NEARING OR SURPASSING: OVERALL EVALUATION OF HUMAN-MACHINE DYNAMIC VISION ABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

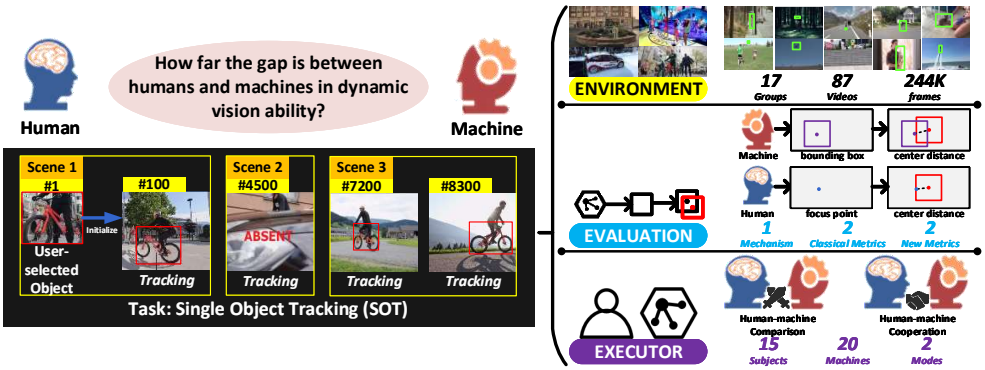
Dynamic visual ability (DVA), a fundamental function of the human visual system, has been successfully modeled by many computer vision tasks in recent decades. However, the prosperity developments mainly concentrate on using deep neural networks (DNN) to simulate the human DVA system, but evaluation systems still simply compare performance between machines, making it tough to determine *how far the gap is between humans and machines in dynamic vision tasks.* In fact, neglecting this issue not only makes it hard to determine the correctness of current research routes, but also cannot truly measure the DVA intelligence of machines. To answer the question, this work designs a comprehensive evaluation system based on the 3E paradigm – we carefully pick 87 videos from various dimensions to construct the environment, confirming it can cover both perceptual and cognitive components of DVA; select 20 representative machines and 15 human subjects to form the task executors, ensuring that different model structures can help us observe the effectiveness of research development; and finally quantify their DVA with a strict evaluation process. Based on detailed experimental analyses, we first determine that the current algorithm research route has effectively shortened the gap. Besides, we further summarize the weaknesses of different executors, and design a human-machine cooperation mechanism with superhuman performance. In summary, the contributions include: (1) Quantifying the DVA of humans and machines, (2) proposing a new view to evaluate DVA intelligence based on the human-machine comparison, and (3) providing a possibility of human-machine cooperation. The datasets, toolkits, codes, and evaluation metrics will be open-sourced to help researchers develop intelligent research on dynamic vision tasks.

1 INTRODUCTION

Research on visual abilities can be dated back to the last century (Hubel & Wiesel (1959; 1962)). Neuroscientists divide the human visual system into two categories, namely the static vision ability (SVA) to perceive the details of static objects (Chan & Courtney (1996)), and the dynamic vision ability (DVA) to resolve spatial information of moving objects (JW et al. (1962)). These two visual abilities are essential in our daily life (Land & McLeod (2000); Beals et al. (1971); Burg (1966); Kohl et al. (1991)), and have been modeled by a series of computer vision tasks. Recently, with the growth of dataset scale and the abundance of computing resources, most data-driven algorithms achieve higher and higher scores in experimental environments and are widely employed in various scenarios (Dankert et al. (2009); Weissbrod et al. (2013); Wei & Kording (2018)).

However, some bad cases hidden under the prosperity development challenge the state-of-the-art (SOTA) algorithms. For example, visual models usually decrease their perception when encountering unknown-category targets under special illumination conditions (*e.g.*, a vision-based autonomous vehicle crashes into a large truck at night). This shortcoming is far from humans' powerful visual abilities and may cause safety hazards, causing us to rethink – with the support of massive datasets and powerful computing resources, why can't SOTA models achieve a similar visual ability to humans? Do existing evaluation methods actually measure the visual intelligence of machines?

Compared to thriving algorithm research, evaluation methods have received far less attention. Traditional evaluations mainly focus on the performance comparison between machines (Deng et al. (2009); Russakovsky et al. (2015)), causing the goal to become *downward* (*i.e.*, a method that exceeds all



(a) Schematic diagram of the evaluation system.

Task Settings			Characteristics		Ability	Group	Frames
			Target Absent	Shot-cut			
Short-term tracking (Target presents from beginning to end)			N	N	Perception	A	500-1,000
						B	1,000-2,000
Long-term tracking (Target may disappear and reappear in a single shot)			Y	N	Perception and cognition	C	1,000-2,000
						D	5,000-10,000
Global instance tracking (Target may disappear and reappear in multiple shots)			Y	Y		E	1,000-2,000
						F	5,000-10,000
						G	15,000-30,000
Short-term tracking with challenging factors (Target presents from beginning to end)	Challenging factors in single frame	Abnormal ratio	N	N	Perception and robustness	H	500-1,000
		Abnormal scale				I	
		Abnormal illumination				J	
		Blur bounding-box				K	
	Challenging factors between consecutive frames	Drastic ratio variation				L	
		Drastic scale variation				M	
		Drastic illumination variation				N	
		Drastic clarity variation				O	
		Fast motion				P	
		Low correlation cofficient				Q	

(b) Information on environment settings. Representative sequences are illuistrated in Appendix C.1.

Figure 1: Illustrations of the evaluation system. We follow 3E paradigm (Hu et al. (2022a)) to design an overall evaluation system of human-machine DVA.

others is considered excellent). In fact, these evaluation mechanisms actually measure the machine's performance rather than intelligence. When we refer to machine intelligence, a natural association is Turing test (Turing (2009)), which indicates that machine intelligence evaluation requires human participation. Using humans as a reference, the goal will become *upward* – exemplary machines must gradually move closer to human capabilities. In other words, *exploring the gap between humans and machines is very crucial for machine intelligence evaluation*.

Based on this idea, recent works have introduced humans to several vision tasks, using human baseline to measure machine intelligence (Geirhos et al. (2018; 2020; 2021); Langlois et al. (2021)). Regrettably, their research scope is mainly restricted to static vision tasks, while neuroscience and cognitive psychology studies have shown that the correlation between SVA and DVA is naturally low (Long & Penn (1987)). Therefore, in the absence of research in DVA, an intuitive question is, *how to compare humans and machines in the dynamic vision ability?* To answer this question, we should first select a suitable computer vision task to model human DVA, then design a comprehensive evaluation system to accomplish human-machine DVA measurement and comparison.

Compared with dynamic vision tasks like multi-object tracking (MOT) (Ciaparrone et al. (2019)) and video object detection (VOD) (Wang et al. (2018)), single object tracking (SOT, *i.e.*, locates a user-specified moving target in a video) (Wu et al. (2015)) is a *category-independent* task with no constraint on motion continuity, scene change, or object category, and can be regarded as the closest task to represent human DVA. Due to the human-like task definition of SOT, excellent task executors should contain two capacities (*perception* and *cognition*). They should not only keep tracking the moving target (*perception*) but also re-locate the target when its position is mutated (*cognition*).

From above analyses, we select SOT as the representative task, and follow 3E paradigm (Hu et al. (2022a)) to design an overall evaluation system from three aspects (Figure 1): (1) **Experimental**

environment construction. To entirely reflect the task characteristics and thoroughly compare the human-machine DVA, we choose 87 videos with 244,455 frames to construct the experimental environment. All videos are carefully picked from various dimensions, ensuring the environment can cover the perceptual and cognitive components of DVA (Section 3.1). (2) **Experimental executor selection.** Task executors included 20 representative algorithms and 15 subjects (all subjects are computer vision researchers aged 20-30). All experiments are managed in a strict process (Section 3.2). (3) **Evaluation method design.** All executors are evaluated under the one-pass evaluation (OPE) mechanism (Wu et al. (2015)). The positional relationship between tracking result and ground-truth is selected to evaluate performance. Especially, we consider the influence of sequence length and revise indicators to generate a more appropriate evaluation (Section 3.3).

Based on these three steps, we quantify the DVA of executors and conduct detailed comparisons from task constraints, video contents, challenging factors, video duration, *etc*.(Section 4) By analyzing and summarizing the results (Figures 2), we find that: (1) **Synthetically, humans have stronger dynamic vision ability than machines.** Experimental results demonstrate that humans outperform machines when both perceptual and cognitive abilities are required. Particularly, when we mainly examine the perceptual ability in dynamic vision tasks, we also notice that **the gap of human-machine perceptual ability is closing.** (2) **Human-machine cooperation is possible and effective in dynamic vision tasks.** Based on human-machine differences in perception and cognition, we design a set of cooperative experiments. Results indicate that the SOTA machines with human-machine cooperation achieve even better performance than humans in some cases.

In summary, this work starts with designing a more intelligent evaluation system, measures the DVA gap between humans and machines, determines the effectiveness of current algorithm research routes, finds the weaknesses of different executors, and notes the possibility of human-machine cooperation. We will open-source the toolbox, code, and metrics used in the experiments to help researchers further develop intelligent research on dynamic vision tasks.

2 Related Works

2.1 Human Vision Abilities

Static vision ability (SVA) and dynamic vision ability (DVA) are two basic functions of human visual acuity. Related research has discovered individual differences in DVA in subjects with a similar SVA – the low correlation between these two abilities causes researchers to measure them separately.

Static Vision Ability. For SVA research, minimum detectable threshold, minimum resolution threshold, minimum perceptible alignment threshold, and minimum recognition threshold are four basic thresholds in measurement (Chan & Courtney (1996)). Researchers usually use the above thresholds for experimental design, such as the widely used optotype for visual acuity tests is designed by the minimum recognition threshold. However, some researchers indicate only using SVA to assess the human visual system is limited (Long & Zavod (2002)). First, many visual stimuli in actual application scenes are often in motion. Besides, the SVA tests are displayed under conditions of maximum contrast (*e.g.*, black symbols on white background), which is seldom observed in daily life.

Dynamic Vision Ability. For DVA research, researchers initially reference the SVA task and try to require subjects to observe moving objects on a high-contrast background. For example, an early work displays some moving plus (+) signs on the screen – subjects need to track the signs that have been flashing and to indicate whenever one of those targets flashed (Pylyshyn & Storm (1988)). Another DVA research avenue primarily serves athletes. Since several works have revealed that the DVA of humans is essential for their sports performance, a machine named Bernell's Rotator has been used in sports optometry. Recently, some researchers have tried to use computer programs to evaluate the DVA of athletes (Quevedo et al. (2012)). This evaluation mechanism includes various stimuli with different color and motion trajectories. Although the evaluation methods have developed from simple designs to comprehensive computer programs, they always have limited target categories (*e.g.*, specific symbols), simple video scenes (*e.g.*, colors or some static photographs), and limited motion patterns (*e.g.*, lateral, vertical, oblique, and linear trajectories).

2.2 COMPUTER VISION TASKS

Encouraged by the eye-catching development of artificial intelligence research, a series of visual tasks have been proposed to model the powerful human visual system. This section includes some computer vision tasks and their characteristics (Appendix A).

Image Classification. As a static vision task, image classification (Nath et al. (2014)) allows for classifying a given image as belonging to one of the labeled and pre-defined categories. An excellent model should understand the target features and have a robust classification ability in facing challenging factors like target deformation or image blur.

Video Object Detection (VOD). VOD (Esteva et al. (2021)), an essential task for multiple computer vision applications, desires to accurately determine the category and location of each target in videos. However, the target category is typically limited to pre-defined categories in the training dataset.

Multi-object Tracking (MOT). MOT (Geuther et al. (2019)) in research usually combines with detection. MOT algorithms detect the object's position in the first frame, then calculate the similarity to determine instances belonging to the same target in consecutive frames. Thus, MOT is a model-specific task and mainly focuses on tracking specific categories (*e.g.*, pedestrians or vehicles).

Single Object Tracking (SOT). Unlike the above vision tasks, SOT (Wu et al. (2015)) is *category-independent*, which means it intends to track a moving target without any assumption about the target category. This characteristic allows SOT to be suitable for open-set testing with broad prospects. Besides, some recent research proposes a new task named Global Instance Tracking (GIT) (Hu et al. (2022b)) to cancel the continuous motion assumption. Thus, the SOT definition is additionally extended, making it a further step toward human dynamic vision tasks.

2.3 HUMAN-MACHINE COMPARISONS

The idea of human-machine comparison originates from the Turing test, which has gradually attracted the attention of scholars and aroused a series of essential works in decision-making problems (*e.g.*, AlphaGo (Silver et al. (2017)) and DeepStack (Brown & Sandholm (2018))). Recently, some researchers have proposed a series of works to analyze the gap between human-machine based on the image classification task (Geirhos et al. (2018; 2020; 2021)). They first tested subjects on multiple out-of-distribution (OOD) datasets, and then examined representative machines with different model architectures. Results indicate that the gap in image classification ability between humans and machines is narrowing. For dynamic vision tasks, some researchers have proposed Navigation Turing Test (NTT) (Devlin et al. (2021)) to quantify the human-machine operation in navigation games. Another research group has attempted to evaluate the human visual tracking ability on 15 videos, and indicated that machines are still far from humans in GIT task (Hu et al. (2022b)).

Although some works have tried to compare human-machine vision capabilities, they are either concentrated on SVA tasks or only compare human-machine with a small amount of data on limited DVA tasks. Due to the lack of systematic and comprehensive experimental settings, existing works cannot deeply analyze the gap between humans and machines in DVA tasks.

3 METHODS

3.1 ENVIRONMENT

Unlike toy environments (Pylyshyn & Storm (1988)) that track simple symbols in a high-contrast background, our environment fully models the application scenes and designs various data combinations for ability analyses (*e.g.*, perception, cognition, and robustness). Concretely, the environment should consider the diversity of video contents and includes the variation of video length (especially long sequences). Thus, we refer SOTVerse (Hu et al. (2022a)) and select sequences from representative benchmarks (OTB (Wu et al. (2013; 2015)), VOT series (Kristan et al. (2013; 2014; 2015; 2016; 2017; 2018; 2019)), GOT-10k (Huang et al. (2021)), LaSOT (Fan et al. (2020)), and VideoCube (Hu et al. (2022b))) to form the experimental environment (Figure 1 and Appendix C.1).

Table 1: The information about 20 representative models (SNN-Siamese Neural Network, CF-Correlation Filter, CNN-Convolutional Neural Network). Please refer to Appendix B for details.

Machine	Representation	Update	Aritciture
MixFormer (Cui et al. (2022))	Mixed Attention Module	✓	Transformer
KeepTrack (Mayer et al. (2021))	Res-50	✓	Custom Networks
SiamRCNN (Voigtlaender et al. (2020))	VGG-16	✓	SNN+Re-detection
SuperDiMP (Danelljan et al. (2020))	Res-50	✓	Custom Networks
PrDiMP (Danelljan et al. (2020))	Res-50	✓	Custom Networks
SiamCAR (Guo et al. (2020))	Res-50		SNN
KYS (Bhat et al. (2020))	Res-50	✓	Custom Networks
Ocean (Zhang & Peng (2020))	Res-50	✓	SNN
SiamFC++ (Xu et al. (2020))	AlexNet/GoogleNet	✓	SNN
SiamDW (Zhang & Peng (2019))	Res-22		SNN
GlobalTrack (Huang et al. (2019))	Res-50		Custom Networks
DiMP (Bhat et al. (2019))	Res-50	✓	Custom Networks
SPLT (Yan et al. (2019))	Res-50	✓	SNN
SiamRPN++ (Li et al. (2018a))	Res-50		SNN
ATOM (Danelljan et al. (2018))	Res-18	✓	Custom Networks
DaSiamRPN (Zhu et al. (2018))	Res-50	✓	SNN
ECO (Danelljan et al. (2017))	VGG-M	✓	CNN+CF
SiamRPN (Li et al. (2018b))	AlexNet		SNN
SiamFC (Bertinetto et al. (2016))	AlexNet		SNN
KCF (Henriques et al. (2015))	Histogram of Oriented Gradients	✓	CF

3.2 EXECUTOR

For machines, since we aim to study the DVA gap between humans and machines, there are two concerns in model selection: *trends* (*i.e.*, selecting various algorithms) and *gap* (*i.e.*, focusing on the upper bound of algorithms). Thus, both classic and SOTA methods (Table 1) are selected to explore whether the current research route has effectively shortened the gap.

For humans, we start the measurement of human DVA based on the following concerns:

Organization. 87 videos with different durations, object classes, and scene categories are played at 25FPS to 15 subjects (aged between 20-30, all computer vision researchers) under a high-quality experimental organization – this organization is known as *small-N design* (Smith & Little (2018)), and has been widely used in SVA experiments (Geirhos et al. (2018; 2020; 2021)).

Approval. We have obtained the approvals of all subjects. They all have signed an experimental statement (the template is provided by our IRB, which includes experiment purpose, procedure, risks and discomforts, costs, and confidentiality (Appendix C.2)) before the experiment.

Process. Previous works mainly use eye-tracker Hu et al. (2022b); Xia et al. (2021) or mouse Geirhos et al. (2018; 2020; 2021) to record human visual ability. To select a better measurement method, all subjects tried both mouse and eye-tracker, and they express that the mouse is better. Besides, we have compared these two devices from practical and theoretical views in Appendix C.3, and finally select the mouse-based method. Experiment includes four steps (Appendix C.4): (1) Subjects check instruments; then adjust the seat height, sitting posture, and the distance to the screen. (2) 2 TEST videos occur sequentially in the screen center. Subjects observe and remember the target features in the first frame, then move the mouse to track the target in subsequent frames. (3) 17 FORMAL videos belonging to different groups are played sequentially (a rest time occurs between two videos to relieve visual fatigue); subjects should concentrate on the target and maintain the mouse's position. (4) Subjects fill in a questionnaire to self-evaluate their performance (Appendix C.5).

3.3 EVALUATION

This work evaluates humans and machines under the one-pass evaluation (OPE) mechanism, which initializes an executor in the first frame and continuously records the tracking results.

For the t-th frame F_t in a sequence $s_i = \{F_1, F_2, \ldots, F_t, \ldots\}$, traditional indicators usually select the positional relationship ($e.g.$, intersection over union (IoU) and center distance) between predicted

result p_t and ground-truth g_t to accomplish calculation. However, the tracking result of humans is a coordinate point, which cannot be calculated by IoU (Zheng et al. (2020); Rezatofighi et al. (2019)). Thus, all evaluation indicators in this work are based on two center points (*i.e.*, the target center c_p predicted by humans or algorithms, and the center point c_g of the ground-truth). Note that target absent is regarded as an empty set (*i.e.*, $g_t = \phi$) and excluded by the evaluation process.

We first select the classical *precision* score (RPE) (Wu et al. (2015)) and the recently proposed *normalized precision score* (N-PRE) (Hu et al. (2022b)) as basic indicators (please refer to Appendix D), then weighted them with sequence length. Obviously, longer sequences challenge all executors – humans demand a higher concentration of attention, and algorithms should overcome the cumulative errors. However, existing metrics (PRE and N-PRE) ignore this factor, causing us to weight original indicators with sequence length, and propose *Weighted-PRE* and *Weighted-N-PRE*:

$$\begin{aligned} \mathcal{P}_w(\mathcal{T}) &= \frac{1}{|\mathcal{T}|} \sum_{\mathcal{G} \in \mathcal{T}} \frac{\sum_{s_i \in \mathcal{G}} |\{F_t : d_c \leq \theta_d\}|}{\sum_{s_i \in \mathcal{G}} |s_i|} \\ \mathcal{P}^{'}{}_w(\mathcal{T}) &= \frac{1}{|\mathcal{T}|} \sum_{\mathcal{G} \in \mathcal{T}} \frac{\sum_{s_i \in \mathcal{G}} \left| \left\{ F_t : \mathcal{N}(d_c{}^{'}) \leq \theta_d{}^{'} \right\} \right|}{\sum_{s_i \in \mathcal{G}} |s_i|} \end{aligned} \tag{1}$$

where $|\cdot|$ is the cardinality, θ_d is a threshold to judge whether the tracking result is precise, $\mathcal{G}$ and $\mathcal{T}$ represent different video groups and tasks.

4 EXPERIMENTS

4.1 A COMPREHENSIVE COMPARISON OF HUMAN-MACHINE DYNAMIC VISION ABILITY

We first interest in the comprehensive performance of human-machine DVA. Figure 2 (I) shows the normalized precision score of executors in all sequences. We use Subject_Top and Subject_Bottom to represent the best and worst subjects, and Subject_Mean denotes the average. Clearly, humans score higher than most machines, indicating that the DVA of most machines is still far from humans. Besides, the boxplot distribution demonstrates that the performance of machines is more unstable. It is worth noting that the latest methods have been comparable to Subject_Bottom in most cases, indicating that SOTA methods do significantly close the gap with humans.

As tasks become more complicated (from short-term tracking to global instance tracking), executors not only need to perceive moving targets, but also should have a good cognitive ability to quickly re-locate the target when its position suddenly changes. Figure 2 (II-IV) illustrates the performance of human-machine DVA under various task constraints. Based on experimental results, we find that: (1) A typical short-term tracking task only needs perceptron ability. Thus, both humans and most machines perform well in tracking the moving target; some SOTA machines even outperform humans in accuracy and stability. (2) For the long-term tracking task, since the sequence length increases and the target disappearance is allowed during the tracking process, executors should locate the target when it reappears in the frame. Results demonstrate that humans can maintain high tracking precision, but the performance of most algorithms drops significantly. (3) Global instance tracking, the most difficult task that sequence length increases with shot-cuts and scene transitions, challenges almost all algorithms. However, humans can still sustain efficient and stable tracking, indicating that the cognitive ability of humans is much better than machines.

Specifically, the top three machines represent various model structures, helping us to understand the performance of different modeling processes. MixFormer (Cui et al. (2022)), a simple end-to-end model based on transformer structure, performs well in both short-term and long-time tracking tasks, indicating that the Mixed Attention Module (MAM) and a straightforward detection head can provide powerful tracking ability. The SiamRCNN (Voigtlaender et al. (2020)) model combines a two-stage scheme with a new trajectory-based dynamic planning algorithm, and uses re-detection to accomplish stable tracking. KeepTrack (Mayer et al. (2021)) model improved by SuperDiMP (Danelljan et al. (2020)) has an enhanced ability to discriminate interferers and perform well in short-time tracking, but lacking a mechanism for processing target mutation decreases its execution in longer sequences. However, all these three machines have a performance gap with humans in the global tracking task. Thus, we select MixFormer (Cui et al. (2022)) as a representation to accomplish more detailed analyses in the following sections.

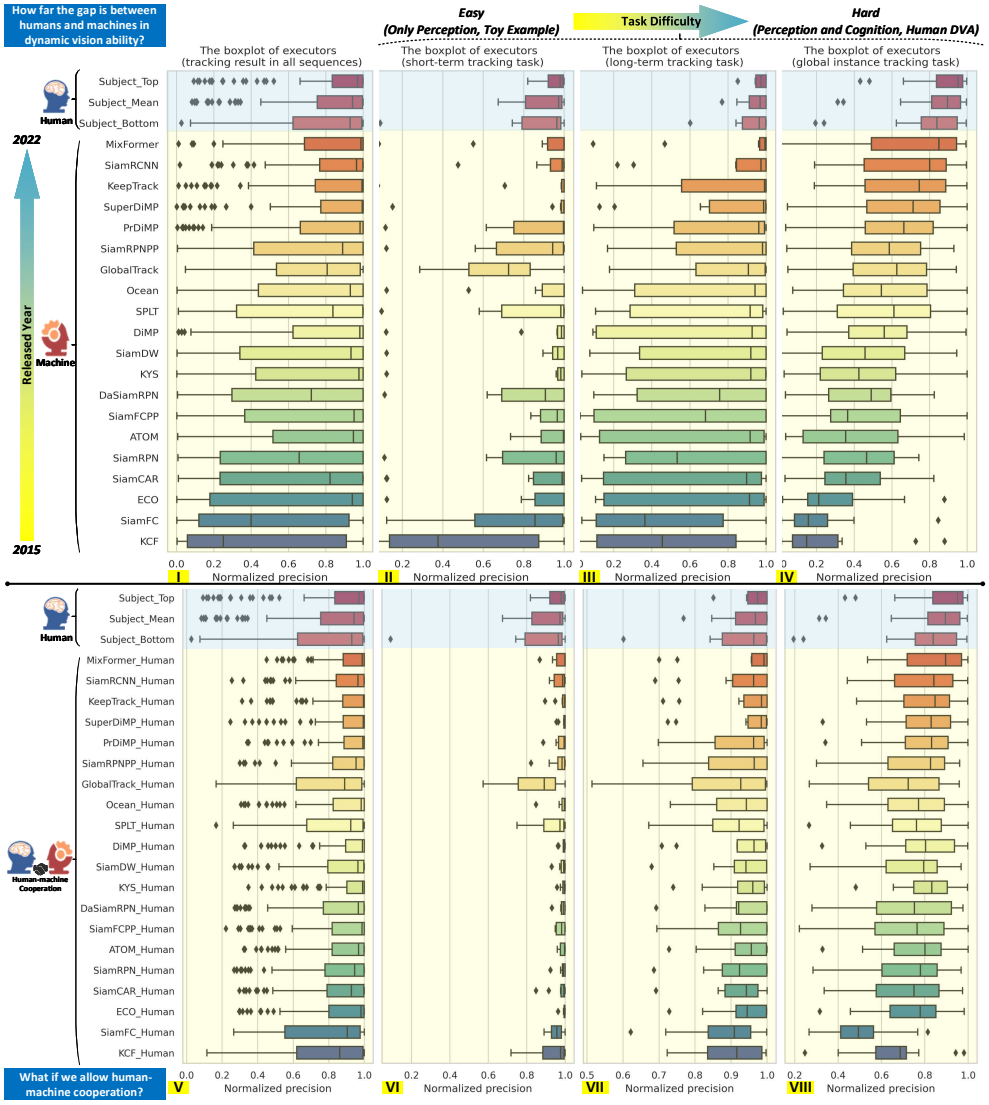


Figure 2: A comprehensive comparison and cooperation of human-machine dynamic vision ability. Boxplots enumerate the distribution of N-PRE scores for all tasks (I&V), short-term tracking task (II&VI), long-term tracking task (III&VII), and global instance tracking task (IV&VIII). The blue area represents the maximum, mean, and minimum scores of human subjects. The yellow area on top (I-IV) represents machines, while the bottom (V-VIII) represents human-machine collaborators.

Cognition. We further explore the effect of sequence length and shot-cuts, which requires the executor to have better cognition ability to re-locate the target in complex situations. Figure 10 in Appendix shows the performance variation of human-machine under different videos. Clearly, the SOTA machine's performance fluctuates wildly when the shot-cut occurs, while the subject can still maintain a stable tracking ability. In addition, machines may ultimately lose the target after frequent shot-cuts, but humans can quickly relocate the target in a new shot. When the sequence length increases, the performance fluctuation of machines will continue to increase, but humans are better than machines in terms of fluctuation amplitude and tracking effect. This result reveals that when the target motion or apparent information is mutated, the performance of machines fluctuates sharply, while humans can keep robust and precise target tracking based on their high cognition ability.

Perceptron. Given that multiple challenging factors may coexist in a video, we show the performance of human and SOTA machine under different factor combinations in Figure 3. Obviously, the purple rectangle represents short-term tracking with challenging factors (*i.e.*, target moves in a single camera

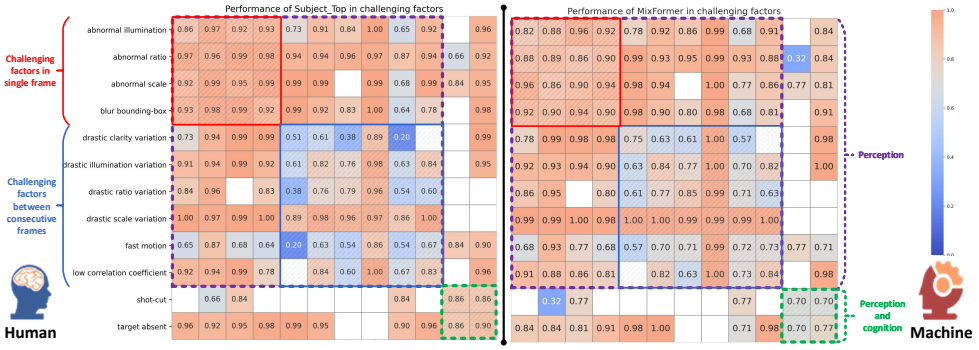


Figure 3: Performance of human and machine (MixFormer (Cui et al. (2022))) on challenging factors. Each cell represents a mixture of challenging factors, and the blank cells represent no such combination. Cell color represents the executor's mean score across all videos that satisfy the combination. The purple rectangle represents short-term tracking with challenging factors (target presents from beginning to end, related to perception), of which the red and blue rectangles represent the combination of static and dynamic challenging factors separately. The green rectangle represents task constraints (long-term tracking and global instance tracking, related to both perception and cognition). For detailed results, please refer to Appendix E.5.

without absence, mainly composed in groups H to Q) and is primarily related to the perceptive ability. The blue rectangle reflects the dynamic changes between consecutive frames. We can find that most cells in the blue rectangle are challenging for both algorithms and humans, but algorithms are slightly better than humans in some factors like fast motion and small targets. Conversely, task constraints (*i.e.*, allowing target absence or shot switching, mainly composed in groups C to G) represented by the green rectangle have less impact on humans, but it powerfully influences machines. This phenomenon is consistent with the above analysis of human-machine cognitive abilities.

Further analyses about humans. Former analyses indicate that *to err is human*. Thus, we provide detailed human performance analysis in Appendix E.6, which indicates that fast-moving targets and small targets are challenging to track, but tracking in a long sequence is not difficult for humans.

The above results show that current research, which mainly focuses on improving tracking robustness under short-term tracking tasks, has progressed. However, machines still have a gap with humans when task constraints are widened to a more general situation. Some bad cases are illustrated in Figure 11; more results based on different indicators are listed in Appendix E.3 to E.4.

4.2 Machine Performance Improvements When Cooperating with Humans

The analysis of Section 4.1 shows that humans and machines can adapt to different challenging factors, so an intuitive idea is *what if we allow human-machine cooperation?*

As shown in Figure 4, the cooperation mechanism is designed as follows: when machines fail to track, subjects can provide the current target position and restart machines from the failure frame. In this cooperation pattern, machines are responsible for performing relatively simple but repetitive tasks, while humans are responsible for supplementing information at critical moments.

To better illustrate the cooperation process, we use video G4 (a long video with 25,397 frames and 450 shot-cuts, describing a table tennis match with two players) as an example. The best human subject scores 0.964 since it is easy to locate a player consistently. However, when the shot is switched, the position variation of the target player (blue-clothed player) will challenge the SOTA algorithm. Thus, it quickly drifts to the interferer (black-clothed player) and keeps wrong tracking, causing the score to decrease to 0.558. But if we allow algorithms to cooperate with humans, when the human subject finds that the algorithm fails (such as drifting to the distractor), the subject will provide correct target position to the algorithm to avoid persistent mistracking. For our experiment in G4, the human subject provides target location information 23 times while the algorithm tracks the remaining 25,374 frames. Although humans only need to participate 0.1% (23/25,397) in cooperation, the human-machine collaborator MixFormer_Human can score 0.977, indicating that cooperation

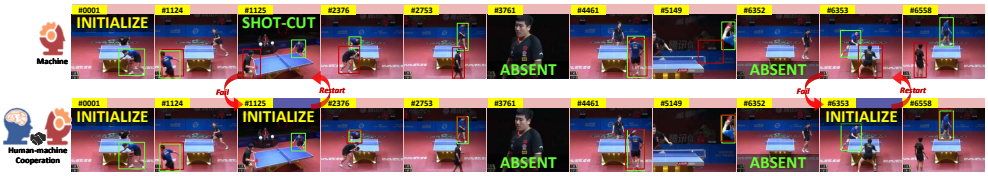


Figure 4: Example of the cooperation mechanism. The top area illustrates machine tracking results on sequence G4, and the bottom is the cooperation results. When machines fail to track (*e.g.*, #1125 and #6353, always due to target absent or shot-cut), subjects can provide the current target position and re-initialize machines to keep tracking from the failure frame.

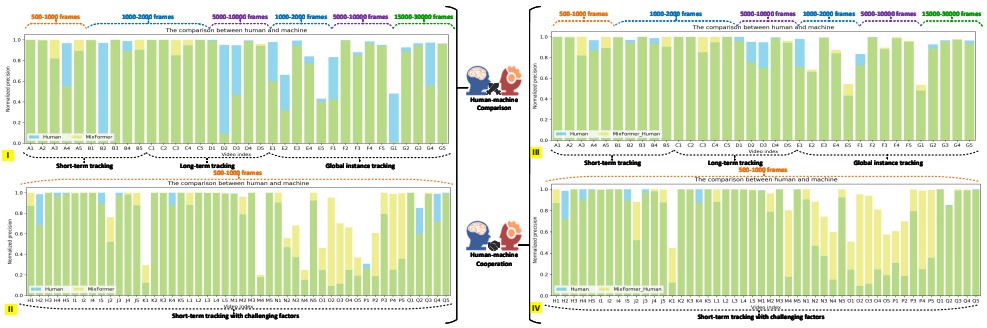


Figure 5: The comparison and cooperation of humans and machines, based on the normalized precision (N-PRE) score. (I-II) Performance of humans (blue bar) and machines (MixFormer (Cui et al. (2022)), yellow bar) on different sequences, and the green bar represents the overlapping part. (III-IV) Performance of humans (blue bar) and human-machine cooperations (MixFormer_Human, yellow bar) on different sequences, and the green bar represents the overlapping part.

can significantly improve the algorithm's tracking performance. This result is also in line with the original intention: algorithms accomplish many simple tasks, while humans focus on the key points.

Figure 5 illustrates the performance of humans, machines, and human-machine collaborators. Compared with machine itself, human-machine collaborator is greatly improved and even outperforms humans in most cases. Besides, subfigures of Figure 2 indicate that human-machine collaborators have significantly improved their performance based on the assistance of humans. Therefore, humans and machines have superb possibilities for cooperation in dynamic vision tasks. This cooperation not only combines strengths and decreases weaknesses, but also save human resources and improve efficiency. For detailed results based on different indicators, please refer to Appendix F.1 to F.3.

5 Conclusions and Future Work

In this paper, we aim to answer *how far the gap is between humans and machines in dynamic vision tasks,* and design an overall evaluation system from three aspects – we choose 87 videos to construct the experimental environment, select 20 representative algorithms and 15 human subjects to form the executors, and quantify their DVA with a strict evaluation process. Results indicate that SOTA machines are close to the lower limit of human subjects, meaning *the DVA gap between humans and machines has significantly closed*. Besides, humans and machines excel at different dynamic vision tasks, as humans can continuously track in general environments and longer sequences based on better cognitive ability, while SOTA machines can maintain efficient and stable tracking ability in short-term tracking with excellent perceptual ability. Finally, we also find that the human-machine cooperative executor can perform even better than humans, which provides a possibility of human-machine cooperation for application scenarios like self-driving systems. In the further, we will select a variety of human subjects (*e.g.*, subjects with different age, gender, and occupation (Burg & Hulbert (1961); Melcher & Lund (1992))) to compose a more comprehensive human baseline, and further explore the human-machine comparison and cooperation in the dynamic vision ability.

REFERENCES

RP Beals, AM Mayyasi, AE Templeton, and WL Johnston. The relationship between basketball shooting performance and certain visual attributes. *American Journal of Optometry and Archives of American Academy of Optometry*, 48(7):585–590, 1971.

Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. *CoRR*, abs/1606.09549, 2016. URL `http://arxiv.org/abs/1606.09549`.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. *CoRR*, abs/1904.07220, 2019. URL `http://arxiv.org/abs/1904.07220`.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 205–221, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58592-1.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Albert Burg. Visual acuity as measured by dynamic and static tests: a comparative evaluation. *Journal of applied psychology*, 50(6):460, 1966.

Albert Burg and Slade Hulbert. Dynamic visual acuity as related to age, sex, and static acuity. *Journal of Applied Psychology*, 45(2):111, 1961.

Alan HS Chan and Alan J Courtney. Foveal acuity, peripheral acuity and search performance: A review. *International Journal of Industrial Ergonomics*, 18(2-3):113–119, 1996.

Gioele Ciaparrone, Francisco Luque Sanchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381: 61–88, 2019.

Yutao Cui, Jiang Cheng, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6931–6939, July 2017. doi: 10.1109/CVPR.2017.733.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: accurate tracking by overlap maximization. *CoRR*, abs/1811.07628, 2018. URL `http://arxiv.org/abs/1811.07628`.

Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Heiko Dankert, Liming Wang, Eric D Hoopfer, David J Anderson, and Pietro Perona. Automated monitoring and analysis of social behavior in drosophila. *Nature Methods*, 6(4):297–303, 2009.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, and Katja Hofmann. Navigation turing test (ntt): Learning to evaluate human-like navigation. In *International Conference on Machine Learning*, pp. 2644–2653. PMLR, 2021.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, Yong Xu, Chunyuan Liao, Lin Yuan, and Haibin Ling. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 2020.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34, 2021.

Brian Q. Geuther, Sean P. Deats, Kai J. Fox, Steve A. Murray, Robert E. Braun, Jacqueline K. White, Elissa J. Chesler, Cathleen M. Lutz, and Vivek Kumar. Robust mouse tracking in complex environments using neural networks. *Communications Biology*, 2(1):124, 2019.

Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6269–6277, 2020.

J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, March 2015. doi: 10.1109/TPAMI.2014.2345390.

Kenneth Holmqvist, Saga Lee Örbom, Ignace TC Hooge, Diederick C Niehorster, Robert G Alexander, Richard Andersson, Jeroen S Benjamins, Pieter Blignaut, Anne-Marie Brouwer, Lewis L Chuang, et al. Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior research methods*, pp. 1–53, 2022.

Shiyu Hu, Xin Zhao, and Kaiqi Huang. Sotverse: A user-defined task space of single object tracking. *CoRR*, abs/2204.07414, 2022a. doi: 10.48550/arXiv.2204.07414. URL https://doi.org/10.48550/arXiv.2204.07414.

Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022b. doi: 10.1109/TPAMI.2022.3153312.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. *arXiv: Computer Vision and Pattern Recognition*, 2019.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (5):1562–1577, 2021. doi: 10.1109/TPAMI.2019.2957464.

David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

MILLER JW et al. The effect of relative motion on visual acuity. *Survey of Ophthalmology*, 7: 83–116, 1962.

P Kohl, B Coffey, A Reichow, W Thompson, and P Willer. A comparative study of visual performance in jet fighter pilots and non-pilots. *Journal of Behavioral Optometry*, 5(2):123–126, 1991.

Matej Kristan, Roman Pflugfelder, Aleš Leonardis, Jiri Matas, and et al. Fatih Porikli. The visual object tracking vot2013 challenge results, Dec 2013.

Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, and et al. Čehovin, Luka. The visual object tracking vot2014 challenge results. 09 2014. doi: 10.13140/2.1.4192.9921.

Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, and et al. Čehovin, Luka. The visual object tracking vot2015 challenge results. In *Visual Object Tracking Workshop 2015 at ICCV2015*, Dec 2015.

Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, and et al. Luka Čehovin Zajc. The visual object tracking vot2016 challenge results. Springer, Oct 2016. URL http://www.springer.com/gp/book/9783319488806.

Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, and et al. Vojir, Tomas. The visual object tracking vot2017 challenge results. pp. 1949–1972, 10 2017. doi: 10.1109/ICCVW.2017.230.

Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, and et al. Roman Pfugfelder. The sixth visual object tracking vot2018 challenge results, 2018.

Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, and et al. Roman Pflugfelder. The seventh visual object tracking vot2019 challenge results, 2019.

Michael F Land and Peter McLeod. From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, 3(12):1340–1345, 2000.

Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. *Advances in Neural Information Processing Systems*, 34, 2021.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *CoRR*, abs/1812.11703, 2018a. URL http://arxiv.org/abs/1812.11703.

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Gerald M Long and David L Penn. Dynamic visual acuity: normative functions and practical implications. *Bulletin of the Psychonomic Society*, 25(4):253–256, 1987.

Gerald M Long and Merrill J Zavod. Contrast sensitivity in a dynamic environment: Effects of target conditions and visual impairment. *Human Factors*, 44(1):120–132, 2002.

Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13424–13434, 2021. doi: 10.1109/ICCV48922.2021.01319.

MH Melcher and DR Lund. Sports vision and the high school student athlete. *Journal of the American Optometric Association*, 63(7):466–474, 1992.

Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 310–327, 2018.

Siddhartha Sankar Nath, Girish Mishra, Jajnyaseni Kar, Sayan Chakraborty, and Nilanjan Dey. A survey of image classification methods and techniques. In *2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT)*, pp. 554–557. IEEE, 2014.

Zenon W Pylyshyn and Ron W Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

Lluïsa Quevedo, José Antonio Aznar-Casanova, Dolores Merindano-Encina, Genís Cardona, and Joan Solé-Fortó. A novel computer software for the evaluation of dynamic visual acuity. *Journal of Optometry*, 5(3):131–138, 2012.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Philip L Smith and Daniel R Little. Small is beautiful: In defense of the small-n design. *Psychonomic bulletin & review*, 25(6):2083–2101, 2018.

Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pp. 23–65. Springer, 2009.

Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6578–6588, 2020.

Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 557–573, Cham, 2018. Springer International Publishing.

Kunlin Wei and Konrad Paul Kording. Behavioral tracking gets real. *Nature Neuroscience*, 21(9): 1146–1147, 2018.

Aharon Weissbrod, Alexander Shapiro, Genadiy Vasserman, Liat Edry, Molly Dayan, Assif Yitzhaky, Libi Hertzberg, Ofer Feinerman, and Tali Kimchi. Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment. *Nature Communications*, 4(1):2018, 2013.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, 2013.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

Chen Xia, Junwei Han, and Dingwen Zhang. Evaluation of saccadic scanpath prediction: Subjective assessment database and recurrent neural network based metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4378–4395, 2020.

Chen Xia, Junwei Han, and Dingwen Zhang. Evaluation of saccadic scanpath prediction: Subjective assessment database and recurrent neural network based metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4378–4395, 2021. doi: 10.1109/TPAMI.2020.3002168.

Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12549–12556, 2020.

Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2385–2393, 2019.

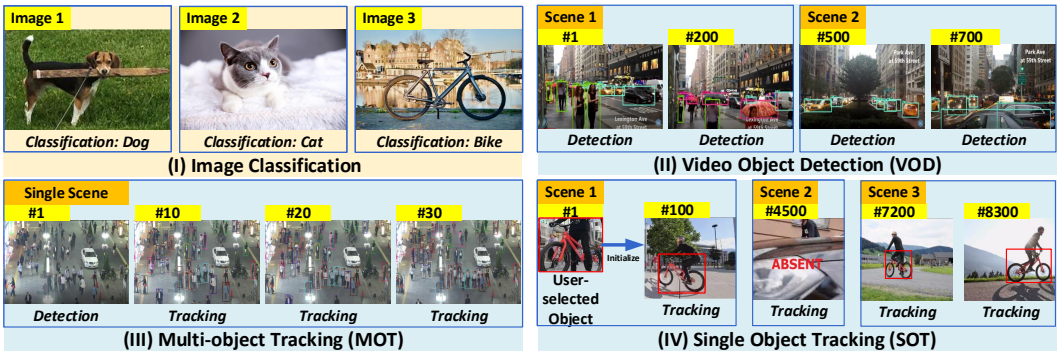
Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4591–4600, 2019.

Zhipeng Zhang and Houwen Peng. Ocean: Object-aware anchor-free tracking. In *ECCV (21)*, pp. 771–787, 2020.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12993–13000, 2020.

Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. *CoRR*, abs/1808.06048, 2018. URL `http://arxiv.org/abs/1808.06048`.

A TASK DESCRIPTION



(a) The execution flow of classical computer vision tasks (Image Classification (I), Video Object Detection (II), Multi-object Tracking (III), and Single Object Tracking (IV)).

Task	Visual Ability	Object		Scenario		Arbitrary Instance	Random Scenario
		Object Category	Obejct Selection	Motion Consistency Assumption	Scene-changing		
Image Classification	SVA	Limited by training set				N	
Video Object Detection	DVA	Limited by training set	Detected	N	Y	N	Y
Multi-object Tracking		Limited, usually pedestrians or vehicles	Detected in the first frame	N	N	N	N
Single Object Tracking		Arbitrary	User-specified in the first frame	N	Y	Y	Y

(b) The comparison of task characteristics.

Figure 6: The execution flow of classical computer vision tasks (a) and comparison of task characteristics (b). Obviously, SOT is closer to human dynamic vision ability.

B LIST OF DATASETS AND MACHINES

Table 2: The URL of open-sourced benchmarks used in this work.

Benchmarks	URL
OTB (Wu et al. (2013; 2015))	http://cvlab.hanyang.ac.kr/tracker_benchmark/index.html
VOT (Kristan et al. (2013; 2014; 2015; 2016; 2017; 2018; 2019))	https://votchallenge.net/
GOT-10k (Huang et al. (2021))	http://got-10k.aitestunion.com
LaSOT (Fan et al. (2020))	https://cis.temple.edu/lasot/
VideoCube (Hu et al. (2022b))	http://videocube.aitestunion.com

We have correctly cited all benchmarks involved in the experiments and carefully checked the licenses involved in each benchmark (CC BY-NC-SA 4.0.). All datasets used in the experiments are open-sourced, and our usage is under the license scope.

Table 3: The URL of open-sourced algorithms used in this work.

Machines	URL
KCF (Henriques et al. (2015))	https://github.com/uoip/KCFpy
SiamFC (Bertinetto et al. (2016))	https://github.com/huanglianghua/siamfc-pytorch
SiamRPN (Li et al. (2018b))	https://github.com/huanglianghua/siamrpn-pytorch
ECO (Danelljan et al. (2017))	https://github.com/visionml/pytracking
e DaSiamRPN (Zhu et al. (2018))	https://github.com/foolwood/DaSiamRPN
ATOM (Danelljan et al. (2018))	https://github.com/visionml/pytracking
SiamRPN++ (Li et al. (2018a))	https://github.com/PengBoXiangShang/SiamRPN_plus_plus_PyTorch
SPLT (Yan et al. (2019))	https://github.com/iiau-tracker/SPLT
DiMP (Bhat et al. (2019))	https://github.com/visionml/pytracking
GlobalTrack (Huang et al. (2019))	https://github.com/huanglianghua/GlobalTrack
SiamDW (Zhang & Peng (2019))	https://github.com/researchmm/TracKit
SiamFC++ (Xu et al. (2020))	https://github.com/MegviiDetection/video_analyst
SiamCAR (Guo et al. (2020))	https://github.com/ohhhyeahhh/SiamCAR
PrDiMP (Danelljan et al. (2020))	https://github.com/visionml/pytracking
SuperDiMP (Danelljan et al. (2020))	https://github.com/visionml/pytracking
SiamRCNN (Voigtlaender et al. (2020))	https://github.com/VisualComputingInstitute/SiamR-CNN
Ocean (Zhang & Peng (2020))	https://github.com/researchmm/TracKit
KYS (Bhat et al. (2020))	https://github.com/visionml/pytracking
KeepTrack (Mayer et al. (2021))	https://github.com/visionml/pytracking
MixFormer (Cui et al. (2022))	https://github.com/MCG-NJU/MixFormer

Our algorithm evaluation experiments are performed on a server with 4 NVIDIA TITAN RTX GPUs and a 64 Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz. We use the parameters provided by the original authors. For detailed information about each executor, please refer to the original introduction via the likes.

C EXPERIMENT ORGANIZATION

C.1 EXPERIMENTAL ENVIRONMENT

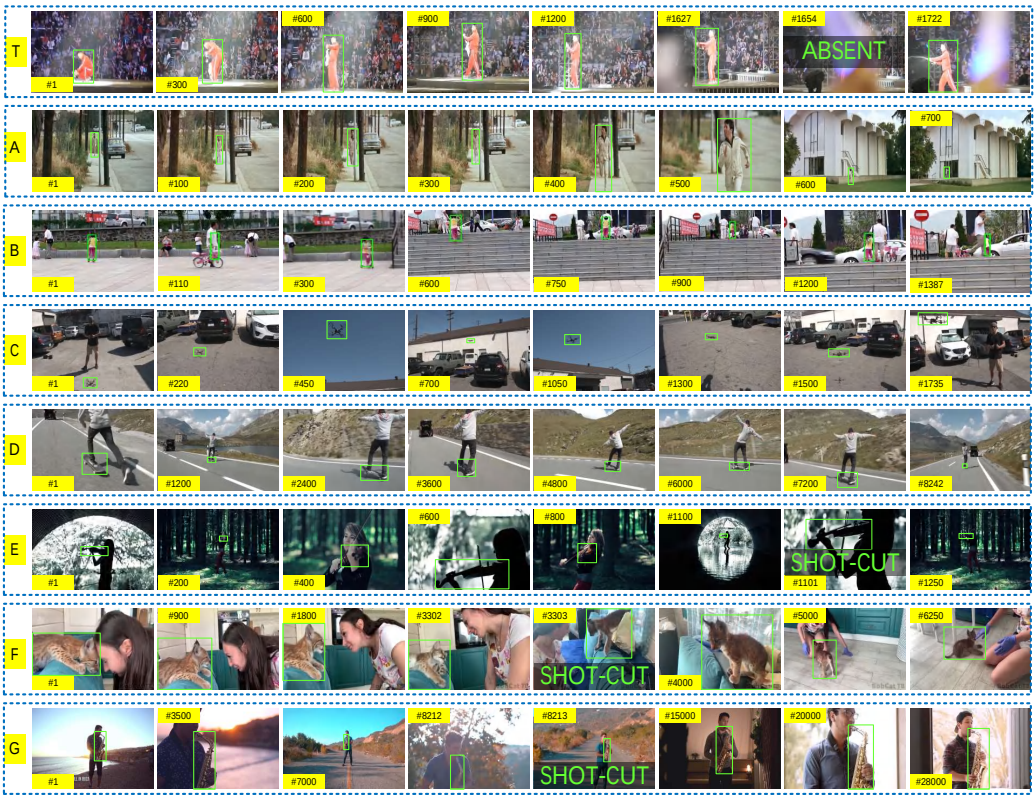


Figure 7: Representative sequences used in experiments. (T) Example of TEST videos, which help the experimenter familiar with the operation. (A) Example of short-term tracking videos, the sequence length of this group is concentrated in 500-1000 frames. (B) Example of short-term tracking videos, the sequence length of this group is concentrated in 1000-2000 frames. (C) Example of long-term tracking videos, the sequence length of this group is concentrated in 1000-2000 frames. (D) Example of long-term tracking videos, the sequence length of this group is concentrated in 5000-10000 frames. (E) Example of global instance tracking videos, the sequence length of this group is concentrated in 1000-2000 frames. (F) Example of global instance tracking videos, the sequence length of this group is concentrated in 5000-10000 frames. (G) Example of global instance tracking videos, the sequence length of this group is concentrated in 15000-30000 frames.

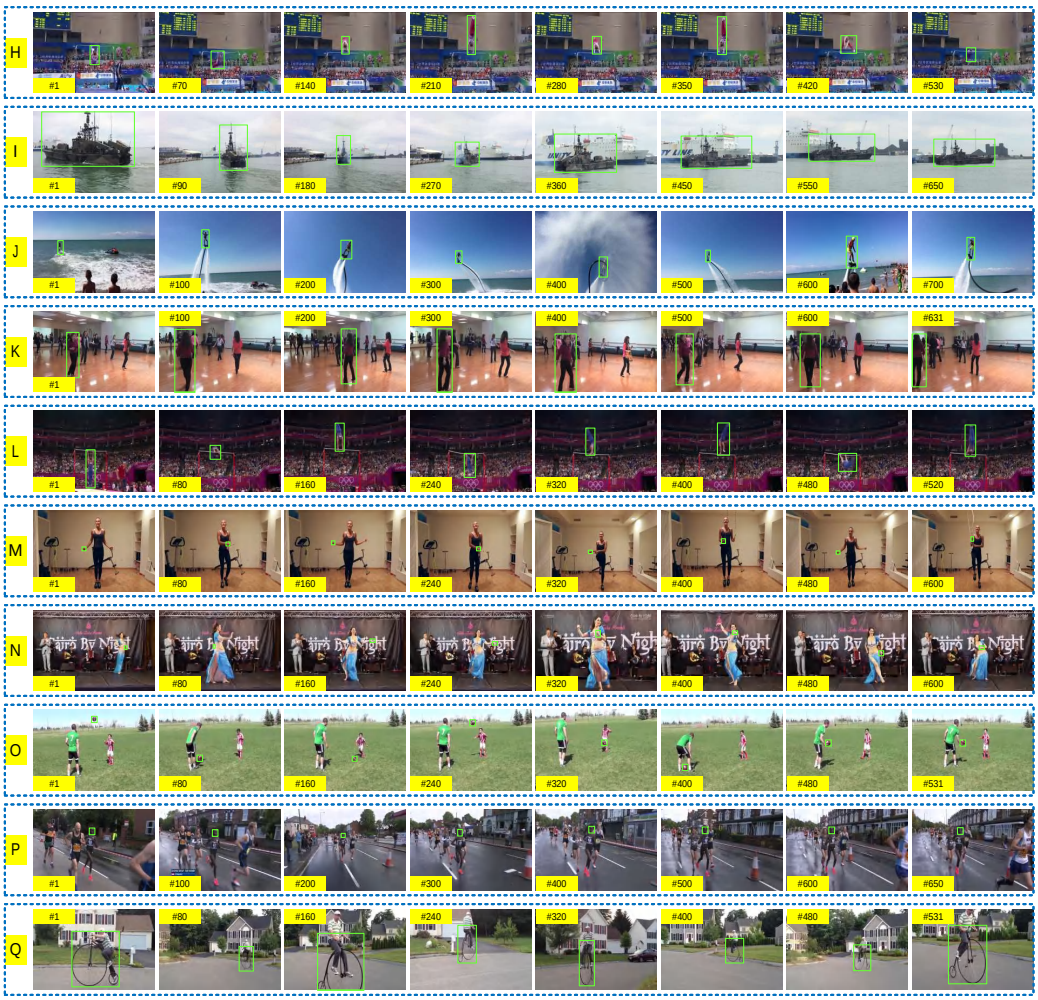


Figure 8: Representative sequences used in experiments, the sequence length in this part is concentrated in 500-1000 frames. (H) Example of tracking the target with abnormal ratio. (I) Example of tracking the target with abnormal scale. (J) Example of tracking the target with abnormal illumination. (K) Example of tracking the target with blur bounding-box. (L) Example of tracking the target with drastic ratio variation. (M) Example of tracking the target with drastic scale variation. (N) Example of tracking the target with drastic illumination variation. (O) Example of tracking the target with drastic clarity variation. (P) Example of tracking the target with fast motion. (Q) Example of tracking the target with low correlation coefficient.

C.2 EXPERIMENTAL STATEMENT

We have obtained the approvals of all experiments, the template is provided by our IRB. Every experiment has signed an experimental statement before the experiment. Here we list the contents:

Purpose *You have been asked to participate in a research study that studies the dynamic visual abilities of humans. We would like your permission to enroll you as a participant in this research study. The instruments involved in the experiment are a computer screen and a mouse. The experimental task consisted of gazing at the presented image sequence and manipulating the mouse to point as consistently as possible at the center of the moving target. You will be given specific instructions for the task before it begins.*

Procedure *In this study, you should read the experimental instructions and ensure that you understand the experimental content. The whole experiment process lasts about one hour, and the experiment is divided into the following steps:*

- *Read and sign the experimental statement;*
- *Test the experimental instrument, and adjust the seat height, sitting posture, and the distance between your eyes and the screen. Please ensure that you are in a comfortable sitting position during the experiment;*
- *Two TEST videos will be played. You should comprehend the specific instrument operation rules and be familiar with the experimental process through the test videos;*
- *Start the FORMAL experiment. Please follow the operation to watch the test sequence and complete the relevant operations. Note that after watching each video sequence, you should have a rest;*
- *After the experiment, you need to fill in a questionnaire.*

Risks and Discomforts *The only potential risk factor for this experiment is trace electron radiation from the computer. Relevant studies have shown that radiation from computers and related peripherals will not cause harm to the human body. To rule out the impact of COVID-19 on the experiment, all participants completed nucleic acid tests before the experiment to ensure their health. Everyone needs to wear a mask throughout the experiment. In addition, the relevant devices are cleaned and disinfected during the experiment.*

Costs *Each participant who completes the experiment will be paid 200 RMB.*

Confidentiality *The results of this study may be published in an academic journal/book or used for teaching purposes. However, your name or other identifiers will not be used in any publication or teaching materials without your specific permission. In addition, if photographs, audio tapes or videotapes were taken during the study that would identify you, then you must give special permission for their use.*

I confirm that the purpose of the research, the study procedures and the possible risks and discomforts as well as potential benefits that I may experience have been explained to me. All my questions have been satisfactorily answered. I have read this consent form. My signature below indicates my willingness to participate in this study.

C.3 DEVICE SELECTION

How to measure the DVA of humans is an important section of our work. Previous works mainly use eye-tracker (Hu et al. (2022b); Xia et al. (2021)) or mouse (Geirhos et al. (2018; 2020; 2021)) to record human visual ability. To select a better measurement method, we divide them into active (mouse) and passive (eye-tracker) methods, then compare them from practical and theoretical views:

- From a practical view, we have organized all subjects to experience both mouse and eye-tracker, and they all express that the user experience of the mouse is better. A possible reason is that the eye-tracker passively records eye movements, causing the effect to be limited by many factors (*e.g.*, the device accuracy, the user's posture, the relative distance between eyes and the eye-tracker, and the user's glasses). Besides, many subjects mention that the eye-tracker can only effectively track eye movements for a short period. When the experimental time is prolonged, the recording result easily shifts, while subjects can only adjust their posture rather than actively correct the capture result. Conversely, using a mouse can minimize the above problems. Although mouse may also have some limitations, its accuracy is greater than eye-tracker in our experiments. For example, when subjects notice that the mouse position is far from the target, they can actively move the mouse and immediately correct the record result.
- From a theoretical view, some researchers (Holmqvist et al. (2022)) also indicate that the eye-tracker is affected by the tracking ratio (*i.e.*, the amount of eye-tracking data lost). Blinks or brief vision drifts can cause a lower tracking ratio, which requires the subject to keep a high degree of concentration when using eye-trackers. To prove this view, we check recent articles and find that experiments based on eye-trackers are either for observing the images (subjects have sufficient time to adjust the position per image) (Xia et al. (2020)) or watching a small amount of video (*e.g.*, 6 videos) (Hu et al. (2022b)). Nevertheless, our experiment requests each participant to watch 17 videos (including long videos). Thus, while maintaining a high concentration of attention, subjects frequently change their sitting position unconsciously, causing the eye-tracker performs poorly in our experiment. What's more, we find that previous works (Geirhos et al. (2018; 2020; 2021)) have allowed subjects to perform image classification tasks with a mouse-based method, and a neuro training company named Reflexion also uses touch screens to provide DVA training for athletes, which means recording the observation position based on hand operation has been applied in academia and industry. Thus, we finally select the mouse-based experimental method.

C.4 EXPERIMENT PROCESS

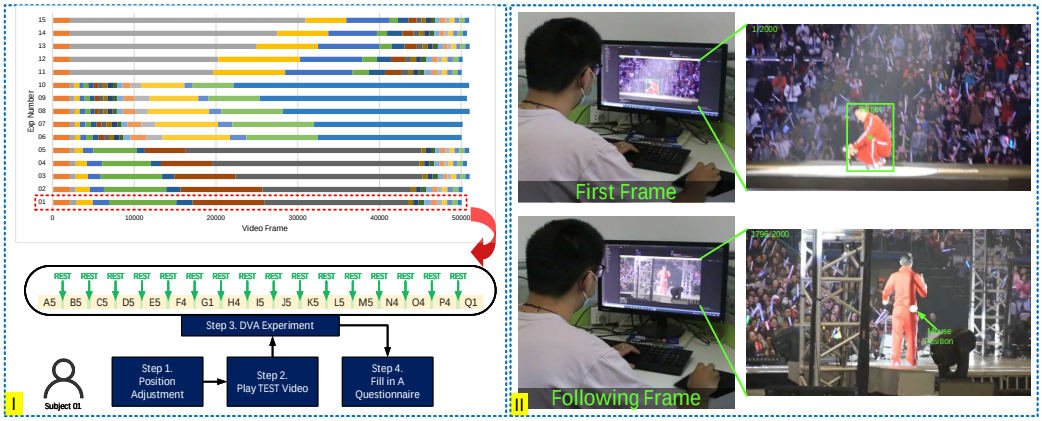


Figure 9: Process of DVA experiment. (I) 15 subjects participated in the DVA experiment. For each video number, the first letter represents the video group, and the second letter represents the index. Each group contains 5 videos (*e.g.*, A1-A5), and each video is observed by 3 subjects (*e.g.*, A1 is observed by subjects 05, 10, 15, respectively). Subjects whose serial numbers differ by 5 should watch the same content but in a different order. (II) The experiment includes two steps: observing and remembering the target features in the first frame, then moving the mouse to track the target in subsequent frames. Note that subject has a rest time to relieve visual fatigue between two videos.

C.5 QUESTIONNAIRE

Subjects should fill in a questionnaire to evaluate their performance and make recommendations. Here we list the contents:

This questionnaire aims to investigate subjects' opinions and self-evaluation.

Q1. Your name:

Q2. Your ID (e.g., subject 1):

Q3. Your age:

Q4. Your vision condition (fill in the left and right eyes separately):

Q5. Please rank the following factors in order of perceived difficulty during your experiment (order from difficult to easy):

- *Video sequence is too long;*
- *Frequent shot changes;*
- *Screen shaking;*
- *Target scale is too small;*
- *Similar objects interference;*
- *Target moves fast;*
- *Target is occluded.*

Q6. In addition to the challenging factors in the previous question, what other factors may influence your observation?

Q7. Please provide your self-evaluation score (0 to 100) in short sequences (1-3 minutes):

Q8. Please provide your self-evaluation score (0 to 100) in medium sequences (3-10 minutes):

Q9. Please provide your self-evaluation score (0 to 100) in long sequences (more than 10 minutes):

Q10. What was the most impressive sequence in the experiment? Why are you impressed by it? (e.g., the video is too long, the content of the video is exceptional, the video has some challenging factors that lead to poor performance, etc.)

Q11. Do you have any suggestions or comments for this experiment?

C.6 GROUP INFORMATION

Table 4: Groups of subjects and corresponding videos. Subjects indicated by the same color watch the same content but in a different order.

(a) Each human subject will watch 19 videos, including 2 TEST videos and 17 FORMAL experiment videos. For each video number, the first letter represents the video group, and the second letter represents the number of videos in that group. Except for the TEST videos, each group contains five videos (e.g., A1-A5), and each video is observed by three subjects (e.g., A1 is observed by subjects 05, 10, and 15, respectively).

Group	**I**					**II**					**III**				
Exp	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1
	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2
	A5	A2	A3	A4	A1	H4	H5	H3	H2	H1	G1	G2	G3	G4	G5
	B5	B4	B3	B2	B1	I5	I1	I2	I3	I4	F4	F5	F3	F2	F1
	C5	C3	C2	C4	C1	J5	J4	J3	J2	J1	D5	D4	D3	D2	D1
	D5	D4	D3	D2	D1	K5	K4	K3	K2	K1	E5	E4	E3	E2	E1
	E5	E4	E3	E2	E1	L5	L4	L1	L2	L3	C5	C3	C2	C4	C1
	F4	F5	F3	F2	F1	M5	M4	M1	M2	M3	B5	B4	B3	B2	B1
	G1	G2	G3	G4	G5	N4	N3	N5	N1	N2	A5	A2	A3	A4	A1
Video	H4	H5	H3	H2	H1	O4	O5	O1	O2	O3	H4	H5	H3	H2	H1
	I5	I1	I2	I3	I4	P4	P5	P1	P2	P3	I5	I1	I2	I3	I4
	J5	J4	J3	J2	J1	Q1	Q2	Q3	Q4	Q5	J5	J4	J3	J2	J1
	K5	K4	K3	K2	K1	A5	A2	A3	A4	A1	K5	K4	K3	K2	K1
	L5	L4	L1	L2	L3	B5	B4	B3	B2	B1	L5	L4	L1	L2	L3
	M5	M4	M1	M2	M3	C5	C3	C2	C4	C1	M5	M4	M1	M2	M3
	N4	N3	N5	N1	N2	D5	D4	D3	D2	D1	N4	N3	N5	N1	N2
	O4	O5	O1	O2	O3	E5	E4	E3	E2	E1	O4	O5	O1	O2	O3
	P4	P5	P1	P2	P3	F4	F5	F3	F2	F1	P4	P5	P1	P2	P3
	Q1	Q2	Q3	Q4	Q5	G1	G2	G3	G4	G5	Q1	Q2	Q3	Q4	Q5

(b) Subjects in group I watch the video length from short to long, then become short again. Subjects in group II watch videos of increasing length. Subjects in group III watch videos of decreasing length.

Group	**I**					**II**					**III**				
Exp	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
	844	700	708	725	610	631	650	630	581	528	17592	18166	22872	25379	28828
	1981	1741	1500	1377	1000	651	570	571	621	645	8750	10000	7500	6250	5000
	1961	1735	1568	1865	1266	701	691	681	600	531	8242	7628	7509	5933	5307
	8242	7628	7509	5933	5307	806	671	631	631	551	2000	1750	1500	1250	1000
	2000	1750	1500	1250	1000	741	688	520	598	671	1961	1735	1568	1865	1266
	8750	10000	7500	6250	5000	659	601	510	531	550	1981	1741	1500	1377	1000
	17592	18166	22872	25379	28828	601	601	585	556	571	844	700	708	725	610
Frame	631	650	630	581	528	601	671	501	531	581	631	650	630	581	528
	651	570	571	621	645	670	681	601	620	650	651	570	571	621	645
	701	691	681	600	531	501	501	531	542	561	701	691	681	600	531
	806	671	631	631	551	844	700	708	725	610	806	671	631	631	551
	741	688	520	598	671	1981	1741	1500	1377	1000	741	688	520	598	671
	659	601	510	531	550	1961	1735	1568	1865	1266	659	601	510	531	550
	601	601	585	556	571	8242	7628	7509	5933	5307	601	601	585	556	571
	601	671	501	531	581	2000	1750	1500	1250	1000	601	671	501	531	581
	670	681	601	620	650	8750	10000	7500	6250	5000	670	681	601	620	650
	501	501	531	542	561	17592	18166	22872	25379	28828	501	501	531	542	561
Total	50052	50165	51038	50710	50970	50052	50165	51038	50710	50970	50052	50165	51038	50710	50970

D CALCULATION PROCESS OF INDICATORS

Precision (PRE). Traditional *precision* score is calculated by:

$$\mathcal{P}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{G} \in \mathcal{T}} \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} \left| \{ F_t : d_c \leq \theta_d \} \right| \tag{2}$$

where $|\cdot|$ is the cardinality, θ_d is a threshold to judge whether the tracking result is precise. The precision score of s_i is defined as the proportion of frames whose center distance $d_c = \|c_p - c_g\|_2 \leq \theta_d$. Calculating the mean value of each sequence s_i under group $\mathcal{G}$ and the mean value of each group under task $\mathcal{T}$ can generate the final precision score $\mathcal{P}(\mathcal{T})$. To illustrate the performance under different thresholds, previous works (Wu et al. (2015); Fan et al. (2020); Müller et al. (2018)) usually draw the statistical results based on different θ_d into a curve named *precision plot*. Typically, $\theta_d = 20$ is wildly used to rank trackers.

Normalized precision (N-PRE). Recent work (Hu et al. (2022b)) indicates that the PRE score ignores the influence of the target scale, and provides a *normalized precision score* named N-PRE to solve this problem. Trackers with a predicted center outside the ground-truth rectangle will add a penalty item ${d_c}^p$ (*i.e.*, the shortest distance between center point c_p and the ground-truth edge). For trackers whose center point falls into the ground-truth rectangle, the center distance ${d_c}^{'}$ equals the original precision d_c (*i.e.*, ${d_c}^p = 0$). Besides, to exclude the influence of target size and frame resolution, N-PRE selects the maximum value in frame F_t to normalize the result. The calculation can be summarized as:

$$\begin{aligned} \mathcal{P}^{'}(\mathcal{T}) &= \frac{1}{|\mathcal{T}|} \sum_{\mathcal{G} \in \mathcal{T}} \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} \left| \left\{ F_t : \mathcal{N}({d_c}^{'}) \leq {\theta_d}^{'} \right\} \right|, {\theta_d}^{'} \in [0, 1] \\ \mathcal{N}({d_c}^{'}) &= \frac{{d_c}^{'} - \min(\{{d_i}^{'} | i \in F_t\})}{\max(\{{d_i}^{'} | i \in F_t\}) - \min(\{{d_i}^{'} | i \in F_t\})} \end{aligned} \tag{3}$$

Draw statistical results based on different ${\theta_d}^{'}$ into a curve generates the *normalized precision plot*. Particularly to overcome the influence of threshold selection, the proportion of frames whose predicted results successfully fall in the ground-truth rectangle is used to rank trackers.

E PERFORMANCE OF HUMANS AND MACHINES

E.1 RESULTS ON DIFFERENT SEQUENCE LENGTHS

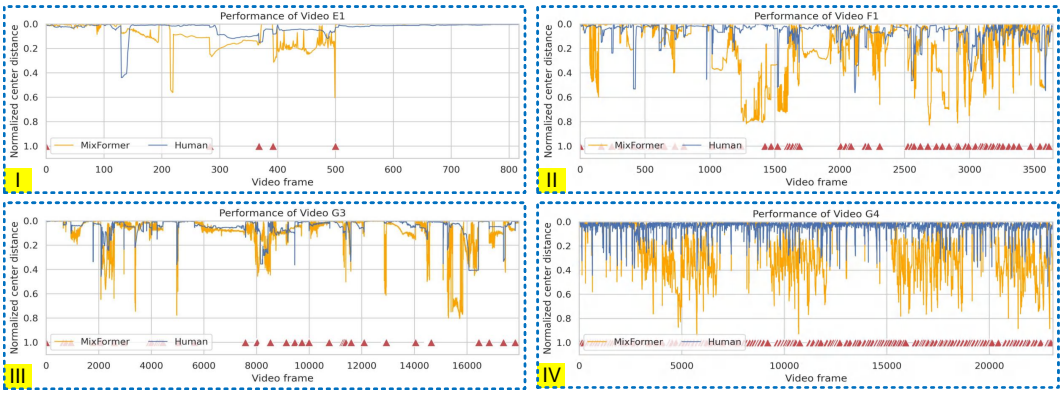


Figure 10: Performance of human-machine on different sequence lengths. The yellow curve represents the SOTA machine (MixFormer (Cui et al. (2022))), the blue curve represents the subject, and the red triangle represents the shot-cut. Subfigures (I-IV) represent four videos with different combinations of lengths and shots. Compared with the SOTA machine, humans can still maintain adequate DVA in the face of lengthy sequences and frequent shot switching.

E.2 EXAMPLES ON DIFFERENT SEQUENCES

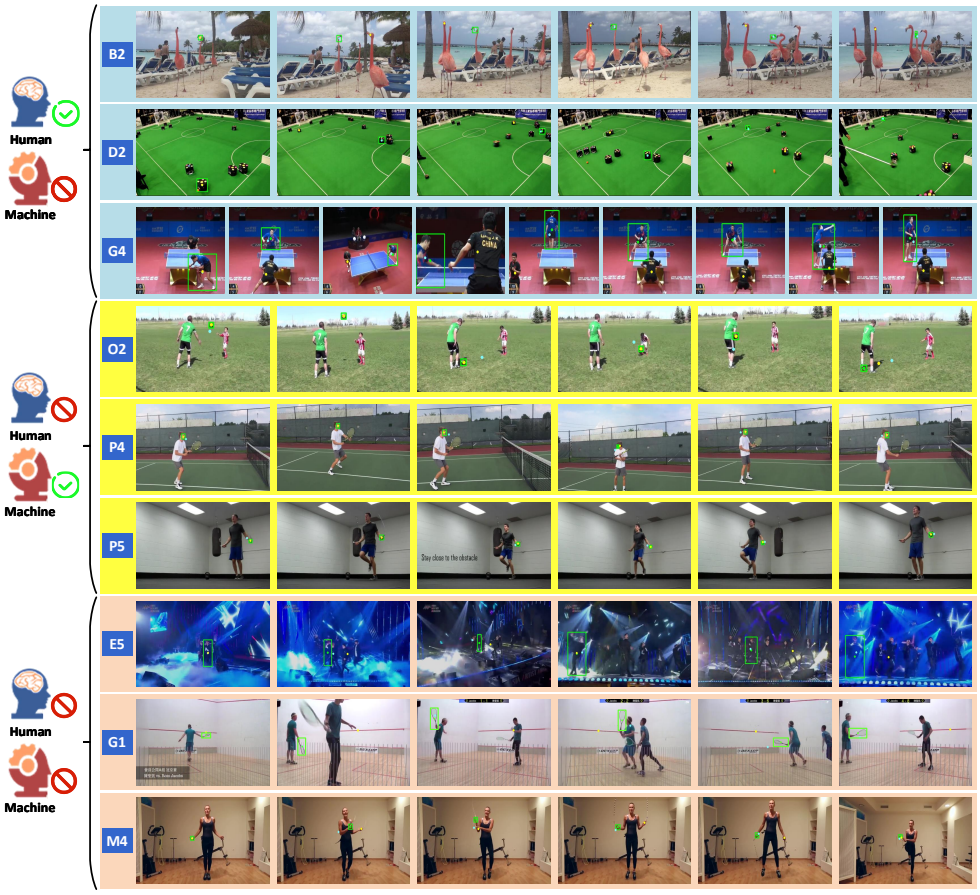


Figure 11: Schematic diagram of human-machine tracking results on partial sequences. The blue sequences (B2, D2, G4) represent that humans are better than machines, the yellow sequences (O2, P4, P5) indicate that machines are better than humans, and the red sequences (E5, G1, M4) indicate that both perform poorly. For each frame, the green rectangle and green point represent the ground-truth, the blue point represents the localization result of humans, and the yellow point represents the tracking result of machines.

Obviously, humans outperform machines in longer sequences with shot-cuts and can better distinguish the target from interfering objects (B2, D2, G4). However, when facing challenges such as tracking small targets with fast movements (O2, P4, P5), limited by hand-brain coordination and mouse movement speed, humans cannot precisely locate such targets at 25FPS. Note that both humans and machines may fail when multiple challenges are superimposed in one sequence (E5, G1, M4).

E.3 BOXPLOTS

Figure 12: Boxplots enumerate the distribution of N-PRE scores for TEST videos and Group A to Q. The blue area represents the maximum, mean, and minimum scores of human subjects, and the yellow area represents machines.

E.4 PRECISION PLOTS

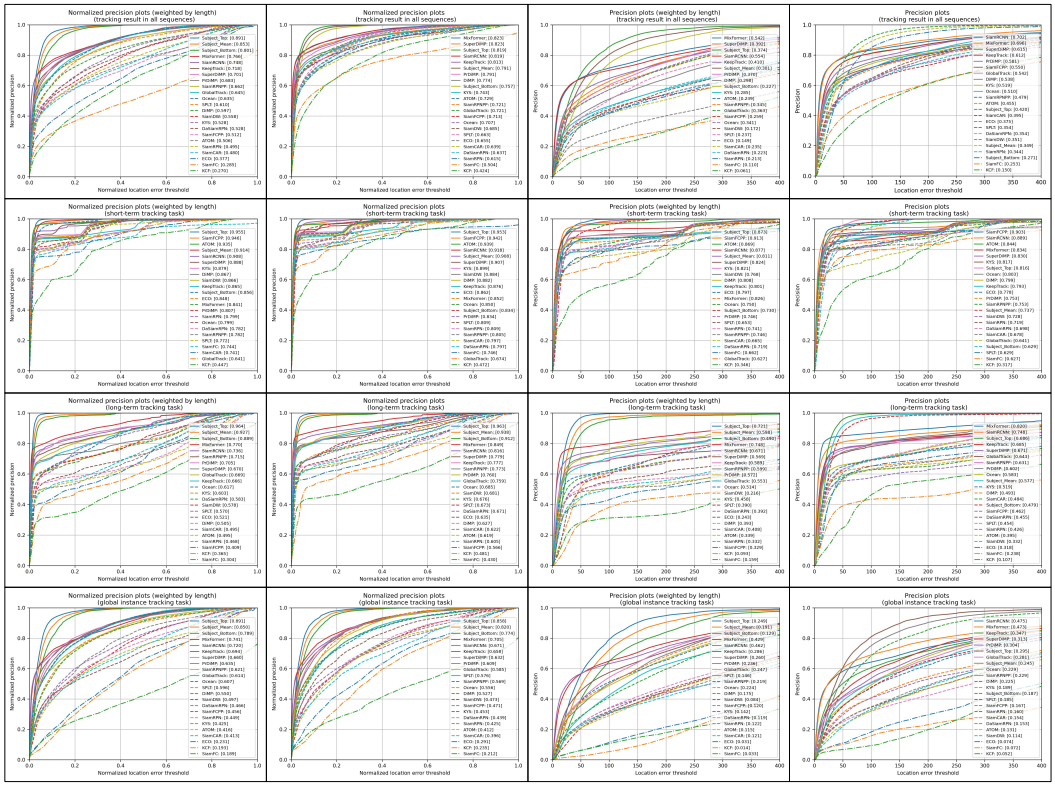


Figure 13: Results for all tasks, short-term tracking task, long-term tracking task, and global instance tracking task based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

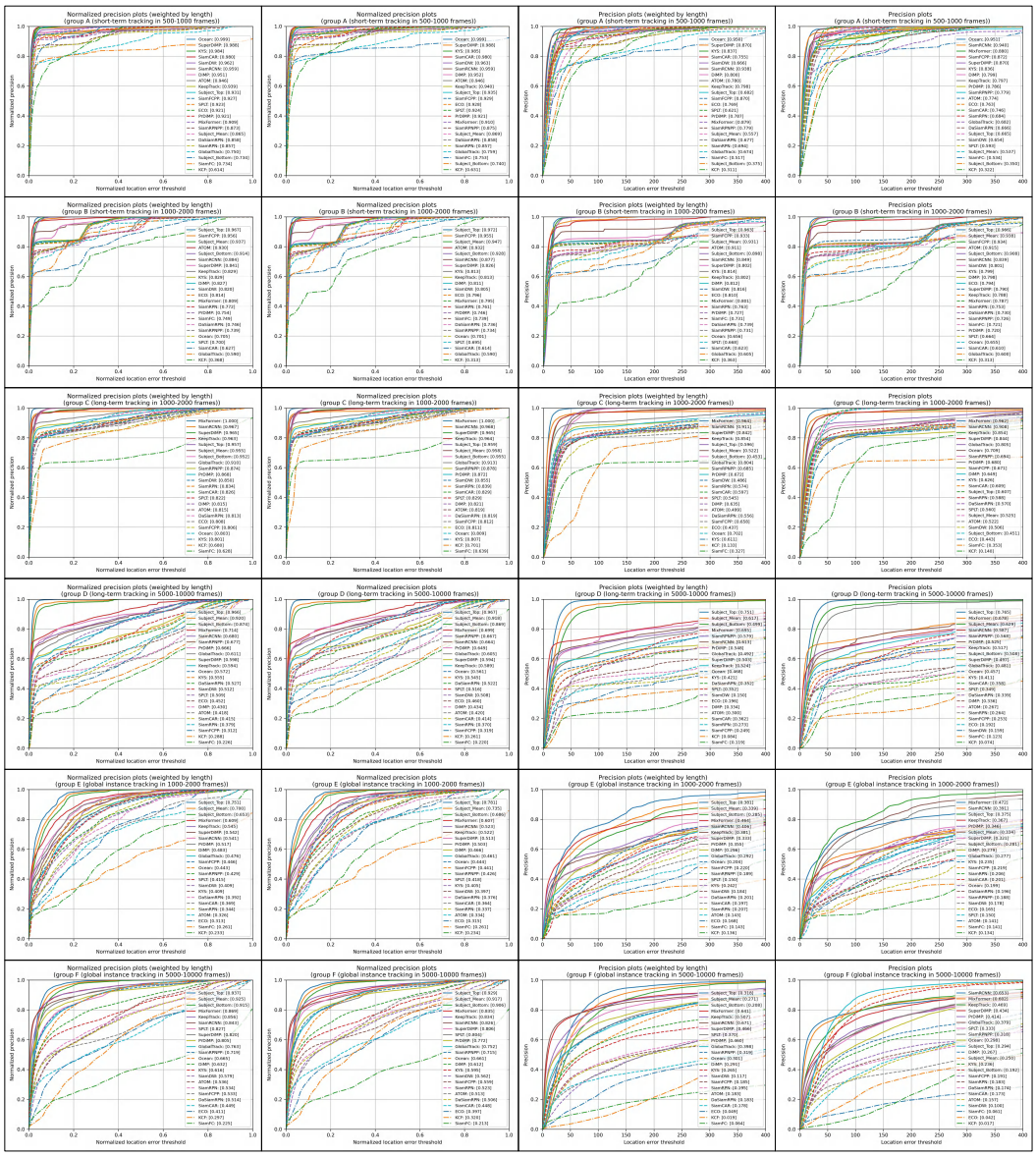


Figure 14: Results for Group A to F based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

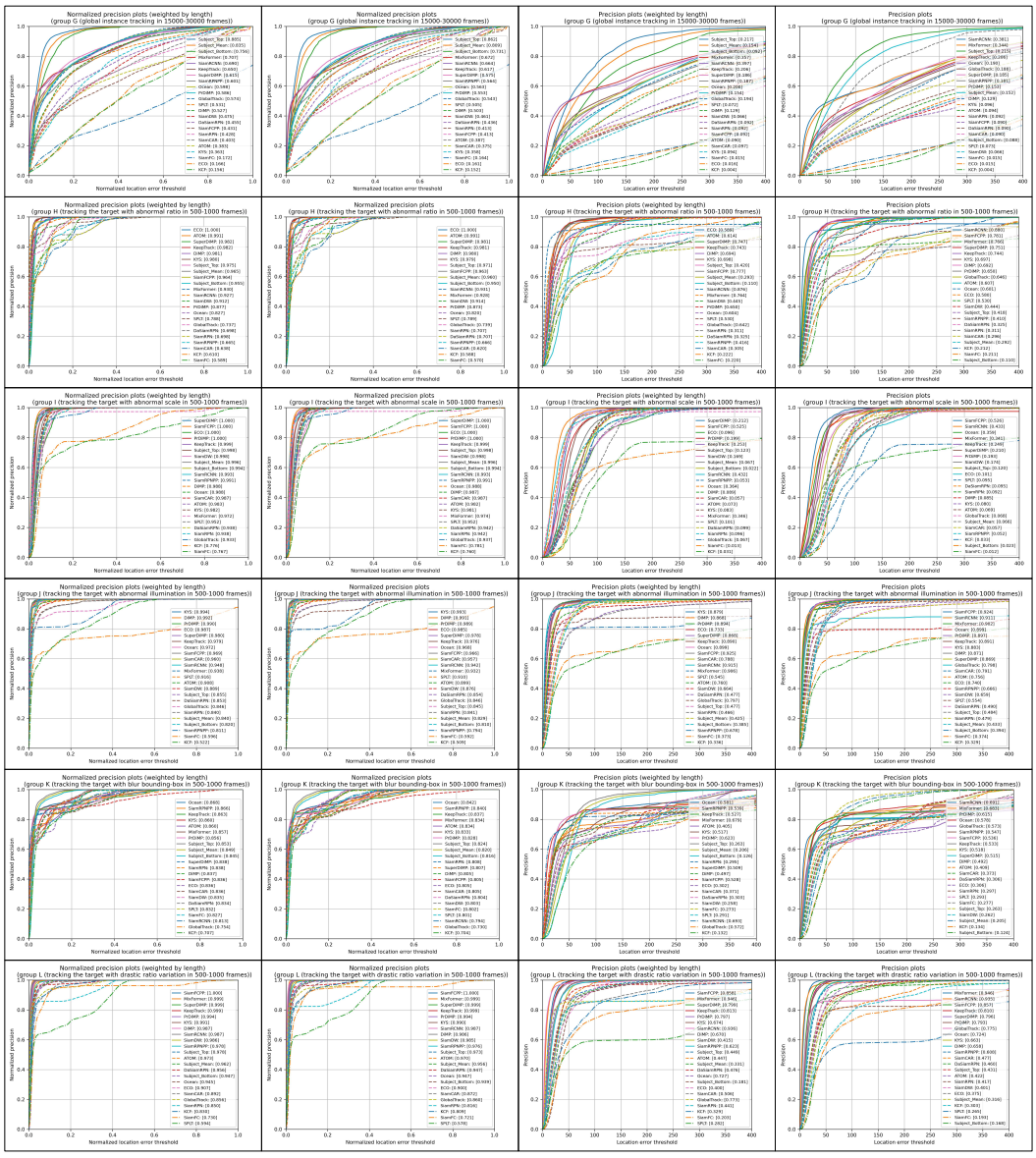


Figure 15: Results for Group G to L based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

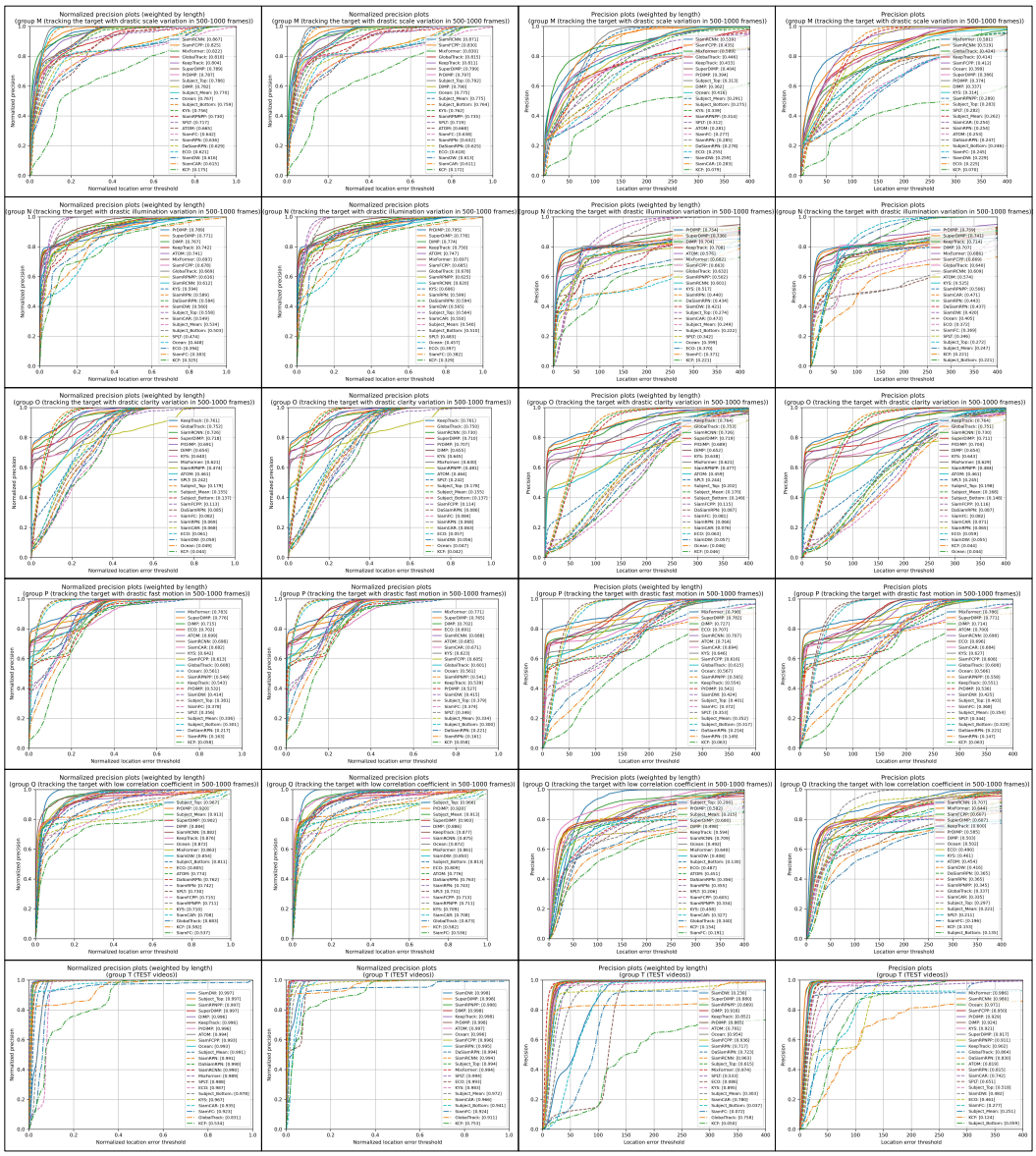


Figure 16: Results for Group M to Q and TEST videos based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

E.5 HEATMAPS

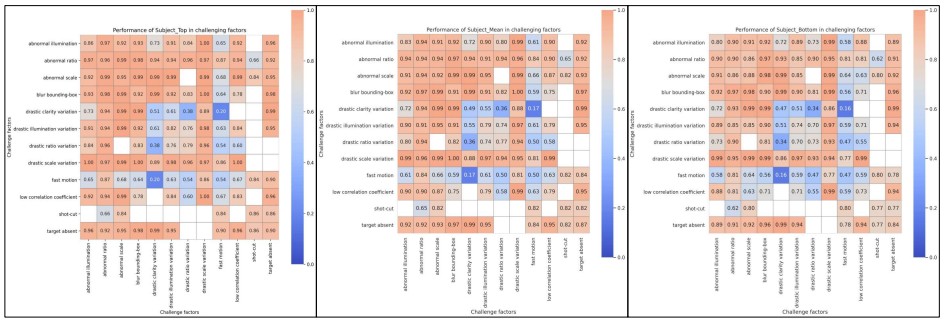


Figure 17: Performance of subjects (the maximum, mean, and minimum scores of human subjects) on multiple challenging factors.

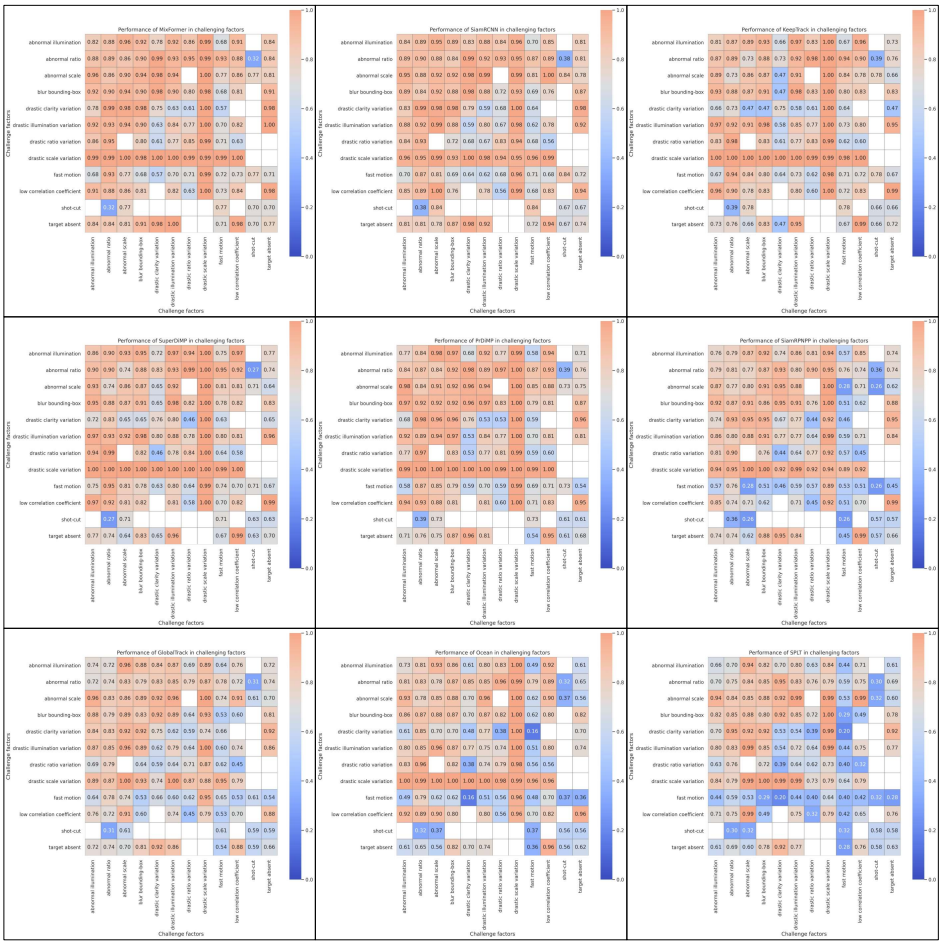


Figure 18: Performance of machines (MixFormer (Cui et al. (2022)), SiamRCNN (Voigtlaender et al. (2020)),KeepTrack (Mayer et al. (2021)), SuperDiMP (Danelljan et al. (2020)), PrDiMP (Danelljan et al. (2020)),SiamRPN++ (Li et al. (2018a)),GlobalTrack (Huang et al. (2019)), Ocean (Zhang & Peng (2020)),SPLT (Yan et al. (2019))) on multiple challenging factors.

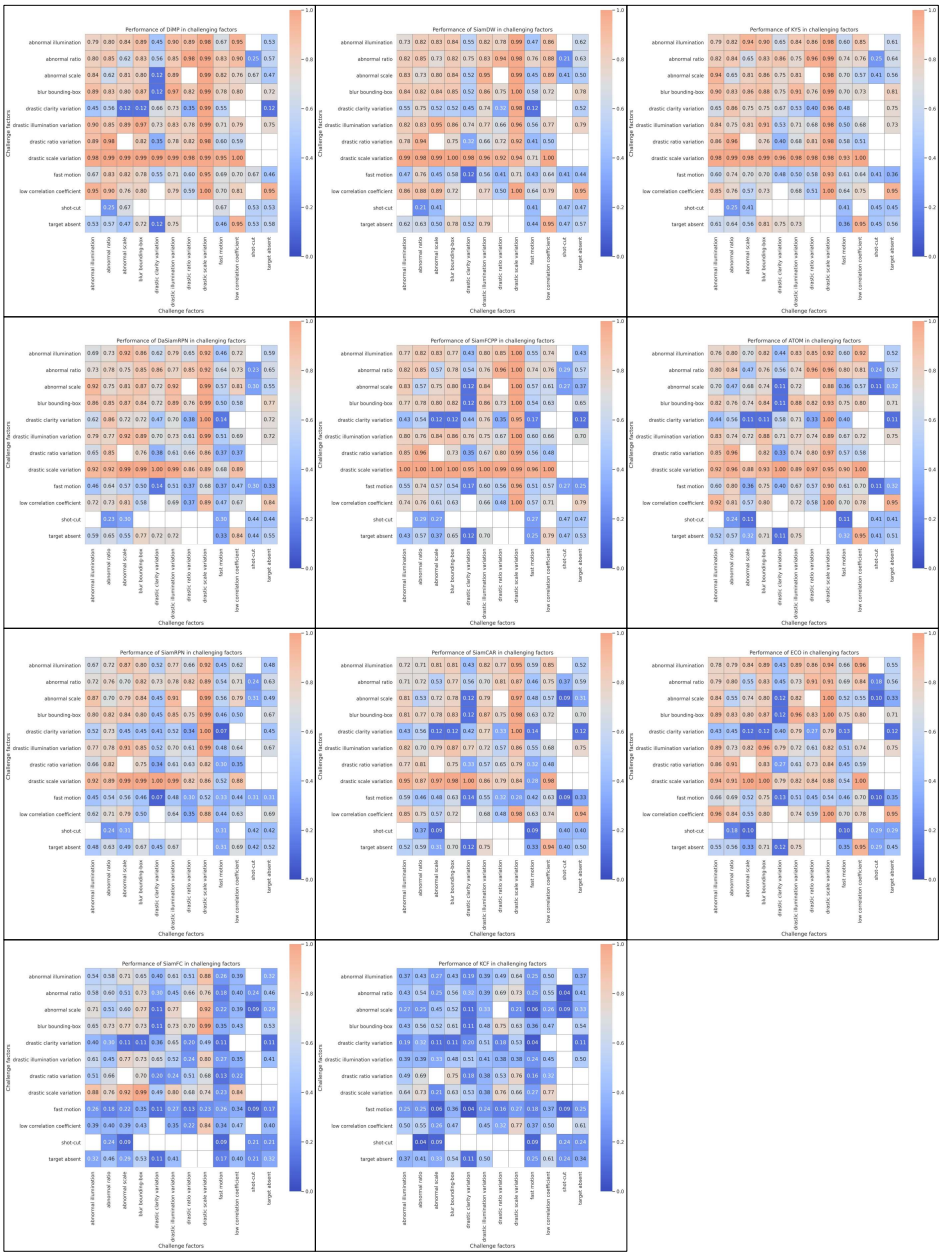


Figure 19: Performance of machines (DiMP (Bhat et al. (2019)), SiamDW (Zhang & Peng (2019)), KYS (Bhat et al. (2020)),DaSiamRPN (Zhu et al. (2018)),SiamFC++ (Xu et al. (2020)),ATOM (Danelljan et al. (2018)),SiamRPN (Li et al. (2018b)),SiamCAR (Guo et al. (2020)), ECO (Danelljan et al. (2017)),SiamFC (Bertinetto et al. (2016)),KCF (Henriques et al. (2015))) on multiple challenging factors.

E.6 ANALYSES OF HUMAN DYNAMIC VISUAL ABILITY

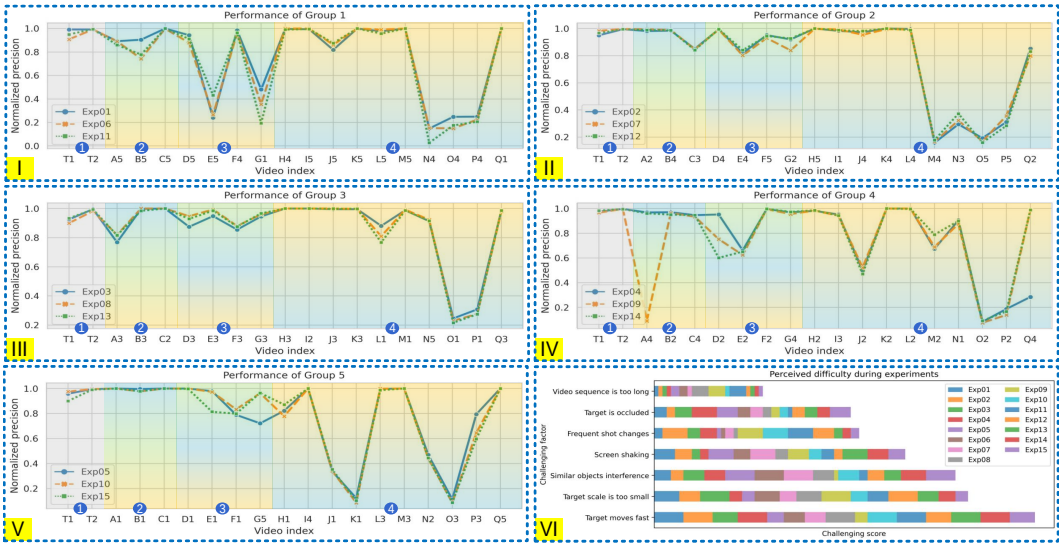


Figure 20: Scores of subject groups and self-perceptions. Subjects with a serial number interval of 5 watches the same videos in different orders. (I-V) demonstrate the performance of subjects. Gray area 1 represents two TEST videos, and colorful areas 2 to 4 represent FORMAL videos. The viewing order of the subjects with blue serial number is 2-3-4, orange is 4-2-3, and green is 3-2-4. (VI) shows their self-perceptions of challenging factors in the questionnaire (a higher score means greater challenge).

We request three subjects to track the same videos in different viewing orders. Figure 20 (I-V) demonstrates that subjects' performance is independent of viewing order – watching long sequences first or last has no significant difference. However, we notice that subjects may also make mistakes, like the performance of Exp09 on A4 is significantly lower than others. A4 is a short video about a basketball game, while Exp09 misadmits the target player and keeps tracking a distractor. It indicates that humans are not absolutely infallible – they may make low-level mistakes due to carelessness.

Figure 20 (VI) shows the self-perceptions of challenging factors in the questionnaire. Most subjects consider that fast-moving targets and small targets are challenging to track, but tracking in a long sequence is not difficult – this is consistent with previous experimental conclusions in this paper.

F PERFORMANCE OF HUMAN-MACHINE COOPERATIONS

F.1 BOXPLOTS

Figure 21: Boxplots enumerate the distribution of N-PRE scores for TEST videos and Group A to Q. The blue area represents the maximum, mean, and minimum scores of human subjects, and the yellow area represents human-machine cooperation.

F.2 PRECISION PLOTS

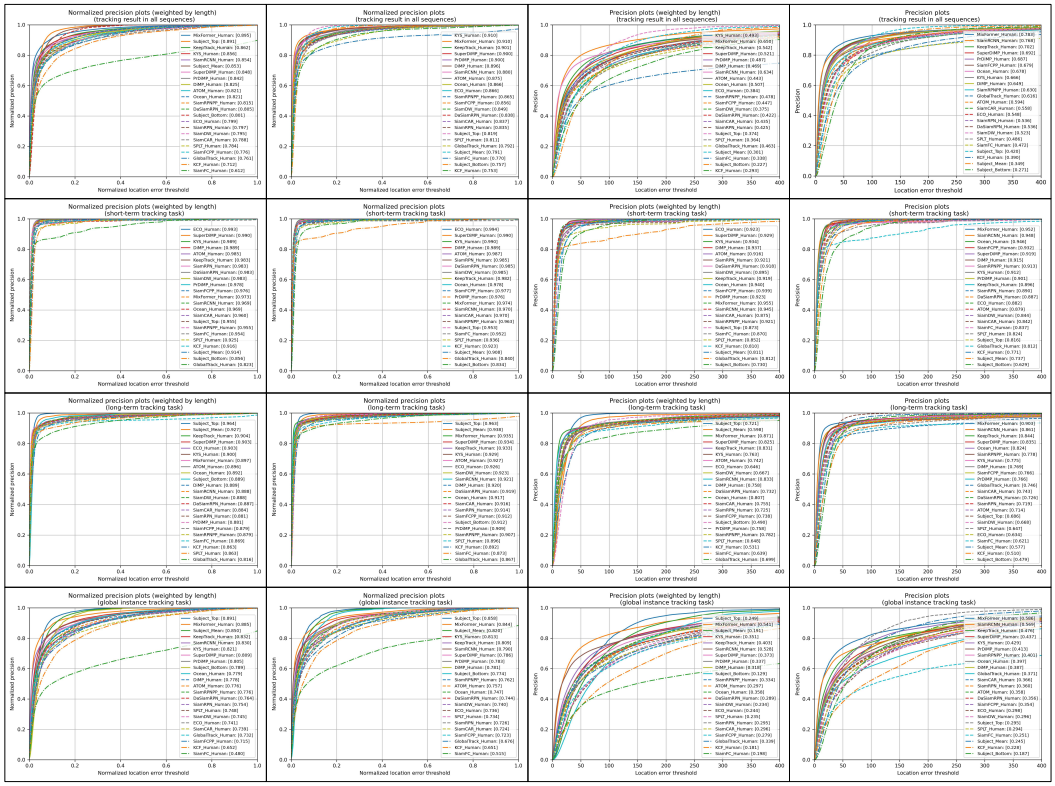


Figure 22: Results for all tasks, short-term tracking task, long-term tracking task, and global instance tracking task based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

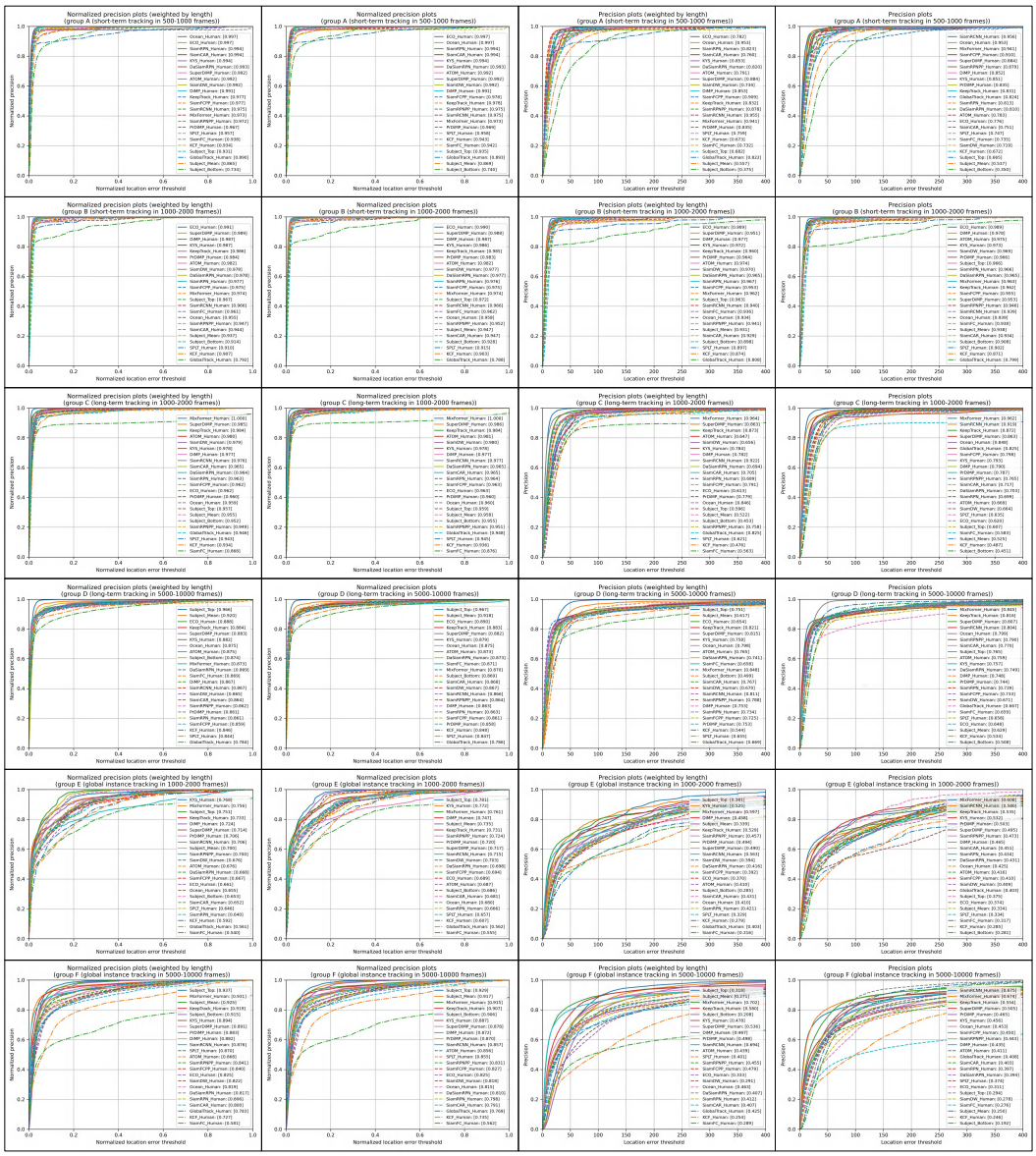


Figure 23: Results for Group A to F based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

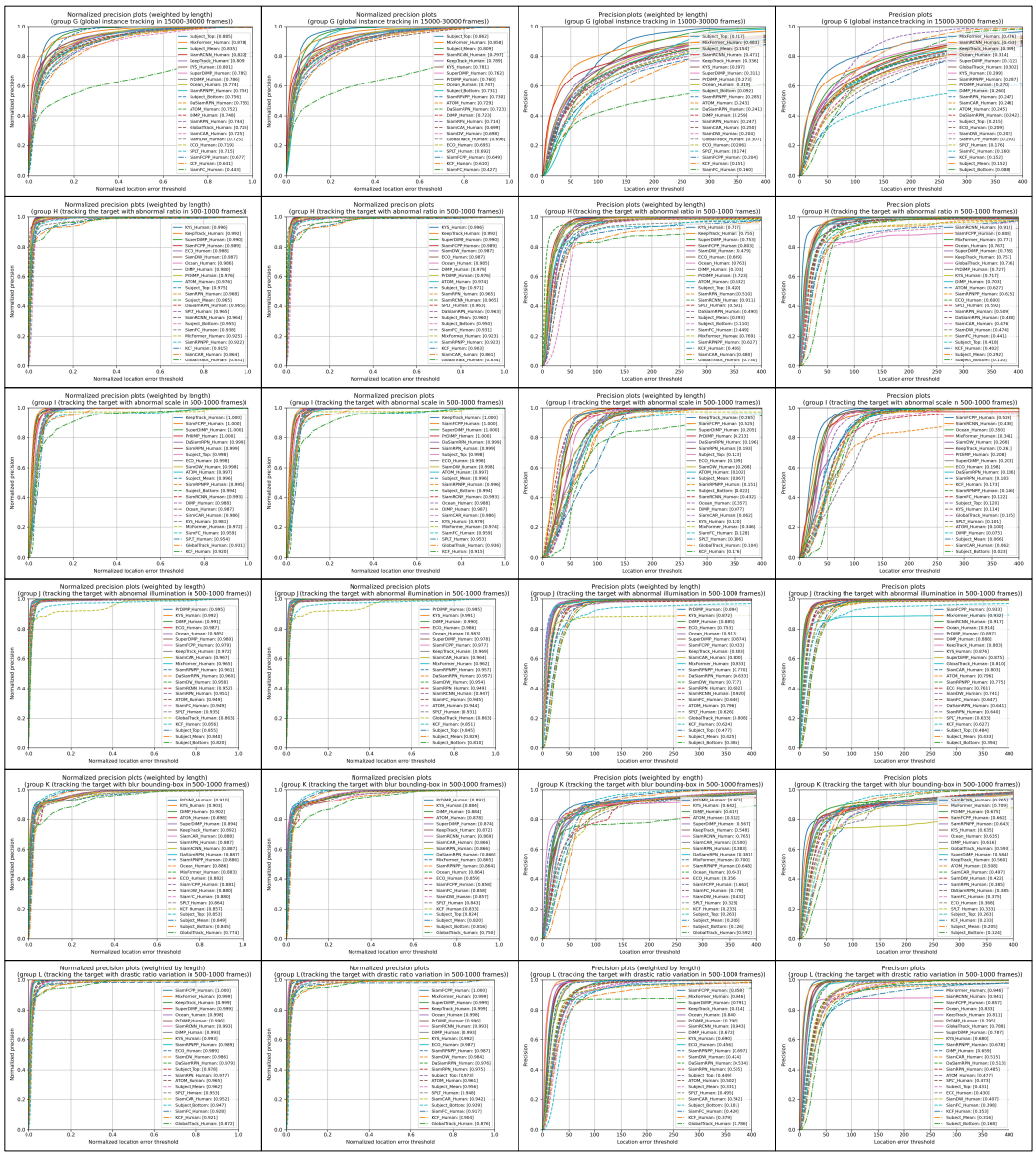


Figure 24: Results for Group G to L based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

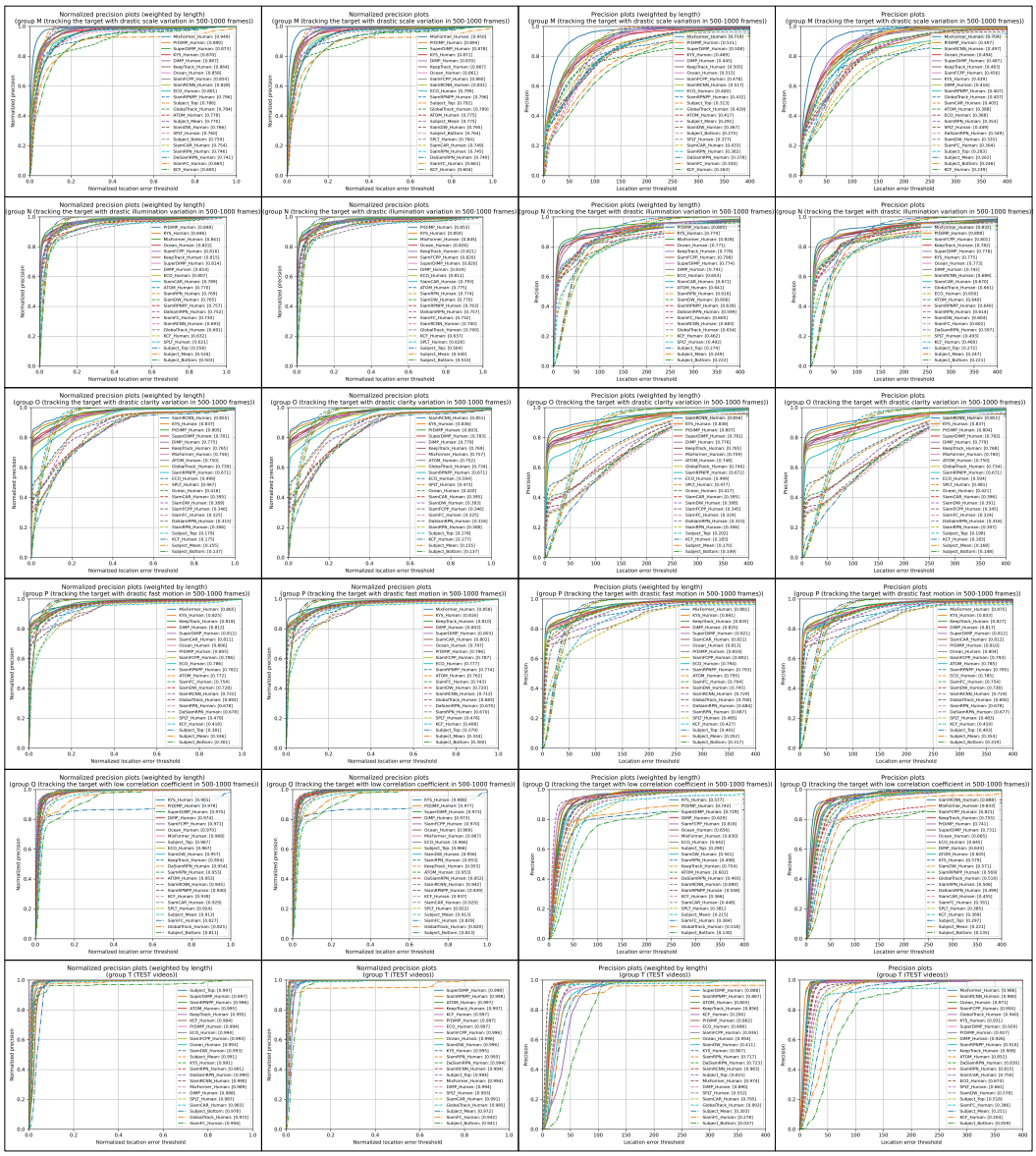


Figure 25: Results for Group M to Q and TEST videos based on various indicators (from left to right: Weighted-N-PRE plots, N-PRE plots, The Weighted-PRE plots and PRE plots).

F.3 HEATMAPS

Figure 26: Performance of human-machine cooperations (MixFormer (Cui et al. (2022)), SiamRCNN (Voigtlaender et al. (2020)),KeepTrack (Mayer et al. (2021)), SuperDiMP (Danelljan et al. (2020)), PrDiMP (Danelljan et al. (2020)),SiamRPN++ (Li et al. (2018a)),GlobalTrack (Huang et al. (2019)), Ocean (Zhang & Peng (2020)),SPLT (Yan et al. (2019)), DiMP (Bhat et al. (2019)), SiamDW (Zhang & Peng (2019)), KYS (Bhat et al. (2020))) on multiple challenging factors.

Figure 27: Performance of human-machine cooperations (DaSiamRPN (Zhu et al. (2018)),SiamFC++ (Xu et al. (2020)),ATOM (Danelljan et al. (2018)),SiamRPN (Li et al. (2018b)),SiamCAR (Guo et al. (2020)), , ECO (Danelljan et al. (2017)), SiamFC (Bertinetto et al. (2016)),KCF (Henriques et al. (2015))) on multiple challenging factors.