2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MATrack: Efficient Multiscale Adaptive Tracker for Real-Time Nighttime UAV Operations

Xuzhao Li[1*], Xuchen Li[1*‡] and Shiyu Hu[1†]

***Abstract*— Nighttime UAV tracking faces significant challenges in real-world robotics operations. Low-light conditions not only limit visual perception capabilities, but cluttered backgrounds and frequent viewpoint changes also cause existing trackers to drift or fail during deployment. To address these difficulties, researchers have proposed solutions based on low-light enhancement and domain adaptation. However, these methods still have notable shortcomings in actual UAV systems: low-light enhancement often introduces visual artifacts, domain adaptation methods are computationally expensive and existing lightweight designs struggle to fully leverage dynamic object information. Based on an in-depth analysis of these key issues, we propose MATrack—a multiscale adaptive system designed specifically for nighttime UAV tracking. MATrack tackles the main technical challenges of nighttime tracking through the collaborative work of three core modules: Multiscale Hierarchy Blende (MHB) enhances feature consistency between static and dynamic templates. Adaptive Key Token Gate accurately identifies object information within complex backgrounds. Nighttime Template Calibrator (NTC) ensures stable tracking performance over long sequences. Extensive experiments show that MATrack achieves a significant performance improvement. On the UAVDark135 benchmark, its precision, normalized precision and AUC surpass state-of-the-art (SOTA) methods by 5.9%, 5.4% and 4.2% respectively, while maintaining a real-time processing speed of 81 FPS. Further tests on a real-world UAV platform validate the system's reliability, demonstrating that MATrack can provide stable and effective nighttime UAV tracking support for critical robotics applications such as nighttime search and rescue and border patrol.**

I. Introduction

As a core task of modern robotic vision systems, unmanned aerial vehicle (UAV) object tracking plays an irreplaceable role in critical applications such as border patrol [1], nighttime search and rescue [2] and aerial reconnaissance [3]. This ability to automatically follow moving objects from an aerial platform provides vital technical support for real-world applications. However, as UAV operations expand into nighttime environments, traditional tracking technologies are facing unprecedented challenges. This operational shift is crucial for missions that require continuous surveillance capabilities. In recent years, single object tracking technology [4] has made significant progress, driven by advancements in deep learning [5] and the Transformer architecture [6], [7]. Algorithms like MixFormer [8], ODTrack [9] and VideoTrack [10] have demonstrated excellent performance in daytime conditions by modeling global context and learning sequence-level representations. However, when these advanced techniques are applied to nighttime UAV tracking, their limitations are quickly exposed. The low-light conditions of nighttime environments cause a sharp drop in image signal-to-noise ratio, while frequent viewpoint changes and motion blur further degrade feature quality. Meanwhile, complex background clutter increases the risk of mistracking. The UAV's inherent motion, combined with these visual challenges, real-time requirements and resource constraints, creates a complex technical problem for stable nighttime object tracking.

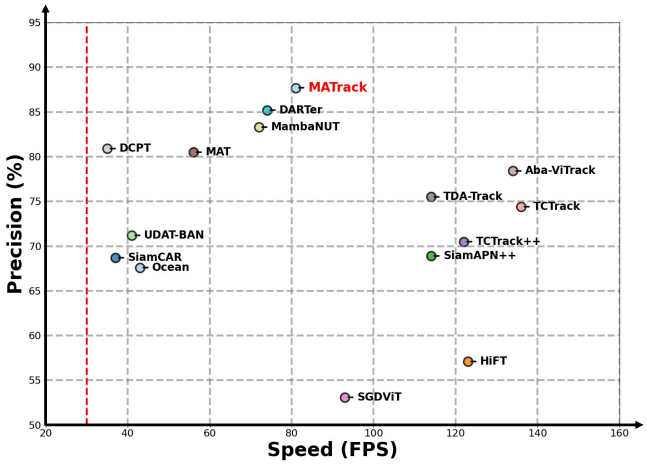

Fig. 1. This scatter plot illustrates the balance between tracking speed (FPS) and precision (%) for various methods on NAT2024-1 [11] benchmark. The red dashed line marks the 30 FPS threshold for real-time performance. Our proposed MATrack (highlighted in red) achieves the highest precision (87.7%) while maintaining a speed (81 FPS) well above the real-time requirement, demonstrating a superior efficiency-performance trade-off compared to other state-of-the-art (SOTA) methods.

To tackle this challenge, researchers explore three main directions. The first category of methods uses low-light enhancement techniques (e.g., HighlightNet [12], Darklighter [13]) to improve image quality. However, this pre-processing often introduces visual artifacts that, in the highly dynamic flight environment of an UAV, can interfere with tracking accuracy. For instance, boosting low-light signals can inadvertently amplify sensor noise or generate false edges, which confuse the tracker. The second category employs domain adaptation techniques (e.g., UDAT [14], SAM-DA [15]) to narrow the distribution gap between daytime and nighttime features. While these methods can improve tracking performance to some extent, their high training costs and computational demands are difficult to meet on resource-constrained UAV platforms. The offline nature of these

[1]Nanyang Technological University (NTU), Singapore.
*Equal contribution. ‡Project lead.
†Corresponding author's e-mail address: shiyu.hu@ntu.edu.sg.

methods also limits their adaptability to unforeseen real-world scenarios. The third category focuses on lightweight and efficient network designs (e.g., DCPT [16], MambaNUT [17], DARTer [18]) to improve computational efficiency, but they still fall short in adapting to dynamic environments, resisting noise and maintaining long-term stability. A common limitation of all these methods is that they primarily seek algorithmic solutions while overlooking the fundamental nature of nighttime UAV tracking as a system-level problem. The core issue is not just about a single-component solution but about creating a robust, end-to-end framework that addresses the holistic set of challenges.

Recognizing this, we propose MATrack—a multiscale adaptive tracking framework designed from a system perspective. MATrack integrates three synergistic core modules: the Multiscale Hierarchical Blende (MHB) enhances feature consistency and robustness by unifying static and dynamic template information; the Adaptive Key Token Gate (AKTG) dynamically identifies and strengthens object-related visual cues in complex nighttime environments; and the Nighttime Template Calibrator (NTC) ensures the stability of the tracking system over long sequences through an intelligent update mechanism. This collaborative design allows MATrack to generate highly discriminative object representations even under severe light degradation, while meeting the real-time and resource constraints of UAV platforms.

Extensive experiments fully validate MATrack's superiority. In a comprehensive evaluation across five nighttime tracking benchmarks, MATrack not only achieves state-of-the-art performance on all metrics—surpassing the best existing methods by 6.0% in precision, 5.4% in normalized precision and 4.2% in AUC on UAVDark135 [19] benchmark—but also strikes an ideal balance between efficiency and performance with a speed of 81 FPS. More importantly, deployment tests on a real-world UAV platform have confirmed MATrack's practical utility, demonstrating that it is not merely an algorithmic advancement but a reliable engineering solution capable of operating effectively in real systems. This real-world validation confirms its robustness beyond theoretical benchmarks, proving its efficiency for practical deployment in missions like search and rescue or surveillance.

In summary, our contributions are as follows:

- We propose the Multiscale Hierarchy Blender (MHB) which hierarchically fuses static and dynamic templates with the search region to enhance multiscale consistency and robustness.
- We introduce the Adaptive Key Token Gate (AKTG) that dynamically balances local and global feature cues, suppresses background noise and emphasizes object-related tokens.
- We design Nighttime Template Calibrator (NTC) module which adaptively updates dynamic templates through an offset-aware mechanism, ensuring reliable long-term tracking under challenging conditions.
- We achieved new state-of-the-art (SOTA) results on five benchmarks while maintaining real-time performance. Furthermore, we validated the system's practicality through real-world UAV deployment, demonstrating a complete chain from algorithmic innovation to practical application.

II. Related Works

A. Single Object Tracking

The purpose of single object tracking is to track an object in challenging scenarios such as those with similar object interference, occlusion and complex backgrounds. With the development of deep learning, MixFormer [8], as a concise end-to-end model based on Transformer, relies on a backbone network that mixes the template and search images together with a regression head to directly output tracking results. ARTrack [20], [21] transforms tracking into a coordinate sequence interpretation task. OSTrack [22] adopts the ViT network architecture to build an efficient visual tracking framework. VideoTrack [10] integrates context information. Similarly, ODTrack [9] proposes a token sequence propagation method to associate various types of context information. OTETrack [6] adds an additional template and continuously updates the additional template to provide more information. To reduce the complexity of the tracker, MCITrack [23] uses the mamba and leverages its linear complexity optimization and long-sequence processing capabilities to build a new tracking framework. The EVPTrack [24] uses a spatio-temporal encoder to propagate information between consecutive frames through tokens and combines a prompt generator to generate multiscale and spatio-temporal explicit visual prompts. LoReTrack [25] improves the tracking performance by enabling the low-resolution tracker to inherit the feature interaction of the high-resolution model from a global perspective.

B. Nighttime UAV Tracking

Due to factors such as lower illumination conditions, nighttime UAV tracking is a much more challenging task. For light enhancement, Highlightnet [12] uses a pixel-level range mask in its adaptive low-light enhancer to focus on targets, Darklighter [13] improves low-light image quality by estimating light and noise maps, Ye et al. [26] train the SCT enhancer via task-inspired perceptual loss for denoising and light adjustment, and ADTrack [27] combines a low-light enhancer with a correlation filter-based framework. Regarding domain adaptation, Fu et al. [11] align day-night spatio-temporal contexts. SAM-DA [15] proposes a training framework. UDAT [14] proposes the first unsupervised domain adaptation nighttime aerial tracking framework. DCPT [16] learns visual prompts iteratively, and MambaNUT [17] leverages Vision Mamba's [28] linear complexity. DARTer [18] is an end-to-end framework for nighttime UAV tracking that improves accuracy and efficiency by adaptively fusing multi-perspective features and activating Vision Transformer layers based on the scene's dynamics. However, these methods rely on extensive training, incur high costs, increase optimization complexity, and fail to fully exploit dynamic information from extreme viewpoint changes. Compared to

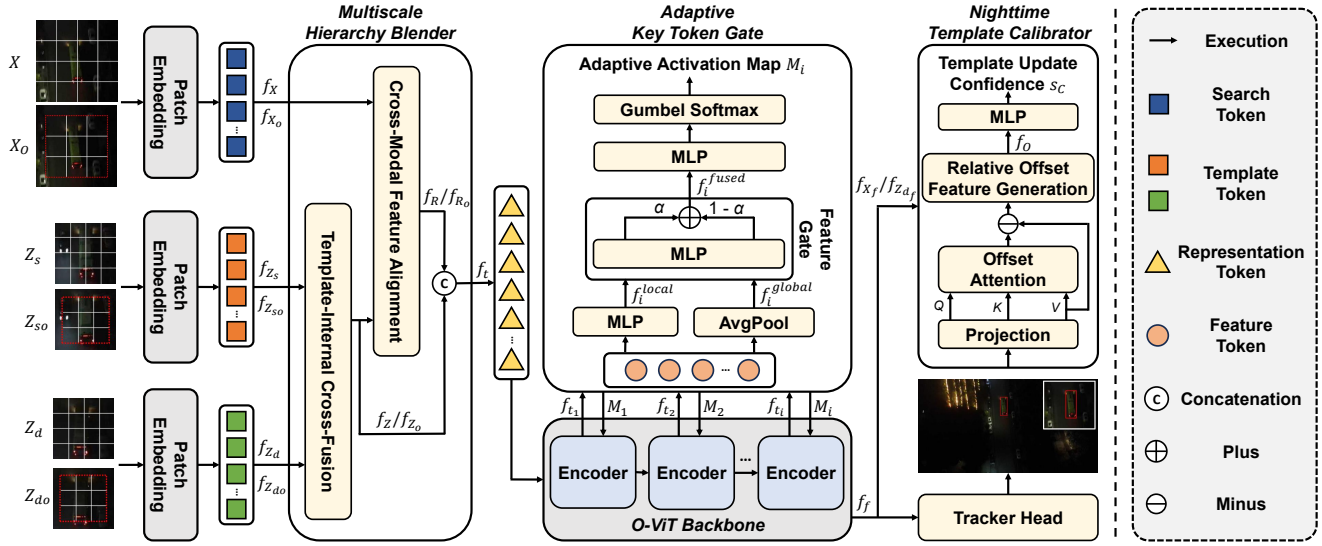

Fig. 2. Overview architecture of MATrack. The nighttime dynamic features of the static, dynamic templates and research region are fused by Multiscale Hierarchy Blender (MHB) module. As the O-ViT backbone performs feature extraction and interaction, the Adaptive Key Token Gate (AKTG) suppresses background noise tokens while emphasizing object-related information. We also designed the Nighttime Template Calibrator (NTC) to adaptively update the dynamic template and ensure reliable long-term tracking under challenging conditions.

existing methods, MATrack is a robust, end-to-end system solution that uses three unique core modules. It effectively suppresses background noise and visual artifacts common in nighttime environments, achieving both high accuracy and real-time efficiency, as validated by real-world UAV deployment.

III. Methodology

A. Overview

We propose a nighttime UAV tracking framework, named MATrack. Its architecture is shown in Fig. 2. The process begins by taking the search image, along with both static and dynamic templates, and slicing them into O-patches [6]. We use a Multiscale Hierarchy Blender (MHB) to align features of different scales from the static and dynamic templates and the search image. Subsequently, all these features are fed into the O-ViT [6]. Furthermore, we use an Adaptive Key Token Gate (AKTG) to suppress background noise tokens while enhancing object-related tokens, thus improving tracking performance. Simultaneously, we use a Nighttime Template Calibrator (NTC) to enable efficient and accurate template updates. The following sections will provide more details on these components.

B. Multiscale Hierarchy Blender

The input images of MATrack include the search images $X \in \mathbb{R}^{3\times H_x\times W_x}$, the static template images $Z_s \in \mathbb{R}^{3\times H_{z_s}\times W_{z_s}}$ and the dynamic template images $Z_d \in \mathbb{R}^{3\times H_{z_d}\times W_{z_d}}$. We adopt the current ViT-based tracking paradigm [22], partitioning images into patches and then converting them into token sequences. After processing, we obtain the initial search features f_X, the static template features f_{Z_s} and the dynamic template features f_{Z_d}. Meanwhile, these images are sliced into O-patches [6], including f_{X_o}, $f_{Z_{so}}$ and $f_{Z_{do}}$, which enhances the correlation between image patches across different scales.

To accurately capture object features at different scales, effectively filter isolated background noise and highlight consistent features between the object and templates, we use the Multiscale Hierarchy Blender (MHB) module to perform hierarchical feature fusion on the static template, dynamic template and search image.

Specifically, we perform Template-Internal Cross-Fusion on the features of the initial static and dynamic templates (f_{Z_s} and f_{Z_d}), as well as their overlapped features ($f_{Z_{s_o}}$ and $f_{Z_{d_o}}$). This process yields the primary blended features, f_Z and f_{Z_o}. The calculation for the initial static and dynamic templates is as follows:

$$\begin{aligned} f_{Z'_s} &= \Phi_{CA}(f_{Z_s}, f_{Z_d}), & f_{Z'_{so}} &= \Phi_{CA}(f_{Z_{so}}, f_{Z_{do}}),\\ f_{Z'_d} &= \Phi_{CA}(f_{Z_d}, f_{Z_s}), & f_{Z'_{do}} &= \Phi_{CA}(f_{Z_{do}}, f_{Z_{so}}),\\ f_Z &= \mathrm{Concat}(f_{Z'_s}, f_{Z'_d}), & f_{Z_o} &= \mathrm{Concat}(f_{Z'_{so}}, f_{Z'_{do}}), \end{aligned} \tag{1}$$

where Φ_{CA} represents the cross-attention operation and Concat represents the concatenation operation. In this operation, the first element functions as Q and the second element is used to acquire K and V [7].

Subsequently, we adopt Cross-Modal Feature Alignment to achieve multiscale feature alignment between the search feature f_X with f_{X_o} and the templates features f_Z with f_{Z_o}, generating two cross-modal interactive feature representations f_R and f_{R_o}. This provides a more comprehensive feature foundation for subsequent matching.

$$f_R = \Phi_{CA}(f_X, f_Z), \quad f_{R_o} = \Phi_{CA}(f_{X_o}, f_{Z_o}). \tag{2}$$

Finally, we perform global feature integration. All fused multiscale features are concatenated to form a feature representation f_t that contains the global information of both the

search frame and the templates:

$$f_t = \text{Concat}(f_R, f_{R_o}, f_Z, f_{Z_o}). \tag{3}$$

C. Adaptive Key Token Gate

We propose the Adaptive Key Token Gate (AKTG) module. This module calculates the Adaptive Activation Map based on the fused features from the previous O-ViT block [6] combined with the feature gate, dynamically adjusts the attention to local and global information, and then suppresses background noise tokens and emphasizes object-related tokens through attention correction.

First, the AKTG module performs fine-grained feature splitting and sub-feature extraction on each output feature f_{t_i} from the i-th O-ViT. These features are split by the number of attention heads h, into sub-features f_i, with each attention head processing a sub-feature independently.

Then, a dual-path feature extraction is performed on each sub-feature f_i. Specifically, we perform local nighttime feature extraction to capture single-token details, denoted as f_i^{local}:

$$f_i^{local} = \text{MLP}(f_i). \tag{4}$$

Simultaneously, we extract global nighttime features to capture overall contextual relationships, denoted as f_i^{global}:

$$f_i^{global} = \text{AvgPool}(\text{MLP}(f_i)). \tag{5}$$

To address the unreliability of local details and the robustness of global information in nighttime UAV tracking, we propose the feature gate mechanism to adaptively weigh and fuse local and global features, which handles complex and changing nighttime environments.

Specifically, we input f_i^{local} and f_i^{global} into the feature gate to obtain the activation weights α:

$$\alpha = \text{MLP}(\text{Concat}(f_i^{local}, f_i^{global})). \tag{6}$$

Using the predicted activation weights α, we perform a weighted sum of the local and global features to obtain the final fused feature f_i^{fused}:

$$f_i^{fused} = \alpha \odot f_i^{local} + (1 - \alpha) \odot f_i^{global}, \tag{7}$$

where $\odot$ denotes the element-wise multiplication.

We then employ Gumbel-Softmax [29] to generate the Adaptive Activation Map $M_i \in \{0 \sim 1\}^N$:

$$M_i = \text{GumbelSoftmax}(\text{MLP}(f_i^{fused})). \tag{8}$$

In complex nighttime scenarios, we dynamically adjust the focus on local and global information using the Adaptive Activation Map M_i.

To continuously suppress background noise tokens and emphasize object-related tokens, i.e. key tokens, we apply attention correction to the attention map A_{map_i} within the O-ViT block:

$$\text{AC}_{map_i} = (A_{map_i} \cdot M_i + A_{map_i}) \cdot V_i, \tag{9}$$

where V_i is the value matrix from i-th O-ViT block.

D. Nighttime Template Calibrator

In complex nighttime environments, previous trackers often rely on fixed time intervals or simple thresholds for dynamic template updates [18], which can easily lead to low-quality or even invalid dynamic templates and in turn reduce tracking accuracy and efficiency. To address this challenge, we propose the Nighttime Template Calibrator (NTC) module, which performs dynamic template calibration through an offset-aware mechanism.

From the output of the final O-ViT block f_f, we partition it by index into f_{X_f}, f_{Z_f} and $f_{Z_{d_f}}$. We then use Offset-Attention to compute the relative offset between the dynamic template and the search frame.

We map the features to Q_n, K_n and V_n matrices:

$$\begin{aligned} Q_n &= \Phi_p(f_{Z_{d_f}}), \\ K_n &= \Phi_p(f_{X_f}), \\ V_n &= \Phi_p(f_{X_f}), \end{aligned} \tag{10}$$

where Φ_p represents the projection operation.

After that, we perform offset attention calculation, which is computed as follows:

$$Attention(Q_n, K_n, V_n) = \text{Softmax}\left(\frac{Q_n \cdot K_n^T}{\sqrt{d_k}}\right) \cdot V_n, \tag{11}$$

where d_k is the dimension of the Q_n and K_n vectors.

We generate the relative offset feature f_O and gain the template update confidence $s_c \in (0, 1)$:

$$f_O = \text{ReLU}\left(\text{InsNorm}\left(\Phi_l(V_n - \text{Attention}(Q_n, K_n, V_n))\right)\right), \tag{12}$$

where Φ_l represents a linear transformation layer and InsNorm is a instance normalization operation.

$$s_c = \text{MLP}(f_O). \tag{13}$$

Let θ be the confidence score threshold. If $s_c \in \theta$, we update the dynamic template.

E. Prediction Head and Training Loss

Following the architecture of models such as MixFormer [8] and DARTer [18], we utilize a prediction head comprising four stacked Conv-BN-ReLU layers. This head first transforms the output tokens of the search image into a 2D spatial feature map. It then processes these features to output three distinct results for each potential object. The final bounding box is located at the position with the peak classification score.

For the training, MATrack's loss function, L_{total}, is a weighted combination of the softmax cross-entropy loss (L_{ce}) [20] and the SloU loss (L_{SloU}) [30], given by the formula $L_{\text{total}} = \lambda_1 L_{ce} + \lambda_2 L_{SloU}$. Both weights ($\lambda_1$ and λ_2) were set to 2 in our experiments.

IV. Experiment

A. Implementation Details

Models. We use Overlapped ViT [6] as the backbone. The head of MATrack consists of a stack of four Conv-BN-Relu

TABLE I

STATE-OF-THE-ART COMPARISON ON THE NAT2024-1 [11], NAT2021 [14] AND UAVDARK135 [19] BENCHMARKS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY. NOTE THAT THE PERCENT SYMBOL (%) IS EXCLUDED FOR PRECISION SCORE (P), NORMALIZED PRECISION (P_{Norm}) AND AREA UNDER THE CURVE (AUC).

Tracker	Source	NAT2024-1			NAT2021			UAVDark135		
		P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC
SiamCAR [31]	CVPR 20	68.7	62.6	51.2	65.8	59.5	45.7	65.8	65.7	52.3
Ocean [32]	ECCV 20	67.6	50.3	44.0	58.1	49.9	38.6	60.1	58.9	47.3
HiFT [33]	ICCV 21	57.1	44.5	40.8	54.5	46.7	37.0	44.8	45.2	35.3
SiamAPN++ [34]	IROS 21	68.9	57.9	47.8	60.2	51.4	41.2	42.7	41.6	33.5
UDAT-BAN [14]	CVPR 22	71.2	64.9	51.1	68.9	58.8	47.2	61.1	61.7	48.4
UDAT-CAR [14]	CVPR 22	68.1	61.6	49.6	68.2	61.3	48.7	60.9	61.3	48.6
TCTrack [35]	CVPR 22	74.4	51.2	47	60.8	51.9	40.8	49.8	50.0	37.7
TCTrack++ [4]	TPAMI 23	70.5	50.8	46.6	61.1	52.8	41.7	47.4	47.4	37.8
MAT [36]	CVPR 23	80.5	76.3	61.9	64.8	58.8	47.7	57.2	57.6	47.1
HiT-Base [37]	ICCV 23	62.7	56.9	48.2	49.3	44.2	36.4	48.9	48.7	41.1
Aba-ViTrack [38]	ICCV 23	78.4	72.2	60.1	60.4	57.3	46.9	61.3	63.5	52.1
SGDViT [39]	ICRA 23	53.1	47.2	38.1	53.1	47.9	37.5	40.2	40.6	32.7
TDA-Track [11]	IROS 24	75.5	53.3	51.4	61.7	53.5	42.3	49.5	49.9	36.9
AVTrack-DeiT [40]	ICML 24	75.3	68.2	56.7	61.5	55.6	45.5	58.6	59.2	47.6
DCPT [16]	ICRA 24	80.9	75.4	62.1	69.0	63.5	**52.6**	69.2	**69.8**	56.7
MambaNUT [17]	IROS 25	**83.3**	**76.9**	**63.6**	**70.1**	**64.6**	52.4	**70.0**	69.3	**57.1**
DARTer [18]	ICMR 25	**85.2**	**80.1**	**65.6**	**70.2**	**63.7**	**53.2**	**71.6**	**72.1**	**58.2**
MATrack	**Ours**	**87.7**	**82.7**	**68.0**	**72.1**	**65.9**	**54.6**	**77.5**	**77.5**	**62.4**

layers. The confidence score threshold is $\theta \in (0.3, 0.8)$. The image sizes of the search and template are 128×128 and 256×256, respectively. The initial and O patches of the search image are 16×16 and 15×15, and the initial and O patches of the template are 8×8 and 7×7, respectively.

Training. For training, we used four common datasets: LaSOT [41], GOT10K [42], COCO [43], and TrackingNet [44]. Additionally, we incorporated three nighttime datasets—BDD100K_Night [45], SHIFT_night [46], and ExDark [47]—to address low-light conditions. The model is trained for 150 epochs using the AdamW optimizer [48], with a batch size of 32. Each epoch involves 60,000 sampling pairs. The initial learning rate is set to 0.0001, and after 120 epochs, the learning rate decays at a rate of 10%. The model is trained on a server with four A40 GPUs.

Evaluation. We evaluate MATrack on five mainstream benchmarks, including NAT2024-1 [11], NAT2021 [14], UAVDark135 [19], NAT2021-L [14] and DarkTrack2021 [26]. We compare MATrack with the state-of-the-art (SOTA) trackers. All evaluation experiments are conducted on an RTX-4090 GPU.

B. Comparison Results

NAT2024-1. NAT2024-1 [11] focuses on simulating long-sequence tracking tasks in real-world low-illumination nighttime scenarios, addressing the insufficiency of existing benchmarks in covering long-term nighttime tracking scenarios. It comprises 40 long-term image sequences, with a total of more than 70,000 frames. As shown in Tab. I, MATrack achieves the best performance across all three metrics on NAT2024-1, it has a precision (P) score of 87.7%, a normalized precision (P_{Norm}) of 82.7% and an area under the curve (AUC) of 68.0%. These results surpass the second-best tracker DARTer [18] by 2.5%, 2.6% and 2.4%, which confirms the robustness of MATrack on challenging UAV tracking sequences.

TABLE II

COMPARISON ON THE NAT2021-L [14] BENCHMARK. THE TOP THREE RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY.

Tracker	Source	NAT2021-L		
		P	P_{Norm}	AUC
SiamRPN++ [49]	CVPR 19	42.9	35.8	30.0
Ocean [32]	ECCV 20	45.1	40.0	31.6
HiFT [33]	ICCV 21	43.0	33.0	28.8
SiamAPN [50]	ICRA 21	37.7	27.7	24.2
SiamAPN++ [34]	IROS 21	40.0	32.7	28.0
UDAT-BAN [14]	CVPR 22	49.4	43.7	35.3
UDAT-CAR [14]	CVPR 22	50.4	44.7	37.8
DCPT [16]	ICRA 24	**58.6**	**54.6**	**47.4**
DARTer [18]	ICMR 25	**64.9**	**58.6**	**50.9**
MATrack	**Ours**	**67.7**	**60.8**	**52.5**

NAT2021. NAT2021 [14] is specifically designed for tracking tasks in nighttime scenarios. It fills the gap of evaluation data in the field of nighttime UAV tracking and covers multiple dimensions including objects, environments and illumination. On NAT2021, MATrack achieves P, P_{Norm} and AUC scores of 72.1%, 65.9% and 54.6%, respectively, outperforming all existing trackers. This highlights MA-

Track's ability to maintain consistent accuracy even when objects undergo significant variations.

NAT2021-L. NAT2021-L [14] is a long-term tracking benchmark that provides sufficient nighttime tracking videos for evaluating the performance of nighttime UAV tracking algorithms in long-term tracking scenarios, with each sequence containing more than 1,400 frames. On the NAT2021-L benchmark, MATrack records 67.7% in P, 60.8% in P_{Norm} and 52.5% in AUC, ranking first across all metrics, showing that our tracker is much less affected by error accumulation in long nighttime scenarios.

TABLE III
COMPARISON ON THE DARKTRACK2021 [26] BENCHMARK. THE TOP THREE RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY.

Tracker	Source	DarkTrack2021		
		P	P_{Norm}	AUC
SiamRPN [51]	CVPR 18	50.9	48.5	38.7
DIMP18 [52]	ICCV 19	62.0	58.9	47.1
PRDIMP50 [53]	CVPR 20	58.0	55.9	46.4
SiamAPN++ [34]	IROS 21	48.9	46.1	37.7
HiFT [33]	ICCV 21	50.3	47.1	37.4
SiamAPN++-SCT [26]	RAL 22	53.7	51.1	40.8
DIMP50-SCT [26]	RAL 22	**67.7**	63.3	52.1
DCPT [16]	ICRA 24	**67.7**	**64.6**	**54.0**
DARTer [18]	ICMR 25	**67.6**	**64.8**	**54.5**
MATrack	**Ours**	**73.1**	**70.2**	**58.6**

UAVDark135. UAVDark135 [19] is a benchmark specifically built for UAV nighttime tracking tasks. It defines a variety of common challenging attributes and consists of 135 sequences, with a total frame count of 125,466. On UAVDark135, MATrack demonstrates superior adaptability to low-light environments with a P of 77.5%, a P_{Norm} of 77.5% and an AUC of 62.4%. Compared to DARTer, MATrack surpasses the SOTA tracker by 6.0%, 5.4% and 4.2% in P, P_{Norm}, and AUC, respectively, suggesting that its design generalizes well to extreme lighting conditions.

DarkTrack2021. DarkTrack2021 [26] provides comprehensive evaluation support for nighttime UAV tracking algorithms. It contains a total of 110 challenging sequences and covers a rich variety of real-world nighttime UAV tracking scenarios. On DarkTrack2021, MATrack reaches the SOTA level in P, P_{Norm} and AUC, delivering a strong lead over all trackers. For instance, compared to DCPT and DARTer, MATrack improves by over 5% in precision and nearly 4% in AUC. These gains highlight MATrack's resilience under both low illumination and cluttered backgrounds, where most trackers tend to fail due to noisy features.

C. Efficiency analysis

As demonstrated in Tab. IV, MATrack achieves a promising trade-off between speed and model size. Specifically, MATrack runs at 81 FPS, which is substantially faster than recent high-parameter trackers such as DCPT [16], while maintaining comparable parameter scale. Compared with lightweight trackers such as MambaNUT [17], MATrack achieves significantly higher FPS than most heavy architectures and provides a stronger balance between efficiency and representational capacity.

TABLE IV
COMPARISON OF OUR TRACKER AND OTHER STATE-OF-THE-ART TRACKERS IN TERMS OF AVERAGE FPS AND PARAMETERS (M).

Tracker	Source	Average FPS	Parameters
SiamCAR [31]	CVPR 20	37	51.3
Ocean [32]	ECCV 20	43	25.8
HiFT [33]	ICCV 21	123	9.9
SiamAPN++ [34]	IROS 21	114	14.7
UDAT-BAN [14]	CVPR 22	41	54.1
UDAT-CAR [14]	CVPR 22	36	54.6
TCTrack [35]	CVPR 22	136	8.5
TCTrack++ [4]	TPAMI 23	122	8.8
MAT [36]	CVPR 23	56	88.4
HiT-Base [37]	ICCV 23	156	42.1
Aba-ViTrack [38]	ICCV 23	134	7.9
SGDViT [39]	ICRA 23	93	23.3
TDA-Track [11]	IROS 24	114	9.2
AVTrack-DeiT [40]	ICML 24	212	7.9
DCPT [16]	ICRA 24	35	92.9
MambaNUT [17]	IROS 25	72	4.1
DARTer [18]	ICMR 25	74	80.9
MATrack	**Ours**	81	76.2

D. Ablation Study

To verify the effectiveness of the proposed modules, we introduce them to the baseline tracker incrementally and provide experimental results on the NAT2024-1 [11] dataset.

TABLE V
ABLATION STUDIES ON NIGHTTIME UAV TRACKING BENCHMARKS NAT2024-1 [11].

Method	P	P_{Norm}	AUC
Base	84.1	79.6	65.1
Base+MHB	85.8 (1.7↑)	80.8 (1.2↑)	66.2 (1.1↑)
Base+MHB+AKTG	86.9 (2.8↑)	81.6 (2.0↑)	67.0 (1.9↑)
Base+MHB+AKTG+NTC	87.7 (3.6↑)	82.7 (3.1↑)	68.0 (2.9↑)

Base+MHB. The MHB module is the core of our tracker. It combines the stability of static templates with the dynamism of dynamic templates through multiscale feature fusion. Simultaneously, it aligns the features of the search frame with the templates at different scales, which boosts the robustness of the feature representation. As shown in Tab.V, the incorporation of the MHB module resulted in a 1.1% increase of AUC.

Base+MHB+AKTG. We further add the AKTG module. As shown in Tab. V, our MATrack achieves a P of 86.9%, surpassing the baseline by 2.8%. This demonstrates that the AKTG module provides robust, noise-resistant tracking by using its feature gate mechanism to adaptively focus on the object and suppress background noise.

Base+MHB+AKTG+NTC We further introduced the NTC module to achieve efficient and accurate dynamic template updates through its offset-aware capability. As detailed

in Tab. V, the NTC module outperforms the baseline on the AUC, P and P_{Norm} metrics, showing the effectiveness of our proposed module.

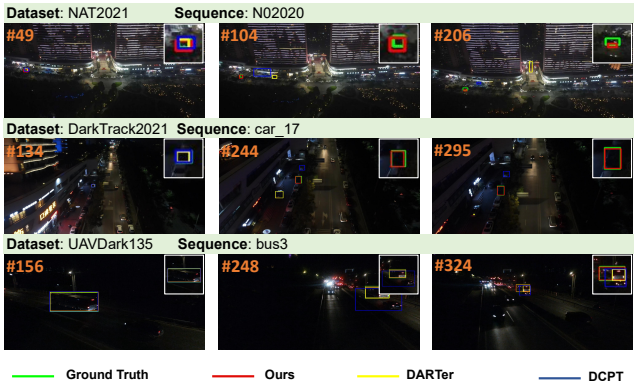

Fig. 3. Qualitative comparison results of our tracker with other two latest trackers (i.e., DCPT [16] and DARTer [18]) in representative nighttime scenarios. Our method maintains its robustness even in complex environments. Better viewed in color with zoom-in.

E. Qualitative Analysis

As shown in Fig. 3, we visualize the tracking results of our model and the previous two SOTA models on three challenging sequences from NAT2021 [14], DarkTrack2021 [26] and UAVDark135 [19]. In these sequences, the scenes contain distractors, and the state of the object undergoes significant changes. It is evident that our model exhibits greater robustness compared to others. This validates that our method contributes to addressing these challenges, further demonstrating the efficiency of our proposed method.

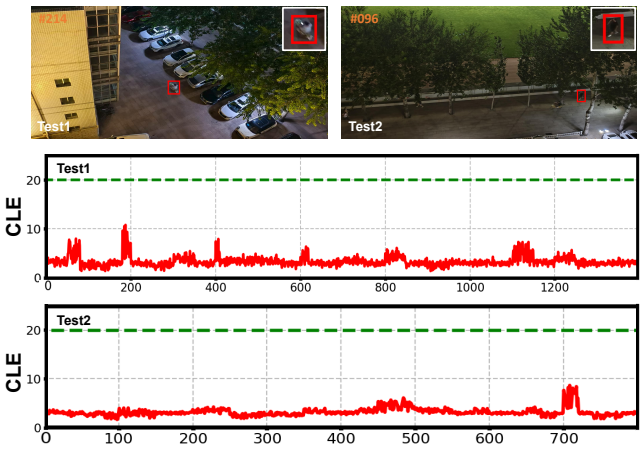

Fig. 4. The reliability of our system is validated through real-world UAV platform tests in nighttime tracking scenarios. The frame-wise performance, represented by Center Location Error (CLE) plots, demonstrates that our tracker's errors are consistently below the green dashed line (CLE = 20 pixels), which is the threshold for acceptable performance.

V. Real-World Testing

As shown in Fig. 4, we conduct real-world tests to verify the performance of MATrack. We use the on-board camera on an UAV to capture nighttime images, and transmit these images to a computer in real time via Wi-Fi communication. The computer is equipped with an Nvidia 2080ti GPU, which can process the received images at a speed of over 30 FPS. The main challenges of these scenarios include viewpoint changes, partial occlusions and background noise. However, MATrack still achieve excellent performance, with the Center Location Error (CLE) of all test frames maintained below 20 pixels. Real-world tests show that MATrack is highly suitable for edge deployment on UAV platforms, delivering robust tracking performance in complex nighttime environments.

VI. Conclusion

We introduce MATrack, a multiscale adaptive tracker that addresses the challenges of nighttime UAV tracking. By combining a Multiscale Hierarchy Blender for robust feature fusion, an Adaptive Key Token Gate for noise-resistant feature selection and a Nighttime Template Calibrator for dynamic template updates, MATrack sets a new state of the art. Extensive experiments show that MATrack consistently outperforms leading trackers in accuracy and robustness. Most importantly, it strikes a crucial balance between performance and efficiency, proving its practicality for real-time operation on UAVs. This makes MATrack a highly valuable solution for real-world low-light surveillance.

References

[1] X. Lei, X. Hu, G. Wang, and H. Luo, "A multi-uav deployment method for border patrolling based on stackelberg game," *Journal of Systems Engineering and Electronics*, vol. 34, no. 1, pp. 99–116, 2023.

[2] A. Al-Kaff, M. J. Gómez-Silva, F. M. Moreno, A. De La Escalera, and J. M. Armingol, "An appearance-based tracking algorithm for aerial search and rescue purposes," *Sensors*, vol. 19, no. 3, p. 652, 2019.

[3] B. Tian, Q. Yao, Y. Gu, K. Wang, and Y. Li, "Video processing techniques for traffic flow monitoring: A survey," in *2011 14th international IEEE conference on intelligent transportation systems (ITSC)*. IEEE, 2011, pp. 1103–1108.

[4] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Towards real-world visual tracking with temporal contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] L. Shen, X. Fan, and H. Li, "Overlapped trajectory-enhanced visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 608–13 618.

[9] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, and X. Li, "Odtrack: Online dense temporal token learning for visual tracking," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 2024, pp. 7588–7596.

[10] F. Xie, L. Chu, J. Li, Y. Lu, and C. Ma, "Videotrack: Learning to track objects via video transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 826–22 835.

[11] C. Fu, Y. Wang, L. Yao, G. Zheng, H. Zuo, and J. Pan, "Prompt-driven temporal domain adaptation for nighttime uav tracking," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9706–9713.

[12] C. Fu, H. Dong, J. Ye, G. Zheng, S. Li, and J. Zhao, "Highlightnet: highlighting low-light potential features for real-time uav tracking," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 146–12 153.

[13] J. Ye, C. Fu, G. Zheng, Z. Cao, and B. Li, "Darklighter: Light up the darkness for uav tracking," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3079–3085.

[14] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised domain adaptation for nighttime aerial tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8896–8905.

[15] C. Fu, L. Yao, H. Zuo, G. Zheng, and J. Pan, "Sam-da: Uav tracks anything at night with sam-powered domain adaptation," in *2024 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2024, pp. 31–38.

[16] J. Zhu, H. Tang, Z.-Q. Cheng, J.-Y. He, B. Luo, S. Qiu, S. Li, and H. Lu, "Dcpt: Darkness clue-prompted tracking in nighttime uavs," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7381–7388.

[17] Y. Wu, X. Yang, X. Wang, H. Ye, D. Zeng, and S. Li, "Mambanut: Nighttime uav tracking via mamba and adaptive curriculum learning," *arXiv preprint arXiv:2412.00626*, 2024.

[18] X. Li, X. Li, and S. Hu, "Darter: Dynamic adaptive representation tracker for nighttime uav tracking," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 2025, pp. 1998–2002.

[19] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "All-day object tracking for unmanned aerial vehicle," *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4515–4529, 2022.

[20] X. Wei, Y. Bai, Y. Zheng, D. Shi, and Y. Gong, "Autoregressive visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9697–9706.

[21] Y. Bai, Z. Zhao, Y. Gong, and X. 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*, 2024, pp. 19 048–19 057.

[22] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *European conference on computer vision*. Springer, 2022, pp. 341–357.

[23] B. Kang, X. Chen, S. Lai, Y. Liu, Y. Liu, and D. Wang, "Exploring enhanced contextual information for video-level object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 4194–4202.

[24] L. Shi, B. Zhong, Q. Liang, N. Li, S. Zhang, and X. Li, "Explicit visual prompts for visual object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 4838–4846.

[25] S. Dong, Y. Feng, Q. Yang, Y. Lin, and H. Fan, "Loretrack: efficient and accurate low-resolution transformer tracking," *arXiv preprint arXiv:2405.17660*, 2024.

[26] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker meets night: A transformer enhancer for uav tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3866–3873, 2022.

[27] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "Adtrack: Target-aware dual filter learning for real-time anti-dark uav tracking," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 496–502.

[28] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[29] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 134–18 144.

[30] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.

[31] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "Siamcar: Siamese fully convolutional classification and regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6269–6277.

[32] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *European Conference on Computer Vision (ECCV)*, 2020.

[33] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "Hift: Hierarchical feature transformer for aerial tracking," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 457–15 466.

[34] ——, "Siamapn++: Siamese attentional aggregation network for real-time uav tracking," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 3086–3092.

[35] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Tctrack: Temporal contexts for aerial tracking," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 778–14 788, 2022.

[36] H. Zhao, D. Wang, and H. Lu, "Representation learning for visual object tracking by masked appearance transfer," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 696–18 705.

[37] B. Kang, X. Chen, D. Wang, H. Peng, and H. Lu, "Exploring lightweight hierarchical vision transformers for efficient visual tracking," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9578–9587, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260887522

[38] S. Li, Y. Yang, D. Zeng, and X. Wang, "Adaptive and background-aware vision transformer for real-time uav tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 989–14 000.

[39] L. Yao, C. Fu, and et al, "Sgdvit: Saliency-guided dynamic vision transformer for uav tracking," *arXiv preprint arXiv:2303.04378*, 2023.

[40] Y. Li, M. Liu, Y. Wu, X. Wang, X. Yang, and S. Li, "Learning adaptive and view-invariant vision transformer for real-time uav tracking," in *Forty-first International Conference on Machine Learning*, 2024.

[41] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5374–5383.

[42] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[44] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 300–317.

[45] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[46] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. 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*, 2022, pp. 21 371–21 382.

[47] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[48] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[49] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4282–4291.

[50] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Siamese anchor proposal network for high-speed aerial tracking," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 510–516.

[51] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[52] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6182–6191.

[53] M. Danelljan, "Probabilistic regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7183–7192.