$_{T,30}$ and transformer$_{150,1}$ initially exhibit lower performance at lower T values, they demonstrate improvement as T increases. This shows that, in our scenario, transformer trained with a small training dataset can exhibit excellent performances on long sequences while the MHT algorithm is better suited for short sequence in comparison.

V. Conclusion and Future Work

Together, our results show that transformers provide a significant improvement in robustness against the combinatorial complexity in tracking multiple particle, rather than the challenges associated with the modeling of the system dynamics or to the clutter alone. While previous studies have highlighted transformers' advantages in the latter, our findings underscore the improved robustness and computational efficiency of transformers in prolonged sequences of clutter-free scenes presenting multiple particle undergoing Brownian motion. In future works, we will focus our research on several developments based on these promising results. We will be working on the scalability of our design, and test if these properties hold when the number of particles is representative of bioimage data, the main application of our methodological research. We will also explore strategies for reducing training cost, or "frugal machine learning", by combining the robustness of transformers trained with small datasets for longer sequences with the properties of optimality of MHT approaches in shorter and isolated trajectories.

References

[1] Cédric Vonesch et al. "The colored revolution of bioimaging". In: *IEEE signal processing magazine* 23.3 (2006), pp. 20–31.

[2] Donald Reid. "An algorithm for tracking multiple targets". In: *IEEE transactions on Automatic Control* 24.6 (1979), pp. 843–854.

[3] Marcel Mettlen and Gaudenz Danuser. "Imaging and modeling the dynamics of clathrin-mediated endocytosis". In: *Cold Spring Harbor perspectives in biology* 6.12 (2014), a017038.

[4] Juliano Pinto et al. "Can deep learning be applied to model-based multi-object tracking?" In: *arXiv preprint arXiv:2202.07909* (2022).

[5] Yudong Zhang and Ge Yang. "A Motion Transformer for Single Particle Tracking in Fluorescence Microscopy Images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 503–513.

[6] Juliano Pinto et al. "Transformer-Based Multi-Object Smoothing with Decoupled Data Association and Smoothing". In: *arXiv preprint arXiv:2312.17261* (2023).

[7] Kaijie He et al. "Target-Aware Tracking with Spatial-Temporal Context Attention". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[8] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[9] Nicolas Chenouard et al. "Objective comparison of particle tracking methods". In: *Nature methods* 11.3 (2014), pp. 281–289.

[10] Roman Spilger et al. "A recurrent neural network for particle tracking in microscopy images using future information, track hypotheses, and multiple detections". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3681–3694.

[11] Roman Spilger et al. "Deep probabilistic tracking of particles in fluorescence microscopy images". In: *Medical image analysis* 72 (2021), p. 102128.

[12] Philippe Roudot et al. "u-track3D: Measuring, navigating, and validating dense particle trajectories in three dimensions". In: *Cell Reports Methods* 3.12 (2023).

[13] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[14] *Flop Counter for PyTorch Model*. https : / / github . com / facebookresearch/fvcore/blob/main/docs/flop_count.md.

[15] *PyPAPI*. https://github.com/flozz/pypapi.

[16] Jordan Hoffmann et al. "An empirical analysis of compute-optimal large language model training". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30016–30030.