IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 10, 2021, accepted November 28, 2021, date of publication December 1, 2021,
date of current version December 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132077

Long-Term Person Tracking for Unmanned Aerial Vehicle Based on Human-Machine Collaboration

TONGTONG ZHOU AND YADONG LIU
College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China
Corresponding author: Yadong Liu (liuyadong@nudt.edu.cn)

This work was supported by the Joint Funds of the National Natural Science Foundation of China under Grant U19A2083.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the School Ethics Committee, and performed in line with the Declaration of Helsinki.

ABSTRACT Unmanned Aerial Vehicle (UAV) has been widely used in military reconnaissance, smart transportation, public security and other fields. UAV-based person tracking is attracting incremental attention for its wide application requirements. Currently, some state-of-the-art visual tracking methods have achieved promising performance in common scenarios. However, in the scene of UAV-based person tracking, there will be long-term target disappearance and unpredictable dramatic target appearance changes, which still pose a huge challenge to UAV-based person tracking. In this work, a human-machine hybrid augmented tracking system based on eye tracking is proposed to cope with the challenge. During tracking, through the interaction between humans and machines, humans can provide real-time guidance and corrections to the tracker, and the tracker can also learn interesting targets from humans to enhance itself. The experimental results show that human-in-the-loop can remarkable improve the success rate and robustness of the tracking and our tracking system outperforms the state-of-the-art tracker in complex environments.

INDEX TERMS UAV-based person tracking, human-in-the-loop, gaze-based human computer interaction, human-machine collaboration, long-term tracking.

I. INTRODUCTION

Unmanned aerial vehicle (UAV) has been widely used in military reconnaissance, smart transportation, public security and other fields., which has the advantages of high concealment, small size, and strong maneuverability. UAV-based person tracking, which greatly expands the application of UAV, is attracting incremental attention. Most studies of visual tracking are based on the assumption that the tracking target changes smoothly and does not disappear for a long time [1]–[4]. However, due to the complex working environment of UAV and the uncertainty of person targets, these assumptions may be broken in UAV-based person tracking. We summarize the challenges of UAV-based person tracking as fast motion, severe occlusion, background clutter, long-term disappearance, and dramatic changes in appearance. Among them, long-term disappearance and dramatic changes in appearance are rarely studied and most trackers can't handle such situations. Figure 1 shows several typical failure cases of some state-of-the-art(SOTA) trackers when the target disappears for a long time and the appearance changes dramatically.

In this paper, we propose a human-in-the-loop tracking framework to handle the challenges of UAV-based person tracking by combining visual attention and context-aware functions of the human visual system with SOTA methods in computer vision. On the one hand, human visual focus is captured by the eye tracker as human predictions of the target position based on empirical knowledge and intuitive reasoning. On the other hand, humans can act as the highest priority decision makers to intervene or correct the computer's tracking results. In the interaction between humans and computers, our method can accurately retrieve the target that has disappeared for a long time and consistently locate the target whose appearance has changed dramatically. Moreover, human demonstrations are also constantly improving the automatic tracking capabilities of computers. Specifically, the main contributions include:

1. A novel human-in-the-loop tracking framework is proposed (Figure2). In this framework, humans will act as an

The associate editor coordinating the review of this manuscript and approving it for publication was Waleed M. Alsabhan.

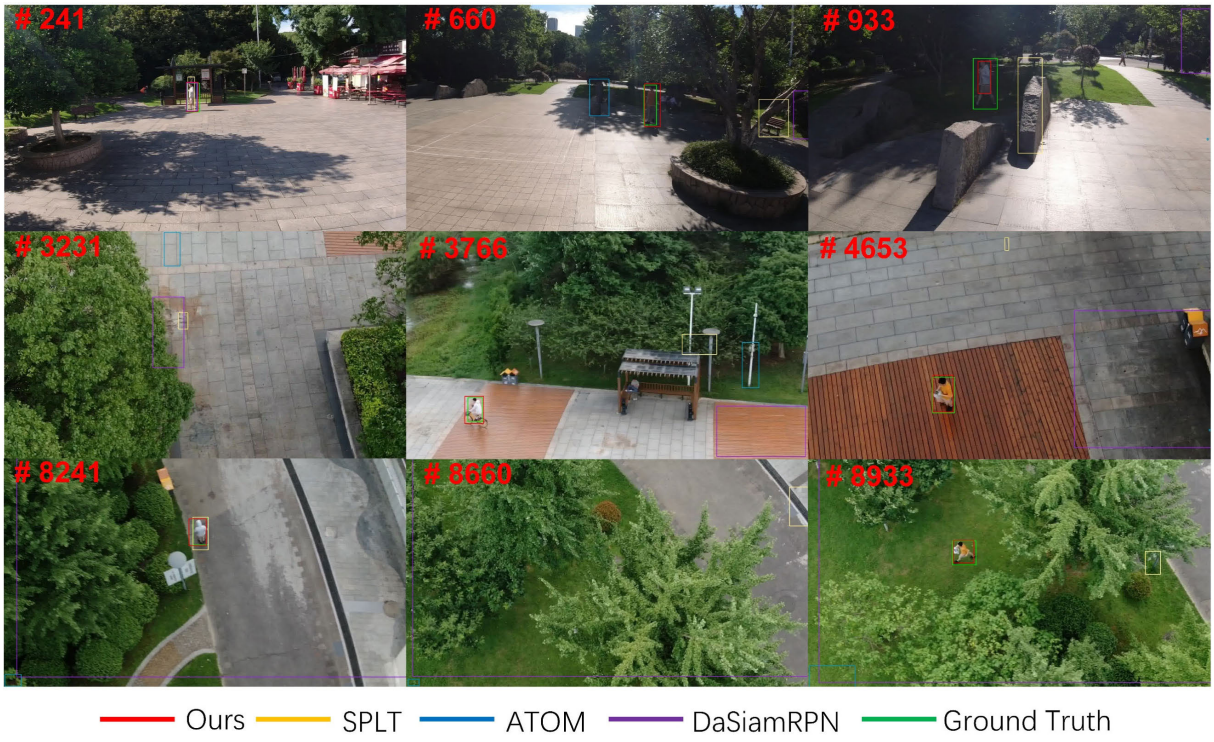

FIGURE 1. A comparison of our approach with state-of-the-art trackers. While a person is walking through the woods, there are long-term disappearance caused by occlusion and dramatic changes in appearance due to illumination variations and changing clothes. SPLT [5], ATOM [6], DaSiamRPN [7] cannot find the target and keep tracking when the target reappears. Our approach successfully handles these challenges and gets robust and accurate tracking results.

instructor and the highest priority decision maker to influence the entire tracking process. The framework contains a local tracking module, a global search module, and a human attention analysis module. As far as we know, this is the first long-term tracking framework based on human-machine collaboration.

2. A novel gaze-based interaction paradigm with computer tracking system is designed in the human attention analysis module. According to this interactive method, human visual attention can be captured naturally. Moreover, the tracking result can be corrected in time by specific eye gestures. The learning cost of this paradigm is very low, and humans only need to keep their attention on the target when they want to intervene in the tracking process. Through this paradigm, the natural advantages of the human visual system in tracking tasks can be fully utilized.

3. A high-quality manually labeled test dataset for UAV-based person tracking is established. Compared with the existing datasets, such as OTB100 [8], VOT2018-LT [9], UAV123 [10], it contains a series of longer image sequences and more complex challenges of tracking, especially the long-term disappearance of the target and the dramatic appearance changes of the target.

II. RELATED WORK

Most trackers and datasets in visual object tracking focus on the short-term tracking problem, which implicitly assumes that the target will not disappear for a long time. Although numerous trackers achieved promising performance in short-term tracking [2], [3], [6], [11], [12], there is still a big gap with application. In order to make the tracker better meet the requirements of the application, more and more researchers begin to pay attention to long-term tracking, which requires the tracker to recognize the disappearance of target and retrieve the target. TLD [13] was the first tracker to re-detect lost targets. It proposed the local tracking and global search paradigm, which is still widely used in long-term tracking recently. MBMD [14] used an offline training regression network to return the bounding box of the target in a local area directly and reused the local regression network in a sliding window to retrieve the lost target. MBMD won the first long-term tracking champion in VOT2018 [15]. SPLT [5] used the SiamRPN [12] as the local tracker and designed a long-term tracking framework called ''Skimming-Perusal''. The framework proposed a lightweight ''skimming'' module to reduce the number of sliding windows, which significantly improved the efficiency of the global search module. LTMU [16] used a deep network tracking algorithm that can be learned online as the local tracker and proposed a meta-update module to predict the update reliability from spatial-temporal multi-cue information which effectively reduced the cumulative noise. LTMU won the VOT2019 [17] and VOT2020 [18] long-term tracking champions. Although these long-term trackers have the ability

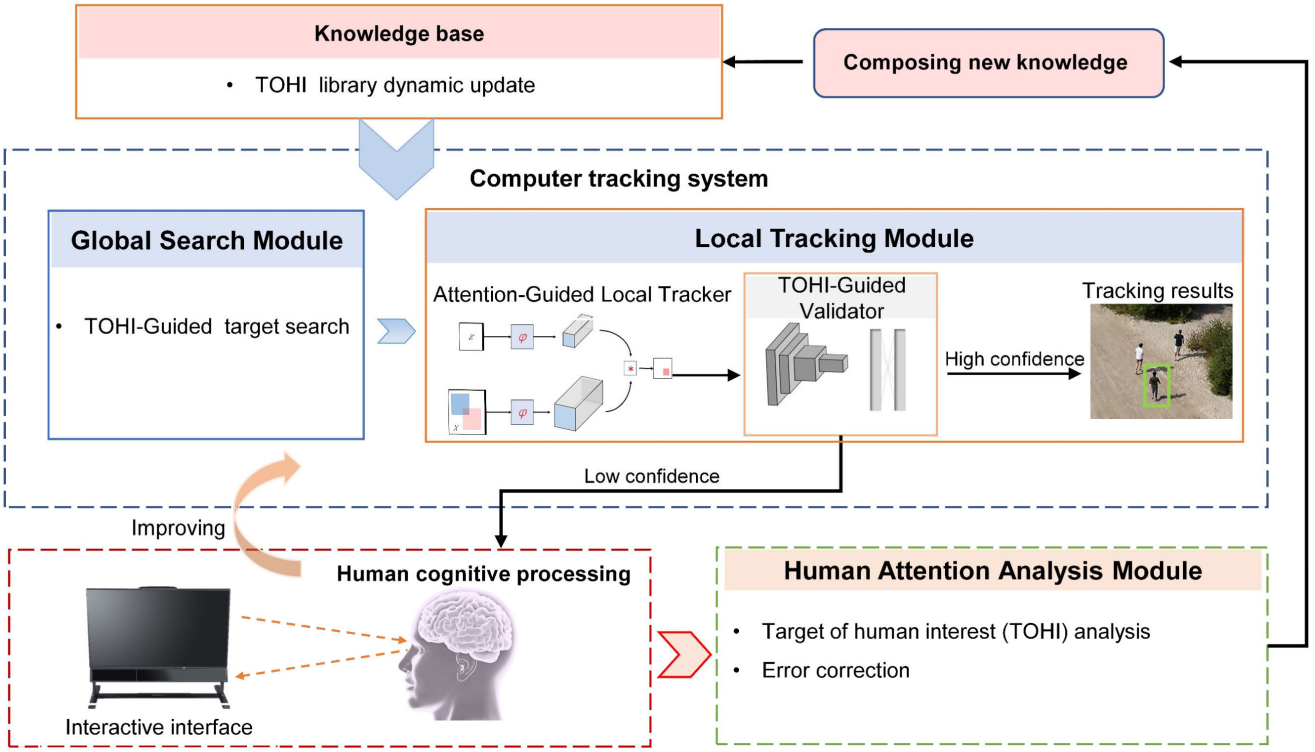

FIGURE 2. Human-in-the-loop tracking framework proposed in this work. The framework includes a local tracking module, a global search module, and an asynchronous human attention analysis module. Humans can guide and correct the tracking in the loop, while the machine will recognize target-of-human-interest (TOHI) by capturing the human visual focus and verify the tracking results by the guidance of TOHI.

to search for targets again, the robustness of these methods is still not enough, particularly when the target appearance changes dramatically, it is hard for these methods to find and track the target again. (Figure 1 and Figure 11 show some typical tracking failure scenarios of SPLT as SOTA long-term tracker.)

Compared with the above long-term tracking methods, the method proposed in this work uses a similar strategy of local tracking combined with global search However, what is new is that our method can accept human guidance at each step of tracking to obtain additional prior knowledge based on human experience and learn the latest target appearance feature from human demonstration, so as to better cope with the changing complex environment.

III. HUMAN-IN-THE-LOOP TRACKING FRAMEWORK

A Human-in-the-loop tracking framework is proposed to address the challenges of UAV-based person tracking, especially the long-term disappearance and the dramatic changes in appearance. The overall framework is shown in Figure 2. The whole frame-work can be divided into three modules: local tracking module, global search module and human attention analysis module. Local tracking module is used for robust tracking in the local search area. Global search module is used to retrieve the target after the target disappears.

Human attention analysis module is used to analyze eye movements to extract target-of-human-interest (TOHI) as well as support human intervention in the framework. In addition, a TOHI-guided validator is embedded to verify the tracking results and the candidate results of the global search.

The tracking workflow is shown in Figure 3. When the tracking starts, the local tracking module first searches for the target in the local search area, and the results are sent to the TOHI guided validator. The validator queries the TOHI library to measure the similarity between the current result and the TOHI. The result with high similarity is considered reliable. If the reliability of the current result exceeds the preset threshold, the local tracking result will be output and used as initial position in the next frame. If the reliability of the current result is below the preset threshold, the local tracking result is considered unreliable, and the tracker will open a window of human intervention. During the intervention window, humans can intervene according to two paradigms described in the human attention analysis module. During the window, if humans intervene, the target area of human visual focus will be output as the tracking result. It is worth noting that if humans do not intervene but are still within the intervention window, the current results will be output, and the local tracking will continue in the next frame. The purpose is to establish a fault-tolerant mechanism for the local tracking results and allow the local tracker to cope with occlusion and short-term disappear. If humans do not intervene and the intervention window is over, the local tracking result will be considered unreliable and judged as target lost in the local area.

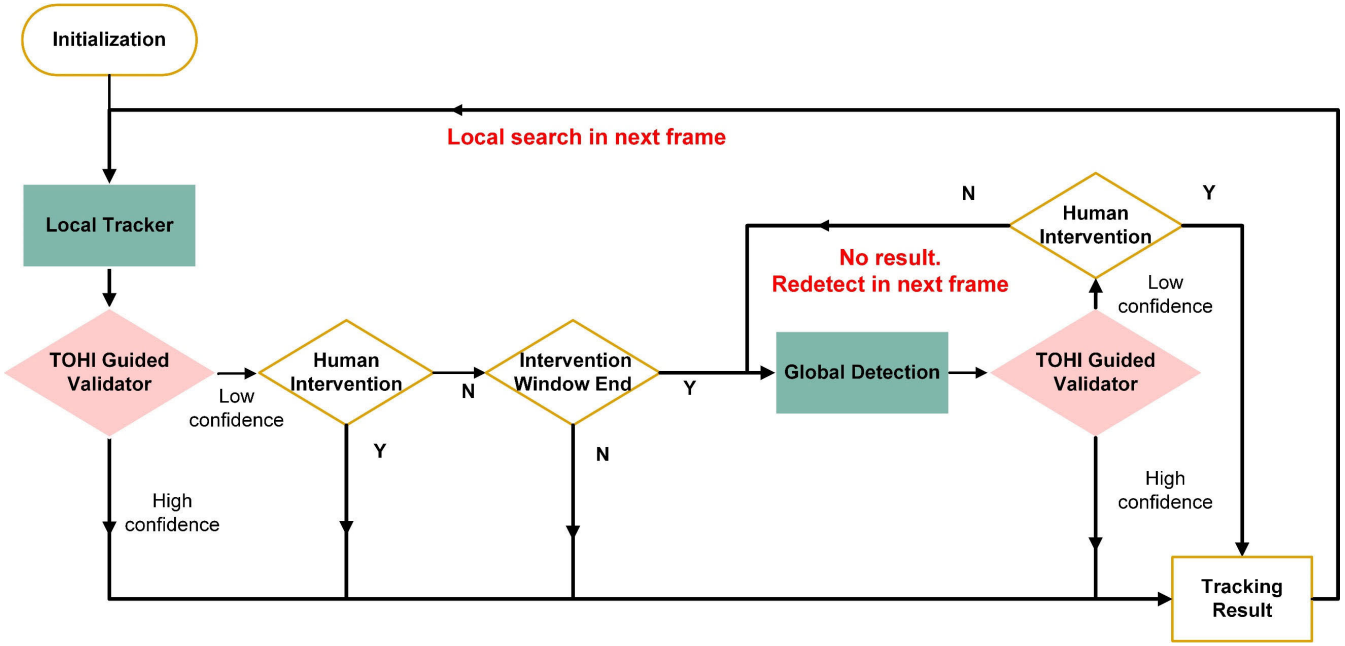

FIGURE 3. Human-in-the-loop tracking system workflow. After tracking initialization, local tracking will be performed first. If local tracking fails and humans do not intervene, global search will start. Local tracking will not continue until human intervention or the target is found globally.

When the target is lost in a local area, global search module will detect people in the whole image and send the detected people targets to TOHI-guided validator for verification one by one. If the candidate human target with the highest score exceeds the preset threshold, it will be output as the result of global search module. Otherwise, the window of human intervention will be open again. If the human intervenes, the target area of human visual focus will be output as the tracking result. If there is no intervention, there will be no tracking result in this frame, and the global search will continue in the next frame.

In this framework, humans not only guide or correct the tracking results in the loop, but also indirectly improve the tracking ability through TOHI. The machine recognizes the TOHI by analyzing human fixation point (details in human attention analysis module) and the validator works under the guidance of these TOHI (details in TOHI-guided validator). With the continuous addition of new TOHI to the TOHI library, the TOHI-guided validator can identify people whose appearance has changed drastically, thus enhancing the discriminant ability of the tracker. The detailed description of each part is as follows.

A. LOCAL TRACKING MODULE

Local tracking module uses the offline trained SiamRPN [12] model, which is an efficient tracking algorithm that introduces region proposal network [19] into SiamFC [11]. The basic principle of SiamRPN is that the model learns a similarity function to judge the similarity between the target template and the candidate bounding box in the search area. The accuracy and efficiency of SiamRPN are sufficiently balanced, which allows our tracking framework to track as accurately as possible while meeting real-time requirements at the same time.

The specific human-attention-guided SiamRPN we proposed in this work is shown in Figure 4. The input of the model is the target template obtained by initialization and the surrounding area of the tracking result. SiamRPN will output the similarity score between the template and the candidate bounding box in the local area. However, the discrimination ability of SiamRPN model is not accurate enough when there is background clutter [7]. As shown in Figure 5a, when the tracked person is blocked by trees and passing pedestrians, SiamRPN cannot make effective judgments based on the similarity score, which leads to abnormal tracking results. Therefore, we introduce the Euclidean distance between the candidate bounding boxes and the user's fixation point to evaluate the attention score of each candidate bounding box. Specifically, an evaluation method is designed to make the candidate bounding boxes closer to the visual focus get a higher attention score. The final score $\boldsymbol{Z_i}$ can be calculated as equation 1:

$$\boldsymbol{Z}_i = \boldsymbol{S}_i + \boldsymbol{\xi} R^{Dst(b_i,g)} \boldsymbol{Dst}\,(b_i, g)\,, \quad (0 < \boldsymbol{\xi} < 1) \qquad (1)$$

where $\boldsymbol{S_i}$ is the similarity score corresponding to the b_i generated by SiamRPN model. $\boldsymbol{Dst}\,(\boldsymbol{b_i}, \boldsymbol{g})$ is the Euclidean distance between the center of the candidate bounding box b_i and the fixation point $\boldsymbol{g}$. R represents the sensitivity value, and a small value corresponds to a higher sensitivity. The specific setting of R should be determined according to the experimental environment setting and the type of tracking target. In our experiment, R is set to 0.8 and $\boldsymbol{\xi}$ is set to 0.3. The final score of each candidate bounding box is expressed as the weighted sum of the similarity score and the attention score, so the

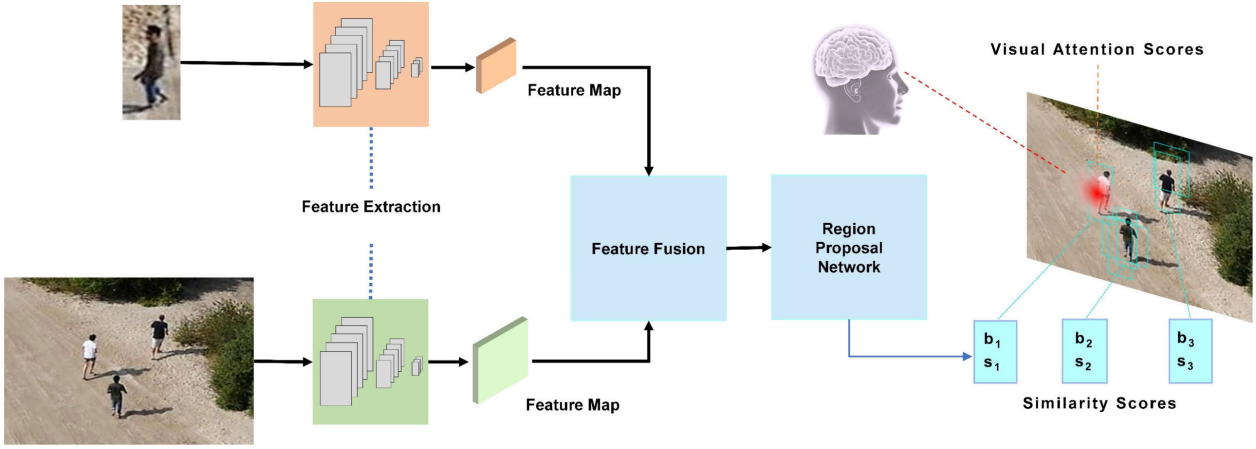

FIGURE 4. Human-attention-guided SiamRPN. The human attention score and the similarity score are fused to determine the final candidate box.

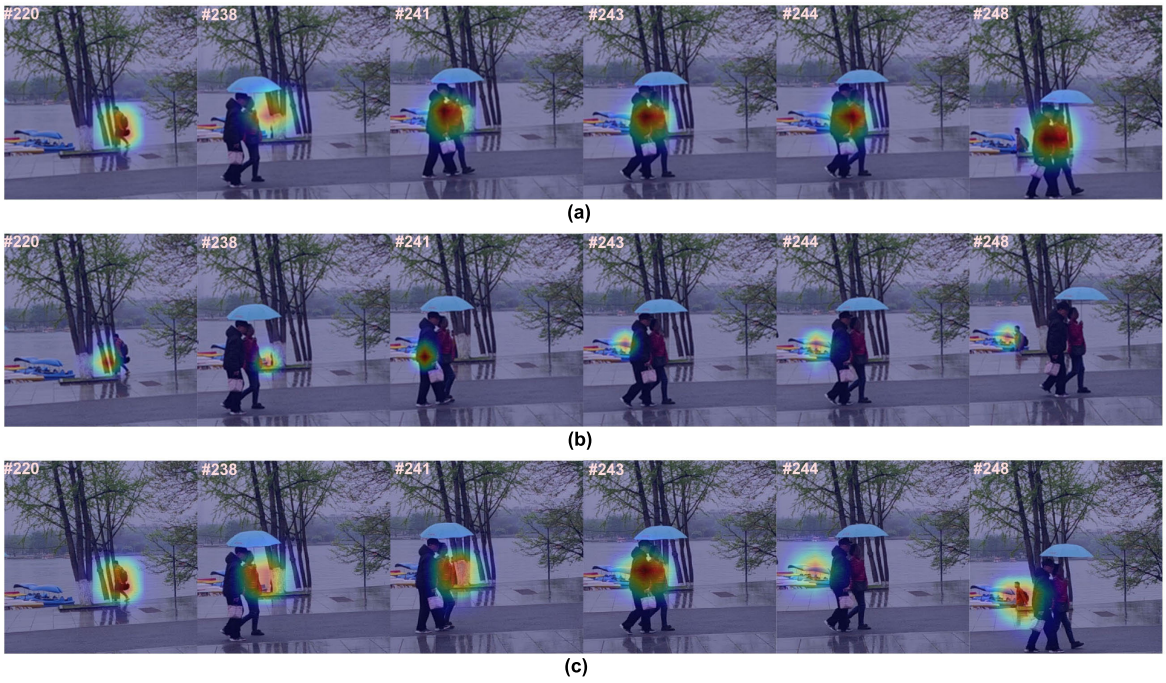

FIGURE 5. The effect of human attention on local tracking(The red area represents the high score and the blue area represents the low score): (a) Tracking process and similarity score of SiamRPN when there is no human guidance (Tracking fails after being interfered by similar targets); (b) Human attention score; (c) Tracking process and fusion score guided by human attention.

candidate bounding box around human visual focus will be more likely to get a high score. It's worth noting that when the human is not in the loop, local tracking module can also work based on the similarity score. Our purpose is not to take over or manually track, but to design a reasonable method to make the tracker more accurate and robust when humans are in the loop.

Figure 5 shows the effect of human attention on local tracking. In a pedestrian tracking scene, the target is blocked by passing pedestrians and trees. As shown in Figure5b, when the target is occluded, humans consciously search for the target and pay attention to the area where the target may appear. Human vision provides the tracker with a strong prior position information and guides the tracker to pay attention to the area of human attention. When the target reappears, our local tracker can continue to locate the target accurately. This tracking process shows that when interference or occlusion occurs in a local area, human attention makes the local tracker more robust.

B. HUMAN ATTENTION ANALYSIS MODULE

In this module, we analyze human attention and identify TOHI through two interaction paradigms. All TOHI that have been successfully identified by these paradigms will be added to the TOHI library and used as the TOHI-guided Validator's query set (details in TOHI-guided validator). The workflow of TOHI identification is shown in Figure 6. Firstly,

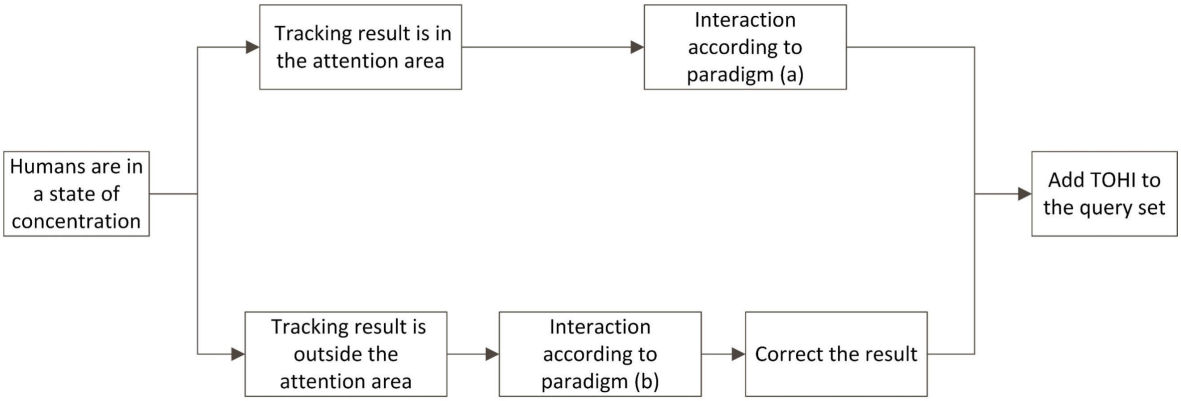

FIGURE 6. TOHI identification workflow. When humans are in a state of concentration, there will be two interactive paradigms to identify TOHI according to the location of the attention area.

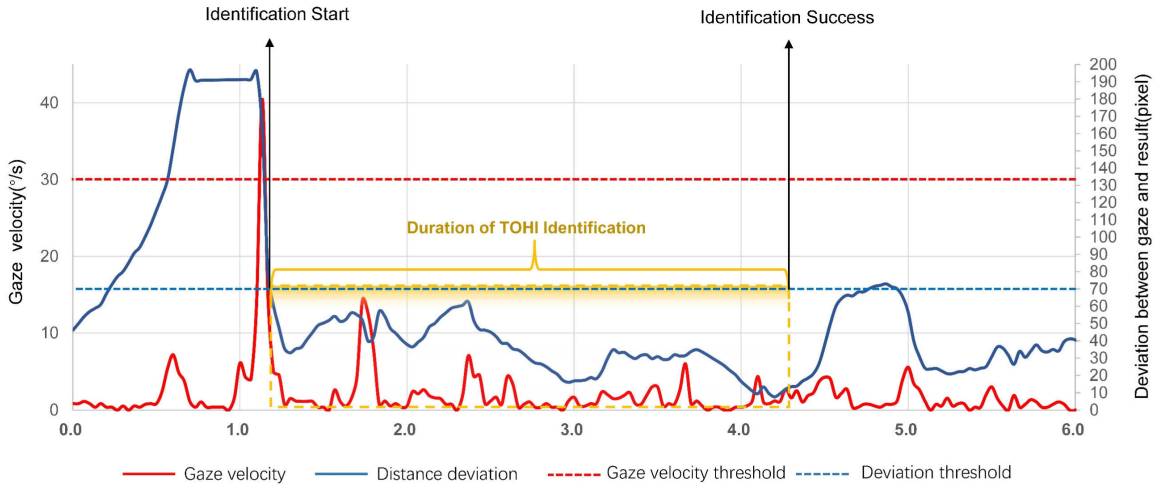

FIGURE 7. The example of TOHI identification process when tracking result within the visual focus area. Red shows gaze velocity, blue shows the distance deviation, and yellow shows the accumulated attention time. The distance deviation threshold is set to 70 pixels, and the gaze velocity threshold is set to 30°/s. The dwell time threshold is set to 3 seconds.

we discriminate the state of human attention by analyzing the patterns of human eye movement and the accumulated attention time. If humans keep gazing or smooth pursuit over a period of time, they are considered to be in a state of concentration. Subsequently, two interaction paradigms are designed to deal with the case when the tracking result is in the human visual focus area and the case when the tracking result is outside the visual focus area.

If the tracking result is in the human visual focus area, the result will be regarded as TOHI and added to the TOHI library. If the tracking result is outside the human visual focus area, a certain local area centered on human fixation point will be identified as TOHI and output as the result of human correction. Human eye movement patterns can be divided into gaze, smooth pursuit and saccade, among which gaze and smooth pursuit reflect the attention state when human tracking a single target. Specifically, ''gaze'' and ''smooth pursuit'' can be distinguished from ''saccades'' by the gaze velocity, which is calculated using human real-time fixation point [20]–[22]. If the gaze velocity is within the preset velocity threshold, we consider it to be ''gaze'' or ''smooth pursuit'', otherwise it is ''saccade''. When the duration of ''gaze'' or ''smooth pursuit'' exceeds the dwell time threshold, we think that human attention is focused on the gaze area. In addition, the distance deviation between the fixation point and center of the output tracking result will be calculated in real time and the tracking result will be divided into within or outside the visual focus area according to the preset deviation range.

Figure 7 shows an example of the TOHI identification process when the tracking result is within the visual focus area. The attention time starts to accumulate when the gaze velocity and the distance deviation are both less than their respective thresholds. If the accumulated attention time exceeds the dwell time threshold, the current tracking result will be successfully identified as TOHI. The corresponding interactive interface is shown in Figure 8a. As the human gaze follows the target, the target bounding box will fade from green to red as feedback, reminding humans that the system is capturing their attention on the target and encouraging

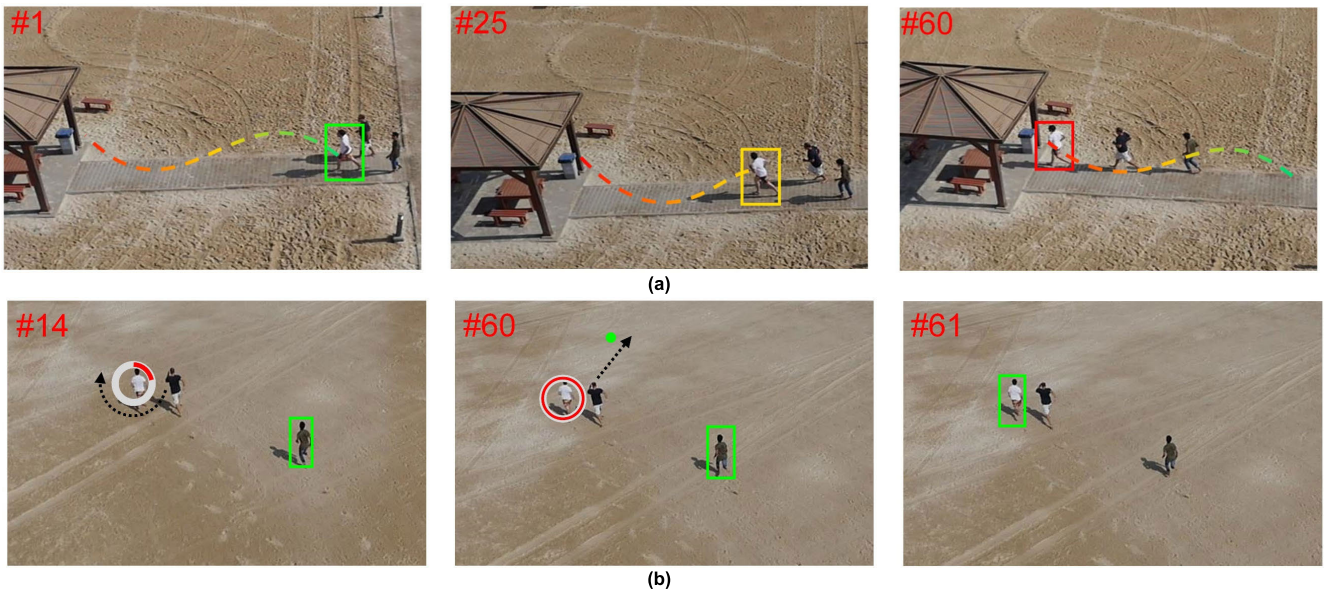

FIGURE 8. The eye movement interaction paradigm: (a) When the tracking result is within the human attention area, the TOHI area can be recognized by smooth pursuit; (b) When the tracking result is outside the human attention area, the selection is confirmed for the second time by capturing eye gestures looking toward the random guide point.

humans continuous focus. The color change of the target box reflects the change of the accumulated attention time, green means that the user is not paying attention to the target, red means that the target is about to be identified as TOHI. Since a complete attention process may include many small gaze processes and micro-saccades, in order to avoid the influence of micro-saccades, we set up an insurance mechanism. Specially, in the duration of gaze or smooth pursuit, a slight glance will not clear the accumulated attention time but punish the accumulated attention time. When the saccade exceeds the tracking result range, or the accumulated attention time is reduced to 0, the attention process is over.

When the tracking result is outside the visual focus area, there are two possibilities. One is the unconscious accidental touch caused by human distraction. The other is that there is a more concerned target or there is an error in the tracking system, and humans want to correct the current tracking result. Since the system cannot confirm whether humans are intentional, we need to obtain further confirmation from humans. Therefore, we design an interaction paradigm that requires humans to consciously trigger a second confirmation. (Figure 8b). When the tracking result outside the visual attention area is sensed, a circle showing accumulated attention time appears on the fixation point. When the accumulated attention time exceeds the dwell time threshold, a random guide point appears near the fixation point. Humans need to look at the guide point to make a second trigger, thus proving its active operation. The tracking result will be corrected to human visual focus area after the second trigger, and the tracking will be reinitialized. The design of this secondary trigger avoids the Midas problem and greatly reduces the misidentification of TOHI. More importantly, this interaction method is asynchronous with the computer tracking system, so that humans have the highest priority to correct the result whenever they want. Once the system has an error which is difficult to recover, this asynchronous interaction provides a reliable way to intervene in the system and ensures absolute human control.

C. TOHI-GUIDED VALIDATOR

In our tracking framework, the local tracker cannot guarantee the tracking result is correct. Therefore, a validator is needed to verify the local tracking results. In this section, the TOHI identified in human attention analysis module will serve as the query set of our validator. We evaluate the reliability of the local tracking result by calculating the similarity between TOHI and the result to be verified. Specifically, an embedding function is learned to embed the local tracking result and TOHI into the same discriminative Euclidean space. Then we calculate the distance in Euclidean space between each TOHI and the result to be verified. If the minimum distance is less than the preset distance threshold, the tracking result is considered to be credible. Otherwise, we believe that the tracking results are unreliable. With the interaction between humans and computers, new TOHI is continuously identified and sent to the query set of the validator, so that the tracker can cope with constantly changing situations. We adopt ResNet50 [23] as the backbone. The parameters of the network are initialized with the ImageNet pretrained models and then finetuned on the Market1501 [24] and DukeMTMC-reID [25], [26]. The discriminative ability is ensured by the TriHard loss [27], [28] denoted as $L_{TriHard}$:

$$L_{TriHard} = \sum_{a \in batch} \left(\max_{p \in A} d_{p,a} - \min_{n \in B} d_{n,a} + m \right)_+ \quad (2)$$

where $d_{p,a}$ and $d_{n,a}$ are feature distances of positive and negative pairs, m is the margin of triplet loss. The feature distance is calculated by a non-squared Euclidean distance.

When the validator judges that the result is unreliable, a short time window for human intervention will be open. During this window, local tracking will continue. Humans can focus on the correct tracking target to intervene, so that the new TOHI is added to the query set. This process will continue until the local tracking result is successfully verified, or the intervention window is over. In our experiment, the intervention window is set to 4 seconds. If the human does not perform any operation on this target during the window, the global search will start.

D. GLOBAL SEARCH MODULE

In this module, YOLOV3 [29] serves as our person detection model, which has been widely recognized for its excellent detection performance and efficiency. All detected candidate person will be sent to TOHI-guided Validator, and the successfully verified candidate person will be considered as the result of global search module.

IV. EXPERIMENT

A. PARTICIPANTS

In this work, there are 5 graduate students as participants, including 3 males and 2 females, aged 23 to 26. Participants have normal or corrected-to-normal vision. They all gave informed consent to participate in the study. This experiment was approved by the Bioethics Committee of National University of Defense Technology.

B. DATASET PREPARATION

In this work, we collected videos with UAV and manually labeled them as our dataset to verify the unique advantages of our tracker in person tracking. The image resolution is 1920 pixels × 1080 pixels, collected at 30fps. The dataset contains 15 sequences, of which the longest sequence is 10084 frames and the shortest sequence is 1608 frames. To further illustrate the characteristics of our dataset, we list several typical challenges in UAV-based person tracking:

- Fast Motion (FM): In two consecutive images, the distance between the center point of the target is more than 20 pixels.
- Appearance Change (AC): Changes in appearance caused by environmental influences or subjective human will. Such as changes in appearance caused by illumination variation or people changing clothes.
- Long-term Disappearance (LD): The tracking object disappears for more than 60 frames.
- Occlusion (OCC): The tracking object is completely or partially occluded
- Background Clutter (BC): There are humans or objects similar to the target in the background.

Figure 9 shows the ratios of representative attributes in different datasets. OTB [8] is the most commonly used dataset for single object tracking which consists of short videos from generic real-world scenarios. UAV123 [10] is the first widely accepted dataset for tracking with UAV. VOT2018LT [9] is the long-term tracking dataset used in VOT2018 challenge. It can be seen from Figure 9 that our dataset contains more of these tracking challenges than the current public test datasets, especially the long-term disappearance and the appearance change. Therefore, the test results on our dataset will more truly reflect the application ability of the tracker in complex scenes of UAV-based person tracking, especially when the target disappears for a long time and the appearance of the target changes.

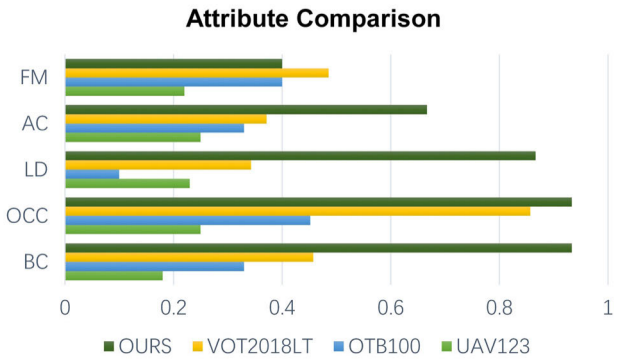

FIGURE 9. The ratios of representative attributes in different datasets.

C. EXPERIMENTAL SETUP AND PROCEDURE

The eye tracker we use is Tobii Pro Nano with a sampling frequency of 60Hz. The screen size of the ground operation station is 23.8in and the screen resolution is 1920 pixels × 1080 pixels. The proposed method is tested on a PC with an Intel i9-10900KF CPU and a NVIDIA GTX2080Ti GPU (11G memory).

This experiment simulates the scene in the UAV ground station. Subjects will act as a UAV operator to observe the images transmitted back in real time. The image captured by the UAV's high-definition camera is displayed on the screen of the ground station. The subject sits about 60cm in front of the monitor with an eye tracker set at the bottom of the screen. To ensure more accurate eye movement data, each subject was asked to place his chin on the bracket, then calibrate the eye tracker.

Before the experiment, we informed each subject that the purpose of the experiment was to track the specified target. Subjects need to be familiar with the paradigm of interaction with the system mentioned in human analysis module. For example, target color change represents that the target is being paid attention to and the meaning of the guide point is to make the subject consciously confirm and choose. After the subject confirmed that the interaction method and the purpose of the experiment were fully understood, we started the experiment.

During the experiment, we played videos one by one to simulate person tracking by UAVs in different scenarios. Before each video played, we stayed in the first frame to

TABLE 1. Experimental results of 5 subjects and the ablation experiment result.

Participant ID	F-score (%)	Sr0.3 (%)	Number of TOHI	Correction Times	Total Attention (%)	Median value of the Deviation (pixel)	Mean value of the Deviation (pixel)
S1	51.4	76	341	14	86.5	71.4	**114.7**
S2	**55.7**	76.1	370	19	86.8	68.5	142.7
S3	55.4	**77.9**	**425**	**20**	**87**	**57.6**	118.6
S4	52.9	72.3	185	17	84.9	90.9	201.8
S5	38.4	41.2	132	11	85.2	107.5	204.3
Ours-without-human	40.4	46.5	-	-	-	-	-

confirm the tracking target to the subjects. Subjects can see their gaze points in real time on the screen. In simple scenarios, subjects can be more relaxed and do not have to keep looking at the target. But in complex scenarios, the subjects were asked to focus on the target as much as possible. If the subject finds that the tracking result is wrong, the subject should make corrections in accordance with the interaction paradigm.

Through our experiment, we want to answer the following questions:

- Ablation experiments: Does the human-in-the-loop improve the tracking ability of the tracker, especially in the complex environment?
- Effect of human visual attention on tracking: In our method, how does human attention affect person tracking?
- Comparison with SOTA trackers: Compared with the SOTA method, does our method perform better in a complex environment?

D. EVALUATION PROTOCOL

We refer to the method in VOT2018-LT challenge [9]. Different trackers are compared using ***F-score*** (Equation 3), where ***Pr*** is tracking accuracy (Equation 4) and ***Re*** is tracking recall rate (Equation 5). $\boldsymbol{G_t}$ is the target position of the ground truth, and $\boldsymbol{A_t}$ is the predicted position of the tracker. $\boldsymbol{\Omega}(\boldsymbol{A_t}, \boldsymbol{G_t})$ is the intersection over union (IOU) which is a measure of the overlap between $\boldsymbol{A_t}$ and $\boldsymbol{G_t}$. If the target is absent, the ground truth is an empty set, namely $\boldsymbol{G_t} = \emptyset$. Similarly, if the tracker did not predict the target, the output is $\boldsymbol{A_t} = \emptyset$. N_p represents the total number of frames in which $\boldsymbol{A_t} \neq \emptyset$, and N_g represents the total number of frames in which $\boldsymbol{G_t} \neq \emptyset$. One reason why we adopted F-score is that it can be used as an evaluation standard for long-term trackers as well as a standard for short-term trackers.

$$F - score = 2Pr \times Re/(Pr + Re) \tag{3}$$

$$Pr = \frac{1}{N_p} \sum\nolimits_{t \in \{t: A_t \neq \emptyset\}} \Omega(A_t, G_t) \tag{4}$$

$$Re = \frac{1}{N_g} \sum\nolimits_{t \in \{t: G_t \neq \emptyset\}} \Omega(A_t, G_t) \tag{5}$$

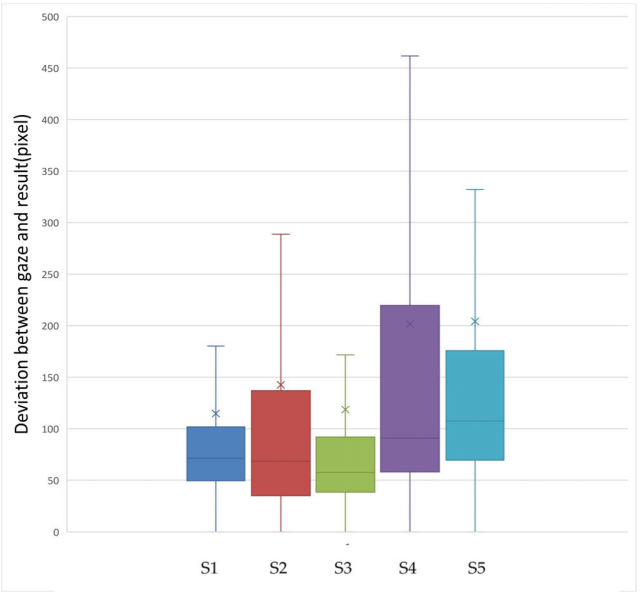

FIGURE 10. Distribution of the deviation between gaze and ground truth.

In addition, we also considered $\boldsymbol{Sr_{0.3}}$ as an evaluation standard, which refers to the proportion of $\boldsymbol{\Omega}(\boldsymbol{A_t}, \boldsymbol{G_t})$ over 0.3. In the UAV tracking task, we are more concerned about whether the target can be successfully tracked and $\boldsymbol{\Omega}(\boldsymbol{A_t}, \boldsymbol{G_t})$ greater than 0.3 is sufficient for locating a target with a UAV [13], [30].

V. EFFECT OF HUMAN VISUAL ATTENTION ON TRACKING

In this part, we show the experimental results of 5 subjects in Table 1 and analyze the relationship between each subject's attention and his tracking results. At the same time, the results of the ablation experiment are also shown in Table 1(Ours-without-human). Specifically, in the ablation experiment, we remove human-related functions in the framework, in which only the local tracker and the global search modules were retained.

All subjects said that they felt natural and comfortable during the experiment and could focus on the target. We evaluate the attention of each subject from four aspects, namely, the number of successfully identified TOHI, the number of human corrections, the proportion of the total attention time

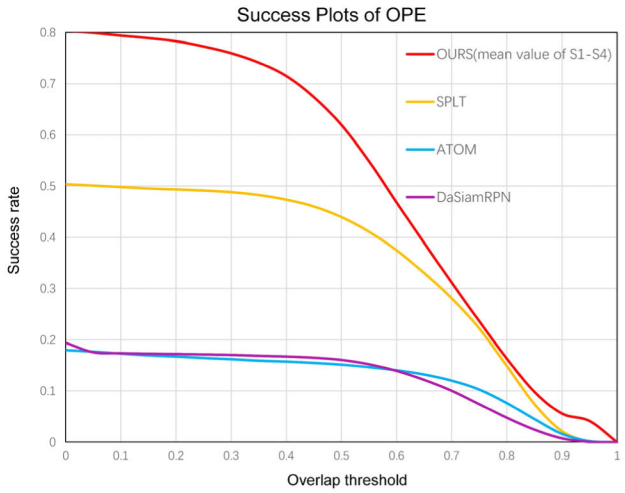

FIGURE 11. Success plots of one pass evaluation on our dataset.

and the deviation between fixation point and ground truth. The total attention time refers to the total time of gaze and smooth pursuit identified in the Human Attention Analysis Module, which reflects the subjects' attention initially. Furthermore, the number of successful TOHI identified can indicate the subjects' attention to the tracking target and their adaptation to the system.

Correction numbers can represent the enthusiasm of humans to make active corrections. The distribution of the deviation intuitively shows the difference of each subject's attention to the target. A smaller and concentrated deviation indicates that the subjects pay more attention to the target. The distribution of the deviation is shown in Figure 10.

As shown in Table 1, S3 has the longest attention time, the most TOHI and the most correction times, which is consistent with the results of $\boldsymbol{Sr_{0.3}}$. In addition, the attention deviation of S3 is relatively concentrated, and the median deviation is the lowest among the five subjects. Based on the above analysis, we believe that S3 has the most concentrated attention to the target and achieve the best $\boldsymbol{Sr_{0.3}}$. S2's attention time, number of corrections, and TOHI number are second only to S3, and S2's attention distribution is more distracted than S3, which may reveal that his concentration is inferior to S3. This still allows him to get the highest ***F-score*** and second $\boldsymbol{Sr_{0.3}}$. From the results of the ablation experiment, we also find that the ***F-score*** and $\boldsymbol{Sr_{0.3}}$ of S1, S2, S3, and S4 far exceed our tracker without humans, which shows that human-in-the-loop can improve the performance of the tracker.

S5 has the least attention time, the least correction times and the least TOHI, which may indicate that S5 did not pay all attention to the target during the experiment. In addition, his attention is not concentrated sufficiently, and the attention deviation is relatively scattered. S5 obtained the lowest score in both ***F-score*** and $\boldsymbol{Sr_{0.3}}$, which is slightly lower than our tracker without humans. It is probably due to the incorrect guidance and inappropriate intervention of S5, which leads to the system's misunderstanding of human intentions. The results of S5 confirm that humans have a dominant role in this method when human-in-the-loop. Actually, our method is designed for operators of UAV ground control stations, who have received professional UAV operation training and are experts in tracking tasks. We believe that in an open task environment, the confidence of human experts' judgments is higher than the results of the machine algorithm itself. Therefore, if the experts' judgments are erroneous, the results of the tracking system will be unstable or even wrong. The user should focus on the task when using this system, otherwise the system should be switched to unmanned mode.

From the analysis of the above results, we believe that in our framework, humans play a decisive role in the loop. When humans focus on the target, the performance of the tracker will be greatly improved. Moreover, the more humans pay attention to the target, the greater the improvement of the tracking system.

VI. COMPARISON WITH STATE-OF-THE-ART TRACKERS

Several typical SOTA trackers are compared with our method. The compared trackers include SPLT [5], ATOM [6] and DaSiamRPN [7]. SPLT is a real-time long-term tracker, while ATOM and DaSiamRPN are short-term trackers. These SOTA methods are all without human intervention

Figure 11 shows the success plot of one pass evaluation on our dataset which illustrates the percentage of successfully

TABLE 2. One pass evaluation for SOTA trackers on our dataset.

Tracker	*F-score* (%)	*Pr* (%)	*Re* (%)	$Sr_{0.3}$ (%)
Ours-*S3*	**55.4**	**57.9**	**53.1**	**77.9**
Ours (Mean value of *S1*-*S4*)	53.9	56.5	51.7	75.6
Ours (Mean value of *S1*-*S5*)	50.8	55.8	47.4	68.7
Ours-without-human	40.4	42.5	38.5	46.5
SPLT	41.7	44.9	39	48.8
ATOM	13.3	12.5	14.1	16.2
DaSiamRPN	13.1	12.3	13.9	17

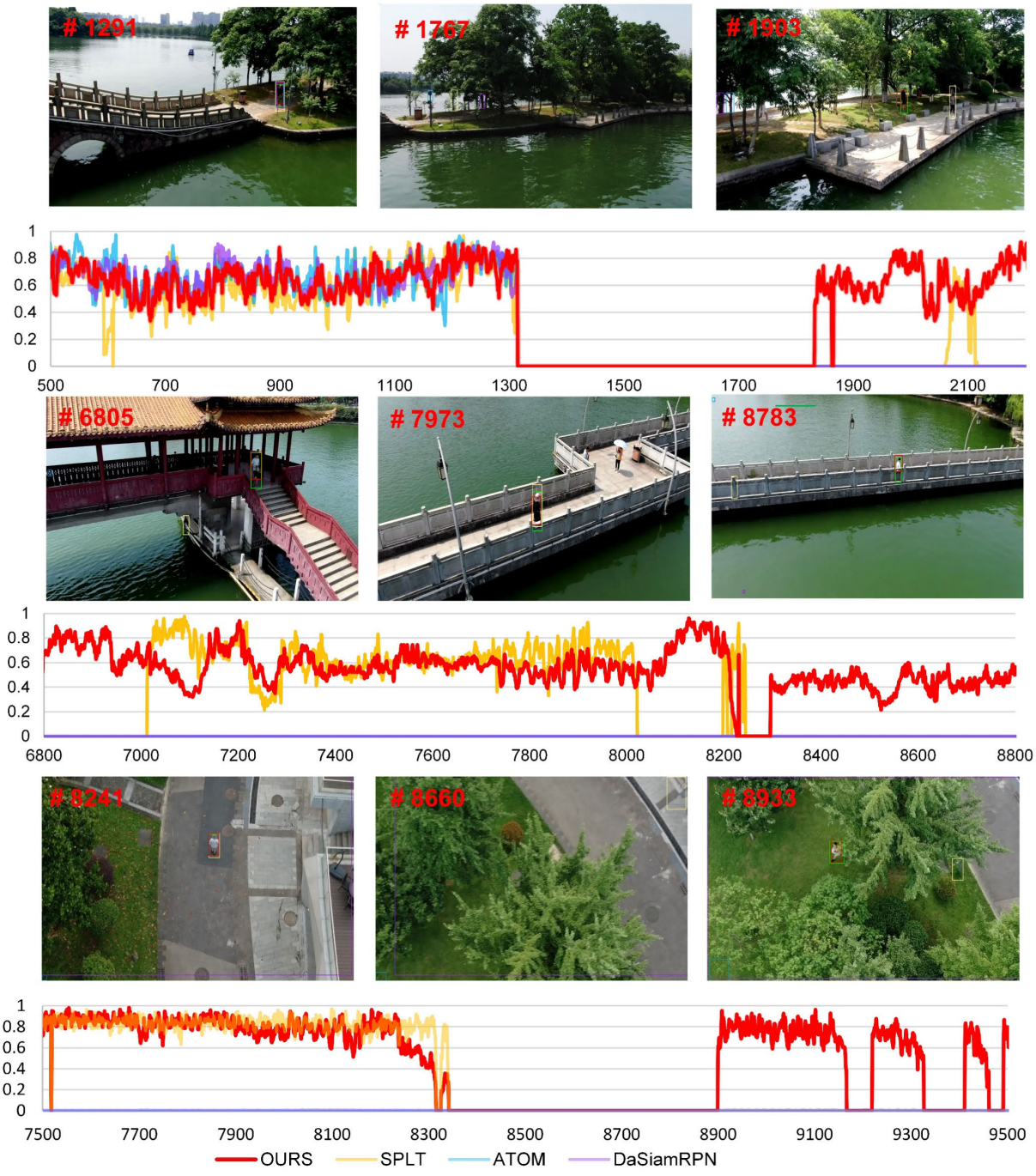

FIGURE 12. IOU over tracking frame.

tracked frames whose overlap is larger than a given threshold. In addition, the recall rate (***Re***), tracking accuracy ***(Pr)***, ***F-score*** and $\boldsymbol{Sr_{0.3}}$ are shown in Table 2. We respectively show the best result(Ours-S3), the average result and the average result without S5. Since S5 is not focused on the target, we believe that the average results of S1-S4 may better reflect the true capabilities of our method. Nevertheless, we still listed the average results of 5 subjects in Table 2.

In our experiment, ATOM and DaSiamRPN lost their targets soon and could not retrieve the target, which led to their poor results. SPLT achieved similar results with Ours-without-human, because it has the ability to retrieve the target. However, since SPLT cannot identify targets whose appearance has dramatically changed, its success rate is far lower than our method. As shown in Figure 11, our tracker has a higher success rate at each threshold. Ours (Mean value of S1-S4) is 26.8% ahead of SPLT in $\boldsymbol{Sr_{0.3}}$ and 12.2% ahead of SPLT in ***F-score***. Furthermore, Ours-S3 is 29.1% ahead of SPLT in $\boldsymbol{Sr_{0.3}}$ and 13.7% ahead of SPLT in ***F-score***, which illustrates the absolute advantages of our approach in

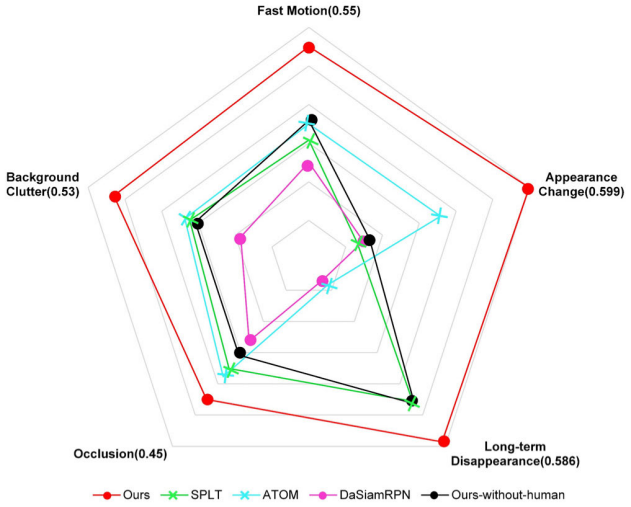

FIGURE 13. Tracking recall rate on different attribute image sequences.

dealing with complex environmental challenges. Even considering S5, Ours(Mean value of S1-S5) still significantly outperformed SPLT.

Figure 12 shows several examples of the tracking process. The top group of pictures shows the person disappearing in the woods. At the beginning of the tracking, all trackers can accurately track the target, but when the target enters the woods, other trackers cannot determine whether the target still exists. When the target reappears, only our tracker can find the target in time and accurately. The middle group of pictures shows pedestrian changing clothes and waking through illumination variations. Only our tracker keeps tracking the target. The bottom group of pictures shows the person changing his clothes and hiding in the woods, then only our tracker retrieves them. We have also calculated the recall rate(***Re***) of Ours(Mean value of S1-S4) on image sequences with different attributes (Figure 13), which proves that our method is more effective in the five challenges, especially in the emphatic challenges of appearance change and long-term disappearance.

Our method needs to ensure smooth interaction between the user and the system when human-in-the-loop, so we pay special attention to real-time performance. Each module in our tracking framework uses the offline trained model to maximize efficiency as much as possible, so as to ensure the real-time performance of the entire framework. We recorded the computation time of each sequence and calculated the average speed. Due to the different intervention processes of users in our method, the algorithm computation time is also different. Finally, the average speed of our method in the experiment is 26.3 FPS, which can meet the real-time requirements of most tracking systems. Ours-without-human reached a speed of 29.1 FPS. It can be seen that human-in-the-loop does not bring too much computational burden, but it can greatly improve the performance of the algorithm.

In our method, human-in-the-loop can deal with more complex challenges. It is worth noting that our method does not require humans to stay in the loop at all times. Our approach without human is enough to deal with simple scenarios. We hope that humans can give the tracker enough guidance in complex scenes, so that the tracker can learn the target of human interest and improve the tracker's ability to respond to environmental changes. On the one hand, machines can reduce the cognitive burden of humans and help humans to complete tasks more easily. On the other hand, human-in-the-loop improves the ability of the tracker, so that the tracker can gradually deal with challenges in complex scenes on its own.

Computer vision is usually used as the basic technology of human-computer interaction [31], but it is rarely found that the performance of computer vision algorithms can be improved through human-computer collaboration. Some existing human-like automatic visual tracking methods [32]–[34] try to analyze contextual information based on human cognition and introduce part of expert experience into the tracking method. The human-like methods and our human-machine-collaboration-based method have the same original intention, which is to improve the ability of the algorithm through human experience. In an open and complex task environment, a large amount of human experience and knowledge is required to support decision-making. However, the human-like methods can only solve problems in very limited scenarios by presupposing some simple human experiences which can be formalized. Therefore, the current human-like tracking methods cannot perform tasks in an open environment. In contrast, the method we proposed introduces human perception and decision-making capabilities of complex changing scenes in real time. Human-in-the-loop will guide and help the machine understand the changing task environment and tracking targets during the tracking process, thereby building an application-oriented human-machine hybrid robust tracking system.

VII. CONCLUSION

This work demonstrates a human-machine hybrid augmented tracking system, which is used to improve the person tracking capabilities of UAV in complex environments. Humans can guide and correct UAV tracking through the eye movements interactive paradigm. Meanwhile, machines recognize and learn TOHI from humans, so as to improve the adaptability to changing environment. The experimental results show that human participation in our framework can effectively improve the tracking success rate and tracking robustness. Compared with the SOTA trackers, our approach has an absolute advantage in complex scenarios, especially when the target disappears for a long time and the target's appearance changes dramatically. We believe that our method is a meaningful attempt of human-machine collaboration in visual tracking and provides a novel and reliable solution for UAV-based person tracking.

REFERENCES

[1] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550, doi: 10.1109/CVPR.2010.5539960.

[2] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015, doi: 10.1109/TPAMI.2014.2345390.

[3] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939, doi: 10.1109/CVPR.2017.733.

[4] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302, doi: 10.1109/CVPR.2016.465.

[5] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'Skimming-Perusal' tracking: A framework for real-time and robust long-term tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2385–2393, doi: 10.1109/ICCV.2019.00247.

[6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4655–4664, doi: 10.1109/CVPR.2019.00479.

[7] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 103–119.

[8] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015, doi: 10.1109/TPAMI.2014.2388226.

[9] A. Lukežič, L. Č. Zajc, T. Vojíř, J. Matas, and M. Kristan, "Now you see me: Evaluating performance in long-term visual tracking," 2018, *arXiv:1804.07056*.

[10] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 445–461.

[11] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, and P. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[12] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980, doi: 10.1109/CVPR.2018.00935.

[13] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012, doi: 10.1109/TPAMI.2011.239.

[14] Y. Zhang, D. Wang, L. Wang, J. Qi, and H. Lu, "Learning regression and verification networks for long-term visual tracking," 2018, *arXiv:1809.04320*.

[15] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Č. Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey, and G. Fernandez, "The sixth visual object tracking VOT2018 challenge results," in *Computer Vision*, L. Leal-Taixé and S. Roth, Eds. Cham, Switzerland: Springer, 2019, pp. 3–53.

[16] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6297–6306, doi: 10.1109/CVPR42600.2020.00633.

[17] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Č Zajc, O. Drbohlav, A. Lukezic, A. Berg, and A. Eldesokey, "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2206–2241, doi: 10.1109/ICCVW.2019.00276.

[18] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. C. Zajc, A. Lukežic, O. Drbohlav, and L. He, "The eighth visual object tracking VOT2020 challenge results," in *Computer Vision*, A. Bartoli and A. Fusiello, Eds. Cham, Switzerland: Springer, 2020, pp. 547–601.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, Palm Beach Gardens, FL, USA, 2000, pp. 71–78, doi: 10.1145/355017.355028.

[21] S. Munn, L. Stefano, and J. Pelz, "Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding," in *Proc. Symp. Appl. Perception Graph. Vis. (ACM)*, 2008, pp. 33–42.

[22] O. V. Komogortsev, D. V. Gobert, U. K. S. Jayarathna, D. H. Koh, and S. M. Gowda, "Standardization of automated analyses of oculomotor fixation and saccadic behaviors," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 11, pp. 2635–2645, Nov. 2010, doi: 10.1109/TBME.2010.2057429.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1116–1124.

[25] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782, doi: 10.1109/ICCV.2017.405.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 17–35.

[27] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[30] M. Kristan, J. Matas, A. Leonardis, T. Vojír, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016, doi: 10.1109/TPAMI.2016.2516982.

[31] S. Nayak, B. Nagesh, A. Routray, and M. Sarma, "A human–computer interaction framework for emotion recognition through time-series thermal video sequences," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107280, doi: 10.1016/j.compeleceng.2021.107280.

[32] J. Gòmez-Romero, M. A. Patricio, J. Garcìa, and J. M. Molina, "Ontology-based context representation and reasoning for object tracking and scene interpretation in video," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7494–7510, Jun. 2011, doi: 10.1016/j.eswa.2010.12.118.

[33] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "A human-like description of scene events for a proper UAV-based video content analysis," *Knowl.-Based Syst.*, vol. 178, pp. 163–175, Aug. 2019, doi: 10.1016/j.knosys.2019.04.026.

[34] D. Cavaliere, V. Loia, and S. Senatore, "Towards an ontology design pattern for UAV video content analysis," *IEEE Access*, vol. 7, pp. 105342–105353, 2019, doi: 10.1109/ACCESS.2019.2932442.