# Live Object Tracking using Efficient Vision Transformers

**Tejas Chintala, Hussein Barakat, Hongyu Chen, Angela Nguyen**

## ABSTRACT

Visual object tracking is an area of computer vision that is concerned with the identification and subsequent movement of particular objects that are present in a video sequence. This field has undergone significant advancements, particularly in tracking methods based on deep learning. Transformers-based models, especially efficient transformers have outperformed traditional CNN-based models in their performance on the benchmark datasets. Blatter et al. introduced an Exemplar Transformer within a Siamese tracking architecture for enhanced visual tracking [2]. Building on that architecture, our research introduces a novel algorithm that integrates additional efficient transformer variants, such as the CMT transformer and WaveViT transformer, into the Siamese network architecture, effectively replacing the Exemplar Transformer. This adaptation seeks to leverage the unique strengths of these variants to further refine and improve the tracking capabilities of the system.

Keywords: Visual Object Tracking, Efficient Transformers, Computer Vision, Siamese Network

## 1 INTRODUCTION

Visual object tracking in real-time applications poses a non-trivial challenge due to the need for high accuracy and robustness against changes in object appearance, occlusions, and environmental conditions while maintaining fast processing speeds to support real-time performance. Traditional deep-learning approaches often compromise either speed or accuracy when scaled to real-time applications. The computational complexity of these methods, especially when involving sophisticated deep learning models like transformers, typically requires substantial hardware capabilities.

Efficient transformers tackle the issue of the self-attention mechanism's quadratic complexity in standard transformers. Techniques such as sparsity, low-rank approximations, and kernel methods were frequently used to decrease memory use and computational expenses [14]. This makes them better suited for handling long sequences and large-scale applications. Additionally, their adaptability is enhanced through specialized architectures that incorporate new attention mechanisms [14]. Within efficient transformers, there is a branch of transformers known as Vision transformers

[13], used for computer vision tasks like image classification.

The paper by Blatter et al. (2023) showcases the integration of Exemplar Attention/Transformer within a Siamese architecture for visual tracking by substituting the convolutional layers in the tracker heads with Exemplar Transformer layers [2]. Given the flexibility and effectiveness of transformers across various architectures, this project extends the study presented in the paper and accomplishes the following tasks:

- Utilizing different variants of efficient vision transformers in light track siamese-based architecture (LT-mobile).
- The efficient transformers under consideration are the Exemplar Transformer from the original paper [2], the CMT transformer, and the WaveViT transformer.

# 2 RELATED WORK

## 2.1 Traditional Approaches

Object tracking is one of the challenging tasks in computer vision as it comprehends object modelling, motion modelling and appearance modelling. Despite the wide variety of techniques used in traditional methods such as discriminative correlation filters (DCF), silhouette tracking, kernel tracking, and point tracking, limited techniques such as DCF-based trackers could provide a decent balance between accuracy and computational efficiency. The main challenge of traditional methods includes forced inflexible assumptions about the appearance and motion of the target in the real world [11]. Therefore, the rise of deep learning set a new horizon for visual object tracking research.

## 2.2 Siamese Architecture

In the evolution of Siamese trackers in object tracking, there has been a general shift towards more sophisticated and computationally intensive models. For instance, early models like SiamFC [1] and SINT [8] initially combined basic feature correspondence with the Siamese framework. However, subsequent developments, like SiamRPN [10] and SiamFC++ [16], introduced improved precision mechanisms for bounding box estimation and utilized more powerful backbones like ResNet-50 to enhance feature representation capabilities. Unfortunately, the increased computational demand and larger memory footprints of these advanced models pose challenges for deployment in real-world applications. This is especially evident on mobile and embedded devices where computational

resources are limited. For example, models like SiamRPN++ [9] exceed typical mobile device computational budgets by a significant margin.

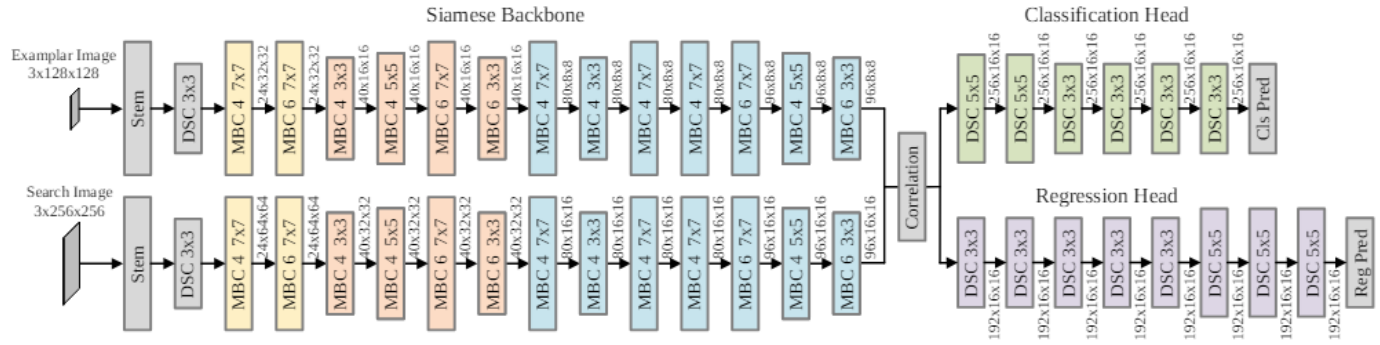


**Fig 1:** Overview of LightTrack Architecture [17]

In response to these challenges, the LightTrack architecture [Fig 1] was designed via Neural Architecture Search (NAS), a methodology aimed at automating the design of efficient neural network architectures [4]. NAS can search simultaneously across the backbone and head network architectures to find optimal solutions tailored for tracking tasks [4]. Unlike earlier methods that required extensive computation, recent NAS approaches use strategies like one-shot weight sharing to reduce costs. The key innovation is to train a hypernetwork whose weights are shared across subnets, enabling efficient exploration of the architecture space.

### 2.3 Efficient Transfomers

Transformers revolutionized signal modeling in machine learning; as it was used firstly for long sequence modeling in Natural language processing [15], then eventually in computer vision [3]. However, the high computational power required by transformers was a constraint to fulfill the low latency requirements of online object tracking. Efficient transformers are a rich topic for research in natural language processing, and recently for computer vision [13]. The different variants of efficient transformer architectures can be classified into 4 main categories: Low rank and Kernel-based methods, Memory-based and downsampling architectures, Factorized and random patterns, and learnable patterns. While low-rank and learnable patterns aim to approximate the self-attention maps; factorized methods and down-sampling-based methods aim to limit the sequence length to reduce the computational complexity [2].

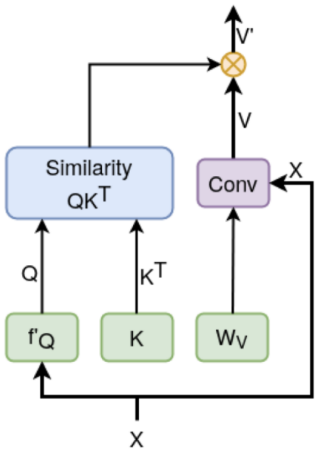


**Fig 2**: Self Attention Block of Exemplar Transformer [2]

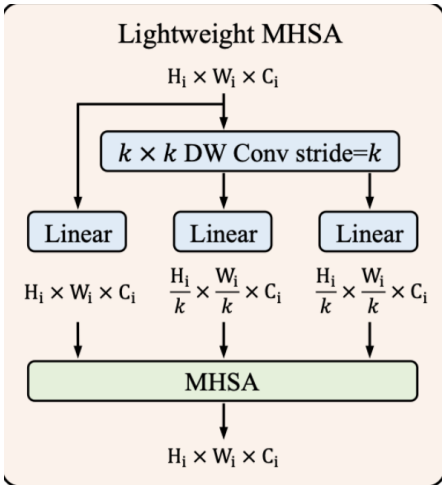


**Fig 3:** Self Attention Block of CMT Transformer [6]

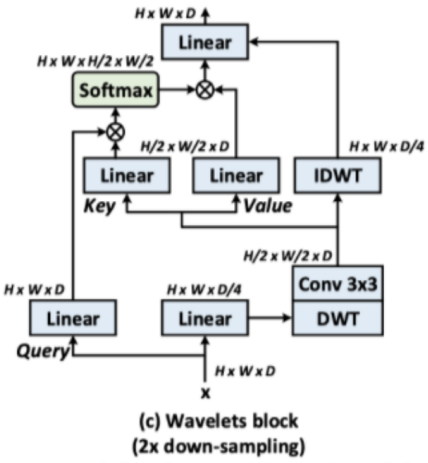


**Fig 4:** Self Attention Block of WaveViT Transformer [18]

## 3 METHODOLOGY

Previous experimental results show that this architecture, termed E.T.Track, not only outperforms other real-time trackers but also competes closely with more advanced transformer-based trackers, significantly closing the performance gap. The success of E.T.Track is evident in its superior performance on the LaSOT dataset compared to established trackers [5]. The choice to replace the Exemplar Transformer with other transformers could potentially offer a significant improvement in handling complex object-tracking scenarios by better capturing and utilizing the spatial relationships within image data.

### 3.1 Transformer Selection

Our project explored the performance of three different transformers on the efficient tracking backbone: Exemplar Transformer [2], CMT Transformer [6] and WaveViT Transformer [18] on an online single object tracking task. These transformers are proposed as strong candidates due to their exceptional performance on the ImageNet dataset [Fig 5] with respect to their total number of parameters [13]. Exemplar transformer utilizes adaptive average pooling with a single global query token and a small set of exemplar key representations to increase compute efficiency [Fig 2]. CMT transformer, on the other hand, processes its input through three consecutive stages: a Local Perception Unit (LPU), a Lightweight Multi-Head self-attention (LMHSA) module [Fig 3] and an Inverted Residual Feed-Forward Network (IRFFN). Depth-wise convolution layers were intensively

used in the CMT attention module to reduce the spatial size for both key and value representations in the attention module, with an additional term for the relative positioning of the token. WaveViT transformer [Fig 4] utilizes an invertible downsampling technique which is the discrete wavelet transform to downsample the signal before the attention layer, with a residual connection with the inverted transform of the signal [6].

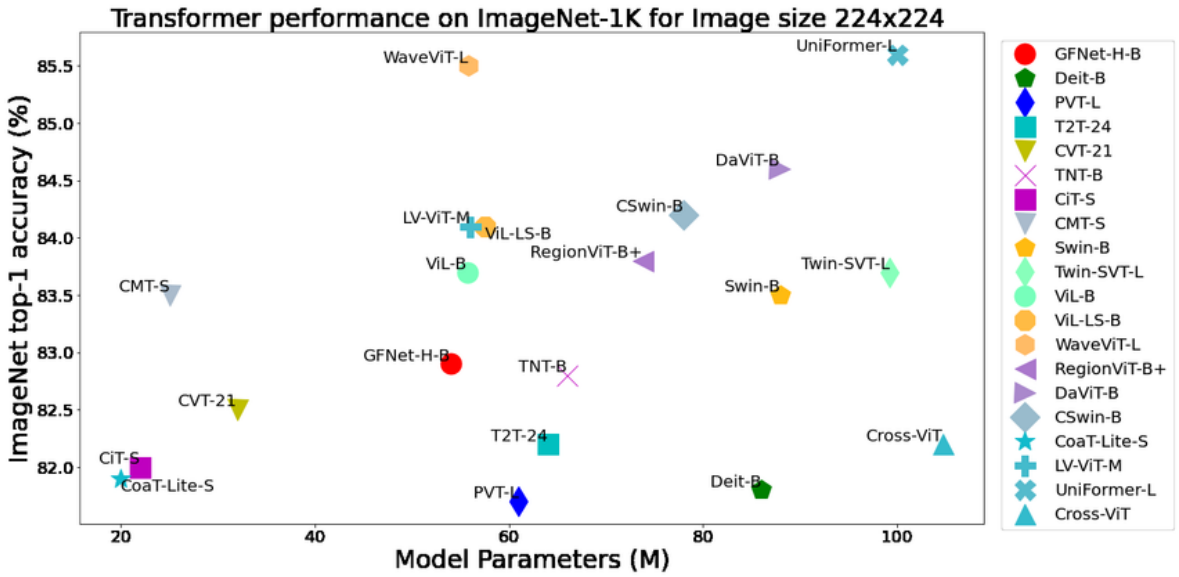


**Fig 5:** Overview of performance of state-of-art vision transformers [13]

### 3.2 Architecture Integration

The architecture of the tracking model utilizes LT-Mobile which passes template and search frame through a convolutional-based siamese-like backbone network for feature extraction [17]. The normalized features afterward are fused through a cross-correlation layer. The generated feature map is then passed through two heads: a classification head and a regression head. The heads are composed of a sequence of convolution layers and self-attention modules. We maintained the number of self-attention modules and the number of attention heads per attention module to generate different models with a close enough number of parameters. The only change with the three models was the attention mechanism. Due to the difference in mechanisms, the CMT-based tracker has a size of 7 million parameters and Exemplar tracker has 9 million parameters while the WaveViT-based tracker contains 12 million parameters. The original implementation by the authors of E.T.Track, Wavelet-L classification model and CMT-S modules were utilized, modified, and adapted to construct the tested models for this project.

### 3.3 Dataset and Experiment Setup

LaSOT is a high-quality benchmark for Large-Scale Single Object Tracking [5]. Due to its extremely long sequences, averaging 2500 frames each, the LaSOT dataset presents significant challenges. Additionally, robustness is crucial for achieving high performance on this dataset. Unfortunately, training each architecture to be evaluated from scratch frequently yields computational demands in the order of thousands of GPU days for NAS [6]. Hence, given our resource constraints, we will focus on training three models specifically on the ”kangaroo” category, which has a total data size of 4GB.

Similar to ET-based tracker, search and template frames are fed through the network. By using point-wise cross-correlation, we can compute the similarity between representations. The result correlation map is fed into the tracker, where a classification branch and a bounding box regression in parallel process it. The bounding box regression branch predicts the distance to all four sides of the bounding box. The classification branch predicts whether each region is part of the foreground or background [3].

All three models have been trained using one single Nvidia RTX 2070, and evaluated on an Inter(R) Core(TM) i7-9700k. Training is based on the OCEAN framework used in Light-Track. Unlike the ET-based Tracker, we did not initialize the backbone with ImageNet pre-trained weights for all three models to make sure the results were comparable. The objective function for the model was a weighted summation of binary cross entropy (BCE) on the classification head and Intersection over Union on the regression head. The optimizer initializes a Stochastic Gradient Descent optimizer with a learning rate of 0.02, momentum of 0.9 and weight decay 1e-4 for training the model. Additionally, it sets up a learning rate scheduler to adjust the learning rate of the optimizer every 5 epochs, with a total of 10 epochs reducing it by a factor of 0.1 each time to optimize convergence.

The sample pairs consist of a 256x256 search frame and a 128x128 template frame, sampled from training splits of LaSOT (kangaroo category) where 15 video clips are used for training and 1 video clip is used for testing [2].

## 4 RESULTS

Table 1 summarizes the comparable results between the three models after 10 epochs of training on a subsample of LaSOT. The performance for ET-Tracker, CMT-Tracker, and WaveViT-Tracker are

0.245, 0.221, and 0.245, respectively, on the training set, and 0.888, 0.856, and 0.938, respectively, on the test set. In light of limited computation power, which prevented further understanding of the behaviour of each model in detail, some patterns might be remarked from the present results. The training of all three models demonstrated nearly identical performance with very similar characteristics in the speed of convergence, likely due to their similar size and initialization strategies. However, early signs of over-fitting could be noticed in Wavelet Tracker in comparison to the ET-Tracker and CMT-Tracker. CMT-Tracker has a slight performance advantage over the other two models, which reflects the power of alternating convolution and attention layers. However, the difference between CMT performance and the other models is questionable due to the relatively low number of parameters of the CMT-based model (7 million parameters vs 9 million parameters for ET-Tracker and 12 million parameters for Wavelet Tracker). Further research is needed to assess the statistical significance of these preliminary findings.

| Model | Avg Train Loss (Last Epoch) | Avg Test loss |
|---|---|---|
| ET-Tracker | 0.245 | 0.888 |
| CMT-Tracker | 0.221 | 0.856 |
| WaveViT-Tracker | 0.245 | 0.938 |

**Table 1.** Train Loss and Test Loss for each of the model

The hyper-parameters and the optimizer used during the training were uniform across all models, derived from the configurations optimized for ET-Tracker's Exemplar Transformer. This uniform approach may not have been ideal, as it could have disadvantages for the CMT-Tracker and WaveViT-Tracker due to potential mismatches in model architecture and hyper-parameter optimization needs. This discrepancy could explain the relatively poor performance of these models compared to ET-Tracker. Further testing should include all categories of the LASOT dataset and with other benchmarks such as GOT10K [7], TrackingNet [12]. This expanded testing would help in validating the initial results and possibly highlight different strengths and weaknesses of the models that were not evident in this constrained testing environment.

## 5 CONCLUSION

This study presented an in-depth analysis in regards to the performance of three transformer-based tracking models: ET-Tracker, CMT-Tracker, and WaveViT-Tracker when handling complex tracking tasks. Our findings revealed that CMT-Tracker, with its alternating depth-wise convolution layers and attention modules, demonstrated the lowest testing loss, which suggests that it may be the most effective model among the three in terms of accuracy. However, the limited testing capacity due to the computation limit as well as the use of uniform hyper-parameters and optimizer settings across all models—settings that were initially optimized specifically for one model may shadow the full capacity of the three models. Model-based hyper-parameter tuning and lack of extensive tests highlight a critical area of improvement for the presented result in this project: measuring the statistical significance of the architecture via a variety of extensive tests on different benchmarks and hyper-parameters. Such tailored adjustments could potentially unveil a more accurate representation of each model's capabilities. Moreover, the constrained resources available for this study limited the depth and breadth of our testing and training phases. To overcome these limitations and to substantiate the preliminary findings, there should be further research involving extended training sessions, comprehensive parameter tuning, and broader dataset applications. Expanding the scope of testing will not only validate the current results but also provide a clearer understanding of how these models perform under varied and more demanding conditions. In conclusion, while CMT-Tracker currently stands out in terms of performance, the true potential of ET-Tracker and WaveViT-Tracker remains to be fully explored. Future research should aim to provide a fair and detailed comparison by optimizing each model's configuration to suit its architectural needs. In doing so, it can be ensured that all models operate under their optimal conditions.

## REFERENCES


[1] BERTINETTO, L., VALMADRE, J., HENRIQUES, J. F., VEDALDI, A., AND TORR, P. H. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14* (2016), Springer, pp. 850–865.

[2] BLATTER, P., KANAKIS, M., DANELLJAN, M., AND GOOL, L. V. Efficient visual tracking with exemplar transformers. In *2023 IEEE/CVF Winter Conference on Applications of Computer*

*Vision (WACV)* (2023), pp. 1571–1581.

[3] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., ET AL. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[4] ELSKEN, T., METZEN, J. H., AND HUTTER, F. Neural architecture search: A survey. *Journal of Machine Learning Research 20*, 55 (2019), 1–21.

[5] FAN, H., LIN, L., YANG, F., CHU, P., DENG, G., YU, S., BAI, H., XU, Y., LIAO, C., AND LING, H. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 5374–5383.

[6] GUO, J., HAN, K., WU, H., TANG, Y., CHEN, X., WANG, Y., AND XU, C. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 12175–12185.

[7] HUANG, L., ZHAO, X., AND HUANG, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence 43*, 5 (2019), 1562–1577.

[8] JIANG, C., XIAO, J., XIE, Y., TILLO, T., AND HUANG, K. Siamese network ensemble for visual tracking. *Neurocomputing 275* (2018), 2892–2903.

[9] LI, B., WU, W., WANG, Q., ZHANG, F., XING, J., AND YAN, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4282–4291.

[10] LI, B., YAN, J., WU, W., ZHU, Z., AND HU, X. High performance visual tracking with siamese region proposal network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 8971–8980.

[11] MARVASTI-ZADEH, S. M., CHENG, L., GHANEI-YAKHDAN, H., AND KASAEI, S. Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems 23*, 5 (2021), 3943–3968.

[12] MULLER, M., BIBI, A., GIANCOLA, S., ALSUBAIHI, S., AND GHANEM, B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 300–317.

[13] PATRO, B. N., AND AGNEESWARAN, V. S. Efficiency 360: Efficient vision transformers. *arXiv preprint arXiv:2302.08374* (2023).

[14] TAY, Y., DEHGHANI, M., BAHRI, D., AND METZLER, D. Efficient transformers: A survey. *ACM Computing Surveys 55*, 6 (2022), 1–28.

[15] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[16] XU, Y., WANG, Z., LI, Z., YUAN, Y., AND YU, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence* (2020), vol. 34, pp. 12549–12556.

[17] YAN, B., PENG, H., WU, K., WANG, D., FU, J., AND LU, H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15180–15189.

[18] YAO, T., PAN, Y., LI, Y., NGO, C.-W., AND MEI, T. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision* (2022), Springer, pp. 328–345.