Tracking Reflected Objects: A Benchmark

Xiaoyu Guo[†], Pengzhi Zhong[†], Lizhi Lin, Hao Zhang, Ling Huang, and Shuiwang Li[⋆]

Guilin University of Technology , China , 541006
{guoxiaoyu, zhongpengzhi}@glut.edu.cn, {linlizhi0204, zhanghao_0618,
huangling240312, lishuiwang0721}@163.com

Abstract. Visual tracking has made great progress recently, thanks to large-scale training datasets. These datasets have enabled the development of highly accurate tracking algorithms. However, most research focuses on tracking general objects, with less attention on specialized challenges like tracking reflected objects. Reflections can distort objects' appearances, making them harder to track, which is a critical issue in areas like autonomous driving, security, smart homes, and industrial settings where tracking reflections in mirrors or glass is important. To fill this gap, we introduce TRO, a benchmark designed for Tracking Reflected Objects. TRO contains 200 sequences with about 70,000 frames, all annotated with bounding boxes, aiming to foster the development of methods tailored for this challenging task. We tested 20 state-of-the-art trackers and found they struggle with the complexities of reflections. To set a better baseline, we propose HIP-HaTrack, a new tracker that uses hierarchical features and outperforms existing algorithms. We hope our benchmark and HIP-HaTrack will inspire further research on tracking reflected objects. The TRO and code are available at https://github.com/Guoxiaoyuy/HIP-HaTrack.

Keywords: Reflected object tracking · Benchmark

1 Introduction

The field of visual tracking has experienced significant advancements over the past decade, largely due to the increasing availability of large-scale datasets and the development of sophisticated algorithms [82,105,56,20,65,31,103,108]. These advancements have enabled tracking systems to achieve remarkable accuracy and robustness in a variety of applications, such as autonomous driving, surveillance, and smart home technologies [12,65]. With improved algorithms and richer datasets, visual tracking systems can now better handle occlusions, variations in object appearance, and environmental changes, which are common challenges in real-world scenarios [82,105,56,58,57,20,89]. Despite these achievements, much of the research in visual tracking has been concentrated on generic objects, leaving more specialized and challenging scenarios relatively underexplored [106,97,36].

[⋆] Corresponding author. [†]These authors contributed equally.

One particularly challenging scenario that has not received adequate attention is the tracking of reflected objects. Reflections can severely distort the appearance of objects, creating ambiguous visual cues that complicate the tracking process. The reflective properties of surfaces like mirrors and glass can lead to multiple, overlapping images of the same object, making it difficult for algorithms to discern the true object from its reflection [40,90,6,33,79,73]. Accurately tracking objects in reflections—whether in mirrors, glass surfaces, or other reflective materials—is essential for improving monitoring systems in various applications like surveillance, smart home, medical procedures, and traffic.
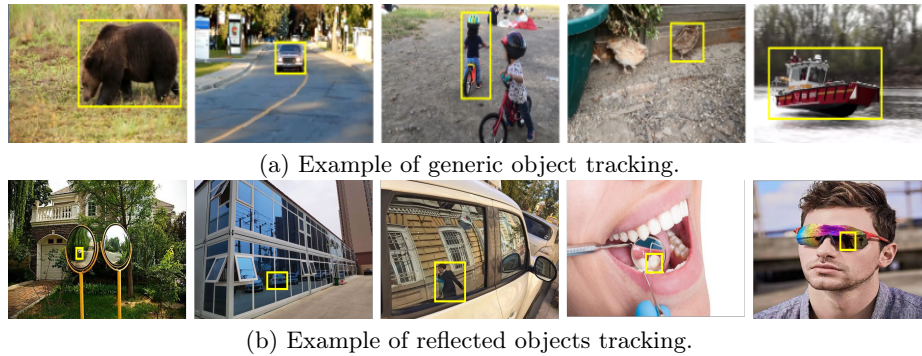

(a) Example of generic object tracking.

(b) Example of reflected objects tracking.

Fig. 1. Generic object tracking (a) and reflected objects tracking (b). Compared with generic object tracking, tracking of reflected objects is more challenging as their tendency to distort the appearance of objects and create ambiguous visual cues.

In surveillance, surveillance systems often face challenges due to obstructions or blind spots that prevent full coverage. By leveraging mirror reflections, "virtual" cameras can be introduced to capture images through reflections, filling these blind spots and providing more comprehensive monitoring [80,81,76]. In smart homes, reflective surfaces such as mirrors, windows, and polished furniture are ubiquitous. Correctly tracking objects reflected in these surfaces can significantly enhance the functionality and reliability of automated systems [38,28,5,48,43]. In smart home, home automation systems with accurate tracking can enhance user experience and safety. By distinguishing between reflections of humans and objects, these systems can make better decisions about turning on lights, adjusting temperature, or locking doors, ensuring they respond to actual human presence [88,86,16]. Smart home assistants can also use reflections to improve awareness, such as detecting a person entering a room through a mirror's reflection and offering a personalized greeting or adjusting the environment to the user's preferences [87,64,86,16]. In the medical field, mirror reflections can provide crucial assistance in tracking surgical instruments and monitoring patients' vital signs [7,75,85,42]. During surgeries, reflections can offer additional angles and perspectives, allowing medical staff to have a comprehensive view of the operating area. This enhanced visibility can improve precision

and safety, ensuring that instruments are used correctly and efficiently [8,7,77]. Similarly, monitoring vital signs through reflections can offer real-time feedback, helping medical professionals respond promptly to any changes in a patient's condition. These extra perspectives are vital for the accuracy and effectiveness of medical procedures, contributing to better patient outcomes and enhanced surgical performance [8,41,1]. In traffic monitoring, strategically placed mirrors at streetlights or intersections can significantly enhance surveillance capabilities [68,3]. These mirrors capture reflected images of vehicles and pedestrians, effectively expanding the range of monitored areas. By using reflections, traffic monitoring systems can gain visibility into spots that are typically out of the direct line of sight of cameras, such as blind corners or areas obscured by other objects [23,78]. This expanded field of view allows for more comprehensive coverage, ensuring that all traffic movements are accurately tracked. The improved detection accuracy helps in better traffic management, accident prevention, and enforcement of traffic laws [24,67,104,50]. Enhanced monitoring through reflections can also aid in incident response, providing valuable information about traffic conditions and potential hazards in real time [60,74].

Given these diverse applications and the critical nature of accurate tracking in these scenarios, it is clear that the challenges posed by reflected objects require dedicated research efforts. Current tracking algorithms, designed primarily for generic objects, often lack the robustness needed to handle these complexities effectively. In light of these challenges, there is a pressing need for the development of specialized benchmarks and algorithms tailored to the tracking of reflected objects. In this work, we introduce the Tracking Reflected Objects (TRO) benchmark, a dedicated dataset designed to stimulate research in this challenging area. TRO comprises 200 sequences, totaling approximately 70,000 frames, each meticulously annotated with bounding boxes. Each sequence undergoes semi-automatic object annotation, possessing various attributes for performance evaluation and analysis. This comprehensive dataset captures a wide range of real-world scenarios involving reflections, providing a robust foundation for developing and evaluating new tracking algorithms tailored to these conditions. In addition, we present a detailed evaluation of 20 state-of-the-art tracking algorithms using the TRO benchmark. Our findings reveal that current methods often struggle with the complexities introduced by reflections, indicating a significant need for improvement. By releasing TRO, we aim to foster advancements in tracking technology that address the unique challenges of reflected objects, ultimately enhancing the accuracy and reliability of visual tracking systems across various applications. The main contributions of this paper are as follows:

– We introduce TRO, which is the first benchmark dedicated to track reflected objects to our knowledge. TRO, containing 200 sequences with meticulous annotations, poses unique challenges not adequately covered by existing benchmarks and may inspire further research and applications in tracking reflected objects.
– We conduct a thorough evaluation of 20 state-of-the-art trackers on TRO. This evaluation contributes to our understanding of the limitations of current

trackers, establishes benchmarks for performance comparison and inspires future research endeavors in the field of tracking reflected objects.
– We propose HIP-HaTrack, a novel tracker that aggregate hierarchical features to enhance performance. HIP-HaTrack surpasses existing state-of-the-art algorithms in TRO, providing a stronger baseline for future research in tracking reflected objects.

2   Related Works

2.1   Visual Tracking Algorithms

There are generally two categories of mainstream visual trackers: DCF-based and DL-based trackers [61,66,94,63,91]. DCF-based trackers treat visual tracking as an online regression problem. Thanks to the utilization of the Parseval theorem and Fast Fourier Transform (FFT), DCF-based trackers demonstrate high speed on CPUs [53]. They initially began with the MOSSE (Minimum Output Sum of Squared Error) filter [9], and since then, significant progress has been made. For example, [55,21] employed an additional scale filter to handle target scale variations. [22,51] explored regularization techniques to enhance robustness, and [52] extended the DCF framework to achieve translation equivariance. However, the limited expressive power of manually crafted features makes it difficult for DCF-based trackers to maintain robustness under challenging conditions. As a result, DL-based trackers have gained significant attention for their remarkable capability to automatically learn features through neural networks. SiamFC [4], a typical representative of DL-based trackers, pioneered the utilization of the Siamese networks [19] in visual tracking and has spurred the development of various sophisticated trackers. For instance, those mentioned in [49,35,99,14,34,17] have demonstrated significant advancements in tracking precision and robustness.

Recently, Vision Transformers (ViTs) [25] have shown significant potential in streamlining and unifying frameworks for visual tracking, as evident in studies like[25,59,101,93]. Based on ViTs, OStrack [101] combines feature extraction and relation modeling, introducing an early elimination module to implement a one-stream tracking framework. DropTrack [93] improves temporal correspondence learning by adaptively performing spatial attention dropout during frame reconstruction. Very recently, HIPTrack [10] offers a historical prompt network constructed with refined historical foreground masks and target visual features to provide comprehensive and precise guidance for the tracker. However, most current research focuses on tracking generic objects, with less attention given to more specialized and challenging scenarios, particularly the tracking of reflected objects. One of the main reasons for this is the lack of dedicated benchmark tests specifically designed for tracking reflective objects.

2.2   Visual Tracking Benchmarks

The development of visual tracking technology relies on accurate measurement and evaluation of tracking method performance, with benchmark datasets serv-

ing as crucial standards for evaluation. Currently, there are two main types of benchmark datasets: general benchmarks and specific benchmarks [27].

General benchmarks. General benchmark datasets are primarily used to evaluate tracking algorithm performance in general scenarios. OTB-2013 [95], a representative dataset, initially contained 50 sequences, later expanded to OTB-2015 [96], covering 100 sequences. VOT [47] organization hosts a series of tracking competitions with up to 60 sequences. NFS [30] focuses on high frame rate videos. Many large-scale benchmarks have been proposed to train DL-based trackers. TrackingNet [70] aimed at providing resources for training and evaluating general object tracking algorithms,featuring over 30,000 sequences. GOT-10k [39] contains over 10,000 video sequences, with each sequence containing rich tracking challenges. LaSOT [26] initially contained 1,400 long-term video sequences and later added 150 additional sequences in subsequent studies, introducing a new evaluation scheme focusing on handling situations where the object becomes invisible during tracking.

Specific benchmarks. Specific benchmark datasets are designed to evaluate the tracking of particular objects. For example, the UAV123 [69] focusing on low-altitude drone object tracking with 123 sequences captured by drone. VOT-TIR [46] utilizes both RGB and thermal infrared images simultaneously to improve object tracking performance in RGB-T sequences. CDTB [62] and PTB [83] aim to evaluate tracking performance in RGB-D videos, where D represents depth images. TOTB [27] focuses on transparent object tracking, collecting 225 videos from 15 transparent object categories. Recently, some new benchmarks have emerged, such as TSFMO [106], specifically objecting tracking small objects and fast-moving objects.

However, existing benchmark datasets primarily focus on tracking physical objects. With the increasing use of reflected objects in fields such as industrial automation, security monitoring, and autonomous driving, the need for specialized tracking algorithms is growing. Reflected objects present unique challenges, including reflection, distortion, and ambiguity, which often cause traditional object tracking algorithms to perform poorly. Additionally, there is a notable lack of dedicated datasets for tracking reflected objects. This paper aims to address this by proposing a novel Reflected Object Tracking Dataset, designed to facilitate the development and evaluation of algorithms specifically for reflected object tracking.

3   Construction of the TRO Dataset

3.1   Data Collection

We collected hundreds of candidate videos through the internet and by recording with a camera. After that, we ranked and filtered the videos based on four criteria of tracking difficulty and some additional challenging cases. Vide os can gain higher rankings through the following ways: 1) sufficient relative motion of the target, 2) high variability of the environment, 3) the target crossing frame borders, and 4) sufficient footage length. In addition to the above criteria, videos
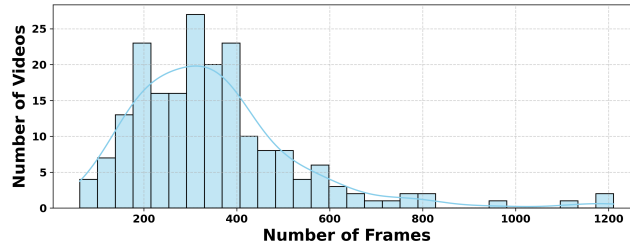
Fig. 2. This figure presents a histogram of the video frame counts in the TRO. The x-axis represents the number of frames per video, divided into bins, and the y-axis shows the number of videos within each bin.

Table 1. Summary of statistics of the proposed TRO.

Number of videos 200	Min frames 62	Max frames 1211
Total frames 69810	Avg frames 349	Frame range 1149

with additional challenges are given higher priority. For example, distinguishing targets from other highly similar objects is a challenge in object detection and tracking. After filtering, videos are further selected and sampled into sequences with a frame number threshold (≤ 1300). Relatively stationary frames were manually discarded. Due to resource constraints in collecting, processing, and annotating video data, we chose 200 sequences for the TRO benchmark. This ensures each video is high quality and thoroughly annotated, while staying within our available time, budget, and manpower. The targets mainly include humans (such as riders, pedestrians and dancers), animals (such as dogs, cats and tigers), and rigid objects (such as cars, airplanes and boats). Our benchmark covers a wide range of targets with high diversity. Table 1 summarizes the TRO dataset, and Fig. 2 presents a histogram of the video frame counts in the TRO. Additionally, to highlight the challenges associated with tracking reflected objects, we evaluate seven state-of-the-art (SOTA) trackers on TRO and two popular generic object tracking benchmarks for comparison. The AUC are presented in Table 2 Remarkably, these tackers' AUC on TrackingNet is more than 15.0% higher than TRO, highlighting the significant challenges posed by tracking reflected object.

Table 2. AUC (%) Comparison of state-of-the-art trackers on the TRO and generic object tracking benchmarks.

Dataset	HIPTrack	ROMTrack	DropTrack	ARTrack	SeqTrack	GRM	SimTrack
TRO	**68.9**	**67.5**	**67.1**	**62.1**	**66.6**	**63.5**	**61.5**
TrackingNet [71]	84.5	84.1	84.1	85.1	83.9	84.0	83.4
LaSOT [26]	72.7	71.4	71.8	72.6	71.5	69.9	70.5

Fig. 3. Some examples of box annotations for TRO.Each sequence is annotated with axis-aligned bounding boxes and attributes.

3.2   Data Annotation

We followed the principles outlined in [72] for sequence annotation. Each video frame was meticulously labeled by a team of expert annotators (i.e., students working on visual tracking). Annotators drew/edited axis-aligned bounding boxes around the visible portion of the object in each frame, representing the tightest bounding box. Objects were labeled if visible, otherwise marked as absent, out of view, or occluded. To improve annotation efficiency and quality, we employed an efficient method called point annotation proposed in [44]. Objects were annotated in each frame using Segment Anything [44], a versatile segmentation method that automatically identifies objects or regions in images and generates annotation information. And than, we conducted visual inspection of the automatically generated annotations from Segment Anything. For cases of poor quality, we performed manual annotation by reviewing and correcting errors or omissions. In addition, a double-check mechanism was implemented where a second annotator reviewed the initial annotations to maintain accuracy and consistency. Any discrepancies were resolved through discussion and consensus. This combination of automatic and manual annotation along with the double-check mechanism resulted in a high-quality annotated dataset, which serves as a reliable foundation for subsequent research. Our annotation methods and strategies ensure both efficiency and accuracy. Examples of box annotations in TRO can be seen in Fig. 3.

3.3   Attributes

In line with the approach of popular tracking benchmarks [110],[39], and [96], we further annotated each video sequence with multiple attributes to better analyze the performance of tracking algorithms. The proposed TRO dataset features 12 attributes, including (1) Illumination Variation (IV), (2) Scale Variation (SV),
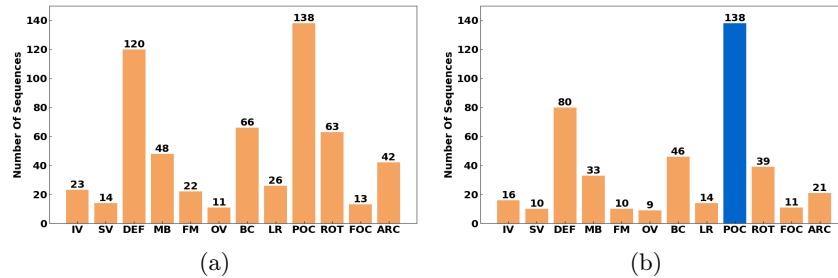
Fig. 4. (a) Attribute distribution of the entire testset, and (b) the distribution of the sequences with Partial Occlusion (POC) attribute.

(3) Deformation (DEF), (4) Motion Blur (MB), (5) Fast Motion (FM), which is assigned when the target center moves by at least 20% of its size in last frame. (6) Out-of-View (OV), (7) Background Clutter (BC), (8) Low Resolution (LR), assigned when the object region is less than 900 pixels, (9) Partial Occlusion (POC), (10) Rotation (ROT), (11) Full Occlusion (FOC), and (12) Aspect Ratio Change (ARC), which assigned when the aspect ratio of the bounding box exceeds the range [0.5, 2]. By considering these attributes, we can gain a more comprehensive understanding of the performance of tracking algorithms under different challenging scenarios.

The distribution of attributes in our dataset is shown in Fig. 4(a). Some attributes occur more frequently, such as DEF (Deformation) and POC (Partial Occlusion). It also indicates that a sequence typically has multiple attribute annotations. In addition to summarizing the performance of the entire testset, we can evaluate tracking methods on a subset corresponding to a particular attribute to report performance on specific challenging conditions. For example, the POC subset contains 138 sequences, which can be used to analyze the performance of trackers in handling occlusions. The attribute distribution in the POC subset is shown in Fig. 4(b), which indicates that the challenge associated with POC frequently co-occurs with other challenges rather than presenting alone. This overlapping of attributes highlights the complex nature of real-world tracking scenarios and underscores the need for robust algorithms capable of handling multiple simultaneous challenges. Such insights are crucial for developing and evaluating advanced tracking methods that can perform effectively in diverse and complicated conditions.

4   A New Baseline: HIP-HaTrack

During our evaluation, we observed that although HIPTrack outperforms other state-of-the-art trackers on TRO, but its performance remains significantly below satisfactory levels. Therefore, based on HIPTrack, we propose a novel baseline tracker called HIP-HaTrack based on aggregating hierarchical features to enhance the ability to address the challenges posed by significant ambiguous vi-
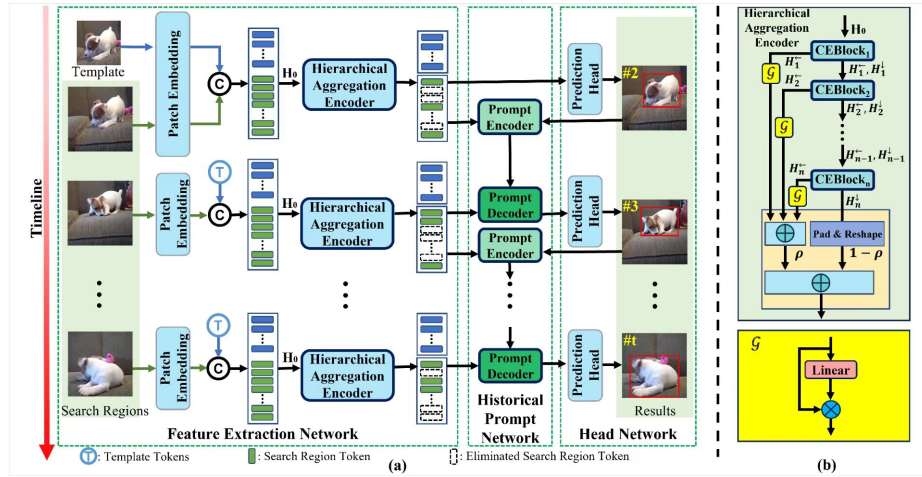
Fig. 5. (a) The overall framework of HIP-HaTrack,which inherited from HIPTrack. The difference lies in the encoder part of the feature extraction network. (b) The structure of the Hierarchical Aggregation Encoder.

sual cues and distortion of object appearance caused by reflections. Reflections and ambiguous visual cues can cause certain features to become less reliable. As multi-level features capture information at various scales and resolutions, aggregating these features enables the tracker to rely on a broader set of data, thereby increasing the likelihood of correctly identifying the target. In addition, aggregating hierarchical features allows the tracker to adaptively utilize the most relevant features for a given scenario.

4.1   Overall Architecture

Like HIPTrack, HIP-HaTrack also comprises three main components, as shown in Fig. 5(a). The feature extraction network is responsible for extracting features from the search region that contain target-template matching information, while filtering out background image patches. The historical prompt network contains a historical prompt encoder and a historical prompt decoder. The historical prompt encoder needs to encode the position information and visual features of the target in the current frame as historical target features, and store them in the memory bank of the historical prompt decoder. And the historical prompt decoder generates historical prompt for each search region and combines these historical prompt with the compressed features of the search region. The head network has a same structure to OSTrack [101]. The difference between HIP-HaTrack and HIPTrack lies in their feature extraction network. HIPTrack utilizes a Vision Transformer (ViT) with early candidate elimination modules, similar to the one used in OSTrack, serving as its feature encoder. Our HIP-HaTrack introduces a novel Hierarchical Aggregation Encoder, as illustrated in Fig. 5(b). This

Hierarchical Aggregation Encoder is designed to integrate features at multiple levels of abstraction. By aggregating hierarchical features, HIP-HaTrack aims to enhance the feature representations, particularly for tracking reflected objects. In the following section, we will provide a detailed description of the proposed . Other architectures inherit from HIPTrack; please refer to [10] for details.

4.2   Hierarchical Aggregation Encoder

Our Hierarchical Aggregation Encoder builds upon the Transformer Encoder architecture of HIPTrack, namely a Vision Transformer (ViT) with early candidate elimination modules. This structure is represented with stacked CEBlocks as shown in Fig. 5(b). The key difference is the proposed aggregation structure for aggregating hierarchical features output by the ViT blocks. Each hierarchical feature is adaptively weighted with a gate module represented by $\mathcal{G}$. The weighted hierarchical features are summed to obtain an aggregated feature, which is finally weighted summed with the feature produced by the original Transformer Encoder. Considering that the early candidate elimination module in each CEBlock may alter the shape of the input feature, we divide the CEBlock into two parts: one representing the ViT block with early candidate elimination, denoted by $CBlock^{\downarrow}i$, and the other representing the ViT block without it, denoted by $CBlock^{\leftarrow}i$. The output features produced by these two parts are represented by $H^{\downarrow}i$ and $H^{\leftarrow}i$, respectively. The effect of the proposed Hierarchical Aggregation Encoder can be formally formulated by

$$F_{ha} = (1 - \rho)\mathcal{P}(H_n^{\downarrow}) + \rho \sum_{i=1}^{n} \mathcal{G}(H_i^{\leftarrow}) \cdot H_i^{\leftarrow}, \tag{1}$$

where $H_i^{\leftarrow} = CBlock_i^{\leftarrow}(H_{i-1}^{\leftarrow})$, $H_i^{\downarrow} = CBlock_i^{\downarrow}(H_{i-1}^{\downarrow})$, $\mathcal{P}$ denotes the padding and reshaping used to recover the shape of $H_n^{\downarrow}$ to that of H_0, $\rho \in [0, 1]$ is a weighting constant used to balance the importance of aggregated feature and $H_n^{\downarrow}$. Note that if $\rho = 0$ $F_h a$ is just the feature produced by the Transformer Encoder of HIPTrack. Thus, our Hierarchical Aggregation Encoder extends and generalizes the feature extraction network of HIPTrack. It enhances feature representation by capturing information at multiple levels of abstraction, thereby improving the richness and comprehensiveness of the features. This approach also provides flexibility for further fine-tuning or extension, making it more effective for tracking objects affected by reflections. Since the proposed Hierarchical Aggregation Encoder does not introduce any additional training loss, we can utilize the same training pipeline as HIPTrack for training HIP-HaTrack. Please refer to HIPTrack [10] for details.

5   Evaluation

In this work, precision and success rate are used as metrics for evaluation.

Precision Plot: The precision plot is a common evaluation metric in tracking studies. It measures tracking precision by calculating the average Euclidean distance between the center of the tracked object and its ground truth position. This gives an average error over all frames in a sequence. However, if the tracker loses the object, the error may become random, making the average less reliable. To address this, the precision plot compares the tracker's estimated position with the ground truth within a set distance threshold. For our evaluation, we use the precision score with a 20-pixel threshold, following standard practices [2].

Success Plot: Another evaluation metric is the success plot, based on the overlap of bounding boxes. Given the tracking bounding box r_t and the ground truth bounding box r_a, the overlap score is defined as $S = \frac{|r_t \bigcap r_a|}{|r_t \bigcup r_a|}$, where $\cap$ and $\cup$ denote the intersection and union of two regions, respectively, and $|\cdot|$ represents the number of pixels in the region. To evaluate tracking performance over multiple frames, we count how many frames have an overlap score S greater than a threshold. The success plot shows the ratio of successful frames at different thresholds from 0 to 1. Since using a single success rate at a specific threshold (e.g. t_o=0.5) may not give a complete picture, the area under the curve (AUC) is typically used to rank tracking algorithms more fairly [2,37].

5.1   Trackers for Comparison

We evaluate our HIP-HATrack and 20 state-of-the-art trackers on TRO, including HIPTrack [10], ROMTrack [11], SeqTrack [18], DropTrack [93], AQATrack [98], SparseTT [29], DCPT [109], GRM [32], Stark [100], ZoomTrack [45], MAT [107], SimTrack [17], ARTrack [92], CSWinTT [84], SiamAPN++ [14], TCTrack [15], SiamGAT [34], UDAT [102], HiFT [13], and Aba-ViTrack [54].

5.2   Evaluation Results

Overall Performance: A comprehensive evaluation of 20 state-of-the-art trackers and our HIP-HATrack was conducted on the TRO dataset. Notably, the existing trackers were used without any modifications for this evaluation. The results are presented with precision and success plots, as displayed in Fig. 6. As can be seen, HIP-HATrack outperforms all other trackers with a precision (PRC) of 0.738, surpassing its baseline HIPTrack, secuing the second place, with a PRC of 0.723 by 1.5%. In terms of AUC, HIP-HATrack also leads with the highest AUC of 0.697, surpassing the second place HIPTrack, which has an AUC of 0.689, by 0.8%. These results demonstrate the superior performance of HIP-HATrack in both precision and success rate metrics, highlighting its effectiveness compared to existing state-of-the-art trackers in handling the challenge associated with tracking reflected objects by aggregating hierarchical features.

Attribute-based performance: We conducted a comprehensive performance evaluation on twelve common attributes to better analyze and understand the performance of different trackers on the TRO dataset. Our HIP-HaTrack achieved the best AUC and PRC for most attributes. Due to the constraints of paper
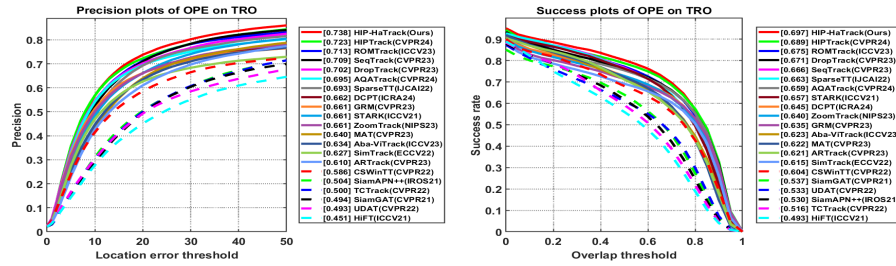
Fig. 6. Overall performance on TRO. Precision and success rate for one-pass evaluation (OPE) [95] are used for evaluation.

length, we present the PRC results for six common challenges in Fig. 7, including background clutter (BC), rotation (ROT), deformation (DEF), illumination variation (IV), partial occlusion (POC), and aspect ratio change (ARC). As can be seen, HIP-HaTrack consistently secures the first place across all attribute subsets, with the baseline HIPTrack taking second place, except for the BC subset where SparseTT ranks second instead. HIP-HaTrack outperforms the second-place trackers by at least 0.7% on all subsets. Notably, the improvements on the BC and ROT subsets are both 3.2%. Specifically, HIP-HaTrack achieves a precision of 0.628 on the BC subset and 0.633 on the ROT subset, surpassing SparseTT's 0.596 and HIPTrack's 0.617 by 3.2%. For the DEF attribute, HIP-HaTrack achieves a precision of 0.732, surpassing HIPTrack's 0.722 by 1.0%. On the IV, POC, and ARC subsets, HIP-HaTrack shows improvements over HIP-Track by 1.6%, 0.7%, and 1.0%, respectively. These results validate the effectiveness of the proposed method of aggregating hierarchical features for tracking reflective objects. Notably, on the BC and IV subsets, all trackers have precision below 0.650, much lower than the other subsets, which are all above 0.70. This can be attributed to the impact of illumination variation on reflections and the degradation of discriminative cues for reflected objects in cluttered backgrounds, emphasizing the need for specialized techniques to handle reflections.

Qualitative evaluation: In Fig. 8, we show some qualitative tracking results of our method in comparison with nine top trackers, including HIPTrack [10], DropTrack [93], AQATrack [98], ROMTrack [11], SeqTrack [18], SparseTT [29], DCPT [109], Stark [100], and ZoomTrack [45]. These examples are influenced by challenges such as illumination variation, deformation, rotation, background clutter, aspect ratio change, and partial occlusion. Specifically, in Car6, only HIP-HaTrack, HIPTrack [10], DropTrack [93], ROMTrack [11], and DCPT [109] successfully track the reflected toy car. In Chicken3, many trackers are influenced by the presence of the real chicken, leading to inaccurate tracking of the reflected one. In Dog1 and Human112, while several trackers manage to track parts of the reflected object, HIP-HaTrack consistently shows visually superior results. These qualitative comparisons support the effectiveness and superiority of our method of aggregating hierarchical features for tracking reflected objects.
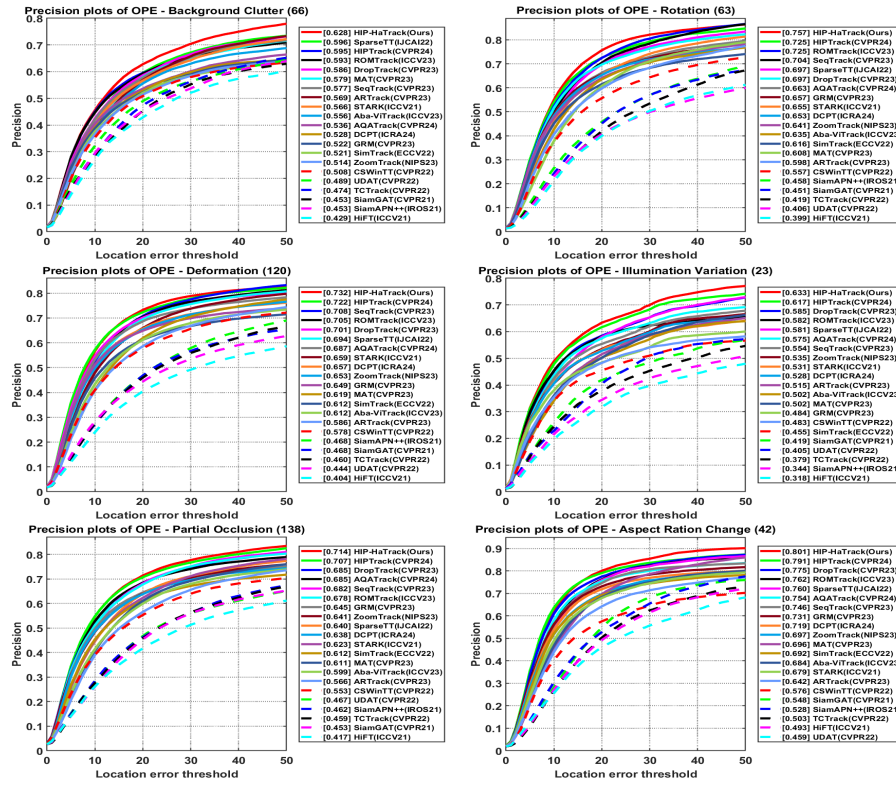
Fig. 7. Attribute-based comparison on illumination variation, Deformation, Rotation, Background Clutter, Aspect Ration Change and Partial Occlusion.

5.3   Ablation Study

Impact of the importance coefficient of aggregated hierarchical feature. To study the impact of the weight coefficient ρ on balancing the contributions of aggregated hierarchical feature and $H_n^{\downarrow}$ in HIP-HaTrack, we evaluated HIP-HaTrack on the TRO dataset with various ρ settings. Specifically, ρ ranges from 0.1 to 1.0 with step of size 0.1. It is important to note that as ρ increases, the contribution of the aggregated feature increases, while the contribution of $H_n^{\downarrow}$ decreases. Table 3 shows HIP-HaTrack's PRC and AUC on the TRO dataset under different ρ values. The results indicate that PRC and AUC reach their highest values when ρ is between 0.3 and 0.7, highlighting the significance of aggregated hierarchical features for improving tracking performance. Specifically, when $\rho = 0.3$, HIP-HaTrack achieved the best PRC and AUC, with values of 0.738 and 0.697, respectively. Therefore, we used $\rho=0.3$ as the default setting for HIP-HaTrack.
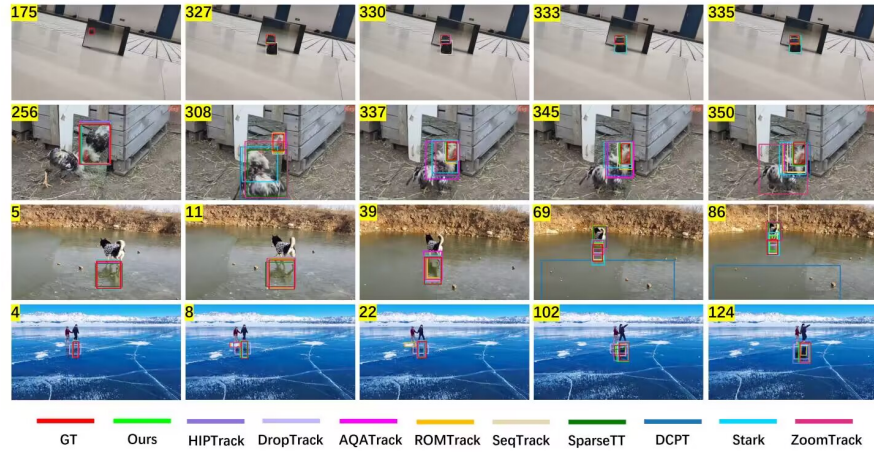
Fig. 8. Qualitative evaluation on 4 sequences from TRO, i.e. Car6,Chicken3,Dog1, and Human112 from top to bottom. The results of different methods have been shown with different colors, and 'GT' denotes the ground truth.

Table 3. Illustrates the impact of the importance coefficient of aggregated feature on HIP-HaTrack's PRC (%) and AUC (%) on TRO. Red, blue, and green represent the first, second, and third place, respectively.

ρ	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
PRC (%)	68.6	68.6	69.1	70.6	**72.3**	71.0	**71.8**	**73.8**	71.2	71.1
AUC (%)	67.9	68.0	68.4	**69.3**	**69.0**	68.1	67.9	**69.7**	69.0	68.9

6   Conclusion

In this work, we investigate a relatively unexplored tracking task, namely, tracking reflected objects. Reflections can significantly distort the appearance of objects, creating ambiguous visual cues that complicate the tracking process. This issue is particularly pertinent in applications where accurately tracking objects reflected in surfaces like mirrors or glass is crucial. However, there hasn't been public benchmark dedicated to tracking reflected objects. To address this gap, we introduce TRO, which is the first benchmark specifically designed for reflected object tracking. Additionally, to assess the performance of existing trackers and establish a baseline for future comparisons, we extensively evaluate 20 SOTA tracking algorithms with in-depth analysis. Furthermore, we propose a novel tracker named HiP-HaTrack, which aggregates hierarchical features to achieve better feature representations, significantly outperforming existing SOTA algorithms.

References

1. Abi-Rafeh, J., Zammit, D., Jaberi, M.M., Al-Halabi, B., Thibaudeau, S.: Nonbiological microsurgery simulators in plastic surgery training: a systematic review. Plastic and Reconstructive Surgery **144**(3), 496e–507e (2019) 3
2. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE transactions on pattern analysis and machine intelligence **33**(8), 1619–1632 (2010) 11
3. Baysal, D.: Assessment of vehicular vision obstruction due to driver-side b-pillar and remediation with blind spot eliminator. arXiv preprint arXiv:2302.07088 (2023) 3
4. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14. pp. 850–865. Springer (2016) 4
5. Bianchi, I., Bertamini, M., Savardi, U.: Differences between predictions of how a reflection behaves based on the behaviour of an object, and how an object behaves based on the behaviour of its reflection. Acta Psychologica **161**, 54–63 (2015) 2
6. Bianchi, I., Savardi, U.: The relationship perceived between the real body and the mirror image. Perception **37**(5), 666–687 (2008) 2
7. Bichlmeier, C., Heining, S.M., Feuerstein, M., Navab, N.: The virtual mirror: a new interaction paradigm for augmented reality environments. IEEE Transactions on Medical Imaging **28**(9), 1498–1510 (2009) 2, 3
8. Bichlmeier, C., Sielhorst, T., Navab, N.: The tangible virtual mirror: New visualization paradigm for navigated surgery. In: International Workshop on Augmented Reality Environments for Medical Imaging and Computer-Aided Surgery, Copenhagen, Denmark (2006) 3
9. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 2544–2550. IEEE (2010) 4
10. Cai, W., Liu, Q., Wang, Y.: Hiptrack: Visual tracking with historical prompts (2024) 4, 10, 11, 12
11. Cai, Y., Liu, J., Tang, J., Wu, G.: Robust object modeling for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9589–9600 (2023) 11, 12
12. Cannons, K.: A review of visual tracking. Dept. Comput. Sci. Eng., York Univ., Toronto, Canada, Tech. Rep. CSE-2008-07 **242** (2008) 1
13. Cao, Z., Fu, C., Ye, J., Li, B., Li, Y.: Hift: Hierarchical feature transformer for aerial tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15457–15466 (2021) 11
14. Cao, Z., Fu, C., Ye, J., Li, B., Li, Y.: Siamapn++: Siamese attentional aggregation network for real-time uav tracking. In: 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 3086–3092. IEEE (2021) 4, 11
15. Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C.: Tctrack: Temporal contexts for aerial tracking. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14778–14788 (2022) 11
16. Chakraborty, A., Islam, M., Shahriyar, F., Islam, S., Zaman, H.U., Hasan, M.: Smart home system: a comprehensive review. Journal of Electrical and Computer Engineering **2023**(1), 7616683 (2023) 2

17. Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W., Ouyang, W.: Backbone is all your need: A simplified architecture for visual object tracking. In: European Conference on Computer Vision. pp. 375–392. Springer (2022) 4, 11

18. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: Sequence to sequence learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14572–14581 (2023) 11, 12

19. Chicco, D.: Siamese neural networks: An overview. Artificial neural networks pp. 73–94 (2021) 4

20. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7183–7192 (2020) 1

21. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Discriminative scale space tracking. IEEE transactions on pattern analysis and machine intelligence **39**(8), 1561–1575 (2016) 4

22. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE international conference on computer vision. pp. 4310–4318 (2015) 4

23. Dhalwar, S., Ruby, S., Salgar, S., Padiri, B.: Image processing based traffic convex mirror detection. In: 2019 Fifth International Conference on Image Information Processing (ICIIP). pp. 41–45. IEEE (2019) 3

24. Díaz, J., Ros, E., Mota, S., Botella, G., Cañas, A., Sabatini, S.: Optical flow for cars overtaking monitor: the rear mirror blind spot problem. Ecovision (European research project) (2003) 3

25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021) 4

26. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019) 5, 6

27. Fan, H., Miththanthaya, H.A., Harshit, Rajan, S.R., Liu, X., Zou, Z., Lin, Y., Ling, H.: Transparent object tracking benchmark. arXiv: Computer Vision and Pattern Recognition (2020) 5

28. Freysteinson, W.M.: Assessing the mirrors in long-term care homes: A preliminary survey. Journal of Gerontological Nursing **36**(1), 34–40 (2010) 2

29. Fu, Z., Fu, Z., Liu, Q., Cai, W., Wang, Y.: Sparsett: Visual tracking with sparse transformers. arXiv preprint arXiv:2205.03776 (2022) 11, 12

30. Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 1134–1143 (2017) 5

31. Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J.: Aiatrack: Attention in attention for transformer visual tracking. In: European Conference on Computer Vision. pp. 146–164. Springer (2022) 1

32. Gao, S., Zhou, C., Zhang, J.: Generalized relation modeling for transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18686–18695 (2023) 11

33. Gregory, E., McCloskey, M.: Mirror-image confusions: Implications for representation and processing of object orientation. Cognition **116**(1), 110–129 (2010) 2

34. Guo, D., Shao, Y., Cui, Y., Wang, Z., Zhang, L., Shen, C.: Graph attention tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9543–9552 (2021) 4, 11
35. Guo, D., Wang, J., Cui, Y., Wang, Z., Chen, S.: Siamcar: Siamese fully convolutional classification and regression for visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6269–6277 (2020) 4
36. Guo, X., Zhong, P., Zhang, H., Huang, L., Huang, D., Li, S.: Camouflaged object tracking a benchmark. ArXiv **abs/2408.13877** (2024), https://api.semanticscholar.org/CorpusID:271957361 1
37. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12. pp. 702–715. Springer (2012) 11
38. Hossain, M.A., Atrey, P.K., El Saddik, A.: Smart mirror for ambient home environment. In: 2007 3rd IET International Conference on Intelligent Environments. pp. 589–596. IET (2007) 2
39. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 1562–1577 (2021) 5, 7
40. Ittelson, W.H., Mowafy, L., Magid, D.F.: The perception of mirror-reflected objects. Perception **20**, 567 – 584 (1991) 2
41. Jansen, J., Dubois, L., Schreurs, R., Gooris, P.J., Maal, T.J., Beenen, L.F., Becking, A.G.: Should virtual mirroring be used in the preoperative planning of an orbital reconstruction? Journal of Oral and Maxillofacial Surgery **76**(2), 380–387 (2018) 3
42. JASMINE, J.L., JINU, M.A., DHANALAKSHMI, T., RAKSHAMBIGAI, B., MADHUMITHA, S.: Medical mirror for health care. In: National Conference on Recent Advancements in Communication. vol. 7 (2020) 2
43. Kilic, D., Sailaja, N.: User-centred repair: From current practices to future design. In: International Conference on Human-Computer Interaction. pp. 52–71. Springer (2024) 2
44. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 7
45. Kou, Y., Gao, J., Li, B., Wang, G., Hu, W., Wang, Y., Li, L.: Zoomtrack: Target-aware non-uniform resizing for efficient visual tracking (2023) 11, 12
46. Kristan, M., Leonardis, A., Matas, J., et al.: The visual object tracking vot2017 challenge results. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) pp. 1949–1972 (2017) 5
47. Kristan, M., Matas, J., Leonardis, A., Vojír, T., Pflugfelder, R.P., Fernandez, G.J., Nebehay, G., Porikli, F.M., Cehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**, 2137–2155 (2016) 5
48. Latif, M.S.A., Ismail, A.A., Zariman, A.: Smart mirror for home automation (2020) 2
49. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4282–4291 (2019) 4

50. Li, D., Hagura, H., Miyabashira, T., Kawai, Y., Ono, S.: Traffic mirror detection and annotation methods from street images of open data for preventing accidents at intersections by alert. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3256–3262. IEEE (2023) 3

51. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H.: Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4904–4913 (2018) 4

52. Li, S.w., Jiang, Q.b., Zhao, Q.j., Lu, L., Feng, Z.l.: Asymmetric discriminative correlation filters for visual tracking. Frontiers of Information Technology & Electronic Engineering **21**(10), 1467–1484 (2020) 4

53. Li, S., Liu, Y., Zhao, Q., Feng, Z.: Learning residue-aware correlation filters and refining scale estimates with the grabcut for real-time uav tracking. In: 2021 International Conference on 3D Vision (3DV). pp. 1238–1248. IEEE (2021) 4

54. Li, S., Yang, Y., Zeng, D., Wang, X.: Adaptive and background-aware vision transformer for real-time uav tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13989–14000 (2023) 11

55. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13. pp. 254–265. Springer (2015) 4

56. Lijia, W., Binbin, W., Xufeng, C.: An adaptive kernel based correlation filter algorithm for real time object tracking. In: 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). vol. 2, pp. 144–148. IEEE (2020) 1

57. Liu, D., Cui, Y., Tan, W., Chen, Y.: Sg-net: Spatial granularity network for one-stage video instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9816–9825 (2021) 1

58. Liu, T., Liu, Y.: Moving camera-based object tracking using adaptive ground plane estimation and constrained multiple kernels. Journal of Advanced Transportation **2021**(1), 8153474 (2021) 1

59. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4

60. Loizides, O.S., Kastrinakis, M., Badawy, G., Smadi, M.N., Murray, D., Koutsakis, P.: Efficient policing for screen mirroring traffic. Human-centric Computing and Information Sciences **8**, 1–14 (2018) 3

61. Lu, H., Wang, D.: Online visual tracking. Springer (2019) 4

62. Lukezic, A., Kart, U., Kapyla, J., Durmush, A., Kamarainen, J.K., Matas, J., Kristan, M.: Cdtb: A color and depth visual object tracking dataset and benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10013–10022 (2019) 5

63. Ma, S., Liu, Y., Zeng, D., Liao, Y., Xu, X., Li, S.: Learning disentangled representation in pruning for real-time uav tracking. In: Asian Conference on Machine Learning. pp. 690–705. PMLR (2023) 4

64. Marikyan, D., Papagiannidis, S., Alamanos, E.: A systematic review of the smart home literature: A user perspective. Technological Forecasting and Social Change **138**, 139–154 (2019) 2

65. Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S.: Deep learning for visual tracking: A comprehensive survey. IEEE Transactions on Intelligent Transportation Systems **23**(5), 3943–3968 (2022) 1

66. Mazzeo, P.L., Ramakrishnan, S., Spagnolo, P.: Visual Object Tracking with Deep Neural Networks. BoD–Books on Demand (2019) 4
67. Mittal, D., Verma, V., Rastogi, R.: A comparative study and new model for smart mirror. International Journal of Scientific Research in Research Paper, Computer Science and Engineering **5**(6), 58–61 (2017) 3
68. Mole, C.D., Wilkie, R.M.: Looking forward to safer hgvs: The impact of mirrors on driver reaction times. Accident Analysis & Prevention **107**, 173–185 (2017) 3
69. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 445–461. Springer (2016) 5
70. Müller, M., Bibi, A., Giancola, S., Al-Subaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV (2018) 5
71. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). pp. 300–317 (2018) 6
72. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). pp. 300–317 (2018) 7
73. Nightingale, S.J., Wade, K.A., Farid, H., Watson, D.G.: Can people detect errors in shadows and reflections? Attention, Perception & Psychophysics **81**, 2917 – 2943 (2019) 2
74. Noriaki, I., Shintaro, O., Yoshihiro, S., Kazuya, O., Grewe, R.: Collision risk prediction utilizing road safety mirrors at blind intersections. In: 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration. No. 23-0164 (2023) 3
75. Poh, M.Z., McDuff, D., Picard, R.: A medical mirror for non-contact health monitoring. In: ACM SIGGRAPH 2011 Emerging Technologies, pp. 1–1 (2011) 2
76. Ray, B.R., Aalsma, M., Zaller, N.D., Comartin, E.B., Sightes, E.: The perpetual blind spot in public health surveillance. Journal of correctional health care : the official journal of the National Commission on Correctional Health Care (2022) 2
77. Roy, A., Mukhopadhyay, M.: Situs inversus totalis: operating on the mirror image. Hellenic Journal of Surgery **92**, 150–152 (2020) 3
78. Sadrhaghighi, S., Dolati, M., Ghaderi, M., Khonsari, A.: Monitoring openflow virtual networks via coordinated switch-based traffic mirroring. IEEE Transactions on Network and Service Management **19**(3), 2219–2237 (2022) 3
79. Sareen, P., Ehinger, K.A., Wolfe, J.M.: Through the looking-glass: Objects in the mirror are less real. Psychonomic Bulletin & Review **22**, 980–986 (2015) 2
80. Shen, Y., Yan, W.Q.: Blind spot monitoring using deep learning. In: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–5. IEEE (2018) 2
81. Singhal, C., Barick, S.: Ecms: Energy-efficient collaborative multi-uav surveillance system for inaccessible regions. IEEE Access **10**, 95876–95891 (2022) 2
82. SM, J.R., Augasta, G.: Review of recent advances in visual tracking techniques. Multimedia tools and applications **80**, 24185–24203 (2021) 1
83. Song, S., Xiao, J.: Tracking revisited using rgbd camera: Unified benchmark and baselines. 2013 IEEE International Conference on Computer Vision pp. 233–240 (2013) 5
84. Song, Z., Yu, J., Chen, Y.P.P., Yang, W.: Transformer tracking with cyclic shifting window attention (2022) 11

85. Soppimath, V.M., Hudedmani, M.G., Chitale, M., Altaf, M., Doddamani, A., Joshi, D.: The smart medical mirror-a review. International Journal of Advanced Science and Engineering **6**(1), 1244–1250 (2019) 2

86. Sovacool, B.K., Del Rio, D.D.F.: Smart home technologies in europe: A critical review of concepts, benefits, risks and policies. Renewable and sustainable energy reviews **120**, 109663 (2020) 2

87. Stojkoska, B.L.R., Trivodaliev, K.V.: A review of internet of things for smart home: Challenges and solutions. Journal of cleaner production **140**, 1454–1464 (2017) 2

88. Saad al sumaiti, A., Ahmed, M.H., Salama, M.M.: Smart home activities: A literature review. Electric Power Components and Systems **42**(3-4), 294–305 (2014) 2

89. Tang, F., Ling, Q.: Ranking-based siamese visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8741–8750 (2022) 1

90. Turnbull, O.H., Mccarthy, R.A.: Failure to discriminate between mirror-image objects: A case of viewpoint-independent object recognition? Neurocase **2**(1), 63–72 (1996) 2

91. Wang, X., Yang, X., Ye, H., Li, S.: Learning disentangled representation with mutual information maximization for real-time uav tracking. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1331–1336. IEEE (2023) 4

92. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9697–9706 (2023) 11

93. Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B.: Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14561–14571 (2023) 4, 11, 12

94. Wu, W., Zhong, P., Li, S.: Fisher pruning for real-time uav tracking. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2022) 4

95. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2411–2418 (2013) 5, 12

96. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015) 5, 7

97. Wu, Y., Wang, Y., Liao, Y., Wu, F., Ye, H., Li, S.: Tracking transforming objects: A benchmark. ArXiv **abs/2404.18143** (2024) 1

98. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporaltransformers (2024) 11, 12

99. Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12549–12556 (2020) 4

100. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10448–10457 (2021) 11, 12

101. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV (2022) 4, 9

102. Ye, J., Fu, C., Zheng, G., Paudel, D.P., Chen, G.: Unsupervised domain adaptation for nighttime aerial tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8896–8905 (2022) 11

103. Yuan, D., Chang, X., Liu, Q., Yang, Y., Wang, D., Shu, M., He, Z., Shi, G.: Active learning for deep visual tracking. IEEE Transactions on Neural Networks and Learning Systems (2023) 1

104. Zhang, C., Steinhauser, F., Hinz, G., Knoll, A.: Traffic mirror-aware pomdp behavior planning for autonomous urban driving. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 323–330. IEEE (2022) 3

105. Zhang, J., Liu, Z., Lin, Y.: Correlation gaussian particle filter for robust visual tracking. In: 2020 Chinese Control And Decision Conference (CCDC). pp. 4854–4857. IEEE (2020) 1

106. Zhang, Z., Wu, F., Qiu, Y., Liang, J., Li, S.: Tracking small and fast moving objects: A benchmark. In: Proceedings of the Asian Conference on Computer Vision. pp. 4514–4530 (2022) 1, 5

107. Zhao, H., Wang, D., Lu, H.: Representation learning for visual object tracking by masked appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18696–18705 (2023) 11

108. Zhu, H., Xue, M., Wang, Y., Yuan, G., Li, X.: Fast visual tracking with siamese oriented region proposal network. IEEE signal processing letters **29**, 1437–1441 (2022) 1

109. Zhu, J., Tang, H., Cheng, Z.Q., He, J.Y., Luo, B., Qiu, S., Li, S., Lu, H.: Dcpt: Darkness clue-prompted tracking in nighttime uavs (2024) 11, 12

110. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7380–7399 (2021) 7