Gaussian Process Assisted Labeling for Video Object Tracking: An Analysis

Felipe J. A. Maia[1], **José E. B. Maia**[1]

[1] Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual do Ceará (UECE) – Fortaleza – CE – Brazil

felipe.maia@aluno.uece.br, jose.maia@uece.br

Abstract. *The development of datasets for video object tracking is inherently resource-intensive, primarily due to the requirement of exhaustive frame-by-frame annotation of target positions. Various strategies have been explored to mitigate these costs. This study evaluates the effectiveness of Gaussian Process (GP) Regression as an auxiliary tool for annotating target trajectories within video tracking benchmarks. By leveraging manual ground-truth data sampled at intervals of k frames, GP models are employed to infer the coordinates for the intervening unlabeled frames. Experimental evaluations conducted on Multi-Object Tracking (MOT) datasets demonstrate a mean Area Under the Curve (AUC) exceeding 90% at $k = 20$ for sequences characterized by uniform target dynamics and stationary camera platforms. Furthermore, the proposed approach sustained a mean precision higher than 80% at same k even when subjected to more challenging and unconstrained scenarios.*

1. Introduction

Constructing large-scale datasets is essential for deep learning development [Chang et al. 2023], yet it often poses a significant bottleneck for the implementation of new machine learning applications [Huang and Zhao 2024].

The development of video tracking datasets necessitates not only the acquisition of representative video sequences but also the exhaustive manual annotation of object coordinates across the entire temporal domain. The annotation phase is a fundamental component, as it establishes the ground truth for individual spatial positioning across temporal sequences. This stage is notoriously labor-intensive. It necessitates the precise marking of numerous targets across extensive frame sequences. The widespread adoption of deep learning in object tracking highlights the necessity of voluminous datasets for the training and refinement of tracking architectures [Muller et al. 2018].

To mitigate this burden, regression techniques can be leveraged to automate portions of the labeling process. Under this framework, manual intervention is limited to annotating every k frames. This study aims to evaluate the efficacy of Gaussian Process (GP) Regression as a framework for assisted labeling in video-based person tracking datasets.

The implementation of assisted labeling holds the potential to reduce manual effort to a mere fraction of the traditional workload. While this method introduces a marginal degree of positional variance, such errors are often statistically insignificant depending on the specific requirements of the downstream application.

The primary contribution of this study lies in the analysis of sparse annotation strategies enhanced by Gaussian Process-based coordinate inference for intermediate frames. Experiments performed on person-tracking datasets, featuring varied target dynamics and camera viewpoints, demonstrate the effectiveness of this procedure under different skip intervals (k). To our knowledge, no previous work has investigated GP performance in this task.

The remainder of this work is structured as follows: Section 2 reviews related literature pertinent to this research. Section 3 outlines the experimental methodology, including Gaussian Process fundamentals and the metrics utilized for evaluation. Section 4 presents the experimental results, followed by a detailed discussion in Section 5. Finally, Section 6 provides the concluding remarks.

2. Related Work

The Multi-Object Tracking (MOT) subfield introduces further complexities to datasets development. Multiple targets (predominantly people) frequently enter and exit the scene or experience periods of occlusion within a single sequence. These dynamic environmental factors significantly complicate the tracking and labeling process. The work by [Gil-Jiménez et al. 2016] investigate the application of cubic spline interpolation for estimating object bounding boxes between keyframes to facilitate video dataset construction. While the authors conduct an extensive analysis of the estimation quality achieved through various interpolators, their study is constrained by a maximum keyframe interval of 20 frames.

The use of Active Learning as a strategy to mitigate annotation costs is proposed in [Maclang et al. 2023]. It is applied to vehicle classification and traffic flow measurement tasks. This approach prioritizes the labeling of the most informative samples, thereby minimizing manual intervention for instances in which the model exhibits high confidence.

Focusing on dataset curation, [Castro-Girón and Shiguihara 2021] introduce a labeling protocol designed for video action recognition tasks. Their framework aims to standardize data organization during the curation process and enhance inter-annotator agreement. Similarly, [Yang 2022] addresses the construction of action recognition benchmarks by proposing a tool that streamlines the labeling pipeline through the automated detection and identification of subjects within a scene.

Advancements in workflow optimization are explored by [Gutiérrez et al. 2025], who evaluate the integration of Artificial Intelligence (AI) driven pre-annotations and Human-in-the-loop systems. Their objective is to reduce labeling duration while maintaining high qualitative standards. [Tarsoly et al. 2025] investigate AI-assisted labeling for pixel-level instance segmentation datasets, providing a comparative analysis of annotation speed and accuracy against conventional manual strategies.

This bottleneck is not unique to computer vision applications but is a pervasive challenge across supervised learning and model evaluation. To address this, the literature proposes several mitigation strategies, including the development of specialized user interfaces designed to streamline human interaction and the deployment of AI models to provide semi-automated assistance to experts during the labeling process [Dai et al. 2021]

The deployment of Large Language Models (LLMs) is investigated by [Horych et al. 2025] as a means of automating data labeling for text classification tasks. This methodology aims to mitigate the total costs of dataset assembly, thereby streamlining the economic overhead of application development.

Dataset annotation becomes significantly more resource-intensive when it necessitates the judgment of domain experts. To address this, [Krenzer et al. 2022] propose a video labeling strategy wherein experts annotate only keyframes, while the remaining data is populated through a combination of machine learning algorithms and non-expert contributors.

The growing body of literature dedicated to facilitating the annotation process underscores the significance of investigating strategies to mitigate the costs of this essential phase of dataset generation.

3. Methodology

3.1. Gaussian Process Regression

A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution [Williams and Rasmussen 2006]. A GP is uniquely and completely characterized by its mean function, $m(x)$, and its covariance function (or kernel), $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')),$$

where:

$$m(x) = \mathbb{E}[f(x)], \quad k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].$$

In the context of video-based object tracking, the variable x denotes the temporal index of the sequence during which the object is tracked, while the function output $f(x)$ represents the spatial coordinates of the target throughout that trajectory, as shown in Figure 1.

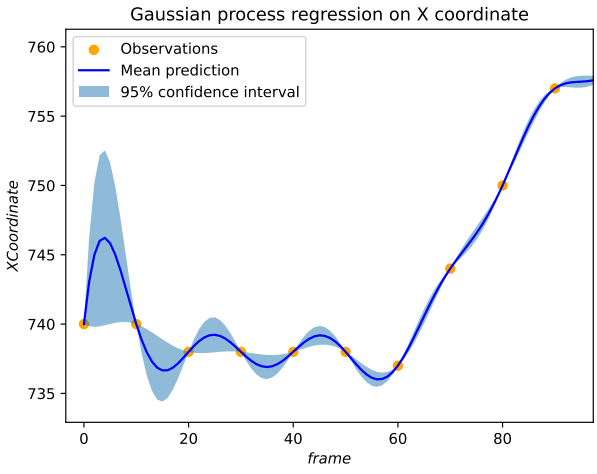

Figure 1. Illustrative Application of Gaussian Processes to Trajectory Estimation.

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, we assume the existence of a latent function f that defines the mapping $y_i = f(x_i) + \epsilon_i$. In this formulation, the mapping error (or observation noise) ϵ_i is assumed to follow a Gaussian distribution with zero mean $\mathcal{N}(0, \sigma_n^2)$.

The joint distribution of the observed inputs $\mathbf{X} = [x_1, \ldots, x_n]^T$, the target vector $\mathbf{y} = [y_1, \ldots, y_n]^T$, and a test input x_* is expressed as:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} K_y & K(\mathbf{X}, x_*) \\ K(x_*, \mathbf{X}) & K(x_*, x_*) \end{bmatrix} \right),$$

in which $K_y = K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I$. The predictive distribution for a novel test point x_* is Gaussian, denoted as $f_* | \mathbf{X}, \mathbf{y}, x_* \sim \mathcal{N}(\mu_*, \sigma_*^2)$, where:

$$\mu_* = K(x_*, \mathbf{X}) K_y^{-1} \mathbf{y},$$

$$\sigma_*^2 = K(x_*, x_*) - K(x_*, \mathbf{X}) K_y^{-1} K(\mathbf{X}, x_*).$$

Various types of kernel functions $k(x, x')$ can be employed in the construction of the model. In our analysis, the Matérn kernel was selected:

$$k(x_i, x_j) = \frac{1}{\Gamma(\nu)\, 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{\ell} d(x_i, x_j) \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} d(x_i, x_j) \right).$$

This kernel function is a generalization of the Radial Basis Function (RBF), also containing the length-scale parameter ℓ, which determines how rapidly the correlation between two points decays as the distance between them increases. In addition, the parameter ν controls the smoothness of the resulting function. Here, Γ denotes the Gamma function, and K_ν represents the modified Bessel function. The optimization of the kernel hyperparameters is achieved by maximizing the marginal log-likelihood [Murphy 2012]:

$$\log p(\mathbf{y} \mid \mathbf{X}) = \log \mathcal{N}(\mathbf{y} \mid 0, \mathbf{K}_y) = -\frac{1}{2} \mathbf{y}^\top \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_y| - \frac{N}{2} \log(2\pi).$$

3.2. Experimental Setup

The quality of the coordinate estimation performed by the GP was assessed using datasets widely adopted as benchmarks for multi-object tracking algorithms. All video sequences providing ground-truth annotations for the localization of individuals within the scene were utilized.

For each video sequence, tracking data were partitioned according to unique subject identifiers (IDs). Each track encapsulates a continuous trajectory for a single individual, spanning from the initial point of entry into the frame until the subject's departure or the conclusion of the video.

The temporal duration of each track is determined by the period during which an individual remains within the field of view. Each observation within a sequence consists of a bounding box vector (x,y,h,w), representing the spatial coordinates and dimensions of the individual at a specific frame.

A sensitivity analysis was conducted to evaluate the effect of different kernel functions, namely RBF, Matérn, and Rational Quadratic. A subset containing 10% of the tracks from all datasets was removed and used exclusively for this purpose. The sensitivity analysis indicated only minor differences among the three kernels, with a precision variation of 0.8% between the worst and best performing configurations. Following this analysis, the Matérn kernel ($\nu = 1.5$) was selected.

For each track, a training set was constructed by subsampling every k-th frame from the original data. This procedure was intended to simulate a sparse annotation workflow, wherein manual labels are provided at an interval of k frames. Four independent Gaussian Process (GP) regression models were fitted, one for each of the subject's positional variables (x, y, h, w). These models were subsequently employed to infer the spatial coordinates for the omitted frames, thereby generating a reconstructed, complete trajectory.

The reconstructed trajectories were subsequently compared against the original ground-truth track to compute the relevant evaluation metrics. This experimental protocol was iterated across multiple values of k to assess the impact of the skip interval on the sequence reconstruction error.

3.3. Performance Metrics

The evaluation of tracking algorithms is typically grounded in metrics designed to quantify both spatial accuracy and temporal robustness. This section details the primary performance indicators employed in this study: Area Under the Curve (AUC) of the Intersection over Union (IoU), Center Location Error and Precision [Wu et al. 2013].

The Intersection over Union (IoU) metric quantifies the degree of spatial overlap between a predicted bounding box and its corresponding ground-truth annotation. Formally, let B_p denote the predicted bounding box and B_g represent the ground-truth bounding box, the IoU is defined as:

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|},$$

where $|B_p \cap B_g|$ denotes the area of intersection between the two regions and $|B_p \cup B_g|$ represents their area of union. The resulting IoU value is constrained to the interval $[0, 1]$, where a value of 1 signifies perfect spatial correspondence between the predicted and actual bounding boxes. For each tracking sequence, the mean IoU value was computed.

The success plot is derived from the tracking success rate evaluated across a range of Intersection over Union (IoU) thresholds. Given a threshold $t \in [0, 1]$, the success rate is formulated as:

$$S(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(IoU_i \geq t).$$

The Area Under the Curve (AUC) of IOU success metric represents the integral of the success curve over the entire threshold range:

$$AUC = \int_0^1 S(t)dt.$$

The Center Location Error (CLE) quantifies the Euclidean distance between the centroids of the predicted and ground-truth bounding boxes. This metric provides a quantitative measure of the target's localization error throughout the duration of the video frames. For each tracking sequence, the mean CLE value was computed. Let (x_p, y_p) denote the center of the predicted box and (x_g, y_g) represent the center of the ground-truth box; the error is then formulated as:

$$d = \sqrt{(x_p - x_g)^2 + (y_p - y_g)^2}.$$

The Precision metric is defined as the proportion of frames in which the Center Location Error (CLE) remains below a specified threshold, τ:

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(d_i < \tau),$$

where N represents the total number of evaluated frames; d_i denotes the center location error for the i-th frame; $\mathbb{I}(\cdot)$ is the indicator function, which yields a value of 1 if the specified condition is satisfied and 0 otherwise. In our experiments, a threshold of $\tau = 20$ pixels was adopted for this calculation.

Performance metrics were computed for each individual tracking sequence. Subsequently, the aggregate mean values for each dataset were calculated using a weighted average based on the length of each sequence.

4. Results

Table 1 summarizes key information derived from widely adopted datasets employed as benchmarks for Multi-Object Tracking (MOT) algorithms [Wojke et al. 2017, Maggiolino et al. 2023, Zhang et al. 2022, Cetintas et al. 2023, Gao et al. 2025]. The data compiled in this table corresponds exclusively to video sequences for which ground-truth annotations are available. Certain datasets include additional sequences within challenge subsets that lack publicly accessible ground-truth files; consequently, these were excluded from the present study.

Both MOT17 and MOT20 feature pedestrian trajectories within urban settings. MOT17 includes three sequences captured via stationary cameras and four sequences involving moving camera settings. This dataset exhibits significant scale variation, with instances of pedestrians appearing in close proximity to the sensor. All four sequences in MOT20 utilize static cameras positioned at elevated, distant vantage points, characteristic of typical surveillance scenarios.

The SportsMOT and SoccerNet datasets feature video sequences captured within athletic environments. SportsMOT includes examples from volleyball, basketball, and football, while SoccerNet is dedicated solely to football. Both benchmarks utilize professional television broadcast framing, occasionally including on-screen graphic overlays

Dataset	Reference	Number of videos	Mean Resolution	Mean Fps	Total Tracks	Mean Track Length (frames)	Mean Boundbox Size
MOT17	[Milan et al. 2016]	7	1737x994	27	796	257	77x166
MOT20	[Dendorfer et al. 2020]	4	1666x1030	25	2332	580	63x140
SportsMOT	[Cui et al. 2023]	90	1280x720	25	1280	475	58x119
SoccerNet	[Cioppa et al. 2022]	106	1920x1080	25	2638	491	48x99
DanceTrack	[Sun et al. 2022]	65	1745x984	20	692	829	160x300

Table 1. Comparison of popular MOT datasets.

such as timers and scoreboards. Notably, while the number of video sequences in these datasets significantly exceeds those of the MOT17 and MOT20 benchmarks, the total number of unique trajectories (tracks) remains comparable.

Each track represents a continuous tracking sequence of a unique subject, spanning from their initial entry into the frame until their eventual egress from the scene. Within the sports-oriented datasets, the number of individuals is strictly constrained to the active participants of the athletic activities. Conversely, the MOT17 and MOT20 benchmarks feature unconstrained environments, resulting in a significantly higher track density per video sequence.

The DanceTrack dataset comprises video sequences of dance performances characterized by significant variation in subject framing. The camera placement ranges from close-range captures to elevated, distant vantage points. Furthermore, the majority of the sequences in this benchmark are recorded using stationary cameras or exhibit minimal camera motion.

k = 5			
Dataset	**Mean CLE**	**Mean Precision**	**Mean AUC**
MOT17	0.72	0.9970	0.9768
MOT20	0.35	0.9999	0.9857
SportsMOT	2.42	0.9972	0.8968
SoccerNet	2.20	0.9918	0.9020
DanceTrack	5.73	0.9530	0.9082
k = 20			
Dataset	**Mean CLE**	**Mean Precision**	**Mean AUC**
MOT17	6.17	0.9414	0.8927
MOT20	1.48	0.9958	0.9515
SportsMOT	12.47	0.8600	0.6846
SoccerNet	11.15	0.8911	0.7131
DanceTrack	19.91	0.6527	0.7517
k = 40			
Dataset	**Mean CLE**	**Mean Precision**	**Mean AUC**
MOT17	17.53	0.8416	0.8161
MOT20	4.11	0.9726	0.8907
SportsMOT	43.86	0.4857	0.4201
SoccerNet	41.12	0.5851	0.4611
DanceTrack	41.76	0.3926	0.6042

Table 2. Test results obtained using k values of 5, 20, and 40.

In MOT17 and MOT20, targets commonly maintain uniform motion with constant velocity and direction, often following linear paths across the field of view. Exceptions occur in MOT17 sequences involving camera motion, where subjects exhibit apparent erratic movement. Conversely, the SportsMOT, SoccerNet, and DanceTrack datasets are

characterized by high-frequency erratic motion. This non-linear behavior is driven by the dynamics of sports and is even more evident in the high-agility maneuvers found in DanceTrack.

Despite their status as primary benchmarks for evaluating state-of-the-art tracking algorithms, these datasets are characterized by limited data volume. This constraint is primarily due to the prohibitive costs and labor-intensive nature of manual subject localization.

Table 2 summarizes the GP performance results for $k \in \{5, 20, 40\}$. The progression of these metrics across other values of k can be observed in Figures 2, 3, and 4. Figure 2 depicts the Mean Center Location Error (CLE) for each dataset as a function of the parameter k. It was observed that the majority of the datasets exhibited significantly elevated Mean CLE values for $k \geq 40$. To facilitate a detailed visualization of the error trends within the range $k < 40$, a logarithmic scale was employed on the vertical axis.

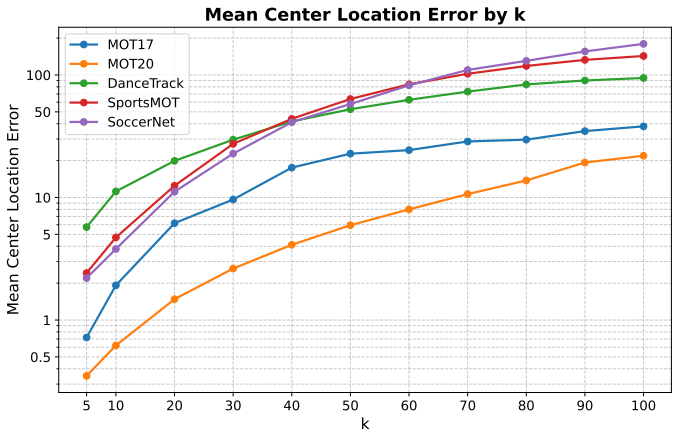

Figure 2. Results of Mean Center Location Error (log scale) for varying k values.

A consistent pattern for Mean CLE was identified within the SportsMOT, SoccerNet, and DanceTrack datasets. The video sequences in these benchmarks are characterized by targets exhibiting more erratic motion profiles, a factor that likely accounts for the error remaining substantially higher than the levels observed in the MOT17 and MOT20 datasets.

Across all datasets, with the exception of DanceTrack, a mean error of less than 13 pixels was observed for $k \leq 20$. Even within the DanceTrack benchmark, the error remained below 20 pixels relative to the ground-truth label center. Considering the mean bounding box sizes detailed in Table 1, these results suggest a tolerable margin of error, implying that more than 80% of the target remains contained within the predicted bounding box during the labeling phase.

The Mean Precision for the tested datasets is presented in Figure 3. The results reinforce the notion that environments within the MOT17 and MOT20 datasets are particularly conducive to GP-based assisted labeling. In this datasets, is observed that substantial errors (exceeding the threshold $\tau = 20$ pixels) only emerge for k values of 30 or higher, where the precision rate first drops below the 90% mark.

The effect of k on the Area Under the IoU Curve is depicted in Figure 4. This

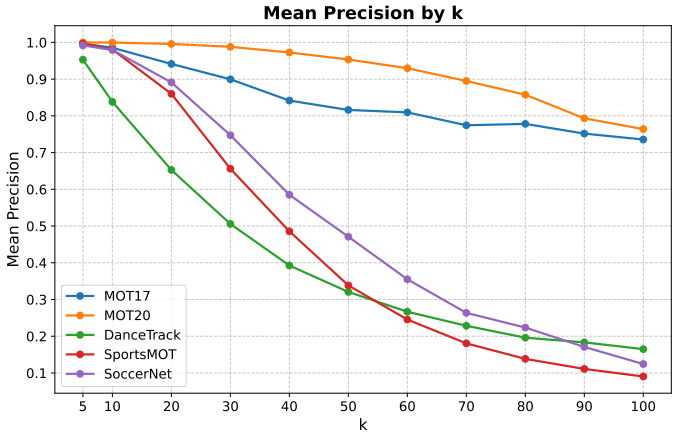

Figure 3. Results of Mean Precision for the values of k.

metric is of high relevance as it evaluates the full region of the bounding box, providing a more comprehensive assessment than metrics based exclusively on the center point. It was observed that the DanceTrack dataset, which generally features larger bounding boxes, exhibited a more gradual decline compared to the SportsMOT and SoccerNet benchmarks. This underscores the utility of this metric for optimizing the selection of k, particularly when considering the spatial occupancy of the tracked object relative to the frame dimensions. For datasets characterized by more consistent motion patterns, results exceeding 80% were maintained for values of k up to 50.

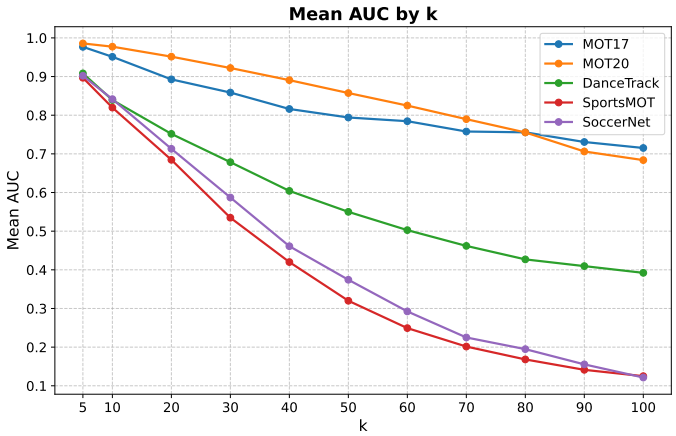

Figure 4. Results of Mean AUC for the values of k.

5. Discussion

In stationary camera sequences with constant subject motion, significantly higher values of k may be utilized with minimal performance degradation. Specifically, the precision for $k = 30$ within the MOT20 dataset remains approximately 98%.

Within datasets characterized by erratic subject motion, a more pronounced decline in precision is observed for values of k exceeding 20. This degradation is particularly evident in the DanceTrack benchmark. This phenomenon suggests a cumulative effect from the inherent motion characteristics of the targets coupled with their greater

proximity to the camera sensor, evidenced by the high ratio between the mean bounding box dimensions and the average video resolution.

A salient feature of the DanceTrack dataset is that all sequences were captured at a frame rate of 20 fps. Consequently, any frame-skipping interval exerts a more pronounced impact on this dataset than on any other tested benchmark, as the temporal gap between frames is larger. This factor renders DanceTrack the most challenging scenario within the current experimental framework.

It is important to note that for $k = 5$, the observed error is negligible, depending on the specific application requirements and recording characteristics. Despite being the smallest tested skip interval, this value already facilitates a significant 80% reduction in manual annotation effort during the dataset construction phase.

Another promising configuration is $k = 20$, which sustained a mean precision exceeding 80% across nearly all evaluated benchmarks. The resulting increase in annotation throughput at this interval may enable the inclusion of iterative review and refinement stages, allowing for the correction of labels where errors are most prominent. The integration of these strategies potentially accelerates the labeling pipeline without compromising the final annotation quality.

Furthermore, the ratio between the mean bounding box dimensions and the video resolution significantly influences the optimal selection of k. The Mean AUC analysis suggests that a higher ratio, indicating that the subject is in closer proximity to the sensor, diminishes the adverse impact of k on the Mean AUC. This phenomenon is clearly reflected in the behavior of the DanceTrack dataset.

The computational complexity associated with kernel hyperparameter optimization is dominated by the matrix inversion operation in $\mathbf{K}_y^{-1}$. In most linear algebra packages, this operation exhibits a computational complexity of $O(N^3)$. Consequently, this complexity may become a limiting factor when dealing with very long tracks. For the examples evaluated in the tested datasets, however, this limitation was not prohibitive. For instance, in the SoccerNet dataset, each analyzed video, containing on average 24 tracks with an mean duration of 491 frames, required 50.66 seconds to fit the GP models under our experimental setup (Intel(R) Xeon(R) CPU @ 2.20GHz running on Google Colab).

6. Conclusions

Based on the empirical evidence, the GP assisted labeling strategy demonstrates significant efficacy across a diverse range of operational scenarios. An analysis of video sequence characteristics indicates that extended labeling intervals of $k \geq 30$ are feasible when the following conditions are met: uniform motion of the tracked targets; utilization of stationary camera; high spatial occupancy of the objects within the image frame.

Even in suboptimal environments, the implementation of conservative intervals, such as k = 5, facilitates a substantial increase in labeling throughput. This configuration introduces a mean error that remains below 1% relative to the total bounding box dimensions, representing a highly favorable trade-off between speed and precision.

Furthermore, the methodology remains viable when integrated with systematic review-and-refinement protocols. In such frameworks, the strategy is applicable even if it

yields perceptible errors, as these can be manually rectified during post-processing while still achieving a significant net reduction in total annotation time.

The limitations of this study include the restricted scope of the evaluated datasets, and the specific focus on Multi-Object Tracking (MOT) tasks. Consequently, future research should prioritize complementary testing across a broader range of benchmarks and extend the methodology to other computer vision domains, such as pose estimation. Exploring alternative performance metrics is also recommended to provide a more comprehensive assessment of the proposed approach.

References

Castro-Girón, E. and Shiguihara, P. (2021). A method for dataset labeling for activity recognition in videos. In *2021 IEEE XXVIII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4.

Cetintas, O., Brasó, G., and Leal-Taixé, L. (2023). Unifying short and long-term tracking with graph hierarchies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22877–22887.

Chang, N., Ferroni, F., Tarr, M. J., Hebert, M., and Ramanan, D. (2023). Thinking like an annotator: Generation of dataset labeling instructions.

Cioppa, A., Giancola, S., Deliège, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., and Van Droogenbroeck, M. (2022). Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3491–3502.

Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., and Wang, L. (2023). Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*.

Dai, K., Zhao, J., Wang, L., Wang, D., Li, J., Lu, H., Qian, X., and Yang, X. (2021). Video annotation for visual tracking via selection and refinement.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes.

Gao, R., Qi, J., and Wang, L. (2025). Multiple object tracking as id prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27883–27893.

Gil-Jiménez, P., Gómez-Moreno, H., López-Sastre, R., and Maldonado-Bascón, S. (2016). Geometric bounding box interpolation: an alternative for efficient video annotation. *EURASIP Journal on Image and Video Processing*, 2016(1):8.

Gutiérrez, J., Gutiérrez, V., Mora, Á., Rodriguez, S., and Blanco, J. L. (2025). An Evaluation of Hybrid Annotation Workflows on High-Ambiguity Spatiotemporal Video Footage. *arXiv e-prints*, page arXiv:2510.21798.

Horych, T., Mandl, C., Ruas, T., Greiner-Petter, A., Gipp, B., Aizawa, A., and Spinde, T. (2025). The promises and pitfalls of llm annotations in dataset labeling: a case study on media bias detection.

Huang, Q. and Zhao, T. (2024). Data collection and labeling techniques for machine learning.

Krenzer, A., Makowski, K., Hekalo, A., Fitting, D., Troya, J., Zoller, W. G., Hann, A., and Puppe, F. (2022). Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine*, 21(1):33.

Maclang, A. R., Orante, M. L., Salvador, R. D., Del Carmen, D. J., and Cajote, R. D. (2023). Video dataset labeling using active learning with applications in vehicle classification and traffic flow rate measurement. In *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, pages 936–941.

Maggiolino, G., Ahmad, A., Cao, J., and Kitani, K. (2023). Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3025–3029.

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking.

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., and Ghanem, B. (2018). Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., and Luo, P. (2022). Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tarsoly, S., Galambos, P., and Károly, A. I. (2025). Comparison of manual and ai-assisted labeling techniques in pixel-wise instance segmentation. In *2025 IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000115–000122.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649.

Wu, Y., Lim, J., and Yang, M.-H. (2013). Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418.

Yang, F. (2022). A multi-person video dataset annotation method of spatio-temporally actions.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, page 1–21, Berlin, Heidelberg. Springer-Verlag.