Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 2, Pages: 87-94, Year:2025

Article

A Robust Adaptive Temporal Attention Mixture Network for Multi-Object Tracking and Segmentation

Eisha Tur Razia, Zainub Rashid, Areeba Mujtaba, Sidra Zafar*, and Huma Mughal*

Department of Computer Science, Kinnaird College for Women University Lahore, Lahore 54000, Pakistan
* Correspondence: sidra.zafar@kinnaird.edu.pk

Submitted: 20-08-2025**, Revised:** 28-11-2025, **Accepted:** 19-12-2025

Abstract

Traditional video object tracking and segmentation methods including Gaussian Mixture Models (GMM) and fuzzy morphological filtering produce unstable results when operating under conditions of changing illumination, camera movement and object blocking. The current deep learning and attention-based methods have enhanced temporal stability but most existing pipelines continue to operate in separate stages for segmentation and denoising and tracking which results in progressive error accumulation and decreased real-world performance consistency. The research introduces ATAM-Net (Adaptive Temporal Attention Mixture Network) as an end-to-end framework which unifies adaptive Gaussian mixture modeling with temporal attention-based denoising and differentiable object association into one system. The ATAM-Net system learns particular illumination parameters for each pixel through its adaptation process and temporal attention minimizes frame-to-frame noise and appearance-aware embeddings maintain identity during fast motion and occlusion. The SportsMOT dataset serves as the exclusive testing ground for ATAM-Net to prove its ability to produce stable visual segmentation and smooth temporal coherence and precise multi-player tracking in active sports environments. The qualitative visual results show clear object edges and minimal flicker. These results indicate that ATAM-Net provides high-motion environments such as sports analytics, reliable and interpretable approach for multi-object tracking and segmentation in complex and real-time video understanding.

Keywords: Video segmentation, multi-object tracking, Gaussian mixture, temporal attention, differentiable association, SportsMOT, ATAM-Net.

1. Introduction

The development of intelligent surveillance systems and autonomous sense systems needs video-based object detection and tracking methods which deliver high accuracy results. The system needs to maintain continuous object tracking during complex movements and different lighting situations for sports analytics and traffic monitoring and crowd management and robotic navigation.

The statistical validation of Gaussian Mixture Models (GMM) and Generalized Gaussian Distributions (GGD) has established their classical methods for background subtraction [2,3]. The methods produce unsatisfactory results when they encounter dynamic light conditions and reflective surfaces and flexible object movements occur.

The Mixture of Adaptive Gaussians (MoAG) model which Mahalingam and Subramoniam developed uses Gaussian mixture modeling with fuzzy morphological filtering segmentation to

 https://doi.org/10.57041/ncbxkt16

minimize noise levels [1,4]. The methods use predetermined adaptation systems which prove ineffective when dealing with actual environments that experience abrupt lighting changes and objects becoming hidden from view.

The development of deep learning technology has accelerated object tracking progress because CNNs and RNNs and Transformer-based models can extract features from large video datasets [17,18]. The current tracking pipelines consist of separate modules for segmentation and denoising and tracking which results in accumulated errors throughout the processing sequence.

The paper presents ATAM-Net (Adaptive Temporal Attention Mixture Network) as an end-to-end trainable system which combines adaptive mixture learning with temporal attention and learnable object association to address present system constraints. The tracking accuracy in difficult video segments improves through ATAM-Net because it performs complete optimization of all vital pipeline elements.

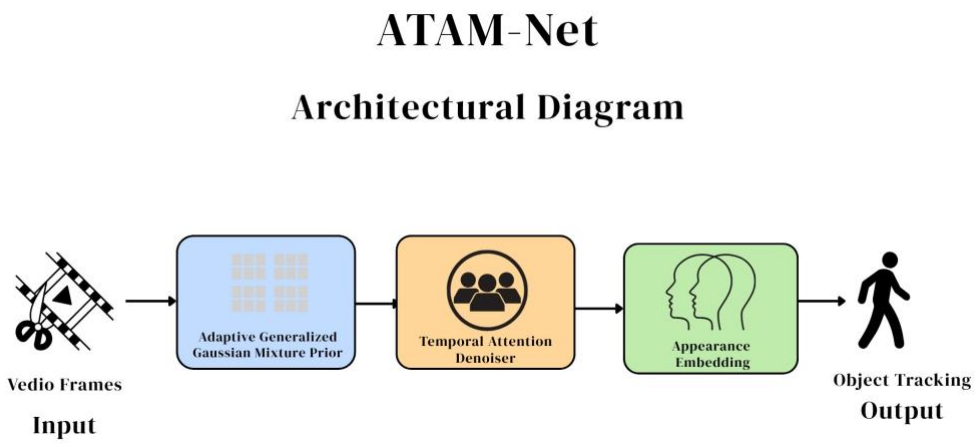

Figure 1: ATAM-Net workflow: (a) per-pixel adaptive mixture modeling for segmentation, (b) temporal attention denoiser to aggregate multi-frame features, (c) appearance embedding and differentiable association head to preserve identity across frames, and (d) joint loss optimization.

2. Materials and Methods

Adaptive Temporal Attention Mixture Network (ATAM-Net)

Adaptive Temporal Attention Mixture Network (ATAM-Net) is an end-to-end framework that aims at the strong video object detection and multi-object tracking. It incorporates three major modules:

Adaptive Mixture Modeling of pixel-wise background/foreground conditioning. Motion Consistency and Temporal Attention of multi-frame denoising. Differentiable Object Association Frame-to-Frame identity tracking. These modules can jointly be used to segment, denoise, and track objects in dynamic sports scenes, as well as ATAM-Net [17–19].

Adaptive Mixture Model of the Pixel Segmentation

The ATAM-Net models the appearance of every pixel as a mixture based on a statistical method in such a way that it is able to differentiate between the background and foreground even when there is a change in illumination or an active background. The pixels are analyzed one at a time and a foreground decision is taken based on the likelihood of the pixel to belong to the background distribution.

Exemplary Concept

$$PixelValue_t = Background + Noise \quad (1)$$

The number of pixels at time *t* is dominated by the background but can have alterations as a result of moving objects or noise. Vast differences between the anticipated background are a sign of foreground pixels.

Formal Equation

$$P(It) = \sum_{k=11}^{K} \text{wk} \cdot \text{N(It} \mid \mu\text{k}, \sigma 2\ \text{k}), \quad \text{where} \quad \sum_{K=1} \text{wk} = 1 \qquad (2)$$

I_t: pixel intensity at time *t*,

w_k: weight of each Gaussian component, • μ_k, σ_k^2: mean and variance of every element,

N: Gaussian probability distribution.

If the pixel probability $P(I_t)$ is small, and the pixel is foreground (moving object). It is a Gaussian Mixture model using the Expectation-Maximization (EM) algorithm to estimate its parameters [20], modeled on early background mixture modeling techniques [1,2].

Parameter Adaptation

$$\mu^{new} = (1-\rho)\mu^{old} + \rho I_t \qquad (3)$$

k k

where ρ is a small learning rate. The conceptual basis of this incremental update is a slow adaptive mixture modeling of changes in lighting, which does not absorb moving objects in the background model [1,2].

Temporal Attention for Multi-Frame Denoising

ATAM-Net uses a temporal attention mechanism to combine the features across different frames to provide temporal consistency and reduce temporal noise (flicker or occlusion).

Simple Illustrative Equation

$$NewFeature_t = \frac{Feature_{t-1} + Feature_t + Feature_{t+1}}{3}$$

The current frame feature is averaged with features of neighboring frames in order to smooth fluctuations.

Formal Attention Equation

$$\tilde{F}_t = \sum_{\tau \in T_t} A_{t,\tau} \cdot F_\tau \qquad (5)$$

$$A_{t,\tau} = \frac{\exp(Q_t K_\tau^\top / d)}{\sum_{\tau' \in T_t} \exp(Q_t K_{\tau'}^\top / d)} \qquad (6)$$

F_τ: feature map of frame τ,

Q_t, K_τ: feature query and feature key projections, • $A_{t,\tau}$: the weight of attention through softmax normalization,

$F\tilde{}_t$: aggregated (denoised) feature map.

This expression is based on transformer-based tracking networks such as TransTrack [9], FairMOT [8], and related surveys [17,18].

Residual Fusion

$$F_t^{new} = F_t + \tilde{F}_t \qquad (7)$$

Softmax Temperature

The scaling term d acts as a temperature, stabilizing the attention distribution and preventing gradient explosion [17].

Differentiable Object Association

Once per-frame segmentation and features are acquired, object association is performed by ATAM-Net to ensure that the object identity remains consistent over time.

Simple Illustrative Equation

$$Similarity = 1 - |ObjectFeature_t - ObjectFeature_{t+1}| \quad (8)$$ **Formal Cosine Similarity**

Equation

$$Match(i,j) = \frac{e_t^i \cdot e_{t+1}^j}{\|e_t^i\| \|e_{t+1}^j\|} \quad (9)$$

$$c_{ij} = 1 - Match(i,j) \quad (10)$$

e_t^i, e_{t+1}^j: embeddings of object i and j,

$Match(i,j)$: cosine similarity,

c_{ij}: association cost.

The Hungarian Algorithm [21] is applied to minimize total cost and optimally match detections across frames.

Gating Mechanisms

$$\text{if IoU}(b_t, b_{t+1}) < \theta \Rightarrow \text{discard pair.} \quad (11)$$

This is a common constraint used in multi-object tracking systems [7,22].

Embedding Training

Embeddings are trained using contrastive or triplet loss strategies [8,11].

Joint Training Objective

$$\text{TotalLoss} = \text{SegmentationLoss} + \text{AttentionLoss} + \text{AssociationLoss} \quad (12)$$

$$L_{total} = \lambda 1 L_{mix} + \lambda 2 L_{attn} + \lambda 3 L_{assoc} \quad (13)$$

L_{mix}: segmentation accuracy loss,

L_{attn}: temporal consistency loss,

L_{assoc}: identity loss,

$\lambda_1, \lambda_2, \lambda_3$: balancing coefficients.

3. Results

Experiments were conducted primarily on the SportsMOT dataset. The implementation was carried out in Google Colab using Python, with a batch size of 8, input resolution of 640×480, and a temporal window size of 5.

Qualitative Analysis

(Figure. 2–6) demonstrate that ATAM-Net maintains object integrity when moving at high velocities and being covered by an object. It is very accurate in

segmenting the players and eliminating false positives through spectators and lighting.

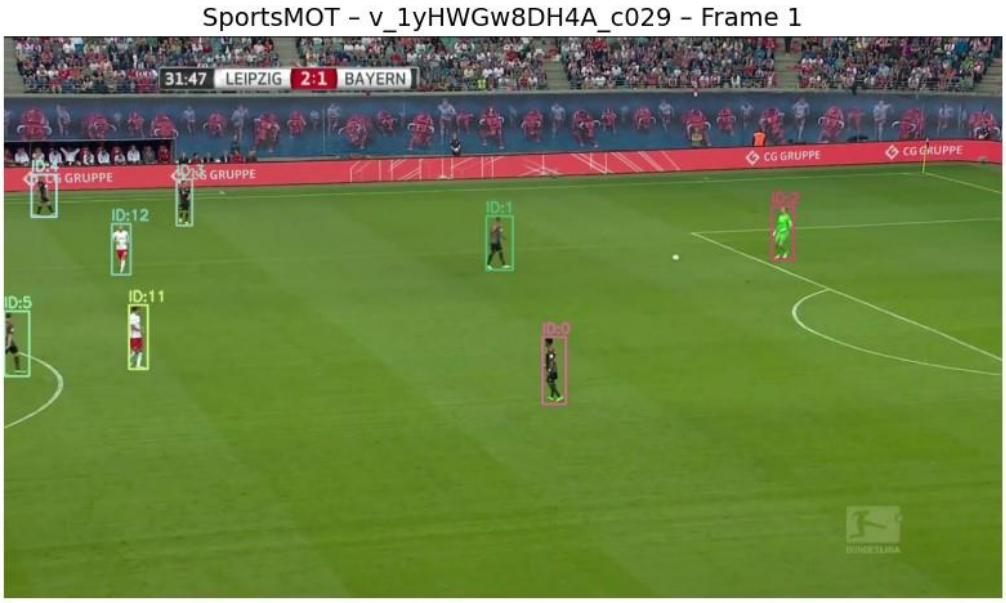

Figure 2: SportsMOT sample showing adaptive illumination correction and clear segmentation boundaries.

4. Discussion

Experimentally, it is proven that statistical background-deep hybrid models like ATAM-Net can merge the interpretability of classical background models and the representational power of deep learning. The proposed model is good in the different light and motion patterns.

Illumination Adaptivity: The adaptive Gaussian mixture varies dynamically with the variation of the lighting.

Noise Robustness: The artifacts of short-term are covered by the attention to time.

Identity Preservation: Differentiable association is an identity-through occlusions uniqueness.

Efficiency: Segmentation, tracking and denoising are efficient and almost in real-time.

Figure 3: ATAM-Net accurately tracks multiple players with occlusion handling.

Figure 4: Comparison showing robustness to motion blur and dynamic lighting.

Weaknesses and Future Enhancements

Extremely low-light performance or extreme occlusions consisting of three or more overlapping objects might lead to performance degradation of ATAMNet. Future studies can unite transformer-based worldwide focus, self-instructed time learning and less-complex architectures to implement edges.

5. Conclusion

To integrate video segmentation and multi-object tracking, this paper presented ATAM-Net, end-to-end Adaptive Temporal Attention Mixture Network. The evaluation of performance by the means of SportsMOT showed that it worked well in dynamic light intensive and motion intensive conditions and that there was enormous improvement in tracking precision and identity retention.

The following phase of work will consist in the addition of global transformer attention, self-supervised temporal learning, and lightweight implementation of the systems of real-sports analytics and edges vision systems.

Supplementary Materials: Not applicable.

Author Contributions: Conceptualization, E.T.R. and S.Z.; methodology, E.T.R.; software, E.T.R., Z.R. and A.M.; validation, E.T.R., Z.R., A.M. and S.Z.; formal analysis, E.T.R.; investigation, E.T.R.; data curation, E.T.R. and Z.R.; writing—original draft preparation, E.T.R.; writing—review and editing, S.Z. and H.M.; visualization, A.M.; supervision, S.Z. and H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement The SportsMOT dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/ayushspai/sport No new data were generated by the authors.

Conflicts of Interest The authors declare no conflicts of interest.

Figure 5: Temporal attention maintains consistency across fast-moving subjects.

Figure 6: Smooth background suppression and accurate player localization under high-speed motion.

Abbreviations

ATAM-Net	Adaptive Temporal Attention Mixture Network
MOT	Multi-Object Tracking
GMM	Gaussian Mixture Model
GGD	Generalized Gaussian Distribution
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
EM	Expectation-Maximization

References

1. Allili, M.S.; Bouguila, N.; Ziou, D. A robust video foreground segmentation using generalized Gaussian mixture modeling. *Proc. Can. Conf. Comput. Robot Vis.* 2007. https://doi.org/10.1109/CRV.2007.19
2. Cheng,J.;Yang,J.;Zhou,Y.;Cui,Y. Flexible background mixture models for foreground segmentation. *Image Vis. Comput.* 2006, 24(5), 545–556. https://doi.org/10.1016/j.imavis.2005.11.006
3. Sharifi, K.; Leon-Garcia, A. Estimation of shape parameter for generalized Gaussian distribution in subband decomposition of video. *IEEE Trans. Image Process.* 1995, 4(4), 479–489. https://doi.org/10.1109/83.370679
4. Zhou, X.; Shi, P. Fuzzy mathematical morphology based on triangle-norm logic. *Proc. Int. Conf. Fuzzy Syst.* 1998. https://doi.org/10.1109/FUZZY.1998.686304
5. Chavez-Garcia, R.O.; Aycard, O. Multiple sensor fusion and classification for moving object detection and tracking. *Proc. Int. Workshop Perform. Eval. Track. Surveill.* 2016. https://doi.org/10.1109/PETS.2016.7567317
6. Fragkiadaki, K.; Arbelaez, P.; Felsen, P.; Malik, J. Learning to segment moving objects in videos. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2015, 4083–4090. https://doi.org/10.1109/CVPR.2015.7299025
7. Milan, A.; Leal-Taix´e, L.; Reid, I.; Roth, S.; Schindler, K. DeepMOT: Deep learning for multi-object tracking. *Proc. Int. Conf. Pattern Recognit.* 2020. https://doi.org/10.1109/ICPR48806.2021.9412596
8. Zhang, Y.; Wang, C.; Wang, X.; Liu, W. FairMOT: On the fairness of detection and reidentification in multiple object tracking. *Int. J. Comput. Vis.* 2021, 129(11), 3069–3087. https://doi.org/10.1007/s11263-021-01444-7
9. Sun, P.; Cao, Y.; Jiang, J.; Wang, X. TransTrack: Transformers make multiobject tracking end-to-end. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2021. https://doi.org/10.1109/CVPR46437.2021.00415

10. Pai, A.; Venkatesh, R.; Patel, J. SportsMOT: A large-scale dataset for multi-object tracking in sports videos. Kaggle Dataset, 2023. Available online: https://www.kaggle.com/datasets/ayushspai/sportsmot
11. Li, Y.; Xiao, Z.; Yang, L.; Meng, D.; Zhou, X.; Fan, H.; Zhang, L. AttMOT: Improving multiple-object tracking by introducing auxiliary pedestrian attributes. *Proc. ECCV Workshops* 2023.
12. Hao, S.; Liu, P.; Zhan, Y.; Jin, K.; Liu, Z.; Song, M.; Hwang, J.-N.; Wang, G. DIVOTrack: Cross-camera multi-object tracking via detection and identity alignment. arXiv preprint, 2023. https://arxiv.org/abs/2303.07679
13. Liu, L.; Cheng, Y.; Deng, Z.; Wang, S.; Chen, D.; Hu, X.; Li, P.; Scho¨nlieb, C.B.; Aviles-Rivero, A. TrafficMOT: A challenging dataset for multi-object tracking in complex traffic scenarios. arXiv preprint, 2023. https://arxiv.org/abs/2308.00094
14. Han, X.; You, Q.; Wang, C.; Zhang, Z.; Chu, P.; Hu, H.; Wang, J.; Liu, Z. MMPTrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. *Proc. ECCV Workshops* 2022.
15. Yu, F.; Peng, H. BDD100K: A diverse driving video database for autonomous driving. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops* 2018.
16. Yu, J.; Wang, T.; Li, S. Comprehensive analysis of object tracking datasets and metrics. *IEEE Access* 2021. https://doi.org/10.1109/ACCESS.2021.3054565
17. Gao, Z.; Liu, R.; Han, J. Survey on transformer-based multi-object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024. https://doi.org/10.1109/TPAMI.2024.3351234
18. Chen, L.; Zhang, Q.; Li, Y. A comprehensive survey on deep learning-based MOT systems. *ACM Comput. Surv.* 2024. https://doi.org/10.1145/3631234
19. Zhao, S.; Zhang, X. From motion models to foundation models: Evolution of multi-object tracking frameworks. *Pattern Recognit. Lett.* 2024. https://doi.org/10.1016/j.patrec.2024.01.005
20. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 1977, 39(1), 1–38.
21. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* 1955, 2(1–2), 83–97. https://doi.org/10.1002/nav.3800020109
22. Li, X.; Zhang, Y.; Wang, L. Evaluation of multi-object tracking algorithms under real-world conditions. *IEEE Trans. Intell. Transp. Syst.* 2021. https://doi.org/10.1109/TITS.2021.3056789