LVPTrack: High Performance Domain Adaptive UAV Tracking with Label Aligned Visual Prompt Tuning

Hongjing Wu[1], Siyuan Yao[2*], Feng Huang[3], Shu Wang[3], Linchao Zhang[4], Zhuoran Zheng[1], Wenqi Ren[1*]

[1]Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
[3]School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China
[4]Artificial Intelligence Institute of China Electronics Technology Group Corporation, Beijing, China
wuhj28@mail2.sysu.edu.cn, yaosiyuan04@gmail.com, huangf@fzu.edu.cn, shu@fzu.edu.cn, hune213@163.com, renwq3@mail.sysu.edu.cn

Abstract

Visual object tracking is essentially crucial for unmanned aerial vehicles (UAVs). Despite the substantial progress, most of the existing UAV trackers are designed for well-conditioned daytime data, while for the scenarios in challenging weather condition, e.g. foggy or nighttime environment, the tremendous domain gap leads to significant performance degradation. To address this issue, in this paper, we propose a novel robust UAV tracker termed LVPTrack, which conducts high quality label-aligned visual prompt tuning to adapt to various challenging weather conditions. Specifically, we first synthesize the sequential foggy and nighttime video frames to assist the model training. A domain adaptive teacher-student network is utilized to distill the hierarchical visual semantic of the target objects in cross-domain scenarios. Then we propose a target-aware pseudo-label voting (PLV) strategy to alleviate the target-level misalignment in the dual domains. Furthermore, we propose a dynamic aggregated prompt (DAP) module to facilitate the appearance variation adaptation of the target object in challenging scenarios. Extensive experiments demonstrate that our tracker achieves superior performance over existing state-of-the-art UAV trackers.

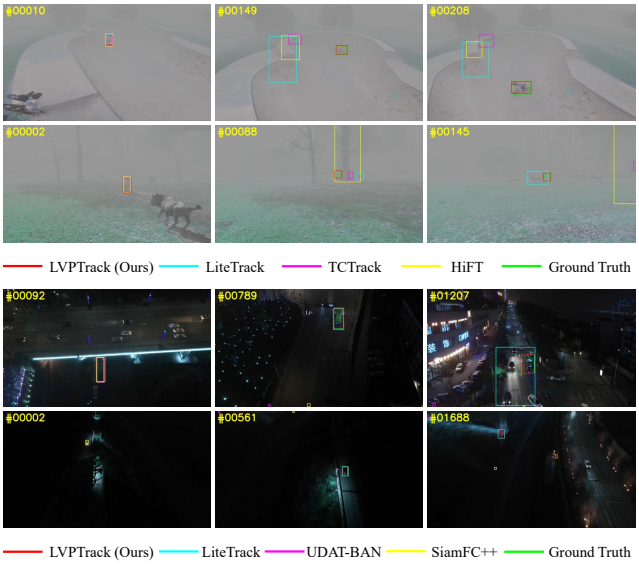


Figure 1: Visualization of tracking results in representative extreme scenarios. The foggy images belong to our synthesized DTB70-Haze dataset. The dark images belong to the NAT2021-test dataset.

Introduction

Visual object tracking is an essential task in various applications of UAVs such as remote sensing, aerial cinematography and GIS. Given the initial annotation of the target template in the first frame, the goal of UAV tracking is to precisely predict the object's location across the entire video frames. Though many trackers have been proposed in recent years, it's still challenging to design a robust UAV tracking system in the real world as the tracking performance is always influenced by the complex environmental factors, e.g. weather condition and background surrounding, etc.

Recently, the template matching based UAV trackers (Bertinetto et al. 2016; Li et al. 2018; Cao et al. 2022; Yao et al. 2021; Tang and Ling 2022) have become the dominant solution due to their well-balanced accuracy and efficiency. These methods generally learn an target-specific matching network on large-scale video dataset, e.g. LaSOT (Fan et al. 2019) or TrackingNet (Müller et al. 2018) to learn the appearance similarity between the paired template-search images. The computed features are then sent into bounding box prediction head to generate response heatmap for target object localization, the maximum peak axis in the heatmap indicates the correct inferred position.

Despite the significant advance, the advanced UAV trackers still suffers from the following limitations. Firstly, existing UAV trackers typically utilize well-conditioned daytime datasets as supervision for model training, while they can hardly adapt to the challenging weather condition, e.g. foggy or nighttime environment, which limits the their deployment in real-world systems. Although some efforts (Ye et al. 2022b,c) have been made to the nighttime environment, their effectiveness has not been well investigated in other weather conditions. Secondly, most of the domain adaptive

*Co-corresponding authors.

UAV trackers produce pseudo labels to maintain the domain transferability on the target data. However, the incorrect label alignment inevitably introduces localization noises in the target domain, leading to significant performance degradation for domain adaptive object tracking.

To address these issues, in this paper, we propose a novel UAV tracker termed **LVPTrack**, which conducts high quality **L**abel-aligned **V**isual **P**rompt **T**uning for domain adaptive UAV tracking. Specifically, we first tackle the daytime video sequences as the source domain, and synthesize the foggy or nighttime video frames as the target domain to assist the model training. The domain shifts of the target's appearance in dual domains are narrowed using a teacher-student network with knowledge transfer. Afterwards, we propose a target-aware pseudo-label voting (PLV) strategy to discard the unreliable tracking results. The background noises in the pseudo-label predicted by the teacher network can be effectively corrected by voting the positions with high confidence, thus the prediction outputs in both domains can be effectively aligned. Finally, to enhance the model's generalizability in challenging scenarios, we propose a dynamic aggregated prompt (DAP) to facilitate the appearance variation adaptation of the target object, yielding the model to be more robust in extremely challenging scenarios. As shown in Fig. 1, our proposed LVPTrack outperforms other state-of-the-art trackers in foggy and nighttime scenarios.

In summary, the main contributions of this work can be concluded in three aspects:

- We design a novel domain adaptive UAV tracker termed LVPTrack, which conducts transferable pseudo label aligned visual prompt tuning for target state inference.
- We propose a target-aware pseudo-label voting strategy to discard the unreliable tracking results, which is capable to alleviate the target-level misalignment in the dual domains.
- We propose a dynamic aggregated prompt (DAP) module to facilitate the significant appearance variation of the target object in challenging scenarios. LVPTrack achieves superior performance to existing state-of-the-art methods at a speed of 110 FPS in our extensive experiments.

Related Work

UAV Tracking Framework. UAV-based tracking models prioritize efficiency due to hardware limitations. Current high-efficiency trackers fall into three categories: DCF-based, Siamese network-based, and lightweight Transformer-based methods. DCF-based methods (Li et al. 2022a; Huang et al. 2019; Danelljan et al. 2015; Li et al. 2020) are known for their computational efficiency but suffer from poor robustness due to their reliance on handcrafted features. Siamese network-based trackers (Fu et al. 2021; Cao et al. 2021b; Li et al. 2018; Xing et al. 2022) have gained popularity in UAV tracking for their high performance and speed, using CNNs to extract and compare features. Consequently, lightweight Transformers have been developed. (Kang et al. 2023) utilizes a Bridge Module to introduce lightweight classification network to the tracking framework successfully and (Ye et al. 2022a) improve inference efficiency with an early elimination module.

Cross-Domain Adaptation. To meet the tracking requirements of UAVs under various weather conditions, domain adaptation gains widespread attention. (Chen et al. 2018) uses adversarial training to adapt the object detector from the source domain to the target domain. (Kim et al. 2019) enhances the cross-domain adaptability of object detection models by generating diversified pseudo-target domain samples and matching feature distributions. In addition to feature matching, self-training methods have also shown promising results. (Yu et al. 2022; Li et al. 2022b) introduce cross-domain adaptive teacher models, leveraging pseudo-labels from the target domain and consistency losses to improve the cross-domain adaptability of object detection models. (Zheng et al. 2021a) introduces Group-aware Label Transfer to refine pseudo-labels for unsupervised domain adaptive person re-identification, improving target domain performance. (He et al. 2022) uses a target-aware dual-branch distillation mechanism to help the source domain model better learn target domain features. (Munir et al. 2021; Ye et al. 2022d) combines self-training with adversarial learning to synergistically optimize the cross-domain adaptability of models. However, to the best of our knowledge, no UAV trackers have been designed for both foggy and nighttime conditions. Therefore, our research fills this gap effectively.

Prompt learning. In NLP research, prompt learning has been proposed to facilitate the rapid application of pre-trained models to various downstream tasks. VPT (Jia et al. 2022) is the first to extend the concept of prompt learning from the language domain to the visual domain, demonstrating its potential in tuning visual models. (Kong et al. 2024) uses prompts to generate natural adversarial patches. (Zhou et al. 2022) introduces conditional prompts, enabling vision-language models to dynamically adjust prompts based on input context, leading to more accurate understanding and generation in diverse tasks. (Yang et al. 2022) integrates cross-modal prompts to guide tracking models in accurately identifying and tracking targets in complex scenes. ViPT (Zhu et al. 2023) encodes auxiliary modality information as visual prompts for tuning RGB modality tracking models, achieving significant results. (Ge et al. 2023) is the first approach to apply prompt learning in domain adaptation tasks, introducing adjustable prompts in cross-domain tasks.

Method

In this section, we present the overall architecture of the proposed LVPTrack in Fig. 2. The LVPTrack consists of three components: a domain adaptive encoder, a pseudo label voting (PLV) module and a dynamic aggregated prompt (DAP) module. LVPTrack firstly utilizes an image generator to synthesize the sequential foggy and nighttime video frames to assist the model training. A domain adaptive encoder following mean teacher training pipeline is employed to narrow the gap between the source and target domains. Then it employs a target-aware pseudo-label voting (PLV) strategy to enforce the predicted consistency. Finally, we design the DAP module to facilitate the appearance variation adaptation of the target object in challenging scenarios.

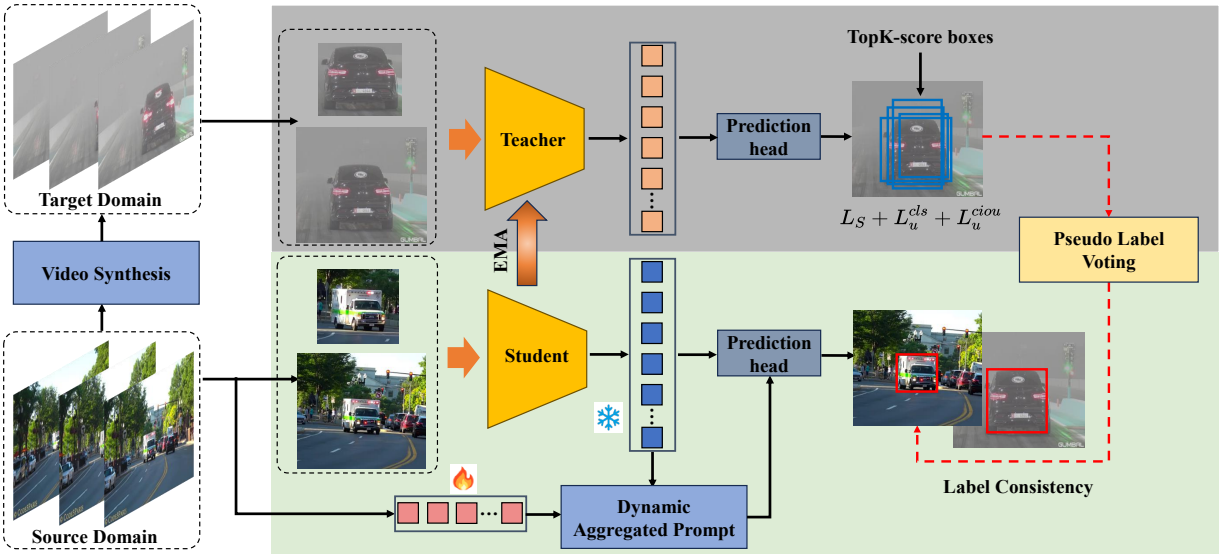


Figure 2: Overview of the proposed LVPTrack. LVPTrack firstly synthesizes the sequential foggy and nighttime video frames to assist the model training, a domain adaptive teacher-student encoder is employed to narrow the gap between the source and target domains. Then it employs a target-aware pseudo-label voting strategy to enforce the predicted consistency. Additionally, a dynamic aggregated prompt (DAP) module is designed to facilitate the appearance variation adaptation of the target object.

Cross-Domain Video Synthesis

To model the target appearance in challenging weather condition, we first synthesize the sequential foggy and nighttime video frames to assist the model training. Specifically, to generate foggy videos, we use an atmospheric scattering model to render haze on a clean image. According to Koschmieder's law (Zheng et al. 2021b), the degradation process of hazy images can be described by the atmospheric scattering model as:

$$H^c(x,y) = J^c(x,y) \cdot t(x,y) + A^c \cdot (1 - t(x,y)), \quad (1)$$

where (x, y) is the coordinate of a pixel within the image, $c \in \{r, g, b\}$ is the color channel, H is the hazy image, J is the corresponding haze-free image, A is the atmospheric light, and t is the medium transmission describing the portion of the light that is not scattered. We employ the transmission map generated by Depth Anything (Yang et al. 2024) to generate the foggy video frames. For the nighttime environment, we employ the unsupervised image-to-image translation framework (Liu, Breuel, and Kautz 2017) based on Coupled GANs to obtain the nighttime videos.

Domain Adaptive Encoder

After obtaining the synthesized video sequences, we design a domain adaptive encoder with teacher-student architecture for model training. Given N_s frames $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ in source domain and N_t unlabeled frames $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ in target domain, where $\mathbf{x}$ denotes the images and $\mathbf{y}$ represents the annotated bounding box coordinates. The goal of LVPTrack is to leverage $\mathcal{D}_s$ and $\mathcal{D}_t$ as training data to learn a domain-invariant tracker. To achieve this, we utilize a domain adaptive encoder following mean teacher pipeline. Formally, the domain adaptive encoder consists of a teacher network and a student network with the same ViT (Dosovitskiy et al. 2021) backbone. The paired template-search images of daytime in source domain are sent into the student network, while the images generated in foggy or nighttime environment are regarded as input to the teacher network. The teacher's knowledge is transferred by updating the weights of the student model in the target domain using the EMA (Exponential Moving Average) as follows:

$$\theta^{\mathcal{T}} \leftarrow \alpha\theta^{\mathcal{T}} + (1-\alpha)\theta^{\mathcal{S}}, \quad (2)$$

where $\theta^{\mathcal{T}}$ and $\theta^{\mathcal{S}}$ denote the parameters of the teacher and student networks, respectively. α is a momentum factor controlling the updating speed of the teacher.

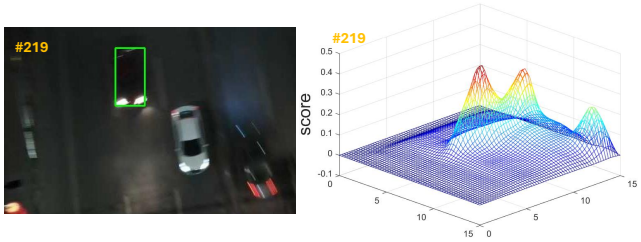
To train the tracking model, the synthesized template-search pairs in target domain are sent into the teacher network to generate pseudo labels, which would be fed back into the student network for model updating. We adopt a supervised loss $\mathcal{L}_s$ and an unsupervised loss $\mathcal{L}_u$ to conduct the knowledge transfer, which can be given by:

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_u, \quad (3)$$

where λ is the hyperparameter to balance the weights of $\mathcal{L}_s$ and $\mathcal{L}_u$. We adopt the supervised loss $\mathcal{L}_s$ following the approach of OSTrack(Ye et al. 2022a). Similarly, the unsupervised loss $\mathcal{L}_u$ is defined using pseudo labels predicted in the target domain. By jointly utilizing ground truth labels and pseudo labels during model training, we align feature representations across the dual domains, which effectively mitigate the semantic domain shift.

Pseudo Label Voting

For object tracking task, the noisy pseudo labels will mislead the target state prediction. To address this issue, we propose



(a) Ground Truth (b) Score map

Figure 3: Illustration of PLV. We show a challenging case with multiple similar targets: (a) the cropped search region with ground truth in green; (b) the corresponding classification score map.

a Pseudo Label Voting (PLV) module to enhance the localization quality. Concretely, we utilize the prediction head (Ye et al. 2022a) consists of three convolutional branches, *i.e.* center classification scores, center point offsets and size to obtain the pseudo labels of tracking object in target domain. As shown in Fig. 3, the peak value and the fluctuation of the predicted classification score map reveal the quality of the tracking results in target domain to some extent. If the predicted bounding box accurately matches to the target object, the ideal score map should have sharp peak. Otherwise, the response fluctuations would be intense, indicating that the pseudo label generated by the teacher network may be inaccurate. Based on this observation, we propose a pseudo label voting module to discard the unreliable tracking results. We first constructed a set, $\mathbf{S}_{topk}$, comprising points selected from the teacher network's results based on the top-k classification scores. To enhance the localization capability, we propose a soft voting mechanism to measure the reliability of the pseudo label, which is given by:

$$\mathbf{M}_i = \frac{\mathbf{D}_{min} \cdot \exp(-\left\|\mathbf{D}_{max}/\sigma_i\right\|^2/2)}{\sum_{i=1}^{K} \mathbf{D}_{min} \cdot \exp(-\left\|\mathbf{D}_{max}/\sigma_i\right\|^2/2)}, \quad (4)$$

where $\mathbf{C}_{max}$, $\mathbf{C}_{min}$ and $\mathbf{C}_i$ denote the maximum, minimum and the i-th peak of the top-k classification set $\mathbf{S}_{topk}$, σ_i is the standard deviation of the i-th peak using K-Means clustering. The distance $\mathbf{D}_{min} = |\mathbf{C}_i - \mathbf{C}_{min}|$ and $\mathbf{D}_{max} = |\mathbf{C}_i - \mathbf{C}_{max}|$. The output measure $\mathbf{M}_i$ indicates the fluctuated degree of response maps and the confidence level of the tracking targets. For sharper peaks and fewer noise, $\mathbf{M}_i$ will become larger and the response map will become smooth except for only one sharp peak. Otherwise, $\mathbf{M}_i$ will drops significantly if the predicted pseudo label is noisy. Finally, we utilize a voting mechanism to adjust the position of the pseudo label as:

$$B^* = \frac{1}{\sum_{i=1}^{K} M_i} \sum_{i=1}^{K} B_i M_i, \quad B = \{x_1, y_1, x_2, y_2\}. \quad (5)$$

Dynamic Aggregated Prompt Learning

The Domain Adaptive Encoder and Pseudo Label Voting can effectively enhance the model's generalizability. However, the dramatic appearance variations still hamper the tracking robustness. To mitigate this issue, we propose a Dynamic Aggregated Prompt (DAP) module to facilitate the appearance variation adaptation of the target object. Suppose the template and search features in the l-th Transformer stage are $\mathbf{Z}_l$ and $\mathbf{X}_l$, respectively. we first utilize the adjacent search features to fuse the hierarchical visual semantics as:

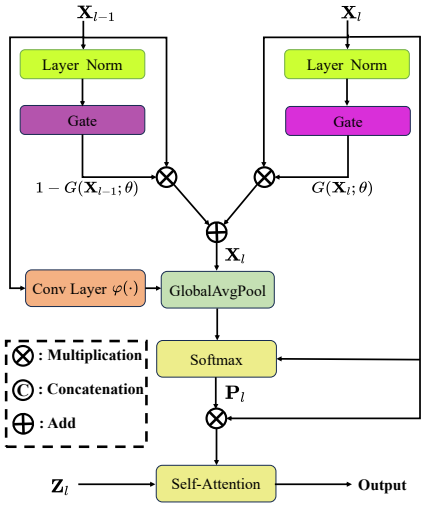


Figure 4: Details of the Dynamic Aggregated Prompt (DAP) module.

$$\mathbf{X}_l = G(\mathbf{X}_l;\theta) \odot \mathbf{X}_l + (1 - G(\mathbf{X}_{l-1};\theta)) \odot \mathbf{X}_{l-1}, \quad (6)$$

where $\mathbf{X}_{l-1}$ denotes the search features in previous Transformer stage. $G(\cdot)$ is a Sigmoid gated function with bilinear upsampling and θ is the learnable parameter. To further enhance the tracker's adaptability to various appearance variations, we introduce a dynamic prompt for search feature aggregation, formulated as $\mathbf{X}_l = \mathbf{X}_l \cdot \mathbf{P}_l$, where $\mathbf{P}_l$ represents the learned dynamic prompt at the l-th stage. Intuitively, this dynamic prompt for target appearance matching is designed to effectively handle unpredictable background distractors. Therefore, we generate the prompt $\mathbf{P}_l$ by passing the input search features $\mathbf{X}_l$ through weighted attention with global average pooling as follows:

$$\mathbf{P}_l = \frac{\exp(\varphi(\mathbf{X}_{l-1}^i;\hat{\theta}) \odot \mathbf{X}_l^i)}{\sum_{i=1} \exp(\varphi(\mathbf{X}_{l-1}^i;\hat{\theta}) \odot \mathbf{X}_l^i))}, \quad (7)$$

where $\mathbf{X}_l^i$ is the i-th element in $\mathbf{X}_l$, $\varphi(\cdot)$ denotes a convolutional layer with global average pooling operator, $\hat{\theta}$ is the trainable parameter. Furthermore, the template feature $\mathbf{Z}_l$ and search feature $\mathbf{X}_l$ are split into patches, which are flattened into 1-dim vector and projected to produce the key-value tokens. Similar to (Ye et al. 2022a), the key-value tokens of the paired template-search features are concatenated for relation modeling, which can be given by:

$$\text{Attn}_{xz} = \text{softmax}(\frac{\mathbf{Q}_x \mathbf{K}_{xz}^\top}{\sqrt{d_k}})\mathbf{V}_{xz}, \quad (8)$$

where $\mathbf{Q}_x$ is the query matrix generated by $\mathbf{X}_l$. $\mathbf{K}_{xz}$ and $\mathbf{V}_{xz}$ denote the concatenated key and value tokens generated

Method	UAV123-Haze		DTB70-Haze		UAV123-Dark		DTB70-Dark	
	AUC	P	AUC	P	AUC	P	AUC	P
LVPTrack(ours)	**61.01**	**81.12**	**63.51**	**83.33**	**62.03**	**81.92**	**64.29**	**82.36**
SimTrack-B/16 (Chen et al. 2022)	<u>60.98</u>	78.83	57.03	73.02	<u>60.70</u>	<u>79.64</u>	55.93	71.50
SeqTrack-B256 (Chen et al. 2023)	60.95	<u>80.56</u>	56.24	74.36	60.30	78.86	58.82	76.64
OSTrack-256 (Ye et al. 2022a)	58.56	75.99	<u>57.07</u>	74.63	58.92	77.01	<u>60.70</u>	<u>77.80</u>
LiteTrack-B8 (Wei et al. 2023)	57.09	74.70	56.80	74.39	59.26	77.55	58.40	74.30
DiMP (Bhat et al. 2019)	57.70	77.00	53.80	69.50	58.60	78.20	55.20	72.30
ATOM (Danelljan et al. 2019)	56.80	76.70	50.20	66.00	56.70	77.90	51.50	69.10
AVTrack-ViT (Li et al. 2024)	55.49	72.27	52.35	68.09	57.63	76.16	56.66	72.21
SiamRPN++ (Li et al. 2019)	54.40	73.00	55.80	<u>74.70</u>	-	-	48.80	70.30
SMAT (Gopal and Amer 2024)	53.19	68.66	53.02	68.25	57.51	74.57	55.43	70.66
SiamRPN (Li et al. 2018)	50.50	69.90	47.40	67.40	50.90	70.10	-	-
SiamTPN (Xing et al. 2022)	47.21	62.40	45.63	58.86	54.20	72.10	47.70	60.30
TCTrack (Cao et al. 2022)	45.70	64.00	50.20	68.80	49.00	69.60	48.20	66.70
HiFT (Cao et al. 2021a)	45.70	62.60	45.40	64.50	48.60	68.00	48.00	66.20
SGDViT (Yao et al. 2023)	44.20	60.50	45.40	64.50	47.80	64.30	48.80	65.20
SiamAPN++ (Cao et al. 2021b)	41.60	61.30	40.80	61.30	49.10	69.20	45.50	65.10
SiamFC++ (Xu et al. 2020)	40.70	56.10	-	-	52.90	71.70	-	-
SiamAPN (Fu et al. 2021)	29.80	47.90	34.10	53.20	45.00	66.10	40.50	63.40
UDAT-BAN (Ye et al. 2022d)	-	-	-	-	54.20	74.40	56.30	75.60
UDAT-CAR (Ye et al. 2022d)	-	-	-	-	53.60	73.30	57.20	75.80

Table 1: Comparison with state-of-the-art visual trackers on UAV123-Haze, DTB70-Haze, UAV123-Dark, DTB70-Dark. The best two results are shown in bold and underline.

by the search and template features. By performing cross-attention of the template-search tokens in Eq. 8, the DAP is capable to capture the information of discriminative target object even with drastic appearance variations.

Experiments

Our method is implemented using python3.8 and pytorch1.11.0. Our tracker is trained with 4 NVIDIA RTX4090 GPUs. All of the inference speed testing are conducted on a single NVIDIA RTX4090 GPU.

Datasets

Synthetic Haze and Night Datasets. The UAV123-Haze, DTB70-Haze and GOT-10k-Haze are synthesized from the original UAV123 (Mueller, Smith, and Ghanem 2016), DTB70 (Li and Yeung 2017) and GOT-10k (Huang, Zhao, and Huang 2019), respectively.

NAT2021-test. NAT2021-test (Ye et al. 2022d) is a challenging dataset for nighttime aerial tracking, containing 180 sequences with over 140,000 frames. All frames are manually annotated with enhanced accuracy using a low-light enhancement approach.

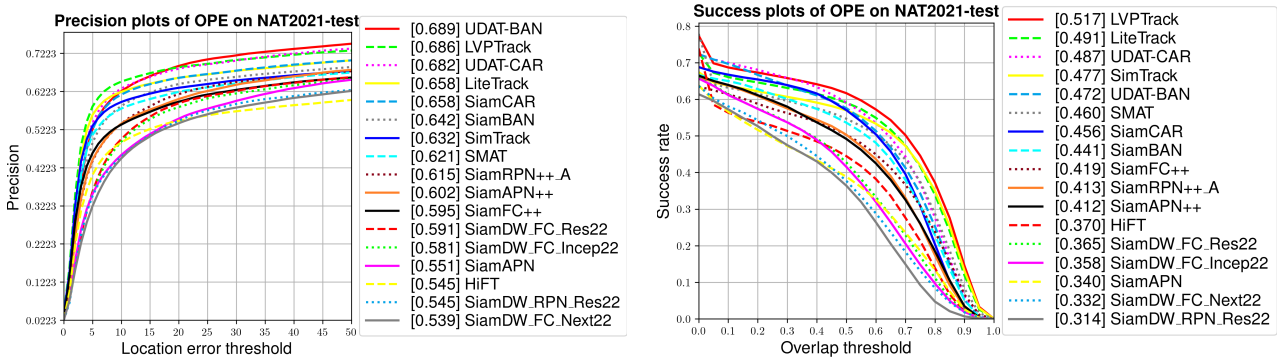
UAVDark70. UAVDark70 (Li et al. 2021) is a dataset with 70 annotated video sequences, designed for testing UAV object tracking in low-light conditions. It contains diverse environments and lighting challenges, providing a robust benchmark for developing and evaluating tracking algorithms in reduced visibility scenarios.

Implementation Details

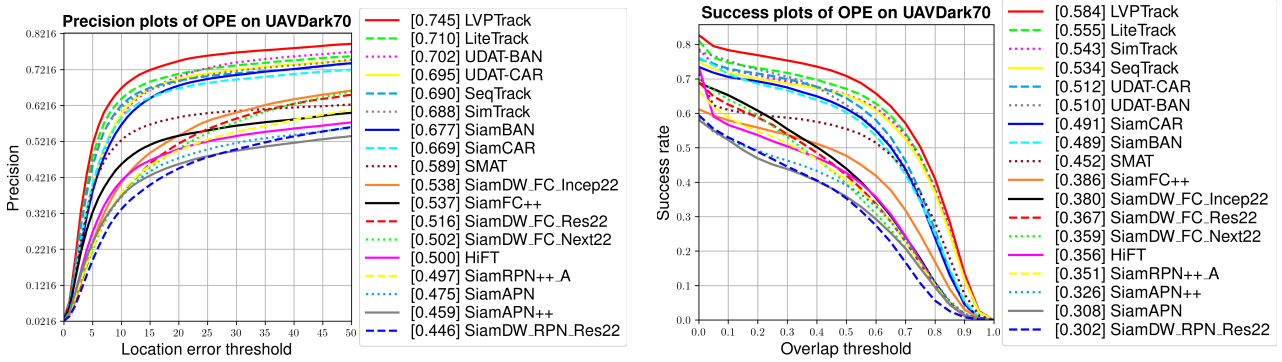
Model Details. We adopt vanilla ViT-Base (Dosovitskiy et al. 2021) model as the backbone of our tracker. Specifically, we take the first 4 layers and the last 4 layers of the ViT-Base (Dosovitskiy et al. 2021) model for feature extraction and relation modeling, respectively. The patch size is set to 16×16. The prediction head is a lightweight FCN consisting 4 stacked Conv-BN-ReLU layers for each output which is same to OSTrack (Ye et al. 2022a). The sizes of the template and search region are resized to 128×128 and 256×256 respectively, corresponding to 2^2 and 4^2 times of the target box area.

Training Details. Our training process is divided into two stages: backbone training and prompt finetune. Firstly, we use the weights of LiteTrack-B8 (Wei et al. 2023) as the initialization weights for LVPTrack, which are obtained from training on the LaSOT (Fan et al. 2019), TrackingNet (Muller et al. 2018), COCO (Lin et al. 2014), and GOT-10k (Huang, Zhao, and Huang 2019) datasets. Then, LVPTrack infers the training set of GOT-10k-Extreme to generate initial pseudo labels. For backbone training, the prompt module is not introduced. We apply weak augmentation which is the same as the test phase to the synthesized GOT-10k-Extreme dataset and use the teacher branch for inference to generate new pseudo labels. Specifically, we randomly update 100,000 samples in the GOT-10k-Extreme dataset every 5 epochs. Four source domain datasets, including LaSOT, TrackingNet, COCO, and GOT-10k, as well as a synthetic dataset, GOT-10k-Extreme, are used for training the student model. The sampling ratio of the datasets is set to 1:1:1:1:2. Backbone training epoch is set to 250. For prompt finetune, we freeze the parameters of the ViT Encoder and prediction head, and train the prompt module parameters for additional 50 epochs. We switch the EMA operation to false, and fully synchronize the parameters of the teacher and student. All the other settings remain consistent with the backbone training.

Inference. To accelerate the inference, the template feature



(a) Precision and success plots on NAT2021-test



(b) Precision and success plots on UAVDark70

Figure 5: Overall performance of SOTA trackers and LVPTrack on nighttime aerial tracking benchmarks. The results indicate that the proposed LVPTrack achieve leading performance.

is initialized using the first frame of each video sequence and stored for relation modeling between the template and search region in subsequent frames, similar to the approach used in LiteTrack (Wei et al. 2023). We use the Hanning window penalty to utilize positional prior in tracking, following standard procedures in (Ye et al. 2022a).

State-of-the-art Comparisons

In this subsection, to demonstrate the effectiveness and high efficiency of our method, we comprehensively compare LVPTrack with SOTA trackers on aerial tracking datasets in extreme scenarios. We evaluated our tracker on foggy benchmarks (UAV123-Haze, DTB70-Haze) and nighttime benchmarks (NAT2021-test (Ye et al. 2022d), UAVDark70 (Li et al. 2021), UAV123-Dark, DTB70-Dark). Compared to Table 1, the additional trackers included for comparison in Fig. 5 are SiamDW (Zhang and Peng 2019), SiamBAN (Chen et al. 2020) and SiamCAR (Guo et al. 2020).

UAV123-Haze. As shown in Table 1, our proposed LVPTrack outperforms other trackers in terms of precision and success rate. Our precision is 5.13% higher than OSTrack-256 (Ye et al. 2022a), which runs at a much lower speed than ours. Compared to LiteTrack-B8 (Wei et al. 2023), which has a similar speed to ours, our AUC and precision are higher by 3.92% and 6.42%, respectively.

DTB70-Haze. LVPTrack performs exceptionally well on the DTB70-Haze dataset, leading the second-best tracker by a significant margin of 6.44% in AUC and 8.7% in precision. Compared to the best-performing UAV tracker, TCTrack (Cao et al. 2022), we achieve over 10% improvement in both AUC and precision at a faster speed as presented in Table 1.

UAV123-Dark. Despite our synthesized UAV123-Dark dataset is challenging, LVPTrack still delivers impressive performance. In terms of AUC, we outperform the second-ranked tracker by 2.77%, and for precision, we see a 4.37% improvement, as shown in Table 1.

DTB70-Dark. LVPTrack achieved the best AUC (64.29) and precision (82.36) on DTB70-Dark, followed by OSTrack with an AUC of 60.7 and precision of 77.8.

NAT2021. On the large-scale night dataset NAT2021 (Ye et al. 2022d), LVPTrack achieved the best AUC as shown in Figure 5. Specifically, in terms of AUC, we outperformed the second tracker LiteTrack by 2.6% and the third tracker UDAT-CAR by 3.0%. This partially proves that our proposed framework helps the model learn effectively from synthetic extreme domain datasets. Moreover, it also demonstrates the effectiveness of our synthetic dataset method.

UAVDark70. As shown in Fig. 5, LVPTrack also outperforms all the other trackers, surpassing the second-best tracker by 2.9 in AUC and by 3.5 in precision on another real-world dataset, UAVDark70.

Ablation Study and Visualization

To validate the effectiveness of our framework and the reasonableness of the parameter settings, we conduct several

Trackers	Precision (%)	Δ_{pre}	AUC (%)	Δ_{AUC}
Baseline-Source	74.70	–	57.09	–
Baseline-Haze	78.03	+3.33	59.03	+1.94
PLV	80.83	+6.13	60.41	+3.32
PLV+Prompt (LVPTrack)	81.12	+6.42	61.01	+3.92
Baseline-Haze	78.03	–	59.03	–
Top K	79.40	+1.37	58.25	-0.78
Top K-nearest	81.12	+3.09	61.01	+1.98

Table 2: Ablation study of the components of LVPTrack. Δ denotes the improvement compared with the Baseline tracker.

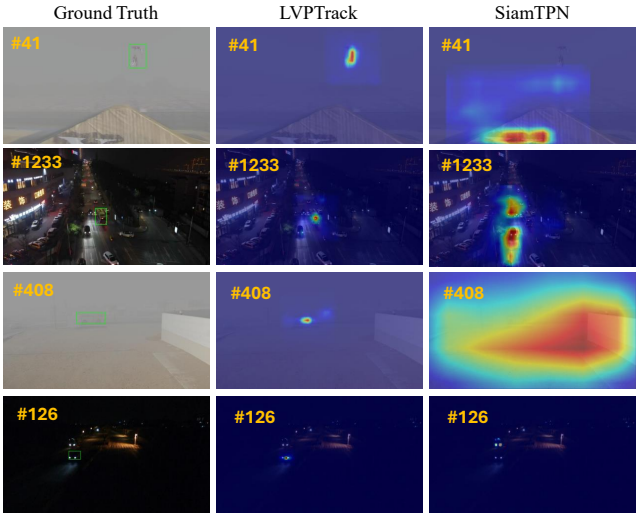


Figure 6: Visual comparison of scoremap generated by SiamTPN and our proposed LVPTrack. The green box indicate the Ground Truth.

key ablation experiments and hyper-parameter selection experiments, and visualize the results of some crucial experiments.

Study on the components of LVPTrack. We conduct ablation experiments on the proposed training framework and two modules to verify their effectiveness. As shown in Table 2, **Baseline-Source** represents training only on the four source domain datasets, which is consistent with the standard training method for trackers. **Baseline-Haze** refers to the results obtained by adding the synthetic haze domain dataset to the training. When our PLV module is added, denoted as **PLV**, the precision and AUC scores increase by 6.13% and 3.32%, respectively. Compared to **Baseline-Haze**, showing a significant performance improvement. Further incorporating the Prompt module, as represented by **PLV+Prompt (LVPTrack)**, would result in even greater enhancement, with precision and AUC scores increasing by 6.42% and 3.92% compared to **Baseline-Haze**, respectively.

Study on the PLV mechanism. As shown in Table 2, **Baseline-Haze** directly updates pseudo labels using predictions from the teacher model without any filtering or fusion. The **Top K** method selects the top K bounding boxes based on classification scores and averages them, leading to a 1.37% precision gain but a 0.78% AUC drop due to merging boxes from similar targets. The **Top K-nearest** method refines this by filtering out boxes farther from the highest-scoring point, resulting in a 3.09% precision and 1.98% AUC improvement.

EMA Frequency	Precision (%)	AUC (%)
Each epoch	81.12	61.01
Every 5 epochs	80.65	60.21
Each batch	79.56	59.44
Dataset Proportion	**Precision (%)**	**AUC (%)**
1:1:1:1:0.5	79.46	60.37
1:1:1:1:1	80.19	60.40
1:1:1:1:2	81.12	61.01
1:1:1:1:4	80.03	60.33

Table 3: Study on the training hyper-parameter of LVPTrack. The dataset proportions are LaSOT: TrackingNet: COCO: GOT-10k: Synthetic. All the data is presented in percentage (%) format.

Study on the training hyper-parameter of LVPTrack. We conduct an ablation study on the update frequency of EMA and the proportion of the training dataset. As shown in Table 3, we experiment by performing EMA after each epoch, every five epochs, and after completing each batch to transfer parameters to the teacher model. The results indicate that performing EMA after each epoch yields the best results. For the dataset proportion settings, we conducted three groups of experiments as shown in the table, and the results indicate that group 3 achieve the best performance. Therefore, we set the training dataset proportion to 1:1:1:1:2.

Heatmap Visualization. As shown in Fig. 6, we visualize the scoremaps output by LVPTrack and SiamTPN. In extreme scenarios, our tracker can accurately locate the motorcycle in heavy fog (#41). Even when the target is barely visible to the human eye, LVPTrack does not lose track of the target (#408). In a real nighttime scene, faced with numerous similar small targets and distracting lights, the attention of SiamTPN is scattered (#1233) or even shifted (#126), whereas LVPTrack accurately captures the target location, achieving robust tracking in extreme scenarios.

Conclusions

In this work, We develop LVPTrack for extreme scenarios using foggy and nighttime tracking datasets. Our Pseudo Label Voting method improves pseudo-label accuracy, and Dynamic Aggregated Prompt Learning enhances cross-domain knowledge transfer. In the future, we will optimize prompt learning for greater efficiency, aiming to advance lightweight visual tracking in extreme conditions.

Acknowledgments

This work is supported by the National Science Fund for Excellent Young Scholars (No. 62322216, 62402055, U24B20175), the Shenzhen Science and Technology Program Project (No. KQTD20221101093559018) and the Fundamental Research Funds for the Central Universities under Grant (No. 2023RC09).

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *European Conference on Computer Vision Workshop*, 850–865.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning Discriminative Model Prediction for Tracking. In *IEEE International Conference on Computer Vision*, 6181–6190.

Cao, Z.; Fu, C.; Ye, J.; Li, B.; and Li, Y. 2021a. HiFT: Hierarchical Feature Transformer for Aerial Tracking. In *IEEE International Conference on Computer Vision*, 15437–15446.

Cao, Z.; Fu, C.; Ye, J.; Li, B.; and Li, Y. 2021b. SiamAPN++: Siamese Attentional Aggregation Network for Real-Time UAV Tracking. arXiv:2106.08816.

Cao, Z.; Huang, Z.; Pan, L.; Zhang, S.; Liu, Z.; and Fu, C. 2022. TCTrack: Temporal Contexts for Aerial Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 14778–14788.

Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022. Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking. *arXiv preprint arXiv:2203.05328*.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 14572–14581.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3339–3348.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese Box Adaptive Network for Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6668–6677.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. ATOM: Accurate Tracking by Overlap Maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4660–4669.

Danelljan, M.; Hager, G.; Shahbaz Khan, F.; and Felsberg, M. 2015. Learning spatially regularized correlation filters for visual tracking. In *IEEE International Conference on Computer Vision*, 4310–4318.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5374–5383.

Fu, C.; Cao, Z.; Li, Y.; Ye, J.; and Feng, C. 2021. Siamese Anchor Proposal Network for High-Speed Aerial Tracking. In *IEEE International Conference on Robotics and Automation*, 510–516.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Gopal, G. Y.; and Amer, M. A. 2024. Separable self and mixed attention transformers for efficient object tracking. 6708–6717.

Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; and Chen, S. 2020. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6268–6276.

He, M.; Wang, Y.; Wu, J.; Wang, Y.; Li, H.; Li, B.; Gan, W.; Wu, W.; and Qiao, Y. 2022. Cross domain object detection by target-perceived dual branch distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9570–9580.

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 43(5): 1562–1577.

Huang, Z.; Fu, C.; Li, Y.; Lin, F.; and Lu, P. 2019. Learning aberrance repressed correlation filters for real-time UAV tracking. In *IEEE International Conference on Computer Vision*, 2891–2900.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727.

Kang, B.; Chen, X.; Wang, D.; Peng, H.; and Lu, H. 2023. Exploring lightweight hierarchical vision transformers for efficient visual tracking. In *IEEE International Conference on Computer Vision*, 9612–9621.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12456–12465.

Kong, D.; Liang, S.; Zhu, X.; and Zhong, W., Yuansheng adn Ren. 2024. Patch is Enough: Naturalistic Adversarial Patch against Vision-Language Pre-training Models. *Visual Intelligence*.

Li, B.; Fu, C.; Ding, F.; Ye, J.; and Lin, F. 2021. ADTrack: Target-Aware Dual Filter Learning for Real-Time Anti-DarkUAV Tracking. In *IEEE International Conference on Robotics and Automation*, 496–502.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4282–4291.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High Performance Visual Tracking With Siamese Region Proposal Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8971–8980.

Li, S.; Liu, Y.; Zhao, Q.; and Feng, Z. 2022a. Learning residue-aware correlation filters and refining scale for real-time uav tracking. *Pattern Recognition*, 127: 108614.

Li, S.; and Yeung, D.-Y. 2017. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI Conference on Artificial Intelligence*.

Li, Y.; Fu, C.; Ding, F.; Huang, Z.; and Lu, G. 2020. AutoTrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 11923–11932.

Li, Y.; Liu, M.; Wu, Y.; Wang, X.; Yang, X.; and Li, S. 2024. Learning Adaptive and View-Invariant Vision Transformer for Real-Time UAV Tracking. In *International Conference on Machine Learning*.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022b. Cross-domain adaptive teacher for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7581–7590.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Liu, M.; Breuel, T. M.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems*, 700–708.

Mueller, M.; Smith, N.; and Ghanem, B. 2016. A benchmark and simulator for UAV tracking. In *European Conference on Computer Vision*, 445–461.

Müller, M.; Bibi, A.; Giancola, S.; Al-Subaihi, S.; and Ghanem, B. 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *European Conference on Computer Vision*, 310–327.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *European Conference on Computer Vision*, 300–317.

Munir, M. A.; Khan, M. H.; Sarfraz, M.; and Ali, M. 2021. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 34: 22770–22782.

Tang, F.; and Ling, Q. 2022. Ranking-Based Siamese Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8731–8740.

Wei, Q.; Zeng, B.; Liu, J.; He, L.; and Zeng, G. 2023. LiteTrack: Layer Pruning with Asynchronous Feature Extraction for Lightweight and Efficient Visual Tracking. arXiv:2309.09249.

Xing, D.; Evangeliou, N.; Tsoukalas, A.; and Tzes, A. 2022. Siamese Transformer Pyramid Networks for Real-Time UAV Tracking. 1898–1907.

Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; and Yu, G. 2020. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI Conference on Artificial Intelligence*, 12549–12556.

Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Prompting for multi-modal tracking. In *Proceedings of the ACM International Conference on Multimedia*, 3492–3500.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, L.; Fu, C.; Li, S.; Zheng, G.; and Ye, J. 2023. SGDViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. arXiv:2303.04378.

Yao, S.; Han, X.; Zhang, H.; Wang, X.; and Cao, X. 2021. Learning Deep Lucas-Kanade Siamese Network for Visual Tracking. *IEEE Transactions on Image Processing*, 30: 4814–4827.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022a. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. In *European Conference on Computer Vision*.

Ye, J.; Fu, C.; Cao, Z.; An, S.; Zheng, G.; and Li, B. 2022b. Tracker Meets Night: A Transformer Enhancer for UAV Tracking. 7(2): 3866–3873.

Ye, J.; Fu, C.; Zheng, G.; Paudel, D. P.; and Chen, G. 2022c. Unsupervised Domain Adaptation for Nighttime Aerial Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8886–8895.

Ye, J.; Fu, C.; Zheng, G.; Paudel, D. P.; and Chen, G. 2022d. Unsupervised Domain Adaptation for Nighttime Aerial Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–10.

Yu, J.; Liu, J.; Wei, X.; Zhou, H.; Nakata, Y.; Gudovskiy, D.; Okuno, T.; Li, J.; Keutzer, K.; and Zhang, S. 2022. MTTrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, 629–645.

Zhang, Z.; and Peng, H. 2019. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zheng, K.; Liu, W.; He, L.; Mei, T.; Luo, J.; and Zha, Z.-J. 2021a. Group-aware label transfer for domain adaptive person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5310–5319.

Zheng, Z.; Ren, W.; Cao, X.; Hu, X.; Wang, T.; Song, F.; and Jia, X. 2021b. Ultra-High-Definition Image Dehazing via Multi-Guided Bilateral Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16185–16194.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9516–9526.