Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.24683

Moving Object Tracking Using Context-Aware Attention Transformer

Yimeng Wang

School of Cyber Security and Computer / Department of Computer Teaching, Hebei University, Baoding, China
2545250732@qq.com

Abstract. In video content analysis, accurate tracking and recognition of objects is a complex task. Current research has primarily focused on the development of complex scenes and fast-moving targets. Yet, there are challenges of small objects, long time-series dependencies, and object occlusion. In this paper, we propose the Intelli-context transformer to detect objects in a dynamic environment. Addressing this challenge, attention mechanisms, contextual information, and semantic information are integrated into Intelli-Context Transformer to enhance the accuracy of video object tracking. Intelli-Context Transformer employs an end-to-end training approach and incorporates a Contextual Spatiotemporal Attention Module, which dynamically adjusts the focus on different information to improve recognition accuracy. The proposed method is capable of capturing and analyzing the spatiotemporal features of a single target in videos in real time, effectively handling tracking tasks in complex scenes. Compared with state-of-the-art methods, Intelli-Context Transformer demonstrates its strong generalization capability in video object recognition. This research provides an efficient and reliable approach for dynamic target tracking in complex scenes and offers technical support for functions such as behavior analysis and anomaly detection, contributing to the development of intelligent video surveillance and navigation.

Keywords: Object Detection, Object Tracking, Transformer Architecture, Attention mechanism, Deep Learning

1. Introduction

Object tracking, as an important branch of computer vision, aims to analyze videos to identify and track objects belonging to one or more categories. In recent years, it has received increasing attention and importance. Traditional object detection aims to identify objects within an image and generate a bounding box to mark the objects of interest. Digital video, however, is a collection of static image frames arranged in a certain temporal sequence. Each frame, as the smallest unit, contains RGB two-dimensional image information. Unlike traditional image object detection, object tracking emphasizes finding the trajectory of an object's position within a sequence [1]. The integration of information across frames, background information of the target, and intrinsic features of the object itself is referred to as context information [2]. The selection of video frames and the

DOI: 10.54254/2755-2721/2025.24683

handling of context information play a crucial role in improving the accuracy of video object detection.

Significant progress has been made in object detection over the past decade, from early CNN-based methods to current approaches such as ResNet [3]. Object tracking, as a subsequent task, includes both traditional and modern optical flow methods (e.g., FGFA), YOLO, and deep learning-based multi-object detection approaches [4]. However, existing object tracking faces challenges such as track creation, re-identification (ReID) [5], and complex detail expression [6], which make video object detection in real-world applications more challenging. The introduction of Transformer architecture has been a recent significant advancement, showing strong robustness in handling variable-length inputs and target occlusion..

The Self-Attention mechanism of the Transformer can capture dependencies between features on a global scale, unlike CNNs, which are limited to local receptive fields. This global perceptual ability is especially important for tasks involving long-range dependencies [7]. Due to the transformer's excellent capability in handling sequential data, it has been applied to object tracking and has demonstrated good performance in visual target tracking [8]. In object tracking, the design of dynamic and static templates aims to address the issue of how the object changes and is effectively tracked throughout the video sequence. The dynamic template is used to capture features under dynamic changes, while the static template captures the fixed parts of the target. By combining the contextual information from both, the Transformer model can update the dynamic changes of the target in each frame while maintaining its static features for better object matching. Subsequently, Transformer-based improved models began to emerge, such as SwinTrack [9], which improved input image size and proposed a local window self-attention mechanism to reduce computational complexity, and MixFormer [10], which adopted a more efficient hybrid mechanism that allows the model to share information across multiple modalities. These networks further optimized feature extraction, temporal modeling, and matching strategies, making Transformer a mainstream approach in the field of object tracking.

Despite significant progress in video object tracking, object detection and tracking remain major challenges. The root cause of the problem lies in attention allocation for contextual information and instability in long video object tracking. To address these issues, we propose a novel adaptive attention allocation strategy, Intelli-Context Transformer (ICT), an end-to-end strategy that leverages a Transformer-based Contextual Spatiotemporal Attention Module (CSAM) to improve object detection accuracy. It can adaptively adjust attention scores for object-related information, enhancing both computational efficiency and accuracy.

The main contributions of this study are as follows:

• We propose an adaptive attention allocation strategy to enhance the accuracy and stability of video object tracking.

• We introduce the CSAM based on the Transformer architecture, which strengthens the model's ability to capture spatiotemporal context information in video sequences.

• We evaluate proposed method on Single Object Tracking (SOT) challenge datasets. The evaluation results show significant advantages in both tracking accuracy and computational efficiency of our method.

2. Related work

Video object recognition is a critical task in computer vision, aiming to identify and track target objects from video sequences. Over the past decade, video object recognition algorithms have made significant progress, largely due to the rapid development of deep learning. In 2015, Shaoqing Ren

DOI: 10.54254/2755-2721/2025.24683

et al. proposed Faster R-CNN based on convolutional neural networks [11], which achieved high detection accuracy but had slower speed and was primarily used for static object detection. In 2016, Bewley et al. introduced the Simple Online and Realtime Tracking (SORT) algorithm [12] for dynamic tracking. This algorithm combines a Kalman filter and the Hungarian algorithm for real-time object tracking and replaces the detection results computed by Faster R-CNN with those obtained using Aggregated Channel Features (ACF) [13].

In 2017, the introduction of Graph Convolutional Networks (GCN) [14] by Thomas Kipf and others marked an expansion of convolutional neural network applications, opening a new chapter in the field of Graph Neural Networks (GNNs). In the same year, Ashish Vaswani et al. proposed the Transformer architecture [15], which discarded traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), using self-attention mechanisms to model the relationships between input data, particularly the long-range dependencies in sequential data. In 2018, Zhang et al. proposed a tracking method based on the Spatio-Temporal Graph Convolutional Networks (ST-MAP) [16], which focused on the temporal continuity of object behaviors. ST-MAP treated video frames as nodes in a graph and connected adjacent frames using graph convolution operations, thereby incorporating more temporal contextual information to improve object detection performance.

Also in 2018, Joseph Redmon et al. introduced YOLOv3 [17], which improved upon the shortcomings of the previous versions in multi-scale feature detection. It adopted the concept of Feature Pyramid Networks (FPN) [18] and used Darknet-53 as the backbone network [19] to extract more features, significantly improving the model's accuracy. In 2020, Nicolas Carion et al. introduced Detection Transformers (DETR) [20], which was the first application of Transformers in video object detection. By utilizing the self-attention mechanism, DETR performed global context modeling, laying the foundation for subsequent applications of Transformers in video object tracking..

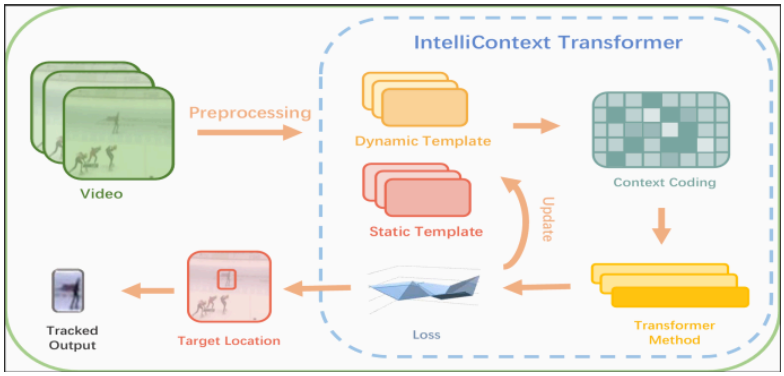

Figure 1: Overall framework of intelli-context transformer

3. Methodology

As shown in Figure 1, this work develops an innovative video object detection and tracking method by utilizing cutting-edge technology. It integrates spatial and temporal information of moving objects through a self-attention module to enhance tracking accuracy. The model incorporates more contextual information into tokens and deeply learns the target's appearance and motion patterns, optimizing the robustness and accuracy of object tracking.

DOI: 10.54254/2755-2721/2025.24683

3.1. Data preprocessing

The preprocessing of video frame data includes grayscale conversion, image masking, and dynamic-static image fusion. The color video is converted into grayscale to reduce the data dimensions. Image masking techniques are applied to enhance the edge information, improving the clarity of object contours and recognition accuracy. To fully utilize the temporal information in the video frames, the frames are divided into dynamic image D and static image V. Based on the tracking box from the previous frame, the search region X is obtained. This fusion strategy allows the model to simultaneously capture the spatial features and temporal dynamics of the target, enabling more accurate object tracking in complex video scenes.

3.2. Intelli-context transformer

This study proposes an innovative Transformer architecture, called Intelli-Context Transformer (ICT), aimed at improving the accuracy of video object tracking by integrating attention mechanisms, contextual information, and semantic cues. The core of ICT lies in its adaptive attention allocation strategy, which dynamically adjusts the attention to different information to adapt to the complexity of the video content. The ICT model is trained end-to-end, combining position prediction errors and feature alignment scores to ensure that the model can effectively learn to track objects across different video sequences.

ICT is based on the standard transformer architecture but introduces the CSAM. The self-attention mechanism in Transformer is as follows:

$$Attention\left(Q,K,V\right)=softmax\left(QK^{T}/\sqrt{d_{k}}\right)V \tag{1}$$

where Q、K and V represent the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vector. This mechanism enables ICT to weigh the importance of different tokens according to their relevance to the tracking task, focusing on the most informative features. This mechanism computes the dot product of the query and all keys, divides by d_k to prevent gradient vanishing or explosion, and applies the softmax function to obtain attention weights, which are used to weight the value matrix V to generate the weighted output.

In ICT, we perform feature extraction through 12 transformer layers, each followed by a residual connection after the MLP to accelerate training speed and improve model generalization. To further reduce the computational load and enhance the model's speed, we prune the computed tokens by calculating the similarity matrix of the tokens and retaining the top N as the pruned tokens. The importance score calculation is as follows:

$$\omega=\phi\left(\text{softmax}\left(Q_{s}K_{x}^{T}/\sqrt{d_{k}}\right)+\text{softmax}\left(Q_{d}K_{x}^{T}/\sqrt{d_{k}}\right)\right)\in\mathbb{R}^{1\times N_{x}} \tag{2}$$

where Q_s and Q_d represent the query vectors of the static and dynamic templates, K_x is the key vector of the search region, and ϕ denotes summing the attention matrix. N_x is the number of tokens in the search region. Finally, we retain the top k elements of the search region tokens with the highest importance scores.

Additionally, ICT normalizes after the LayerNorm (LN) layer to accelerate training speed. For the pruned tokens, we apply zero-padding and flatten them into a 5x5 tensor for further processing. The ICT tracking head consists of a Score Head, Offset Head, and Size Head, responsible for

DOI: 10.54254/2755-2721/2025.24683

calculating the tracking score, optimizing the position on the score map, and adjusting the offset of the bounding box size. Through this structure, ICT effectively captures the dynamic changes of the target and maintains stable tracking performance in complex video environments.

Contextual Spatiotemporal Attention Module

The CSAM module calculates the attention weight for each token as follows:

$$\alpha_i = \exp(e_i) / \sum_{j-1}^{N} \exp(e_i) \tag{3}$$

where e_i is the unnormalized attention score of the i token, and N is the total number of tokens. This calculation allows the model to focus more on regions containing target-related information when processing video frames.

The operation for fusing semantic and contextual features is represented as:

$$Fout = \sigma(W_1 F_{semantic} + W_2 F_{context}) \tag{4}$$

where $F_{semantic}$ and $F_{context}$ are the semantic and contextual features, W_1 and W_2 are learnable weight matrices, and σ is the activation function that introduces nonlinearity.

4. Experiment

4.1. Datasets

GOT-10k is a large-scale generic object tracking dataset, containing over 10,000 video sequences, with each sequence averaging 300 frames, totaling more than 3 million frames. The dataset provides precise annotations for the target bounding boxes and covers a wide range of scenes, object types, and challenges, such as deformation, occlusion, and lighting changes, making it ideal for evaluating and comparing the performance of object tracking algorithms.

Due to the common limitation in temporal sampling of traditional object tracking models, we sampled the dataset every 200 frames and divided the sampled video frames into dynamic and static images through preprocessing methods, incorporating more contextual information to improve tracking accuracy.

4.2. Experimental setup

1) Evaluation Metrics: To evaluate the tracking accuracy, we used the following matching measures between the predicted target bounding boxes and the ground truth boxes.

a) Average Overlap (AO)

$$AO = \frac{1}{N} \sum_{i=0}^{N} IoU_i \tag{5}$$

This metric computes the average overlap between the predicted target box and the ground truth box for each frame, where IoU_i represents the IoU value for the i frame, and N is the total number of frames in the video sequence.

b) Area Under the Curve (AUC)

$$AUC = \int_0^1 SR(IoU) \tag{6}$$

DOI: 10.54254/2755-2721/2025.24683

This metric plots the success rate (SR) against the IoU threshold (> 0.5) and calculates the area under the curve to measure the overall performance of the model on the test set.

In quantifying the results, the following metrics were used to determine the model's accuracy:

c) Success Rate (SR)

$$SR = (Number\ of\ Successful\ Frames / Total\ Number\ of\ Frames) \times 100\% \tag{6}$$

This metric measures the proportion of frames in which the tracking model successfully tracks the object within the video sequence.

2) Implementation Details: All experiments are conducted on the Nvidia GeForce 2080Ti GPU using the PyTorch framework. All models are trained for 300 epochs, and the batch size is set to 32. Each training period samples 60,000 frames, and each validation period samples 10,000 frames. Training results are printed every 50 steps. The learning rate is initialized to 0.0004 with a step decay factor of 0.1, and the AdamW is used. The learning rate decay is enabled after 240 epochs. The loss function uses IoU loss and L1 loss with weights of 2.0 and 5.0, respectively.

4.3. Analysis results

In this study, the proposed ICT is compared with the two methods SwinTrack and AQATrack[21]. SwinTrack uses Swin Transformer as the backbone network and has certain performance in target tracking tasks. AQATrack is a Transformer-based target tracking method. By comparing these two methods on the GOT-10K dataset, the ability of the ICT model in terms of tracking accuracy and model efficiency can be comprehensively evaluated.

Experimental results are shown in Table 1. Proposed ICT shows excellent performance on the GOT-10K dataset. Its average overlap rate (AO) reaches 74.1%, 3.2% and 0.3% higher than SwinTrack and AQATrack, respectively. In terms of success rate metrics, ICT's $SR_{0.5}$ is 84.3% and $SR_{0.75}$ is 72.4%, both better than SwinTrack and AQATrack. This suggests that the ICT model has higher accuracy and robustness in the target tracking task, and can better handle challenges such as deformation, occlusion, and lighting changes of the target.

Table 1: Experimental results on GOT-10K dataset

Method	AO (%)	$\boldsymbol{SR}_{0.5}$ (%)	$\boldsymbol{SR}_{0.75}$ (%)
SwinTrack	70.9	81.2	64.9
AQATrack	73.8	83.2	72.1
ICT (ours)	74.1	84.3	72.4

Table 2: Comparison of architecture

Method	Backbone	Number of Parameters (M)	Speed(fps)
SwinTrack	Swin Transformer	22.7	98
AQATrack	Transformer	72	67.6
ICT (ours)	Transformer	48.9	94.2

Table 2 compares the architectures of each model. The ICT uses Transformer as the backbone, with a parameter number of 48.9M, which is less than 72M of AQATrack, while maintaining a high

DOI: 10.54254/2755-2721/2025.24683

speed, reaching 94.2fps, close to SwinTrack's 98fps. This shows that the ICT model has achieved a good balance between model efficiency and performance and is more practical.

Fig.2 shows the success rate curves of each model on the GOT-10K dataset. It can be seen that the curve of the ICT is above the entire overlap threshold range, especially in the high overlap threshold area, with more obvious advantages. This further verifies the tracking capability of the ICT model in complex scenarios.

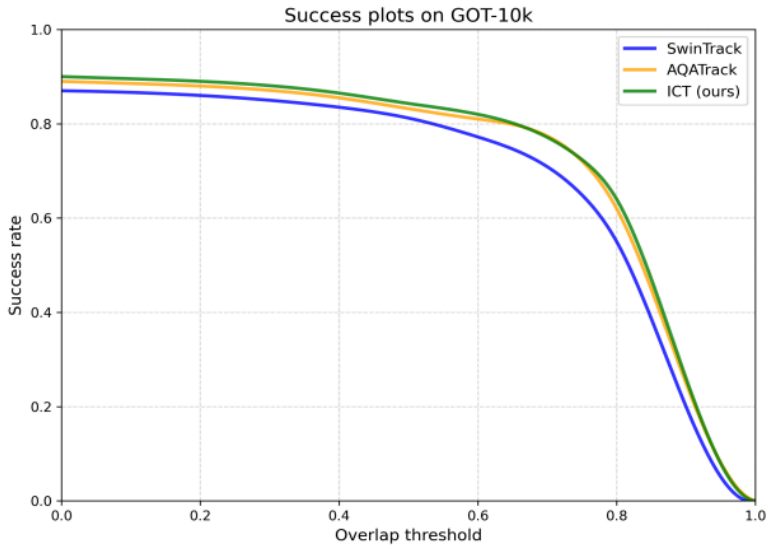

Figure 2: Success plots on GOT-10k

Fig.3 shows the attention graph comparison of CSAM on the GOT-10k dataset. CSAM accurately focuses on the target key areas and suppresses background interference through self-attention. As can be seen from the figure, the self-attention mechanism of the ICT model can focus more accurately on the target area and match the real labels more accurately.

Figure 3: The comparison attention map of CSAM on GOT-10k. The first and third columns are the ground truth, the second and fourth columns are the self-attention of CSAM

5. Conclusion and future works

The proposed ICT in this study is based on the Transformer architecture for video target tracking. It achieves performance superior to existing models through sampling every 200 frames and preprocessing of dynamic and static images on the GOT-10K dataset and combining appropriate evaluation indicators and training strategies. In terms of quantitative indicators, AO, $SR_{0.5}$ and $SR_{0.75}$ have all been improved to achieve a balance of efficiency and accuracy. ICT accurately focuses on the target through the self-attention mechanism to suppress background interference.

DOI: 10.54254/2755-2721/2025.24683

However, it is not robust enough in extremely complex occlusion scenarios. In the future, the adaptability of the model in complex scenarios will be enhanced, and the generalization ability and robustness of the model will be improved.

References

[1] O. Abdelaziz, M. Shehata, and M. Mohamed, “Beyond traditional single object tracking: A survey, ” 2024, arXiv: 2405.10439. [Online]. Available: https: //arxiv.org/abs/2405.10439

[2] X. Wang and Z. Zhu, “Context understanding in computer vision: A survey, ” Comput. Vis. Image Underst., vol. 229, p. 103646, Mar. 2023, doi: 10.1016/j.cviu.2023.103646.

[3] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, “A survey of modern deep learning based object detection models, ” 2021, arXiv: 2104.11892. [Online]. Available: https: //arxiv.org/ abs/2104.11892

[4] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, “Deep learning in video multi-object tracking: A survey, ” Neurocomputing, vol. 381, pp. 61–88, Mar. 2020, doi: 10.1016/j.neucom.2019.11.023.

[5] A. Kamboj, “The progression of transformers from language to vision to MOT: A literature review on multi-object tracking with transformers, ” 2024, arXiv: 2406.16784. [Online]. Available: https: //arxiv.org/ abs/2406.16784

[6] H. Ouyang, Q. Wang, Y. Xiao, Q. Bai, J. Zhang, K. Zheng, X. Zhou, Q. Chen, and Y. Shen, “Codef: Content deformation fields for temporally consistent video processing, ” 2023.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need, " in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS 2017), pp. 6000–6010, Dec. 2017.

[8] W. G. C. Bandara and V. M. Patel, “A transformer-based Siamese network for change detection, ” 2022, arXiv: 2201.01293. [Online]. Available: https: //arxiv.org/abs/2201.01293

[9] L. Lin, H. Fan, Z. Zhang, Y. Xu, and H. Ling, “SwinTrack: A simple and strong baseline for transformer tracking, ” 2022, arXiv: 2112.00995. [Online]. Available: https: //arxiv.org/abs/2112.00995

[10] Y. Cui, C. Jiang, L. Wang, and G. Wu, “MixFormer: End-to-end tracking with iterative mixed attention, ” 2022, arXiv: 2203.11082. [Online]. Available: https: //arxiv.org/abs/2203.11082

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks, " 2016, arXiv: 1506.01497. [Online]. Available: https: //arxiv.org/ abs/1506.01497

[12] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking, " in 2016 IEEE Int. Conf. Image Process. (ICIP), pp. 3464-3468, Sep. 2016, doi: 10.1109/ICIP.2016.7533003.

[13] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey, " Neurocomputing, vol. 381, pp. 61-88, Mar. 2020, doi: 10.1016/j.neucom.2019.11.023.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks, " 2017, arXiv: 1609.02907. [Online]. Available: https: //arxiv.org/abs/1609.02907

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need, " 2017, arXiv: 1706.03762. [Online]. Available: https: //arxiv.org/abs/ 1706.03762

[16] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, " in Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI), pp. 3634-3640, Jul. 2018, doi: 10.24963/ijcai.2018/505.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection, " 2016, arXiv: 1506.02640. [Online]. Available: https: //arxiv.org/abs/1506.02640

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection, " 2017, arXiv: 1612.03144. [Online]. Available: https: //arxiv.org/abs/ 1612.03144

[19] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement, " 2018, arXiv: 1804.02767. [Online]. Available: https: //arxiv.org/abs/ 1804.02767

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers, " 2020, arXiv: 2005.12872. [Online]. Available: https: //arxiv.org/abs/ 2005.12872

[21] J. Xie, B. Zhong, Z. Mo, S. Zhang, L. Shi, S. Song and R. Ji, “Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers, ” IEEE, 2024, DOI: 10.1109/CVPR52733.2024.01826.