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

Supplementary Material

In this supplementary material, we provide more details of the proposed UMDATrack. Specifically, in section A, we display the synthesizing datasets, along with the mathematical formulations of Optimal Transport in section B. In section C, we also include speed test results on embedded device to demonstrate the practical efficiency of our approach. Furthermore, in section D, we add more ablation studies. Finally, we show extensive visualization results of UMDATrack in diverse challenging scenarios in section E.

A. Visualization of the Synthesizing Datasets

To evaluate the robustness of our tracker in adverse conditions, we visualize the frames synthesized by CSG in various weather conditions, including darkness, fog, rain, and snow. As illustrated in Fig. 1, by introducing different captions for domain-specific translation, CSG can flexibly transfer arbitrary video frames to the desired target domain via changing the text prompts.

B. Optimal Transport

The mathematical formulations of Optimal Transport (OT) are described as follows in detail. Suppose $\mathbf{p} \in \mathbb{R}^m$ and $\mathbf{q} \in \mathbb{R}^n$ represent two discrete probabilistic distributions in different domains. OT aims to find a transportation plan that minimizes the transportation cost as follows:

$$\min \langle \mathbf{C}, X \rangle \,, \text{s.t. } X\mathbf{1} = \mathbf{p}, X^\top \mathbf{1} = \mathbf{q}, \tag{1}$$

where $X \in \mathbb{R}^{m \times n}$ is the transportation plan from $\mathbf{p}$ to $\mathbf{q}$, $\mathbf{C} \in \mathbb{R}^{m \times n}$ is the costmap. The OT problem can be converted to its dual formulation, which is given by:

$$\mathbf{W}_{ot}(\mathbf{p}, \mathbf{q}) = \max_{\boldsymbol{\mu}, \boldsymbol{\nu}} \langle \boldsymbol{\mu}, \mathbf{p} \rangle + \langle \boldsymbol{\nu}, \mathbf{q} \rangle \,, \text{s.t. } \mu_i + \nu_j \leq \mathbf{C}_{i,j}, \forall i, j, \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \in \mathbb{R}^n$ are the solutions of the OT problem. In TCA, we reshape the response maps of the teacher-student networks into vector representation and use $\mathbf{C} = \mathbf{C}^{\text{Conf}} + \mathbf{C}^{\text{Pos}}$ as the total cost matrix. The OT problem can be optimized using a fast Sinkhorn distances algorithm.

C. Deployment on Embbed Device

To validate the effectiveness of our model in real-world mobile embedded systems, we deployed UMDATrack on the NVIDIA Jetson AGX Orin for speed testing. The NVIDIA Jetson AGX Orin is an embedded AI computing platform designed for edge AI inference and compute-intensive tasks.

Table 1. Inference speed tested under different power settings on the NVIDIA Jetson AGX Orin.

Power mode	15W	30W	60W
FPS	21.07	35.48	46.32

Table 2. Experiments on the hyperparameter and the composition of costmap $\mathbf{C}$. The results are evaluated on NAT2021 dataset.

λ	$\mathbf{C}^{\text{Conf}}$	$\mathbf{C}^{\text{Pos}}$	AUC (%)	Precision (%)
10	-	-	52.24	67.49
10	✓	-	54.01	69.21
10	-	✓	53.69	68.76
10	✓	✓	**54.58**	**70.78**
1	✓	✓	52.14	68.55
100	✓	✓	52.76	68.36

It has different power modes, allowing us to easily test the performance of models under low power and low computational conditions. Speed test results are shown in Table 1. Although the power setting is reduced to 30W, UMDATrack is able to achieve real-time tracking (35.48 FPS), validating the effectiveness of our model's lightweight design. Under full power (60W) condition, we achieve the highest speed of 46.32 FPS.

D. More Ablation Studies

Our hybrid supervision loss $\mathcal{L}$ consists of two parts: L_t and L_p. L_t follows the loss design used by most trackers, such as OSTrack [2], while L_p is our proposed position-sensitive optimal transport (PSOT) loss. Here we analyze the effects of different compositions of costmap $\mathbf{C}$ in L_p and the hyperparameter λ. As shown in Table 2, we can observe that both confidence and position costs ($\mathbf{C}^{\text{Conf}}$ and $\mathbf{C}^{\text{Pos}}$) are essential to L_p. Besides, setting $\lambda = 10$ yields the highest AUC and Precision scores.

E. More Visualization Results of UMDATrack

E.1. Tracking Results Visualization

In Fig. 2, we present some challenging videos with extreme weather scenarios, including dark, foggy, and rainy conditions. We can see that in these cases, the generic trackers like ARTrackV2 [1] and ODTrack [3] fail to capture the indistinguishable target object due to the extreme low light or fog, etc. However, UMDATrack can maintain high-quality target state prediction even in these extreme scenarios.

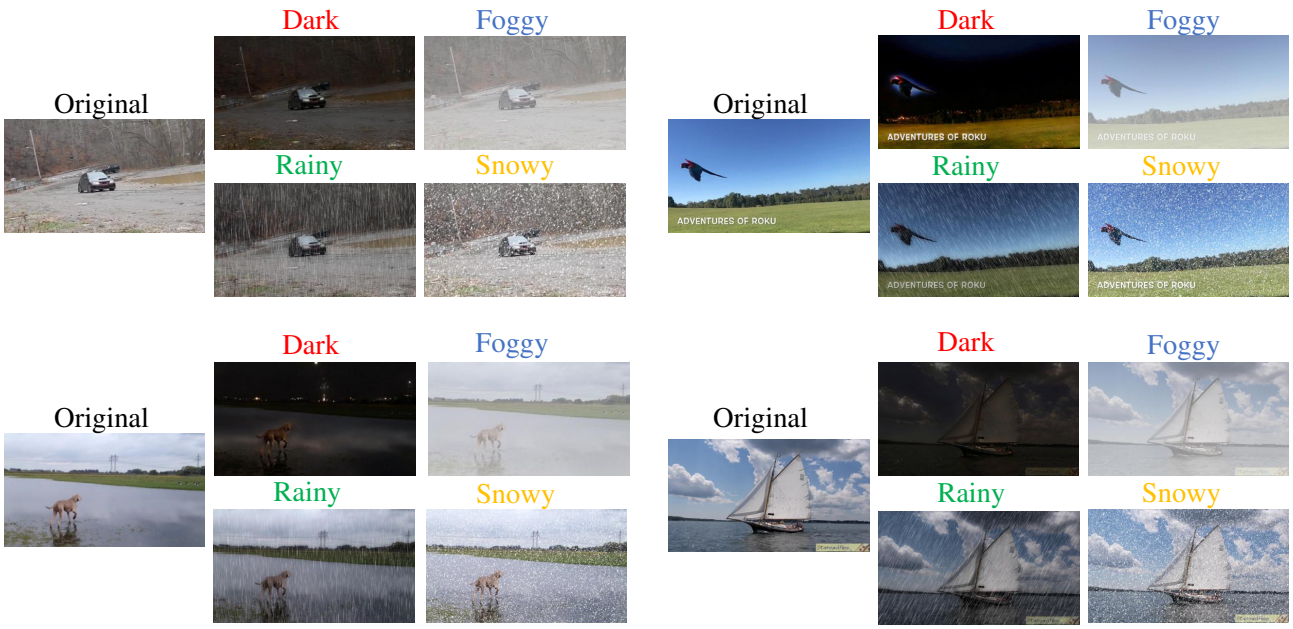

Figure 1. Visualization of the synthesized video frames under adverse weather conditions, *i.e.* dark, foggy, rainy and snowy scenarios. Please zoom in for details.

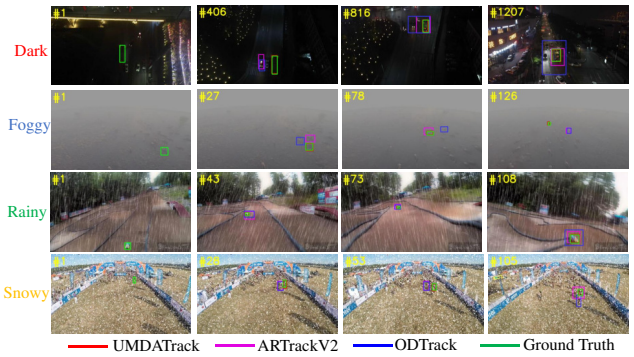

Figure 2. Comparative tracking results of the proposed UMDATrack and other state-of-the-art trackers.

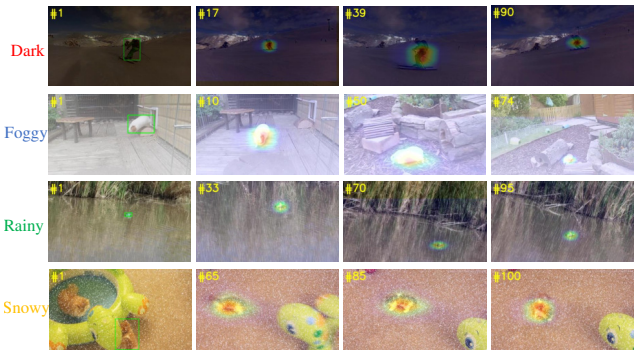

Figure 3. Visualization of the tracking heatmaps predicted by UMDATrack in adverse weather conditions.

E.2. Heatmap Visualization in Adverse Conditions

We present the visualization results of tracking heatmaps in Fig. 3. In extreme dark scenarios, our tracker can accurately locate the target object, even when the target is barely visible to the human eyes. In the foggy scene, although there are significant appearance variation and poor visibility, UMDATrack still computes correct prediction outputs. Other scenarios like rain and snow also show that UMDATrack can generate high-quality response heatmaps across multiple video frames in challenging scenarios, even when visibility is severely limited or the target is obscured by environmental factors.

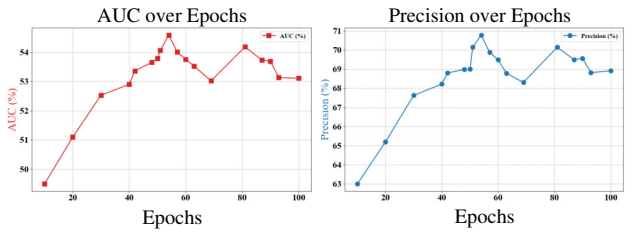

Figure 4. The convergence speed of DCA. Please zoom in for details.

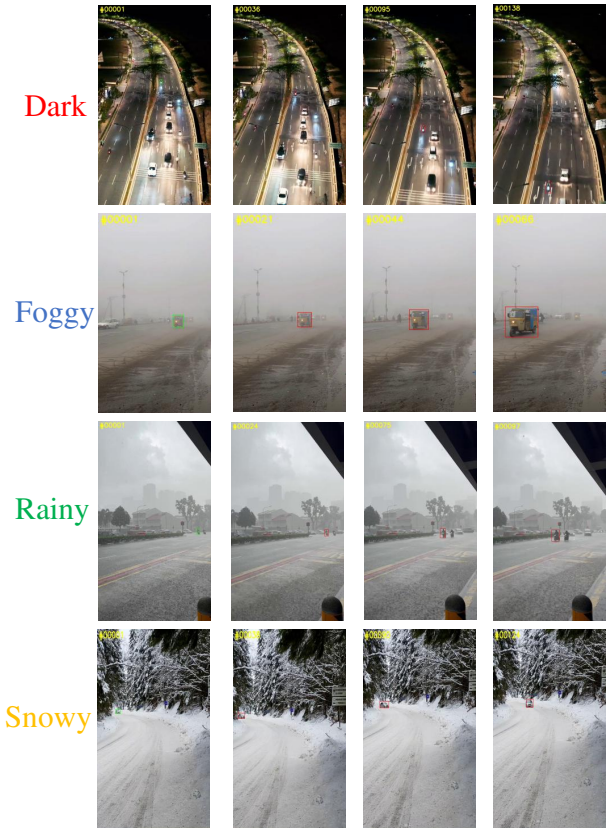

Figure 5. Visualized tracking results of UMDATrack in real-world scenarios, including nighttime, foggy, rainy and snowy videos.

E.3. Study on the speed of DCA convergence

We analyze the convergence speed in which the DCA achieves its optimal performance during training. As shown in Fig. 4, around 50 epochs, the DCA has already obtained encouraging performance. Beyond this point, performance increases only slightly, and may even decline with additional epochs. Therefore, we suggest a trade-off between performance and training time to achieve efficiency.

E.4. Real-World Test

To further validate the tracking performance of our UMDATrack in real adverse scenarios, we test it on the videos collected from real-world imaging systems. The results are shown in Fig. 5, we can see that the real-world tests confirm the effectiveness of UMDATrack. Note that our UMDATrack is not limited to the aforementioned extreme scenarios, it can be rapidly trained using a small partition of synthesized dataset under the guidance of weather-specific text prompts. Our UMDATrack leads new SOTA performance for object tracking under adverse weather conditions.

References

[1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19048–19057, 2024.

[2] 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.

[3] 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.