

# UMDATrack: Unified Multi-Domain Adaptive Tracking Under Adverse Weather Conditions

Siyuan Yao<sup>1,2</sup>, Rui Zhu<sup>2</sup>, Ziqi Wang<sup>2</sup>, Wenqi Ren<sup>1,4,5</sup>, Yanyang Yan<sup>3</sup>, Xiaochun Cao<sup>1\*</sup>

<sup>1</sup> Sun Yat-sen University <sup>2</sup> Beijing University of Posts and Telecommunications

<sup>3</sup> University of Chinese Academy of Sciences <sup>4</sup> MoE Key Laboratory of Information Technology

<sup>5</sup> Guangdong Key Laboratory of Information Security Technology

yaosiyuan04@gmail.com, ruizhu@bupt.edu.cn, zq\_wang@bupt.edu.cn,

yanyanyang@ict.ac.cn, rwq.renwenqi@gmail.com, caoxiaochun@mail.sysu.edu.cn

## Abstract

Visual object tracking has gained promising progress in past decades. Most of the existing approaches focus on learning target representation in well-conditioned daytime data, while for the unconstrained real-world scenarios with adverse weather conditions, e.g. nighttime or foggy environment, the tremendous domain shift leads to significant performance degradation. In this paper, we propose UMDATrack, which is capable of maintaining high-quality target state prediction under various adverse weather conditions within a unified domain adaptation framework. Specifically, we first use a controllable scenario generator to synthesize a small amount of unlabeled videos (less than 2% frames in source daytime datasets) in multiple weather conditions under the guidance of different text prompts. Afterwards, we design a simple yet effective domain-customized adapter (DCA), allowing the target objects' representation to rapidly adapt to various weather conditions without redundant model updating. Furthermore, to enhance the localization consistency between source and target domains, we propose a target-aware confidence alignment module (TCA) following optimal transport theorem. Extensive experiments demonstrate that UMDATrack can surpass existing advanced visual trackers and lead new state-of-the-art performance by a significant margin. Our code is available at <https://github.com/Z-Z188/UMDATrack>.

## 1. Introduction

Visual object tracking (VOT) is a fundamental visual task of computer vision over the past decades, aiming to estimate the state of arbitrary target objects in video sequences given the initial annotation. Existing mainstream methods formulate object tracking as a target matching problem, which

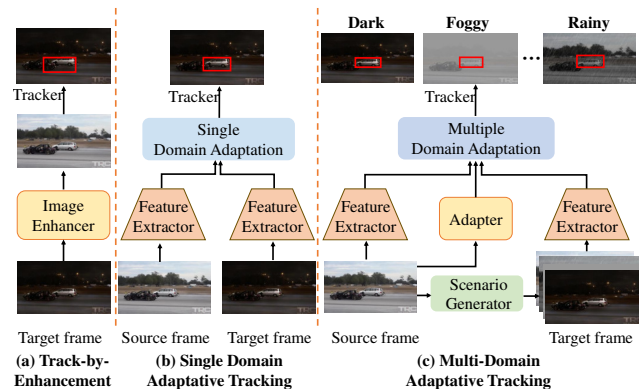


Figure 1. Three representative tracking pipelines under adverse weather conditions. (a) “Track-by-Enhancement” pipeline [48]. (b) Single domain adaptation pipeline [49]. (c) The proposed unified multi-domain adaptive tracking (UMDATrack) pipeline. UMDATrack utilizes controllable scenarios generator to synthesize unlabeled video frames and employ a flexible domain-customized adapter to transfer the knowledge to multi-domain.

constructs template-search pairs to learn a position-sensitive matching network for target localization. Owing to the promising advances of recent deep learning architectures, VOT has achieved remarkable success in terms of accuracy and efficiency.

Recent advanced object trackers typically utilize well-conditioned daytime datasets, e.g. LaSOT [10] or TrackingNet [29] as supervision for model training, however, the performance of these SOTA trackers is unsatisfactory in real-world scenarios with adverse weather conditions (e.g. nighttime or foggy environment) due to the tremendous domain gap. To address this issue, some efforts have explored to introduce synthesized datasets [40, 52] or domain adaptive discriminator [49, 55] to enhance the cross-domain transferability. Despite the significant advances, they potentially suffer from two drawbacks. First, most of the existing ap-

\*Corresponding Author.

proaches are designed for single weather condition, while the generalization abilities are greatly limited in various scenarios where multiple target domains are available. For example, as shown in Fig. 1, the nighttime tracker UDAT [49] is capable of predicting the target state in nighttime data, but its performance drops significantly when the environment changed to another foggy weather condition. Besides, recent domain adaptive trackers generate large amounts of target domain samples for model knowledge transfer, the sample generation process is time-consuming and the intrinsic relationship of the target objects in multiple domains has been overlooked. For different weather conditions in multiple target domains, existing approaches require to introduce redundant parameters to conduct feature alignment separately, which fails to perform cross-domain interaction in an efficient manner.

In this paper, we propose a unified multi-domain adaptive tracker termed UMDATrack, which is capable of maintaining high-quality target state prediction under various adverse weather conditions. Inspired by the great success of the controllable text-to-image generation technique, we first utilize a text-conditioned diffusion model to synthesize unlabeled videos in multiple weather conditions under the guidance of different text prompts. Afterwards, to flexibly transfer the target objects' representation from source domain to multiple target domains, we froze the backbone feature extractor and design a simple yet effective domain-customized adapter (DCA) to remedy the tracking model, allowing it to be rapidly adapted to various weather conditions without redundant model updating. Furthermore, we propose an target-aware confidence alignment module (TCA) with optimal transport theorem, which enhances the localization consistency between source and target domains by measuring the discrepancies of the localization confidence at the candidate positions. Experiments show that by only synthesizing a small partition of videos (less than 2% frames in source domain) at arbitrary weather conditions, UMDATrack can surpass existing advanced visual trackers and lead new state-of-the-art performance on either real-world or synthesized datasets by a significant margin. *To the best of our knowledge, this is the first unified multi-domain adaptation tracker in VOT community.*

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

- We propose a unified multi-domain adaptive tracking framework termed UMDATrack, which conducts multi-domain transfer using text-conditioned diffusion model and maintains high-quality target state prediction under various adverse weather conditions.
- We design a simple yet effective domain-specific adapter (DCA) to remedy the tracking model, which can flexibly transfer the target objects' representation from original daytime scenario to various weather conditions without redundant

model updating.

- We propose a target-aware confidence alignment module (TCA) with optimal transport theorem to enhance the localization consistency in source and target domains. Extensive experiments demonstrate that UMDATrack achieves superior performance to existing state-of-the-art methods.

## 2. Related Work

### 2.1. Tracking in Adverse Weather Conditions

Recently, object tracking in adverse weather conditions has attracted increasing interest due to a variety of practical applications. The classical methods employ multi-modal sensors, e.g. Visible+Depth (RGB-D) [43] Visible+Thermal (RGB-T) [38] for target appearance modeling in complex scenarios. However, these methods require to collect large amount of labelled examples to learn the cross-modal target representation. To address this issue, some works explore to use the RGB images only to transfer the knowledge to unlabelled target domains. Existing methods generally [48, 52] perform image enhancement to unify target object's representation. For example, Zhang *et al.* [52] combine RGB images and the corresponding depth maps to synthesize the foggy images. The feature alignment is conducted on Siamese trackers [6, 45, 46] using the synthesized foggy datasets to eliminate the semantic-level domain shift. HighlightNet [11] adapts to illumination variation and excavates the potential object for low-light UAV tracking. UDAT [49] proposes a transformer-based bridging layer to transfer the semantic knowledge from daytime domain to the nighttime domain. Though effective, the aforementioned trackers are designed for single weather condition, while the generalization abilities are greatly limited in various weather conditions where multiple target domains are available.

### 2.2. Controllable Text-to-Image Generation

To transfer the knowledge in various weather conditions, the scene translation technique has been introduced to synthesize high-quality images. The early efforts use Generative Adversarial Networks (GANs) [18] to transform images from source domain to target domain by modifying image style. However, these GAN-based methods typically require training from scratch on the specific domains. Recently, the advanced text-to-image (T2I) diffusion models [13, 50] have shown impressive controllable flexibilities using text descriptions. GLIDE [30] trains a CLIP model in noisy image space to provide CLIP guidance for image generation and editing. DALL-E [32] employs an autoregressive transformer to combine both text and image tokens, which demonstrates remarkable zero-shot translation capabilities without using large-scale training samples. ControlNet [50] treats the pre-trained model as a strong backbone and finetune the trainable copy connected with zero convolution layers, allowing users

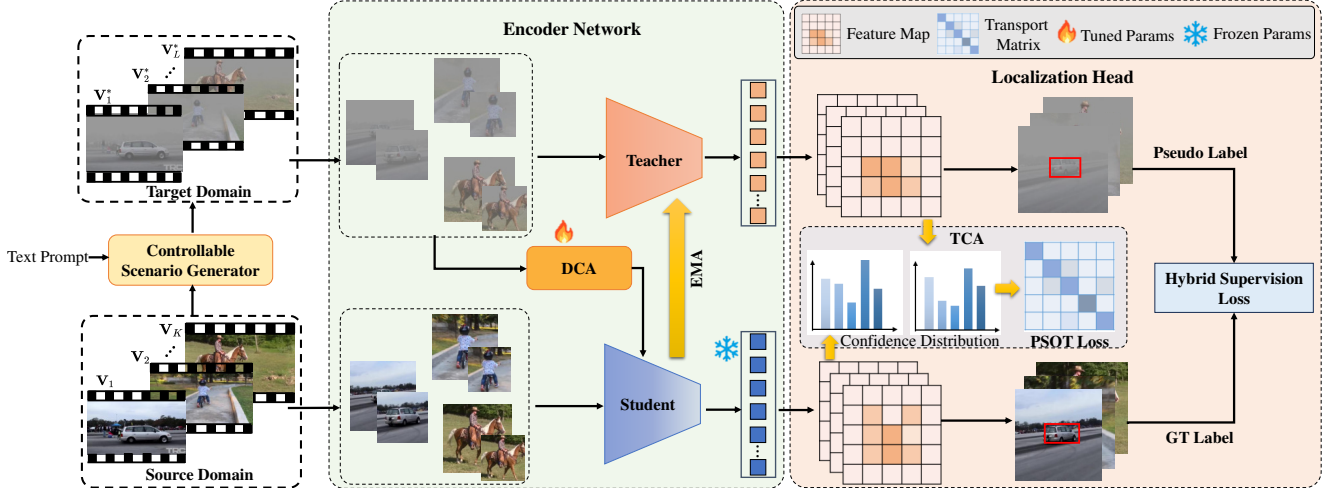


Figure 2. Overview of the proposed UMDATrack. It first utilizes a controllable scenarios generator (CSG) to synthesize the video frames in arbitrary adverse weather conditions. The cropped template-candidate pairs are sent into a student-teacher network, which transfers the target objects’ representation to multiple weather conditions using an encoder network with domain customized adapter (DCA) and a localization head with target-aware confidence alignment module (TCA). Here we only demonstrate the *daytime*  $\rightarrow$  *foggy* environment translation for simplicity.

to add various spatial conditions to control the image generation. Inspired by the success of these text-to-image (T2I) generation models, in this work, we utilize text-conditioned diffusion model to synthesize unlabeled videos in multiple weather conditions for target feature translation.

### 2.3. Multi-Target Domain Adaptation

Recently, various techniques have been employed for Multi-Target Domain Adaptation (MTDA) to enhance cross-domain robustness and generalization. For example, curriculum learning and feature aggregation have been combined to align similar features and adapt models gradually to domain complexities [35]. Other approaches[23] have explored merging independently adapted models from distinct domains by combining model parameters and buffer merging. Additionally, graph matching techniques [24] have been applied to improve generalization in cross-domain object detection, with self-training methods also showing promising potential. Optimal transport theory has been widely studied and applied across various domains. A regularized unsupervised optimal transport model[7] has been proposed to align source and target domain representations, using a transport plan that enhances cross-domain robustness. In particular, SOOD [16] uses optimal transport to ensure global layout consistency between pseudo-labels and predictions. Despite the aforementioned efforts, it is still challenging to design a unified tracker to conduct MTDA in adverse weather conditions like fog, nighttime, and rain. Our research effectively fills this gap by leveraging optimal transport theory to improve tracking robustness in these challenging scenarios.

## 3. Method

In this section, we describe the overall architecture of the proposed UMDATrack, which consists of three main components: a controllable scenarios generator (CSG), an encoder network with domain customized adapter (DCA) and a localization head with target-aware confidence alignment module (TCA).

### 3.1. Controllable Scenario Generator

As it is not trivial to collect large number of video sequences in adverse weather conditions, we first synthesize a small amount of training data to conduct domain knowledge transfer. Inspired by recent advances of text-to-image (T2I) techniques, we utilize a controllable scenario generator (CSG) for data synthesis. Let  $\mathbb{V} = \{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_K\}$  denotes the videos in source domain and  $\mathbb{V}^* = \{\mathbf{V}_1^*, \mathbf{V}_2^*, \dots, \mathbf{V}_L^*\}$  denotes the videos in target domain, here  $L \ll K$  indicates the size of  $\mathbb{V}^*$  is significantly smaller compared to  $\mathbb{V}$ . Our goal is to randomly select the videos in  $\mathbb{V}$  and translate them to arbitrary weather conditions, e.g. hazy, dark and rainy, etc. To achieve this, we use the T2I model, *i.e.* *Stable Diffusion-Turbo* [36] to translate the scenarios using different text prompts. As shown in Fig. 3, the text prompt  $c_X$ , *e.g.* "Car in the night/haze/rain/snow" and the video images  $x \in \mathbb{V}$  in source domain are fed into the text encoder and image encoder respectively. We generate the output video frames  $y \in \mathbb{V}^*$  in target domain by integrating video frame  $x$  with conditional controls  $c_X$  and the noise  $\epsilon$  as:

$$y = G_{\text{SDT}}(x, c_X, \epsilon), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

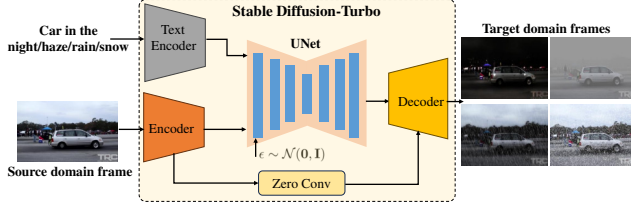


Figure 3. Details of the Controllable Scenario Generation (CSG) module.

where  $G_{\text{SDT}}(x, c_X, \epsilon)$  denotes the Stable Diffusion-Turbo generator,  $\epsilon$  is the noise map. The skip connections and Zero-Convs are used to preserve the essential structural details of the images. Benefited from the powerful transferability of T2I model, the video frames in target domains can be rapidly generated within only 1-4 iteration steps by simply changing the text prompts.

### 3.2. Tracking in Multiple Weather Conditions

Though CSG can generate continuous video frames in multiple weather conditions, the appearance discrepancies of target objects between the source daytime videos and the synthesized videos still limit the tracker’s generalization ability. To address this issue, we design a unified domain adaptation framework following the teacher-student pipeline, which can be flexibly deployed to various domain-customized scenarios. Specifically, given  $N_S$  video frames  $\mathcal{D}_S = \{(\mathcal{I}_i^S, \mathbf{b}_i^S)\}_{i=1}^{N_S}$  in source domain and  $N_T$  unlabeled frames  $\mathcal{D}_T = \{\mathcal{I}_i^T\}_{i=1}^{N_T}$ , where  $\mathcal{I}_i^S$  and  $\mathbf{b}_i^S$  denotes the images and annotated bounding boxes in the source domain,  $\mathcal{I}_i^T$  denotes the images in multiple target domains. We crop the paired template-search images of  $\mathcal{D}_S$  and  $\mathcal{D}_T$  and then send them into the student and teacher network, respectively. The **student**  $\rightarrow$  **teacher** knowledge transfer is conducted by updating the weights of the teacher model using the EMA (Exponential Moving Average) as:

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

where  $\theta^T$  and  $\theta^S$  denote the learnable parameters of the teacher and student networks.  $\alpha$  is the momentum coefficient controlling the updating rate of the teacher.

**Domain-Customized Adapter** The student-teacher training paradigm allows the tracker to gradually propagate source domain information to target domain. However, as the data distributions in different weather conditions vary greatly, it’s time-consuming to generate large amounts of multi-domain samples and would inevitably introduce redundant parameters if we conduct domain knowledge transfer separately. Considering this, we propose a Domain Customized Adapter (DCA) to transfer the target object’s representation to arbitrary weather conditions in an efficient fashion.

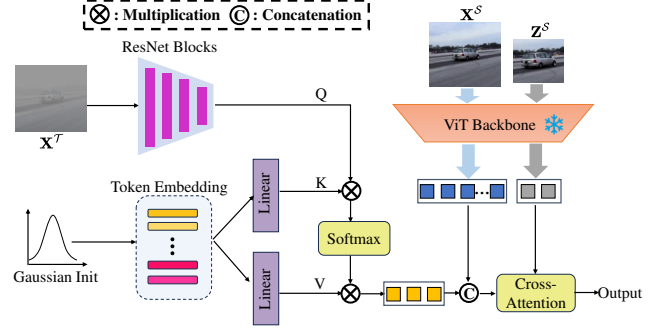


Figure 4. Details of the Domain-Customized Adapter (DCA) module.

We present the detailed structure of DCA in Fig. 4. Formally, suppose the cropped template-search images in source domain are  $\mathbf{Z}^S$  and  $\mathbf{X}^S$ , respectively. While the image pairs in target domain are  $\mathbf{Z}^T$  and  $\mathbf{X}^T$ . We first use a lightweight ResNet block to transform and reshape  $\mathbf{X}^T$  as query  $\mathbf{Q} \in \mathbb{R}^{K \times C}$ . Then we initialize a Gaussian random variable and embed it to be learnable token bank  $\mathbf{B} \in \mathbb{R}^{L' \times C}$  that consists of  $L'$  learnable feature vectors with channel dimension  $C$ . The token bank  $B$  is further projected as key-value tokens  $K$  and  $V$  with the size of  $L' \times C$  by two FC layers, respectively. We compute an structural token  $\mathbf{S}$  between the query and embedded key-value tokens as follows:

$$\mathbf{S} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (3)$$

the structural token  $\mathbf{S} \in \mathbb{R}^{K \times C}$  encodes the latent image content representation, which shares similar contextual structure to  $\mathbf{X}^S$  in the embedding space. The structural token  $\mathbf{S}$  are subsequently fed into the frozen vision transformer and concatenated with the encoded template-search tokens of the source domain images, allowing the model to rapidly find the optimal convergence checkpoints in various adverse weather conditions.

### 3.3. Target-Aware Confidence Alignment

Since the annotations are only available in the source domain, we train the tracker following a pseudo-label propagation strategy. Specifically, we send the synthesized template-search pairs into the teacher network to generate pseudo labels. These pseudo labels are fed back into the student network as supervision to update the weights of the tracking model. However, as the pseudo labels may be noisy, the incorrect pseudo labels will mislead the target state prediction. To address this problem, we propose a Target-Aware Confidence Alignment (TCA) module using optimal transport theory (OT) to enhance localization consistency in both domains by measuring the discrepancies in localization confidence at the candidate positions.

To be concrete, suppose the regressed response maps of student and teacher network are  $\mathbf{r}^S \in \mathbb{R}^{N \times (H' \times W')}$  and  $\mathbf{r}^T \in \mathbb{R}^{N \times (H' \times W')}$ , where  $N$  denotes the number of image samples in a mini-batch,  $H'$ ,  $W'$  represent the height and width of the response maps. We construct confidence distributions  $\mathbf{d}^S \in \mathbb{R}^N$  and  $\mathbf{d}^T \in \mathbb{R}^N$  for each sample in a mini-batch as:

$$\mathbf{d}^S = \exp(\mathbf{r}_{i, \mathbf{p}_i}^S), \mathbf{d}^T = \exp(\mathbf{r}_{i, \mathbf{p}_i}^T), \quad (4)$$

where for the  $i$ -th sample,  $\mathbf{p}_i = \arg \max_{j=1 \dots H' \times W'} \mathbf{r}_{i, j}^T$  denotes the spatial index of the response map with the highest confidence score.

To construct the costmap  $\mathbf{C}_{i, j}$  for the OT problem, we simultaneously consider the spatial and confidence discrepancies of each sample. Here we introduce two cost to measure the matching cost:

$$\mathbf{C}_{i, j}^{\text{Conf}} = \frac{\|\mathbf{r}_{i, \mathbf{p}_i}^S - \mathbf{r}_{j, \mathbf{p}_j}^T\|_1}{\max_{1 \leq m, n \leq N} \|\mathbf{r}_{m, \mathbf{p}_m}^S - \mathbf{r}_{n, \mathbf{p}_n}^T\|_1}, \quad (5)$$

$$\mathbf{C}_{i, j}^{\text{Pos}} = \frac{\|\mathbf{p}_i^S - \mathbf{p}_j^T\|_2}{\max_{1 \leq m, n \leq N} \|\mathbf{p}_m^S - \mathbf{p}_n^T\|_2}, \quad (6)$$

$$\mathbf{C}_{i, j} = \mathbf{C}_{i, j}^{\text{Conf}} + \mathbf{C}_{i, j}^{\text{Pos}} \quad (7)$$

where  $\mathbf{C}^{\text{Conf}}$  and  $\mathbf{C}^{\text{Pos}}$  represent the confidence and position cost between the distribution  $\mathbf{d}^S$  to  $\mathbf{d}^T$ .

Based on what we discussed above, we design a **position-sensitive optimal transport (PSOT)** loss to measure the cost for moving the confidence distribution from  $\mathbf{d}^S$  to  $\mathbf{d}^T$ , which can be defined as the OT problem's **dual** formulation:

$$L_p = \left\langle \boldsymbol{\mu}, \frac{\mathbf{d}^T}{\|\mathbf{d}^T\|_1} \right\rangle + \left\langle \boldsymbol{\nu}, \frac{\mathbf{d}^S}{\|\mathbf{d}^S\|_1} \right\rangle. \quad (8)$$

where  $\boldsymbol{\mu}$  and  $\boldsymbol{\nu}$  are the solutions of the OT problem. The details can be found in appendix B.

During training, we jointly adopt the target supervision loss and position-sensitive optimal transport loss as hybrid supervision loss to train the whole student-teacher model, which is given by:

$$\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_p, \quad (9)$$

where  $\lambda$  is the hyperparameter to balance the weights of the loss terms. We solve the OT problem by a fast Sinkhorn distances algorithm [8]. Similar to [47], the target supervision loss consists of the classification loss, localization  $L_1$  loss and generalized GIoU loss as below:

$$L_t = \mathcal{L}_{cls} + \beta L_1 + \gamma L_{GIoU}. \quad (10)$$

By minimizing the target supervision loss and position-sensitive optimal transport loss, the feature representations and localization response can be effectively aligned to alleviate the domain shift.

## 4. Experiments

In this section, we conduct several experiments to evaluate the effectiveness of our proposed method. Our method is implemented based on python 3.10 and pytorch 2.1.1. Our tracker is trained with 4 NVIDIA RTX 3090 GPUs. All of the inference speed testing are conducted on a single NVIDIA RTX 3090 GPU.

### 4.1. Implementation Details

**Model settings.** We adopt vanilla ViT-Base [9] model as the backbone of our tracker, similar to OSTrack[47]. The patch size is set to  $16 \times 16$ . We adopt a lightweight FCN consists of 4 stacked Conv-BN-ReLU layers as prediction head for both teacher and student branches. The sizes of the template and search region are resized to  $128 \times 128$  and  $256 \times 256$  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 stage and domain customized training stage. We first synthesize the videos in adverse weather conditions only using GOT-10k dataset, the synthesized datasets includes GOT-10k-Dark, GOT-10k-Foggy and GOT-10k-Rainy. For backbone training, the DCA module is not introduced, we employ target supervision loss and position-sensitive optimal transport loss to perform domain adaptation between the teacher and student networks. Four source domain datasets, including LaSOT [10], TrackingNet [29], COCO [26], and GOT-10k [17], as well as three synthetic datasets train the student model. The sampling ratio of the datasets is set to 1:1:1:1:4:4:4. The backbone training takes 250 epochs. The learning rate is  $4 \times 10^{-4}$  and decreased with weight decay  $1 \times 10^{-4}$ . The EMA hyperparameter  $\alpha$  is set to 0.99. For domain customized training stage, we froze the backbone feature extractor and train the DCA module for an additional 50 epochs. Both two stages optimize the model with ADAMW. Note that our UMDATrack does not require repetitive backbone training stage, we only need to train the DCA module for each weather condition. Therefore, it only takes one and a half days to train UMDATrack in all weather conditions. This approach significantly improves training efficiency while maintaining superior model performance.

**Loss Function.** In our implementation, we utilize focal loss [34] for foreground-background classification and employ L1 loss and GIoU loss [33, 44] for bounding box regression. Additionally, PSOT (Position-Sensitive Optimal Transport) loss is applied to align the distributions between the teacher and student networks. The weighting coefficients for the focal loss, L1 loss, GIoU loss, and PDOT loss are set to 1.0, 5.0, 2.0, and 10.0, respectively.

**Inference.** To accelerate the inference, the template feature is initialized using the first frame of each video sequence and stored for relation modeling between the template and search region in subsequent frames. As demonstrated in Tab 3, we

Table 1. Comparison with state-of-the-art visual trackers on synthetic datasets: GOT-10k-Foggy, DTB70-Foggy, GOT-10k-Dark, DTB70-Dark, GOT-10k-Rainy and DTB70-Rainy. The top two results are highlighted with red and blue fonts, respectively. The double line above represents the **cross-domain** trackers, while the line below represents the **generic** trackers.

Tracker	GOT-10k-Foggy			DTB70-Foggy		GOT-10k-Dark			DTB70-Dark		GOT-10k-Rainy			DTB70-Rainy	
	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>	AUC	P	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>	AUC	P	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>	AUC	P
<b>UMDATrack</b>	<b>66.6</b>	<b>75.8</b>	<b>62.2</b>	<b>66.21</b>	<b>86.05</b>	<b>65.4</b>	<b>75.3</b>	<b>57.3</b>	<b>66.07</b>	<b>85.72</b>	<b>68.5</b>	<b>78.4</b>	<b>63.2</b>	<b>66.75</b>	<b>87.60</b>
DCPT[55]	61.6	70.2	56.9	58.31	75.33	62.4	70.5	54.2	61.87	80.11	62.3	70.1	59.8	61.68	82.56
UDAT-CAR[49]	51.5	60.3	45.2	50.21	69.41	56.8	64.2	49.1	57.20	75.80	59.5	65.2	55.3	56.42	75.36
SAM-DA[12]	50.2	60.5	48.3	51.33	69.89	55.4	63.1	48.3	57.15	75.12	60.2	66.1	57.6	57.63	76.12
MLKD-Track[28]	52.3	62.3	49.1	52.46	70.32	53.8	61.6	46.9	55.21	73.68	57.3	64.8	57.1	56.89	74.12
ARTrackV2[1]	64.8	73.0	<b>59.9</b>	<b>62.25</b>	<b>80.15</b>	<b>63.1</b>	<b>72.8</b>	53.9	62.87	80.56	<b>66.2</b>	<b>75.8</b>	61.2	63.84	83.32
EVPTrack[37]	63.5	70.7	56.5	57.96	75.45	62.7	71.8	53.9	<b>63.01</b>	81.12	65.5	75.2	60.5	<b>64.03</b>	<b>84.11</b>
ODTrack[53]	65.1	74.5	56.0	61.12	79.32	62.5	71.5	53.1	62.21	80.23	64.8	74.5	59.5	63.95	83.56
HipTrack[3]	63.3	72.0	59.6	60.52	78.22	62.9	72.4	53.8	62.48	80.57	65.6	75.4	60.2	63.57	83.36
DropTrack[41]	64.9	73.8	58.5	59.95	77.66	62.2	72.5	<b>54.3</b>	61.98	80.21	65.3	75.3	60.4	62.87	83.13
SeqTrack[5]	<b>65.2</b>	<b>74.6</b>	56.3	60.21	78.70	61.4	70.5	52.3	62.84	<b>81.57</b>	65.1	75.0	60.3	63.75	83.28
AQATrack[42]	64.9	72.8	59.7	57.28	75.61	61.7	70.6	52.5	61.17	79.87	63.4	72.3	<b>61.8</b>	63.12	83.55
ROMTrack[4]	63.6	70.9	56.7	59.05	76.59	60.8	71.1	51.7	60.80	77.95	62.7	73.4	60.1	63.21	83.25
OSTrack[47]	61.9	71.7	59.7	56.23	77.43	61.3	70.9	51.5	59.23	77.43	61.6	71.0	58.6	59.23	77.43
AVTrack[25]	56.9	63.5	49.5	52.35	68.09	55.3	62.3	46.2	56.66	72.21	57.5	63.4	48.1	60.21	79.53
DiMP[2]	57.6	64.2	50.4	53.80	69.50	56.9	60.4	44.3	55.20	72.30	57.9	63.8	49.2	57.32	75.21
SiamRPN++[20]	58.4	64.9	51.2	55.80	74.70	56.6	60.8	45.1	48.80	70.30	56.2	61.4	46.8	51.52	71.96
SiamRPN[21]	51.7	55.6	32.5	47.40	67.40	49.2	53.2	31.4	43.70	60.30	50.1	54.6	35.1	48.25	68.22

compared inference speed, MACs, and parameter counts with those of state-of-the-art trackers, showing that UMDATrack achieves the highest inference speed with relatively low computational costs and parameter counts.

## 4.2. Comparisons with State-of-the-arts

In this subsection, we comprehensively compare UMDATrack with SOTA trackers in both real-world and synthesized adverse weather conditions to demonstrate the effectiveness and high efficiency of our method. It's worth noting that our task is focused on cross-domain tracking, rather than being a generic one. However, we have observed significant performance improvement compared to the current state-of-the-art in generic trackers.

Specifically, for nighttime conditions, we use the real-world NAT2021-test[49], UAVDark70[19], and two synthesized datasets, i.e. GOT-10k-Dark, and DTB70-Dark. For foggy environment, we evaluate the tracking performance using the GOT-10k-Foggy and DTB70-Foggy datasets. For rainy conditions, we use the GOT-10k-Rainy and DTB70-Rainy datasets. Finally, we use the real-world AViT [31] dataset to evaluate the tracking performance under various adverse weather conditions in natural environment.

**Synthetic GOT-10k and DTB70[22].** As shown in Table 1, UMDATrack performs exceptionally well across all three challenging conditions (foggy, dark, and rainy) on both the synthetic GOT-10k and DTB70 datasets. Under dark conditions, UMDATrack achieved the highest AUC (66.07) and

precision (85.72) on the DTB70-Dark dataset, outperforming the second-best results by a notable margin of 3.06% in AUC and 4.15% in precision. A similar trend is observed on the GOT-10k-Dark dataset, where UMDATrack leads both AUC and precision. In foggy conditions, UMDATrack outperforms the second-best results obtained by other trackers by 3.96% in AUC and 5.90% in precision on the DTB70-Foggy dataset. In rainy conditions, UMDATrack also demonstrates superior performance to the advanced SOTA trackers. e.g. ARTrackV2 or ODTrack.

**Results on Real-World datasets** To further verify the effectiveness of the proposed UMDATrack, we conduct experiments on the real-world datasets with adverse weather conditions for comparison. As shown in Table 2, on the large-scale night dataset NAT2021, UMDATrack achieved the best AUC (54.58) and precision (70.78). Specifically, in terms of AUC, we outperformed the second tracker ARTrackV2 (53.13) by 1.45 points. This partially proves that our proposed framework helps the model learn effectively from synthetic extreme domain datasets. For the challenging UAV tracking dataset UAVDark70, UMDATrack outperforms all other trackers on the UAVDark70 real-world dataset, achieving an AUC score 1.83 points higher and a precision 1.4 points greater than the second-best tracker. Note that most of the reported trackers in the table can not directly deployed run for UAV system. However, UMDATrack obtains the best performance with real-time speed, shown great potential in real-world UAV tracking. Furthermore, we also test UM-

Table 2. Comparison with state-of-the-art visual trackers on real-world datasets: NAT2021, UAVDark70, and AVisT. The top two results are highlighted in red and blue, respectively. The double line above represents the **cross-domain** trackers, while the line below represents the **generic** trackers.

Tracker	NAT2021		UAVDark70		AVisT	
	AUC	P	AUC	P	AUC	P
<b>UMDATrack</b>	<b>54.58</b>	<b>70.78</b>	<b>60.05</b>	<b>73.35</b>	<b>60.50</b>	<b>59.01</b>
DCPT[55]	52.55	69.01	56.86	70.16	55.66	52.41
UDAT-CAR[49]	48.75	65.96	51.25	70.22	38.91	33.65
SAM-DA[12]	47.31	65.50	49.52	65.59	37.36	34.29
MLKD-Track[28]	44.31	60.21	47.27	61.54	33.62	30.26
ARTrackV2[1]	<b>53.13</b>	<b>69.72</b>	<b>58.22</b>	<b>71.95</b>	58.52	57.65
ODTrack[53]	53.11	69.68	58.07	71.11	58.63	57.36
EVPTrack[37]	53.08	69.51	57.47	71.10	57.31	55.55
DropTrack[41]	52.98	69.11	58.13	71.86	<b>59.56</b>	<b>57.97</b>
ROMTrack[4]	51.57	68.75	53.77	69.80	56.12	55.09
SeqTrack[5]	51.65	67.97	53.88	66.88	57.15	55.30
AQATrack[42]	51.33	67.03	58.18	70.98	57.32	56.60
SMAT[14]	45.96	59.87	45.19	56.71	50.35	49.58
AVTrack[25]	45.41	59.51	46.91	59.49	49.21	48.50

Table 3. Comparison of inference speed, FLOPs, and model parameters across different trackers.

Tracker	Speed (FPS)	MACs (G)	Params (M)
<b>UMDATrack</b>	<b>138</b>	<b>18</b>	<b>65</b>
ARTrackV2[1]	95	45	126
EVPTrack[37]	71	22	74
AQATrack[42]	68	26	72
DropTrack[41]	52	48	92
SeqTrack[5]	40	66	89

DATrack on AVisT dataset, which is specifically collected for tracking in diverse scenarios with adverse visibility. The various weather conditions such as rain, snow, fog and camouflage are included in this dataset, UMDATrack also obtains the leading performance in both precision and AUC metrics.

**Inference Speed.** Since UMDATrack does not require to introduce heavy blocks for target appearance model, the computational cost of UMDATrack is limited. As demonstrated in Table 3, we compared inference speed, MACs, and parameter counts with those of state-of-the-art trackers, showing that UMDATrack achieves the highest inference speed with relatively low computational costs and parameter counts.

### 4.3. Ablation Studies and Visualization

**Study on the components of UMDATrack.** We conducted ablation experiments on the proposed three modules to verify their effectiveness. As shown in Table 4, the baseline approach doesn't introduce any modules, thus it is only trained

Table 4. Ablation study on the individual impact of each module (CSG, DCA, and TCA) in our model. The presence or absence of each module is marked with a check or dash, respectively. Results are reported in terms of AUC and Precision for each configuration, evaluated on the NAT2021 dataset.

Modules			Indicators	
CSG	DCA	TCA	AUC (%)	Precision (%)
-	-	-	49.11	63.52
✓	-	-	50.90	65.38
-	✓	-	50.56	65.50
✓	-	✓	52.27	67.10
✓	✓	-	52.24	67.49
✓	✓	✓	<b>54.58</b>	<b>70.78</b>

only on the four source domain datasets. When the CTG module is introduced, the model achieves the AUC of 50.90% and Precision of 65.38%. Adding the TCA module improves these results, bringing the AUC to 52.27% and precision to 67.10%. Further including the DCA module increases performance to the AUC of 54.58% and Precision of 70.78%. These results demonstrate that each module provides a significant performance gain, with the full model configuration yielding the highest scores in both metrics on the NAT2021 dataset.

Table 5. Effect of different EMA (Exponential Moving Average) update frequencies on model performance.

EMA Frequency	AUC (%)	Precision (%)
Each epoch	<b>54.48</b>	<b>70.78</b>
Every 3 epochs	53.65	68.99
Every 5 epochs	52.90	68.22
Each batch	53.89	69.57

Table 6. Different dataset proportions used for training, with LaSOT, GOT-10k, TrackingNet, COCO, and Synthetic datasets in the specified ratios.

Dataset Proportion	AUC (%)	Precision (%)
1 : 1 : 1 : 1 : 1 : 1 : 1	53.13	68.72
1 : 1 : 1 : 1 : 2 : 2 : 2	53.68	69.01
1 : 1 : 1 : 1 : 4 : 4 : 4	<b>54.58</b>	<b>70.78</b>
1 : 1 : 1 : 1 : 6 : 6 : 6	54.26	70.44

### Study on the training hyper-parameter of UMDATrack.

We conducted two ablation studies on the update frequency of EMA and the proportion of the training dataset. As shown in Table 5, we experimented with performing EMA after each epoch, every three epochs, every five epochs, and after completing each batch to transfer student network's weight

to the teacher network. The results indicate that performing EMA after each epoch yields the best results. For the dataset proportion settings, we conducted four groups of experiments as shown in the Table 6, and the results indicate that group 3 achieve the best performance. Therefore, we set the training dataset proportion to 1:1:1:1:4:4.

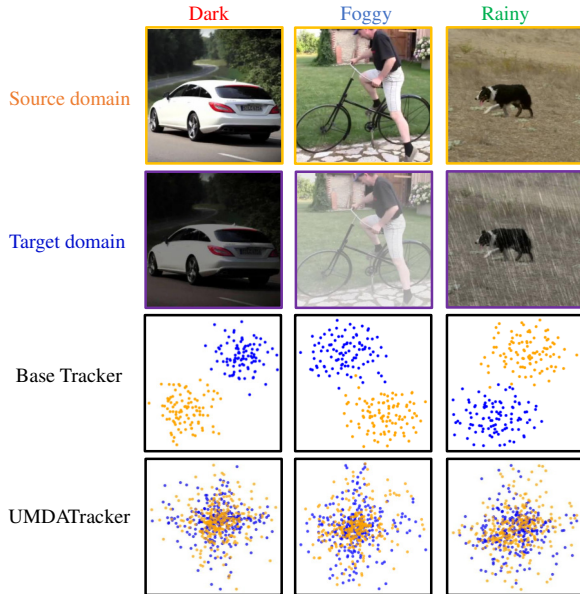


Figure 5. Feature visualization by t-SNE of dark, foggy, and rainy scenes compared to normal (daytime) scenes. Orange and blue indicate source domain and target domains, respectively. The scattergrams depict the feature distributions of the base tracker and UMDATracker across different weather conditions. The results show that UMDATracker effectively narrows the domain discrepancy in various challenging weather conditions.

Table 7. Quality comparison of synthetic datasets generated by different generators. AUC is evaluated on NAT2021 dataset.

Method	SSIM $\uparrow$	LPIPS $\downarrow$	Time (h)	AUC (%)
<b>CSG with Text</b>	<b>0.920</b>	<b>0.086</b>	<b>24</b>	<b>54.58</b>
CSG without Text	0.902	0.104	20	52.53
CycleGAN [54]	0.895	0.119	30	51.10
UNIT[27]	0.875	0.136	14	50.23
Gamma(only for dark)	0.787	0.216	5	-

**Study on the impact of the synthetic datasets.** We use SSIM [39] and LPIPS[51] to evaluate image quality in the second and third columns of Table 7, Compared to other methods like CycleGAN, UNIT, or simply using Gamma, CSG especially with text prompt achieves the best generation quality. Although our generator requires slightly more time to synthesize datasets, this is a trade-off between data generation quality and computational time. The use of text prompts

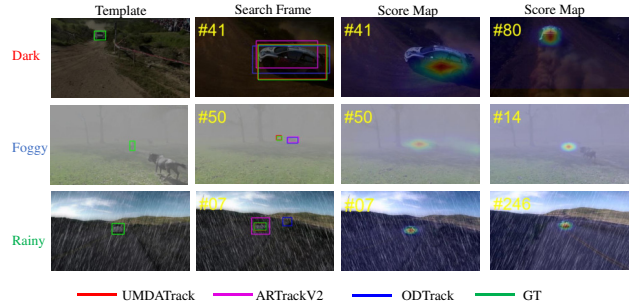


Figure 6. Visualization comparison of our approach and other excellent trackers and results of the scoremaps.

improves the quality and relevance of the generated datasets, leading to better downstream performance. As a result, the tracker achieves the best AUC performance.

**Visualizing Robustness in Adverse Conditions.** Fig. 5 shows feature distributions using t-SNE[15], where UMDATrack better aligns source domain and target domain across dark, foggy, and rainy conditions, reducing domain discrepancy. Fig. 6 presents tracking results, with UMDATrack achieving higher accuracy and significantly stronger resistance compared to other trackers in extreme scenarios.

## 5. Conclusion

In this paper, we propose a unified multi-domain adaptive tracker termed UMDATrack to predict target state under various adverse weather conditions. We first use a controllable scenario generator to synthesize unlabeled videos in multiple weather conditions under the guidance of different text prompts. Afterwards, we propose a simple yet effective domain-customized adapter to remedy the tracking model, allowing it to rapidly adapt to various weather conditions without redundant model updating. Furthermore, we propose a target-aware confidence alignment module (TCA) with optimal transport theorem, which enhances the localization consistency between source and target domains by measuring the discrepancies of the localization confidence at the candidate positions. Experiments show that UMDATrack leads new state-of-the-art performance on either real-world or synthesized datasets by a significant margin.

## 6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62402055, 62322216, 62025604, 62302480), the Shenzhen Science and Technology Program (CJGJZD20240729141505007).

## References

- [1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Ar-trackv2: Prompting autoregressive tracker where to look and

- how to describe. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19048–19057, 2024.
- [2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *IEEE International Conference on Computer Vision*, pages 6182–6191, 2019.
  - [3] Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19258–19267, 2024.
  - [4] Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *IEEE International Conference on Computer Vision*, pages 9589–9600, 2023.
  - [5] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14572–14581, 2023.
  - [6] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6667–6676, 2020.
  - [7] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.
  - [8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Annual Conference on Neural Information Processing Systems*, pages 2292–2300, 2013.
  - [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
  - [10] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2019.
  - [11] Changhong Fu, Haolin Dong, Junjie Ye, Guangze Zheng, Sihang Li, and Jilin Zhao. Highlightnet: Highlighting low-light potential features for real-time UAV tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 12146–12153, 2022.
  - [12] Changhong Fu, Liangliang Yao, Haobo Zuo, Guangze Zheng, and Jia 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*, pages 1–8, 2024.
  - [13] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *IEEE International Conference on Computer Vision*, pages 7511–7522, 2023.
  - [14] Goutam Yelluru Gopal and Maria A Amer. Separable self and mixed attention transformers for efficient object tracking. In *IEEE Winter Conference on Applications of Computer Vision*, pages 6708–6717, 2024.
  - [15] G Hinton and L Van Der Maaten. Visualizing data using t-sne journal of machine learning research. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
  - [16] Wei Hua, Dingkang Liang, Jingyu Li, Xiaolong Liu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. SOOD: towards semi-supervised oriented object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15558–15567, 2023.
  - [17] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019.
  - [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017.
  - [19] Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. Adtrack: Target-aware dual filter learning for real-time anti-dark UAV tracking. In *IEEE International Conference on Robotics and Automation*, pages 496–502, 2021.
  - [20] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.
  - [21] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.
  - [22] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI Conference on Artificial Intelligence*, pages 4140–4146, 2017.
  - [23] Wenyi Li, Huan-ang Gao, Mingju Gao, Beiwen Tian, Rong Zhi, and Hao Zhao. Training-free model merging for multi-target domain adaptation. *arXiv preprint arXiv:2407.13771*, 2024.
  - [24] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5291–5300, 2022.
  - [25] Yongxin Li, Mengyuan Liu, You Wu, Xucheng Wang, Xi-angyang Yang, and Shuiwang Li. Learning adaptive and view-invariant vision transformer for real-time uav tracking. In *International Conference on Machine Learning*.
  - [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.
  - [27] Ming-Yu Liu, Thomas M. Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Annual Conference on Neural Information Processing Systems*, pages 700–708, 2017.
  - [28] Yufeng Liu. Mutual-learning knowledge distillation for nighttime UAV tracking. *arXiv preprint arXiv:2312.07884*, 2023.
  - [29] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *European Conference on Computer Vision*, pages 310–327,

- 2018.
- [30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804, 2022.
- [31] Mubashir Noman, Wafa Al Ghallabi, Daniya Kareem, Christoph Mayer, Akshay Dudhane, Martin Danelljan, Hisham Cholakkal, Salman Khan, Luc Van Gool, and Fahad Shahbaz Khan. Avist: A benchmark for visual object tracking in adverse visibility. In *British Machine Vision Conference*, page 817, 2022.
- [32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021.
- [33] Hamid Rezaatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.
- [34] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.
- [35] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5351–5360, 2021.
- [36] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103, 2024.
- [37] Liangtao Shi, Bineng Zhong, Qihua Liang, Ning Li, Shengping Zhang, and Xianxian Li. Explicit visual prompts for visual object tracking. In *AAAI Conference on Artificial Intelligence*, pages 4838–4846, 2024.
- [38] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for RGB-T tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7062–7071, 2020.
- [39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [40] Hongjing Wu, Siyuan Yao, Feng Huang, Shu Wang, Linchao Zhang, Zhuoran Zheng, and Wenqi Ren. Lvprack: High performance domain adaptive UAV tracking with label aligned visual prompt tuning. In *AAAI Conference on Artificial Intelligence*, pages 8395–8403, 2025.
- [41] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B. Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023.
- [42] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19300–19309, 2024.
- [43] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Ales Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of RGBD tracking. In *IEEE International Conference on Computer Vision*, pages 10705–10713, 2021.
- [44] Siyuan Yao, Yang Guo, Yanyang Yan, Wenqi Ren, and Xiaochun Cao. Untrack: Reliable visual object tracking with uncertainty-aware prototype memory network. *IEEE Transactions on Image Processing*, 34:3533–3546, 2025.
- [45] Siyuan Yao, Xiaoguang Han, Hua Zhang, Xiao Wang, and Xiaochun Cao. Learning deep lucas-kanade siamese network for visual tracking. *IEEE Transactions on Image Processing*, 30:4814–4827, 2021.
- [46] Siyuan Yao, Hua Zhang, Wenqi Ren, Chao Ma, Xiaoguang Han, and Xiaochun Cao. Robust online tracking via contrastive spatio-temporal aware network. *IEEE Transactions on Image Processing*, 30:1989–2002, 2021.
- [47] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357, 2022.
- [48] Junjie Ye, Changhong Fu, Guangze Zheng, Ziang Cao, and Bowen Li. Darklighter: Light up the darkness for UAV tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3079–3085, 2021.
- [49] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022.
- [50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*, pages 3813–3824, 2023.
- [51] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [52] Zhongzhou Zhang and Lei Zhang. Domain adaptive siamrpn++ for object tracking in the wild. *arXiv preprint arXiv:2106.07862*, 2021.
- [53] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *AAAI Conference on Artificial Intelligence*, 2024.
- [54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2242–2251, 2017.
- [55] Jiawen Zhu, Huayi Tang, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Shihao Qiu, Shengming Li, and Huchuan Lu. DCPT: darkness clue-prompted tracking in nighttime uavs. In *IEEE International Conference on Robotics and Automation*, pages 7381–7388, 2024.