SFDATrack: Generalized Source-Free Domain Adaptive Tracking Under Adverse Weather Conditions

Siyuan Yao[1], Ziqi Wang[1,2], Ruiqi Yu[3], Junqi Huang[1], Wenqi Ren[1†], and Xiaochun Cao[1†]

[1] Shenzhen Campus of Sun Yat-sen University
[2] Beijing University of Posts and Telecommunications
[3] Nanyang Technological University

{yaosiyuan04,zqwatcherbr0,yuruiqi422,huangjq3632,rwq.renwenqi}@gmail.com, caoxiaochun@mail.sysu.edu.cn

Abstract. Domain adaptive visual object tracking under adverse weather conditions has garnered significant attention in recent years. Despite the impressive performance, existing methods heavily rely on the large-scale video frames from both source and target domains, which is impractical under rigid resource constraints where source data is unavailable. To overcome this limitation, we propose SFDATrack, a generalized source-free domain adaptive tracker that merely leverages adverse weather samples from the target domain for robust state estimation. Specifically, SFDATrack first employs a mean-teacher backbone with Dual Interactive Mamba (DIM) blocks to distill the candidate target tokens that are resilient to weather variations from classified, augmented samples. Afterwards, we introduce a hyperspherical prototype projection (HPP) module to project these tokens onto multi-domain prototypes within a latent hyperspherical space. By enforcing both domain-specific and domain-invariant properties of the multi-domain prototypes, SFDATrack can be seamlessly adapted to diverse weather conditions with powerful generalizability. Extensive experiments evaluated on various benchmarks demonstrate that SFDATrack achieves superior performance compared to state-of-the-art approaches. The code is available at https://github.com/watcherBR0/sfdatrack.

Keywords: Visual Object Tracking · Source-Free Domain Adaptation · Hyperspherical Embedding

1 Introduction

Visual object tracking (VOT) is an essential computer vision task, which aims to estimate the trajectory of an arbitrary target in video sequences under the guidance of first frame initialization. It serves as a critical component for a

† Corresponding authors.

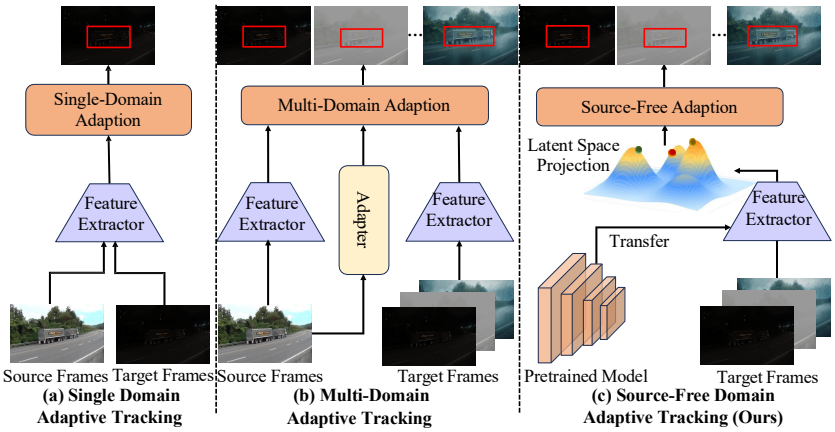

Fig. 1: Comparison of the domain adaptive tracking frameworks under adverse weather conditions. (a) Single domain adaptive tracking [50]. (b) Multi-domain adaptive tracking [48]. (c) The proposed source-free domain adaptive tracker (SFDATrack). SFDATrack adapts a pre-trained model to multiple target domains via latent space projection, enabling seamless adaptation to diverse weather conditions without requiring any original source data.

wide range of applications, spanning from autonomous driving, video surveillance and embodied intelligence. Driven by the powerful deep learning architectures, modern trackers have shown remarkable capability to predict the target state in controlled environments.

While the deep trackers have achieved impressive performance on standard benchmarks, a critical challenge persists for modern VOT, *i.e.*, when a tracking model trained on a curated source domain (*e.g.*, daytime) is deployed to an unseen target domain (*e.g.*, nighttime or fog), it often suffers severe performance degradation. Such adverse weather conditions introduce complex, non-uniform perturbations like low contrast and atmospheric noise, which severely disrupt the model's identification ability to track the target object.

To overcome this challenge, recent studies have proposed Domain Adaptive Visual Object Tracking (DAVOT) [13, 39, 50, 51, 53] as a promising solution. As shown in Fig. 1, the prevailing DAVOT methods typically follow an unsupervised single domain adaptation or multi-domain adaptation paradigms, which utilizes annotated source data and unlabeled target data to learn domain-invariant features across different weather conditions. However, these standard domain adaptive tracking framework depend on a critical yet often impractical assumption: the perpetual availability of the source domain data during adaptation. In real-world scenarios, this assumption is frequently invalid due to storage limitations, computational costs, or data privacy regulations. For instance, it is infeasible to collect large-scale source videos on a device with limited memory or to transmit them over bandwidth-constrained networks. This limitation raises a pivotal question: ***can a pre-trained source domain model be successfully adapted to***

arbitrary challenging target domains for object tracking without any access to the original source data?

In this paper, we propose SFDATrack, a novel source-free domain adaptive tracker that achieves robust visual tracking in adverse weather using only unlabeled target data. Our approach is built upon the idea that an object's essential visual semantics are resilient to weather variations in the embedding space, even as its low-level appearance shifts. Motivated by this, we design a mean-teacher network that performs multi-domain knowledge distillation for target state prediction. It comprises two novel components: 1) a Dual Interactive Mamba (DIM) module establishes bidirectional sequence interactions and facilitates context token exchange between differently augmented views of the target, effectively narrowing domain discrepancies. 2) a Hyperspherical Prototype Projection (HPP) module projects these tokens onto a set of multi-domain prototypes within a latent hyperspherical space. A dual alignment strategy is designed to minimize distribution differences among these prototypes, concurrently enforcing both domain-specific and domain-invariant properties for model generalization, enabling the tracker to adapt seamlessly across diverse weather conditions. To the best of our knowledge, this is the first work to address source-free domain adaptation in the VOT community.

In summary, the main contributions of this work are:

• We introduce SFDATrack, a novel framework that addresses the critical yet under-explored problem of source-free domain adaptive tracking under adverse weather conditions.

• We design a Dual Interactive Mamba (DIM) module to enhance the model's resilience to multiple weather conditions via bidirectional sequence interaction and context token exchange.

• We propose a Hyperspherical Prototype Projection (HPP) module to enforce both domain-specific and domain-invariant properties for model generalization. Extensive experiments demonstrate that SFDATrack achieves superior performance to existing state-of-the-art methods.

2 Related Work

2.1 Visual Object Tracking

Visual object tracking (VOT) aims to localize target objects given the initial annotation throughout a video sequence. With the success of deep convolutional networks, Siamese-based trackers [2, 7, 18, 24, 46] have achieved significant progress. Following the success of Transformers in vision tasks, several works unified feature extraction and target localization into end-to-end attention frameworks. STARK [44] proposed a spatio-temporal Transformer that directly regresses the target box via an encoder-decoder structure. MixFormer [9] and OSTrack [49] further adopted single-stream architectures to jointly learn global context and target-specific representation, achieving state-of-the-art performance on LaSOT [12] and GOT-10k [19]. Recently, Mamba-based models

[16,17] have introduced sequence-state modeling to overcome the quadratic complexity of attention and improve temporal modeling efficiency. MambaVLT [31] introduces time-evolving multimodal state-space modeling for efficient vision-language tracking. MambaLCT [28] employs a state-space model to dynamically update long-term contextual states. Despite this progress, most of these modern trackers are still trained and evaluated within a single domain, lacking sufficient generalization capability to unseen scenarios and suffering from significant performance degradation when facing domain shifts. To combat this, a series of domain adaptation (DA) tracking methods have been proposed. UDAT [51] develops a transformer-based unsupervised adaptation framework that bridges the feature discrepancy between daytime and nighttime domains. UMDATrack [48] adopts a mean-teacher framework with domain-specific adapters to achieve unified multi-domain adaptation under adverse weather conditions. However, such domain-adaptive trackers still rely on access to source-domain samples during adaptation, which may not always be feasible.

2.2 Source-Free Domain Adaptation

In many real-world scenarios, sharing source data is infeasible due to privacy, copyright, and transmission constraints, while retaining complete source datasets on resource-constrained devices is also impractical. To overcome these limitations of traditional unsupervised domain adaptation (UDA), source-free domain adaptation (SFDA) [8,30] has been proposed, which adapts a pretrained source model to a target domain without any access to source data, using only unlabeled target samples. Following [26], existing methods can be grouped into data-centric and model-centric paradigms. Data-centric approaches compensate for the missing source by reconstructing or transforming target data, including image restoration and enhancement techniques under adverse conditions [36]. SF(DA)^2 [20] constructs an augmentation graph to implicitly augment and cluster target features, thereby tightening the decision boundary, whereas SF-YOLO [37] introduces a learnable target-specific augmentation network in the image space to transform target data as a surrogate for the missing source. In the model-centric paradigm, SFDA primarily relies on self-training methods such as a teacher-student framework to ensure robustness and stability. However, noisy pseudo labels from the teacher can mislead the student and eventually cause the whole system to collapse through EMA feedback. DRU [22] dynamically retrains the student and selectively updates the EMA teacher based on the student's evolution. LPLD [52] distills knowledge from low-confidence predictions to enhance model generalization. While prior SFDA studies focus on static recognition tasks, tracking introduces unique challenges: predictions must remain temporally coherent, and noisy pseudo labels that accumulate across frames will cause drift.

3 Method

In this section, we present the overall architecture of SFDATrack. The whole model follows a mean-teacher pipeline, which conducts multiple-domains knowl-

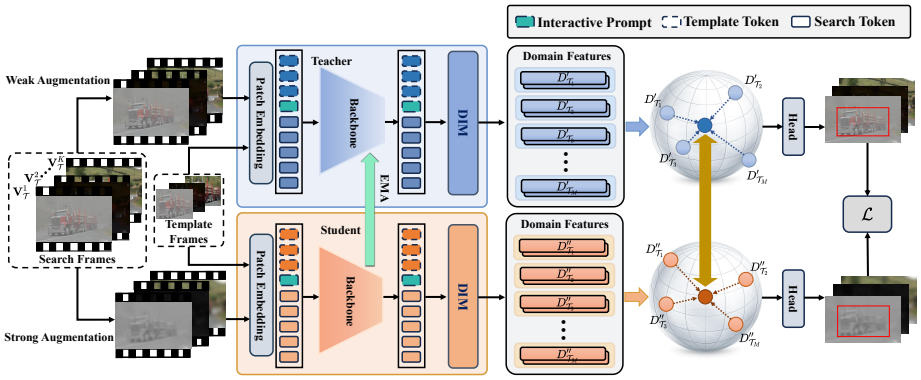

Fig. 2: Overview of the proposed SFDATrack. The input target video is first augmented into weak and strong views, which are processed by the teacher and student encoders, respectively. The extracted features are further fed into the Dual Interactive Mamba (DIM) module and Hyperspherical Prototype Projection (HPP) module to enhance domain consistency and representation robustness. Finally, a total loss $\mathcal{L}$ is applied to jointly optimize the tracking framework, ensuring stable and consistent feature learning across domains.

edge distillation for target object localization. SFDATrack consists of two crucial components: a Dual Interactive Mamba (DIM) module and a Hyperspherical Prototype Projection (HPP) module.

3.1 Problem Definition

Formally, let D_S denotes the source domain data and $\mathcal{M}_\mathcal{S}$ is the source domain pretrained model, the set of data from the target domains are denoted as $\mathcal{D}_\mathcal{T} = \{D_{\mathcal{T}_1}, D_{\mathcal{T}_2}, \cdots, D_{\mathcal{T}_M}\}$, where $\mathcal{D}_{\mathcal{T}_i} = \left\{x^j_{\mathcal{T}_i}\right\}_{j=1}^{N_{\mathcal{T}_i}}$ is the i-th target domain and $x^j_{\mathcal{T}_i}$ is the corresponding j-th image, M and $N_{\mathcal{T}_i}$ denote the number of target domains and target domain samples, respectively. In the source-free setting, the source data D_S are inaccessible during adaptation, and only the pre-trained source model $\mathcal{M}_S$ is available. Our goal is to adapt the pretrained tracking model $\mathcal{M}_S$ to the set of target domains $\mathcal{D}_\mathcal{T}$.

As shown in Fig. 2, SFDATrack follows a mean teacher pipeline to achieve source-free domain adaptation. Given K videos $\mathbb{V}_\mathcal{T} = \left\{\mathbf{V}^1_\mathcal{T}, \mathbf{V}^2_\mathcal{T}, \cdots, \mathbf{V}^K_\mathcal{T}\right\}$ from the target domains, we simultaneously perform weak and strong augmentations on the video frames. The teacher branch generates pseudo-labels from the weakly augmented video frames, while the student branch receives another strongly augmented version. Both networks share the same architecture but have different parameter sets $\theta^\mathcal{T}$ and $\theta^\mathcal{S}$. The teacher's parameters are updated through an Exponential Moving Average (EMA), *i.e.*, $\theta^\mathcal{T} \leftarrow \alpha\,\theta^\mathcal{T} + (1-\alpha)\,\theta^\mathcal{S}$, where α is the momentum coefficient controlling the update rate of the teacher model.

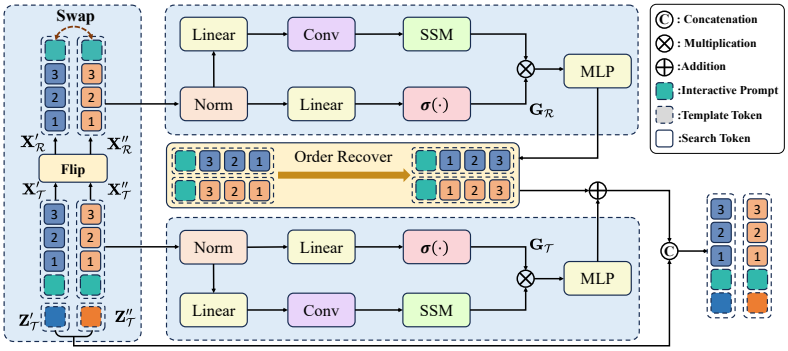

Fig. 3: Illustration of the Dual Interactive Mamba (DIM) module. The DIM module processes feature tokens from two different augmentations through bidirectional Mamba blocks with prompt tokens exchange. For simplicity, all prime and double-prime notations (e.g., $\mathbf{G}'_{\mathcal{T}}$, $\mathbf{G}''_{\mathcal{R}}$) are omitted in the figure.

3.2 Dual Interactive Mamba

Although the generic mean-teacher framework provides an efficient way for domain knowledge transfer, their effectiveness and generalizability in multiple weather conditions are greatly limited due to the significant discrepancies of the different target domains. To address the issue, we introduce a Dual Interactive Mamba (DIM) block to distill the candidate target tokens that are resilient to weather variations from classified, augmented samples.

Specifically, suppose we have obtained the encoded features $\mathbf{F}'_{\mathcal{T}} = [\mathbf{Z}'_{\mathcal{T}}, \mathbf{X}'_{\mathcal{T}}]$ and $\mathbf{F}''_{\mathcal{T}} = [\mathbf{Z}''_{\mathcal{T}}, \mathbf{X}''_{\mathcal{T}}]$ in the teacher and student networks, where $\mathbf{Z}'_{\mathcal{T}}$, $\mathbf{Z}''_{\mathcal{T}} \in \mathbb{R}^{L_Z \times C}$ denote the template tokens, $\mathbf{X}'_{\mathcal{T}}$, $\mathbf{X}''_{\mathcal{T}} \in \mathbb{R}^{L_X \times C}$ denote the candidate search tokens, L_Z, L_X and C are length and feature dimension of the template-search tokens, respectively. Note that the weak and strong augmented branches share the similar target object's visual semantics, we introduce a set of learnable interactive prompt tokens $\mathbf{P}'_{\mathcal{T}}$ and $\mathbf{P}''_{\mathcal{T}}$ to establish bidirectional sequence interactions and facilitates context token exchange. As shown in Fig. 3, we flip the dual search tokens and concatenate them with the interactive prompts for knowledge exchange as follows:

$$\begin{aligned} \mathbf{X}'_{\mathcal{R}} &= \text{Concat}(\mathbf{P}''; \mathcal{R}(\mathbf{X}'_{\mathcal{T}})), \\ \mathbf{X}''_{\mathcal{R}} &= \text{Concat}(\mathbf{P}'; \mathcal{R}(\mathbf{X}''_{\mathcal{T}})). \end{aligned} \tag{1}$$

where $\mathcal{R}(\cdot)$ denotes the reverse function of the search tokens. Then we normalize and linearly project $\mathbf{X}'_{\mathcal{T}}$ and $\mathbf{X}'_{\mathcal{R}}$ to obtain intermediate token representations. These representations are then passed through a sigmoid function to generate the gating vectors $\mathbf{G}'_{\mathcal{T}}$ and $\mathbf{G}'_{\mathcal{R}}$. The same process is applied to the corresponding tokens in the second branch to obtain $\mathbf{G}''_{\mathcal{T}}$ and $\mathbf{G}''_{\mathcal{R}}$, then we fed these tokens into a shared Mamba block, which can be given by:

$$\mathbf{Y}'_{\mathcal{T}} = \mathcal{SSM}(\psi(\mathbf{X}'_{\mathcal{T}})), \mathbf{Y}'_{\mathcal{R}} \quad = \mathcal{SSM}(\psi(\mathbf{X}'_{\mathcal{R}})). \tag{2}$$

where $\psi(\cdot)$ performs channel mixing through a point-wise convolution and a SiLU activation to refine feature channels and $\mathcal{SSM}(\cdot)$ denotes the Mamba state-space operator. Similarly, the second branch follows the same procedure to produce $\mathbf{Y}''_{\mathcal{T}}$ and $\mathbf{Y}''_{\mathcal{R}}$. Finally, we fuse the gated features from both branches to obtain the enhanced search tokens:

$$\begin{aligned} \mathbf{X}'_{\mathcal{T}} &= \mathcal{MLP}(\mathbf{G}'_{\mathcal{T}} \odot \mathbf{Y}'_{\mathcal{T}}) + \mathcal{R}(\mathcal{MLP}(\mathbf{G}'_{\mathcal{R}} \odot \mathbf{Y}'_{\mathcal{R}}))), \\ \mathbf{X}''_{\mathcal{T}} &= \mathcal{MLP}(\mathbf{G}''_{\mathcal{T}} \odot \mathbf{Y}''_{\mathcal{T}}) + \mathcal{R}(\mathcal{MLP}(\mathbf{G}''_{\mathcal{R}} \odot \mathbf{Y}''_{\mathcal{R}}))), \end{aligned} \tag{3}$$

By stacking multiple Dual Interactive Mamba blocks, the model progressively enhances the cross-view semantic alignment and strengthens the bidirectional information flow between the weak and strong augmentations.

3.3 Hyperspherical Prototype Projection

As the source domain data is not accessible in the adaptation process, distilling the original pretrained model into different target domains inevitably suffers significant performance degradation. To address this issue, we propose a *Hyperspherical Prototype Projection* (HPP) module to enhance generalizability. Specifically, we perform average pooling on the output search features $\mathbf{X}'_{\mathcal{T}}$ of DIM to be $\mathbf{X}'_{\mathcal{T}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$ and project these features onto a learnable prototype space $\mathcal{P} = \{\mathbf{p}_k\}_{k=1}^{K}$ [45], which can be given by:

$$s_{i,k} = \frac{\exp\left(\frac{1}{\tau}\,\mathbf{x}_i^\top \mathbf{p}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\,\mathbf{x}_i^\top \mathbf{p}_{k'}\right)}, \tag{4}$$

where τ is a temperature parameter. Thus we can construct a prototype similarity matrix $\mathbf{S} = \{s_{i,k}\}_{i=1,k=1}^{N,K} \in \mathbb{R}^{N \times K}$, where each row corresponds to the similarity distribution of a feature to all prototypes. We define a cost map $\mathbf{C}$ to measure the dissimilarity between features and prototypes as:

$$\mathbf{C}_{i,k} = 1 - \frac{\mathbf{x}_i^\top \mathbf{p}_k}{\|\mathbf{x}_i\|_2 \, \|\mathbf{p}_k\|_2}. \tag{5}$$

As it's not trivial to learn the prototype space representation, similar to [5], we introduce a soft assignment matrix $\mathbf{Q} \in \mathbb{R}^{N \times K}$ that enforces the features can be projected to the unit hypersphere as below:

$$\mathcal{Q} = \left\{ \mathbf{Q} \in \mathbb{R}_+^{N \times K} \mid \mathbf{Q}\mathbf{1}_K = \tfrac{1}{K}\mathbf{1}_N,\ \mathbf{Q}^\top \mathbf{1}_N = \tfrac{1}{N}\mathbf{1}_K \right\}, \tag{6}$$

where $\mathbf{1}_K$ and $\mathbf{1}_N$ denote all-one vectors of dimension K and N, respectively. This constraint ensures that each prototype can be equally assigned within a mini batch. The soft assignment matrix $\mathbf{Q}$ is obtained by minimizing the assignment cost and encourages smooth, balanced distributions:

$$\min_{\mathbf{Q} \in \mathcal{Q}} \langle \mathbf{Q}, \mathbf{C} \rangle - \varepsilon\, \mathcal{H}(\mathbf{Q}), \tag{7}$$

where $\mathcal{H}(\mathbf{Q}) = -\sum_{i,k} \mathbf{Q}_{i,k} \log \mathbf{Q}_{i,k}$ is the entropy term, and ε controls the degree of assignment smoothness. This optimization problem is efficiently solved via the Sinkhorn-Knopp algorithm [10].

Domain-Specific Alignment. For multi-domain adaptation tracking task, the augmented samples related to the same weather in both teacher-student networks should maintain the uniform localization capabilities. To achieve this, we propose a Domain-Specific Alignment (DSA) strategy to ensure the localization consistency. Suppose the augmented samples are collected from M domains, as described in Eq. 4, we send them into the student network and construct M domain-specific similarity matrix $\{\mathbf{S}_m^* \in \mathbb{R}^{N\times K}\}_{m=1}^M$ in the prototype space. Correspondingly, the teacher provides the soft assignment targets $\{\mathbf{Q}_m^* \in \mathbb{R}^{N\times K}\}_{m=1}^M$ as pseudo labels for M domains. We adopt a domain-specific loss function to align each student prediction with its corresponding teacher model at all of the weather domains:

$$\mathcal{L}_{\text{DSA}} = -\frac{1}{MNK} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{Q}_{m,i,k}^* \log \mathbf{S}_{m,i,k}^*. \tag{8}$$

Domain-Invariant Alignment. We propose a Domain-Invariant Alignment (DIA) strategy to enforce the consistency of multiple domains' representation in the latent hyperspherical space. Specifically, we first perform k-means clustering on the encoded DIM output $\mathbf{X}'_{\mathcal{T}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times D}$ to be M clusters $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]^\top \in \mathbb{R}^{M\times D}$, we calculate the normalized clustering weight $\mathbf{W} = \{w_{i,m}\}_{i=1,m=1}^{N,M}$ by:

$$w_{i,m} = \frac{\left(1 + \|\mathbf{x}_i - \mathbf{v}_m\|^2\right)^{-1}}{\sum_j \left(1 + \|\mathbf{x}_i - \mathbf{v}_j\|^2\right)^{-1}}, \tag{9}$$

where $\mathbf{v}_m$ is the m-th clustered center, $w_{i,m}$ indicates the weight of the i-th sample belonging to the m-th domain. Then we construct a clustered domain-invariant representation $\mathbf{X}^* = \text{MLP}(\mathbf{WV}) \in \mathbb{R}^{N\times D}$ by fusing the cluster-weighted features from different domains using a lightweight multilayer perceptron.

Similarly to DSA, we project the clustered domain-invariant feature $\mathbf{X}^*$ into latent hyperspherical space [33, 47], and construct a domain-invariant similarity matrix $\hat{\mathbf{S}}^* \in \mathbb{R}^{N\times K}$. The teacher network generates the soft assignment targets $\mathbf{Q}^* \in \mathbb{R}^{N\times K}$. The domain-invariant alignment loss is defined as:

$$\mathcal{L}_{\text{DIA}} = \text{KL}(\mathbf{Q}^* \| \mathbf{S}^*) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{Q}_{i,k}^* \log \frac{\mathbf{Q}_{i,k}^*}{\mathbf{S}_{i,k}^*}. \tag{10}$$

The loss function in Eq. 10 enforces the distribution alignment of the clustered domain-invariant features in teacher-student networks, allowing the tracker to alleviate the model biases to a specific domain.

Table 1: Performance comparisons with state-of-the-art trackers on synthetic datasets: GOT-10k-Foggy, DTB70-Foggy, GOT-10k-Dark, DTB70-Dark, GOT-10k-Rainy and DTB70-Rainy. The top three results are highlighted with red, blue and green fonts, respectively. Trackers above the double line denote **generic** methods, and those below denote **cross-domain** methods.

Tracker	GOT-10k-Foggy			DTB70-Foggy		GOT-10k-Dark			DTB70-Dark		GOT-10k-Rainy			DTB70-Rainy	
	AO	$SR_{0.50}$	$SR_{0.75}$	AUC	P	AO	$SR_{0.50}$	$SR_{0.75}$	AUC	P	AO	$SR_{0.50}$	$SR_{0.75}$	AUC	P
SiamRPN++ [23]	58.4	64.9	51.2	55.80	74.70	56.6	60.8	45.1	48.80	70.30	56.2	61.4	46.8	51.52	71.96
DiMP [3]	57.6	64.2	50.4	53.80	69.50	56.9	60.4	44.3	55.20	72.30	57.9	63.8	49.2	57.32	75.21
AVTrack [29]	56.9	63.5	49.5	52.35	68.09	55.3	62.3	46.2	56.66	72.21	57.5	63.4	48.1	60.21	79.53
OSTrack [49]	61.9	71.7	59.7	57.54	73.95	61.3	70.9	51.5	59.23	77.43	61.6	71.0	58.6	63.16	83.60
ROMTrack [4]	63.6	70.9	56.7	59.05	76.59	60.8	71.1	51.7	60.80	77.95	62.7	73.4	60.1	63.21	83.25
AQATrack [42]	64.9	72.8	59.7	57.28	75.61	61.7	70.6	52.5	61.17	79.87	63.4	72.3	61.8	63.12	83.55
SeqTrack [6]	65.2	74.6	56.3	60.21	78.70	61.4	70.5	52.3	62.84	81.57	65.1	75.0	60.3	63.75	83.28
DropTrack [40]	64.9	73.8	58.5	59.95	77.66	62.2	72.5	54.3	61.98	80.21	65.3	75.3	60.4	62.87	83.13
EVPTrack [35]	63.5	70.7	56.5	57.96	75.45	62.7	71.8	53.9	63.01	81.12	65.5	75.2	60.5	64.03	84.11
ARTrackV2 [1]	64.8	73.0	59.9	62.25	80.15	63.1	72.8	53.9	62.87	80.56	66.2	75.8	61.2	63.84	83.32
UncTrack [45]	63.9	72.4	59.0	59.02	76.79	63.5	72.8	56.2	59.72	77.14	63.8	71.8	57.6	65.09	84.57
ORTrack [41]	60.7	70.0	51.2	54.94	71.92	56.9	67.2	43.4	56.29	71.75	63.3	72.6	55.7	65.40	84.70
SGLATrack [43]	62.4	72.7	54.4	55.28	71.64	56.0	65.9	43.8	58.79	74.97	63.8	73.8	57.5	63.93	83.30
UETrack [21]	63.1	71.9	56.5	59.19	74.64	65.3	75.1	58.0	60.95	77.02	66.3	75.7	59.8	66.73	84.42
MLKD-Track [32]	52.3	62.3	49.1	52.46	70.32	53.8	61.6	46.9	55.21	73.68	57.3	64.8	57.1	56.89	74.12
SAM-DA [14]	50.2	60.5	48.3	51.33	69.89	55.4	63.1	48.3	57.15	75.12	60.2	66.1	57.6	57.63	76.12
UDAT-CAR [51]	51.5	60.3	45.2	50.21	69.41	56.8	64.2	49.1	57.20	75.80	59.5	65.2	55.3	56.42	75.36
DCPT [55]	61.6	70.2	56.9	58.31	75.33	62.4	70.5	54.2	61.87	80.11	62.3	70.1	59.8	61.68	82.56
LVPTrack [39]	64.5	74.0	58.4	63.51	83.33	62.2	72.6	53.4	64.29	82.36	67.9	77.0	60.6	65.32	84.65
UMDATrack [48]	66.6	75.8	62.2	66.21	86.05	65.4	75.3	57.3	66.07	85.72	68.5	78.4	63.2	66.75	87.60
SFDATrack	70.1	80.6	67.1	65.58	86.47	67.3	78.0	62.2	66.18	85.96	72.1	82.6	70.1	67.22	88.32

To train SFDATrack, we integrate the target supervision loss, the domian-specific and domain-invariant alignment loss to achieve robust source-free domain adaptation. The target supervision loss $\mathcal{L}_S$ consists of the classification, localization L1, and generalized IoU terms following [49]:

$$\mathcal{L}_S = \mathcal{L}_{CLS} + \beta\,\mathcal{L}_1 + \gamma\,\mathcal{L}_{GIoU}, \tag{11}$$

The whole training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_S + \omega\,\mathcal{L}_{DSA} + \lambda\,\mathcal{L}_{DIA}, \tag{12}$$

where ω and λ balances the contribution of the above losses.

4 Experiments

4.1 Implementation Details

Experiments Settings. Our SFDATrack is implemented using Python 3.9 and PyTorch 2.4.1. The model is trained with 2 NVIDIA A100 GPUs and tested on a single NVIDIA A6000 GPU. We employ a vanilla ViT-Base [11] network as the backbone of our tracker. The patch size is set to 16×16 and the feature dimension is 768. We insert three Dual Interactive Mamba (DIM) modules into the network. The template and search region sizes are set to 128×128 and 256×256, respectively. We train the model for 300 epochs and the network is

Table 2: Comparisons with state-of-the-art trackers on real-world datasets: NAT2021, UAVDark70, and AVisT. The best results are shown in bold. Trackers above the double line denote **generic** methods, and those below denote **cross-domain** methods.

Tracker	NAT2021		UAVDark70		AVisT	
	AUC	P	AUC	P	AUC	P
AVTrack [29]	45.41	59.51	46.91	59.49	49.21	48.50
SMAT [15]	45.96	59.87	45.19	56.71	50.35	49.58
AQATrack [42]	51.33	67.03	58.18	70.98	57.32	56.60
SeqTrack [6]	51.65	67.97	53.88	66.88	57.15	55.30
ROMTrack [4]	51.57	68.75	53.77	69.80	56.12	55.09
EVPTrack [35]	53.08	69.51	57.47	71.10	57.31	55.55
ODTrack [54]	53.11	69.68	58.07	71.11	58.63	57.36
ARTrackV2 [1]	53.13	69.72	58.22	71.95	58.52	57.65
ORTrack [41]	49.62	63.90	47.43	59.83	48.17	49.10
SGLATrack [43]	47.52	61.75	53.06	67.32	50.24	50.82
MLKD-Track [32]	44.31	60.21	47.27	61.54	33.62	30.26
SAM-DA [14]	47.31	65.50	49.52	65.59	37.36	34.29
UDAT-CAR [51]	48.75	65.96	51.25	70.22	38.91	33.65
DCPT [55]	52.55	69.01	56.86	70.16	55.66	52.41
LVPTrack [39]	51.74	68.61	58.40	74.53	52.93	51.63
UMDATrack [48]	54.58	70.78	60.05	73.35	60.50	59.01
SFDATrack	**56.78**	**73.60**	**61.22**	**76.51**	**60.87**	**59.70**

optimized using AdamW with a learning rate of 4×10^{-4} and a weight decay of 1×10^{-4}. The exponential moving average (EMA) momentum α is set to 0.99. The weighting coefficients for the L1, GIoU, DSA and DIA losses are set to 5.0, 2.0, 0.5 and 0.5, respectively.

Datasets. For the source-free domain adaptation training, we use three synthetic datasets: GOT-10k-Dark, GOT-10k-Foggy, and GOT-10k-Rainy generated in [48]. Each dataset corresponds to a specific adverse condition (darkness, fog, and rain), and the video frames are sampled with a balanced sampling ratio of 1:1:1. A total of 60,000 samples are uniformly drawn for training. These datasets simulate different adverse weather scenarios and are used as unlabeled target domains.

4.2 Comparisons with State-of-the-art Trackers

This subsection provides a comprehensive comparison of SFDAtrack with recent state-of-the-art (SOTA) trackers under diverse real-world and synthetic adverse weather conditions. Although our approach is specifically designed for cross-domain tracking, it consistently outperforms existing SOTA generic trackers, highlighting its strong generalization and robustness.

Specifically, we use synthetic datasets that include three variants from GOT-10k and DTB70, namely Dark, Foggy, and Rainy, resulting in six weather-specific

Table 3: Comparison of inference speed, MACs, and model parameters across different trackers.

Tracker	Speed (FPS)	MACs (G)	Parameters (M)
SeqTrack [6]	40	66	89
DropTrack [40]	52	48	92
AQATrack [42]	68	26	72
EVPTrack [35]	71	22	74
ARTrackV2 [1]	95	45	126
DCPT [55]	30	30	95
SFDATrack	**91**	**30**	**93**

domains in total. These variants are designed to simulate nighttime, foggy, and rainy environments, respectively. In addition, two real-world datasets, NAT2021-test [51] and UAVDark70 [25], are employed to evaluate tracking performance under realistic nighttime conditions. Finally, we evaluate the tracking performance on the real-world AVisT dataset [34], which contains various challenging weather conditions in natural scenes.

Synthetic GOT-10k and DTB70 [27]. As shown in Table 1, SFDATrack exhibits outstanding performance across various adverse weather conditions on both synthetic datasets. On the synthetic GOT-10k dataset, SFDATrack consistently achieves the best results across all adverse weather conditions, leading the second-best tracker by notable margins, such as 1.9% in AO, 2.7% in $SR_{0.5}$, and 4.9% in $SR_{0.75}$ under dark scenarios. Similarly, on the synthetic DTB70 dataset, SFDATrack achieves the highest AUC and precision scores under almost all weather conditions.

Results on Real-World Datasets Besides achieving outstanding performance on synthetic datasets, we further evaluate SFDATrack on real-world benchmarks to verify its generalization ability in practical scenarios. As shown in Table 2,

Table 4: Ablation study on the individual impact of each module (DIM, HPP) in our model. The presence or absence of each module is marked with a check or dash, respectively. The performance is evaluated on the NAT2021 dataset.

Modules		**Indicators**	
DIM	HPP	AUC (%)	Precision (%)
-	-	51.22	67.62
✓	-	53.56	70.88
-	✓	53.24	70.37
✓	✓	**56.78**	**73.60**

Table 5: Ablation study on different alignment strategies within the HPP module. The compared variants include Domain-Specific Alignment (DSA) and Domain-Invariant Alignment (DIA). Results are evaluated on the UAVDark70 dataset.

Modules		**Indicators**	
DSA	DIA	AUC (%)	Precision (%)
-	-	58.29	72.37
✓	-	59.65	74.07
-	✓	59.55	73.91
✓	✓	**61.22**	**76.51**

Table 6: Ablation study on the impact of different architectures within the DIM module. The compared variants include Transformer, Uni-Mamba, Bi-Mamba, and our full DIM module. The results are evaluated on the UAVDark70 dataset.

Architecture	AUC (%)	Precision (%)
Transformer [38]	56.77	70.45
Uni-Mamba [16]	58.85	74.58
Bi-Mamba [56]	60.33	75.60
DIM (ours)	**61.22**	**76.51**

Table 7: Experiments on the weighting coefficients of Domain-Specific Alignment (DSA) ω and Domain-Invariant Alignment (DIA) λ. Results are reported in terms of AUC and Precision on the NAT2021 dataset.

ω	λ	AUC (%)	Precision (%)
0.1	0.1	56.16	73.20
0.1	0.5	55.09	71.67
0.5	0.1	55.37	72.12
0.5	**0.5**	**56.78**	**73.60**

SFDATrack exhibits superior performance across all three real-world datasets. On the NAT2021 dataset, it achieves the highest AUC and precision scores of 56.78 and 73.60, respectively, outperforming the second tracker UMDATrack by 2.20% and 2.82%. On UAVDark70 dataset, SFDATrack again achieves the best AUC (61.22) and precision (76.51), surpassing the previous best tracker by 1.17% and 3.16%. Furthermore, on the challenging AVisT dataset, which involves various adverse weather and illumination conditions, SFDATrack attains the best AUC (59.64) and precision (58.25), clearly outperforming other advanced trackers. These results demonstrate the strong generalization ability and robustness of SFDATrack when applied to real-world tracking scenarios.

Table 8: The study on the update frequency of EMA (Exponential Moving Average).

EMA Frequency	AUC (%)	Precision (%)
Each batch	55.20	72.47
Every 5 epochs	54.32	70.41
Each epoch	**56.78**	**73.60**

Inference Speed. As demonstrated in Table 3, SFDATrack achieves a good balance between accuracy and efficiency. Compared with existing state-of-the-art trackers, SFDATrack achieves a similar inference speed to ARTrackV2, while requiring fewer parameters and lower computational cost. Moreover, SFDATrack runs notably faster than most other trackers, such as SeqTrack (40 FPS) and DropTrack (52 FPS), demonstrating its superior efficiency and real-time capability.

4.3 Ablation Studies

Study on the Components of SFDATrack. We conducted ablation experiments on the proposed two modules to verify their effectiveness. As shown in

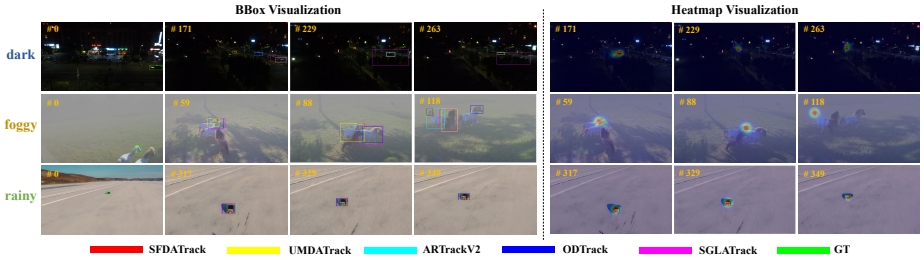

Fig. 4: Visualization results of SFDATrack under three challenging weather conditions: dark, foggy, and rainy scenes. The left part shows the bounding box (BBox) visualizations, while the right part presents the corresponding heatmaps.

Table 4, when the DIM module is introduced, the model achieves an AUC of 2.34 points and a Precision of 3.26 points higher than the baseline. When adding the HPP module, it improves the AUC to 53.24 and the Precision to 70.37. While both the DIM module and HPP module are introduced, it achieves the best performance with an AUC of 56.78 and Precision of 73.60. These results indicate that each module plays an important role in the whole model.

Study on the Architecture of HPP. We further conduct ablation experiments to validate the effectiveness of the proposed HPP module. As shown in Table 5, when the Domain-Specific Alignment (DSA) is introduced, the AUC and Precision are improved by 1.36% and 1.70%, respectively. And when the Domain-Invariant Alignment (DIA) is introduced, the AUC and Precision are improved by 1.26% and 1.54%. However, when both DSA and DIA strategies are incorporated, the model achieves the best performance with an AUC of 61.22 and a Precision of 76.51 on the UAVDark70 dataset.

Study on the Architecture of DIM. We investigate the effect of different architectural designs within the DIM module. As shown in Table 6, replacing the Transformer [38] with the Mamba-based [16] structure brings a clear performance gain. Specifically, Uni-Mamba (Unidirectional Mamba) achieves an improvement of 2.08% in AUC and 4.15% in Precision over the Transformer baseline. Introducing bidirectional modeling in Bi-Mamba [56] further improves AUC by 1.48% and Precision by 1.18% over the Uni-Mamba. Finally, our full DIM achieves the best results in the UAVDark70 dataset, with AUC of 61.22 and Precision of 76.51, confirming the effectiveness of the proposed bidirectional Mamba design.

Study on Training Hyperparameters. We further analyze the loss weights of Domain-Specific Alignment (DSA) and Domain-Invariant Alignment (DIA), as well as the EMA update frequency. As shown in Table 7, a balanced weighting between DSA and DIA (0.5, 0.5) performs best, while imbalanced settings degrade performance, indicating that both terms are necessary for effective adaptation. Moreover, Table 8 indicates that updating EMA once per epoch yields the most favorable performance.

4.4 Visualization Results and Generalized Ability

Visualization in Adverse Conditions. Fig. 4 illustrates the qualitative tracking results, where SFDATrack exhibits higher accuracy and stronger robustness than other trackers under extreme weather conditions. Even in real nighttime scenarios with small, barely visible targets, SFDATrack is still able to localize the object clearly. The corresponding heatmaps further demonstrate that SFDATrack focuses more precisely on target regions across frames, maintaining concentrated attention in challenging scenes.

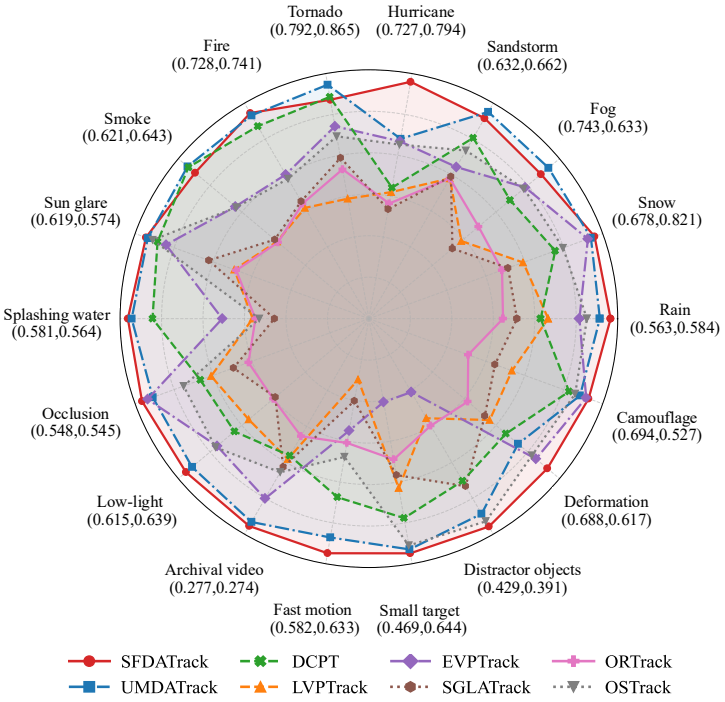

Fig. 5: The performance of our method compared with other trackers in terms of AUC and Precision on eighteen attributes of the real-world AVisT dataset. Each axis corresponds to one attribute, with the radius representing the AUC value under the corresponding condition.

Generalized Ability. As shown in Fig. 5, we further analyze the performance of SFDATracker on different challenging attributes of the AVisT dataset. Specifically, SFDATrack achieves the best performance across most adverse weather scenarios and challenging attributes, such as hurricane, rain, snow, low-light, fast motion, among others. For several other challenging conditions, such as fog and sandstorm, it achieves the second-best performance, demonstrating strong robustness and generalization ability under diverse adverse visibility settings.

5 Conclusion

In this work, we introduce a generalized source-free domain adaptive tracker termed SFDATrack that only leverages adverse weather samples from the target domain for robust state estimation. We propose a Dual Interactive Mamba module that uses bidirectional sequence interaction and prompt token exchange strategies to narrow domain discrepancies. Furthermore, we design a Hyperspherical Prototype Projection module to project tokens onto multi-domain prototypes within a latent hyperspherical space, enforcing both domain-specific and domain-invariant properties for model generalization. Extensive experiments on multiple adverse weather datasets demonstrate the effectiveness of the proposed SFDATrack.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62402055, 62302053, 62322216, and U2541229) and the Shenzhen Science and Technology Program (Grant No. KQTD20221101093559018 and SYSRD20250529113401002).

References

1. Bai, Y., Zhao, Z., Gong, Y., Wei, X.: Artrackv2: Prompting autoregressive tracker where to look and how to describe. In: CVPR. pp. 19048–19057 (2024)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCVW. pp. 850–865 (2016)
3. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: ICCV. pp. 6182–6191 (2019)
4. Cai, Y., Liu, J., Tang, J., Wu, G.: Robust object modeling for visual tracking. In: ICCV. pp. 9589–9600 (2023)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS. pp. 9912–9924 (2020)
6. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: Sequence to sequence learning for visual object tracking. In: CVPR. pp. 14572–14581 (2023)
7. Chen, Z., Shi, F.: Double template fusion based siamese network for robust visual object tracking. Journal of Image and Graphics **27**(4), 1191–1203 (2022)
8. Chidlovskii, B., Clinchant, S., Csurka, G.: Domain adaptation in the absence of source domain data. In: ACM KDD. pp. 451–460 (2016)
9. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: CVPR. pp. 13598–13608 (2022)
10. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NeurIPS. pp. 2292–2300 (2013)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

12. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: CVPR. pp. 5374–5383 (2019)
13. Fu, C., Dong, H., Ye, J., Zheng, G., Li, S., Zhao, J.: Highlightnet: Highlighting low-light potential features for real-time UAV tracking. In: IROS. pp. 12146–12153 (2022)
14. Fu, C., Yao, L., Zuo, H., Zheng, G., Pan, J.: SAM-DA: UAV Tracks Anything at Night with SAM-Powered Domain Adaptation. In: ICARM. pp. 31–38 (2024)
15. Gopal, G.Y., Amer, M.A.: Separable self and mixed attention transformers for efficient object tracking. In: WACV. pp. 6708–6717 (2024)
16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
17. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: CVPR. pp. 25261–25270 (2025)
18. He, A., Luo, C., Tian, X., Zeng, W.: A twofold siamese network for real-time object tracking. In: CVPR. pp. 4834–4843 (2018)
19. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE TPAMI **43**(5), 1562–1577 (2019)
20. Hwang, U., Lee, J., Shin, J., Yoon, S.: Sf(da)^2: Source-free domain adaptation through the lens of data augmentation. In: ICLR (2024)
21. Kang, B., Zhao, J., Chen, X., Geng, W., Zhang, B., Zhang, L., Wang, D., Lu, H.: Uetrack: A unified and efficient framework for single object tracking. In: CVPR. pp. 20890–20901 (2026)
22. Khanh, T.L.B., Nguyen, H., Pham, L.H., Tran, D.N., Jeon, J.W.: Dynamic retraining-updating mean teacher for source-free object detection. In: ECCV. pp. 328–344 (2024)
23. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
24. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR. pp. 8971–8980 (2018)
25. Li, B., Fu, C., Ding, F., Ye, J., Lin, F.: Adtrack: Target-aware dual filter learning for real-time anti-dark UAV tracking. In: ICRA. pp. 496–502 (2021)
26. Li, J., Yu, Z., Du, Z., Zhu, L., Shen, H.T.: A comprehensive survey on source-free domain adaptation. IEEE TPAMI **46**(8), 5743–5762 (2024)
27. Li, S., Yeung, D.: Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In: AAAI. pp. 4140–4146 (2017)
28. Li, X., Zhong, B., Liang, Q., Li, G., Mo, Z., Song, S.: Mambalct: Boosting tracking via long-term context state space model. In: AAAI. pp. 4986–4994 (2025)
29. Li, Y., Liu, M., Wu, Y., Wang, X., Yang, X., Li, S.: Learning adaptive and view-invariant vision transformer for real-time uav tracking. In: ICML. pp. 28403–28420 (2024)
30. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML. pp. 6028–6039 (2020)
31. Liu, X., Zhou, L., Zhou, Z., Chen, J., He, Z.: Mambavlt: Time-evolving multimodal state space model for vision-language tracking. In: CVPR. pp. 8731–8741 (2025)
32. Liu, Y.: Mutual-learning knowledge distillation for nighttime UAV tracking. arXiv preprint arXiv:2312.07884 (2023)
33. Mettes, P., van der Pol, E., Snoek, C.: Hyperspherical prototype networks. In: NeurIPS. pp. 1485–1495 (2019)

34. Noman, M., Ghallabi, W.A., Kareem, D., Mayer, C., Dudhane, A., Danelljan, M., Cholakkal, H., Khan, S., Gool, L.V., Khan, F.S.: Avist: A benchmark for visual object tracking in adverse visibility. In: BMVC. p. 817 (2022)
35. Shi, L., Zhong, B., Liang, Q., Li, N., Zhang, S., Li, X.: Explicit visual prompts for visual object tracking. In: AAAI. pp. 4838–4846 (2024)
36. Shi, Z., Liu, C., Ren, W., Du, S., Zhao, M.: Convolutional neural networks for sand dust image color restoration and visibility enhancement. Journal of Image and Graphics **27**(5), 1493–1508 (2022)
37. Varailhon, S., Aminbeidokhti, M., Pedersoli, M., Granger, E.: Source-free domain adaptation for yolo object detection. In: ECCVW. pp. 218–235 (2024)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
39. Wu, H., Yao, S., Huang, F., Wang, S., Zhang, L., Zheng, Z., Ren, W.: Lvptrack: High performance domain adaptive UAV tracking with label aligned visual prompt tuning. In: AAAI. pp. 8395–8403 (2025)
40. Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B.: Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: CVPR. pp. 14561–14571 (2023)
41. Wu, Y., Wang, X., Yang, X., Liu, M., Zeng, D., Ye, H., Li, S.: Learning occlusion-robust vision transformers for real-time uav tracking. In: CVPR. pp. 17103–17113 (2025)
42. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporal transformers. In: CVPR. pp. 19300–19309 (2024)
43. Xue, C., Zhong, B., Liang, Q., Zheng, Y., Li, N., Xue, Y., Song, S.: Similarity-guided layer-adaptive vision transformer for uav tracking. In: CVPR. pp. 6730–6740 (2025)
44. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: ICCV. pp. 10428–10437 (2021)
45. Yao, S., Guo, Y., Yan, Y., Ren, W., Cao, X.: Unctrack: Reliable visual object tracking with uncertainty-aware prototype memory network. IEEE TIP **34**, 3533–3546 (2025)
46. Yao, S., Han, X., Zhang, H., Wang, X., Cao, X.: Learning deep lucas-kanade siamese network for visual tracking. IEEE TIP **30**, 4814–4827 (2021)
47. Yao, S., Sun, H., Xiang, T., Wang, X., Cao, X.: Hierarchical graph interaction transformer with dynamic token clustering for camouflaged object detection. IEEE TIP **33**, 5936–5948 (2024)
48. Yao, S., Zhu, R., Wang, Z., Ren, W., Yan, Y., Cao, X.: Umdatrack: Unified multi-domain adaptive tracking under adverse weather conditions. In: ICCV. pp. 6466–6475 (2025)
49. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV. pp. 341–357 (2022)
50. Ye, J., Fu, C., Zheng, G., Cao, Z., Li, B.: Darklighter: Light up the darkness for UAV tracking. In: IROS. pp. 3079–3085 (2021)
51. Ye, J., Fu, C., Zheng, G., Paudel, D.P., Chen, G.: Unsupervised domain adaptation for nighttime aerial tracking. In: CVPR. pp. 8886–8895 (2022)
52. Yoon, I., Kwon, H., Kim, J., Park, J., Jang, H., Sohn, K.: Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In: ECCV. pp. 337–353 (2024)
53. Zhang, Z., Zhang, L.: Domain adaptive siamrpn++ for object tracking in the wild. arXiv preprint arXiv:2106.07862 (2021)

54. Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X.: Odtrack: Online dense temporal token learning for visual tracking. In: AAAI. pp. 7588–7596 (2024)
55. Zhu, J., Tang, H., Cheng, Z., He, J., Luo, B., Qiu, S., Li, S., Lu, H.: DCPT: darkness clue-prompted tracking in nighttime uavs. In: ICRA. pp. 7381–7388 (2024)
56. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: ICML. pp. 62429–62442 (2024)