

Prompt-Driven Temporal Domain Adaptation for Nighttime UAV Tracking

Changhong Fu^{1*}, Yiheng Wang¹, Liangliang Yao¹, Guangze Zheng², Haobo Zuo², and Jia Pan²

Abstract—Nighttime UAV tracking under low-illuminated scenarios has achieved great progress by domain adaptation (DA). However, previous DA training-based works are deficient in narrowing the discrepancy of temporal contexts for UAV trackers. To address the issue, this work proposes a prompt-driven temporal domain adaptation training framework to fully utilize temporal contexts for challenging nighttime UAV tracking, *i.e.*, TDA. Specifically, the proposed framework aligns the distribution of temporal contexts from daytime and nighttime domains by training the temporal feature generator against the discriminator. The temporal-consistent discriminator progressively extracts shared domain-specific features to generate coherent domain discrimination results in the time series. Additionally, to obtain high-quality training samples, a prompt-driven object miner is employed to precisely locate objects in unannotated nighttime videos. Moreover, a new benchmark for long-term nighttime UAV tracking is constructed. Exhaustive evaluations on both public and self-constructed nighttime benchmarks demonstrate the remarkable performance of the tracker trained in TDA framework, *i.e.*, TDA-Track. Real-world tests at nighttime also show its practicality. The code and demo videos are available at <https://github.com/vision4robotics/TDA-Track>.

I. INTRODUCTION

Visual object tracking on intelligent unmanned aerial vehicles (UAVs) has caught widespread interest for its versatility in various real-world applications, *e.g.*, navigation [1], search and rescue mission [2], as well as security and surveillance [3]. Significant advancements in UAV tracking performance have been made in favorably-illuminated scenarios [4]–[7]. However, the images captured by UAVs at night have much lower contrast, brightness, and signal-to-noise ratio [8] than ones captured in the daytime, resulting in a huge domain discrepancy between nighttime and daytime visual feature spaces. Hence, state-of-the-art (SOTA) trackers suffer from severe tracking capability degradation. Moreover, the temporal contexts, vital information contained in the consecutive frames, haven't been utilized in nighttime UAV tracking. Consequently, robust nighttime UAV tracking is far from being properly addressed.

One promising solution to the domain discrepancy challenge encountered by UAV tracking is domain adaptation (DA) [8], [10]. Specifically, the DA training framework includes a feature generator and a discriminator. Trained

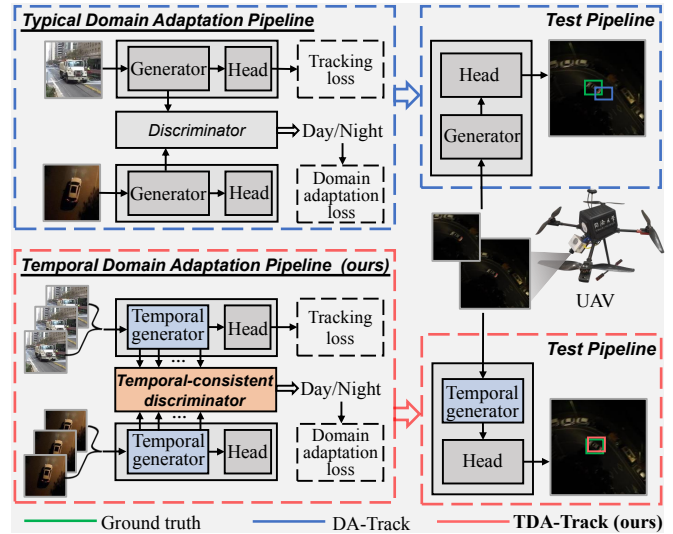


Fig. 1. Comparison with previous domain adaptation (DA) training framework for nighttime UAV tracking. The proposed temporal domain adaptation (TDA) training framework generates the temporal contexts among daytime and nighttime images, and then narrows the feature discrepancy of temporal contexts from different domains with the temporal-consistent discriminator. (Image frames are from GOT-10k [9] and NAT2021-train [8].)

to deceive the discriminator, the generator learns to extract domain-invariant features from unordered daytime and nighttime training samples. Meanwhile, the classification and localization capabilities are trained under the supervision of daytime annotations. However, the DA training on static image pairs neglects the strong temporal contexts that inherently exist among consecutive frames, which is unsatisfactory for UAV tracking which highlights the continuity. The feature distribution is not consistently aligned among consecutive frames, leading to the failure in long-term object tracking. Moreover, the plain Transformer discriminators [8], [10] fail to generate coherent discrimination results due to context changes among consecutive frames, which undermines the practicality of the DA framework for UAV tracking. **Therefore, how to consistently align temporal contexts and to obtain coherent discrimination is worth attention in DA training-based nighttime UAV tracking.**

Previous DA training frameworks for nighttime UAV tracking [8], [10] discover nighttime training samples with video saliency detection model [11] and segment anything model (SAM) [12]. However, these object discovery approaches focus on the object position in the single frame and are unsatisfactory in building smooth trajectories of uncovered objects, which results in inconsistent training samples and suboptimal performances. **Hence, how to generate high-**

*Corresponding author

¹C. Fu, Y. Wang, and L. Yao are with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China. Email: changhongfu@tongji.edu.cn

²G. Zheng, H. Zuo, and J. Pan are with the Department of Computer Science, the University of Hong Kong, Hong Kong, China.

continuity training samples for nighttime UAV tracking remains unresolved.

This work proposes a novel framework, *i.e.*, TDA, which explores temporal contexts in DA training for nighttime UAV tracking and utilizes prompts to obtain high-quality training samples. The Baseline tracker [13] trained in TDA framework is symbolized as TDA-Track. As illustrated in Fig. 1, to alleviate the inconsistency of temporal contexts, the proposed training framework aligns the distribution of temporal contexts and image features from daytime and nighttime domains. Specifically, given a sequence with t frames, the temporal generator maps frames into temporal contexts and image features, which are classified into different domains by discriminator. In the adversarial training manner, the temporal generator is trained to deceive the discriminator and consistently obtain domain-invariant representations across the time series. To improve domain adaptability, a novel temporal-consistent discriminator is designed to make more coherent discrimination results with common features, which are robust representations progressively extracted from temporal contexts. Moreover, benefited from progress in multi-modal learning [14], [15], a prompt-driven object mining approach is proposed to acquire high-quality training samples. Objects are mined by detection [16] with text prompt and associated into smooth trajectories in the time series. The main contributions of this work are summarized as follows:

- A novel temporal domain adaptation training framework is proposed for nighttime UAV tracking. To the best of our knowledge, our method is the first work to explore the power of temporal contexts in training nighttime UAV trackers.
- An innovative temporal-consistent Transformer discriminator is designed to boost the coherency of discrimination results by progressively extracting common features from temporal contexts.
- A new prompt-driven object mining approach is put forward to mine highly-continuous training patches from nighttime videos. It surpasses previous pre-processing methods by excellent consistency and valuable object categories.
- A novel long-term nighttime UAV tracking benchmark is constructed to enlarge the existing benchmarks for a fair comparison. Evaluations on both public and self-constructed benchmarks show the remarkable nighttime UAV tracking performance of TDA-Track. Real-world deployment on the UAV platform demonstrates its robustness and practicality.

II. RELATED WORKS

A. Nighttime UAV Tracking

Real-world UAV tracking applications are severely impeded by low illumination at nighttime. Recently, the approaches to boost nighttime UAV tracking performance are founded on either low-light enhancement or DA. For enhancement-based nighttime UAV tracking [17], [18], different enhancers are proposed to improve the image illumination ahead of trackers. However, due to the limited

relationship between low-light image enhancement and UAV tracking, integrating enhancers and trackers in the plug-and-play manner leads to suboptimal performance. For DA training-based nighttime UAV tracking [8], [10], trackers are trained on static image pairs to obtain domain-invariant features for predicting the object location in upcoming frames. However, existing DA training frameworks are insufficient in aligning the distribution of temporal contexts, restricting performance improvement of nighttime UAV tracking.

B. Temporal Contexts

Temporal contexts have aroused much research interest for the effectiveness in object tracking. B. Yan *et al.* [19] explicitly replace the template with high-confidence patches. Z. Fu *et al.* [20] and N. Wang *et al.* [21] design architectures to fuse previous template features into robust template representations. To exploit temporal contexts more comprehensively, Z. Cao *et al.* [13], [22] introduce temporal knowledge into feature extraction and similarity map refinement. In challenging nighttime UAV tracking, temporal contexts among consecutive frames are valuable hints for the possible object location. However, existing DA methods for nighttime UAV tracking have not taken temporal contexts into consideration. Moreover, the lack of high-quality training samples that follow objects with smooth trajectories poses great challenges for trackers to learn temporal contexts.

C. Temporal Domain Adaptation

To achieve favorable performance on the target domain with the power of temporal contexts, mappings from source domain representations to target domain temporal contexts should be coherent in the time series, which remains a problem in temporal domain adaptation. To alleviate the problem, B. Pan *et al.* [23] design a discriminator architecture for domain distribution matching and temporally aligned distribution matching. I. Shin *et al.* [24] adopt a sequence discriminator to take soft segmentation maps in a sequence for domain classification. X. Feng *et al.* [25] and D. Guan *et al.* [26] design two discriminators to classify spatial features from a single image and spatial-temporal features from consecutive frames. However, these works adopt simple discriminator architectures for temporal domain adaptation. The incoherent discrimination results misguide the generator in adversarial training and undermines the domain adaptability. Moreover, temporal domain adaptation for nighttime UAV tracking has not been investigated yet.

III. PROPOSED METHOD

The overview of the TDA training pipeline is shown in Fig. 2. The proposed framework innovatively introduces temporal contexts into domain adaptation from two perspectives: 1) To boost nighttime UAV tracking with robust temporal representations, the temporal contexts of consecutive frames from different domains are aligned jointly. 2) The discriminator progressively extracts common features from temporal contexts, which benefits domain classification by highly representative features for nighttime attributes.

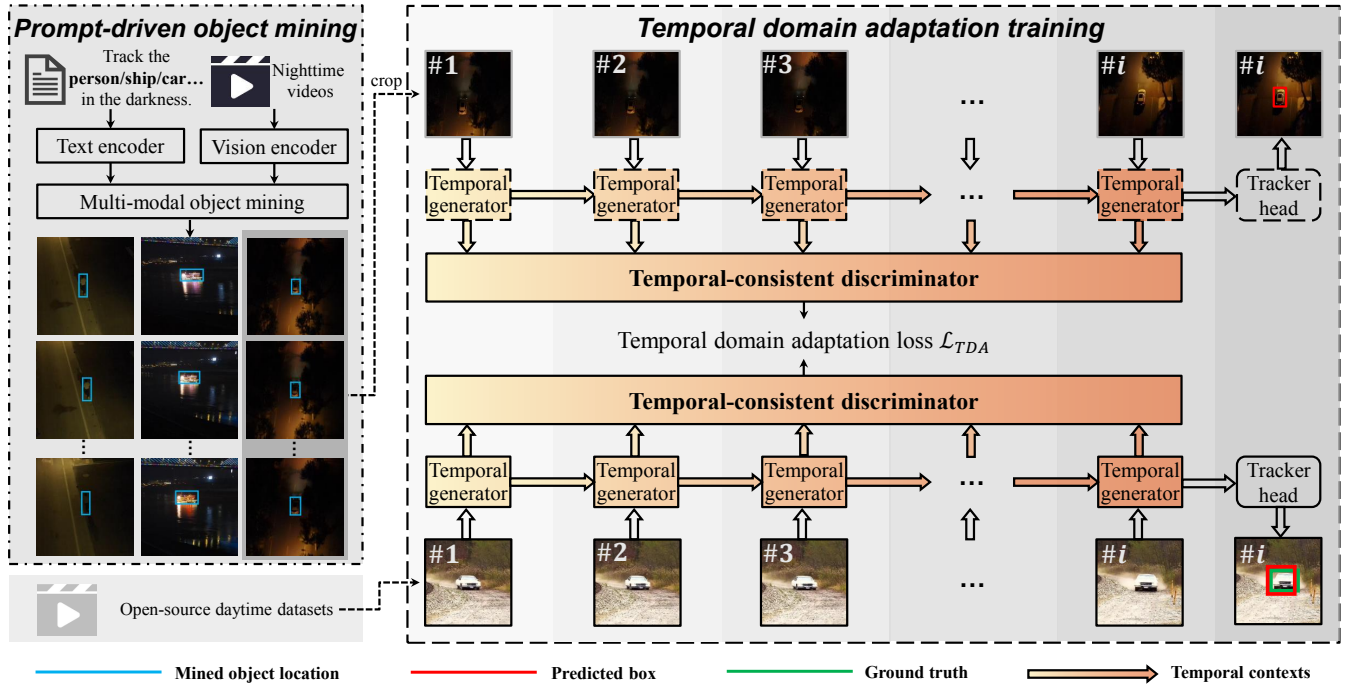


Fig. 2. Overview of the temporal day-to-night domain adaptation framework for nighttime UAV tracking. The *temporal generator* learns to generate temporal contexts that are more adaptive to the nighttime domain. The *temporal-consistent discriminator* is trained to classify features and temporal contexts into different domains based on progressively extracted domain-specific representations. Prompt-driven object mining locates valuable objects with text prompts and builds their smooth trajectories in the time series. (Image frames are from GOT-10k [9] and NAT2021-train [8].)

Moreover, to obtain high-quality training samples, a prompt-driven object mining approach is provided to locate objects from unlabelled videos and build smooth trajectories.

A. Temporal Domain Adaptation Training

The temporal domain adaptation training framework (TDA) consists of a temporal tracker and a domain discriminator. Generally, a temporal tracker includes a temporal generator and tracker head. The temporal feature generation can be described as:

$$\mathbf{F}_t = \varphi(I_t, \mathbf{M}), \quad (1)$$

where \mathbf{F}_t denotes the temporal features generated from current frame I_t and temporal contexts $\mathbf{M} \in \mathbb{R}^{(t-1) \times C}$ by feature extraction and temporal modelling function φ .

For the huge discrepancy between daytime and nighttime domains, the temporal generator trained on well-illuminated datasets can hardly extract satisfactory temporal representations from nighttime images captured by UAVs. Previous DA methods [8], [10] narrow feature distribution discrepancy of \mathbf{F}_t . However, inappropriate temporal contexts are likely to misguide temporal generators into generating unsatisfactory representations of upcoming frames, resulting in tracking failure. Consequently, the proposed framework aims to take full advantage of robust temporal contexts by aligning their distribution. The loss functions of the discriminator and

tracker, denoted as E_D and E_G , are formulated as follows:

$$E_D = \mathcal{L}_D(\theta_1(\mathbf{F}_t), l_t) + \sum_{i=1}^{t-1} \mathcal{L}_D(\theta_2(\mathbf{M}_i), l_t), \quad (2)$$

$$E_G = \mathcal{L}_{gt} + \mathcal{L}_G(\theta_1(\mathbf{F}_t), l_f) + \sum_{i=1}^{t-1} \mathcal{L}_G(\theta_2(\mathbf{M}_i), l_f),$$

where θ_1, θ_2 denote the discriminators for temporal features and temporal contexts, and l_t, l_f represent the true or false bool label for whether the domain classification is correct. $\mathcal{L}_D, \mathcal{L}_G$, and \mathcal{L}_{gt} respectively denote the loss for domain discrimination, feature alignment, as well as classification and regression.

The adversarial training process is formulated as:

$$\begin{aligned} (\hat{\theta}_1, \hat{\theta}_2) &= \min_{\theta_1, \theta_2} E_D(\hat{\varphi}, \theta_1, \theta_2), \\ (\hat{\varphi}, \hat{\psi}) &= \min_{\varphi, \psi} E_G(\varphi, \psi, \hat{\theta}_1, \hat{\theta}_2), \end{aligned} \quad (3)$$

where ψ denotes the tracker head. $\hat{\varphi}, \hat{\psi}$ represent the temporal generator and tracker head with learned parameters, while $\hat{\theta}_1, \hat{\theta}_2$ denote the learned discriminators. In cases where \mathbf{F}_t and \mathbf{M} are mapped into shared space by the same temporal generator, only one discriminator is necessary.

With contradictory training objectives, the temporal tracker and the discriminator gradually reach convergence. The reduction of discrepancy in both temporal contexts and image features guarantees robust representations vital for long-term nighttime UAV tracking.

Remark 1: While previous DA frameworks overlook the

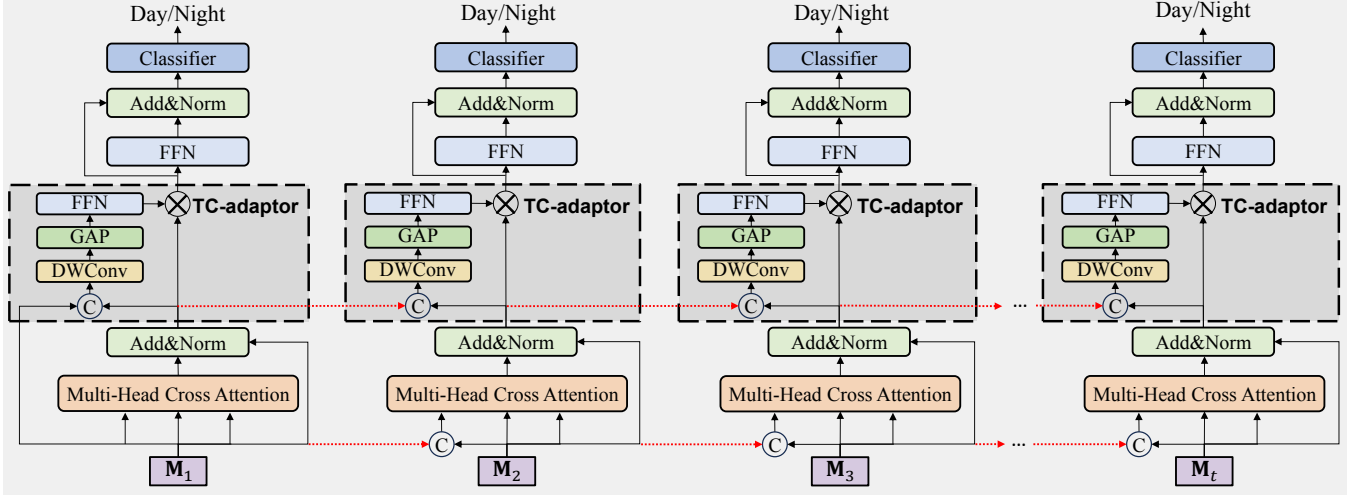


Fig. 3. The structure of the temporal-consistent discriminator. M_i denotes the temporal contexts encoded from the first i frames. The utilization of temporal contexts is marked with red dotted lines. Better representations oriented for daytime or nighttime attributes are progressively extracted, which enables more robust discrimination.

discrepancy of temporal contexts, the proposed novel framework narrows the distribution gap of temporal contexts from different domains. Aligned representations for both image features and temporal contexts enable TDA-Track to perform robustly in poorly illuminated nighttime scenarios.

B. Temporal-Consistent Discriminator

The precision and robustness of the discriminator substantially contribute to the effectiveness of adversarial training. However, discrimination is interfered by inaccurate features extracted from a single frame with noises, resulting in incoherent discrimination results in the time series. To alleviate the problem, shared features among consecutive frames are more reliable grounds for domain discrimination, since they tend to be more robust and representative for domain-specific distributions. Hence, a temporal-consistent discriminator is designed for the proposed TDA framework, as presented in Fig. 3. It progressively extracts and refines common features from temporal contexts and image features with the cross attention mechanism and a novel adaptor, to boost the accuracy and coherence of domain discrimination.

Specifically, multi-head cross attention first extracts the common features from consecutive temporal contexts. Given temporal contexts M_{i-1} and M_i , they are concatenated as input for query projection. The cross attention mechanism can be formulated as:

$$M'_i = \text{Norm}(M_i + \text{Attn}((M_{i-1}, M_i), M_i, M_i)), \quad (4)$$

where Norm represents the layer normalization, while Attn denotes the multi-head cross attention. M'_i denotes the extracted common features.

It's noted that when encountered with severe nighttime UAV tracking challenges, the features change greatly across the time series, resulting in the inaccuracy of the common features extracted. Hence, we propose a temporal-consistent adaptor (TC-adaptor) to refine the common features by concentrating on channels insensitive to context changes.

This work concatenates (Concat) the features of two consecutive image frames, and utilizes a depthwise separable convolution layer (DWConv) to discover the differences. The extent of feature difference is obtained by global average pooling (GAP) to compress the size of features, *i.e.*, $D_i = \text{GAP}(\text{DWConv}(\text{Concat}(M'_{i-1}, M'_i)))$. The feed-forward network (FFN) and multiplication follow to enhance the feature robustness:

$$M''_i = M'_i * \text{FFN}(D_i), \quad (5)$$

where M''_i is the refined common representations.

Then, a linear classifier tells which domain the extracted common features belong to.

Remark 2: The temporal-consistent discriminator progressively encodes domain-specific representation and denoises irrelevant information, which helps narrow the domain discrepancy of temporal contexts and image features.

C. Prompt-Driven Object Mining

To improve the coherence of cropped training patches for better temporal tracking performance, this work develops a two-stage strategy to alleviate the problem: 1) utilize multi-modal object detection to find object candidates specified by text prompt in each frame. 2) build smooth trajectories with tracking by detection method.

Given a $(Video, Text)$ pair, text features and visual features are extracted by a text encoder and an image encoder, respectively:

$$(\mathbf{E}_i^v, \mathbf{E}^t) = (\chi(\text{Image}_i), \phi(\text{Text})), \quad (6)$$

where \mathbf{E}_i^v denotes the visual features of the i_{th} frame Image_i extracted by the image encoder χ , while \mathbf{E}^t denotes the text features of the given Text extracted by the text encoder ϕ .

Then text features and visual features are fused within a deep network to strengthen their correlation in vision-text space. A predictor follows to output pairs of object boxes

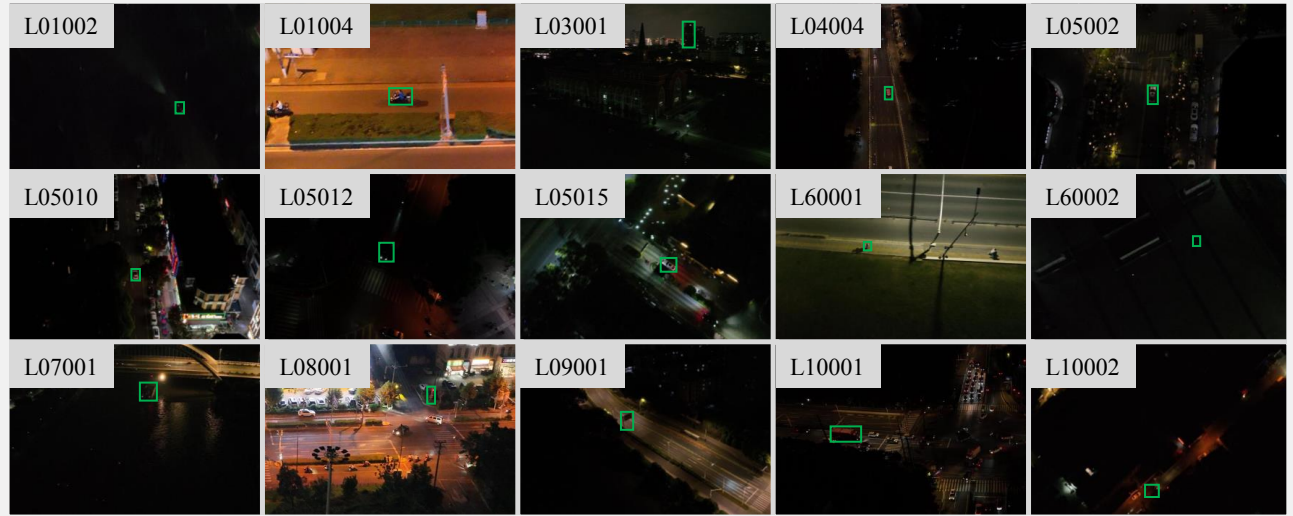


Fig. 4. The first frames of typical scenes in NAT2024-1. The tracking objects are marked with green boxes. The dark environments pose a great challenge to nighttime UAV tracking.

and category names. The object boxes in the time series are associated into smooth trajectories $Trajs$. The process can be formulated as:

$$Trajs = Associate(\vartheta(\mathbf{E}^v, \mathbf{E}^t)), \quad (7)$$

where ϑ represents the deep vision-text fusion and prediction [16], while $Associate(\cdot)$ denotes the association function [27].

High-quality training patches are cropped based on $Trajs$:

$$\begin{aligned} [Z_1^t, \dots, Z_n^t] &= ZCrop(I_t; [B_1^t, \dots, B_n^t]), \\ [X_1^t, \dots, X_n^t] &= XCrop(I_t; [B_1^t, \dots, B_n^t]), \end{aligned} \quad (8)$$

where B_i^t denotes the predicted bounding box of i_{th} trajectory in t_{th} frame I_t . $ZCrop$, $XCrop$ denote cropping images into different sizes with object bounding boxes in the patch center. Z_i^t and X_i^t are respectively the cropped search patch and template patch corresponding to B_i^t .

Remark 3: The prompt-driven object mining approach generalizes well in the nighttime scenarios and builds smooth trajectories from temporal information, which generates high-continuity samples for the temporal domain adaptive training of nighttime UAV trackers.

IV. NOVEL BENCHMARK: NAT2024-1

A large-scale long-term nighttime UAV tracking benchmark, *i.e.*, NAT2024-1, is constructed to realistically evaluate the tracking performance and application value of tracking models. As illustrated in TABLE I, the benchmark consists of 40 long-term image sequences with over 70K frames in total. The sequences are newly captured in diverse scenes by a DJI Mavic 3 Classic UAV. The benchmark includes various typical UAV tracking targets, *e.g.*, electric bikes, pedestrians, and vehicles. Several typical nighttime UAV tracking cases are shown in Fig. 4.

Five attributes are annotated, *i.e.*, aspect ratio change (ARC), fast motion (FM), illumination variation (IV), low ambient illumination (LAI), and scale variation (SV). It's

worth noting that 35 sequences in NAT2024-1 feature as low ambient intensity, as shown in TABLE I, which are representative scenes in nighttime UAV tracking.

V. EXPERIMENTS

In this section, the detailed implementation is provided. Long-term tracking performance evaluation and illumination-oriented analysis demonstrate that TDA-Track surpasses other lightweight trackers in long-term nighttime UAV tracking. Moreover, an ablation study is introduced to testify the effectiveness of the proposed temporal domain adaptation framework. Finally, real-world tests validate the remarkable performance of TDA-Track in UAV applications.

A. Implementation Details

The proposed TDA training framework is implemented using PyTorch and is trained for 25 epochs on an A100 GPU. In terms of data preparation, the daytime training samples are extracted from ImageNet VID [29] and GOT-10K [9], while the nighttime ones are cropped from NAT2021-*train* [8]. The multi-modal tracking model [27] is used to implement the prompt-driven object mining approach. For faster training, TDA-Track is initialized with pre-trained parameters of its Baseline tracker [13]. Note that the backbone parameters in TDA-Track are frozen for the first 10 epochs. Moreover, the temporal-consistent Transformer discriminator is optimized by Adam Optimizer [30] with a base learning rate of 0.005 and poly learning rate policy.

TABLE I
COMPARISON OF NAT2024-1 WITH LONG-TERM SUBSETS OF EXISTING NIGHTTIME UAV TRACKING BENCHMARKS.

Benchmarks	Long Sequences	Frames	LAI attribute
NAT2021- <i>L-test</i> [8]	23	53.6K	18
UAVDark135 [28]	19	41.7K	15
NAT2024-1	40	70.0K	35

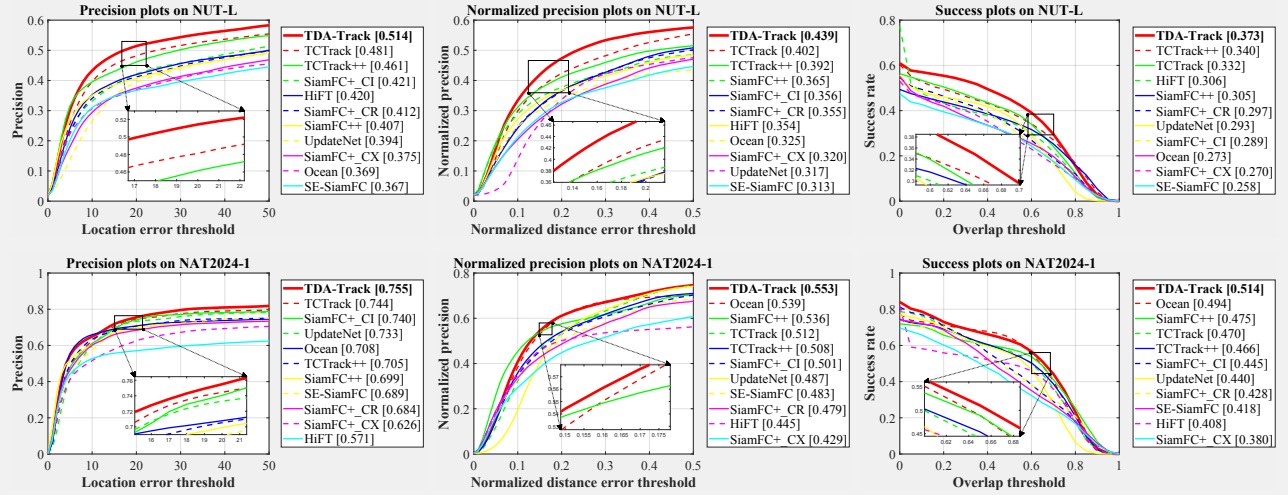


Fig. 5. Long-term nighttime UAV tracking performance of TDA-Track and lightweight trackers on NUT-L and NAT2024-1. TDA-Track ranks first in all three metrics with remarkable improvement.

B. Evaluation Metrics

The one-pass evaluation (OPE) [31] is adopted to rank tracking performances with precision, normalized precision, and success rate. The precision is calculated by the Euclidean distance between the ground truth centers and predicted bounding box centers, *i.e.*, the center location error (CLE). The precision plot demonstrates the percentage of frames where CLE is less than the threshold. The normalized precision plot normalizes precision with ground truth sizes to alleviate the influence of object sizes. The success rate plot computes the percentage of frames where the intersection over union (IoU) of the ground truth and predicted bounding box is greater than a given threshold.

C. Overall Performance

TDA-Track has a hybrid architecture with a lightweight CNN backbone and an efficient Transformer neck. Hence it's compared with lightweight CNN trackers [7], [32]–[35] and efficient CNN-Transformer hybrid trackers [13], [22], [36]. The evaluation is conducted on long-term tracking benchmarks to rate the power of temporal contexts for nighttime UAV tracking. The results are shown in Fig. 5.

NUT-L: NUT-L [10] is a long-term nighttime UAV tracking benchmark collecting the long sequences of NAT2021-*L-test* [8] and UAVDark135 [28]. TDA-Track surpasses the Baseline tracker [13] by a large margin and achieves first rank in precision (**0.514**), normalized precision(**0.439**), and success rate (**0.373**), which makes **6.9%**, **9.2%**, and **12.3%** improvement over Baseline tracker [13] on the three metrics respectively.

NAT2024-1: TDA-Track wins first price in precision (**0.755**), normalized precision(**0.553**), and success rate (**0.514**), which achieves **0.011**, **0.041**, **0.044** gains in three metrics compared with Baseline tracker [13].

As shown in Fig. 6, several confidence maps of TDA-Track and Baseline tracker [13] are visualized using Grad-Cam [37]. While the Baseline tracker has difficulty con-

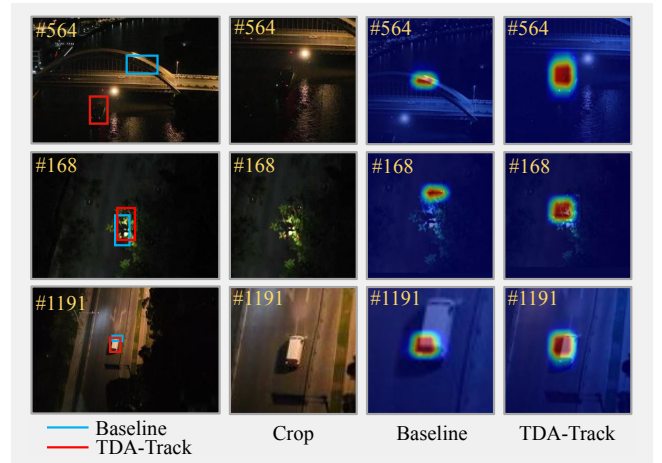


Fig. 6. Visualization of the confidence map and tracking performance on several sequences. The first row showcases that TDA-Track (red) tracks the ship robustly despite the severe low ambient intensity (LAI). More cases prove that TDA-Track achieves better performance than Baseline tracker [13] (blue) at nighttime.

centrating on objects at nighttime, TDA-Track has stronger nighttime perception ability and boosts nighttime tracking performance.

Remark 4: The proposed framework brings about favorable performance improvement on the benchmarks. The results prove the generalization capability and application potential of TDA-Track in various nighttime conditions.

D. Illumination-Oriented Evaluation

The discrepancy between daytime and nighttime domains is mainly caused by different illumination conditions. To investigate the performance of TDA-Track at nighttime, we perform an analysis oriented for the low ambient illumination (LAI) and the illumination variation (IV) challenge on NUT-L. As shown in TABLE II, concerning the LAI attribute, TDA-Track achieves remarkable **13.8%** and **17.6%** improvement in precision and success rate compared to Baseline tracker [13]. Moreover, TDA-Track is far ahead of other

light-weight trackers on IV challenge in precision(**0.467**) and success rate (**0.334**).

E. Ablation Study

In this section, contributions of the proposed methodology are verified on NUT-L, as demonstrated in TABLE III. To prove the superiority of prompt-driven object mining, domain adaptation training on the uncovered samples (Baseline tracker+OM) aligns image features (IF) by learning against the plain discriminator (PD) [8], which achieves **4.7%** and **7.5%** improvement in normalized precision and success rate compared to Baseline tracker [13]. Then to introduce temporal context (TC) into framework, the UAV tracker is trained to learn domain-invariant temporal contexts (Baseline tracker+OM+TC) against PD, but fails to achieve substantial further performance improvement due to incoherent discrimination results of PD. Finally, the proposed TDA framework replaces PD with the temporal-consistent discriminator (TCD) and exploits full potential of temporal contexts. TDA-Track achieves favorable nighttime tracking performance with normalized precision (**0.439**) and success rate (**0.373**).

VI. REAL-WORLD TESTS

Extensive real-world experiments are conducted on a UAV platform equipped with an edge smart camera powered by NVIDIA Jetson Orin NX, as shown in Fig. 7, to evaluate real-world performance. Several challenging long sequences in unfavorably illuminated scenarios are presented. In Test 1, the drone captures the scene from a high altitude, where the truck intermittently drives into the darkness. In Test 2, the drone follows the car in a horizontal flight for a considerable distance. The camera motion and background inference pose great challenges to tracking. In Test 3, the

TABLE II

ILLUMINATION-ORIENTED EVALUATION OF LIGHTWEIGHT TRACKERS. THE TOP 2 PERFORMANCES ARE HIGHLIGHTED IN RED, AND BLUE.

Challenges Trackers	LAI		IV	
	Prec.	Succ.	Prec.	Succ.
SiamFC+_CX [35]	0.347	0.254	0.351	0.239
SiamFC+_CI [35]	0.385	0.269	0.379	0.248
SiamFC+_CR [35]	0.376	0.287	0.362	0.246
UpdateNet [32]	0.396	0.309	0.351	0.262
Ocean [33]	0.313	0.238	0.318	0.237
SiamFC++ [7]	0.365	0.275	0.410	0.307
SE-SiamFC [34]	0.328	0.232	0.325	0.217
HiFT [36]	0.391	0.290	0.389	0.276
TCTrack [13]	0.450	0.319	0.446	0.304
TCTrack++ [22]	0.431	0.313	0.411	0.295
TDA-Track	0.512	0.375	0.467	0.334

TABLE III

ABLATION STUDY OF THE PROPOSED METHODOLOGY ON NUT-L.

Trackers	Align		Disc.		Norm.Prec.	Succ.
	IF	TC	PD	TCD		
Baseline tracker					0.402	0.332
Baseline tracker+OM	✓		✓		0.421	0.357
Baseline tracker+OM+TC	✓	✓	✓		0.419	0.357
TDA-Track	✓	✓		✓	0.439	0.373

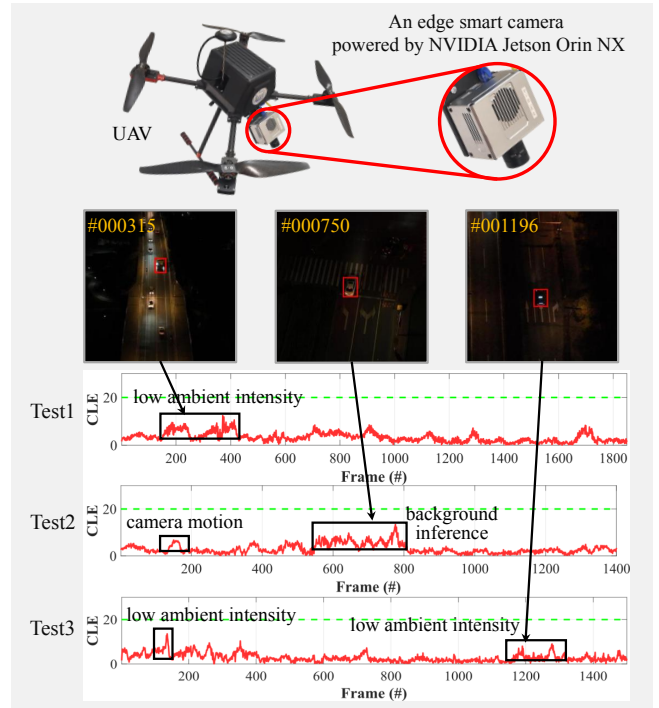


Fig. 7. Real-world tests on UAV platform equipped with NVIDIA Jetson Orin NX prove the robustness and precision of TDA-Track. The red bounding boxes represent the tracking results.

low ambient intensity exacerbates the tracking difficulty. To assess tracking performance, the center location error (CLE) is adopted as the criteria. Despite the challenges of the three sequences, TDA-Track tracks the objects robustly. In addition, TDA-Track is real-time running with 32 frames per second. We are confirmed that TDA-Track is remarkable in both low latency and strong long-term nighttime tracking performance, which is suitable for edge deployment on UAV platforms.

VII. CONCLUSIONS

In this work, a temporal DA training framework for nighttime UAV tracking, namely TDA, is developed. The temporal generator learns to coherently map consecutive frames into domain-invariant temporal contexts and image features by adversarial training. The domain discriminator is designed to extract common representations from encoded features, which enables more accurate domain classification and better domain adaptability. Furthermore, a prompt-driven object mining method obtains high-quality training samples of valuable objects in smooth trajectories. Moreover, a long-term nighttime UAV tracking benchmark, namely, NAT2024-1 is constructed. Evaluation of TDA-Track, the corresponding tracker trained under TDA framework, proves the effectiveness of our framework in boosting nighttime tracking performance. We're convinced that the temporal domain adaptation framework can boost nighttime UAV tracking with temporal context alignment.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No. 62173249) and the Natural Science

REFERENCES

- [1] X. Xiao, J. Dufek, T. Woodbury, and R. Murphy, "UAV Assisted USV Visual Navigation for Marine Mass Casualty Incident Response," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 6105–6110.
- [2] M. Mittal, R. Mohan, W. Burgard, and A. Valada, "Vision-Based Autonomous UAV Navigation and Landing for Urban Search and Rescue," in *Proceedings of the International Symposium of Robotics Research (ISRR)*, 2019, pp. 575–592.
- [3] I. Bozcan and E. Kayacan, "UAV-AdNet: Unsupervised Anomaly Detection using Deep Neural Networks for Aerial Surveillance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 1158–1164.
- [4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshop*, 2016, pp. 850–865.
- [5] 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 (CVPR)*, 2018, pp. 8971–8980.
- [6] 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 (CVPR)*, 2019, pp. 4282–4291.
- [7] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 12 549–12 556.
- [8] 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 (CVPR)*, 2022, pp. 8896–8905.
- [9] 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.
- [10] C. Fu, L. Yao, H. Zuo, G. Zheng, and J. Pan, "SAM-DA: UAV Tracks Anything at Night with SAM-Powered Domain Adaptation," in *Proceedings of the IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*, 2024, pp. 1–8.
- [11] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, "Dynamic Context-Sensitive Filtering Network for Video Salient Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1553–1563.
- [12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.
- [13] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "TC-Track: Temporal Contexts for Aerial Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 798–14 808.
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.
- [15] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded Language-Image Pre-Training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10965–10975.
- [16] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding Dino: Marrying Dino with Grounded Pre-Training for Open-Set Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1–17.
- [17] J. Ye, C. Fu, G. Zheng, Z. Cao, and B. Li, "Darklighter: Light up the Darkness for UAV Tracking," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3079–3085.
- [18] 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.
- [19] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning Spatio-Temporal Transformer for Visual Tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 448–10 457.
- [20] Z. Fu, Q. Liu, Z. Fu, and Y. Wang, "STMTrack: Template-Free Visual Tracking with Space-Time Memory Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 774–13 783.
- [21] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1571–1580.
- [22] Z. Cao, Z. Huang, L. Pan, S. Zhang, and C. Fu, "Towards Real-World Visual Tracking with Temporal Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 834–15 849, 2023.
- [23] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial Cross-Domain Action Recognition with Co-Attention," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 11 815–11 822.
- [24] I. Shin, K. Park, S. Woo, and I. S. Kweon, "Unsupervised Domain Adaptation for Video Semantic Segmentation," *arXiv preprint arXiv:2107.11052*, pp. 1–12, 2021.
- [25] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional Transformer Based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 5546–5554.
- [26] D. Guan, J. Huang, A. Xiao, and S. Lu, "Domain Adaptive Video Segmentation via Temporal Consistency Regularization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8053–8064.
- [27] G. Zheng, S. Lin, H. Zuo, C. Fu, and J. Pan, "NetTrack: Tracking Highly Dynamic Objects with A Net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19 145–19 155.
- [28] 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, 2023.
- [29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the International Conference on Learning Representations (ICML)*, 2015, pp. 1–15.
- [31] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 445–461.
- [32] L. Zhang, A. Gonzalez-Garcia, J. V. D. Weijer, M. Danelljan, and F. S. Khan, "Learning the Model Update for Siamese Trackers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4010–4019.
- [33] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-Aware Anchor-Free Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 771–787.
- [34] I. Sosnovik, A. Moskalev, and A. W. Smeulders, "Scale Equivariance Improves Siamese Tracking," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2765–2774.
- [35] Z. Zhang and H. Peng, "Deeper and Wider Siamese Networks for Real-Time Visual Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4591–4600.
- [36] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical Feature Transformer for Aerial Tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021, pp. 15 457–15 466.
- [37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.