METFormer: A Motion Enhanced Transformer for Multiple Object Tracking

Jianjun Gao*, Kim-Hui Yap*, Yi Wang*, Kratika Garg†, Boon Siew Han†

Email: gaoj0018@e.ntu.edu.sg, { ekhyap,wang_yi}@ntu.edu.sg, {gargkat, hanbon}@schaeffler.com

* School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

†Schaeffler Hub for Advanced Research at NTU, Singapore

***Abstract*—Multiple object tracking (MOT) is an important task in computer vision, especially video analytics. Transformer-based methods are emerging approaches using both tracking and detection queries. However, motion modeling in existing transformer-based methods lacks effective association capability. Thus, this paper introduces a new METFormer model, a Motion Enhanced TransFormer-based tracker with a novel global-local motion context learning technique to mitigate the lack of motion information in existing transformer-based methods. The global-local motion context learning technique first centers on difference-guided global motion learning to obtain temporal information from adjacent frames. Based on global motion, we leverage context-aware local object motion modelling to study motion patterns and enhance the feature representation for individual objects. Experimental results on the benchmark MOT17 dataset show that our proposed method can surpass the state-of-the-art Trackformer [21] by 1.8% on IDF1 and 21.7% on ID Switches under public detection settings.**

Keywords—Multiple Object Tracking, Motion Modeling, Tracking by Attention, Transformer

I. INTRODUCTION

Multiple object tracking (MOT) is a challenging and long-standing task. It aims to detect different objects and keep their trajectories, which means the task focuses on spatial and temporal domain object association. It is a basis for high-level computer vision tasks like pose estimation, action localization, and recognition. Meanwhile, it is widely applied to solve real-world problems, such as safeguarding in factories and autonomous driving.

Multiple object tracking can be categorized into tracking-by-detection methods, joint tracking and detection methods, and tracking-by-attention methods. Tracking-by-detection methods [1, 2, 3, 4, 5, 6, 7, 8] follow the concept of detecting objects with off-the-shelf detectors and tracking them with association algorithms in two stages. Joint tracking and detection methods [10, 12, 13, 14, 15, 16] detect and link objects simultaneously. Recently, transformer-based trackers or tracking-by-attention methods have presented a new paradigm for joint detection and tracking methods. Transformer-based object detectors [17, 18, 19, 20] have shown advanced performance by adopting transformer encoders, decoders, and object queries. Thus, it is intuitive to modify transformer-based object detectors into tracking-by-attention models [21, 22, 23] by introducing additional tracking queries besides original detection queries in transformer decoders, where the tracking queries aim to link objects that appear in the previous frame while the detection queries detect new-born objects.

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its IAF-ICP Programme I2001E0067 and the Schaeffler Hub for Advanced Research at NTU.

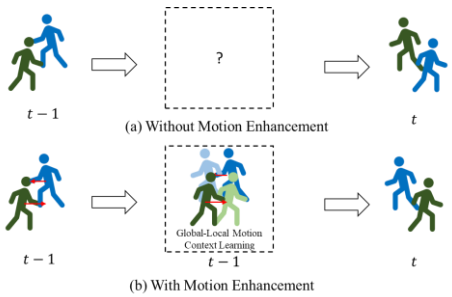

Fig. 1. Illustration of the motivation of our proposed METFormer. Existing transformer-based trackers are sensitive to complex conditions like occlusion due to lacking motion information and enhancement, as shown in (a). As for our METFormer it learns global and local motion information to mitigate the lacking of motion issues. With motion enhancement, METFormer is able to associate objects more accurately in two adjacent frames, as shown in (b).

Motion is an important cue for a tracker to localize and track objects accurately in upcoming frames. Tracking-by-attention methods have outperformed traditional joint tracking and detection methods. However, they leverage less motion modeling, which is essential for re-identification and association. The current transformer-based tracker also suffers from severe ID switch problems because of lacking motion information.

Based on Trackformer [21], we developed a model called METFormer. The motivation is demonstrated in Fig. 1. It introduces a global-local motion context learning technique to enhance its motion feature representation by difference-guided global motion learning and context-aware local motion modelling. As for difference-guided global motion learning, we propose a cross attention and difference feature pyramid network (DFPN) to learn global motion patterns from perspectives of temporal feature differences and correlations. Then, a guided squeeze and excitation module is applied to the global correlation features to generate the final global motion features with the difference features. For local motion extraction, we use a deformable cross-attention network to transfer object queries into motion queries and provide local motion context information for every detected object. After that, we update motion features to the object features for better associations in the next decoder layer for the next frame. An offset loss is also proposed to supervise the training process of the global-local motion learning technique.

We evaluate our METFormer on MOT datasets. Compared with the Trackformer [21] baseline under public detection settings, our METFormer improves the IDF1 and ID Switches by 1.8% and 21.7%, respectively. The experimental results from public detection demonstrate that our methods take advantage of the global-local motion context learning technique.

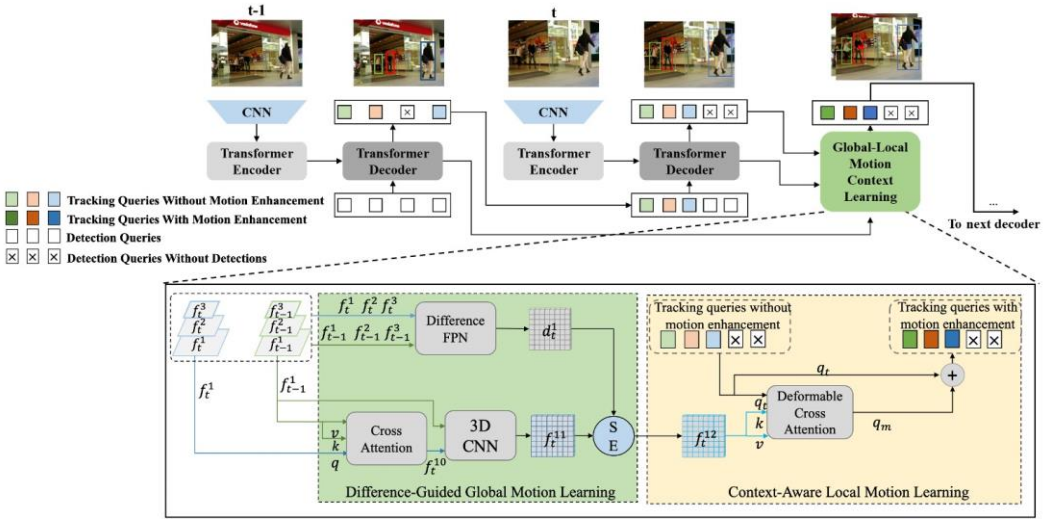

Fig. 2. The overall architecture of our proposed METFormer. The METFormer is developed with a novel global-local motion context learning technique to learn both global motion and local motion. For difference-guided global motion learning, our METFormer uses two methods to learn global motion features. The first method learns global motion information from temporal feature differences with a DFPN, and the second method learns from temporal correlations with a cross attention and 3D CNN. A guided squeeze and excitation module is proposed to excite global correlation features with difference features. For context-aware local motion modelling, Our METFormer utilizes a deformable cross attention to learn local motion information for every object query.

II. Related Work

A. Tracking by Detection

Many trackers adopt the concept of tracking by detection. The basic idea is to detect objects in adjacent frames with an off-the-shelf object detector followed by a tracking method to associate the same object in different frames. IOUTracker [1] first follows the paradigm and uses intersection over union (IoU) to connect the moving objects. IoU is an efficient measurement for associating the same object but can easily introduce false tracking. SORT [2] combines the detector with Kalman Filter (KF) to predict the motion of every detected object and update the tracking with the Hungarian Algorithm. DeepSort [3] introduces a re-identification (Re-ID) module to find the best-matched objects along with motion information. Recently, many methods [4, 5, 6, 33] have attempted to mitigate the weakness of the original SORT by associating both high-confidence and low-confidence detections, combining observations, introducing more Kalman Filter parameters and motion compensation, and using more accurate association criterion. Besides SORT-based methods, MOTDT [7] achieves real-time tracking by using a shared region-of-interest (ROI) feature. Besides, FairMOT [8] uses an anchor-free detector CenterNet [9], to obtain the ID embeddings for Re-ID.

B. Joint Detection and Tracking

Apart from tracking by detection, joint detection and tracking methods integrate tracking into the detection models. [10] proposed a tracker based on the RCNN network [11] to learn the motion and offset from region of interests (ROIs) from continuous frames and associate the tracklets by offsets. SiamMOT [12] follows a similar idea to [10] but explores how to extract more accurate motion information with an implicit and explicit motion module. Integrated detection [13] detects the objects from the current frame conditioned on the tracklets from the previous frame. Tracker [14] is more straightforward as it directly regresses the objects in the current frame from previous observations. CenterTrack [15] also uses CenterNet [9] as the detector. Different from FairMOT, CenterTrack learns the motion directly from the object heatmap. To learn better motion information, TraDeS [16] leverages a cost volume with correlation learning on CenterTrack. [34] changes the association cost matrix from center distance to IoU distance.

C. Tracking by Attention

Detection Transformer (DETR) [17] has presented a new paradigm for object detection. Naturally, tracking can also be tackled by DETR in a tracking-by-attention way. In DETR, the image will go through a Convolution Neural Network (CNN) and encoder layers. In the following decoder layers, DETR generates some random queries with position embeddings and sends them into the subsequent decoder layers. The final feed-forward network (FFN) and Hungarian matching algorithm are designated to regress and learn the final bounding boxes. However, training DETR requires a long convergence period because DETR calculates the correlations from object queries to the feature maps in a one-to-one manner. [18, 19, 20] utilize sparse features or prior knowledge to accelerate the learning speed. For tracking-by-attention methods, [21, 22, 23] share a similar concept by introducing tracking queries into Deformable DETR [18] in addition to detection queries. The tracking queries are responsible for associating objects in adjacent frames, while detection queries are responsible for new-born objects.

III. Methodology

A. Overall Architecture

The overall architecture of our proposed METFormer is illustrated in the upper part of Fig. 2. We use two adjacent frames as inputs, and they will go through a ResNet-50 [24] backbone and transformer encoder first. In the decoder, we also follow the design in Trackformer, which separates object queries into two parts: tracking and detection queries. The tracking queries are from the previous frames with tracked objects and aim to link the same objects in the current and previous frames. And the detection queries detect new-born objects. After the decoder layers, we introduce a global-local

motion context learning technique to enhance the motion representation for each object query. We also adopt an offset loss function to supervise the motion learning.

B. Global-Local Motion Context Learning

Based on Trackformer, we develop our METFormer with a global-local motion context learning technique to conduct both difference-guided global motion learning and context-aware local motion modelling, as shown in Fig. 2. As for the difference-guided global motion learning, we use two branches to learn global motion features. The Two branches learn global information from two aspects: differences and correlations. Once two global motion features are obtained, a guided squeeze and excitation module is proposed to excite global correlation features guided by difference features. As for local motion extraction, it learns and updates the local motion context information for every detected object. We use a deformable cross attention to project the positional features to motion features and update the motion features to positional features. More details will be given in the following subsections.

1) Difference-Guided Global Motion Learning: Difference-guided global motion learning consists of three parts, as shown in the green dashed box of Fig. 2: a cross attention with a 3D CNN, a difference FPN, and a guided squeeze and excitation module. Here we make use of the feature pyramid from the output of the previous decoder, but we only adopt the last three levels, which can be represented as f_t^i for current t frame from top to down in size of $H^i \times W^i \times C$, where $i \in [1, 2, 3]$.

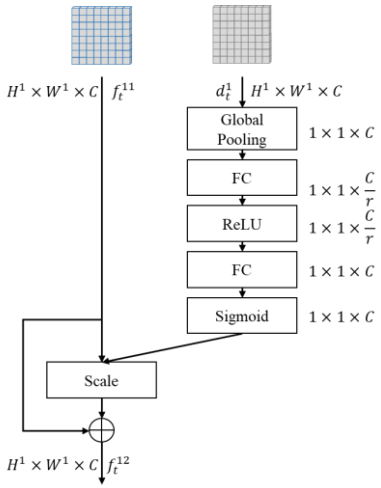

Fig. 3. The structure of the guided squeeze and excitation module.

First, the lower branch formulates motion learning from the correlation aspect by using cross attention and a simple 3D CNN. The cross attention calculates the correlations between two feature maps f_t^1 and f_{t-1}^1 to obtain motion features f_t^1 from temporal dependencies by

$$f_t^{10} = softmax\left(\frac{f_{t-1}^1 f_{t-1}^{1\,T}}{\sqrt{d_k}}\right) f_t^1, \tag{1}$$

After the cross attention, we obtain a correlation feature map f_t^{10} for time t, which contains global temporal correlation information. Afterward, we use a 3D CNN to smooth the f_t^{10} with f_{t-1}^1 by using a $2 \times 3 \times 3 \times C$ kernel and a ReLU activation function generating f_t^{11}.

The upper branch aims to learn global motion features from feature differences. It first calculates the differences d_t^i between f_t^i and f_{t-1}^i, $i \in [1, 2, 3]$ by

$$d_t^i = f_t^i - f_{t-1}^i. \tag{2}$$

Based on the difference pyramid, we propose DFPN to learn motion features from multiple scales:

$$d_t^i = d_t^i + convolution(upsample(d_{t-1}^i)). \tag{3}$$

To enhance the global motion representation, we generate the final global motion features f_t^{12} by combining the global correlation feature map f_t^{11} and the difference feature map d_t^1 with a guided squeeze and excitation module inspired by [25], as shown in Fig. 3. Specifically, the module squeezes the difference feature map d_t^1 to conduct channel-wise attention with the global correlation feature map f_t^{11}, generating the enhanced global motion feature f_t^{12} by adding excited feature map to the global correlation feature map f_t^{11}.

2) Context-Aware Local Motion Modelling: Context-aware local motion modelling aims to learn every object's motion features by a deformable cross attention. Since we already have the global feature map f_t^{12} from the global motion aggregation part. Meanwhile, we have the detection queries $q_{detection}$ and tracking queries $q_{tracking}$ from the previous decoder layers in DETR. Thus, we can transfer and update these object queries q with motion context.

Since f_t^{12} already contains motion features while q has object positions p. A deformable cross attention is implemented to transfer q from positional feature space to motion feature space by

$$q_m = DeformAttn(q, p_q, f_t^{12}) = \sum_{n=1}^{N} W_m[\sum_{k \in \Omega_k} A_{mqf_t^{12}} \cdot W'_m f_t^{12}(p_q + \Delta p_{nf_t^{12}p})]. \tag{4}$$

where A is the deformable attention map, p_q represents the 2-D reference points, Δp means the offsets to sampling points, W represents the learnable weights and $m = H^1 \times W^1$, k is the index for sampled keys in space Ω_k.

Consequently, the objects queries q can be updated by adding motion queries q_m

$$q = q + q_m. \tag{5}$$

Thus, the updated objects queries q contains both positional and motion context information. Then, they will be sent to the subsequent transformer decoder layers for the next frame learning. With motion queries q_m, we can also learn the 2-D offsets $\hat{o}$ by a feed-forward network (FFN). The offsets can be used to supervise motion learning, and the loss function will be introduced in the next section.

C. Loss Function

Since our method is developed on Trackformer, we keep the detection loss $l_{detection}$,

$$l_{detection}(y, \hat{y}, \pi) = \sum_i^N l_{query}(y, \hat{y}_i, \pi), \tag{6}$$

where y is the detection ground truth, $\hat{y}$ and $\hat{y}_i$ are the detection results, and π is a mapping relation. We introduce another offset loss based on detection loss to supervise global-local motion context learning. Since MOT datasets only provide detection and tracking labels without offset labels, we need to generate the offset ground truth o first by

$$o_i^t = c_i^t - c_i^{t-1}, \quad (7)$$

where c_i^t is the center coordinates for object i at time t, and o_i^t is the offset ground truth for object i at time t. Then, the offset loss l_{offset} can be calculated with a $l1$ loss by

$$l_{offset} = \sum_{i=0}^{N} l_1(o_i^t, \widehat{o_i^t}). \quad (8)$$

Overall, the final loss function is the summation of the detection loss $l_{detection}$, and offset loss l_{offset}:

$$l = l_{detection} + l_{offset}. \quad (9)$$

IV. Experiments

A. Settings

We follow the experiment settings in Trackformer. Our METFormer adopts deformable DETR as the detector, which consists of a ResNet-50 as the backbone, six layers of deformable encoders, and six layers of deformable decoders. And for queries, we set the number of detection queries $N_{detection}$ to 500. For tracking queries, they are from both active and inactive queries in the previous frame. We conduct experiments on the MOT17 dataset under public detection settings. For public detection, we train our model on $3 \times 24G$ GPUs for 50 epochs. For the first 40 epochs, the learning rate is 2×10^{-4}. And after that, we drop the learning rate to 10% of the initial learning rate. And we adopt pre-trained weights from the COCO dataset [32] for deformable DETR.

B. MOT17 Dataset

MOT17 includes 14 video sequences with full-body annotation. Seven sequences are used as the training set with provided ground truth. Seven sequences are used for testing without ground truth. For public detection, MOT17 provides detection results from existing detectors, including DPM [29], Faster RCNN [30], and SDP [31]. As shown in Table I, MOT17 can be evaluated on different metrics focusing on different aspects. Multiple Object Tracking Accuracy (MOTA) focuses more on object coverage, while identity F1 score (IDF1) emphasizes the ID perspective. Other metrics like ID Switches, False Positives, and False Negatives are also important to indicate detection and tracking performance, and their definitions can be referred to [28].

TABLE I. The Comparisons among different online tracking methods on the MOT17 test set under public detection Settings.

Methods	MOTA↑	IDF1↑	FP↓	FN↓	ID Sw.↓
FAMNeT [26]	52.0	48.7	14138	253616	3072
GSM [27]	56.4	57.8	14379	230174	**1485**
CenterTrack [15]	60.5	55.7	**11599**	208577	2540
Trackformer [21]	62.3	57.6	16591	192123	4018
METFormer	**62.9**	**59.4**	16710	**189694**	2827

C. Public Detection

Public detection aims to evaluate the performance of the tracker mainly. The principle of public detection is that the detection produced by our detectors should be associated with the detections from three existing detectors [29, 30, 31]. And the tracker is required to link the associated objects rather than all the objects from our detector. For the association, we follow the method in Trackformer, which calculates the IoU distance between detections from our detector and three existing detectors and selects the detections from our detectors with IoUs larger than 0.5.

D. Benchmark Results

As shown in Table I, our proposed METFormer performs better on multiple aspects than our baseline and other leading methods under public detection on the MOT17 dataset, including MOTA, IDF1, False Negatives, and ID Switches.

Our proposed METFormer improves MOTA by 0.6% compared with Trackformer, which means our method has better object coverage. And the main contributions are from the decreased number of False Negatives and ID Switches. It is noticeable that the ID Switches in our method show 21.7% superiority compared to Trackformer. IDF1 is improved by 1.8% compared with Trackformer, which means our method achieves better performance on tracking by introducing the novel global-local motion context learning technique. This also emphasizes that motion is an essential clue for object tracking. We also show two examples to demonstrate the advantage of our METFormer over Trackformer in Fig. 4.

For ablation study, we compare the results for our METFormer with and without difference guidance in Table Ⅱ. The Results show that our final model with DFPN and guided squeeze and excitation module performs better on most comparison metrics, which also indicates that difference guidance is nonnegligible to learn global motion information.

TABLE II. Ablation study on METFormer with (+) and without (-) differnce guidance on MOT17 Test set.

Methods	MOTA↑	IDF1↑	FP↓	FN↓	ID Sw.↓
METFormer (-)	62.2	58.3	12714	197328	3045
METFormer (+)	**62.9**	**59.4**	16710	**189694**	**2827**

Fig. 4. Visualized comparisons between our METFormer and Trackformer. In scenario 1, our METFormer can track the person with ID 106, but Trackformer cannot, and its ID changes from 96 to 219 when occlusion happens. In scenario 2, our METFormer also can keep tracing the object with ID 17, while Trackformer cannot make it when occlusion occurs.

V. Conclusion

In this paper, we developed a tracker named METFormer and introduced a global-local motion learning technique to enhance tracking by global motion aggregation and local motion extraction. Difference-guided global motion learning is to learn the overall motion information. And context-aware local motion modelling is to understand the motion feature of every individual object. Besides, we also introduce an offset loss function to supervise the motion learning process. Referring to the results, our proposed method shows better performance on object coverage and tracking, demonstrating motion as a crucial cue for object tracking.

REFERENCES

[1] E. Bochinski, V. Eiselein and T. Sikora, "High-Speed tracking-by-detection without using image information," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, pp. 1-6.

[2] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 3464-3468.

[3] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 3645-3649.

[4] Y. Zhang et al., '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, 2022, pp. 1–21.

[5] J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani, 'Observation-centric sort: Rethinking sort for robust multi-object tracking', arXiv preprint arXiv:2203. 14360, 2022.

[6] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, 'BoT-SORT: Robust associations multi-pedestrian tracking', arXiv preprint arXiv:2206. 14651, 2022.

[7] L. Chen, H. Ai, Z. Zhuang and C. Shang, "Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification," 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 2018, pp. 1-6.

[8] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, 'Fairmot: On the fairness of detection and re-identification in multiple object tracking', International Journal of Computer Vision, vol. 129, pp. 3069–3087, 2021.

[9] X. Zhou, D. Wang, and P. Krähenbühl, 'Objects as points', arXiv preprint arXiv:1904. 07850, 2019.

[10] C. Feichtenhofer, A. Pinz and A. Zisserman, "Detect to Track and Track to Detect," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 3057-3065, doi: 10.1109/ICCV.2017.330.

[11] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[12] B. Shuai, A. Berneshawi, X. Li, D. Modolo and J. Tighe, "SiamMOT: Siamese Multi-Object Tracking," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 12367-12377.

[13] Z. Zhang, D. Cheng, X. Zhu, S. Lin, and J. Dai, 'Integrated object detection and tracking with tracklet-conditioned detection', arXiv preprint arXiv:1811. 11167, 2018.

[14] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking Without Bells and Whistles," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 941-951, doi: 10.1109/ICCV.2019.00103.

[15] X. Zhou, V. Koltun, and P. Krähenbühl, 'Tracking objects as points', in Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part IV, 2020, pp. 474–490.

[16] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang and J. Yuan, "Track to Detect and Segment: An Online Multi-Object Tracker," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 12347-12356.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, 'End-to-end object detection with transformers', in Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part I 16, 2020, pp. 213–229.

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, 'Deformable detr: Deformable transformers for end-to-end object detection', arXiv preprint arXiv:2010. 04159, 2020.

[19] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan and L. Zhang, "Dynamic DETR: End-to-End Object Detection with Dynamic Attention," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 2968-2977.

[20] D. Meng et al., "Conditional DETR for Fast Training Convergence," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 3631-3640, doi: 10.1109/ICCV48922.2021.00363.

[21] T. Meinhardt, A. Kirillov, L. Leal-Taixé and C. Feichtenhofer, "TrackFormer: Multi-Object Tracking with Transformers," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 8834-8844, doi: 10.1109/CVPR52688.2022.00864.

[22] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, 'Motr: End-to-end multiple-object tracking with transformer', in Computer Vision--ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23--27, 2022, Proceedings, Part XXVII, 2022, pp. 659–675.

[23] P. Sun et al., 'Transtrack: Multiple object tracking with transformer', arXiv preprint arXiv:2012. 15460, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[25] J. Hu, L. Shen, and G. Sun, 'Squeeze-and-excitation networks', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[26] P. Chu and H. Ling, "FAMNet: Joint Learning of Feature, Affinity and Multi-Dimensional Assignment for Online Multiple Object Tracking," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 6171-6180, doi: 10.1109/ICCV.2019.00627.

[27] Q. Liu, Q. Chu, B. Liu, and N. Yu, 'GSM: Graph Similarity Model for Multi-Object Tracking', in IJCAI, 2020, pp. 530–536.

[28] P. Dendorfer et al., 'Motchallenge: A benchmark for single-camera multiple target tracking', International Journal of Computer Vision, vol. 129, pp. 845–881, 2021.

[29] X. Wang, M. Yang, S. Zhu and Y. Lin, "Regionlets for Generic Object Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pp. 2071-2084, 1 Oct. 2015, doi: 10.1109/TPAMI.2015.2389830.

[30] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[31] F. Yang, W. Choi and Y. Lin, "Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2129-2137, doi: 10.1109/CVPR.2016.234.

[32] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in Springer, 2014, pp. 740–755.

[33] Vehicle Tracking Using Deep SORT with Low Confidence Track Filtering

[34] N. Yang, Y. Wang and L. -P. Chau, "Multi-Object Tracking with Tracked Object Bounding Box Association," 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021, pp. 1-6.