

INTEGRATING VISUAL FEATURES IN MULTIPLE HYPOTHESIS TRACKING THROUGH SELF-SUPERVISED LEARNING

Raphael Reme^{*†} Alasdair Newson[†] Elsa Angelini[†]
Jean-Christophe Olivo-Marin^{*} Thibault Lagache^{*}

^{*} Institut Pasteur, Université de Paris-Cité, CNRS UMR 3691, BioImage Analysis Unit F-75015 Paris, France

[†] LTCI, Telecom Paris, Institut Polytechnique de Paris, France

Corresponding author: thibault.lagache@pasteur.fr

ABSTRACT

Multiple Object Tracking (MOT) is a crucial step in the automatic analysis of biological processes. In this work, we introduce VISUALMHT, a Bayesian Multiple Hypothesis Tracking algorithm integrating self-supervised visual features to leverage robust target identity and consistently track biological objects. This novel tracking algorithm achieves state-of-the-art results on the SINETRA synthetic datasets and demonstrates robust neuron tracking in the freshwater cnidarian *Hydra vulgaris*, highlighting the benefit of combining probabilistic motion models with learned visual representations.

Index Terms— Multiple Hypothesis Tracking, Contrastive learning, Hydra Vulgaris

1. INTRODUCTION

The gold standard for studying the *in vivo* dynamics of biological processes is time-lapse fluorescence imaging, followed by quantitative analysis. A crucial step in this pipeline is *Multiple Object Tracking* (MOT), which aims to reconstruct the trajectories of individual targets (*e.g.*, fluorophores, cells, pathogens) over time [1, 2].

Most tracking methods adopt a *tracking-by-detection* paradigm: objects are first detected in each frame, and detections are subsequently linked to form trajectories. In fluorescence microscopy, labeled objects typically appear as bright spots on a noisy background [3], and their detection can be performed using analytical techniques, such as wavelet-based filtering and statistical thresholding [4], or by data-driven approaches based on deep learning [5, 6].

Several strategies have been developed to robustly link noisy detections into coherent trajectories. Graph-based optimization methods formulate tracking as a global association problem over time, allowing for consistent trajectory reconstruction [7, 8, 9]. Alternatively, *Multiple Hypothesis Tracking* (MHT) [10, 2, 11] adopts a fully Bayesian formulation within a sliding temporal window, explicitly maintaining and updating multiple competing association hypotheses to identify the most probable tracking configuration. When the MHT is restricted to frame-by-frame optimization, the problem reduces to a *Linear Assignment Problem* (LAP) [12, 13] enabling real-time performance at the cost of reduced long-term tracking accuracy.

The quality of any linking algorithm critically depends on the cost function used to compare detections. Simple geometric costs (Euclidean, Intersection-Over-Union), often fail when targets undergo large displacements. Therefore, more robust probabilistic models (*e.g.* Kalman filtering, Interactive Multiple Model) have been used to anticipate such motion [14, 13]. Advanced Bayesian tracking frameworks extend these costs with modeled priors such as a target life-cycle [10, 2].

Recently deep learning-based approaches have emerged as powerful alternatives, learning association costs directly from annotated data [15, 8]. In natural image tracking, neural networks are used to learn target identity through appearance embeddings [16, 17], and we recently showed that such identity embeddings could also be achieved for biological targets [18].

In this work, we introduce VISUALMHT, a fully Bayesian Multiple Hypothesis Tracking framework that integrates learned visual identity features alongside probabilistic motion modeling. We design robust target embeddings using self-supervised contrastive learning with carefully chosen data augmentations, enabling to learn consistent target identity without any manual annotation. Our approach achieves state-of-the-art performance on the SINETRA synthetic tracking benchmark [19] and demonstrates high robustness for neuron tracking in the freshwater cnidarian *Hydra vulgaris* [20], confirming that the incorporation of visual features in our tracking framework significantly improves target re-identification across time.

2. METHOD

In the following, we denote by $I_{1:T} = (I_1, \dots, I_T)$ a time-lapse image sequence of length T , where each frame $I_t \in [0, 1]^{H \times W}$ represents a two-dimensional fluorescence image of size $H \times W$ pixels at time t . For each frame I_t , a detection algorithm yields a set of M_t detections, denoted $Z_t = \{\mathbf{z}_t^j\}_{j=1}^{M_t}$, where $\mathbf{z}_t^j \in \mathbb{R}^2$ corresponds to the pixel-wise localization of the j^{th} detection. Our tracking framework is agnostic to the choice of detector; in this work, we employ either a trained StarDist model [5] or the unsupervised wavelet-based spot detector [4] implemented in the Icy software [21].

We first describe how each detection $\mathbf{z}_t^j$ is encoded into an identity feature vector using a self-supervised learning approach. Then, we introduce VISUALMHT, a Bayesian Multiple Hypothesis Tracking algorithm that leverages these features to robustly associate detections across time.

2.1. Self-supervised learning of a target identity

We aim to encode the visual identity of each detected target into a discriminative feature vector implicitly learned without any track label. For each detection $\mathbf{z}_t^j$ in frame I_t , we extract a square image patch $P_t^j \subset I_t$ of size $L \times L$ centered on $\mathbf{z}_t^j$, that captures the pixel information about the target identity. Figure 1 shows typical patch evolution through time, indicating a preserved visual identity for each target.

We propose to learn this visual identity conservation through a convolutional neural network (CNN) encoder φ that projects the raw patch into a feature vector $\mathbf{f}_t^j = \varphi(P_t^j) \in \mathbb{R}^d$. This encoder

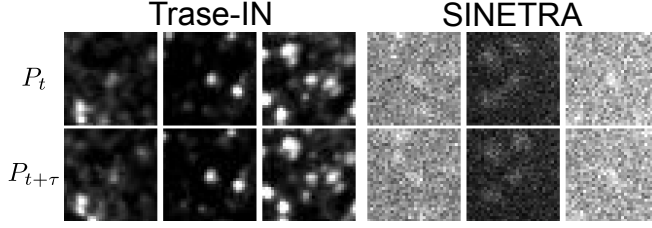


Fig. 1. (Top row) Examples of target patches extracted from Hydra neurons in Trase-IN [20] and synthetic targets in SINETRA [19]. **(Bottom row)** The same target is shown after $\tau = 8$ frames, highlighting the slow temporal evolution of its visual appearance.

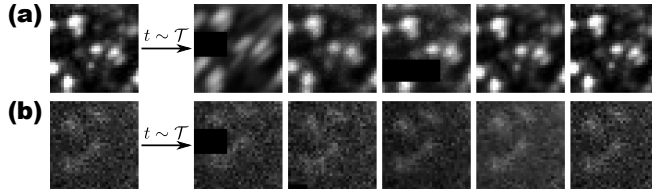


Fig. 2. Patch-based augmentation strategy for learning robust and invariant target identity embeddings. (a) Five random augmentations $t \sim \mathcal{T}$ applied to the same patch from Trase-IN [20]. **(b)** Similar augmentations applied to a patch from SINETRA [19].

is trained to produce identity feature vectors robust to carefully designed random image augmentations $t \sim \mathcal{T}$ (Figure 2) that randomly modify patch intensities, while preserving the expected target identity. The model optimizes a triplet loss [22] to encourage feature proximity for identical targets and separation for distinct ones.

For each imaging and target modality, we extract training patches from three representative videos and train a dedicated encoder specific to that modality. Since our approach is fully self-supervised and requires no manual annotations, an encoder can be trained for any imaging condition or biological target of interest. For each detection $\mathbf{z}_t^j$ included in the training set, we then adopt the following unsupervised triplet sampling strategy. **(1)** The anchor and positive patches are generated as two independently augmented views of the same detection: $P_a = t_a(P_t^j)$ and $P_p = t_p(P_t^j)$, where $t_a, t_p \sim \mathcal{T}$ denote random augmentations drawn from our augmentation distribution described below. **(2)** The negative patch is sampled from another detection $\mathbf{z}_{t'}^{j'}$ within the same frame or from a different video to ensure distinct target identities: $P_n = t_n(P_{t'}^{j'})$ with $t_n \sim \mathcal{T}$.

The encoder learns to minimize the triplet loss of the embeddings $\mathbf{f}_a$, $\mathbf{f}_p$, and $\mathbf{f}_n$ of these patches:

$$L(\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n) = \max(\|\mathbf{f}_a - \mathbf{f}_p\|_2 - \|\mathbf{f}_a - \mathbf{f}_n\|_2 + m, 0) \quad (1)$$

where $m = 1$ is a fixed margin controlling typical inter-class separation. The network thus learns to minimize intra-target distances while maximizing inter-target distances in the feature space.

2.1.1. Data augmentations

To learn robust identity vectors, each patch is transformed by a randomly sampled augmentation operator $t \sim \mathcal{T}$ that is designed to alter the patch appearance while preserving the target identity (Figure 2). The sampled transformations typically simulate realistic changes that may occur between consecutive frames, and include:

(i) small affine translations and rotations ($\pm 5^\circ$), (ii) small elastic deformations, (iii) motion blur (maximum displacement of 10 pixels), (iv) Gaussian blurring ($\sigma \in [0.5, 1.0]$) and Poisson noise (PSNR 5–20 dB), (v) contrast jittering (scaling in $[0.8, 1.25]$), and (vi) random sub-patch erasing to encourage spatial invariance. Each random transformation has a probability $p = 0.5$ of being applied for each sampled patch.

Our augmentation strategy $\mathcal{T}$ is designed to generalize across imaging modalities and biological contexts, but it can be adapted to specific applications by tuning the transformation parameters. When reliable tracking annotations are available, they can be seamlessly integrated as additional supervision signals—either in combination with or as a substitute for random augmentations—to further improve temporal consistency and identity preservation.

2.1.2. Implementation details

Across our modalities, we set $L = 32$, which ensures that the patch captures both the target and a few of its closest neighbors, providing local spatial context that helps disambiguate visually similar targets. For each modality, we extract patches around all the detections of 30 frames distributed among 3 different videos, resulting in around 30k training patches.

We use a lightweight ResNet32 [23] architecture. After the last convolutional layer, adaptive average pooling is used, and a final linear layer projects the features onto a d -dimensional space. In practice, we found that $d = 50$ was sufficient to learn robust identity embeddings for our biological targets.

A batch size of 512 is used, and the learning rate is set at 0.001. We use AdamW [24] optimizer with a weight decay of 0.01 to reduce overfitting and momenta set to 0.9 and 0.999. The CNN is trained for 30,000 steps.

2.2. Multiple Hypothesis Tracking with Visual Features

2.2.1. Problem formulation

We design a MHT tracking algorithm that integrates our learned identity features, in order to preserve the target visual identity through the optimization process. Formally, on each image I_t , the detection algorithm produces M_t detections $Z_t = \{\mathbf{z}_t^j\}_{j=1}^{M_t}$, where $\mathbf{z}_t^j \in \mathbb{R}^2$ is the pixel-wise localization of the j^{th} detection. Additionally, each detection $\mathbf{z}_t^j$ is characterized by a feature vector $\mathbf{f}_t^j \in \mathbb{R}^d$, encoded by our trained CNN encoder φ . We denote by $F_t = \{\mathbf{f}_t^j\}_{j=1}^{M_t}$ and $F_{1:T} = \{F_t\}$.

The MHT produces tracks, *i.e.* assignments between detections. We denote a set of N tracks as $\Gamma_{1:T} = \{\gamma_{1:T}^i\}_{i=1}^N$, where the i^{th} track $\gamma_{1:T}^i = \{\gamma_t^i\}_{t=1}^T$ corresponds to a detection assignment $\gamma_t^i \in [0, M_t]$ for each frame t , where $\gamma_t^i = 0$ indicates that the track is not assigned to any detection in Z_t .

As in previous MHT approaches [10, 2], our method estimates trajectories by maximizing the *Maximum A Posteriori* (MAP) probability. In contrast to classical formulations that rely solely on positional observations $Z_{1:T}$, our MAP estimator additionally conditions on learned visual features $F_{1:T}$. This writes:

$$\Gamma_{1:T}^* \in \arg \max_{\Gamma_{1:T} \subset \Omega(T)} \mathbb{P}(\Gamma_{1:T} | Z_{1:T}, F_{1:T}) \quad (2)$$

where $\Omega(T)$ is the set of all feasible assignments. As $\Omega(T)$ grows exponentially with time, the exact optimization is intractable, and the MHT addresses this by solving associations recursively over short temporal windows of size $\tau \ll T$. Let $\Gamma_{1:t}$ denote the set of confirmed tracks up to frame t . For a look-ahead horizon τ , the set of

feasible track continuations is denoted $\Omega(t+\tau \mid t)$ and at each frame t , MHT solves:

$$\Gamma_{1:t+\tau}^* \in \arg \max_{\Gamma_{1:t+\tau} \subset \Omega(t+\tau \mid t)} \mathbb{P}(\Gamma_{1:t+\tau} \mid Z_{1:t+\tau}, F_{1:t+\tau}), \quad (3)$$

then advances the time index $t \leftarrow t+1$ and repeats up to $t+\tau = T$.

We use a mixture of Poisson and Bernoulli priors on tracks [10, 25]: **(1)** The number of newly appearing targets on each frame follows a Poisson distribution; **(2)** Each active track may die with an independent Bernoulli model; **(3)** The number of false positive detections follows a Poisson distribution; **(4)** Each target may be missed with an independent Bernoulli model.

Moreover, given the assignments $\Gamma_{1:T}$, we assume that the position and features of each track is independent of other tracks positions and features. Under these assumptions, the posterior probability can be factorized following:

$$\mathbb{P}(\Gamma_{1:t+\tau} \mid Z_{1:t+\tau}, F_{1:t+\tau}) \propto \prod_{i=1}^N \prod_{k=1}^{t+\tau} g(\gamma_k^i) \ell(\gamma_k^i) \quad (4)$$

where $g(\gamma_k^i)$ measures the a-priori on the track (we refer the reader to more detailed MHT papers [10, 25]) and $\ell(\gamma_k^i)$ is the per-track likelihood which will be detailed in the following Section 2.2.2.

Under such factorization, the MHT optimization can be framed as a multi-dimensional linear assignment problem [25, 26, 2] and we use the open-source HiGHs optimization solver [27] to find the optimal tracks prolongation.

2.2.2. Tracks likelihood model

For unassociated tracks at frame t , we have $\ell(\gamma_t^i=0) = 1$. Otherwise, for any associated track with $\gamma_t^i > 0$, we denote by $\mathbf{z}_t^{\gamma_t^i}$ the position of its associated detection and by $\mathbf{f}_t^{\gamma_t^i}$ its corresponding feature vector. We also define the sets of previously assigned measurements as $Z^{\gamma_{1:t-1}^i}$ and $F^{\gamma_{1:t-1}^i}$.

Since our encoder φ is position-agnostic, positional and feature-based components are statistically independent. Formally:

$$\ell(\gamma_t^i) = p(\mathbf{z}_t^{\gamma_t^i} \mid Z^{\gamma_{1:t-1}^i}) p(\mathbf{f}_t^{\gamma_t^i} \mid F^{\gamma_{1:t-1}^i}), \quad (5)$$

where $p(\mathbf{z})$ and $p(\mathbf{f})$ denote the likelihood densities of the position and visual features, respectively, for a single track. In the standard MHT formulation [10, 2, 26], only the positional likelihood term is considered. For notational simplicity, in the following we omit the track superscript γ_t^i when referring to single-target likelihoods.

Motion modeling. We adopt the classical near-constant-velocity (directed) motion model implemented through a Kalman filter [13, 14]. Each target is represented by a hidden position $\mathbf{w}_t \in \mathbb{R}^2$ and velocity $\mathbf{dw}_t \in \mathbb{R}^2$, forming the state vector $\mathbf{x}_t = (\mathbf{w}_t, \mathbf{dw}_t)^\top$. Kalman filtering enables the recursive estimation [13] of the hidden state, as well as the positional likelihood $p(\mathbf{z}_t \mid \mathbf{z}_{t_0:t-1})$, where t_0 denotes the time of birth of the track.

Feature modeling. Through contrastive self-supervised training (see Section 2.1), we expect feature embeddings to form compact clusters in the feature space, with each cluster corresponding to a distinct target identity. As the visual appearance of a target may evolve slowly over time, we estimate the current cluster center $\hat{\mathbf{f}}_t$ using an exponential moving average (EMA) with coefficient α :

$$\hat{\mathbf{f}}_t = \alpha \hat{\mathbf{f}}_{t-1} + (1 - \alpha) \mathbf{f}_t = \text{EMA}_\alpha(\mathbf{f}_{t_0:t}). \quad (6)$$

The likelihood of the next feature vector is then modeled as a Gaussian centered on the estimated cluster mean:

$$p(\mathbf{f}_{t+1} \mid \mathbf{f}_{t_0:t}) = \mathcal{N}(\mathbf{f}_{t+1}; \hat{\mathbf{f}}_t, \sigma_{\text{vis}}^2 \mathbf{I}_d), \quad (7)$$

where σ_{vis} is estimated from the intra-cluster variability in the feature space.

Since features are expected to form distinct clusters, we define the initial prior as:

$$p(\mathbf{f}_{t_0} \mid \emptyset) = \frac{1}{2M_1} \mathcal{N}(\mathbf{f}_{t_0}; \mathbf{f}_{t_0}, \sigma_{\text{vis}}^2 \mathbf{I}_d), \quad (8)$$

which approximates a uniform mixture over $2M_1$ potential cluster centers, where M_1 is the number of targets detected in the first frame. We found this approximation to be both efficient and robust in practice.

3. EXPERIMENTS AND RESULTS

3.1. Datasets and Evaluation

We evaluated VISUALMHT on challenging neuron-tracking datasets. Tracking neurons in behaving animals is particularly demanding due to large non-rigid deformations, abrupt global motion, and high target density.

Trase-IN. We first evaluated our method on TdTomato-labeled fluorescence recordings of the freshwater polyp *Hydra vulgaris*, where TdTomato fluorophores mark neuronal nuclei (see Figure 1). The publicly available Trase-IN data [20] consist of six time-lapse sequences of 1000 confocal fluorescence frames (1024×1024 pixels). Since no ground-truth annotations are available, we performed a manual evaluation procedure. Each algorithm was run on a 100-frame sequence corresponding to a contraction event of the *Hydra*. We randomly selected 100 tracks spanning the entire sequence and visually inspected them. A track was considered erroneous if it switched between two distinct neurons at any point in time. We then computed the proportion of correct tracks for each tracking algorithm. Neuron detections were obtained using the StarDist detector [5] provided in Trase-IN [20].

SINETRA. To further assess generalization and quantitative performance, we evaluated VISUALMHT on synthetic datasets generated with the SINETRA simulator [19]. We focused on 2D scenarios, where each video sequence consists of $T = 200$ frames at a resolution of 1024×1024 pixels, containing approximately $N \simeq 800$ targets to track. Tracking accuracy was quantified using the Higher-Order Tracking Accuracy (HOTA) metric [28], which jointly evaluates detection, localization, and association performance. We adopted a matching tolerance of 2.0 pixels, as larger deviations would compromise downstream analyses. For each experimental setting, five videos were generated with identical parameters but different random seeds. We report the mean and standard deviation of HOTA scores across these five videos. In SINETRA, detections were obtained using the wavelet-based spot detector [4]. To ensure fair comparison, all methods employed the same post-processing step based on the Rauch–Tung–Striebel (RTS) smoother [29] to refine trajectory localization.

3.2. Evaluated algorithms

We compared VISUALMHT with several state-of-the-art biological tracking algorithms (see Table 1): First, we benchmarked against TRACKMATE [30] from the Fiji software [31], a widely used bioimage tracking tool based on a two-step Linear Assignment Problem (LAP) optimization that approximates an MHT formulation. We employed its advanced directed Kalman filter configuration. Closest to

Method	Motion Model	Optimization	Cost
TRACKMATE [30]	Directed	LAP	Euclidean
EMHT [2]	IMM	MHT ($\tau=3$)	Posterior
KOFT [13]	Directed + Flow	LAP	Likelihood
VISUALMHT	Directed	MHT ($\tau=3$)	Posterior (with visual)

Table 1. Overview of the evaluated tracking methods and their main characteristics.

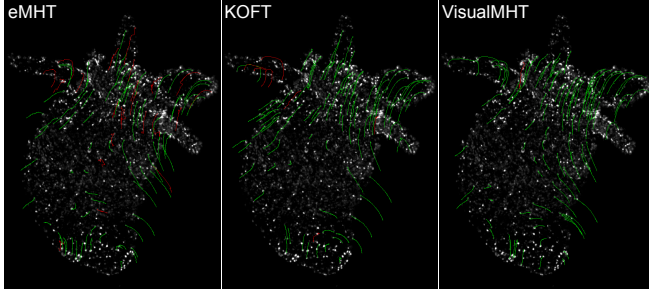


Fig. 3. Neuron tracking in *Hydra vulgaris* during a 100-frame contraction sequence. For each algorithm, 100 randomly sampled tracks were manually validated. Tracks without any identity-switching errors are shown in green, while erroneous tracks are highlighted in red.

our framework, we included EMHT [2], an efficient implementation of Multiple Hypothesis Tracking for biological targets. While EMHT provides robust motion handling with an Interactive Multiple Model (IMM) filter, it does not incorporate visual identity features. Finally, we evaluated the recent KOFT [13] algorithm, which combines Kalman filtering with optical flow to handle large and abrupt displacements, and solves data association through a dedicated LAP optimization.

For all experiments, we used consistent parameter settings across methods when applicable. In particular, the Kalman filters are set with a detection noise of 1.0 pixel and a motion uncertainty of 2.0 pixels. The visual feature variance is estimated at $\sigma_{\text{vis}} = 0.4$ and the feature cluster mean is computed with an EMA smoothing factor $\alpha = 0.5$. For MHT approaches, we use a sliding window size of $\tau = 3$, effectively considering two additional frames over LAP-based approaches. The experiments and code are available at https://github.com/bia-pasteur/visual_mht.

3.3. Results

Figure 3 illustrates qualitative results on *Hydra vulgaris*. Our method substantially improves the accuracy of neuron tracking during rapid contractions. The encoder φ learns robust identity embeddings that are effectively leveraged by the MHT framework to maintain track consistency. On the same 100-frame sequence, EMHT [10], KOFT [13], and our VISUALMHT produced comparable numbers of long trajectories (623, 586, and 595, respectively), defined as tracks persisting for at least 90 frames. However, manual evaluation of 100 randomly selected tracks revealed 23 identity-switching erroneous tracks for EMHT, that are reduced to 6 for KOFT thanks to its refined motion model. In contrast, VISUALMHT achieved only a single erroneous track, highlighting the benefit of incorporating visual identity features for biological targets.

On the SINETRA synthetic benchmark [19], where targets exhibit low signal-to-noise ratios and weakly distinctive appearances

Scenario	Springs 2D	Hydra Flow	FPS
TRACKMATE [30]	$83.7 \pm 3.6\%$	$71.5 \pm 5.8\%$	25
EMHT [2]	$86.9 \pm 3.7\%$	$80.7 \pm 7.3\%$	2
KOFT [13]	$94.3 \pm 1.0\%$	$89.8 \pm 2.5\%$	10
VISUALMHT (ours)	$96.4 \pm 0.7\%$	$93.9 \pm 1.5\%$	5

Table 2. Tracking performance (HOTA at 2 pixels tolerance [28]) on SINETRA synthetic datasets. Frame rates (FPS) are measured on the same laptop and exclude the detection stage.

(see Figure 1), we expected visual features to be less discriminative. Nevertheless, the proposed self-supervised augmentations enabled the encoder to learn invariant identity representations, allowing VISUALMHT to outperform all competing methods. Table 2 reports the quantitative HOTA scores. While EMHT surpasses TRACKMATE due to its more robust multiple-hypothesis association, and KOFT achieves strong results by integrating optical flow, our VISUALMHT further improves performance by leveraging learned appearance features. It achieves 96.4% on the Springs 2D scenario and 93.9% on Hydra Flow, surpassing KOFT’s previous state-of-the-art performance (94.3% and 89.8%, respectively). Although the relative gain on SINETRA is smaller than on real data—likely due to less distinctive simulated appearances—the trend confirms the consistent benefit of integrating visual features.

Finally, we compared runtime performance on a Lenovo Legion laptop equipped with an RTX 2070 GPU. TRACKMATE remains the fastest approach, processing 25 frames per second (FPS). KOFT achieves 10 FPS because of the additional optical flow computation. MHT-based methods are more computationally demanding due to the size of the hypothesis trees: EMHT runs at 2 FPS, while our VISUALMHT reaches 5 FPS. Interestingly, the integration of visual embeddings improves pruning efficiency within MHT, reducing the number of active hypotheses and compensating for the added feature-extraction overhead.

4. CONCLUSION

We presented VISUALMHT, a fully Bayesian Multiple Hypothesis Tracking framework that integrates self-supervised visual features to preserve target identity in biological imaging. By combining probabilistic motion modeling with learned appearance embeddings, our method improves both tracking accuracy and computational efficiency through early pruning of unlikely hypotheses.

VISUALMHT achieves state-of-the-art performance on both synthetic (SINETRA) and experimental (*Hydra vulgaris*) datasets, demonstrating its robustness under severe motion, high target density, and low signal-to-noise conditions.

Currently designed for 2D imaging and validated on neuron-tracking datasets, the proposed tracking framework will be extended to 3D data, and thanks to its self-supervised training scheme, we expect its generalization to broader biological tracking applications.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

6. ACKNOWLEDGMENTS

This work is supported by the Institut Pasteur and France-BioImaging Infrastructure (ANR-10-INBS-04). R.R. and T.L. are supported by the ANR (ANR-21-CE45-0020-01 REBIRTH).

None of the authors declare to have a financial conflict of interest in the results of this study.

7. REFERENCES

- [1] I. Smal and E. Meijering, “Quantitative comparison of multi-frame data association techniques for particle tracking in time-lapse fluorescence microscopy,” *Medical Image Analysis*, vol. 24, no. 1, pp. 163–189, 2015.
- [2] N. Chenouard, I. Bloch, and J.-C. Olivo-Marin, “Multiple hypothesis tracking for cluttered biological image sequences,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2736–3750, 2013.
- [3] I. Smal, M. Loog, W. Niessen, and E. Meijering, “Quantitative comparison of spot detection methods in fluorescence microscopy,” *IEEE transactions on Medical Imaging*, vol. 29, no. 2, pp. 282–301, 2009.
- [4] J.-C. Olivo-Marin, “Extraction of spots in biological images using multiscale products,” *Pattern Recognition*, vol. 35, no. 9, pp. 1989–1996, 2002.
- [5] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, “Cell detection with star-convex polygons,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, proceedings, part II 11*. Springer, 2018, pp. 265–273.
- [6] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, “Cellpose: a generalist algorithm for cellular segmentation,” *Nature Methods*, vol. 18, no. 1, pp. 100–106, 2021.
- [7] C. Malin-Mayor, P. Hirsch, L. Guignard, et al., “Automated reconstruction of whole-embryo cell lineages by learning from sparse annotations,” *Nature Biotechnology*, vol. 41, no. 1, pp. 44–49, 2023.
- [8] B. Gallusser and M. Weigert, “Trackastra: Transformer-based cell tracking for live-cell microscopy,” in *European Conference on Computer Vision*. Springer, 2025, pp. 467–484.
- [9] T. Ben-Haim and T. R. Raviv, “Graph neural network for cell tracking in microscopy videos,” in *European Conference on Computer Vision*. Springer, 2022, pp. 610–626.
- [10] D. Reid, “An algorithm for tracking multiple targets,” *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.
- [11] T. Kaiser, M. Schier, and B. Rosenhahn, “Cell tracking according to biological needs—strong mitosis-aware multi-hypothesis tracker with aleatoric uncertainty,” *arXiv preprint arXiv:2403.15011*, 2024.
- [12] K. Jaqaman, D. Loerke, M. Mettlen, et al., “Robust single-particle tracking in live-cell time-lapse sequences,” *Nature Methods*, vol. 5, no. 8, pp. 695–702, 2008.
- [13] R. Reme, A. Newson, E. Angelini, et al., “Particle tracking in biological images with optical-flow enhanced kalman filtering,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.
- [14] A. Genovesio, T. Liedl, V. Emiliani, et al., “Multiple particle tracking in 3-d+ t microscopy: method and application to the tracking of endocytosed quantum dots,” *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1062–1070, 2006.
- [15] Y. Yao, I. Smal, I. Grigoriev, et al., “Deep-learning method for data association in particle tracking,” *Bioinformatics*, vol. 36, no. 19, pp. 4935–4941, 2020.
- [16] N. Wojke, A. Bewley, and D. Paulus, “Simple online and real-time tracking with a deep association metric,” in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.
- [17] Y. Du, Z. Zhao, Y. Song, et al., “Strongsort: Make deepsort great again,” *IEEE Transactions on Multimedia*, 2023.
- [18] R. Reme, V. Piriou, A. Hanson, et al., “Tracking intermittent particles with self-learned visual features,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.
- [19] R. Reme, A. Newson, E. Angelini, et al., “Sinetra: A versatile framework for evaluating single neuron tracking in behaving animals,” in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, pp. 1–5.
- [20] A. Hanson, R. Reme, N. Telerman, et al., “Automatic monitoring of neural activity with single-cell resolution in behaving hydra,” *Scientific Reports*, vol. 14, no. 1, pp. 5083, 2024.
- [21] F. de Chaumont, S. Dallongeville, N. Chenouard, et al., “Icy: an open bioimage informatics platform for extended reproducible research,” *Nature Methods*, vol. 9, no. 7, pp. 690–696, 2012.
- [22] J. Wang, Y. Song, T. Leung, et al., “Learning fine-grained image similarity with deep ranking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [24] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [25] A. B. Poore, “Multidimensional assignment formulation of data association problems arising from multitarget and multisensor tracking,” *Computational Optimization and Applications*, vol. 3, no. 1, pp. 27–57, 1994.
- [26] L. Feng, Y. Xu, Y. Yang, and X. Zheng, “Multiple dense particle tracking in fluorescence microscopy images based on multidimensional assignment,” *Journal of Structural Biology*, vol. 173, no. 2, pp. 219–228, 2011.
- [27] Q. Huangfu and J. J. Hall, “Parallelizing the dual revised simplex method,” *Mathematical Programming Computation*, vol. 10, no. 1, pp. 119–142, 2018.
- [28] J. Luiten, A. Osep, P. Dendorfer, et al., “Hota: A higher order metric for evaluating multi-object tracking,” *International Journal of Computer Vision*, vol. 129, pp. 548–578, 2021.
- [29] H. E. Rauch, F. Tung, and C. T. Striebel, “Maximum likelihood estimates of linear dynamic systems,” *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.
- [30] J.-Y. Tinevez, N. Perry, J. Schindelin, et al., “Trackmate: An open and extensible platform for single-particle tracking,” *Methods*, vol. 115, pp. 80–90, 2017.
- [31] J. Schindelin, I. Arganda-Carreras, E. Frise, et al., “Fiji: an open-source platform for biological-image analysis,” *Nature Methods*, vol. 9, no. 7, pp. 676–682, 2012.