VCoT: Visual Chain-of-Thought for Continual Learning in Day-Night Object Tracking

Anonymous authors
Paper under double-blind review

Abstract

Stable tracking in both daytime and nighttime is essential for applying single object tracking to real-world scenarios. Traditional daytime trackers mainly rely on clear appearance features, which leads to significant performance degradation under nighttime conditions. Conversely, nighttime trackers often incorporate low-light enhancement techniques to improve robustness but struggle to maintain comparable accuracy in daytime environments. To address this challenge, we propose a novel framework, termed Visual Chain-of-Thought (VCoT), which reformulates object tracking as a structured reasoning process. VCoT follows a three-stage cognitive path of Observe–Recall-Infer–Memorize: it first observes and extracts the appearance and motion features of the current frame; then retrieves and fuses relevant historical prompts from a memory pool via an attention mechanism to enable context-aware reasoning; and finally employs gradient-based importance evaluation to update the memory by selectively retaining the most valuable knowledge. This design allows the model to integrate real-time observations with historical experiences, while achieving continual learning and effective knowledge transfer across tasks. Extensive experiments on multiple challenging benchmarks demonstrate that VCoT consistently outperforms existing methods under diverse illumination conditions. Codes will be available at `https://github.com/Gkk10/VCoT`.

1 Introduction

Object tracking plays a vital role in applications such as search-and-rescue drones Mishra et al. (2020); Martinez-Alpiste et al. (2021); Abdelnabi & Rabadi (2024), nighttime border patrol Bhanuprakash et al. (2025); Sharma et al. (2021), and urban surveillance Mohanty et al. (2025); Abba et al. (2024); Liu et al. (2021b). These tasks usually require tracking systems to maintain stable and reliable perception across two drastically different lighting environments: daytime and nighttime. For example, in earthquake rescue missions Calamoneri et al. (2022); Papyan et al. (2024), drones need to continuously search suspicious areas across day and night; in nighttime security or border patrol tasks Koslowski & Schulzke (2018); Misbah et al. (2023), the system must still accurately localize targets even under poor appearance visibility. If a system only works under a single lighting condition, its practicality in real-world scenarios will be severely limited. Therefore, developing a unified tracking mechanism with strong generalization across both daytime and nighttime scenes has become a key step toward making visual tracking truly applicable in practice.

Existing object tracking algorithms often perform well under specific lighting conditions such as daytime or nighttime. For example, ProContEXT Lan et al. (2023) achieves precise target localization in daylight scenes with sufficient illumination and clear textures by relying on appearance features. In contrast, DCPT Zhu et al. (2024a) improves tracking robustness in low-light environments by introducing the mechanism of darkness clue prompts. However, these methods are usually designed exclusively for either daytime or nighttime scenarios: daytime trackers are typically effective only under bright conditions, while nighttime trackers are tailored to low-light settings. When such single-condition methods are deployed in real-world applications that require continuous operation across day and night—such as search-and-rescue drones or surveillance systems—their performance may degrade rapidly under unseen lighting conditions. This limitation significantly restricts the reliability and practicality of these systems. These challenges highlight the necessity of design-

ing a tracking mechanism that can maintain stable performance across both daytime and nighttime environments.

From the perspective of human-like cognition, current mainstream tracking methods Zhou et al. (2020); Voigtlaender et al. (2020) face two major limitations. First, most approaches Bertinetto et al. (2016); Li et al. (2019a) focus only on single-frame information and lack the ability to model temporal continuity. As a result, they struggle to reason about motion changes through multi-step inference in the way humans do. Second, when leveraging historical information Danelljan et al. (2019); Bhat et al. (2019), these models cannot selectively retain or flexibly transfer knowledge, which makes it difficult for them to accumulate experience and adapt quickly when the environment changes. In contrast, humans, when facing uncertain situations such as occlusion, blur, or incomplete information, typically observe the motion trend of the target, recall past experiences, infer the potential position, and remember the most critical information. This process illustrates how humans integrate observations with memory, enabling them to accumulate knowledge while improving perception across different scenarios and tasks.

Motivated by the success of the Chain-of-Thought (CoT) Wei et al. (2022) mechanism in large language models for complex reasoning tasks, this paper introduces a Visual Chain-of-Thought (VCoT) framework to enable unified cognitive reasoning across both daytime and nighttime tracking scenarios. We further formulate day–night tracking as a continual learning problem, where the model must adapt between tasks under different illumination conditions while avoiding the loss of previously acquired knowledge. The overall architecture of VCoT is illustrated in Fig. 1. VCoT unfolds along a three-stage cognitive pathway of Observe–Recall-Infer–Memorize: Observe: extract appearance features from the current frame to encode the target's visual state and motion trend, generating observation prompts; Recall-Infer: use the current observation prompt as a query to retrieve and integrate relevant historical knowledge from the prompt pool, producing context-aware reasoning signals that guide Transformer sub-modules for structural modeling and decision making; Memorize: apply gradient-based importance weighting to evaluate newly generated prompts, selectively retaining the most representative knowledge to support accumulation and preservation across tasks.

The main contributions of this work are summarized as follows: **1)** VCoT framework. We are the first to introduce the concept of chain-of-thought reasoning into visual object tracking. By designing a cognitive process of Observe–Recall-Infer–Memorize, our model can maintain effective feature extraction and deliver stable tracking performance across both daytime and nighttime environments. **2)** Prompt-based continual learning. We construct a prompt pool as a memory unit and employ gradient-based importance weighting for selective updating, enabling dynamic adaptation to new tasks while effectively alleviating catastrophic forgetting. **3)** Extensive evaluation. Experiments on multiple daytime and nighttime benchmarks demonstrate that VCoT achieves superior cross-scenario generalization and overall performance compared to state-of-the-art methods.

2 Related Works

2.1 Object Tracking Across Daytime and Nighttime Scenes

Maintaining stable tracking performance under varying illumination remains a long-standing challenge in single object tracking. Most existing methods Held et al. (2016); Bertinetto et al. (2016) rely heavily on appearance-based similarity learning. These approaches typically achieve strong results in daytime scenarios with good lighting and clear textures. For instance, representative trackers such as OSTrack Ye et al. (2022a); Chen et al. (2024) leverage rich visual appearance cues to deliver accurate performance on standard well-lit benchmarks. However, in nighttime or low-light environments, image quality degrades severely, with blurred contours and missing texture information, making appearance-driven trackers struggle to maintain robustness. To address this issue, some studies Li et al. (2019b); Luo et al. (2025) introduce external prompts as complementary signals. For example, DCPT Zhu et al. (2024a) incorporates darkness-related prompts to enhance noise resistance in low-light conditions. Despite these efforts, such methods are often optimized for a single illumination domain. When deployed in real-world applications like inspection drones or surveillance systems that require continuous operation across both daytime and nighttime, their performance typically drops significantly in non-target illumination scenarios.

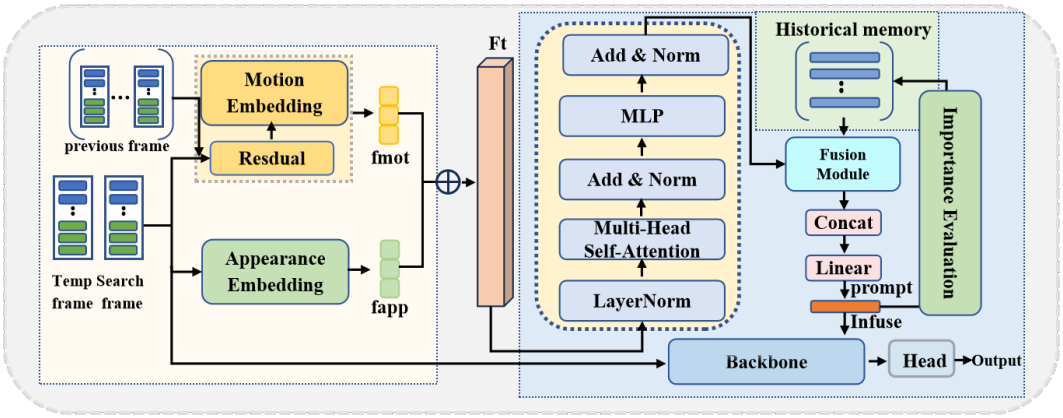


Figure 1: Schematic illustration of the overall VCoT framework. The current frame's appearance features and residual-based motion features are first extracted and fused. The fused features then interact with historical prompts in the memory via a query function, enabling the "recall-infer" process. The generated prompts are injected into the backbone, and finally, based on importance weight estimation, selectively stored in the prompt pool.

2.2 Continual Learning in Vision Tasks

The core problem that continual learning (CL) Rebuffi et al. (2017); Shin et al. (2017) aims to solve is how to enable a model to continuously accumulate and transfer knowledge while adapting to new tasks or environments, with minimal forgetting of previously learned information. In recent years, researchers have explored several strategies for mitigating forgetting in tasks such as classification and detection Li & Hoiem (2017); Shmelkov et al. (2017). These include using regularization techniques to preserve prior knowledge and employing sample replay to reduce forgetting. However, in the field of visual object tracking, related studies remain relatively limited. Existing trackers are often tailored to a single task, and their performance tends to degrade significantly when the environment changes. Introducing continual learning into tracking Liu et al. (2023b); Choi et al. (2022) not only helps alleviate performance degradation when switching between daytime and nighttime tasks but also provides new insights for addressing similar cross-domain challenges in future research.

2.3 Inspiration from Prompt Learning and Chain-of-Thought

Over the past few years, prompt learning and chain-of-thought (CoT) techniques have achieved remarkable success in natural language processing Wei et al. (2022) and multimodal tasks Liu et al. (2023a). Prompt learning Li & Liang (2021); Liu et al. (2021a) guides models to adapt to different task scenarios by embedding learnable prompt vectors into the current task. Chain-of-thought reasoning Yao et al. (2023), on the other hand, tackles complex problems by decomposing them into multiple steps, enabling models to gradually analyze and derive results in a step-by-step manner. Some prior studies Wang et al. (2022) have applied prompt mechanisms to continual learning, while others Hao et al. (2024) have explored the use of CoT for handling complex scenarios. However, most of these efforts remain limited to static image settings, lacking effective modeling of temporal dynamics and historical experience. Motivated by these insights, this work integrates prompt learning with chain-of-thought reasoning and introduces a Visual Chain-of-Thought (VCoT) framework. Furthermore, we incorporate continual learning into the design, allowing the model to adapt to nighttime tasks while still retaining knowledge and performance on daytime tasks.

3 Method

3.1 Visual Prompt Generator

This paper introduces the Visual Chain-of-Thought (VCoT) framework, which draws inspiration from human cognitive processes to achieve unified target modeling and reasoning across both

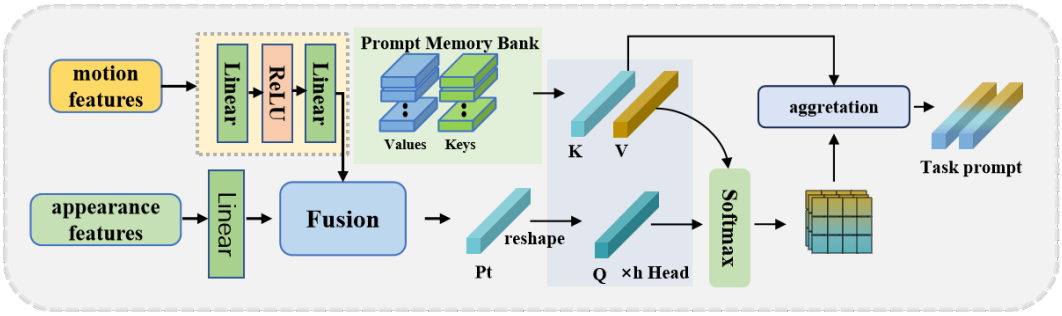


Figure 2: The Prompt Generation and Fusion Process: First, the current task prompts are extracted from the appearance and motion features. Then, they are used as queries to perform attentional fusion with the historical memory prompts (keys/values), resulting in enhanced prompt information.

daytime and nighttime environments. VCoT formulates tracking as a cognitive cycle of "Observe–Recall &Infer–Memorize". Unlike conventional trackers that rely solely on the current frame, VCoT actively leverages motion cues and historical experiences to compensate for degraded appearance information in nighttime scenarios, while its dynamic memory update mechanism supports continual learning. This enables the model to maintain stable performance across different illumination domains. the prompt generation and fusion process is illustrated in Fig. 2. Specifically, appearance features are first encoded through a linear transformation, and motion features are extracted by computing residuals between the current frame and the previous k frames. These appearance and motion representations are then concatenated along the sequence dimension and fed into a Transformer encoder, where their relationships are jointly modeled to produce fused prompt vectors. This process can be expressed as:

$$
\begin{aligned}
p_t^{\text{app}} &= W_{\text{app}} x_t, \\
p_t^{\text{mot}} &= \text{MLP}(x_t - x_{t-1}), \\
\hat{p}_t &= \text{TransEnc}([p_t^{\text{app}}, p_t^{\text{mot}}]) \in \mathbb{R}^{1\times D},
\end{aligned}
$$

where, x_t denotes the feature vector of the current frame, while x_{t-k} represents the feature vector from the previous k frames. The terms p_t^{app} and p_t^{mot} correspond to the appearance prompt and the motion prompt, respectively, and $\hat{p}_t$ denotes the fused prompt representation. D is the feature dimension. The generated prompt vectors are stored in a prompt pool, which serves as a memory buffer to maintain previously generated prompts. In this way, when new prompts are created, the model can refer to past information and integrate prior experience into the current task as guidance. This process can be formulated as follows:

$$
\begin{aligned}
\text{M}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O, \\
\text{head}_i &= \text{softmax}\left(\frac{Q W_i^Q (K W_i^K)^T}{\sqrt{d}}\right) V W_i^V,
\end{aligned}
$$

where Q is$\hat{p}_t$, K and V is the historical memory, and W_i^Q, W_i^K, W_i^V are learnable projection matrices. This process enables the model to aggregate contextual information across time and generate updated prompts. The linear layer projects them into a fixed-length prompt sequence $P_t \in \mathbb{R}^{L\times D}$. The prompt is then expanded along the batch dimension and concatenated to the front of the main input token sequence:

$$
X_t' = [P_t^{(D)}; X_t].
$$

3.2 Prompt Memory and Importance Evaluation

Traditional object tracking algorithms typically rely on modeling the current target, while overlooking the role of historical experience in assisting the ongoing tracking task. Prior studies Cai et al. (2024) have shown that accumulating historical knowledge can not only improve recognition performance but also help mitigate catastrophic forgetting. This aligns with the goal of day–night tracking,

where the objective is to fully leverage complementary information from both daytime and nighttime while avoiding the loss of nighttime knowledge during training. To this end, we maintain a prompt memory module that stores previously generated prompts. When a new prompt is generated, the system determines whether it should be added to memory. Since storage capacity is limited, it is impossible to retain all prompts, making it essential to establish a mechanism for selecting the most valuable ones. We propose a gradient-guided prompt scoring mechanism to evaluate the contribution of each prompt to the reduction of the loss function during model updates. Unlike traditional similarity-based metrics, our method dynamically prioritizes memory retention based on the verified utility of prompts, thereby improving both knowledge consolidation and evolution. The core idea is intuitive: if a prompt has a stronger influence on the loss reduction during training, it is more valuable to retain. Let $P_t \in \mathbb{R}^{L \times D}$ denote the prompt sequence at time step t. During backpropagation, we compute the gradient $\nabla_{P_t}\mathcal{L}$ with respect to each token in the sequence. The sensitivity score is obtained by averaging the ℓ_2-norm of the gradients across all tokens:

$$g_t = \frac{1}{L}\sum_{l=1}^{L} \|\nabla_{p_{t,l}}\mathcal{L}\|_2,$$

which reflects the overall sensitivity of the prompt to parameter updates in a single optimization step. To suppress noise fluctuations and ensure stable scoring, we apply a sliding window of size K to smooth the sensitivity values. The final importance score is then defined as the average over the most recent K steps:

$$s_t = \frac{1}{K}\sum_{k=1}^{K} g_t^{(k)}.$$

After computing s_t, each prompt is assigned an importance score and stored in the prompt memory. To prevent unbounded growth, the memory retains only the top-M prompts with the highest importance scores at any given time.

3.3 Continual Learning Modeling

For a tracker to adapt robustly across day and night conditions, it must maintain stable perception under varying illumination. Addressing the plasticity–stability dilemma in continual learning is therefore critical. If the model relies solely on the features of the current task without retaining past knowledge, new training will inevitably overwrite previous representations, leading to severe forgetting. To alleviate this, we introduce a memory mechanism that supports selective storage and updating of prompts across tasks. Prompts serve both as contextual cues for the current task and as transferable knowledge units that accumulate experience over time. This allows the model to continually leverage prior knowledge while adapting to new environments, thus achieving cross-task consistency and stability. Specifically, when encountering new tasks, the model evaluates the importance of newly generated prompts and updates the memory pool accordingly. The retained prompts can then be recalled and fused with current observations, enabling knowledge transfer between day and night domains. Formally, let P_t denote the prompt at time t and $s(P_t)$ its gradient-based importance score. The memory update rule is:

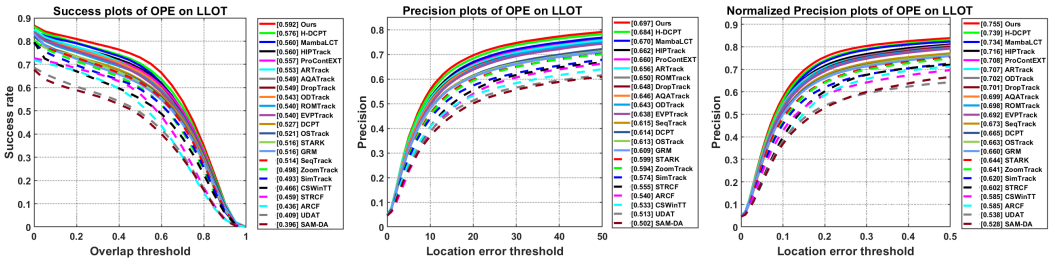
$$M_{t+1} = \text{Top}_M\big(M_t \cup \{(P_t, s(P_t))\}\big),$$

where Top_M selects the top-M prompts with the highest importance scores.

At inference, the current observation o_t (encoded from external inputs) is used as the query, while prompts in the memory serve as keys and values. A multi-head attention module retrieves and integrates the most relevant historical prompts with the current observation:

$$r_t = \text{MHA}(Q = o_t,\ K = P_{\text{hist}},\ V = P_{\text{hist}}),$$

where P_{hist} denotes the set of prompts stored in memory. In this design, gradient sensitivity governs memory writing and updating, while attention drives memory retrieval and fusion. This complementary mechanism allows the tracker to accumulate knowledge incrementally while maintaining stability, ensuring robust adaptation across day and night tasks.



(a) Success (b) Precision (c) NormPrcesion

Figure 3: Overall performance of VCoT and other SOTA trackers (DropTrack Durve et al. (2022), ROMTrack Cai et al. (2023), SeqTrack Chen et al. (2023), OSTrack Ye et al. (2022a), GRM Gao et al. (2023), STARK Yan et al. (2021), ZoomTrack Kou et al. (2023), SimTrack Chen et al. (2022), STRCF Li et al. (2018), CSWinTT Song et al. (2022), ARCF Huang et al. (2019b), UDAT Ye et al. (2022c)) on LLOT. View enlarged image for clarity.

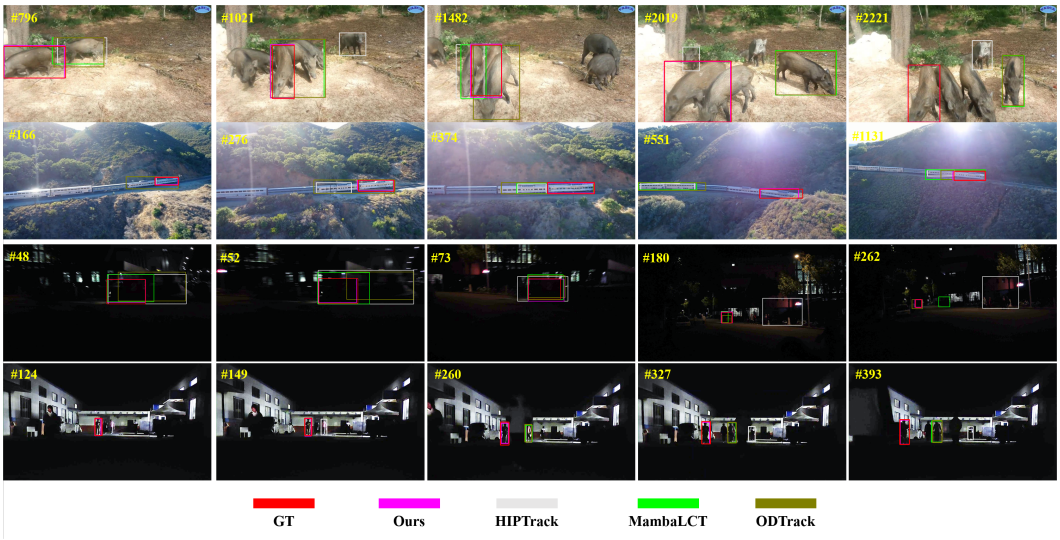


Figure 4: Visual comparison between our tracker and three SOTA methods on LaSOT and LLOT benchmarks.

4 EXPERIMENTS

We treat daytime and nighttime tracking as two sequential tasks in a continual learning setting. For training, we construct task-specific datasets under two illumination conditions. The daytime task integrates LaSOT Fan et al. (2019), GOT-10k Huang et al. (2019a), COCO Lin et al. (2014), and TrackingNet Muller et al. (2018) to cover diverse daytime scenarios, while the nighttime task leverages BDD100K-Night Yu et al. (2020) and SHIFT-Night Sun et al. (2022), which contain low-light conditions. For evaluation, DTB70 Li & Yeung (2017), VisDrone2018 Zhu et al. (2018), UAVDT Du et al. (2018), and OTB100 Wu et al. (2013) are used as benchmarks for daytime tasks, whereas UAVDark135 Li et al. (2022), NAT2024-1 Fu et al. (2024a), DarkTrack2021 Ye et al. (2022b), and LLOT Zhong et al. (2024) are employed for nighttime tasks. To further validate the generalization ability of our method, we additionally conduct large-scale experiments on GOT-10k Huang et al. (2019a) and TrackingNet Muller et al. (2018). Performance is evaluated following standard single-object tracking protocols, using Area Under the Curve (AUC), Precision (P), and Normalized Precision (P_{norm}) as the main metrics. To ensure fair comparisons, strong baselines such as ODTrack, HIPTrack, and AQATrack are retrained on the same datasets using their official default settings before testing on nighttime benchmarks.

4.1 Implementation Details

We adopt HiViT-Base as the backbone network, with input sizes of 224×224 for the search region and 112×112 for the template region. The model is trained using the AdamW optimizer Loshchilov & Hutter (2017) for 300 epochs with a batch size of 16. The initial learning rate is set to 1×10^{-4} and the weight decay to 1×10^{-4}. Each epoch contains approximately 60,000 image pairs. To stabilize training, the learning rate is reduced to 1×10^{-5} after 250 epochs. All experiments are conducted on a workstation equipped with an Intel i9-10850K CPU, 16 GB of memory, and an NVIDIA Titan X GPU.

Table 1: Comparison of tracking performance on nighttime datasets. The top three results are highlighted in red, blue, and green.

		UAVDark135			NAT2024-1			DarkTrack2021		
Method	Source	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}
UniTrack (Ours)	-	68.6	82.6	83.9	73.3	94.5	90.7	62.6	74.2	75.0
MambaLCT Li et al. (2025)	AAAI2025	64.4	77.9	80.6	60.8	88.9	85.8	61.5	74.4	75.1
ODTrack Zheng et al. (2024)	AAAI2024	63.2	77.8	78.1	69.1	89.6	85.5	60.5	72.2	72.3
HIPTrack Cai et al. (2024)	CVPR2024	59.7	72.0	70.0	69.4	88.4	84.1	57.1	68.5	68.7
EVPTtrack Shi et al. (2024)	AAAI2024	58.1	69.2	70.5	65.0	83.7	78.1	53.7	64.8	64.7
AQATrack Xie et al. (2024)	CVPR2024	58.2	69.2	70.7	64.2	82.1	77.1	55.0	66.1	66.8
SAM-DA Fu et al. (2024b)	ICARM24	47.6	60.4	59.4	53.4	75.3	64.9	44.7	55.5	54.6
DCPT Zhu et al. (2024b)	ICRA2024	56.7	69.2	69.8	62.1	80.9	75.4	54.0	66.7	64.6
AVTrack Li et al. (2024)	ICML2024	47.6	58.6	59.2	56.7	68.2	75.3	46.1	55.1	54.9
LiteTrack Wei et al. (2024)	CVPR2024	53.9	63.6	65.9	61.8	79.7	74.1	52.8	63.5	62.8

Table 2: Comparison of tracking performance on daytime datasets. The top three results are highlighted in red, blue, and green.

		DTB70			VisDrone2018			UAVDT		
Method	Source	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}
UniTrack (Ours)	-	70.1	90.7	85.3	70.6	90.0	86.8	67.2	88.8	77.6
MambaLCT Li et al. (2025)	AAAI2025	68.7	88.3	84.0	65.4	88.1	84.3	63.6	84.4	75.8
ODTrack Zheng et al. (2024)	AAAI2024	70.0	90.0	86.1	64.7	85.6	83.1	63.8	85.8	74.7
HIPTrack Cai et al. (2024)	CVPR2024	68.6	81.2	76.2	67.1	86.7	83.9	60.9	81.2	76.2
EVPTtrack Shi et al. (2024)	AAAI2024	66.6	86.7	81.7	66.6	87.0	82.6	60.2	80.0	71.3
AQATrack Xie et al. (2024)	CVPR2024	66.1	86.3	80.7	66.9	87.2	89.8	63.7	84.7	75.9
SAM-DA Fu et al. (2024b)	ICARM24	63.0	82.2	76.3	53.1	71.4	67.0	61.3	82.6	73.3
DCPT Zhu et al. (2024b)	ICRA2024	64.6	83.7	77.6	64.2	83.1	79.7	56.9	76.8	66.0
AVTrack Li et al. (2024)	ICML2024	65.0	84.3	80.0	64.2	84.8	80.3	58.7	82.1	68.6
LiteTrack Wei et al. (2024)	CVPR2024	64.7	83.5	77.6	61.8	79.8	75.7	62.1	84.3	71.6

4.2 Comparison with State-of-the-art Methods

Quantitative Comparison On nighttime benchmarks, our method demonstrates clear performance advantages. As shown in Table 1 and Fig. 3, VCoT achieves 68.61% AUC and 82.69% P on UAVDark135, outperforming all competing state-of-the-art methods. On Nat2024-1, VCoT ranks first with 73.33% AUC, and 94.57% P, highlighting its robustness to appearance degradation under low-light conditions. Furthermore, on two particularly challenging datasets, DarkTrack2021 and LLOT, our method consistently maintains leading results across all three metrics, validating the stability and generalization ability of VCoT under diverse illumination scenarios.

On daytime benchmarks, VCoT likewise surpasses existing SOTA trackers. As presented in Table 2, on DTB70, our method achieves 70.13% AUC and 90.75% P. On OTB100, it further obtains 71.76% AUC, achieving the best overall tracking performance. On UAVDT, VCoT delivers significant improvements with 67.25% AUC and 88.81% P, reflecting strong adaptability to complex real-world conditions. On VisDrone2018, it reaches 70.64% AUC, 90.02% P, and 86.84% P_{Norm}, outperforming all competitors and confirming its effectiveness under scale variation and motion blur. VCoT

Table 3: Performance comparison on TrackingNet dataset.

Method	AUC	P	P_{Norm}
UniTrack (Ours)	85.02	84.60	89.52
ManBaLCT	84.30	83.90	89.20
HIPTrack	84.50	83.80	89.10
LoRAT	83.50	82.10	87.90
ARTrackV2	84.90	84.50	89.30
EVPTtrack	83.50	-	88.30
AQATrack	83.80	83.10	88.60
LiteTrack	80.80	78.20	85.70

Table 4: Performance comparison on GOT-10K dataset.

Method	AO	$\mathbf{SR}_{0.5}$	$\mathbf{SR}_{0.75}$
UniTrack(Ours)	77.10	87.01	76.70
MambaLCT	74.80	85.40	72.10
HIPTrack	77.40	88.70	74.50
LoRAT	72.10	81.80	70.70
ARTrackV2	75.90	85.40	72.70
EVPTrack	73.30	83.60	70.70
AQATrack	73.80	83.20	73.10
LiteTrack	68.70	78.20	64.20

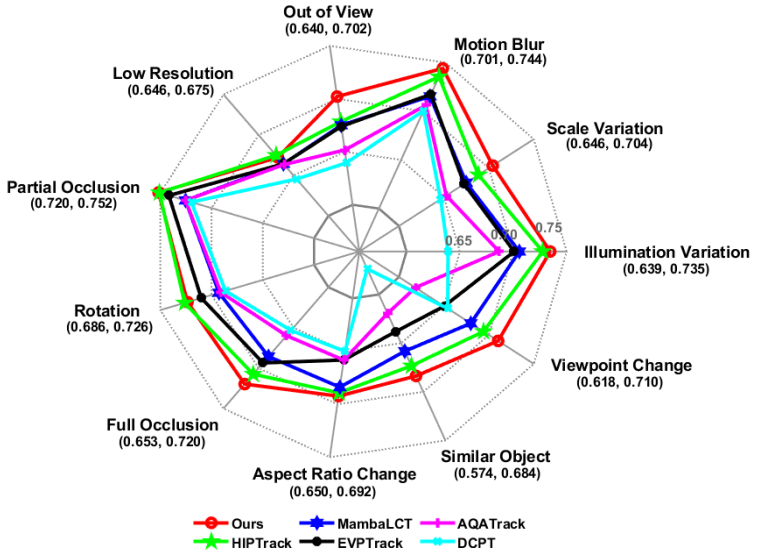


Figure 5: Performance Comparison Across OTB100 Challenge Attributes.

demonstrates remarkable performance in the OTB100 radar chart (Fig. 5), exhibiting outstanding advantages in handling core challenges such as partial occlusion and motion blur.

To further verify generalization and scalability, we conduct experiments on two large-scale datasets, GOT-10k and TrackingNet. As reported in Table 3 and 4, VCoT consistently achieves strong results. On GOT-10k, it achieves higher AO and SR compared to MambaLCT and LoRAT Lin et al. (2024). On TrackingNet, it surpasses HIPTrack and ARTrackV2 Bai et al. (2024) across AUC, P, and P_{Norm}, demonstrating robust generalization in large-scale real-world scenarios.

Qualitative Comparison To further validate the effectiveness of the proposed method under varying illumination conditions, we conduct qualitative comparisons with representative state-of-the-art trackers across diverse day and night scenarios, as illustrated in Fig. 4. In daytime scenes (first and second rows), mainstream methods such as HIPTrack, MambaLCT, and ODTrack perform reasonably well when the target appearance is clear. However, they often suffer from boundary shifts or target loss under strong illumination or background distractions. In contrast, our method consistently maintains accurate boundary alignment, leading to more stable and precise tracking. In nighttime scenarios (third and fourth rows), low illumination and noise pose significant challenges, where existing methods commonly exhibit bounding box drift or incorrect matches. For example, HIPTrack tends to drift under weak illumination, MambaLCT struggles when the contrast between target and background is low, and ODTrack fails to capture the target under complex lighting conditions. By leveraging the proposed Visual Chain-of-Thought reasoning mechanism to effectively integrate historical memory, our method is able to robustly localize the target even in extremely poor illumination and heavy occlusion cases.

Table 5: Stepwise Ablation Results of VCoT Components on LLOT and DTB70. O, R, and M correspond to the Observe, Recall-Infer, and Memorize stages in our VCoT framework.

Method	LLOT		DTB70	
	Succ.	Prec.	Succ.	Prec.
Baseline	55.88	62.90	67.85	87.39
Baseline+O	57.52	63.85	68.72	89.02
Baseline+O+R	58.47	64.91	69.20	88.90
Baselinee+O+R+M	**59.38**	**65.42**	**70.13**	**90.75**

Table 6: Ablation study on continual learning ability in LaSOT benchmark.

Train Data	Method	**Succ.**	**Prec.**
Daytime only	baseline	70.99	77.04
Daytime+Nighttime	baseline	67.49	72.71
Daytime +Nighttime	VCoT	**72.74**	**78.87**

Table 7: Ablation study on the role of the historical prompt pool on DTB70 dataset.

Method	**Succ.**	**Prec.**
Baseline	67.85	87.39
Baseline+Prompt	68.15	87.24
VCoT	**70.13**	**90.75**

4.3 Ablation Study

Stepwise Ablation of VCoT Components. As shown in Table 5. We first evaluate the independent contributions of the three stages: Observe (O), Recall-Infer (R), and Memorize (M) on the LLOT and DTB70 datasets. The baseline model (B) achieves only 55.88%/62.9% success and precision on LLOT. Adding the Observe module (B+O), then the Recall-Infer stage (B+O+R), and finally the full three-stage pipeline (B+O+R+M), the performance improves step by step, reaching 59.38%/65.42% on LLOT. On DTB70, the complete model achieves 70.13% success and 90.75% precision, surpassing the baseline by 2.28% and 3.36%, respectively. Each stage proves useful, and Memorize especially strengthens stability across day and night.

Continual Learning Capability. As presented in Table 6. To evaluate the model's continual learning ability under day-to-night task switching, we conduct staged training experiments on LaSOT. When trained only on daytime data, the model achieves 70.99%/77.04% in success and precision. However, when subsequently trained on nighttime data without an effective cross-task memory mechanism, the performance drops to 67.49%/72.71%, indicating clear forgetting. In contrast, with our proposed continual learning scheme, the model is able to retain both daytime and nighttime knowledge, achieving 72.74%/78.87%.

Effect of the Historical Prompt Pool. As illustrated in Table 7. We further evaluate the impact of the historical prompt pool on model performance. As shown in the experiments, when relying only on the current frame's appearance and motion prompts (B), the model achieves 67.85%/87.39% in success and precision on DTB70. Simply introducing prompts without retaining historical memory (B+Prompt) yields little improvement. In contrast, the full method, which leverages the prompt pool to accumulate and selectively retain historical information, improves performance to 70.13%/90.75%. This result underscores the prompt pool's role in knowledge selection and transfer.

5 Conclusion

This study is the first to introduce continual learning and the Chain-of-Thought (CoT) reasoning mechanism into visual object tracking. We propose a novel framework, termed Visual Chain-of-Thought (VCoT), which models the human cognitive process of "observe–recall–infer–memorize." By effectively integrating real-time observations with historical experience, VCoT maintains stable and accurate tracking performance across both daytime and nighttime scenarios. Furthermore, we design a continual learning mechanism that employs gradient-guided importance evaluation to update and retain critical historical information, enabling the model to adapt to new tasks while alleviating the common problem of catastrophic forgetting. Extensive experiments demonstrate that VCoT consistently outperforms current state-of-the-art methods in overall performance.

REFERENCES

Sani Abba, Ali Mohammed Bizi, Jeong-A Lee, Souley Bakouri, and Maria Liz Crespo. Real-time object detection, tracking, and monitoring framework for security surveillance systems. *Heliyon*, 10(15), 2024.

Ahmad A Bany Abdelnabi and Ghaith Rabadi. Human detection from unmanned aerial vehicles' images for search and rescue missions: a state-of-the-art review. *IEEE Access*, 2024.

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.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pp. 850–865. Springer, 2016.

M Bhanuprakash, Vaibhav Buyya, B Ravinaik, et al. Intelligent surveillance and night patrolling drone. *Indo-American Journal of Mechanical Engineering*, 14(2):61–66, 2025.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6182–6191, 2019.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19258–19267, 2024.

Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9589–9600, 2023.

Tiziana Calamoneri, Federico Corò, and Simona Mancini. A realistic model to support rescue operations after an earthquake via uavs. *IEEE access*, 10:6109–6125, 2022.

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, 2022.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5343–5353, 2024.

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.

Janghoon Choi, Sungyong Baik, Myungsub Choi, Junseok Kwon, and Kyoung Mu Lee. Visual tracking by adaptive continual meta-learning. *IEEE Access*, 10:9022–9035, 2022.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4660–4669, 2019.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 370–386, 2018.

Mihir Durve, Adriano Tiribocchi, Fabio Bonaccorso, Andrea Montessori, Marco Lauricella, Michał Bogdan, Jan Guzowski, and Sauro Succi. Droptrack—automatic droplet tracking with yolov5 and deepsort for microfluidic applications. *Physics of Fluids*, 34(8), 2022.

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.

Changhong Fu, Yiheng Wang, Liangliang Yao, Guangze Zheng, Haobo Zuo, and Jia Pan. Prompt-driven temporal domain adaptation for nighttime uav tracking. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9706–9713. IEEE, 2024a.

Changhong Fu, Liangliang Yao, Haobo Zuo, Guangze Zheng, and Jia Pan. Sam-da: Uav tracks anything at night with sam-powered domain adaptation. In *2024 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 31–38. IEEE, 2024b.

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.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European conference on computer vision*, pp. 749–765. Springer, 2016.

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, 2019a.

Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2891–2900, 2019b.

Rey Koslowski and Marcus Schulzke. Drones along borders: Border security uavs in the united states and the european union. *International Studies Perspectives*, 19(4):305–324, 2018.

Yutong Kou, Jin Gao, Bing Li, Gang Wang, Weiming Hu, Yizheng Wang, and Liang Li. Zoomtrack: Target-aware non-uniform resizing for efficient visual tracking. *Advances in Neural Information Processing Systems*, 36:50959–50977, 2023.

Jin-Peng Lan, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Xu Bao, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Procontext: Exploring progressive context transformer for tracking. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4282–4291, 2019a.

Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. All-day object tracking for unmanned aerial vehicle. *IEEE Transactions on Mobile Computing*, 22(8):4515–4529, 2022.

Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019b.

Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4904–4913, 2018.

Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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, 2025.

Yongxin Li, Mengyuan Liu, You Wu, Xucheng Wang, Xiangyang Yang, and Shuiwang Li. Learning adaptive and view-invariant vision transformer for real-time uav tracking. In *Forty-first International Conference on Machine Learning*, 2024.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023a.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021a.

Ze Liu, Yingfeng Cai, Hai Wang, Long Chen, Hongbo Gao, Yunyi Jia, and Yicheng Li. Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions. *IEEE Transactions on Intelligent Transportation Systems*, 23 (7):6640–6653, 2021b.

Zhizheng Liu, Mattia Segu, and Fisher Yu. Cooler: class-incremental learning for appearance-based multiple object tracking. In *DAGM German Conference on Pattern Recognition*, pp. 443–458. Springer, 2023b.

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

Binxin Luo, Dongxu Liu, Xianrong Peng, Haorui Zuo, Jianlin Zhang, Meihui Li, and Yuxing Wei. Progressive transformer with multi-modality adaptation for rgb-t tracking. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Ignacio Martinez-Alpiste, Gelayol Golcarenarenji, Qi Wang, and Jose Maria Alcaraz-Calero. Search and rescue operation using uavs: A case study. *Expert Systems with Applications*, 178:114937, 2021.

Maham Misbah, Misha Urooj Khan, Zhaohui Yang, and Zeeshan Kaleem. Tf-net: Deep learning empowered tiny feature network for night-time uav detection. In *International Conference on Wireless and Satellite Systems*, pp. 3–18. Springer, 2023.

Balmukund Mishra, Deepak Garg, Pratik Narang, and Vipul Mishra. Drone-surveillance for search and rescue in natural disaster. *Computer Communications*, 156:1–10, 2020.

Anita Mohanty, Ambarish G Mohapatra, and Subrat Kumar Mohanty. Real-time traffic monitoring with ai in smart cities. In *Internet of Vehicles and Computer Vision Solutions for Smart City Transformations*, pp. 135–165. Springer, 2025.

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.

Narek Papyan, Michel Kulhandjian, Hovannes Kulhandjian, and Levon Aslanyan. Ai-based drone assisted human rescue in disaster environments: Challenges and opportunities. *Pattern Recognition and Image Analysis*, 34(1):169–186, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Manish K Sharma, Gaurav Singal, Suneet K Gupta, Basa Chandraneil, Saksham Agarwal, Deepak Garg, and Debajyoti Mukhopadhyay. Intervenor: Intelligent border surveillance using sensors and drones. In *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–7. IEEE, 2021.

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.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pp. 3400–3409, 2017.

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.

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21371–21382, 2022.

Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6578–6588, 2020.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qingmao Wei, Bi Zeng, Jianqi Liu, Li He, and Guotian Zeng. Litetrack: Layer pruning with asynchronous feature extraction for lightweight and efficient visual tracking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4968–4975. IEEE, 2024.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2411–2418, 2013.

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.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

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, 2022a.

Junjie Ye, Changhong Fu, Ziang Cao, Shan An, Guangze Zheng, and Bowen Li. Tracker meets night: A transformer enhancer for uav tracking. *IEEE Robotics and Automation Letters*, 7(2): 3866–3873, 2022b.

Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8896–8905, 2022c.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7588–7596, 2024.

Pengzhi Zhong, Xiaoyu Guo, Defeng Huang, Xiaojun Peng, Yian Li, Qijun Zhao, and Shuiwang Li. Low-light object tracking: A benchmark. *arXiv preprint arXiv:2408.11463*, 2024.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pp. 474–490. Springer, 2020.

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. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7381–7388. IEEE, 2024a.

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. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7381–7388. IEEE, 2024b.

Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.