RESEARCH ARTICLE

Boosting Multi-Object Tracking Performance with Advanced Attention Mechanism based on Transformer

Pimpa Cheewaprakobkit[1,*]

[1]Faculty of Information Technology, Asia-Pacific International University, Saraburi 18180, Thailand

*Corresponding author: Pimpa Cheewaprakobkit, pimpa@apiu.edu

Article Information

Article History:
Received: 16 August 2024
Revised: 25 October 2024
Accepted: 5 November 2024
Published: 19 October 2025

Keywords:
Advanced Attention Mechanism
Attention Mechanisms
Multi-Object Tracking
Transformer

Abstract

The Transformer architecture has been highly successful in natural language processing and is increasingly being applied to computer vision tasks, such as medical image analysis, traffic light monitoring, surveillance, and object tracking, due to its self-attention mechanism enabling global interactions between image patches. However, the quadratic complexity in time and memory limits its scalability for high-resolution images. To address this, we propose an advanced attention mechanism for multi-object tracking, incorporating Transposed Self-Attention (TSA) and Cross Patch Interaction (CPI) modules. TSA reduces computational complexity by capturing feature dependencies across the entire channel space instead of image patches, resulting in linear complexity relative to the number of patches. CPI enhances cross patch communication, improving the model's learning efficiency. Our method reduces computational costs by approximately 13% and achieves state-of-the-art performance, with a multi-object tracking accuracy of 72.8% on MOT17 and 63.4% on MOT20. This represents a 10.3% improvement over the baseline method on MOT17, while also reducing training time per epoch by nearly 7 minutes and increasing frame per second from 7 to 8. These results demonstrate the effectiveness and efficiency of our approach for multi-object tracking.

1. Introduction

Object tracking is considered one of the important tasks in computer vision and image processing (Kadam et al., 2024), and Multiple Object Tracking (MOT) is a category within object tracking (Park et al., 2021). MOT has gained increasing interest in recent years, and use in many applications such as automating robotics navigation (Gad et al., 2022), autonomous driving, aerial surveillance (Wang et al., 2024), and pedestrian tracking. Its main objective is to estimate the trajectories of multiple objects in a video sequence while maintaining their identities across frames (Zhang et al., 2022). The main challenges of object tracking are occlusion, illumination conditions, deformation, cluttered or textured background, and viewpoint variation.

Most MOT methods, such as those mentioned in (Fang et al., 2018; Mahmoudi et al., 2018), employ the tracking-by-detection paradigm and aim to address the issue through the use of two separate models. Initially, the detection model is used to localize the target objects with bounding boxes in individual video frames. Subsequently, an association model connects the set of detections across frames through re-identification (re-ID) to establish target trajectories (Zhang et al., 2021). Noteworthy advancements have been made in object detection, enhancing overall tracking accuracy. However, these two-step approaches face challenges in achieving real-time inference speed in environments with a high number of objects. This challenge arises due to the lack of feature sharing between the detection and association models, necessitating the independent application of re-ID models for each bounding box in the video (Fu et al., 2019). One-shot trackers have gained increased attention by introducing a single network to estimate objects and learn re-identification features (Wang et al., 2020). For example, Voigtlaender et al. (2019) incorporated a re-ID branch into Mask R-CNN to extract re-ID features for each proposal, reducing inference time by reusing backbone features for the re-ID network. However, de-

ISSN 3088-1803

spite improvements in inference speed, these one-shot trackers exhibit a significant drop in performance compared to the two-step models. While detection accuracy remains high, tracking performance experiences a considerable decline.

Lately, significant advancements have been observed in the realm of computer vision, with deep neural networks, particularly Convolutional Neural Networks (CNNs), playing a pivotal role (Chuangju, 2023; Zhao et al., 2024). To further enhance performance, Google AI has employed the Transformer architecture, renowned for its self-attention mechanisms in NLP (Vaswani et al., 2017), showcasing its potential for computer vision tasks. Subsequently, the Transformer migrated to computer vision, specifically image classification, detection, and segmentation (Dosovitskiy et al., 2020). Despite it achieves remarkable performance in several tasks, Transformers face significant challenges due to the quadratic time and memory complexity of their core self-attention operation relative to the input's length or the number of patches. Consequently, processing high-resolution feature maps and long sequences is slow. Although several strategies have been introduced to address this complex issue, none of the existing solutions are fully satisfactory, such as hierarchical Transformer architectures using shifted windows (Liu et al., 2021) to avoid non-overlapping local windows to reduce computation, but it may limit the interactions between patches within the same window. This limitation could potentially affect the model's ability to capture fine-grained details and long-range dependencies in the image. Zhu et al. (2020) incorporated the Transformer architecture into the Deformable DETR model, enabling the model to capture complex relationships between objects and their contexts while predicting objects in images of various sizes. The deformable attention mechanism in DETR helps reduce the complexity of global attention compared to conventional attention mechanisms. However, deformable attention requires dynamically computing the attention locations and scores for each query, which involves multiple iterations of attention computation. This process can become computationally intensive, particularly in high-resolution images or datasets with numerous objects, where many features must be tracked and processed simultaneously.

Existing multi-object tracking methods, such as those based on Transformers, face significant challenges due to the quadratic computational complexity of the global self-attention mechanism. This complexity makes these models inefficient when processing high-resolution inputs or large-scale datasets, leading to slow training convergence. To address these limitations, we propose a novel multi-object tracking framework that incorporates an advanced attention mechanism consisting of two key components: Transposed Self-Attention (TSA) and Cross Patch Interaction (CPI) modules. TSA focuses on feature dependencies across the channel space rather than individual patch comparisons, reducing computational complexity from quadratic to linear with respect to the number of patches. This significantly lowers the computational overhead and accelerates the training process, particularly for high-resolution inputs. Meanwhile, CPI enhances communication between patches by enabling the model to share information across patches more effectively. This leads to better feature interaction and representation learning, improving the model's ability to handle complex scenarios such as occlusions or closely interacting objects.

In summary, our main contributions are:

1) We propose a novel multi-object tracking framework that incorporates an advanced attention mechanism, effectively addressing slow training convergence by reducing computational complexity and improving feature interactions through TSA and CPI modules.

2) We compare the performance of our proposed method with state-of-the-art approaches on the MOT17 and MOT20 datasets, demonstrating notable improvements in tracking accuracy.

The subsequent sections of this paper are structured as follows. Section 2 reviews the related works. Section 3 introduces the proposed method and provides detailed insights into its formulation and implementation. Section 4 presents the experimental results. Finally, Sections 5 and 6 provide the discussions and conclusions.

2. Related work

2.1 Tracking-by-Detection

The Tracking-by-detection paradigm has emerged as a powerful approach, combining the strengths of state-of-the-art object detection algorithms with data association techniques to create coherent object trajectories. The process begins with the application of advanced object detection algorithms, such as Faster R-CNN (Region-based Convolutional Neural Network) (Ren et al., 2017), YOLO (You Only Look Once) (Nazir and Wani, 2023), or SSD (Single Shot Multibox Detector) (Magalhães et al., 2021). These algorithms identify and localize objects in individual frames, providing bounding boxes, class labels, and confidence scores. Object detection serves as the foundational step in the Tracking-by-detection pipeline, enabling the extraction of rich spatial and semantic information about objects within each frame. The next process of Tracking-by-detection lies in the association of detected objects across consecutive frames. Various data association techniques, including the Hungarian algorithm, Kalman filtering, and deep association methods, are employed to establish correspondences between detections and form coherent object trajectories. This step is pivotal in addressing the temporal aspect of tracking, allowing for

the seamless tracking of objects as they move through the video sequence. Some approaches, such as those that tackle data association as a graph optimization problem (Brasó and Leal-Taixé, 2020), consider each detection a graph node. However, separating detection and tracking can lead to higher computational costs and inefficiencies, particularly in complex scenes with occlusions. This separation also results in suboptimal performance when dealing with dense or overlapping objects, as detection and data association occur independently.

In contrast, our approach unifies detection and tracking, offering greater efficiency by reducing computational complexity and enhancing feature interaction. This integration improves tracking accuracy, especially in crowded or occluded environments, addressing common limitations in traditional tracking-by-detection methods.

2.2 Multi-Object Tracking (MOT)

Multi-Object Tracking (MOT) involves detecting and tracking objects of interest across a sequence of images or video frames. This requires associating the same object across frames, even when objects are in motion, occlude one another, or appear and disappear from the scene. This task becomes more challenging when objects are visually similar or moving rapidly. Traditional MOT methods rely heavily on separate stages of detection and data association, with techniques like feature-based tracking (Nazir and Wani, 2023), appearance-based tracking, and data association methods (Li et al., 2023). Feature-based tracking uses elements such as edges, corners, or regions with similar colors to track objects. Appearance-based tracking involves learning the appearance of objects and matching them across frames. Data association methods use probabilistic algorithms to associate objects across frames based on their location, motion, and appearance.

Moreover, MOT trackers often follow a tracking-by-detection approach (Bochinski et al., 2017), where objects of interest are detected independently in each frame, and data association is performed across frames using algorithms such as the Hungarian algorithm (Fang et al., 2018). Graph models are also used to represent the temporal connections and positions of objects to associate sets of detections into trajectories (Tian et al., 2019). Although these methods have improved performance, they are not entirely satisfactory, as they create ambiguity in complex scenes with occlusions and struggle to handle real-time tracking efficiently. Additionally, they can be computationally expensive, especially with large graphs.

2.3 Transformer

The Transformer architecture, first introduced by Vaswani et al. (2017), has brought significant advancements in natural language processing and has recently been extended to various computer vision tasks. It leverages a unique query-key mechanism and relies on attention mechanisms to process extracted deep features. This has allowed the Transformer to achieve remarkable success in tasks such as detection (Chen et al., 2016; Voigtlaender et al., 2019), segmentation (Zheng et al., 2020), and backbone construction (Wu et al., 2021). The Vision Transformer (ViT) architecture was proposed to utilize self-attention mechanisms for capturing spatial relationships and dependencies among image patches, enabling competitive performance in image classification benchmarks (Huo et al., 2023). However, there are a few limitations to consider. First, ViT can be computationally expensive compared to CNNs. Second, ViT may face challenges in capturing long-range dependencies effectively.

Our work builds on existing MOT approaches, particularly Transformer-based models. Unlike traditional methods that separate detection and tracking, our framework integrates both into a unified architecture. We leverage TSA to significantly reduce computational overhead and use the CPI module to enhance patch-level communication and handle spatial relationships between objects more effectively. This makes our approach more robust in dealing with occlusions and complex object interactions in dense scenes.

3. Proposed Method

We proposed an advanced attention mechanism aimed at improving the performance of the Transformer. The Transformer architecture consists of both an encoder and a decoder. To capture frame-level features, we utilize ResNet-50 as the underlying backbone. These frame features undergo encoding through a multi-head advanced attention mechanism within the Transformer encoder. This advanced attention mechanism is designed to capture meaningful relationships across different parts of the input. Subsequent to the advanced attention mechanism, we apply layer normalization, a feed-forward neural network, and residual connections. These elements serve to augment non-linearity and refine the encoded information. Ultimately, the output of the encoder is a sequence of vectors, each representing a context-aware embedding for the corresponding input token. This output comprehensively encapsulates both semantic meaning and positional information.

Within the Transformer decoder, we leverage information obtained from the encoded vectors to generate the output through the multi-head advanced attention. This advanced attention mechanism facilitates the decoder in selectively attending to different parts of the encoded input sequence, enabling the generation of contextually rich and accurate predictions. Subsequently, to produce bounding box and class predictions, we employ multilayer perceptron (MLP).

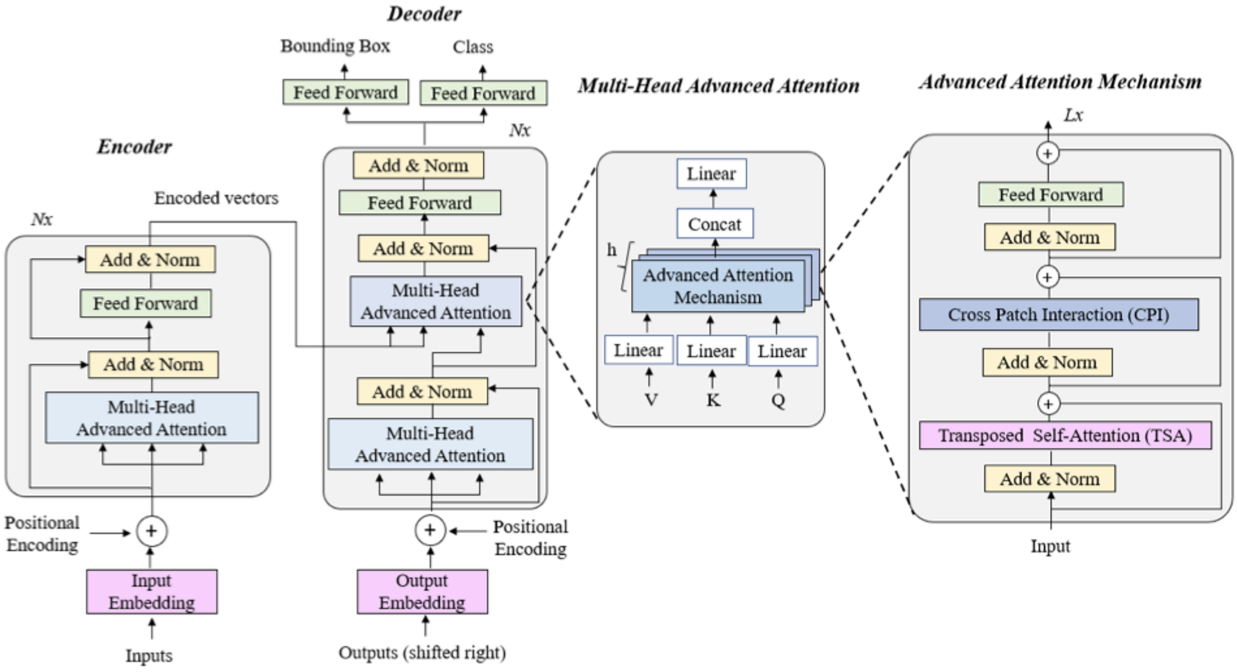

Figure 1. The proposed architecture.

These MLP serve as robust neural networks, extracting complex features and relationships from the decoder's output, contributing to the final predictions of bounding boxes and class labels with enhanced precision and expressiveness.

We have modified the conventional self-attention mechanism to an advanced attention mechanism, which comprises two modules: Transposed Self-Attention (TSA) and Cross Patch Interaction (CPI). The TSA module operates along the feature channel, while the CPI module enables communication among patches, allowing inputs to interact with each other. The proposed architecture is shown in Figure 1.

3.1 Multi-Head Advanced Attention Mechanism

The Multi-Head Advanced Attention in the Transformer divides the advanced attention into parallel subsets or heads, independently applying attention to each subset. This approach enables the model to capture a variety of relationships among input elements and learn distinct features by attending to different parts of the input sequence simultaneously.

The multi-head advanced attention is defined in Eq. 1 and Eq. 2. Where H_i represents the attention calculation for the $i-th$ head, while w_i^Q, w_i^K, w_i^V are learnable weight matrices applied to the input sequence to obtain query, key, and value projections specific to the $i-th$ head. The attention function computes the weighted sum of the value vectors based on the dot product similarity between the query and key vectors. W^O is the learnable weight matrix applied to the concatenated tensor to produce the final output.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, \ldots, H_{nh})W^O \quad (1)$$

$$H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

3.2 Conventional Self-Attention Mechanism

The Transformer relies on the conventional self-attention mechanism for input interactions, but its computational complexity is quadratic due to pairwise interactions between all patches. The time complexity is $\mathcal{O}(N^2 d)$ and the memory complexity is $\mathcal{O}(hN^2 + Nd)$, where h is the number of heads, N is the number of patches, and d is the embedding dimension. The conventional self-attention function (Vaswani et al., 2017) can be formulated as Eq. 3.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3)$$

here, Q, K, and V represent queries, keys, and values, and d_k refers to the dimensionality of K. The attention weights, obtained by scaled dot-product, are applied to V to produce the final output.

3.3 Advanced Attention Mechanism (AAM)

Conventional self-attention relies on heavy calculations, repetitive matrix multiplications. We propose mitigating this by reducing matrix dimensionality, allowing the

model to adjust its attention patterns and extract valuable key information. The advanced attention mechanism that we proposed consists of TSA and CPI modules as shown in Figure 1.

3.3.1 Transposed Self-Attention (TSA) Module

The primary motivation for introducing TSA is to reduce the high computational complexity of conventional self-attention, particularly in multi-object tracking tasks that involve processing high-resolution images. TSA operates across feature channels instead of image patches (El-Nouby et al., 2021), which significantly reduces computational costs by avoiding the quadratic scaling associated with traditional self-attention on patches, achieving linear complexity relative to the number of patches while maintaining strong feature representation.

The calculation of the advanced attention mechanism can be performed using Eq. 4. The definitions of queries Q, keys K, and values V are the same as those used in the self-attention mechanism within the Transformer.

$$\text{Attention}_{\text{AAM}}(Q, K, V) = \text{Softmax}\left(\frac{K^{\top}Q}{\sqrt{d_k}}\right)V \quad (4)$$

The attention output is a matrix $A_{\text{AAM}} \in R^{d_k \times d_q}$, which reduced the dimensionality and computations. The results in a time complexity of $\mathcal{O}\left(\frac{Nd_2}{h}\right)$ and the memory complexity of $\mathcal{O}\left(\frac{d_2}{h} + Nd\right)$, which address the quadratic complexity problem. Where h represents the number of heads, N denotes the number of patches, and d represents the dimensionality of the input embeddings.

3.3.2 Cross Patch Interaction (CPI) Module

To facilitate communication across patches, we utilized a CPI module following the TSA module. This two-stage process ensures improvements in both speed and accuracy. As shown in Table 1, our method increases MOTA by 10.3% and reduces training time by nearly 7 minutes per epoch, highlighting the practical benefits of our design compared to conventional attention mechanisms.

The CPI module consists of two depth-wise convolution layers of size 3×3, with Batch Normalization (BN) and Gaussian Error Linear Unit (GELU) activation functions in between (El-Nouby et al., 2021), as depicted in Figure 2. Importantly, the CPI module exhibits a low computational cost in terms of parameters.

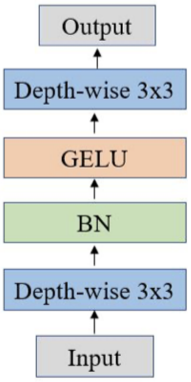

Figure 2. The Cross Patch Interaction module.

The utilization of depth-wise convolution layers within this module allows the model to capture interactions between local features across patches. This helps in incorporating contextual surrounding information, not just the patches themselves. Batch Normalization (BN) is employed to normalize the input to each layer, reducing internal covariate shift (distribution changes during training) and providing stable training. BN often enables faster learning rates, leading to quicker convergence during training. Gaussian Error Linear Unit (GELU) introduces non-linearity to the network, allowing it to model complex relationships within the data. GELU is known for its smooth activation function, and it helps in preserving gradient information during backpropagation. This activation function enhances the expressive power of the network, enabling it to capture intricate patterns and relationships.

4. Experimental Results

4.1 Datasets

Our model was trained on the standard splits of the MOT17 dataset (Dendorfer et al., 2020), which contains 14 different sequences of public places with annotated pedestrian bounding boxes. The training set consists of 15,948 frames, while the test set comprises 17,757 frames. We initialized our model with pre-trained weights from the COCO dataset. To enhance the model's generalization, we applied data augmentation techniques such as random horizontal flipping, cropping, and resizing, randomly transforming each frame. While the training set size remained unchanged, these augmentations introduced variability, improving the model's ability to generalize. All methods in our experiments were trained on the same dataset, ensuring consistency in training set size. We trained our model for 200 epochs using a batch size of 2 and an initial learning rate of 0.00001, which decreased by a factor of 10 after 10 epochs. The training was conducted on RTX 3090 GPUs.

4.2 Comparison

In Table 1, We present a comparison between our proposed method and the baseline, which is the Vanilla Transformer using conventional self-attention, both evaluated on the MOT17 dataset. Our approach replaces conventional self-attention with an advanced attention mechanism, resulting in several improvements. Firstly, we reduced the total number of parameters by 1.2 M. Additionally, this modification led to a reduction in training time by nearly 7 minutes per epoch, or approximately 13%. Furthermore, our proposed method exhibited a notable increase in the multi-object tracking accuracy (MOTA) during testing, achieving a boost of 10.3%, while the frame rate increased by 1 frame per second.

Table 1. Comparison between our proposed method and baseline method evaluated on MOT17 dataset.

Method	Total Parameters	Time/Epoch (min)	MOTA	FPS
Baseline	39.82 M	51:18	62.5%	7
Ours	38.62 M	44:30	72.8%	8

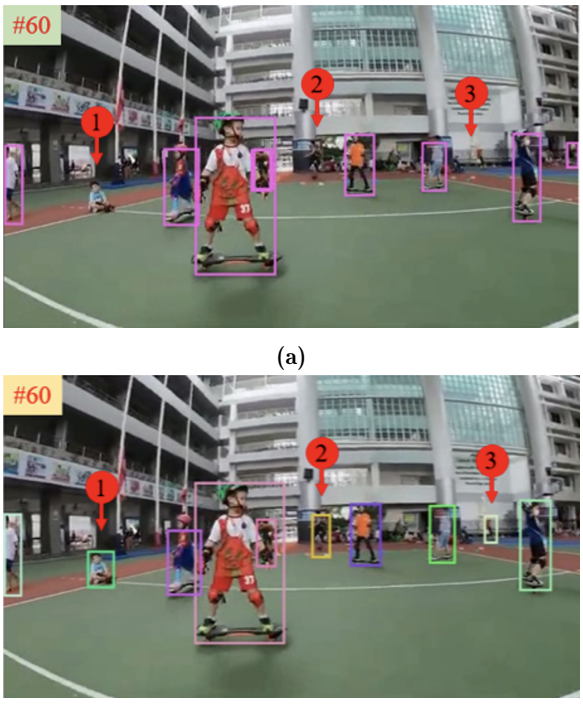

Figure 3. An illustration of tracking within a video frame: (a) Baseline method and (b) our method.

Figure 3 illustrates an example of the tracking results. In frame number 60 of the video clip, we compare the baseline method with our proposed method. Firstly, the baseline method is unable to track a boy seated on the floor (indicated as number 1), whereas our proposed method successfully tracks him. Secondly, the baseline method fails to track a person who is walking and wearing a sportwear of dark blue color, which blends with the background (indicated as number 2). In contrast, our proposed method effectively tracks this person. Thirdly, the baseline method struggles to track two people who are walking at the back due to their distance and small size, while our tracker accurately tracks them (indicated as number 3). The video available via this link: https://drive.google.com/drive/folders/14Wa0adz0ev0VddQxdH4o3kndenB8TTdi

In addition, we compare our proposed method with other methods, which are evaluated on MOT17 as shown in Table 2. We provide results based on the CLEAR MOT metrics (Bernardin Stiefelhagen, 2008), which are widely recognized as standard evaluation criteria for multiple target tracking algorithms. These metrics include MOTA (Multiple object tracking accuracy), IDF1 (Identity preservation score), MT (Mostly tracked targets), ML (Mostly lost targets), and ID Sw. (ID-switch).

In Table 2, our method outperforms other approaches on several key metrics, including MOTA, IDF1, and MT, which are crucial for assessing performance in complex tracking scenarios such as occlusion, clutter, and motion. MOTA accounts for missed targets, false positives, and ID switches. Our approach achieves the highest MOTA score of 72.8%, indicating superior overall tracking accuracy. This high score suggests that our model excels in detecting and tracking objects across frames, even in challenging conditions like cluttered environments or occlusion, showing that the model reliably tracks multiple objects with minimal errors.

IDF1 measures how well the model maintains object identities over time. Our high score of 70.4% demonstrates the model's ability to track identities in complex situations, like occlusion or fast motion, re-identifying objects when they reappear. This is crucial for applications like surveillance.

ID switches count how often the model incorrectly assigns a new identity to a tracked object. While our method has a higher ID switch count (5,121), the strong MOTA and IDF1 scores reflect robust overall performance. In crowded scenes, higher ID switches are expected, but our model compensates by maintaining accurate detection and minimizing false positives. The high number of ID switches is due to the global attention mechanism in our Transformer architecture, which covers a large area and improves its ability to detect multiple objects. However, in crowded scenes or when objects are closely clustered or occlude each other, this wide coverage can lead to misidentification. As objects temporarily overlap or pass each other, the model may switch their identities, resulting in more ID switches.

Table 2. The comparison between our method and other methods evaluated on MOT17.

Method	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	ID Sw. ↓
Tracktor++ (Bergmann et al., 2019)*	56.3	55.1	498	831	1,987
GSM (Bergmann et al., 2019)**	56.4	57.8	523	813	1,485
Baseline*	62.5	60.7	702	632	3,917
TubeTK (Pang et al., 2020)**	63.0	58.6	734	468	4,137
UTM (You et al., 2023)**	63.5	65.1	881	635	1,686
CTracker (Pang et al., 2020)**	66.6	57.4	758	570	5,529
TrajE (Girbau et al., 2021)**	67.4	61.2	820	587	4,019
CenterTrack (Girbau et al., 2021)*	67.8	64.7	814	579	3,039
TransCenter (Xu et al., 2023)*	68.8	61.4	867	564	4,102
TraDes (Wu et al., 2021)*	69.1	63.9	857	506	3,555
NCT (Zeng et al., 2023)*	69.5	68.5	1092	399	4,919
PixelGuide (Boragule et al., 2022)*	69.7	68.4	903	615	3,639
Ours*	**72.8**	**70.4**	**972**	**102**	**5,121**

Note: Arrows indicate whether higher (↑) or lower (↓) values are better; * indicates use of data augmentation; ** indicates no explicit mention of data augmentation.

Table 3. The comparison between our method and other methods evaluated on MOT20.

Method	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	ID Sw. ↓
IQHAT (He et al., 2022)*	57.1	57.7	507	249	1,875
LMOT (Mostafa et al., 2022)**	59.1	61.1	759	312	1,398
OCSORTpublic (Cao et al., 2023)**	59.9	67.0	478	330	554
MPTC (Stadler and Beyerer, 2021)*	60.6	59.7	635	208	4,533
TransCenter (Xu et al., 2023)*	61.0	49.8	601	192	4,493
RETracker (Kawanishi, 2022)*	62.4	53.0	605	192	3,804
Ours*	**63.4**	**57.5**	**642**	**173**	**3,133**

Note: Arrows indicate whether higher (↑) or lower (↓) values are better; * indicates use of data augmentation; ** indicates no explicit mention of data augmentation.

While the model maintains overall tracking accuracy, its tendency to switch identities may be problematic in applications where long-term identity tracking is critical, such as behavioral analysis. A high ID switch count could lead to confusion in situations requiring consistent identity tracking across sequences. In Table 3, our method outperforms existing approaches in overall tracking accuracy, achieving the highest MOTA (63.4%) and the lowest ML (173), making it highly reliable in crowded and occluded scenes. It also performs well in MT (642), demonstrating its ability to track objects through most of the sequence. However, it faces challenges in consistently preserving object identities (IDF1) and managing ID switches during complex interactions or occlusions, indicating areas where further improvements are needed, particularly in refining the model's identity management mechanisms.

5. Discussion

The experimental results demonstrate that the TSA module successfully reduces the computational complexity of traditional self-attention from quadratic to linear, significantly lowering overall computational overhead. This leads to faster training convergence, as evidenced by the results in Table 1, where our method shows a notable reduction in time per epoch compared to the baseline. Additionally, the CPI module enhances feature interactions between patches, improving the model's ability to handle challenging tracking scenarios, such as occlusions and closely interacting objects, which are common in multi-object tracking.

Tables 2 and 3 compare our method against state-of-the-art approaches on the MOT17 and MOT20 datasets. Our method achieves the highest MOTA scores of 72.8% on MOT17 and 63.4% on MOT20, demonstrating superior overall tracking accuracy. These results confirm that our approach effectively tracks objects in cluttered environments and during occlusions, aligning with our goal of improving tracking performance. Furthermore, the lowest ML values in both tables underscore the model's reliability in maintaining object tracks throughout the sequence.

While our method outperforms existing approaches in key metrics such as MOTA and ML, we recognize that there are areas for further improvement, particularly in

managing ID switches, as our method shows a higher ID switch count compared to some other approaches. This suggests that, despite the strengths of the TSA and CPI modules, additional refinement is needed in handling identity preservation during complex interactions and occlusions.

6. Conclusion

In this work, we proposed a novel multi-object tracking architecture based on an advanced attention mechanism, integrating Transposed Self-Attention (TSA) and Cross Patch Interaction (CPI) modules. Our method achieves state-of-the-art performance, with a MOTA of 72.8% on the MOT17 dataset and 63.4% on MOT20. These results highlight the effectiveness of our approach in improving tracking accuracy, reducing computational costs, and speeding up training time by optimizing the attention mechanism. Specifically, the TSA module significantly reduces computational complexity from quadratic to linear, while the CPI module enhances feature interaction across patches, improving tracking in complex scenarios like occlusions and crowded scenes. Overall, our method achieves a 10.3% accuracy improvement over conventional attention-based tracking models.

Beyond the technical contributions, our approach holds potential for real-world applications in various industries. In autonomous driving, for example, the improved accuracy and efficiency of our method can enhance object tracking for vehicles, pedestrians, and cyclists, even in challenging environments. Similarly, in surveillance, the ability to maintain identity consistency across frames could improve the monitoring of crowded areas, making the system more reliable for security purposes.

Looking ahead, several opportunities for future research exist. One promising direction is the integration of additional modalities, such as depth data or LiDAR, into the multi-object tracking pipeline. This could enhance the system's ability to understand spatial relationships between objects, providing better detection and tracking in 3D environments. Another area for improvement is addressing specific failure cases, such as high ID switch rates in crowded scenes. Future work could focus on refining identity preservation mechanisms or incorporating temporal consistency checks to reduce ID switches and improve object re-identification.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bergmann, P., Meinhardt, T., and Leal-Taixé, L. (2019). Tracking without bells and whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951. DOI: 10.1109/ICCV.2019.00103.

Bochinski, E., Eiselein, V., and Sikora, T. (2017). High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, page 1–6. IEEE. DOI: 10.1109/avss.2017.8078516.

Boragule, A., Jang, H., Ha, N., and Jeon, M. (2022). Pixel-guided association for multi-object tracking. *Sensors*, 22(22):8922. DOI: 10.3390/s22228922.

Brasó, G. and Leal-Taixé, L. (2020). Learning a neural solver for multiple object tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6246–6256. IEEE. DOI: 10.1109/cvpr42600.2020.00628.

Cao, J., Pang, J., Weng, X., Khirodkar, R., and Kitani, K. (2023). Observation-centric sort: Rethinking sort for robust multi-object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696. DOI: 10.1109/CVPR52729.2023.00934.

Chen, X., Li, Y., Li, Y., Yu, J., and Li, X. (2016). A novel probabilistic data association for target tracking in a cluttered environment. *Sensors*, 16(12):2180. DOI: 10.3390/s16122180.

Chuangju, W. (2023). Comprehensive survey of deep learning-based approaches for aerial visual tracking. *Journal of Optics*, 53(3):1906–1913. DOI: 10.1007/s12596-023-01357-w.

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., and Leal-Taixé, L. (2020). MOTChallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881. DOI: 10.1007/s11263-020-01393-0.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., and Jégou, H. (2021). XCiT: Cross-covariance image transformers. *CoRR*, abs/2106.09681.

Fang, K., Xiang, Y., Li, X., and Savarese, S. (2018). Recurrent autoregressive networks for online multi-object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 466–475. DOI: 10.1109/WACV.2018.00057.

Fu, J., Zong, L., Li, Y., Li, K., Yang, B., and Liu, X. (2019). Model adaption object detection system for robot. *CoRR*, abs/1911.02718.

Gad, A., Basmaji, T., Yaghi, M., Alheeh, H., Alkhedher, M., and Ghazal, M. (2022). Multiple object tracking in robotic applications: Trends and challenges. *Applied Sciences*, 12(19):9408. DOI: 10.3390/app12199408.

Girbau, A., Giró-i-Nieto, X., Rius, I., and Marqués, F. (2021). Multiple object tracking with mixture density networks for trajectory estimation. *CoRR*, abs/2106.10950.

He, Y., Wei, X., Hong, X., Ke, W., and Gong, Y. (2022). Identity-quantity harmonic multi-object tracking. *IEEE Transactions on Image Processing*, 31:2201–2215. DOI: 10.1109/tip.2022.3154286.

Huo, Y., Jin, K., Cai, J., Xiong, H., and Pang, J. (2023). Vision transformer (ViT)-based applications in image classification. In *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 135–140. DOI: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033.

Kadam, P., Fang, G., and Zou, J. J. (2024). Object tracking using computer vision: A review. *Computers*, 13(6):136. DOI: 10.3390/computers13060136.

Kawanishi, Y. (2022). Label-based multiple object ensemble tracking with randomized frame dropping. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 900–906. DOI: 10.1109/ICPR56361.2022.9956158.

Li, Z., Chen, J., and Bi, J. (2023). Multiple object tracking with appearance feature prediction and similarity fusion. *IEEE Access*, 11:52492–52500. DOI: 10.1109/ACCESS.2023.3279868.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9992–10002. IEEE. DOI: 10.1109/iccv48922.2021.00986.

Magalhães, S. A., Castro, L., Moreira, G., dos Santos, F. N., Cunha, M., Dias, J., and Moreira, A. P. (2021). Evaluating the single-shot multibox detector and YOLO, deep learning models for the detection of tomatoes in a greenhouse. *Sensors*, 21(10):3569. DOI: 10.3390/s21103569.

Mahmoudi, N., Ahadi, S. M., and Rahmati, M. (2018). Multi-target tracking using CNN-based features: CNNMTT. *Multimedia Tools and Applications*, 78(6):7077–7096. DOI: 10.1007/s11042-018-6467-6.

Mostafa, R., Baraka, H., and Bayoumi, A. (2022). LMOT: Efficient light-weight detection and tracking in crowds. *IEEE Access*, 10:83085–83095. DOI: 10.1109/access.2022.3197157.

Nazir, A. and Wani, M. A. (2023). You only look once - object detection models: A review. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1088–1095.

Pang, B., Li, Y., Zhang, Y., Li, M., and Lu, C. (2020). TubeTK: Adopting tubes to track multi-object in a one-step training model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6307–6317. DOI: 10.1109/CVPR42600.2020.00634.

Park, Y., Dang, L. M., Lee, S., Han, D., and Moon, H. (2021). Multiple object tracking in deep learning approaches: A survey. *Electronics*, 10(19):2406. DOI: 10.3390/electronics10192406.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149. DOI: 10.1109/tpami.2016.2577031.

Stadler, D. and Beyerer, J. (2021). Multi-pedestrian tracking with clusters. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–10. DOI: 10.1109/AVSS52988.2021.9663829.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635. DOI: 10.1109/ICCV.2019.00972.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, page 6246–6256. DOI: 10.1109/cvpr42600.2020.00628.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., and Leibe, B. (2019). MOTS: Multi-object tracking and segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7934–7943. IEEE. DOI: 10.1109/cvpr.2019.00813.

Wang, P., Wang, Y., and Li, D. (2024). DroneMOT: Drone-based multi-object tracking considering detection difficulties and simultaneous moving of drones and objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, page 7397–7404. IEEE. DOI: 10.1109/icra57147.2024.10610941.

Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2020). Towards real-time multi-object tracking. In *Computer Vision – ECCV 2020*, page 107–122. Springer International Publishing. DOI: 10.1007/978-3-030-58621-8_7.

Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., and Yuan, J. (2021). Track to detect and segment: An online multi-object tracker. *CoRR*, abs/2103.08808.

Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., and Alameda-Pineda, X. (2023). TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7820–7835. DOI: 10.1109/tpami.2022.3225078.

You, S., Yao, H., Bao, B.-k., and Xu, C. (2023). Utm: A unified multiple object tracking model with identity-aware feature enhancement. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21876–21886. DOI: 10.1109/CVPR52729.2023.02095.

Zeng, K., You, Y., Shen, T., Wang, Q., Tao, Z., Wang, Z., and Liu, Q. (2023). NCT: noise-control multi-object tracking. *Complex & Intelligent Systems*, 9(4):4331–4347. DOI: 10.1007/s40747-022-00946-9.

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*, page 1–21. Springer Nature Switzerland. DOI: 10.1007/978-3-031-20047-2_1.

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087. DOI: 10.1007/s11263-021-01513-4.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., and Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4). DOI: 10.1007/s10462-024-10721-6.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., and Zhang, L. (2020). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *CoRR*, abs/2012.15840.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable DETR: deformable transformers for end-to-end object detection. *CoRR*, abs/2010.04159.