SPATIAL-FREQUENCY COLLABORATIVE MAMBA TRACKER FOR VISUAL OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual object tracking has achieved substantial progress, yet robust tracking in complex dynamic scenes remains challenging because spatial-domain appearance representations are vulnerable to fast motion, illumination variations, occlusions, and background distractors. Existing Mamba-based trackers improve long-range dependency modeling and computational efficiency, but mainly focus on spatial-temporal representations, leaving the temporal evolution of frequency-domain motion information insufficiently explored. To address this limitation, we propose a spatial-frequency collaborative tracking framework based on Mamba, termed SFMT, which jointly models spatial-temporal appearance representations and frequency-domain motion cues. First, a Frequency-Domain Awareness (FDA) module employs Mamba to model the temporal evolution of phase spectra and uses the resulting motion cues to guide amplitude filtering. Second, a Spatial-Frequency Fusion (SFF) module performs global sequence-level and local point-level interaction to effectively integrate complementary information across the two domains. Finally, a pyramid multi-scale decoder enhances high- and low-frequency representations and dynamically balances their contributions through gated adaptive fusion for accurate target localization. Extensive experiments on six challenging benchmarks demonstrate the effectiveness of SFMT, achieving competitive tracking performance, including an AUC of 74.1% on LaSOT and an AO of 77.2% on GOT-10K.

1 INTRODUCTION

Visual object tracking Chen et al. (2022b); Wei et al. (2023) is one of the fundamental tasks in computer vision. Given the target location in the first frame of a video, the tracker is required to continuously predict the target's spatial position and bounding box in subsequent frames. This technology has been widely applied in areas such as intelligent surveillance Zheng et al. (2021), autonomous driving Chen et al. (2015), robotic perception Tuscher et al. (2021), and human–computer interaction.

Although deep learning has significantly advanced object tracking, robust tracking in complex dynamic scenarios remains challenging. During video sequences, targets may experience fast motion, severe deformation, scale variations, partial occlusions, and viewpoint changes. Meanwhile, environmental factors such as background clutter, similar objects, illumination variations, and motion blur can substantially degrade target representations, resulting in unreliable feature matching and potential tracking drift Zhang et al. (2022); Zhu et al. (2023). Existing tracking methods Chen et al. (2023); Gao et al. (2022) typically concatenate the template and search region into a unified visual sequence and employ self-attention mechanisms for feature interaction. Although these approaches provide effective global interaction, their computational cost increases rapidly with sequence length. Recently, State Space Models (SSMs) Gu et al. (2021) have attracted increasing attention due to their efficient long-range dependency modeling with near-linear complexity. Mamba-based trackers Hu et al. (2025) further exploit selective state space modeling and multi-directional scanning to improve temporal information utilization and visual representation learning. However, existing Mamba-based trackers mainly focus on spatial-domain appearance modeling, while the potential of frequency-domain motion information remains largely unexplored. Since spatial representations are vulnerable to appearance variations caused by fast motion, illumination changes, background

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

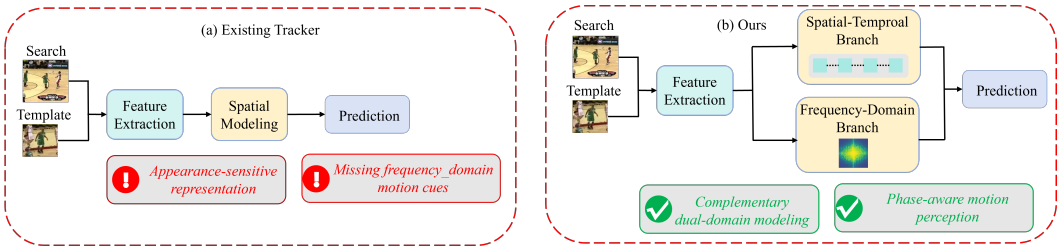

Figure 1: Comparison between existing spatial-domain trackers and ours. We introduce frequency-domain motion modeling to complement spatial representations.

clutter, and occlusions, relying solely on spatial-domain features is insufficient for robust tracking in complex scenarios.

As illustrated in Fig. 1, existing trackers mainly perform target localization based on spatial-domain representations, which may suffer from limited motion perception and sensitivity to appearance changes. In contrast, our method introduces an additional frequency-domain motion modeling branch to complement spatial representations, enabling the tracker to exploit discriminative motion cues for robust target localization.

Motivated by the above observation, we explore frequency-domain representations as a complementary information source beyond conventional spatial features. The frequency domain provides rich structural and motion-related cues for visual tracking Qin et al. (2021); Cui et al. (2009). Specifically, low-frequency components mainly capture the global structure and slowly varying semantic information of the target, while high-frequency components preserve detailed information such as edges, textures, and local variations. Furthermore, the Fourier spectrum can be decomposed into amplitude and phase components, where the amplitude spectrum mainly reflects appearance-related information and the phase spectrum is closely associated with structural layouts and positional changes. According to the Fourier shift property Fleet & Jepson (1990), target displacement in the spatial domain induces systematic variations in the phase spectrum, making phase information particularly suitable for capturing motion dynamics. Based on this insight, we propose a Frequency-Domain Awareness (FDA) module that explicitly models frequency-domain motion cues and exploits phase evolution to complement spatial representations, thereby improving tracking robustness under challenging scenarios.

Although frequency-domain representations provide complementary motion cues, effectively integrating them with spatial features remains challenging. Frequency-domain and spatial-domain features exhibit different distributions and semantic characteristics, and naive addition or single-level concatenation may lead to feature misalignment and information interference. To address this issue, we design a Spatial-Frequency Fusion (SFF) module to facilitate effective interaction between spatial-temporal and frequency-domain representations.

In the target localization stage, we further propose a pyramid multi-scale decoder with adaptive high-low frequency fusion to enhance localization capability under scale variations, occlusions, and motion blur.

The main contributions of this paper are summarized as follows:

- We propose a spatial-frequency collaborative tracking framework based on Mamba, which integrates spatial-temporal representations with frequency-domain motion cues for robust visual tracking.
- We design a Frequency-Domain Awareness (FDA) module to exploit phase-based motion cues and guide frequency representation learning, improving the robustness against motion variations and appearance interference.
- We develop a Spatial-Frequency Fusion (SFF) module and a pyramid multi-scale decoder with gated adaptive fusion to effectively integrate complementary spatial-frequency information and enhance target localization under challenging scenarios.

2 Related Work

2.0.1 Visual Mamba.

State Space Models (SSMs) Gu et al. (2021), as a family of sequence modeling approaches, have recently attracted significant attention due to their ability to capture long-range dependencies while maintaining linear computational complexity. Mamba Gu & Dao (2024) was initially introduced as a language sequence modeling approach within the domain of natural language processing. Mamba has also been widely applied to tasks such as multimodal learning Qiao et al. (2024) and image classification Gong et al. (2025). Vim Zhu et al. (2024) designs positional embeddings and bidirectional state-space modeling to achieve better long-sequence modeling. VideoMamba Li et al. (2024) introduces Mamba into video tasks to address local redundancy and global dependency problems.

In visual tracking tasks, MambaLCT Li et al. (2025a) leverages Mamba to scan search frames along the temporal axis, selectively compressing target-relevant information from historical frames into hidden states to construct long-term context spanning from the initial frame to the current frame. However, existing Mamba-based tracking methods mainly focus on spatial-temporal feature modeling, while the potential of frequency-domain temporal evolution information for motion perception has not yet been fully explored. In this paper, we introduce frequency-domain motion modeling into the Mamba tracking framework, enabling the joint utilization of spatial-temporal information and frequency-temporal information, thereby improving the robustness of object tracking in complex scenes.

2.0.2 Frequency Learning.

Frequency-domain representations Qin et al. (2021); Cui et al. (2009) provide complementary information beyond spatial features for visual understanding. Low-frequency components mainly preserve the overall structures and semantic information of objects, while high-frequency components contain detailed textures and local variations. In addition, amplitude and phase in the frequency spectrum exhibit different representation characteristics, where amplitude reflects the energy distribution related to object appearance, while phase contains structural and positional information.

Recently, some methods Zhong et al. (2022) introduce explicit frequency cues and frequency-aware losses to provide additional supervision in the frequency domain. Other approaches employ frequency decomposition strategies, such as Octave Convolution Cong et al. (2023), to decouple high-frequency and low-frequency representations for more effective feature learning. Furthermore, attention-based methods Sun et al. (2024) further model the interactions among different frequency components.

Existing methods mainly rely on convolution or self-attention mechanisms for frequency learning, but overlook the long-range dependencies within frequency component sequences. In visual tracking tasks, the phase variations caused by target motion contain important temporal motion cues, which have not been fully exploited. Therefore, this work introduces a frequency-domain motion perception module into the tracking framework and constructs online frequency-domain representations by incorporating historical target features, thereby improving the stability of target localization in complex scenarios.

3 Method

3.0.1 Overall Architecture.

The proposed SFMT is a spatial-frequency collaborative tracking framework that jointly exploits spatial-temporal appearance representations and frequency-domain motion cues. The detailed network architecture of SFMT is illustrated in Fig. 2. As can be observed, the design discards a purely serial stacking scheme and instead adopts a symmetric dual-branch structure: the upper branch focuses on capturing spatial structural information through spiral scanning, while the lower branch extracts motion cues in the frequency domain via the FDA module. This parallel design ensures that the model can still reliably lock onto the target by leveraging phase evolution in the frequency domain, even when spatial features become degraded.

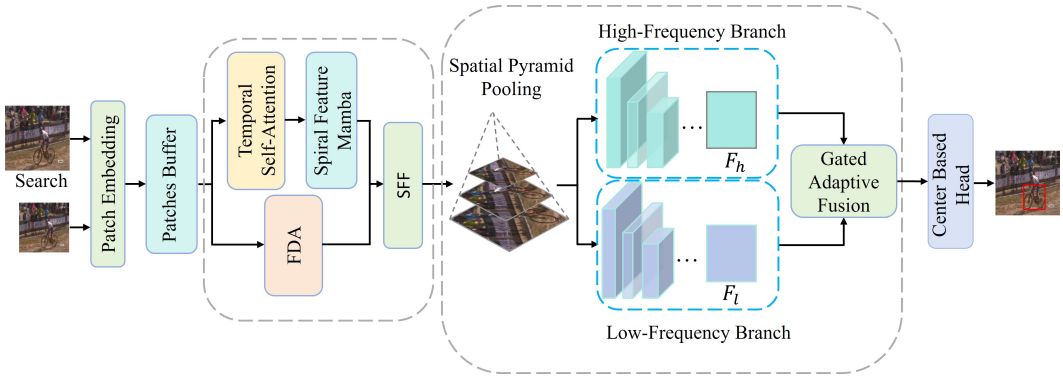

Figure 2: Overall architecture of SFMT. The framework consists of a spatial-temporal branch and a frequency-domain branch. The two branches are fused by the spatial-frequency fusion module and then fed into the pyramid multi-scale decoder for target localization.

Specifically, at the input stage, the network takes a pair of images as input: a template image and a search region image. Both images are first partitioned into non-overlapping patches of size $P \times P$, producing N_T patches for the template image and N_S patches for the search region image. The resulting search tokens are then fed into a Patches Buffer, which maintains a sequence of search tokens from the most recent N frames in a sliding-window fashion. The search features of the current frame, together with the historical information stored in the buffer, are subsequently passed into two parallel branches: the spatial-temporal branch and the frequency-domain branch.

The spatial-temporal branch first aggregates historical information through Temporal Self-Attention to produce enhanced features for the current frame. These features are then processed by a Spiral Feature Mamba module, which models the spatial structure while preserving the distributional characteristics of the target and background. In parallel, the frequency-domain branch employs the Frequency-Domain Awareness (FDA) module to extract phase-based motion features and guide amplitude filtering, yielding frequency-enhanced features. The outputs of both branches are effectively fused by the Spatial-Frequency Fusion (SFF) module to obtain a unified representation F_{fused}, which is subsequently reshaped into a 2D feature map $F_{\text{fused}} \in \mathbb{R}^{H \times W \times C}$ and fed into a pyramid multi-scale decoder. The decoder first extracts multi-scale context via spatial pyramid pooling, then passes the features through a high-frequency enhancement branch and a low-frequency enhancement branch to strengthen edge details and global structure, respectively. The outputs of these two branches are fused to produce the decoded feature F_{dec}. Finally, a center-based prediction head regresses the target bounding box from F_{dec}.

3.0.2 Spatial-Temporal Branch.

The temporal module maintains a sliding-window buffer that stores the search-region token sequences from the most recent T frames. Upon processing a new frame, the earliest entry is evicted from the buffer and the token sequence of the current frame is appended, following a first-in-first-out (FIFO) policy. This mechanism enables the model to adaptively retrieve strongly discriminative cues from historical frames and perform feature compensation for the current frame when the target undergoes transient occlusion or motion blur. Within the SFMT architecture, this design plays a complementary role: when the frequency-domain branch exhibits fluctuations due to abrupt target appearance changes, the spatiotemporal branch can flexibly invoke historical appearance representations through the self-attention mechanism, thereby providing stable feature compensation for the current frame and effectively mitigating tracking drift.

3.0.3 Frequency-Domain Branch.

The core objective of the frequency-domain branch is to overcome the bottleneck of spatial-domain motion modeling by exploiting the unique properties of the frequency domain. The Frequency-Domain Awareness (FDA) module serves as the central component for capturing dynamic cues in SFMT, and its detailed internal information flow is illustrated in Fig. 3. Conventional spatial-domain

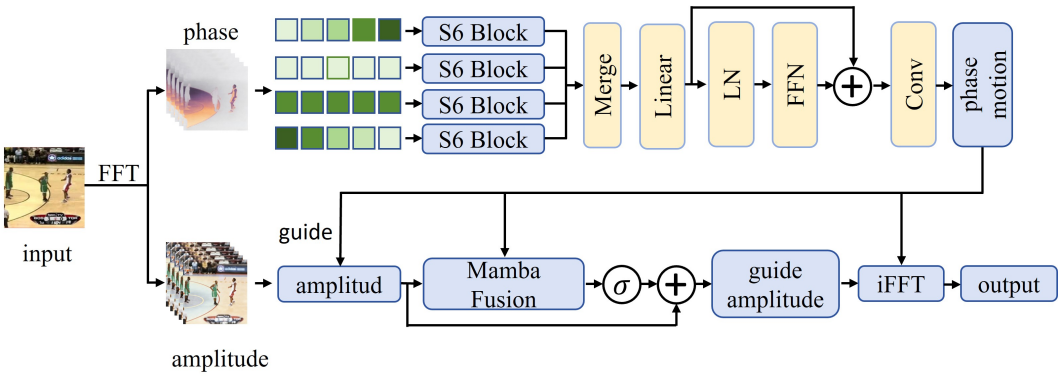

Figure 3: The architecture of the Frequency-Domain Awareness (FDA) module. The input features are transformed into amplitude and phase components through FFT. Phase variations are modeled for motion perception, while amplitude features are adaptively filtered and reconstructed through iFFT.

convolutions struggle to directly perceive the absolute displacement of the target. According to the Fourier shift theorem, a spatial translation introduces a linear phase shift in the Fourier domain. We employ Mamba to model the temporal evolution of phase spectra and encode long-range motion dynamics. Meanwhile, to prevent interference from high-frequency background noise, we design a side-path guidance mechanism that employs the extracted phase motion features to adaptively filter the amplitude spectrum, thereby achieving background denoising and target enhancement.

First, a 2D Fast Fourier Transform (FFT) is applied over the spatial dimensions H and W of the input feature F_{in} to obtain the complex spectrum $F = FFT(F_{in}) \in \mathbb{C}^{H\times W\times C}$. The amplitude spectrum $A = |F|$ and phase spectrum $P = \angle F$ are then extracted from F.

According to the Fourier shift theorem, spatial motion predominantly manifests as changes in the phase spectrum, and thus the phase spectrum is adopted as the carrier for motion perception. To exploit multi-frame motion information, the phase spectra of historical frames are stacked with that of the current frame along the temporal dimension, forming a frequency-time sequence. After flattening, this sequence is fed into a Mamba block for causal modeling, yielding phase motion features $P_m \in \mathbb{R}^{H\times W\times C}$. This design is analogous to the idea in FPDT Qi et al. (2025) of utilizing frequency-domain analysis for feature separation and interaction, but further incorporates temporal modeling.

Meanwhile, the amplitude spectrum contains rich appearance information but may also be affected by background noise. We construct a MambaFusion module, where P_m and A are concatenated along the sequence dimension and fed into a Mamba block to model their interaction and generate guidance weights:

$$A_g = \sigma\left(\text{MambaFusion}(P_m, A)\right) \odot A, \tag{1}$$

where σ denotes the Sigmoid function and $\odot$ represents element-wise multiplication. Finally, the guided amplitude spectrum A_g is combined with the original phase spectrum P and transformed back to the spatial domain via inverse FFT, yielding the reconstructed feature F_{freq}. Through temporal phase modeling and phase-guided amplitude filtering, FDA enhances motion-relevant frequency responses while suppressing background interference.

3.0.4 Spatial-Frequency Fusion Module.

After extracting the structural feature F_{spa} that encodes the spatial-temporal dimension and the motion feature F_{fre} from the frequency dimension, the key to improving discriminative power lies in achieving conflict-free alignment of cross-domain information. Because the spatial and frequency features reside in different mathematical spaces, simple additive fusion tends to provoke feature conflicts. To tackle this, we design a Spatial-Frequency Fusion module (SFF), whose detailed structure is shown in Fig. 4. Departing from conventional methods, the module adopts a dual-path strategy that concurrently performs sequence-level (global) concatenation and point-level (local) concatena-

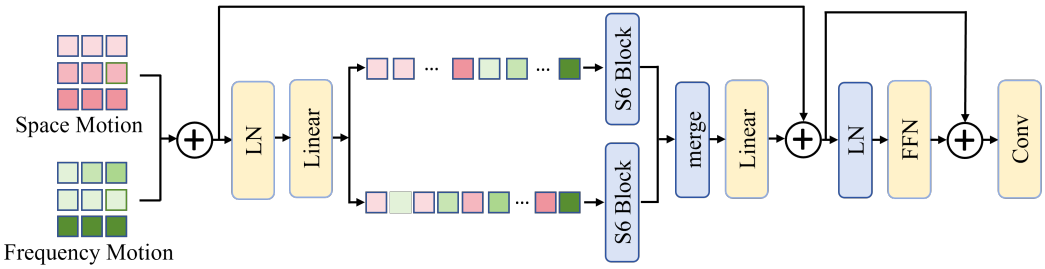

Figure 4: The architecture of the Spatial-Frequency Fusion (SFF) module. The module adopts global and local interaction paths to model cross-domain dependencies and spatial correspondence between spatial-temporal and frequency-domain features.

tion. Concretely, F_{spa} and F_{fre} are first flattened into 1D sequences $S_{spa}, S_{fre} \in \mathbb{R}^{N_x \times C}$.Two fusion paths are then constructed. Sequence-level concatenation: the two sequences are concatenated end-to-end to produce a global sequence $S_{global} \in \mathbb{R}^{2N_x \times C}$, allowing the subsequent Mamba block to model cross-domain global dependencies. This design enables the model to capture the correlations between spatial-temporal context and frequency motion cues at a global scale. Point-level concatenation: tokens at the same spatial position are concatenated to form a local sequence $S_{local} \in \mathbb{R}^{N_x \times 2C}$, which is then projected back to C dimensions through a linear layer. This allows the model to perceive dual-domain information at each location, establishing pixel-level structural correspondences and achieving fine-grained fusion. After dual-path Mamba modeling, the network produces F_{global}, which captures global cross-domain dependencies, and F_{local}, which emphasizes local point-wise correspondence. Since global dependencies and local spatial correspondence contribute differently across feature channels, we introduce learnable channel-wise re-weighting vectors to balance the two representations:

$$\begin{aligned} F_{\text{global}} &= \text{Reshape}\left(\text{Mamba}(S_{\text{global}})\right) \in \mathbb{R}^{H \times W \times C}, \\ F_{\text{local}} &= \text{Reshape}\left(\text{Mamba}(S_{\text{local}})\right) \in \mathbb{R}^{H \times W \times C}, \\ F_{\text{fused}} &= \alpha \odot F_{\text{global}} + \beta \odot F_{\text{local}}. \end{aligned} \tag{2}$$

Here, $\alpha, \beta \in \mathbb{R}^C$ are learnable channel-wise re-weighting vectors. Both vectors are initialized to 0.5 and jointly optimized with the entire network. The symbol $\odot$ denotes channel-wise element-wise weighting.

3.0.5 Pyramid Multi-Scale Decoder.

To handle target scale variations and enhance target localization, we design a pyramid multi-scale decoder. The input is the fused feature $F_{fused} \in \mathbb{R}^{H \times W \times C}$. First, spatial pyramid pooling is applied to generate feature maps with different scales. These feature maps are then upsampled and concatenated along the channel dimension, producing a multi-scale feature $S_{pyramid} \in \mathbb{R}^{H \times W \times (K \cdot C)}$. Subsequently, a 1×1 convolution is employed to compress the channel dimension back to C, obtaining $S'_{pyramid} \in \mathbb{R}^{H \times W \times C}$. This pyramid structure enables the decoder to capture contextual information from global structures to local details.

The obtained multi-scale representation is further processed by two complementary branches. The high-frequency enhancement branch employs small-kernel convolutions with residual connections to refine detailed representations, such as edges and textures. Meanwhile, the low-frequency enhancement branch utilizes average pooling to capture global structural information and improve robustness against scale variations and deformations.

Since the importance of detailed and structural information varies across different tracking scenarios, we introduce a gated adaptive fusion mechanism to dynamically adjust the contributions of high- and low-frequency information based on the input features. Specifically, the high-frequency and low-frequency features are first concatenated and fed into a gating network to generate adaptive weights:

$$G = \sigma\left(\text{Conv}\left([F_h, F_l]\right)\right), \tag{3}$$

Table 1: Performance comparison with state-of-the-art trackers on LaSOT, LaSOT$_{ext}$, TrackingNet, and GOT-10K.

Tracker	Source	LaSOT			LaSOT$_{ext}$			TrackingNet			GOT-10K		
		AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$
SFMT	Ours	**74.1**	**84.2**	81.5	52.8	63.8	60.0	85.5	89.5	84.8	77.2	87.0	73.2
TemTrack-256 Xie et al. (2025)	AAAI2025	72.0	82.1	79.1	52.4	63.3	60.2	84.3	88.8	83.5	74.9	84.8	71.7
SUTrack-B224 Chen et al. (2025)	AAAI2025	73.2	83.4	80.5	**53.1**	**64.2**	**60.5**	85.7	**90.3**	85.1	**77.9**	87.5	**78.5**
RSTrack-256 Zeng et al. (2025)	MM2025	73.4	84.1	**81.7**	52.7	63.8	60.3	85.3	89.9	84.7	76.6	86.4	76.9
DreamTrack-256 Guo et al. (2025)	CVPR2025	73.8	83.4	80.6	**53.1**	64.1	59.8	**85.8**	90.0	**85.3**	77.5	87.1	74.2
DUTrack-256 Li et al. (2025b)	CVPR2025	73.0	83.8	81.1	50.5	61.5	58.1	-	-	-	76.7	-	-
ARPTrack-256 Liang et al. (2025)	CVPR2025	72.6	81.4	78.5	52.0	62.9	58.7	85.5	91.1	87.4	81.5	**90.6**	81.7
MambaLCT Li et al. (2025a)	AAAI2025	71.8	83.0	79.4	51.6	64.0	59.0	84.3	89.2	83.9	74.8	85.4	72.1
LMTrack Xu et al. (2025)	AAAI2025	69.8	79.1	76.3	49.0	59.6	55.8	84.2	89.0	82.8	76.3	87.1	73.9
ARTrackV2 Bai et al. (2024)	CVPR2024	71.6	80.2	77.2	50.8	61.9	57.7	84.9	89.3	84.5	75.9	85.4	72.7
EVPTrack Shi et al. (2024)	AAAI2024	70.4	80.9	77.2	48.7	59.5	55.1	-	-	-	73.1	83.6	70.7
LoRAT-B Lin et al. (2024)	ECCV2024	71.7	79.5	76.6	46.4	56.6	52.3	84.2	88.7	83.5	73.5	82.2	70.9
AQATrack Xie et al. (2024)	CVPR2024	71.4	81.9	78.6	51.2	62.2	58.9	-	-	-	73.8	83.2	71.1
SeqTrack Chen et al. (2023)	CVPR2023	69.9	79.7	76.3	49.5	60.8	56.3	83.3	88.3	82.2	74.7	84.7	71.8
VideoTrack Xie et al. (2023)	CVPR2023	70.2	-	76.4	-	-	-	-	-	-	72.9	81.9	69.8
OSTrack Ye et al. (2022)	ECCV2022	69.1	78.7	75.2	47.4	57.3	53.3	83.1	87.8	82.0	71.0	80.4	68.2
GRM-B Gao et al. (2023)	CVPR2023	69.9	79.3	75.8	-	-	-	84.0	88.7	83.3	73.4	82.9	70.4
ToMP Mayer et al. (2022)	CVPR2022	67.6	78.0	72.2	-	-	-	81.2	86.2	78.6	-	-	-
CSWinTT Song et al. (2022)	CVPR2022	66.2	75.2	70.9	-	-	-	81.9	86.7	79.5	69.4	63.6	80.4
AiATrack Gao et al. (2022)	ECCV2022	69.0	79.4	73.8	-	-	-	82.7	87.8	80.4	68.6	63.2	80.0
SwinTrack-T Lin et al. (2021)	NeurIPS2022	67.2	70.8	47.6	47.6	-	53.9	81.1	-	78.4	71.3	81.9	64.5
SimTrack Chen et al. (2022a)	ECCV2022	69.3	78.5	-	-	-	-	82.3	86.5	-	68.6	-	-
TransT Chen et al. (2021)	CVPR2021	64.9	73.8	69.0	-	-	-	81.4	86.7	80.3	67.1	60.9	76.8
STARK Yan et al. (2021)	ICCV2021	66.4	76.3	71.2	-	-	-	81.3	86.1	78.1	68.0	62.3	77.7

Table 2: Comparison with state-of-the-art trackers on UAV123 and TNL2K in terms of AUC.

Dataset	STARK	ARTrack	SeqTrack	MambaLCT	MCITrack	RSTrack	DUTrack-256	SFMT
UAV123	68.2	67.7	69.2	70.1	70.5	70.5	69.3	70.4
TNL2K	-	57.5	54.9	58.5	62.9	60.5	64.9	59.8

where G denotes the weight for high-frequency features and σ represents the Sigmoid activation function. A complementary weighting strategy is then adopted to obtain the final fused feature:

$$F_{\text{out}} = G \odot F_h + (1 - G) \odot F_l. \tag{4}$$

Through this adaptive fusion strategy, the decoder can dynamically select more informative frequency representations according to the target state. Specifically, high-frequency features receive larger weights in scenarios with clear target boundaries and rich textures, while low-frequency structural information is emphasized in smooth regions, blurred areas, or occluded scenarios. Finally, the decoded feature is fed into the prediction head for target localization.

3.0.6 Prediction Head and Loss Function.

We adopt a center-based prediction head, which consists of a classification branch and a regression branch. During training, we jointly optimize the classification loss and the regression loss:

$$L = L_{cls} + \lambda_1 L_{bbox} + \lambda_2 L_{giou}, \tag{5}$$

where, L_{cls} is the focal loss Lin et al. (2017), which alleviates the imbalance between positive and negative samples. L_{bbox} is the L1 loss, supervising the bounding box regression. L_{giou} is the generalized IoU loss Rezatofighi et al. (2019), which optimizes the overlap between the predicted and ground-truth boxes. The balancing coefficients are set to $\lambda_1 = \lambda_2 = 2$. By jointly optimizing these losses, the model learns to emphasize both classification accuracy and localization precision.

4 Experiment

4.0.1 Experimental details.

To comprehensively evaluate SFMT, we conduct extensive experiments on six mainstream benchmarks: LaSOT Fan et al. (2019), LaSOT$_{ext}$ Fan et al. (2021), TrackingNet Muller et al. (2018), GOT-10K Huang et al. (2019), UAV123 Benchmark (2016), and TNL2K Wang et al. (2021). The model is trained on the training sets of COCO Lin et al. (2014), LaSOT, GOT-10K, and TrackingNet. In terms of model configuration, SFMT employs patch embedding to divide the template

Table 3: Ablation study of different components in SFMT on the LaSOT benchmark.

FDA	SFF	Multi-Scale Decoder	AUC(%)
			71.0
✓			71.8
✓	✓		73.4
✓	✓	✓	74.1

Table 4: Comparison of inference efficiency among different trackers.

Model	FLOPs (G)	Speed (FPS)	Params (M)
MambaLCT	61	58.6	72
MCITrack	19	78.5	45
SUTrack-B224	23	55	70
DreamTrack	37.6	139	87
SFMT	40.8	77.2	70.5

image (128×128) and the search region image (256×256) into patches of size 16×16, with an embedding dimension of 256. In the Frequency-Domain Awareness (FDA) module, the hidden dimension of the Mamba block is 256, and the frequency-temporal sequence length is also set to 5. The pyramid multi-scale decoder employs spatial pyramid pooling at three scales ($K = 3$) to extract local, intermediate, and global contextual information.

During training, we adopt the AdamW optimizer Loshchilov & Hutter (2017) for end-to-end optimization. The initial learning rate is set to 1×10^{-5} for the encoder and 1×10^{-4} for the decoder and other components, with a weight decay of 1×10^{-4}. The model is trained on an NVIDIA RTX 4090 GPU with a batch size of 16 for a total of 500 epochs, and the learning rate is decayed by a factor of 0.1 at the 400th epoch.

4.0.2 Comparison with State-of-the-Art Trackers.

Table 1 presents the full comparison on LaSOT, LaSOT$_{ext}$, TrackingNet, and GOT-10K. SFMT obtains 74.1% AUC on LaSOT, exceeding the listed methods, and 52.8% on LaSOT$_{ext}$. It achieves 85.5% AUC on TrackingNet and 77.2% AO on GOT-10K. Table 2 reports 70.4% and 59.8% AUC on UAV123 and TNL2K, respectively.

The four benchmarks probe different tracking conditions. LaSOT emphasizes long sequences with deformation, occlusion, and target disappearance; SFMT's 74.1% AUC improves on the 73.8% of DreamTrack-256 and the 73.4% of RSTrack-256 in Table 1. The result is consistent with the intended role of SFF in combining spatial-temporal context with frequency cues, although the aggregate benchmark score alone does not isolate the source of the gain. On the more difficult LaSOT$_{ext}$, SFMT obtains 52.8% AUC, close to the 53.1% of SUTrack-B224 and DreamTrack-256. This smaller margin suggests that extreme long-term scenarios remain challenging for the proposed design.

TrackingNet includes illumination changes, scale variation, and motion blur. SFMT reaches 85.5% AUC, comparable to the 85.7% of SUTrack-B224 and 85.8% of DreamTrack-256. GOT-10K applies a class-separation protocol that tests generalization to unseen object categories; SFMT records 77.2% AO, above the 74.8% of MambaLCT but below the 81.5% of ARPTrack-256. These results show a strong but nonuniform performance profile across benchmarks. On UAV123, SFMT obtains 70.4% AUC, and on TNL2K it obtains 59.8% AUC. The latter is lower than several recent methods in Table 2, indicating room for improvement on language-annotated sequences.

4.0.3 Ablation Studies.

Effectiveness of each component. Table 3 reports the component ablation on LaSOT. The baseline reaches 71.0% AUC; FDA improves it to 71.8%, SFF to 73.4%, and the multi-scale decoder to 74.1%. These cumulative gains support the contribution of each module.

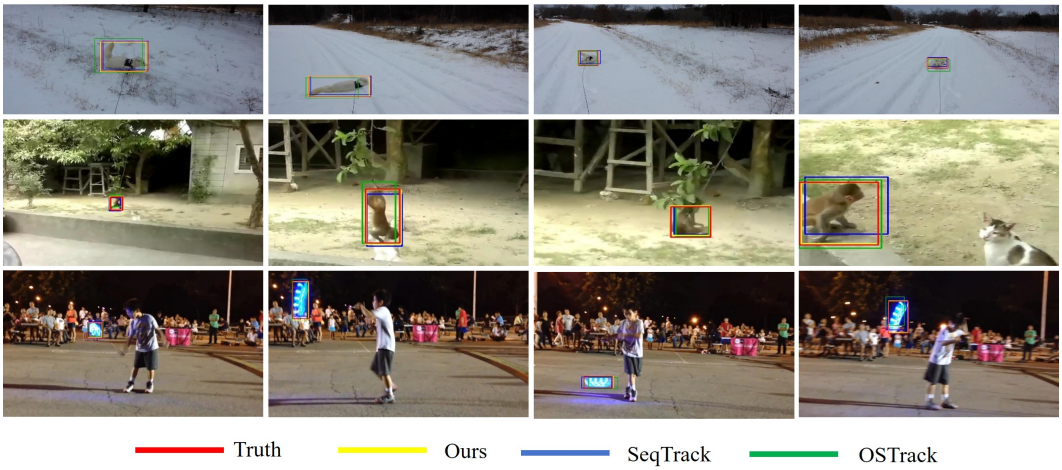

Figure 5: Qualitative comparison between SFMT and other trackers on representative tracking sequences.

FDA models the temporal evolution of phase spectra and uses the resulting motion cues to guide amplitude filtering. Its 0.8-point improvement indicates that frequency-domain information is useful even before explicit interaction with the spatial-temporal branch. Adding SFF yields a further 1.6-point increase, the largest increment in this sequence of ablations. SFF combines sequence-level interaction with point-wise fusion, so the two branches can exchange context while preserving local target detail. The final 0.7-point increase from the decoder is consistent with its use of both high-frequency edge information and low-frequency structure for localization. Because the ablation is cumulative, these numbers quantify the gains of adding modules in the reported order; they do not independently measure every possible combination. In particular, the table does not isolate SFF or the decoder when FDA is absent.

Inference Efficiency and Model Complexity. Table 4 compares model size, computation, and speed. SFMT has 70.5M parameters, requires 40.8G FLOPs, and runs at 77.2 FPS.

Compared with MambaLCT, SFMT reduces computation from 61G to 40.8G FLOPs and increases speed from 58.6 to 77.2 FPS, while using a similar number of parameters (72M versus 70.5M). The comparison also shows that SFMT is not uniformly the lightest or fastest tracker: MCITrack uses 19G FLOPs, and DreamTrack reaches 139 FPS. The efficiency claim is therefore specific to the trade-off between accuracy, computation, and throughput in the listed settings. All speed comparisons should be interpreted alongside the implementations and evaluation hardware used by the respective methods.

Qualitative results. Figure 5 compares representative tracking sequences. SFMT maintains target localization under appearance changes and background distractors in these examples. These qualitative observations complement the benchmark scores reported above.

5 Conclusion

In this paper, we propose SFMT, a spatial-frequency collaborative tracking framework that integrates spatial-temporal representations with frequency-domain motion cues. By introducing the Frequency-Domain Awareness (FDA) module, Spatial-Frequency Fusion (SFF) module, and pyramid multi-scale high-low frequency enhanced decoder, SFMT effectively exploits complementary information from spatial and frequency domains for robust target localization.Extensive experiments on six challenging benchmarks demonstrate the effectiveness of SFMT, achieving competitive performance under various challenging scenarios. Our work highlights the potential of frequency-domain motion modeling and provides a new perspective for multi-domain feature fusion in visual tracking.

AI USE STATEMENT

The authors used generative AI (GenAI) tools to assist with manuscript preparation and to correct minor grammatical errors in the manuscript. No experimental data or reported results were generated using GenAI tools. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19048–19057, 2024.

U Benchmark. A benchmark and simulator for uav tracking. In *European conference on computer vision*, volume 7, 2016.

Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *European conference on computer vision*, pp. 375–392. Springer, 2022a.

Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pp. 2722–2730, 2015.

Fei Chen, Xiaodong Wang, Yunxiang Zhao, Shaohe Lv, and Xin Niu. Visual object tracking: A survey. *Computer Vision and Image Understanding*, 222:103508, 2022b.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8126–8135, 2021.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14572–14581, 2023.

Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 2239–2247, 2025.

Runmin Cong, Mengyao Sun, Sanyi Zhang, Xiaofei Zhou, Wei Zhang, and Yao Zhao. Frequency perception network for camouflaged object detection. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 1179–1189, 2023.

Xinyi Cui, Qingshan Liu, and Dimitris Metaxas. Temporal spectral residual: fast motion saliency detection. In *Proceedings of the 17th ACM international conference on Multimedia*, pp. 617–620, 2009.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin 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*, pp. 5374–5383, 2019.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129(2):439–461, 2021.

David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International journal of computer vision*, 5(1):77–104, 1990.

Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European conference on computer vision*, pp. 146–164. Springer, 2022.

Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18686–18695, 2023.

Haifan Gong, Luoyao Kang, Yitao Wang, Yihan Wang, Xiang Wan, Xusheng Wu, and Haofeng Li. nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Mingzhe Guo, Weiping Tan, Wenyu Ran, Liping Jing, and Zhipeng Zhang. Dreamtrack: Dreaming the future for multimodal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7210, 2025.

Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3581–3589, 2025.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pp. 237–255. Springer, 2024.

Xiaohai Li, Bineng Zhong, Qihua Liang, Guorong Li, Zhiyi Mo, and Shuxiang Song. Mambalct: Boosting tracking via long-term context state space model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4986–4994, 2025a.

Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19165–19174, 2025b.

Shiyi Liang, Yifan Bai, Yihong Gong, and Xing Wei. Autoregressive sequential pretraining for visual tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7254–7264, 2025.

Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *arXiv preprint arXiv:2112.00995*, 2021.

Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets lora: Faster training, larger model, stronger performance. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8731–8740, 2022.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 300–317, 2018.

Lihua Qi, Haijun Wang, Haoyu Qu, and Zihao Su. Learning adaptive frequency-prompt denoising transformer for uav nighttime tracking. *Knowledge-Based Systems*, pp. 114341, 2025.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 783–792, 2021.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

Liangtao Shi, Bineng Zhong, Qihua Liang, Ning Li, Shengping Zhang, and Xianxian Li. Explicit visual prompts for visual object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4838–4846, 2024.

Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8791–8800, 2022.

Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pp. 343–360. Springer, 2024.

Marc Tuscher, Julian Hörz, Danny Driess, and Marc Toussaint. Deep 6-dof tracking of unknown objects for reactive grasping. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14185–14191. IEEE, 2021.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng 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*, pp. 13763–13773, 2021.

Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9697–9706, 2023.

Fei Xie, Lei Chu, Jiahao Li, Yan Lu, and Chao Ma. Videotrack: Learning to track objects via video transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22826–22835, 2023.

Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19300–19309, 2024.

Jinxia Xie, Bineng Zhong, Qihua Liang, Ning Li, Zhiyi Mo, and Shuxiang Song. Robust tracking via mamba-based context-aware token learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8727–8735, 2025.

Chenlong Xu, Bineng Zhong, Qihua Liang, Yaozong Zheng, Guorong Li, and Shuxiang Song. Less is more: Token context-aware learning for object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8824–8832, 2025.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10448–10457, 2021.

Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pp. 341–357. Springer, 2022.

Fansheng Zeng, Bineng Zhong, Haiying Xia, Yufei Tan, Xiantao Hu, Liangtao Shi, and Shuxiang Song. Explicit context reasoning with supervision for visual tracking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8067–8076, 2025.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8886–8895, 2022.

Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3529–3537, 2021.

Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4504–4513, 2022.

Jiawen Zhu, Huayi Tang, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Shihao Qiu, Shengming Li, and Huchuan Lu. Dcpt: Darkness clue-prompted tracking in nighttime uavs. *arXiv preprint arXiv:2309.10491*, 2023.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.