Digit. Signal Process. 173 (2026) 105935

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp

FETrack: One-stream framework-based feature enhancement for object tracking

Yue Chen [a,b,c,1], Huiying Xu [a,b,1,*], Xinzhong Zhu [a,b,d,*], Xuedong He [a,b], Hongbo Li [d,*], Yi Li [a,b]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, JinHua, 321004, China
[b] School of Computer Science and Technology of Zhejiang Normal University, JinHua, 321004, China
[c] Zhejiang Industry & Trade Vocational College, WenZhou, 325026, China
[d] Beijing Geekplus Technology Co., Ltd, Beijing, 100101, China

ARTICLE INFO

Keywords:
Object tracking
One-Stream
Feature enhancement
Hard sample learning

ABSTRACT

Vision Transformer (ViT)-based one-stream architectures have emerged as the dominant framework for object tracking. However, their performance is hampered by similar object interference and background distractions. To address these limitations, this paper proposes FETrack, a one-stream tracker designed to enhance feature discriminability for improved object tracking. The core innovations of FETrack are as follows: 1) Global Enhancement (GE) and Cross-Depth Template Fusion (CDTF) modules, where the GE module adopts a novel global feature extraction mechanism to suppress background interference, and the CDTF module ensures efficient propagation of contextual information via cross-depth template fusion. 2) An unsupervised hard sample learning strategy, which introduces contrastive learning and treats each candidate token as an independent instance by leveraging its inherent hard sample properties, thereby enhancing feature discriminability. 3) A distillation-based fine-tuning approach that guides parameter optimization for the entire backbone network through feature distillation, enabling efficient tuning of newly integrated modules and ensuring their synergy with the original architecture. Experimental results on six benchmark datasets demonstrate the effectiveness of FETrack and confirm its state-of-the-art performance. Furthermore, the transferability of the proposed approaches for enhancing other one-stream trackers is validated.

1. Introduction

Object tracking is a fundamental task in computer vision, which aims to determine the positions of a target in subsequent frames based on an arbitrarily initialized target in a video sequence. Continuous variations in object states and the modeling of arbitrary object appearance features represent key challenges for target-background discrimination. In the development of tracking frameworks, Siamese network-based trackers [1–5] have attained remarkable success by calculating the similarity between template features and search features. This performance has spurred extensive exploration of the Siamese paradigm in contemporary tracker designs. The advent of the Transformer network has led to the widespread adoption of the derived ViT [6] in diverse image-related tasks, where it has yielded significant performance improvements. Consequently, ViT has been effectively employed as the feature extraction backbone in object tracking, and this has yielded notable performance enhancements.

Research on Transformer-based tracking methodologies has broadly diverged into two main categories: one-stream trackers [7–10] and two-stream trackers [11–14]. Two-stream trackers, which are represented by Siamese trackers, independently extract features from template and search patches prior to similarity computation. In contrast, one-stream Transformer trackers process template and search patches as a unified input, and they perform joint feature extraction. Compared to two-stream trackers, one-stream trackers offer a more concise architecture. They holistically integrate feature extraction, which enables enhanced and more comprehensive feature interactions between templates and search patches. As a representative of one-stream trackers, OSTrack [9] has been widely explored by researchers, and it has extended to ODTrack [10] for temporal propagation of tokens. ODTrack inherits the candidate

* Corresponding authors.
E-mail addresses: chenyue@zjnu.edu.cn (Y. Chen), xhy@zjnu.edu.cn (H. Xu), zxz@zjnu.edu.cn (X. Zhu), hexuedong@zjnu.edu.cn (X. He), Jason.li@geekplus.com (H. Li), leeyee@zjnu.edu.cn (Y. Li).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.dsp.2026.105935

Available online 22 January 2026

elimination module and backbone structure from OSTrack. The candidate elimination module effectively eliminates a large number of low-quality tokens, thereby substantially enhancing training and inference efficiency. Currently, several notable issues in object tracking warrant consideration: 1) A defining characteristic of the object tracking task lies in the uniqueness of the tracked target, and thus the ideal distribution pattern of its features approximates that of a Gaussian kernel. The Gaussian kernel priors assist trackers in acquiring more robust features [15,16]. However, the uncertainty of target positions may cause a misalignment between the peaks of the prior Gaussian kernel and the actual targets. Furthermore, existing multi-head attention mechanisms, serving as general-purpose feature extractors, lack dedicated designs tailored to the characteristics of tracking tasks. 2) Modified tracker backbones often retain pre-trained parameters from their original configurations without specific fine-tuning for the adapted structures prior to training. Consequently, these modified backbones are typically trained jointly with the head. 3) Within the one-stream framework, existing candidate elimination modules are limited to eliminating simple samples, but they lack an elaborate design for distinguishing hard samples. Furthermore, deeper template tokens are more strongly influenced by the remaining candidate tokens, which potentially leads to suboptimal contextual understanding.

Therefore, the objective of this paper is to leverage Gaussian-like unimodal characteristics to enhance feature discriminability while refraining from the use of average pooling. Although average pooling is widely adopted as a global feature extraction strategy in feature enhancement schemes and attention mechanisms [15,17–19], it may weaken the distinctive information of the target relative to the background. Moreover, the aforementioned issues 2 and 3 stem primarily from the following factors. Firstly, the pre-training of the modified backbone involves additional training tasks and consumes substantial time, and a dedicated approach for fine-tuning backbone parameters prior to the main training phase is currently lacking [20,21]. Secondly, the candidate tokens processed by the candidate elimination module are underutilized, particularly in terms of the identification and leveraging of hard samples. Thirdly, while it is beneficial for efficiency, the aggressive elimination of candidate tokens can restrict the interactive context available to deeper template tokens [9].

In this paper, we propose the FETrack tracker, which is an improvement upon the ODTrack tracker. Within FETrack, to generate more robust features for object tracking, we design a Global Enhancement module to improve discriminability between tokens. Compared with the commonly used average pooling function, GE employs a unimodal highlighting function, which is more aligned with the actual requirements of tracking tasks. Instead of simply appending GE modules externally to the backbone, we embed them into each encoder block of ViT. To propagate template information interacted with by candidate tokens across different depths, we design the CDTF module. To enable hard sample learning, we adopt the InfoNCE loss [22,23] from contrastive learning. Originally designed for self-supervised tasks, this loss inherently constructs pseudo-labels for positive and negative samples. The novelty of our approach lies in adapting the InfoNCE loss to unsupervised learning by leveraging the instance discrimination inherent in tracking tasks, where candidate tokens are treated as independent samples.

The aforementioned GE and CDTF module introduce new parameters to the ViT backbone network that require effective initialization. While optimal performance typically relies on backbones pre-trained on ImageNet or with MAE [20,21], such approaches often entail additional learning tasks and considerable computational expense. Therefore, conventional pre-training approaches are not well suited for partial architectural modifications to the backbone network. To address the challenge of effectively initializing and optimizing these newly introduced parameters, we propose a distillation-based fine-tuning approach. This approach leverages feature distillation to optimize these untrained parameters, exclusively utilizing four existing tracking task datasets: LaSOT, GOT10k, TrackingNet, and COCO, without introducing external training data.

The contributions of this paper are summarized as follows:

(1) To enhance the discriminative capability of feature representation, we design a Global Enhancement module, which employs an innovative global feature extraction mechanism to suppress background distractions. Furthermore, we adopt cross-layer deep fusion of template tokens to actively participate in the feature extraction process of the backbone network.
(2) By leveraging the inherent advantages of candidate tokens, we integrate contrastive learning into hard sample learning. Candidate tokens are treated as individual instances, thereby enabling the unsupervised learning of hard samples.
(3) Inspired by feature distillation, we propose a distillation-based fine-tuning approach to adaptively initialize the modified ViT backbone, thereby enhancing the performance of the tracker. Unlike the conventional large-scale pre-training approach on ImageNet and MAE, our approach eliminates the requirement for additional training datasets and offers a significantly simpler implementation.
(4) Experiments are conducted on six benchmarks, where our tracker achieves state-of-the-art performance at 32 fps. Further, we verify the feasibility of transferring the proposed approaches to enhance other one-stream trackers.

2. Related works

In recent years, the field of object tracking has received strong support from advancements in tracker design [1,3,7,9,11,12,24] and the emergence of more tracking benchmarks [25–29]. Researchers have been exploring effective structures that can enable accurate object tracking, such as early correlation filters [30–32], which utilize hand-crafted features to calculate the corresponding template for target localization. With the success of deep learning in computer vision, deep features have replaced hand-crafted features. Currently, numerous tracker studies are based on Siamese and Transformer networks, and related research has significantly advanced the development of object tracking.

2.1. Two-stream based object tracking

With the successful application of Siamese networks in object tracking, trackers within deep learning frameworks leverage a shared-parameter backbone to independently extract feature maps from template and search regions. These feature maps are nonlinearly fused into a unified feature map to enable target-background discrimination and predict bounding boxes. The SiamFC tracker [1] is a typical example of such frameworks, and it estimates target size on multi-scale feature maps. Following SiamFC's success, subsequent Siamese-based trackers have continuously emerged [24,33,34], and AlexNet and VGG convolutional networks serve as the feature extraction backbones. To address the additional computation caused by multi-scale matching, researchers integrated RPN [35] into the Siamese framework, and this led to the development of SiamRPN [2]. Based on the SiamRPN paradigm, numerous subsequent trackers [36–38] further advanced performance. Further, researchers overcame the limitation of not being able to use deeper networks by randomly shifting the target's position within the search area to eliminate center bias. This breakthrough led to the proposal of SiamRPN++ [3] based on ResNet50. Subsequently, many excellent trackers [4,5,39–42] have been derived from this foundational work. In recent years, Transformer-based models have achieved success in numerous image-related tasks. Consequently, visual object tracking has also started to employ Transformer networks [43]. Notably, several trackers have achieved significant performance improvements by integrating Transformer-based attention mechanisms for feature fusion [11,12].

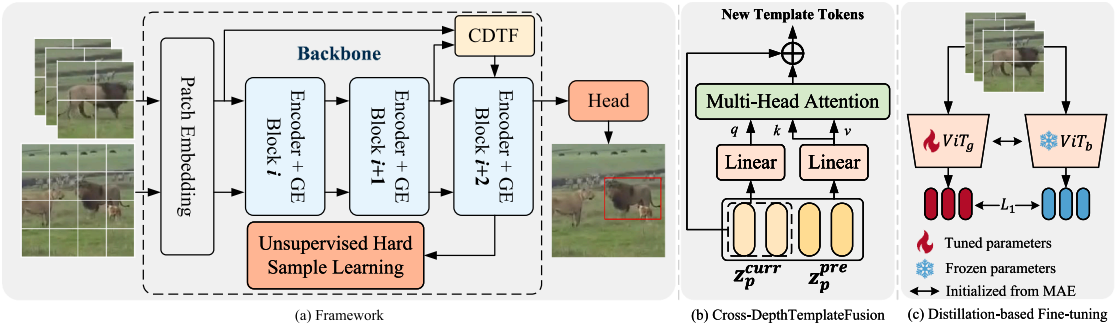

Fig. 1. (a) Framework of the proposed FETrack. The key components include the Global Enhancement (GE) module, the Cross-Depth Template Fusion (CDTF) module, and Unsupervised Hard Sample Learning (UHSL). (b) Architecture of the proposed CCDTF module. (c) Framework of the distillation-based fine-tuning approach. For the identical components of ViT_g and ViT_b, the MAE pre-training parameters of ViT are adopted as their shared initialization parameters. Fine-tune the overall backbone parameters of ViT_g using the output of ViT_b as the reference.

2.2. One-stream based object tracking

As research pivoted towards Transformer architectures, ViT was adopted as the primary backbones for feature extraction. Notably, models such as SwinTrack [13] leveraged ViT. However, even with ViT, early Transformer-based trackers followed the two-stream paradigm where features from the template and search region were extracted independently. This independent processing inherently limited mutual interaction between the features, thereby constraining the full exploitation of the Transformer's potential. To address this shortcoming, the one-stream framework has been proposed to process the template and search region as an integrated input. This joint feature learning approach has yielded significant breakthroughs in tracker performance, exemplified by OSTrack [9] and MixFormer [7]. Currently, auxiliary methods for enhancing the training and inference of primary tasks have been widely adopted [44,45]. In the field of object tracking, relevant studies improve tracker performance by designing distillation methods and auxiliary functions [46–48]. For instance, AVTrack [46] constructs a viewpoint-invariant learning mechanism to enhance model robustness via maximizing mutual information; ORTrack [47] employs the spatial Cox process to model random masking operations, guiding the model to learn more discriminative features. Notably, both AVTrack and ORTrack adopt the teacher-student distillation strategy to optimize the inference process.

Within the one-stream framework, most trackers focus on further integrating features after the backbone and designing effective prediction heads. Concurrently, several approaches attempt to further extract global feature information based on Transformer [13,49,50]. Moreover, to optimize the calculation efficiency, the token selection mechanism is regarded as an optimization strategy, which eliminates tokens to reduce the calculation burden.

While existing trackers often prioritize components external to the backbone network, any modifications introduced within the backbone typically necessitate jointly fine-tuning the newly added parameters along with the prediction head during formal training. The pre-training of ViT networks mostly relies on ImageNet image classification [20] or MAE [21], which increases the difficulty of parameter learning after modifying the backbone. To enhance tracking performance, we propose the GE and CDTF modules for feature enhancement. To optimize the parameters of the newly added modules in the backbone network, we design a distillation-based fine-tuning strategy to assist in fine-tuning the backbone parameters. Additionally, to refine the candidate elimination module and enhance the distinguishability of hard samples, we implement unsupervised hard sample learning.

3. Method

This section details the overall framework of FETrack, along with its four improved components: the GE module, CDTF module, distillation-based fine-tuning, and unsupervised hard sample learning for candidate tokens. Its framework is illustrated in Fig. 1(a). Specifically, the GE module enhances features of the search region, the CDTF module fuses features of the template region, distillation-based fine-tuning optimizes backbone parameters post integration of the GE and CDTF modules, and unsupervised hard sample learning is exclusively applied in the training phase. As these components target distinct problems and application scenarios, there is no design redundancy among them. The image frame input to the FETrack is defined as follows, template patches $z \in \mathbb{R}^{3\times H_z\times W_z}$, search patches $x \in \mathbb{R}^{3\times H_x\times W_x}$. Transformer trackers receive a sequence of non-overlapping image patches (each with a resolution of p × p) as input, which are linear, flatten, and transposed to obtain $z_p^0 \in \mathbb{R}^{N_z\times D}, x_p^0 \in \mathbb{R}^{N_x\times D}$,where D is the token dimension, $N_z = nH_zW_z/p^2$, $N_x = H_xW_x/p^2$, n is number of template image patches. $H^0 = [z_p^0, x_p^0]$, $H^0 \in \mathbb{R}^{(N_z+N_x)\times D}$, it as input features of Transformer encoder.

3.1. Feature enhancement driven by global features and template interaction

This subsection introduces how global features enhance feature representations in FETrack. Our approach focuses on identifying representative features by extracting global information facilitated through token comparison. This process is primarily intended to identify information with unimodal highlighting attributes and emphasize the distinction between the target and background. In the ViT, H^{i-1} passes through the $(i-1)$th encoder to obtain $H^i = [z_p^i; x_p^i]$. To reduce interference from background information while acquiring global information in the search region, we extract information from the remaining candidate tokens. We define the extraction method as follows:

$$F_j^i = \max(\text{mean}(\max(x_{pj}'^i) - x_{pj}'^i), \text{mean}(\max(-x_{pj}'^i) + x_{pj}'^i)) \quad (1)$$

where i is the number of the encoder and j is the index of token dimension, $x_{pj}'^i \in \mathbb{R}^{N_x'\times 1}$, $F^i = [F_0^i; F_1^i; \ldots; F_{D-1}^i] \in \mathbb{R}^{1\times D}$. H^i is calculated by the multi-head attention module and the candidate elimination module. The first 30% of tokens are retained and normalized, as illustrated in Fig. 2, to obtain $x_p'^i$. $N_x' = 30\% \cdot N_x$. The weights of candidate tokens are calculated based on the Q and K of Multi-Head Attention (MHA) [9]. 30% is selected based on the remaining ratio of the reference candidate elimination module [9].

The aforementioned global feature F^i reflects the unimodality degree of feature patterns across different channels. Taking global features as a reference, we achieve feature enhancement through Global Feature Matching and Unimodal-Aware processing, with the specific process illustrated in Fig. 2. The progressive reduction in token count due to candidate elimination during training renders fixed-input-dimension Multi-Layer Perceptrons (MLP) unsuitable. To address this, we design a specialized $\text{Conv}(\cdot)$, which consists of two 1×5 convolution kernels and a ReLU activation function. Symmetric zero-padding of 4 is applied

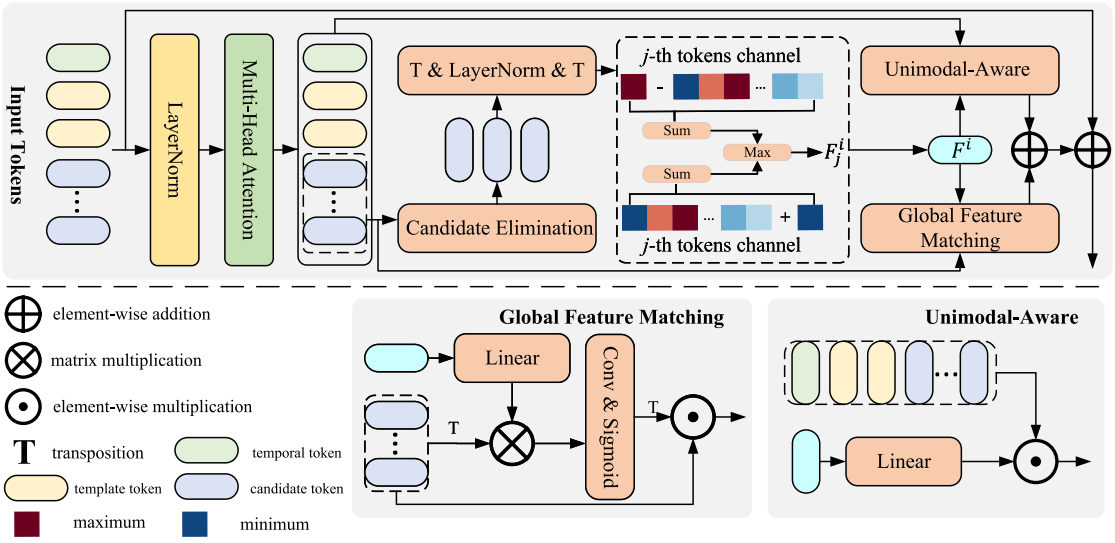

Fig. 2. The encoder architecture incorporates the details of the GE module. F^i represents the global feature that is calculated from a subset of the remaining candidate tokens. The global feature matching is conducted using F^i, together with utilization of unimodal-aware channels.

along the longer edge of the convolution kernels. In the encoder, H^i is first calculated via MHA, as follows:

$$Q^i, K^i, V^i = \text{Linear}(H^i) \tag{2}$$

$$\begin{aligned} H_m^i &= [z_m^i; x_m^i] \\ &= \text{Linear}(\text{Softmax}\left(Q^i K^{iT}/\sqrt{d_k}\right)V^i) \end{aligned} \tag{3}$$

where Linear is a linear change, Q^i, K^i, V^i are query, key and value independent matrices, d_k is scale factor. The calculation for Global Feature Matching and Unimodal-Aware are as follows:

$$C^i = \text{Sigmoid}(F^i W_1^i) \tag{4}$$

$$H_m'^i = H_m^i \odot C^i \tag{5}$$

$$Attn = \text{Conv}(F^i W_2^i x_m^{iT}) \tag{6}$$

$$Attn = Attn[:; 8 : end] \tag{7}$$

$$x_m'^i = \text{Sigmoid}(Attn^T) \odot x_m^i + x_m'^i \tag{8}$$

$$P^i = H^i + H_m'^i \tag{9}$$

$$H^{i+1} = \text{MLP}(\text{LayerNorm}(P^i)) + P^i \tag{10}$$

where $W_1^{i-1}, W_2^{i-1} \in \mathbb{R}^{D\times D}$, $\odot$ is element-wise multiplication. If $\mathbb{R}^{a\times b} \odot \mathbb{R}^{1\times b}$ occurs, the $\mathbb{R}^{1\times b}$ is repeated a times to form $\mathbb{R}^{a\times b}$. $\text{Conv}(\cdot)$ will convert $\mathbb{R}^{1\times N_x}$ to $\mathbb{R}^{1\times N_x+8}$. $[:; 8 : end]$ is used to remove the first 8 weights on the left which ensures the dimensions of $Attn$ are consistent with number of tokens, it will convert $\mathbb{R}^{1\times N_x+8}$ to $\mathbb{R}^{1\times N_x}$. Candidate tokens are ranked by weight in descending order, and high-weight tokens serve as the focus of subsequent computations. Therefore, filtering out tokens affected by left-padded zeros is critical.

It is important to note that the aforementioned approaches are integrated into each encoder of the ViT backbone, and they actively participate in feature extraction.

Fig. 3 presents the visualization results of the output features from GE module and multi-head attention module. The multi-head attention module extracts global features by leveraging the similarity between tokens. The GE module proposed in this study is developed based on this mechanism, and it aims to identify unimodal prominent features in the global context. This identification of unimodal features helps reduce information ambiguity. From Fig. 3, we can observe that as the layer progresses, our method is able to reduce irrelevant responses.

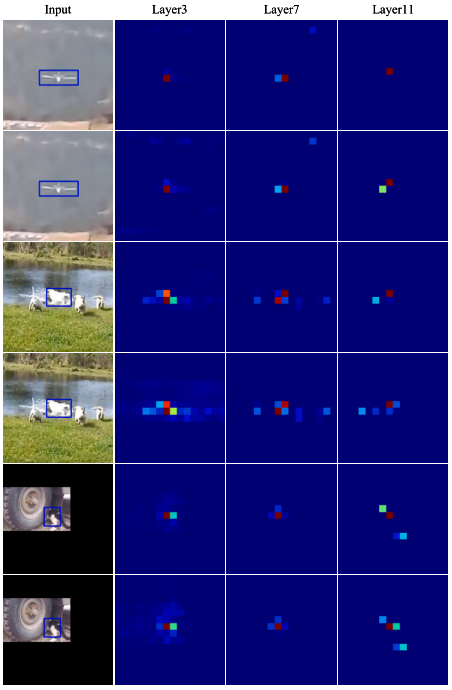

Fig. 3. Visualize the attention weights of the search region. These weights correspond to the central part of the templates in different layers. The odd-numbered rows of images are the visualizations of GE module output, and even-numbered rows of images represent the visualizations of the multi-head attention module output.

The candidate elimination module [9] enables template tokens at different depths to interact with candidate tokens of varying quantities. Consequently, template tokens at different depths acquire distinct structural and semantic representations. To enrich the contextual information of template tokens while conforming to the Transformer architecture, CDTF adopts the following approach:

$$Q = \text{Linear}(z_p^i) \quad K, V = \text{Linear}([z_p^i, z_p^{i-2}]) \tag{11}$$

$$z_h^i = \text{Linear}(\text{Softmax}(QK^T/\sqrt{d_k})V) + z_p^i \tag{12}$$

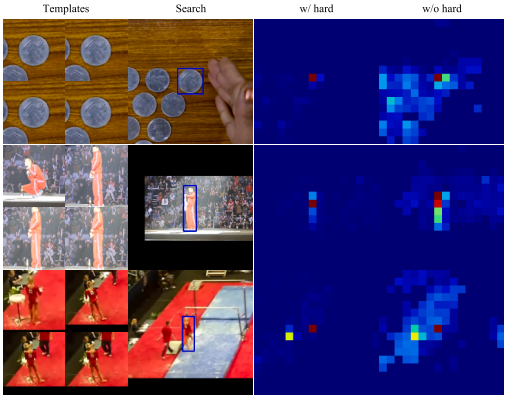

Fig. 4. Visualize the proposed unsupervised hard sample learning approach, with identical input samples and search regions employed to ensure fairness and comparability. "w/ hard" denotes training with the proposed approach.

where Linear is a linear change, $Q \in \mathbb{R}^{N_z \times D}$, $K, V \in \mathbb{R}^{(2N_z) \times D}$, z_h^i is repalce z_p^i to constitute H^i, $i = \{6, 8, 10\}$. The visualization of the above calculation process is shown in Fig. 1(b).

3.2. Distillation-based fine-tuning for backbone

In visual tracker design, ViT backbones typically leverage pre-trained networks. Consequently, structural modifications or the integration of bespoke backbone modules are infrequently adopted. This is primarily because structural modifications to the original architecture tend to induce performance instability. Several trackers involve modifications to the backbone; the untrained parts of these trackers are usually jointly trained with the head. When addressing these backbone modifications, the traditional pre-training approaches rely on ImageNet image classification [20] or unsupervised objectives such as MAE [21]. However, for modifications to existing backbone architectures (e.g., ViT), we propose leveraging the original pre-trained parameters. This approach enables efficient parameter fine-tuning through feature distillation, which allows the joint optimization of both newly integrated modules and the backbone structure. We denote the original ViT backbone architecture as ViT_b, which consists of 12 encoders. The improved backbone, integrated with the modules introduced in Section 3.1, is defined as ViT_g. The overall distillation-based fine-tuning framework is illustrated in Fig. 1(c), and its formulation is given below:

$$H_b^i = \text{ViT}_b(H^0) \tag{13}$$

$$H_g^i = \text{ViT}_g(H^0) \tag{14}$$

$$L = \sum_i \lambda_i \| H_b^i - H_g^i \| \tag{15}$$

where H_0 denotes the combination of template and search region images, with specific reference to the definition of H_0 in Section 3.1. The i denotes the index corresponding to the encoder whose output is expected to be retrieved. For the identical components of ViT_g and ViT_b, the MAE pre-training parameters of ViT_b are adopted as their shared initialization parameters. During the fine-tuning process, we adjust the overall parameters of ViT_g. In subsequent experiments, we will validate the effectiveness of the proposed approach.

3.3. Unsupervised learning of hard samples

Our tracker is improved upon ODTrack, while retaining its prediction head. The classification branch obtains the classification score map$\mathbb{R}^{1 \times H_x/p \times W_x/p}$, bounding box size$\mathbb{R}^{2 \times H_x/p \times W_x/p}$, and offset size $\mathbb{R}^{2 \times H_x/p \times W_x/p}$. In addition to the loss functions composed of the aforementioned three branches, we conduct unsupervised learning of hard samples on the candidate tokens x_p^{end} obtained through the ViT_g. The tokens of x_p^{end} are sorted by the candidate elimination module in descending order of weight to form $\mathbb{R}^{N_{x'} \times D}$, and $N_{x'}$ is the number left after eliminating N_x. As described in the candidate elimination module [9], the weights are obtained by similarity calculation, so tokens with high weights inherently represent hard-to-distinguish samples.

Different tokens actually map to different $p \times p$ regions, so we can reasonably assume that different candidate tokens are independent instances. We select the first n tokens from N_x' tokens as the hard sample $K_n \in \mathbb{R}^{n \times D}$. We introduce InfoNCE loss to distinguish hard samples, which is expressed as follows:

$$\alpha = K_n K_n^T \tag{16}$$

$$\beta_i = -\log(\exp(\alpha_{i,i}/\tau) / \sum_j \exp(\alpha_{i,j}/\tau)) \tag{17}$$

$$\beta_i = \begin{cases} \beta_i, & \beta_i > \sigma \\ 0, & other \end{cases} \tag{18}$$

$$\mathcal{L}_{con} = \sum_i \beta_i \tag{19}$$

where each instance corresponds to an independent sample. The setting of σ in formula (18) aims to avoid performance degradation caused by uncontrolled divergence between tokens. The hard sample learning only takes effect during the training process and does not impact the inference process. Fig. 4 presents the visualization results of our proposed approach.

We optimize the following objective function through joint training:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{GIoU} + \lambda_3 \mathcal{L}_{con} \tag{20}$$

where $\lambda_1 = 5, \lambda_2 = 2, \lambda_3 = 0.1$ are the regularization parameters. $\mathcal{L}_{cls}$, $\lambda_1 \mathcal{L}_1$, $\lambda_2 \mathcal{L}_{GIoU}$ reference ODTrack [10].

4. Experiment

In this section, we present the performance results of the proposed tracker on multiple tracking benchmarks and compare it with other state-of-the-art trackers. We also conduct an ablation study to analyze the roles of each approach and module in the model.

4.1. Implementation details

Training. Our primary reference backbone is a ViT-Base architecture [6], whose parameters are pre-initialized with MAE [21]. The initialization of the backbone parameters uses the distillation-based fine-tuning approach in Section 3. The training dataset includes LaSOT [25], GOT10k [27], TrackingNet [62] and COCO [63]. The data input method is to use three template frames with 128×128 pixels and two search frames with 256×256 pixels. We set the ratio factor between the search area and the target size to 5. Optimization is performed using the AdamW optimizer with an initial learning rate of 2×10^{-5} for the backbone and 2×10^{-4} for the head; and the weight decay is set to 10^{-4}. We train the FETrack with 300 epochs and 60k image pairs for each epoch. After 240 epochs, all learning rates are multiplied by 0.1. The hyperparameter settings of unsupervised hard sample learning are as follows: $n = 24, \sigma = 0.3, \tau = 1$. Distillation-based fine-tuning adopts the same parameters as aforementioned, with the training epoch set to 10. The tracker is trained on three 24G RTX3090 GPUs with a batch size of 16. The distillation-based fine-tuning is done on a single 24G RTX3090 GPU with a batch size of 16.

Inference. Our tracker performs inference on a single RTX 3090 GPU. The inference protocol follows that of ODTrack [10], achieving a running speed of 32 fps.

4.2. Comparison with the SOTA

To validate the performance of our proposed tracker, we conduct evaluations across six tracking benchmarks. The authors of ODTrack

Table 1
Comparison results with state-of-the-art trackers on the three tracking benchmarks: LaSOT, LaSOT$_{ext}$ and TrackingNet. The best two results are highlighted in **bold** and *italic*, respectively .

Method	LaSOT			LaSOT$_{ext}$			TrackingNet		
	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P
SiamFC [1]	33.6	42.0	33.9	23.0	31.1	26.9	57.1	66.3	53.3
ECO [51]	32.4	33.8	30.1	22.0	25.2	24.0	55.4	61.8	49.2
SiamRPN++ [3]	49.6	56.9	49.1	34.0	41.6	39.6	73.3	80.0	69.4
ATOM [52]	51.5	57.6	50.5	37.6	45.9	43.0	70.3	77.1	64.8
DiMP [53]	56.9	65.0	56.7	39.2	47.6	45.1	74.0	80.1	68.7
SiamR-CNN [40]	64.8	72.2	–	–	–	–	81.2	85.4	80.0
Ocean [54]	56.0	65.1	56.6	–	–	–	–	–	–
TrDiMP [43]	63.9	–	61.4	–	–	–	78.4	83.3	73.1
TransT [11]	64.9	73.8	69.0	–	–	–	81.4	86.7	80.3
Stark [12]	67.1	77.0	–	–	–	–	82.0	86.9	–
SwinTrack-224 [13]	67.2	–	70.8	47.6	–	53.9	81.1	–	78.4
Mixformer-22k [7]	69.2	78.7	74.7	–	–	–	83.1	88.1	81.6
OSTrack-256 [9]	69.1	78.7	75.2	47.4	57.3	53.3	83.1	87.8	82.0
AiATrack [14]	69.0	79.4	73.8	47.7	55.6	55.4	82.7	87.8	80.4
TATrack-B224 [55]	69.4	78.2	74.1	–	–	–	83.5	88.3	81.8
GRM-256 [56]	69.9	79.3	75.8	–	–	–	84.0	88.7	83.3
SeqTrack-B256 [57]	69.9	79.7	76.3	49.5	*60.8*	56.3	83.3	88.3	82.2
ARTrack-256 [58]	70.4	79.5	76.6	48.4	57.7	53.7	84.2	88.7	83.5
SPformer-256 [15]	70.7	*80.7*	77.0	*50.1*	60.7	*56.9*	83.6	88.5	82.8
DiffusionTrack-256 [59]	70.7	80.0	77.3	–	–	–	83.6	88.1	82.0
ODTrack-256 [10]	69.3	79.4	76.1	47.1	57.5	53.1	*84.4*	**89.7**	83.4
LMTrack-256 [60]	69.8	79.2	76.3	49.0	59.6	55.8	84.2	89.0	82.8
UncTrack-B [61]	*70.9*	*80.7*	*77.5*	–	–	–	**85.0**	*89.6*	**84.5**
FETrack	**71.1**	**81.9**	**77.6**	**50.5**	**62.0**	**57.9**	84.3	**89.7**	*83.8*

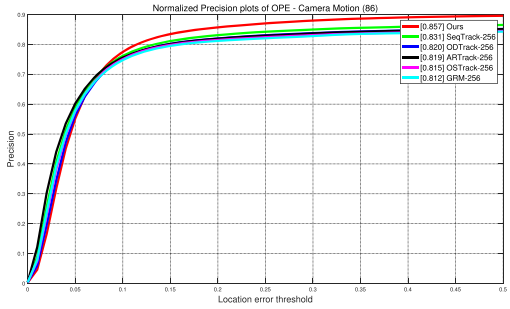

Fig. 5. P_{norm} evaluation of one-pass evaluation (OPE) for camera motion challenges on LaSOT.

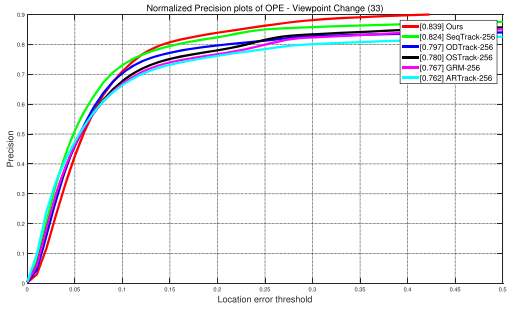

Fig. 6. P_{norm} evaluation of OPE for viewpoint change challenges on LaSOT.

have made the size 256 model parameters available on their GitHub project page.

LaSOT. LaSOT is a challenging large-scale visual tracking benchmark. Its test set comprises 280 video sequences. Compared with other datasets, LaSOT has an average of 2500 frames per video sequence. These sequences are longer and more challenging than those in other datasets. The excellent performance of FETrack is verified by the results presented in Table 1. For instance, in comparison with ODTrack-256, our tracker achieves 1.8%, 1.5%, 2.5% improvements in AUC, P and P_{norm}, respectively. This confirms that our approaches can effectively enhance tracking performance. Compared with ARTrack-256, our tracker achieves 0.7%, 1.0%, 2.4% improvements in AUC, P and P_{norm}, respectively. Compared with the DiffusionTrack-256, our tracker achieves 0.4%, 1.3%, 1.9% improvements in AUC, P and P_{norm}, respectively. To explore specific details, we conducted separate statistics on different challenge attributes in LaSOT. The corresponding results are illustrated in Fig. 7. We observe that our tracker outperforms ODTrack-256 in most challenge scenarios. Additionally, in scenarios such as fast motion, camera motion, and viewpoint change, our method achieves significant improvements. It also compensates for the limitations of ODTrack-256. As shown in Figs. 5 and 6, we present the plots of P_{norm} corresponding to the two challenge attributes with the highest AUC values.

LaSOT$_{ext}$. LaSOT$_{ext}$ [26] is an extended dataset of LaSOT. It consists of 150 more challenging long-term video sequences. LaSOT$_{ext}$ is more challenging than LaSOT. It includes 15 new categories and 15 new videos. From Table 1, our tracker achieves 50.5%, 57.9%, 62.0% in AUC, P and P_{norm}, respectively. It is an improvement of 3.4%, 4.8%, 4.5% compared to ODTrack-256. Compared with AiATrack, our tracker achieves 2.8%, 2.5%, 6.4% improvements in AUC, P and P_{norm} , respectively. Compared with SeqTrack-B256, our tracker achieves 1.0%, 1.6%, 1.2% improvements in AUC, P and P_{norm}, respectively. Our tracker achieves the most significant improvement on the LaSOT$_{ext}$ dataset. This is because state-of-the-art trackers generally exhibit limited performance on LaSOT$_{ext}$, which leaves substantial room for performance enhancement. As shown in Fig. 8, our tracker demonstrates competitive performance in most challenge scenarios on LaSOT$_{ext}$.

TrackingNet. TrackingNet is a large-scale collection of short-term video sequences. Its test set comprises 511 short video sequences. In addition, the annotations for the TrackingNet test set are not publicly available; they need to be submitted via the evaluation server. As presented in Table 1, our tracker exhibits limited performance improvement compared to ODTrack-256 on the TrackingNet test set. Several state-of-the-art trackers in Table 1 achieve high values in AUC, P, and P_{norm}. This indicates that the video sequences in TrackingNet have relatively lower difficulty compared with other test benchmarks. Compared

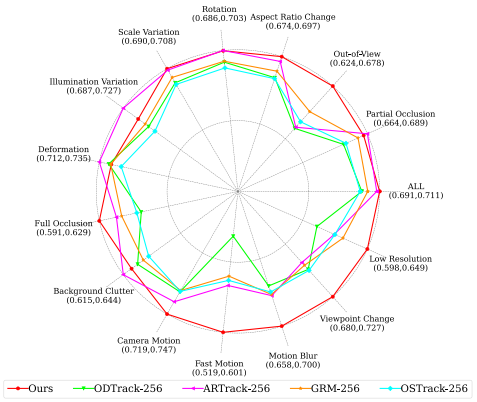

Fig. 7. AUC evaluation for different challenges on LaSOT.

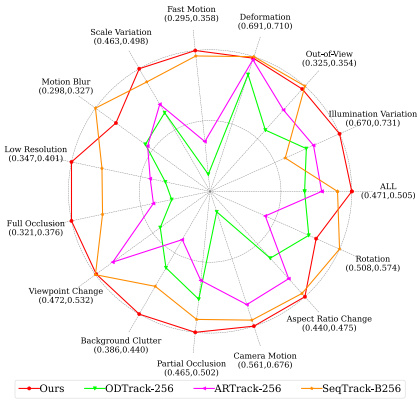

Fig. 8. AUC evaluation for different challenges on $\text{LaSOT}_{\text{ext}}$.

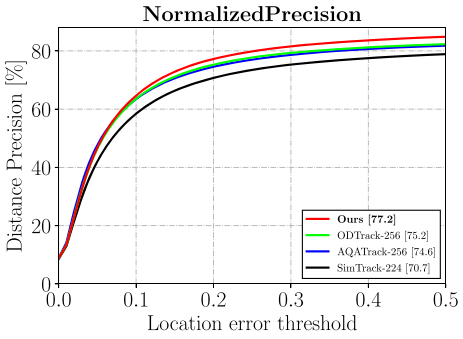

Fig. 9. Normalized Precision plots of OPE on TNL2K.

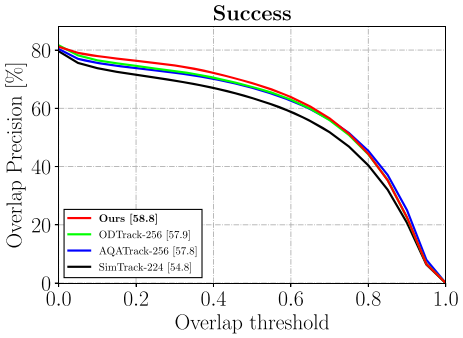

Fig. 10. Success plots of OPE on TNL2K.

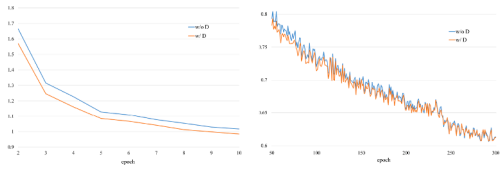

Fig. 11. Variation of training loss, where D is Distillation-based Fine-tuning.

with DiffusionTrack-256, our tracker achieves 0.7%, 1.8% and 1.6% improvements in AUC, P and P_{norm}, respectively. As shown in Table 1, compared to other advanced trackers, our tracker still has advantages.

GOT10k. GOT10k is a large-scale video dataset with more than 10,000 videos, and its test set consists of 180 videos. The training and testing protocols of GOT10k are unique. Specifically, models are trained exclusively on the GOT10k training dataset. The evaluation server provides corresponding test data for upload; GOT10k adopts the one-pass evaluation (OPE) protocol. This protocol uses three metrics: average overlap (AO), and success rates (SR) at thresholds of 0.5 and 0.75. From Table 2, our tracker achieves 1.0%, 1.5%, 1.6% improvements over ODTrack-256 in AO, $SR_{0.5}$, $SR_{0.75}$, respectively. Compared with DiffusionTrack-256, our tracker achieves 1.4%, 2.1%, 3.8% improvements in AO, $SR_{0.5}$ and $SR_{0.75}$, respectively. In addition, compared with UncTrack-256, our tracker has improvements of 1.0%, 2.9%, 0.6% in AO, $SR_{0.5}$ and $SR_{0.75}$, respectively.

NFS and TNL2K. We evaluate our tracker on two benchmarks: NFS [64], TNL2K [29]. These two benchmarks contain 100 and 700 test video sequences, respectively. NFS is a dedicated benchmark for fast-moving objects. It includes two versions with different frame rates: 30 fps and 240 fps. We mainly evaluate our tracker on its low-frame-rate version (30 fps). TNL2K is a recently released high-quality dataset for natural language-based tracking. From the results in Table 2, we can observe that our proposed strategies yield significant performance improvements. On the NFS test set, our tracker achieves 0.9%, 2.5%, 2.0% improvements in AUC, P, P_{norm}, respectively, compared with ODTrack-256. On the TNL2K test set, the tracker proposed in this study achieves 0.7%, 1.2%, 0.8% improvements in AUC, P, P_{norm}, respectively, compared with ODTrack-256. Compared with DiffusionTrack-256, our tracker achieves 2.3%, 4.7%, 4.6% improvements in AUC, P and P_{norm} , respectively. In the TNL2K testing evaluation, our tracker demonstrates leading performance compared with other advanced trackers, such as AQATrack-256 [65], SeqTrack-B256 [57] and ARTrack [58]. To intuitively demonstrate the effectiveness of our FETrack, Figs. 9 and 10 present the relevant curves of partial P_{norm} and AUC metrics.

Through comparisons across the aforementioned six test benchmarks, our tracker achieves state-of-the-art performance. Compared with the original model, our tracker also achieves an overall performance improvement.

4.3. Ablation study

To investigate the individual contribution of each component within our proposed tracker, we conduct comprehensive ablation experiments. Furthermore, we evaluate different feature extraction functions to identify those better suited for the single object tracking (SOT) task. Comparative experiments are performed to analyze the effectiveness of distillation-based fine-tuning. These extensive experiments are performed on two benchmarks: LaSOT [25] and $\text{LaSOT}_{\text{ext}}$ [26].

Global Enhancement. From Table 3, we prune GE modules. This operation is conducted to verify the effectiveness of these modules. The results indicate that the integration of GE module can yield positive performance gains. In terms of AUC, GE module can bring 0.7% and 0.9% improvements on the LaSOT and $\text{LaSOT}_{\text{ext}}$ test sets. In terms of P, GE module can bring 0.7% and 1.6% improvements on the LaSOT and $\text{LaSOT}_{\text{ext}}$ test sets. As shown in Table 4, we conducted comparative experiments on the remaining ratio mentioned in Section 3.1. The exper-

Table 2
Comparison results with state-of-the-art trackers on the three tracking benchmarks: GOT10K, NFS and TNL2K. The best two results are highlighted in **bold** and *otalic, respectively*.

Method	GOT10K			NFS			TNL2K		
	AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{norm}	P	AUC	P_{norm}	P
SiamRPN++ [3]	51.7	61.6	32.5	50.2	–	–	40.3	–	41.2
ATOM [52]	55.6	63.4	40.2	58.3	–	–	40.1	–	–
DiMP [53]	61.1	71.7	49.2	61.8	–	–	44.7	–	–
Ocean [54]	61.1	72.1	47.3	57.4	–	–	38.4	–	–
TransT [11]	67.1	76.8	60.9	65.3	–	78.8	50.7	–	51.7
SwinTrack-224 [13]	71.3	81.9	64.5	66.0	–	–	53.0	–	53.2
SimTrack-B/16 [66]	68.6	78.9	62.4	–	–	–	54.8	70.7	53.2
OSTrack-256 [9]	71.0	80.4	68.2	64.7	–	–	54.3	–	–
F-BDMTrack-256 [67]	72.7	82.0	69.9	–	–	–	56.4	–	56.5
ARTrack-256 [58]	73.5	82.2	70.9	64.3	–	–	57.5	–	–
SeqTrack-B256 [57]	74.7	84.7	71.8	66.1	84.0	81.9	56.4	–	–
DiffusionTrack-256 [59]	75.2	85.9	72.0	–	–	–	56.5	72.6	57.3
AQATrack-256 [65]	73.8	83.2	72.1	–	–	–	57.8	74.6	59.4
ODTrack-256 [10]	*75.6*	*86.5*	74.2	*66.9*	*84.5*	*82.5*	*57.9*	75.2	59.9
TemTrack-256 [68]	74.9	84.8	71.7	–	–	–	**58.8**	*75.6*	*60.9*
UncTrack-B [61]	*75.6*	85.1	*75.2*	–	–	–	–	–	–
FETrack	**76.6**	**88.0**	**75.8**	**67.8**	**86.5**	**85.0**	**58.8**	**77.2**	**62.0**

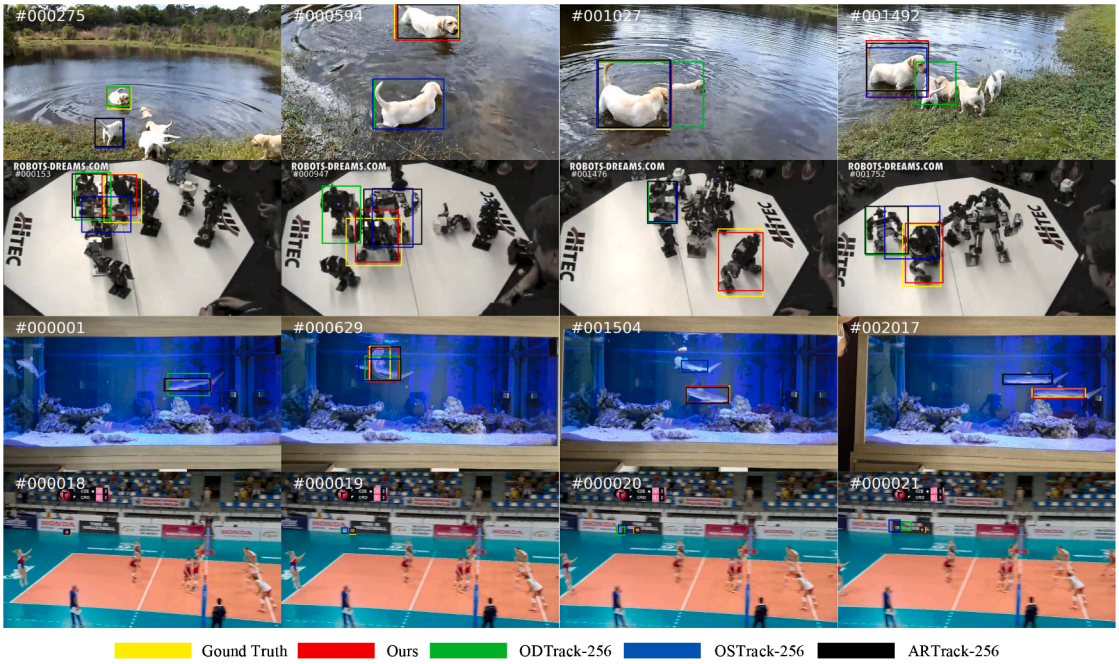

Fig. 12. Qualitative comparison of FETrack with three other SOTA trackers on LaSOT.

Table 3
Ablation studies of different approaches on LaSOT and $LaSOT_{ext}$ benchmarks. Boldface indicates the first place.

Method	LaSOT			$LaSOT_{ext}$		
	AUC	P_{norm}	P	AUC	P_{norm}	P
FETrack	**71.1**	**81.9**	**77.6**	**50.5**	**62.0**	**57.9**
w/o Global Enhancement	70.4	81.7	76.9	49.6	61.3	56.3
w/o Unsupervised Hard Sample Learning	70.1	80.9	76.6	49.0	60.3	55.3
w/o Cross-Depth Template Fusion	70.0	80.9	76.7	**50.5**	**62.0**	57.7
w/o Distillation-based Fine-tuning	70.3	81.4	77.0	50.0	61.6	56.7

imental results confirm that 30% is the optimal value of this parameter, which is consistent with relevant reference criteria.

Unsupervised Hard Sample Learning. To fully utilize the candidate tokens obtained from the candidate elimination module, we treat these candidate tokens as independent hard instances for unsupervised learning. To verify the feasibility of this approach, we conduct an ablation experiment on the approach. The corresponding results are presented in Table 3. Ablation experiments demonstrate that unsupervised learning of candidate tokens yields positive gains, with 1.0% and 1.5% improvements in the AUC of the LaSOT and $LaSOT_{ext}$ test sets, respectively. In terms of P, this approach achieves 1.0% and 2.6% improvements on the LaSOT and $LaSOT_{ext}$ test sets, respectively. To explore

Table 4
Comparative studies on different remaining ratio for GE on the LaSOT and $\text{LaSOT}_{\text{ext}}$. Boldface indicates the first place.

Ratio	LaSOT		$\text{LaSOT}_{\text{ext}}$	
	AUC	*P*	*AUC*	*P*
30%	**71.1**	**77.6**	50.5	**57.9**
20%	70.3	77.0	**50.7**	**57.9**
10%	70.3	76.9	49.8	56.4

Table 5
Comparative studies on different hyperparameter for UHSL on the LaSOT and $\text{LaSOT}_{\text{ext}}$. Boldface indicates the first place.

Hyperparameter	LaSOT		$\text{LaSOT}_{\text{ext}}$	
	AUC	*P*	*AUC*	*P*
$\lambda_3 = 0.1, n = 24, \sigma = 0.3$	**71.1**	**77.6**	50.5	57.9
$\lambda_3 = 1.0, n = 24, \sigma = 0.3$	69.7	76.0	50.2	57.1
$\lambda_3 = 0.01, n = 24, \sigma = 0.3$	69.7	75.7	49.2	55.7
$\lambda_3 = 0.1, n = 6, \sigma = 0.3$	70.1	76.6	**51.1**	**58.2**
$\lambda_3 = 0.1, n = 12, \sigma = 0.3$	69.8	76.1	50.0	56.8
$\lambda_3 = 0.1, n = 48, \sigma = 0.3$	70.5	77.0	50.2	57.0
$\lambda_3 = 0.1, n = 24, \sigma = 0.0$	69.8	76.1	49.8	56.4
$\lambda_3 = 0.1, n = 24, \sigma = 0.1$	70.0	76.3	49.6	56.2
$\lambda_3 = 0.1, n = 24, \sigma = 0.5$	70.2	76.6	50.9	57.7

Table 6
Comparative studies on different cross-layer selection for CDTF on the LaSOT and $\text{LaSOT}_{\text{ext}}$. Boldface indicates the first place.

Select	LaSOT		$\text{LaSOT}_{\text{ext}}$	
	AUC	*P*	*AUC*	*P*
6,8,10	**71.1**	**77.6**	**50.5**	**57.9**
4,7,10	69.4	75.7	50.2	57.1
4,6,8,10	69.3	75.5	49.4	56.1
all	69.8	75.9	49.4	56.4

the impact of different hyperparameters in unsupervised hard sample learning, we conducted comparative experiments with varied hyperparameter values, as shown in Table 5. Experimental results show that λ_3 tends to adversely affect the experimental outcomes when its value is either too large or too small. Regarding other hyperparameters, alternative configurations outperform the current settings under specific conditions.

Cross-Depth Template Fusion. To investigate the role of CDTF modules, we conduct module addition and ablation operations, followed by separate training. Ablation experiments demonstrate that CDTF yields positive gains, with results presented in Table 3. Specifically, results on the LaSOT test set indicate that the involvement of cross-depth template tokens in subsequent template token interactions contributes positive gains, with 1.1%, 0.9% and 1.0% improvements in AUC, P and P_{norm}, respectively. To verify the impact of selecting different layers, we perform template fusion using different layers, as shown in Table 6. Herein,"all" indicates that the template of each layer is fused with that of the previous layer. As observed, the optimal performance is achieved when the layer fusion positions are set to [6, 8, 10].

Distillation-based Fine-tuning. To verify the impact of distillation-based fine-tuning on the backbone prior to backbone-head joint training, we adopt the fine-tuned parameters of ViT-B as the baseline for comparison. Both Table 3 and Fig. 11 demonstrate that distillation-based fine-tuning enhances model performance and accelerates the decline in training loss. Furthermore, our fine-tuning method is straightforward and requires only a small batch size, providing valuable insights for integrating modules into existing backbones for object tracking tasks.

Table 7
Comparative studies of different global feature extraction strategies on LaSOT and $\text{LaSOT}_{\text{ext}}$. Boldface indicates the first place.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$		
	AUC	P_{norm}	*P*	*AUC*	P_{norm}	*P*
Unimodal Highlighting	**71.1**	**81.9**	**77.6**	**50.5**	**62.0**	**57.9**
w/o Max	69.7	80.3	76.0	49.8	61.3	57.1
Average Pooling	70.0	80.8	76.6	50.4	61.9	57.4

Table 8
Compare the results of OSTrack-256 and its improved tracker on five test benchmarks. Boldface indicates the first place.

Method	LaSOT		$\text{LaSOT}_{\text{ext}}$		TrackingNet		TNL2K	NFS
	AUC	P_{norm}	*AUC*	P_{norm}	*AUC*	P_{norm}	*AUC*	*AUC*
baseline	**69.1**	78.7	47.4	57.3	**83.1**	87.8	54.3	64.7
Improvement	68.6	**78.8**	**49.4**	**60.2**	82.7	**88.0**	**55.9**	**66.8**

Table 9
Comparison of model parameters, FLOPs, and inference speed.

Method	Device	Speed(FPS)	MACs(G)	Params(M)
GRM-256	3090	28	30.9	99.4
SeqTrack-B256	3090	37	65.8	89.1
ODTrack-256	3090	58	32.5	92.1
FETrack	3090	33	34.6	108.6
w/o GE	3090	55	34.6	94.5
w/o CDTF	3090	35	32.5	106.3

Global Feature Extraction Function. To verify the advantages of the feature extraction module proposed in this paper, we conduct comparative experiments using alternative global feature extraction modules: w/o Max and average pooling. w/o Max corresponds to the first half of the calculation of max in formula (1) [39]. From Table 7, we can observe that the function proposed in this study exhibits overall excellent performance. Compared with average pooling, our function achieves 1.1%, 0.9% and 1.0% improvements in the AUC, P and P_{norm} on the LaSOT test set, respectively. Compared with w/o Max, our study adopts a max strategy to balance different unimodal distributions. This operation yields positive performance gains. Specifically, our function achieves 1.4% and 0.7% improvements in the AUC on the LaSOT and $\text{LaSOT}_{\text{ext}}$ test sets, respectively. Through comparative experiments, we can verify that the proposed strategy exerts certain positive effects.

4.4. Approach transfer

To verify that the approaches proposed in this study are applicable not only to improving ODTrack but also to extending to other one-stream trackers, we apply these approaches to the OSTrack tracker. OSTrack [9] is a representative one-stream tracker. It also serves as the baseline for numerous one-stream trackers. As presented in Table 8, our approaches enable OSTrack-256 to achieve significant performance improvements on the $\text{LaSOT}_{\text{ext}}$, TNL2K and NFS test sets. Meanwhile, they avoid notable performance degradation on LaSOT and TrackingNet. This result demonstrates two key points: the transferability of our strategies, and the feasibility of enhancing the performance of one-stream trackers based on OSTrack.

4.5. Visualization and limitation

Visualization. To visually evaluate our tracker's performance in challenging scenarios, we present its visualization results on the LaSOT test set. As shown in Fig. 12, our tracker demonstrates stronger robustness in complex environments than three state-of-the-art counterparts. Specifically, the first row of Fig. 12 verifies its superior reliability over the other three trackers; in the third row, object deformation and sim-

ilar distractors drastically impair the performance of ODTrack-256 and OSTrack-256; the fourth row further shows that our tracker outperforms ODTrack-256 in high-speed motion scenarios. Combined with the results in Figs. 6 and 12, FETrack excels over the three comparison trackers in addressing motion blur, fast motion, occlusion and viewpoint variation, with both superior visualization performance and more prominent AUC metrics across these specific challenges.

Limitation. For the sake of fairness, we conducted a speed evaluation against SOTA trackers under identical testing conditions, as shown in Table 9. Analysis reveals that the tracking speed degradation of FETrack is mainly attributed to the GE module. Further analysis reveals that this phenomenon stems from the token weight ranking implemented during candidate elimination. As shown in Table 1, FETrack achieves more favorable performance metrics than ODTrack-256, GRM-256 and SeqTrack-B256 across all three primary test sets. Moreover, under the same evaluation environment, FETrack attains a comparable speed to GRM-256 and SeqTrack-B256. For complex architectures, we have not yet developed a more effective solution. However, two promising directions are identified: the design of lightweight Transformer architectures and the development of an efficient token pruning approach.

5. Conclusion

This paper presents an improved FETrack based on a one-stream framework. In terms of architecture, we fully integrate the Global Enhancement module into the backbone network, where a novel global feature extraction method is designed to guide the global feature extraction process. Furthermore, we propagate template tokens across layers to construct the Cross-Depth Template Fusion module. By fully exploiting the hard-sample characteristics of candidate tokens, we implement unsupervised hard sample learning. We also propose a distillation-based fine-tuning approach to fine-tune the newly integrated components in the backbone. Extensive experiments demonstrate that our tracker verifies the feasibility of the proposed approaches on six tracking benchmarks. Additionally, we transfer these approaches to OSTrack to verify the feasibility of the proposed approaches in enhancing the performance of one-stream trackers.

CRediT authorship contribution statement

Yue Chen: Conceptualization, Methodology, Data curation, Investigation, Validation, Software, Writing – original draft, Writing – review & editing; **Huiying Xu:** Formal analysis, Data curation, Data curation, Investigation, Project administration; **Xinzhong Zhu:** Project administration, Funding acquisition; **Xuedong He:** Formal analysis, Writing – review & editing; **Hongbo Li:** Supervision; **Yi Li:** Writing - review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

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.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056); Zhejiang Province Province-Land Synergy Program (2025SDXT004-3); Zhejiang Provincial Natural Science Foundation of China (Grant no. LQN25F030016); Jinhua Science and Technology Bureau (Grant no. 2024-4-006); and Teacher Technology Innovation Program of Zhejiang Industry & Trade Vocational College (G230210).

References

[1] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, 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, Springer, 2016, pp. 850–865.
[2] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[3] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, 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.
[4] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6668–6677.
[5] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, SiamCAR: siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6269–6277.
[6] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale,(2020). arXiv:2010.11929
[7] Y. Cui, C. Jiang, L. Wang, G. Wu, Mixformer: end-to-end tracking with iterative mixed attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13608–13618.
[8] Y. Cui, T. Song, G. Wu, L. Wang, Mixformerv2: efficient fully transformer tracking, Adv. Neural Inf. Process. Syst. 36 (2023) 58736–58751.
[9] B. Ye, H. Chang, B. Ma, S. Shan, X. Chen, Joint feature learning and relation modeling for tracking: a one-stream framework, in: European Conference on Computer Vision, Springer, 2022, pp. 341–357.
[10] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, X. Li, Odtrack: online dense temporal token learning for visual tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 7588–7596.
[11] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8126–8135.
[12] B. Yan, H. Peng, J. Fu, D. Wang, H. Lu, Learning spatio-temporal transformer for visual tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10448–10457.
[13] L. Lin, H. Fan, Z. Zhang, Y. Xu, H. Ling, Swintrack: a simple and strong baseline for transformer tracking, Adv. Neural Inf. Process. Syst. 35 (2022) 16743–16754.
[14] S. Gao, C. Zhou, C. Ma, X. Wang, J. Yuan, Aiatrack: attention in attention for transformer visual tracking, in: European Conference on Computer Vision, Springer, 2022, pp. 146–164.
[15] F. Cheng, G. Peng, J. Li, B. Zhao, J.-S. Pan, H. Li, A transformer-based network with adaptive spatial prior for visual tracking, Neurocomputing 614 (2025) 128821.
[16] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, S. Maybank, Learning attentions: residual attentional siamese network for high performance online visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4854–4863.
[17] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[18] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[19] Z. Zhang, Y. Liu, Z. Zhou, G. Yang, X. Liao, Q.M.J. Wu, DGeC: dynamically and globally enhanced convolution, IEEE Transact. Artifi. Intell. 6 (4) (2025) 921–933.
[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[22] O.A. van den, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, (2018). arXiv:1807.03748
[23] T. Park, A.A. Efros, R. Zhang, J.-Y. Zhu, Contrastive learning for unpaired image-to-image translation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 319–345.
[24] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H.S. Torr, End-to-end representation learning for correlation filter based tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2805–2813.
[25] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, 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.
[26] H. Fan, H. Bai, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, Harshit, M. Huang, J. Liu, et al., Lasot: a high-quality large-scale single object tracking benchmark, Int. J. Comput. Vis. 129 (2021) 439–461.

[27] L. Huang, X. Zhao, K. Huang, Got-10k: a large high-diversity benchmark for generic object tracking in the wild, IEEE Trans. Pattern Anal. Mach. Intel.l 43 (5) (2019) 1562–1577.
[28] M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for UAV tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 445–461.
[29] X. Wang, X. Shu, Z. Zhang, B. Jiang, Y. Wang, Y. Tian, F. Wu, Towards more flexible and accurate object tracking with natural language: algorithms and benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13763–13773.
[30] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intel.l 37 (3) (2014) 583–596.
[31] H. Possegger, T. Mauthner, H. Bischof, In defense of color-based model-free tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2113–2120.
[32] D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, Visual object tracking using adaptive correlation filters, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2544–2550.
[33] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4834–4843.
[34] Z. Zhang, H. Peng, Deeper and wider siamese networks for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4591–4600.
[35] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intel.l 39 (6) (2016) 1137–1149.
[36] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.
[37] H. Fan, H. Ling, Siamese cascaded region proposal networks for real-time visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7952–7961.
[38] G. Wang, C. Luo, Z. Xiong, W. Zeng, Spm-tracker: series-parallel matching for real-time visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3643–3652.
[39] X. He, Y. Chen, Z. Zheng, P. Bian, Feature complement for visual tracking based on global feature comparison, IEEE Access 11 (2023) 11840–11848.
[40] P. Voigtlaender, J. Luiten, P.H.S. Torr, B. Leibe, Siam r-cnn: visual tracking by re-detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6578–6588.
[41] W. Zhou, Y. Liu, N. Wang, Y. Wang, B. Peng, Dual-branch collaborative siamese network for visual tracking, Digit. Signal Process. 155 (2024) 104716.
[42] Z. Zhu, X. Zhao, K. Tang, S. Zhang, W. Yang, W. He, Efficient long-term tracking with local–global similar object interference suppression, Digit. Signal Process. 160 (2025) 105065.
[43] N. Wang, W. Zhou, J. Wang, H. Li, Transformer meets tracker: exploiting temporal context for robust visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1571–1580 .
[44] M. Xiong, Q. Zhang, D. Li, W. Wang, Z. Zhang, K. Zhang, C. Liu, D. Chen, J. Zhang, Fine-tuning feature interaction for unsupervised domain adaptive low-light object detection, Neurocomputing 658 (2025) 131717.
[45] Q. Chen, Z. Zhang, Z. Zhang, K. Zhang, D. Li, W. Wang, J. Zhang, C. Liu, Distilled large language model-driven dynamic sparse expert activation mechanism, Appl. Soft Comput. 185 (2025) 114037.
[46] Y. Wu, Y. Li, M. Liu, X. Wang, X. Yang, H. Ye, D. Zeng, Q. Zhao, S. Li, Learning an adaptive and view-invariant vision transformer for real-time UAV tracking, IEEE Trans. Circuit. Syst. Video Technol. (2025) 1–1.
[47] Y. Wu, X. Wang, X. Yang, M. Liu, D. Zeng, H. Ye, S. Li, Learning occlusion-robust vision transformers for real-time UAV tracking, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17103–17113.
[48] C. Xue, B. Zhong, Q. Liang, Y. Zheng, N. Li, Y. Xue, S. Song, Similarity-guided layer-adaptive vision transformer for UAV tracking, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6730–6740.
[49] Y. Liang, Q. Li, F. Long, Global dilated attention and target focusing network for robust tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 1549–1557.
[50] X. Zhou, P. Guo, L. Hong, J. Li, W. Zhang, W. Ge, W. Zhang, Reading relevant feature from global representation memory for visual object tracking, Adv. Neural Inf. Process. Syst. 36 (2023) 10814–10827.
[51] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, Eco: efficient convolution operators for tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6638–6646.
[52] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Atom: accurate tracking by overlap maximization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4660–4669.
[53] G. Bhat, M. Danelljan, L.V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6182–6191.
[54] Z. Zhang, H. Peng, J. Fu, B. Li, W. Hu, Ocean: object-aware anchor-free tracking, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 771–787.
[55] K. He, C. Zhang, S. Xie, Z. Li, Z. Wang, Target-aware tracking with long-term context attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 773–780.
[56] S. Gao, C. Zhou, J. Zhang, Generalized relation modeling for transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18686–18695.
[57] X. Chen, H. Peng, D. Wang, H. Lu, H. Hu, Seqtrack: sequence to sequence learning for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14572–14581.
[58] X. Wei, Y. Bai, Y. Zheng, D. Shi, Y. Gong, Autoregressive visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9697–9706.
[59] F. Xie, Z. Wang, C. Ma, Diffusiontrack: point set diffusion model for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19113–19124.
[60] C. Xu, B. Zhong, Q. Liang, Y. Zheng, G. Li, S. Song, Less is more: token context-aware learning for object tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 8824–8832.
[61] S. Yao, Y. Guo, Y. Yan, W. Ren, X. Cao, Unctrack: reliable visual object tracking with uncertainty-aware prototype memory network, IEEE Trans. Image Process. 34 (2025) 3533–3546.
[62] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, B. Ghanem, 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.
[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[64] H. Kiani Galoogahi, A. Fagg, C. Huang, D. Ramanan, S. Lucey, Need for speed: a benchmark for higher frame rate object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1125–1134.
[65] J. Xie, B. Zhong, Z. Mo, S. Zhang, L. Shi, S. Song, R. Ji, Autoregressive queries for adaptive tracking with spatio-temporal transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19300–19309.
[66] B. Chen, P. Li, L. Bai, L. Qiao, Q. Shen, B. Li, W. Gan, W. Wu, W. Ouyang, Backbone is all your need: a simplified architecture for visual object tracking, in: European Conference on Computer Vision, Springer, 2022, pp. 375–392.
[67] D. Yang, J. He, Y. Ma, Q. Yu, T. Zhang, Foreground-background distribution modeling transformer for visual object tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10117–10127.
[68] J. Xie, B. Zhong, Q. Liang, N. Li, Z. Mo, S. Song, Robust tracking via mamba-based context-aware token learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 8727–8735.

Yue Chen received the M.S. degree in Computer Science and Technology from College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua China, in 2021, where he is currently pursuing the Ph.D. degree in Computer Science and Technology. His current research interests include Object Tracking and Deep Learning.

Huiying Xu received the M.S. degree from National University of Defense Technology (NUDT), China. She is an associate professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd, China. Her research interests include Kernel learning and feature selection, Object Detection, Vision SLAM, Computer vision, Image processing, Pattern recognition, Computer simulation, Deep clustering, Generative Adversarial Network, Diffusion Model, Clustering Ensemble, Multiple Kernel Learning, Learning with incomplete data and their applications. She is a member of the China Computer Federation. She has published papers, including those in highly regarded journals such International Journal of Intelligent Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, etc.

Xinzhong Zhu received the Ph.D. degree from Xidian University and M.S. degree from National University of Defense Technology (NUDT), China. He is a professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the chief scientist of Beijing Geekplus Technology Co., Ltd. and president of Research Institute of Ningbo Cixing Co., Ltd., China. His research interests include Machine learning, Deep clustering, Computer vision, Object detection, Segmentation, Recognition and Tracking, Diffusion Model, Manufacturing informatization, Manufacturing informatization, Robotics and System integration, Laser SLAM, Vision SLAM, Low Quality Data Learning, Multiple Kernel Learning, and Intelligent manufacturing. He is a member of the ACM and certified as CCF distinguished member. Dr. Zhu has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Knowledge and Data Engineering, CVPR, NeurIPS, AAAI, IJCAI, etc. He served on the Technical Program Committees of IJCAI 2020 and AAAI 2020.

Xuedong He received the Ph.D. degree in School of Intelligent Systems Engineering, Sun Yat-sen University, China. Currently, He is with the School of Computer Science and Technology, Zhejiang Normal University, as a University Associate Professor. His current research interests include visual object tracking, video object segmentation, object detection and deep learning.

Hongbo Li received his Ph.D. degree in computer science from Tsinghua University in 2009. Currently, he holds the position of the Chief Technology Officer and Co-founder of Beijing Geek+ Technology Co., Ltd China. In addition, he also serves as the secretary-general of Chinese Intelligent service Society and is an Editorial Board Member of several high-profile journals. His research interests include the design and application of intelligent robots, intelligent information process, and intelligent logistic systems. He has pub-

lished more than 70 papers in prestigious journals and conference, and has been awarded more than 120 patents, including 46 international invention patents.

Yi Li received the M.S. degree in Software Engineering from College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua China, in 2021, where he is currently pursuing the Ph.D. degree in Computer Science and Technology. He is a student member of CFF. His current research interests include deep learning, computer vision tasks, Object Detection, Instance Segmentation, Object Tracking.